MULTIPLES MÉTAMODÈLES POUR L'APPROXIMATION ET L'OPTIMISATION DE FONCTIONS NUMÉRIQUES MULTIVARIABLES

David Ginsbourger

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

N° d'ordre : **519 MA**

THÈSE

présentée par

David GINSBOURGER

pour obtenir le grade de
Docteur de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : Mathématiques Appliquées

MULTIPLES MÉTAMODÈLES POUR L'APPROXIMATION ET L'OPTIMISATION DE FONCTIONS NUMÉRIQUES MULTIVARIABLES

soutenue à Saint-Etienne, le 26 Mars 2009

Membres du jury

Président :	Alain VAUTRIN	Professeur, Ecole des Mines, Saint-Etienne
Rapporteurs :	Raphael HAFTKA	Professeur, University of Florida, Gainesville
	Joachim KUNERT	Professeur, Technische Universität Dortmund
Examinateurs :	Rodolphe LE RICHE	Chargé de Recherches CNRS, Saint-Etienne
	Michel SCHMITT	Professeur, Ecole des Mines, Paris
Directeurs de thèses :	Laurent CARRARO	Professeur, Télécom Saint-Etienne
	Anestis ANTONIADIS	Professeur, Univ. Joseph Fourier, Grenoble

Spécialités doctorales :	**Responsables :**
SCIENCES ET GENIE DES MATERIAUX	J. DRIVER Directeur de recherche Centre SMS
MECANIQUE ET INGENIERIE	A. VAUTRIN Professeur Centre SMS
GENIE DES PROCEDES	G. THOMAS Professeur Centre SPIN
SCIENCES DE LA TERRE	B. GUY Maître de recherche Centre SPIN
SCIENCES ET GENIE DE L ENVIRONNEMENT	J. BOURGOIS Professeur Centre SITE
MATHEMATIQUES APPLIQUEES	E. TOUBOUL Ingénieur Centre G2I
INFORMATIQUE	O. BOISSIER Professeur Centre G2I
IMAGE, VISION, SIGNAL	JC. PINOLI Professeur Centre CIS
GENIE INDUSTRIEL	P. BURLAT Professeur Centre G2I
MICROELECTRONIQUE	Ph. COLLOT Professeur Centre CMP

Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d un doctorat d État ou d une HDR)

AVRIL	Stéphane	MA	Mécanique & Ingénierie	CIS
BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	SITE
BENABEN	Patrick	PR 2	Sciences & Génie des Matériaux	CMP
BERNACHE-ASSOLANT	Didier	PR 0	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 2	Informatique	G2I
BOUCHER	Xavier	MA	Génie Industriel	G2I
BOUDAREL	Marie-Reine	MA	Génie Industriel	DF
BOURGOIS	Jacques	PR 0	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	MR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 0	Génie des Procédés	SPIN
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	ICM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 1	Sciences & Génie de l'Environnement	SITE
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DELAFOSSE	David	PR 1	Sciences & Génie des Matériaux	SMS
DOLGUI	Alexandre	PR 1	Génie Industriel	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	SMS
DRIVER	Julian	DR	Sciences & Génie des Matériaux	SMS
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FORTUNIER	Roland	PR 1	Sciences & Génie des Matériaux	SMS
FRACZKIEWICZ	Anna	MR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	CR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GOEURIOT	Patrice	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUILHOT	Bernard	DR	Génie des Procédés	CIS
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
KLÖCKER	Helmut	MR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LERICHE	Rodolphe	CR	Mécanique et Ingénierie	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
LONDICHE	Henry	MR	Sciences & Génie de l'Environnement	SITE
MOLIMARD	Jérôme	MA	Mécanique et Ingénierie	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	PR1	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 1	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	CR	Sciences & Génie de l'Environnement	SITE
THOMAS	Gérard	PR 0	Génie des Procédés	SPIN
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VAUTRIN	Alain	PR 0	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	MR	Génie des procédés	SPIN
WOLSKI	Krzysztof	CR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 0	Professeur classe exceptionnelle
PR 1	Professeur 1ère catégorie
PR 2	Professeur 2ème catégorie
MA(MDC)	Maître assistant
DR (DR1)	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
ICM	Ingénieur en chef des mines

Centres :

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
SITE	Sciences Information et Technologies pour l Environnement
G2I	Génie Industriel et Informatique
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Table des matières

Remerciements

Pour sa disponibilité scientifique et humaine lors de la direction de ce travail à l'EMSE, ses compétences nombreuses et multiformes, ainsi que la passion contagieuse qu'il voue à la transmission des mathématiques, je remercie chaleureusement Laurent Carraro. Je suis aussi très reconnaissant à Anestis Antoniadis d'avoir participé à mon encadrement de manière efficace et bienveillante lors de nos discussions scientifiques à Grenoble et ailleurs, et de m'être venu amicalement en aide lors d'un épisode délicat de la thèse. Deux autres chercheurs doivent être remerciés ici pour leur rôle dans la supervision de ce travail : il s'agit d'Olivier Roustant et Rodolphe Le Riche. Olivier, merci infiniment pour ta grande disponibilité, ton soutien, et les nombreuses heures passées ensemble à faire des calculs ou à programmer nos packages en R. Un grand merci à Rodolphe pour ses conseils formateurs en écriture scientifique, pour m'avoir sensibilisé avec bonheur à l'optimisation globale, et pour avoir cru en mes travaux du début à la fin.

C'est un immense plaisir de pouvoir remercier ici Raphaël Haftka et Joachim Kunert, avant tout pour la patience et la bonne volonté inébranlable dont ils ont fait preuve en tant que rapporteurs de cette thèse, et bien sûr pour avoir apporté à l'évaluation de ce mémoire leur expertise en optimisation et en statistique. Je souhaite aussi adresser toute ma reconnaissance à Alain Vautrin et Michel Schmitt, qui m'ont fait l'honneur de participer à ce jury de thèse en tant d'examinateurs. Merci à Yann Richet et Patrick Cousinou de l'IRSN pour leur présence scientifique et humaine au fil de la thèse et de la soutenance. Je tiens aussi à souligner le rôle de Philippe Renard, qui n'a pas pu se libérer pour la soutenance, mais qui a indirectement participé à l'aboutissement de cette thèse en ayant parfaitement respecté les phases intenses de rédaction lors de mes premiers mois à Neuchâtel. André Journel a lui aussi joué un rôle positif, par l'excellent accueil dont il m'a gratifié à Stanford, ainsi que par les échanges de mail que nous avons pu avoir depuis lors. Je regrette qu'il n'ait pû se joindre au jury, d'autant plus que cela tient pour beaucoup à un décalage de soutenance auquel je ne suis pas étranger...

Ces trois années de doctorat à l'Ecole des Mines de Saint-Etienne ont été pour moi l'occasion de faire quelques belles recontres. Tout d'abord au sein de l'équipe 3MI, département de mathématiques haut en couleur, et où s'applique à merveille l'adage (*quelque peu adapté*) « un mathématicien (*appliqué*) est une machine à transformer le café (*et les sucreries*) en théorèmes (*et surtout en méthodes et modèles mathématiques pour l'industrie*) ». En particulier (pour revenir sur les sucreries au café), un grand merci à Xavier Bay pour le temps qu'il m'a accordé sur quelques points de mathématiques et

de probabilités (cette fois-ci plutôt fondamentales), ainsi qu'à Delphine Dupuy, Eric Touboul, Céline Helbert, et Anca Badea pour nos nombreuses discussions sur des sujets variés. La gentillesse et le professionnalisme à toute épreuve de Liliane Brouillet et Christine Exbrayat on également contribué au bon déroulement de cette thèse, y compris sur le plan humain. J'ai aussi eu la chance de travailler régulièrement en collaboration avec plusieurs personnes de la Direction de la Formation, au premier plan desquelles Sophie Peillon, Marie-Reine Boudarel, et Bernadette Zold pour ce qui concerne l'équipe management, ainsi que Bertrand Jullien, Michel Cournil, Yves Barbry, et Laurent Perrier-Camby. Merci beaucoup à Sophie et Bernadette pour leur coup de main salutaire lors des préparatifs de la soutenance, ainsi qu'à Axel Momm, qui n'a pas compté son temps pour faire des essais de vidéoconférence dans des conditions d'une telle urgence qu'elles en furent ludiques ! Victor Picheny et Nicolas Durrande ont eux aussi joué un rôle de premier plan dans les préparatifs scientifiques et techniques, ainsi que Matthieu Canaud qui s'est dévoué pour prendre les photos. Merci aussi à celles et ceux qui comme Chris Yukna, Bernard Guy, ou Olga Guschinskaya —je ne peux pas citer tout le monde ici !— m'ont honoré de leur présence à l'exposé et/ou au pot de soutenance. Pour ce qui est du service reprographie, ma gratitude va à Pierre Igier, qui s'est adapté à des contraintes de temps ric-rac pour l'impression de la version finale de ce mémoire. Du côté grenoblois, merci de nouveau à Anestis ainsi qu'à Jacques Istas pour m'avoir accueilli dans leur équipe, à Claudine Meyrieux qui a admirablement géré les aspects pratiques, et à Claire Tauvel et Robin Girard qui m'ont fait une place dans leur bureau.

La période doctorale m'a par ailleurs permis de rencontrer des chercheurs, scientifiques, et ingénieurs venus d'horizons divers. Tout d'abord par le biais du consortium DICE et du projet OMD, dont je salue les membres avec qui j'ai pu avoir de nombreuses discussions intéressantes, mais aussi au travers du Groupement de Recherche Mascot Num ou encore du réseau ENBIS. Outre Philippe et André, mon séjour doctoral aux USA m'a aussi permis de faire la connaissance de Lin-Ying Hu et Alexandre Boucher, dont la vivacité et la profondeur intellectuelles lors de nos séances de *brainstorming* m'ont durablement marqué. Dans un autre registre, je tiens à remercier chaque étudiant(e) que j'ai pu croiser sur mon chemin au cours de cette période à Saint-Etienne et ailleurs, en tant que collègue ou élève. Enseigner durant ces années a été pour moi non seulement une grande source de satisfaction, mais aussi un moyen de progression scientifique tout à fait saisissant. Cette expérience pédagogique s'est prolongée avec bonheur à Saint-Etienne, Paris, et avant tout à l'Institut de Mathématiques de l'Université de Neuchâtel depuis Septembre 2008. Merci à Bruno Colbois et à Maria Paula Gomez Aparicio qui ont toléré mon emploi du temps infernal lorsque j'étais leur assistant, à l'équipe des profs

pour m'avoir fait confiance et chargé d'un cours d'école doctorale pour le printemps 2009, ainsi qu'a Paul Jolissaint de m'avoir invité au séminaire « Mathématiques et Société ». Je ne remercierai jamais assez mes camarades assistants, post-docs (ou assimilé), et maître-assistantes pour m'avoir accepté dans leur équipe, grâce à laquelle l'ambiance de travail fût au global vraiment excellente tout au long de cette année, pour moi de transition. Dans le désordre : merci à Béa, Mumu, So-Young, Maria, Agnès, Olivier, Kola, Greg, Mat, Dany, Cédric, Lionel. Merci encore aux trois irréductibles venus en visite à Saint-Etienne pour la soutenance. Un clin d'oeil aux collègues de l'équipe d'hydro stochastique, dont j'espère avoir le temps de mieux faire connaissance dans un futur proche : Greg, Damian, Julien, Alessandro, Andrea, Andres. Ah, j'oubliais : il reste sûrement pas mal de coquilles et autres fautes de frappe dans la thèse, mais il y en auraît certainement eu encore beaucoup plus si Laurent Carraro, Olivier Roustant, Olivier Isely, Victor Picheny, Yacine Barhoumi-Andréani, et Cédric Boutillier n'avaient pas pris le temps de relire un chapitre ou plus. Merci aussi à Ben Smarslok, Yann Richet, et Alain Valette pour leur relecture d'un ou plusieurs articles relatifs à cette thèse.

Je souhaite maintenant faire des remerciements un peu spéciaux aux nombreuses personnes (potes, proches, et famille) qui m'ont hébergé avant ou pendant la rédaction, lorsque j'étais dans une de mes phases de balade (souvent assez studieuses !) : Clarisse, Robin, Paola, Nino, et Aimé qui m'ont très souvent accueilli aux alentours de Grenoble, en particulier au début de la thèse ; Séverine, Olivier, et Lizoé, à Valence, Nantes, et Poissy ; Clarence, à Brest ; Guillaume, à Saint-Etienne ; Natacha et Victor (+Picheny family), à Saint-Etienne et Orvault ; Julie et Olivier, à Saint-Etienne ; Céline et Moregg', à Grenoble ; Maité, à Grenoble ; Marine et Eric à Grenoble ; Rémi et Jean-Luc, à Grenoble ; Gaëtane et ses colocs, à Londres ; Fabienne, à Grenoble ; Arnaud, à Berlin ; Odile et Jérôme, à Paris et Sausset-les-Pins ; Ma maman Patricia, à Mooréa ; Handan et Michaël, à Vienne ; Christine, Claude, Thomas, Samuel, et Simon aux alentours de Toulouse ; Philippe, à Sarrebrück ; Bernard, à Brême ; Ma grand-mère Janine, à Grenoble ; Ma grand-mère Huguette, à Paris et Ovronnaz ; Monique, Antoine, et mon papa Francis, à Paris et Bréchamps ; Damian, à Neuchâtel ; Lulu, Blaise, et François, à Neuchâtel ; et last but not least, la plus belle, Başak, Bursa (Turquie), avec qui j'ai en bonne partie à distance partagé ma vie, et qui a gagné au fil de ces années une place dans mon coeur.

Chapitre 1

Introduction

1.1 Contexte, besoins, et apports de la thèse

1.1.1 Contexte applicatif et scientifique

Montée en puissance de la simulation numérique

Depuis l'ENIAC (*Electronic Numerical Integrator And Calculator*, développé pendant la seconde guerre mondiale avec le concours du mathematicien John Von Neumann), ses 30 tonnes, et ses 5000 opérations arithmétiques par seconde rendues possibles par une fréquence d'horloge de 100kHz, la simulation numérique s'est développée à un rythme effréné (exponentiel, selon la *loi de Moore*) jusqu'à permettre l'avènement des capacités de calcul actuelles. Ces dernières permettent aujourd'hui —via des techniques d'analyse numérique telles que les éléments finis, ou les méthodes de Monte-Carlo— de simuler avec précision une large gamme de phénomènes physiques, en autorisant à la fois la prise en compte d'un nombre croissant de paramètres, et toujours plus d'adéquation aux phénomènes originaux dans les résultats obtenus. Les simulateurs numériques apparaîssent ainsi comme le moyen de mettre en oeuvre des modèles mathématiques de la réalité, et sont utilisés comme de très utiles compléments à l'expérimentation physique (e.g. crash-tests automobiles, prospection de ressources naturelles), voire même dans certains cas comme des substituts (sûreté nucléaire, prédictions météorologiques, etc).

Un des problèmes majeurs qui se posent aujourd'hui aux utilisateurs de simulateurs numériques de pointe est le temps de calcul. Les ingénieurs et chercheurs prenant part à des projets industriels, le plus souvent avec de fortes contraintes temporelles et de budget, se retrouvent à devoir fournir des solutions technologiques sur base de simulation numérique en limitant drastiquement le nombre d'évaluations du simulateur numérique.

Ce constat est particulièrement frappant dans le cadre des avant-projets de conception automobile, dans lesquelles un petit nombre de semaines de calcul (typiquement de l'ordre de six) sont allouées pour les simulations de crash-tests, alors même qu'une évaluation du simulateur commercial utilisé « coûte » une douzaine d'heures de calcul. Il est ainsi nécessaire (et cela devient crucial lorsque le nombre variables est grand, phénomène souvent mentionné suite aux travaux de Richard Bellman sous le nom de *malédiction de la dimension*) d'adopter des stratégies d'évaluation mûrement réfléchies, ce qui nécessite un recours supplémentaire à l'analyse déterministe et probabiliste via le développement d'outils spécifiques tels que les *métamodèles*. Ces *modèles de modèles*, utilisés comme des approximations abstraites des modèles mathématiques qu'incarnent les simulateurs, jouent un rôle d'aide à la décision. Ils permettent, comme nous allons le voir au chapitre 4, de produire des stratégies d'évaluation automatiques, servant des objectifs tels que l'optimisation d'une réponse particulière — par exemple la résolution du problème de minimisation globale de la masse d'un véhicule, sous des contraintes de sécurité, de coûts de fabrication et/ou fonctionnement, et de pollution.

Modélisation et optimisation des simulateurs numériques

Suivant les considérations exposées dans le dernier paragraphe, cette thèse s'inscrit dans la thématique de planification d'expériences numériques. Elle porte plus précisément sur l'optimisation de simulateurs numériques *coûteux à évaluer* par des stratégies de planification basées sur des représentations simplifiées du simulateur (*metamodèles*, dits encore *surrogates*). Quitte à considérer des combinaisons des composantes dans le cas de simulateurs à réponses vectorielles, nous ferons dans la suite sans restriction de généralité l'hypothèse que le simulateur numérique étudié est à réponse scalaire, et qu'il constitue la fonction objectif y d'un problème d'approximation et/ou d'optimisation posé.

L'utilisation de métamodèles revient à remplacer une telle fonction y par une approximation de forme pré-établie, paramétrique ou semi-paramétrique. Le point de départ est de procéder à un petit nombre de simulations (évaluation de y en un *plan d'expériences initial*, Cf. [Fra08] pour un passage en revue des plans *ad hoc* pour la phase d'apprentissage initial). On dispose alors généralement de deux formes d'informations : les réponses — ou observations— recueillies au plan d'expériences initial, et d'éventuelles informations acquises *a priori* traduisant une expertise "métier" (fonctions de base, tendances, ordre de grandeur des longueurs de corrélation, etc). Une fois choisi un metamodèle parmi les familles existant (polynômes, splines, modèles additifs, Krigeage, réseaux de neurones ... Cf. chapitres 2 et 3), on estime les paramètres du metamodèle. On dispose alors

d'une première représentation simplifiée du simulateur, que l'on pourra faire évoluer en fonction des informations apportées par de nouvelles évaluations.

Parmi les choix qui se posent en amont de cette démarche, la sélection d'un type de metamodèle est de première importance puisqu'elle va fortement conditionner notre représentation du simulateur. Cette question est pour autant souvent shuntée dans la littérature de l'optimisation de simulateurs numériques, où il est d'usage de raisonner à metamodèle fixé. L'algorithme EGO [JSW98], faisant aujourd'hui référence en optimisation séquentielle de simulateurs, en est un bon exemple : le métamodèle est donné par un Krigeage Ordinaire avec covariance exponentielle généralisée, dont les paramètres de moyenne et de covariance sont obtenus en maximisant une fonction de vraisemblance gaussienne (calculée sur la base du plan d'expériences courant). Or, rien ne garantit que cela soit un metamodèle capable de remplir aux mieux les objectifs poursuivis.

Etant donné qu'il est difficile de savoir *a priori* quel sera le type de metamodèle capable de guider au mieux un algorithme d'optimisation, une des motivations de ce travail est d'examiner comment une construction ad hoc de la structure du metamodèle, voire la prise en compte de plusieurs metamodèles, peuvent améliorer les méthodes d'approximation et les stratégies d'optimisation globale actuellement employées. Cela soulève à la fois des questions mathématiques et statistiques de sélection de modèle (Chapitres 2, 3 : quelles familles de métamodèles considérer ? Chapitres 5, 6 : comment estimer les termes de covariance et/ou de tendance d'un métamodèle de Krigeage, et selon quels critères les évaluer ? Chapitre 7 : comment prendre en compte certaines formes d'instationnarité dans la covariance de Krigeage que sont les symétries et la présence de bruits d'observation hétérogènes ?), de combinaison de modèles (Chapitre 8 : une fois un ensemble de metamodèles choisis, comment agrège-t-on les pseudo-informations qu'ils nous apportent ?), et de définition de critères décisionnels pour guider les évaluations de y au sein d'algorithmes d'optimisation (Chapitre 9 : comment paralléliser EGO ou des procédures similaires d'exploration sur base de Krigeage).

La thèse vise à apporter des éléments de réponse à ces questions en s'appuyant à la fois sur des considérations d'ordre théorique, sur des expérimentations basées sur des cas abstraits (fonctions analytiques, simulations de processus stochastiques), et sur des applications concrètes (deux cas d'étude industriels, Cf. chapitres 6 et 7). Un effort particulier a été entrepris pour exhiber et tirer bénéfice des parallèles existant entre la littérature connue par l'auteur dans le domaine de l'approximation et de l'optimisation sur base de Krigeage, et les ré-interprétations « modernes » de ce métamodèle en termes d'interpolation dans les espaces de Hilbert à noyau reproduisant (RKHS, Cf. Sec. 2.3

et [Wah90, Ver07, Vaz05, Lar08]), et surtout en termes d'apprentissage statistique par processus gaussiens et extensions bayésiennes associées (Cf. Sec. 3.3 et [RW06]).

1.1.2 Résumé des apports de la thèse

Le présent document se veut avant tout une thèse de généraliste, c'est à dire d'un candidat qui disposait au début de son doctorat d'une certaine culture scientifique et mathématique, mais qui n'était spécialiste d'aucune des disciplines abordées durant le doctorat : approximation, optimisation, probabilités, etc. Un long chemin a donc été parcouru en termes de compréhension et de synthèse des notions rencontrées dans les problèmes posés, et une des vocations évidentes du manuscript de thèse est de retranscrire ce cheminement, dans l'espoir qu'il pourra bénéficier un jour à d'autres chercheurs en formation ou non-spécialistes désireux d'approfondir un des points traités.

Cette remarque vaut en particulier pour les notions d'estimation statistique présentées au chapitre 5, pour lesquelles l'auteur est parti de zéro (dans sa recherche comme dans la façon dont sont présentées les notions) et a nettement privilégié le point de vue pédagogique, au risque de faire quelques rappels élémentaires paraissant superflus au lecteurs initiés. La démarche de synthèse entreprise a par ailleurs occasionné quelques remarques vraisemblablement assez originales, comme la seconde interprétation des équations de régression avec le théorème du représentant (Issue d'une séance de travail avec B. Gauthier, Cf. fin du chapitre 2), ou encore l'explicitation du noyau de covariance conditionnelle associé au Krigeage Universel dans son interprétation bayésienne (Cf. fin du chapitre 3).

Pour venir à des apports méthodologiques plus concrets, la thèse réunit quelques résultats orginaux de nature expérimentale et/ou théorique sur les sujets suivants :

- Sur la loi de l'EMV des paramètres de covariance du Krigeage (Chap. 5) : constat expérimental que l'approximation de Fisher est abusive lorsque le nombre d'observation est très petit, mais peut éventuellement servir d'aide à la décision dans la conception de plans d'expériences.
- Sur l'utilisation de tendances non-linéaires dans l'approximation par Krigeage (Chap. 6) : proposition sur un cas d'étude d'une méthode heuristique reposant sur le découpage de l'ensemble d'apprentissage, pour estimer à la fois la tendance et les paramètres de covariance lorsque le problème classique de circularité ne peut être résolu par concentration de la vraisemblance comme en Krigeage Universel.

- Sur la prise en compte de symétries dans le Krigeage (Chap. 7) : caractérisation en termes de noyaux de covariance des processus aléatoires centrés et de carré intégrable (non nécessairement gaussiens) dont les trajectoires sont invariantes sous l'action d'un groupe fini, à une modification près. Proposition de différentes méthodes pour prendre en compte des symétries dans le Krigeage. Présentation d'un cas d'étude industriel ayant inspiré ces travaux, et comparaisons des différentes méthodes proposées sur un cas test analytique.
- Sur la prise en compte d'un bruit de simulation stochastique réglable dans le Krigeage (Chap. 7) : proposition d'un modèle statistique de processus gaussien avec observations entâchées d'un bruit hétéroscédastique pour les simulateurs à fidélité réglable, et dérivation des équations de Krigeage Ordinaire associées.
- Sur les mélanges de métamodèles possédant différentes structures pour l'approximation et l'optimisation (Chap. 8) : proposition d'un modèle de mélange discret de noyaux de covariance, et d'une procédure de pondération adaptative pour la gestion de multiples métamodèles de Krigeage sous un tel mélange au sein de procédures d'exploration séquentielle. Dérivation des équations de mélanges discrets de Krigeage et de critères tels que l'amélioration espérée sous un mélange de Krigeages, discussion sur les rapports avec la modélisation bayésienne, et illustration de l'algorithme EGO avec mélange adaptatif de deux noyaux.
- Sur la parallélisation de l'algorithme EGO et du critère d'amélioration espérée (Chap. 9) : calcul analytique de l'amélioration espérée à deux points, et explicitation d'une méthode de calcul Monte-Carlo pour le cas où il y a plus de deux points. Proposition, avec une justification mathématique, de deux stratégies heuristiques permettant d'approcher la maximisation (difficile) du critère d'amélioration espérée à q points. Illustrations et comparaisons sur des réalisations de processus aléatoires gaussiens et sur deux exemples analytiques de dimension 2 et 6.

1.2 Formalisme et notations employées

1.2.1 Fonctions numériques déterministes multivariables

On étudiera le plus souvent dans cette thèse des simulateurs numériques de phénomènes déterministes. Il est d'usage de modéliser un tel simulateur par une fonction y :

$$y : \mathbf{x} \in D \subset \mathbb{R}^d \rightarrow y(\mathbf{x}) \in \mathbb{R} \tag{1.1}$$

prenant en argument un vecteur $\mathbf{x}$ de nombres réels (d représente le nombre de paramètres -ou *entrées*- du simulateur), et retournant un scalaire $y(\mathbf{x})$. Les simulateurs utilisés dans l'industrie retournent bien souvent une réponse vectorielle. Comme précisé plus haut, nous supposons ici que le travail de transformation d'une telle réponse en fonction objectif a déjà été fait, et y représente alors cette dernière. Cette modélisation fonctionnelle est souvent qualifiée de modélisation "boîte noire". Notons que dans certains cas, la fonction déterministe y ne peut pas être évaluée directement mais plutôt via un calcul de type Monte-Carlo.

1.2.2 Plans d'expériences et métamodèlisation

La construction d'un métamodèle passe par une phase initiale d'expérimentation (ou *apprentissage*). On évalue le simulateur pour un certain nombre n de valeurs de $\mathbf{x}$ (le *plan d'expériences* ou *ensemble d'apprentissage* initial), et on observe le vecteur des n réponses associées :

$$\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\} \tag{1.2}$$

$$\mathbf{Y} = \{y_1, ..., y_n\} = \{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\} \tag{1.3}$$

où chaque $\mathbf{x}^i = (x_1^i, ..., x_d^i)$ est une instance du vecteur de paramètres. Dans le cas où la fonction déterministe y est évaluée via un calcul de type Monte-Carlo, le vecteur des observations est du type :

$$\mathbf{Y}_B = \{y(\mathbf{x}^1) + \epsilon_1, ..., y(\mathbf{x}^n) + \epsilon_n\} \tag{1.4}$$

où les ϵ_i sont des réalisations de variables aléatoires indépendantes identiquement distribuées, généralement supposées être des variables gaussiennes centrées, et dont la variance dépend du nombre de tirages Monte-Carlo (dans le cas de simulateurs stochastiques tels que présentés dans la première section du chapitre 7). Nous porterons une attention particulière au moyen de construire un métamodèle de Krigeage adapté à ce

cas de figure dans la section 3 du chapitre 7.

La métamodélisation consiste alors à rechercher une approximation de y, sur la base des données $(\mathbf{X}, \mathbf{Y})$ (ou $(\mathbf{X}, \mathbf{Y}_B)$) et éventuellement de quelques informations *a priori* sur la nature de la réponse du simulateur. Même si cette thèse n'aborde que très peu les passionnantes et difficiles questions de sélection du plan d'expériences initial en fonction du métamodèle employé (on renvoit une nouvelle fois à [Fra08] pour plus de détails sur ce sujet), les algorithmes d'optimisation étudiés dans les chapitres 4, 8, et 9 peuvent être vus commes de procédures séquentielles adaptatives de construction de plan d'expériences. La problématique de parallélisation d'EGO à laquelle est consacrée le chapitre 9 peut d'ailleurs être revisitée en termes de sélection de petits plans d'expériences additionnels (à q points) au sein d'une procédure itérative d'optimisation. Par ailleurs, quelques ouvertures sont suggérées dans le chapitre 5 pour ce qui est de la prise en compte de l'information sur les paramètres de modèle inconnus au sein même du protocole de choix du plan d'expériences. La section ci-après donne quelques premières précisions sur la distinction entre métamodèles déterministes et probabilistes.

1.2.3 Métamodèles déterministes *versus* probabilistes

L'idée que l'on peut construire une approximation réaliste de la fonction y en se basant **uniquement** sur les observations au plan d'expériences ne fait pas sens. Il se cache en effet toujours quelques hypothèses mathématiques plus ou moins fortes derrière la démarche de métamodélisation. Nous proposons dans la première section du chapitre suivant (Sec. 2.1) un passage en revue synthétique de quelques métamodèles couramment recontrés dans les travaux d'analyse et optimisation de simulateurs numériques déterministes. Un effort particulier sera fait pour préciser les hypothèses qui sous-tendent les différentes familles de métamodèles.

On distinguera ensuite les métamodèles déterministes (délivrant une fonction candidate pour représenter y) des métamodèles probabilistes (délivrant des prédictions de y hors du plan d'expériences, avec une quantification des incertitudes associées). La section 2.2 ainsi que le chapitre 3 sont consacrés à l'approximation basée sur ces deux types de métamodèles, et reposent donc fortement sur ce qui suit.

Même si la fonction y étudiée est déterministe par nature, il est possible de traduire l'inconnaissance que l'on a de y hors du plan d'expériences en utilisant des outils probabilistes. De manière générale, chaque $y(\mathbf{x})$ ($\mathbf{x} \in D$) est vue en modélisation probabiliste

comme *réalisation* d'une *variable aléatoire* $Y(\mathbf{x})$. L'entreprise des métamodèles probabilistes se résume alors à prédire le comportement de chaque variable aléatoire $Y(\mathbf{x})$ (sa moyenne, sa dispersion ... voire sa loi, si possible) en se basant sur l'information disponible (i.e. le fait que $Y(\mathbf{x}) = y(\mathbf{x})$ pour $\mathbf{x} \in \mathbf{X}$).

Autrement dit, les métamodèles probabilistes fournissent non seulement des prédictions de ce que vaut y hors du plan d'expériences, mais aussi une quantification des *incertitudes* associées à ces prédictions. Nous présentons aux chapitres 2 et 3 la régression linéaires et le Krigeage comme deux manières distinctes (mais voisines, Cf. [RW06] chap. 2 pour comprendre leur profonde interconnection) de tirer parti des informations disponibles pour prédire au mieux les valeurs des $\{Y(\mathbf{x}), \mathbf{x} \notin \mathbf{X}\}$.

Le chapitre 4, portant sur l'utilisation de ces deux grandes familles de métamodèles pour l'optimisation, montre en quoi l'approche probabiliste permet de dépasser certaines limites des stratégies sur base de métamodèles déterministes. En revanche, les chapitres 5 à 9, dédiés à des approches probabilistes d'approximation et d'optimisation basées sur le Krigeage, font parfois apparaître les métamodèles déterministes dans le rôle de tendances structurales.

Première partie

De l'approximation à l'optimisation : un état de l'art

Chapitre 2

Quelques techniques classiques d'approximation fonctionnelle

Ce chapitre a pour vocation de présenter un échantillon choisi de méthodes classiques d'approximation. Sans bien sûr prétendre à l'exhaustivité (l'approximation étant un monde en soi), on s'est attaché ici à faire ressortir certains principes fondamentaux qui apparaissent dans la plupart des approches connues par l'auteur. On a pris soin de présenter séparément le cadre déterministe et le cadre probabiliste, même si les outils abordés de part et d'autre sont bien souvent similaires comme nous le verrons. La troisième section, consacrée à l'approximation dans les espaces de Hilbert à noyau reproduisant, ne constitue pas une alternative aux deux premières sections mais plutôt un complément théorique permettant de revisiter plusieurs méthodes de manière générique. Les passages les plus techniques peuvent naturellement être ignorés en première lecture.

2.1 Approximation dans un cadre déterministe

2.1.1 Généralités

L'objet central d'étude est ici une fonction numérique

$$y : x \in D \longmapsto y(x) \in \mathbb{R}$$

où D est une partie[1] de $\mathbb{R}^d$. On entend ici par approximation déterministe le fait de remplacer la fonction objectif y appartenant à $\mathcal{E} \subset \mathbb{R}^D$ (un sous-espace de l'espace des applications de D dans $\mathbb{R}$), par une autre fonction $m : D \longmapsto \mathbb{R}$, cette dernière étant

[1] Souvent supposée compacte et connexe, mais qui ne l'est pas toujours dans les applications.

choisie de manière à « ressembler » le plus possible à y, ce dont le sens mathématique est discuté au fil du chapitre. Remarquons d'emblée que cette tentative de définition englobe ce que l'on fait usuellement en mathématiques lorsque l'on projette une fonction y connue sur un sous-espace de dimension finie de $\mathcal{E}$, par exemple quand on tronque un développement en série de Fourier [2], ou bien encore tout simplement lorsque l'on fait des approximations locales via l'utilisation de développements limités.

Ce n'est pas tout à fait de cela dont il s'agit dans le cadre de cette thèse : on se place ici plutôt dans le cadre d'une fonction y **inconnue** [3] , et que l'on voudrait estimer globalement en partant des rares informations disponibles, à savoir les données observées $(\mathbf{X}, \mathbf{Y})$ et éventuellement quelques informations *a priori*.

L'approximation de y se fait alors en cherchant une fonction m permettant d'obtenir le meilleur ajustement possible aux observations tout en étant contrainte à rester dans un certain espace de fonctions de dimension finie (ex : polynômes de degré p, combinaisons linéaires de fonctions de base données, etc...) ou encore à posséder certaines propriétés pré-établies (ex : $\mathcal{C}^2$ et de courbure aussi petite que possible). L' *a priori* concernant y peut ainsi être pris en compte soit en rendant le problème paramétrique (Cf. 2.1.2.), soit en utilisant des critères de pénalisation (Cf. 2.1.3.).

2.1.2 Approximation paramétrique

Considérons $\mathcal{F} \subset \mathcal{E}$, un sous-ensemble de $\mathcal{E}$ ($\mathcal{F}$ peut être l'ensembles des polynômes de degré donné, les fonctions deux fois continûment dérivables, etc...). Supposons dans un premier temps que $\mathcal{F}$ est une famille de fonctions paramétrées par un vecteur α à composantes réelles, et de taille finie. Ce cas inclut les ensembles de fonctions affines, polynômiales, polynômiales trigonométriques, réseaux de neurones dont le nombre de couches cachées est fini, ou bien tout simplement l'ensemble des combinaisons linéaires d'une famille de fonctions $\{f_1, ..., f_b\}$ fixée. Rechercher la "meilleure" approximation de y parmi les fonctions de $\mathcal{F}$ signifie souvent minimiser (dans l'espace $\mathcal{F}$) l'*erreur d'apprentissage* E_a :

[2] Cela renvoie au cadre général de l'approximation dans des espaces de Hilbert, faisant notamment l'objet d'avancées récentes telles que l'approximation par ondelettes (Cf. [Mal99], [AAP06])

[3] Remarquons que cette distinction n'exclut pas une approche par séries de Fourier tronquées pour l'estimation d'une fonction inconnue...mais les coefficients seraient alors eux-mêmes approximés.

$$\left\{ \begin{array}{c} \min_{m\in\mathcal{F}} E_a(m) \\ \text{où } E_a(m) = \frac{1}{n}\sum_{i=1}^{n}(m(\mathbf{x}^i) - y_i)^2 = \frac{1}{n}||\mathbf{Y} - m(\mathbf{X})||^2_{\mathbb{R}^n} \end{array} \right. \tag{2.1}$$

E_a est parfois aussi appelée l'*erreur quadratique moyenne d'apprentissage* (traduction de *Mean Squared Error*, Cf. [DM02]). On remarque que toute fonction de $\mathcal{F}$ interpolant $(\mathbf{X}, \mathbf{Y})$ est solution du problème (2.1). Autrement dit, si l'espace de fonctions est riche[4], on a des chances de trouver une approximation qui « colle »parfaitement aux observations. Il n'est pour autant pas du tout évident sans une hypothèse très forte sur le rapport entre la fonction y, l'espace de fonctions $\mathcal{F}$, et le plan d'expériences $\mathbf{X}$ qu'une faible erreur d'apprentissage E_a entraîne une approximation globalement « bonne », via par exemple une faible erreur quadratique intégrée[5] $\int_{x\in D}(y(x)-m(x))^2dx$. Remarquons que cette dernière quantité est similaire à ce qui est appelé *erreur de généralisation* en théorie de l'apprentissage statistique.

Minimiser E_a a tout prix n'est ainsi pas systématiquement souhaitable : une approximation précise sur le plan d'expériences est succeptible de donner de mauvais résultats en généralisation. On fait souvent référence à ce phénomène sous le nom de *sur-apprentissage* (ou *overfitting*, Cf. le chapitre 2 de [HTF01] pour une discussion illustrée).

En **approximation linéaire**, on se place dans l'espace vectoriel

$$\mathcal{F} = \left\{ m_\alpha(\mathbf{x}) = \sum_{j=1}^{b} \alpha_j f_j(\mathbf{x}),\ \alpha \in \mathbb{R}^b \right\} \tag{2.2}$$

engendré par les fonctions $\{f_1, ..., f_b\}$. On note $f(\mathbf{x})$ le vecteur $[f_1(\mathbf{x}), ..., f_b(\mathbf{x})]^T$. Le programme de minimisation (2.1) est alors quadratique[6] en α, et possède à ce titre au moins une solution. On peut calculer le gradient et la Hessienne de l'erreur d'apprentissage :

$$\left\{ \begin{array}{c} \nabla_\alpha E_a(m_\alpha) = -\frac{2}{n}\sum_{i=1}^{n}[y_i - m_\alpha(\mathbf{x}^i)]f(\mathbf{x}^i) \\ \nabla^2_\alpha E_a(m_\alpha) = \frac{2}{n}\sum_{i=1}^{n} f(\mathbf{x}^i)f(\mathbf{x}^i)^T \end{array} \right. \tag{2.3}$$

Si la matrice $\sum_{i=1}^{n} f(\mathbf{x}^i)f(\mathbf{x}^i)^T$ est inversible, le problème admet une unique solution α^*, obtenue en annulant le gradient. On obtient le système linéaire :

$$\left[\sum_{i=1}^{n} f(\mathbf{x}^i)f(\mathbf{x}^i)^T\right]\alpha^* = \sum_{i=1}^{n} f(\mathbf{x}^i)y_i \implies \boxed{\alpha^* = \left[\sum_{i=1}^{n} f(\mathbf{x}^i)f(\mathbf{x}^i)^T\right]^{-1}\sum_{i=1}^{n} f(\mathbf{x}^i)y_i} \tag{2.4}$$

[4] c.f. par exemple [Vap98] pour donner un sens mathématique précis au mot « riche ».

[5] On a fait ici implicitement l'hypothèse qu'une telle quantité existe bien et est finie.

[6] La hessienne $\nabla^2_\alpha E_a(m_\alpha)$ est même semi définie positive (s.d.p.), et d.p. lorsque la famille est libre.

L'unique solution de (2.1) s'écrit alors

$$\begin{aligned} m_{\alpha^*}(\mathbf{x}) &= \alpha^{*T} f(\mathbf{x}) \\ &= f(\mathbf{x})^T \alpha^* \\ &= f(\mathbf{x})^T (F^T F)^{-1} F^T \mathbf{Y} \end{aligned} \tag{2.5}$$

où $F = [f(\mathbf{x}^1), ..., f(\mathbf{x}^n)] \in \mathcal{M}_{b,n}(\mathbb{R})$. Remarquons qu'en chaque $\mathbf{x} \in D$ fixé, l'approximation optimale obtenue est linéaire en les observations $\{y_i, i \in [1, n]\}$.

Exemples de modèles d'approximation linéaire reposant sur l'eq. (2.4) :

Fonctions polynômiales par morceaux et splines : le choix des f_j ($j \in [1, b]$) d'une approximation linéaire joue un rôle essentiel dans la forme des solutions obtenues. Il est par exemple bien connu que l'utilisation de polynômes comme fonctions de base peut occasionner des comportements oscillatoires indésirables. Les fonctions polynômiales par morceaux offrent davantage de flexibilité. Le domaine (on prend pour cet exemple un ensemble monodimensionnel, disons $D = [0, 1]$) est subdivisé en sous-intervalles $[\gamma_i, \gamma_{i+1}]$ ($i \in [0, N_\gamma + 1]$), avec pour convention $\gamma_0 = 0$ et $\gamma_{N_\gamma+1} = 1$; les γ_i $(1 \leq i \leq N_\gamma)$ sont appelés des « noeuds » (*knots*). On définit une première base de fonctions polynômiales par morceaux en posant

$$f_1(x) = \mathbb{1}_{x<\gamma_1},\ f_2(x) = \mathbb{1}_{\gamma_1 \leq x < \gamma_2},\ ...\ f_{N_\gamma}(x) = \mathbb{1}_{x \geq \gamma_{N_\gamma}}$$

Approximer une fonction y dans cette base, i.e. minimiser l'erreur d'apprentissage en supposant que $y = \sum_{j=1}^{b} \alpha_j f_j$, mène directement à une approximation constante par morceaux, où chaque constante est la moyenne empirique des observations sur le morceau correspondant. On peut évidemment obtenir des approximations de plus en plus proches des observations en enrichissant la base de fonctions. Incorporer les fonctions

$$x\mathbb{1}_{x<\gamma_1}, x\mathbb{1}_{\gamma_1 \leq x < \gamma_2},\ ...\ x\mathbb{1}_{x \geq \gamma_{N_\gamma}}$$

permet par exemple d'obtenir des polynômes de degré 1 sur chaque intervalle de la subdivision. Cela dit, rien n'impose jusqu'ici à l'approximation obtenue d'être continue aux noeuds $\{\gamma_i,\ i \in [1, N_\gamma]\}$ (Cf. graphes [HTF01] p. 118). Une manière élégante de forcer la continuité aux noeuds est de remplacer la base précédemment définie par la base des *splines linéaires* :

$$f_1(x) = 1,\ f_2(x) = x,\ f_3(x) = (x - \gamma_1)_+,\ ...\ f_{N_\gamma+2}(x) = (x - \gamma_{N_\gamma})_+ \tag{2.6}$$

On préfère souvent avoir recours à des approximations plus lisses, quitte à augmenter le degré des polynômes par morceaux considérés. La base des *splines cubiques* permet

d'obtenir des approximations dont les dérivées première et seconde sont continues sur l'ensemble du domaine :

$$\begin{cases} f_1(x) = 1,\ f_2(x) = x,\ f_3(x) = x^2,\ f_4(x) = x^3, \\ f_5(x) = (x - \gamma_1)_+^3,\ \ldots\ f_{N_\gamma+4}(x) = (x - \gamma_{N_\gamma})_+^3 \end{cases} \tag{2.7}$$

Les splines cubiques sont les plus utilisées dans la pratique, ainsi que leur variante appelée *splines cubiques naturelles* pour lesquelles est rajoutée la contrainte d'être affine aux régions extrêmes, ce qui permet de réduire les effets de bords tout en gagnant quatre degrés de liberté. Précisons enfin que différents modèles de splines ont été introduits pour les problèmes multivariables (Cf [HTF01] pour plus de détails) ; une approche possible —bien que souvent peu raisonnable— consiste à prendre comme fonctions de base des produits de fonctions de base monodimensionnelles. Le nombre de combinaisons possibles croissant exponentiellement en la dimension, il devient rapidement indispensable d'utiliser des méthodes automatiques de sélection de fonctions de base. La procédure MARS (*Multivariate Adaptive Regression Splines*), implémentée en langage R, est un exemple d'approche aujourd'hui couramment utilisée dans les applications.

En **approximation non-linéaire**, les fonctions m_α sont quelconques en α (ex : *projection pursuit*, réseaux de neurones). La minimisation de l'erreur d'apprentissage se fait alors généralement en utilisant des méthodes numériques basées sur les gradients[7] (Cf. 3.1.2.). De telles méthodes délivrent —après convergence— des minima locaux.

"Projection pursuit" (PP) et réseaux de neurones : l'idée centrale de ces méthodes est de s'appuyer sur certaines combinaisons linéaires bien choisies des d composantes de x, puis de modéliser la réponse comme une fonction non-linéaire des combinaisons caractéristiques en question. En particulier, la *projection pursuit* consiste à rechercher parmi les fonctions de type

$$m_{\mathbf{w}}(x) = \sum_{j=1}^{b} f_j(w_j^T x), \tag{2.8}$$

celles qui présentent le meilleur ajustement aux observations. Chaque $f_j(w_j^T x)$ est appelée fonction *ridge*. Les f_j sont à déterminer au cours de l'estimation, ainsi que les vecteurs $w_j \in \mathbb{R}^d$. L'estimation des fonctions ridge se fait généralement séquentiellement (une seule fonction ridge estimée pour de bon, puis une nouvelle est introduite, etc...), et pour chaque fonction ridge fixée, f_j et w_j sont estimées tour à tour en ré-itérant jusqu'à convergence (Cf. [HTF01] p.349). Les *réseaux de neurones*, développés dans le

[7] l'algorithme utilisé pour les réseaux de neurones est appelé *rétro-propagation d'erreur*

domaine de l'intelligence artificielle, constituent une méthode populaire et très répandue en apprentissage machine (*machine learning*). A titre d'exemple, l'équation du réseau de neurones à une couche cachée s'écrit exactement comme l'eq. (2.8) avec comme fonctions de base f_j des fonctions sigmoïdes ou gaussiennes paramétrées par trois coefficients scalaires. Ce dernier modèle apparaît ainsi comme un cas particulier de *projection pursuit*.

Dans le cas linéaire comme dans le cas non-linéaire, l'approximation obtenue dépend de l'*a priori* du modélisateur via le choix de l'espace de fonctions $\mathcal{F}$ considéré. Ce choix est loin d'être sans conséquence sur l'approximation obtenue (et ses extrema, Cf. chapitre 4). Voyons maintenant comment cet *a priori* peut se transcrire d'une manière alternative, non plus en restreignant l'espace de recherche $\mathcal{F}$ mais en changeant le critère E_a à minimiser via l'ajout d'un terme de pénalité.

2.1.3 Approximation non-paramétrique : régularisation

On a évoqué plus haut le fait que le problème

$$\min_{m\in\mathcal{E}} E_a(m) \tag{2.9}$$

ne pouvait généralement pas être résolu de manière univoque sans un ajout d'information de la part du modélisateur, tel qu'une restriction de l'espace de recherche $\mathcal{E}$ à une famille de fonctions plus modeste (un espace vectoriel de dimension finie, par exemple). Une autre manière de rendre le problème *bien posé*, dont le développement est attribué au mathématicien russe Thikonov, consiste à changer la fonction objectif de l'éq. (2.9) en introduisant un terme de *pénalité*, et tout en laissant l'espace de recherche $\mathcal{E}$ inchangé. On cherche ainsi maintenant dans l'espace $\mathcal{E}$ tout entier à minimiser l'erreur d'apprentissage E_a pénalisée [8]par un *critère de régularité* p :

$$\min_{m\in\mathcal{E}}[E_a(m) + \lambda p(m)] \tag{2.10}$$

où le *paramètre de lissage* λ (aussi appelé *paramètre de complexité*) contrôle le compromis entre la fidélité aux données et la régularité souhaitée de la fonction solution. La pénalité p est une fonction $f \in \mathcal{E} \longmapsto p(f) \in \mathbb{R}^+$ (il se peut que l'ensemble de définition de p soit strictement inclus dans $\mathcal{E}$), choisie de telle manière à ce que $p(f)$ proche de zéro signifie que f est proche d'avoir exactement les propriétés de régularité voulues, et qu'inversement une grande valeur $p(f)$ signifie une fonction f peu compatible avec ces

[8]Remarque : on peut voir le problème $\min_{m\in\mathcal{F}} E_a(m)$ comme le problème pénalisé $\min_{m\in\mathcal{E}}[E_a(m) + J(m)]$ où $J(m) = 0$ si $m \in \mathcal{F}$ et $J(m) = +\infty$ si $m \notin \mathcal{F}$.

dernières. Un exemple de famille de semi-normes (au carré) très couramment utilisées comme pénalités est donné par $\forall k \in \mathbb{N}^*,\ \forall f \in \mathcal{C}^k([0,1],\mathbb{R}),\ \ p_k(f) = \int_0^1 (f^{(k)}(x))^2 dx$, où $f^{(k)}$ est la dérivée k-ième de f. En particulier, la semi-norme au carré p_2 traduit la notion de courbure moyenne d'une fonction de classe $\mathcal{C}^2([0,1],\mathbb{R})$, et le problème ci-dessous est bien connu en approximation fonctionnelle :

$$(\mathcal{A}_2) \qquad \min_{m\in\mathcal{E}} \left[\frac{1}{n}\sum_{i=1}^{n} (m(\mathbf{x}^i) - y_i)^2 + \lambda \int_{t\in D} [m^{(2)}(t)]^2 dt \right] \tag{2.11}$$

La résolution de l'eq. (2.11) mène aux splines de lissage, couramment utilisées dans les applications des mathématiques, par exemple en statistiques pour approcher une fonction sur la base de données bruitées. On pourra consulter l'un des ouvrages de référence sur le sujet, [Wah90], au prix d'un investissement mathématique certain ([HTF01] donne aussi un aperçu plutôt pédagogique de ces techniques). Remarquons en reprenant les notations de [Car] que le problème d'approximation $(\mathcal{A}_2)$ peut en fait être vu comme un cas particulier des problèmes d'approximation $\mathcal{A}_k$ $(k \in \mathbb{N}^*)$, où la pénalité p_2 de $(\mathcal{A}_2)$ est généralisée à p_k $(k \in \mathbb{N}^*)$, l'espace de recherche étant alors bien sûr restreint à des fonctions au moins k fois dérivables [9] . Lorsque le paramètre de complexité $\lambda \to 0^+$, les $\mathcal{A}_k$ deviennent des problèmes d'interpolation, notés $\mathcal{I}_k$: on cherche à minimiser p_k sous contrainte d'annuler $\sum_{i=1}^n (m(\mathbf{x}^i) - y_i)^2$, c'est à dire d'interpoler les observations au plan d'expériences [10]. Pour revenir au très populaire cas $k = 2$, l'analogue de (2.11) en interpolation est le problème de minimisation contrainte

$$(\mathcal{I}_2) \qquad \begin{cases} \min_{m\in\mathcal{E}} \left[\int_{t\in\mathbb{D}} [m^{(2)}(t)]^2 dt \right] \\ \text{sous les contraintes : } \forall i \in [1,n],\ m(\mathbf{x}^i) = y_i \end{cases} \tag{2.12}$$

dont les solutions sont les *splines cubiques d'interpolation*. Cela est rigoureusement mis en lumière dans [Wah90], où l'on peut aussi trouver quelques remarques sur les rapports entre splines, krigeage, et processus stochastiques (voir le chapitre suivant pour plus de détails sur le krigeage). Signalons aussi qu'une synthèse des méthodes non-paramétriques (incluant les polynômes locaux, les estimateurs à noyaux,...) est par ailleurs présentée en langue française dans la thèse [AF04] dédiée aux modèles additifs parcimonieux.

[9] Par soucis de concision et de simplicité, nous ne rentrerons pour le moment pas dans les détails des espaces de classes d'équivalence de fonctions traités usuellement en *analyse fonctionnelle* (e.g. les "L^p"), ainsi que des notions de dérivation (au sens des distributions) qui leur sont associées.

[10] A l'inverse lorsque $\lambda \to +\infty$, on cherche à minimiser $\sum_{i=1}^n (m(\mathbf{x}^i) - y_i)^2$ sous contrainte d'annuler p_k. Pour $k = 1$ par exemple, la solution est une fonction de valeur constante égale à $\frac{1}{n}\sum_{i=1}^n y_i$ (régression sur les constantes). Cela revient en fait à restreindre l'espace de recherche à $\mathcal{F} = \{f \in \mathcal{E},\ p(f) = 0\}$.

2.2 Modèles de régression : un cadre statistique pour l'approximation

Le terme générique de *régression* est employé lorsque l'on cherche à expliquer, sur la base d'observations, le comportement d'une variable aléatoire Y (dite *principale* ou endogène) en fonction d'autres variables aléatoires X_i (dites *explicatives*, *exogènes*, ou encore *predictives*) et d'un terme de **bruit** ε. La relation fonctionnelle liant la variable principale aux variables explicatives peut être de plusieurs natures : linéaire (paramétrique), non-linéaire (idem), voire non-paramétrique. Différentes hypothèses peuvent aussi être faites sur le terme de bruit, terme dont l'importance est capitale pour la validation et l'interprétation des modèles de régression. Nous présentons ici succintement le cas linéaire gaussien puis abordons quelques cas non-linéaires (parmi la pléthore des modèles existants), le but étant d'évoquer certaines propriétés incontournables de la régression.

2.2.1 Le modèle linéaire gaussien multivarié

On considère ici que chaque $y_i = y(\mathbf{x}^i)$ s'exprime comme somme d'une combinaison linéaire de fonctions de base f_k connues (cas le plus courant : $f_k(\mathbf{x}) = x_k$) évaluées en les points $\mathbf{x}^i$ du plan d'expériences et d'une réalisation ε_i d'un bruit gaussien centré $\boldsymbol{\epsilon_i}$ de variance σ^2. On suppose de plus que les variables aléatoires $\boldsymbol{\epsilon_i}$ sont indépendantes dans leur ensemble. On peut résumer ces différentes hypothèses ainsi :

$$\begin{cases} \forall i \in [1,n],\ y_i = \sum_{k=1}^{b} \beta_k f_k(\mathbf{x}^i) + \varepsilon_i \\ \epsilon = (\epsilon_i)_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \end{cases} \tag{2.13}$$

Matriciellement, on obtient le système

$$\mathbf{Y} = \mathbf{F}\beta + \varepsilon, \tag{2.14}$$

où $\beta = (\beta_1, ..., \beta_n)^T$ est le vecteur des *coefficients de régression* et $\mathbf{F} = \{f_k(\mathbf{x}^i)\}_{i,k}$ est la *matrice d'expériences*. Remarquons que le vecteur de bruit ε est défini comme différence entre les observations et la partie déterministe du modèle de régression. C'est en imposant au bruit $\mathbf{Y} - \mathbf{F}\beta$ de respecter certaines conditions que l'on estime les coefficients β_k. On cherche classiquement les coefficients inconnus de manière à ce que les β_k rendent l'observation du vecteur de bruit $\mathbf{Y} - \mathbf{F}\beta$ la plus vraisemblable possible sous l'hypothèse gaussienne, c'est à dire qu'ils maximisent la fonction de vraisemblance :

$$L(\beta; \varepsilon) = f(\varepsilon|\beta) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2}\varepsilon^T \varepsilon} = (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2}[\mathbf{Y}-\mathbf{F}\beta]^T[\mathbf{Y}-\mathbf{F}\beta]}, \tag{2.15}$$

ce qui conduit directement à minimiser $||\mathbf{Y} - \mathbf{F}\beta||^2$ comme fonction de β. Puisque le critère est convexe, on peut trouver le vecteur de paramètres optimaux $\hat{\beta}$ en faisant

directement un calcul de point critique : $\hat{\beta} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{Y}$ (remarquons au passage la similarité avec l'équation (2.4)). Le modèle probabiliste permet de préciser davantage ce résultat en donnant la loi du $\hat{\beta}$ estimé en fonction du "vrai" β :

$$\hat{\beta} \sim \mathcal{N}(\beta, \Sigma_\beta) := \mathcal{N}\left(\beta, \sigma^2(\mathbf{F}^T\mathbf{F})^{-1}\right). \tag{2.16}$$

Les estimations $\widehat{\mathbf{Y}}$ et $\hat{\varepsilon}$ de la réponse et du bruit au plan d'expériences s'écrivent alors :

$$\widehat{\mathbf{Y}} := \mathbf{F}\hat{\beta} \sim \mathcal{N}\left(\mathbf{Y}, \sigma^2\mathbf{H}\right) \text{ où } \mathbf{H} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T, \tag{2.17}$$

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{M}) \text{ où } \mathbf{M} = I - H. \tag{2.18}$$

C'est sur la base de l'eq. (2.18) que sont conçus la plupart des tests statistiques permettant d'évaluer la pertinence d'un modèle de régression [ABC92, Car03b].

Une fois le vecteur de paramètres $\hat{\beta}$ estimé et le modèle statistique jugé acceptable, on peut faire des prédictions de la réponse pour tout nouveau jeu de valeurs des variables explicatives. Soit en effet $\mathbf{F_{new}} = \{f_k(\mathbf{x}^i_{new})\}_{i,k}$ la matrice d'expériences associée à un ensemble "test", i.e. pour $\mathbf{X}_{new} = \{\mathbf{x}^1_{new}, ..., \mathbf{x}^m_{new}\}$. On peut alors utiliser le modèle de régression linéaire pour construire un prédicteur des valeurs de la réponse en $\mathbf{X}_{new}$:

$$\widehat{\mathbf{Y_{new}}} = \mathbf{F_{new}}\hat{\beta} \sim \mathcal{N}\left(\mathbf{F_{new}}\beta, \sigma^2\mathbf{F_{new}}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F_{new}}^T\right). \tag{2.19}$$

La connaissance de cette loi permet par exemple de construire des intervalles de prévision pour les prédictions données par le modèle de régression. Cela peut aussi s'avérer très utile pour propager des incertitudes ou estimer des probabilités de dépassement de seuil. Rappelons à quel point tout ceci est conditionné par le choix du modèle, et qu'il faut donc rester prudent dans l'analyse des résultats (comme dans toute modélisation probabiliste).

Cas particulier : polynômes de régression de degré 2 avec interactions

En prenant comme fonctions de base tous les monômes de degré inférieur ou égal 2 en les composantes du vecteur $\mathbf{x}$, on a que $b = 1 + 2d + \frac{d(d-1)}{2}$ et :

$$\begin{cases} f_1(\mathbf{x}) = 1 \\ f_2(\mathbf{x}) = x_1, .., f_{d+1}(\mathbf{x}) = x_d \\ f_{d+2}(\mathbf{x}) = x_1^2, ..., f_{2d+1}(\mathbf{x}) = x_d^2 \\ f_{2d+2}(\mathbf{x}) = x_1x_2, ..., f_b(\mathbf{x}) = x_{d-1}x_d. \end{cases} \tag{2.20}$$

Ce modèle de régression est le plus employé dans la pratique. Notons que lorsque d augmente, il n'est pas toujours possible de considérer l'ensemble des b variables explicatives

(e.g. lorsque $d = 10$, on a besoin d'au moins $b = 66$ données pour estimer β). Nous donnerons dans le chapitre 3 un exemple d'optimisation reposant sur un tel modèle.

2.2.2 Quelques autres modèles de régression

Régression non-linéaire

L'objet de la régression non-linéaire est d'expliquer Y en fonction de variables explicatives d'une manière plus générale qu'en régression linéaire. La régression paramètrique non-linéaire vise à estimer au mieux les coefficients intervenant dans des relations fonctionnelles quelconques entre Y et les X_k ($k \in [1, d]$). On pourrait par exemple rechercher α tel que Y soit approchée la plus fidèlement possible par $e^{-\frac{\alpha}{X}}$. Cela nous mènerait à un problème d'optimisation numérique non-linéaire et même non-convexe. Ces questions délicates sont abordées en détail dans [ABC92].

Exemple de régression non-paramétrique : Nadaraya-Watson

Une autre approche généralisant la régression linéaire consiste à rechercher une approximation non-paramétrique de la réponse. Cela ouvre un champ d'étude mathématique très vaste qu'il serait difficile de résumer en quelques lignes. Citons tout de même le cas des techniques de lissage en dimension 1 : au lieu de présupposer une forme paramétrique de dépendance en des variables explicatives, on prédit la réponse au point $\mathbf{x}$ comme une somme pondérée des observations faites pour des abscisses voisines :

$$m(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_\lambda(\mathbf{x}, \mathbf{x}^i) y_i}{\sum_{i=1}^{n} K_\lambda(\mathbf{x}, \mathbf{x}^i)} \tag{2.21}$$

où $K_\lambda : (\mathbf{x}, \mathbf{x}') \in D \times D \longmapsto K_\lambda(\mathbf{x}, \mathbf{x}') \in \mathbb{R}$ est une fonction (souvent appelée *noyau*) servant à quantifier la similarité attendue entre la réponse en deux points quelconques $\mathbf{x}$ et $\mathbf{x}'$ de D. Les poids intervenant dans le calcul de la fonction de régression m dans la méthode de Nadaraya-Watson sont ainsi calculés en se basant sur un noyau K_λ, ce dernier étant traditionnellement une densité de probabilité gaussienne évaluée en $\mathbf{h} := \mathbf{x} - \mathbf{x}'$. On rencontre parfois le terme de *fenêtre de lissage* pour désigner une telle fonction de pondération. Remarquons qu'en dépit de l'appellation « non-paramétrique », on retrouve malgré tout des paramètres (ici λ) lors du dimensionnement de la fenêtre. La *largeur de fenêtre* joue ainsi un rôle important quant à la variabilité de l'approximation obtenue : elle résume —lorsque la densité utilisée est à support borné— l'éloignement à partir duquel on peut considérer que les observations en deux points cessent de s'influencer mu-

tuellement. Ce type de régression est particulièrement pertinent lorsque l'on dispose de beaucoup de données (on retombe en fait sur les splines de lissage évoquées en 2.1.2.). Or on peut souvent se ramener à ce cas lorsque l'on traite des probèmes multidimensionnels pour lesquels on peut faire une hypothèse simplificatrice telle que l'**additivité**.

Exemple de régression semi-paramétrique : modèles additifs

Les modèles additifs constituent une classe de modèles de régression bien spécifique, particulièrement commode pour faire chuter la complexité de certains problèmes en grande dimension : la tendance de la réponse Y y est supposée dépendre des prédicteurs $\mathbf{X}_k$ de manière non nécessairement linéaire mais additive, c'est à dire via une relation du type $Y = \sum_{k=1}^{b} f_k(X_k) + \varepsilon$ avec f_k des fonctions monodimensionnelles quelconques. En poursuivant avec les notations employées dans la section 2.2.1, on peut développer les observations y_i issues d'un modèle additif de la manière suivante :

$$\begin{cases} \forall i \in [1,n],\ y_i = \sum_{k=1}^{b} f_k(\mathbf{x}_k^i) + \varepsilon_i \\ \epsilon = (\epsilon_{\mathbf{i}})_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \end{cases} \tag{2.22}$$

où chaque f_k est une fonction monodimensionnelle paramétrique ou non-paramétrique fixée par le modélisateur. Les grands avantages de ces modèles sont de permettre une visualisation fidèle de la relation entrées-sortie (la donnée d'un graphe par dimension permet de résumer complètement le modèle), mais aussi d'aller de pair avec un algorithme d'estimation efficace des fonctions f_k : le *backfitting*. Ce dernier repose sur des résultats mathématiques solides (Cf. [HT95], chap. 5), et permet sous des hypothèses assez faibles de garantir que le fait de procéder dimension par dimension est suffisant pour converger vers la meilleure approximation additive possible. Suivant précisément la description de l'algorithme donnée dans ([HT95], chap. 4), le backfitting se résume aux étapes suivantes :

Algorithm 1 Idée de l'algorithme de backfitting

Initialiser $f_j = f_j^0$, $j = 1, ..., d$

Cycler : $j = 1, ..., d, 1, ..., d, ...$

$$f_j = \mathcal{S}_j \left(\mathbf{Y} - \sum_{k \neq j} f_k \Big| x_j \right)$$

Continuer (2) jusqu'à ce que les fonctions f_j se stabilisent

où la notation $\mathcal{S}_j \left(\mathbf{Y} - \sum_{k \neq j} f_k | x_j \right)$ signifie un lissage (de type Nadaraya-Watson) du nuage de points obtenu par projection du résidu $\mathbf{Y} - \sum_{k \neq j} f_k$ sur l'espace engendré par

x_j. Cette étape de lissage peut en fait être remplacée par toute forme d'approximation monodimensionnelle de $\mathbf{Y} - \sum_{k \neq j} f_k$ en fonction de x_j, telle qu'une régression linéaire. Nous appliquerons dans le chapitre 6 un modèle additif avec des splines de lissage comme fonctions monodimensionnelles. Mentionnons tout de suite l'existence d'un package R développé par les auteurs de [HT95, HTF01].

2.3 Un mot sur les espaces de Hilbert à noyau reproduisant

Il serait peu raisonnable de faire un tour d'horizon des méthodes d'approximation sans mentionner les espaces de Hilbert à noyau reproduisant (souvent appelés RKHS, pour *Reproducing Kernel Hilbert Spaces*), tant la vision et le formalisme qui leur sont associés ont pris de l'importance dans de nombreux travaux récents en apprentissage fonctionnel [RW06, Vap98, Wah90]. Donnons tout d'abord quelques définitions indispensables, dans l'esprit de la présentation faite dans le livre de Wahba [Wah90]. Par souci de simplicité et pour ne pas trop s'éloigner de nos objectifs pratiques, on se restreindra à $\mathbb{R}$ comme corps de base tout au long de l'exposé.

2.3.1 Définitions de base et exemples

Noyaux positifs et RKHS

Soit $\mathcal{T}$ un ensemble, que l'on supposera par défaut être $\mathbb{R}^d$ ($d \in \mathbb{N}^*$). Nous nous intéresserons ici —et en de multiples occasions par la suite— à la notion de *type positif* et de *définie positivité* pour les fonctions $\mathcal{T} \times \mathcal{T} \longrightarrow \mathbb{R}$ (ou « *noyaux* ») symétriques.

Définition ([Aro50, Wah90]) : un noyau $k : (s,t) \in \mathcal{T} \times \mathcal{T} \longmapsto k(s,t) \in \mathbb{R}$, symétrique[11] (i.e. tel que $\forall s,t \in \mathcal{T},\ k(s,t) = k(t,s)$), est dit *de type positif* lorsque pour tous réels $a_1, ..., a_n$ et $t_1, ..., t_n \in \mathcal{T}$ on a

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(t_i, t_j) \geq 0 \tag{2.23}$$

Lorsque l'inégalité est stricte pour tout $(a_1, \ldots, a_n) \in \mathbb{R}^n \backslash \{(0, \ldots, 0)\}$, on parle de noyau *défini positif*.

On peut remarquer que les propriétés de symétrie et de positivité font d'un tel noyau un objet conceptuellement assez proche d'un produit scalaire, même s'il lui manque en

[11] Nous ne rencontrerons ici que des noyaux symétriques, si bien que nous ne le signalerons plus.

premier lieu la propriété de *bilinéarité*. Nous allons voir un peu plus loin que malgré cette apparente lacune, tout noyau de type positif k permet de définir un espace de fonctions muni d'un produit scalaire particulier, directement lié à k. Rappelons qu'un espace vectoriel E est dit *préhilbertien* dès lors qu'il est muni d'un produit scalaire $\langle .,. \rangle$, et que l'on appelle *Espace de Hilbert* tout espace préhilbertien $(E, \langle .,. \rangle)$ complet pour la norme $||.|| : x \in E \longmapsto ||x|| = \sqrt{\langle x, x \rangle} \in \mathbb{R}^+$ (i.e. tel que toute suite de Cauchy pour cette norme converge dans E). On essaye souvent en analyse de se placer dans des espaces de Hilbert de manière à pouvoir utiliser un ensemble de propriétés extrêmement commodes (Cf. [Rud77]), permettant en particulier de transposer des résultats classiques de géométrie euclidienne à des espaces fonctionnels.

Définition : on appelle *espace de Hilbert réel à noyau reproduisant* [12] tout espace de Hilbert $(\mathcal{H}, \langle .,. \rangle)$ formé de fonctions $\mathcal{T} \longrightarrow \mathbb{R}$ telles que pour tout $t \in \mathcal{T}$, la fonctionnelle d'**évaluation** $L_t : f \in \mathcal{H} \longmapsto f(t) \in \mathbb{R}$ soit **continue**. Cette dernière condition est équivalente à l'existence $\forall t \in \mathcal{T}$ d'un nombre $M_t \in]0, +\infty[$ tel que

$$\forall f \in \mathcal{H},\ |L_t f| = |f(t)| \leq M_t ||f||. \tag{2.24}$$

La condition de continuité (2.24) est plus subtile qu'elle ne pourrait paraître de prime abord : elle signifie qu'en tout point $t \in \mathcal{T}$ considéré, la quantité $f(t)$ varie continûment lorsque l'on fait varier f dans $\mathcal{H}$ (et ce en relation avec la distance entre fonctions canoniquement associée à la norme de $\mathcal{H}$). Elle n'est par exemple pas satisfaite pour l'espace $(L^2, ||.||_2)$, qui n'est d'ailleurs même pas un espace de fonctions.

Propriété : soit $(\mathcal{H}, \langle .,. \rangle)$ un RKHS. $\forall t \in \mathcal{T}$, il existe un unique élément $k_t \in \mathcal{H}$ tel que

$$\forall f \in \mathcal{H},\ L_t f = \langle k_t, f \rangle = f(t). \tag{2.25}$$

La fonction k_t est appelé "représentant de l'évaluation en t".

Ce résultat découle directement de l'application du *théorème de représentation de Riesz* (Cf. [Rud77]) aux formes linéaires continues $L_t,\ t \in \mathcal{T}$. Notons que les $k_t,\ t \in \mathcal{T}$ jouissent de certaines propriétés particulières : on obtient par exemple la propriété dite d'*autoreproduction* en écrivant que

$$\forall s, t \in \mathcal{T},\ k_t(s) = L_s k_t = \langle k_s, k_t \rangle = \langle k_t, k_s \rangle = L_t k_s = k_s(t). \tag{2.26}$$

Le résultat fondamental qui suit relie les RKHS et les noyaux de type positif :

[12] dénoté par son abréviation anglaise RKHS dans la suite, en se restreignant toujours au cas réel.

Théorème. *A tout RKHS correspond un unique noyau de type positif, appelée noyau reproduisant. Réciproquement, partant d'un noyau $k : \mathcal{T} \times \mathcal{T} \longrightarrow \mathbb{R}$ de type positif, on peut construire un unique RKHS $\mathcal{H}$ (sur $\mathcal{T}$) dont le noyau reproduisant soit k.*

Démonstration. En posant $k(s,t) = \langle k_s, k_t \rangle$, il est facile de voir que k est de type positif. On a en effet pour tous réels $a_1, ..., a_n$ et $t_1, ..., t_n \in \mathcal{T}$:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(t_i, t_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \langle k_{t_i}, k_{t_j} \rangle \\ &= \left\langle \sum_{i=1}^{n} a_i k_{t_i}, \sum_{j=1}^{n} a_j k_{t_j} \right\rangle = \left\| \sum_{i=1}^{n} a_i k_{t_i} \right\|^2 \geq 0\end{aligned}$$

par la propriété de *positivité* de $\langle ., . \rangle$. Réciproquement, si k est un noyau de type positif, on définit les fonctions k_t en posant $\forall t \in \mathcal{T},\ k_t(.) = k(t, .)$. Considérons l'espace vectoriel $V := vect\{k_t,\ t \in \mathcal{T}\}$ des combinaisons linéaires de type $\sum_{i=1}^{n} a_i k_{t_i}$ ($n \in \mathbb{N},\ \forall i \in \{1, ..., d\},\ a_i \in \mathbb{R},\ t_i \in \mathcal{T}$). On vérifie que la forme bilinéaire symétrique définie par

$$\left\langle \sum_{i=1}^{n} a_i k_{t_i}, \sum_{j=1}^{m} b_j k_{t_j} \right\rangle_V = \sum_{i=1}^{n}\sum_{i=1}^{m} a_i b_j k(t_i, t_j)$$

est un produit scalaire sur V. La propriété de séparation découle du fait que

$$\begin{aligned}\left(\| \sum_{i=1}^{n} a_i k_{t_i} \| = 0 \right) &\Longrightarrow \left(\forall t \in \mathcal{T},\ \langle \sum_{i=1}^{n} a_i k_{t_i}, k_t \rangle = 0 \right) \\ &\Longrightarrow \left(\forall t \in \mathcal{T},\ \sum_{i=1}^{n} a_i k_{t_i}(t) = 0 \right) \Longrightarrow \sum_{i=1}^{n} a_i k_{t_i}(t) = 0\end{aligned} \tag{2.27}$$

D'autre part $\langle ., . \rangle_V$ est bien tel que $\forall t \in \mathcal{T},\ \langle k_t, f \rangle_V = f(t)$ quelque soit $f \in V$, c'est à dire que k_t est bien un représentant de l'évaluation sur V. Le seul problème est que V n'est a priori pas complet. On va en fait obtenir un RKHS défini de manière unique par complétion de cet espace. Voyons ce point un peu plus en détail que cela n'est présenté dans [Wah90], en nous inspirant partiellement du cours [GA] et du rapport [Lar08] : Soit $\mathcal{C}_V$ l'espace vectoriel des suites de Cauchy de l'espace normé $(V, ||.||_V)$, où $||.||_V$ est la norme canoniquement associée au produit scalaire $\langle ., . \rangle_V$. $f = (f_n)_{n \in \mathbb{N}^*} \in \mathcal{C}_V$ signifie que (f_n) est de Cauchy dans $(V, ||.||_V)$, ce qui implique en particulier —par inégalité triangulaire « renversée »— que $(||f_n||_V)_{n \in \mathbb{N}^*}$ est de Cauchy et donc convergente dans $\mathbb{R}$. Notons alors sa limite

$$q(f) := \lim_{n \to +\infty} ||f_n||_V < +\infty \tag{2.28}$$

On vérifie sans encombre que q définit une semi-norme sur $\mathcal{C}_V$. Notons alors $\mathcal{C}_V^0$ le sous-espace vectoriel de $\mathcal{C}_V$ constitué des suites de Cauchy de V qui annulent q :

$$\mathcal{C}_V^0 := \{f \in \mathcal{C}_V : q(f) = 0\} \tag{2.29}$$

On peut alors montrer (Cf. par exemple [GA]) que le séparé de C_V pour q,

$$\widehat{C_V} := C_V/C_V^0, \tag{2.30}$$

est d'une part un espace complet, et qu'il existe d'autre part une application

$$i : (V, ||.||_V) \longrightarrow (\widehat{C_V}, q) \tag{2.31}$$

linéaire continue isométrique et d'image dense. En bref, $(\widehat{C_V}, q)$ est un espace de Banach dans lequel on peut plonger V de manière particulièrement commode. Remarquons que le produit scalaire sur V peut être prolongé à $\widehat{C_V}$ de la manière suivante : soient $\hat{f}, \hat{g} \in \widehat{C_V}$ et $(f_n)_{n\in\mathbb{N}^*}, (g_n)_{n\in\mathbb{N}^*}$ deux suites de Cauchy de V telle que $(f_n) \in \hat{f}$, et $(g_n) \in \hat{g}$. On pose alors

$$\langle \hat{f}, \hat{g} \rangle_{\widehat{C_V}} := \lim_{n\to+\infty} \langle f_n, g_n \rangle_V \tag{2.32}$$

Assurons-nous dans un premier temps que la limite de l'éq. (2.32) a bien un sens, en montrant que la suite définie par $\forall n \in \mathbb{N}^*,\ u_n = \langle f_n, g_n \rangle$ est de Cauchy dans $\mathbb{R}$ et converge donc bien (puisque $\mathbb{R}$ est complet). Soient $n, p \in \mathbb{N}^*$; on a que

$$\begin{aligned}
&|\langle f_{n+p}, g_{n+p}\rangle_V - \langle f_n, g_n\rangle_V| \\
=&|\langle f_{n+p}, g_{n+p}\rangle_V - \langle f_{n+p}, g_n\rangle_V + \langle f_{n+p}, g_n\rangle_V - \langle f_n, g_n\rangle_V| \\
\leq&|\langle f_{n+p}, g_{n+p} - g_n\rangle_V| + |\langle f_{n+p} - f_n, g_n\rangle_V| \\
\leq&\underbrace{||f_{n+p}||_V}_{\text{borné}}\ \underbrace{||g_{n+p} - g_n||_V}_{\underset{n\to+\infty}{\longrightarrow 0}} + \underbrace{||g_n||_V}_{\text{borné}}\ \underbrace{||f_{n+p} - f_n||_V}_{\underset{n\to+\infty}{\longrightarrow 0}} \overset{n\to+\infty}{\longrightarrow} 0
\end{aligned} \tag{2.33}$$

$(u_n)_{n\in\mathbb{N}^*}$ étant de Cauchy, elle converge donc bien dans $\mathbb{R}$, et c'est en fait sa limite $u \in \mathbb{R}$ que l'on a appelé $\langle \hat{f}, \hat{g} \rangle_{\widehat{C_V}}$. Précisons que l'on peut montrer de manière analogue (Cf. [Lar08]) que $\langle \hat{f}, \hat{g} \rangle_{\widehat{C_V}}$ ne dépend pas des suites $(f_n)_{n\in\mathbb{N}^*}, (g_n)_{n\in\mathbb{N}^*} \in V^{\mathbb{N}^*}$ choisies, du moment qu'elles appartiennent respectivement aux classes $\hat{f}$ et $\hat{g}$. Il est finalement capital de remarquer que la norme q rendant $\widehat{C_V}$ complet découle en fait du produit scalaire $\langle ., . \rangle_{\widehat{C_V}}$, si bien que

$$\boxed{(\widehat{C_V},\ \langle ., . \rangle_{\widehat{C_V}}) \text{ est un espace de Hilbert}}$$

Il reste à voir comment on peut maintenant obtenir un espace de Hilbert **de fonctions** de noyau reproduisant k. Disons d'emblée que la clef de ce résultat est l'existence d'une *correspondance biunivoque* (Cf. [Lar08]) entre les éléments de $\widehat{C_V}$ et les fonctions "limites ponctuelles" naturellement associées aux suites de Cauchy de V. C'est ce que nous allons détailler au cours de deux lemmes suivants :

Premier lemme. *Toute $f = (f_n) \in \mathcal{C}_V$ converge simplement vers une fonction $l_f \in \mathbb{R}^{\mathcal{T}}$. De plus, toute $g = (g_n) \in \mathcal{C}_V$ telle que $g \in \hat{f}$ admet la même limite simple que f.*

Démonstration. Soit $f = (f_n) \in \mathcal{C}_V$, $x \in \mathcal{T}$, et $n, p \in \mathbb{N}^*$ quelconques. Comme $\langle f_n, k_x \rangle_V = f_n(x)$ et $\langle f_{n+p}, k_x \rangle_V = f_{n+p}(x)$, il vient que

$$\begin{aligned} |f_{n+p}(x) - f_n(x)| &= |\langle f_{n+p} - f_n, k_x \rangle_V| \\ &\leq ||f_{n+p} - f_n||_V \, ||k_x||_V, \end{aligned} \tag{2.34}$$

et comme (f_n) est de Cauchy dans $(V, ||.||_V)$, il ressort de l'éq. (2.34) que $(f_n(x))$ est de Cauchy dans $(\mathbb{R}, |.|)$. Elle converge donc par complétude de $(\mathbb{R}, |.|)$ vers une limite que l'on appelle $l_f(x)$. En faisant de même pour tout $x \in \mathcal{T}$, on obtient la fonction l_f, limite simple de la suite f. Considérons maintenant une suite $g = (g_n) \in \hat{f}$. On a de nouveau

$$\begin{aligned} |f_n(x) - g_n(x)| &= |\langle f_n - g_n, k_x \rangle_V| \\ &\leq ||f_n - g_n||_V \, ||k_x||_V \end{aligned} \tag{2.35}$$

Or $||f_n - g_n||_V$ tend lorsque $n \to +\infty$ vers $q(f-g)$, qui est nul puisque $g \in \hat{f}$. Cela permet de conclure que les limites simples de f et g coïncident bien en tout point $x \in \mathcal{T}$. □

Deuxième lemme. *Si $f = (f_n),\ g = (g_n) \in \mathcal{C}_V$ sont deux suites de Cauchy de V telles que $\forall x \in \mathcal{T},\ l_f(x) = l_g(x)$, alors $\hat{f} = \hat{g}$ (i.e. $q(f - g) = 0$).*

Démonstration. Commençons par introduire la suite r définie par $\forall n \in \mathbb{N}^*,\ r_n = f_n - g_n$. On va montrer que $q(r) = \lim_{n \to +\infty} ||r_n||_V = 0$. Il est clair que r est de Cauchy comme différence de suites de Cauchy. En particulier, (r_n) est bornée et il existe donc $A > 0$ tel que $\forall n \in \mathbb{N}^*,\ ||r_n||_V < A$. Soit maintenant $\epsilon > 0$. Comme r est de Cauchy, il existe $N_0 \in \mathbb{N}^*$ tel que $\forall n \geq N_0,\ ||r_n - r_{N_0}||_V < \frac{\epsilon}{2A}$. Par ailleurs, $r_{N_0} \in V$ implique —par définition de V— qu'il existe un entier $p \in \mathbb{N}$, un ensemble de points $\{t_i,\ 1 \leq i \leq p\} \in \mathbb{R}^p$ et de coefficients $\{a_i,\ 1 \leq i \leq p\} \in \mathbb{R}^p$ tels que

$$r_{N_0} = \sum_{i=1}^{p} a_i k_{t_i}$$

$\forall n \geq N_0$, on a $\langle r_{N_0}, r_n\rangle_V = \sum_{i=1}^p \langle a_i k_{t_i}, r_n\rangle_V = \sum_{i=1}^p a_i r_n(t_i)$. Sachant que (r_n) converge ponctuellement vers la fonction nulle, on vérifie simplement qu'il existe $N_1 \in \mathbb{N}^*$ tel que

$$\forall n \geq N_1,\ \sum_{i=1}^{p} |a_i||r_n(t_i)| < \frac{\epsilon}{2}$$

En posant $N = \max\{N_0, N_1\}$, on obtient alors que $\forall n \geq N$:

$$\begin{aligned} ||r_n||_V^2 &= \langle r_n - r_{N_0}, r_n\rangle_V + \langle r_{N_0}, r_n\rangle_V \\ &\leq |\langle r_n - r_{N_0}, r_n\rangle_V| + |\langle r_{N_0}, r_n\rangle_V| \\ &\leq \underbrace{||r_n - r_{N_0}||_V}_{<\frac{\epsilon}{2A}} \underbrace{||r_n||_V}_{<A} + \underbrace{\sum_{i=1}^{p} |a_i||r_n(t_i)|}_{<\frac{\epsilon}{2}} < \epsilon \end{aligned} \tag{2.36}$$

On a ainsi montré que $\lim_{n\longrightarrow+\infty} ||r_n||_V = 0$ c'est-à-dire que $q(f-g) = 0$. □

Pour conclure la preuve du théorème, l'espace de fonctions V_{lim} formé par les limites ponctuelles de suites de V et muni du produit scalaire de l'éq. (2.32) est bien un RKHS de noyau reproduisant k. En remarquant que pour tout $t \in \mathcal{T}$, $k_t \in V_{lim}$ comme limite ponctuelle de la suite de terme constant k_t, on a en effet la relation caractéristique :

$$\begin{aligned} \forall f \in V_{lim},\ \langle f, k_t\rangle &= \lim_{n\longrightarrow+\infty} \langle f_n, k_t\rangle_V \\ &= \lim_{n\longrightarrow+\infty} f_n(t) = f(t) \end{aligned} \tag{2.37}$$

où (f_n) est une suite quelconque de fonctions de V admettant f comme limite simple. □

Un premier exemple (fondamental) de RKHS

L'un des grands classiques de la littérature des RKHS est essentiellement basé sur la formule de Taylor avec reste intégral (dite de *Taylor-Laplace*). Si f est une fonction de $[0,1]$ dans $\mathbb{R}$, $m-1$ fois continûment dérivable ($m \in \mathbb{R}$) et telle que $f^{(m)} \in L^2([0,1])$, on rappelle qu'on a l'identité suivante :

$$\begin{aligned} \forall t \in [0,1],\ f(t) &= \sum_{\nu=1}^{m-1} \frac{t^\nu}{\nu!} f^{(\nu)}(0) + \int_0^t \frac{(t-u)^{m-1}}{(m-1)!} f^{(m)}(u)du \\ &= \sum_{\nu=1}^{m-1} \frac{t^\nu}{\nu!} f^{(\nu)}(0) + \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} f^{(m)}(u)du \end{aligned} \tag{2.38}$$

où $(x)_+ = x$ pour $x > 0$ et $(x)_+ = 0$ sinon. Suivant les notations de [Wah90], notons $\mathcal{B}_m$ l'espace des fonctions $(m-1)$ fois dérivables f telles que $\forall \nu \in [0, m-1],\ f^{(\nu)}(0) = 0$, et

W_m^0 le sous-espace vectoriel de $\mathcal{B}_m$ défini ci-dessous :

$$W_m^0 = \{f : [0,1] \longrightarrow \mathbb{R} \text{ t.q. } f \in \mathcal{B}_m, \text{ et } f, f', ..., f^{(m-1)} \text{ abs. cont. }, f^{(m)} \in L^2([0,1])\}$$

Pour tout $f \in \mathcal{B}_m$, et *a fortiori* $\forall f \in W_m^0$, on a alors :

$$\begin{aligned} \forall t \in [0,1],\ f(t) &= \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} f^{(m)}(u)du \\ &= \int_0^1 G_m(t,u) f^{(m)}(u)du \end{aligned} \tag{2.39}$$

où l'on a posé $\forall t, u \in [0,1],\ G_m(t,u) := \frac{(t-u)_+^{m-1}}{(m-1)!}$. On peut vérifier les résultats classiques que constituent le fait que la forme bilinéaire symétrique

$$(f,g) \in W_m^0 \times W_m^0 \longrightarrow \langle f, g \rangle = \int_0^1 f^{(m)}(u) g^{(m)}(u)du$$

définisse bien un produit scalaire sur W_m^0, et que W_m^0 soit complet pour la norme définie par $||f||^2 = \int_0^1 (f^{(m)}(u))^2 du$, faisant ainsi de $(W_m^0, \langle ., . \rangle)$ un espace de Hilbert. Il est maintenant crucial de s'assurer de la continuité de la fonctionnelle d'évaluation L_t. On obtient cette dernière en appliquant l'inégalité de Cauchy-Schwarz à partir de l'eq. (2.39) :

$$\begin{aligned} \forall t \in [0,1],\ |L_t(f)| = |f(t)| &= \left| \int_0^1 G_m(t,u) f^{(m)}(u)du \right| \\ &\leq \sqrt{\left(\int_0^1 (G_m(t,u))^2 du \right)} ||f|| \end{aligned} \tag{2.40}$$

Il est ainsi établi que W_m^0 est un RKHS ; il reste à expliciter son noyau k^1. En utilisant à nouveau l'éq. (2.39) ainsi que la définition du représentant de l'évaluation, il vient que

$$\forall f \in W_m^0,\ \forall t \in [0,1],\ \int_0^1 f^{(m)}(u) k_t^{1,(m)}(u)du = \langle f, k_t^1 \rangle = f(t) = \int_0^1 G_m(t,u) f^{(m)}(u)du$$

On trouve alors par identification que $\forall u \in [0,1],\ k_t^{1,(m)}(u) = G_m(t,u)$, et on obtient par intégrations successives que

$$\forall u \in [0,1],\ k^1(t,u) := k_t^1(u) = \int_0^1 G_m(t,v) G_m(u,v)dv \tag{2.41}$$

On vérifie enfin sans peine que $k_t^1(.) = \int_0^1 G_m(t,v) G_m(.,v)dv \in W_m^0$, et on peut ainsi conclure que k^1 est le noyau reproduisant de W_m^0.

En pratique, on est plus souvent confronté à des espaces de fonctions plus vastes tels que

$$W_m = \{f : [0,1] \longrightarrow \mathbb{R} \text{ t.q. } f, f', ..., f^{(m-1)} \text{ abs. cont., et } f^{(m)} \in L^2([0,1])\}$$

Le travail que nous venons de faire sur W_m^0 est heureusement loin d'être perdu puisque l'on peut munir W_m d'une structure de RKHS, en lien étroit avec celle de W_m^0. Considérons les m fonctions polynômes $\phi_\nu(t) = \frac{t^{\nu-1}}{(\nu-1)!}$ ($\nu \in \{1, ..., m\}$, et soit $\mathcal{H}_0 = vect\{\phi_\nu,\ \nu \in [1, m]\}$ le sous-espace vectoriel de W_m engendré par les ϕ_ν ($\nu \in \{1, ..., m\}$). En appelant D l'opérateur de dérivation sur W_m, on peut définir un produit scalaire sur $\mathcal{H}_0$ comme suit :

$$\forall \phi, \chi \in \mathcal{H}_0,\ \langle \phi, \chi \rangle_{\mathcal{H}_0} = \sum_{\nu=0}^{m-1} \left[(D^\nu \phi)(0)\right] \left[(D^\nu \chi)(0)\right], \tag{2.42}$$

et il vient directement que $\mathcal{H}_0$ est un espace de Hilbert de dimension $m < +\infty$ (un *espace euclidien*), que $\{\phi_1, ..., \phi_n\}$ forme une base orthonormale de cet espace, et surtout que $\mathcal{H}_0$ est un RKHS de noyau

$$\forall s, t \in \mathcal{T},\ k^0(s,t) = \sum_{\nu=1}^{m} \phi_\nu(s)\phi_\nu(t) \tag{2.43}$$

On vérifie en effet que $k_s^0(.) := k^0(s,.) = \sum_{\nu=1}^m \phi_\nu(s)\phi_\nu(.)$ est bien le représentant de l'évaluation en $s \in \mathcal{T}$ en remarquant que

$$\forall \alpha \in \{1, ..., m\},\ \langle k_s^0(.), \phi_\alpha(.) \rangle = \sum_{\nu=1}^{m} \phi_\nu(s) \langle \phi_\nu(.), \phi_\alpha(.) \rangle = \phi_\alpha(s) \tag{2.44}$$

Nous pouvons desormais munir l'espace W_m d'une structure de RKHS sur la base de celles de $\mathcal{H}_0$ et W_m^0, c'est à dire (selon [Wah90], p.7) de construire l'*espace de Sobolev-Hilbert*. En notant $\mathcal{H}_1 = W_m^0$, l'éq. (2.38) revient en effet à la décomposition suivante

$$W_m = \mathcal{H}_0 \oplus \mathcal{H}_1, \tag{2.45}$$

c'est à dire que le développement limité donné par la formule de Taylor-Laplace permet d'exprimer de manière unique toute fonction $f \in W_m$ comme la somme d'une fonction polynôme $f_0 \in \mathcal{H}_0$ et d'une fonction dont les dérivées d'ordre 0 à $m-1$ s'annulent en 0, $f_1 \in \mathcal{H}_1$. Le caractère "direct" de cette somme provient du fait que $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0_{W_m}\}$. On peut en fait dire plus sur cette somme directe, à savoir qu'elle est orthogonale lorsque W_m est muni du produit scalaire "somme" des produits scalaire de $\mathcal{H}_0$ et $\mathcal{H}_1$,

$$\forall f, g \in W_m,\ \langle f, g \rangle_{W_m} = \sum_{\nu=0}^{m-1} \left[(D^\nu f)(0)\right] \left[(D^\nu g)(0)\right] + \int_0^1 [(D^m f)(u)][(D^m g)(u)] du \tag{2.46}$$

Le fait que la somme directe soit alors bien orthogonale est simplement dû au fait que $\forall \phi \in \mathcal{H}_0,\ \int_0^1 [(D^m\phi)(u)]^2 du = 0$ et $\forall f \in \mathcal{H}_1,\ \sum_{\nu=0}^{m-1} [(D^\nu f)(0)]^2 = 0$. En appliquant finalement le résultat classique [Aro50] selon lequel la somme directe de deux RKHS est un RKHS dont le noyau est la somme des deux premiers noyaux, il vient que

$$\begin{aligned} k : (s,t) \in \mathcal{T}^2 \longmapsto k(s,t) &= k^0(s,t) + k^1(s,t) \\ &= \sum_{\nu=1}^{m} \phi_\nu(s)\phi_\nu(t) + \int_0^1 G_m(s,v)G_m(t,v)dv \end{aligned} \tag{2.47}$$

est le noyau reproduisant de W_m. Remarquons au passage —cela nous servira en particulier dans la prochaine section— que la pénalisation $J_m : f \in W_m \longmapsto J_m(f) = \int_0^1 (f^{(m)}(u))^2 du$ peut désormais être interprétée en termes géométriques :

$$J_m(f) = ||P_1 f||^2_{W_m}, \tag{2.48}$$

où P_1 est la projection sur $\mathcal{H}_1$ orthogonalement à $\mathcal{H}_0$.

2.3.2 Un résultat central d'approximation dans les espaces de Hilbert à noyau reproduisant : le théorème du représentant

Le théorème du représentant constitue un résultat d'analyse incontournable dans la mesure où il donne la solution d'une vaste classe de problèmes d'approximation dans les RKHS. Il permet en effet d'exhiber une correspondance entre certains problèmes d'approximation fonctionnelle posés en dimension infinie et des problèmes équivalents de dimension finie, dont la solution est facilement accessible.

Commençons par donner une version très générale de ce théorème, dans l'esprit de ([Ver07], pp.73-76), avant d'en présenter une application dans le contexte de l'approximation de données observationnelles par splines de lissage. Plaçons-nous dans le cadre d'un ensemble $\mathcal{T}$ muni d'un noyau d.p. $k : \mathcal{T} \times \mathcal{T} \longrightarrow \mathbb{R}$, du RKHS $\mathcal{H}$ correspondant, et d'un plan d'expériences fini $P = \{t_1, ..., t_n\} \in \mathcal{T}^n$ $(n \in \mathbb{N}^*)$. Considérant une application $\Psi : \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}$ strictement croissante par rapport à sa dernière variable, le théorème du représentant peut être énoncé comme suit :

Théorème. *Toute solution au problème d'optimisation*

$$\min_{f \in \mathcal{H}} \Psi\left(f(t_1), ..., f(t_n), ||f||_{\mathcal{H}}\right) \tag{2.49}$$

admet une représentation de la forme

$$f(.) = \sum_{i=1}^{n} \alpha_i k(t_i, .) \qquad (\alpha_1, ..., \alpha_n) \in \mathbb{R}^n \tag{2.50}$$

Démonstration. Posons $\mathcal{H}_{\mathcal{P}} := vect\{k(t_i,.), i \in \{1,...,n\}\}$. En tant que sous-espace vectoriel de dimension finie de $\mathcal{H}$, $\mathcal{H}_{\mathcal{P}}$ est un sous-espace vectoriel fermé de $\mathcal{H}$ et admet à ce titre un supplémentaire orthogonal $\mathcal{H}_{\mathcal{P}}^{\perp}$:

$$\mathcal{H} = \mathcal{H}_{\mathcal{P}} \oplus \mathcal{H}_{\mathcal{P}}^{\perp} \tag{2.51}$$

Toute fonction $f \in \mathcal{H}$ peut alors s'écrire de manière unique sous la forme $f = f_{\mathcal{H}_{\mathcal{P}}} + f_{\perp}$, avec $f_{\mathcal{H}_{\mathcal{P}}} \in \mathcal{H}_{\mathcal{P}}$ et $f_{\perp} \in \mathcal{H}_{\mathcal{P}}^{\perp}$. Remarquons que

$$\forall i \in \{1,...,n\},\ f(t_i) = \langle k(t_i,.), f_{\mathcal{H}_{\mathcal{P}}} \rangle + \underbrace{\langle k(t_i,.), f_{\perp} \rangle}_{0} = f_{\mathcal{H}_{\mathcal{P}}}(t_i) \tag{2.52}$$

En notant $\zeta(f)$ la fonctionnelle à minimiser dans l'éq. (2.49), on a ainsi que $\zeta(f)$ ne dépend de f que via sa norme $||f||_{\mathcal{H}}$. Or on obtient par le théorème de Pythagore que

$$||f||_{\mathcal{H}}^2 = ||f_{\mathcal{H}_{\mathcal{P}}}||_{\mathcal{H}}^2 + ||f_{\perp}||_{\mathcal{H}}^2, \tag{2.53}$$

si bien qu'à $f_{\mathcal{H}_{\mathcal{P}}}$ fixé, $||f||_{\mathcal{H}}$ est minimale lorsque $f_{\perp} = 0$, i.e. lorsque $f \in \mathcal{H}_{\mathcal{P}}$. Ψ étant supposée croissante en la norme $||f||_{\mathcal{H}}$, le problème admet donc nécessairement sa ou ses solution(s) dans $\mathcal{H}_{\mathcal{P}}$. □

Dans la pratique, le critère à minimiser est bien souvent de la forme

$$\Psi\left(f(t_1),...,f(t_n),||f||_{\mathcal{H}}\right) = c(f(t_1),...,f(t_n)) + \lambda||f||_{\mathcal{H}},$$

où c est un critère de *coût* quantifiant l'ajustement de f (le *fit*) à des observations, et $\lambda \in [0,+\infty[$ est un paramètre de régularisation permettant de contrôler un compromis entre ajustement aux données et régularité (i.e. faible valeur de la norme $||f||_{\mathcal{H}}$).

A ce sujet, l'application suivante aux splines de lissage permet de relier la problématique de l'approximation non-paramétrique évoquée à la section 2.1.3., l'espace de Sobolev-Hilbert explicité précedemment, et le théorème du représentant.

Application à l'approximation sur base de données bruitées : splines de lissage

Plaçons-nous dans un premier temps dans le cadre du *problème "restreint" des splines de lissage* (appelé "special spline smoothing problem" dans [Wah90]) : $y \in W_m$ est observée dans un bruit gaussien homoscédastique, en un nombre fini $n \in \mathbb{N}$ de points $\{t^1,...,t^n\}$ de l'ensemble $\mathcal{T} = [0,1]$. On note les observations bruitées comme suit :

$$\begin{cases} \forall i \in \{1,...,n\},\ y_\epsilon^i = y(t_i) + \varepsilon_i = L_i y + \varepsilon_i \\ \epsilon = (\epsilon_i)_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \end{cases} \tag{2.54}$$

où les L_i ($i \in \{1, ..., n\}$) sont les fonctionnelles d'évaluation aux points $\{t^1, ..., t^n\}$. Rappelons que le problème classique des splines de lissage revient à la minimisation du critère

$$\begin{aligned} \zeta(y) &= \frac{1}{n}\sum_{i=1}^{n}(y_\epsilon^i - y(t_i))^2 + \lambda\int_0^1 (y^{(m)}(u))^2 du \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_\epsilon^i - L_i y)^2 + \lambda\int_0^1 (y^{(m)}(u))^2 du \end{aligned} \tag{2.55}$$

Dans le *problème "général" des splines de lissage*, on s'intéresse à un ensemble $\mathcal{T}$ quelconque, une fonction y d'un RKHS $\mathcal{H}_k$ donné de fonctions de $\mathcal{T}$ dans $\mathbb{R}$, et L_i ($i \in \{1, ..., n\}$) des formes linéaires continues. Notons que les L_i peuvent tout aussi bien être des évaluations de dérivées ou d'intégrales que des évaluations ou encore des combinaisons linéaires d'évaluations (Cf. [Wah90], p.11). A l'instar de l'espace de Sobolev-Hilbert donné en exemple, $\mathcal{H}_R$ est supposé admettre une décomposition en somme directe :

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1 \tag{2.56}$$

où $\mathcal{H}_0$ est un espace de dimension finie $M \leq n$. Le critère à minimiser —généralisation de l'éq. (2.55)— est défini en faisant apparaître P_1, la projection orthogonale sur $\mathcal{H}_1$:

$$\frac{1}{n}\sum_{i=1}^{n}(y_\epsilon^i - \langle \eta_i, y\rangle)^2 + \lambda||P_1 y||_{\mathcal{H}}^2 \tag{2.57}$$

où les $\eta_i \in \mathcal{H}$ sont les représentants de Riesz des formes linéaires continues L_i ($i \in \{1, ..., n\}$). La variante donnée ci-dessous du théorème du représentant donne une solution explicite au problème de minimisation dans $\mathcal{H}$ du critère (2.57).

Théorème (Kimeldorf & Wahba, 1971) 1. *Soient $f_1, ..., f_M$ une base de l'espace nul ($\mathcal{H}_0$) de P_1 et $F \in \mathcal{M}_{n\times M}(\mathbb{R})$, définie par $F := (L_i f_\nu)_{(i,\nu)\in[1,n]\times[1,M]}$, une matrice de rang M. Alors y_λ, le minimiseur du critère (2.57) est donné par*

$$y_\lambda = \sum_{\nu=1}^{M} d_\nu f_\nu + \sum_{i=1}^{n} c_i \zeta_i \tag{2.58}$$

$$\text{où} \begin{cases} \zeta_i := P_1\eta_i, \\ K := (\langle \zeta_i, \zeta_j\rangle)_{(i,j)\in[1,n]^2}, \\ M := K + n\lambda I, \\ d = (d_1, ..., d_M)^T := (F^T M^{-1} F)^{-1} F^T M^{-1} \mathbf{Y}_\epsilon, \\ c = (c_1, ..., c_M)^T := M^{-1}(I - F(F^T M^{-1} F)^{-1} F^T M^{-1})\mathbf{Y}_\epsilon. \end{cases} \tag{2.59}$$

Démonstration. De la même manière que dans la preuve précédente du théorème du représentant, on peut affirmer l'existence d'un unique élément $\rho \in \mathcal{H}$ orthogonal aux $\{f_\nu, \nu \in \{1, ..., M\}\}$ et aux $\{\zeta_i, i \in \{1, ..., n\}\}$ tel que y_λ s'écrive sous la forme

$$y_\lambda = \sum_{\nu=1}^{M} d_\nu f_\nu + \sum_{i=1}^{n} c_i \zeta_i + \rho.$$

En se servant de la décomposition de l'identité en $I = P_0 + P_1$ (par somme directe) et du caractère auto-adjoint de P_0, on remarque que

$$\begin{aligned} \forall i \in [1, n],\ \langle \rho, \eta_i \rangle &= \langle \rho, P_0 \eta_i \rangle + \langle \rho, P_1 \eta_i \rangle \\ &= \langle \underbrace{P_0 \rho}_{0}, \eta_i \rangle + \underbrace{\langle \rho, \zeta_i \rangle}_{0} = 0 \end{aligned} \tag{2.60}$$

Le critère à minimiser peut ainsi se mettre sous la forme matricielle suivante :

$$\frac{1}{n} ||\mathbf{Y}_\epsilon - (Kc + Fd)||^2 + \lambda (c^T K c + ||\rho||^2)$$

et il vient que ρ est nécessairement nul. Il nous faut alors trouver les vecteurs c et d qui minimisent la quantité $\mathcal{C}(c, d) = \frac{1}{n} ||\mathbf{Y}_\epsilon - (Kc + Fd)||^2 + \lambda (c^T K c)$. Après vérification par calcul direct que sa matrice hessienne est positive et que le critère est donc convexe en (c, d), il reste à calculer c et d en annulant le gradient de $\mathcal{C}$. Nous dérivons $\mathcal{C}$ ci-dessous par rapport à c, puis par rapport à d :

$$\begin{aligned} \nabla_c \mathcal{C}(c, d) = 2\lambda c^T K - \frac{2}{n} \mathbf{Y}_\epsilon^T K + \frac{2}{n} c^T K^T K + \frac{2}{n} d^T F^T K = 0 \\ \Rightarrow (K + n\lambda I) c = (\mathbf{Y}_\epsilon - Fd) \\ \Rightarrow c = M^{-1} \mathbf{Y}_\epsilon - M^{-1} F d \end{aligned} \tag{2.61}$$

$$\begin{aligned} \nabla_d \mathcal{C}(c, d) = -\frac{2}{n} \mathbf{Y}_\epsilon^T F + \frac{2}{n} c^T K^T F + \frac{2}{n} d^T F^T F = 0 \\ \Rightarrow F^T F d = F^T \mathbf{Y}_\epsilon - F^T K c \end{aligned} \tag{2.62}$$

En revenant à l' éq. (2.61) que l'on multiplie à gauche par $F^T M$, il vient que $F^T M c + F^T F d = F^T \mathbf{Y}_\epsilon$ d'où l'on tire grâce à l'éq. (2.62) que $F^T M c = F^T K c$, ce qui implique $F^T c = 0$. On obtient finalement $d = (F^T M^{-1} F)^{-1} F^T M^{-1} \mathbf{Y}_\epsilon$ par multiplication à gauche de l'éq. (2.61) par F^T, et il suit par une dernière substitution de d dans l'éq. (2.61) que $c = M^{-1}(I - F(F^T M^{-1} F)^{-1} F^T M^{-1}) \mathbf{Y}_\epsilon$. □

2.3.3 Quelques liens avec l'approximation et la régression linéaires

Nous considérons ici à nouveau le problème d'approximation d'une fonction y par une combinaison linéaire de fonctions de base $\{f_1, ..., f_b\}$ linéairement indépendantes, tel que présenté dans la section 2.1.2. de ce chapitre. Ce problème peut être interprété en termes de RKHS, comme on peut le comprendre à la lecture de l'un des articles pionniers de N. Aronszajn [Aro50]. Munissons en effet l'espace

$$\mathcal{F} = \left\{ m_\alpha(\mathbf{x}) = \sum_{j=1}^{b} \alpha_j f_j(\mathbf{x}), \alpha \in \mathbb{R}^b \right\} \tag{2.63}$$

du produit scalaire défini par

$$\langle m_\alpha, m_{\alpha'} \rangle = \sum_{j=1}^{b} \alpha_j \alpha'_j, \tag{2.64}$$

faisant ainsi de la famille $\{f_1, ..., f_b\}$ une base orthonormale de $\mathcal{F}$. On définit k à partir des fonctions de base, en utilisant de nouveau la notation vectorielle $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), ..., f_b(\mathbf{x})]^T$:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{b} f_j(\mathbf{x}) f_j(\mathbf{x}') = \mathbf{f}(\mathbf{x})^T \mathbf{f}(\mathbf{x}) \tag{2.65}$$

On peut alors vérifier que $\mathcal{F}$ est un RKHS de noyau reproduisant k. Il est facile de voir que k est un noyau de type positif (comme somme de noyaux de type positif). La propriété de reproduction vient du fait que pour toute fonction $g = \sum_{i=1}^{b} \alpha_j f_j \in \mathcal{F}$ et tout point $\mathbf{x_0} \in \mathcal{T}$,

$$\begin{aligned} \langle g, k_{\mathbf{x_0}} \rangle &= \left\langle \sum_{i=1}^{b} \alpha_i f_i, \sum_{j=1}^{b} f_j(\mathbf{x}) f_j(\mathbf{x_0}) \right\rangle \\ &= \sum_{i=1}^{b} \sum_{j=1}^{b} \alpha_i f_j(\mathbf{x_0}) \langle f_i, f_j \rangle \\ &= \sum_{i=1}^{b} \sum_{j=1}^{b} \alpha_i f_j(\mathbf{x_0}) \delta_{i,j} = \sum_{j=1}^{b} \alpha_j f_j(\mathbf{x_0}) = g(\mathbf{x_0}) \end{aligned} \tag{2.66}$$

On peut voir dans un premier temps qu'une application du théorème du représentant lorsque l'on pose $\mathcal{H} = \mathcal{H}_0$ (i.e. $\mathcal{H}_1 = \{0\}$) permet de retrouver les résultats déjà obtenus par calcul direct dans la section 2.1 : la meilleure approximation de y dans $\mathcal{H}$ est donnée

par $\sum_{\nu=1}^{b} d_\nu f_\nu$ avec

$$\begin{aligned} d = (d_1, ..., d_b)^T &= (F^T M^{-1} F)^{-1} F^T M^{-1} \mathbf{Y}_\epsilon \\ &= (F^T (n\lambda I)^{-1} F)^{-1} F^T (n\lambda I)^{-1} \mathbf{Y}_\epsilon \\ &= (F^T F)^{-1} F^T \mathbf{Y}_\epsilon, \end{aligned} \tag{2.67}$$

c'est à dire que la solution du problème d'approximation (2.59) s'écrit

$$y_\lambda(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T (F^T F)^{-1} F^T \mathbf{Y}_\epsilon. \tag{2.68}$$

Ce résultat est à rapprocher des équations (2.5) et (2.19). Nous proposons de retrouver le même résultat (les équations « normales » de la régression linéaire) en adoptant une approche d'approximation dans le RKHS associé au noyau k défini dans l'eq. (2.65). Posons maintenant que $\mathcal{H} = \mathcal{H}_1$, i.e. que l' « espace nul »$\mathcal{H}_0$ se réduit à $\{0\}$. En reprenant les notations de la démonstration du théorème du représentant, on s'intéresse maintenant au programme de minimisation :

$$\min_{c \in \mathbb{R}^n} \left\{ \frac{1}{n} ||\mathbf{Y}_\epsilon - Kc||^2 + \lambda (c^T K c + ||\rho||^2) \right\}$$

On retrouve par le même argument que précédemment que $\rho = 0$. Si l'on considère à présent le cas particulier où $\lambda = 0$, on se ramène simplement à minimiser $||\mathbf{Y}_\epsilon - Kc||^2$. Comme K n'est pas toujours inversible (par exemple lorsque le nombre de fonctions de base est petit), ce problème n'admet pas nécessairement une solution unique. Il existe en revanche une unique solution de norme minimale, reposant sur la notion de "pseudo-inverse" de Moore-Penrose :

$$c = K^\dagger \mathbf{Y}_\epsilon \tag{2.69}$$

où l'on rappelle que $K^\dagger$ est définie comme étant l'unique solution du système

$$\begin{cases} K K^\dagger K = K \\ K^\dagger K K^\dagger = K^\dagger \end{cases} \tag{2.70}$$

Avant d'aller plus loin au sujet de l'eq. (2.69) et de l'application du théorème du représentant dans $\mathcal{F}$, donnons quelques formules de passage entre les notations de régression et celles utilisées dans le problème de RKHS associé :

$$\begin{cases} \forall i \in \{1, ..., n\},\ \zeta_i(t) = k(t^i, t) = \sum_{j=1}^{b} f_j(t^i) f_j(t) = (\mathbf{F}\mathbf{f}(\mathbf{t}))_i \Leftrightarrow \zeta(t) = \mathbf{F}\mathbf{f}(\mathbf{t}) \\ K = (\langle \zeta_i, \zeta_j \rangle)_{i,j} = \left(\sum_{k=1}^{b} f_k(t^i) f_k(t^i) \right)_{i,j} = \mathbf{F}\mathbf{F}^T \end{cases} \tag{2.71}$$

Remarquons enfin que la pseudo-inverse de Moore-Penrose de K peut s'écrire

$$K^{\dagger} = \left(\mathbf{F}\mathbf{F}^T\right)^{\dagger} = \mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T, \tag{2.72}$$

et la solution y_0 donnée par le théorème du représentant s'exprime alors comme suit

$$\begin{aligned} \forall t \in D,\ y_0(t) &= \langle \zeta(t), c \rangle = \left(\mathbf{F}\mathbf{f}(\mathbf{t})\right)^T \left(\mathbf{F}\mathbf{F}^T\right)^{\dagger} \mathbf{Y}_{\epsilon} \\ &= \mathbf{f}(\mathbf{t})^T \mathbf{F}^T \mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}_{\epsilon} \\ &= \mathbf{f}(\mathbf{t})^T\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}_{\epsilon} \end{aligned} \tag{2.73}$$

On retrouve bien l'équation solution du problème classique de régression linéaire. En bref, nous avons montré ici que les équations normales de l'approximation (ou de la régression) linéaire pouvaient être obtenues de deux manières bien distinctes en utilisant le théorème de Kimeldorf et Wahba : soit en se restreignant à un "espace nul" $\mathcal{H}_0$ engendré par les fonctions de base, soit se plaçant dans le RKHS $\mathcal{H}_1$ de noyau (2.65) et en cherchant parmi les meilleures approximations de y dans $\mathcal{H}_1$ celle de norme minimale.

Chapitre 3

Géostatistique et apprentissage par processus gaussiens

Nous nous intéressons depuis le chapitre précédent à différentes méthodes pour l'approximation d'une fonction déterministe y. Parmi celles que nous avons évoquées, certaines sont purement déterminististes (splines d'interpolation, approximations linéaires), quand d'autres font intervenir le hasard (régression(s)). Les techniques issues de la géostatistique se démarquent de ces dernières par une conception bien particulière de la notion de fonction inconnue : la fonction objectif y est modélisée comme réalisation d'un processus aléatoire spatial. A l'origine, les outils mathématiques utilisés en géostatistique étaient réservés à des applications en petite dimension algébrique (espace usuel à 2 ou 3 dimensions), avec des observations nombreuses. Ces outils sont aujourd'hui repris dans différentes branches des mathématiques, avec toute une terminologie moderne mais aussi de nouveaux problèmes, liés entre autres à la rareté des observations relativement au nombre de paramètres pris en compte. Ces conditions contraignent les modélisateurs à épurer les modèles et à faire des hypothèses davantage guidées par la nécessité pratique que par l'exactitude. Après avoir donné un aperçu des grandes notions de géostatistique classique, nous proposerons quelques développements et réinterprétations modernes du Krigeage, reposant sur les processus gaussiens, le conditionnement et l'inférence bayésienne. En guise de préliminaire, la première section est dédiée au rappel de quelques fondamentaux sur les variables et processus aléatoires. Les lecteurs désireux de se concentrer d'emblée sur les aspects de modélisation plus spécifiques à cette thèse pourront sauter les deux premières sections —synthèses dont la vocation est plutôt d'ordre pédagogique— et directement passer à la Section 3.3.

3.1 Processus aléatoires

3.1.1 Définitions et propriétés de base

On considère un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et un espace d'états mesuré $(E, \mathcal{E})$ (Cf. [LG06] pour des définitions complètes). Dans la majorité des cas abordés, $(E, \mathcal{E})$ sera $\mathbb{R}^d$ ($d \in \mathbb{N}^*$) muni de sa tribu borelienne, notée $\mathcal{B}(\mathbb{R}^d)$. On rencontrera aussi parfois des situations dans lesquelles E est un espace fonctionnel, mais on évitera alors autant que possible de rentrer dans des détails techniques (Cf. par exemple [RY91]).

Définitions et notations au sujet des variables aléatoires

On appelle *variable aléatoire* toute application mesurable $Y : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (E, \mathcal{E})$. Lorsque $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, on parle de *Vecteur aléatoire réel* (V.a.r.), ou encore de *variable aléatoire réelle* (v.a.r.) dans le cas particulier où $d = 1$. Les variables aléatoires complexes peuvent s'identifier à des V.a.r. de dimension 2. L'espérance d'une v.a.r. Y, notée $\mathbb{E}[Y]$, est l'intégrale de Y par rapport à la mesure de probabilité $\mathbb{P}$:

$$\mathbb{E}[Y] = \int_\Omega Y(\omega) d\mathbb{P}(\omega) \tag{3.1}$$

On appelle $\mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, souvent simplement noté $\mathcal{L}^1$, l'espace des v.a.r. Y telles que $\mathbb{E}[|Y|] < +\infty$. On note de même $\mathcal{L}^p$ ($p \in \mathbb{N} - \{0\}$) l'espace des v.a.r. Y telles que $\mathbb{E}[|Y|^p] < +\infty$, et les nombres $\mathbb{E}[|Y|^p]$ sont appelés les *moments d'ordre p* de la v.a.r. Y. La définition de l'espérance s'étend sans difficulté aux vecteurs aléatoires, en raisonnant composante par composante. En munissant $\mathbb{R}^d$ du produit scalaire euclidien $\langle , \rangle$ ainsi que de la norme $||.||$ canoniquement associée, on appelle $\mathcal{L}^p_d$ ($p \in \mathbb{R} - \{0\}$) l'ensemble des V.a.r. d-dimensionnels Y tels que $\mathbb{E}[||Y||^p] < +\infty$, ou de manière équivalente tels que chacune des v.a.r. composantes $Y_1,\ Y_2, \ldots,\ Y_d$ soit dans $\mathcal{L}^p_1 = \mathcal{L}^p$. Remarquons à titre d'exemple que l'espérance d'une variable aléatoire complexe $Y = Y_1 + \mathrm{i}Y_2$, où Y_1 et Y_2 sont des v.a.r. de $\mathcal{L}^1$, est donnée par $\mathbb{E}[Y] = \mathbb{E}[Y_1] + \mathrm{i}\mathbb{E}[Y_2]$. A ce propos, la *fonction caractéristique* Φ_Y d'un V.a.r. $Y \in \mathcal{L}^1_d$ est définie comme suit :

$$\Phi_Y : \mathbf{u} \in \mathbb{R}^d \longrightarrow \Phi_Y(\mathbf{u}) = \mathbb{E}[e^{\mathrm{i}<Y,\mathbf{u}>}] \in \mathbb{C} \tag{3.2}$$

Cette notion joue un rôle majeur en probabilités et statistiques, et permet notamment de faciliter l'étude des sommes de variables aléatoires indépendantes (preuves simples et élégantes du Théorème "Central Limit" (TCL) et de la Loi des Grands Nombres (LGN), introduction des lois stables, etc...Cf. [Lé65]).

Considérons maintenant une v.a. Y. La *mesure image de* $\mathbb{P}$ *par* Y, définie par

$$\mu_Y : A \in \mathcal{E} \longrightarrow \mathbb{P}(Y \in A) \tag{3.3}$$

est aussi appelée *loi de* Y. Dans le cas particulier où $E = \mathbb{R}^d$ ($d \in \mathbb{N}^*$), l'application

$$F_Y : (y_1, ..., y_d) \in \mathbb{R}^d \longrightarrow F_Y(\mathbf{y}) = \mu_Y(]-\infty, y_1[\times \ldots \times]-\infty, y_d[) \in [0,1] \tag{3.4}$$

est appelée *fonction de répartition de* Y. Si μ_Y est absolument continue par rapport à une mesure de référence μ, supposée σ-finie, μ_Y admet une densité f_Y (unique à une égalité μ-p.p. près, par le théorème de Radon-Nikodym, Cf. [LG06]) appelée *densité de probabilité de* Y *par rapport à* μ. S'il arrive que l'on parle —dans le cas de vecteurs aléatoires réels— d'une densité sans donner plus de précisions sur μ, on raisonne généralement par rapport à la mesure de Lebesgue. Lorsqu'une telle fonction de densité f_Y existe, on dit souvent que Y est un V.a.r. *absolument continu*, ou plus simplement *continu*. La loi μ_Y, ou encore la densité f_Y lorsqu'elle existe, contient non seulement toute l'information sur les lois respectives des v.a.r composantes $\{Y_j, j \in [1,d]\}$ (les *lois marginales*), mais aussi la structure de dépendance entre ces dernières (résumée par la notion de *copule*). En particulier, la matrice des covariances entre les couples de composantes d'un V.a.r. $Y \in \mathcal{L}^2_d$ est au coeur de nombreuses applications statistiques. On définit la covariance entre deux v.a.r. $Y_1, Y_2 \in \mathcal{L}^2$ comme l'espérance du produit des variables centrées [1] :

$$\begin{aligned} Cov[Y_1, Y_2] &= \mathbb{E}\left[(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2])\right] \\ &= \mathbb{E}\left[Y_1 Y_2\right] - \mathbb{E}[Y_1]\mathbb{E}[Y_2] \end{aligned} \tag{3.5}$$

Remarquons que la covariance ne résume pas toute la dépendance entre deux variables aléatoires : elle mesure seulement la dépendance linéaire. On a d'ailleurs que l'application $(Y_1, Y_2) \in \mathcal{L}^2 \times \mathcal{L}^2 \longrightarrow \mathbb{E}\left[Y_1 Y_2\right]$ est bilinéaire, symétrique, et constitue un produit scalaire sur $\mathcal{L}^2$ à condition de quotienter l'espace par la relation d'égalité presque sûre, ce qui permet d'affirmer que $(\mathbb{E}[Y^2] = 0) \Rightarrow (Y$ est nulle dans l'espace quotienté $L^2)$. On peut alors montrer (Cf. [LG06]) que L^2 muni du produit scalaire $\langle ., . \rangle = \mathbb{E}\left[. \times .\right]$ est un espace pré-hilbertien complet, et bénéficier ainsi des résultats et de la vision géométrique classiquement associés aux espaces de Hilbert. Par exemple, l'inégalité de Cauchy-Schwarz appliquée aux variables centrées $Y_1 - \mathbb{E}[Y_1]$ et $Y_2 - \mathbb{E}[Y_2]$ devient

$$\forall Y_1, Y_2 \in L^2,\ Cov[Y_1, Y_2] \leq \sqrt{Var[Y_1]}\sqrt{Var[Y_2]} \tag{3.6}$$

où Var est la forme quadratique associée à Cov. Si l'on considère maintenant un V.a.r. $Y \in L^2_d$ ($d \in \mathbb{N} - \{0\}$), on a vu que chacune de ses v.a.r. composantes est dans L^2, et on

[1] On montre sans difficulté que $(\mathbb{E}[|Y|^2] < +\infty) \Rightarrow (\mathbb{E}[|Y|] < +\infty)$, i.e. que $\mathcal{L}^2 \subset \mathcal{L}^1$

peut ainsi définir autant de covariances qu'il y a de couples de v.a.r. composantes. On appelle *matrice de covariance* de Y, notée $Cov[Y]$, la matrice de taille $d \times d$ de terme général $Cov[Y_i, Y_j]$. $Cov[Y]$ est clairement une matrice symétrique, et on a de plus que

$$\begin{aligned} \forall a = (a_1, ..., a_d)^T \in \mathbb{R}^d,\ a^T Cov[Y] a &= \sum_{i=1}^{d} \sum_{j=1}^{d} a_i a_j Cov[Y_i, Y_j] \\ &= Var\left[\sum_{i=1}^{d} a_i Y_i\right] \geq 0, \end{aligned} \tag{3.7}$$

i.e. que la matrice $Cov[Y]$ est semi-définie positive. En tant que matrice symétrique rélle semi-définie positive, une caractérisation classique —théorème spectral— nous permet d'affirmer que toute matrice de covariance est diagonalisable en base orthonormée, de valeurs propres positives ou nulles.

Les vecteurs gaussiens constituent une classe particulièrement intéressante de vecteurs aléatoires. Rappelons qu'une v.a.r. est dite *gaussienne standard*, ou encore de loi $\mathcal{N}(0,1)$, lorsqu'elle admet pour densité la fonction $f_{\mathcal{N}(0,1)}(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$. Une v.a.r. Y est dite *gaussienne* [2] lorsqu'il existe $m \in \mathbb{R}$, $\sigma \in [0, +\infty[$, et $N \sim \mathcal{N}(0,1)$ tels que $Y = m + \sigma N$. On a alors que $Y \in \mathcal{L}^2$ avec $\mathbb{E}[Y] = m$ et $Var[Y] = \sigma^2$, et on utilise la notation $Y \sim \mathcal{N}(m, \sigma^2)$. On appelle *vecteur gaussien standard* un V.a.r. $\mathbf{N}$ dont les composantes sont indépendantes et de loi $\mathcal{N}(0,1)$. La loi d'un tel vecteur est notée $\mathcal{N}(\mathbf{0}, I)$, où la matrice de covariance I représente la variance unité des marginales et l'orthogonalité des composantes distinctes, conséquence directe de l'hypothèse d'indépendance (On montre en annexe que les hypothèses d'orthogonalité et d'indépendance des composantes d'un vecteur gaussien sont en fait équivalentes). De manière plus générale, un V.a.r. $\mathbf{Y} \in \mathcal{L}^2_d$ est dit *gaussien* lorsqu'il existe $\mathbf{m} \in \mathbb{R}^d$, $A \in \mathcal{M}_d(\mathbb{R})$, et $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, I)$ tels que $\mathbf{Y} = \mathbf{m} + A\mathbf{N}$. Il vient alors par calcul direct que $Cov(\mathbf{Y}) = AA^T$, et on note $\mathbf{Y} \sim \mathcal{N}(\mathbf{m}, K)$, avec $K = AA^T$. La forme de la matrice de covariance obtenue est loin d'être anodine. Si l'on considère un vecteur gaussien quelconque $\mathbf{Y} \sim \mathcal{N}(\mathbf{m}, K)$ (K étant une matrice symétrique semi-définie positive arbitraire), il existe $A \in \mathcal{M}_d(\mathbb{R})$ tel que K sécrive $K = AA^T$. On obtient par exemple une décomposition unique en imposant à A d'être triangulaire inférieure (*décomposition de Cholesky*). Ce résultat permet entre autres de simuler tout vecteur gaussien à partir d'un vecteur gaussien standard $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, I)$, i.e. en se basant uniquement sur des v.a.r. gaussiennes standard. On a de plus que tout

[2] Nous incluons volontairement le cas $\sigma = 0$ de manière à considérer les variables aléatoires constantes comme des v.a. gaussiennes dégénérées.

vecteur $\mathbf{Y} \sim \mathcal{N}(\mathbf{m}, K)$ possède une fonction caractéristique de la forme suivante :

$$\forall \mathbf{u} \in \mathbb{R}^d,\ \Phi_{\mathbf{Y}}(\mathbf{u}) = \mathbb{E}[e^{\mathrm{i}<\mathbf{u},\mathbf{Y}>_{\mathbb{R}^d}}] = e^{\mathrm{i}<\mathbf{m},\mathbf{Y}>_{\mathbb{R}^d} - \frac{1}{2}<\mathbf{u},K\mathbf{u}>_{\mathbb{R}^d}} \tag{3.8}$$

et dans le cas où K est inversible on a la densité de probabilité dite *multigaussienne* :

$$\forall \mathbf{y} \in \mathbb{R}^d,\ f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{(\det K)}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{m})^T K^{-1}(\mathbf{y}-\mathbf{m})} \tag{3.9}$$

Parmi les propriétés incontournables des vecteurs gaussiens (Cf. section 12.2 en annexe), mentionnons en priorité le fait que lorsque l'on conditionne un sous-vecteur d'un vecteur gaussien par rapport à un autre de ses sous-vecteurs, la loi conditionnelle obtenue est toujours gaussienne, telle que précisé ci-dessous :

$$\boxed{\begin{array}{c} \text{Si } \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix}, \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \right) \\ \\ \text{alors } \mathbf{Y}_1|\mathbf{Y}_2 \sim \mathcal{N}\left(\mathbf{m}_1 + K_{12}K_{22}^{-1}(\mathbf{Y}_2 - \mathbf{m}_2), K_{11} - K_{12}K_{22}^{-1}K_{21}\right) \end{array}} \tag{3.10}$$

Nous reviendrons plus tard sur ce résultat important. Contentons-nous pour le moment de souligner que l'on retrouve bien a posteriori l'équivalence entre orthogonalité et indépendance dans le cas gaussien puisque si $\mathbb{E}[\mathbf{Y}_1\mathbf{Y}_2^T] = K_{12} = 0$ alors $\mathbf{Y}_1|\mathbf{Y}_2 \sim \mathcal{N}(\mathbf{m}_1, K_{11})$ i.e. $\mathbf{Y}_2$ n'apporte aucune information sur la loi de $\mathbf{Y}_1$. La notion générale de conditionnement sera rappelée et discutée plus en détail dans la section 3.1.3.

Définitions de base au sujet des processus aléatoires.

Revenons au cas général de variables aléatoires abstraites, définies comme applications mesurables $Y : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (E, \mathcal{E})$, où $(\Omega, \mathcal{A}, \mathbb{P})$ est un espace probabilisé et $(E, \mathcal{E})$ un espace mesurable quelconque. Un *processus stochastique* est défini dans le livre de Paul Lévy ([Lé65], chapitre II, p.27) comme étant « un procédé de définition d'une fonction aléatoire $Y(t)$ du temps t dans lequel le hasard intervient à chaque instant ». L'auteur dit plus loin que « l'on peut aussi généraliser la notion de processus stochastique en remplaçant t par un système de plusieurs variables ». Nous utiliserons une approche générique, semblable à celle développée dans l'annexe de ce même livre, écrite par M. Loève, dans laquelle une *fonction aléatoire* (ou encore un *processus aléatoire*) est définie comme étant une famille de variables aléatoires $\{Y(t)\}_{t\in D}$ où D est un ensemble quelconque « pourvu que les opérations sur les t y aient un sens », et telle que pour toute

partie finie $D_n = \{t_1, ..., t_n\} \subset D$ la loi de probabilité de $\{Y(t)\}_{t \in D_n} = \{Y(t_1), ..., Y(t_n)\}$ soit connue. De telles lois sont souvent appelées *distributions finies-dimensionnelles* du processus Y ; dire que l'on connaît ces lois revient à dire (Cf. Lévy, p.30) que l'on connaît pour tout $D_n = \{t_1, ..., t_n\} \subset A$ la fonction de répartition [3] :

$$U_n : (t_1, y_1, ..., t_n, y_n) \in D^{2n} \longrightarrow \mathbb{P}(Y(t_1) < y_1, ..., Y(t_n) < y_n) \tag{3.11}$$

Dans la pratique, il arrive bien souvent que l'on étudie un processus aléatoire sans parfaitement connaître sa loi de manière analytique, et en s'appuyant plutôt sur des propriétés plus faibles telles que des relations entre passé et futur dans le cas où D est monodimensionnel (égalités et inégalités pour les *processus de Markov* ou les *martingales* lorsque D représente le temps), ou encore des tendances et des relations de *corrélations spatiales* lorsque l'ensemble D est multidimensionnel (variables spatiales et spatio-temporelles en géostatistique, météorologie, etc.).

Remarquons qu'en définissant un processus aléatoire comme une famille de variables aléatoires $\{Y(t)\}_{t \in D}$, on ne fait pas apparaître l'espace Ω explicitement. Il est fréquent de rencontrer les notations $\{Y(t;\omega),\ t \in D,\ \omega \in \Omega\}$ ou encore $\{Y_t(\omega),\ t \in D,\ \omega \in \Omega\}$, et nous les utiliserons par la suite. On appelle *trajectoire* ou *réalisation* d'un processus aléatoire toute fonction $Y(.;\omega) : t \in D \longrightarrow Y(t;\omega) \in E$ associée à un certain événement $\omega \in \Omega$. On peut ainsi voir un processus aléatoire comme une application de Ω dans E^D, l'espace des applications de D dans E. L'étude des propriétés des trajectoires de processus aléatoires est un sujet délicat que nous ne ferons qu'effleurer ici (Cf. [Lé65, RY91], en particulier en ce qui concerne les travaux de Kolmogorov sur ce sujet).

Exemples :

1. *Le mouvement Brownien (MB) est un processus stochastique* $\{B_t\}_{t \in \mathbb{R}^+}$ *tel que* $B_0 = 0$ *(départ à l'origine),* $\forall t \in \mathbb{R}^+,\ B_t \sim \mathcal{N}(0, t)$ *(marginales gaussiennes), et* $\forall\ t_1, t_2, t_3, t_4 \in \mathbb{R}^+$ *t.q.* $t_1 < t_2 < t_3 < t_4$, $B_{t_4} - B_{t_3}$ *indépendant de* $B_{t_2} - B_{t_1}$ *(accroissements indépendants). Il joue un rôle clef en théorie des processus aléatoires et en calcul stochastique. Il possède de nombreuses propriétés remarquables, parmi lesquelles la propriété de Markov : si* $0 \leq s < t$, *la loi de* B_t *ne dépend des valeurs* $\{B_u\}_{u \leq s}$ *prises par le processus* B *avant le temps* s *que via la valeur* B_s.
2. *On appelle souvent champ aléatoire ou encore processus aléatoire spatial un processus* $\{Y_x\}_{x \in D}$ *défini sur un ensemble multidimensionnel* $D \subset \mathbb{R}^d$ $(d \in \mathbb{N} \backslash \{0, 1\})$. *De*

[3] Mentionnons sans entrer dans plus de détails que pour qu'une famille de fonctions U_n définisse une famille de fonctions de répartitions d'un processus aléatoire, elles doivent vérifier certaines *conditions de compatibilité*, présentées et expliquées dans ([Lé65], pp. 31-32).

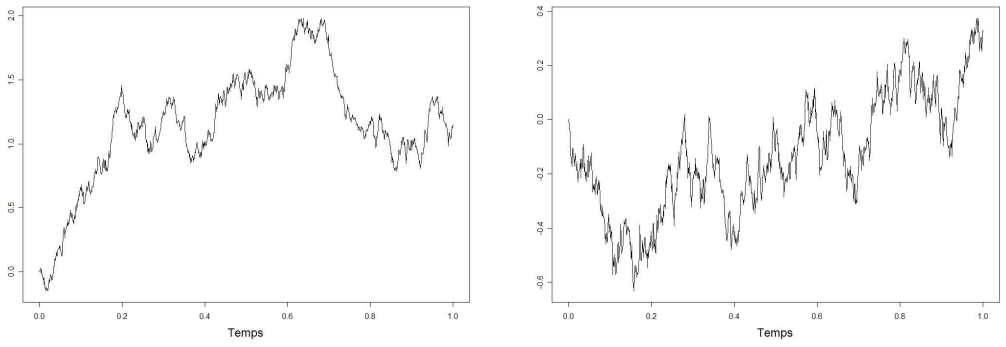


FIG. 3.1 – Deux réalisations simulées du MB sur $[0, 1]$.

tels objets apparaissent de manière assez évidente en sciences de la terre (concentrations diverses dans un sous-sol, champs de pression ou de température), et peuvent être utilisés de manière très générale pour décrire des phénomènes jouissant d'une certaine régularité spatiale ou spatio-temporelle (météorologie, océanographie, imagerie médicale).

Dans la suite de ce mémoire, nous porterons essentiellement notre attention sur des processus spatiaux tels que précédemment évoqués, en substituant à l'espace tri-dimensionnel usuel des espaces de paramètres de dimension algébrique quelconque. Nous aurons de plus besoin de recourir à des hypothèses sur la *stationnarité* des processus considérés, ou encore d'autres proprits spcifiques sur leur loi de probabilité. Nous donnons ci-dessous quelques définitions supplémentaires qui pourront nous être bien utiles.

Définitions : Stationnarité(s).

Suivant ([Lé65], p.91), un processus aléatoire $\{X(t)\}_{t\in D}$ est dit (fortement) *stationnaire* si quelque soit $n \in \mathbb{N}\backslash\{0\}$ et $t_1, ..., t_n \in D$, la loi de $(X(t_1), ..., X(t_n))$ ne dépend que des différences $t_i - t_j$ $(i, j \in [1, n])$. Une formulation équivalente est de dire que la loi de $(X(t_1 + h), ..., X(t_n + h))$ est la même que celle de $(X(t_1), ..., X(t_n))$, quelque soit h.

La stationnarité apparaît ainsi comme une invariance de la loi d'un processus aléatoire par changement d'origine de l'espace D. En particulier, la loi de $X(t)$ (X évalué au point t) ne dépend pas de t. Ainsi, s'ils existent, les moments de $X(t)$ sont constants sur D. De même, si elle existe, la covariance entre $X(t_1)$ et $X(t_2)$ $(t_1, t_2 \in D)$ ne dépend que de $t_1 - t_2$. La notion de stationnarité à l'ordre p permet de décrire d'intéressantes classes de processus aléatoires, non nécessairement stationnaires, mais pour lesquels les moments

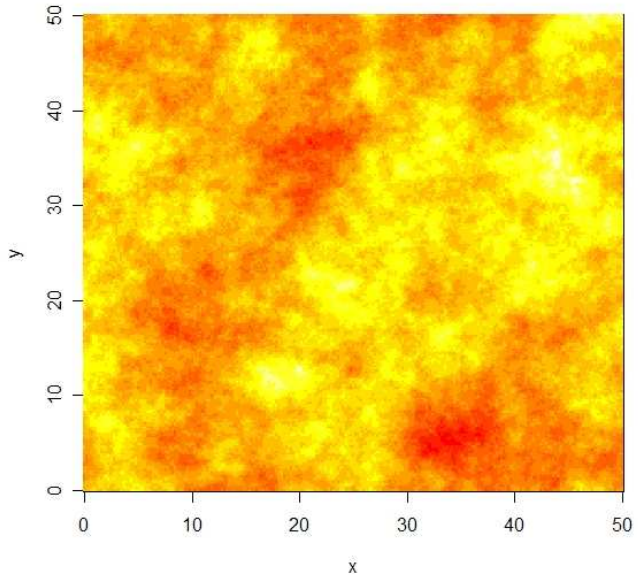


FIG. 3.2 – Réalisation d'un champ aléatoire (gaussien) de covariance exponentielle de portée 5 et de variance 10. Cette réalisation a été obtenue par simulation en utilisant le package R *Random Fields*.

jusqu'à l'ordre p existent et possèdent la propriété d'invariance par translation :

$\{X(t)\}_{t\in D}$ est dit *stationnaire à l'ordre* p ($p \in \mathbb{N}\backslash\{0\}$) si quelque soit $n \in \mathbb{N}\backslash\{0\}$ et $t_1, ..., t_n \in D$ les moments jusqu'à l'ordre p de la loi de $(X(t_1), ..., X(t_n))$ ne dépendent que des différences $t_i - t_j$ $(i, j \in [1, n])$.

Remarquons qu'un processus aléatoire stationnaire n'admet pas nécessairement de moments d'ordre p finis. En revanche, il est clair qu'un processus stationnaire dont les moments d'ordre p sont finis est en particulier un processus stationnaire à l'ordre p. Revenons sur nos deux exemples afin d'illustrer quelques notions de stationnarité :

Retour sur les exemples :

1. *Le mouvement Brownien* $\{B_t\}_{t\in\mathbb{R}^+}$ *n'est pas stationnaire. On a en effet que* $\forall t \in \mathbb{R}^+$, $B_t \sim \mathcal{N}(0, t)$, *ce qui entraîne que* $Var[B_t] = t$ *et donc que la loi de* B_t *dépend de* t. *Par contre, l'espérance de* B_t *est finie puisque* $\forall t \in \mathbb{R}^+$, $\mathbb{E}[B_t] = 0$, *et on remarque que le MB est stationnaire à l'ordre* 1.

2. *Lorsque l'on étudie un processus aléatoire spatial* $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ *en sciences de la terre, il est assez courant de supposer* Y *stationnaire à l'ordre* 2 *(on rencontre parfois la terminlogie "faiblement stationnaire" pour de tels processus). La fonction d'autocovariance* $C : (\mathbf{x}^1, \mathbf{x}^2) \in D^2 \longrightarrow C(\mathbf{x}^1, \mathbf{x}^2) = cov[Y_{\mathbf{x}^1}, Y_{\mathbf{x}^2}]$ *ne dépend alors que de la différence* $\mathbf{h} = \mathbf{x}^1 - \mathbf{x}^2$*, et est appelée "autocovariance stationnaire".*

3.1.2 Processus aléatoires gaussiens (PG)

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ la droite réelle munie de la tribu des boréliens, et $D \subset \mathbb{R}^d$ $(d \in \mathbb{N}\backslash\{0\})$. $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ est appelé *processus aléatoire gaussien* (réel) ou encore *champ gaussien* lorsque toutes ses loi finies-dimensionnelles sont gaussiennes, i.e. $\forall n \in \mathbb{N}\backslash\{0\},\ \forall \mathbf{x}^1, ..., \mathbf{x}^n \in D$ le V.a.r. $(Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n})$ est gaussien.

On a en particulier que quelque soit $\mathbf{x} \in D$, la v.a.r. $Y_{\mathbf{x}}$ suit une loi gaussienne. Il découle de cette définition que tout processus gaussien $Y_{\mathbf{x}}$ admet des moments d'ordre 1 et 2 finis. Notons-les comme suit, à l'instar des auteurs du récent livre [RW06] :

$$\begin{cases} m(\mathbf{x}) := \mathbb{E}(Y_{\mathbf{x}}) \\ k(\mathbf{x}, \mathbf{x}') := cov[Y_{\mathbf{x}}, Y'_{\mathbf{x}}] = \mathbb{E}[Y_{\mathbf{x}} - m(\mathbf{x}), Y_{\mathbf{x}'} - m(\mathbf{x}')] \end{cases} \tag{3.12}$$

On a de plus que les fonctions m et k déterminent complètement la loi de $Y_{\mathbf{x}}$, et on utilise parfois la notation synthétique

$$\boxed{Y_{\mathbf{x}} \sim \mathcal{PG}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))} \tag{3.13}$$

pour signifier que $Y_{\mathbf{x}}$ est un processus gaussien de *tendance* $m(\mathbf{x})$ (moment d'ordre 1) et de fonction d'autocovariance $k(\mathbf{x}, \mathbf{x}')$, encore appelée *noyau de covariance*. Dire que le processus gaussien $Y_{\mathbf{x}}$ est stationnaire à l'ordre 1 signifie que m est constante, i.e. $\exists \mu \in \mathbb{R} :\ \forall \mathbf{x} \in D,\ m(\mathbf{x}) = \mu$. On sera souvent amener à travailler avec de tels processus, assez souvent d'ailleurs avec des PG de moyenne nulle.

Pour un processus stationnaire à l'ordre 1, la stationnarité à l'ordre deux s'obtient par l'existence et la stationnarité du noyau de covariance k. Dans le cas d'un processus gaussien $Y_{\mathbf{x}}$, si $m(\mathbf{x}) = \mu \in \mathbb{R}$ et $k(\mathbf{x}, \mathbf{x}')$ ne dépend que de $\mathbf{h} = \mathbf{x} - \mathbf{x}'$, alors $Y_{\mathbf{x}}$ est de plus fortement stationnaire (Cf [Lé65] ou [Abr97]). Cela illustre en partie à quel point le noyau de covariance k joue un rôle prépondérant dans la définition d'un PG. Nous reviendrons sur l'importance du noyau de covariance dans la suite. Donnons tout d'abord quelques illustrations de la notion de PG sur la base de nos deux exemples.

1. *On a vu que le mouvement Brownien* $\{B_t\}_{t\in\mathbb{R}^+}$ *était un processus aléatoire stationnaire à l'ordre* 1*, à accroissements indépendants, et que* $\forall t \in \mathbb{R}^+,\ B_t \sim \mathcal{N}(0,t)$*. Montrons que* $\{B_t\}_{t\in\mathbb{R}^+}$ *est un PG et exhibons son noyau de covariance. Soit* $n \in \mathbb{N}\backslash\{0\}$*,* $t^0 = 0$*, et* $t^1, ..., t^n \in \mathbb{R}^+$*. On déduit des propriétés ci-dessus que* $(B_{t^0}, B_{t^1} - B_{t^0}, ..., B_{t^n} - B_{t^{n-1}})$ *est un vecteur gaussien, puisqu'il a des composantes gaussiennes indépendantes. Il reste a remarquer que* $(B_{t^0}, B_{t^1}, ..., B_{t^n})$ *s'obtient par une transformation linéaire de* $(B_{t^0}, B_{t^1} - B_{t^0}, ..., B_{t^n} - B_{t^{n-1}})$ *et il vient que* $(B_{t^0}, B_{t^1}, ..., B_{t^n})$*, et a fortiori* $(B_{t^1}, ..., B_{t^n})$*, est un vecteur gaussien. Soit maintenant* $t, t' \in \mathbb{R}^+$ *tels que* $t \leq t'$*. On a*

$$\begin{aligned} k_B(t,t') :&= cov[B_t, B_{t'}] = \mathbb{E}[B_t B_{t'}] \\ &= \mathbb{E}[B_t^2] + \mathbb{E}[B_t(B_{t'} - B_t)] = t + 0 = \min(t,t') \end{aligned} \tag{3.14}$$

Le mouvement Brownien est ainsi un processus gaussien centré non-stationnaire, de noyau $k_B(t,t') = \min(t,t')$ *(pour la non-stationnarité, on aurait pu se contenter du fait que* $Var[B_t] = t$*, déjà vu précédemment et confirmé par 3.14).*

2. *On peut définir un PG spatial* $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ *remarquable en lui imposant d'être de moyenne nulle et de noyau de covariance* k_Y *défini par* $\forall \mathbf{x}, \mathbf{x}' \in D,\ k_Y(\mathbf{x}, \mathbf{x}') = \sigma^2 e^{-\left(\frac{||\mathbf{x}-\mathbf{x}'||}{p}\right)^2}$ *(*$\sigma \in \mathbb{R}^+, p \in]0, +\infty[$*). La fonction d'autocovariance* k_Y*, dite "gaussienne isotrope", est stationnaire et le PG* $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ *l'est donc lui-aussi.* σ^2 *est la variance du processus et* p *est un paramètre de longueur de corrélation.*

3.1.3 Conditionnement des V.a.r. et processus aléatoires

Le conditionnement est un sujet central en probabilités et constitue le fondement théorique de nombreuses méthodes (prédiction, estimation, etc.) utilisées dans les applications, concernant des domaines tels que la médecine (imagerie), l'économie (séries temporelles), ou les sciences de la terre (géostatistique). Le conditionnement d'un processus aléatoire $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ —de loi connue— consiste à prendre en compte certaines informations disponibles à son sujet ou au sujet de variables dépendantes de Y pour affiner la connaissance que l'on a de son comportement : cela peut signifier s'intéresser aux trajectoires d'un mouvement Brownien sachant qu'il n'a pas dépassé un seuil donné au cours d'un certain intervalle de temps, ou encore établir une carte de prédictions météorologiques pour le lendemain en partant des températures du jour et éventuellement d'un historique. Nous aurons essentiellement besoin dans le cadre de cette thèse du conditionnement [4]

[4] Cf. ([LG06], chap. 11) pour une présentation mathématique précise des notions d'espérance conditionnelle et de conditionnement en termes de théorie de la mesure.

d'un processus aléatoire $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ par rapport à la donnée d'un nombre fini d'observations ponctuelles $\{Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}\}$. Avant d'en dire un peu plus sur le conditionnement des processus aléatoires, nous rappelons ci-dessous quelques définitions et propriétés élémentaires au sujet du conditionnement des variables aléatoires.

Rappels de conditionnement (dans $L^2(\mathbb{P})$)

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et $A, B \in \mathcal{A}$ deux événements. On rappelle que $\mathbb{P}(A \cap B)$ est la probabilité que A et B se réalisent simultanément. Après observation de l'évènement B, la probabilité de l'évènement A, $\mathbb{P}(A)$, se trouve modifiée en *probabilité conditionnelle* $\mathbb{P}(A|B)$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3.15}$$

Ce résultat élémentaire est le point de départ du conditionnement, et permet de définir la notion de loi conditionnelle. En utilisant le fait que $\mathbb{P}(A|B) = \mathbb{E}[\mathbb{1}_A|B]$, où $\mathbb{1}_A$ est la v.a.r. qui vaut 1 si A est réalisée et 0 sinon, on se ramène classiquement à l'étude de l'*espérance conditionnelle* comme nous le faisons ci-dessous.

Considérons deux variables aléatoires $Y_1, Y_2 \in L^2(\mathbb{P})$. Suivant la présentation proposée par Van der Vaart dans son chapitre lumineux sur la projection [VdV98], $\mathbb{E}[Y_1|Y_2]$ est la v.a.r. image $g^*(Y_2)$ de Y_2 par une fonction borélienne $g^*(.)$ qui minimise

$$\mathbb{E}\left[(Y_1 - g(Y_2))^2\right] \tag{3.16}$$

parmi toutes les fonctions mesurables g. Cela revient à dire que $\mathbb{E}[Y_1|Y_2]$ est la projection de Y_1 sur l'ensemble des variables aléatoires images de Y_2 par des fonctions boréliennes, au sens de la norme $||.|| : X \in L^2(\mathbb{P}) \longrightarrow ||X|| = \mathbb{E}[X^2] \in [0, +\infty[$ induite par le produit scalaire usuel de $L^2(\mathbb{P})$. Il s'en suit que $\mathbb{E}[Y_1|Y_2]$ est l'unique fonction de Y_2 (à une égalité $\mathbb{P}_{Y_2}$-p.p. près) qui satisfasse pour tout fonction mesurable g la relation d'orthogonalité

$$\mathbb{E}\left[(Y_1 - \mathbb{E}[Y_1|Y_2])g(Y_2)\right] = 0 \tag{3.17}$$

Bon nombre de propriétés classiques sur l'espérance conditionnelle découlent de 3.17. A ce propos, les trois propriétés ci-dessous sont fondamentales dès que l'on utilise le conditionnement, comme nous le ferons en particulier dans la section 3.3 pour de l'interprétation bayésienne du Krigeage puis au chapitre 8 au sujet des mélanges de modèles.

Formule de l'espérance totale. *Soient* $Y_1, Y_2 \in L^2(\mathbb{P})$.

$$\mathbb{E}[\mathbb{E}[Y_1|Y_2]] = \mathbb{E}[Y_1] \tag{3.18}$$

Démonstration. Découle directement de l'application de 3.17 avec $g \equiv 1$. □

Formule de la covariance totale. *Soient* $Y_1, Y_2, Y_3 \in L^2(\mathbb{P})$.

$$\boxed{Cov[Y_1, Y_2] = \mathbb{E}[Cov[Y_1, Y_2|Y_3]] + Cov[\mathbb{E}[Y_1|Y_3], \mathbb{E}[Y_2|Y_3]]} \tag{3.19}$$

Démonstration.

$$\begin{aligned}
Cov[Y_1, Y_2] &= \mathbb{E}[(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2])] \\
&= \mathbb{E}[\mathbb{E}[(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2])|Y_3]] \\
&= \mathbb{E}\left[\mathbb{E}[(Y_1 - \mathbb{E}[Y_1|Y_3] + \mathbb{E}[Y_1|Y_3] - \mathbb{E}[Y_1]) \times (Y_2 - \mathbb{E}[Y_2|Y_3] + \mathbb{E}[Y_2|Y_3] - \mathbb{E}[Y_2])|Y_3]\right] \\
&\quad \text{on utilise le fait que } \mathbb{E}[(Y_i - \mathbb{E}[Y_i|Y_3])|Y_3] = 0 \text{ pour } i \in \{1,2\} \\
&= \mathbb{E}\left[\mathbb{E}[(Y_1 - \mathbb{E}[Y_1|Y_3])(Y_2 - \mathbb{E}[Y_2|Y_3])|Y_3]\right] + \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y_1|Y_3] - \mathbb{E}[Y_1])(\mathbb{E}[Y_2|Y_3] - \mathbb{E}[Y_2])|Y_3]] \\
&\quad \text{on se sert de la définition de la covariance conditionnelle et de 3.18} \\
&= \mathbb{E}[Cov[Y_1, Y_2|Y_3]] + \mathbb{E}[(\mathbb{E}[Y_1|Y_3] - \mathbb{E}[Y_1])(\mathbb{E}[Y_2|Y_3] - \mathbb{E}[Y_2])] \\
&\quad \text{on se sert de } \mathbb{E}[\mathbb{E}[X_i|Y_3]] = \mathbb{E}[X_i] \text{ pour } i \in \{1,2\} \\
&= \mathbb{E}[Cov[Y_1, Y_2|Y_3]] + Cov[\mathbb{E}[Y_1|Y_3], \mathbb{E}[Y_2|Y_3]]
\end{aligned}$$

□

Cette propriété admet enfin comme corollaire la formule de la variance totale :

Formule de la variance totale. *Soient* $Y_1, Y_2 \in \mathcal{L}^2(\mathbb{P})$.

$$Var[Y_1] = \mathbb{E}[Var[Y_1|Y_2]] + Var[\mathbb{E}[Y_1|Y_2]] \tag{3.20}$$

Lorsque l'on s'intéresse à la valeur prise par un processus Y en un nouveau plan d'expériences, disons en un point $\mathbf{x}^{new}$ comme cela sera souvent le cas dans les sections suivantes, on se ramène à l'étude du vecteur aléatoire $(Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}, Y_{\mathbf{x}^{new}})$. Notons pour compléter ce qui vient d'être dit concernant le conditionnement d'une variable aléatoire par une autre, que lorsque le V.a.r. considéré est dans $L^2(\Omega, \mathcal{A}, \mathbb{P})$, $\mathbb{E}[Y_{\mathbf{x}^{new}}|Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}]$ n'est autre que la meilleure approximation (au sens « L^2 ») de $Y_{\mathbf{x}^{new}}$ dans $L^2(\Omega, \mathcal{B}, \mathbb{P})$, où $\mathcal{B} = \sigma(Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n})$ est la *tribu engendrée par les* $\{Y_{\mathbf{x}^i}, i \in [1, n]\}$.

Il est généralement loin d'être évident d'avoir accès à la loi conditionnelle de $Y_{\mathbf{x}^{new}}$ sachant $Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}$, voire même à $\mathbb{E}[Y_{\mathbf{x}^{new}}|Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}]$. Dans les cas où l'on ne sait pas calculer l'espérance conditionnelle $\mathbb{E}[Y_{\mathbf{x}^{new}}|Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}]$, on utilise souvent par commodité l'espérance conditionnelle linéaire $\mathbb{E}_L[Y_{\mathbf{x}^{new}}|Y_{\mathbf{x}^1}, ..., Y_{\mathbf{x}^n}]$ présentée ci-dessous.

L'espérance conditionnelle linéaire

L'espérance conditionnelle linéaire est définie comme la meilleure approximation de $Y_{\mathbf{x}^{new}}$ par une combinaison linéaire des $Y_{\mathbf{x}^i}$:

$$\begin{aligned} &\mathbb{E}_L[Y_{new}|Y_1,\dots,Y_n] = \sum_{j=1}^{n} \lambda_j Y_j \\ &\text{où } \lambda = argmin_\lambda \mathbb{E}\left[\left(Y_{new} - \sum_{j=1}^{n} \lambda_j Y_j\right)^2\right] \end{aligned} \tag{3.21}$$

En reprenant les notations précédentes, $\mathbb{E}_L[Y_{new}|Y_1,\dots,Y_n] = \sum_{j=1}^{n} \lambda_j Y_j$ apparaît ainsi dans $L^2(\Omega, \mathcal{A}, \mathbb{P})$ comme la projection orthogonale de Y_{new} sur l'espace vectoriel des variables aléatoires combinaisons linéaires des $\{Y_i, i \in \{1, ..., n\}\}$. Notons que la qualité de l'espérance conditionnelle linéaire comme approximation de l'espérance conditionnelle dépend fortement de la loi du V.a.r. considéré. Le cas gaussien est particulièrement favorable, comme on le résume le résultat suivant (théorème 11.4.3, [LG06]) :

Théorème. *Soit $(Y_1,\dots,Y_n,Y_{new})$ un vecteur gaussien centré. L'espérance conditionnelle $\mathbb{E}[Y_{new}|Y_1,\dots,Y_n]$ coïncide alors avec la projection orthogonale de Y_{new} sur l'espace vectoriel engendré par $Y_1,\dots,Y_n$, c'est-à-dire avec l'espérance conditionnelle linéaire $\mathbb{E}_L[Y_{new}|Y_1,\dots,Y_n]$. Il existe donc des réels $\lambda_1,\dots,\lambda_n$ tels que*

$$\mathbb{E}[Y_{new}|Y_1,\dots,Y_n] = \sum_{j=1}^{n} \lambda_j Y_j \tag{3.22}$$

De plus, pour toute fonction borélienne $h : \mathbb{R} \longrightarrow \mathbb{R}^+$,

$$\begin{aligned} &\mathbb{E}[h(Y_{new})|Y_1,\dots,Y_n] = \int_{\mathbb{R}} h(x) f_{\mathcal{N}(\sum_{j=1}^{n} \lambda_j Y_j, \sigma^2)}(x)dx \\ &\text{où } \sigma^2 = \mathbb{E}\left[\left(Y_{new} - \sum_{j=1}^{n} \lambda_j Y_j\right)^2\right] \end{aligned} \tag{3.23}$$

Il apparaît ainsi que *l'espérance conditionnelle et l'espérance conditionnelle linéaire coincident dans le cas gaussien.* De plus, la loi conditionnelle d'une composante d'un vecteur gaussien sachant les autres composantes est une gaussienne. La loi conditionnelle multivariables donnée par 3.10 en est une généralisation directe. Nous utiliserons ces résultats à maintes reprises dans la suite de ce mémoire, en particulier en ce qui concerne le Krigeage et la géostatistique linéaire.

Conditionnement des processus gaussiens

Lorsque l'on conditionne un processus aléatoire $\{Y_{\mathbf{x}}\}_{\mathbf{x}\in D}$ par rapport à un événement $A \in \mathcal{A}$, on obtient une loi conditionnelle définie par les lois $\mathbb{P}\left(Y_{\mathbf{x}^1} \in I_1, ..., Y_{\mathbf{x}^p} \in I_p | A\right)$, $p \in \mathbb{N}^*$, $(\mathbf{x}^1, ..., \mathbf{x}^p) \in D^p$, $I_1, ..., I_p \in \mathcal{E}$, appelées *lois finies-dimensionnelles*. Dans le cas où Y est gaussien et A est un événement de type $(Y_{\mathbf{x}'^1} \in I'_1, ..., Y_{\mathbf{x}'^{p'}} \in I'_{p'})$ ($p' \in \mathbb{N}^*$, $(\mathbf{x}'^1, ..., \mathbf{x}'^{p'}) \in D^{p'}$, $I'_1, ..., I_{p'} \in \mathcal{E}$), les lois des V.a.r $(Y_{\mathbf{x}^1} \in I_1, ..., Y_{\mathbf{x}^p} \in I_p | A)$ sont gaussiennes, et le le « processus aléatoire conditionnel »$\{Y_{\mathbf{x}}|A\}_{\mathbf{x}\in D}$ est donc un PG. Le processus de *pont brownien* constitue un exemple fondamental de processus gaussien conditionné par un événement de type « passage en un point ».

L'exemple du pont brownien : on appelle loi du pont brownien la loi d'un mouvement brownien B_t sur $[0, 1]$ contraint à passer par par 0 en $t = 1$ (Cf. fig. 3.1.3). Nous allons calculer la loi du pont brownien en conditionnant B_t par rapport à l'évènement $B_1 = 0$. Notons $\mathbb{Y}_1 = (B_{t_1}, \ldots, B_{t_k})$, où $k \in \mathbb{N}^*$, $0 \leq t_1 \leq \ldots \leq t_k \leq 1$, $\mathbb{Y}_2 = B_1$, et $\mathbb{Y} = (\mathbb{Y}_1, \mathbb{Y}_2) = (B_{t_1}, \ldots, B_{t_k}, B_1)$. $\mathbb{Y}$ a pour de matrice de covariance

$$K_{\mathbf{Y}} = \begin{pmatrix} \min(t_1,t_1) & \min(t_1,t_2) & \ldots & \min(t_1,1) \\ \min(t_2,t_1) & \min(t_2,t_2) & \ldots & \min(t_2,1) \\ \vdots & \vdots & \vdots & \vdots \\ \min(t_k,t_1) & \min(t_k,t_2) & \ldots & \min(1,1) \end{pmatrix} = \begin{pmatrix} t_1 & t_1 & \ldots & t_1 \\ t_1 & t_2 & \ldots & t_2 \\ \vdots & \vdots & \vdots & \vdots \\ t_1 & t_2 & \ldots & 1 \end{pmatrix} \tag{3.24}$$

Selon la loi donnée par 3.10, on a l'espérance conditionnelle :

$$\mathbb{E}\left[\begin{pmatrix} B_{t_1} \\ \vdots \\ B_{t_k} \end{pmatrix} | B_1\right] = \mathbb{E}\left[\begin{pmatrix} B_{t_1} \\ \vdots \\ B_{t_k} \end{pmatrix}\right] + \left(\frac{B_1 - \mathbb{E}[B_1]}{1}\right) \times \begin{pmatrix} B_{t_1} \\ \vdots \\ B_{t_k} \end{pmatrix} \tag{3.25}$$

Et comme $\mathbb{E}\left[\begin{pmatrix} B_{t_1} \\ \vdots \\ B_{t_k} \end{pmatrix}\right] = 0$ et $\mathbb{E}[B_1] = 0$, il vient que

$$\mathbb{E}\left[\begin{pmatrix} B_{t_1} \\ \vdots \\ B_{t_k} \end{pmatrix} \mid B_1 = 0\right] = 0 \tag{3.26}$$

et on voit que le pont brownien est centré.

Le calcul de la matrice de covariance conditionnelle donne :

$$
\begin{aligned}
Var\left[\begin{pmatrix} B_{t_1} \\ \vdots \\ B_{t_k} \end{pmatrix} | B_1\right] &= \begin{pmatrix} t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & \dots & t_2 \\ \vdots & \vdots & \vdots & \vdots \\ t_1 & t_2 & \dots & t_k \end{pmatrix} - \begin{pmatrix} t_1 \\ \vdots \\ t_k \end{pmatrix} \times (1)^{-1} \times \begin{pmatrix} t_1 & \dots & t_k \end{pmatrix} \\
&= \begin{pmatrix} t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & \dots & t_2 \\ \vdots & \vdots & \vdots & \vdots \\ t_1 & t_2 & \dots & t_k \end{pmatrix} - \begin{pmatrix} t_1^2 & t_1 t_2 & \dots & t_1 t_k \\ t_2 t_1 & t_2^2 & \dots & t_2 t_k \\ \vdots & \vdots & \vdots & \vdots \\ t_k t_1 & t_k t_2 & \dots & t_k^2 \end{pmatrix} \\
&= \begin{pmatrix} t_1(1-t_1) & t_1(1-t_2) & \dots & t_1(1-t_k) \\ t_2(1-t_1) & t_2(1-t_2) & \dots & t_2(1-t_k) \\ \vdots & \vdots & \vdots & \vdots \\ t_k(1-t_1) & t_k(1-t_2) & \dots & t_k(1-t_k) \end{pmatrix}
\end{aligned}
\tag{3.27}
$$

Il ne reste plus qu'à lire un des termes non-diagonaux de cette matrice pour trouver que pour tout couple $(t, t') \in \{t_1, \dots, t_k\}^2$ tel que $t \leq t'$,

$$
Cov[B_t, B_{t'}|B_1] = t(1-t') = min(t, t') \times (1 - max(t, t')). \tag{3.28}
$$

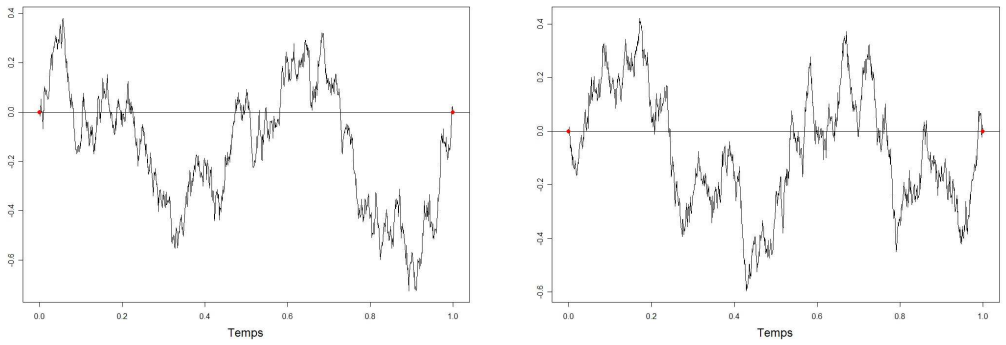


FIG. 3.3 – Deux réalisations du Pont Brownien.

Comme on sait que le pont brownien est centré et puisqu'il est gaussien en tant processus gaussien conditionné par une observation ponctuelle, il vient finalement que

$$
P_t \sim \mathcal{PG}\left(0, min(t, t') \times (1 - max(t, t'))\right) \tag{3.29}
$$

3.2 Eléments de géostatistique linéaire classique

3.2.1 Modélisation géostatistique

Développée dans les années 1950 par l'ingénieur minier sud-africain Daniel Krige [Kri51] puis mathématisée à l'école des mines de Paris par le français Georges Matheron [Mat69, Mat70] pendant toute la seconde moitiée du XX$^{\text{ème}}$ siècle, la géostatistique connaît aujourd'hui un nouvel essor. Elle permet en effet de modéliser des fonctions multivariables en prenant en compte des structures de dépendance spatiale. Nous rappelons ici comment les notions de champs aléatoires et de covariance spatiale ont permis de construire les outils traditionnels de la géostatistique linéaire, tels que le Krigeage.

Des mines d'or aux champs aléatoires

Les problématiques d'extraction des minéraux précieux offrent un cadre simple et naturel pour illustrer les notions géostatistiques de base. Considérons par exemple une mine d'or : sans trop restreindre la généralité, on peut supposer que le domaine est un cube $D = [0,1]^3$. On modélise alors la concentration locale en or comme une fonction $y : D \longrightarrow \mathbb{R}^+$ (teneur moyenne en or dans un petit bloc de roche centré sur x). L'expérience montre que cette fonction possède certaines propriétés de régularité. Informellement, deux concentrations $y(x)$ et $y(x')$ ont d'autant plus de chances d'être proches que x et y sont proches, c'est à dire que $h = y - x$ est petit en norme. En d'autres termes, si l'on mesure une concentration de 10 g. d'or par tonne de minerai en un site x, la concentration au site $x + h$ (t.q. $x + h \in D$) sera d'autant plus probablement proche de 10 que $||h||$ est petite. Ainsi, si l'on mesure $(y(x), y(x+h))$ en un très grand nombre de sites x [5] , on obtient un nuage de points positivement corrélés, avec un coefficient de corrélation ρ dépendant de $||h||$: lorsque $||h|| = 0$ le coefficient de corrélation $\rho(0)$ est de 1, et lorsque $||h|| \to +\infty$, $\rho(||h||) \to 0$. La manière dont la corrélation décroit avec $||h||$ est synthétisée par le comportement de $\rho(h)$, vu comme fonction de h. On appelle à ce propos souvent l'application $\rho : h \to \rho(h)$ *fonction d'autocorrélation*. Lorsque ρ est connue, la donnée des valeurs de y en un ensemble de sites $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ permet de dire des choses sur les valeurs de y aux sites non encore explorés. C'est sur la base de cette idée que sont apparus les concepts fondateurs de la géostatistique. La notion de fonction d'autocorrélation est d'ailleurs marginalement utilisée en géostatistique traditionnelle, au sein de laquelle on privilégie plutôt l'usage du *variogramme*.

[5] On devrait en fait les mesurer en un même x -relatif au centre du domaine- pour un grand nombre de mines d'or du même type. Dire que l'on se base sur des $(y(x), y(x+h))$ pour différents x d'un unique gisement cache en fait une hypothèse d'*ergodicité*, encore plus forte que la stationnarité.

Notions sur les processus intrinsèques, le variogramme, et la variographie

A l'instar du mouvement Brownien, bien des processus aléatoires apparaissant dans les applications ne jouissent pas de la stationnarité à l'ordre 2. La stationnarité intrinsèque est une notion plus faible que la stationnarité à l'ordre 2, donnant à la géostatistique un domaine de pertinence plus vaste.

Définitions ([GG08]) : Y est un *processus intrinsèquement stationnaire*, ou encore un *processus intrinsèque*, si $\forall h \in D$ le processus

$$\Delta Y^{(h)} = \{\Delta Y_x^{(h)} = Y_{x+h} - Y_x,\ x \in D\} \tag{3.30}$$

est stationnaire au second ordre. Le *semi-variogramme* de Y est la fonction $\gamma : D \longrightarrow \mathbb{R}$ définie par :

$$2\gamma(h) = var\left[(Y_{x+h} - Y_x)\right] \tag{3.31}$$

Il est clair que la propriété de stationnarité intrinsèque impose que la fonction $x \longrightarrow \mathbb{E}[Y_{x+h} - Y_x]$ soit constante. Comparé au cas stationnaire à l'ordre deux, cela n'impose ni que $\mathbb{E}[Y_x]$ soit constante (cas particulier où $\mathbb{E}[Y_{x+h} - Y_x]$ est nulle) ni même que $\mathbb{E}[Y_x]$ soit finie ! Remarquons aussi que la stationnarité d'ordre deux de $\Delta Y^{(h)}$ revient à l'existence du semi-variogramme γ mais n'implique pas l'existence d'une fonction de covariance stationnaire pour le processus Y ni même que Y soit dans L^2. Par exemple, le Mouvement Brownien B_t, dont nous avons vu qu'il n'est pas stationnaire à l'ordre 2, est bien intrinsèquement stationnaire et possède comme variogramme

$$\begin{aligned} var[B_{t+h} - B_t] &= var[B_{t+h}] + var[B_t] - 2cov[B_{t+h}, B_t] \\ &= (t+h) + t - 2\min(t+h, t) = |h| \end{aligned} \tag{3.32}$$

Nous reviendrons plus loin sur les liens entre processus intrinsèques et stationnaires à l'ordre deux. Voyons d'abord quelques propriétés incontournables du semi-variogramme.

Quelques propriétés de γ (adaptées de [GG08]) :

1. $\gamma(h) = \gamma(-h)$, $\gamma(h) \geq 0$ et $\gamma(0) = 0$.
2. Un variogramme est conditionnellement de type négatif, i.e. $\forall a \in \mathbb{R}^n$ t.q. $\sum_{i=1}^n a_i = 0$, $\forall \mathbf{x}^1, ..., \mathbf{x}^n \in D$,
$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(\mathbf{x}^i - \mathbf{x}^j) \leq 0. \tag{3.33}$$
3. Si A est une transformation linéaire sur $\mathbb{R}^d$, alors il est suffisant que γ soit un variogramme pour que $h \longrightarrow \gamma(Ah)$ en soit un.

4. Si γ est continue en 0, alors γ est continue en tout site $\mathbf{x}$ où γ est localement borné.
5. Si γ est borné au voisinage de 0, $\exists a$ et $b \geq 0$ tels que, pour tout $h : \gamma(h) \leq a||h||^2 + b$.

Pour revenir aux processus L^2 stationnaires à l'ordre 2, on a que tout Y de la sorte est intrinsèque, et son variogramme s'exprime en fonction de sa covariance k sous la forme :

$$\begin{aligned} 2\gamma(h) &= var\left[Y_{x+h} - Y_x\right] \\ &= var[Y_{x+h}] + var[Y_x] - 2cov[Y_{x+h}, Y_x] \\ &= 2\sigma^2 - 2k(h) \end{aligned} \tag{3.34}$$

Le variogramme apparaît ainsi comme un outil plus général que la fonction de covariance, présentant de plus l'avantage de pouvoir être estimé sans avoir besoin de recourir à une estimation de la moyenne, contrairement à k.

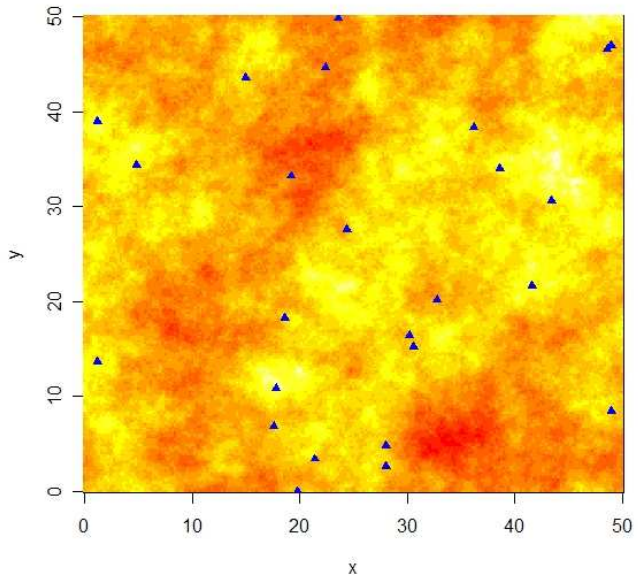


FIG. 3.4 – Réalisation (la même que sur 3.2) d'un champ gaussien de covariance exponentielle isotrope de portée 5 et de variance 10. Les triangles bleus représentent les 25 points d'un plan d'expériences, sur lequel se base l'analyse variographique de 3.5.

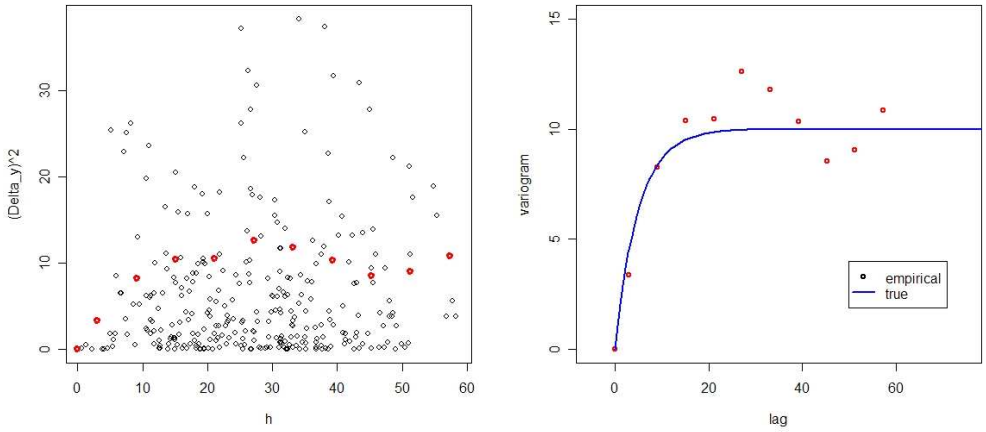


FIG. 3.5 – A gauche : nuée variographique associée aux 25 observations $\mathbf{x}^i$, $i \in [1, 25]$ (triangles bleus sur 3.4). Les 300 cercles noirs représentent les carrés des écarts observés $\frac{1}{2}\left(y(\mathbf{x}^i) - y(\mathbf{x}^j)\right)^2$, $i < j$, $(i, j) \in [1, 25]^2$ contre les $|\mathbf{x}^i - \mathbf{x}^j|$, et les 11 points rouges représentent les valeurs du variogramme empirique obtenues avec le package *RandomFields*. A droite : variogramme exponentiel effectivement utilisé (en bleu) versus les valeurs du variogramme empirique précédemments obtenues (toujours en rouge).

Quelques modèles usuels de variogrammes et de noyaux de covariance

Commençons par quelques variogrammes en dimension 1. Le *variogramme exponentiel* ou encore *variogramme de Ornstein-Uhlenbeck*, représenté à droite sur la figure 3.5, peut s'écrire sous la forme paramétrique

$$\forall h \in \mathbb{R},\ \gamma(h; \sigma^2, l) := \sigma^2 \left(1 - e^{-\frac{|h|}{l}}\right), \tag{3.35}$$

Le paramètre l est homogène à une *longueur de corrélation* ou *portée*, terme que nous retrouverons à de multiples reprises pour divers modèles de variogramme. σ^2 est appelé *pallier*, et représente la valeur limite de γ lorsque $||\mathbf{h}||$ tend vers l'infini, i.e. quantifie la dissemblance entre deux points suffisamment éloignés[6]. La portée donne justement un ordre de grandeur de la distance critique à partir de laquelle le variogramme atteint ce pallier (selon qu'il s'agisse de 100% ou de 95% de σ^2, on parle de *portée théorique* ou de *portée pratique*), i.e. à partir de laquelle deux valeurs prises par le champ aléatoire

[6]La notion de *pallier* n'a pas lieu d'être avec tous les variogrammes. On peut s'en convaincre aisément en pensant au variogramme $\gamma(h) = |h|$ du Mouvement Brownien.

cessent d'avoir de l'influence l'une sur l'autre. Comme on peut le constater en particulier sur la figure 3.4, le variogramme exponentiel va de pair avec des réalisations de champ aléatoire possédant des variations brusques (on peut même montrer d'un point de vue théorique qu'elles sont presque sûrement continues mais non dérivables, Cf. [RY91]). En faisant tendre la portée du variogramme exponentiel vers 0, on peut approcher le comportement du *variogramme pépitique* :

$$\forall h \in \mathbb{R},\ \gamma(h;\tau^2) = \tau^2(1-\delta_0(h)), \tag{3.36}$$

où δ_0 est un Dirac en 0 (Cf. [GG08], p.12). Ce variogramme correspond à un champ stationnaire dont les variables aléatoires ne sont pas du tout corrélées, même lorsques prises très voisines l'une de l'autre —sans être confondues—. Il possède ainsi un pallier σ^2, et une portée nulle, ce qui peut signifier en pratique « négligeable à l'échelle à laquelle on observe la grandeur d'intérêt ». Ainsi, si l'on s'intéresse à la teneur en or dans un sous-sol, la présence d'une pépite dans un bloc peut apparaître comme une discontinuité par rapport à la teneur dans un bloc voisin ne possédant pas de pépite. C'est autour de ces questions qu'a émergé l'adjectif pépitique et surtout le fameux *effet de pépite*. On peut par exemple introduire un effet de pépite dans le variogramme exponentiel :

$$\forall h \in \mathbb{R},\ \gamma(h;\sigma^2,l,\tau^2) := \tau^2(1-\delta_0(h)) + \sigma^2\left(1-e^{-\frac{|h|}{l}}\right), \tag{3.37}$$

Remarquons que l'effet de pépite n'influe pas sur la portée du variogramme, mais modifie le pallier de σ^2 en $\sigma^2+\tau^2$. En pratique, cela signifie qu'il existe une dissemblance entre des concentrations prises en deux sites distincts, même s'ils sont très proches. Cela se traduit aussi en terme de discontinuité en moyenne quadratique du processus étudié.

Pour revenir au variogramme exponentiel et à ses extensions, on appelle variogramme *exponentiel généralisé* le modèle

$$\forall h \in \mathbb{R},\ \gamma(h;\sigma^2,l,p) := \sigma^2\left(1-e^{-\left(\frac{|h|}{l}\right)^p}\right), \tag{3.38}$$

où $p \in]0,2]$ est un paramètre permettant de régler le comportement de γ, et en particulier la manière dont il croît avec h au voisinage de 0. On retrouve évidemment le variogramme exponentiel en prenant $p=1$, et le modèle obtenu pour $p=2$ est appelé *variogramme gaussien*. Ce dernier possède une très grande régularité en 0 ($\mathcal{C}^\infty$), et les processus qui lui sont associés sont dérivables en moyenne quadratique —et possèdent même des réalisations presque sûrement $\mathcal{C}^\infty$ lorsqu'il s'agit de processus gaussiens—.

Lorsque l'on passe à un cadre multidimensionnel ($d \geq 2$) et à mesure que d augmente, il est nécessaire de rechercher des variogrammes admissibles. En effet, il n'est

malheureusement pas exact qu'un variogramme $\gamma(.)$ admissible en dimension 1 donne systématiquement naissance à un variogramme admissible en dimension d, en prenant $\mathbf{h} \in \mathbb{R}^d \longrightarrow \gamma(||\mathbf{h}||)$ (Cf. chap.1 de [GG08] pour une discussion à ce sujet [7]). Pour autant, le modèle gaussien défini par $h \in \mathbb{R} \longrightarrow \gamma(\mathbf{h}; \sigma^2, l, p) = \sigma^2 \left(1 - e^{-\left(\frac{||\mathbf{h}||}{l}\right)^2}\right)$ est bien conditionnellement de type négatif et convient donc comme variogramme en dimension d. Il peut en fait être vu comme un cas particulier du modèle de Matèrn.

Le *modèle de Matèrn* est défini par

$$\forall \mathbf{h} \in \mathbb{R}^d, \gamma(\mathbf{h}; \sigma^2, l, \nu) := \sigma^2 \left\{1 - \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{||\mathbf{h}||}{l}\right)^{\nu} K_\nu \left(\frac{||\mathbf{h}||}{l}\right)\right\}, \tag{3.39}$$

où $\nu \in]-1, +\infty[$ et K_ν est la fonction de Bessel modifiée de deuxième espèce de paramètre ν. Ce variogramme est apprécié pour la flexibilité apportée par le paramètre ν, qui permet de régler finement la régularité de γ en 0, et donc la dérivabilité en moyenne quadratique des processus associés (Cf. [GG08] pour plus de détails). Citons enfin l'existence d'autres modèles, tels que les variogrammes *cubique*, *sphérique*, *circulaire*, *puissance*, qui ont été développés et étudiés en détail dans le cadre de la géostatistique traditionnelle [Cre93].

La difficulté majeure lorsque l'on veut utiliser une fonction comme variogramme tient au fait que cette fonction doive impérativement être conditionnellement de type négatif, ce qui peut être difficile à vérifier. On se ramène ainsi très souvent à des familles de variogrammes bien connues, telles que présentées jusqu'ici.

A ce stade, nous n'avons rencontré que des variogrammes dépendant de $||\mathbf{h}||$, i.e. des modèles de variogrammes *isotropes*. Or l'hypothèse d'isotropie n'est pas toujours tenable, et l'étude de phénomènes multidimensionnels ($d \geq 2$) nécessite souvent d'utiliser des modèles *anisotropes*. On parle d'*anisotropie géométrique* lorsque γ n'est pas une fonction de $||\mathbf{h}||$ mais peut s'écrire comme une fonction de $||M\mathbf{h}||$, avec $M \in \mathcal{M}_d(\mathbb{R})$ une matrice carrrée.En notant $||\mathbf{h}||_A = \sqrt{\mathbf{h}^T A \mathbf{h}}$, qui est une norme lorsque A est définie positive, on a que $||M\mathbf{h}|| = ||\mathbf{h}||_{M^T M}$. Un variogramme sujet à une anisotropie géométrique est donc un variogramme isotrope modulo un changement de norme quadratique. Ce résultat permet de montrer sans difficulté (Cf. [GG08], prop. 3, p.50) que les variogrammes isotropes classiques restent des variogrammes admissibles lorsque l'on remplace la norme euclidienne usuelle par une norme quadratique, i.e. lorsque l'on introduit une anisotropie géométrique. En pratique, toute la difficulté est de connaître une transformation linéaire

[7] En revanche, si $\gamma(||\mathbf{h}||)$ est un variogramme en dimension d, alors il est aussi admissible en dimension d', quel que soit $d' \in \mathbb{N}$ tel que $d' < d$.

M permettant de se ramener à l'isotropie ; nous y reviendrons à plusieurs reprises dans la suite de l'exposé. Mentionnons aussi l'existence d'anisotropies plus complexes, telles que l'*anisotropie zonale* développée par Journel et Matheron (Cf. par exemple [Mat70], fascicule 5, p.57).

Même si les variogrammes anisotropes ont été introduits et largement étudiés en géostatistique traditionnelle pour le cas $d = 2$, et dans une moindre mesure pour le cas $d = 3$, l'approche classique par processus stationnaires intrinsèques et variographie reste très marginale dans le domaine plus récent de l'*apprentissage machine* [RW06], où d est quelconque et où l'on raisonne plus volontiers en termes de processus stationnaire à l'ordre 2 et de noyaux de covariance. L'équation 3.34 permet d'obtenir des noyaux de covariance stationnaires à partir de la plupart des variogrammes classiques, pour autant qu'ils soient compatibles avec l'hypothèse de stationnarité à l'ordre 2 : à part pour le variogramme $\gamma(h) = |h|$ (sans pallier) associé au mouvement brownien, c'est le cas avec tous les modèles de variogramme évoqués dans cette section. En particulier, la fonction de covariance exponentielle généralisée isotrope s'écrit

$$k(\mathbf{h}) = \sigma^2 e^{-\sum_{j=1}^{d} \left(\frac{|\mathbf{h}_j|}{l} \right)^p} \tag{3.40}$$

La figure 3.9 représente cette fonction pour trois valeurs du paramètre p ($p \in \{0.4, 1.2, 2\}$, $\sigma^2 = 1, l = 1$), ainsi que quatre réalisations d'un processus gaussien centré et stationnaire pour chaque noyau de covariance. Similairement à ce qui est fait en variographie, on peut montrer que la régularité des réalisations sont contrôlées par la valeur de la dérivée en 0 de la fonction d'autocovariance, Cf. [SWN03].

La version anisotrope de l'autocovariance exponentielle généralisée est donnée par

$$k(\mathbf{h}) = \sigma^2 e^{-\sum_{j=1}^{d} \left(\frac{|\mathbf{h}_j|}{l_j} \right)^{p_j}} \tag{3.41}$$

Comme dans le cas isotrope, les p_j contrôlent la régularité du processus aléatoire. Chaque p_j permet ainsi de régler la forme de la décroissance du noyau de covariance dans la direction j. Les l_j jouent quant à eux le rôle de paramètres d'échelle : ils agissent sur la distance à partir de laquelle les variables aléatoires du champ aléatoire cessent de s'influencer dans la direction j (Cf. 3.6 pour une illustration dans les cas bidimensionnels gaussien et exponentiel). Le cas où $\forall i \in [1, d], p_i = 2$, correspondant à un noyau de covariance dit *gaussien anisotrope*, est très répandu dans les applications faisant intervenir un grand nombre de variables d'entrée. Signalons que ce dernier peut s'écrire en fonction

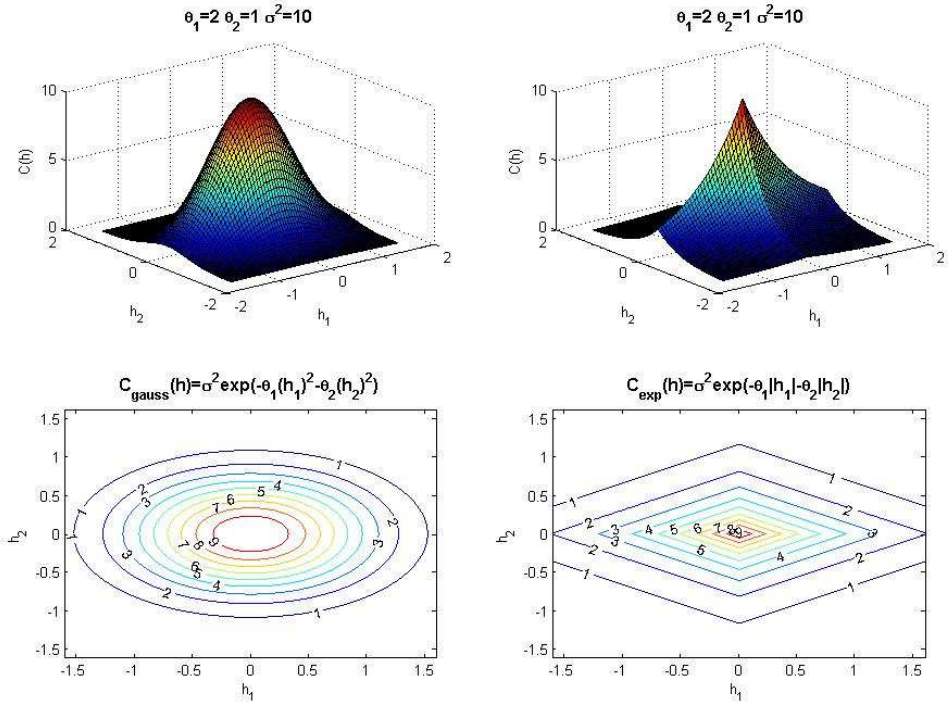


FIG. 3.6 – Surfaces (en haut) et lignes de niveau (en bas) de noyaux de covariance stationnaires anisotropes gaussien (à gauche) et exponentiel (à droite).

d'une norme quadratique particulièrement simple :

$$k(\mathbf{h}) = \sigma^2 e^{-\sum_{j=1}^{d}\left(\frac{|\mathbf{h}_j|}{l_j}\right)^2} = \sigma^2 e^{-\left\{\mathbf{h}^T \begin{pmatrix} \frac{1}{l_1} & 0 & \dots & 0 \\ 0 & \frac{1}{l_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{l_d} \end{pmatrix} \mathbf{h}\right\}} = \sigma^2 e^{-||\mathbf{h}||_A^2}, \tag{3.42}$$

avec $A = D^T D$, où $D = diag\left(\frac{1}{\sqrt{l_i}}, i \in [1,d]\right)$. Ce type de noyaux de covariance fait l'objet d'un paragraphe dans ([RW06], section 4.2) : on peut y trouver des références sur l'utilisation du noyau $\sigma^2 e^{-||\mathbf{h}||_A^2}$ avec A matrice symétrique semi-définie positive quelconque, et sur l'estimation optimale de A à partir d'observations. On se retrouve alors avec un problème d'estimation à $\frac{d(d+1)}{2}$ paramètres, dont la résolution devient rapidement déraisonnable lorsque la dimension d croît. Parmi les techniques de réduction de dimension envisageables, on peut décider par exemple (Cf. [RW06], p. 89) de restreindre

la forme de la matrice A à une famille du type

$$A = D + \Lambda\Lambda^T \tag{3.43}$$

où D est une matrice diagonale et $\Lambda \in \mathcal{M}_{d,k}(\mathbb{R})$ est une matrice dont les colonnes définissent k directions préférentielles, censées capturer au mieux la variabilité du processus considéré, tout en limitant le nombre de paramètres (ramené ici à $(k+1)d$, inférieur à $\frac{d(d+1)}{2}$ si et seulement si $k < \frac{d-1}{2}$). La distance associée à ce type de matrice est souvent appelée *distance d'analyse factorielle.*

Les noyaux de covariance exponentiels généralisés et de Matèrn (avec ou sans changement de métrique quadratique) ne constituent en fait qu'une maigre sélection d'exemples parmi les covariances stationnaires admissibles en dimension $d \in \mathbb{N}$. Ces dernières sont en effet exactement les fonctions de *type positif* [CG47] sur $\mathbb{R}^d$, i.e. les fonctions $k : \mathbf{h} \in \mathbb{R}^d \longrightarrow \mathbb{R}$ telles que $\forall a \in \mathbb{R}^n$, $\forall \mathbf{x}^1, ..., \mathbf{x}^n \in D$, $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(\mathbf{x}^i - \mathbf{x}^j) \geq 0$ (ceci est à rapprocher de l'équation 3.33 et de la propriété c.d.n. des variogrammes). Les fonctions de type positif se caractérisent avec des outils d'analyse harmonique (transformée de Fourier), comme l'illustre le résultat fondamental suivant.

Théorème de Bochner. *Soit k une fonction de type positif sur $\mathbb{R}^d$ et continue à l'origine. Alors il existe une unique mesure (positive, finie) μ sur $\mathbb{R}^d$ telle que*

$$\forall \mathbf{h} \in \mathbb{R}^d,\ k(\mathbf{h}) = \int_{\mathbb{R}^d} e^{\mathbf{i}<\mathbf{h},\mathbf{x}>} d\mu(\mathbf{x}) \tag{3.44}$$

Lorsque la mesure μ est dominée par la mesure de Lebesgue λ, on appelle *densité spectrale* la densité S de μ par rapport à λ (S est définie à une égalité λ-p.p. près). On obtient ainsi une classe très large de fonctions de type positif en considérant les transformées de Fourier de fonctions S positives. C'est l'approche principale développée dans [Ste99], où les propriétés des interpolateurs spatiaux de Krigeage sont étudiées en fonction de certaines caractéristiques de la densité spectrale associée à différents noyaux de covariance. Notons que l'obtention de noyaux de covariance admissibles par ce procédé requiert le calcul d'une transformée de Fourier, ce qui est d'une part contraignant et est d'autre part succeptible de ne pas mener à une forme explicite pour le noyau correspondant.

Une alternative simple pour concevoir de nouveaux noyaux de covariance admissibles (i.e. de type positif, noté ici « d.t.p. ») est de combiner les modèles classiques. La section *Making new kernels from old* de [RW06] est dédiée à cette idée. On peut y lire par exemple (dans le cas plus général de noyaux non-stationnaires, et avec des démonstrations élémentaires basées sur les processus aléatoires) que la somme de deux

noyaux d.t.p. est un noyau d.t.p., le produit d'un noyau d.t.p. par un nombre réel positif est un noyau d.t.p., ou encore que le produit de deux noyaux d.t.p. est un noyau d.t.p. De plus, si k_1 et k_2 sont des noyaux d.t.p. respectivement sur deux espaces E_1 et E_2, alors $k_1 + k_2$ est un noyau d.t.p. sur l'espace somme $E_1 + E_2$ et $k_1 k_2$ est un noyau d.t.p. sur l'espace produit $E_1 \times E_2$. On peut illustrer cette propriétés en considérant à nouveau la covariance exponentielle généralisée anisotrope (Cf. eq.(3.2.1)), dont le type positif sur $\mathbb{R}^d = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{d \text{ fois}}$ découle directement du fait que quel que soit $j \in [1, d]$, la fonction $h_j \in \mathbb{R} \longrightarrow k_j(h_j) = e^{-\left(\frac{|\mathbf{h}_j|}{l_j}\right)^{p_j}}$ $(l_j > 0, p_j \in]0, 2])$ est d.t.p. sur $\mathbb{R}$ et que

$$k(\mathbf{h}) = \sigma^2 e^{-\sum_{j=1}^{d}\left(\frac{|\mathbf{h}_j|}{l_j}\right)^{p_j}} = \sigma^2 \Pi_{j=1}^{d} k_j(h_j)$$

Citons aussi pour exemple l'annexe 5.6 de la thèse [Vaz05], dans laquelle sont proposées des combinaisons linéaires de noyaux. Les coefficients des combinaisons linéaires de noyaux sont estimés en utilisant un algorithme de *selection de caractéristiques par maximum de vraisemblance* (p. 171). Selon son auteur, « il s'agit d'une procédure itérative où de nouveaux noyaux ... sont inclus dans la combinaison linéaire s'ils sont jugés pertinents. Le critère de pertinence est obtenu à partir de la vraisemblance des données ». Cette approche est illustrée dans [Vaz05] sur deux exemples de synthèse, sur lesquels elle apparaît bien fondée comme alternative au choix arbitraire d'un unique noyau de covariance paramétré. L'approche proposée au chapitre 8 constitue une autre alternative.

Point de départ de l'approximation linéaire en les observations

Lorsque l'on se retrouve dans la situation représentée sur la figure 3.7, i.e. que l'on connaît une fonction y en un nombre fini de points $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ et que l'on souhaite avoir une idée ce que y vaut en un point $\mathbf{x} \in D$ quelconque , il semble naturel d'exploiter autant que possible les observations disponibles $\mathbf{Y} = \big(y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\big)$. Une approche particulièrement simple serait de se contenter de faire la moyenne arithmétique des observations $m = \sum_{j=1}^{n} \frac{1}{n} y(\mathbf{x}^j)$, et de proposer m comme première approximation de $y(\mathbf{x})$, indépendamment de $\mathbf{x}$. L'approximation serait alors bien grossière puisqu'elle donnerait la même prédiction sur tout le domaine, y compris à proximité de $\mathbf{X}$ (et même en les points d'observation). Cette approche revient en fait à supposer que $\forall \mathbf{x} \in D,\ y(\mathbf{x}) \approx m = \lambda^T \mathbf{Y}$, où $\lambda^T = (\frac{1}{n}, \ldots, \frac{1}{n})$, i.e. qu'il est raisonnable d'approcher $y(\mathbf{x})$ par une constante obtenue par combinaison linéaire des $\{y(\mathbf{x}^i), i \in [1, n]\}$. La faiblesse évidente d'une telle approximation est de ne pas du tout prendre en compte la régularité spatiale de y : si y ne varie pas trop « vite », on peut en fait s'attendre à ce que

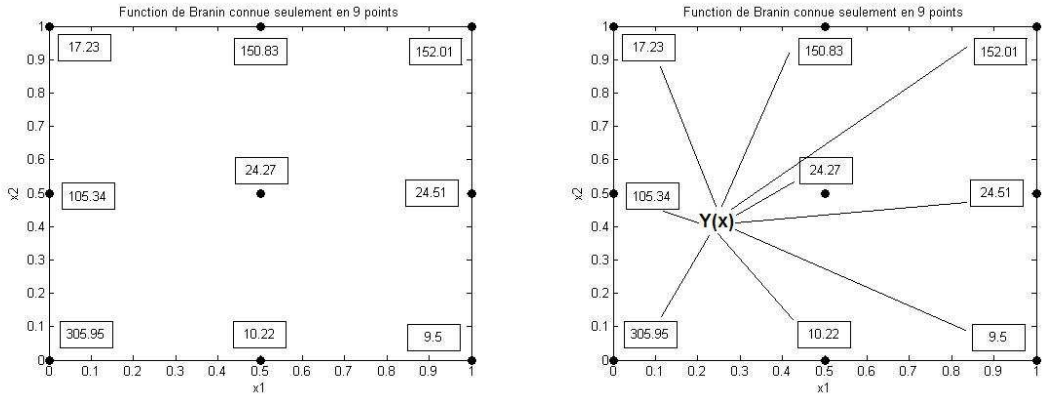


FIG. 3.7 – A gauche : évaluation en 9 points $\{\mathbf{x}^i, i \in [1,9]\}$ de la fonction de Branin-Hoo. A droite : la valeur de la fonction en un point $\mathbf{x}$ quelconque est vue comme la réalisation d'une variable aléatoire $Y(\mathbf{x})$, corrélée avec les variables aléatoires $\{Y(\mathbf{x}^i), i \in [1,9]\}$ dont on connaît les valeurs réalisées.

la valeur de $y(\mathbf{x})$ soit d'autant plus proche des valeurs $y(\mathbf{x}^i)$ que le point $\mathbf{x}$ est proche des $\mathbf{x}^i$. Il paraît alors sensé de prolonger l'idée de prédire $y(\mathbf{x})$ via une combinaison linéaire des $\{y(\mathbf{x}^i), i \in [1,n]\}$, mais en autorisant les poids de cette combinaison à varier avec $\mathbf{x}$ afin de respecter l'intuition que l'influence de la valeur prise par y en un point sur celle en un point voisin dépend —*a priori* de manière décroissante— de leur éloignement. L'approximation linéaire consiste à rechercher en chaque point $\mathbf{x}$ du domaine d'étude une approximation de $y(\mathbf{x})$ comme barycentre des observations :

$$\forall \mathbf{x} \in D,\ y(\mathbf{x}) \approx \sum_{j=1}^{n} \lambda_j(\mathbf{x}) y(\mathbf{x}^j),$$

où les $\lambda_j(\mathbf{x})$ sont des coefficients réels, souvent appelés *poids*. Les poids peuvent alors être choisis en suivant différents procédés. Pour prendre un exemple historique, le procédé dit *méthode des inverses des distances*(Cf. [RMC07, MdBMV94]) consiste à prendre des poids $\lambda_j(\mathbf{x})$ proportionnels aux $\frac{1}{d_j(\mathbf{x})}$ où $d_j(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}^i||$, c'est à dire après normalisation à approximer y en tout $\mathbf{x}$ par $\sum_{j=1}^{n} \left(\frac{\frac{1}{d_j(\mathbf{x})}}{\sum_{j=1}^{n} \frac{1}{d_j(\mathbf{x})}} \right) y(\mathbf{x}^j)$, en prolongeant l'expression par continuité lorsque $\mathbf{x} \in \mathbf{X}$. Une généralisation immédiate est de remplacer $d_j(\mathbf{x})$ par $d_j(\mathbf{x})^p$ avec $p \in]0, +\infty[$, le cas $p = 2$ permettant de faire de nombreuses analogies avec les lois de la physique *en inverse du carré de la distance* (classiques en gravitation, électrostatique, acoustique). De nombreuses fonctions de poids alternatives existent (Cf. [MdBMV94], p.123) et permettent de construire une vaste classe de méthodes d'ap-

proximation. Toujours employées « sur le terrain », ces méthodes possèdent malgré tout quelques défauts rédhibitoires : par exemple, même si plusieurs observations très voisines apportent pratiquement la même information, chacune d'elles sera prise en compte lors de prédictions et elles recevront ainsi au total un poids exagérément élevé par rapport à d'autres observations isolées.

Une des vertus de la méthode dite de *Krigeage* est de remédier à ce problème, comme nous le verrons dans la prochaine section. La base de la méthode développée par D.G. Krige est de construire des fonctions de poids à partir de variogrammes. Les poids sont calculés de manière à ce que l'approximation soit la meilleure possible sous des hypothèses d'ordre probabiliste sur la nature de la fonction y. En effet, une des idées fondamentales de la géostatistique est de supposer que la fonction y est une *réalisation* d'un champ aléatoire Y intrinsèquement stationnaire. Les poids $\lambda_j(\mathbf{x})$ sont alors choisis de telle sorte que la prédiction soit la moins risquée possible compte tenu des informations apportées par le variogramme, i.e. en cherchant en chaque point $\mathbf{x} \in D$ à rendre minimale l'*erreur quadratique moyenne de prédiction*[8],

$$\mathbb{E}\left[\left(Y(\mathbf{x}) - \sum_{j=1}^{n} \lambda_j(\mathbf{x}) Y(\mathbf{x}^j)\right)^2\right],$$

où l'espérance se réfère au vecteur aléatoire $\big(Y(\mathbf{x}), Y(\mathbf{x}^1), \ldots, Y(\mathbf{x}^n)\big)$, dont la loi jointe découle de la loi du processus Y. Avant d'expliciter les solutions de ce problème d'optimisation sous différentes hypothèses faites au sujet du processus Y (correspondant à autant de modèles de Krigeages), donnons quelques indispensables définitions et notations concernant les matrices de covariance et/ou de variogramme qui jouent un rôle central dans les équations de Krigeage.

Matrices et vecteurs de covariance stationnaire et de variogramme

Pour une fonction de covariance stationnaire k donnée, on note K la matrice de terme général $\big(k(\mathbf{x}^i - \mathbf{x}^j)\big)_{i,j\in[1,n]}$

$$K = \begin{pmatrix} k(\mathbf{0}) & k(\mathbf{x}^1 - \mathbf{x}^2) & \ldots & k(\mathbf{x}^1 - \mathbf{x}^n) \\ k(\mathbf{x}^2 - \mathbf{x}^1) & k(\mathbf{0}) & \ldots & k(\mathbf{x}^2 - \mathbf{x}^n) \\ \ldots & \ldots & \ldots & \ldots \\ k(\mathbf{x}^n - \mathbf{x}^1) & k(\mathbf{x}^n - \mathbf{x}^2) & \ldots & k(\mathbf{0}) \end{pmatrix}$$

[8] Nous écrivons ici le prédicteur de Krigeage comme combinaison linéaire des $Y(\mathbf{x}^j)$. La dépendance en les observations sera en fait selon les modèles considérés par la suite tantôt linéaire, tantôt affine.

K est la matrice de covariance du vecteur aléatoire $\big(Y(\mathbf{x}^1), Y(\mathbf{x}^2), ..., Y(\mathbf{x}^n)\big)$, où Y est par définition un processus aléatoire stationnaire à l'ordre 2 et de noyau de covariance k. Au titre de matrice de covariance, il est immédiat que K est symétrique réelle de valeurs propres positives ou nulles. On aura aussi besoin dans la suite de la fonction vectorielle $\mathbf{x} \longrightarrow \mathbf{k}(\mathbf{x})$ dont les fonctions composantes sont les covariances entre $Y(\mathbf{x})$ et les $\{Y(\mathbf{x_j}),\ j \in [1,n]\}$:

$$\forall \mathbf{x} \in D,\ \mathbf{k}(\mathbf{x}) = \Big(\begin{array}{cccc} k(\mathbf{x}-\mathbf{x}^1) & k(\mathbf{x}-\mathbf{x}^2) & ... & k(\mathbf{x}-\mathbf{x}^n) \end{array} \Big)^T$$

Une propriété remarquable liant K et $\mathbf{k}(\mathbf{x})$ dans le cas où K est inversible est que

$$\forall i \in [1,n],\ K^{-1}\mathbf{k}(\mathbf{x}^i) = \mathbf{e}^i, \tag{3.45}$$

où $\mathbf{e}^i = (0, \ldots, \underbrace{1}_{i^{\text{ème}}\ \text{coordonnée}}, \ldots, 0)^T$ est un vecteur de base canonique de $\mathbb{R}^n$.

Dans le cadre intrinséquement stationnaire, on définit de même à γ connu les matrices :

$$\begin{cases} \Gamma = \begin{pmatrix} \gamma(\mathbf{0}) & \gamma(\mathbf{x}^1-\mathbf{x}^2) & ... & \gamma(\mathbf{x}^1-\mathbf{x}^n) \\ \gamma(\mathbf{x}^2-\mathbf{x}^1) & \gamma(\mathbf{0}) & ... & \gamma(\mathbf{x}^2-\mathbf{x}^n) \\ ... & ... & ... & ... \\ \gamma(\mathbf{x}^n-\mathbf{x}^1) & \gamma(\mathbf{x}^n-\mathbf{x}^2) & ... & \gamma(\mathbf{0}) \end{pmatrix} \\ \\ \forall \mathbf{x} \in D,\ \gamma(\mathbf{x}) = \Big(\begin{array}{cccc} \gamma(\mathbf{x}-\mathbf{x}^1) & \gamma(\mathbf{x}-\mathbf{x}^2) & ... & \gamma(\mathbf{x}-\mathbf{x}^n) \end{array} \Big)^T \end{cases}$$

Ces différentes matrices jouent un rôle récurrent en géostatistique linéaire, et particulièrement au sein des équations de Krigeage présentées dans la section suivante.

3.2.2 Panorama des différents types classiques de Krigeage

Le terme de *Krigeage* est le plus souvent utilisé lorsque l'on cherche un prédicteur $\widehat{Y}_{\mathbf{x}^0}$ de la valeur en un point $\mathbf{x}^0 \in \mathbb{R}^d$ d'un processus aléatoire $(Y_{\mathbf{x}})_{\mathbf{x}\in\mathbb{R}^d}$, sous la forme d'une fonction affine des observations $\mathbf{Y} = \big(y(\mathbf{x}^1), ...,\ y(\mathbf{x}^n)\big)$ faites aux points d'un plan d'expériences $\mathbf{X} = \big(\mathbf{x}^1, ...,\ \mathbf{x}^n\big)$:

$$\begin{cases} \widehat{Y}_{\mathbf{x}^0} = a + \sum_{i=1}^n \lambda_i(\mathbf{x}^0) Y_{\mathbf{x}^i} = a + \lambda(\mathbf{x}^0)^T\mathbf{Y} \\ a \in \mathbb{R},\ \forall i \in [1,n]\ \lambda_i(\mathbf{x}^0) \in \mathbb{R} \end{cases} \tag{3.46}$$

La détermination de la constante a et des poids $\lambda(\mathbf{x}^0)$ se fait sur la base d'hypothèses concernant le processus aléatoire Y. On distingue différents jeux d'hypothèses classiques, ayant donné naissance à plusieurs types de Krigeage, avec chacun leur terminologie associée. Le *Krigeage Simple* est sans doute le plus fondamental d'entre eux.

Krigeage Simple

On suppose ici que le processus Y est stationnaire à l'ordre 2, de fonction d'autocovariance $k(\mathbf{h})$ et de moyenne $m(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}]$ connus. Quitte à considérer le processus centré $Y - m(\mathbf{x})$ [9], on peut ainsi toujours se ramener au cas d'un processus Y de moyenne nulle. Précisons aussi que le variogramme $\gamma(\mathbf{h})$ existe toujours dans le cadre stationnaire à l'ordre 2 —hypothèse plus forte que dans le cadre stationnaire intrinsèque, comme on l'a vu dans la section précédente— et que l'on a la relation de passage $\gamma(\mathbf{h}) = k(\mathbf{0}) - k(\mathbf{h})$. Les équations du Krigeage Simple en $\mathbf{x}^0 \in D$ s'obtiennent alors en recherchant la constante a et les poids $\lambda(\mathbf{x}^0)$ de manière à obtenir un prédicteur $\widehat{Y}_{\mathbf{x}^0}$ sans biais et de variance minimale. La première condition impose que

$$\mathbb{E}[Y_{\mathbf{x}^0} - \widehat{Y}_{\mathbf{x}^0}] = 0, \tag{3.47}$$

ce qui implique $a = 0$. La seconde condition permet de calculer le vecteur de poids optimaux $\lambda(\mathbf{x}^0)$ en résolvant

$$\min_{\lambda}\{Var[Y_{\mathbf{x}^0} - \lambda^T\mathbf{Y}]\} \tag{3.48}$$

On fait apparaître la convexité du problème en remarquant que :

$$\begin{aligned} Var[Y_{\mathbf{x}^0} - \lambda^T\mathbf{Y}] &= Var[Y_{\mathbf{x}^0}] + Var[\lambda^T\mathbf{Y}] - 2Cov[Y_{\mathbf{x}^0}, \lambda^T\mathbf{Y}] \\ &= k(\mathbf{0}) + \lambda^T K\lambda - 2\lambda^T\mathbf{k}(\mathbf{x}^0) \end{aligned} \tag{3.49}$$

Et il ne reste plus qu'à exhiber l'unique point critique. On obtient ainsi que $K\lambda(\mathbf{x}^0) - \mathbf{k}(\mathbf{x}^0) = 0$ par annulation du gradient de l'éq. 3.49, et il suit directement dans le cas où K est inversible que $\lambda(\mathbf{x}^0) = \mathbf{k}(\mathbf{x}^0)^T K^{-1}$.

Le prédicteur de Krigeage Simple en $\mathbf{x}^0$ s'écrit ainsi $\widehat{Y}_{\mathbf{x}^0} = \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbf{Y}$ dans le cas où Y est centrée. On obtient au passage dans le cas général :

$$\widehat{Y}_{\mathbf{x}^0} = m(\mathbf{x}^0) + \mathbf{k}(\mathbf{x}^0)^T K^{-1}(\mathbf{Y} - m(\mathbf{X})) \tag{3.50}$$

Dand un cas comme dans l'autre, la variance introduite dans l'éq. 3.49 vaut alors

$$Var[Y_{\mathbf{x}^0} - \lambda^T\mathbf{Y}] = k(\mathbf{0}) - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbf{k}(\mathbf{x}^0) \tag{3.51}$$

Cette quantité, notée ici $s^2(\mathbf{x}^0)$, est souvent appelée *Variance de Krigeage*. Nous verrons dans la suite (section 3.3) que $s^2(\mathbf{x}^0)$ peut en effet être vue comme une variance conditionnelle, modulo quelques hypothèses supplémentaires au sujet de Y.

[9] On devrait en toute rigueur noter $Y - m(\mathbf{x})\mathbb{1}_\Omega$ de manière à bien indiquer que le second terme est un processus aléatoire constant. On conservera tout de même cette notation simplifiée dans la suite.

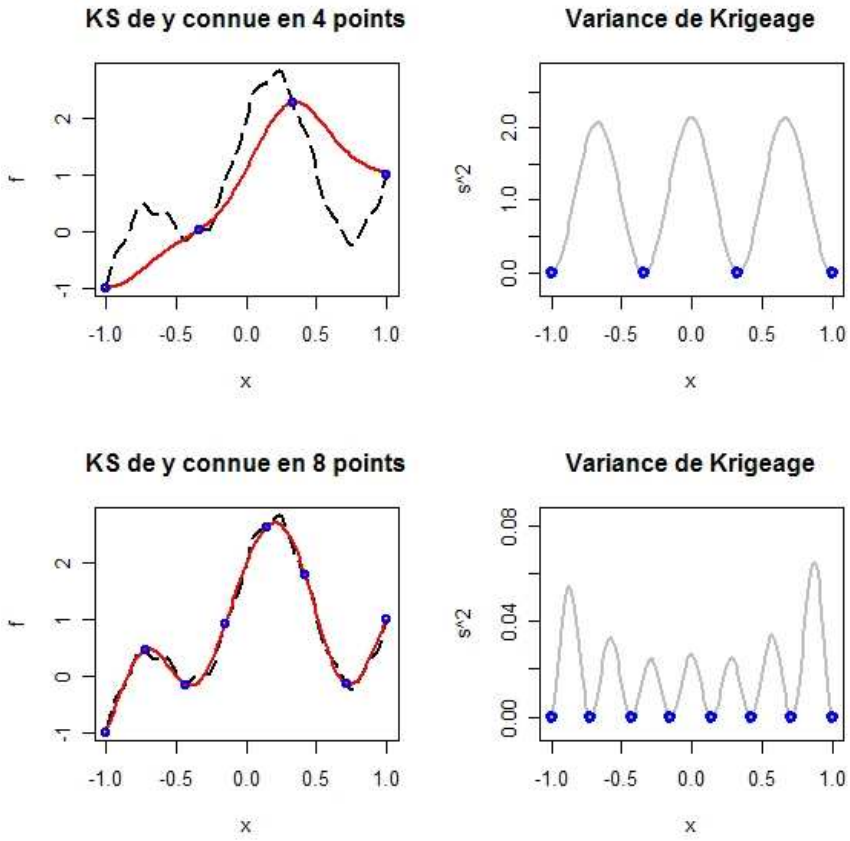


FIG. 3.8 – Krigeage Simple de la fonction $y(x) = 2\exp(-5x^2)+\sin(2\pi x)+0.1\sin(10\pi x)+x$, avec covariance gaussienne de variance 4 et de portée 0.4.

La figure 3.8 illustre l'application du Krigeage Simple à la prédiction d'une fonction déterministe monodimensionnelle. Deux plans d'expériences sont utilisés, respectivement une subdivision à 4 et 8 points de l'intervalle d'étude $[-1, 1]$ (Cf. fig. 3.8, haut et bas). On peut constater dans les deux cas que le prédicteur de Krigeage Simple interpole les observations et que la variance de Krigeage s'annule aux points du plan d'expériences. Dans le cas du plan à 4 points, l'interpolateur (en trait continu) approxime la fonction objectif assez grossièrement hors du plan ; la variance de Krigeage reflète ce phénomène, et donne des ordres de grandeurs assez réalistes pour anticiper les erreurs de prévision. Remarquons d'une part que cette variance ne dépend pas des observations (Cf. 3.51), et d'autre part que les ordres de grandeur obtenus dépendent du modèle et des paramètres de covariance choisis ici (le choix de la covariance est un vaste sujet auquel nous re-

viendrons à plusieurs reprises). Dans le cas du plan à 8 points, l'approximation obtenue est globalement bien meilleure, et la variance de Krigeage évolue dans ce sens. On peut remarquer les niveaux de variance sensiblement plus importants aux extrémités du domaine : le processus est mieux prévu aux points du centre du domaine, où il existe des points influents (à des distances relativement petites devant la portée) de part et d'autre du point considéré, ce qui permet d'avoir des prévisions plus précises.

Krigeages Ordinaire (KO) et Universel (KU)

On suppose maintenant que Y est la somme d'un partie déterministe $\mu(\mathbf{x})$ dépendant linéairement de paramètres inconnus, et d'un processus aléatoire intrinsèquement stationnaire δ d'espérance nulle [10] et de variogramme γ connu. Dans le *Krigeage Ordinaire*, $m(\mathbf{x})$ est une constante $\mu \in \mathbb{R}$:

$$\begin{cases} Y(\mathbf{x}) = \mu + \delta(\mathbf{x}), \mu \in \mathbb{R} \\ \delta \text{ stationnaire intrinsèque et } \forall \mathbf{x} \in D,\ \mathbb{E}[\delta(\mathbf{x})] = 0, \\ \forall \mathbf{h} \in D,\ \forall \mathbf{x} \in D,\ \gamma(h) = Var[\delta(\mathbf{x}+\mathbf{h}) - \delta(\mathbf{x})] \end{cases} \tag{3.52}$$

En plus des contraintes de linéarité et de non-biais déjà recontrées en Krigeage Simple, le Krigeage Ordinaire exige de vérifier que l'erreur de prévision $\hat{Y}_{\mathbf{x}^0} - Y_{\mathbf{x}^0}$ possède bien des moments d'ordre 1 et 2 (*contrainte d'autorisation*), puisque l'on va chercher à minimiser la variance de cette dernière. En notant comme dans le KS $\hat{Y}_{\mathbf{x}^0} = a + \sum_{i=1}^n \lambda_i \hat{Y}_{\mathbf{x}^i}$, l'erreur de prévision peut s'écrire

$$\begin{aligned} \hat{Y}_{\mathbf{x}^0} - Y_{\mathbf{x}^0} &= a + \sum_{i=1}^n \lambda_i Y_{\mathbf{x}^i} - Y_{\mathbf{x}^0} \\ &= a + \mu \left(\sum_{i=1}^n \lambda_i - 1 \right) + \sum_{i=1}^n \lambda_i \left(\delta_{\mathbf{x}^i} - \delta_{\mathbf{x}^0} \right) + \left(\sum_{i=1}^n \lambda_i - 1 \right) \delta_{\mathbf{x}^0} \end{aligned} \tag{3.53}$$

Comme cela est expliqué dans [Bai05], il est nécessaire d'imposer que $\sum_{i=1}^n \lambda_i - 1$ soit nulle pour assurer l'existence du moment d'ordre 2 de l'erreur de prévision. En gardant cette première condition à l'esprit, la contrainte de non-biais devient :

$$\begin{aligned} \mathbb{E}[\hat{Y}_{\mathbf{x}^0} - Y_{\mathbf{x}^0}] &= a + \mathbb{E}\left[\sum_{i=1}^n \lambda_i Y_{\mathbf{x}^i} - Y_{\mathbf{x}^0} \right] \\ &= a + \sum_{i=1}^n \lambda_i \mathbb{E}[Y_{\mathbf{x}^i} - Y_{\mathbf{x}^0}] = 0 \end{aligned} \tag{3.54}$$

[10] δ est nécessairement stationnaire à l'ordre 1, et Y est ainsi un processus intrinsèquement stationnaire et stationnaire à l'ordre 1 d'espérance $\mu(\mathbf{x})$, mais pas forcément stationnaire à l'ordre 2.

ce qui entraîne que $a = 0$ puisque $\forall i \in [1, n],\ \mathbb{E}[Y_{\mathbf{x}^i} - Y_{\mathbf{x}^0}] = 0$. Il reste alors pour calculer le prédicteur de Krigeage Ordinaire à minimiser la variance de l'erreur de prévision, dont nous donnons ici les étapes « clef » du développement classique[11] :

$$\begin{aligned} var[\hat{Y}_{\mathbf{x}^0} - Y_{\mathbf{x}^0}] &= \mathbb{E}\left[\left(\hat{Y}_{\mathbf{x}^0} - Y_{\mathbf{x}^0}\right)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \lambda_i \delta_{\mathbf{x}^i} - \delta_{\mathbf{x}^0}\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j \delta_{\mathbf{x}^i}\delta_{\mathbf{x}^j} - 2\delta_{\mathbf{x}^0}\sum_{j=1}^{n}\lambda_j\delta_{\mathbf{x}^j} + \delta_{\mathbf{x}^0}^2\right] \\ &= \mathbb{E}\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j(\delta_{\mathbf{x}^i} - \delta_{\mathbf{x}^j})^2\right] + \mathbb{E}\left[\sum_{i=1}^{n}\lambda_i(\delta_{\mathbf{x}^i} - \delta_{\mathbf{x}^0})^2\right] \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\frac{1}{2}\mathbb{E}\left[(\delta_{\mathbf{x}^i} - \delta_{\mathbf{x}^j})^2\right] + 2\sum_{i=1}^{n}\lambda_i\frac{1}{2}\mathbb{E}\left[(\delta_{\mathbf{x}^i} - \delta_{\mathbf{x}^0})^2\right] \\ &= -\lambda^T\Gamma\lambda + 2\lambda^T\gamma(\mathbf{x}^0) \end{aligned} \tag{3.55}$$

Compte-tenu de la contrainte $\sum_{i=1}^{n}\lambda_i = 1$, on obtient alors le prédicteur de Krigeage Ordinaire en résolvant le problème de minimisation contrainte suivant :

$$\begin{cases} \min\limits_{\lambda\in\mathbb{R}^n}\left\{-\lambda^T\Gamma\lambda + 2\lambda^T\gamma(\mathbf{x}^0)\right\} \\ \text{s.c. } \lambda^T\mathbb{1} = 1 \end{cases} \tag{3.56}$$

Suivant le cheminement de [Bai05], l'équation 3.56 peut être écrite en formulation lagrangienne en introduisant

$$\begin{aligned} L_{OK} : \mathbb{R}^n \times \mathbb{R} &\longrightarrow \mathbb{R} \\ (\lambda, l) &\longrightarrow L_{OK}(\lambda, l) = -\lambda^T\Gamma\lambda + 2\lambda^T\gamma(\mathbf{x}^0) + 2l(\lambda^T\mathbb{1} - 1) \end{aligned} \tag{3.57}$$

On trouve ainsi la solution du problème 3.56 en exhibant l'unique point critique ($\forall l \in \mathbb{R},\ L_{OK}(., l)$ est convexe puisque Γ est semi-définie négative) :

$$\begin{cases} \dfrac{\partial L_{OK}}{\partial\lambda}(\hat{\lambda}, l) = -2\Gamma\hat{\lambda} + 2\gamma(\mathbf{x}^0) + 2l\mathbb{1} = 0 \Rightarrow \forall l \in \mathbb{R},\ \hat{\lambda}(l) = \Gamma^{-1}(\gamma(\mathbf{x}^0) + l\mathbb{1}) \\ \dfrac{\partial L_{OK}}{\partial\lambda}(\hat{\lambda}, \hat{l}) = \mathbb{1}^T\Gamma^{-1}(\gamma(\mathbf{x}^0) + \hat{l}\mathbb{1}) = 1 \Rightarrow \hat{l} = (\mathbb{1}^T\Gamma^{-1}\mathbb{1})^{-1}\left(1 - \mathbb{1}^T\Gamma^{-1}\gamma(\mathbf{x}^0)\right) \end{cases} \tag{3.58}$$

On trouve en injectant l'expression de $\hat{l}$ dans $\hat{\lambda}$ que

$$\hat{\lambda} = \Gamma^{-1}\left(\gamma(\mathbf{x}^0) + \frac{1 - \mathbb{1}^T\Gamma^{-1}\gamma(\mathbf{x}^0)}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\mathbb{1}\right) \tag{3.59}$$

[11] La 4ème étape, fastidieuse mais sans difficulté majeure, est développée dans [Cre93] ou [Bai05]

et on obtient finalement la prévision par KO et la variance associée :

$$\hat{Y}(\mathbf{x}^0) = \left(\gamma(\mathbf{x}^0) + \frac{1 - \mathbb{1}^T\Gamma^{-1}\gamma(\mathbf{x}^0)}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\mathbb{1}\right)^T \Gamma^{-1}\mathbf{Y} \tag{3.60}$$

$$var[\hat{Y}(\mathbf{x}^0) - Y(\mathbf{x}^0)] = \gamma(\mathbf{x}^0)^T\Gamma^{-1}\gamma(\mathbf{x}^0) - \frac{\left(1 - \mathbb{1}^T\Gamma^{-1}\gamma(\mathbf{x}^0)\right)^2}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}} \tag{3.61}$$

Pour aller plus loin, le Krigeage Universel (KU) est préféré au KO lorsque le processus étudié comporte une tendance déterministe que le modélisateur pense pouvoir approcher linéairement. Le KU permet en effet de prendre en compte dans le Krigeage de Y l'estimation des paramètres d'une tendance déterministe, pourvu que la tendance $\mu(\mathbf{x})$ soit linéaire par rapport à un nombre fini de fonctions de base $\{f_1, ..., f_b\}$ données :

$$\mu(\mathbf{x}) = \sum_{j=1}^{b} \beta_j f_j(\mathbf{x}) \tag{3.62}$$

Notons $\mathcal{F} = \{f_1, ..., f_b\}$ la famille de fonctions choisies, $f(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_b(\mathbf{x}))^T$ la fonction vectorielle de composantes les f_j ($j \in [1, b]$), et $\mathbb{F} = (f_j(\mathbf{x}^i))_{i\in[1,n],j\in[1,b]}$ la matrice d'expériences. Comme dans le cas du Krigeage Ordinaire, la contrainte d'autorisation impose que $\sum_{i=1}^n \lambda_i = 1$ (Cf. [Bai05]). En revanche, la contrainte de non-biais s'écrit

$$\begin{aligned}\mathbb{E}[\hat{Y}_{\mathbf{x}^0} - Y_{\mathbf{x}^0}] &= a + \sum_{i=1}^{n}\lambda_i \sum_{j=1}^{b}\beta_j f_j(\mathbf{x}^i) - \sum_{j=1}^{b}\beta_j f_j(\mathbf{x}^0) \\ &= a + \sum_{j=1}^{b}\left(\sum_{i=1}^{n}\lambda_i f_j(\mathbf{x}^i) - f_j(\mathbf{x}^0)\right)\beta_j = 0.\end{aligned} \tag{3.63}$$

Cela entraîne que $a = 0$ et $\forall j \in [1, b]$, $\sum_{i=1}^n \lambda_i f_j(\mathbf{x}^i) - f_j(\mathbf{x}^0) = 0$, ce que l'on peut aussi résumer sous une forme matricielle plus compacte :

$$\lambda^T\mathbb{F} = f(\mathbf{x}^0) \tag{3.64}$$

On peut noter que dans le cas où $f_1 \equiv 1$, la première ligne de l'équation vectorielle 3.64 correspond exactement à la contrainte $\sum_{i=1}^n \lambda_i = 1$. En convenant que l'on a bien $f_1 \equiv 1$, et en remarquant que l'expression formelle de la variance de prévision (3.49 reste inchangée par rapport à celle du Krigeage Ordinaire, on obtient le prédicteur de Krigeage Universel en résolvant le problème de minimisation contrainte suivant

$$\begin{cases} \min\limits_{\lambda\in\mathbb{R}^n} \left\{-\lambda^T\Gamma\lambda + 2\lambda^T\gamma(\mathbf{x}^0)\right\} \\ \text{s.c. } \lambda^T\mathbb{F} = f(\mathbf{x}^0) \end{cases} \tag{3.65}$$

Comme dans le cas du KO, 3.65 peut être écrite en formulation lagrangienne en introduisant

$$\begin{aligned} L_{UK} : \mathbb{R}^n \times \mathbb{R}^b &\longrightarrow \mathbb{R} \\ (\lambda, \mathbf{l}) &\longrightarrow L_{UK}(\lambda, l) = -\lambda^T \Gamma \lambda + 2\lambda^T \gamma(\mathbf{x}^0) + 2(\lambda^T \mathbb{F} - f(\mathbf{x}^0))\mathbf{l}, \end{aligned} \tag{3.66}$$

où $\mathbf{l}$ est un vecteur de b multiplicateurs de Lagrange. On obtient la solution de l'équation 3.65 en exhibant l'unique point critique :

$$\begin{cases} \dfrac{\partial L_{UK}}{\partial \lambda}(\hat{\lambda}, l) = -2\Gamma\hat{\lambda} + 2\gamma(\mathbf{x}^0) + 2\mathbb{F}\mathbf{l} = 0 \Rightarrow \hat{\lambda} = \Gamma^{-1}(\gamma(\mathbf{x}^0) + \mathbb{F}\mathbf{l}) \\ \dfrac{\partial L_{UK}}{\partial \mathbf{l}}(\hat{\lambda}, \hat{\mathbf{l}}) = \mathbb{F}^T\Gamma^{-1}(\gamma(\mathbf{x}^0) + \mathbb{F}\mathbf{l}) = f(\mathbf{x}^0) \\ \Rightarrow \hat{\mathbf{l}} = (\mathbb{F}^T\Gamma^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0) - \mathbb{F}^T\Gamma^{-1}\gamma(\mathbf{x}^0)) \end{cases} \tag{3.67}$$

On trouve en injectant l'expression de $\hat{\mathbf{l}}$ dans $\hat{\lambda}$ que

$$\hat{\lambda} = \Gamma^{-1}\left(\gamma(\mathbf{x}^0) + \mathbb{F}(\mathbb{F}^T\Gamma^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0) - \mathbb{F}^T\Gamma^{-1}\gamma(\mathbf{x}^0))\right) \tag{3.68}$$

et on obtient finalement la prévision par Krigeage Universel :

$$\hat{Y}(\mathbf{x}^0) = \left(\gamma(\mathbf{x}^0) + \mathbb{F}(\mathbb{F}^T\Gamma^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0) - \mathbb{F}^T\Gamma^{-1}\gamma(\mathbf{x}^0))\right)^T \Gamma^{-1}\mathbf{Y}, \tag{3.69}$$

et la variance associée s'écrit :

$$\begin{aligned} var[\hat{Y}(\mathbf{x}^0) - Y(\mathbf{x}^0)] &= \gamma(\mathbf{x}^0)^T\Gamma^{-1}\gamma(\mathbf{x}^0) \\ &\quad - (f(\mathbf{x}^0) - \mathbb{F}^T\Gamma^{-1}\gamma(\mathbf{x}^0))^T(\mathbb{F}^T\Gamma^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0) - \mathbb{F}^T\Gamma^{-1}\gamma(\mathbf{x}^0)) \end{aligned} \tag{3.70}$$

On peut vérifier au passage que les deux dernières expressions coïncident bien, lorsque $\mathcal{F} = vect\{f_1\}$ ($f_1 \equiv 1$), avec celles du Krigeage Ordinaire.

Dans le cas où le processus Y étudié est stationnaire à l'ordre 2, les équations du KU (et *a fortiori* celles du KO) peuvent aussi s'écrire en termes de covariance :

$$\begin{aligned} \hat{Y}(\mathbf{x}^0) &= \left(\mathbf{k}(\mathbf{x}^0) + \mathbb{F}(\mathbb{F}^T K^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0) - \mathbb{F}^T K^{-1}\mathbf{k}(\mathbf{x}^0))\right) K^{-1}\mathbf{Y} \\ var[\hat{Y}(\mathbf{x}^0) - Y(\mathbf{x}^0)] &= k(0) - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbf{k}(\mathbf{x}^0) \\ &\quad + (f(\mathbf{x}^0)^T - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbb{F})(\mathbb{F}^T K^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0)^T - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbb{F})^T \end{aligned} \tag{3.71}$$

L'hypothèse de variogramme connu est généralement abusive. Nous abordons quelques questions relatives au choix et à l'estimation de γ (resp. de k) dans la section suivante ainsi que dans les chapitres 5 à 8.

3.2.3 Questions de choix de modèle, et en particulier du variogramme

On a vu dans la section précédente que différents types de Krigeage étaient usuellement employés, en faisant une distinction entre des termes de tendance de plusieurs natures (allant d'une tendance connue à une combinaison linéaire de fonctions de base avec coefficients β inconnus). La question du choix entre structures de tendance et/ou de la sélection des fonctions de base seront abordées un peu plus en détail au chapitre 6.

La prise en compte des incertitudes liées au terme de tendance est un problème déjà en partie résolu grâce au Krigeage Universel, qui permet d'intégrer l'incertitude sur β dans les équations du prédicteur et de la variance associée ; on verra d'ailleurs dans la prochaine section qu'il est possible de faire cette intégration d'une manière encore plus aboutie en suivant une démarche bayésienne. Convenons pour la suite de cette discussion que les fonctions de base, s'il y a lieu d'en considérer, sont choisies *a priori* et fixées une fois pour toutes. Que l'on fasse alors du KS, KO, ou KU, il est une hypothèse commune qui sous-tend l'ensemble des résultats obtenus : tous les modèles abordés lors de notre panorama des méthodes classiques de Krigeage pré-supposent le variogramme γ ou encore le noyau de covariance k connu.

Dans la pratique, il est pourtant très rare de vraiment connaître le variogramme ou le noyau de covariance *ad hoc* avant toute analyse des observations disponibles. Le choix et/ou l'estimation du variogramme (resp. du noyau) s'avèrent être des étapes cruciales de la démarche géostatistique, d'autant plus qu'elles peuvent avoir une influence considérable sur les prédictions de Krigeage. Si de nombreuses techniques existent à ces sujets, deux paradigmes principaux semblent ressortir en ce qui concerne la marche à suivre pour déterminer γ : l'analyse structurale classique —basée sur la « variographie »—, *versus* les méthodes contemporaines de sélection et estimation automatiques de noyaux de covariance. Nous en donnons ci-dessous un rapide aperçu, dans l'objectif de permettre au lecteur de se faire une première idée sur leurs forces et faiblesses respectives. Certains points seront sensiblement approfondis dans les prochains chapitres (chap. 5 concernant l'estimation de paramètres par maximum de vraisemblance, chap. 7 pour l'injection d'informations de nature algébrique à l'étape de conception du noyau).

Méthodes d'analyse structurale pour le choix et l'estimation du variogramme

L'outil de travail de base de l'analyse structurale est la nuée variographique (on se place dans un premier temps en contexte isotrope, à l'image de l'exemple illustré par 3.5). Il

s'agit alors de remonter à un variogramme γ à partir de la nuée variographique

$$\mathcal{V} = \left\{ \left(||\mathbf{x}^i - \mathbf{x}^j||, \frac{1}{2}(y(\mathbf{x}^i) - y(\mathbf{x}^j))^2 \right), \ (i,j) \in [1,n]^2 \right\} \tag{3.72}$$

en gardant bien à l'esprit que

$$\mathbb{E}\left[\frac{1}{2} \left(Y(\mathbf{x}^i) - Y(\mathbf{x}^j) \right)^2 \right] = \gamma(\mathbf{x}^i - \mathbf{x}^j) \tag{3.73}$$

En dépit de l'égalité 3.73, ce problème est très difficile à résoudre pour plusieurs raisons :

- d'une part, on possède généralement trop peu de données pour que $\gamma(||\mathbf{h}||)$ puisse être convenablement estimé comme moyenne des $\{\frac{1}{2}(y(\mathbf{x}^i) - y(\mathbf{x}^j))^2$ t.q. $||\mathbf{x}^i - \mathbf{x}^j|| = ||\mathbf{h}||\}$ (il y a même peu de chances qu'un seul couple tel que $||\mathbf{x}^i - \mathbf{x}^j|| = ||\mathbf{h}||$ existe dans le plan d'expériences ; on utilise des couples de points du plan pour lesquels l'égalité est approchée).
- d'autre part, on a vu que le variogramme devait impérativement être conditionnellement de type négatif, ce qui est considéré comme pratiquemment impossible à vérifier directement sur la base du graphe de γ. Il est alors nécessaire de se restreindre à des classes de fonctions paramétriques admissibles comme variogrammes, et d'estimer les paramètres en question en se basant sur $\mathcal{V}$.

Le choix d'un modèle paramétrique de variogramme $\gamma(.;\psi)$ est alors fait sur la base d'informations « métier »(expertise du géologue, pratiques issues de l'état de l'art, etc.) ou graphiquement à partir de la nuée $\mathcal{V}$ (*eye fitting*). Une fois cette *structure* fixée, on cherche les paramètres ψ permettant d'obtenir la meilleure adéquation possible avec les données observées ; là encore, les procédés employés pour estimer les paramètres variographiques à partir des données sont multiples.

Pour donner une vue d'ensemble des techniques usitées —en étant loin de prétendre à l'exhaustivité—, on peut citer l'ajustement direct de $\gamma(.;\psi)$ à $\mathcal{V}$ par moindres carrés ordinaires, par moindres carrés pondérés (Cf. [Cre93]), par maximum de vraisemblance en supposant que les écarts quadratiques suivent une loi du χ^2 centrée en 2γ (conséquence d'une hypothèse de gaussiannité sur Y), où encore en ajustant un modèle paramétrique à un *variogramme empirique* déduit de la nuée variographique par moyennes locales. On peut trouver de nombreuses variantes au sujet de l'estimation des paramètres d'un variogramme sur la base de données expérimentales dans les travaux de Matheron [Mat70], Journel [Jou88], Stein [Ste99], Cressie [Cre93] ou encore Gaetan & Guyon [GG08].

Mentionnons aussi l'approche spectrale, chère à Stein [Ste99], et ayant donné lieu à la thèse [Yao98] sur l'utilisation de la transformée de Fourier rapide, permettant une estimation fidèle de variogrammes $2D$ (anisotropes) à partir d'observations récoltées sur des grilles. L'idée de passer dans l'espace des fréquences pour l'estimation du variogramme est profonde puisqu'elle permet de changer une contrainte de négativité conditionnelle du type en simple contrainte de signe sur la densité spectrale ; elle permet donc une approche non-paramétrique, inenvisageable dans le domaine spatial. L'inconvénient majeur d'une telle méthode est de nécessiter beaucoup de points d'observations, suivant une répartition spatiale régulière très contraignante. En particulier, l'usage de grilles devient rapidement inabordable lorsque la dimension d dépasse 3.

Pour revenir aux outils classiques de l'analyse structurale, c'est justement autour des questions de montée en dimension et d'anisotropies qu'ils se révèlent souvent limités pour les applications en approximation de codes de calcul à entrées multivariables. En effet, les méthodes d'analyse variographique permettent —au prix d'un grand nombre d'observations— de distinguer au sein de nuées variographiques $2D$ (voire $3D$) des comportements hétérogènes entre des coupes de variogramme prises selon des orientations différentes (par exemple en distinguant Sud/Nord, Sud-Ouest/Nord-Est, etc.).

L'inférence de ces *variogrammes directionnels* nécessite de disposer dans la nuée variographique $\mathcal{V}$, pour chaque direction considérée, de suffisamment de couples de points $(\mathbf{x}^i, \mathbf{x}^j)$ tels que les $\mathbf{x}^i - \mathbf{x}^j$ soient approximativement dans la direction en question, et que les modules $||\mathbf{x}^i - \mathbf{x}^j||$ couvrent de manière raisonnable la plage d'interdistances sur laquelle ce variogramme directionnel varie (de 0 à une valeur sensiblement au-delà de la portée pratique). Ces conditions sont difficiles à réunir en pratique, fautes d'observations en nombre suffisant, sans même rajouter la difficulté de trouver les directions d'anisotropie, qui sont généralement renseignées à l'avance sur la base d'avis d'experts ;

Il est alors clair que l'approche par nuées variographiques n'est pas indiquée pour les cas d'estimation de variogrammes anisotropes en plus grandes dimensions, à moins de posséder des informations *a priori* fiables permettant de simplifier l'estimation, ou encore de développer des outils spécifiques de fouille de données pour ce problème (techniques n'ayant visiblement pas encore été pleinement transposées en analyse structurale). La section suivante présente des méthodes automatiques alternatives à l'analyse structurale, aujourd'hui majoritairement employées en *apprentissage statistique*, plus précisément dans le domaine de l'*apprentissage machine par processus gaussiens* [RW06].

Méthodes automatiques pour l'estimation des paramètres de covariance

Dans les techniques d'apprentissage statistique dédiées à l'approximation de fonctions déterministes multidimensionnelles, on fait presque systématiquement l'hypothèse que le processus aléatoire sous-jacent Y est de carré intégrable, remplaçant ainsi l'estimation d'un variogramme par celle d'un noyau de covariance k (parfois même un noyau non-stationnaire). Comme en géostatistique, l'estimation de k est généralement faite en deux phases : le choix d'une structure paramétrique, i.e. une famille de noyaux de type positif $k(.;\psi)$ paramétrés par un vecteur ψ de dimension fini, puis l'estimation des paramètres en question sur la base des données disponibles.

Contrairement à ce qui est fait dans le cadre variographique, dans lequel il est parfois possible de choisir la structure de variogramme de manière graphique, la première phase repose essentiellement dans le cadre des noyaux de covariance sur l'*a priori* du modélisateur, via le choix d'une famille de noyaux dont les réalisations de processus associées possèdent des propriétés (régularité, stationnarité, etc.) correspondant au comportement attendu de la fonction y. De même, les moyens d'estimer des paramètres de

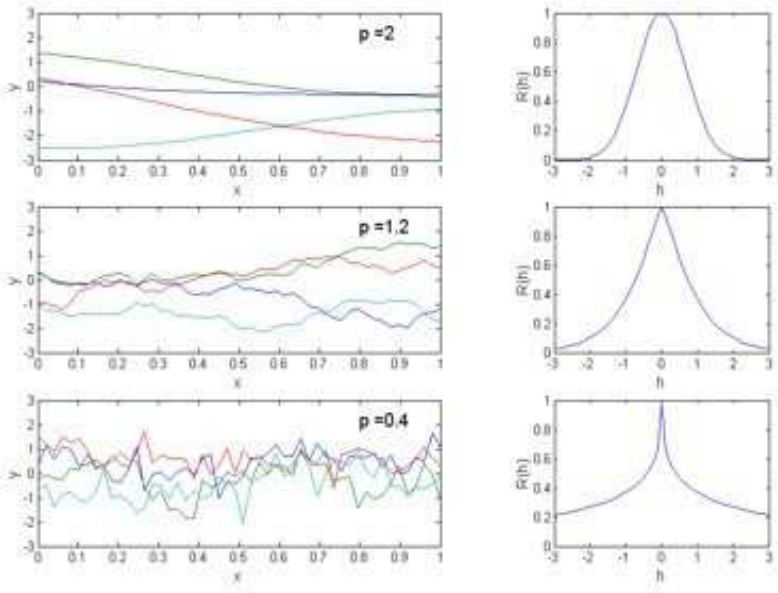


FIG. 3.9 – Réalisations de PG (à gauche) et leurs fonctions d'autocorrélation exponentielles généralisées (à droite, Cf. 3.40) pour $l = 1$ et $p = 2, 1.2, 0.4$ (de haut en bas)

covariance diffèrent de ceux employés en géostatistique pour les paramètres variographiques : même s'il on fait dans un cas comme dans l'autre appel aux observations, il n'est pas question d'utiliser la nuée variographique avec peu d'observations en dimensions supérieures à deux, comme on l'a expliqué à la fin du dernier paragraphe.

A modèle paramétrique choisi, l'estimation des paramètres de covariance repose en fait sur ce que Vazquez appelle dans [Vaz05] une « fonction d'attache aux données », i.e. elle passe par la maximisation en ψ d'un critère d'adéquation entre les données observées et celles qu'un processus de noyau de covariance $k(.;\psi)$ est susceptibe d'engendrer. Cette fonction d'attache aux données peut être choisie de plusieurs natures ; dans la pratique, suivant la littérature du Krigeage pour expériences numériques, on a le plus souvent recours lorsque l'on souhaite remonter aux paramètres ψ directement à partir des observations $(\mathbf{X}, \mathbf{Y})$ soit à une estimation de type "maximum de vraisemblance" (EMV), soit à une minimisation de l'erreur de validation croisée (Cf. [RW06, Vaz05]) :

- L'EMV repose sur l'hypothèse que Y est un processus gaussien de noyau $k(;\psi)$, et ainsi que $\mathbf{Y}$ est une réalisation d'un vecteur gaussien de matrice de covariance construite à partir de $k(;\psi)$. Il s'agit alors de trouver ψ tel que l'observation du vecteur $\mathbf{Y}$ soit la plus vraisemblable possible sous le modèle gaussien paramétré par ψ, i.e. de maximiser la densité de probabilité du vecteur $\mathbf{Y}$ en tant que fonction de ψ. Nous reviendrons brièvement dans la prochaine section sur la notion de vraisemblance, à laquelle sera par ailleurs consacré l'ensemble du chapitre 5.

- La validation croisée (*cross-validation*) vise à quantifier la capacité de généralisation d'un modèle en se basant uniquement sur les données d'apprentissage $(\mathbf{X}, \mathbf{Y})$, i.e. sans faire appel à un nouvel ensemble test. Elle consiste, pour un noyau $k(;\psi)$ donné, à retirer une partie $\mathbf{X}_{cv}$ des observations connues, à prédire y en $\mathbf{X}_{cv}$ avec le modèle de Krigeage considéré, et à mesurer la distance entre observations et prédictions (via par exemple la somme des écarts au carré). Comme la quantité obtenue est fortement dépendante de l'ensemble $\mathbf{X}_{cv}$ choisi, on réitère le procédé pour $\mathbf{X}_{cv}$ parcourant une partition de $\mathbf{X}$. La variante la plus célèbre est sans doute la validation croisée *leave-one-out* (LOO), pour laquelle $\mathbf{X}_{cv}$ est réduit à un seul point et parcourt les singletons de $\mathbf{X}$. Une fois les observations et prédictions comparées sur tous les ensembles de validation croisée, la somme des distances obtenues donne un indicateur global de cohérence du modèle de Krigeage associé au noyau $k(;\psi)$. L'estimation de ψ consiste finalement à trouver la valeur de ψ permettant de minimiser cette erreur globale de validation croisée.

KO de la fonction de Branin, avec paramètres ψ estimés par MV

La fonction de Branin-Hoo y_{BH} est une fonction test classique en optimisation globale (Cf. [JSW98]). Elle est définie pour $(x_1, x_2) \in [-5, 10] \times [0, 15]$ comme suit :

$$y_{BH}(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)cos(x_1) + 10$$

Nous avons ici normalisé les variables entre 0 et 1.

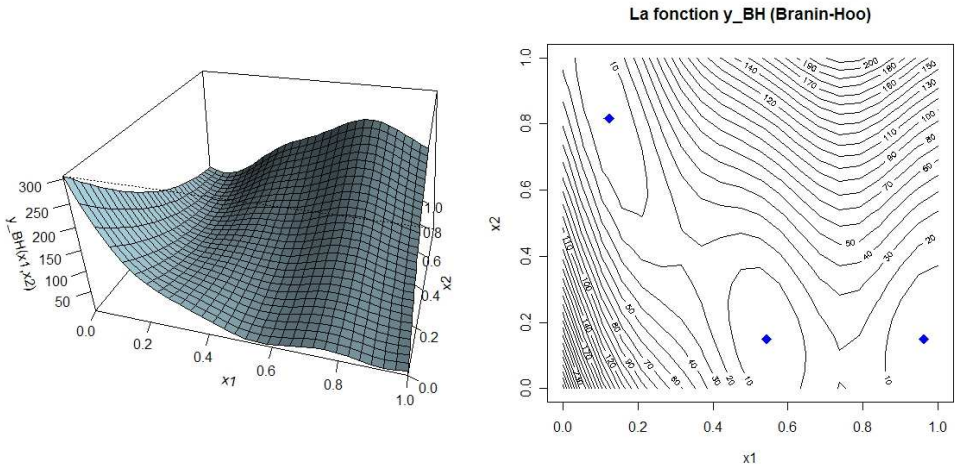


FIG. 3.10 – Surface (à gauche) et lignes de niveaux (à droite) de la fonction de Branin-Hoo. Les points bleus du graphe de droite représentent les optimiseurs globaux de y_{BH}.

La figure 3.10 représente la fonction y_{BH} normalisée, sous forme de surface en $3D$ et de lignes de niveaux en $2D$. Remarquons, même si nous n'en avons pas encore besoin ici, que y_{BH} possède trois minimiseurs globaux $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$ (valeurs non-normalisées), et que la valeur du minimum global associé est de approximativement de 0.4. La figure 3.11 représente les surfaces de moyenne et de variance de Krigeage obtenues par un Krigeage Ordinaire de y_{BH} sur la base d'un plan factoriel complet $\mathbf{X}$ à neuf points (en rouge sur le graphe de gauche), et avec un noyau de covariance gaussien anisotrope, de paramètres ψ estimés par maximum de vraisemblance.

On observe sur la figure (3.11, graphique de gauche) que la surface moyenne de Krigeage Ordinaire interpole bien les observations au plan d'expériences, et possède globalement un forme très proche de la surface que l'on cherche à approcher. Le graphique de droite représente la surface de variance de Krigeage Ordinaire pour ce même plan. On constate

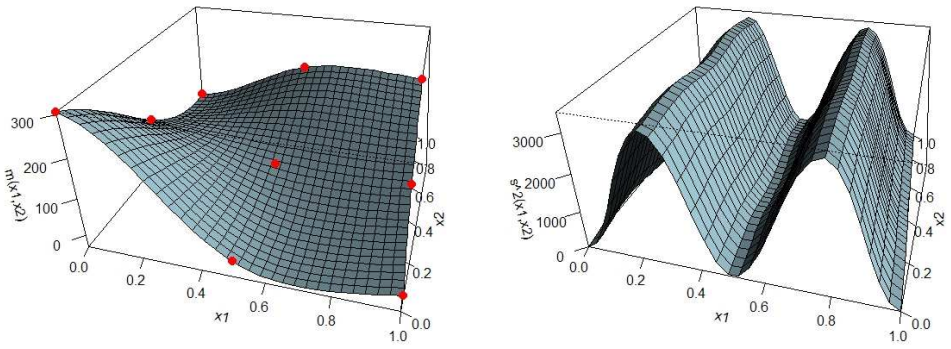


FIG. 3.11 – KO de la fonction de Branin-Hoo. La covariance est gaussienne anisotrope et ses paramètres sont estimés par maximum de vraisemblance à partir du plan d'expériences factoriel complet (points en rouge). A gauche : surface de moyenne de Krigeage. A droite : surface de variance de KO.

d'une part que la variance s'annule aux points du plan, et d'autre part que la surface ne varie pas de la même manière dans les directions des deux axes canoniques : cela reflète une anisotropie prononcée, conséquence des valeurs très dissemblables des deux paramètres de portée estimés par maximum de vraisemblance.

3.3 Krigeage, conditionnement des processus gaussiens, et méthodes bayésiennes

Les techniques de Krigeage ont connu un fort regain d'intérêt dans la littérature du *machine learning* depuis les années 1990, où elles sont présentées avec un vocabulaire et une interprétation sensiblement différents de ceux de la géostatistique. Le Krigeage tel qu'il est aujourd'hui employé —parfois sous d'autres noms— dans des contextes multidimensionnels (d dépasse largement 2 ou 3) s'appuie bien souvent sur des hypothèses fortes, telles que la "gaussiannité" du processus aléatoire duquel y est censé être une réalisation. Une manière d'élargir le champ de pertinence de cette hypothèse est de prendre en compte la méconnaissance que l'on a des paramètres de covariance, et d'adopter une démarche bayésienne les concernant. Cela revient à voir le processus Y comme un mélange continu de processus gaussiens.

3.3.1 Eléments de statistique bayésienne

La statistique bayésienne a pour objet l'étude des modèles paramétriques. Elle s'attache particulièrement à prendre en compte les incertitudes sur les paramètres de modèle, en substituant à la traditionnelle étape d'estimation paramètrique une intégration par rapport à une certaine loi de probabilité concernant les paramètres inconnus. La loi de probabilité en question dépend à la fois des observations collectées, et d'une *loi de probabilité a priori* reflétant l'idée que l'on se fait de la distribution des paramètres inconnus avant toute observation. Comme nous allons le voir ci-dessous, la pierre angulaire de la statistique Bayesienne est sans aucun doute la formule due au pasteur britannique Thomas Bayes ($1702 - 1761$), déjà rencontrée en présentant la notion de conditionnement :

$$\forall A, B \in \mathcal{A},\ \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3.74}$$

Remarquons avant d'aller plus loin que les critiques parfois formulées à l'encontre de l'approche bayésienne ciblent la notion de loi *a priori* mais ne remettent nullement en cause la formule de Bayes, ni le calcul conditionnel. Plaçons-nous maintenant dans la situation d'un observateur ayant recueilli des données $\mathbf{Y}$, supposées avoir été tirées au hasard selon une loi de probabilité P_ψ, où ψ est un paramètre fixe mais inconnu de l'observateur. On suppose de plus ici que la mesure de probabilité P_ψ est continue par rapport à la mesure dominante, et on note $p_\psi(\mathbf{Y})$ ou $p(\mathbf{Y}|\psi)$ la densité de probabilité associée. Cette densité, fonction des observations —connues ici—, peut aussi être vue comme une fonction du paramètre ψ : on appelle *vraisemblance* (notion longuement

commentée au chapitre 5) l'application

$$\begin{aligned} \Psi &\longrightarrow \mathbb{R} \\ \psi &\longrightarrow L(\psi;\mathbf{Y}) := p(\mathbf{Y}|\psi) \end{aligned} \tag{3.75}$$

Supposons maintenant que le paramètre ψ possède une loi a priori de densité $\pi(\psi)$. Le fait d'avoir observé $\mathbf{Y}$ nous permet d'affiner notre connaissance de ψ, et on obtient par conditionnement la *loi a posteriori de ψ sachant les observations* $\mathbf{Y}$:

$$\pi(\psi|\mathbf{Y}) := \frac{p(\mathbf{Y}|\psi)\pi(\psi)}{\mathbb{P}(\mathbf{Y})} = \frac{L(\psi;\mathbf{Y})\pi(\psi)}{\int_{\psi\in\Psi} L(\psi;\mathbf{Y})\pi(\psi)d\psi} \tag{3.76}$$

Remarquons que dans certaines situations, il n'est pas nécessaire que $\pi(.)$ soit une densité de probabilité bien définie (d'intégrale unité en particulier) pour que $\pi(.|\mathbf{Y})$ en soit une. Nous verrons à la fin de ce chapitre un exemple pour lequel $\int_\psi \pi(\psi)d\psi = +\infty$ et $\int_\psi \pi(\psi|\mathbf{Y})d\psi = 1$.

Considérons maintenant le cas —qui se pose souvent en pratique— d'une fonction de ψ notée $f : \psi \in \Psi \longrightarrow f(\psi) \in E$ (disons par soucis de simplicité que $E \subset \mathbb{R}$), dont on aimerait estimer la valeur $f(\psi)$ en un certain ψ, inconnu, mais pour lequel on a observé une réalisation $\mathbf{Y}$ tirée selon la loi P_ψ. Une approche souvent retenue en contexte classique, appelée *plug in*, consiste à faire une estimation $\psi^*(\mathbf{Y})$ de ψ (par exemple par maximum de vraisemblance, par méthode des moments, etc. Cf. [VdV98]) puis à injecter $\psi^*(\mathbf{Y})$ dans f, obtenant $f(\psi^*(\mathbf{Y}))$ comme estimation de $f(\psi)$. On constate que la réussite de cette approche est fortement dépendante de la qualité de l'estimation $\psi^*(\mathbf{Y})$, en regard bien sûr de la sensibilité de f au voisinage de ψ. L'approche bayésienne préfère au *plug in* une philosophie en un certain sens plus prudente : pour une loi *a priori* de densité π choisie et compte tenu d' observations $\mathbf{Y}$, $f(\psi)$ y est estimé par sa *moyenne a posteriori*

$$\boxed{\mathbb{E}\left[f(\psi)|\mathbf{Y}\right] = \int_\Psi f(\psi)\pi(\psi|\mathbf{Y})d\psi = \frac{\int_\Psi f(\psi)L(\psi;\mathbf{Y})\pi(\psi)d\psi}{\int_\Psi L(\psi;\mathbf{Y})\pi(\psi)d\psi}} \tag{3.77}$$

Ainsi la valeur ponctuelle $f(\psi^*(\mathbf{Y}))$ est-elle remplacée par une moyenne pondérée de f, où la pondération proportionnelle à $L(\psi;\mathbf{Y})\pi(\psi)$ favorise les différentes valeurs de ψ suivant un compromis entre leur vraisembance et la valeur de densité *a priori* qui leur est associée. Si π n'est pas mal choisie, la moyenne *a posteriori* de ψ lorsque le nombre de données est peu élevé et que la maximisation de $L(.;\mathbf{Y})$ n'est donc pas très fiable permet de ne pas être tributaire des fluctuations d'un estimateur par *plug in*, et de prendre en compte notre incapacité à connaître ψ précisément en intégrant f par

rapport à une mesure d'autant plus proche de π que $L(\psi; \mathbf{Y})$ est plate. A l'inverse lorsque beaucoup de données sont disponibles, la vraisemblance $L(.; \mathbf{Y})$ est censée se concentrer autour de ψ (Cf. chapitre 5), et le produit $L(.; \mathbf{Y})\pi(.)$ se rapprocher d'un Dirac en ψ, c'est à dire la moyenne *a posteriori* se rapprocher de l'estimation *plug in* par maximum de vraisemblance lorsque les conditions sont favorables pour cette dernière. L'approche bayésienne apparaît ainsi en premier lieu comme un moyen de se protéger contre les flucuations trop importantes dans le cas d'estimateurs *plug in* utilisés avec un nombre insuffisant d'observations. Elle offre aussi un cadre très pratique pour quantifier l'incertitude associée à l'estimation, en donnant l'accès à la loi $f(\psi)|\mathbf{Y}$, ce qui permet notamment le calcul d'indicateurs de précision tels que $Var[f(\psi)|\mathbf{Y}]$.

Il y a évidemment un prix à payer pour une utilisation réussie de l'approche bayésienne. D'une part, le choix de π n'est pas du tout neutre, et un choix inadéquat peut mener à des résultats décevants, voire très mauvais dans la situation où peu de données sont disponibles. C'est là probablement le talon d'Achille bayésien, et un argument de poids sous la plume des détracteurs de cette approche. D'autre part, le calcul de la loi *a posteriori* et l'intégration par rapport à cette dernière ont un coût computationnel non-négligeable, qui font du calcul bayésien un outil relativement délicat à mettre en oeuvre.

- Le premier point a fait l'objet de nombreuses contributions, au sujet desquelles l'ouvrage de référence [Rob92] donne une vision de synthèse. La question du choix du *prior* soulève en effet plusieurs types de problèmes concernant la calculabilité des lois *a posteriori* (notion de *lois conjuguées*, Cf. [Rob92]), l'injection d'informations « métiers » à l'étape de définition de la loi *a priori*, ou à l'extrême inverse la recherche de lois *a priori* dites *non-informatives*, et parmi ces dernières des lois possédant si possible la propriété d'*invariance par reparamétrisation*. Nous aborderons de nouveau ces questions au sujet de l'interprétation bayésienne du Krigeage (fin de ce chapitre) et des mélanges de modèles (chapitre 8).

- Le deuxième point est parfois contourné en évitant le calcul complet de la moyenne *a posteriori* et en lui substituant un *plug in* du *mode a posteriori*, encore appelé *maximum a posteriori* (MAP) : le MAP est tout simplement un maximiseur de $L(.; \mathbf{Y})\pi(.)$. La méthode MAP apparaît ainsi comme une hybridation du *plug in* et de l'approche bayésienne : on injecte bien une seule valeur estimée du paramètre ψ, mais elle dépend d'une loi *a priori* (On peut en fait même dire qu'il s'agit d'un simple *plug in* avec un estimateur particulier mêlant vraisemblance et *a priori*). Remarquons enfin à ce sujet que la maximisation de vraisemblance revient à une estimation par MAP avec un *prior impropre uniforme* sur Ψ, i.e. $\forall \psi \in \Psi,\ \pi(\psi) = 1$.

3.3.2 Krigeage(s) et conditionnement de PG : approche bayésienne

Avantages de l'hypothèse gaussienne pour la prédiction linéaire

On a vu dans la section 3.2.1 que la prédiction par Krigeage Simple (à moyenne nulle) reposait sur la recherche de la meilleure combinaison linéaire $\hat{Y}(\mathbf{x}^0)$ en les observations $Y(\mathbf{X})$ pour prédire $Y(\mathbf{x}^0)$:

$$\hat{Y}(\mathbf{x}^0) = \sum_{j=1}^{n} \lambda_j Y(\mathbf{x}^j) \text{ où } \lambda = argmin_\lambda \mathbb{E}\left[\left(Y(\mathbf{x}^0) - \sum_{j=1}^{n} \lambda_j Y_j\right)^2\right] \tag{3.78}$$

D'après les définitions introduites dans la section 3.1.3, cela correspond à rechercher l'espérance conditionnelle linéaire $\mathbb{E}_L[Y(\mathbf{x}^0)|Y(\mathbb{X})]$. On a vu d'autre part (Cf. 3.23) que l'espérance conditionnelle linéaire coïncide dans le cas gaussien avec l'espérance conditionnelle $\mathbb{E}[Y(\mathbf{x}^0)|Y(\mathbb{X})]$, et que le fait de conditionner un processus gaussien par rapport à un ensemble fini d'observations ponctuelles donne un « processus conditionnel » lui-même gaussien. Lorsque l'on fait l'hypothèse que Y est gaussien, on retrouve ainsi non-seulement le prédicteur de Krigeage comme espérance conditionnelle, mais aussi la loi conditionnelle du processus sachant les observations. En particulier, la variance de Krigeage apparaît alors comme une variance conditionnelle

$$Var[Y(\mathbf{x}^0)|Y(\mathbb{X})] = Var[Y_{\mathbf{x}^0} - \lambda^T \mathbf{Y}] = k(\mathbf{0}) - \mathbf{k}(\mathbf{x}^0)^T K^{-1} \mathbf{k}(\mathbf{x}^0) \tag{3.79}$$

De plus, la loi de conditionnelle de $Y(\mathbf{x}^0)|Y(\mathbb{X})$ permet d'accompagner les prédictions d' intervalles de confiance, ainsi que de faire des simulations conditionnelles (Cf. 3.12, et [Rip87, MS04b, RW06, Sch01]. Nous proposons ci-après de revisiter le Krigeage Simple avec l'hypothèse gaussienne, puis détaillons une interprétation bayésienne du KO et du KU permettant d'obtenir des lois conditionnelles intégrant le fait que certains paramètres du modèle de Krigeage sont inconnus. Ces lois conditionnelles nous permettront dans le chapitre 4 de calculer explicitement certains critères d'exploration voués à l'optimisation de y, tels que l'*amélioration espérée*.

Une ré-interprétation du Krigeage Simple

On fait l'hypothèse que

$$Y \sim \mathcal{PG}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{3.80}$$

où $\mu : \mathbf{x} \in D \longrightarrow \mu(\mathbf{x}) \in \mathbb{R}$ est une fonction connue, et $k : (\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow k(\mathbf{x}, \mathbf{x}') \in \mathbb{R}$ est un noyau de type positif connu, auquel nous n'imposons pas ici d'être stationnaire. Supposons que ce processus Y ait été observé en $Y(\mathbf{X}) = \big(Y(\mathbf{x}^1), Y(\mathbf{x}^2), ..., Y(\mathbf{x}^n)\big)$, et

que l'on souhaite dans un premier temps prédire sa valeur en un point $\mathbf{x}^0$. On obtient alors la loi conditionnelle

$$[Y(\mathbf{x}^0)|Y(\mathbb{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{KS}(\mathbf{x}^0), s^2_{KS}(\mathbf{x}^0)) \tag{3.81}$$

où l'espérance conditionnelle m_{KS} et la variance conditionnelle s^2_{KS} sont données par les équations du Krigeage Simple :

$$m_{KS}(\mathbf{x}^0) = \mu(\mathbf{x}^0) + \mathbf{k}(\mathbf{x}^0)^T K^{-1}(\mathbf{Y} - \mu(\mathbf{X})) \tag{3.82}$$

$$s^2_{KS}(\mathbf{x}^0) = k(\mathbf{x}^0, \mathbf{x}^0) - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbf{k}(\mathbf{x}^0) \tag{3.83}$$

Pour des prédictions ponctuelles, le résultat donné par 3.81 permet en fait de connaître des informations bien plus précises qu'une simple prédiction par moyenne de Krigeage. On peut par exemple affirmer que

$$\mathbb{P}\left(Y(\mathbf{x}^0) \in [m_{KS}(\mathbf{x}^0) - 1.96 s_{KS}(\mathbf{x}^0), m_{KS}(\mathbf{x}^0) + 1.96 s_{KS}(\mathbf{x}^0)] | Y(\mathbb{X}) = \mathbf{Y}\right) \approx 0.95 \tag{3.84}$$

Ce qui signifie que conditionnellement à $Y(\mathbf{X}) = \mathbf{Y}$), Y prend en $\mathbf{x}^0$ une valeur dans l'intervalle $[m_{KS}(\mathbf{x}^0) - 1.96 s_{KS}(\mathbf{x}^0), m_{KS}(\mathbf{x}^0) + 1.96 s_{KS}(\mathbf{x}^0)]$ avec une probabilité proche de 95%. Il est important de préciser que cette probabilité n'a de sens que sur un très grand nombre de répétitions de $Y(\mathbf{x}^0)|Y(\mathbb{X}) = \mathbf{Y}$. Elle ne signifie pas du tout que pour une réalisation donnée de Y, les valeurs réalisées des $Y(\mathbf{x})$ sont dans les intervalles de confiance sur 95% du domaine D (Cf. fig. 3.12 pour une illustration).

Pour aller au bout de l'interprétation du KS comme conditionnement d'un processus gaussien, rappelons que le « processus conditionnel » $Y(\mathbf{x}^0)|Y(\mathbb{X}) = \mathbf{Y}$ est nécessairement gaussien. On peut alors dire beaucoup plus que les lois marginales de 3.81 : pour un jeu d'observations $(\mathbb{X}, \mathbf{Y})$ fixé, il existe un noyau de type positif k_{KS} : $(\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow k_{KS}(\mathbf{x}, \mathbf{x}') \in \mathbb{R}$ tel que

$$[Y|Y(\mathbb{X}) = \mathbf{Y}] \sim \mathcal{PG}(\mu(\mathbf{x}), k_{KS}(\mathbf{x}, \mathbf{x}')) \tag{3.85}$$

Pour calculer ce noyau, considérons deux points $\mathbf{x}, \mathbf{x}' \in D$ et calculons la covariance $k_{KS}(\mathbf{x}, \mathbf{x}') = Cov[Y(\mathbf{x}), Y(\mathbf{x}')|Y(\mathbb{X}) = \mathbf{Y}]$. En se souvenant des règles matricielles pour le conditionnement des vecteurs gaussiens, nous savons écrire la matrice de covariance conditionnelle du vecteur d'observations $(Y(\mathbf{x}), Y(\mathbf{x}'))$ sachant $Y(\mathbf{X}) = \mathbf{Y}$:

$$\begin{aligned} Var[Y(\mathbf{x}), Y(\mathbf{x}')|Y(\mathbb{X}) = \mathbf{Y}] &= K_{new} - \left[\mathbf{k}(\mathbf{x}), \mathbf{k}(\mathbf{x}')\right]^T K^{-1} \left[\mathbf{k}(\mathbf{x}), \mathbf{k}(\mathbf{x}')\right] \\ &= K_{new} - \begin{pmatrix} \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}) & \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}') \\ \mathbf{k}(\mathbf{x}')^T K^{-1}\mathbf{k}(\mathbf{x}) & \mathbf{k}(\mathbf{x}')^T K^{-1}\mathbf{k}(\mathbf{x}') \end{pmatrix} \end{aligned} \tag{3.86}$$

où $K_{new} = \begin{pmatrix} k(\mathbf{x},\mathbf{x}) & k(\mathbf{x},\mathbf{x}') \\ k(\mathbf{x}',\mathbf{x}) & k(\mathbf{x}',\mathbf{x}') \end{pmatrix}$ et $\mathbf{k}(\mathbf{x}) = \begin{pmatrix} k(\mathbf{x},\mathbf{x}^1) & \dots & k(\mathbf{x},\mathbf{x}^n) \end{pmatrix}^T$.

On obtient ainsi l'expression du noyau de covariance conditionnelle associé au KS :

$$k_{KS}(\mathbf{x},\mathbf{x}') = k(\mathbf{x}',\mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1} \mathbf{k}(\mathbf{x}'), \tag{3.87}$$

et on peut vérifier au passage que l'on retrouve bien l'expression de la variance conditionnelle du Krigeage Simple en prenant $\mathbf{x} = \mathbf{x}'$ dans 3.87. Cette dernière équation nous donne accès à la loi conditionnelle explicite du processus sachant les observation, ce qui va nous permettre entre autres de réaliser des simulations conditionnelles (Cf. 3.12 pour un exemple de simulation conditionnelle en Krigeage Ordinaire).

Dans la pratique, il est souvent peu raisonnable de supposer que le processus est centré ou bien que sa moyenne est connue. On doit alors estimer sa moyenne dans la foulée, ce qui a une incidence sur la loi du prédicteur de Krigeage. Cette incidence peut pour autant être maîtrisée dans certains cas particuliers : lorsque μ est constante et inconnue ou plus généralement lorsque μ est linéaire en des fonctions connues de $\mathbf{x}$ mais que les coefficients sont inconnus. Il s'agit respectivement des Krigeages Ordinaire et Universel.

Ré-interprétation des Krigeages Ordinaire et Universel

Comme on l'a vu dans la section 3.2.2, les Krigeages Ordinaire et Universel se distinguent du Krigeage Simple par l'intégration d'une étape d'estimation de la moyenne $\mu(\mathbf{x})$ du processus Y. Dans ces deux cas, la moyenne est supposée inconnue mais de forme paramétrique simple. En KU, $\mu(\mathbf{x})$ est supposée être une combinaison linéaire de b fonctions de base connues : $\mu(\mathbf{x}) = \sum_{j=1}^{b} \beta_j f_j(\mathbf{x})$. On rappele qu'en KO, on suppose que $\mu(\mathbf{x}) = \mu \in \mathbb{R}$, ce qui revient à faire du KU avec une seule fonction de base $f_1 \equiv 1$ et un paramètre inconnu $\beta_1 = \mu$. Le fait d'avoir à estimer les paramètres de tendance est loin d'être sans effet sur les prédictions par Krigeage. Nous allons voir qu'en nous plaçant dans un cadre bayésien, et avec des lois *a priori* bien particulières, il est possible de propager l'incertitude sur les paramètres inconnus tout en conservant la propriété commode d'avoir une loi conditionnelle explicite pour le Krigeage.

Intéressons-nous dans un premier temps au KO, en faisant l'hypothèse que Y est un processus gaussien de noyau connu, et que la moyenne $\mu \in \mathbb{R}$ a pour loi *a priori* $\mu \sim \Pi$, où Π est une loi continue de densité π :

$$\begin{cases} Y \sim \mathcal{PG}(\mu, k(\mathbf{x},\mathbf{x}')) \\ \mu \sim \Pi, \text{ loi de densité } \pi(\mu) \end{cases} \tag{3.88}$$

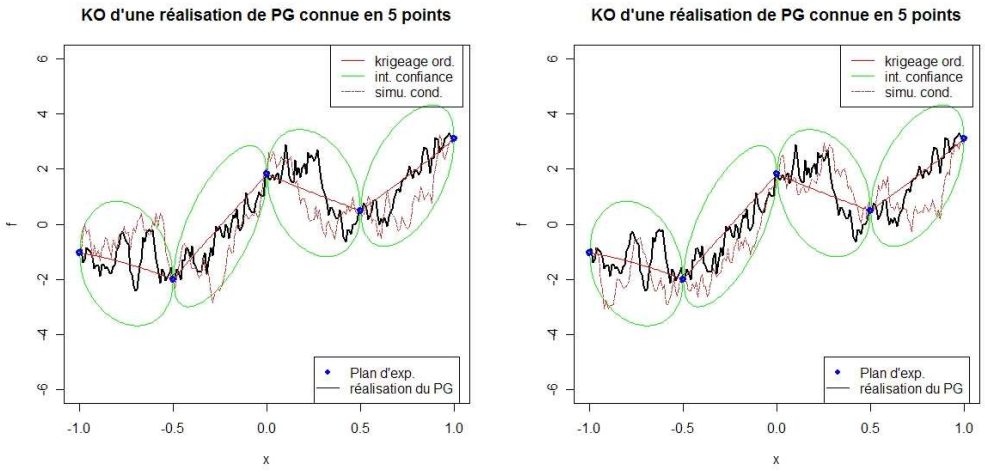


FIG. 3.12 – Krigeage Ordinaire et Simulation Conditionnelle d'une réalisation de Processus Gaussien centré de covariance exponentielle (Ornstein-Uhlenbeck). La simulation de départ ainsi que les deux simulations conditionnelles ont été produites à l'aide du package R « Random Fields »[Sch01], avec des paramètres $\sigma^2 = 4$ et $l = 0.8$.

On souhaite connaître la loi du processus $[Y|Y(\mathbb{X} = \mathbf{Y})]$ sous ces hypothèses. Il paraît alors raisonnable de se ramener à des résultats connus en reliant la loi de $Y|Y(\mathbf{X}) = \mathbf{Y}$ à celle de $Y|Y(\mathbf{X}) = \mathbf{Y}, \mu$, ce qui peut se faire par l'intégration suivante :

$$f(Y(\mathbf{x}^0)|Y(\mathbb{X}) = \mathbf{Y}) = \int_{\mu\in\mathbb{R}} f(Y(\mathbf{x}^0)|Y(\mathbb{X}) = \mathbf{Y}, \mu)\pi(\mu|Y(\mathbb{X}) = \mathbf{Y})d\mu \tag{3.89}$$

où la lettre f est utilisée pour désigner des densités de probabilité, et $\pi(\mu|Y(\mathbb{X}) = \mathbf{Y})$ est la loi *a posteriori* de μ sachant les observations. L'équation 3.91 a le grand avantage de faire apparaître la loi $[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}, \mu]$, qui est une gaussienne bien connue puisqu'elle nous ramène aux conditions précédentes du Krigeage Simple. La question est maintenant de savoir ce que vaut le *posterior* $\pi(\mu|Y(\mathbb{X}) = \mathbf{Y})$ en fonction du *prior* π, et de la nature de la loi conditionnelle finale du processus —présentée comme une intégrale dans 3.91— en fonction de $\pi(\mu|Y(\mathbb{X}) = \mathbf{Y})$. La propriété suivante illustre le cas particulièrement favorable d'un *prior* bien spécial, avec lequel on retrouve des résultats analogues à ceux du Krigeage Simple.

Propriété. *Lorsque* Π *est la loi impropre uniforme sur* $\mathbb{R}$ *i.e.* $\forall\mu \in \mathbb{R}\ \pi(\mu) = 1$*, la loi a posteriori de* μ *est gaussienne de loi*

$$[\mu|Y(\mathbb{X}) = \mathbf{Y}] \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right) \overset{déf}{\equiv} \mathcal{N}\left(\frac{\mathbb{1}^T K^{-1}\mathbf{Y}}{\mathbb{1}^T K^{-1}\mathbb{1}}, \frac{1}{\mathbb{1}^T K^{-1}\mathbb{1}}\right) \tag{3.90}$$

De plus, la loi de $Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}$ *est gaussienne*

$$[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{KO}(\mathbf{x}^0), s^2_{KO}(\mathbf{x}^0)) \tag{3.91}$$

où l'espérance conditionnelle $m_{KO} = \mathbb{E}[Y(\mathbf{x}^0)|Y(\mathbb{X}) = \mathbf{Y}]$ *et la variance conditionnelle* $s^2_{KO} = var[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}]$ *sont données par les équations du Krigeage Ordinaire :*

$$m_{KO}(\mathbf{x}^0) = \bar{\mu} + \mathbf{k}(\mathbf{x}^0)^T K^{-1}(\mathbf{Y} - \bar{\mu}\mathbb{1}_n) \tag{3.92}$$

$$s^2_{KO}(\mathbf{x}^0) = k(\mathbf{x}^0, \mathbf{x}^0) - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbf{k}(\mathbf{x}^0) + \frac{(1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}^0))^2}{\mathbb{1}_n^T K^{-1}\mathbb{1}_n} \tag{3.93}$$

De manière plus générale, $Y|Y(\mathbb{X}) = \mathbf{Y}$ *est un PG de moyenne* m_{KO} *et de noyau*

$$\begin{aligned} k_{KO}(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}', \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}') + \frac{(1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}))(1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}'))}{\mathbb{1}_n^T K^{-1}\mathbb{1}_n} \\ &= k_{KS}(\mathbf{x}', \mathbf{x}) + \frac{(1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}))(1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}'))}{\mathbb{1}_n^T K^{-1}\mathbb{1}_n} \end{aligned} \tag{3.94}$$

Démonstration. On a tout d'abord par application de l'éq. 3.76 que [12]

$$\begin{aligned} f(\mu|Y(\mathbf{X}) = \mathbf{Y}) &\propto \pi(\mu) \times f(Y(\mathbf{X}) = \mathbf{Y}|\mu) \\ &\propto 1 \times e^{-\frac{1}{2}(\mathbf{Y} - \mu\mathbb{1}_n)^T K^{-1}(\mathbf{Y} - \mu\mathbb{1}_n)} \end{aligned}$$

En remarquant que

$$\begin{aligned} (\mathbf{Y} - \mu\mathbb{1}_n)^T K{-}1(\mathbf{Y} - \mu\mathbb{1}_n) &= \mu^2 \mathbb{1}_n^T K^{-1}\mathbb{1}_n - 2\mu\mathbb{1}_n^T K^{-1}\mathbf{Y} + \mathbf{Y}K^{-1}\mathbf{Y} \\ &= (\mathbb{1}_n^T K^{-1}\mathbb{1}_n) \times \left(\mu - \frac{\mathbb{1}_n^T K^{-1}\mathbf{Y}}{\mathbb{1}_n^T K^{-1}\mathbb{1}_n}\right), \end{aligned}$$

il vient que $[\mu|Y(\mathbf{X}) = \mathbf{Y}]$ est gaussienne, de loi notée $\mathcal{N}\left(\bar{\mu}, \sigma^2_\mu\right)$ où l'on obtient $\bar{\mu} = \frac{\mathbb{1}^T K^{-1}\mathbf{Y}}{\mathbb{1}^T K^{-1}\mathbb{1}}$ et $\sigma^2_\mu = (\mathbb{1}^T K^{-1}\mathbb{1})^{-1}$ par identification. Focalisons maintenant notre attention sur les lois conditionnelles du Krigeage Ordinaire, et considérons pour ce faire un ensemble de points $\mathbf{X}_{new} = \{\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+q}\}$ ($q \in \mathbb{N} - \{0\}$) auxquels on souhaite prédire les valeurs prises par le processus Y. Dans les conditions où μ est connu, on peut se ramener au Krigeage Simple et obtenir que $[[Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}]|\mu] \equiv [Y(\mathbb{X}_{new})|Y(\mathbb{X}) = \mathbf{Y}, \mu]$ est gaussien de loi connue, d'espérance affine en μ et de variance indépendante de μ (Cf. paragraphes précédents). Par ailleurs, on vient de voir que $[\mu|Y(\mathbf{X}) = \mathbf{Y}]$ est gaussien.

[12] On ne s'intéresse pas au dénominateur de 3.76, qui ne dépend pas de μ et joue ici un rôle de coefficient de normalisation. On s'intéressera de plusieurs fois dans la suite à des relations de proportionnalité entre lois, en négligeant les termes multiplicatifs ne dépendant pas de la variable d'intérêt.

On peut montrer par application du lemme présenté ci-dessous (éq. (3.95) en prenant $Y_1 = [\mu|Y(\mathbf{X}) = \mathbf{Y}]$ et $Y_2 = [Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}]$) que le vecteur $(Y(\mathbf{X}_{new}), \mu)|Y(\mathbf{X}) = \mathbf{Y}$ est gaussien. Il ressort en particulier que $Y(\mathbf{X}_{new})|Y(\mathbf{X}) = \mathbf{Y}$ est gaussien, ce qui suffit au passage à montrer que le processus $Y|Y(\mathbf{X}) = \mathbf{Y}$ est un PG. Il reste alors à trouver les moments de ce PG, ce que l'on fait ci-dessous en utilisant les formules de l'espérance et de la covariance totales (3.18 & 3.19). Pour tout $\mathbf{x}$ dans D, on a :

$$\begin{aligned} m_{OK}(\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \mathbb{E}\left[\mathbb{E}[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] \\ &= \mathbb{E}\left[\mu + \mathbf{c}(\mathbf{x})^T K^{-1}(\mathbf{Y} - \mu \mathbb{1}_n)\right] = \bar{\mu} + \mathbf{c}(\mathbf{x})^T K^{-1}(\mathbf{Y} - \bar{\mu}\mathbb{1}_n) \end{aligned}$$

Le noyau de covariance conditionnelle est donné par

$$\begin{aligned} k_{KO}(\mathbf{x}, \mathbf{x}') &= cov[Y(\mathbf{x}), Y(\mathbf{x}')|Y(\mathbf{X}) = \mathbf{Y}] \\ &= \mathbb{E}\left[cov[Y(\mathbf{x}), Y(\mathbf{x}')|Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] \\ &+ cov\left[\mathbb{E}[Y(\mathbf{x})|Y(\mathbf{X}), \mu], \mathbb{E}[Y(\mathbf{x}')|Y(\mathbf{X}), \mu]\right] \\ &= \mathbb{E}\left[k_{kS}(\mathbf{x}, \mathbf{x}')\right] \\ &+ cov[\mu + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbf{Y} - \mu\mathbb{1}_n), \mu + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbf{Y} - \mu\mathbb{1}_n)|\mu Y(\mathbf{X}) = \mathbf{Y}] \\ &= k_{kS}(\mathbf{x}, \mathbf{x}') + var[\mu|Y(\mathbf{X}) = \mathbf{Y}](1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}))(1 - \mathbb{1}_n^T K^{-1}\mathbf{k}(\mathbf{x}')) \end{aligned}$$

Et on trouve bien le résultat annoncé pour k_{KO}. 3.91 est une conséquence directe de la gaussianité du processus $Y|Y(\mathbf{X}) = \mathbf{Y}$, où l'expression s^2_{OK} découle de 3.94. □

Lemme. *Si $Y_1 \sim \mathcal{N}(m_1, K_1)$ et $Y_2|Y_1 \sim \mathcal{N}(AY_1 + b, K)$, (Y_1, Y_2) est un vecteur gaussien, de loi*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_1 \\ Am_1 + b \end{pmatrix}, \begin{pmatrix} K_1 & K_1 A^T \\ AK_1^T & K + AK_1^T A^T \end{pmatrix}\right) \tag{3.95}$$

Démonstration. Cf. annexe 12.2.2. □

Pour finir, la construction précédente permet d'obtenir une distribution conditionnelle gaussienne de manière similaire pour le Krigeage Universel, modulo un *prior* impropre uniforme sur l'ensemble des coefficients de la tendance linéaire. On rappelle que $\mathcal{F} = \{f_1, ..., f_b\}$ est la famille de fonctions choisies, $f(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_b(\mathbf{x}))^T$ la fonction vectorielle de composantes les f_j ($j \in [1, b]$), et $\mathbb{F} = (f_j(\mathbf{x}^i))_{i\in[1,n], j\in[1,b]}$ la matrice d'expériences.

Propriété. *Lorsque Π est la loi impropre uniforme sur $\mathbb{R}^b$ i.e. $\forall \beta \in \mathbb{R}^b$ $\pi(\beta) = 1$, la loi a posteriori de β est gaussienne de loi*

$$[\beta|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}\left(\bar{\beta}, K_\beta\right) \overset{déf}{\equiv} \mathcal{N}\left(\frac{\mathbb{F}^T K^{-1}\mathbf{Y}}{\mathbb{F}^T K^{-1}\mathbb{F}}, (\mathbb{F}^T K^{-1}\mathbb{F})^{-1}\right) \tag{3.96}$$

De plus, la loi de $Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}$ *est gaussienne*

$$[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{KU}(\mathbf{x}^0), s^2_{KU}(\mathbf{x}^0)) \tag{3.97}$$

où l'espérance conditionnelle $m_{KU} = \mathbb{E}[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}]$ *et la variance conditionnelle* $s^2_{KU} = var[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}]$ *sont donnés par les équations du Krigeage Universel :*

$$m_{KU}(\mathbf{x}^0) = f(\mathbf{x}^0)^T\bar{\beta} + \mathbf{k}(\mathbf{x}^0)^T K^{-1}(\mathbf{Y} - f(\mathbf{x}^0)^T\bar{\beta}) \tag{3.98}$$

$$\begin{aligned} s^2_{KU}(\mathbf{x}^0) = {} & k(\mathbf{x}^0, \mathbf{x}^0) - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbf{k}(\mathbf{x}^0) \\ & + (f(\mathbf{x}^0)^T - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbb{F})(\mathbb{F}^T K^{-1}\mathbb{F})^{-1}(f(\mathbf{x}^0)^T - \mathbf{k}(\mathbf{x}^0)^T K^{-1}\mathbb{F})^T \end{aligned} \tag{3.99}$$

De manière plus générale, $Y|Y(\mathbf{X}) = \mathbf{Y}$ *est un PG de moyenne* m_{KU} *et de noyau*

$$\begin{aligned} k_{KU}(\mathbf{x}, \mathbf{x}') = {} & k(\mathbf{x}', \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}') \\ & + (f(\mathbf{x})^T - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbb{F})(\mathbb{F}^T K^{-1}\mathbb{F})^{-1}(f(\mathbf{x}')^T - \mathbf{k}(\mathbf{x}')^T K^{-1}\mathbb{F})^T \end{aligned} \tag{3.100}$$

Démonstration. Similaire à celle de la propriété précédente sur l'interprétation bayésienne du Krigeage Ordinaire. □

On remarque que les expressions trouvées pour le Krigeage Universel dans le cas où la seule fonction considérée est $f_1 \equiv 1$ coïncident bien avec celles du Krigeage Ordinaire.

Discussion : limites de validité de ces modélisations

Nous venons dans cette partie de montrer comment l'ajout de deux hypothèses, la gaussiannité de $Y|\beta$ et une loi *a priori* uniforme impropre sur β, permettait de donner au Krigeage une signification profonde en rapport avec le conditionnement, et lui faisant largement dépasser le statut —déjà honorable— de prédicteur linéaire sans biais de variance minimale sous lequel nous l'avions présenté au départ, dans le cadre de l'étude de processus aléatoires de carré intégrable (resp. intrinsèquement stationnaires). Il faut cependant bien garder à l'esprit que ces hypothèses sont fortes, souvent notoirement abusives, et davantage introduites de manière à pouvoir construire et/ou justifier l'emploi de critères d'échantillonnage calculables analytiquement (tels que présentés dans la suite de cette thèse) ou encore de permettre des simulations conditionnelles à coût computationnel abordable, plutôt que pour faire des prédictions de grande précision destinées à être employées ensuite sans réserve par des utilisateurs finaux.

Pour prendre un exemple concret, le choix d'un prior impropre sur μ dans le modèle de Krigeage Ordinaire implique que l'on considère a priori que $Y = \mu + \delta$ est non-gaussien et surtout non-intégrable, ce qui reflète assez peu souvent la volonté du modélisateur. Ce même modélisateur est en revanche plus souvent enclin à accepter un résultat de type $[Y(\mathbf{x}^0)|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{KO}(\mathbf{x}^0), s^2_{KO}(\mathbf{x}^0))$ lorsqu'il s'agit de concevoir des critères pour l'évaluation de plans d'expériences (la loi conditionnelle servant alors plutôt comme aide à la décision pour guider de nouvelles évaluations de y). Un des objectifs de la section qui précède était de clarifier le lien entre les hypothèses et les résultats, qui respectivement sous-tendent et découlent de ces choix de modélisation.

Remarquons enfin que la démarche bayésienne présentée ici concernant uniquement des paramètres de tendance linéaire, et en se restreignant à des lois a priori bien particulières, est en fait complètement généralisable [MS04b, O'H06], et mène au Krigeage Bayésien (KB), à la fois beaucoup plus modulaire que les modèles classiques et source de dilemmes pour le choix des prior ainsi que de nouvelles difficultés computationnelles (MCMC). Le modèle de mélange discret présenté au chapitre 8 et dans [GHC08] en est une instance simplifiée, traitant toutefois plus spécifiquement de structures de modèles que de valeurs de paramètres continus.

Chapitre 4

Optimiser avec un métamodèle

L'optimisation de fonctions coûteuses à évaluer ne permet généralement pas l'utilisation des méthodes tradionnelles reposant sur le calcul différentiel (descentes de gradient de type BFGS, quasi-Newton, etc. Cf. annexes) ni des algorithmes de recherche aléatoire très gourmands en nombre d'évaluations (stratégies évolutionnaires, recuit simulé, etc.) : optimiser une fonction sans aucune information différentielle et en devant payer chaque évaluation au prix fort requiert l'utilisation de techniques *ad hoc* (Cf. [KO96]).

Nous considérons dans ce chapitre des méthodes d'optimisation non-contrainte sur base de métamodèles globaux, en portant particulièrement notre attention sur le cas où aucune information n'est disponible sur les gradients et les différentielles d'ordre supérieurs de la fonction objectif (méthodes d'ordre 0). Le chapitre se découpe en trois sections. Les deux premières comparent les approches par surfaces de réponses déterministes (Cf. chap. 2) versus par métamodèles probabilistes (Cf. chap. 3) : on y montre que l'optimisation basée sur la substitution de la fonction objectif par une surface de réponse déterministe présente des modes de défaillance majeurs, puis en quoi l'approche probabiliste permet de les contourner. On étudie enfin dans la troisième section l'algorithme d'optimisation globale EGO, basé sur le Krigeage Ordinaire, et point de départ de plusieurs développements présentés dans la partie 3 de cette thèse (Cf. chap. 8 et 9).

Précisons d'emblée que l'approche employée ici et tout au long de ce mémoire su sujet de l'optimisation de fonctions coûteuses est volontairement restrictive, à la fois pour retranscrire une certaine réalité applicative et pour des raisons pratiques : dans presque tous les cas, les algorithmes étudiés le seront en fixant le nombre d'évaluations à l'avance, et le critère employé pour juger de leurs performances sera la distance entre la meilleure valeur trouvée de y et l'optimum global, i.e. $min_{\mathbf{x}\in\mathbf{X}}(y(\mathbf{x})) - min_{\mathbf{x}\in D}(y(\mathbf{x}))$.

4.1 Optimisation sur base de métamodèles déterministes

Comme décrit dans [QHS+05], l'optimisation non-contrainte de base assistée par métamodèle (*Basic unconstrained Surrogate Based Analysis and Optimization*) peut être résumée sous la forme algorithmique suivante :

1. Construction d'un métamodèle à partir d'un plan d'expériences initial
2. Estimation du minimum de la fonction objectif en utilisant le métamodèle
3. Evaluation de la fonction objectif au minimum estimé précédemment
4. Test de convergence. Arrêt de l'algorithme si la convergence est (jugée) atteinte.
5. Actualisation du métamodèle en utilisant les nouvelles observations.
6. Itération jusqu'à convergence.

Nous nous intéressons dans cette section à de tels algorithmes, faisant intervenir toutes sortes de métamodèles déterministes, et où la phase 2 consiste en une optimisation pure et simple du métamodèle. Aprés avoir évoqué quelques techniques reposant sur des métamodèles de l'état de l'art, nous discuterons la pertinence de l'approche déterministe pour optimiser sur la base d'un métamodèle global[1].

4.1.1 Quelques exemples

L'article [Jon01], publié par Donald Jones en 2001, propose une taxonomie des méthodes d'optimisation sur base de surfaces de réponse. Il y distingue l'optimisation sur base d'approximateurs (surfaces de réponses polynômiales et autres régressions linéaires, etc.), dite de type 1, et celle sur base d'interpolateurs (RBF, splines, moyenne de Krigeage), de type 2. Nous donnons ici quelques exemples de type 1, laissant le lecteur consulter l'excellent [Jon01] pour s'assurer que l'attitude de réserve que l'on recommande pour le type 1 reste bien pertinente pour le type 2.

Minimisation directe d'une surface de réponse polynômiale

Les surfaces de réponses polynômiales, que nous avons déja abordées aux chapitre 2, constituent sans doute les surfaces de réponse les plus populaires en ingénierie. Une fois estimée à partir des observations au plan d'expériences, il n'est pas rare que l'on utilise la surface de réponse m directement comme un substitut à la fonction y, que cela soit pour l'optimisation ou pour toutes autres opérations dépendant de y. Voyons maintenant sur

[1] global par opposition aux stratégies basées sur des approximations locales, ce que nous ne traiterons pas ici. Remarquons que bon nombre des méthodes traditionnelles citées en annexe reposent sur des développements limités à l'ordre 1 ou 2, ce que l'on peut assimiler à des métamodèles locaux.

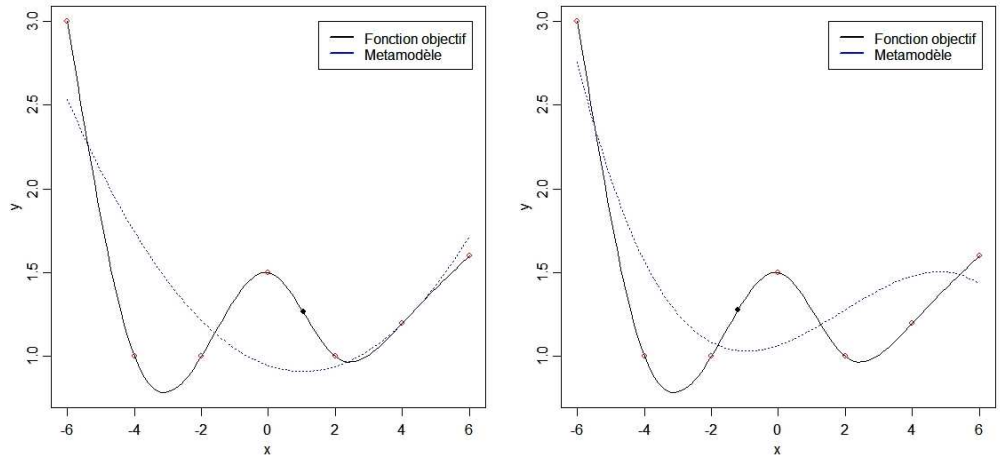


FIG. 4.1 – Approximation par surfaces de réponses polynômiales (en bleu) de degrés 2 (à gauche) et 3 (à droite) d'une fonction déterministe (une spline, en noir), en considérant comme plan d'expériences (en rouge) 7 points équi-espacés de l'intervalle d'étude $[-6, 6]$. Les points noirs représentent le minimum global de chacune des approximations.

quelques exemples simples si cette approche est bien raisonnable.

Considérons tout d'abord le cas d'une fonction objectif y déterministe, mono-dimensionnelle, connue en 7 points équi-espacés, et approchée par des surfaces de réponse polynômiales de degrés 2 et 3 respectivement. Comme l'illustre le graphe (fig. 4.1, à gauche), la surface de réponse de degré 2 admet son minimum en un point qui est bien loin d'être un minimum pour y. En augmentant le degré du polynôme d'approximation à 3, on obtient (Cf. fig. 4.1, à droite) un phénomène analogue, et le remplacement pur et simple de y par son approximation m semble peu propice à une optimisation réussie.

Quid d'une approche itérative ?

La mise en garde précédente porte sur la méthode statique consistant à optimiser m et à se contenter du résultat $(\mathbf{x}^*(m))$. Il semble naturel de s'intéresser au procédé qui consiste à évaluer le simulateur y au point $\mathbf{x}^*(m)$, à ré-estimer le métamodèle m en prenant la nouvelle observation en compte, et à itérer jusqu'à ce qu'une condition de convergence soit remplie. Nous allons dans un premier temps appliquer ce procédé itératif à l'exemple précédent ainsi qu'à la fonction de Branin-Hoo (déjà rencontrée au chapitre 3, Cf. fig. 3.2.3). On peu constater sur la figure 4.2 que les itérations successives n'ont pas permis

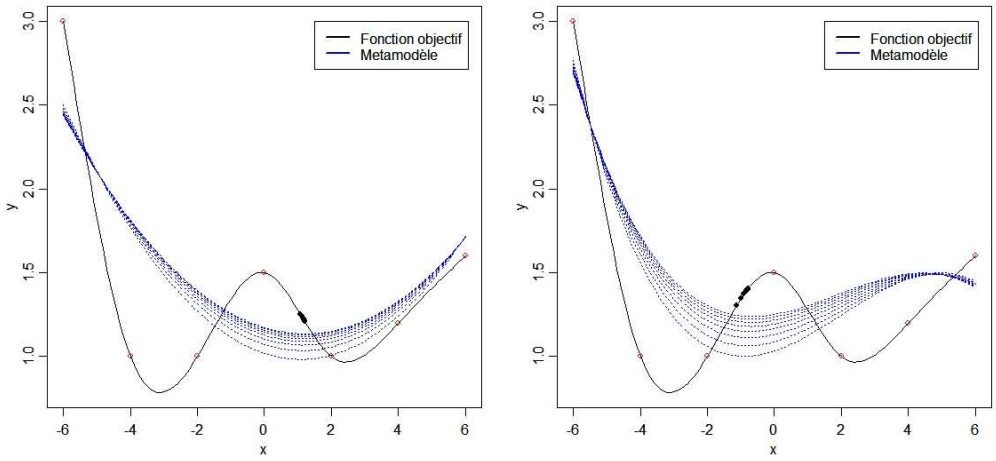


FIG. 4.2 – Optimisation séquentielle de la fonction objectif de la figure 4.1, basée sur une alternance entre estimation de surfaces de réponses polynômiales (en bleu, de degrés 2 à gauche, et 3 à droite) et optimisation des surfaces de réponse, avec intégration dans le plan d'expériences des minima obtenus à chaque itération (points noirs).

d'améliorer sensiblement l'optimisation de la fonction objectif : qu'il s'agisse de la surface de degré 2 ou de celle de degré 3, l'ajout du minimum de la surface de réponse à chaque pas ne permet pas ici de se rapprocher du vrai minimiseur recherché. Le problème lors de l'optimisation séquentielle avec un métamodèle global aussi simple (polynômial de degré 2) est que l'ajout de nouvelles observations ne mène pas nécessairement à une surface de réponse plus précise, en tout cas en ce qui concerne le voisinage du vrai minimum. Nous nous pencherons au prochain paragraphe sur la question de savoir si le cas de métamodèles plus souples (splines de lissage) est plus favorable. Voyons avant tout quel résultat donne l'optimisation de la fonction de Branin-Hoo sur base de surface de réponses polynômiales de degré 2, avec et sans terme d'interaction.

On peut constater sur 4.3 que la démarche séquentielle appliquée à la fonction de Branin-Hoo permet d'améliorer les résultats que l'on aurait obtenus en une seule itération, en particulier dans le cas sans interaction : la prise en compte de l'observation au point minimiseur de la première surface de réponse occasionne ici un changement des surfaces de réponse suivantes et réoriente la sequence de points vers des zones plus pertinentes pour la minimisation de la vraie fonction. En revanche, cela n'a pas permis de visiter l'ensemble des zones de minima. De même dans le cas avec interactions : on reste coincé

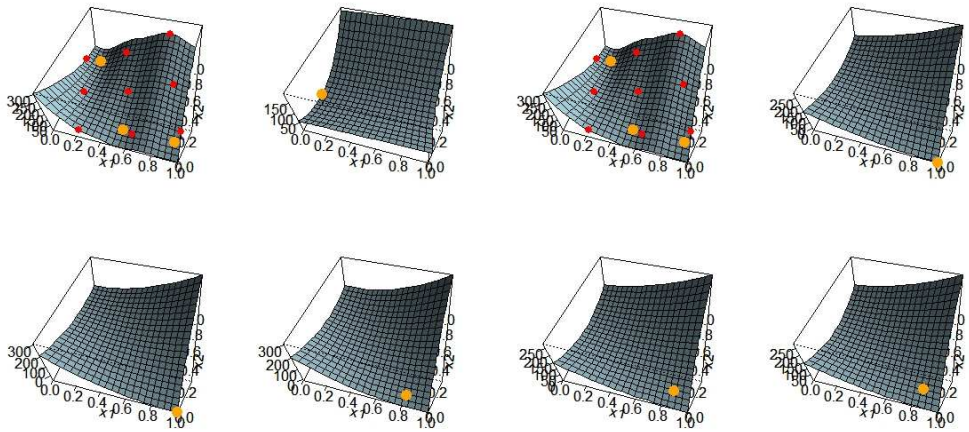


FIG. 4.3 – 3 itérations d'une optimisation séquentielle de la fonction de Branin-Hoo (en haut à gauche de chaque quadrant) sur base de surfaces de réponses polynômiales, en considérant comme plan d'expériences initial un plan factoriel tourné à 9 points (en rouge). A gauche : surface de degré 2 sans interaction. A droite : surface de degré 2 avec interaction. Les sphères oranges indiquent les minima globaux des surfaces représentées.

à proximité d'un minimum local (on ne visite pas les autres locaux, même après un grand nombre d'itérations). Cela n'est pas satisfaisant du point de vue de l'optimisation globale. Sachant que le régression avec interaction présente un coefficient R^2_{adj} de 0.9824, on constate ici que le bon ajustement d'une approximation aux observations est clairement insuffisant pour garantir une optimisation réussie. Concluons ces premiers exemples sur une citation :

> This example(s) shows that even for non-pathological functions, the method can fail abismally. *Donald R. Jones*, [Jon01]

Optimisation séquentielle à base de splines de lissage

Pour finir sur ce sujet, la figure 4.4 montre le résultat obtenu avec un métamodèle « souple », i.e. une spline d'approximation. On obtient cette fois un minimum local, mais on néglige toute une zone de l'espace, non visitée. On trouvera dans [Jon01] d'autres exemples illustrant la non-efficacité des splines comme métamodèle déterministe pour l'optimisation. L'exemple peut sembler caricatural, mais il faut bien garder à l'esprit que l'existence de grandes zones non visitées sera inévitable en plus grandes dimensions.

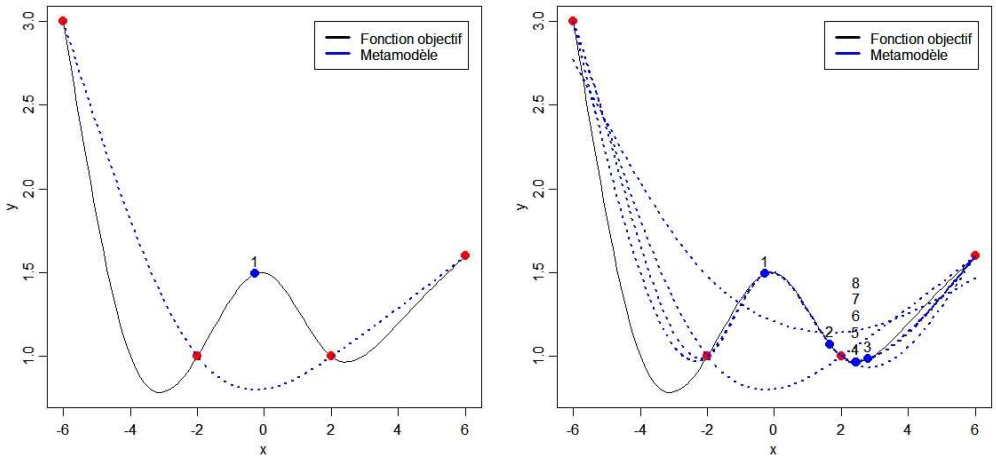


FIG. 4.4 – Optimisation séquentielle d'approximations par splines (en bleu) d'une spline (en noir), en considérant un plan d'expériences initial (en rouge) à 4 points. Les points bleus représentent les minimiseurs des approximations succesives.

4.1.2 Mise en garde

Optimiser sur la base d'un métamodèle déterministe consiste à remplacer le problème

$$\mathbf{x}^*(y) = argmin_{\mathbf{x}\in D} y(x) \tag{4.1}$$

par un problème d'optimisation sur la fonction m, métamodèle de y :

$$\mathbf{x}^*(m) = argmin_{x\in D} m(x) \tag{4.2}$$

On peut alors éventuellement s'intéresser à la proximité entre les deux solutions $\mathbf{x}^*(y)$ et $\mathbf{x}^*(m)$ afin de de juger de la pertinence du remplacement de y par m lors de l'optimisation. Une alternative est de porter son attention à la différence entre $y(\mathbf{x}^*(m))$ et $y(\mathbf{x}^*(y)) = \min_{\mathbf{x}\in D} y(x)$: on attend d'un procédé satisfaisant qu'il fournisse un $\mathbf{x}$ dont l'image par y est proche de la réponse optimale, peu importe la proximité de $\mathbf{x}$ au meilleur $\mathbf{x}$ possible *dans l'espace des variables.*

Il n'est malheureusement pas chose aisée de contrôler $y(\mathbf{x}^*(m)) - \min_{\mathbf{x}\in D} y(x)$ en fonction des critères utilisés habituellement pour évaluer la qualité des métamodèles. On a par exemple pu constater sur les figures 13.1 et 13.2 que les résultats de minimisation avec modèles quadratiques n'étaient pas satisfaisants au sens du critère que l'on vient

de définir. Ce constat négatif est en fait généralisable à de nombreux problèmes : la qualité d'une approximation (interpolation ou non) au sens de l'erreur d'apprentissage —ou autres mesures de l'adéquation globale entre modèles— ne garantit aucunement la qualité de l'optimisation obtenue en remplaçant la fonction objectif par le métamodèle en question (Cf.[Jon01] pour plus de détails concernant l'exploitation directe de polynômes, d'interpolations par fonctions de base, et de splines pour l'optimisation).

La faiblesse majeure du remplacement pur et simple de y par m lors de l'optimisation semble ainsi être d'accorder une confiance abusive au métamodèle. Il y a cependant des cas ou cette confiance est méritée. Par exemple, dès lors que y possède des tendances très prononcées (e.g. monotonies), il peut être indiqué d'optimiser directement sur la base d'une régression. Cela vaut aussi dans le cas particulièrement favorable d'une réponse additive. En résumé, optimiser directement sur m n'est justifié que lorsque l'on dispose d'une connaissance raisonnable de la réponse (et plus particulièrement concernant ses optima), soit par les données nombreuses soit par des connaissances *a priori* sur sa nature mathématique. Nous allons voir dans ce qui suit comment il est possible d'optimiser sur base de métamodèles tout en gardant une confiance réservée en ces derniers.

4.2 Avantages du probabiliste sur le déterministe

4.2.1 Prise en considération de l'erreur de modèle

Nous avons vu dans le chapitre précédent qu'optimiser directement sur la base d'un métamodèle déterministe présente le danger de rester bloqué dans des zones sub-optimales, même lorsque l'on procède itérativement en ré-estimant le métamodèle à chaque nouvelle évaluation de la fonction objectif. Une des insuffisances de ce type de démarche est de ne pas du tout favoriser l'exploration de régions de l'espace qui n'ont pas encore été explorées auparavant. Certains métamodèles probabilistes —et particulièrement le Krigeage— donnent des solutions naturelles à ce problème en prenant l'erreur de modèle en considération. Nous donnons ici quelques repères essentiels sur l'utilisation qui peut être faite des métamodèles probabilistes en optimisation au travers d'un exposé de critères classiques d'exploration pour l'optimisation globale sur base de Krigeage. On présentera ensuite l'algorithme EGO, avec une application détaillée à la fonction de Branin-Hoo. Cet algorithme fait l'objet dans les chapitres 8 et 9 de deux extensions, respectivement pour prendre simultanément en compte plusieurs modèles de Krigeage, et pour choisir un nombre arbitraire de points à chaque itération de l'algorithme.

4.2.2 Critères de sampling pour l'optimisation sur base de Krigeage

Les stratégies séquentielles d'optimisation sur base de Krigeage (telles que développées dans [JSW98] et commentées dans [Jon01]) traitent le problème de la convergence prématurée vers des zones non optimales en remplaçant l'optimisation directe d'un métamodèle par la recherche de sites à la fois prometteurs (au sens du métamodèle employé) et méconnus (au sens de la variance de Krigeage). Cela permet de forcer l'algorithme à explorer des zones distantes de celles qui ont déjà été visitées précédemment (puisque la variance de Krigeage y est nécessairement nulle[2]). De telles procédures d'optimisation reposent généralement sur l'évaluation à chaque itération de la vraie fonction objectif y en un point maximisant une *figure de mérite* (ou critère) basé sur la loi conditionnelle $[Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}]$. Les critères usuels présentés ci-dessous reposent sur différents compromis entre la prédiction moyenne de Krigeage et l'incertitude associée.

Maximiser l'incertitude via s_{KO}

Le problème fondamental de l'optimisation directe de la moyenne de krigeage m_{KO} lorsque l'on veut optimiser la fonction y est (on l'a vu de manière générale dans le chapitre précédent) que l'on ne prend pas en compte l'erreur de modèle. A l'extrême inverse, il est possible de visiter à chaque itération le point de D le plus mal connu au sens du modèle de Krigeage :

$$\mathbf{x}' = argmax_{\mathbf{x}\in D} s_{KO}(\mathbf{x}) \tag{4.3}$$

où $s_{KO}(\mathbf{x})$ est l'écart-type de prédiction en $\mathbf{x}$ fourni par le krigeage (la racine carrée de la variance de Krigeage). Une telle procédure permet d'obtenir un suite de points qui remplissent l'espace (une suite dense dans D). Utiliser cette stratégie fournira donc nécessairement *in fine* les optima globaux de la fonction (puisqu'elle aura visité tout l'espace). Cela dit, elle ne tire absolument pas avantage des informations collectées au fil de l'algorithme, i.e. les images $y(\mathbf{x}^i)$ (Cf. 3.93 : on observe que la variance de Krigeage Ordinaire ne dépend pas des observations [3]). Il n'y a ainsi aucune incitation à visiter les zones de haute performance. Maximiser l'écart-type de Krigeage comme stratégie d'optimisation est jugé inéfficace en pratique.

[2] s'il n'y a pas d'effet de pépite

[3] on trouve parfois le terme *homoscédasticité en les observations* pour qualifier ce phénomène.

Optimisation multi-critère avec m_{KO} and s_{KO}

La façon la plus générale de formuler le compromis entre l'exploitation des précédents résultats —au travers de m_{KO}— et l'exploration de l'espace D —basée sur s_{OK}— est sans doute le problème bi-critère suivant :

$$\begin{cases} \min_{\mathbf{x}\in D} m_{OK}(\mathbf{x}) \\ \text{and } \max_{\mathbf{x}\in D} s_{OK}(\mathbf{x}) \end{cases} \tag{4.4}$$

Soit $\mathbb{P}$ le front de Pareto des solutions [4]. Trouver et choisir un élément (ou un nombre fini d'éléments) de $\mathbb{P}$ reste un problème difficile puisque $\mathbb{P}$ contient typiquement un nombre infini de points. Une approche comparable —bien que non basée sur le Krigeage— est développée dans [JPS93] : le métamodèle est constant par morceaux et l'incertitude est simplement quantifiée par la distance euclidienne aux points déjà explorés. L'espace D est discretisé et les éléments du front de Pareto définissent des zones où la discrétisation est raffinée. Le coût de calcul de cette méthode devient prohibitif avec l'augmentation du nombre d'itérations, et à plus forte raison avec la dimension de l'espace. Précisons que [BWG+01] propose une version parallélisée de cette méthode.

Maximiser la probabilité d'amélioration

Parmi les nombreux critères présentés dans [Jon01] et [VVW09], la probabilité d'améliorer la fonction au-delà du minimum courant $min(\mathbf{Y}) = \min\{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\}$ semble la plus fondamentale :

$$PI(\mathbf{x}) := P(Y(\mathbf{x}) \leq \min(Y(\mathbf{X}))|Y(\mathbf{X}) = \mathbf{Y}) \tag{4.5}$$

$$= \mathbb{E}[\mathbb{1}_{Y(\mathbf{x})\leq min(Y(\mathbf{X}))}|Y(\mathbf{X}) = \mathbf{Y}] = \Phi\left(\frac{\min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right) \tag{4.6}$$

$\min(\mathbf{Y})$ est parfois remplacée par une valeur cible arbitraitre $T \in \mathbb{R}$. Le critère PI est connu pour fournir une recherche très locale lorsque la valeur de T est proche de $min(\mathbf{Y})$. Prendre plusieurs T est une des solutions pour forcer l'exploration, évoquée dans [Jon01].

Maximiser l'*expected improvement*

Une autre solution est de maximiser l'amélioration espérée (*expected improvement*)

$$EI(\mathbf{x}) := \mathbb{E}[\max\{0, \min(Y(\mathbf{X})) - Y(\mathbf{x})\}|Y(\mathbf{X}) = \mathbf{Y}] \tag{4.7}$$

qui prend non seulement en compte la probabilité de progrès mais aussi **l'amplitude** de ce dernier. l'EI mesure le progrès espérée lorsque l'on évalue y en $\mathbf{x}$. *In fine*, le progrès en

[4] Definition du front de Pareto de $(s_{KO}, -m_{KO})$: $\forall x \in \mathbb{P}, \nexists\, \mathbf{y} \in D : (m_{KO}(\mathbf{y}) < m_{KO}(\mathbf{x})$ et $s_{KO}(\mathbf{y}) \geq s_{KO}(\mathbf{x}))$ ou $(m_{KO}(\mathbf{y}) \leq m_{KO}(\mathbf{x})$ et $s_{KO}(\mathbf{y}) > s_{KO}(\mathbf{x}))$

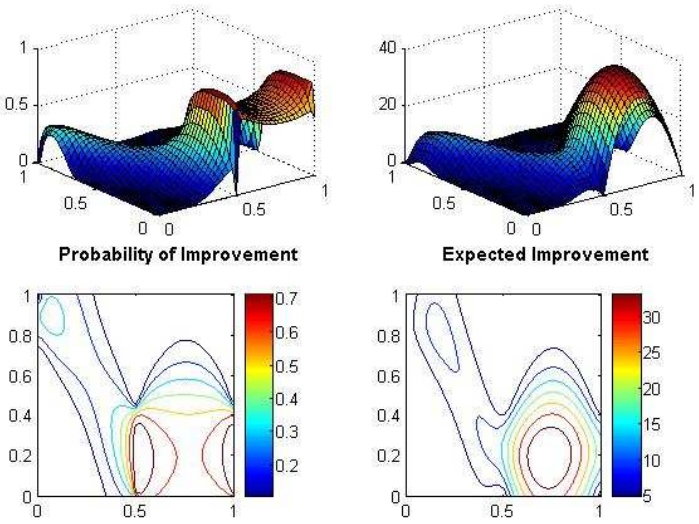


FIG. 4.5 – Surfaces de probabilité d'amélioration et d'amélioration espérée pour la fonction de Branin-Hoo (même plan d'expériences initial, modèle de Krigeage, et paramètres de covariance que dans la figure 3.11). Maximiser PI mène à évaluer y en des points proches des « meilleurs points » (i.e. ceux associés aux plus basses observations), alors que maximiser l'EI mène à évaluer y **entre** les meilleurs points. Par construction, ces deux critères s'annulent aux points d'expérimentation, mais la probabilité d'amélioration devient très proche de $\frac{1}{2}$ au voisinage des meilleurs points.

question vaudra 0 si la vraie valeur de $y(\mathbf{x})$ est supérieure ou égale à $min(\mathbf{Y})$ et vaudra $min(\mathbf{Y})-y(\mathbf{x}) > 0$ dans le cas contraire. Comme on connaît la distribution conditionnelle de $Y(\mathbf{x})$ sachant les observations, on peut calculer EI analytiquement (voir [JSW98]) :

Calcul de l'*amélioration espérée*.

$$EI(\mathbf{x}) = (min(\mathbf{Y}) - m_{KO}(\mathbf{x}))\Phi\left(\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right) + s_{KO}(\mathbf{x})\phi\left(\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right) \qquad (4.8)$$

où ϕ et Φ représentent respectivement la densité de probabilité et la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$.

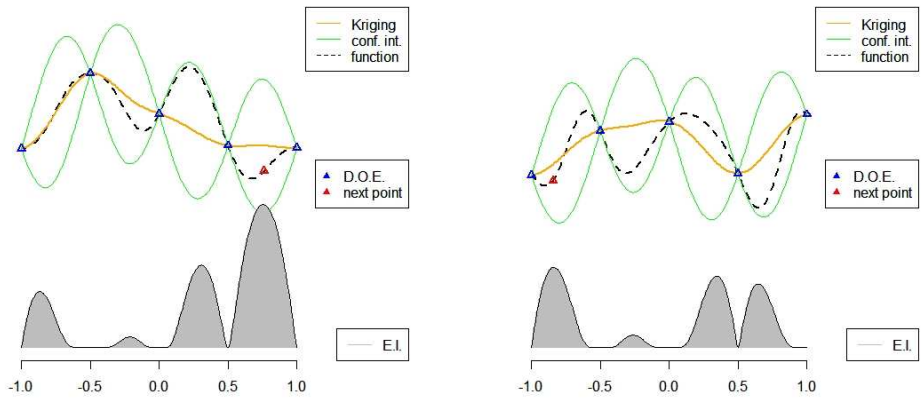


FIG. 4.6 – Deux réalisations (en pointillés) d'un processus gaussien de covariance gaussienne, de portée 0.3 et de variance 1. Les courbes jaunes représentent les moyennes de Krigeage, encadrées par les courbes de quantiles à 2.5% et 97.5% (en vert). Les deux fonctions d'amélioration espérée sont schématiquement représentées par des courbes « pleines », grisées. Le plan d'expériences initial et le maximiseur global de l'EI sont représentés par des triangles, respectivement bleus et rouge.

Démonstration.

$$
\begin{aligned}
EI(\mathbf{x}) &= \mathbb{E}[(min(\mathbf{Y}) - Y(\mathbf{x}))\mathbb{1}_{Y(\mathbf{x})\leq min(\mathbf{Y})}|Y(\mathbf{X}) = \mathbf{Y}] \\
&= \int_{-\infty}^{min(\mathbf{Y})} (min(\mathbf{Y}) - y)\, f_{\mathcal{N}(m_{KO}(\mathbf{x}),s^2_{KO}(\mathbf{x}))}(y)dy \\
&= \int_{-\infty}^{\frac{min(\mathbf{Y})-m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}} (min(\mathbf{Y}) - m_{KO}(\mathbf{x}) - s_{KO}(\mathbf{x}) \times u)\, f_{\mathcal{N}(0,1)}(u)du \\
&= (min(\mathbf{Y}) - m_{KO}(\mathbf{x})) \int_{-\infty}^{\frac{min(\mathbf{Y})-m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}} f_{\mathcal{N}(0,1)}(u)du \\
&- s_{KO}(\mathbf{x}) \int_{-\infty}^{\frac{min(\mathbf{Y})-m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}} u \times f_{\mathcal{N}(0,1)}(u)du \\
&= (min(\mathbf{Y}) - m_{KO}(\mathbf{x}))\, \Phi\left(\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right) + s_{KO}(\mathbf{x})\phi\left(\frac{min(\mathbf{Y}) - m_{KO}(\mathbf{x})}{s_{KO}(\mathbf{x})}\right)
\end{aligned}
$$

où la dernière égalité découle du fait que $\frac{f_{\mathcal{N}(0,1)}}{du}(u) = -uf_{\mathcal{N}(0,1)}(u)$. □

On peut remarquer que cette expression fait apparaître le compromis entre zones prometteuses et incertaines. L'EI possède certaines propriétés importantes pour l'exploration séquentielle : il est nul aux points déjà explorés et strictement positif partout

ailleurs, avec une amplitude croissante en la variance de Krigeage et décroissante en la moyenne de krigeage (les maximiseurs de l'EI font d'ailleurs partie du front de Pareto de $(s_{KO}, -m_{KO})$).

La stratégie SUR : *Stepwise Uncertainty Reduction*

La stratégie SUR (*Stepwise Uncertainty Reduction*) a été introduite en 1995 dans [GJ95], puis étendue au domaine de l'optimisation globale dans [VVW09]. En adoptant une vision de y basée sur la loi conditionnelle du processus aléatoire Y, $[Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}]$, il devient possible de définir $[\mathbf{x}^*|Y(\mathbf{X}) = \mathbf{Y}]$, la loi du vecteur aléatoire de l'emplacement du minimiseur de $Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}$, de densité notée $p_{\mathbf{x}^*|Y(\mathbf{X})=\mathbf{Y}}$. L'incertitude sur la position du minimiseur $\mathbf{x}^*$ est alors quantifiée par l'entropie conditionnelle $H(\mathbf{x}^*|Y(\mathbf{X}) = \mathbf{Y})$ associée à la densité $p_{\mathbf{x}^*|Y(\mathbf{X})=\mathbf{Y}}(\mathbf{x})$. $H(\mathbf{x}^*|Y(\mathbf{X}) = \mathbf{Y})$ diminue à mesure que la distribution de $\mathbf{x}^*|Y(\mathbf{X}) = \mathbf{Y}$ devient « pointue » (resserrée autour d'une (de) certaine(s) valeur(s)). En substance, la stratégie SUR pour l'optimisation globale choisit comme prochain itéré le point qui donne le plus d'information sur la position du minimiseur,

$$\mathbf{x}' = argmin_{\mathbf{x}\in D} H(\mathbf{x}|Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x})) \tag{4.9}$$

Dans la pratique, $p_{\mathbf{x}^*|Y(\mathbf{X})=\mathbf{Y}}(\mathbf{x})$ est estimée par tirages Monte-Carlo de $Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}$ aux points d'une grille de D, ce qui est succeptible de devenir problématique pour les cas où l'on a beaucoup de variables d'entrée puisque le nombre de points de la grille doit augmenter géométriquement en la dimension. Le critère SUR est par nature bien différent des autres critères présentés ci-dessus puisqu'il ne se focalise pas sur un gain immédiat (quel progrès va-t-on obtenir à la prochaine itération), mais plutôt sur un gain retardé, en privilégiant l'apprentissage global de Y en réduisant l'entropie associée à la position de son minimiseur. L'amélioration espérée multi-points (*q-EI*) —exposée en détail au chapitre 9— présente quelques similarités avec SUR dans le sens qu'elle favorise simultanément le gain à court terme et l'exploration globale.

4.3 Autour de l'algorithme *Efficient Global Optimization*

4.3.1 Présentation de l'algorithme EGO

EGO (*Efficient Global Optimization*) est un algorithme d'optimisation basé sur le Krigeage, et plus particulièrement sur le critère d'*amélioration espérée*. Partant d'un plan d'expériences initial $\mathbf{X}$ (typiquement un hypercube latin), EGO forme une séquence de points en alternant à chaque itération la visite du maximiseur courant de l'amélioration espérée (disons le premier visité s'il y en a plusieurs) et l'actualisation du modèle de Krigeage, y compris via ré-estimation des paramètres de covariance :

Algorithm 2 L'algorithme EGO

1: **function** EGO($\mathbf{X}$, $\mathbf{Y}$, n)
2: **for** $i \leftarrow 1, n$ **do**
3: $\psi^* = argmax_{\psi \in \Psi} L(\psi; Y(\mathbf{X}) = \mathbf{Y})$ ▷ Estimation de ψ par MV
4: $\mathbf{x}^* = argmax_{\mathbf{x} \in D}\{EI_{\psi^*}(\mathbf{x})\}$ ▷ Maximisation de l'amélioration espérée
5: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^*\}$ *and* $\mathbf{Y} = \mathbf{Y} \bigcup \{y(\mathbf{x}^*)\}$ ▷ Actualisation du plan d'exp.
6: **end for**
7: **end function**

Dans son contexte originel [JSW98], l'algorithme boucle jusqu'à ce que $max_{\mathbf{x} \in D}\{EI_{\psi^*}(\mathbf{x})\} < \delta$ pour un δ fixé par l'utilisateur, et dépendant des précédentes évaluations de la fonction objectif (dans la version d'origine, δ est 1% *de la meilleure valeur connue au temps courant* [JSW98]). Nous considérons ici le cas dans lequel le nombre d'itérations $n \in \mathbb{N}$ est fixé à l'avance en fonction du budget de l'utilisateur (rappelons que dans certains cas, le temps de simulation se compte en jours de calcul). Après avoir été développé et appliqué à des problèmes de conception de circuits électroniques ([SWJ97], [Sch97]), puis d'ingénierie automobile et aerospatiale, EGO est devenu une référence en optimisation de simulateurs numériques et a inspiré de nombreux travaux contemporains en optimisation sur base de métamodèles (voir [Jon01], [Kra06]). Cet algorithme est aujourd'hui employé pour l'optimisation globale de simulateurs numériques possédant jusqu'à une dizaine de paramètres d'entrée.

Mise en oeuvre

Comme on peut le constater en consultant sa version compacte ci-dessus, EGO jouit d'une grande simplicité conceptuelle : on répète un petit nombre d'opérations à chaque itération sans que ces dernières ne soient modifiées ou adaptées au fil de la procédure. Ces trois opérations sont la ré-estimation du métamodèle de Krigeage, la recherche du maxi-

miseur global de l'amélioration espérée, et l'actualisation du couple (plan d'expériences, observations). Cette dernère opération a un coup négligeable en comparaison avec les deux autres. La première concerne principalement la maximisation de vraisemblance, qui nécessite d'avoir recours à un algorithme d'optimisation globale. Nous considérerons ce problème plus en détail dans le chapitre 5. Concernant le second problème, il est sans doute le plus difficile de l'algorithme. L'amélioration espérée est non seulement coûteuse à évaluer (tout comme la vraisemblance), mais elle est aussi hautement multimodale et pas nécessairement différentiable sur tout son domaine de définition. Nous avons implémenté une version d'EGO dans laquelle l'optimisation de l'EI se base sur un algorithme génétique utilisant les gradients [MS08]. Cela a nécessité le calcul analytique du gradient de l'EI.

Propriété : gradient de l'amélioration espérée.

$$\nabla EI(\mathbf{x}) = \frac{1}{2s^2(\mathbf{x})}\left\{EI(\mathbf{x}) - z(\mathbf{x})\Phi(z(\mathbf{x}))\right\}\nabla s^2(\mathbf{x}) - \frac{\nabla m(\mathbf{x})}{s(\mathbf{x})} \tag{4.10}$$

$$o\grave{u}\ \begin{cases} z(\mathbf{x}) = \frac{(\min(\mathbf{Y})-m(\mathbf{x}))}{s(\mathbf{x})} \\ \nabla m(\mathbf{x}) = \mathbf{Y}^T K^{-1}\nabla\mathbf{k}(\mathbf{x}) \\ \nabla s^2(\mathbf{x}) = -2\left(\mathbf{k}(\mathbf{x})^T K^{-1}\nabla\mathbf{k}(\mathbf{x}) + \frac{(1-\mathbb{1}^T K^{-1}\mathbf{k}(\mathbf{x}))}{\mathbb{1}^T K^{-1}\mathbb{1}}\mathbb{1}^T K^{-1}\nabla\mathbf{k}(\mathbf{x})\right) \end{cases} \tag{4.11}$$

Démonstration. En écrivant $EI(\mathbf{x}) = s(\mathbf{x})\left(z(\mathbf{x})\Phi(z(\mathbf{x})) + \phi(z(\mathbf{x}))\right)$, on a :

$$\begin{aligned} \nabla EI(\mathbf{x}) &= \left\{z(\mathbf{x})\Phi(z(\mathbf{x})) + \phi(z(\mathbf{x})\right\}\nabla s(\mathbf{x}) + s(\mathbf{x})\nabla\left\{z(\mathbf{x})\Phi(z(\mathbf{x})) + \phi(z(\mathbf{x}))\right\} \\ &= \frac{EI(\mathbf{x})}{s(\mathbf{x})}\nabla s(\mathbf{x}) + \underbrace{\nabla\left\{z(\mathbf{x})\Phi(z(\mathbf{x})) + \phi(z(\mathbf{x}))\right\}}_{A}, \end{aligned} \tag{4.12}$$

qui peut être simplifié en remarquant que

$$\begin{aligned} A &= \nabla\left\{z(\mathbf{x})\Phi(z(\mathbf{x}))\right\} + \nabla\left\{\phi(z(\mathbf{x}))\right\} \\ &= \left\{\Phi(z(\mathbf{x}))\nabla z(\mathbf{x}) + z(\mathbf{x})\phi(z(\mathbf{x}))\right\} + \left\{-z(\mathbf{x})\phi(z(\mathbf{x}))\right\} \\ &= \Phi(z(\mathbf{x}))\nabla z(\mathbf{x}) \end{aligned} \tag{4.13}$$

Il reste à développer

$$\nabla z(\mathbf{x}) = \frac{-s(\mathbf{x})\nabla m(\mathbf{x}) - (\min(\mathbf{Y}) - m(\mathbf{x}))\nabla s(\mathbf{x})}{s^2(\mathbf{x})} = -\frac{\nabla m(\mathbf{x})}{s(\mathbf{x})} - z(\mathbf{x})\frac{\nabla s(\mathbf{x})}{s(\mathbf{x})} \tag{4.14}$$

puis à utiliser que $\nabla s^2(\mathbf{x}) = 2s(\mathbf{x})\nabla s(\mathbf{x})$, et enfin à injecter les équations 4.13 et 4.14 dans l'éq. 4.12 pour obtenir que

$$\begin{aligned} \nabla EI(\mathbf{x}) &= \left\{EI(\mathbf{x}) - z(\mathbf{x})\Phi(z(\mathbf{x}))\right\}\frac{\nabla s(\mathbf{x})}{s(\mathbf{x})} - \frac{\nabla m(\mathbf{x})}{s(\mathbf{x})} \\ &= \frac{1}{2s^2(\mathbf{x})}\left\{EI(\mathbf{x}) - z(\mathbf{x})\Phi(z(\mathbf{x}))\right\}\nabla s^2(\mathbf{x}) - \frac{\nabla m(\mathbf{x})}{s(\mathbf{x})} \end{aligned} \tag{4.15}$$

Concernant les gradients de l'éq. 4.11, il sont obtenus par différentiation directe des équations 3.93. □

4.3.2 Application de l'algorithme EGO à la fonction de Branin-Hoo

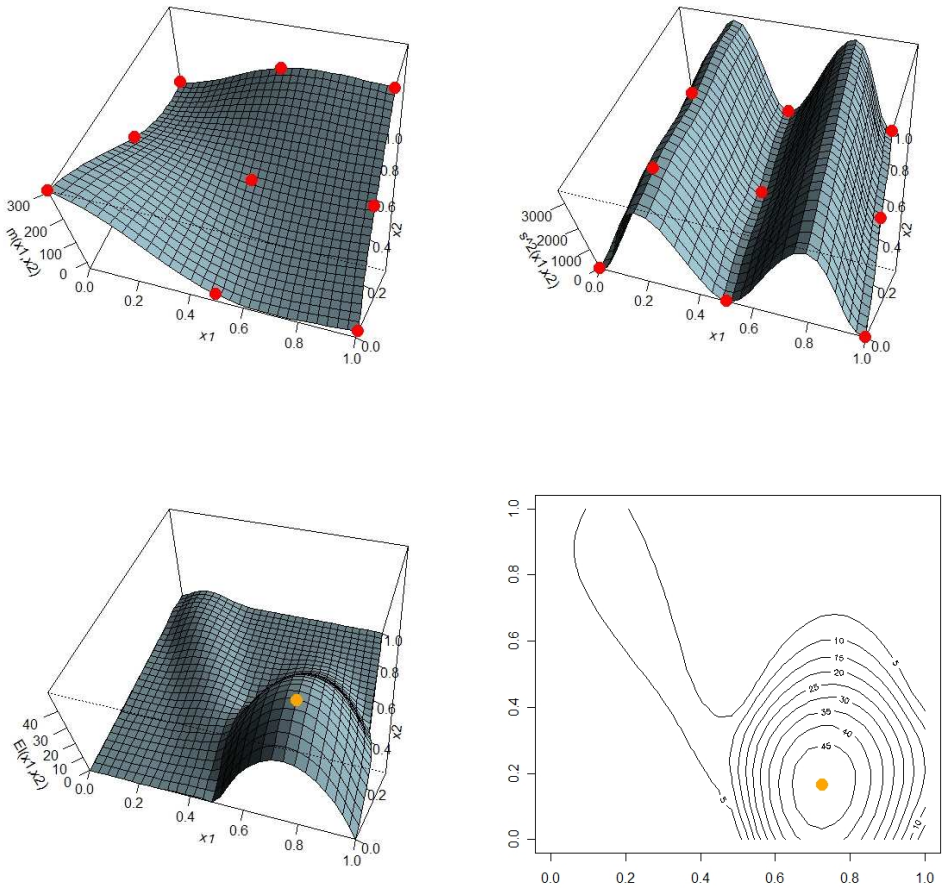


FIG. 4.7 – Première itération d'EGO sur la fonction de Branin-Hoo, avec pour plan d'expérience initial un plan complet 3×3, une covariance gaussienne anisotrope, et des paramètres de covariances estimés par EMV. En haut : moyenne de KO (à gauche) et variance de KO (à droite) ; les points rouges représentent le plan d'expériences initial. En bas : surface et lignes de niveaux de l'amélioration espérée ; le point orange représente le maximum courant de l'amélioration espérée.

La fonction de Branin-Hoo, déjà rencontrée dans la partie 3, constitue un cas-test d'optimisation globale intéressant dans la mesure où elle présente trois optima globaux (trois locaux auxquels la valeur prise par la fonction objectif est la même). Nous illustrons

dans cette section comment l'algorithme EGO permet de trouver ces trois optima. La figure 4.3.2 représente le métamodèle de Krigeage Ordinaire à l'état initial de l'algorithme, ainsi que la surface d'amélioration espérée associée. Cette surface présente deux zones de maxima locaux et un seul maximum global, matérialisé par un point orange. La situation de ce point illustre bien le compromis entre une moyenne de Krigeage basse et une variance élevée. Remarquons au passage la forme particulière de la surface de variance de Krigeage : elle varie beaucoup plus selon x_1 que selon x_2. Cela illustre une forte anisotropie décelée par maximum de vraisemblance Cf. 9.3.2.

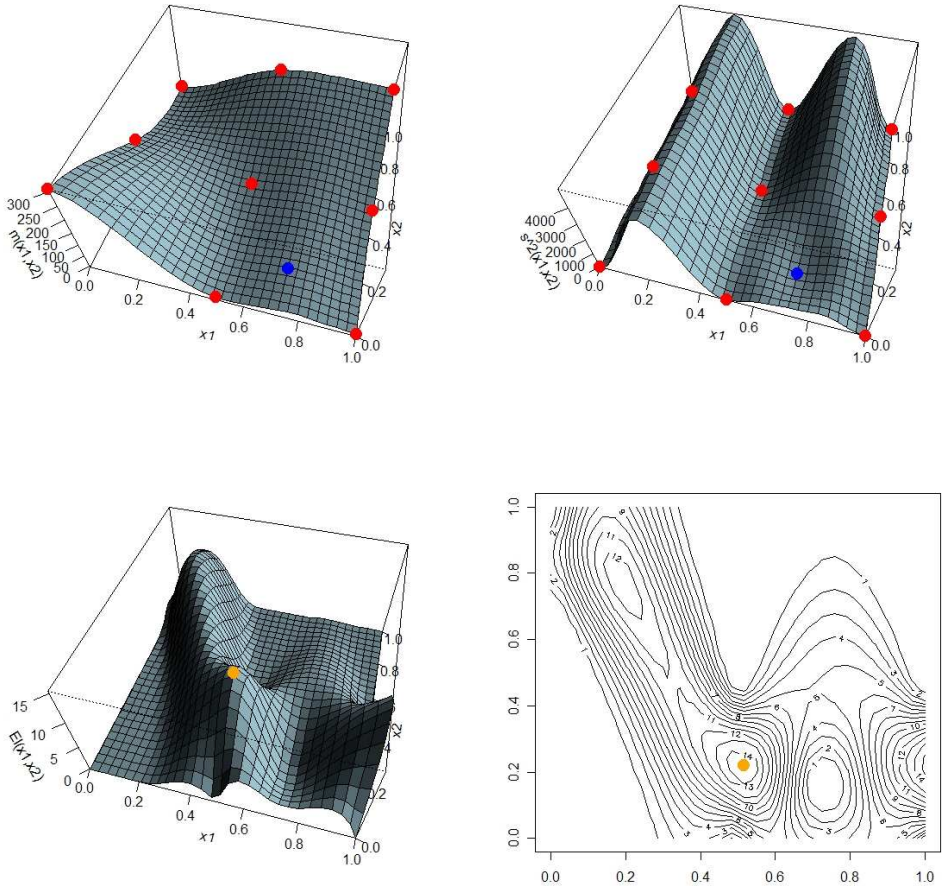


FIG. 4.8 – Deuxième itération d'EGO (Cf. 4.3.2 pour la légende). Le(s) point(s) bleu(s) symbolise(nt) le(s) point(s) visité(s) durant l'(les) itération(s) précédente(s).

Le graphique correspondant à la deuxième itération (Cf. 4.3.2) présente le métamodèle de Krigeage après intégration du maximiseur de l'EI de l'itération précédente et réestimation des paramètres de covariance, ainsi que la moyenne de Krigeage (Cf. 9.3.2 et 4.12 pour observer l'évolution des valeurs numériques). On peut aussi observer que la variance et l'amélioration ont sensiblement évolué, en particulier au voisinage du point précédemment visité, où les deux s'annulent. Le nouveau maximiseur de l'EI est malgré tout relativement proche du dernier point visité, dont l'image basse a confirmé l'intérêt de la zone. Remarquons qu'une seconde zone d'intérêt se profile déjà.

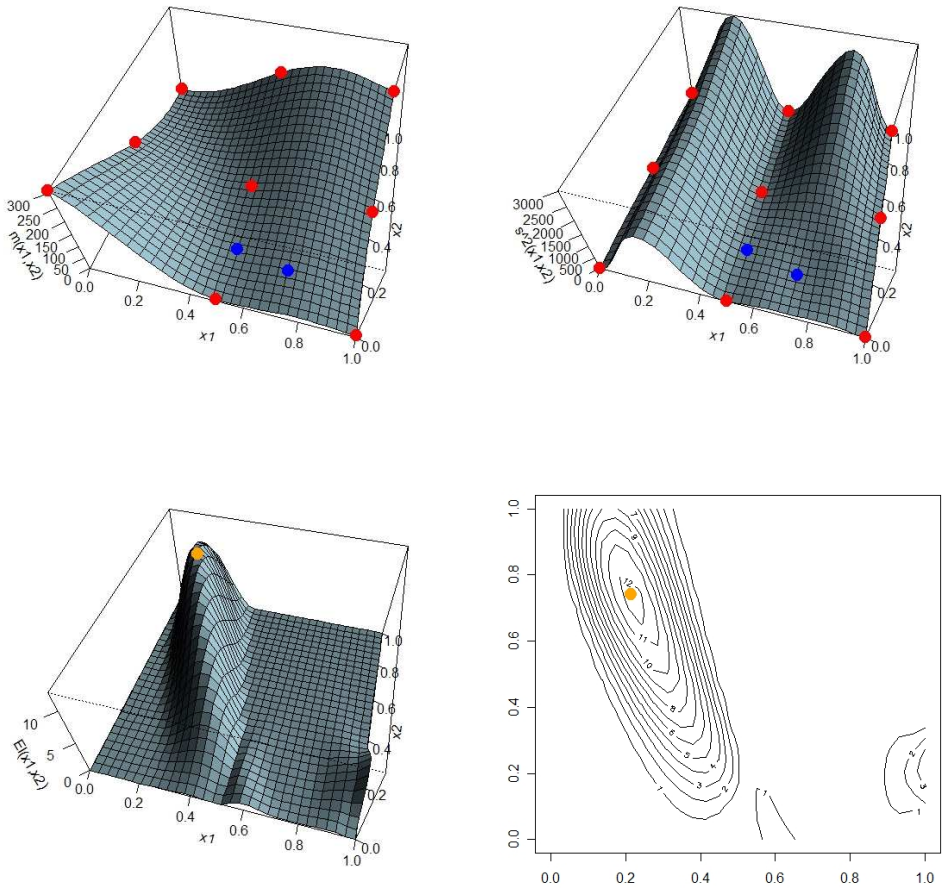


FIG. 4.9 – Troisième itération d'EGO sur la fonction de Branin-Hoo. Cf. 4.3.2 et 4.3.2 pour les légendes.

La zone pressentie à la deuxième itération comme potentiellement intéressante devient incontournable à la troisième itération (Cf. 4.3.2 où ce point s'avère être le nouveau maximiseur de l'EI), après que la visite du point de la deuxième itération a eu lieu. Vues les surfaces de moyenne et de variance de Krigeage, cette zone apparaît clairement comme la seule en laquelle on a à la fois des prédictions optimistes et une grande incertitude (ce qui s'explique par l'éloignement aux observations, toutes confondues).

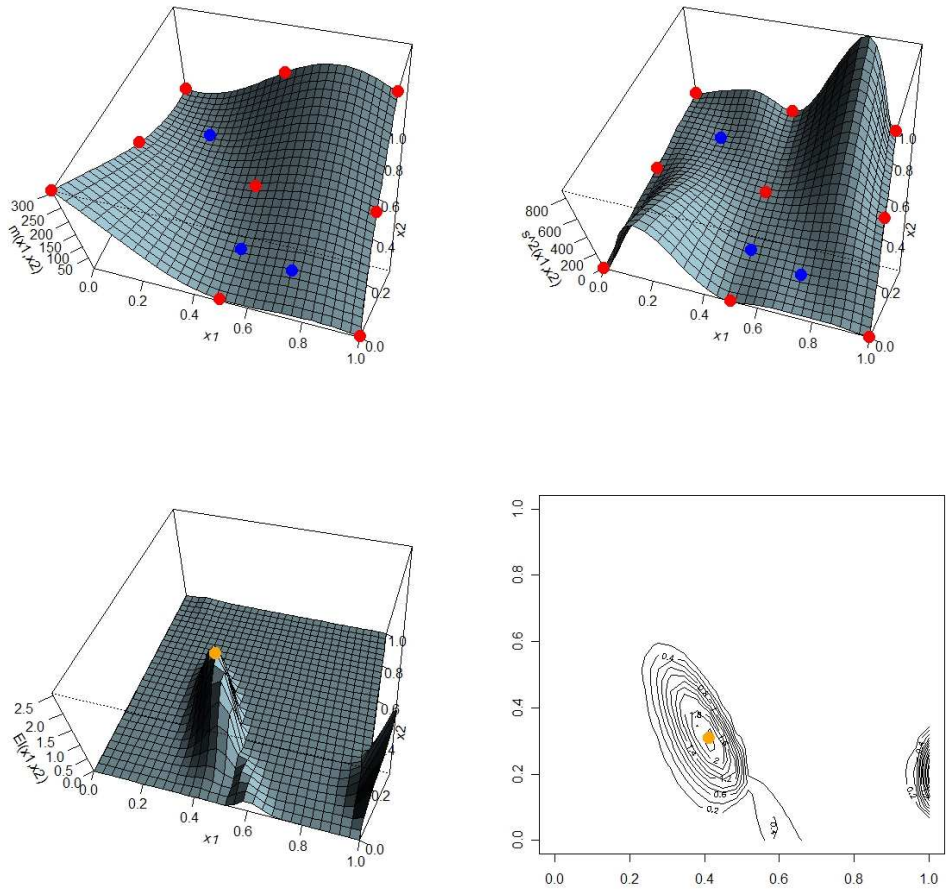


FIG. 4.10 – Quatrième itération d'EGO sur la fonction de Branin-Hoo. Cf. fig. (4.3.2) et (4.3.2) pour les légendes.

Remarquons enfin avec le graphique de la quatrième itération (Cf. fig. (4.3.2) et tab. (9.3.2) que l'ordre de grandeur de l'amélioration espérée a nettement baissé depuis la

première itération (divisé par plus de 10). Cela est dû au fait que l'acquisition de nouveaux points a permis de mieux connaître la vraie fonction, et ainsi de simultanément diminuer la variance de Krigeage et de supprimer des zones d'amélioration possible (au sens où une fois connues, elle ne représentent plus de gain potentiel). La figure 4.3.2 résume les points visités et l'état du métamodèle de Krigeage après 10 itérations d'EGO.

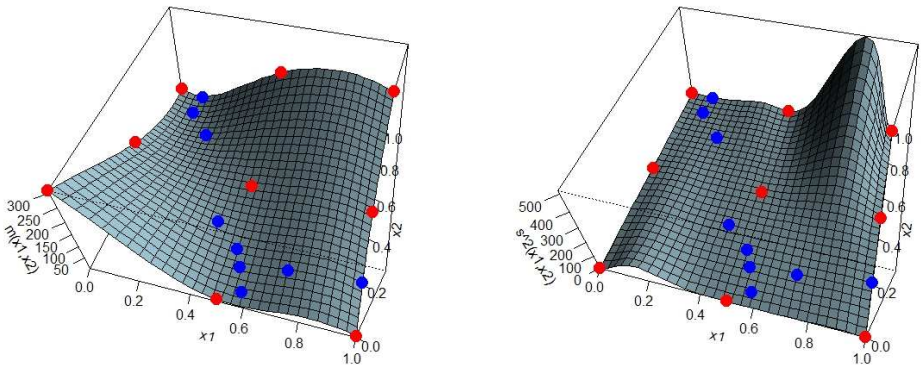


FIG. 4.11 – Surfaces de moyenne et de variance de Krigeage Ordinaire, et répartition spatiale des points visités après 10 itfations d'EGO. Les trois régions de minimum ont bien été visitées, et le minimum courant est de 0.47 (à comparer au vrai minimum global de la fonction de Branin-Hoo, à savoir 0.4). On peut remarquer que la variance reste élevée dans les régions qui n'ont pas encore été visitées par l'algorithme.

Les graphiques de 4.12 et 9.3.2 nous permettent d'observer l'évolution des paramètres de modèle au fil de l'algorithme [5]. Notons que la trajectoire des paramètres de covariance est assez variable entre différents lancements de l'algorithme (non montré ici). En revanche les résultats d'optimisation sont robustes, et on trouve généralement comme sur cet exemple un minimum très proche du vrai minimum global (0.47) et une visite des trois zones de minima au bout des 10 itérations. On voit ainsi comment l'introduction d'une modélisation probabiliste —en particulier via l'utilisation de la variance de Krigeage— offre une alternative puissante aux techniques présentées dans la section précédente.

[5] Il se peut que certaines données de 4.12 et 9.3.2 soient légèrement différentes de celles représentées sur les graphiques de l'évolution du Krigeage et de l'EI puisqu'ils correspondent à deux applications successives de l'algorithme. Même en prenant une population de 100 individus dans l'algorithme génétique avec gradients utilisé pour maximiser L et EI, il subsiste une légère variabilité.

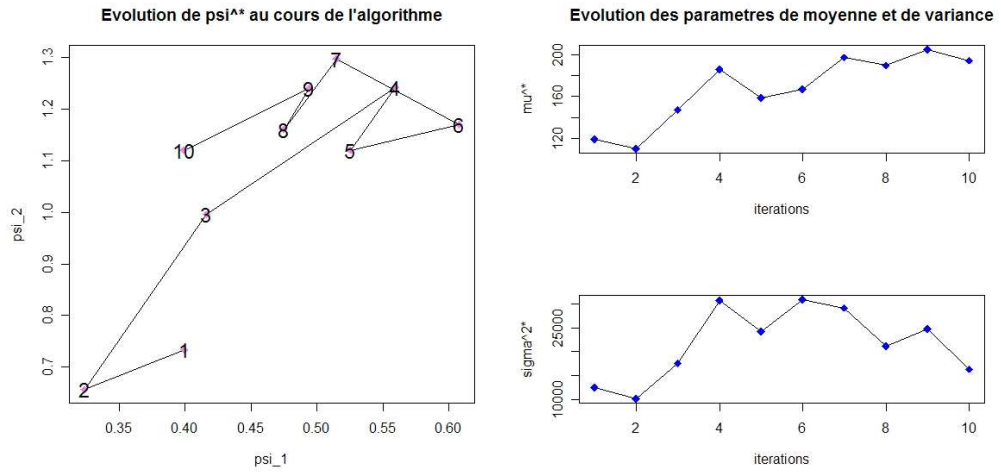


FIG. 4.12 – Evolution des paramètres de covariances (ψ_1, ψ_2) (à gauche), de la moyenne μ^* (en haut à droite) et σ^{2*} (en bas à droite) durant 10 itérations de l'algorithme EGO appliqué à la fonction de Branin-Hoo.

	μ^*	σ^{2*}	ψ_1^*	ψ_2^*	$\mathbf{x}^*$	EI^*	$y(\mathbf{x}^*)$	$min(y(\mathbf{X}\backslash\{\mathbf{x}^*\}))$
it 1	119	12472	0.40	0.73	$(0.73, 0.17)$	49.84	21.05	9.5
it 2	110	10127	0.32	0.66	$(1, 0.22)$	15.40	2.30	9.5
it 3	147	17545	0.42	0.99	$(0.21, 0.75)$	12.16	11.63	2.3
it 4	185	30692	0.56	1.24	$(0.43, 0.28)$	2.94	11.11	2.3
it 5	158	24212	0.53	1.12	$(0.6, 0.0)$	1.55	6.40	2.3
it 6	167	30884	0.61	1.17	$(0.11, 0.99)$	0.021	5.84	2.3
it 7	197	29115	0.52	1.30	$(0.10, 0.90)$	0.22	1.07	2.3
it 8	189	21173	0.47	11.41	$(0.55, 0.16)$	1.18	0.47	1.07
it 9	205	24752	0.49	1.24	$(0.97, 0.18)$	$2e^{-5}$	0.61	0.47
it 10	194	16348	0.4	1.12	$(0.99, 0.11)$	$5e^{-2}$	0.75	0.47

TAB. 4.1 – Evolution des paramètres de modèle et des points visités par l'algorithme EGO appliqué à la fonction de Branin-Hoo. Les valeurs initiales des paramètres de covariance sont fixées à $(0.5, 0.5)$. La dernière colonne représente la plus petite valeur de $y(\mathbf{x})$ connue avant l'itération courante.

4.3.3 Extensions connues, limites, et pistes d'amélioration

Quelques extensions d'EGO

On peut trouver dès la thèse de Schonlau des propositions de généralisations d'EGO à différentes fins. La section 5.2 de [Sch97] est en effet dédiée à une extension de l'EI, le *generalized expected improvement*, dans lequel le terme $(min(\mathbf{Y}) - Y(\mathbf{x}))^+$ de l'EI est remplacé par une puissance entière de ce dernier, $((min(\mathbf{Y}) - Y(\mathbf{x}))^+)^g$, $g \in \mathbb{N}$. L'auteur affirme que l'EI marche bien dans le cas où la fonction objectif se laisse bien modéliser comme réalisation d'un processus gaussien de covariance choisie, mais que dans le cas contraire l'exploration par EI est trop locale. En fonction de la valeur du paramètre g, le *generalized expected improvement* permetrait ainsi de forcer EGO à un comportement plus exploratoire, et donc d'obtenir des résultats meilleurs que l'EI dans le cas où le modèle de processus gaussien est peu adapté à la fonction traitée.

Dans un registre différent, [HANZ06] propose une adaptation d'EGO aux simulateurs non-déterministes, avec une version particulière de l'EI prenant en compte le fait que le minimum courant n'est pas parfaitement connu dans le cas où les observations sont bruitées. On peut aussi trouver dans l'article [HANM06] un algorithme dérivé d'EGO pour le cas d'expériences numériques à plusieurs niveaux de fidélité ; un critère prenant en compte le coût des évaluations, l'*augmented expected improvement*, y est introduit. On peut encore citer [Kno05], dédié à l'optimisation multi-objectifs sur base de Krigeage, [HGKK05], qui traite aussi d'optimisation multi-objectifs et propose une adaptation d'EGO pour les expériences physiques, et l'excellent mémoire [Kra06] proposant une introduction générale au Krigeage, à EGO, des critères permettant respectivement une optimisation à la fois robuste, sous contraintes, et multicritères, avec une application au laminage de tôles. Mentionnons enfin l'existence de travaux en optimisation évolutionnaire, notamment [EGN06], dans lesquels un métamodèle de Krigeage est utilisé au sein d'un algorithme évolutionnaire pour faire une pré-selection parmi les points d'une génération, permettant ainsi de réduire sensiblement les coûts computationnels.

Limites connues et pistes suivies

EGO tel qu'il est présenté dans la littérature se base exclusivement sur le Krigeage Ordinaire, et présuppose donc que la fonction étudiée est la réalisation d'un processus au moins stationnaire à l'ordre 1. Cela peut être un important manque à gagner lorsque l'hypothèse est clairement abusive, i.e. dans les cas où y possède une tendance prononcée. Par ailleurs, la fonction de covariance utilisée dans EGO est de type exponentielle généralisée

3.40 avec paramètres estimés par maximisation de vraisemblance. La pertinence du type de covariance et surtout du protocole d'estimation des paramètres sont des questions épineuses qui sont rarement discutées lors de l'application d'EGO. Nous les considérons pour autant d'importance, et elles font l'objet de la dernière section du chapitre 5 et de plusieurs perspectives de travail.

Les limites auxquelles nous proposons des éléments de réponse à ce stade concernent la parallélisation d'EGO, mais aussi la prise en compte de plusieurs métamodèles de Krigeage en simultané durant l'optimisation. Le premier point (Cf. chapitre 9, ainsi que [GLRC07] reproduit en annexe) est destiné à pallier la sequentialité d'EGO, et à adapter l'algorithme à un context de calcul distribué sur plusieurs processeurs i.e. à fournir à chaque appel un nombre arbitraire de points destinés à lancer plusieurs simulations de manière simultanée, dans l'esprit de [QVPH06]. Le second point (chapitre 8) vise à proposer un cadre formel permettant d'intégrer de multiples modèles de Krigeage —avec différents noyaux de covariance, mais aussi potentiellement plusieurs structures de tendances—, ainsi qu'une variante de l'algorithme EGO basée sur des mélanges de noyaux. Cette denière a fait l'objet de [GHC08], également reproduite en annexe.

Deuxième partie

Contributions à l'étude des métamodèles probabilistes

Chapitre 5

Variabilité de l'EMV des paramètres de covariance du KS

Lorsque l'on définit le Krigeage Simple à partir du conditionnement d'un processus gaussien —tel que cela est présenté dans la section 3 du chapitre 3—, la fonction y est vue comme une réalisation d'un PG Y centré (quitte à considérer le processus centré $Y-\mu(x)$) de noyau de covariance k_ψ de forme paramétrique fixée, où $\psi \in \Psi$ est un paramètre fini-dimensionnel. $\mathbf{Y} = \big(y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\big)$ est une réalisation de $\mathbb{Y} = \big(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n)\big)$, vecteur aléatoire des valeurs prises par Y au plan d'expériences $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$. Par définition des processus gaussiens, nous savons que le vecteur aléatoire $\mathbb{Y}$ est de loi

$$\mathbb{Y} \sim \mathcal{N}(\mathbf{0}, K(\psi)) \tag{5.1}$$

où $K(\psi)$ est la matrice de covariance des observations, dépendant à la fois de $\mathbf{X}$ et du noyau de covariance k_ψ. Nous insistons ici plus particulièrement sur la dépendance en ψ, puisque c'est le paramètre auquel nous souhaitons remonter à partir des observations $\mathbf{Y}$. L'estimation par maximum de vraisemblance (EMV) consiste à rechercher la valeur de ψ qui rende maximale la densité de probabilité des observations :

$$p_\psi(\mathbf{Y}) := f(\mathbf{Y}|\psi) = (2\pi)^{-\frac{n}{2}} det[K(\psi)]^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{Y}^T K(\psi)^{-1}\mathbf{Y}} \tag{5.2}$$

Nous rappelons ci-dessous quelques éléments de théorie de la vraisemblance ainsi que des résultats asymptotiques associés, puis les applications qui en sont classiquement faites dans un contexte géostatistique. Une étude expérimentale est ensuite rapportée et analysée au sujet de la variabilité —et donc de la robustesse— de l'*estimateur du maximum de vraisemblance* des paramètres de covariance du KS lorsque le nombre d'observations est peu élevé. Nous proposons enfin une discussion au sujet de méthodes basées sur l'EMV, et spécifiquement adaptées au cas où n est petit.

5.1 Elements de théorie de la vraisemblance

La théorie de la vraisemblance [Lin96] porte principalement sur l'étude des modèles probabilistes paramétriques (nous nous y restreignons ici complètement), et plus particulièrement sur l'estimation des paramètres de ces derniers à partir de données observées. Nous considérons dans cette section le cadre général d'un vecteur aléatoire n-dimensionnel $\mathbb{Y}$ distribué selon une loi de probabilité P_ψ ($\psi \in \Psi$, où Ψ est [1] une partie de $\mathbb{R}^p$, $p \in \mathbb{N}$). Lorsque P_ψ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, nous utilisons la notation p_ψ pour la densité de probabilité correspondante.

5.1.1 Principes de base

Principe de vraisemblance

Le principe de vraisemblance (« Likelihood principle », Cf. [Lin96]) stipule qu'un modèle de paramètre ψ_1 est plus *vraisemblable* qu'un modèle de paramètre ψ_2 — à la lumière d'observations $\mathbf{Y}$ — si ce premier rend l'observation de $\mathbf{Y}$ plus probable. En d'autres termes, un modèle (auquel est associé une loi P_ψ) est jugé d'autant plus vraisemblable qu'il est probable d'observer $\mathbf{Y}$ si les données ont été générées selon P_ψ.

Ce principe a été maintes fois justifié[2], en particulier d'un point de vue décisionnel : considérons le cas où l'on cherche sur la base d'observations $\mathbf{Y}$, à distinguer parmi deux valeurs possibles $\{\psi_1, \psi_2\}$ le paramètre ψ ayant effectivement servi à générer les données. En appelant $\mathcal{Y}$ l'ensemble des observations $\mathbf{Y}$ possibles (souvent $\mathbb{R}^n$ par la suite), on peut résumer une règle de décision à une partition $\{\mathcal{Y}_1, \mathcal{Y}_2\}$ de $\mathcal{Y}$, chaque sous-ensemble $\mathcal{Y}_i$ correspondant à la décision que le paramètre ψ vaut ψ_i ($i \in \{1, 2\}$). La question d'une règle de décision optimale peut se poser comme le problème suivant : comment choisir $\mathcal{Y}_1$ de manière à ce que $\mathbf{Y}$ soit aussi souvent que possible dans $\mathcal{Y}_1$ lorsque $\psi = \psi_1$ et qu'inversement, $\mathbf{Y}$ ne soit pas dans $\mathcal{Y}_1$ lorsque $\psi = \psi_2$. Formellement, cela revient à rechercher $\mathcal{Y}_1$ de manière à maximiser l'application $Q \in [0, 2]^{\mathcal{P}(\mathcal{Y})}$ définie par

$$Q(\mathcal{Y}_1) := P_{\psi_1}(\mathcal{Y}_1) + P_{\psi_2}(\mathcal{Y} \backslash \mathcal{Y}_1) = P_{\psi_1}(\mathcal{Y}_1) + 1 - P_{\psi_2}(\mathcal{Y}_1), \tag{5.3}$$

i.e. à minimiser l'intégrale suivante :

$$\int_{\mathcal{Y}_1} [p_{\psi_2}(y) - p_{\psi_1}(y)] dy. \tag{5.4}$$

[1] On distinguera parfois dans le cas stationnaire le paramètre de variance σ^2 des paramètres de corrélation $(\psi_1, ..., \psi_p)$, la notation ψ_0 étant alors utilisée pour σ^2 en convenant que $\Psi \subset \mathbb{R}^{p+1}$.

[2] mais aussi controversé ! (Cf. [Rob92])

Cela nous mène clairement à choisir $\mathcal{Y}_1^* = \{\mathbf{Y} \in \mathcal{Y} : p_{\psi_2}(\mathbf{Y}) < p_{\psi_1}(\mathbf{Y})\}$, c'est à dire à choisir $\psi = \psi_1$ lorsque la valeur en y de la densité p_{ψ_1} est supérieure à celle de p_{ψ_2}.

Rappels : fonctions de vraisemblance et de log-vraisemblance

La fonction de vraisemblance en ψ basée sur l'événement $\mathbb{Y} = \mathbf{Y}$ est donnée par

$$L : \psi \in \Psi \longrightarrow L(\psi) := L(\psi; \mathbf{Y}) := p_\psi(\mathbf{Y}) \in \mathbb{R}^+ \tag{5.5}$$

On remarque que L n'est rien d'autre que la densité de probabilité de la variable aléatoire $\mathbb{Y}$, vue comme une fonction de ψ (et non plus comme une fonction de $\mathbf{Y}$). Dans le cas où la variable est discrète, la vraisemblance de ψ basée sur l'observation du fait que $\mathbb{Y} = \mathbf{Y}$ est définie comme $L(\psi) := P_\psi(\mathbf{Y})$, la probabilité que $\mathbb{Y}$ prenne la valeur $\mathbf{Y}$ lorsque la valeur du paramètre d'intérêt est effectivement ψ.

On définit de même la fonction de log-vraisemblance

$$\mathcal{L} : \psi \in \Psi \longrightarrow \mathcal{L}(\psi) := \mathcal{L}(\psi; \mathbf{Y}) := \log[L(\psi; \mathbf{Y})] \in \mathbb{R} \tag{5.6}$$

où log représente la fonction logarithme néperien. Il est courant d'étudier $\mathcal{L}$ (voire $-2\mathcal{L}$) en lieu et place de L. Cela permet d'éviter de travailler avec des valeurs de densité dont l'ordre de grandeur est usuellement très petit, et n'occasionne aucune perte d'information puisqu'il y a une bijection monotone entre $\mathcal{L}$ (ou $-2\mathcal{L}$) et L.

Estimation et estimateur du maximum de vraisemblance

L'estimation par maximum de vraisemblance de ψ associée à l'observation $\mathbf{Y}$ est une[3] valeur ψ^* du paramètre de Ψ qui maximise la fonction $L(.; \mathbf{Y})$ ou, de manière équivalente, un maximiseur global de la fonction $\mathcal{L}(.; \mathbf{Y})$:

$$\psi^* := \psi^*(\mathbf{Y}) := arg\,[sup_{\psi \in \Psi} L(\psi; \mathbf{Y})] = arg\,[sup_{\psi \in \Psi} \mathcal{L}(\psi; \mathbf{Y})] \tag{5.7}$$

L'estimateur du maximum de vraisemblance (EMV) est l'analogue aléatoire de la definition 5.7, i.e. l'application qui à une observation aléatoire $\mathbb{Y}$ associe $\psi^*(\mathbb{Y})$. Comme nous allons l'exposer dans les prochains paragraphes, l'EMV jouit sous certaines conditions de propriétés statistiques (consistance, normalité et efficacité asymptotiques) qui en font un estimateur satisfaisant pour l'esprit et très largement répandu dans la pratique.

[3] Il n'y en a souvent heureusement qu'une, mais il est théoriquement possible que $L(.; \mathbf{Y})$ ait plusieurs maximiseurs globaux. Il faut dans ce cas préciser duquel on parle (en donnant par exemple un voisinage).

Exemple : échantillon d'observations gaussiennes i.i.d.

Prenons l'exemple d'une variable aléatoire gaussienne N de loi $\mathcal{N}(\mu, \sigma^2)$ ($\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*$), et de n réplications indépendantes de N désignées par $Y_1, ..., Y_n$. On note $\mathbb{Y} = (Y_1, ..., Y_n)$ le vecteur aléatoire des réplications, et $\mathbf{Y} = (y_1, ..., y_n)$ une réalisation quelconque de $\mathbb{Y}$. Par indépendance des Y_i, il vient directement que

$$L(\mu, \sigma^2; \mathbf{Y}) = \Pi_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \right) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2} \tag{5.8}$$

d'où l'on tire que $-2\mathcal{L}(\mu, \sigma^2; \mathbb{Y}) = n \times ln(2\pi) + n \times ln(\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu)^2$.

$$\begin{cases} -2\frac{\partial \mathcal{L}(\mu,\sigma^2;\mathbb{Y})}{\partial \mu} = \frac{2}{\sigma^2}\sum_{i=1}^n (\mu - y_i) \longrightarrow \mu^* = \frac{1}{n}\sum_{i=1}^n y_i \\ -2\frac{\partial \mathcal{L}(\mu,\sigma^2;\mathbb{Y})}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{1}{\sigma^4}\sum_{i=1}^n (\mu - y_i)^2 \longrightarrow \sigma^{2*}(\mu) = \frac{1}{n}\sum_{i=1}^n (y_i - \mu)^2 \end{cases} \tag{5.9}$$

On trouve que les estimateurs du maximum de vraisemblance de μ et σ^2 coïncident dans le cas gaussien avec les estimateurs classiques de la moyenne et de la variance d'un échantillon. Remarquons au passage que l'EMV de la variance σ^2 est biaisé [4] , mais que son biais $\frac{-1}{n}\sigma^2$ tend asymptotiquement à s'annuler .

5.1.2 Dérivées de la log-vraisemblance et propriétés

Comme nous venons de le voir sur un exemple, la mise en oeuvre d'une estimation par maximum de vraisemblance repose sur la recherche des points critiques de la fonction $\mathcal{L}$ (ceux qui annulent son gradient, et sont donc des candidats à l'optimalité). Le gradient de la log-vraisemblance, appelé *score*, joue un rôle essentiel en estimation paramétrique.

Le "score", ou gradient de la log-vraisemblance

Le *score* associé au paramètre[5] ψ et aux observations y est ici noté U

$$U(\psi) := U(\psi; y) := \nabla_\psi \mathcal{L}(\psi; y) = \left(\frac{\partial \mathcal{L}(\psi; y)}{\partial \psi_i} \right)_{1 \le i \le p} \tag{5.10}$$

Sous des hypothèses standard sur l'ensemble Ψ et la régularité de $\mathcal{L}$ (Cf. [ABC92]), L'EMV du paramètre ψ est donc une solution de l'*equation normale*

$$U(\psi^*) = \nabla_\psi \mathcal{L}(\psi^*; \mathbf{Y}) = \mathbf{0} \tag{5.11}$$

[4] On a en effet le résultat classique : $\mathbb{E}\left[\sum_{i=1}^n \left(Y_i - \frac{1}{n}\sum_{j=1}^n Y_j\right)^2\right] = (n-1)\sigma^2$.

[5] On se place sans perte de généralité dans le cas où $\Psi \subset \mathbb{R}^p$. Toutes les notations vectorielles introduites dans les pages à venir restent valables en prenant $0 \le i \le p$ dans le cas stationnaire où $\psi_0 \equiv \sigma^2$.

Précisons que le paramètre ψ^0 de la loi selon laquelle les observations $\mathbf{Y}$ ont été générées n'est —malheureusement— pratiquemment jamais solution de l'eq. 5.11. On peut en revanche affirmer que ψ^0 annule U « en moyenne » lorsque $\mathbb{Y}$ est tiré selon P_{ψ^0} :

$$\begin{aligned}\mathbb{E}_{\psi^0}[U(\psi^0;\mathbb{Y})] &= \mathbb{E}_{\psi^0}[\nabla_{\psi^0}\mathcal{L}(\psi^0;\mathbb{Y})] \\ &= \mathbb{E}_{\psi^0}\left[\frac{1}{L(\psi^0;\mathbb{Y})}\nabla_{\psi^0}L(\psi^0;\mathbb{Y})\right] \\ &= \int_{\mathbf{Y}\in\mathcal{Y}}\left[\frac{1}{L(\psi^0;\mathbf{Y})}\nabla_{\psi^0}L(\psi^0;\mathbf{Y})\right]L(\psi^0;\mathbf{Y})d\mathbf{Y} \\ &= \int_{\mathbf{Y}\in\mathcal{Y}}\nabla_{\psi^0}L(\psi^0;\mathbf{Y})d\mathbf{Y} = \nabla_{\psi^0}\left[\int_{\mathbf{Y}\in\mathcal{Y}}L(\psi^0;\mathbf{Y})d\mathbf{Y}\right] = \nabla_{\psi^0}[1] = \mathbf{0},\end{aligned} \tag{5.12}$$

où $\mathbb{E}_{\psi^0}$ dénote l'opérateur d'espérance sous la loi de probabilité P_{ψ^0}, et où l'échange gradient-intégrale de la dernière ligne est licite pourvu que L soit suffisamment régulière. L'équation 5.12 joue un rôle d'importance pour l'obtention et la compréhension des résultats de consistance et de normalité asymptotique de l'estimateur du maximum de vraisemblance ; elle permet en effet d'établir que la vraisemblance moyenne admet un point critique en le "vrai" paramètre ψ^0. Cela n'est a priori pas suffisant pour affirmer que $\mathbb{E}_{\psi^0}[\mathcal{L}(.;\mathbb{Y})]$ admet son maximum en ψ^0. Par bonheur, cette dernière assertion est en fait bien fondée comme nous allons le montrer ci-dessous après avoir introduit la *divergence de Küllback-Leibler* (quelquefois abusivement appelée *distance K-L*) :

$$\begin{aligned}KL(\psi||\psi^0) :&= \mathbb{E}_{\psi^0}[\mathcal{L}(\psi^0;\mathbb{Y}) - \mathcal{L}(\psi;\mathbb{Y})] \\ &= \mathbb{E}_{\psi^0}\left[\log\left(\frac{L(\psi^0;\mathbb{Y})}{L(\psi;\mathbb{Y})}\right)\right] = -\mathbb{E}_{\psi^0}\left[\log\left(\frac{L(\psi;\mathbb{Y})}{L(\psi^0;\mathbb{Y})}\right)\right]\end{aligned} \tag{5.13}$$

En utilisant la dernière égalité de l'éq. 5.13 ainsi qu'une application de l'inégalité de convexité de Jensen à l'opposé du logarithme néperien, il vient que

$$KL(\psi||\psi^0) \geq -\log\left(\underbrace{\mathbb{E}_{\psi^0}\left[\left(\frac{L(\psi;\mathbb{Y})}{L(\psi^0;\mathbb{Y})}\right)\right]}_{=\int_{\mathcal{Y}}L(\psi;\mathbf{Y})d\mathbf{Y}=1}\right) = 0 \tag{5.14}$$

On constate que $KL(\psi||\psi^0) \geq 0$, avec égalité lorsque $L(\psi;.) = L(\psi^0;.)$, i.e. si et seulement si $\psi = \psi^0$ dès lors que le modèle est identifiable. On peut ainsi obtenir un résultat encore plus fort que l'éq. 5.12 en ce qui concerne le comportement de $\mathbb{E}_{\psi^0}[\mathcal{L}(.;\mathbb{Y})]$ en ψ^0, comme énoncé ci-dessous.

Maximum global de la log-vraisemblance moyenne.

$$\forall \psi \neq \psi^0,\ \mathbb{E}_{\psi^0}[\mathcal{L}(\psi;\mathbb{Y})] < \mathbb{E}_{\psi^0}[\mathcal{L}(\psi^0;\mathbb{Y})], \tag{5.15}$$

c'est-à-dire que ψ^0 *est l'unique maximiseur global de* $\mathbb{E}_{\psi^0}[\mathcal{L}(.;\mathbb{Y})]$.

Or comme nous allons le détailler ci-après, la surface de log-vraisemblance associée à un grand nombre d'observations indépendantes est une somme de surfaces de log-vraisemblance respectivement associées à chacune de ces observations. La loi des grands nombres et le théorème « central limit »(TCL) vont nous permettre de tirer des conclusions intéressantes sur le comportement asymptotique de l'EMV. Intéressons-nous maintenant pour ce faire à la notion de matrice d'information.

Matrices d'information de Fisher observée et théorique

La matrice hessienne (dérivées partielles secondes) de $-\mathcal{L}(\psi;y)$, définie par Fisher en 1925 (Cf. par exemple [Lin96]), est appelée *matrice d'information observée* :

$$\mathcal{J}(\psi;\mathbf{Y}) := -\nabla^2_\psi \mathcal{L}(\psi;\mathbf{Y}) = -\left(\frac{\partial^2 \mathcal{L}(\psi;\mathbf{Y})}{\partial\psi_i \partial\psi_j}\right)_{1\leq i,j\leq p} \tag{5.16}$$

En tant que hessienne, il est clair que $\mathcal{J}$ est symétrique (par le théorème de Schwarz, pourvu que la condition suffisante classique sur le caractère ouvert du domaine Ψ soit remplie). Revenons à l'exemple d'un échantillon d'observations gaussiennes i.i.d. On obtient la matrice d'information observée par calcul direct :

$$\mathcal{J}(\mu,\sigma^2;\mathbf{Y}) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{\sum_{i=1}^n (y_i-\mu)}{\sigma^4} \\ \frac{\sum_{i=1}^n (y_i-\mu)}{\sigma^4} & -\frac{n}{2\sigma^4} + \frac{\sum_{i=1}^n (y_i-\mu)^2}{\sigma^6} \end{pmatrix} \tag{5.17}$$

Il apparaît clairement que l'information disponible sur les paramètres μ et σ^2 décroît avec la variance σ^2 (Cf. [Lin96] p. 97 pour de plus amples commentaires sur cet exemple). Il est courant d'utiliser l'inverse $\mathcal{J}^{-1} = (\mathcal{J}^{ij})_{i,j}$ pour étudier le comportement de la fonction de log-vraisemblance : les $\sqrt{\mathcal{J}^{ii}}$ fournissent des approximations (à l'ordre 2) de la largeur de $\mathcal{L}$ dans chaque direction i. La quantité $\frac{1}{\sqrt{|\mathcal{J}|}}$ est aussi parfois employée comme une mesure approximative de la « largeur globale »de la fonction de log-vraisemblance au point considéré. Pour revenir à l'éq. 5.17, remarquons enfin qu'en vertu de la loi des grands nombres, les termes non-diagonaux tendent à s'annuler lorsque n croît.

La *matrice d'information théorique* de Fisher est définie comme l'espérance de $\mathcal{J}$:

$$\mathcal{I}(\psi) := \mathbb{E}_\psi\left[-\nabla^2_\psi \mathcal{L}(\psi;\mathbb{Y})\right] = \left(-\mathbb{E}_\psi\left[\frac{\partial^2 \mathcal{L}(\psi;\mathbb{Y})}{\partial\psi_i \partial\psi_j}\right]\right)_{1\leq i,j\leq p} \tag{5.18}$$

Elle joue un rôle clef dans l'étude des propriétés asymptotiques de l'EMV. Pour poursuivre notre exemple gaussien (Cf. 5.17), le calcul de l'information théorique donne

$$\mathcal{I}(\mu, \sigma^2) = \mathbb{E}[\mathcal{J}(\mu, \sigma^2; \mathbb{Y})] = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{5.19}$$

La prochaine propriété va nous permettre de faire le lien entre la matrice d'information théorique et les moments d'ordre 2 du score.

Identité de Bartlett. $\forall \psi \in \Psi$,

$$\mathcal{I}(\psi) = \left(\mathbb{E}_\psi \left[\frac{\partial \mathcal{L}(\psi; \mathbb{Y})}{\partial \psi_i} \frac{\partial \mathcal{L}(\psi; \mathbb{Y})}{\partial \psi_j} \right] \right)_{1 \leq i,j \leq p} = \mathbb{E}_\psi \left[(U(\psi; \mathbb{Y})) \, (U(\psi; \mathbb{Y}))^T \right] \tag{5.20}$$

Démonstration. Il suffit de montrer que pour $i, j \in [1, p]$ quelconques, $\mathbb{E}_\psi \left[\frac{\partial^2 \mathcal{L}(\psi; \mathbb{Y})}{\partial \psi_i \partial \psi_j} \right] = \mathbb{E}_\psi \left[\frac{\partial \mathcal{L}(\psi; \mathbb{Y})}{\partial \psi_i} \frac{\partial \mathcal{L}(\psi; \mathbb{Y})}{\partial \psi_j} \right]$. Or on a $\frac{\partial \mathcal{L}(\psi; \mathbf{Y})}{\partial \psi_j} = \frac{1}{L(\psi; \mathbf{Y})} \frac{\partial L(\psi; \mathbf{Y})}{\partial \psi_j}$, et une deuxième dérivation par rapport à ψ_i donne :

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\psi; \mathbf{Y})}{\partial \psi_i \partial \psi_j} &= -\left(\frac{1}{L(\psi; \mathbf{Y})^2} \frac{\partial L(\psi; \mathbf{Y})}{\partial \psi_i} \right) \frac{\partial L(\psi; \mathbf{Y})}{\partial \psi_j} + \frac{1}{L(\psi; \mathbf{Y})} \frac{\partial^2 L(\psi; \mathbf{Y})}{\partial \psi_i \partial \psi_j} \\ &= -\frac{\partial \mathcal{L}(\psi; \mathbf{Y})}{\partial \psi_i} \frac{\partial \mathcal{L}(\psi; \mathbf{Y})}{\partial \psi_j} + \frac{1}{L(\psi; \mathbf{Y})} \frac{\partial^2 L(\psi; \mathbf{Y})}{\partial \psi_i \partial \psi_j} \end{aligned}$$

Il reste alors à passer à l'espérance et à remarquer que sous des hypothèses (de convergence dominée à l'ordre 2) sur $L(\psi; .)$, $\mathbb{E}_\psi \left[\frac{1}{L(\psi; \mathbb{Y})} \frac{\partial^2 L(\psi; \mathbb{Y})}{\partial \psi_i \partial \psi_j} \right] = 0$ (analogue à l'éq. 5.12). La dernière égalité de l'éq. 5.20 est immédiate en appliquant la définition de U. □

5.1.3 Asymptotique pour des observations indépendantes

Le cas d'observations i.i.d. se prête particulièrement bien à l'étude du comportement de la log-vraisemblance puisqu'il fait apparaître des sommes de variables aléatoires indépendantes. Nous allons voir ci-dessous comment la loi des grands nombres et le TCL permettent d'obtenir des résultats asymptotiques remarquables au sujet de l'estimateur du maximum de vraisemblance.

Log-vraisemblance, score, et information de Fisher dans le cas i.i.d.

On suppose ici que la fonction de log-vraisemblance est suffisamment régulière pour que toutes les écritures aient un sens. En vertu de la multiplicativité de la densité d'un ensemble de v.a. indépendantes, la log-vraisemblance d'un vecteurs d'observations $\mathbb{Y} =$

$(Y_1, ..., Y_n)$ indépendantes identiquement distribuées (i.i.d.) s'écrit

$$\mathcal{L}_n(\psi) := \mathcal{L}(\psi; \mathbb{Y}) = \log\left(\prod_{i=1}^{n} L(\psi; Y_i)\right) = \sum_{i=1}^{n} \mathcal{L}(\psi; Y_i) \tag{5.21}$$

En particulier, une application de la loi faible des grands nombres nous enseigne que

$$\forall \psi \in \Psi,\ \frac{1}{n}\mathcal{L}_n(\psi) - \mathbb{E}_{\psi^0}[\mathcal{L}(\psi, Y_1)] \xrightarrow{\mathbb{P}} 0 \tag{5.22}$$

où $\mathbb{E}_{\psi^0}[\mathcal{L}(\psi, Y_1)]$ est la vraisemblance moyenne d'une observation tirée selon la loi de paramètre ψ^0. En d'autres termes, la surface de vraisemblance (normalisée par un facteur $\frac{1}{n}$) associée à n observations indépendantes tirées selon la loi de paramètre ψ^0 converge vers la surface de vraisemblance moyenne $\mathbb{E}_{\psi^0}[\mathcal{L}(., Y_1)]$ associée à une seule observation [6].

Pour ce qui est des dérivées de la log-vraisemblance, on a de même :

$$\begin{cases} \nabla_\psi \mathcal{L}_n(\psi) = \sum_{i=1}^{n} \nabla_\psi \mathcal{L}(\psi; Y_i) \\ \nabla^2_\psi \mathcal{L}_n(\psi) = \sum_{i=1}^{n} \nabla^2_\psi \mathcal{L}(\psi; Y_i) \end{cases} \tag{5.23}$$

Par conséquent, la matrice d'information empirique —resp. théorique— associée à l'échantillon $\{Y_1, ..., Y_n\}$ est la somme des matrices d'information empiriques —resp. théoriques— associées à chacune des observations. On notera désormais $\mathcal{I}_n(\psi)$ la matrice d'information théorique associée à n observations i.i.d. tirées selon la loi de paramètre ψ. Les résultats suivants jouent un rôle clef dans la distribution asymptotique de l'EMV :

Convergences asymptotiques du score et de l'information empirique.

$$\forall \psi' \in \Psi,\ \frac{1}{n}\nabla^2_\psi \mathcal{L}_n(\psi') \xrightarrow{\mathbb{P}} -\mathcal{I}_1(\psi') \tag{5.24}$$

$$\frac{1}{\sqrt{n}}\nabla_\psi \mathcal{L}_n(\psi^0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_1(\psi^0)\right) \tag{5.25}$$

Démonstration. 5.24 découle de la deuxième ligne de l'équation 5.23 et d'une application de la loi des grands nombres à $\frac{1}{n}\sum_{i=1}^{n} \nabla^2_\psi \mathcal{L}(\psi; Y_i)$, en se rappelant que par définition $\mathbb{E}_\psi[\nabla^2_\psi \mathcal{L}(\psi; Y_1)] = -\mathcal{I}_1(\psi)$. De manière analogue, 5.25 est le résultat direct d'une application du TCL à la quantité $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_{\psi^0} \mathcal{L}(\psi; Y_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_{\psi^0} U(\psi; Y_i)$. La moyenne et la matrice de covariance asymptotiques du score proviennent respectivement de 5.12 et de l'identité de Bartlett 5.20. □

[6] Le mode de convergence reste à préciser, en particulier en fonction de la topologie de Ψ

Généralités sur la consistance asymptotique des estimateurs

Une des propriétés souhaitables d'un estimateur est qu'il permette de retrouver les vrais paramètres (quels qu'ils soient) aussi précisément que voulu lorsque n croît. Cette propriété porte le nom de *consistance*. L'estimateur classique de la moyenne fournit une illustration fondamentale de cette propriété : si $X \in \mathcal{L}^2$, $\psi^0 = \mathbb{E}[X]$, et $\{X_i, i \in \mathbb{N}^*\}$ est une famille de réplications indépendantes de X, l'estimateur défini par $\widehat{\psi_n} = \frac{X_1+...+X_n}{n}$ est asymptotiquement consistant : on a en effet $\widehat{\psi_n} \xrightarrow{\mathbb{P}} \psi^0$ d'après la loi des grands nombres. Depuis les travaux fondateurs de Sir Fisher dans les années 1920, une vaste littérature s'est dévelopée sur ce sujet, à laquelle la *preuve de consistance de Wald* (1949) constitue une des contributions majeures. Nous suivons plutôt ici l'approche générique employée dans [VdV98] pour traiter le sujet de la consistance des *M-estimateurs*. Le résultat suivant donne une condition suffisante pour qu'une suite d'estimateurs définis à partir de fonctions aléatoires M_n (comme la vraisemblance) convergent en probabilité vers le maximiseur de leur fonction limite M :

Une condition suffisante de consistance. *Soient M_n des fonctions aléatoires et M une fonction fixée de ψ telle que pour tout $\epsilon > 0$* [7]

$$\begin{cases} sup_{\psi\in\Psi}|M_n(\psi) - M(\psi)| \xrightarrow{\mathbb{P}} 0, \\ sup_{\{\psi\in\Psi:\ d(\psi,\psi_0)\geq\epsilon\}} M(\psi) < M(\psi_0) \end{cases} \tag{5.26}$$

Alors toute séquence d'estimateurs $\widehat{\psi_n}$ telle que $M_n(\widehat{\psi_n}) \geq M_n(\psi_0) - o_{\mathbb{P}}(1)$ converge vers ψ_0 en probabilité.

Démonstration. On peut trouver une preuve très accessible dans ([VdV98], p.46) □

On pourra remarquer que la condition sur la fonction déterministe M garantit que si $M(\psi)$ est suffisament proche de $M(\psi_0)$, ψ est nécessairement proche de ψ_0. Cela permet d'écarter les cas pathologiques où le maximiseur est mal séparé (Cf. [VdV98], p. 45).

Dans le cas de l'EMV pour des observations i.i.d., la deuxième condition de la propriété 5.26 est assurée par l'éq. 5.15 sous réserve que le modèle soit identifiable. Si de plus la convergence de la surface de vraisemblance vers la surface de vraisemblance espérée est uniforme (au sens de la seconde condition de 5.26), on peut conclure que l'estimateur du maximum de vraisemblance est consistant (Cf. [VdV98] p. 62).

[7] Comme précisé dans [VdV98], certaines expressions de ce théorème peuvent être non-mesurables. Les probabilités sont alors à comprendre en termes de mesure extérieure.

Normalité asymptotique de l'EMV. *Soit ψ_n^* la suite des estimateurs du maximum de vraisemblance associés à des observations i.i.d. tirées selon la loi de paramètre ψ^0, un point intérieur de Ψ. Dans des conditions où l'EMV est consistant et pour une fonction de vraisemblance suffisament régulière, on a*

$$\sqrt{n}\left(\psi_n^* - \psi^0\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathcal{I}_1(\psi^0)^{-1}), \text{ i.e. } \psi_n^* \overset{\mathcal{L}}{\approx} \mathcal{N}(\psi_0, \mathcal{I}_n(\psi^0)) \tag{5.27}$$

Démonstration. Comme par hypothèse $\psi_n^* \xrightarrow{\mathbb{P}} \psi^0$, ψ_n^* est intérieur à Ψ pour n suffisament grand, et la régularité des fonctions de vraisemblance garantit que ψ_n^* est un point critique de $\mathcal{L}_n$. Un développement limité (développement de Taylor exact en version multivariables) de $\nabla_\psi \mathcal{L}_n$ autour de ψ^0 donne

$$0 = \nabla_\psi \mathcal{L}_n(\psi_n^*) = \nabla_\psi \mathcal{L}_n(\psi^0) + \nabla_\psi^2 \mathcal{L}_n(\psi_n^1)(\psi_n^* - \psi^0), \tag{5.28}$$

où ψ_n^1 a ses coordonnées sur la corde joignant les coordonnées respectives de ψ^0 et ψ^*. Il vient alors après multiplication par $\frac{1}{\sqrt{n}}\mathcal{I}_1^{-\frac{1}{2}}$ de part et d'autre de l'éq. 5.28, que :

$$\underbrace{-\mathcal{I}_1^{-\frac{1}{2}}\left(\frac{1}{n}\nabla_\psi^2 \mathcal{L}_n(\psi_n^1)\right)}_{Y_n}\left\{\sqrt{n}(\psi_n^* - \psi_0)\right\} = \underbrace{\frac{1}{\sqrt{n}}\mathcal{I}_1^{-\frac{1}{2}}\nabla_\psi \mathcal{L}_n(\psi_0)}_{X_n} \tag{5.29}$$

Or, 5.25 entraine que $X_n \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{I})$, et 5.24 combinée avec l'hypothèse de consistance de ψ_n^* —qui implique que $\psi_n^1 \xrightarrow{\mathbb{P}} \psi^0$, par encadrement— ainsi que la continuité de l'inverse, nous donnent que $Y_n \xrightarrow{\mathbb{P}} \mathcal{I}_1^{\frac{1}{2}}$. Une application du point (*iii*) du lemme de Slutsky (Cf. ci-dessous) permet d'obtenir

$$\sqrt{n}(\psi_n^* - \psi_0) = Y_n^{-1} X_n \xrightarrow{\mathcal{L}} \mathcal{I}_1^{-\frac{1}{2}} \mathcal{N}(\mathbf{0}, \mathbf{I}) \equiv \mathcal{N}(\mathbf{0}, \mathcal{I}_1^{-1}) \tag{5.30}$$

Lemme de Slutsky ([VdV98], p. 11). *Soient X_n, X et Y_n des variables ou vecteurs aléatoires (X_n et Y_n étant définis sur le même espace de probabilité Ω_n pour chaque $n \in \mathbb{N}$). Si $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathcal{L}} c$ où c est une constante, alors*

(i) $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$

(ii) $X_n Y_n \xrightarrow{\mathcal{L}} cX$ (5.31)

(iii) $Y_n^{-1} X_n \xrightarrow{\mathcal{L}} c^{-1} X$ *pourvu que c soit inversible.*

□

5.2 EMV des paramètres de covariance du Krigeage Simple

De même que dans le cas i.i.d. précédemment traité, l'estimateur du maximum de vraisemblance peut être utilisé pour estimer les paramètres de covariance d'un processus aléatoire Y sur la base d'un ensemble d'observations. Le fait que ces observations soient issues d'une unique réalisation de Y est loin d'être sans conséquence sur les propriétés de l'estimateur : le type de résultats obtenus dans le cas i.i.d. ne peuvent se transposer au cas de processus que si certaines hypothèses très fortes sur Y et le plan d'expériences $\mathbf{X}$ sont vérifiées (e.g. stationnarité, ergodicité de Y, plan d'expériences adapté, etc.).

L'EMV reposant sur une famille paramétrique de lois de probabilité, on adopte l'hypothèse gaussienne pour le processus Y ; cela revient en pratique à supposer que quelque soit $\mathbf{X}$, $Y(\mathbf{X})$ et un vecteur gaussien (de moyenne nulle dans le cadre du KS, quitte à centrer les observations). Une fois observé $Y(\mathbf{X}) = \mathbf{Y}$ et choisie une famille de noyaux de covariance $\{\sigma^2 r_\psi,\ \psi \in \Psi,\ \sigma \in [0, +\infty[\}$[8] , on se ramène alors à estimer σ^2 et ψ en maximisant la vraisemblance L, ou encore la log-vraisemblance $\mathcal{L}$:

$$\begin{cases} L(\sigma^2, \psi; \mathbf{Y}) := (2\pi)^{-\frac{n}{2}} det[K_{\sigma^2,\psi}]^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{Y}^T K^{-1}_{\sigma^2,\psi}\mathbf{Y}} \\ \mathcal{L}(\sigma^2, \psi; \mathbf{Y}) := -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(det[K_{\sigma^2,\psi}]) - \frac{1}{2}\mathbf{Y}^T K^{-1}_{\sigma^2,\psi}\mathbf{Y}, \end{cases} \tag{5.32}$$

ce qui est équivalent à résoudre le problème de minimisation suivant

$$\min_{\psi\in\Psi, \sigma^2\in[0,+\infty[} \left\{ \log(det[K_{\sigma^2,\psi}]) + \mathbf{Y}^T K^{-1}_{\sigma^2,\psi}\mathbf{Y} \right\}. \tag{5.33}$$

Remarquons que la résolution de ce programme de minimisation n'est pas évidente : la fonction objectif n'a aucune raison d'être convexe en (σ^2, ψ) (même s'il est tentant de la regarder comme une fonction de $\mathbf{Y}$), et nécessite de mettre en œuvre des méthodes d'optimisation numérique (Cf. section 5.2.1). Une fois franchies les difficultés numériques, les estimations obtenues par EMV sont succeptibles de différer largement des vrais paramètres de covariance : comme dans le cas d'observations i.i.d., l'EMV possède ici une loi de probabilité, dépendant du processus considéré et du plan d'expériences. Nous aborderons dans la section 5.2.2 la question de ce que l'on peut dire et ne pas dire sur la consistance et la normalité de l'EMV dans les conditions du Krigeage Simple, et discuterons le cas non-asymptotique à l'aune de résultats expérimentaux sur la variabilité de l'EMV lorsque le nombre d'observations est petit (section 5.2.3).

[8] On va se restreindre ici par commodité à des noyaux de covariance stationnaires (les r_ψ étant des fonctions de corrélation). La démarche présentée reste cependant valable dans un cadre plus général.

5.2.1 Aspects pratiques : un problème d'optimisation non-convexe

Résoudre 5.33 n'est généralement pas possible de manière analytique. On a alors affaire à un problème d'optimisation numérique. Voyons dans un premier temps comment ce problème peut souvent être simplifié en introduisant la vraisemblance concentrée.

Log-vraisemblance concentrée

Lorsque la matrice de covariance s'écrit $K_{\sigma^2,\psi} = \sigma^2 R_\psi$ —ce qui est le cas lorsque l'on observe un processus stationnaire Y non-bruité—, où R_ψ est la matrice de corrélation des observations, l'équation 5.33 peut s'écrire

$$\min_{\psi\in\Psi,\sigma^2\in[0,+\infty[} \left\{ n\log(\sigma^2) + \log(det[R_\psi]) + \frac{1}{\sigma^2}\mathbf{Y}^T R_\psi^{-1}\mathbf{Y} \right\} \tag{5.34}$$

Pour $\psi \in \Psi$ quelconque fixé, on remarque que l'on peut obtenir une expression du σ^2 optimal correspondant, comme fonction de ψ :

$$\left(\frac{n}{\widehat{\sigma^2}} - \frac{1}{\widehat{\sigma^2}^2}\mathbf{Y}^T R_\psi^{-1}\mathbf{Y} = 0 \right) \implies \boxed{\widehat{\sigma^2}(\psi) = \frac{\mathbf{Y}^T R_\psi^{-1}\mathbf{Y}}{n}} \tag{5.35}$$

En injectant l'expression du σ^2 optimal dans 5.32, on peut définir la fonction de *log-vraisemblance concentrée*

$$\mathcal{L}_c(\psi;\mathbf{Y}) := \mathcal{L}(\widehat{\sigma^2}(\psi),\psi;\mathbf{Y}), \tag{5.36}$$

et l'EMV s'obtient alors en résolvant une optimisation portant uniquement sur ψ :

$$\min_{\psi\in\Psi}\{\mathcal{L}_c(\psi;\mathbf{Y}) - n\log(2\pi e)\} \equiv \min_{\psi\in\Psi}\left\{ n\log\left(\frac{\mathbf{Y}^T R_\psi^{-1}\mathbf{Y}}{n}\right) + \log(det[R_\psi]) \right\} \tag{5.37}$$

Une fois $\widehat{\psi}$ trouvé, il suffit ainsi de calculer σ^2 en utilisant l'éq. 5.35. Voyons maintenant comment maximiser $\mathcal{L}_c$ en pratique. Commençons par un calcul de gradient.

Gradient de la vraisemblance concentrée.

$$\forall i \in [1,p],\ \frac{\partial\mathcal{L}_c(\psi;\mathbf{Y})}{\partial\psi_i} = -n(\mathbf{Y}^T R_\psi^{-1}\mathbf{Y})^{-1}\mathbf{Y}^T R_\psi^{-1}\frac{\partial R_\psi}{\partial\psi_i}R_\psi^{-1}\mathbf{Y} + tr\left(R_\psi^{-1}\frac{\partial R_\psi}{\partial\psi_i}\right) \tag{5.38}$$

Démonstration. Les deux termes de la somme sont obtenus directement par différentiation des deux termes de 5.37, en utilisant pour chacun la règle de différentiation des fonctions composées (*chain rule*, souvent attribuée à Leibniz). Le premier terme nécessite

d'employer la différentielle de l'inverse ; on rappelle que pour E et F deux espaces de Banach, l'application $inv : Isom(E,F) \to Isom(F,E)$ qui u associe u^{-1} est C^1 avec $Dinv(u)(v) = -u^{-1}.v.u^{-1}$. Il vient alors que $d(R_\psi^{-1}) = -R_\psi^{-1}\frac{\partial R_\psi}{\partial \psi_i}R_\psi^{-1}$. Pour le second terme, on utilise la différentielle du déterminant (Cf. annexe), et on obtient $d\left(\log(det[R_\psi])\right) = \frac{1}{det[R_\psi]}\left(det[R_\psi]\mathrm{tr}\left(R_\psi^{-1}\frac{\partial R_\psi}{\partial \psi_i}\right)\right) = \mathrm{tr}\left(R_\psi^{-1}\frac{\partial R_\psi}{\partial \psi_i}\right)$ □

Algorithmes d'optimisation employés

La littérature de l'EMV pour les processus gaussiens est vaste, et nous ne pouvons garantir une vision exhaustive des techniques d'optimisation employées pour résoudre 5.37. On peut cependant affirmer avoir souvent rencontré le nom de *scoring*, encore appelé *algorithme de Newton-Raphson* ; il se résume à un algorithme de recherche de zéro (la méthode de Newton en version multivariable, Cf. annexe) appliquée au score afin de trouver un point critique de la vraisemblance. Il est important de préciser que cette méthode repose sur l'inversion d'une matrice d'information à chaque itération. Elle semble néanmoins s'être imposée comme l'une des routines d'optimisation les plus populaires dans les applications statistiques. L'article de référence [MM84] suggère d'augmenter la robustesse de cette technique en incorporant un paramètre de Levenberg-Marquart afin de garantir une amélioration de la vraisemblance à chaque itération. Les auteurs évoquent aussi la possibilité d'utiliser une algorithme de quasi-Newton.

Nous avons choisi ici, comme pour la maximisation de l'EI au chapitre 4, d'utiliser un algorithme génétique hybride avec optimisations locales d'ordre 1 (*genoud*, Cf. [MS08]). Cette méthode a donné des résultats bien meilleurs que les techniques de descente de type BFGS (Implémentées dans les méthodes de base de [dCT06]), à la fois en termes de performances extrêmes que de robustesse. Le prix à payer pour une robustesse accrue est bien sûr le temps de calcul, qui augmente avec la taille de la population.

Exemples

Ex. 1 : On considère un PG tri-dimensionnel Y, centré, de noyau gaussien isotrope et de paramètres $(\sigma^2, \psi) \equiv (\sigma^2, \theta) = (1, 0.5)$. Les plans d'expériences $\{\mathbf{X}_i, i \in \{1, ..., 27\}\}$ sont ici des réalisations i.i.d. d'un plan aléatoire $\mathbb{X}$ de loi uniforme sur les 10-uplets de points de $[0,1]^3$. On s'intéresse aux fonctions de vraisemblance associées à l'observation de Y en les $\mathbf{X}_i$. La figure 5.2.1 illustre les 27 surfaces de log-vraisemblance obtenues en représentant la surface de vraisemblance associée aux observations en $\mathbf{X}_i$ d'une réalisation y_i du processus Y (les 27 y_i étant i.i.d. et tirées indépendamment des plans).

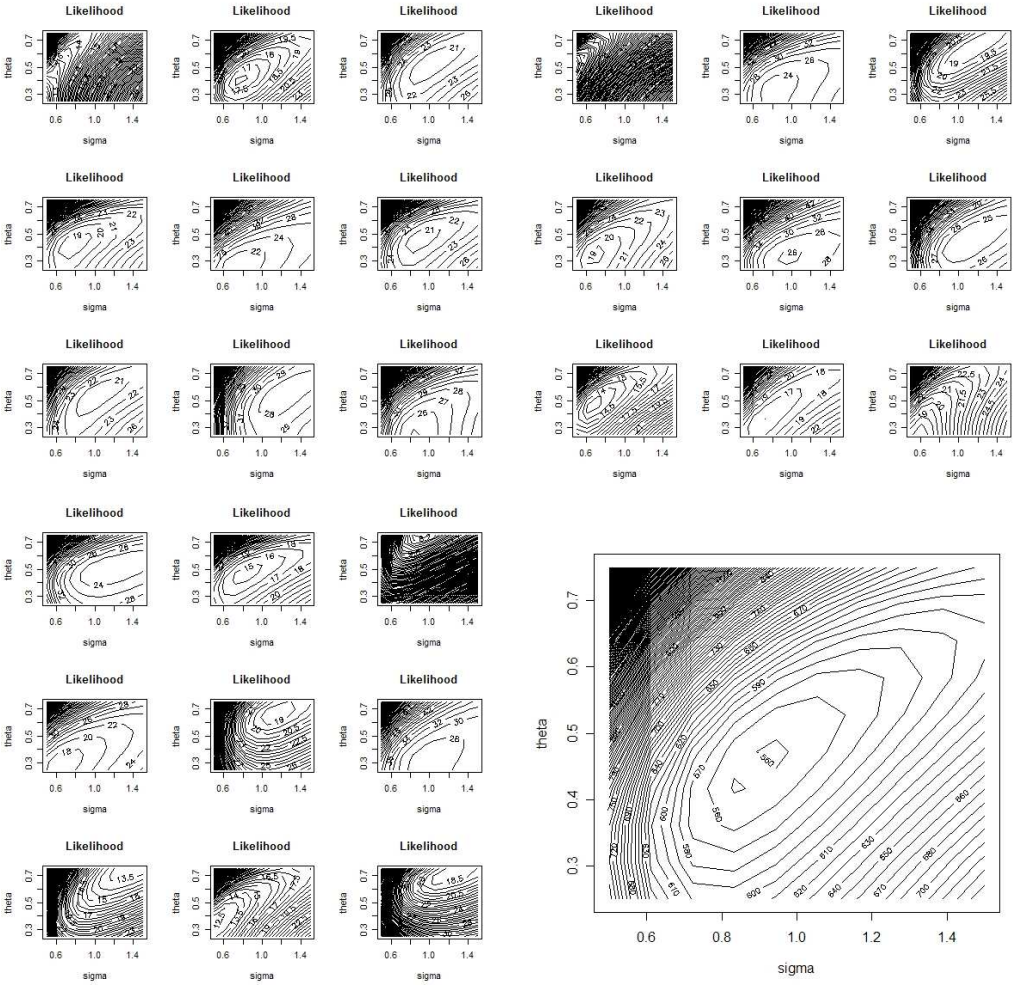


FIG. 5.1 – On observe que l'échantillon des 27 fonctions de $(-2\times)$log-vraisemblance obtenues présente une grande variabilité ; en particulier, les minimiseurs globaux des ces surfaces —i.e. les estimations par MV de (σ^2, θ)— sont dispersés spatialement et peuvent parfois être très différents des valeurs (σ^2, θ) des paramètres de covariance effectivement utilisés pour générer les observations. En revanche la somme des 27 surfaces (graphe en bas à droite) présente bien nettement son minimum global au niveau de (σ^2, θ), ce qui peut être vu comme une illustration du résultat 5.15 dans le contexte du KS.

Ex. 2&3 : on reprend l'exemple 1 avec des plans de taille 30, respectivement en dimension 3 (exemple 2) et en dimension 10. Comme dans l'exemple 1, 27 réalisations de $Y(\mathbb{X})$ sont considérées, mais seules 9 d'entre elles sont représentées sur la figure 5.2.1.

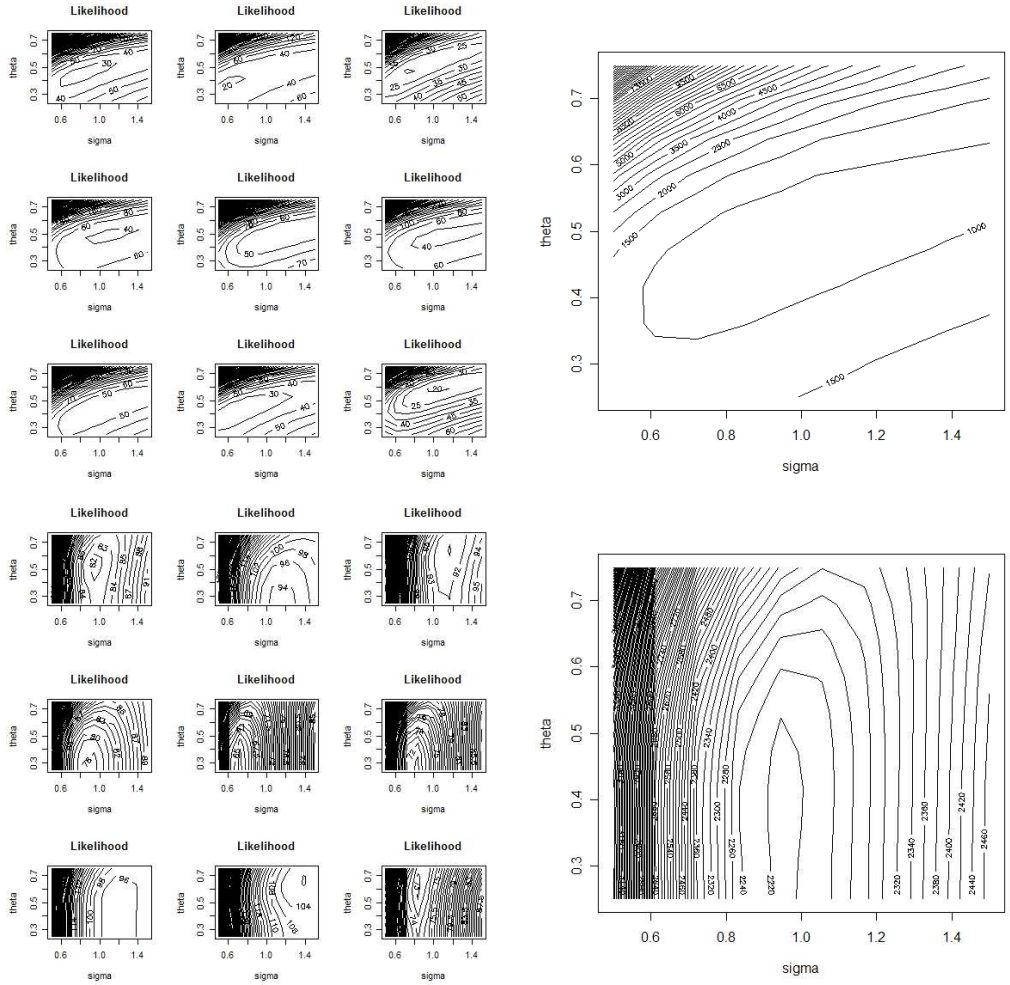


FIG. 5.2 – 9 des 27 fonctions de $(-2\times)$log-vraisemblance obtenues dans les exemples 2 (en haut à gauche) et 3 (en bas à gauche), ainsi que les sommes des deux ensembles de 27 surfaces (respectivement en haut à droite et en bas à droite). Les graphes relatifs à l'exemple 2 illustrent le fait que 30 points choisis uniformément dans le cube suffisent pour estimer convenablement (en prenant peu de risques, Cf. la variabilité des 9 surfaces) les paramètres de covariance $(\sigma^2, \theta) = (1, 0.5)$ du processus Y considéré ; par ailleurs, la surface somme indique une estimation assez précise du paramètre θ mais une estimation plus grossière du paramètre σ^2 (variation lente de L dans la direction horizontale). Les graphes relatifs à l'exemple 3 illustrent que la procédure d'estimation avec 30 points tirés dans l'hypercube unité fonctionne toujours en dimension 10, mais que la variabilité de l'EMV se fait ressentir cette fois-ci plutôt au niveau de l'estimation de θ que de celle de σ^2 ; cela conforte l'intuition selon laquelle il devient plus difficile d'estimer la valeur d'un paramètre de portée à mesure que les interdistances du même ordre de grandeur que cette portée se font rares dans le plan d'expériences.

5.2.2 Aspects théoriques : consistance et normalité asymptotique ?

Les propriétés asymptotiques de l'EMV dans le cadre d'observations dépendantes ont été abondamment étudiées depuis le milieu des années 1970. En particulier l'article de Sweeting [Swe80] donne un jeu de conditions suffisantes pour garantir la consistance et la normalité asymptotiques de l'EMV des paramètres d'un processus stochastique. Ces conditions très générales (*Growth and convergence* "C1" et *Continuity* "C2" dans [Swe80], p.1376) portent sur le comportement asymptotique des matrices d'information empiriques lorsque le nombre d'observations augmente.

Les conditions de [Swe80] sont formulées d'une manière assez abstraite ; l'information théorique de Fisher n'y apparaît d'ailleurs que comme un cas particulier dans le rôle de limite de l'information empirique. L'article de Mardia et Marshall [MM84] apporte quant à lui un regard plus spécifique aux applications en statistique spatiale et propose deux reformulations relativement simplifiées des conditions suffisantes de consistance et normalité asymptotiques énoncées dans [Swe80]. La première de ces reformulations consiste en les trois points suivants :

1. Continuité : noyau de covariance $\mathcal{C}^2$ en ψ sur le domaine Ψ ($\forall \mathbf{x},\ \mathbf{y} \in D$)
2. Information croissante : $\mathcal{I}_n^{-1} \xrightarrow{\mathbb{P}} \mathbf{0}$
3. Convergence de l'information empirique : $\mathcal{I}_n^{-\frac{1}{2}} \mathcal{J}_n^{-1} \mathcal{I}_n^{-\frac{1}{2}} \xrightarrow{\mathbb{P}} I$

Le théorème 1 de [MM84] établit que les trois conditions ci-dessus sont suffisantes pour que la relation $\psi_n^* \overset{\mathcal{L}}{\approx} \mathcal{N}(\psi^0, \mathcal{I}_n(\psi^0))$ rencontrée dans le cas i.i.d. reste valable. La convergence de l'information (point 2) signifie que la plus petite valeur propre de $\mathcal{I}_n$ tend vers $+\infty$, i.e. que l'échantillon considéré devient infiniment informatif au sujet de tous les paramètres du modèle. Remarquons d'autre part que le point 3 peut être remplacé par une condition de convergence en moyenne quadratique (puisqu'elle est plus forte que la convergence en probabilité), i.e. $\lim \mathbb{E}\left[||\mathcal{I}_n^{-\frac{1}{2}} \mathcal{J}_n^{-1} \mathcal{I}_n^{-\frac{1}{2}} - I||^2 \right] = 0$. Suivant ([MM84], p. 138), cette dernière s'écrit aussi sous la forme développée

$$\lim \sum_{i,j,k,l=1}^{p} \mathcal{I}_n^{ki} \mathcal{I}_n^{lj} \mathrm{tr}\left(K^{kj} K K^{li} K \right) = 0, \tag{5.39}$$

où $\forall i, j \in [1, p]$, $\mathcal{I}_n^{ij} = (\mathcal{I}_n^{-1})_{ij}$ et $K^{ij} = \left(\frac{\partial K_\psi}{\partial \psi}(\psi^0) \right)_{ij}$. Le lecteur non-spécialiste de ce sujet conviendra aisément qu'il n'est pas évident de se faire une intuition sur la base de l'éq. 5.39. Le deuxième théorème issu de [MM84] —qui constitue le coeur de l'article en

question— donne un jeu de conditions équivalentes, vraisemblablement plus propices à l'interprétation qualitative :

Deuxième théorème de Mardia & Marshall ([MM84], p.139).

Soient $\lambda_1 \leq \ldots \leq \lambda_n$ les valeurs propres ordonnées de K, et soient λ_k^i et λ_k^{ij} $(k = 1, \ldots, n)$ celles, ordonnées selon les modules croissants, des matrices dérivées première et seconde K_i et $K_{i,j}$ $(1 \leq i, j \leq p)$. Si les conditions suivantes sont remplies pour $n \to +\infty$:

1. $\forall i, j \in [1, p]$, $\lim \lambda_n = C < \infty$, $\lim |\lambda_n^i| = C_i < \infty$, $\lim |\lambda_n^{ij}| = C_{ij} < \infty$
2. $\forall i \in [1, p]$, $||K_i||^{-2} = O(n^{-\frac{1}{2}-\delta})$ *avec* $\delta > 0$,
3. $\forall i, j \in [1, p]$, $a_{ij} = \lim \left(\left\{ \frac{t_{ij}}{t_i t_j} \right\}^{\frac{1}{2}} \right)$ *existe, où* $t_{ij} = tr\big(K^{-1} K_i K^{-1} K_j\big)$, *et la matrice* $A = (a_{ij})_{i,j \in [1,p]}$ *est inversible,*

alors l'EMV est consistant et asymptotiquement normal, avec $\psi_n^* \overset{\mathcal{L}}{\approx} \mathcal{N}(\psi^0, \mathcal{I}_n(\psi^0))$.

Remarquons que la condition 3 revient à l'existence d'une matrice de corrélation asymptotique pour l'EMV, et que les conditions 2 et 1 imposent respectivement que l'information croisse, et que le rythme de cette croissance soit suffisamment rapide —en évitant les situations pathologiques avec accumulation d'observations de moins en moins informatives à l'infini. Ce résultat permet par exemple à Mardia et Marshall d'énoncer des conditions très simples pour le cas particulier d'un treillis régulier ([MM84], section 4). Ces même auteurs proposent ensuite une étude expérimentale par simulation numérique (section 5) ; elle illustre le fait que l'approximation asymptotique est déjà plutôt satisfaisante pour des réalisations de processus gaussiens bi-dimensionnels isotropes (4 exemples avec différents noyaux) observés sur une grille 10×10 ($n = 100$), et que la variance asymptotique constitue même souvent une approximation grossière mais acceptable (de la variance observée) lorsque le plan d'expériences est une grille 6×6 ($n = 36$).

Partant du fait que l'inverse de la matrice d'information de Fisher est communément utilisée comme approximation de la matrice de covariance de l'EMV, l'article de Abt et Welch [AW98] propose quelques développements théoriques et des études par simulation numérique, visant à évaluer la pertinence et la qualité de cette approximation dans le cas où le domaine d'observation est compact et où la taille de l'échantillon augmente par densification du nombre d'observations au sein de ce domaine (*infill asymptotics*). Ils calculent les matrices d'information de Fisher et leurs inverses pour des vecteurs aléatoires

issus de processus gaussiens avec différents noyaux de covariance (calculs explicites de l'inverse de la matrice d'information pour les noyaux de covariance triangulaire et exponentiel ; approximations dans le cas gaussien), et montrent que —pour que les paramètres soient identifiables— l'EMV est consistant sous des asymptotiques *infill*.

Si les résultats asymptotiques sont indiscutables, les exemples produits par les auteurs pour démontrer l'applicabilité de l'approximation asymptotique dans le cas d'un petit nombre d'observations sont très partiels ; nous reviendrons sur ce point important dans la section suivante. Dans le même registre, Abt développe dans un article de 1999 [Abt99] une approximation de la variance de Krigeage prenant en compte l'estimation des paramètres de covariance. Cette dernière permet de pallier au moins partiellement la sous-estimation notoire de la variance de Krigeage "classique". Même si les résultats obtenus permettent effectivement d'affiner la connaissance de l'erreur quadratique moyenne de prédiction, ils reposent sur l'approximation asymptotique gaussienne maintenant bien connue. Or, comme le souligne l'auteur avec honnêteté dans la conclusion de l'article, utiliser l'approximation gaussienne non-biaisée avec variance asymptotique dans les cas pratiques peut fort bien être grossier, voire abusif. Nous présentons dans la section suivante une analyse de résultats obtenus par simulation au sujet de la loi effective de l'EMV avec très peu d'observations et leur comparaison avec l'approximation asymptotique.

5.2.3 Une étude expérimentale de l'EMV pour de petits échantillons

Portons maintenant de nouveau notre attention sur une fonction numérique y que l'on souhaite approximer par Krigeage, sur la base d'observations en un plan d'expériences $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$. On a vu au chapitre 3 que dans les cas où d est supérieur à 2 ou 3 —i.e. lorsque l'on sort du champ de la géostatistique traditionnelle pour appliquer le Krigeage à des fonctions numériques multivariables de type boîte noire—, on estime dans de nombreux travaux les paramètres de covariance en utilisant des méthodes automatiques, et plus spécifiquement via l'EMV.

Lorsque de plus y est coûteuse à évaluer, le nombre d'observations n est drastiquement limité. Les propriétés de l'EMV dans ces conditions extrêmes constitue un sujet d'importance : une mauvaise estimation des paramètres de covariance peut en effet conduire à un Krigeage de mauvaise qualité. Nous examinons ici la variabilité de l'EMV lorsqu'il n'y a pas d'erreur de modèle, c'est-à-dire lorsque les observations sont effectivement issues d'une réalisation de processus gaussien de fonction de covariance connue.

L'étude qui suit a été réalisée en 2006 (avant d'avoir pris connaissance des travaux [Ste99, AW98, Abt99, LS05]) et a constitué une partie d'une présentation donnée en

conférence [GDB+07], aujourd'hui publiée comme article de revue ([GDB+09], joint en annexe). On y compare la distribution de l'EMV obtenue par simulation numérique à la distribution normale donnée par les résultats asymptotiques précédemment exposés, et ce en fonction de deux tailles de plans d'expériences $1D$ et de différentes fonctions de covariances (gaussienne et exponentielle, avec différentes valeurs de paramètres de variance et de portée). Les résultats asymptotiques reposent sur l'information théorique de Fisher, dont on donne ci-dessous le calcul dans le cas du Krigeage Simple.

Calcul de l'information de Fisher théorique.

$$\boxed{\forall i,j \in [0,p]\ (\mathcal{I}(\psi))_{ij} = \frac{1}{2} tr\left(K_\psi^{-1}\frac{\partial K_\psi}{\partial \psi_i}K_\psi^{-1}\frac{\partial K_\psi}{\partial \psi_j}\right)} \tag{5.40}$$

Démonstration. Partant à nouveau de $-2\times\mathcal{L}(\psi) = n\log(2\pi)+\log(\det[K_\psi])+\mathbf{Y}^T K_\psi^{-1}\mathbf{Y}$, on a que $-2\frac{\mathcal{L}}{\partial\psi_j}(\psi) = \mathrm{tr}\left(\partial K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}\right) - \mathbf{Y}^T K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}K_\psi^{-1}\mathbf{Y}$. Une seconde dérivation par rapport à ψ_i permet alors de développer le terme général de la matrice d'information :

$$\begin{aligned}
2(\mathcal{I}(\psi))_{ij} &= -2\mathbb{E}\left[\frac{\partial^2\mathcal{L}}{\partial\psi_i\partial\psi_j}(\psi)\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial\psi_i}\left(\mathrm{tr}\left(K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}\right) - \mathbb{Y}^T K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}K_\psi^{-1}\mathbb{Y}\right)\right] \\
&= -\mathrm{tr}\left(K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_i}K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}\right) + \mathrm{tr}\left(K_\psi^{-1}\frac{\partial^2 K_\psi}{\partial\psi_i\psi_j}\right) \\
&\quad + \underbrace{2\,\mathbb{E}\left[\mathbb{Y}^T K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_i}K_\psi^{-1}\mathbb{Y}\right]}_{A} - \underbrace{\mathbb{E}\left[\left(\mathbb{Y}^T K_\psi^{-1}\frac{\partial^2 K_\psi}{\partial\psi_i\psi_j}K_\psi^{-1}\mathbb{Y}\right)\right]}_{B}
\end{aligned}$$

Le calcul de A et B peut être clarifié à l'aide du petit lemme suivant : si $\mathbb{Y}\sim\mathcal{N}(0,K)$ et S est une matrice réelle symétrique semi-définie positive, alors $\mathbb{E}[\mathbb{Y}^T S\mathbb{Y}] = \mathrm{tr}(KS)$. On peut démontrer ce lemme aisément en considérant la décomposition $S = PDP^T$, avec P orthogonale et $D = diag(\{\lambda_i, i\in[1,n]\})$. En effet, dans le cas où $K = I$, on a

$$\begin{aligned}
\mathbb{E}[\mathbb{Y}^T S\mathbb{Y}] &= \mathbb{E}[\mathbb{Y}^T PDP^T\mathbb{Y}] \\
&= \mathbb{E}[\mathbb{W}^T D\mathbb{W}] \text{ (avec } \mathbb{W} := P^T\mathbb{Y}\text{, d'où } \mathbb{W}\sim\mathcal{N}(0,I)) \\
&= \sum_{i=1}^n \lambda_i\mathbb{E}[\mathbb{W}_i^2] = \sum_{i=1}^n\lambda_i = \mathrm{tr}(S) = \mathrm{tr}(KS)
\end{aligned}$$

On peut généraliser au cas où $K\neq I$ en utilisant la décomposition de Mahalanobis $\mathbb{Y} = K^{\frac{1}{2}}\mathbb{N}$ avec $\mathbb{N}\sim\mathcal{N}(0,I)$ — ou encore celle de Cholesky avec $\mathbb{Y} = L\mathbb{N}$, L triangulaire

inférieure telle que $LL^T = K$, et il vient que

$$\mathbb{E}[\mathbb{Y}^T S \mathbb{Y}] = \mathbb{E}[\mathbb{N}^T K^{\frac{1}{2}} S K^{\frac{1}{2}} \mathbb{N}]$$

On obtient en se ramenant au cas précédent que

$$\mathbb{E}[\mathbb{Y}^T S \mathbb{Y}] = \mathrm{tr}(K^{\frac{1}{2}} S K^{\frac{1}{2}}) = \mathrm{tr}(KS),$$

où l'on a utilisé l'invariance de la trace d'un produit de matrices carrées par commutation de ces dernières. Pour revenir à la propriété principale, on peut maintenant effectuer le calcul de A et B en utilisant le lemme, et on trouve de manière immédiate que $A = \mathrm{tr}\left(K_\psi^{-1} \frac{\partial K_\psi}{\partial \psi_i} K_\psi^{-1} \frac{\partial K_\psi}{\partial \psi_j}\right)$ et $B = \mathrm{tr}\left(K_\psi^{-1} \frac{\partial^2 K_\psi}{\partial \psi_i \psi_j}\right)$, ce qui permet de conclure. □

Expériences réalisées

La deuxième section de l'article ([GDB+09], Cf. annexe) est dédiée à une série d'expériences sur la variabilité de l'EMV lorsque les données observées sont les valeurs prises par un processus gaussien monodimensionnel sur des plans d'expériences de taille 5 à 10. Les processus gaussiens étudiés sont stationnaires, de moyenne nulle, et de noyaux de covariance $k_g(h) = \sigma^2 e^{-\frac{h^2}{2p^2}}$ (*gaussien*) ou $k_e(h) = \sigma^2 e^{-\frac{|h|}{2p}}$ (*exponentiel*). Les paramètres de covariance se résument ainsi à un vecteur de $]0, +\infty[^2$, $\psi = (\sigma^2, p) \in \mathbb{R}^+ \times \mathbb{R}^+$. Les plans d'expériences sont des subdivisions régulières de $D := [-1, 1]$, et sont notés $\mathbf{X}_{n+1} := \{-1, -1 + \frac{2}{n}, -1 + \frac{4}{n}, ..., -1 + \frac{2(n-1)}{n}, 1\}$ ($n \in \mathbb{N}^*$). Nous nous sommes limités à des expériences avec les configurations suivantes : plans d'expériences $\mathbf{X_5}$ et $\mathbf{X_{10}}$, avec chacun des deux noyaux de covariance k_g et k_e, et leurs paramètres $\psi_1 = \sigma^2 \in \{5, 10\}$, et $\psi_2 \in \{0.3, 0.4, 0.5, 0.6\}$.

Pour chaque configuration, on a simulé 1000 réalisations de processus gaussiens au plan d'expériences correspondant, et calculé autant de réalisations de l'estimateur du maximum de vraisemblance de (ψ_1, ψ_2), permettant en particulier de produire une estimation de la moyenne et de la variance de l'estimateur. On a par ailleurs calculé pour chaque réalisation y l'erreur quadratique intégrée $\tilde{e}$ (ou « ISE », pour *Integrated Squared Error*) occasionnée en approximant y par la moyenne de Krigeage m_{OK} :

$$\tilde{e} := \frac{1}{\mathrm{vol}(D)} \int_D (y(\mathbf{x}) - m_{KO}(\mathbf{x}))^2 \, d\mathbf{x}, \tag{5.41}$$

où $\mathrm{vol}(D)$ est la mesure de Lebesgue de l'ensemble D. ISE a ici été approximée en prenant la moyenne des carrés des erreurs sur une grille de 200 points. On a finalement conservé pour chaque configuration expérimentale les moyennes et matrices de covariance

empiriques des valeurs relatives $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$, la moyenne et la variance de l'ISE $\tilde{e}$ (qui peut être considérée comme aléatoire puisque variant avec chaque réalisation y du processus sous-jacent Y), et des covariances empiriques entre $\tilde{e}$ et les ψ_i^{rel}.

Afin d'obtenir des résultats comparables pour différentes valeurs de ψ, nous avons utilisé une inverse de matrice d'information de Fisher relative :

$$(\mathcal{I}_{rel}(\boldsymbol{\psi}))_{ij} = (\mathcal{I}^{-1}(\boldsymbol{\psi}))_{ij}/(\psi_i \psi_j) \tag{5.42}$$

Remarquons que $\mathcal{I}_{rel}$ est en fait la matrice de covariance asymptotique de $\frac{\widehat{\boldsymbol{\psi}}}{\boldsymbol{\psi}}$, où la division est faite composante par composante.

Résumé des résultats obtenus

TAB. 5.1 – Valeurs de l'ISE et valeurs relatives des estimations par maximum de vraisemblance des paramètres de covariance, $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$, $i = 1, 2$, pour 1000 processus gaussiens de covariance gaussienne, échantillonés en $\mathbf{X} = \mathbf{X_5}$. La deuxième colonne illustre le fait que les estimateurs MV sont ici pratiquemment non-biaisés, même avec un plan à 5 points. A l'inverse, une comparaison entre la troisième et la quatrième colonne fait ressortir le fait que les ψ_i^{rel} sont clairement plus dispersés que ce que l'on obtiendrait en prenant la variance asymptotique basée sur la matrice d'information de Fisher.

(cov, ψ)	$E[\frac{\psi - \widetilde{\psi^*}}{\psi}]$	$Var\left[\frac{\psi - \widetilde{\psi^*}}{\psi}\right]$	FIM relative	$E[\widetilde{e}]$	$Var[\widetilde{e}]$	$Cov\left[\frac{\psi - \widetilde{\psi^*}}{\psi}, \widetilde{e}\right]$
gau, 5, 0.3	$\begin{pmatrix} -0.034 \\ 0.018 \end{pmatrix}$	$\begin{pmatrix} 1.105 & 0.277 \\ 0.277 & 1.270 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	4.976	16.172	$\begin{pmatrix} -0.193 \\ 0.439 \end{pmatrix}$
gau, 5, 0.4	$\begin{pmatrix} -0.147 \\ 0.042 \end{pmatrix}$	$\begin{pmatrix} 1.329 & 0.501 \\ 0.501 & 0.976 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	3.287	13.294	$\begin{pmatrix} -0.076 \\ 0.503 \end{pmatrix}$
gau, 5, 0.5	$\begin{pmatrix} -0.222 \\ 0.033 \end{pmatrix}$	$\begin{pmatrix} 4.037 & 0.757 \\ 0.757 & 0.679 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	1.947	8.149	$\begin{pmatrix} 0.028 \\ 0.537 \end{pmatrix}$
gau, 5, 0.6	$\begin{pmatrix} -0.187 \\ 0.027 \end{pmatrix}$	$\begin{pmatrix} 2.058 & 0.504 \\ 0.504 & 0.421 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	0.706	2.072	$\begin{pmatrix} 0.107 \\ 0.549 \end{pmatrix}$
gau, 10, 0.3	$\begin{pmatrix} -0.131 \\ 0.006 \end{pmatrix}$	$\begin{pmatrix} 3.334 & 0.867 \\ 0.867 & 1.564 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	10.138	61.959	$\begin{pmatrix} -0.097 \\ 0.401 \end{pmatrix}$
gau, 10, 0.4	$\begin{pmatrix} -0.083 \\ 0.106 \end{pmatrix}$	$\begin{pmatrix} 1.645 & 0.484 \\ 0.484 & 0.862 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	6.398	47.838	$\begin{pmatrix} -0.068 \\ 0.511 \end{pmatrix}$
gau, 10, 0.5	$\begin{pmatrix} -0.166 \\ 0.024 \end{pmatrix}$	$\begin{pmatrix} 1.343 & 0.440 \\ 0.440 & 0.629 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	3.678	32.761	$\begin{pmatrix} 0.025 \\ 0.557 \end{pmatrix}$
gau, 10, 0.6	$\begin{pmatrix} -0.256 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} 14.960 & 0.963 \\ 0.963 & 0.392 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	1.459	9.357	$\begin{pmatrix} 0.047 \\ 0.535 \end{pmatrix}$

On observe dans le cas d'un noyau gaussien un biais relatif négatif (de -3.4% to -25.6%, i.e. une surestimation du paramètre) dans l'estimation de $\psi_1 = \sigma^2$. Ce biais est

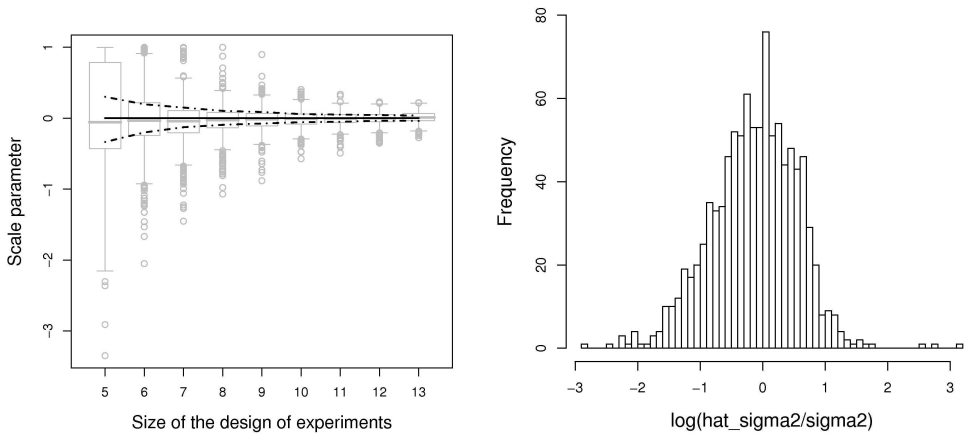


FIG. 5.3 – A gauche : Comparaisons entre la distribution empirique (boxplots gris) et la loi asymptotique (en noir) pour l'estimateur du paramètre de portée ψ_2 (en échelle relative) lorsque la taille du plan d'expériences $\mathbf{X}$ augmente. Les boxplots pour les distributions empiriques ont été obtenus sur la base de 1000 simulations de GP, avec un noyau de covariance gaussien de paramètres $\boldsymbol{\psi} = (5, 0.5)$. Concernant la distribution asymptotique de l'EMV, la médiane est représentée sous la forme d'une ligne continue et les quantiles à 25% et 75% en lignes brisées. A droite : Histogramme du logarithme des erreurs relatives obtenues en estimant ψ_1 par EMV sur la base de 1000 réalisations de PG. La forme de cet histogramme suggère que la distribution des erreurs relatives est beaucoup plus proche d'une loi lognormale que d'une gaussienne.

décroissant avec le nombre de points du plan, $\#\mathbf{X}$ (Cf. 5.2 où il varie entre -0.06% and -11.9%) , ce qui semble en accord avec le non-biais asymptotique de l'EMV. Par ailleurs, le biais relatif de $\hat{\psi}_2$ est déjà faible lorsque $\#\mathbf{X} = 5$, et devient négligeable pour $\#\mathbf{X} = 10$.

Les matrices de covariance empiriques de l'estimateur du maximum de vraisemblance offrent quelques résultats surprenants. En particulier, les variances relatives des $\widehat{\psi}_1$ présentent d'importantes fluctuations : elles varient parfois d'un ordre de grandeur de plus de 10 entre deux échantillons de 1000 réalisations du même processus gaussien ; on a par exemple obtenu, en faisant un nouvel ensemble de simulations avec un PG de paramètres $\psi = (10, 0.4)$ et un plan $\#\mathbf{X} = 5$, la matrice de covariance $Var\left[\left(\psi_i^{rel}\right)_i\right] = \begin{pmatrix} 43.555 & 3.242 \\ 3.242 & 0.971 \end{pmatrix}$. Comme ce résultat est en contradiction avec la normalité et l'ordre de grandeur de dispersion donné par l'équation 5.40, il semble pertinent d'examiner le phénomène plus en détail.

TAB. 5.2 – Mesures des erreurs relatives de l'EMV et valeurs moyennes de l'ISE associées à 1000 réalisations de PG de noyau de covariance gaussien, pour $\mathbf{X} = \mathbf{X_{10}}$. L'approximation basée sur l'inverse de la matrice d'information de Fisher sous-estime toujours la variance des estimateurs du maximum de vraisemblance mais est tout de même moins imprecise que sur le tableau 5.2.3.

(cov, ψ)	$E\left[\left(\frac{\psi_i-\widehat{\psi}_i^*}{\psi_i}\right)_i\right]$	$Var\left[\left(\frac{\psi_i-\widehat{\psi}_i^*}{\psi_i}\right)_i\right]$	relative FIM	$E[\widetilde{e}]$	$Var[\widetilde{e}]$	$Cov\left[\frac{\psi_i-\widehat{\psi}_i^*}{\psi_i},\widetilde{e}\right]$
gau, 5, 0.3	$\begin{pmatrix}-0.054\\0.012\end{pmatrix}$	$\begin{pmatrix}0.432 & 0.105\\0.105 & 0.085\end{pmatrix}$	$\begin{pmatrix}0.297 & 0.057\\0.057 & 0.033\end{pmatrix}$	0.177	0.356	$\begin{pmatrix}0.16\\0.625\end{pmatrix}$
gau, 5, 0.4	$\begin{pmatrix}-0.042\\-0.019\end{pmatrix}$	$\begin{pmatrix}0.424 & 0.058\\0.058 & 0.024\end{pmatrix}$	$\begin{pmatrix}0.340 & 0.044\\0.044 & 0.014\end{pmatrix}$	0.009	0.007	$\begin{pmatrix}0.055\\0.300\end{pmatrix}$
gau, 5, 0.5	$\begin{pmatrix}-0.067\\-0.013\end{pmatrix}$	$\begin{pmatrix}0.46 & 0.051\\0.051 & 0.013\end{pmatrix}$	$\begin{pmatrix}0.362 & 0.036\\0.036 & 0.008\end{pmatrix}$	0.0004	3.e-07	$\begin{pmatrix}0.08\\0.211\end{pmatrix}$
gau, 5, 0.6	$\begin{pmatrix}-0.075\\-0.007\end{pmatrix}$	$\begin{pmatrix}0.728 & 0.059\\0.059 & 0.012\end{pmatrix}$	$\begin{pmatrix}0.375 & 0.032\\0.032 & 0.005\end{pmatrix}$	4.e-05	5.e-09	$\begin{pmatrix}0.076\\0.263\end{pmatrix}$
gau, 10, 0.3	$\begin{pmatrix}-0.067\\0.003\end{pmatrix}$	$\begin{pmatrix}0.432 & 0.089\\0.089 & 0.079\end{pmatrix}$	$\begin{pmatrix}0.297 & 0.057\\0.057 & 0.033\end{pmatrix}$	0.345	1.701	$\begin{pmatrix}0.098\\0.562\end{pmatrix}$
gau, 10, 0.4	$\begin{pmatrix}-0.097\\0.015\end{pmatrix}$	$\begin{pmatrix}0.495 & 0.071\\0.071 & 0.028\end{pmatrix}$	$\begin{pmatrix}0.340 & 0.044\\0.044 & 0.014\end{pmatrix}$	0.03	0.097	$\begin{pmatrix}0.075\\0.409\end{pmatrix}$
gau, 10, 0.5	$\begin{pmatrix}-0.06\\-0.009\end{pmatrix}$	$\begin{pmatrix}0.491 & 0.046\\0.046 & 0.011\end{pmatrix}$	$\begin{pmatrix}0.362 & 0.036\\0.036 & 0.008\end{pmatrix}$	0.0008	1.e-06	$\begin{pmatrix}-0.086\\0.15\end{pmatrix}$
gau, 10, 0.6	$\begin{pmatrix}-0.119\\-0.009\end{pmatrix}$	$\begin{pmatrix}0.582 & 0.05\\0.05 & 0.011\end{pmatrix}$	$\begin{pmatrix}0.375 & 0.032\\0.032 & 0.005\end{pmatrix}$	0.0001	3.e-08	$\begin{pmatrix}0.064\\0.314\end{pmatrix}$

Premièrement, on observe que les valeurs extrêmes de $Var[\widehat{\psi}_1]$ sont causées par quelques *outliers*, perturbant fortement l'estimateur (bien peu robuste) de la variance. Deuxièmement, l'histogramme de la figure 5.3 illustre bien le fait que la distribution de probabilité des $\widehat{\psi}_1$'s est plutôt log-normale que normale. Au final, la comparaison avec la matrice d'information de Fisher relative montre que la variance empirique de $\widehat{\psi}_1$ est clairement plus grande qu'indiqué par l'approximation de Fisher à l'ordre 2, en particulier avec les plus petits plans d'expériences étudiés.

Concernant les variances relatives de $\widehat{\psi}_2$, les résultats sont beaucoup plus ordinaires : elles décroissent de manière monotone avec ψ_2 et avec $\#\mathbf{X}$, à la fois en ce qui concerne les quantités théoriques et empiriques. Une fois de plus, les variances empiriques tendent à se rapprocher des variances asymptotiques lorsque $\#\mathbf{X}$ croît, même si les premières restent typiquement deux fois plus élevées que les secondes pour un plan à 10 éléments. Par ailleurs, les deux tableaux illustrent quelques propriétés fondamentales de l'erreur quadratique moyenne. Evidemment décroissante en $\#\mathbf{X}$, l'ISE moyenne $\mathbb{E}[\tilde{e}]$ est aussi

décroissante en la longueur de corrélation ψ_2 et linéairement croissante en la variance ψ_1. Nous avons finalement étudié quantitativement la dépendance linéaire entre la sous-estimation des deux paramètres de covariance par EMV et l'ISE. Il est remarquable que ψ_1 and ψ_2 jouent ici des rôles sensiblement différents : une mauvaise estimation de ψ_1 est en effet très faiblement corrélée avec l'ISE. Ce résultat apparaît en fait rassurant lorsque l'on considère que la moyenne de Krigeage Ordinaire ne dépend pas de la variance du processus (lorsqu'il n'y a pas d'effet de pépite, Cf. [Cre93] ou encore les chapitres 3 et 7 de cette thèse.) A l'inverse, la corrélation entre l'ISE et l'erreur relative d'estimation par MV de ψ_2 est significativement positive : elle varie entre 40.1% et 55.7% pour $\#\mathbf{X} = 5$ et entre 15% et 62.5% pour $\#\mathbf{X} = 10$. Cela coïncide avec nos observations qualitatives antérieures concernant une ISE particulièment élevée lorsque la portée est sous-estimée.

Une étude similaire avec un noyau exponentiel a donné des résultats notablement différents à la fois en ce qui concerne le biais et la variance des estimateurs par MV (les tableaux correspondant ne sont pas présentés ici). On a en effet observé des variances d'estimation beaucoup plus proches de l'approximation asymptotique que dans le cas du noyau gaussien, et à l'inverse un biais relatif beaucoup plus important. Cependant, le comportement de l'ISE et les corrélations entre l'ISE et les erreurs relatives dans l'estimation des paramètres de covariance sont similaires à ceux du cas du noyau gaussien.

Pour résumer notre première étude empirique sur la variabilité de l'EMV des paramètres de covariance du KS pour de petits échantillons :

- L'approximation asymptotique de Fisher doît être appliquée avec reserve, en particulier lorsque la taille des plans d'expériences est sévèrement limitée. Plus précisément, il a été observé dans les cas traités avec $n \leq 5$ que la distribution de l'estimateur du paramètre de portée est asymétrique avec une variance plus grande que l'inverse de l'information de Fisher, mais qu'elle se stabilise assez rapidement vers une distribution gaussienne lorsque n croît (de 5 à 13).

- Par ailleurs, la distribution de l'estimateur du maximum de vraisemblance du paramètre de variance du processus a une queue lourde et sa forme est loin d'être gaussienne lorsque n est très petit ($n \leq 5$). De plus, ce phénomène persiste lorsque n croît (ici de 5 à 13) et il semble que l'approximation gaussienne ne devient raisonnable qu'à partir de valeurs de n plus élevées.

5.3 Discussion sur l'estimation de paramètres du noyau de covariance d'un modèle de Krigeage

5.3.1 L'EMV est-il raisonnable lorsque l'on a très peu d'observations ?

Nous avons vu dans la section 5.2. que les propriétés de l'EMV rappelées auparavant (section 5.1.) n'étaient que partiellement exportables au cadre d'un petit nombre d'observations dépendantes. En effet, si la propriété de non-biais semble bien respectée (du moins pour le noyau gaussien), la loi des estimateurs du MV peut lorsque n est petit présenter une dispersion sensiblement plus importante que celle indiquée par l'approximation asymptotique, et prendre des formes mal maîtrisées (lois clairement non-gaussiennes). Ces constats sont d'autant moins rassurants que les cas d'étude sont en général multidimensionnels et anisotropes, la taille du plan drastiquement limitée, et que certaines plages de valeurs des paramètres de covariance sont pratiquemment non-identifiables (e.g. portées mal représentées par les interdistances du plan d'expériences). Accéder à une approximation correcte de la loi de l'EMV de manière analytique semble ainsi hors d'atteinte, alors que c'est justement dans ces conditions extrêmes qu'il est crucial de prendre en compte l'incertitude sur ψ dans les prédictions par Krigeage.

A supposer que l'on passe outre ces considérations sur la difficulté de bien connaître la loi de l'EMV, un problème de taille subsiste toutefois pour le modélisateur : il doit se contenter d'une seule réalisation du processus gassien étudié, et par là-même d'une unique réalisation de l'EMV. Il ne lui importe pas vraiment par exemple que l'estimateur soit non-biaisé s'il sait qu'il peut obtenir avec une forte probabilité des valeurs de ψ faisant totalement échouer le Krigeage (typiquement avec une portée estimée en bord de domaine). Les questions abordées au fil de cette section visent d'une part à discuter de perspectives au sujet de l'utilisation pratique de l'EMV, et d'autre part à évoquer des directions actuelles proposées pour obtenir des estimateurs alternatifs, construits à partir de l'EMV, et plus adaptés à la pratique du Krigeage avec peu d'observations.

La simulation comme garde-fou

A plan $\mathbf{X}$ et paramètres de covariance ψ^0 connus, la simulation permet de calculer (sous l'hypothèse gaussienne) la loi μ_{EMV,ψ^0} de l'EMV des paramètres du modèle, Cf. Alg. 3. Cela peut être utile pour l'aide à la décision, par exemple pour raffiner les plans d'expériences de manière à réduire l'incertitude dans l'estimation de certains paramètres, ou encore pour renoncer à un modèle anisotrope en faveur d'un modèle isotrope pour cause de plan jugé non suffisamment informatif.

Algorithm 3 Approximation de μ_{EMV,ψ^0} **en dimension quelconque**, à ψ^0 connu.

1: $K(\psi^0) = (k_{\psi^0}(\mathbf{x}^i, \mathbf{x}^j))_{i,j}$ ▷ Calculer $K(\psi^0)$
2: Pour $i = 1, \ldots, N_{sim}$, faire
3: Tirer une réalisation $\mathbf{Y}$ de $\mathbb{Y} \sim \mathcal{N}(0, K(\psi^0))$ ▷ Simuler un V.a.r. $\mathbb{Y}$ de loi $\mathcal{N}(0, K(\psi^0))$
4: $\psi^{*i} = \text{argmax} L(\psi; \mathbf{Y})$ ▷ Maximiser la vraisemblance de ψ sachant que $\mathbb{Y} = \mathbf{Y}$
5: Fin de boucle
6: $\widehat{\mu_{EMV,\psi^0}} := \sum_{i=1}^{N_{sim}} \frac{1}{N_{sim}} \delta_{\psi^{*i}}$ ▷ Loi empirique de l'EMV $\psi^*(\mathbf{Y})$ lorsque $\psi = \psi^0$

La donnée de la loi $\widehat{\mu_{EMV,\psi^0}}$ permet ainsi d'approcher la loi μ_{EMV,ψ^0} de l'EMV , et en particulier son espérance et sa matrice de covariance, avec une précision aussi grande que voulue en fonction de N_{sim}. Si le calcul de $\widehat{\mu_{EMV,\psi^0}}$ est rendu tout à fait abordable par l'hypothèse que $\mathbb{Y}$ est gaussien, il subsiste un problème non-négligeable : le vrai ψ^0, nécessaire pour la construction de K, n'est évidemment pas connu *a priori*. La solution souvent retenue en pratique consiste à estimer ψ^0 par maximum de vraisemblance sur la base d'observations $\mathbf{Y}$, d'appliquer Alg. 3 en substituant $\psi^*(\mathbf{Y})$ à ψ^0 dans le calcul de K, et en approchant finalement μ_{EMV,ψ^0} par $\widehat{\mu_{EMV,\psi^*(\mathbf{Y})}}$. La qualité de l'approximation obtenue est alors crucialement dépendante de celle de l'estimation de ψ^0 par $\psi^*(\mathbf{Y})$ [9]. En anticipant un peu le passage à venir sur les méthodes bayésiennes, une des solutions possibles pour réduire l'incertitude sur $\psi^*(\mathbf{Y})$ dans l'approximation de la loi de l'EMV par simulation avec *plug-in* est d'approcher μ_{EMV,ψ^0} par $\int_\Psi \widehat{\mu_{EMV,\psi}} \pi(\psi|\mathbb{Y} = \mathbf{Y}) d\psi$.

L'idée de Stein : recycler l'information de Fisher pour évaluer des plans

Utiliser la simulation pour améliorer des plans est possible, mais n'est pas du tout commode en pratique pour mener une optimisation numérique avec les outils classiques. L'approximation de Fisher, sans se contraindre à y croire, peut constituer une fonction objectif de substitution durant une phase de conception du plan d'expériences. L'idée proposée par Stein [Ste99] consiste à se servir de l'approximation asymptotique de la variance de l'EMV comme critère pour construire ou raffiner des plans d'expériences : sans avoir besoin de croire que l'approximation est fidèle —on a vu dans la section prédédente que l'ordre de grandeur pouvait être peu réaliste—, on peut raisonnablement utiliser l'information de Fisher comme « proxy » de la vraie variance d'estimation. Un usage possible est alors d'utiliser la l'approximation analytique de la variance d'estima-

[9] Mentionnons sans plus de détails que des questions similaires, sinon identiques, apparaîssent dans l'étude du *bootstrap* [ET93]

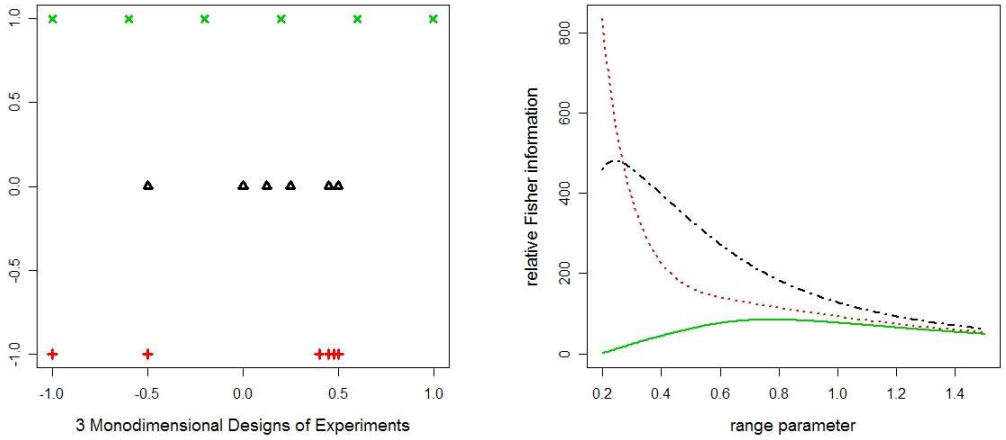


FIG. 5.4 – A gauche : trois plans d'expériences à six éléments. A droite : courbes d'information de Fisher relatives associées au paramètre de portée d'un PG de covariance gaussienne (de variance 5) observé respectivement en chacun des plans du graphe de gauche. La courbe pleine est associée aux plan reprśenté avec des croix obliques, la courbe en pointillés avec les croix droites, et la courbe mixte avec les triangles.

tion au sein de problèmes de planification expérimentale, par exemple pour décider quel plan est le meilleur pour réduire l'incertitude sur ψ parmi plusieurs plans candidats : Cf. 5.4, où l'information de Fisher relative apparaît comme un critère potentiel de sélection de plans dans le but d'estimer un paramètre de portée (et ce en fonction de la valeur attendue de ce paramètre).

5.3.2 Quelques alternatives possibles à l'EMV

Méthodes bayésiennes

Dans le même ordre d'idée que la construction bayésienne du Krigeage Ordinaire (et Universel) présentée à la fin du chapitre 3, il est devenu assez courant [KO96, MS04b, O'H06, RW06] de remplacer l'approche par *plug-in* pour les paramètres de covariance par une intégration par rapport à une loi *a posteriori*, construite comme toujours en bayésien sur la base d'une loi *a priori* π et des observations $(\mathbf{X}, \mathbf{Y})$ disponibles :

$$\pi(\psi|Y(\mathbf{X}) = \mathbf{Y}) \propto L(\psi; Y(\mathbf{X}) = \mathbf{Y})\pi(\psi).$$

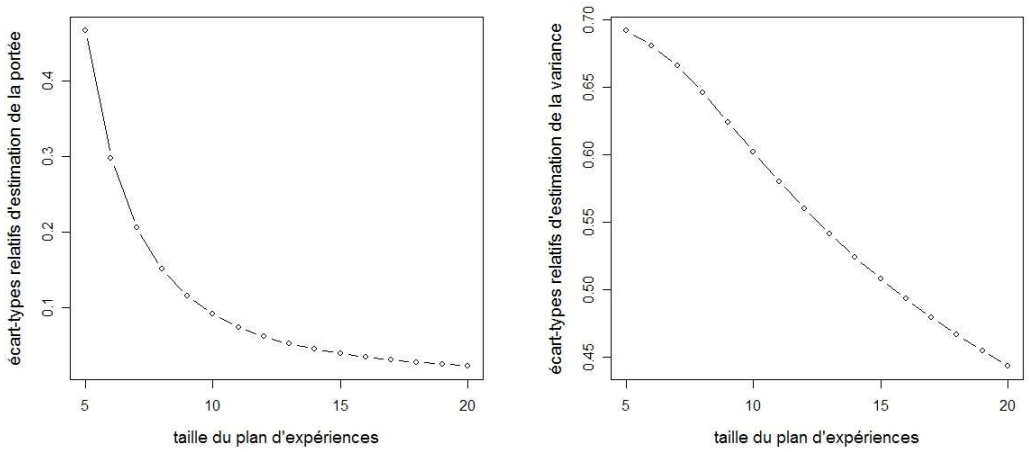


FIG. 5.5 – Ecart-types asymptotiques (basés sur l'information de Fisher) de l'EMV des paramètres de portée et de variance d'un PG de covariance gaussienne (de variance 5) observé en des subdivisions régulières de $[-1, 1]$ de tailles 5 à 20. Le graphe de gauche représente l'écart-type approché du paramètre de portée en fonction de la taille du plan. Le graphe de droite est l'analogue pour le paramètre de variance.

Conformément à l'habitude, le choix de la loi *a priori* est un sujet délicat : choisies uniformes, non-informatives (e.g. Jeffrey, Cf. [Rob92]), ou informatives (injection d'« informations métier »), les lois *a priori* conditionnent d'autant plus sensiblement les résultats du Krigeage Bayésien que les observations sont peu nombreuses. Nous ne rentrerons ici faute de temps pas plus en détail concernant ces techniques, mais signalons que les mélanges discrets de noyaux développés au chapitre 8 présentent bon nombre de similarités avec ces dernières. Voyons maintenant une autre manière d'injecter de l'information sans faire directement appel au formalisme bayésien : la pénalisation.

EMV pénalisé

Une des méthodes employées pour réduire la variance de l'EMV, appelée *pénalisation*[10], consiste à ajouter un terme à la fonction objectif (ici la log-vraisemblance) de manière à rendre le problème d'optimisation plus robuste aux fluctuations d'échantillonnage. En pratique, cela revient à remplacer la maximisation de $\mathcal{L}(\psi; \mathbf{Y})$ par celle de

[10] L'idée de pénalisation n'est pas du tout propre à l'EMV et se retrouve dans différentes branches de l'optimisation. Elle est à rapprocher avec la méthode des *multiplicateurs de Lagrange*.

$$\mathcal{L}^{pen}(\psi;\mathbf{Y}) = \mathcal{L}(\psi;\mathbf{Y}) - P(\psi), \tag{5.43}$$

où $P(\psi)$ est une fonction de ψ appelée *fonction de pénalité*, et prise dans l'article [LS05] auquel nous nous référons principalement dans cette section sous la forme additive

$$P_n(\psi) = n\sum_{i=1}^{d} p_\lambda(\psi_i), \tag{5.44}$$

où p_λ est une fonction à valeurs positives et λ un coefficient de régularisation.

En adoptant des notations similaires aux sections précédentes et en supposant que la fonction de pénalisation est suffisamment régulière, un développement limité à l'ordre 2 nous donne — en reprenant les notations de la section 5.2. :

$$0 = \nabla_\psi \mathcal{L}_n^{pen}(\psi_n^*) = \nabla_\psi \mathcal{L}_n^{pen}(\psi^0) + \nabla_\psi^2 \mathcal{L}_n^{pen}(\psi_n^1)(\psi_n^* - \psi^0), \tag{5.45}$$

ce que l'on peut aussi écrire, sous l'hypothèse de consistance et avec n suffisamment grand de manière à ce que $\nabla_\psi^2 \mathcal{L}_n^{pen}(\psi^1)$ soit inversible :

$$\begin{aligned}(\psi_n^* - \psi^0) &= -[\nabla_\psi^2 \mathcal{L}_n^{pen}(\psi_n^1)]^{-1} \nabla_\psi \mathcal{L}_n^{pen}(\psi^0) \\ &= -[\nabla_\psi^2 \mathcal{L}_n^{pen}(\psi_n^1)]^{-1} \nabla_\psi \mathcal{L}_n(\psi^0) - [\nabla_\psi^2 \mathcal{L}_n^{pen}(\psi_n^1)]^{-1} \nabla_\psi P(\psi^0).\end{aligned} \tag{5.46}$$

Or $\nabla_\psi^2 \mathcal{L}_n^{pen} = \nabla_\psi^2 \mathcal{L}_n - \nabla_\psi^2 P_n$ et on a, sous les conditions sur l'information empirique évoquées dans la section précédente, les convergences

$$\begin{aligned}&-(\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0))[\nabla_\psi^2 \mathcal{L}_n^{pen}(\psi^0)]^{-1} \xrightarrow{\mathbb{P}} I \\ &\text{i.e. } -[\nabla_\psi^2 \mathcal{L}_n^{pen}(\psi^0)]^{-1} \approx [\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]^{-1}\end{aligned} \tag{5.47}$$

et

$$\begin{aligned}&\mathcal{I}_n(\psi^0)^{-1}\left(\nabla_\psi \mathcal{L}_n^{pen}(\psi^0) + \nabla P_n(\psi^0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I) \\ &\text{i.e. } \nabla_\psi \mathcal{L}_n^{pen}(\psi^0) \approx \mathcal{N}(-\nabla P_n(\psi^0), \mathcal{I}_n(\psi^0)).\end{aligned} \tag{5.48}$$

A l'instar du cas non pénalisé, on obtient alors la normalité asymptotique de l'EMV pénalisé en utilisant les convergences 5.47 et 5.48 dans l'équation 5.46 :

$$[\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]\left(\psi_n^* - \psi^0 + [\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]^{-1} \nabla P_n(\psi^0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I), \tag{5.49}$$

ce que l'on peut aussi informellement écrire

$$\psi_n^* \approx \mathcal{N}(M_n^{pen}, V_n^{pen}) \tag{5.50}$$

$$\text{où } \begin{cases} M_n^{pen} = \psi^0 - [\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]^{-1} \nabla P_n(\psi^0) \\ V_n^{pen} = [\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]^{-1} \mathcal{I}_n(\psi^0) [\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]^{-1} \end{cases} \tag{5.51}$$

De manière générale, le prix à payer pour la modification de la variance de l'EMV par pénalisation est l'introduction d'un biais indésirable. On voit dans l'éq. 5.51 que l'emploi d'une pénalisation a eu pour effet de modifier le comportement moyen de l'EMV par ajout du terme $-[\mathcal{I}_n(\psi^0) + \nabla^2 P_n(\psi^0)]^{-1} \nabla P_n(\psi^0)$. Notons que ce terme de biais est d'autant plus « grand » que $\nabla P_n(\psi^0)$ l'est. Il y a ainsi un compromis à trouver lors du choix de la fonction de pénalité, l'idéal étant de s'arranger pour que la courbure soit importante et le gradient faible autour de ψ^0. Le problème est bien entendu que l'on ne connaît pas ψ^0 à l'avance.

Le choix de P est donc un problème en soi, tout comme l'est le choix du prior en inférence bayésienne. La pénalité SCAD (*smoothly clipped absolute deviation*) est proposée dans [LS05] comme alternative intéressante aux pénalités classiques dites L^1 et L^2. Les vertus pratiques de SCAD[11] y sont illustées sur deux exemples, et les auteurs mentionnent que cette pénalité satisfait des propriétés suffisantes pour retrouver dans le cas d'observations très nombreuses la consistance et la normalité asymptotique habituelles de l'EMV.

[11] Une procédure reposant sur la validation croisée reste nécessaire pour régler λ.

Chapitre 6

Autour du choix de la structure d'un métamodèle de Krigeage

6.1 Quelques limites liées à l'hypothèse de stationnarité

6.1.1 Deux questions d'instationnarité concernant le Krigeage

Nous avons vu au chapitre 3 que certaines hypothèses de stationnarité étaient faites au sujet du processus Y pour pouvoir appliquer les différents types classiques de Krigeage. Les Krigeages Simple et Ordinaire reposent par exemple tous deux sur l'hypothèse que le processus $(Y_{\mathbf{x}})_{\mathbf{x}\in D}$ est L^1 et d'espérance constante ($\forall \mathbf{x} \in D,\ \mathbb{E}[Y_{\mathbf{x}}] = \mu \in \mathbb{R}$), le KS exigeant de plus que Y soit L^2 de noyau de covariance $k : (\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow k(\mathbf{x}, \mathbf{x}')$ ne dépendant que de l'accroissement $\mathbf{h} := \mathbf{x} - \mathbf{x}'$ (*noyau stationnaire*), et le KO imposant comme condition plus souple que le variogramme existe, i.e. que $\forall \mathbf{x} \in D,\ \forall \mathbf{h} \in D - D,\ var[Y_{\mathbf{x}+\mathbf{h}} - Y_{\mathbf{x}}] < +\infty$ ne dépende que de $\mathbf{h}$ (*hypothèse intrinsèque*).

Ces hypothèses ne sont pour autant pas toujours tenables. Nous nous intéressons dans le présent chapitre ainsi que dans le suivant à différentes motivations pratiques pour ne pas se cantonner au cadre(s) stationnaire(s), et proposons une revue de travaux existants ([Mat69], *KU* avec résolution du problème de circularité par maximisation d'une fonction de vraisemblance concentrée) ainsi que des travaux originaux (Krigeage avec tendance additive, Krigeage avec symétries, Krigeage avec bruit de simulation contrôlé hétérogène) visant à étendre le domaine d'applicabilité du Krigeage à des situations ou les hypothèses classiques de stationnarité sont clairement mises en défaut.

6.1.2 Prise en compte de tendances déterministes inconnues

La mise en défaut de l'hypothèse de stationnarité à l'ordre 1 est sans doute le premier enrichissement non-stationnaire (ni intrinsèquemment stationnaire) notable du Krigeage, dû à Georges Matheron et André Journel (sous le nom de « Krigeage Universel » [Mat69], Cf. les équations données au chapitre 3), et visant à la prise en compte, au sein de l'interpolation par Krigeage, d'une tendance déterministe connue à un ensemble de coefficients linéaires près. Comme expliqué dans la suite de ce chapitre, l'obtention des équations de KU, leur rapport avec le KS et les équations de la régression linéaire [Bai05], ainsi que la question difficile de l'estimation des paramètres de covariance en présence d'une tendance sont aujourd'hui bien connus, et des solutions existent. En revanche, beaucoup moins de résultats sont semble-t-il connus concernant la marche à suivre pour estimer la forme des fonctions de base à utiliser dans le KU sur la base de données d'observations, en particulier lorsque la dimension $d \geq 2$ ne permet pas d'intuiter la tendance directement par visualisation. Cela nous a amené à proposer l'utilisation de modèles additifs non-linéaires comme tendance d'un modèle de Krigeage (KAT, Cf. section 3), allant de pair avec un procédé original pour l'estimation des paramètres de covariance.

6.1.3 Utilisation de noyaux de covariance non-stationnaires

Si l'utilisation de noyaux de covariances non-stationnaires ne pose pas de problème conceptuel majeur dans le cadre d'une modélisation par processus gaussiens [Pac03], il est plutôt rare d'en trouver dans les applications classiques du Krigeage —en dehors peut-être du cas de processus intrinsèquement stationnaires mais non-stationnaires tels que le mouvement Brownien, de noyau $k(\mathbf{x}, \mathbf{x}') = min(x, x')$ mais de variogramme $\gamma(h) = |h|$, ou encore les mouvements Browniens multi-fractionnaires [Bai05]—. Nous nous sommes particulièrement intéressés à l'existence et à la caractérisation de noyaux de covariance non-stationnaires dont les processus gaussiens (centrés) associés possèdent des trajectoires symétriques. Cela a donné lieu à l'initiation d'un travail de recherche, dont les premiers résultats sur les processus centrés de trajectoires invariantes sous l'action d'un groupe fini sont présentés dans le chapitre 7.

6.2 Kriging with trends : a bless or a curse ?

6.2.1 Preliminaries : context and notations (following [GDB+09])

We study a deterministic numerical simulator as a function $y : D \subset \mathbb{R}^d \to \mathbb{R}$, where $\mathbf{x} \in D$ is the vector of inputs variables. We denote the set of the design points (or "design") by $\mathbf{X} = \{\mathbf{x^1}, ..., \mathbf{x^n}\}$ and by $\mathbf{Y} = \{y(\mathbf{x^1}), ..., y(\mathbf{x^n})\}$ the set of simulator responses associated with $\mathbf{X}$. Kriging is a class of methods coming from the field of geostatistics [Mat63, Cre93], known as *linear optimal prediction* in classical statistics. It provides at each point $\mathbf{x} \in D$ a prediction $\hat{Y}(\mathbf{x})$ linearly depending on $\mathbf{Y}$, where the weights depend on the design and on the Kriging model but not on the observations. The way the weights are defined varies as a function of the type of Kriging —Simple (SK), Ordinary (OK), Universal (UK), etc— and many parameters such as the trend functions, the covariance kernel and their own parameters : threshold (or « sill »), scales, nugget, etc... denoted by the r-dimensional vector $\boldsymbol{\psi}$. In the following, we will concentrate on the parameters of sill and scale ($r = 2$), denoted respectively either by $\psi_1,\ \psi_2$ or by $\sigma^2,\ p \in [0, +\infty[$. Most classic Kriging types (including SK, OK, UK, and more) can be interpreted as random process interpolation relying on the assumption that :

$$\forall \mathbf{x} \in D,\ y(\mathbf{x}) = t(\mathbf{x}) + \varepsilon(\mathbf{x}) \tag{6.1}$$

where t is a numerical deterministic function and $\varepsilon(\mathbf{x})$ is one path of a centered stationary Gaussian Process (GP) with known stationary covariance kernel $k : h \in \mathbb{R}^d \longrightarrow k(h) \in \mathbb{R}$. t is generally known up to a set of parameters or a semi-parametric structure to be estimated within Kriging. Several founder works [SWMW89, JSW98] on the application of Kriging to computer simulations start off with an extremely simplified version of equation 6.1. They assume that the trend is an unknown constant (Ordinary Kriging, i.e. $t(x) = \mu \in \mathbb{R}$) and that k is a generalized exponential kernel [SWN03], letting the stochastic part of equation 6.1 account for the variability of y. Then the covariance parameters $\boldsymbol{\psi}$ are estimated by maximizing the Gaussian likelihood of the observations $\mathbf{Y}$. On the other hand, recent approaches [Jou02, MS05] try to take advantage of more complex trends, from linear and polynomial functions to Fourier series. In other respects, [MS04b] as well as [O'H06] present an application of Bayesian analysis to Kriging interpolation of computer codes.

The motivation of this chapter is to raise some basic questions that should become crucial when applying Kriging techniques with few observations regarding the dimension of inputs, which is quite often the case in numerical simulation. The currect section,

based on toy experiments, puts a focus on the choice of the trend t and some aspects of its relation with the estimation of the covariance parameters $\boldsymbol{\psi}$. The two following sections are dedicated to presenting an original combination of additive models and Simple Kriging, with a heuristic fitting methodology. The efficiency of this technique is illustrated on a 3-dimensional example from a porous media simulation test case.

6.2.2 On the selection of deterministic trends

Now we wish to examine a major difficulty encountered when kriging based on few data : the selection and the estimation of deterministic trends. In computer experiments, the most commonly used Kriging model seems to be OK. However, OK reaches one of its limits when the first order stationarity assumption does not hold any longer, i.e. when non constant trends $t(\mathbf{x})$ are impossible to ignore. In this case, we are back to the general decomposition of equation 6.1, where y is assumed to be the sum of a deterministic trend t and one realization of a centered GP ε. At this stage, we may consider several subcases.

If t is known and the parameters of ε have to be estimated, a straightforward solution is to perform Simple Kriging of the residuals $\{y(\mathbf{x}) - t(\mathbf{x})\}_{\mathbf{x}\in D}$.

If t is unknown, it is common to distinguish between a linear and a more general non-linear framework. The case in which t depends linearly on its parameters and ε has a kwown covariance structure has been intensively studied : it is well known as Universal Kriging [MS05]. When the covariance parameters $\boldsymbol{\psi}$ are known and the trend is a linear combination of some chosen basis functions f_j ($j \in [1, b],\ b \in \mathbb{N}\backslash\{0\}$), the only unknowns are the parameters of the trend ($\forall j \in [1, b],\ \beta_j \in \mathbb{R}$) ; indeed, if $t(\mathbf{x}) = \sum_{j=1}^{b} \beta_j f_j(\mathbf{x})$, the β_j's can directly be estimated by Generalized Least Squares (GLS) :

$$\hat{\beta}(\psi_2) = (\mathbf{F}^T K_{\boldsymbol{\psi}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T K_{\boldsymbol{\psi}}^{-1} \mathbf{Z} = (\mathbf{F}^T R_{\psi_2}^{-1} \mathbf{F})^{-1} \mathbf{F}^T R_{\psi_2}^{-1} \mathbf{Z} \tag{6.2}$$

where $\mathbf{F}$ denotes the evaluation of $\mathbf{f}(x) = [f_1(x), \ldots, f_b(x)]$ at the n design points and $R_{\psi_2} = (1/\psi_1) K_{\boldsymbol{\psi}}$ (proportionality since the observations are here assumed to be noise-free) is the correlation matrix of the random vector $Y(\mathbf{X})$.

In practice, however, one has seldom the value of the covariance parameters at disposal previous to performing UK. So one has to estimate a model with linear trend and unknown covariance parameters $\boldsymbol{\psi}$ (in the following we will also refer to this case as « UK », like many practitioners do). Hence $\boldsymbol{\psi}$ and β have to be estimated within Kriging. At a first sight, this is likely to create a circularity problem : one needs a known trend t to work on the residuals $\{y(\mathbf{x}) - t(\mathbf{x})\}_{\mathbf{x}\in D}$ and thus estimate $\boldsymbol{\psi}$. On the other hand,

estimating t without taking the residuals into account may lead to unadapted trends (the estimation of the trend parameters would rely on Ordinary Least Squares instead of GLS). Fortunately, ML estimation gives a way to escape this vicious circle. Assuming that the covariance parameters to be estimated are $\boldsymbol{\psi} = (\sigma^2, p)$, and using MLE (and the same formula 6.2 for $\widehat{\beta}$), one can get a straightforward formula for $\widehat{\sigma^2}$, explicitly depending on ψ_2 :

$$\widehat{\sigma^2}(\psi_2) = (1/n)(\mathbf{Y} - \mathbf{F}\widehat{\beta}(\psi_2))^T R_{\psi_2}^{-1}(\mathbf{Y} - \mathbf{F}\widehat{\beta}(\psi_2)) \tag{6.3}$$

By injecting 6.2 and 6.3 in the expression of the likelihood, one can then obtain the so-called *concentrated likelihood* function $L(\psi_2, \widehat{\sigma^2}(\psi_2), \widehat{\beta}(\psi_2))$ which clearly depends only on ψ_2 and which has to be maximized in order to get $\widehat{\psi}_2$. The Kriging predictor with plugged-in covariance parameters is hence given by :

$$\widehat{Y}_{\widehat{\psi}_2}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\widehat{\beta}(\widehat{\psi}_2) + r^T(\mathbf{x})R_{\widehat{\psi}_2}^{-1}\left(\mathbf{Y} - \mathbf{F}\widehat{\beta}(\widehat{\psi}_2)\right) \tag{6.4}$$

where $r(\mathbf{x})$ is a vector of correlation values between Y at an unknown point $\mathbf{x}$ and at the points of the design $\mathbf{X}$. Most of the time (apart in Bayesian Kriging) the variability due to the estimation ψ_2 is not propagated, and one uses the regular UK prediction variance.

UK appears as a very convenient means to incorporate known deterministic trends within Kriging. By the way, we will see in the next section that overcoming the circularity problem is not easy in a more general non-linear framework. Now we would like to go one step deeper in practical considerations and raise a naive but complex question which has to be handled in real-world applications, and particularly in high-dimensional problems : how can one come back to the nature of a deterministic trend from raw data ?

As soon as neither prior information nor obvious graphical clue is available, one has indeed to select a trend on the basis of $(\mathbf{X}, \mathbf{Y})$. What means does one have to do so, and what risk does one run in case of a bad choice ? In order to show that these questions are crucial, let us first perform some toy experiments. The set-up is the following. A realization of a one-dimensional GP with known covariance function and parameters is simulated on a regular grid (401 points on $[-1, 1]$) and an affine trend is added ; From this set we choose different subsets of points and perform three types of Kriging : OK, UK with linear trend and UK with quadratic trend.

We choose at first a subset of 5 regularly distributed points. Due to the fact that the points are regularly spaced on the grid, all the three Kriging models give similar good results, even if in two of the three cases the trend is misspecified (See 6.1 left, for the

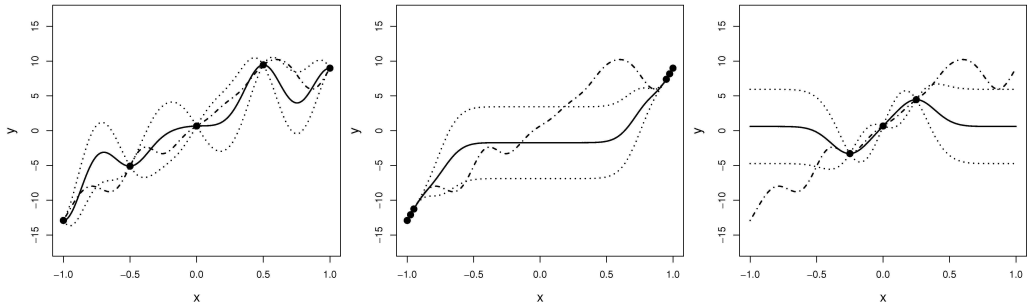


FIG. 6.1 – One realization (dashdot line) of a GP with linear trend is interpolated by Ordinary Kriging, based on 3 designs. The OK mean and 95% confidence intervals are represented by bold lines and dotted lines, respectively. The first design (left) is a regular grid ; the associated OK prediction seems satisfying, even if the trend model is misspecified. The second design (center) is formed by 6 points concentrated at the boundaries of the domain ; The Kriging predictor fails to capture the shape of the realization at the center of the domain. The third design is made of three points clustered at the center of the domain ; OK automatically comes back to the mean value outside of the design and dramatically misses the actual trend.

case of Ordinary Kriging). This might lead to the conclusion that specifying the trend is not very important and we could obtain good results using OK. But if we perform the same Krigings on different designs, where there are few points concentrated either on the boundaries or in the center of the domain, then the results are very bad (due to the ratio between the parameter ψ_2 and the subdivision length) when the trend is misspecified, see (6.1, middle and right). The covariance parameters used for the simulated process in 6.1 and 6.2 are $\psi = (5, 0.2)$. The results are even worse if we use the Kriging predictor given by OK or by UK with quadratic trend in extrapolation (6.2, left and right). In the one-dimensional case, the choice of the trend doesn't seem to be essential while interpolating data which are not very distant one from another with respect to the frequency of variation of the process. On the contrary, when the design is not regular and when we are in extrapolation, the performances of Kriging are very sensitive to the adequacy between the actual trend of the process and the Kriging trend.

Hence it seems (ironically) enough to properly fill the space to avoid the risks caused by the choice of trend functions. But what is possible in one or two dimensions becomes unrealistic when the dimension increases : a design with only one point at each vertex of a cubic domain $[0, 1]^d$ has 2^d points, i.e. 1024 points in 10 dimensions and more than a billion points in 30 dimensions (« curse of dimensionality », Cf. [HTF01]). As we usually

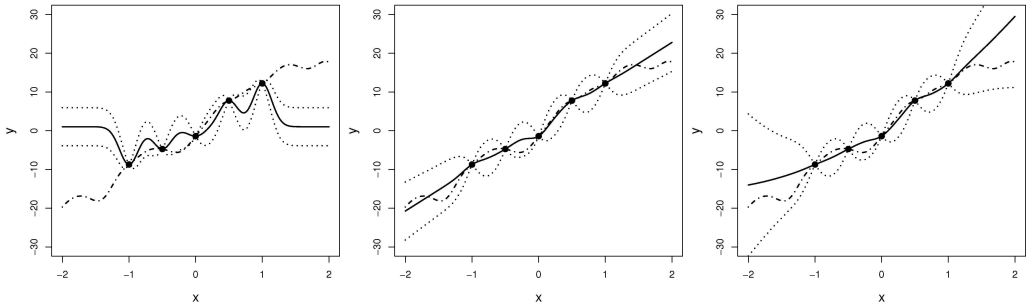


FIG. 6.2 – One realization (dashdot line) of a GP with linear trend is interpolated by three different Krigings, based on a regular grid (5 points evenly spaced between -1 and 1). The GP is here represented between -2 and 2, such that all three cases can be referred to as extrapolations. UK with affine trend (center) gives accurate results on the whole domain. On the contrary, both OK (left) and UK with quadratic trend (right) give good results between -1 and 1 but dramatically fail in extrapolation.

dispose of 10 observations per dimension, which is already an optimistic case, choosing a trend based on data only appears as a very difficult task.

Let us see nevertheless what would be possible in order to choose a trend starting from a data set $(\mathbf{X},\ \mathbf{Y})$: the classical frame of linear regression offers a panel of diagnostic tools [ABC92, HTF01] dedicated to validating both assumptions on the trend and on the model of residuals. For instance, commonly used indicators include R^2 (and adjusted R^2), the F-ratio and the p-values for each estimated regression coefficients, and numerous criteria to check the adequacy of the residuals to the underlying model. In most cases the Gaussian likelihood of the residuals is considered among the relevant criteria of model selection (some model testing and comparison techniques or even based upon it).

Now it seems necessary to recall that the latter measures are exclusively done at the design of experiments, also called « training sample »or « learning set » in the literature of statistical learning, see [HTF01]. Selecting the trend only on the basis of a R^2 fit would lead for instance to the systematic choice of models interpolating $(\mathbf{X},\ \mathbf{Y})$. However such models are not meant to be good in prediction outside the design of experiments. This warning leads to the double message :

- Model complexity must be taken into account in selection procedures
- Testing the model at some test points not used in the model fitting could be worth : this is for instance what cross-validation does.

The following experiment is performed in an intent to illustrate the first point. The second point will be illustrated in the next section. Here we investigate on a simple case how trend selection may be misleading when likelihood is the only criterion, without any consideration of model complexity. To do so, we compute, for each trend form of the Kriging model, the optimal parameter $\hat{p}$ by ML, we compare the corresponding values of the likelihoods and we select the Kriging model having the highest value of likelihood. In table 6.1, we compare three Kriging models (OK, UK with linear trend, UK with quadratic trend) for three different functions : one realization of a one-dimensional GP with 11 points and with Gaussian covariance function ($\psi = (5, 0.4)$), the same realization plus a linear trend $0.5+5x$, and the same realization plus a quadratic trend $0.5+5x+5x^2$.

TAB. 6.1 – Comparison of minimum $-2\log(L)$ values obtained by fitting three different Kriging models (OK, UK affine, UK quadratic) to three GP realizations. Each realization is drawn from the GP underlying one of the Kriging models. The design is a regular grid on $[-1, 1]$. The results illustrate that adding degrees of freedom to a Kriging model always lead to a larger value of the maximum likelihood.

	GP		GP +linear $\mathbf{t}$		GP+quadratic $\mathbf{t}$	
Kriging type	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$
OK	0.4082	**32.07**	0.4445	36.90	0.4595	38.80
UK, linear $\mathbf{t}$	0.4085	**31.89**	0.4085	31.89	0.4387	35.80
UK, quadratic $\mathbf{t}$	0.4084	31.89	0.4084	31.89	0.4084	31.89

Here it is essential to notice that the likelihood values are necessarily larger when adding more degrees of freedom to a statistical model. This constitutes a misleading incentive to always choose the model with the largest number of parameters within a given family. This happens for instance between Kriging models with first order and second order polynomial trends. As can be observed in table 6.1, L always increases (i.e. the values of $-2\ln(L(\hat{p}))$ decreases) with the complexity of nested model. What we should really compare are maximum likelihood values between models with the same number of degrees of freedom. On the last line of table 6.1, in the cases of the GP without trend and of the GP with linear trend, the estimated values $\hat{\beta}$ are very close to but different from zero. Thus the model obtained by automatically selecting the Kriging with highest likelihood will perform badly in extrapolation because of the higher order terms of the polynomial. The same phenomenon applies in an even more pronounced way with a linear trend in the case of a centered GP (first column, second row).

As a conclusion to this section, we have pointed out that OK and UK may seem to

deliver similar results when the design is dense [Ste99], but modeling the trend matters in extrapolation situations [JR89]. Since working in high-dimensional spaces means that we will practically always be in extrapolation, we need exploratory and visualization tools dedicated to finding trends in multivariate data. Recent methods of data mining and functional analysis may help [HTF01]. We propose now to use additive models within spatial interpolation.

6.3 Splitting the DOE : a strategy to overcome the circularity problem in Kriging with non-linear trends

6.3.1 Using non-linear additive models as external drift

Linear models are often used by practitioners of quantitative disciplines since they are simple to interpret and to assess. Additive models (AM's) can be seen as an extension of linear models. A precise description of these models —as well as an illustration of their practical interest— can be found for example in the book [HT91]. The advantage of AM's is to conserve the feature of non-interacting predictors, but they allow much more flexible inference for each univariate problem, using kernel smoothers for instance [Wah90]. The generic expression for an additive model is the following :

$$\begin{cases} Y_i = y(\mathbf{x}^i) + \varepsilon_i \ (1 \leq i \leq n) \\ \forall \mathbf{x} \in D, \ y(\mathbf{x}) = \alpha + \sum_{j=1}^{d} f_j(\mathbf{x}_j) \\ \varepsilon_i \text{ are realizations of } i.i.d. \text{ Gaussian random variables} \end{cases} \tag{6.5}$$

and the f_js are arbitrary univariate functions, one for each predictor[1] but possibly not the same kind of function for each dimension. Hence additive models deal with additive functions observed in a Gaussian noise. $\mathbf{x}$ is in fact assumed here to be a control variable, and not a random variable as in [HT91]. This model may be used to approximate deterministic computer experiments, provided that the response surface can reasonably be decomposed in an additive way. Once the nature of the f_js is chosen, they can be estimated using a powerful iterative procedure called *backfitting algorithm*, see [HTF01]. Backfitting means that f_1 is estimated on the basis of all data $(\mathbf{X}, \mathbf{Y})$, then f_2 is fitted to the residuals $\mathbf{Y} - f_1(\mathbf{X})$, and so on. Under mild assumptions, the backfitting algorithm converges and finds the unique best solution (in the L^2 sense) of the additive decomposition of equation 6.5. In this section, we propose a combination of additive model and

[1] It is also possible to include univariate functions for the interactions, see [HT91]

Kriging that offers the great flexibility of $AM's$ and yet interpolates the data. It seems very natural to combine both models by using the following decomposition :

$$\begin{cases} y(\mathbf{x}) = t(\mathbf{x}) + \varepsilon_{SK}(\mathbf{x}) \\ t(\mathbf{x}) = \alpha + \sum_{j=1}^{d} f_j(\mathbf{x}_j) \\ \varepsilon_{SK}(\mathbf{x}) \text{ is a GP realization like in equation 6.1} \end{cases} \tag{6.6}$$

This identity may at first seem similar to the equation of Universal Kriging. However, in this case the non-linear nature of the trend prevents one from solving the estimation globally. Indeed, a likelihood maximization would lead to an optimization problem in infinite dimension :

$$\max_{\boldsymbol{\psi},(f_j)_{j\in[1,d]}} L(\boldsymbol{\psi}, \mathbf{t}; \mathbf{Y}) \tag{6.7}$$

To our knowledge such a problem is analytically intractable, without strong assumptions concerning the space of functions in which the f_j's lie (an analysis in terms of RKHS seems likely to clarify this ... but is out of scope of the heuristic method proposed for now). On the other hand, the backfitting algorithm is not suited anymore if we take the Kriging part into account. Indeed, Kriging the residuals after fitting a smoother in one dimension would lead to an interpolation and thus end the iterative procedure without fitting the additive parts in the other dimensions.

Kriging with external trend [Cre93] seems to constitute a good alternative for solving both the problem of the « general » form of trend and the one of circularity. Consequently, we consider now a two-step approach (see Alg. 4) : first, the additive trend $t(\mathbf{x})$ is estimated using the backfitting algorithm, and then Simple Kriging is applied to $(\mathbf{Y} - \mathbf{t})$ with covariance parameters estimated on the basis of those residuals, by ML or other.

Algorithm 4 A first two-step approach to fit a Kriging with additive trend

1: Estimate the trend t by backfitting
2: Estimate the covariance parameters and fit a SK model on the basis of the residuals at $\mathbf{X}$

Unfortunately, there are significant drawbacks in the latter procedure, mainly related to the uncontrolled trade-off between deterministic and stochastic parts. Hence, the whole uncertainty reduces here —when plugging-in the trend estimated by backfitting and then performing Simple Kriging— to the Kriging variance estimated on the residuals ; there is indeed no global uncertainty on the trend. This may cause a large underestimation of the process variance associated with the model. Furthermore, these residuals may be not

well-suited to estimate the Gaussian Process part : the additive model is constructed to fit y accurately at the design —possibly leading to *overfitting*—, and thus the residuals at $\mathbf{X}$ are likely to vary with a smaller magnitude than in prediction. Since we look for a model with reasonable generalization properties, it is necessary to find an alternative way of estimating the covariance parameters. We propose here a sequential estimation technique for combined Kriging models like equation 6.5. It is based on the idea that when the trend is non-linear, the parameters of the GP model should be estimated on a validation set rather than on the set at which the trend is fitted.

Algorithm 5 An alternative two-step approach to fit a Kriging with additive trend

1: Consider two designs $\mathbf{X}_1$ and $\mathbf{X}_2$ ▷ possibly obtained by splitting $\mathbf{X}$
2: Estimate the trend t by backfitting, based on the data $(\mathbf{X}_1, z(\mathbf{X}_1))$
3: Estimate the SK covariance parameters on the basis of the residuals at $\mathbf{X}_2$, $\{t(\mathbf{x}) - z(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}_2}$
4: Fit the SK model on the basis of all residuals ▷ with parameters estimated at the previous step

6.3.2 A $3d$ application of Kriging with Additive Trend (KAT)

The previous approach is applied to a 3-dimensional example from an industrial test case. The data are obtained with a flow simulator and the numerical response z, standing for the outcome of interest, is studied as a function of three physical parameters characterizing the porous media and denoted by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3 \in [-1, 1]$. The response is simulated at 1331 locations corresponding to a 11-level full factorial design, denoted by « F » in the sequel. Our goal is to provide a surrogate of the simulator on the basis of a poor design of experiments. The metamodel should interpolate the data (to respect the determinism of the underlying simulation) and provide a prediction uncertainty that allows statistical-based exploration, for instance to solve optimization problems. Furthermore, it should take into account a prior knowledge inherited from a previous study : the phenomenon is almost additive in its parameters.

Our initial design, $\mathbf{X}_1$, is a 20-elements Hammersley sequence. We first perform a graphical analysis (Cf. 6.3) of the response at $\mathbf{X}_1$, discuss the hypothesis of additivity, and propose several kinds of linear and additive trends to model the data. Algorithm 1 is tested with the design $\mathbf{X}_1$ (Cf. 6.2). Then a second design, $\mathbf{X}_2$, is used for an intermediate validation of the covariance parameters of the model previously obtained. $\mathbf{X}_2$ is made of 14 points taken from a 40-elements D-optimal design (see 6.4). Algorithm 2 is then performed by re-estimating the covariance parameters of the previous SK model on the basis of the residuals at $\mathbf{X}_2$. An original estimation method is proposed, which differs

from the traditional MLE : the process variance σ^2 is fixed such that the standardized residuals have most of their values between -2 and 2 [JSW98] and the range parameter p is chosen in order to minimize the ISE at the design $\mathbf{X}_2$ 6.5. The full factorial design F is finally used for a phase of model validation 6.6.

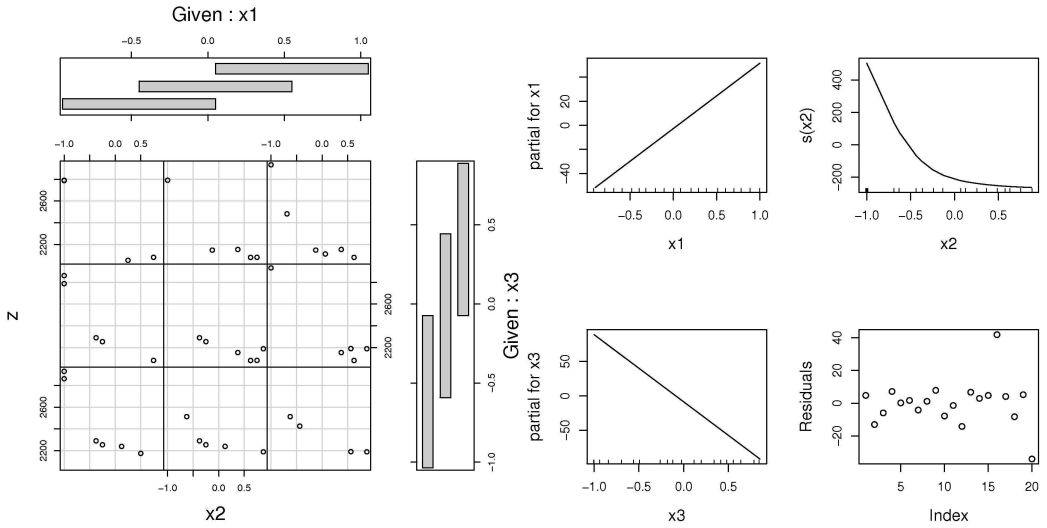


FIG. 6.3 – Coplots of y on the Hammersley design $\mathbf{X}_1$ (left) and summary of the additive components and the residuals obtained after application of the backfitting algorithm (right). The additive model is here chosen with a linear function in both directions of $\mathbf{x}_1$ and $\mathbf{x}_3$, and a smoothing spline in $\mathbf{x}_2$'s direction.

A graphical analysis of the coplots at $\mathbf{X}_1$ (Cf. 6.3) does not reject the prior belief of additivity. A first additive decomposition is then estimated using splines in all directions (referred to as « GAM splines » in the following). We observe that we might take a linear trend in the directions of $\mathbf{x}_1$ and $\mathbf{x}_3$, and a non-linear trend in $\mathbf{x}_2$ without loosing much accuracy (see Table 6.2 for a quantitative validation). Hence we choose to fit an additive model with mixed trends, called « GAM mixed » in the sequel.

Different Krigings with external trend are fitted to the observed data at the design $\mathbf{X}_1$. In all cases, the SK part has a structure of isotropic GP with Gaussian covariance. We focus on the two additive trends defined above and on two additional linear trends : a first and a second order regression polynomials. For each model, we fit the trends respectively by OLS and backfitting, and we measure their relevance using indicators computed with the residuals at the design $\mathbf{X}_1$ (residuals deviance and p-values when available). Then we fit a Kriging to the residuals, as explained in Algorithm 4. For each Kriging model,

we store the maximum reached value of the log-likelihood and the corresponding range and variance values. The results are listed in table 6.2.

TAB. 6.2 – Optimal loglikelihood values and estimated covariance parameters associated with the residuals provided by Algorithm 4 at $\mathbf{X}_1$ with different trend structures. The R^2 values are computed by comparing the residual deviance after fitting the trend only, to the total variance of the response at the design $\mathbf{X}_1$.

Model	Loglikelihood	Range	σ^2	R^2_{adj}	R^2	p-value
1st order Linear + SK	-121.91	1.04	26101.89	0.78	0.82	4.03e-06
2st order Linear + SK	-100.69	0.048	1381.13	0.97	0.98	6.44e-08
GAM splines + SK	-76.01	0.048	117.10	-	0.99	-
GAM mixed +SK	-80.62	0.16	185.71	-	0.99	-

These results support the belief that a general additive trend is adapted for these data : both the variance of residuals and the values of their likelihood (compared to the 2nd order linear model, which uses more degrees of freedom) indicate their good fit to the data.

In practice, however, we care more about the model's abilities to make correct predictions at new points than about its mean squared error at the design. Hence, model validation should not be blindly supported by the indicator R^2 or the likelihood of the residuals at $\mathbf{X}_1$. First, we should consider the number of degrees of freedom of the model. Second, it may be worth validating the model outside of the design. Indeed, the residuals drawn from fig. 6.3 are computed at the same locations as those used to fit the trend.

Concerning the first point, we can compare the degrees of freedom of both « 2nd order linear » and « GAM mixed » : respectively 10 and 7. Regarding the second point, we conduct a validation test on some additional data, inspired by the cross-validation procedure. Following Algorithm 5, $\mathbf{X}_2$ is used to valid and update the parameters associated with the model fitted at $\mathbf{X}_1$.

The points of $\mathbf{X}_2$ are used to test the validity of the covariance parameters of the model « GAM mixed », previously estimated by ML. Figure 6.5 shows the associated residuals standardized by the ML variance (left), and the behaviour of the ISE at $\mathbf{X}_2$ as a function of the parameter p (right). We recall that the residuals should satisfy the assumption of normality in order to get a relevant Kriging modeling, which seems a reasonable effort in the intent to get acceptable statistical predictions.

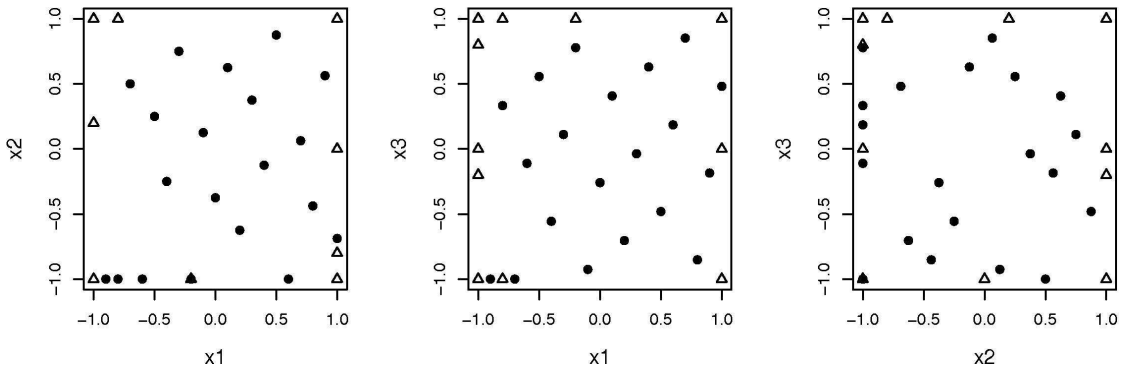


FIG. 6.4 – Coplots representing both $\mathbf{X}_1$ (dots) and $\mathbf{X}_2$ (triangles) designs in projection on all pairs of coordinates. The 3 graphics illustrate the space-filling behaviour of $\mathbf{X}_1$ and the D-optimal nature of $\mathbf{X}_2$. $\mathbf{X}_2$ appears to be reasonably disconnected from $\mathbf{X}_1$.

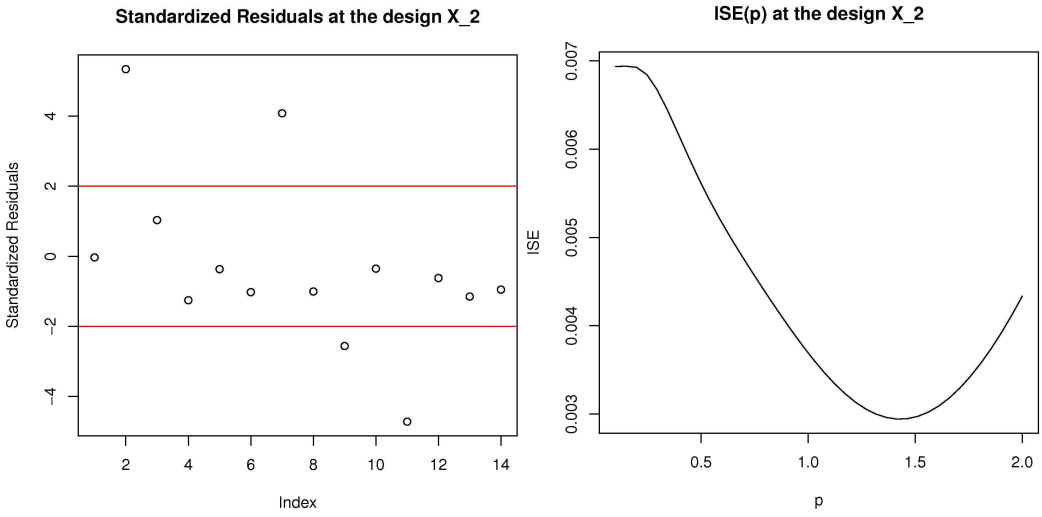


FIG. 6.5 – Standardized residuals at the validation design $\mathbf{X}_2$ (left) and variation of the ISE with respect to the covariance parameter p (right). One can observe that the value of p found by ML at $\mathbf{X}_1$ ($p_1 = 0.16$) is clearly suboptimal to get an accurate Kriging predictor when extrapolating to $\mathbf{X}_2$.

Figure (6.5, right) shows that the ISE at the validation design $\mathbf{X}_2$ can be significantly reduced by increasing the range p. Following Algorithm 2, we re-estimate the covariance parameters based on these residuals at $\mathbf{X}_2$. Instead of using ML however, we prefer to directly use the work done hereabove to compute the ISE as a function of the range. It appears indeed that the optimal range to accurately fit the residuals at the validation design is given by $p_2 = 1.4$. Concerning the variance, we observe more satisfying stan-

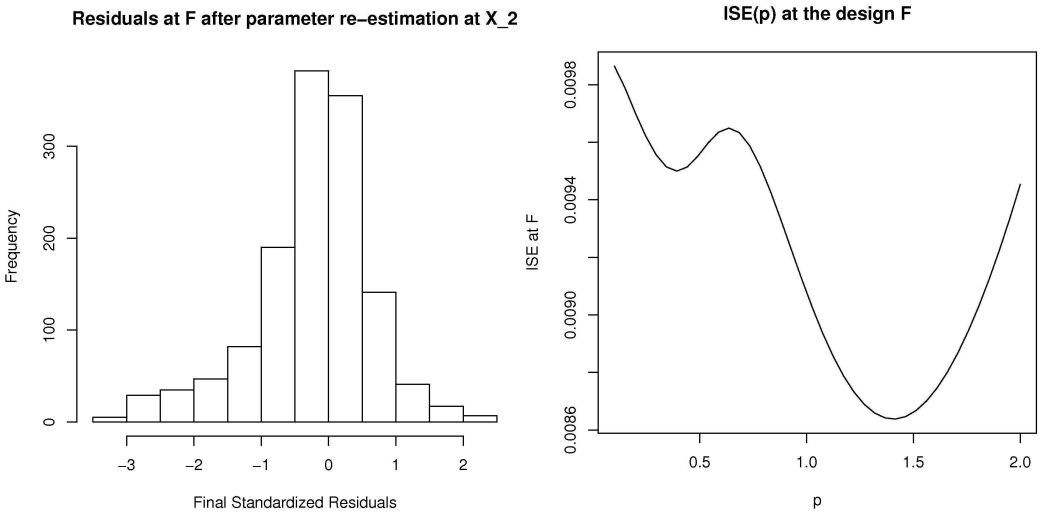


FIG. 6.6 – **Left :** Histogram of the standardized residuals at the test design F with the model previously obtained by Algorithm 2 (GAM mixed, $\sigma^2 = \sigma_2^2,\ p = p_2$). **Right :** Variation of the ISE with respect to the covariance parameter p : the value $p_2 = 1.4$ previously chosen at $\mathbf{X}_2$ is almost optimal again.

dardized residuals with $\sigma_2^2 = (2 \times \sigma_{ML})^2$. So we keep this new value as process variance. [2].

We finally test the model of Algorithm 2 at the full factorial design F 6.6. The standardized residuals (with the variance σ_2^2) and the ISE as a function of p validate our empirical decisions made on the basis of the intermediate design $\mathbf{X}_2$ (note that ML on $\mathbf{X}_2$ —see remark hereabove— gives also better results than ML on $\mathbf{X}_1$ but the cross-validating strategy minimizing the ISE at $\mathbf{X}_2$ remains the best).

To conclude with, the algorithm investigated performed well on this example : Simple Kriging seems to constitute a good complement to additive models in an intent to interpolate data and also possibly explain some non-additive part. The method we used here allows inference of covariance parameters with well-suited values for a correct quantification of uncertainty. This seems encouraging to develop further « cross-validation-like » methods for the hybridation of additive models and Kriging.

[2] *Remark :* A ML estimation with the residuals at $\mathbf{X}_2$ delivers $p = 0.97$.

6.3.3 Conclusions and perspectives of the study

Toy experiments on the topic of trend selection illustrated the fact that the likelihood cannot be considered as only criterion when comparing different functional families. This is suggesting methods penalizing complexity (like AIC and BIC). But we mainly wish to place emphasis on the risks took when predicting with trended Kriging : in higher dimensions, we will allways be in an extrapolation situation. Choosing a trend with the help of a small design of experiments thus seems very risky. This appears as a suitable argument to consider Ordinary Kriging (possibly with local neighborhoods [Cre93]) in the cases where no prior information concerning the trend is available.

In other respects, we proposed a model combining an additive model and Simple Kriging. The application to a simple industrial test case confirmed that directly Kriging the residuals by ML gives a poor result. Our attempt to adapt a method inspired by cross-validation with a single test set gave here a Kriging model which estimated parameters have different features from ML, apparently accounting well both for the additive part and for the non-additive part of the response. However, the question of the robustness to a change of design has not been raised yet. This is a subject to be treated in further works. Mixing Kriging models (see chapter 8 for more technical detail on mixed Krigings) obtained using different partitions of the design of experiments might constitute a suitable candidate method to address this issue.

Chapitre 7

Approximation du code MORET. Symétries et bruits hétérogènes : prise en compte dans le Krigeage

Ce chapitre se démarque des précédents par une structure bien particulière : on commence ici par passer en revue un cas d'application étudié en 2006 – 2007 lors d'une collaboration avec l'IRSN (*Institut de Radioprotection et de Sûreté Nucléaire*) dans le cadre du Consortium DICE, puis on présente des résultats et des modèles plus généraux obtenus *a posteriori* en réponse à des questions s'étant posées lors de cette première étude. Il s'agit d'une part de résultats concernant les processus aléatoires à trajectoires invariantes par symétrie et la prise en compte de symétries physiques dans le Krigeage, et d'autre part de l'écriture d'un modèle statistique rigoureux pour permettre l'utilisation des modèles classiques de Krigeage lorsque les observations sont entâchées d'un bruit d'observation réglable par le modélisateur. C'est en particulier le cas des simulateurs stochastiques avec possibilité de multiples évaluations pour un même vecteur d'entrée.

7.1 Présentation générale du cas d'étude IRSN-SEC

7.1.1 Contexte industriel et scientifique

Criticité neutronique et coefficient « k-effectif »

Les réactions de fission entretenues dans les réacteurs de centrales nucléaires afin de produire de la chaleur reposent sur une réaction neutronique en chaîne. Les atomes d'uranium, présents dans le coeur du réacteur sous forme d'un assemblage de « crayons »

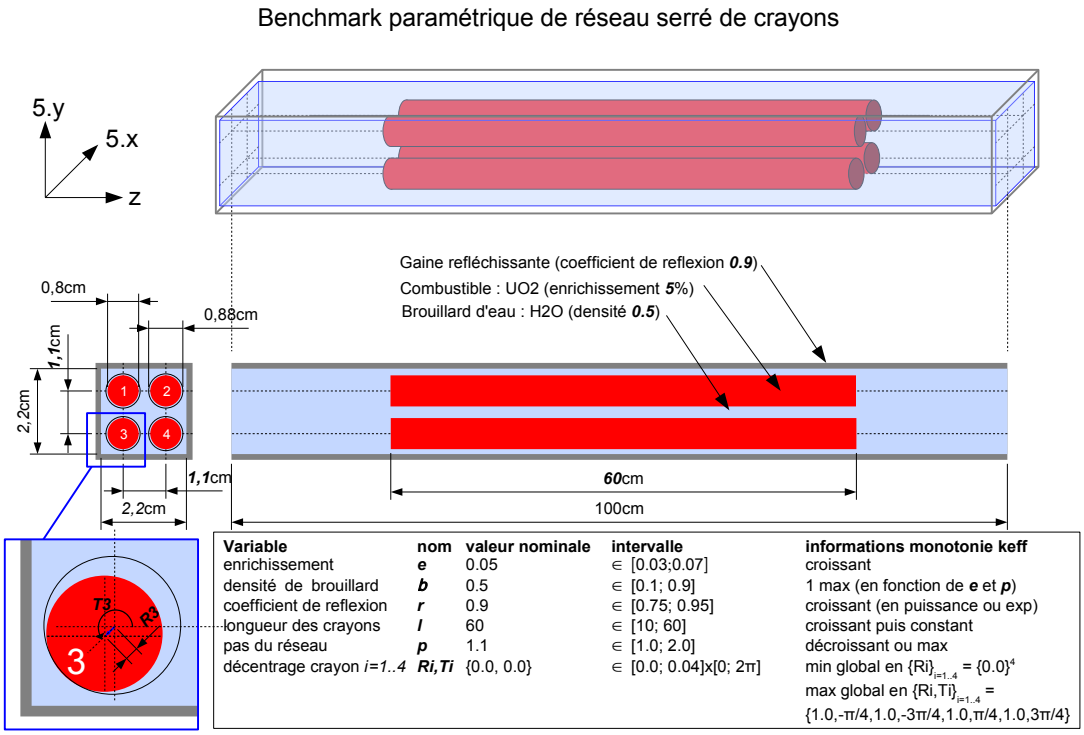
baignés dans fluide modérateur, sont bombardés de neutrons afin d'initier la réaction. Chaque collision d'intensité ad hoc[1] entre un neutron et un atome d'uranium provoque alors la fission du noyau de ce dernier, donnant des produits de fission (deux noyaux moyennement lourds, tels que ceux du Krypton et du Baryum), ainsi que de nouveaux neutrons, appelés *neutrons prompts*. Les produits de fission sont eux aussi susceptibles de se désintégrer suite à la première réaction, produisant d'autres nouveaux neutrons appelés *neutrons retardés*. Au total, chaque neutron engagé dans la fission d'un noyau d'uranium va engendrer la libération d'un certain nombre de nouveaux neutrons dans le milieu. Bien entendu, tous les nouveaux neutrons produits ne sont pas destinés à occasionner une nouvelle réaction de fission ; le système serait alors complètement explosif. Les neutrons engendrés par les fissions ont différentes vitesses, et différentes trajectoires, si bien que l'on peut considérer en toute généralité qu'un neutron libéré a une certaine probabilité de provoquer une nouvelle fission.

Les grandeurs d'intérêt pour décrire la stabilité de la réaction en chaîne apparaîssent alors comme des expressions telles que le nombre moyen de réactions de fission occasionnées suite à une réaction de fission. Plus précisément (Cf. [MR07], p. 158), le *coefficient de multiplication effectif des neutrons* (k_{eff}) est défini comme "le rapport du nombre de neutrons produits sur le nombre de neutrons perdus (par fuite et absorbtion)". Pour reprendre le rapport de l'IRSN 2007 précédemment cité, "Ce coefficient caractérise donc « l'état de criticité » du milieu fissile considéré. Dans ces configurations, souvent complexes, le k_{eff} est généralement estimé au moyen d'un code de calcul, permettant de modéliser les matériaux et leurs géométries ainsi que les lois physiques régissant le comportement des particules". En particulier, le modèle mathématique classiquement retenu pour décrire l'évolution de la population neutronique est l'équation de Boltzmann [Bel01]. Or comme souligné dans [MR07], le problème est que l'on ne sait résoudre analytiquement cette équation que dans des cas d'école. La simulation numérique apparaît alors comme une nécessité pour étudier le transport neutronique en milieu complexe, que l'on fasse appel à des codes déterministes (par discrétisation des équations) ou probabilistes (par méthodes dites « de Monte-Carlo »). MORET est un code Monte-Carlo développé par l'IRSN, dédié aux études de criticité neutronique, et souvent utilisé de pair avec un code déterministe de manière à optimiser les performances de calcul à la fois en termes de robustesse et de compromis précision-rapidité.

[1] Pour qu'une réaction de fission nucléaire survienne lors de la collision d'un neutron avec un noyau d'uranium, il faut que le neutron ne soit ni trop lent ni trop rapide. La gamme d'énergies propices à une entrée en réaction est souvent résumée en physique des particules à l'aide de graphes donnant la section efficace (homogène à une aire, et traduisant une probabilité de réaction) en fonction de l'énergie.

Problème abordé ici

Le code MORET permet d'estimer le k_{eff} associé à des géométries complexes, prenant en compte de nombreuses variables de nature physico-chimique, de position, ainsi que divers paramètres issus de banques de données internationales (sections efficaces, lois de probabilité de la direction de particules après un choc, etc.). Une configuration étant donnée, la résolution approchée de l'équation de Boltzmann —grâce à une méthode de Monte-Carlo par Chaînes de Markov— délivre une distribution statistique de la répartition spatiale des neutrons, permettant d'obtenir une estimation moyenne du k_{eff} ainsi qu'une valeur d'écart-type associée.



Variable	nom	valeur nominale	intervalle	informations monotonie keff
enrichissement	*e*	0.05	∈ [0.03;0.07]	croissant
densité de brouillard	*b*	0.5	∈ [0.1; 0.9]	1 max (en fonction de *e* et *p*)
coefficient de reflexion	*r*	0.9	∈ [0.75; 0.95]	croissant (en puissance ou exp)
longueur des crayons	*l*	60	∈ [10; 60]	croissant puis constant
pas du réseau	*p*	1.1	∈ [1.0; 2.0]	décroissant ou max
décentrage crayon *i=1..4*	*Ri,Ti*	{0.0, 0.0}	∈ [0.0; 0.04]x[0; 2π]	min global en $\{Ri\}_{i=1..4} = \{0.0\}^4$
				max global en $\{Ri,Ti\}_{i=1..4}$ = {1.0,-π/4,1.0,-3π/4,1.0,π/4,1.0,3π/4}

Le travail présenté ici constitue une première étape dans la résolution d'un problème d'optimisation et de quantification d'incertitudes sur le coefficient de criticité neutronique d'un réseau serré de crayons d'uranium. Il s'agit, à paramètres physico-chimiques fixés, d'étudier la manière dont évolue le k_{eff} en fonction de la position des crayons d'uranium à l'intérieur de leurs compartiments respectifs. Il existe en effet un jeu de nature géométrique sur leurs positions, dont les conséquences doivent être maîtrisées. Par

soucis de simplicité, l'étude a porté dans cette première phase sur un réseau de quatre crayons. La position de chacun des crayons de combustible est paramétrée par une distance de décentrage (elle aussi fixée) ainsi qu'un angle, censés décrire le jeu géométrique des crayons dans leur emplacement.

Le but est précisément de caractériser l'évolution du k_{eff} en fonction des positions relatives des crayons, et plus particulièrement de se donner les moyens d'étudier la loi de probabilité et les maxima du k_{eff} lorsque les positions varient aléatoirement, en supposant par exemple que les angles $(t_i, 1 \leq i \leq 4)$ sont indépendants et uniformément distribués sur $[0, 2\pi[$. Précisons que, par mesure de confidentialité, les unités du k_{eff} ont été modifiées. Précisons aussi que les valeurs de k_{eff} simulées (et après changement d'échelle) sont entâchées d'un bruit blanc gaussien de variance 10^{-8}.

7.1.2 Résultats obtenus par régressions linéaire et additive

Etude préliminaire sur un plan factoriel complet à 4 niveaux

Nous désignerons dans la suite par la lettre C le plan factoriel complet à 4 niveaux fourni par l'IRSN, et ayant servi de point de départ à la présente étude. Chacune des quatre variables $\{t_i |\ i \in [1,4]\}$ y prend les valeurs $\{\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}\}$, et le plan complet compte ainsi $4^4 = 256$ éléments. Le graphique de gauche de la figure 7.1 représente C en projection sur les plans (géométriques) formés par toutes les paires de variables $\{(t_i, t_j) |\ i, j \in [1,4],\ i \neq j\}$. Sur ces plans, C est ainsi vu comme un plan d'expériences à deux dimensions pour lequel chacune des expériences à été répétée 16 fois.

Régression additive non-paramétrique sur C

Les tendances non-paramétriques que l'on visualise par application d'une méthode « gam » (Cf. graphe de droite sur la fig. 7.1) et le taux de variance expliquée de 97% obtenu peuvent inciter à conclure hâtivement que la réponse k_{eff} est bel est bien additive en les t_i, et que les effets principaux sont affines par morceaux. Nous allons voir sur la base de deux modèles de régression linéaire en quoi la forme du plan C rend difficile à identifier précisément l'allure réelle des effets principaux.

Régressions linéaires sur C

On se propose d'essayer un modèle de régression sans interactions avec comme prédicteurs les fonctions valeurs absolues $|t_1 - 3\pi/4|,\ |t_2 - 5\pi/4|,\ |t_3 - 5\pi/4|,\ |t_4 - 3\pi/4|$ (modèle appelé ici *Vabs*). La table d'analyse de variance ci-dessous montre que le modèle est

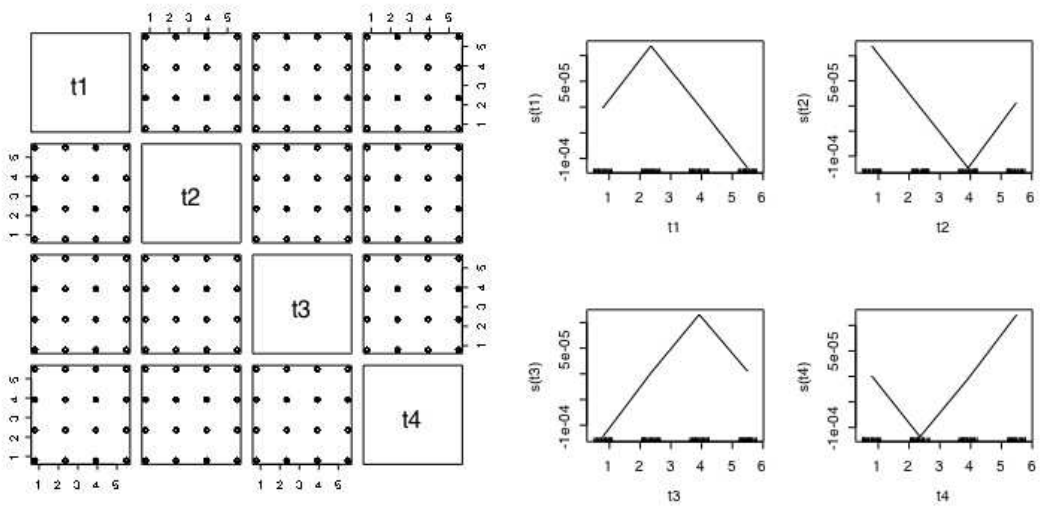


FIG. 7.1 – Représentation de C en projection sur les plans formés par les paires de variables $\{(t_i, t_j)|\ i, j \in [1, 4],\ i \neq j\}$ (à gauche), et splines de lissage monodimensionnelles issues de l'estimation d'un modèle additif à partir du plan C (à droite).

précis (en interne), au vu de l'indice R^2 proche de 97 %. Par ailleurs, les très faibles p-valeurs indiquent que les coefficients sont bien estimés, et leur signe est cohérent avec la courbure des fonctions estimées par le modèle « gam ».

Il ne faudrait cependant pas oublier que la nature du plan C a pour conséquence que chacun des effets principaux visualisés avec le modèle « gam » a été estimé à partir d'un plan d'expériences monodimensionnel formé de 64 répétitions en 4 points régulièrement espacés sur $[\frac{\pi}{4}, \frac{7\pi}{4}]$. Il subsiste donc une forte ambiguïté sur la forme fonctionnelle la plus adaptée. Par exemple, s'agissant de grandeurs angulaires, il est naturel de songer à des fonctions trigonométriques. Il n'est donc pas surprenant qu'un modèle de type polynôme trigonométrique de degré 1 (*trigo1*) soit aussi précis que le modèle précédent.

Sur la table d'Anova (7.1.2), on remarque que les coefficients obtenus sont tous pratiquement égaux au signe près, modulo l'erreur d'estimation. En particulier, cela indique que les fonctions trigonométriques peuvent être réarrangées de façon à faire apparaître les points de cassure visibles sur la représentation gam. Par exemple, on sait que $cos(t - \frac{3\pi}{4}) = -\frac{\sqrt{2}}{2} * (cos(t) - sin(t))$. Cela suggère un nouveau modèle de régression dont les prédicteurs sont $z_1 = cos(t_1 - \frac{3\pi}{4}),\ z_2 = cos(t_2 - \frac{5\pi}{4}),\ z_3 = cos(t_3 - \frac{5\pi}{4}),\ z_4 = cos(t_4 - \frac{5\pi}{4})$. La table d'analyse de variance (7.1.2) confirme cette intuition. On peut également ob-

```
lm(formula = keff ~ I(abs(t1 - 3 * pi/4)) + I(abs(t2 - 5 * pi/4)) +
    I(abs(t3 - 5 * pi/4)) + I(abs(t4 - 3 * pi/4)))

Residuals:
      Min        1Q    Median        3Q       Max
-9.608e-05 -2.149e-05 -2.937e-06  1.678e-05  9.293e-05

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)         1.080e-01  5.725e-06 18865.45   <2e-16 ***
I(abs(t1 - 3 * pi/4)) -7.565e-05  1.719e-06   -44.00   <2e-16 ***
I(abs(t2 - 5 * pi/4))  7.762e-05  1.718e-06    45.18   <2e-16 ***
I(abs(t3 - 5 * pi/4)) -7.579e-05  1.718e-06   -44.12   <2e-16 ***
I(abs(t4 - 3 * pi/4))  7.634e-05  1.719e-06    44.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.052e-05 on 251 degrees of freedom
Multiple R-squared: 0.9692,     Adjusted R-squared: 0.9687
F-statistic:  1974 on 4 and 251 DF,  p-value: < 2.2e-16
```

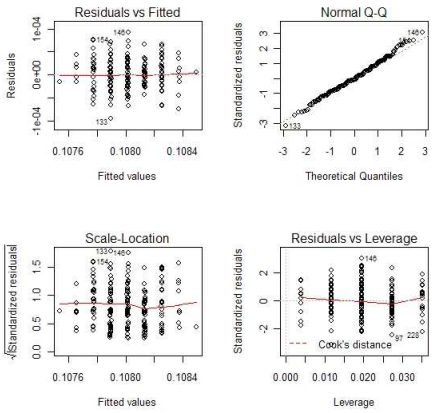


FIG. 7.2 – Table d'ANOVA du modèle de régression linéaire *Vabs* sur le plan C

server que le signe des coefficients estimés est en adéquation avec la représentation gam.

La similarité des résultats obtenus par ces 2 modèles en terme de R^2 s'explique par le fait qu'ils sont pratiquement équivalents sur le plan d'expériences utilisé et que le R^2 est un indicateur de précision interne, c'est à dire limité aux points de ce plan. Par contre, cette analyse suggère que les valeurs prédites par les modèles peuvent être sensiblement différentes pour des points tests externes. En outre, sur ces points tests, les modèles peuvent s'avérer très imprécis comme on va le voir maintenant[2].

Des validations externes décevantes sur un plan test uniforme

On se propose de comparer les valeurs prédites par les modèles précédents avec le résultat du code MORET sur des points externes, i.e. n'ayant pas servi lors de la phase d'estimation. Nous avons généré un second plan d'expériences, noté U, constitué de 100 points tirés uniformément dans $[0, 2\pi]^4$. Après avoir lancé le simulateur MORET au plan U, on a comparé les résultats de simulation à ceux prédits par le modèle additif et les deux modèles de régression *Vabs* et *trigo2* estimés sur C.

Nous avons calculé le taux de variance expliquée par chaque modèle. En notant « m » un modèle quelconque et « k_{eff} » la vraie réponse, ce taux s'écrit :

[2]Remarque : la proximité des valeurs des coefficients de la table d'ANOVA suggère que y ne dépend ici que d'une seule variable, i.e. $z_1 - z_2 + z_3 - z_4$. Ceci est précisé plus en détail dans le rapport [DHB07].

```
lm(formula = keff ~ I(cos(t1)) + I(sin(t1)) + I(cos(t2)) + I(sin(t2)) +
    I(cos(t3)) + I(sin(t3)) + I(cos(t4)) + I(sin(t4)))

Residuals:
      Min        1Q    Median        3Q       Max
-9.603e-05 -1.885e-05 -1.281e-06  1.713e-05  9.204e-05

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.080e-01  1.904e-06 56736.92   <2e-16 ***
I(cos(t1))  -8.547e-05  2.693e-06   -31.73   <2e-16 ***
I(sin(t1))   8.236e-05  2.691e-06    30.60   <2e-16 ***
I(cos(t2))   8.961e-05  2.693e-06    33.27   <2e-16 ***
I(sin(t2))   8.274e-05  2.691e-06    30.74   <2e-16 ***
I(cos(t3))  -8.320e-05  2.693e-06   -30.89   <2e-16 ***
I(sin(t3))  -8.508e-05  2.691e-06   -31.61   <2e-16 ***
I(cos(t4))   8.636e-05  2.693e-06    32.06   <2e-16 ***
I(sin(t4))  -8.303e-05  2.691e-06   -30.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046e-05 on 247 degrees of freedom
Multiple R-squared: 0.9698,    Adjusted R-squared: 0.9688
F-statistic: 991.1 on 8 and 247 DF,  p-value: < 2.2e-16
```

```
lm(formula = keff ~ I(cos(t1 - 3 * pi/4)) + I(cos(t2 - 5 * pi/4)) +
    I(cos(t3 - 5 * pi/4)) + I(cos(t4 - 3 * pi/4)))

Residuals:
      Min        1Q    Median        3Q       Max
-9.632e-05 -2.132e-05 -2.999e-06  1.678e-05  9.293e-05

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)         1.080e-01  1.907e-06 56632.37   <2e-16 ***
I(cos(t1 - 3 * pi/4))  1.187e-04  2.697e-06    44.00   <2e-16 ***
I(cos(t2 - 5 * pi/4)) -1.219e-04  2.697e-06   -45.18   <2e-16 ***
I(cos(t3 - 5 * pi/4))  1.190e-04  2.697e-06    44.11   <2e-16 ***
I(cos(t4 - 3 * pi/4)) -1.198e-04  2.697e-06   -44.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.052e-05 on 251 degrees of freedom
Multiple R-squared: 0.9692,    Adjusted R-squared: 0.9687
F-statistic:  1974 on 4 and 251 DF,  p-value: < 2.2e-16
```

FIG. 7.3 – Table d'ANOVA des modèles de régression linéaire *trigo1* et *trigo2* sur C

$$\frac{Var\left(k_{eff}(U)\right) - Var\left(m(U) - k_{eff}(U)\right)}{Var\left(k_{eff}(U)\right)} = 1 - \frac{Var\left(m(U) - k_{eff}(U)\right)}{Var\left(k_{eff}(U)\right)} \tag{7.1}$$

Cette quantité[3] est souvent présentée comme l'analogue externe du R^2. Notons que cette analogie est limitée, puisqu'il est par exemple possible que le taux en question soit négatif. Nous considérerons aussi la MSE (*Mean Squared Error*), qui est classiquement employée comme critère de comparaison des métamodèles :

$$MSE(U) := \frac{1}{|U|} \sum_{\mathbf{x} \in U} \left(m(\mathbf{x}) - k_{eff}(\mathbf{x})\right)^2 . \tag{7.2}$$

Comme l'illustre le graphe de droite de la figure 7.4, les prédictions faites au plan U par le modèle gam (splines monodimensionnelles) estimé sur C diffèrent nettement des vraies réponses obtenues au plan U par le code Moret (la situation idéale serait celle d'un nuage de point confondu avec la première bissectrice). Le taux de variance expliquée de 51.58% vient renforcer l'idée que le modèle est médiocre en performances externes, et ce en dépit de l'excellent R^2 relevé sur le plan C. Dans le même esprit, les prédictions obtenues sur U avec les modèles de régression linéaire (*Vabs* : 63.9%, *trigo1* : 68.15%) se sont avérées moins enthousiasmantes que la validation interne optimiste des modèles.

Le comportement médiocre des modèles sur des points tests s'explique ici non pas par un mauvais choix de modèle, mais par le choix du plan d'expériences. Celui-ci occasionne

[3] $Var\left(k_{eff}(U)\right)$ signifie ici la variance empirique du vecteur des valeurs de k_{eff} observées en U.

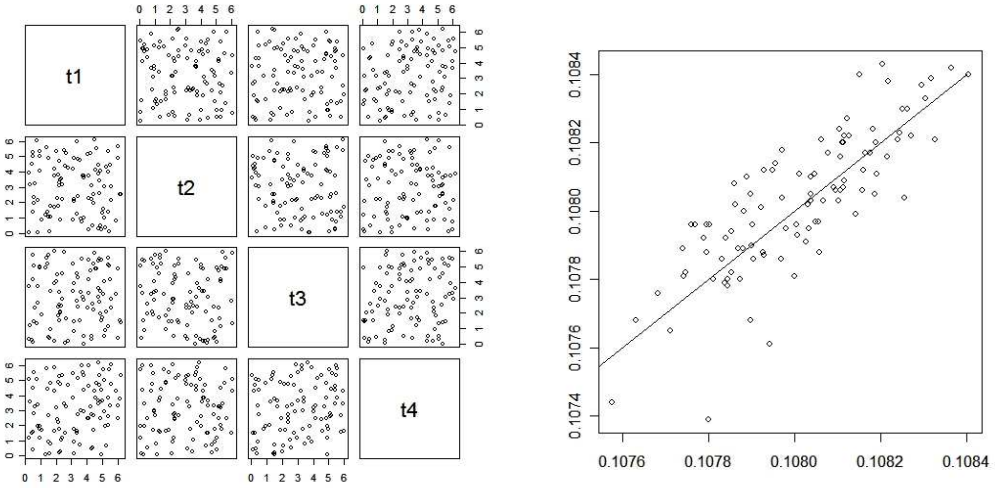


FIG. 7.4 – U, i.e. 100 points tirés uniformément dans $[0, 2\pi]^4$ (à gauche), et les prédictions au plan U fournies par le gam estimé sur C v.s. les vraies réponses au plan U (à droite).

trop de redondances en projection, ce qui fait que la question de l'estimation d'un modèle se trouve mal posée (difficulté de discriminer des modèles différents).

Ré-estimation des modèles sur un plan SFD

50 nouveaux points ont été choisis de façon à bien couvrir l'espace. Pour éviter le problème observé au paragraphe précédent, on souhaite que le nouveau plan n'ait pas trop de redondances en projections sur les axes factoriels. Comme plans possibles figurent donc les hypercubes latins ou les plans de Strauss (Cf. thèse [Fra08]). Nous avons ici opté pour un plan de Strauss. Les projections en dimension 2 permettent de visualiser la différence par rapport au plan factoriel, et la bonne répartition dans l'espace.

On réessaye alors un modèle GAM non-paramétrique. On peut noter qu'avec seulement 50 points bien répartis (au lieu des 256 du plan factoriel), on dispose cette fois de 50 points environ (en projection sur chaque axe de coordonnées) pour estimer les splines, ce qui fait que l'allure est nécessairement plus précise.

Les quatre graphiques évoquent une fonction trigonométrique, avec des effets de bord. Pour en avoir confirmation, on peut ré-estimer les modèles précédents (prédicteurs en valeur absolue, prédicteurs fonctions trigo). On constate que le modèle avec prédicteurs

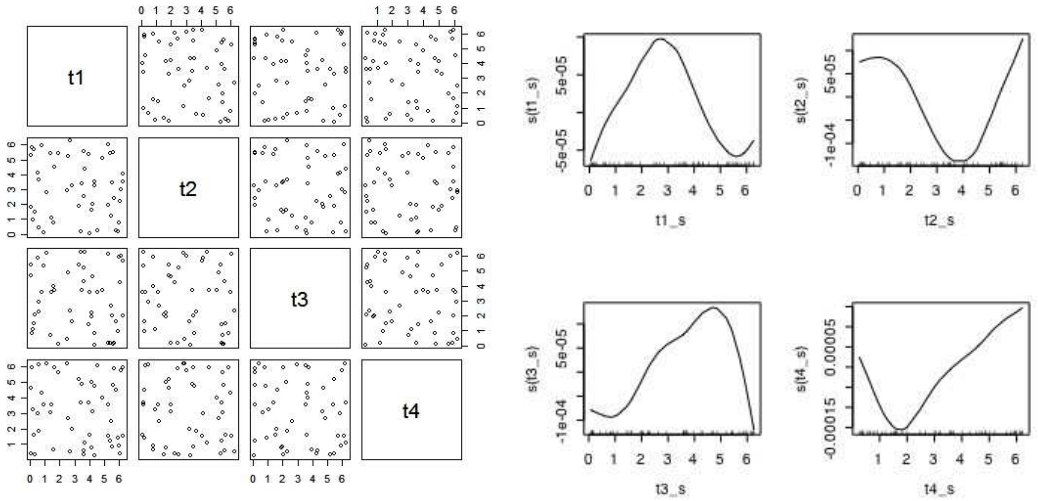


FIG. 7.5 – Plan de Strauss S à 50 éléments, vu en projection sur les plans engendrés par les paires d'axes canoniques (à gauche), et splines de lissage monodimensionnelles obtenues par estimation d'un modèle additif pour k_{eff} sur la base de S (à droite).

trigonométriques est le plus précis, ce qui confirme l'information donnée par les splines monodimensionnelles. Le R^2 de 54% est beaucoup moins bon, mais c'est plutôt rassurant : en effet, cette valeur est comparable au R^2 externe de 51.58% obtenu en prédiction sur le plan test U. On est donc bien réaliste en comparaison avec les résultats exagérément optimistes obtenus dans la sous-section précédente.

Prise en compte de symétries du problème par enrichissement des plans

Nous proposons ici de prendre en compte certaines invariances géométriques du problème étudié pour tirer au mieux parti des informations recueillies. On commence par observer que le k_{eff} associé à une configuration angulaire (t_1, t_2, t_3, t_4) n'a aucune raison d'être changé par une rotation d'angle $\frac{\pi}{2}$ du système : tourner d'un bloc le réseau de crayons n'influe pas sur les propriétés physiques de l'ensemble (cela peut être vu comme un changement de référentiel).

Il se trouve que lorsque cette rotation (dans l'espace usuel où « vit » matériellement le réseau de crayons) est d'angle un multiple de $\frac{\pi}{2}$, on retombe sur la configuration initiale modulo permutation des angles et ajout de $\frac{\pi}{2}$ aux quatre t_j ($j \in \{1, ..., 4\}$). On obtient en réitérant ce procédé un quadruplet de configurations angulaires "équivalentes", en ce

```
lm(formula = keff_s ~ I(abs(t1_s - 3 * pi/4)) + I(abs(t2_s -
    5 * pi/4)) + I(abs(t3_s - 5 * pi/4)) + I(abs(t4_s - 3 * pi/4)))

Residuals:
      Min        1Q    Median        3Q       Max
-4.915e-04 -9.785e-05  1.626e-05  9.082e-05  3.528e-04

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.080e-01  9.037e-05 1194.833  < 2e-16 ***
I(abs(t1_s - 3 * pi/4)) -5.780e-05  2.388e-05   -2.420  0.01961 *
I(abs(t2_s - 5 * pi/4))  6.648e-05  2.326e-05    2.858  0.00644 **
I(abs(t3_s - 5 * pi/4)) -6.175e-05  2.251e-05   -2.743  0.00872 **
I(abs(t4_s - 3 * pi/4))  7.982e-05  2.305e-05    3.462  0.00119 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm(formula = keff_s ~ I(cos(t1_s - 3 * pi/4)) + I(cos(t2_s -
    5 * pi/4)) + I(cos(t3_s - 5 * pi/4)) + I(cos(t4_s - 3 * pi/4)))

Residuals:
      Min        1Q    Median        3Q       Max
-4.282e-04 -7.977e-05  2.354e-06  1.194e-04  2.660e-04

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.080e-01  2.411e-05 4481.131  < 2e-16 ***
I(cos(t1_s - 3 * pi/4))  1.012e-04  3.593e-05    2.817 0.007183 **
I(cos(t2_s - 5 * pi/4)) -1.300e-04  3.444e-05   -3.775 0.000465 ***
I(cos(t3_s - 5 * pi/4))  1.038e-04  3.349e-05    3.099 0.003339 **
I(cos(t4_s - 3 * pi/4)) -1.306e-04  3.462e-05   -3.771 0.000471 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0001592 on 45 degrees of freedom
Multiple R-squared: 0.5471,     Adjusted R-squared: 0.5068
F-statistic: 13.59 on 4 and 45 DF,  p-value: 2.423e-07
```

FIG. 7.6 – Table d'ANOVA des modèles de régression linéaire *Vabs* et *trigo2* sur S

sens qu'elles correspondent exactement aux mêmes conditions physiques :

$$\begin{aligned} k_{eff}(t_1,t_2,t_3,t_4) &= k_{eff}\left(t_2+\frac{\pi}{2},t_4+\frac{\pi}{2},t_1+\frac{\pi}{2},t_3+\frac{\pi}{2}\right) \\ &= k_{eff}(t_4+\pi,t_3+\pi,t_2+\pi,t_1+\pi) \\ &= k_{eff}\left(t_3+\frac{3\pi}{2},t_1+\frac{3\pi}{2},t_4+\frac{3\pi}{2},t_2+\frac{3\pi}{2}\right) \end{aligned} \tag{7.3}$$

Comparatif des modèles par validation externe sur U et U_{sym}

Nom	Nature du plan	Taille	Fonction
C	Plan factoriel complet à 4 niveaux	256	Apprentissage
S	Plan aléatoire généré par processus de Strauss	50	Apprentissage
S_{sym}	Plan S "symétrisé"	200	Apprentissage
U	Plan aléatoire uniforme	100	Test
U_{sym}	Plan U "symétrisé"	400	Test

TAB. 7.1 – Liste des plans d'expériences considérés pour estimer les modèles de régression

Afin de quantifier l'intérêt de la prise en compte des symétries, on a ré-estimé les trois modèles GAM splines, régression VA, et régression cos avec le plan S_{sym}. Rappelons que ce plan qui contient 4 fois plus de points que S a été obtenu sans aucune simulation supplémentaire, mais simplement en exploitant certaines symétries [4]. On a ensuite calculé

[4]Comme on le verra dans la suite de cette étude, on peut aller plus loin en termes de symétries.

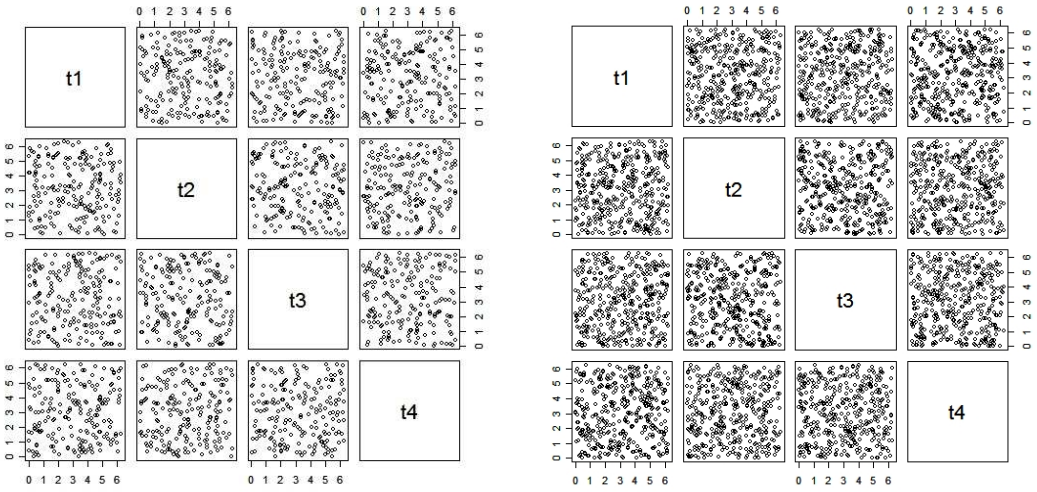


FIG. 7.7 – A gauche : plan S_{sym} obtenu à partir de S après prise en compte de l'invariance du k_{eff} par rotations du système d'angles multiples de $\frac{\pi}{2}$ (Cf. eq. (7.3) pour la traduction de ces rotations avec les variables du problème). A droite : plan U_{sym} obtenu à partir de U par le même procédé.

les critères de validation externe sur le plan U.

Par ailleurs, on peut aussi utiliser le plan symétrisé U_{sym} comme plan test. L'ensemble des modèles étudiés a donc aussi été testé en validation externe sur U_{sym}. Le récapitulatif des modèles et les résultats de validation sont présentés dans les tables 7.2.

La qualité des métamodèles peut être examinée par comparaison de la précision interne avec la précision externe. Ainsi, un modèle précis sur le plan d'apprentissage mais peu précis en validation externe peut être le signe d'un danger de sur-apprentissage (cas typique d'un interpolateur) ou de mauvais échantillonnage (comme dans le paragraphe précédent). Un modèle satisfaisant devrait donner des prévisions d'un niveau correct et d'un même ordre de grandeur sur les points d'apprentissage et les points test. Au vu des tables 7.2, on observe que les 3 premiers modèles trahissent une nette dégradation en validation externe, ce qui avait déjà été remarqué et qui est dû au choix du plan C.

Les autres modèles sont plutôt satisfaisants. Assez curieusement, le niveau d'erreur est plus petit sur les points de test qu'en apprentissage. Cela pourrait s'expliquer par la présence d'*outliers* dans les plans d'apprentissage. Enfin, on observe que les modèles

Modèle	Plan d'exp.	MSE int.	MSE (U)	MSE (U_{sym})
GAM splines	C	0.88×10^{-9}	14.47×10^{-9}	14.46×10^{-9}
Régression "VA"	C	0.91×10^{-9}	14.99×10^{-9}	14.99×10^{-9}
Régression "cos"	C	0.91×10^{-9}	13.33×10^{-9}	13.31×10^{-9}
GAM splines	S	16.46×10^{-9}	19.82×10^{-9}	17.71×10^{-9}
Régression "VA"	S	27.50×10^{-9}	15.91×10^{-9}	15.62×10^{-9}
Régression "cos"	S	22.82×10^{-9}	13.88×10^{-9}	13.52×10^{-9}
GAM splines	S_{sym}	21.62×10^{-9}	13.74×10^{-9}	14.02×10^{-9}
Régression "VA"	S_{sym}	25.97×10^{-9}	14.59×10^{-9}	14.77×10^{-9}
Régression "cos"	S_{sym}	23.16×10^{-9}	13.10×10^{-9}	13.10×10^{-9}

Modèle	Plan d'exp.	Var. expl. (sur U)	idem (sur U_{sym})
GAM splines	C	65.38 %	65.53 %
Régression "VA"	C	63.9 %	64.13 %
Régression "cos"	C	68.15 %	68.19 %
GAM splines	S	51.58 %	56.97 %
Régression "VA"	S	61.64 %	62.48%
Régression "cos"	S	66.33 %	67.17 %
GAM splines	S_{sym}	66.50 %	65.87 %
Régression "VA"	S_{sym}	64.36 %	63.99 %
Régression "cos"	S_{sym}	67.98 %	67.98 %

TAB. 7.2 – Comparaison des modèles par validation externe sur U et U_{sym}

estimés sur S_{sym} sont systématiquement meilleurs en test que ceux estimés sur S. ceci est particulièrement clair pour le modèle non-paramétrique GAM splines, pour lequel on passe d'un R^2 externe de 51.58% à 66.50% (MSE de 19.82 à 13.74) sur U et de 56.97 % à 65.87% sur U_{sym}.

Précisons avant de conclure plus formellement qu'il aurait était illusoire de viser un pourcentage de variance expliquée nettement plus important puisque l'écart-type d'estimation du k_{eff} par Monte-Carlo donné par le code Moret était de 10^{-4}, i.e. 10^{-8} de variance d'estimation. En considérant que la variance du k_{eff} estimée sur U vaut 4.12×10^{-8}, on obtient en effet un rapport signal-bruit d'environ $3.12/4.12 \approx 76\%$.

7.1.3 Quelques conclusions et problèmes généraux posés par l'étude

Ce premier cas d'application sur le simulateur MORET nous a permis de faire certains constats. Dans un premier temps, il est clairement apparu que le plan complet était un plan initial tout à fait inadapté dès lors que la réponse pouvait se décomposer de manière additive : en projection sur les axes canoniques, un plan complet 4×4 tel que C apparaît en effet comme un plan complet à 4 niveaux avec 64 répétitions par niveau, ce qui est loin d'être optimal pour l'estimation d'approximations monodimensionnelles (le même raisonnement tient pour une réponse additive par blocs). L'utilisation de plans d'expériences *Space Filling*, tels que présentés dans [Fra08], semble constituer une alternative pertinente aux plans classiques dans le cadre d'approximations non-linéaires.

Nous avons ensuite pu constater que la prise en compte de certaines symétries du problème permettait d'enrichir les plans d'expériences sans simulation supplémentaire, et autorisait ainsi la construction de bien meilleures approximations « à coût nul ». Ceci vaut en particulier en ce qui concerne les modèles non-paramétriques, tels qu'un modèle additif avec une spline de lissage dans chacune des directions. Remarquons à ce sujet que toutes les symétries présentes n'ont pas encore été prises en compte dans cette étude (il reste la périodicité des angles, ainsi que l'invariance du problème par retournement du réseau de crayons). Par ailleurs, la prise en compte du bruit de simulation a été ici peu discutée. C'est pour autant un enjeu de taille, et ce d'autant plus qu'il est possible avec un simulateur probabiliste tel que Moret de planifier des expériences à la fois dans l'espace des variables d'entrée et en termes de crédit de calcul.

Il devient alors nécessaire de développer des métamodèles à la fois

- capables de prendre en compte les invariances du problème traité, et
- capables de prendre en compte des observations bruitées de qualité hétérogène

Ces points sont ceux que nous nous proposons d'intégrer dans le cadre du Krigeage au cours des deux prochaines sections, respectivement sur les processus aléatoires à trajectoires invariantes, et sur le Krigeage avec bruit de simulation hétérogène.

7.2 Noyaux pour le krigeage de fonctions symétriques

7.2.1 Rappels sur les actions de groupe et les processus aléatoires

Nous donnons ici quelques définitions fondamentales qui nous servirons dans la prochaine section pour caractériser les processus aléatoires de réalisations invariantes, et du même coup pour adapter les outils du Krigeage à l'approximation de fonctions symétriques.

Un peu de vocabulaire des actions de groupe

Soient $(G, *)$ un groupe —simplement noté G dans la suite— et E un ensemble. On notera e l'élément neutre de G. Les définitions et exemples ci-dessous suivent d'assez près l'ouvrage pédagogique [Sch98] (on pourra aussi consulter par exemple les livres [Lan65] ou [Rob96] pour plus de détails techniques, en Anglais).

Définition. On rappelle qu'une *action (à gauche) du groupe G sur l'ensemble E* est une application $\Phi : (g, x) \in G \times E \longrightarrow g.x := \Phi(g, x) \in E$ telle que

- L'application $x \in E \longmapsto \Phi(e, x)$ est l'identité de E, i.e. $\forall x \in E,\ \Phi(e, x) = x$,
- $\forall x \in E,\ \forall g, g' \in G,\ \Phi(gg', x) = \Phi(g, \Phi(g', x))$.

Une telle action peut aussi être vue comme la donnée d'un homomorphisme $g \in G \longmapsto \Phi_g \in \mathcal{P}(E)$ entre G et le groupe $\mathcal{P}(E)$ des bijections de E. On montre en effet sans encombre (Cf. [Sch98]) d'une part que quel que soit $g \in G$, l'application $\Phi_g : x \in E \longmapsto \Phi(g, x)$ est bijective, d'inverse $\Phi_{g^{-1}}$, et d'autre part que $g \in G \longmapsto \Phi_g \in \mathcal{P}(E)$ est bien compatible avec les lois de groupes associées à G et $\mathcal{P}(E)$.

Exemples. L'application $\Phi^{trans} : (g, h) \in G \times G \longmapsto \Phi^{trans}(g, h) = gh \in G$ définit une action de groupe, appelée *action par translation à gauche de G sur lui-même*. Dans le cas où $(G, *)$ est la droite réelle munie de l'addition $(\mathbb{R}, +)$, les Φ^{trans}_x $(x \in \mathbb{R})$ sont les applications $y \in \mathbb{R} \longmapsto \Phi^{trans}_x(y) = x + y$. On s'intéresse souvent dans la pratique à l'action par translation sur $\mathbb{R}$ de sous-groupes de $(\mathbb{R}, +)$, tels que $(2\pi\mathbb{Z}, +)$; les fonctions 2π-périodiques peuvent ainsi être vues comme les fonctions invariantes sous une telle action (Cf. ci-dessous pour plus de détails sur les fonctions invariantes par action de groupe). Pour donner un autre exemple élémentaire de famille d'actions de groupe, on peut faire agir $(\mathbb{Z}/2\mathbb{Z}, +)$ sur l'ensemble $\mathbb{R}^2$ en associant à $\bar{1}$ la symétrie par rapport à la première bissectrice (droite d'équation $x_2 = x_1$), ou encore $(\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, +)$ sur $\mathbb{R}^2$ en associant à $(\bar{1}, \bar{0})$ la symétrie par rapport à la première bissectrice, et à $(\bar{0}, \bar{1})$ celle par rapport à la seconde bissectrice (droite d'équation $x_2 = -x_1$). Nous noterons dans la suite ces deux actions Φ^{sym1} et Φ^{sym2}.

Revenons au cas général d'un groupe G agissant sur un ensemble E via une action Φ (où $\Phi(g,x)$ est noté $g.x$ lorsqu'il n'y a pas de confusion possible), et notons $x \in E$ et $A \subset E$ respectivement un élément et une partie arbitraires de l'ensemble E. Nous allons maintenant définir les notions d'orbite, de stabilisateur, et de fixateur, en suivant de nouveau la présentation faite dans [Sch98].

Définitions. L'*orbite* d'un point $x \in E$ par l'action Φ est l'ensemble

$$\mathcal{O}(x) := \{g.x,\ g \in G\},$$

formé des images de x par l'action de G[†]. On dit que x est un point fixe de l'action lorsque $\forall g \in G,\ g.x = x$. Le *fixateur* de $A \subset E$ dans G est défini par $Fix_\Phi(x) := \{g \in G \mid \forall a \in A,\ g.a = a\}$, et le stabilisateur de A par $Stab_\Phi(A) := \{g \in G \mid \forall a \in A,\ g.a \in A\}$.

Remarquons que de manière tout à fait générale, les stabilisateurs et fixateurs d'une partie sont des sous-groupes de G. De plus, le fixateur est clairement inclus dans le stabilisateur, et il n'y a pas nécessairement égalité. Les deux notions coïncident en revanche dans le cas où la partie A est réduite à un élément $\{a\}$, ce qui sera le cas pour l'usage que nous ferons de ces notions dans la partie principale de cette section.

Les orbites jouent un rôle important dans l'étude des actions de groupes[5]. En remarquant que l'appartenance à une même orbite constitue une relation d'équivalence sur E, on obtient une partition de E selon les orbites de Φ, i.e. une réunion de parties disjointes recouvrant E. Il est ainsi possible de choisir un élément dans chacune de ces orbites — modulo une utilisation éventuelle de l'axiome du choix dans les cas où le nombre d'orbites est infini— et d'obtenir une partie $A \subset E$ de cardinalité minimale telle que $G.A = E$. Suivant la terminologie classiquement employée en géométrie, nous appelerons une telle partie A un *domaine fondamental* pour l'action de G sur E.

Retour sur les exemples. Considérons de nouveau l'action par translation Φ^{trans}. Dans le cas où $(\mathbb{R},+)$ agit sur $(\mathbb{R},+)$, l'orbite d'un point $x \in \mathbb{R}$ est $\mathbb{R}$ tout entier, et $\{x\}$ constitue un domaine fondamental de l'action. On dit, lorsqu'il n'y a comme dans ce cas qu'une seule orbite, que l'action est *transitive*. Si l'on restreint maintenant l'action Φ^{trans} au sous-groupe $(2\pi\mathbb{Z},+)$, l'orbite de tout élément $x \in \mathbb{R}$ est $\mathcal{O}(x) = \{x + 2k\pi,\ k \in \mathbb{Z}\}$, et tout intervalle de la forme $[a, a + 2\pi[$ $(a \in \mathbb{R})$ constitue un domaine fondamental

[†] autrement dit $\Phi(G,x)$, ou encore $G.x$.

[5] et par ce biais dans l'étude des groupes, en particulier en considérant l'action par conjugaison d'un groupe sur lui-même (Cf. par exemple [Sch98, Lan65, Rob96] pour des approfondissements sur ce sujet).

connexe de l'action. Pour finir, l'orbite d'un point $x = (x_1, x_2) \in \mathbb{R}^2$ par l'action Φ^{sym1} est donnée par $\mathcal{O}(x) = \{(x_1, x_2),\ (x_2, x_1)\}$, qui a un ou deux éléments selon que $x_1 = x_2$ ou $x_1 \neq x_2$. Le stabilisateur d'un point $x = (x_1, x_2) \in \mathbb{R}^2$ est ainsi $\mathbb{Z}/2\mathbb{Z}$ si x est sur la première bissectrice, et réduit à l'élément neutre $\{\bar{0}\}$ si x en est en dehors. Un domaine fondamental connexe est donné par $A := A_1 \cup A_2$, où $A_1 = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid x_1 < x_2\}$ (demi-plan ouvert) et $A_2 = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid x_1 = x_2\}$ (première bissectrice).

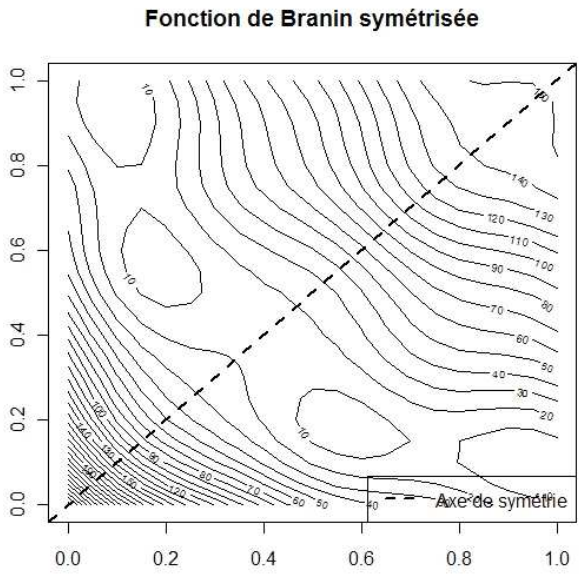


FIG. 7.8 – Fonction de Branin symétrisée par rapport à la première bissectrice (y_{BHS})

Définition. Soit F un ensemble quelconque. On dit qu'une application $f : E \longrightarrow F$ est *invariante par* Φ, ou encore *invariante sous l'action du groupe* G, lorsque

$$\forall x \in E,\ \forall g \in G,\ f(g.x) = f(x).$$

Cela revient à dire que f est constante sur les orbites de Φ.

La figure 7.8 représente une fonction invariante par Φ^{sym1}. Il s'agit de la fonction de Branin y_{BH} (Cf. chaps. 3,4,8,9) symétrisée, c'est à dire de l'application

$$x \in \mathbb{R}^2 \longrightarrow y_{BHS}(x) := \frac{1}{2}\left[y_{BH}(x) + y_{BH}(s(x))\right] = \frac{1}{2}\left[y_{BH}(x_1, x_2) + y_{BH}(x_2, x_1)\right] \in \mathbb{R},$$

où s est la symétrie par rapport à A_2. Il est remarquable que la fonction y_{BHS}, a l'instar de toute fonction invariante, peut en fait s'écrire complètement à partir de sa restriction à un domaine fondamental (ici l'ensemble A détaillé ci-dessus). On a en effet

$$\forall x \in \mathbb{R}^2,\ y_{BHS}(x) = \frac{f(x) + f(s.x)}{1 + \mathbb{1}_{A_2}(x)}, \text{où } f(x) = y_{BHS}(x)\mathbb{1}_A(x)$$

Cette décomposition, bien que pouvant apparaître un peu artificielle en première lecture, va nous servir pour traiter du sujet des processus aléatoires de réalisations invariantes.

Un peu de vocabulaire des processus aléatoires

Nous reprenons ici les définitions de l'ouvrage [RY91], avec quelques changement de notations mineurs. Pour tous les processus considérés, l'espace des temps —le plus souvent multidimensionnel ici— est noté D.

Définition. Deux processus aléatoires Y et Y' définis respectivement sur des espaces probabilisés $(\Omega, \mathcal{F}, \mathbb{P})$ et $(\Omega', \mathcal{F}', \mathbb{P}')$, et possédant le même espace d'états $(E, \mathcal{E})$, sont dits *équivalents* si pour toutes suites finies $x^1, \ldots, x^n \in D$ et $A_1, \ldots, A_n \in \mathcal{E}$,

$$\mathbb{P}\left(Y_{x^1} \in A_1, \ldots, Y_{x^n} \in A_n\right) = \mathbb{P}'\left(Y'_{x^1} \in A_1, \ldots, Y'_{x^n} \in A_n\right) \tag{7.4}$$

On dit aussi dans ce cas que chacun des processus est une *version* de l'autre, ou encore qu'ils sont des *versions* du même processus. Autrement dit, deux processus sont des versions l'un de l'autre s'ils ont les mêmes lois finies-dimensionnelles.

Définition. Deux processus Y et Y' définis sur le même espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ sont dits être des *modifications* l'un de l'autre lorsque pour tout $x \in D$,

$$Y_x = Y'_x \ \ \mathbb{P}\text{-p.s.} \tag{7.5}$$

Ils sont dits *indistinguables* lorsque pour $\mathbb{P}$-presque tout $\omega \in \Omega$,

$$\forall x \in D,\ Y_x(\omega) = Y'_x(\omega) \tag{7.6}$$

Comme précisé dans ([RY91], p. 18), si Y et Y' sont des modifications l'un de l'autre, ils sont clairement des versions du même processus. Un résultat moins trivial est que si deux processus modifications l'un de l'autre sont continus p.s., alors ils sont indistinguables.

Exemples. soient B^1 et B^2 deux Mouvements Browniens sur $[0, 1]$, X une v.a.r. de loi uniforme sur $[0, 1]$, et α une v.a.r. valant 1 sur une partie de Ω de mesure nulle et 0

sinon, où les processus et les v.a.r. sont indépendants dans leur ensemble. Il est clair que B^1 et B^2 sont équivalents mais n'ont aucune raison a prori d'être des modifications l'un de l'autre. En revanche, B^1 et $B^1 + \mathbb{1}_X$ sont bien des modifications l'un de l'autre, mais ne sont pas indistinguables. Enfin, B^1 et $B^1 + \alpha\mathbb{1}_{\mathbb{R}^+}$ sont indistinguables.

7.2.2 Noyaux de covariance et processus aléatoires de réalisations invariantes sous l'action d'un groupe fini

Préliminaire sur la relation entre noyaux et invariances : il semble important de préciser d'emblée que le travail qui va suivre ne cherche pas *a priori* à répondre à la question « quelles sont les propriétés d'un processus dont le noyau de covariance présente certaines invariances donnés ? » mais plutôt en quelque sorte au problème inverse « quelles propriétés du noyau de covariance caractérisent (si possible) l'invariance des réalisations d'un processus par une action de groupe donné ? ». La première question est relativement ancienne dans l'étude des processus aléatoires spatiaux, et rassemble déjà bon nombre de résultats classiques (Cf. [Abr97, Ste99]). Ainsi l'invariance d'un noyau $k : (x, y) \in (\mathbb{R}^d)^2 \longrightarrow k(x, y) \in \mathbb{R}$ par action du groupe des *translations* de type $\tau_h : (x, y) \in (\mathbb{R}^d)^2 \longrightarrow (x+h, y+h) \in (\mathbb{R}^d)^2$ correspond-elle à la *stationnarité* d'un processus gaussien centré, ou encore l'invariance d'un noyau stationnaire $k : h \in \mathbb{R}^d \longrightarrow k(h) \in \mathbb{R}$ par *rotation* (i.e. précisément par action du groupe des isométries de $\mathbb{R}^d$) exprime-t-elle la notion d'*isotropie* (Cf. [Ste99], pp. 16-17). Nous portons ci-dessous notre attention sur la deuxième question, objet —à notre surprise, et d'après nos recherches bibliographiques— de beaucoup moins de travaux aujourd'hui aboutis que la première.

Nous considérons dans ce qui suit un ensemble $D \subset \mathbb{R}^d$ muni de sa tribu borélienne $\mathcal{B}(D)$, $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, $Y = (Y_x)_{x\in D}$ un processus aléatoire centré de carré intégrable ($\forall x \in D,\ Y_x \in L^2(\mathbb{P})$), de noyau de covariance noté $k_Y : D \times D \longrightarrow \mathbb{R}$, et G un groupe fini d'ordre $r \in \mathbb{N}^*$ agissant sur D via une action quelconque $\Phi : (g, x) \in G \times D \longrightarrow \Phi(g, x) := g.x \in D$.

Définition. On dit que Y a toutes ses réalisations invariantes sous l'action de G lorsque

$$\forall \omega \in \Omega,\ \forall x \in D,\ \forall g \in G,\ Y_x(\omega) = Y_{g.x}(\omega) \tag{7.7}$$

Propriété. Si Y a toutes ses réalisations invariantes sous l'action de G, alors k_Y satisfait

$$\forall x, x' \in D,\ k_Y(x, x') = \frac{1}{r^2} \sum_{(g,g')\in G^2} k_Y(g.x, g'.x') \tag{7.8}$$

Démonstration. Par définition du noyau de covariance k_Y et par invariance de Y_x sous l'action de G, on a que

$$\begin{aligned}\sum_{(g,g')\in G^2} k_Y(g.x, g'.x') &= \sum_{(g,g')\in G^2} Cov[Y_{g.x}, Y_{g'.x'}] \\ &= \sum_{(g,g')\in G^2} Cov[Y_x, Y_{x'}] = r^2 k_Y(x,x').\end{aligned}$$

Remarque. Cette décomposition de k_Y en double somme indexée par les éléments de G est loin d'être unique. Soit en effet $\mathcal{R}$ la relation d'équivalence sur D qui lie deux éléments appartenant à une même orbite de Φ et $\pi : D \longrightarrow D/\mathcal{R}$ la surjection canonique associée. Soit $A \subset D$ un système mesurable de représentants des orbites de Φ ; Y peut alors s'écrire comme le symétrisé d'un certain processus Z :

$$\forall x \in D,\ Y_x = \sum_{g\in G} Y_x \frac{\mathbb{1}_{g.A}(x)}{\#Stab_\Phi(x)} = \sum_{g\in G} Y_{g^{-1}.x} \frac{\mathbb{1}_A(g^{-1}.x)}{\#Stab_\Phi(g^{-1}.x)} = \sum_{g\in G} Z_{g.x} \tag{7.9}$$

où $\#Stab_\Phi(x)$ est l'ordre du sous-groupe stabilisateur de x par Φ, et $Z_x := Y_x \frac{\mathbb{1}_A(x)}{\#Stab_\Phi(x)}$. Il vient en particulier, en notant k_Z le noyau de covariance de Z, que

$$\forall x, x' \in E,\ k_Y(x,x') = Cov\left[\sum_{g\in G}^{m} Z_{g.x}, \sum_{g'\in G}^{m} Z_{g'.x'}\right] = \sum_{(g,g')\in G^2} k_Z(g.x, g'.x') \tag{7.10}$$

Corollaire. Si Y a toutes ses réalisations invariantes sous l'action de G, alors il existe un noyau de type positif $\overline{k_Y} : (D/\mathcal{R})^2 \longrightarrow \mathbb{R}$ tel que k_Y s'écrive :

$$\forall x, x' \in D,\ k_Y(x,x') = \overline{k_Y}(\pi(x), \pi(x')) \tag{7.11}$$

Démonstration. En notant $R_A : D/\mathcal{R} \longrightarrow A$ la bijection entre les orbites de Φ et le système de représentants A, on introduit $\overline{k_Y} : (o_1, o_2) \in (D/\mathcal{R})^2 \longrightarrow k_Y(R_A(o_1), R_A(o_2))$. Il est clair d'une part que $\overline{k_Y}$ est de type positif sur $(D/\mathcal{R})^2$, puisque $\forall (o_1, ..., o_n) \in (D/\mathcal{R})^n$ la matrice $\left(\overline{k_Y}(o_i, o_j)\right)_{1\leq i,j\leq n}$ est par construction égale à la matrice semi-définie positive $K := \left(k_Y(R_A(o_i), R_A(o_j))\right)_{1\leq i,j\leq n}$. D'autre part, pour tout $x \in D$ donné, il existe $g, g' \in G$ tels que $R_A(\pi(x)) = g.x$ et $R_A(\pi(x')) = g'.x$. Il vient finalement que $k_Y(x,x') = k_Y(g.x, g'.x') = k_Y(R_A(\pi(x)), R_A(\pi(x'))) = \overline{k_Y}(\pi(x), \pi(x'))$.

Un premier exemple. Soit X un processus gaussien centré sur $\mathbb{R}$ de noyau $k_X : x, x' \in \mathbb{R} \longrightarrow k_X(x,x') = e^{-|x-x'|} \in \mathbb{R}$ (processus d'Ornstein-Uhlenbeck, Cf. [RY91] sec. 1.3), et $\Phi : (g,x) \in (\mathbb{Z}/2\mathbb{Z}) \times \mathbb{R} \longrightarrow \mathbb{R}$ l'action qui a $(\overline{1}, x)$ associe $-x$ (la symétrie par rapport à l'origine). Le processus Y obtenu par symétrisation de la restriction de X à

$A = [0, +\infty[$, défini par $Y_x = \frac{1}{1+\mathbb{1}_{\{0\}}(x)} X_x \mathbb{1}_{[0,+\infty[}(x) + \frac{1}{1+\mathbb{1}_{\{0\}}(x)} X_x \mathbb{1}_{[0,+\infty[}(-x)$, a toutes ses réalisations invariantes sous l'action de G. Son noyau de covariance est donné par $\forall x, x' \in \mathbb{R},\ k_Y(x, x') = e^{-||x|-|x'||}$. Remarquons que le processus Y, symétrisé du processus stationnaire X, n'est évidemment pas stationnaire à l'ordre 2.

Un second exemple. Soit X un processus gaussien centré sur $\mathbb{R}^2$, de noyau $k_X : x, x' \in \mathbb{R}^2 \longrightarrow k_X(x, x') = e^{-||x-x'||^2} \in \mathbb{R}$, et $\Phi : (g, x) \in (\mathbb{Z}/2\mathbb{Z}) \times \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ l'action qui à $(\overline{1}, x)$ associe $s(x) = (x_2, x_1)$, le symétrique de $x = (x_1, x_2)$ par rapport à la première bissectrice (l'action Φ^{sym1} vue plus haut). Le processus Y obtenu par symétrisation de la restriction de X à $A = \{x \in \mathbb{R}^2 : x_1 \leq x_2\}$ est défini par $Y_x = \frac{1}{1+\mathbb{1}_{\{x\in\mathbb{R}^2 : s(x)=x\}}(x)} X_x \mathbb{1}_A(x) + \frac{1}{1+\mathbb{1}_{\{x\in\mathbb{R}^2 : s(x)=x\}}(x)} X_x \mathbb{1}_A(s(x))$. Remarquons en particulier que Y conserve la régularité de X en dehors de l'ensemble $\{x \in \mathbb{R}^2 : s(x) = x\}$.

Supposons maintenant qu'un processus centré Y —gaussien ou non— possède un noyau de covariance s'écrivant, à l'instar de k_Y dans l'eq. (7.10), comme la somme sur $G \times G$ d'un noyau de type positif quelconque. Nous allons montrer ci-dessous qu'il est alors possible de remonter aux propriétés d'invariance par Φ des réalisations de Y.

Théorème. *Soit Y un processus aléatoire centré de carré intégrable possédant un noyau de covariance de la forme $k_Y(x, x') = \sum_{(g,g')\in G^2} k_Z(g.x, g'.x')$, où $k_Z : E \times E \longrightarrow \mathbb{R}$ est un noyau de type positif . Y est alors équivalent à un processus de réalisations invariantes par Φ. Il existe de plus une modification de Y dont toutes les réalisations sont invariantes par Φ.*

Démonstration. Considérons dans un premier temps un processus gaussien Z centré de noyau k_Z, et Z^Φ le processus défini par $\forall x \in D,\ Z^\Phi_x = \sum_{g\in G} Z_{g.x}$. Il est clair que Z^Φ est aussi un processus gaussien centré, et que ses réalisations sont toutes invariantes sous l'action de G. Par ailleurs, $\forall x, x' \in E,\ Cov[Z^\Phi_x, Z^\Phi_{x'}] = Cov[\sum_{g\in G} Z_{g.x}, \sum_{g'\in G} Z_{g'.x}] = \sum_{(g,g')\in G^2} k_Z(g.x, g'.x') = k_Y(x, x')$. Comme Y et Z^Φ sont deux processus gaussiens centrés de même noyau de covariance, ils sont de même loi, et on a en particulier

$$P(Y_{x_1} \in B_1, ..., Y_{x_m} \in B_m) = P(Z^\Phi_{x_1} \in B_1, ..., Z^\Phi_{x_m} \in B_m)$$

quels que soient $m \in \mathbb{N}$ et $\{B_i \in \mathcal{B}(\mathbb{R}), i \in [1, m]\}$, c'est à dire que Y et Z^Φ sont bien équivalents. On peut en fait donner un résultat en termes de modifications, plus fort que ce qui vient d'être montré en termes de versions, et sans même utiliser d'hypothèse gaussienne concernant Z. En reprenant les notations de la démonstration précédente,

introduisons maintenant le processus $\tilde{Y}$ défini comme suit : $\forall x \in D,\ \tilde{Y}_x = Y_{R_A(\pi(x))}$. Par construction, $\tilde{Y}$ a toutes ses réalisations invariantes par Φ. Enfin, pour tout $x \in D$, il existe $g \in G$ tel que $R_A(\pi(x)) = g.x$, et il vient que

$$Var[Y_x - \tilde{Y}_x] = Var[Y_x - Y_{g.x}] = k_Y(x,x) + k_Y(g.x, g.x) - 2k_Y(x, g.x) = 0.$$

On a ainsi $\forall x \in D,\ P(Y_x = \hat{Y}_x) = 1$ et $\tilde{Y}$ est bien une modification de Y.

Remarque. En pratique, les réalisations de Y sont souvent simulées sur une partie finie $S = \{x_1, ..., x_m\}$ d'éléments de E en se basant sur une décomposition (Cholesky, Mahalanobis) de la matrice de covariance $K = (k_Y(x_i, x_j))_{1 \leq i,j \leq m}$. L' invariance par Φ des vecteurs ainsi obtenus est donc certaine (i.e. $\forall \omega \in \Omega$).

L'exemple suivant illustre le fait qu'un processus Y dont presque toutes les réalisations sont non invariantes par Φ peut posséder une modification de réalisations toutes invariantes par Φ.

Exemple. Soient $\Omega =]0,1[,\ \mathcal{A} = \mathcal{B}(]0,1[)$, $\mathbb{P}$ la mesure de Lebesgue sur Ω, $D = \mathbb{R}$, $G = \{e, s_0\}$ (s_0 symétrie par rapport à 0), $F : x \in \mathbb{R} \longrightarrow \int_{-\infty}^{x} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} du \in]0,1[$, $\varepsilon : \omega \in \Omega \longrightarrow \varepsilon(\omega) = F^{-1}(\omega) \in \mathbb{R}$, et $Y : (x, \omega) \in E \times \Omega \longrightarrow Y_x(\omega) = |x|\varepsilon(\omega)\mathbb{1}_{x \neq \varepsilon(\omega)}$. Le processus défini par $\tilde{Y}_x(\omega) = |x|\varepsilon(\omega)$ a clairement toutes ses réalisations invariantes par la symétrie s_0, et on a bien que $\tilde{Y}$ est une modification de Y puisque pour tout $x \in D$, $P(Y_x = \tilde{Y}_x) = P(\varepsilon \neq x) = 1$. En revanche, $\left\{\omega \in \Omega \ / \ (\forall x \in D,\ Y_x(\omega) = \tilde{Y}_x(\omega))\right\} = \left\{\frac{1}{2}\right\}$ est négligeable, et les deux processus ne sont donc pas indistinguables.

Pour finir, l'exemple suivant illustre la possibilité de construire une classe particulière de processus invariants en symétrisant des processus stationnaires.

Exemple. Reprenons les notations du deuxième exemple de processus symétrisé. On peut construire un processus de réalisations invariantes par Φ sur la base du processus stationnaire X en posant $\forall x \in D,\ X_x^{\Phi} = \frac{1}{2}(X_x + X_{s(x)}) = \frac{1}{2}(X_{(x_1,x_2)} + X_{(x_2,x_1)})$. Le noyau de covariance du processus gaussien X^{Φ} ainsi construit est donné par

$$\begin{aligned} k_{X^{\Phi}}(x, x') &= \frac{1}{4}[k_X(x - x') + k_X(s(x) - x') + k_X(x - s(x')) + k_X(s(x) - s(x'))] \\ &= \frac{1}{4}e^{-||(x_1 - x'_1, x_2 - x'_2)||^2} + \frac{1}{4}e^{-||(x_2 - x'_1, x_1 - x'_2)||^2} \\ &+ \frac{1}{4}e^{-||(x_1 - x'_2, x_2 - x'_1)||^2} + \frac{1}{4}e^{-||(x_2 - x'_2, x_1 - x'_1)||^2} \end{aligned} \tag{7.12}$$

Remarquons qu'avec ce procédé de construction, le processus symétrisé X^{Φ} conserve la régularité de X, y compris sur l'axe de symétrie $\{x \in \mathbb{R}^2 : s(x) = x\}$.

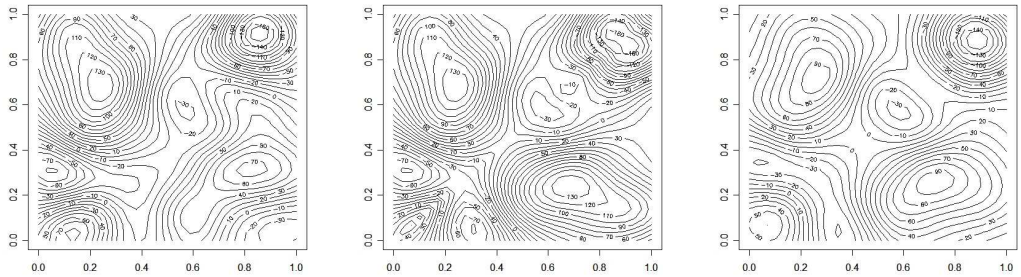


FIG. 7.9 – A gauche : une réalisation simulée sur une grille 30×30 d'un processus gaussien réel centré X de covariance $k_X(x, x') = \sigma^2 e^{-\theta_1(x_1 - x'_1)^2 - \theta_2(x_2 - x'_2)^2}$ (ici $\sigma^2 = 10,\ \theta_1 = \theta_2 = 10$). Au centre : la symétrisée de la restriction à $\{x \in [0,1]^2 : x_1 \leq x_2\}$ de la réalisation de gauche (Exemple 2). A droite : une réalisation simulée (en utilisant le même aléa que précédemment, i.e. en transformant la même réalisation d'un vecteur gaussien centré réduit) d'un processus gaussien stationnaire symétrisé Y (Exemple 4), de covariance construite à partir de k_X selon le procédé de l'eq. (7.12).

7.2.3 Application : Krigeage avec noyau symétrisé

Nous présentons maintenant les résultats d'une toute première expérience illustrant le Krigeage de fonctions symétriques. L'application étudiée est la fonction de Branin symétrisé y_{BHS} introduite ci-avant, et les techniques comparées s'inspirent à la fois des méthodes mises en oeuvre dans la première section de ce chapitre sur l'approximation du code Moret, et des résultats théoriques sur les noyaux des processus à réalisations symétriques présentés dans la section précédente.

Le point de départ est le suivant : on observe y_{BHS} sur un plan $\mathbf{X}$ à 9 points obtenu par tirages indépendants selon la loi uniforme sur le carré unité, et on essaye de reconstruire la fonction par cinq méthodes différentes. Les surfaces de réponse obtenues sont alors évaluées sur une grille régulière $\mathbf{T}$ à 21×21 éléments, et on calcule pour chacune d'entre elles l'erreur quadratique moyenne d'approximation

$$EQM_i := \frac{1}{441} \sum_{x \in \mathbf{T}} (y_{BHS}(x) - m_i(x))^2 \quad (1 \leq i \leq 5),$$

où les m_i sont les moyennes de Krigeage respectives. Ces dernières sont conçues selon les procédés ci-après :

1. m_1 correspond à un krigeage ordinaire de y_{BHS} avec noyau gaussien anisotrope, sur la base du plan d'expériences $\mathbf{X}$ à 9 points. Les paramètres de covariance sont estimés à partir des observations faites en $\mathbf{X}$.

2. m_2 correspond à un krigeage ordinaire de y_{BHS} avec noyau gaussien anisotrope, sur la base du plan d'expériences $\mathbf{X}_{sym}$ à 18 points, symétrisé de $\mathbf{X}$ par rapport à la première bissectrice. Les paramètres de covariance sont estimés à partir des observations faites en $\mathbf{X}_{sym}$.

3. m_3 correspond à un krigeage ordinaire de y_{BHS} avec noyau gaussien anisotrope, fait sur la base de la restriction $\mathbf{X}'_{sym}$ du plan d'expériences $\mathbf{X}_{sym}$ à un demi-plan (domaine fondamental de Φ^{sym1}), et resymétrisé a posteriori (les points de $\mathbf{T}$ sont si besoin symétrisés à l'étape des prédictions par Krigeage). Les paramètres de covariance sont estimés à partir des observations faites en $\mathbf{X}'_{sym}$.

4. m_4 correspond à un krigeage ordinaire de y_{BHS} avec le noyau Beta, sur la base du plan d'expériences $\mathbf{X}$ à 9 points. Le noyau Beta est exactement le noyau somme de l'équation 7.12 ci-dessus, dont les paramètres de covariance sont estimés à partir des observations faites en $\mathbf{X}$ (fixées pour le moment aux valeurs obtenues lors de la construction de m_1).

5. m_5 correspond à un krigeage ordinaire de y_{BHS} avec le noyau Alpha, sur la base du plan d'expériences $\mathbf{X}$ à 9 points. Le noyau Alpha est une composition du noyau gaussien anisotrope k_g avec le projecteur $R_{A_1^c}$ sur le système de représentants $A_1^c = \{x = (x_1, x_2) \mid x_1 \geq x_2\}$, i.e. $\forall x, x' \in [0,1]^2,\ k_\alpha(x, x') = k_g(R_{A_1^c}(x), R_{A_1^c}(x'))$. Les paramètres de covariance sont ici aussi fixés —pour des raisons pratiques temporaires— aux valeurs obtenues lors de la construction de m_1.

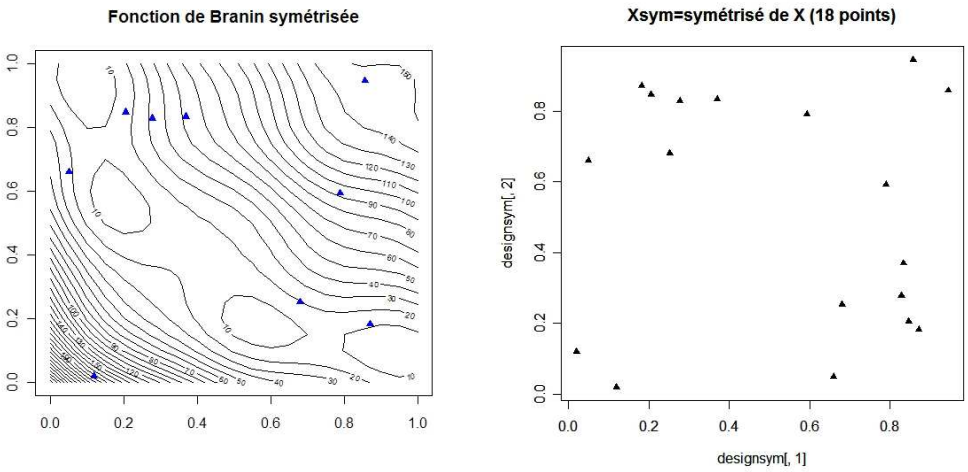


FIG. 7.10 – Fonction de Branin symétrisée avec le plan $\mathbf{X}$ à 9 points, puis le plan $\mathbf{X}_{sym}$ symétrisé de $\mathbf{X}$ par rapport au même axe que la fonction de Branin symétrisée

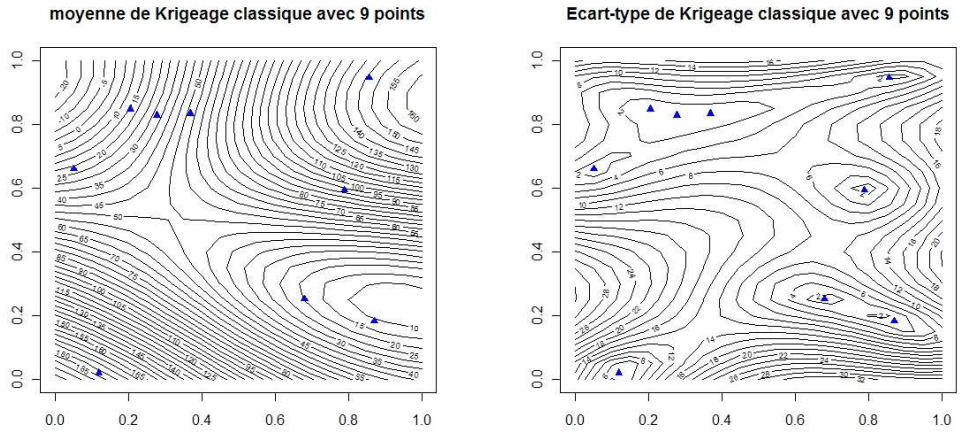


FIG. 7.11 – Krigeage sur la base du plan $\mathbf{X}$ de Branin symétrisée (m_1), avec covariance gaussienne. Erreur quadratique moyenne (EQM_1) sur le plan test 21×21 : **820.93**

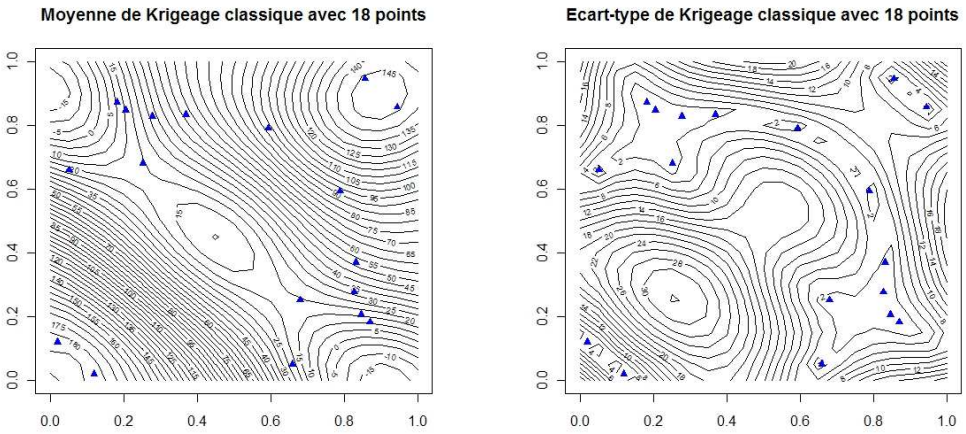


FIG. 7.12 – Krigeage sur la base du plan $\mathbf{X}_{sym}$ de Branin symétrisée (m_2), avec covariance gaussienne. EQM_2 : **694.11**

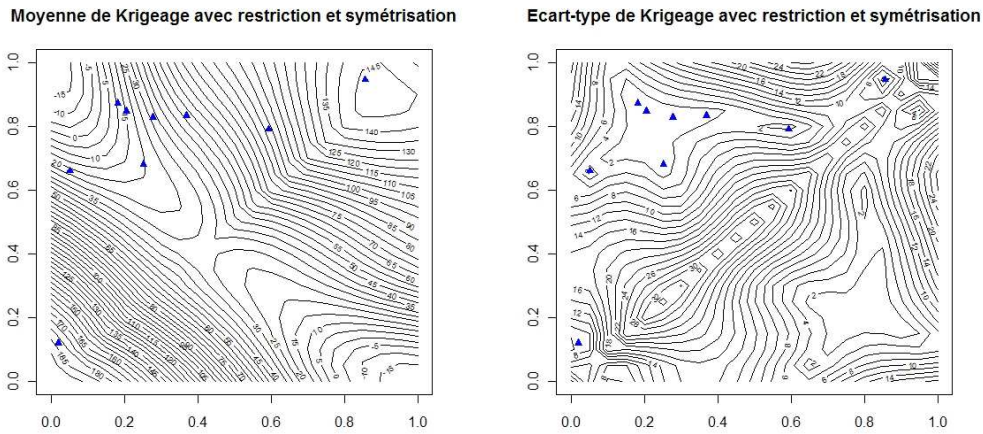


FIG. 7.13 – Krigeage sur la base du projeté du plan $\mathbf{X}$ sur $\{x_1 \geq x_2\}$ de Branin symétrisée (m_3), avec restriction et symétrisation. EQM_3 : **544.83**

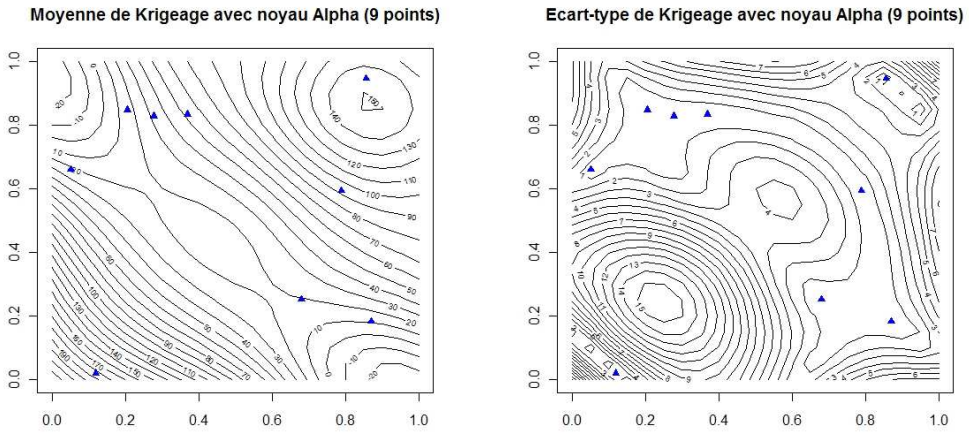


FIG. 7.14 – Krigeage sur la base du plan **X** de Branin symétrisée, avec covariance gaussienne somme (m_4). EQM_4 : **330.21**

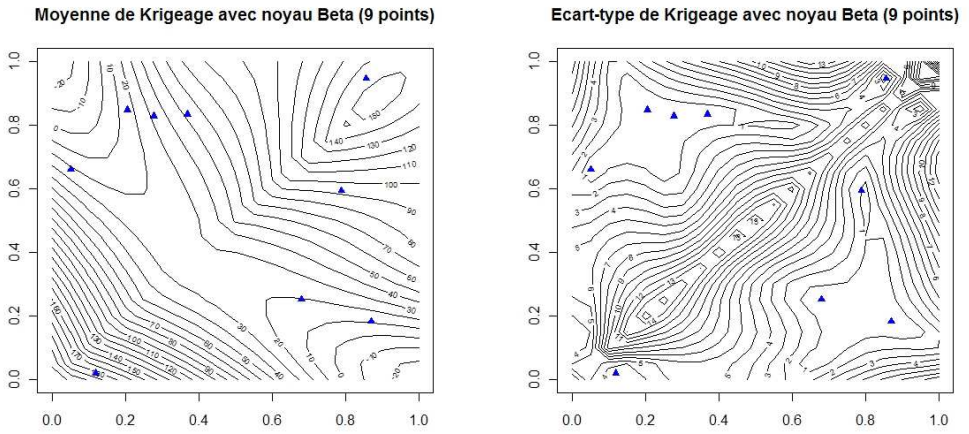


FIG. 7.15 – Krigeage sur la base du plan **X** de Branin symétrisée, avec covariance gaussienne en projection sur le domaine fondamental $\{x_1 \leq x_2\}$ (m_5). EQM_5 : **297.61**

Modèle	Symétries	Plan	Spécificités	EQM
m_1	Non	$\mathbf{X}$	noyau gaussien anisotrope	820.93
m_2	Oui	$\mathbf{X}_{sym}$	—	694.11
m_3	Oui	$P(\mathbf{X})$	P : projection sur un domaine fondamental	544.83
m_4	Oui	$\mathbf{X}$	Noyau somme sur les orbites de Φ	330.31
m_5	Oui	$\mathbf{X}$	Noyau incluant P	297.61

TAB. 7.3 – Comparaison des modèles de Krigeage pour l'approximation de y_{BHS}

Ce premier essai est tout à fait encourageant pour le développement et l'utilisation de modèles de Krigeage incorporant des symétries, et plus généralement des invariances par action de groupe. Si une symétrisation du plan d'expériences ($\mathbf{X}_{sym}$ Cf. 7.12) a pu ici sensiblement améliorer la précision des prédictions comparées à celles d'un Krigeage construit sur la base du plan $\mathbf{X}$ (Cf. 7.11), les résultats résumés dans le tableau 7.3 indiquent que les méthodes proposées dans cette section permettent de tirer encore davantage profit de la connaissance *a priori* concernant les invariances de la fonction étudiée. Malgré certaines similarités apparaissant respectivement entre 7.12 & 7.14 et 7.13 & 7.2.3, l'approche par symétrisation du noyau semble être porteuse d'informations additionnelles sur la fonction invariante étudiée, notamment ici en ce qui concerne son comportement autour de l'axe de symétrie. Remarquons au sujet de la conception du noyau Alpha, composition d'un noyau stationnaire avec un projecteur, que le domaine fondamental A_1^c choisi n'est pas arbitraire : simuler un processus anisotrope sur A_1^c puis symétriser n'est pas équivalent à faire de même en se basant sur $\overline{A_1}$; en quelque sorte, l'anisotropie et la symétrie ne commutent pas ! Ce deuxième noyau —expérience non représentée ici— donnerait en fait un EQM de 753.45. Cela n'empêche bien entendu nullement les deux noyaux d'être invariants à gauche et à droite par l'action Φ^{sym1}.

Pour revenir sur le bilan général de l'approximation de la fonction de Branin symétrisée par Krigeage, il ne faut pas oublier qu'il s'agit ici d'une première illustration, et que les résultats positifs obtenus mériteraient d'être confirmés (et étendus) dans le cadre d'applications ultérieures. Le simulateur MORET s'y prête particulièrement bien, et une étude est en cours pour définir un noyau permettant de prendre en compte les symétries du k_{eff}, i.e. à la fois sa 2π-périodicité en chacune des variables, et son invariance par rotations d'angles $\{k\frac{\pi}{2},\ 1 \leq k \leq 3\}$ et retournement du système de crayons —non pris en compte au début de ce chapitre. Ces huit dernières invariances sont en fait directement reliées à une action du groupe diédral D_8, dit aussi *groupe du carré* (groupe des transformations du plan laissant un carré invariant), sur l' hypercube $[0, 2\pi]^4$.

7.3 Prise en compte d'un bruit de simulation hétérogène

7.3.1 Simulations stochastiques avec fidélité réglable

Comme nous l'avons évoqué dans au début de la première section de ce chapitre, un appel au simulateur Moret pour calculer le k_{eff} correspondant à une configuration donnée des variables d'entrée ne rend pas un nombre, mais un couple formé d'une valeur estimée et d'un écart-type d'estimation. Ils résument en fait l'estimation statistique du k_{eff} par une méthode de Monte-Carlo, délivrant une réalisation d'une variable aléatoire approximativement gaussienne centrée sur le k_{eff} et dont la variance est fonction du nombre de tirages, donc du temps de calcul alloué. Les outils de Krigeage traditionnels, employés jusqu'ici essentiellement pour l'approximation sur la base d'observations déterministes ponctuelles, peuvent sembler mal adaptés pour travailler avec un champ de distributions de probabilité. Nous allons voir qu'il est en fait possible de détourner ces derniers de manière quasiment immédiate pour intégrer l'incertitude occasionnée par un bruit de simulation hétérogène, au sein d'un unique modèle probabiliste.

Une des spécificités d'un simulateur numérique stochastique —par opposition aux expériences issues des sciences expérimentales, naturellement bruitées—, est que l'utilisateur a un contrôle sur le niveau de bruit, i.e. sa variance. Le prix à payer pour la réduction du niveau d'incertitude est bien sûr un temps de simulation accru. La variance Monte-Carlo varie typiquement en $\frac{1}{N}$, où N est le nombre de tirages, i.e. comme l'inverse du temps de calcul. La simulation probabiliste n'est d'ailleurs pas le seul cadre où la fonction d'intérêt y ne peut évaluée avec une précision illimitée mais doit plutôt être estimée en chaque point du plan d'expériences avec une précision réglable. Les méthodes par éléments finis peuvent par exemple être vues sous un angle très similaire en autorisant le raffinement des fonctions de base. Dans un cas comme dans l'autre (MC ou EF), la planification expérimentale n'apparait alors plus seulement comme le choix d'un ensemble de points de l'espace des variables, mais comme celui d'un ensemble de couples (point, précision), ou encore d'une séquence de points dans l'*espace - temps de calcul*.

La modélisation qui suit répond à une question posée début 2007 par Yann Richet et Eric Letang (IRSN), et apporte une formalisation au cadre de travail décrit ci-dessus en termes d'observations bruitées et de processus aléatoires, permettant ainsi d'obtenir une adaptation rigoureuse du Krigeage aux besoins constatés dans l'utilisation du simulateur MORET. Nous allons considérer dans ce qui suit que les simulations sont faites par « périodes » et utiliser l'indice $i \in [1, n]$ pour les décrire. Chaque période concerne un

point (une valeur des variables d'entrée) $\mathbf{x}^i \in D$ et un temps de calcul $t_i \in \mathbb{R}^+$. Les observations bruitées faites en chaque période sont notées Y_N^i, et nous allons supposer dans la suite que ces variables suivent des lois gaussiennes, centrées sur les $y(\mathbf{x}^i)$, et avec une variance $\tau^2(t_i)$ dépendant du temps de calcul, et éventuellement de $\mathbf{x}^i$:

$$Y_N^i \sim \mathcal{N}\left(y(\mathbf{x}^i), \tau^2(t_i, \mathbf{x}^i)\right) \tag{7.13}$$

Considérer les $\mathbf{x}^i$ et les t_i comme connus, et même contrôlés, mène à différentes questions emergeant assez naturellement dans l'étude des simulateurs à fidélité variable. Nous donnons ici des éléments de réponse à quelques-unes d'entre elles. L'outil proposé pour ce faire est une variante du Krigeage, le Krigeage avec bruit d'observation hétérogène, construit et explicité en cette fin de chapitre. Avant de rentrer plus en détail dans la construction de ce modèle, décrivons plus précisément un exemple de situation pratique dans laquelle le modèle 7.13 est particulièrement pertinent.

Exemple « canonique » : le cadre des simulations par tirages Monte-Carlo

On suppose ici qu'il est possible d'échantillonner un nombre arbitraire de réplications, i.e. de tirages Monte-Carlo, d'une variable aléatoire gaussienne centrée en $y(\mathbf{x})$:

$$Y_{\mathbf{x}}^j = y(\mathbf{x}) + \tau(\mathbf{x})N_j \sim \mathcal{N}(y(\mathbf{x}), \tau^2(\mathbf{x})) \quad (1 \leq j \leq m) \tag{7.14}$$

En faisant ainsi $m \in \mathbb{N}^*$ tirages indépendants en $\mathbf{x}$, on peut obtenir une estimation de $y(\mathbf{x})$, dont la précision dépend à la fois de $\tau^2(\mathbf{x})$ et de m : $\frac{1}{m}\sum_{j=1}^m Y_{\mathbf{x}}^j \sim \mathcal{N}\left(y(\mathbf{x}), \frac{\tau^2(\mathbf{x})}{m}\right)$.

Pour reprendre le formalisme associé à des jeux d'observations obtenus par périodes, le nombre de tirages alloué et les réplications du bruit d'observation sont notés pour chaque $i \in [1, n]$ respectivement $m_i \in \mathbb{N}^*$ et $\{N_i^1, ..., N_i^{m_i}\}$. L'estimation Monte-Carlo Y_i^{MC} obtenue au cours d'une période $i \in [1, n]$ peut alors être explicitée sans difficulté :

$$\begin{aligned} Y_i^{MC} &:= \frac{1}{m_i}\sum_{j=1}^{m_i}(y(\mathbf{x}^i) + \tau(\mathbf{x}^i)N_i^j) \\ &= y(\mathbf{x}^i) + \tau(\mathbf{x}^i)\left(\frac{N_i^1 + ... + N_i^{m_i}}{m_i}\right) \sim \mathcal{N}\left(y(\mathbf{x}^i), \frac{\tau^2(\mathbf{x}^i)}{m_i}\right) \end{aligned}$$

Pour des raisons pratiques, nous serons éventuellement amenés à faire l'approximation que τ ne dépend pas de $\mathbf{x}$, par exemple lors de la recherche de plans optimaux.

Retour sur le Krigeage Ordinaire

Reprenons pour un instant les hypothèses du Krigeage Ordinaire (Cf. chapitre 3). La fonction d'intérêt y est supposée être une réalisation d'un processus gaussien stationnaire

Y avec pour moyenne[6] une constante inconnue $\mu \in \mathbb{R}$ et pour noyau de covariance une fonction de type positif $k : \mathbf{h} \in \mathbb{R}^d \longrightarrow k(\mathbf{h}) = \sigma^2 r_\psi(\mathbf{h})$ de structure de corrélation r et de paramètres $\sigma > 0$, $\psi \in \mathbb{R}^p$ **connus**. On suppose aussi que y a déjà été parfaitement observée en un plan d'expériences à n points, $\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\} \in D^n$. Comme nous l'avons vu en détail au chapitre 3, le Krigeage Ordinaire consiste à prédire la valeur de $y(\mathbf{x})$ en un point quelconque $\mathbf{x} \in D$ en conditionnant le processus Y aux observations $\mathbb{Y} = y(\mathbf{X})$ faites au plan d'expériences :

$$Y^{OK}(\mathbf{x}) = [Y(\mathbf{x}) | Y(\mathbf{X}) = \mathbb{Y}] \sim \mathcal{N}\left(m(\mathbf{x}), s^2(\mathbf{x})\right)$$

L'espérance conditionnelle et la variance conditionnelle de Y aux observations, alias la moyenne et la variance de Krigeage Ordinaire, apparaissent comme des conséquences vertueuses de la vision par conditionnement d'un processus aléatoire gaussien :

$$\begin{cases} m(\mathbf{x}) = \widehat{\mu} + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbb{Y} - \widehat{\mu}\mathbb{1}) \\ \\ s^2(\mathbf{x}) = \sigma^2 \left[1 - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}) + \frac{\left(1 - \mathbb{1}^T K^{-1}\mathbf{k}(\mathbf{x})\right)^2}{\mathbb{1}^T K^{-1}\mathbb{1}}\right] \end{cases} \tag{7.15}$$

où les notations employées sont toujours :

$$\widehat{\mu} = \frac{\mathbb{1}^T K^{-1}\mathbb{Y}}{\mathbb{1}^T K^{-1}\mathbb{1}} \quad K = \begin{pmatrix} k(0) & \ldots & k(\mathbf{x}^1 - \mathbf{x}^n) \\ k(\mathbf{x}^2 - \mathbf{x}^1) & \ldots & k(\mathbf{x}^2 - \mathbf{x}^n) \\ \ldots & \ldots & \ldots \\ k(\mathbf{x}^n - \mathbf{x}^1) & \ldots & k(0) \end{pmatrix} \quad \mathbf{k}(\mathbf{x}) = \begin{pmatrix} k(\mathbf{x} - \mathbf{x}^1) \\ k(\mathbf{x} - \mathbf{x}^2) \\ \ldots \\ k(\mathbf{x} - \mathbf{x}^n) \end{pmatrix}$$

Rappelons quelques propriétés fondamentales de ce métamodèle dans le cadre d'observations déterministes :

- $m(\mathbf{x})$ interpole y aux points du plan d'expériences
- $s^2(\mathbf{x})$ vaut zero en ces mêmes points
- Les équations de Krigeage peuvent être écrites de manière équivalente en termes de vecteurs et matrices de covariance ou de corrélation :

$$\begin{aligned} m(\mathbf{x}) &= \widehat{\mu} + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbb{Y} - \widehat{\mu}\mathbb{1}) \\ &= \widehat{\mu} + \mathbf{r}(\mathbf{x})^T R^{-1}(\mathbb{Y} - \widehat{\mu}\mathbb{1}), \text{ où } K =: \sigma^2 R \end{aligned}$$

[6] Le travail d'adaptation des équations de Krigeage Ordinaire présenté dans cette section peut être directement étendu au cas de tendances quelconques (e.g. KU) d'une part, et de noyaux non-stationnaires d'autre part. Le KO est le modèle le plus simple permettant de voir à quels stades intervient le bruit.

– Lors de l'estimation de ψ et de σ^2 par maximum de vraisemblance, $\widehat{\mu}$ et $\widehat{\sigma^2}$ peuvent être calculés explicitement en fonction de $\widehat{\psi}$ via les relations

$$\widehat{\mu}_{\widehat{\psi}} = \frac{\mathbb{1}^T R^{-1} \mathbb{Y}}{\mathbb{1}^T R^{-1} \mathbb{1}} \qquad \text{et} \qquad \widehat{\sigma^2} = \frac{1}{n} (\mathbb{Y} - \hat{\mu}_{\widehat{\psi}} \mathbb{1})^T R_{\widehat{\psi}}^{-1} (\mathbb{Y} - \hat{\mu}_{\widehat{\psi}} \mathbb{1}) \tag{7.16}$$

7.3.2 Approximation par Krigeage de fonctions déterministes observées dans un bruit gaussien hétéroscédastique

Etat de l'art

L'effet de pépite (mentionné au chapitre 3) permet de modéliser la discontinuité des fonctions intrinsèques étudiées en géosciences, en forçant le variogramme à croître très rapidement au voisinage de l'origine. Contrairement à ce que l'on pourrait penser en première analyse, ce dernier n'est pas vraiment adapté à la prise en compte d'observations bruitées : il n'est pas conçu pour tolérer des réponses différentes en un même $\mathbf{x} \in D$, mais plutôt en de très proches voisins [Mat63, Mat70, Cre93]. Comme nous allons le préciser ci-dessous, notre cadre d'étude particulier nécessite de séparer dans la modélisation la fonction inconnue des variables aléatoires liées au bruit d'observation (Cf. discussion sur ce sujet dans [Ste99], pp. 95-95). Par ailleurs, la question du traitement d'un bruit hétéroscédastique au sein d'un modèle de Krigeage n'est pas tout à fait nouvelle. Depuis l'article [GWB98], plusieurs modèles tels que ceux de [KPPB07, LSC05] ou [KVB05, VBK08] ont été développés concernant la prise en compte d'un bruit de simulation avec un niveau d'incertitude fonction du point considéré dans l'espace des variables d'éntrée. Cependant, ces travaux ne répondent pas exactement à nos attentes, pour au moins l'une des deux raisons suivantes :

– Le bruit y est vu dans la plupart des cas comme une réalisation d'un processus spatial, éventuellement avec une forme paramétrique —par exemple de covariance— prédéfinie (un deuxième processus gaussien). Une telle approche ne convient pas ici, en premier lieu car elle ne permet pas la prise en compte d'observations répétées en un même point de D.

– Des techniques sophistiquées y sont proposées pour l'estimation des paramètres du processus de bruit sous-jacent. Cela dépasse clairement nos ambitions, puisque nous nous intéressons au cas d'un bruit complètement contrôlé par l'utilisateur. Le cadre théorique dans lequel nous nous plaçons ici est à mi-chemin entre celui de la régression et celui du Krigeage, dans la même veine que ce que propose [Wah90] pour motiver le modèle des splines lissage.

Krigeage avec bruit de simulation hétérogène : hypothèses proposées

Le modèle considéré ici est une hybridation des modèles de Krigeage Ordinaire et de Régression : on fait l'hypothèse que y est une réalisation d'un processus spatial (de paramètres partiellement connus), et que les observations faites aux différentes périodes $i \in [1, n]$ sont les valeurs $y(\mathbf{x}^i)$ entâchées d'un bruit gaussien :

$$\begin{cases} y \text{ est une réalisation de } Y \sim \mathcal{GP}(\mu, \sigma^2 r(.)) \\ \sigma^2 r(.) \text{ est connu et } \mu \sim \mathcal{U}(\mathbb{R}) \text{ (prior impropre)} \\ \mathbb{Y}_\epsilon = \mathbb{Y} + \boldsymbol{\varepsilon} = (y(\mathbf{x}^1) + \boldsymbol{\varepsilon}_1, ..., y(\mathbf{x}^n) + \boldsymbol{\varepsilon}_n) \\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \Delta) \text{ où } \Delta = \text{diag}\{\tau_i^2,\ i \in [1, n]\} \\ \boldsymbol{\varepsilon} \text{ et Y sont indépendents, de même que } \boldsymbol{\varepsilon} \text{ et } \mu \end{cases}$$

En outre, le développement reste valable dans le cas où $cov(\boldsymbol{\varepsilon})$ est une matrice de covariance connue de forme non diagonale, i.e. lorsque le bruit lui-même est autocorrélé.

Calcul des équations de Krigeage Ordinaire avec bruit hétérogène

Comme dans le cas d'observations déterministes, nous allons considérer le vecteur aléatoire « augmenté » $(\mathbb{Y}_\epsilon, Y(\mathbf{x})) = (Y(\mathbf{X}) + \boldsymbol{\varepsilon}, Y(\mathbf{x}))$, où $\mathbf{x} \in D$ est un point en lequel on souhaite prédire Y (i.e. le processus non-bruité). Les propriétés clef pour la construction des équations de Krigeage Ordinaire sont les suivantes :

- $\boldsymbol{\varepsilon}$ est indépendant de μ, d'où $\left(Y(\mathbf{X}) + \boldsymbol{\varepsilon}, Y(\mathbf{x}^{n+1})\right) \Big| \mu = \left((Y(\mathbf{X}), Y(\mathbf{x}^{n+1})\right) \Big| \mu + (\boldsymbol{\varepsilon}, 0)$ est gaussien en tant que somme de deux vecteurs gaussiens indépendants.

- Par suite, dans le cas où μ est connu (Krigeage Simple), $\left[Y(\mathbf{x}^{n+1}) | Y(\mathbf{X}) + \boldsymbol{\varepsilon} = \mathbf{Y}_\epsilon, \mu\right]$ est clairement gaussien comme sous-partie d'un vecteur gaussien conditionnée à une autre sous-partie de ce même vecteur.

- On dispose alors à la fois d'équations de Krigeage Simple, et de la loi *a prosteriori* de μ après avoir pris connaissance des observations bruitées $\mathbb{Y}_\epsilon = \mathbf{Y}_\epsilon$:

$$\begin{cases} [\mu | Y(\mathbf{X}) + \boldsymbol{\varepsilon} = \mathbf{Y}_\epsilon] \sim \mathcal{N}\left(\widehat{\mu}_\Delta, \frac{\sigma^2}{\mathbb{1}^T(\Sigma+\Delta)^{-1}\mathbb{1}}\right) \\ \text{où } \widehat{\mu}_\Delta = \frac{\mathbb{1}^T(\Sigma+\Delta)^{-1}\mathbf{Y}_\epsilon}{\mathbb{1}^T(\Sigma+\Delta)^{-1}\mathbb{1}} \end{cases}$$

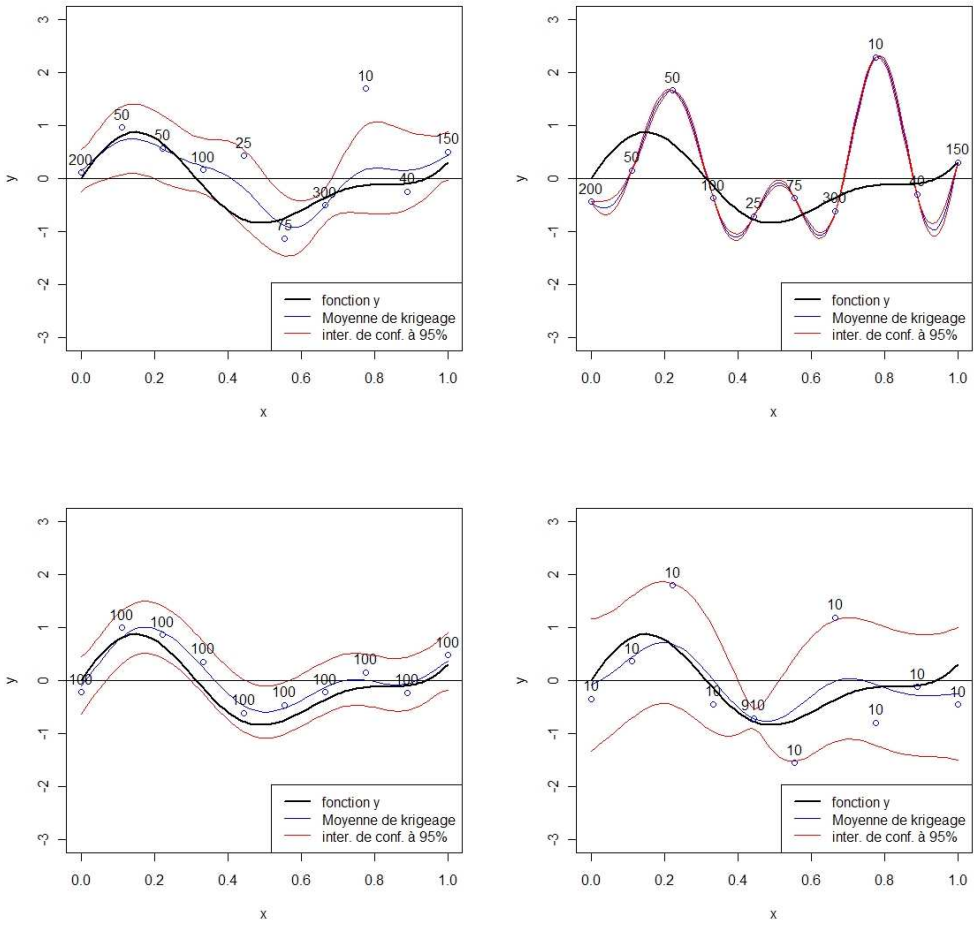


FIG. 7.16 – Krigeage Ordinaire de la fonction f (Cf. 7.19) avec bruit contrôlé.

On peut montrer —exactement de la même manière que ce qui a été fait avec le KO et le KU « classiques » à la fin du chapitre 3— que la loi conditionnelle de $Y(\mathbf{x})$ sachant les observations bruitées $[Y(\mathbf{x})|Y(\mathbf{X}) + \boldsymbol{\varepsilon} = \mathbb{Y}_\epsilon]$ est elle aussi gaussienne. On obtient finalement ses moments en appliquant les lois de l'espérance totale et de la variance totale :

$$m_\Delta(\mathbf{x}) = \widehat{\mu}_\Delta + \mathbf{k}(\mathbf{x})^T (K + \Delta)^{-1} (\mathbf{Y}_\epsilon - \widehat{\mu}_\Delta \mathbb{1}) \tag{7.17}$$

$$s^2_\Delta(\mathbf{x}) = \sigma^2 \left[1 - \mathbf{k}(\mathbf{x})^T (K + \Delta)^{-1} \mathbf{k}(\mathbf{x}) + \frac{\left(1 - \mathbb{1}^T (K + \Delta)^{-1} \mathbf{k}(\mathbf{x})\right)^2}{\mathbb{1}^T (K + \Delta)^{-1} \mathbb{1}} \right] \tag{7.18}$$

Nous proposons ci-dessus (Cf. figure 7.16) une illustration du Krigeage avec bruit d'observation hétérogène, obtenue en appliquant la méthode à la fonction déterministe

$$f : x \in \mathbb{R} \longrightarrow f(x) := \frac{sin(10x)}{(1+x)} + 2cos(5x)x^3 \tag{7.19}$$

observée en 10 points de $[0, 1]$ dans un bruit gaussien contrôlé. Les deux graphes du haut représentent l'approximation de f sur la base de dix observations bruitées hétérogènes, avec (à gauche) et sans prise en compte (à droite) du bruit dans les équations de KO. Les nombres indiqués au niveau de chacune des observations traduisent des temps de calcul virtuels , de somme constante. Les deux graphes du dessous représentent le KO avec prise en compte du bruit obtenu en observant f aux mêmes points que précédemment, avec deux profils de variance différents : uniforme à gauche, et avec la majorité du temps de calcul réparti en un point à droite.

Pour revenir sur les équations 7.17, remarquons que comparé aux équations du KO classique, seules la matrice K est modifiée, changée en $K+\Delta$. En particulier, le spectre de la matrice des observation bruitées est décalé vers la droit par rapport à celui de la matrice K, en vertu du théorème de Weyl. Les vecteurs de covariance restent eux inchangés. Les changements occasionnés sur le modèle de KO concernent la perte d'interpolation de la moyenne de Krigeage (effet de lissage, sans lequel deux observations distinctes en un même point donneraint un système non-inversible), et une inflation globale du terme de variance : $\forall \mathbf{x} \in D, s^2_\Delta(\mathbf{x}) > s^2(\mathbf{x})$. En particulier, les points auxquels sont observés (uniquement) des valeurs bruitées reçoivent des variances de Krigeage strictement positives.

La possibilité de distribuer des observations de précisions contrôlables dans l'espace des variables d'éntrée révèle de nouveaux problèmes en planification expérimentale pour les modèles de Krigeage. Comme illustré sur la figure 7.16, la variabilité des temps de calcul pose la question de la meilleure répartition possible sur un plan donné. Plus dûr, on peut considérer des problèmes d'allocation optimale de ressources dans l'espace-temps de calcul, en se donnant des contraintes sur le temps de calcul total. C'est une des pistes explorées par Victor Picheny dans son travail de thèse (à paraître fin 2009).

On remarque pour finir que les simplifications valables dans le cas non-bruité pour exprimer les équations de Krigeage à parti de matrices et vecteurs de corrélation, et pour expliciter l'estimateur du maximum de vraisemblance de la variance à partir de celui des paramètres de corrélation, ne sont plus valables : $\widehat{\sigma^2} \neq \frac{1}{n}(\mathbb{Y} - \hat{\mu}_{\widehat{\psi}}\mathbb{1})^T R^{-1}_{\widehat{\psi}}(\mathbb{Y} - \hat{\mu}_{\widehat{\psi}}\mathbb{1})$ et $m(\mathbf{x}) \neq \widehat{\mu} + \mathbf{r}(\mathbf{x})^T R^{-1}(\mathbb{Y} - \widehat{\mu}\mathbb{1})$. Il est ainsi nécessaire de prendre garde dans l'adaptation

des codes de Krigeage existants, et aussi de modifier les routines d'optimisation (max. de vraisemblance concentré) pour l'estimation des paramètres de covariance.

Troisième partie

Contributions aux stratégies d'optimisation sur base de métamodèles probabilistes

Chapitre 8

Mélanges discrets adaptatifs de Krigeages pour l'optimisation

Comme on l'a vu au chap. 4, les métamodèles probabilistes tels que le Krigeage permettent de construire une classe de stratégies d'optimisation globale, incluant par exemple les algorithmes EGO [JSW98] ou SUR [VVW09]. Ces derniers reposent sur l'optimisation séquentielle de critères d'échantillonnage, définis à partir des lois conditionnelles $Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}$ issues de l'interprétation du Krigeage en termes de processus gaussiens (Cf. chap. 3). Or l'étape de définition du modèle de processus utilisé est généralement traitée avec son lot d'arbitraire. En amont de l'estimation d'éventuels paramètres de covariance (Cf. chap. 5) et/ou de paramètres de tendance (Cf. chap. 6), la sélection de la structure d'un métamodèle de Krigeage —forme fonctionnelle du noyau, fonctions de base intervenant dans la tendance— est en effet souvent faite *a priori*, ou sur la base d'observations en un plan d'expériences initial (Cf. chap. 3 et 4).

Les conséquences d'un mauvais choix de structure et/ou d'une méthode d'estimation inadaptée peuvent être très préjudiciables à la qualité du prédicteur de Krigeage obtenu. Ce constat reste d'autant plus pertinent en optimisation sur base de métamodèles probabilistes que le modèle de Krigeage est fixé en début d'algorithme, et qu'un mauvais choix de structure sur le plan initial peut altérer la qualité de toute la séquence de points visités. Il apparait ainsi souhaitable d'une part de ne pas négliger l'incertitude de modèle dans les critères d'exploration, et d'autre part de prendre en considération le fait que les algorithmes d'optimisation sur base de Krigeage sont des stratégies adaptatives, avec ré-estimation du modèle au fil des nouvelles évaluations de la fonction objectif.

Ce chapitre présente une méthode permettant d'intégrer l'incertitude de modèle au sein d'algorithmes d'optimisation tels qu'EGO. L'idée maîtresse est de réunir la structure et les paramètres —ici restreints à la covariance, par souci de simplicité— au sein de la notion d'estimateur fonctionnel, puis de probabiliser une famille donnée d'estimateurs fonctionnels sur la base des observations disponibles. Le concept statistique directement associée à une telle construction est celui de *mélanges de lois de probabilités*, que nous nous proposons de contextualiser dans notre cadre de travail.

Après quelques préliminaires au sujet de la dépendance en la structure de covariance dans les équations de Krigeage, et sur quelques méthodes existantes pour la prise en compte de multiples métamodèles en prédiction et optimisation, nous présenterons la méthode et illustrerons son intérêt sur un exemple classique. En particulier, nous préciserons la notion d'estimateur fonctionnel (ici dans le cadre d'estimateurs du noyau de covariance pour le Krigeage Ordinaire), donnerons les équations des mélanges discrets de Krigeages, et présenterons le critère d'amélioration espérée sous un mélange de métamodèles.

8.1 Back to the notations

The objective function $y : \mathbf{x} \in D \subset \mathbb{R}^d \to y(\mathbf{x}) \in \mathbb{R}$ is here assumed known at first at an initial Design of Experiments $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^{n_0}\}$, where $n_0 \in \mathbb{N}$ is the number of initial runs. We denote by $\mathbf{Y} = \{y(\mathbf{x}^1), ..., y(\mathbf{x}^{n_0})\}$ the set of observations made by evaluating y at the points of $\mathbf{X}$. Let us recall that in almost all Kriging models, the starting point is to assume that y is one realization of a random process of the following form :

$$Y(x) = \mu(x) + \varepsilon(x) \tag{8.1}$$

where $\mu(x)$ is a deterministic trend function, $\varepsilon(x)$ is a centered stationary random field with covariance kernel k. The covariance kernel k is here assumed to belong to a set of positive-definite stationary kernels :

$$\mathcal{K} = \{k_{(r,\sigma^2,\psi)} : h \in D - D \longrightarrow \sigma^2 r(h;\psi),\ r \in \mathcal{R},\ \sigma^2 \in \mathbb{R}^+,\ \psi \in \Psi_r\} \tag{8.2}$$

$\mathcal{K}$ is indexed by a finite set $\mathcal{R}$ of correlation kernel parametric families, by their respective continous hyperparameters $\psi \in \Psi_r$ (e.g. correlation lengths), and by a positive parameter σ^2 (the process variance). In the following, $\chi = (r, \sigma^2, \psi)$ denotes the tree-structured covariance parameters. In many industrial applications, r is arbitrarily chosen to belong to a parametric family (exponential, Gaussian, Matèrn, etc...), (σ^2, ψ) are then fitted to the data using automatic estimation procedures, and χ is finally plugged in as if it

were known. Our particular concern here is to review and extend the EGO Algorithm ([JSW98]) in taking the risk of model into account. After recalling some basics about Gaussian processes and metamodel-based optimization, we propose an adaptation of Ordinary Kriging (OK) with mixed kernels. We then derive an optimization criterion based on a discrete mixture of Kriging models, and finally illustrate its efficiency by applying EGO with two simultaneous kernels to a classical test case function.

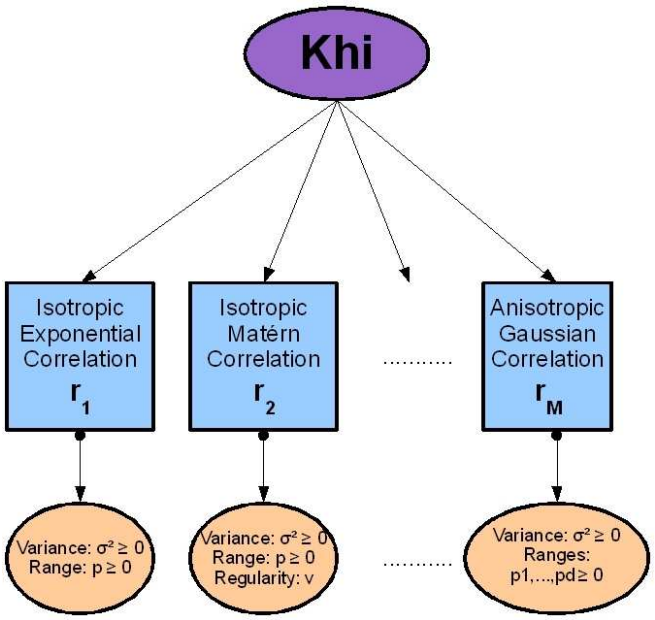


FIG. 8.1 – Schematic representation of the tree strucure of χ, which depends on a correlation kernel in the first place, and on suitable correlation coefficients in the second place. The family of kernels represented here includes (among others) isotropic exponential, isotropic Matérn, and anisotropic gaussian stationary kernels. A similar description could be done when considering a finite family of non-stationary covariance structures or a familly of trend functions, both with sets of hyperparameters to be tuned.

OK provides at each point $\mathbf{x} \in D$ a prediction of Y as a linear combination of the observed values $\mathbf{Y}$. The weights depend on the distance between the prediction point $\mathbf{x}$ and the design of experiments $\mathbf{X}$ through the chosen covariance kernel. Here we recall the equations of Kriging for a fixed kernel $k_\chi(.) = \sigma^2 r(.;\psi)$, adding an unusal emphasis on the tree-structured covariance parameter χ. The Kriging mean m_χ and mean squared error (or variance) s^2_χ at $\mathbf{x}$ are the following functions of $\mathbf{x}$:

$$\begin{cases} m_\chi(\mathbf{x}) = \widehat{\mu}_\chi + \mathbf{k}_\chi(\mathbf{x})^T \mathbf{K}_\chi^{-1} \left(\mathbf{Y} - \widehat{\mu}_\chi \mathbf{1}_n\right) \\ s_\chi^2(x) = \sigma^2 - \mathbf{k}_\chi(\mathbf{x})^T \mathbf{K}_\chi^{-1} \mathbf{k}_\chi(\mathbf{x}) + \left(\mathbf{1}_n^T \mathbf{K}_\chi^{-1} \mathbf{1}_n\right)^{-1} \left(1 - \mathbf{1}_n^T \mathbf{K}_\chi^{-1} \mathbf{k}_\chi(\mathbf{x})\right)^2 \end{cases} \tag{8.3}$$

where we recall that $\chi = (r, \sigma^2, \psi)$. $\mathbf{K}_\chi$ and $\mathbf{k}_\chi(\mathbf{x})$ are the matrices

$$\mathbf{K}_\chi = \begin{pmatrix} k_\chi(0) & k_\chi(\mathbf{x}_1 - \mathbf{x}_2) & ... & k_\chi(\mathbf{x}_1 - \mathbf{x}_n) \\ k_\chi(\mathbf{x}_2 - \mathbf{x}_1) & k_\chi(0) & ... & k_\chi(\mathbf{x}_2 - \mathbf{x}_n) \\ ... & ... & .... & .... \\ k_\chi(\mathbf{x}_n - \mathbf{x}_1) & ... & .... & k_\chi(0) \end{pmatrix} \text{ and } \mathbf{k}_\chi(\mathbf{x}) = \begin{pmatrix} k_\chi(\mathbf{x} - \mathbf{x}_1) \\ k_\chi(\mathbf{x} - \mathbf{x}_2) \\ ... \\ k_\chi(\mathbf{x} - \mathbf{x}_n) \end{pmatrix}$$

and $\widehat{\mu}_\chi$ is given by : $\widehat{\mu}_\chi = (\mathbf{1}^T \mathbf{K}_\chi^{-1} \mathbf{1})^{-1} \mathbf{1}^T \mathbf{K}_\chi^{-1} \mathbf{Y}$. Here we consider an interpretation of OK in terms of Gaussian processes, in the flavour of [RW06] (see chap. 3 for more details). Assuming that $\varepsilon(\mathbf{x})$ is a centered stationary GP with known covariance function $k_\chi(.)$ and that μ is an unknown constant with improper uniform [Rob92] prior distribution $\mu \sim \mathcal{U}(\mathbb{R})$, one obtains the following conditional distribution for $Y(\mathbf{x})$

$$Y_\chi^{OK}(\mathbf{x}) := [Y(\mathbf{x}) | Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}\left(m_\chi(\mathbf{x}), s_\chi^2(\mathbf{x})\right) \tag{8.4}$$

As we seen in chap. 4, this approach allows the analytical calculation of various quantities involving $Y(\mathbf{x})$ knowing the observations, as well as conditional simulations[1] of Y, which ensures for instance that the expectation of any function involving $Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}$ can be estimated by Monte Carlo. In the practice, one chooses a parametric correlation kernel r, then estimates the parameters (σ^2, ψ), and finally plugs in the estimated values in the formulas. It seems however that such a sketch forgets the uncertainty associated with the choice of r and the estimation of (σ^2, ψ) and hence underestimates the modeling uncertainty. Assessing uncertainty with a variance s_χ^2 obtained by plugging in χ in the Kriging equations entails a hidden risk of trusting too much a "bad" model.

When estimating Kriging parameters by *Maximum Likelihood*, on assumes that r is known and searches for the parameters $(\widehat{\mu}, \widehat{\psi}, \widehat{\sigma^2})$ that give the largest density to $\mathbf{Y}$. Noting [2] $R_\chi = \frac{1}{\sigma^2} K_\chi$, MLE then relies on the maximization of the Gaussian likelihood :

$$L(\sigma^2, \psi, \mu; \mathbf{Y}) = f(\mathbf{Y}|\sigma^2, \psi, \mu) = \frac{1}{(2\pi)^{\frac{d}{2}} (\sigma^2)^{\frac{d}{2}} det(R_\chi)^{\frac{1}{2}}} e^{-\left[\frac{(\mathbf{Y} - \mu\mathbf{1})' R_\chi^{-1} (\mathbf{Y} - \mu\mathbf{1})}{2\sigma^2}\right]} \tag{8.5}$$

[1] Conditional simulations on a fine grid covering D is only feasible when D is low-dimensional. However, conditional simulations at a small set of points are affordable whatever the dimension d.

[2] We are implicitely working here on deterministic experiments without observational noise

or equivalently on the minimization of $-2 \times \mathcal{L}(\sigma^2, \psi, \mu; \mathbf{Y})$. It can be shown that for every fixed ψ, the optimal μ and σ^2 are $\mu = \widehat{\mu_\chi}$ [3] , which doesn't depend on σ^2, and :

$$\widehat{\sigma^2}(\psi) = \frac{(\mathbf{Y} - \widehat{\mu_\chi}\mathbf{1})^T R_\chi^{-1} (\mathbf{Y} - \widehat{\mu_\chi}\mathbf{1})}{n} \tag{8.6}$$

After some calculations, ML can be restricted to the p_r-dimensional problem :

$$\min_{\psi \in \Psi_r} \{log(|K_{(r,\widehat{\sigma^2}(\psi),\psi)}|)\} \tag{8.7}$$

This non-convex optimization problem is generally solved numerically, which adds both computational complexity, dependence on the starting point(s), and randomness in the result in the case of a stochastic optimization algorithm (like genetic algorithms using derivatives, see [MS08]). Furthermore, there is an inherent, non-reducible uncertainty due to estimating ψ from a limited number of observations. The variability of the parameters obtained by ML on the basis of data actually sampled from a Gaussian process has been studied in detail in the theory of likelihood [Swe80, MM84, AW98], and more recently discussed in this particular framework in the fifth chapter of the present dissertation. Alternative estimation procedures dedicated to cope with this variability include restricted maximum likelihood methods [Cre93], penalized maximum likelihood estimation [LS05], as well as fully Bayesian approaches [O'H06, MS04b]. However, we will restrict ourselves in the following to the classical ML framework.

8.2 Mutiple metamodels : from aggregation to mixtures

Let us come back to a very general setting, in which several metamodels are in competition to be used as surrogate of the objective function y. Without giving an exhaustive presentation of what could be possible in order to combine several metamodels, we propose here a short discussion about the basic principles of aggregation and mixing, who interestingly appear to have deep differences despite obvious similarities.

8.2.1 Aggregation and other classical « multimodel » techniques

The idea of taking averaged sums of quantities estimated in different ways can be found in many fields of modeling litterature. Let us forget about probabilsitic models for some time, and come back to the deterministic metamodels of chapter 2 (sections 1 and 3).

[3] Directly maximizing the likelihood with respect to μ and (σ^2, ψ) delivers the same value of μ as in the frame of OK, $\widehat{\mu} = \widehat{\mu_\chi}$. Note however that OK includes the variability due to μ's estimation in its prediction variance.

Assume that M functions $m_1, \ldots, m_M$ are in competition to model y. When wantig to approximate $y(\mathbf{x})$ for some $\mathbf{x} \in D$, M values $\{m_1(\mathbf{x}), \ldots, m_M(\mathbf{x})\}$ are thus potentially considered. If all metamodels are treated equally, without any prior belief of any competitor being more adequate for approximating y than the others, it makes sense to introduce the arithmetic average $m_{ave}(\mathbf{x}) := \frac{1}{M}\sum_{i=1}^{M} m_j(\mathbf{x})$. The latter is the point that minimizes the sum of square distances to the $m_j(\mathbf{x})'s$, i.e. the isobarycenter of the guesses given by the M metamodels. Naturally, if some of the guesses are presumed better than others, it has to be taken into account in the way the $m_j(\mathbf{x})'s$ are weighted : what we mean here by aggregating guesses is simply taking a weighted sum of them, $\sum_{i=1}^{M} w_j m_j(\mathbf{x})$, where the w_j are by convention a set of real scalars adding to one. Now, when $\mathbf{x}$ varies, the aggregated guesses become aggregated functions, i.e. aggregated deterministic metamodels. Letting the weights w_j depend on $\mathbf{x}$ provides a very broad class of aggregated metamodels, as proposed in [GHSQ06]. The next step in their construction is how to estimate optimal weights, which happens to be a very complex case-dependent problem, addressed in [GHSQ06] with the help of cross-validation-based metrics.

In a more classical statistical framework, *model averaging* in the sense of ([BA98], p.150) relies on the use of one parametric model with averaged parameters. The best analogue in our framework is maybe the research undertaken about hyperkernels in ([Vaz05], section 5.6) or the more recent developments about multiple kernel learning, i.e. searching for optimal linear combinations of kernels, in works like [RBCG08]. In other respects [BTW07] studies statistical aggregation procedures in the regression setting. The authors consider three different types of aggregation (model selection aggregation, convex aggregation, and linear aggregation ; not that the word *aggregation* is hence far less restrictive compared to the way we used it before). They wish to evaluate the rates of convergence of the excess risks of these respective estimators, and to investigate nearly optimal aggregation schemes with weights obtained by penalized least squares. Other contemporary works like [ST07] propose to use aggregation (in the wide sense) to try selecting the best estimate of an unknown density knowing some data and based on several available estimators. This, of course, is just a very short overview of the existing research in the fertile field of model combination ; without going into more detail, let us mention the more theoretical works of O. Catoni and his collaborators (Cf. for instance [Ver00]). In other respects, it seems worth evokating some of the model combination techniques presented in [HTF01]. Chapter 8 of this book is dedicated to model inference and averaging, including bayesian methods & MCMC, model averaging (section 8.8 is rather similar to what we will propose here) and stacking (which has more to to with

[GHSQ06]). Breiman's Bagging [Bre94] studies the virtues of aggregating predictions obtained using the same model but with several bootstrapped learning sets. Chapter 9 and 10 of [HTF01] finally deals respectively with trees & related methods and with boosting, among which the section 9.5 is about Hierarchical Mixtures of Experts [JJ93], closely connected to what we will develop in the next sections.

We now come back to very concrete issues concerning our Ordinary Kriging models, for which we assume that $M \in \mathbb{N}^*$ instances of χ (say $\{\chi_1, \ldots, \chi_M\}$) are at disposal, and in competition to best model the function y. Each value χ_j $(1 \leq j \leq M)$ entails an Ordinary Kriging conditional process $Y^{OK}_{\chi_j}$. For pedagogical purpose, let us try here to brutally aggregate the conditional processes $Y^{OK}_{\chi_j}$'s —say with constant weights w_j's— and see what comes out of it. By defining the aggregated probabilistic metamodel as a weighted sum of N metamodels, we get :

$$Y^{OK}_{agg}(\mathbf{x}) := \sum_{i=1}^{N} w_i Y^{OK}_{\chi_j}(\mathbf{x}) \tag{8.8}$$

Linearity of the expectation straightforwardly implies that :

$$m^{OK}_{agg}(\mathbf{x}) = \mathbb{E}[Y^{OK}_{agg}(\mathbf{x})] = \sum_{i=1}^{N} w_i \mathbb{E}[Y^{OK}_{\chi_j}(\mathbf{x})] = \sum_{i=1}^{N} w_i m^{KO}_{\psi_i}(\mathbf{x}) \tag{8.9}$$

where $m^{OK}_{agg}(\mathbf{x})$ stands for the mean value of the aggregated predictor at $\mathbf{x}$. Only the calculation of the variance of the aggregated predictor will be more problematic. Indeed, by applying a classical development, we get

$$Var[Y^{OK}_{agg}(\mathbf{x})] = Cov\left[\sum_{i=1}^{N} w_i Y^{OK}_{\chi_j}(\mathbf{x}), \sum_{i=1}^{N} w_i Y^{OK}_{\chi_j}(\mathbf{x})\right] \tag{8.10}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j Cov\left[Y^{OK}_{\chi_i}(\mathbf{x}), Y^{OK}_{\chi_j}(\mathbf{x})\right] \tag{8.11}$$

Now, evaluating the terms $Cov\left[Y^{OK}_{\chi_i}(\mathbf{x}), Y^{OK}_{\chi_j}(\mathbf{x})\right]$ is problematic since we only know the conditional processes up to an equivalence. Handling sums of conditional processes such as in the previous definition of aggregated Kriging predictors is meaningless. This seems a good motivation to introduce mixtures of statistical distribution, who will allow us to achieve a suitable combined metamodel addressing the latter problem, while conserving the nice property of 8.9 and additionally providing a well-defined variance term.

8.2.2 Discrete mixtures of distributions

Let us consider a real random variable X, and μ_1, μ_2 two probability measures on $\mathbb{R}$. We say that X is following a mixture of the two distributions μ_1 and μ_2 when $\mu_X = p\mu_1 + (1-p)\mu_2$ for some $p \in [0,1]$. This is equivalent to saying that there exist some random variables X_1 and X_2 with respective distributions μ_1 and μ_2 such that $X = AX_1 + (1-A)X_2$ (see figure 8.5), where $A \sim \mathbb{B}(p)$ (binomial distribution) independently on X_1 and X_2. The general definition of a discrete mixture follows by extending the latter to the case of M random variables (and of their respective measures $\{\mu_1, \ldots, \mu_M\}$) and with a « one shot » multinomial distribution, or equivalently by writing $\mu_X = p_1\mu_1 + \ldots + p_M\mu_M$ where $(p_1, \ldots, p_M)$ lies on the unit M-simplex.

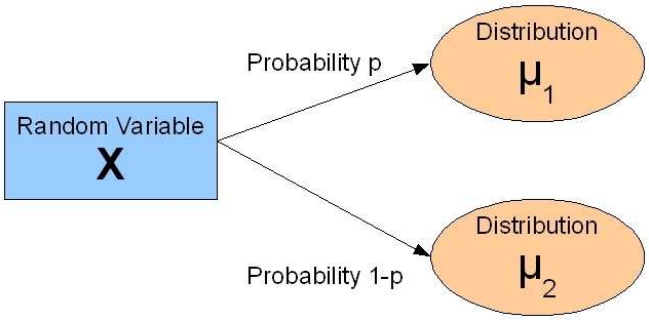


FIG. 8.2 – Illustration of a random variable X which law is a mixture between two given probability distributions μ_1 and μ_2, with weights $p \in [0,1]$ and $1-p$.

Assuming that $X_1, \ldots, X_M$ are square integrable, we directly get that

$$\begin{aligned} \mathbb{E}[X] :&= \int_{\omega\in\Omega} X(\omega)d\mathbb{P}(\omega) = \int_{x\in\mathbb{R}} x d\mu_X(x) \\ &= \int_{x\in\mathbb{R}} x\left(p_1 d\mu_1(x) + \ldots + p_M d\mu_M(x)\right) = \sum_{i=1}^{M} p_i \mathbb{E}[X_i], \end{aligned} \tag{8.12}$$

and a straightforward calculation delivers the following expression for the variance :

$$\begin{aligned} Var[X] :&= \int_{x\in\mathbb{R}} \left(x - \sum_{i=1}^{M} p_i\mathbb{E}[X_i]\right)^2 \left(p_1 d\mu_1(x) + \ldots + p_M d\mu_M(x)\right) \\ &= \sum_{i=1}^{M} p_i Var[X_i] + \sum_{i=1}^{M} p_i \left(\mathbb{E}[X_i] - \sum_{j=1}^{M} p_j \mathbb{E}[X_j]\right)^2 \end{aligned} \tag{8.13}$$

Discrete mixtures of distributions are often used to model phenomena arising in an hererogeneous populations, like mixtures of two gaussians to model the overall distribution

of weights in a country (A plays here the rôle of the gender variable). Note also that continuous mixtures of distributions are of constant use in bayesian statistics.

8.3 Discrete mixtures of kernels for Ordinary Kriging

8.3.1 Modeling and construction

We now come back to Kriging and focus on the situation in which several kernels are in competition to model the unnknown function from a vector of observed data. This is typically the case when different methods are available for the estimation of (σ^2, ψ) (e.g. ML with different initial values, ML and cross-validation, penalized ML with different penalty functions, etc.), or when there is a choice to make between a set of functional forms for the correlation kernel r. Let us assume that the function y has already been evaluated at a finite set of points $\mathbf{X}_{obs}$ (not necessarily the initial design), and denote by $\mathbf{Y}_{obs}$ the associated responses. We now consider that $M \in \mathbb{N}$ *experts* $\{\mathcal{E}_i,\ i \in [1, M]\}$ are at disposal to estimate χ on the basis of the observations. The $\mathcal{E}_i$'s are functional estimators, providing at the same time a correlation structure and its associated parameters. They are defined as follows [4] :

$$\forall i \in [1, M],\ \mathcal{E}_i : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \in \bigcup_{k=n_0}^{+\infty} \left(D^k \times \mathbb{R}^k\right) \longrightarrow \chi_i = \mathcal{E}_i(\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \in \mathcal{K} \quad (8.14)$$

For the sake of convenience, we identify $\mathcal{K}$ here with the set of possible χ's (there is an obvious one-to-one mapping between both sets). Given the set of kernels $\{\chi_1, ..., \chi_M\}$ delivered by the experts $\{\mathcal{E}_i, i \in [1, M]\}$ and $\mathcal{W} = \{w_1, ..., w_M\}$ (s.t. $\sum_{i=1}^{M} w_i = 1$) a set of weights meant to quantify the respective relevance levels of the M experts, we study the idea of replacing the classical approach of kernel selection by a mixture of kernels : instead of keeping the best kernel and dropping off the others, we propose to keep them all and integrate them within OK in probabilizing χ.

Discrete mixture of Gaussian processes : The unknown function y is now seen as one path of a random field associated with an OK model, which underlying kernel is independently chosen at random following a discrete law supported by the set of kernels delivered by the M experts :

$$\begin{cases} [Y_{mix}^{OK}|\chi] = Y_{\chi}^{OK} \\ P(\chi = \chi_i) = w_i \end{cases} \quad (8.15)$$

[4] More formally, The $\mathcal{E}_i$ may also be seen as applications from the set of discrete measures on D to the considered subset $\mathcal{K}$ of the set of kernels of positive type.

Note that the proposed approach is not strictly Bayesian : the *prior distribution* on χ, in the sense of a Bayesian framework, would depend on the data in this case. In this work, we first restrict ourselves to a small set of kernels selected on the basis of the available data, and choose to mix them afterwards : the observations are here used at every step, unlike what could be expected from a straight Bayesian procedure (More detail about that is given at the end of this section). Following Equation 8.15, the conditional distribution of $Y(\mathbf{x})$ ($\mathbf{x} \in D$) is a mixture of Gaussians :

$$\begin{aligned} &Y_{mix}^{OK}(\mathbf{x}) := [Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}] \\ &\text{with density function } \sum_{j=1}^{M} w_j p_{\mathcal{N}\left(m_{\chi_j}(\mathbf{x}), s^2_{\chi_j}(\mathbf{x})\right)}(.) \end{aligned} \tag{8.16}$$

Ordinary Kriging with a mixed kernel : Following Equation (8.16), Y_{mix}^{OK} is a field of Gaussian mixtures. This entails the equations of the mixed mean [5] and variance [6] :

$$m_{mix}(\mathbf{x}) = \mathbb{E}[\mathbb{E}[Y_{mix}^{OK}(\mathbf{x})|\chi]] = \sum_{i=1}^{M} w_i m_{\chi_i}(\mathbf{x}) \tag{8.17}$$

Hence, the mean of the resulting metamodel is the weighted average of the means associated with the different Krigings (which coincides with the concept of weighted average surrogate model developped in [GHSQ06]). Furthermore, the corresponding variance is given by

$$\begin{aligned} s^2_{mix}(\mathbf{x}) &= Var[Y_{mix}^{OK}(\mathbf{x})] = \mathbb{E}[Var[Y_{mix}^{OK}(\mathbf{x})|\chi]] + Var[\mathbb{E}[Y_{mix}^{OK}(\mathbf{x})|\chi]] \\ &= \sum_{i=1}^{M} w_i s^2_{\chi_i}(\mathbf{x}) + \sum_{i=1}^{M} w_i[(m_{\chi_i}(\mathbf{x}) - m_{mix}(\mathbf{x}))^2] \end{aligned} \tag{8.18}$$

The first term is a linear combination of the model variances weigthed by the $w_i's$, whereas the second term reflects the dispersion between the different Kriging means. The latter plays a capital role since it introduces data dependence in the Kriging variance [7] : contrarily to the case of regular OK, the variance now depends on the observations $\mathbf{Y}$ through the second term. This is in fact a particular case of a well-known property in the frame of Bayesian Kriging [Gor04, O'H06].

[5] Using the law of total expectation : $\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1|X_2]]$.

[6] Using the law of total variance : $Var[X_1] = \mathbb{E}[Var[X_1|X_2]] + Var[\mathbb{E}[X_1|X_2]]$

[7] Sometimes referred to as *heteroskedasticity of the variance with respect to the data*

8.3.2 Elementary properties of a discrete mixture of Krigings

Selecting a benchmark of experts : replacing the step of model selection by a step dedicated to choosing a set of models may seem at first to create more problems than it solves indeed. For instance, if we consider several families of correlation kernels (e.g. a Gaussian, an exponential, and even nonstationary correlation kernels —as in [Pac03]—) and estimate each set of kernel parameters by ML, it naturally increases the computational amount. The price for mixing is in that case to multiply the time needed for model inference by the number of experts. In this flavor, possible approaches would be to simultaneously consider several experts relying on the same correlation kernel but with hyperparameters inferred using different methods (e.g. mixing the ML and the Leave-One-Out (LOO) "best" models), or even getting several candidate hyperparameter sets by parametric bootstrap. Ideally, we would like to have all relevant classes of experts represented in a small set. One of the future issues to be addressed seems to be the construction of benchmarks with dissimilar good experts, i.e. of multiple kernels that fit the data well but provide very different results in prediction. In the frame of stationary Gaussian Processes, both regularity and anisotropy properties of the kernels consitute interesting features since they closely condition the behaviour of the corresponding process realizations (see [RW06] for instance). The example of section 4 presents a mixture bewteen one smooth (Gaussian) and one non-smooth (exponential) kernel.

Setting a probability measure over the set of experts : Once a set of experts is chosen, probability weights have to be defined. The most naïve way of probabilizing the models is to put a uniform distribution on them. This approach may be relevant when mixing models obtained by maximizing different criteria (e.g. : a $50\% - 50\%$ mix of the "best ML model" and the "best LOO model"). On the contrary, it is also possible to consider the density of the mixture of models as a function of both the covariance parameters and parametric weights and then to perform likelihood maximization over all parameters including the $w_i's$. Such problems are typically numerically solved using an *Expectation-Maximization* (EM) algorithm [MK97]. We propose a way inbetween, more informative than a raw uniform distribution and yet computationally cheaper than EM. Since we have a criterion of fit (the Gaussian likelihood), why not use it to weight models ? As a first step, we propose here a Kriging mixture with weights based on the likelihood criterion. In what follows, we consider Akaike weights (see [BA98]) :

$$\forall i \in [1, M],\ w_i = \frac{L(\chi_i; \mathbf{Y})}{\sum_{j=1}^{N} L(\chi_j; \mathbf{Y})} \tag{8.19}$$

The w_i's may be simply interpreted as likelihood[8] profile values divided by a normalization coefficient. Note that they may also be interpreted as conditional probabilities : putting a prior distribution π on χ, an application of Bayes' rule and the total probability formula gives

$$\begin{aligned} P(\chi = \chi_i|\mathbf{Y}) &= \frac{p(\mathbf{Y}|\chi = \chi_i)\pi(\chi_i)}{p(\mathbf{Y})} \\ &= \frac{p(\mathbf{Y}|\chi = \chi_i)\pi(\chi_i)}{\int p(\mathbf{Y}|\chi)d\pi(\chi)} \end{aligned} \tag{8.20}$$

Let us remark that the weigths w_i of Equation 8.19 coincide with $P(\chi = \chi_i|\mathbf{Y})$ when $\pi(\chi) = \frac{1}{M}\sum_{j=1}^{M}\delta_{\chi_i}(\chi)$. Recalling that $\chi_i = \mathcal{E}_i(\mathbf{X}_{obs}, \mathbf{Y}_{obs})$, the latter prior is clearly dependent on the observations so that the approach is not really Bayesian. At this stage, we do not know yet if any interpretation with a suitable prior can be found in order to describe this Akaike weighting in a fully Bayesian setting.

8.4 Mixtures and Optimization

8.4.1 Integration of optimization criteria when multiple metamodels are considered

Optimization under a mixture of kernels : when several values of χ are possible, finding the next most promizing point with a kriging-based optimization criterion (say C_{χ}^{EI}) becomes a multicriteria decisional problem. Our approach here is to combine all $C_{\chi_i}^{EI}$'s in order to provide a unified criterion that takes into account both sources of randomness. Using again the so-called law of total expectation, we derive the *mixed* EI :

$$\begin{aligned} C_{mix}^{EI}(\mathbf{x}) : &= \mathbb{E}\left[\left(y_{min} - Y_{mix}^{OK}(\mathbf{x})\right)^{+}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(y_{min} - Y_{mix}^{OK}(\mathbf{x})\right)^{+}|\chi\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(y_{min} - Y_{\chi}^{OK}(\mathbf{x})\right)^{+}\right]\right] = \sum_{i=1}^{M} w_i C_{\chi_i}^{EI}(\mathbf{x}) \end{aligned} \tag{8.21}$$

and hence, the expected improvement function under a mixture of kernels is simply the convex combination of the M expected improvement functions weighted by the $\{w_i, \in [1, M]\}$. Note that any integral criterion under a mixture of Krigings can be calculated in the same manner.

[8] The value of μ is here implicitely set to $\widehat{\mu_\chi}$ in the likelihood functions depending on χ.

8.4.2 EGO with mixed kernels, applied to Branin's function

The Branin-Hoo function has been intensively studied in the litterature of global optimization of black-box functions [JSW98]. We recall that

$$y(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)cos(x_1) + 10,\ (x_1, x_2) \in [-5, 10] \times [0, 15],$$

and that y has three global minimizers $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$, with a global minimum approximately equal to 0.4. The variables are normalized between 0 and 1. Now we wish to illustrate, and compare EGO with different kernels and kernel mixtures.

The experimental set-up is the following : the initial design of experiments is a three-level full factorial design $\mathbf{X} \in ([0, 1] \times [0, 1])^{n_0}$ $(n_0 = 9)$. Two correlation kernels are selected :

Gaussian correlation	Exponential correlation
$r_1(h) = e^{-\frac{\|\|h\|\|^2}{p^2}}$	$r_2(h) = e^{-\frac{\|\|h\|\|}{p}}$

The experts considered here are the two parametric correlation kernels r_1 and r_2 with their respective correlation parameters and associated variances estimated by ML :

$$\begin{cases} \mathcal{E}_1 : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \rightarrow \chi_1 = (r_1, \sigma_1^{2*}, \psi_1^*), \text{ where } (\sigma_1^{2*}, \psi_1^*) = argmax_{\sigma^2,\psi}[L(\sigma^2, \psi; \mathbf{Y}_{obs}, r_1)] \\ \mathcal{E}_2 : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \rightarrow \chi_2 = (r_2, \sigma_2^{2*}, \psi_2^*), \text{ where } (\sigma_2^{2*}, \psi_2^*) = argmax_{\sigma^2,\psi}[L(\sigma^2, \psi; \mathbf{Y}_{obs}, r_2)] \end{cases}$$

All the algorithms and computations are implemented in the frame of the MatLab Simple Kriging free toolbox "Gaussian Processes for Machine Learning" (illustrating the book [RW06]). In both cases, the kernel hyperparameters intitial values are fixed to $(p, \sigma^2) = (0.1, 10)$.

The results are summarized in figure 8.5. The left figure illustrates $n = 25$ iterations of EGO with mixed experts. The pattern of the visited points is close to the trajectory of EGO with Gaussian expert. In particular, the three zones of local optima are visited during the first 25 iterations. This similarities can be understood by looking at the sequences of weights plotted on the right figure. The two curves on the graphic below represent the log-likelihood associated with both experts as functions of the number of EGO iterations. Note that the likelihood of expert 2 is greater than the likelihood of expert 1 until the number of iterations reaches 6, and then becomes significantly lower than the other one. The likelihood ratios plotted on figure 8.5 show more precisely how the exponential kernel prevails at the beginning of the EGO algorithm and is later dropped in favour of the Gaussian kernel. This kind of *automatic selection* seems due

Algorithm 6 The E.G.O. Algorithm with 2 mixed kernels weighted by their likelihood ratios

function EGO($\mathbf{X}$, $\mathbf{Y}$, n)
  for $i \leftarrow 1, n$ **do**
    for $j \leftarrow 1, 2$ **do**
      $(\sigma_j^{2*}, \psi_j^*) = argmax_{(\sigma_j^2, \psi_j) \in \mathbb{R}^+ \times \Psi_{r_j}} L(\sigma_j^2, \psi_j; Y(\mathbf{X}) = \mathbf{Y}, r_j)$ ▷ MLE
    end for
    for $j \leftarrow 1, 2$ **do**
      $w_j = \frac{L(\sigma_j^{2*}, \psi_j^*; Y(\mathbf{X}) = \mathbf{Y}, r_j)}{\sum_{l=1}^{2} L(\sigma_l^{2*}, \psi_l^*; Y(\mathbf{X}) = \mathbf{Y}, r_l)}$ ▷ Computing the mixing weights
    end for
    $\mathbf{x}^* = argmax_{\mathbf{x} \in D} \sum_{l=1}^{2} w_j C^{EI}_{(r_j, \sigma_j^{2*}, \psi_j^*)}(\mathbf{x})$ ▷ Maximizing the mixed EI
    $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^*\}$ *and* $\mathbf{Y} = \mathbf{Y} \bigcup \{y(\mathbf{x}^*)\}$ ▷ Updating the Design of Experiments
  end for
end function

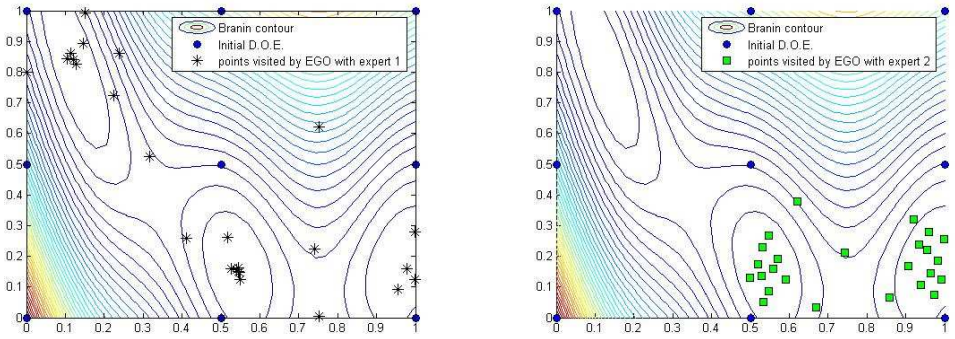


FIG. 8.3 – 25 iterations of EGO applied to the Branin-Hoo function, with both experts $\mathcal{E}_1$ and $\mathcal{E}_2$ and initial design $\mathbf{X}$ (in dark blue dots). Left : the path of EGO with expert $\mathcal{E}_1$ is represented by black thin stars. Right : the path of EGO with expert $\mathcal{E}_2$ is represented by light green squares. One of the three zones of minimum (upper left) is not visited.

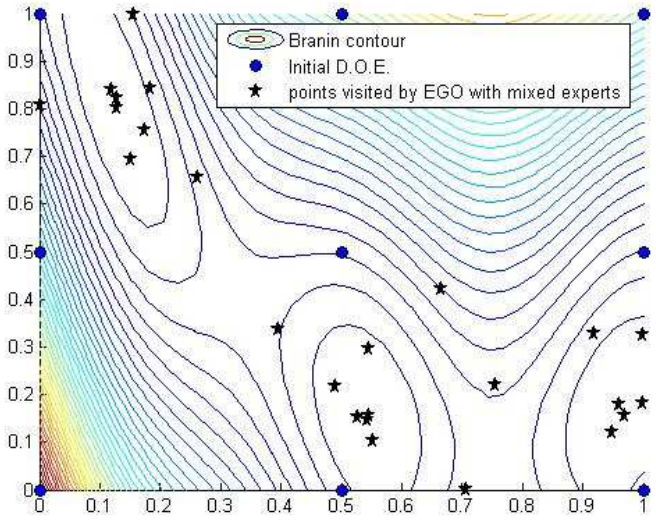


FIG. 8.4 – 25 iterations of EGO applied to the Branin-Hoo function, with a mixture of experts $\mathcal{E}_1$ and $\mathcal{E}_2$ weighted by Akaike weights.

to an asymptotical steep decrease of the likelihood ratio. Using Akaike weights to mix kernels within a sequential exploration seems here to be a useful means to automatically select an expert without making a decision based on the initial design of experiments only. Hence, the proposed approach appears to be a sound option to increase EGO's robustness to modeling uncertainty.

Conclusions concerning the method ans its first application to a toy example

We have derived and discussed an optimization criterion, the *mixed expected improvement*, relying on discrete mixtures of Kriging metamodels with different covariance kernels. The presented framework of multiple experts allows one to handle several parametric correlation structures and/or different parameter estimation techniques within the same Kriging-based procedure. The application of the latter to the optimization of the Branin-Hoo function provided a first example, where mixing appears to be a succesful alternative to model selection based on initial data. It is illustrated that, possibly after an oscillating behaviour for a few iterations, mixing two *experts* within the EGO algorithm (Krigings with Gaussian and Exponential correlation structures, with Akaike weights) may ultimately lead to an automatic selection of a unique metamodel. The issues of selecting parsimonious benchmarks of experts, and of using weighting methods

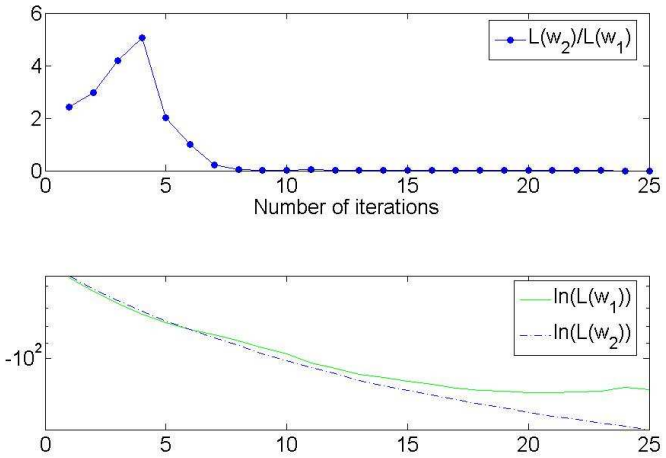


FIG. 8.5 – Evolution of both sequences of log-weights (lower graphic), and the associated series of likelihood ratios (upper graphic).

dedicated to different purposes are to be addressed in forthcoming works.

Perspectives for discrete mixtures of Kriging and their use in optimization

The idea of simultaneously considering multiple probabilistic metamodels with different structures seems valuable, and susceptible of being developed further. Enriching the work proposed here by considering more kernels, but also multiple possible structures for the trend functions in Universal Kriging, as well as adapting Hierarchical Mixtures of Experts [JJ93] to the frame of Kriging with multiple structures naturally appear to constitute potential tracks for future resarches. Weighting Krigings in a reasonable way entails defining suitable similarity measures and metric structures on the set of models, which could be investigated for instance with the help of information geometry tools [AN00].

To conclude with some remarks concerning optimization issues, we have here directly jumped from a set of weighted optimization criteria —the multiple expected improvement surfaces— to an aggregate criterion, the mixed expected improvement. Optimizing based on multiple Krigings can however be seen in a much more general manner, as a multicriteria problem : each expected improvement surface is a potential criterion. Not only criteria averaging but also minmax procedures or input-dependent selection may

be preferred, depending of course on the nature of the considered application.

Chapitre 9

Stratégies parallèles pour explorer un Krigeage

9.1 Parallélisation de critères sur base de Krigeage(s)

9.1.1 Motivations de la parallélisation d'algorithmes de type EGO

Comme on l'a vu dans les chapitres 4 et 7, l'algorithme EGO repose sur l'enrichissement **séquentiel** d'un plan d'expériences numérique, en s'appuyant sur un métamodèle de Krigeage actualisé après chaque simulation. Rappelons que le critère décisionnel sur lequel se base EGO, l'*expected improvement* (EI), a pour fonction de quantifier l'amélioration potentielle associée à tout point candidat à l'évaluation ; la maximisation de l'EI permet de guider l'échantillonnage de y en fournissant un unique point à chaque itération. Dans un contexte industriel où les clusters de machines et/ou les processeurs en parallèles sont devenus d'utilisation courante, il est absurde de devoir attendre le résultat de l'évaluation courante pour en lancer de nouvelles. Ce problème est d'autant plus important que les délais d'étude d'ingénierie se font toujours plus exigeants (en conception par exemple), ce qui fait entrer le temps d'étude effectif —nombre de jours « calendaires » nécessités par l'étude, et non pas seulement le temps de calcul CPU— au premier plan des préoccupations économiques.

Il semble ainsi pertinent de chercher à définir un (ou des) critère(s) pour mesurer le potentiel d'amélioration conjoint associé à un ensemble arbitraire de points, c'est-à-dire de quantifier l'intêret en termes d'optimisation que représente l'évaluation de y en un **plan d'expériences supplémentaire** donné. L'objectif de cette section est de proposer et d'analyser un critère d'optimisation directement inspiré de l'*expected improvement*,

l'EI « multipoints », permettant d'obtenir un nombre arbitraires de points (disons q points, $q \in \mathbb{N}\backslash\{0\}$) sur la base d'un métamodèle de Krigeage. Un tel critère est un premier pas vers la parallélisation d'EGO. Il se distingue aussi —tout comme le critère « SUR » de [VVW09]— des critères classiques (Cf. chapitre 4) dont l'objet est un gain immédiat. Nous proposons ci-dessous une construction de l'EI multipoints (il avait déjà été évoqué sous le nom de « q-step EI » dans [Sch97]), puis un développement sur deux manières de calculer ce critère selon que $q = 2$ ou $q \geq 3$. La section 9.2 est consacrée à l'optimisation approchée de l'EI multipoints. Quelques cas test illustrant les performances absolues et relatives des algorithmes d'optimisation parallèles proposés dans la section 2 sont enfin présentés dans la section terminale.

9.1.2 Le critère d'amélioration espérée à q points

On se souvient que le critère d'EI avait été construit au chapitre 4 comme l'espérance conditionnelle —sachant que $[Y(\mathbf{X}) = \mathbf{Y}]$— de la variable aléatoire d'amélioration (*improvement*) : $\forall \mathbf{x} \in D,\ I(\mathbf{x}) := (min(Y(\mathbf{X})) - Y(\mathbf{x}))^+$. On peut en fait étendre cette définition au cas de q points, en choisissant un point de vue selon lequel seule compte la meilleure amélioration réalisée par l'un des points considérés.

Définition. $\forall \mathbf{x}^{n+1}, \ldots, \mathbf{x}^{n+q} \in D$, on définit l'amélioration multipoints comme suit :

$$\begin{aligned} I(\mathbf{x}^{n+1}, \ldots, \mathbf{x}^{n+q}) : &= \max\left(I(\mathbf{x}^{n+1}), \ldots, I(\mathbf{x}^{n+q})\right) \\ &= \max\left((min(Y(\mathbf{X})) - Y(\mathbf{x}^{n+1}))^+, \ldots, (min(Y(\mathbf{X})) - Y(\mathbf{x}^{n+q}))^+\right) \\ &= \left(\min(Y(\mathbf{X})) - \min(Y(\mathbf{x}^{n+1}), \ldots, Y(\mathbf{x}^{n+q}))\right)^+, \end{aligned} \tag{9.1}$$

où la dernière égalité découle du fait que

$$\forall a, b, c \in \mathbb{R},\ \max((a-b)^+, (a-c)^+) = (a-b)^+ \text{ si } b \leq c \text{ et } (a-c)^+ \text{ sinon.}$$

Remarquons que la manière d'unifier les q critères d'amélioration (à un point) employés dans 9.1 peut être qualifiée d'élitiste : on ne juge la qualité de l'ensemble des q points qu'en fonction du point le plus performant d'entres eux. Ceci est à distinguer par exemple des sommes pondérées de critères, souvent rencontrées lorsque l'on s'intéresse à des applications économiques.

L e critère d'amélioration espérée à q points est alors défini sans détour comme l'espérance conditionnelle de l'amélioration à q points (Cf. 9.2).

$$\begin{aligned} EI(\mathbf{x}^{n+1},...,\mathbf{x}^{n+q}) :&= \mathbb{E}[\max\{(min(Y(\mathbf{X})) - Y(\mathbf{x}^{n+1}))^+,...,(min(\mathbf{Y}) - Y(\mathbf{x}^{n+q}))^+\}/Y(\mathbf{X}) = \mathbf{Y}] \\ &= \mathbb{E}\left[(\min(Y(\mathbf{X})) - \min\left(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q})\right))^+/Y(\mathbf{X}) = \mathbf{Y}\right] \\ &= \mathbb{E}\left[(\min(\mathbf{Y}) - \min\left(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q})\right))^+/Y(\mathbf{X}) = \mathbf{Y}\right] \end{aligned} \tag{9.2}$$

L'EI à q points peut ainsi être vu comme l'EI classique appliqué à la variable aléatoire $\min(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q}))$. On a dès lors affaire à une expression faisant intervenir le minimum de q variables aléatoires dépendantes, ce qui est susceptible d'occasionner des difficultés dans les calculs. Les résultats de loi conditionnelle conjointe obtenus à la fin de la section 3 du chapitre 3 vont heureusement nous permettre de débloquer la situation, comme précisé ci-dessous.

9.1.3 Comment calculer l'EI à q points ?

Calcul analytique de l'EI à deux points

Nous intéressons dans un premier temps au calcul de l'EI à deux points, défini comme

$$EI(\mathbf{x}^{n+1},\mathbf{x}^{n+2}) := \mathbb{E}[(\min(Y(\mathbf{X})) - min(Y(\mathbf{x}^{n+1}),Y(\mathbf{x}^{n+2})))^+|Y(\mathbf{X}) = \mathbf{Y}],$$

où $\mathbf{x}^{n+1},\mathbf{x}^{n+2} \in D$ sont deux points quelconques de D. Remarquons qu'en reformulant la fonction *partie positive*, cette expression peut encore s'écrire :

$$EI(\mathbf{x}^{n+1},\mathbf{x}^{n+2}) = \mathbb{E}[(\min(Y(\mathbf{X})) - min(Y(\mathbf{x}^{n+1}),Y(\mathbf{x}^{n+2})))\mathbb{1}_{min(Y(\mathbf{x}^{n+1}),Y(\mathbf{x}^{n+2}))\leq\min(\mathbf{Y})}|Y(\mathbf{X}) = \mathbf{Y}].$$

Nous allons maintenant voir que l'EI à deux points peut être développé comme somme des deux EI à un point et d'un terme correctif faisant intervenir les fonctions de répartition gaussiennes uni- et bi-dimensionnelles.

Quelques résultats classiques de calcul conditionnel nous permettent avant toute chose de préciser la relation de dépendance entre $Y(\mathbf{x}^{n+1})$ et $Y(\mathbf{x}^{n+2})$ conditionnellement à $Y(\mathbf{X}) = \mathbf{Y}$ et de fixer les notations pour le calcul à suivre.

$$\forall i \in \{1,2\},\ \begin{cases} m_i := m_{KO}(\mathbf{x}^i) = \mathbb{E}[Y(\mathbf{x}^i)|Y(\mathbf{X}) = \mathbf{Y}], \\ \sigma_i := s_{KO}(\mathbf{x}^i) = \sqrt{Var[Y(\mathbf{x}^i)|Y(\mathbf{X}) = \mathbf{Y}]}, \\ c_{1,2} := \rho_{1,2}\sigma_1\sigma_2 := cov[Y_{OK}(\mathbf{x}^1),Y_{OK}(\mathbf{x}^2)|Y(\mathbf{X}) = \mathbf{Y}] \end{cases} \tag{9.3}$$

Des résultats bien connus de conditionnement gaussien nous donnent alors les espérances et variances conditionnelles de chaque réponse inconnue connaissant l'autre :

$$
\begin{aligned}
m_{2|1} &= \mathbb{E}[Y(\mathbf{x}^2)|Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x}^1)] = m_2 + \frac{c_{1,2}}{\sigma_1^2}(Y(\mathbf{x}^1) - m_1),\\
\sigma^2_{2|1} &= \sigma_2^2 - \frac{c_{1,2}^2}{\sigma_1^2} = \sigma_2^2(1-\rho_{12}^2)\\
m_{1|2} &= \mathbb{E}[Y(\mathbf{x}^1)|Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x}^2)] = m_1 + \frac{c_{1,2}}{\sigma_2^2}(Y(\mathbf{x}^2) - m_2),\\
\sigma^2_{1|2} &= \sigma_1^2 - \frac{c_{1,2}^2}{\sigma_2^2} = \sigma_1^2(1-\rho_{12}^2)
\end{aligned}
$$

Nous noterons desormais Y_i pour $Y(\mathbf{x}^i)$ et ne **signalerons plus le conditionnement** par rapport à $Y(\mathbf{X}) = \mathbf{Y}$ par soucis de lisibilité. Nous sommes maintenant à même de calculer l'amélioration espérée à deux points $EI(\mathbf{x}^1, \mathbf{x}^2)$.

Phase 1

$$
\begin{aligned}
EI(\mathbf{x}^1, \mathbf{x}^2) &= \mathbb{E}[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbb{1}_{min(Y_1,Y_2)\leq min(\mathbf{Y})}]\\
&= \mathbb{E}[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbb{1}_{min(Y_1,Y_2)\leq min(\mathbf{Y})}(\mathbb{1}_{Y_1\leq Y_2} + \mathbb{1}_{Y_2\leq Y_1})]\\
&= \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1\leq min(\mathbf{Y})}\mathbb{1}_{Y_1\leq Y_2}] + E[(min(\mathbf{Y}) - Y_2)\mathbb{1}_{Y_2\leq min(\mathbf{Y})}\mathbb{1}_{Y_2\leq Y_1}]
\end{aligned}
$$

Comme les deux termes de la dernière somme sont similaires (à une permutation près entre $\mathbf{x}^1$ et $\mathbf{x}^2$), nous restreignons dans un premier temps notre attention à la première. En utilisant la relation $\mathbb{1}_{Y_1\leq Y_2} = 1 - \mathbb{1}_{Y_2\leq Y_1}$[1] , on obtient :

$$
\begin{aligned}
\mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1\leq min(\mathbf{Y})}\mathbb{1}_{Y_1\leq Y_2}] &= \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1\leq min(\mathbf{Y})}(1 - \mathbb{1}_{Y_2\leq Y_1})]\\
&= EI(\mathbf{x}^1) - \mathbb{E}[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1\leq min(\mathbf{Y})}\mathbb{1}_{Y_2\leq Y_1}]\\
&= EI(\mathbf{x}^1) + B(\mathbf{x}^1, \mathbf{x}^2)
\end{aligned}
$$

où $B(\mathbf{x}^1, \mathbf{x}^2) = E[(Y_1 - min(\mathbf{Y}))\mathbb{1}_{Y_1\leq min(\mathbf{Y})}\mathbb{1}_{Y_2\leq Y_1}]$. Informellement, $B(\mathbf{x}^1, \mathbf{x}^2)$ représente l'opposé de l'amélioration apportée par Y_1 lorsque $Y_2 \leq Y_1$, i.e. qui ne contribue pas à l'EI à deux points. Notre but dans les prochaines phases de calcul sera d'expliciter $B(\mathbf{x}^1, \mathbf{x}^2)$.

[1] Cette expression devrait en toute rigueur être notée $1 - \mathbb{1}_{Y_2<Y_1}$. Comme nous travaillons ici avec des v.a.r. gaussiennes, il suffit cependant que leur corrélation soit différente de 1 pour que l'expression soit exacte ($\{Y_1 = Y_2\}$ est alors négligeable). Nous faisons implicitement cette hypothèse ici et dans la suite.

Phase 2

$$B(\mathbf{x}^1,\mathbf{x}^2) = \mathbb{E}[Y_1 \mathbb{1}_{Y_1 \leq min(\mathbf{Y})} \mathbb{1}_{Y_2 \leq Y_1}] - min(\mathbf{Y}) \mathbb{E}[\mathbb{1}_{Y_1 \leq min(\mathbf{Y})} \mathbb{1}_{Y_2 \leq Y_1}]$$

A ce stade, il est bon de remarquer que $Y_1 = m_1 + \sigma_1 N_1$, où $N_1 \sim \mathcal{N}(0,1)$. On obtient en injectant cette décomposition dans la dernière expression de $B(\mathbf{x}^1,\mathbf{x}^2)$:

$$B(\mathbf{x}^1,\mathbf{x}^2) = \sigma_1 \mathbb{E}[N_1 \mathbb{1}_{Y_1 \leq min(\mathbf{Y})} \mathbb{1}_{Y_2 \leq Y_1}] + (m_1 - min(\mathbf{Y})) \mathbb{E}[\mathbb{1}_{Y_1 \leq min(\mathbf{Y})} \mathbb{1}_{Y_2 \leq Y_1}]$$

Les deux termes de cette somme requièrent qu'on y consacre de notre attention. Nous les calculons tous deux respectivement dans les phases 3 et 4.

Phase 3

En appliquant une propriété fondamentale de calcul conditionnel [2], il vient que :

$$\mathbb{E}[N_1 \mathbb{1}_{Y_1 \leq min(\mathbf{Y})} \mathbb{1}_{Y_2 \leq Y_1}] = \mathbb{E}[N_1 \mathbb{1}_{Y_1 \leq min(\mathbf{Y})} \mathbb{E}[\mathbb{1}_{Y_2 \leq Y_1} | Y_1]]$$

En utilisant de plus le fait que $Y_2|Y_1 \sim \mathcal{N}(m_{2|1}(Y_1), s^2_{2|1}(Y_1))$, on obtient

$$\mathbb{E}[\mathbb{1}_{Y_2 \leq Y_1} | Y_1] = \Phi\left(\frac{Y_1 - m_{2/1}}{s_{2/1}}\right) = \Phi\left(\frac{Y_1 - m_2 - \frac{c_{1,2}}{\sigma_1^2}(Y_1 - m_1)}{\sigma_2\sqrt{1-\rho_{12}^2}}\right)$$

Un retour au terme principal en utilisant de nouveau la décomposition de Y_1 donne :

$$\begin{aligned}\mathbb{E}\left[N_1 \mathbb{1}_{Y_1 \leq min(\mathbf{Y})} 1_{Y_2 \leq Y_1}\right] &= \left[N_1 1_{N_1 \leq \frac{min(\mathbf{Y}) - m_1}{\sigma_1}} \Phi\left(\frac{m_1 - m_2 + (\sigma_1 - \rho_{12}\sigma_2)N_1}{\sigma_2\sqrt{1-\rho_{12}^2}}\right)\right] \\ &= \mathbb{E}\left[N_1 \mathbb{1}_{N_1 \leq \gamma_1} \Phi(\alpha_1 N_1 + \beta_1)\right]\end{aligned}$$

$$\text{où } \gamma_1 = \frac{min(\mathbf{Y}) - m_1}{\sigma_1},\ \beta_1 = \frac{m_1 - m_2}{\sigma_2\sqrt{1-\rho_{12}^2}} \text{ et } \alpha_1 = \frac{\sigma_1 - \rho_{12}\sigma_2}{\sigma_2\sqrt{1-\rho_{12}^2}} \tag{9.4}$$

$\mathbb{E}[N_1 \mathbb{1}_{N_1 \leq \gamma_1} \Phi(\alpha_1 N_1 + \beta_1)]$ peut finalement être calculée par intégration par parties :

$$\begin{aligned}\int_{-\infty}^{\gamma_1} u\phi(u)\Phi(\alpha_1 u + \beta_1)du &= [-\phi(u)\Phi(\alpha_1 u + \beta_1)]_{-\infty}^{\gamma} + \int_{-\infty}^{\gamma_1} \alpha_1 \phi(u)\phi(\alpha_1 u + \beta_1)du \\ &= -\phi(\gamma_1)\Phi(\alpha_1\gamma_1 + \beta_1) + \frac{\alpha_1}{2\pi}\int_{-\infty}^{\gamma_1} e^{\frac{-u^2-(\alpha_1 u+\beta_1)^2}{2}}du\end{aligned}$$

[2] $\forall \phi \in L^2,\ \mathbb{E}[X\phi(Y)] = \mathbb{E}[\mathbb{E}[X|Y]\phi(Y)]$.

Et comme $u^2+(\alpha_1 u+\beta_1)^2 = \left(\sqrt{(1+\alpha_1^2)}u+\frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)^2+\frac{\beta_1^2}{1+\alpha_1^2}$, la dernière intégrale se réduit à :

$$\sqrt{2\pi}\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)\int_{-\infty}^{\gamma_1} e^{\frac{-\left(\sqrt{(1+\alpha_1^2)}u+\frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)^2}{2}}du = \frac{2\pi\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)}{\sqrt{(1+\alpha_1^2)}}\int_{-\infty}^{\sqrt{(1+\alpha_1^2)}\gamma_1+\frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}}\frac{e^{\frac{-v^2}{2}}}{\sqrt{2\pi}}dv$$

On conclut en reconnaissant la définition d'une fonction de répartition :

$$\mathbb{E}[N_1\mathbb{1}_{Y_1\leq min(\mathbf{Y})}1_{Y_2\leq Y_1}] = -\phi(\gamma_1)\Phi(\alpha_1\gamma_1+\beta_1)+\frac{\alpha_1\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)}{\sqrt{(1+\alpha_1^2)}}\Phi\left(\sqrt{(1+\alpha_1^2)}\gamma_1+\frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)$$

Phase 4

Il reste pour finir à calculer le terme :

$$\mathbb{E}[\mathbb{1}_{Y_1\leq min(\mathbf{Y})}\mathbb{1}_{Y_2\leq Y_1}] = E[\mathbb{1}_{X\leq min(\mathbf{Y})}\mathbb{1}_{Z\leq 0}]$$

où $(X,Z) := (Y_1, Y_2-Y_1)$ suit une loi gaussienne bi-dimensionnelle d'espérance $M = (m_1, m_2-m_1)$, et de matrice de covariance $\Gamma := \begin{pmatrix} \sigma_1^2 & c_{1,2}-\sigma_1^2 \\ c_{1,2}-\sigma_1^2 & \sigma_2^2+\sigma_1^2-2c_{1,2}\end{pmatrix}$. Le résultat final repose sur le fait que :

$$\mathbb{E}[\mathbb{1}_{X\leq min(\mathbf{Y})}\mathbb{1}_{Z\leq 0}] = CDF(M,\Gamma)(min(\mathbf{Y}),0)$$

où la notation « CDF » est utilisée pour la fonction de répartition bi-gaussienne.

Proposition : formule de l'EI à deux points.

$$EI(\mathbf{x}^1,\mathbf{x}^2) = EI(\mathbf{x}^1)+EI(\mathbf{x}^2)+B(\mathbf{x}^1,\mathbf{x}^2)+B(\mathbf{x}^2,\mathbf{x}^1) \tag{9.5}$$

$$\text{où } \begin{cases} B(\mathbf{x}^1,\mathbf{x}^2) = (m_{OK}(\mathbf{x}^1)-min(\mathbf{Y}))\delta(\mathbf{x}^1,\mathbf{x}^2)+\sigma_{OK}(\mathbf{x}^1)\epsilon(\mathbf{x}^1,\mathbf{x}^2) \\ \epsilon(\mathbf{x}^1,\mathbf{x}^2) = \alpha_1\phi\left(\frac{|\beta_1|}{\sqrt{(1+\alpha_1^2)}}\right)\Phi\left(\frac{\gamma+\frac{\alpha_1\beta_1}{1+\alpha_1^2}}{(1+\alpha_1^2)^{-\frac{1}{2}}}\right)-\phi(\gamma)\Phi(\alpha_1\gamma+\beta_1) \\ \delta(\mathbf{x}^1,\mathbf{x}^2) = CDF(\Gamma)\begin{pmatrix} min(\mathbf{Y})-m_1 \\ m_1-m_2\end{pmatrix}\end{cases} \tag{9.6}$$

Illustrations de l'*EI* à deux points

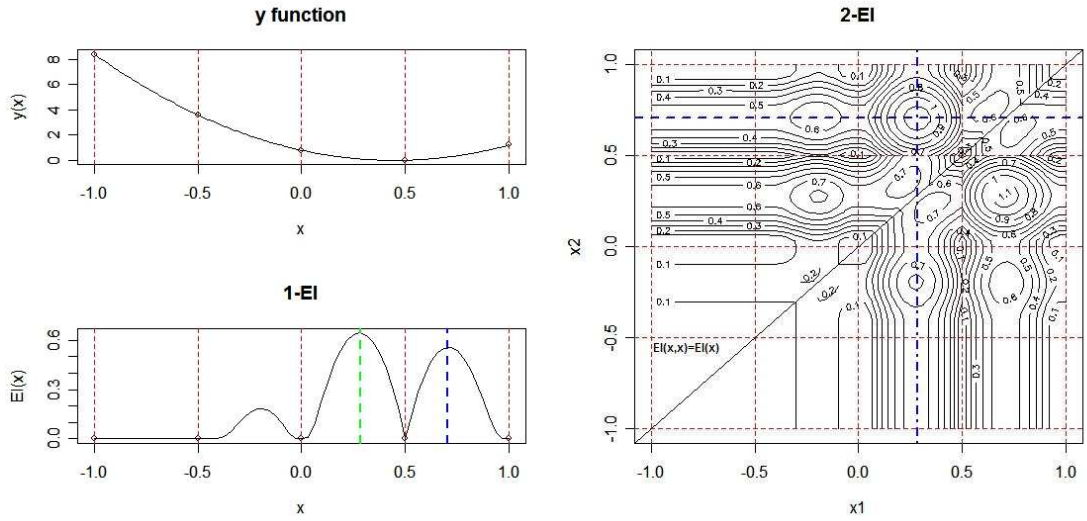


FIG. 9.1 – EI à 1 point (en bas à gauche) et EI à deux points (à droite) associés à une fonction polynômiale de degré 2 en une dimension ($y(x) = 4 \times (x - 0.45)^2$). y est ici connue sur le plan $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$ et on utilise un Krigeage Ordinaire avec noyau exponentiel, de paramètres $\sigma^2 = 10,$ et portée $= 0.9$).

On représente ci-dessus les fonctions d'EI à un et deux points associées à un polynôme de degré 2 connu en un plan d'expériences à 5 points. L'EI à un point nous incite ici à échantillonner entre les « meilleurs points » du plan d'expériences initial. Le graphe de l'EI à deux points met en lumière certaines propriétés générales : l'EI à deux points est symétrique et ses valeurs sur la diagonale sont égales à l'EI à un point, ce que l'on peut aisément voir de manière analytique en revenant aux définitions. Dit simplement, l'EI associé à un couple de points est grand lorsque les deux EI sont grands et les points raisonnablement distants l'un de l'autre (précisément au sens de la métrique employée dans le Krigeage).

De plus, maximiser l'EI à deux points sélectionne dans cet exemple les deux meilleurs optimiseurs locaux de l'EI à un point. Ce n'est en fait pas une propriété générale. D'autres cas d'étude illustrent par exemple comment la maximisation de l'EI à deux points peut fournir deux points situés de part et d'autre de l'optimiseur global de l'EI lorsque ce dernier n'a qu'un seul pic de grande amplitude (Cf. fig. 9.2).

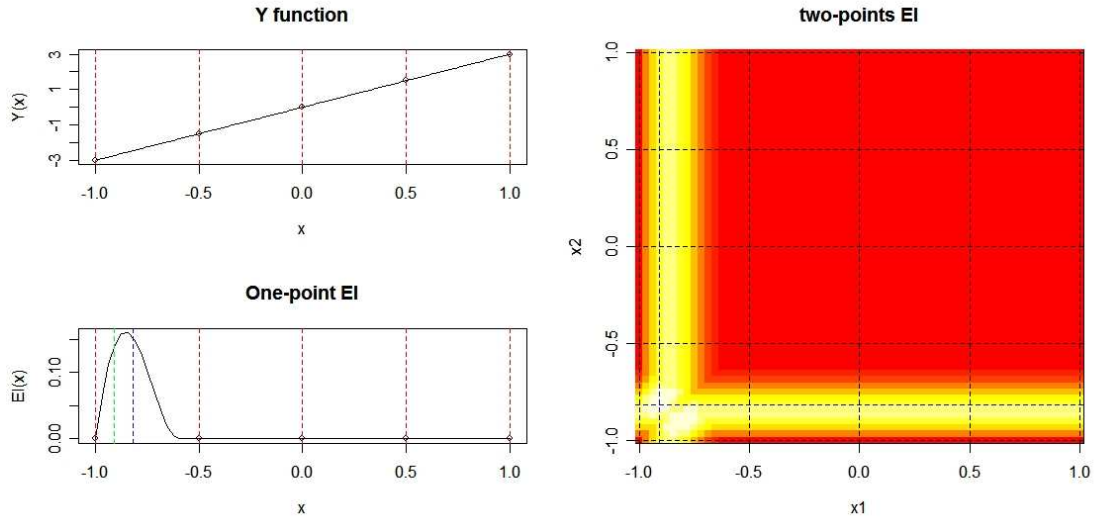


FIG. 9.2 – EI à un point (en bas à gauche) et à 2 points (à droite) associés à une application affine monodimensionnelle ($y(x) = 3 \times x$) connue au plan $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. Le Krigeage Ordinaire a ici une covariance cubique avec pour paramètres $\sigma^2 = 10$, et portée $= 1.4$).

Calcul de l'EI à q points par intégration Monte-Carlo

L'extrapolation du calcul de l'EI à deux points au cas général (q points) fait apparaître des expressions complexes où interviennent des fonctions de répartitions gaussiennes à q dimensions. Il semble ainsi que le calcul de l'EI à q points doive nécessiter d'une façon ou d'une autre un recours à des techniques d'intégration numérique, par quadrature où de type Monte-Carlo.

A ce stade, Il paraît ainsi raisonnable de passer outre tout développement analytique de l'EI à q points et d'utiliser directement la simulation Monte-Carlo pour l'évaluer. Grace à la loi conditionnelle donnée par 3.94 aux équations 3.93, le vecteur aléatoire $[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))|Y(\mathbf{X}) = \mathbf{Y}]$ peut être aisément simulé en utilisant une propriété fondamentale des vecteurs gaussiens :

$$\begin{aligned} &\forall k \in [1, n_{sim}],\ M_k = (m_{OK}(\mathbf{x}^{n+1}), ..., m_{OK}(\mathbf{x}^{n+q})) + [S_q^{\frac{1}{2}} N_k]^T, \\ &N_k \text{ réalisations de vecteurs aléatoires i.i.d. de loi } \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q) \end{aligned} \tag{9.7}$$

où $S_q^{\frac{1}{2}}$ est une »racine carrée « (obtenue par exemple par transformation de Mahalanobis ou par décomposition de Cholesky) de la matrice de covariance conditionnelle (Cf. 3.94) :

$$S_q := Var[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))|Y(\mathbf{X}) = \mathbf{Y}] = \left(k_{KO}(\mathbf{x}^{n+i}, \mathbf{x}^{n+j})\right)_{1 \leq i,j \leq q}.$$

Remarquons que l'on peut calculer l'intégrale de n'importe quelle fonction —non nécessairement linéairement— dépendant de $[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))|Y(\mathbf{X}) = \mathbf{Y}]$ en se basant sur les M_k simulés ci-dessus (probabilités de dépassement de seuil, etc.). Nous donnons en particulier ci-dessous une description algorithmique du calcul de l'amélioration espérée à q points par méthode Monte-Carlo :

function Q-$\mathbb{E}$I($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X}^{new}$, n_{sim})
  $L = \text{chol}(Var[Y(\mathbf{X}^{new})|Y(\mathbf{X}) = \mathbf{Y}])$ ▷ Décomposition de Cholesky de S_q
  for $i \leftarrow 1, n_{sim}$ **do**
    $N \sim \mathcal{N}(0, I_q)$ ▷ Tirage aléatoire d'un vecteur N
    $M_i = m_{OK}(\mathbf{X}^{new}) + LN$ ▷ Simulation de Y en $\mathbf{X}^{new}$
    $qI_{sim}(i)$=$[\min(\mathbf{Y}) - \min(M_i)]^+$ ▷ Simulation de l'amélioration en $\mathbf{X}^{new}$
  end for
  $qEI_{sim} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} qI_{sim}(i)$ ▷ Estimation de l'amélioration espérée
end function

Une application directe de la loi forte des grands nombres donne

$$\sum_{k=1}^{n_{sim}} \frac{[\min(\mathbf{Y}) - min\{M_k(i), i \in [1, q]\}]^+}{n_{sim}} \xrightarrow[n_{sim} \to +\infty]{} EI(\mathbf{x}^1, ..., \mathbf{x}^q) \ p.s. \tag{9.8}$$

On peut finalement utiliser les résultats liés au théorème *central limit* pour contrôler la précision de l'intégration Monte-Carlo en fonction de n_{sim} (voir par exemple [Rip87] pour des détails concernant l'estimation de la variance) :

$$\sqrt{n_{sim}} \left(\frac{qEI_{sim} - EI(\mathbf{x}^1, ..., \mathbf{x}^q)}{\sqrt{Var[I(\mathbf{x}^1, ..., \mathbf{x}^q)|Y(\mathbf{X}) = \mathbf{Y}]}} \right) \xrightarrow[n_{sim} \to +\infty]{} \mathcal{N}(0, 1) \text{ en loi.} \tag{9.9}$$

9.2 Heuristiques basées sur des points pilotes

9.2.1 Optimisations approchées de l'EI à q points

On a présenté dans la dernière sous-section un critère multi-points développé avec pour objectif final de fournir des plan d'expériences additionnels d'un bloc, en résolvant des problèmes d'optimisation de la forme suivante :

$$(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')] \tag{9.10}$$

Le temps de calcul nécessaire pour estimer l'EI à q points avec précision est cependant non-négligeable. De plus, le problème d'optimisation 9.10 est en dimension $d \times q$. Nous essayons ici de trouver des stratégies séquentielles pour approcher la solution de ce problème en évitant le coût computationnel lié à sa résolution directe.

Re-formalisation de l'EI pour l'intégration d'évènements virtuels

Revenons avant toute chose sur les notations. Nous utiliserons tout au long de cette section la notation compacte :

$$EI[Y(\mathbf{Z}) = z](\mathbf{x}) := \mathbb{E}[(min(Y(\mathbf{X}), Y(\mathbf{Z})) - Y(\mathbf{x}))^+ | Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{Z}) = z] \tag{9.11}$$

où $\mathbf{Z}$ représente un ensemble de points de D et z est un vecteur d'images (réelles ou virtuelles) de $\mathbf{Z}$ par y. Par exemple, utiliser ce formalisme pour exprimer q itérations d'EGO (sans actualisation des paramètres de covariance) donne le système

$$\begin{cases} \mathbf{x}^{n+1} = argmax_{\mathbf{x} \in D} EI(\mathbf{x}) = argmax_{\mathbf{x} \in D} EI[\,](\mathbf{x}) \\ \forall j \in [1, q-1],\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x} \in D} EI[Y(\mathbf{x}^{n+j}) = y(\mathbf{x}^{n+j}), \quad \ldots, Y(\mathbf{x}^{n+1}) = y(\mathbf{x}^{n+1})](\mathbf{x}) \end{cases} \tag{9.12}$$

Précisons que ce formalisme reste valable lorsque l'évènement « $Y(\mathbf{Z}) = z$ » est remplacé par un évènement de la forme « $Y(\mathbf{Z})$ ». Par exemple, si $Y(\mathbf{Z})$ est aléatoire,

$$EI[Y(\mathbf{Z})](\mathbf{x}) = \mathbb{E}[(min(Y(\mathbf{X}), Y(\mathbf{Z})) - Y(\mathbf{x}))^+ | Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{Z})] \tag{9.13}$$

devient à son tour une variable aléatoire, dépendant du vecteur aléatoire $Y(\mathbf{Z})$ (conditionnellement à $Y(\mathbf{X}) = \mathbf{Y}$). Ceci constitue la base des stratégies exposées ci-après.

Un plan à q points construit avec l'EI à un point

Au lieu de rechercher le plan des q expériences optimales (au sens de l'EI à q points), i.e. $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q})$, une manière intuitive de le remplacer par une approche séquentielle est de considérer tout d'abord un point qui maximise l'EI à un point,

$$\mathbf{x}^{n+1} = argmax_{\mathbf{x} \in D} EI(\mathbf{x}),$$

puis d'actualiser le modèle, de rechercher $\mathbf{x}^{n+2} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+1})](x)$, etc. Bien entendu, la valeur de $Y(\mathbf{x}^{n+1})$ n'est pas connue à la seconde itération (sinon nous serions dans un algorithme réellement séquentiel, comme EGO). Nous disposons néanmoins de certaines informations à ne pas négliger : le point $\mathbf{x}^{n+1}$ a déjà été visité, et l'on connaît la loi de la v.a. $Y(\mathbf{x}^{n+1})$ conditionnellement à $Y(\mathbf{X}) = \mathbf{Y}$. Plus précisément, cette dernière est $[Y(\mathbf{x}^{n+1})|Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{OK}(\mathbf{x}^{n+1}), s^2_{OK}(\mathbf{x}^{n+1}))$. Le second site $\mathbf{x}^{n+2}$ peut ainsi, en choisissant ici de se ramener à une maximisation de l'espérance du critère conditionnellement à $Y(\mathbf{X}) = \mathbf{Y}$, être pris comme solution du problème

$$\mathbf{x}^{n+2} = argmax_{\mathbf{x}\in D} \mathbb{E}\left[EI[Y(\mathbf{x}^{n+1})](\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}\right] \tag{9.14}$$

On peut ensuite appliquer la même procédure de manière à délivrer q points, en calculant itérativement $\forall j \in [1, q-1]$:

$$\begin{aligned} \mathbf{x}^{n+j+1} &= argmax_{\mathbf{x}\in D} \mathbb{E}\left[EI[Y(\mathbf{x}^{n+j}), \ldots, Y(\mathbf{x}^{n+1})](\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}\right] \\ &= argmax_{\mathbf{x}\in D} \int_{\mathbf{u}\in\mathbb{R}^j} \left(EI[(Y(\mathbf{x}^{n+1}), \ldots, Y(\mathbf{x}^{n+j-1})) = \mathbf{u}](\mathbf{x})\right) f^{1:j}(\mathbf{u}) d\mathbf{u} \end{aligned} \tag{9.15}$$

où $f^{1:j}(.) := f_{(Y(\mathbf{x}^{n+1}),\ldots,Y(\mathbf{x}^{n+j}))|Y(\mathbf{X})=\mathbf{Y}}(.)$ est la densité multi-gaussienne des prédictions conjointes données par le krigeage en $(\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+j})$. Même si l'équation 9.15 permet de définir une version séquentialisée de la maximisation de l'EI à q points, elle ne remplit pas complètement nos objectifs. On se retrouve de nouveau avec une densité multi-gaussienne à intégrer, ce qui semble être un souci typique lorsque l'on traite de problèmes faisant intervenir des vecteurs aléatoires de composantes interdépendantes.

9.2.2 Deux stratégies heuristiques : Constant Liar et Kriging Believer

Le mesonge comme alternative à l'intégration

Afin de contourner la complexité des calculs recontrés dans le paragraphe précédent, nous proposons d'affaiblir l'information prise en compte lors des conditionnements faits à chaque étape. Cette idée a insipiré deux stratégies heuristiques présentées et testées dans les prochaines sous-sections : le *Kriging Keliever* (KB) et le *Constant Liar* (CL).

L'heuristique du « Kriging Believer »

La stratégie du *Kriging Believer* (introduite dans [GLRC07]) consiste à remplacer les lois conditionnelles des réponses aux points choisis au cours des dernières itérations par des valeurs déterministes égales à la moyenne de Krigeage prise aux points en question. En gardant les mêmes notations que précédemment, la stratégie peut être résumée ainsi :

$$\left\{ \begin{array}{r} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}),\ m^n_{OK}(\mathbf{x}^{n+1}) = \mathbb{E}[Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}] \text{ et } \forall j \in [1, q-1] : \\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = m^{n+j-1}_{OK}(\mathbf{x}^{n+j}), ..., Y(\mathbf{x}^{n+1}) = m^n_{OK}(\mathbf{x}^{n+1})](x) \\ m^{n+j}_{OK}(\mathbf{x}^{n+j+1}) = \mathbb{E}[Y(\mathbf{x}^{n+j+1})|Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x}^{n+j}) = m^{n+j-1}_{OK}(\mathbf{x}^{n+j}), \ldots, Y(\mathbf{x}^{n+1}) = m^n_{OK}(\mathbf{x}^{n+1})] \end{array} \right. \tag{9.16}$$

Algorithm 7 L'algorithme du « Kriging Believer » : une première solution approchée au problème multipoints $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')]$

1: **function** KB($\mathbf{X}$, $\mathbf{Y}$, q)
2: **for** $i \leftarrow 1, q$ **do**
3: $\mathbf{x}^{n+i} = argmax_{\mathbf{x} \in D} EI(\mathbf{x})$
4: $m_{OK}(\mathbf{x}^{n+i}) = \mathbb{E}[Y(\mathbf{x}^{n+i}) | Y(\mathbf{X}) = \mathbf{Y}]$
5: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
6: $\mathbf{Y} = \mathbf{Y} \bigcup \{m_{OK}(\mathbf{x}^{n+i})\}$
7: **end for**
8: **end function**

Cette stratégie (pseudo-) séquentielle fournit un plan à q points, et reste abordable puisqu'elle repose sur la maximisation en d dimensions du critère d'EI à un point, connu annalytiquement. Il y a cependant un danger puisque la méthode est susceptible de se laisser emporter par une surface de krigeage qui extrapolerait exagérément les observations (Cf. prochaine sous-section), et on se retrouverait alors avec une suite de points capturés pour un certain temps par une zone artificiellement prometteuse (sans « rappel à l'ordre » puisqu'il n'y a justement pas d'évaluation intermédiaire de la vraie fonction y). On propose maintenant une seconde stratégie dédiée à la réduction de ce danger.

L'heuristique du « Constant liar » :

Considérons maintenant une stratégie dans laquelle le modèle est actualisé à chaque itération avec une valeur fixée de manière exogène par l'utilisateur, et non nécessairement en lien avec le prédicteur de Krigeage. La stratégie appelée « Constant liar » (menteur constant) consiste à mentir avec la même valeur L à chaque itération : maximiser l'EI à un point (trouver x^{n+1}), actualiser le modèle comme si $y(x^{n+1}) = L$, et ainsi de suite, toujours avec le même $L \in R$:

$$\left\{ \begin{array}{c} \mathbf{x}^{n+1} = argmax_{\mathbf{x} \in D} EI(\mathbf{x}) \text{ et } \forall j \in [1, q-1] : \\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x} \in D} EI[Y(\mathbf{x}^{n+j}) = L, ..., Y(\mathbf{x}^{n+1}) = L](\mathbf{x}) \end{array} \right. \tag{9.17}$$

L'effet de la valeur choisie pour L sur les performances de l'algorithme d'optimisation est étudié dans la prochaine sous-section. L devrait en toute logique être déterminé sur la base des valeurs prises par y au plan d'expériences initial. On considère ici trois valeurs : $min\{\mathbf{Y}\}$, $mean\{\mathbf{Y}\}$, et $max\{\mathbf{Y}\}$. On peut s'attendre à ce que l'algorithme soit d'autant plus exploratoire que L est grand (effet de répulsion).

Algorithm 8 L'algorithme du « Constant Liar » : une seconde solution approchée au problème multipoints $(\mathbf{x}^{'n+1}, \mathbf{x}^{'n+2}, ..., \mathbf{x}^{'n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')]$

1: **function** CL($\mathbf{X}$, $\mathbf{Y}$, L, q)
2: **for** $i \leftarrow 1, q$ **do**
3: $\mathbf{x}^{n+i} = argmax_{\mathbf{x} \in D} EI(\mathbf{x})$
4: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
5: $\mathbf{Y} = \mathbf{Y} \bigcup \{L\}$
6: **end for**
7: **end function**

9.3 Comparaisons expérimentales

9.3.1 Application à la fonction de Branin-Hoo

Nous allons maintenant comparer les quatre stratégies d'optimisation sur la fonction de Branin-Hoo ([JSW98],[Sch97],[QVPH06]).

$$\begin{cases} y_{BH}(x_1, x_2) = (x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1 - \frac{1}{8\pi})cos(x_1) + 10 \\ x_1 \in [-5, 10],\ x_2 \in [0, 15] \end{cases} \tag{9.18}$$

y_{BH} possède trois optimiseurs globaux $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$, et le minimum global est approximativement égal à 0.4. Les variables sont ici normalisées par les transformations affines $x_1^{'} = \frac{x_1+5}{15}$ and $x_2^{'} = \frac{x_2}{15}$. Le plan d'expériences initial $\mathbf{X}_9$ est un plan factoriel complet 3×3 (Cf. 9.3), et ainsi $\mathbf{Y} = y_{BH}(\mathbf{X}_9)$. Le Krigeage Ordinaire est fait avec un noyau de covariance gaussienne stationnaire anisotrope.

$$\forall h = (h_1, h_2) \in \mathbb{R}^2,\ k(h_1, h_2) := \sigma^2 e^{-\theta_1 h_1^2 - \theta_2 h_2^2} \tag{9.19}$$

où les paramètres (θ_1, θ_2) sont fixés à leur valeur $(5.27, 0.26)$ estimée par MV, et σ^2 est estimé comme dans [JSW98] via l'équation 8.6. Nous avons construit un plan d'expériences (supplémentaire) à 10 points pour chacune des stratégies. Nous avons de plus estimé la probabilité d'amélioration et l'amélioration espérée associées aux q premiers points (avec $q \in \{2, 6, 10\}$) de chaque stratégie par simulations Monte-Carlo.

Les quatre stratégies (KB et les trois variantes de CL) ont donné des résultats (plans d'exploration) et des performances (en optimisation) hétérogènes. On a observé d'une part dans la construction des séquences données par le *Constant Liar* (CL) que les points déjà choisis exerçaient une forme de répulsion lors du choix des nouveaux points.

Des valeurs du mensonge L raisonnablement grandes (telles que les valeurs $\max(\mathbf{Y})$ et $mean(\mathbf{Y})$ considérées ici) forcent la séquence à explorer l'espace en évitant $\mathbf{X_9}$. Les deux

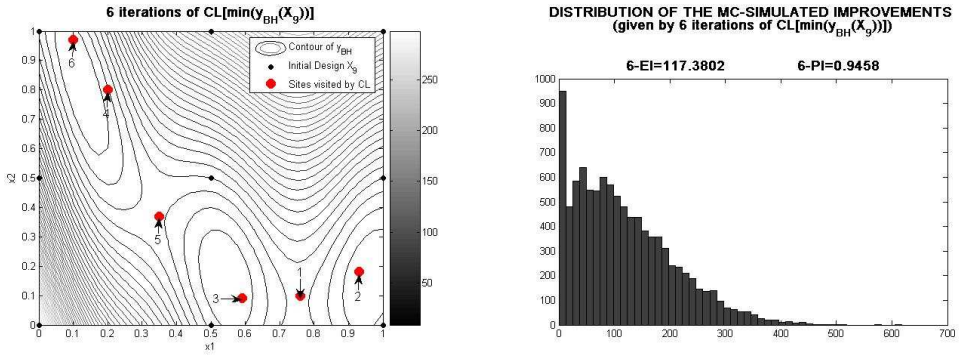


FIG. 9.3 – A gauche : contours de y_{BH} avec le plan initial $\mathbf{X}_9$ (points noirs) et les 6 premiers points donnés par la stratégie CL[min($y_{BH}(\mathbf{X}_9)$)] (points rouges). A droite : histogramme de 10^4 valeurs d'amélioration apportées par les 6 premiers points de la stratégie CL[min($y_{BH}(\mathbf{X}_9)$)], simulées par Monte Carlo. Les estimations correspondantes de l'amélioration espérée et de la probabilité d'amélioration sont données au dessus.

stratégies ont fourni des plans d'exploration avec de bonnes propriétés de remplissage de l'espace, de grandes probabilités d'amélioration (PI à 10 points proche de 100%), et des valeurs prometteuses pour l'EI à q points (Cf. 9.3.1). *In fine*, ils ont apporté des améliorations respectives de 7.86 et 6.25.

De toutes les stratégies testées, c'est CL[min($\mathbf{Y}$)] qui a ici donné les meilleurs résultats. En 6 itérations, elle a visité les trois bassins des minima de la fonction y_{BH}. En 10 itérations, elle a donné la meilleure amélioration globale parmi les stratégies, ce qui est de plus en accord avec les améliorations espérées à 10 points simulées par Monte-Carlo. Il semble en effet que la répulsion "douce" lorsque $L = \min(\mathbf{Y})$ soit le bon réglage pour l'optimisation de la fonction de Branin-Hoo avec $\mathbf{X}_9$ pour plan d'expériences initial.

D'autre part, la stratégie du *Kriging Believer* a donné ici des résultats décevants. Tous les points (sauf un, le dernier visité) se sont retrouvés agglomérés autour de $\mathbf{x}^{n+1}$, le premier point visité (le même que dans la stratégie CL, par construction). Cela peut s'expliquer par la prédiction exagérément basse donnée par le Krigeage au point en question : le prédicteur m_{OK} passe largement en dessous des observations (à cause d'une certaine rigidité de la covariance gaussienne), et l'amélioration espérée devient abusivement grande au voisinage de $\mathbf{x}^{n+1}$. Le point $\mathbf{x}^{n+2}$ est ensuite choisi proche de $\mathbf{x}^{n+1}$, et ainsi de suite. L'algorithme reste bloqué au premier point visité. Le KB se comporte en fait comme

	CL[$min(\mathbf{Y})$]	CL[$mean(\mathbf{Y})$]	CL[$max(\mathbf{Y})$]	KB
PI (2 prem. points)	87.7%	87%	88.9%	65%
EI (2 prem. points)	114.3	114	113.5	82.9
PI (6 prem. points)	94.6%	95.5%	92.7%	65.5%
EI (6 prem. points)	117.4	115.6	115.1	85.2
PI (10 prem. points)	99.8%	99.9%	99.9%	66.5%
EI (10 prem. points)	122.6	118.4	117	85.86
Improvement (6 prem. points)	7.4	6.25	7.86	0
Improvement (10 prem. points)	8.37	6.25	7.86	0

TAB. 9.1 – PI, EI, et améliorations réelles multipoints pour les 2, 6, et 10 premières itérations des straégies heuristiques CL[min($\mathbf{Y}$)], CL[$mean(\mathbf{Y})$], CL[max($\mathbf{Y}$)], et KB (avec ici min($\mathbf{Y}$) = min($y_{BH}(\mathbf{X}_9)$)). Les critères PI et EI à q points sont évalués par simulation Monte-Carlo (Cf. 9.8).

le ferait CL avec une constante L largement en-dessous de min($\mathbf{Y}$). Comme on peut le voir dans le tableau 9.3.1 (dernière colonne), le phénomène peut être anticipé à l'aide des critères PI et EI à q points : ils restent presque constants lorsque q augmente. Cela illustre en particulier en quoi les critères à q points permettent de rejeter des stratégies inappropriées.

Les résultats présentés dans le tableau 9.3.1 font ressortir un inconvénient majeur du critère de PI à q points. Lorsque q augmente, les probabilités d'amélioration associées aux trois stratégies CL convergent rapidement vers 100%, de telle manière qu'il n'est pas possible de discriminer les "bons plans" d'expériences des "très bons plans" d'expériences. L'EI à q points est un critère plus fin grace à la prise en compte de l'amplitude des améliorations potentielles apportées par les plans. Il semblerait néanmoins que l'EI surévalue ici l'amélioration apportée par les plans considérés. Cet effet, déjà mis en évidence dans [Sch97], peut être expliqué en se référant à la fois à la valeur élevée du σ^2 estimé sur la base de $\mathbf{Y}$ et à la faible différence entre le minimum atteint sur le plan d'expériences initial $\mathbf{X}_9$ (9.5) et le vrai minimum de y_{BH} (0.4).

9.3.2 Kriging-based optimization of gaussian process realizations

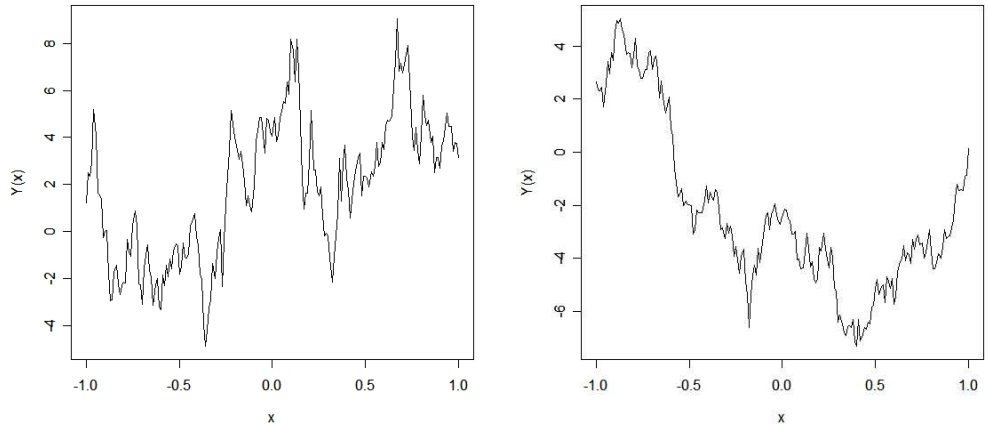


FIG. 9.4 – Two stationary Gaussian Process paths (both centered, with variance 10 and exponential covariance structure with respective correlation lengths 0.2 and 0.7). This family of Gaussian Process is often referred to as *Ornstein-Uhlenbeck* Process [RW06].

With the intent to produce general results, we chose to study and compare the 3 heuristics KB, CL[min $\mathbf{Y}$], and CL[max $\mathbf{Y}$] presented in 9.2.2. in applying them to random functions. Gaussian Process simulation is a handy way to work with such functions.[3]. We considered four experimental configurations (denoted by $k \in [1,4]$) involving Gaussian Processes $Y^k(x)$ and 1000 realizations $\{y_i^k(x), i \in [1,1000]\}$ of them for each configuration. For all configurations, the outputs varied between -1 and 1 ($D = [-1,1]$), and the initial design of experiments was fixed to the set $\mathbf{X} = \{-1,0,1\}$ (see 9.5). The other experimental parameters varied accordingly to the values specified in Table 9.3.2.

Let us now fix $k \in \{1,2,3,4\}$ for the sake of clarity. Formally, each heuristic strategy $\mathcal{S}$ (here $\mathcal{S} \in \{KB, CL[min], CL[max]\}$) provides a sequence of points $X^{k,1}(\mathcal{S}), \ldots, X^{k,N_k}(\mathcal{S})$. These points are random variables since they closely depend on the process Y^k which is itself random. Here we wish to study the performances of each strategy (given a configuration) by looking at the behavior of the random variable $\Delta_k(\mathcal{S})$:

[3] simulating mono- or multi-dimensional GPs on a grid (having m elements) is theorically (but not always numerically) straightforward, the cost being the inversion of an $m \times m$ covariance matrix.

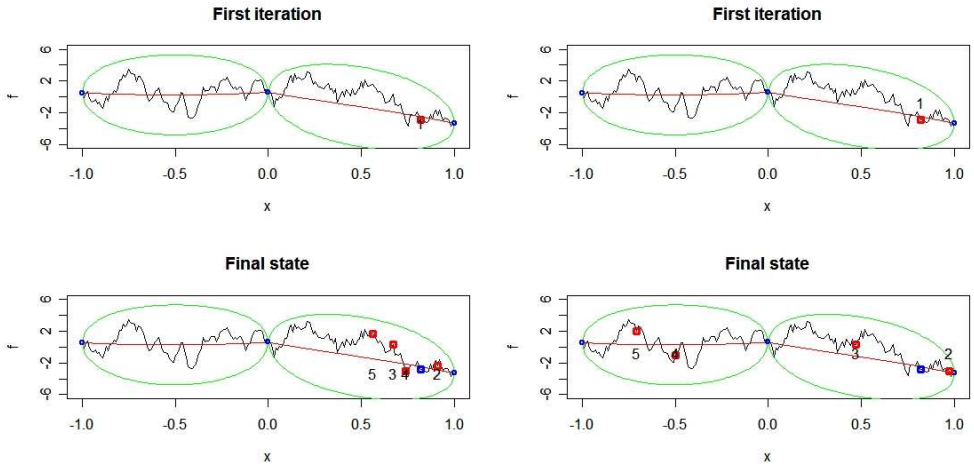


FIG. 9.5 – 5 iterations of CL[min(**Y**)] (left) and CL[max(**Y**)] (right) to one Gaussian Process path (scale 0.7, variance 10, exponential covariance). This example clearly illustrates how the CL strategy privileges local search whenever $L = \min(\mathbf{Y})$, and has a more space-filling behaviour when $L = \max(\mathbf{Y})$.

k	covariance	correlation length	variance	N_k
1	Exponential	0.3	40	2
2	Exponential	1	40	2
3	Exponential	0.3	40	10
4	Exponential	1	40	10

TAB. 9.2 – Design of experiments for a comparison between the 3 heuristics

$$\Delta_k(\mathcal{S}) := \min\{Y^k(\mathbf{X}), Y^k(X^{k,1}(\mathcal{S})), ..., Y^k(X^{k,N_k}(\mathcal{S}))\} - \min_{x\in D}[Y^k(x)] \geq 0, \qquad (9.20)$$

which measures how far we are from having perfectly optimized the process $Y^k(x)$ after having ran N_k iterates of the strategy $\mathcal{S}$. Hence, the closer the realizations of $\Delta_k(\mathcal{S})$ are to 0, the better $\mathcal{S}$ fullfils its goals as optimizer. We studied the experimental performances of the three algorithms applied to the 1000 realizations ran for every configuration k.

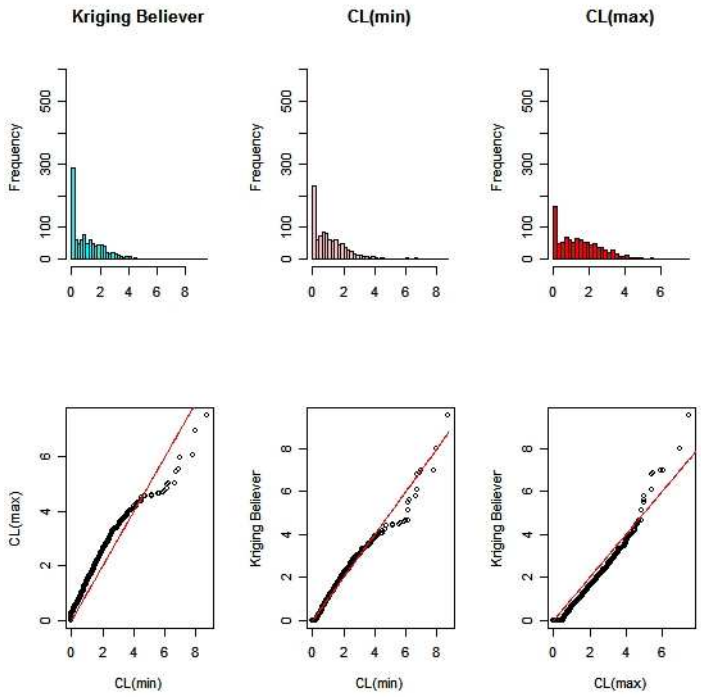


FIG. 9.6 – Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configuration 4. $CL[max]$ and KB keep their positions of best performers, respectively at the right and left extremes. Note the particular shape of the qq-plot between $CL[min]$ and $CL[max]$. The first one is statistically more likely to perform very well, but also more likely to fail dramatically. Conversely to configuration 1, $CL[max]$ is here a good challenger for a risk averse user.

We considered the corresponding 1000 realizations of $\Delta_k(\mathcal{S})$, denoted by

$$\delta_k^i(\mathcal{S}) = \min\{y_i^k(\mathbf{X}), y_i^k(x_i^{k,1}(\mathcal{S})), ..., y_i^k(x_i^{k,N_k}(\mathcal{S}))\} - \min_{x\in D}[y_i^k(x)] \geq 0, \tag{9.21}$$

where the $x_i^{k,j}(\mathcal{S})$'s stand for the 1000 realizations of the $X^{k,j}(\mathcal{S})$'s. The results are summarized in fig. 9.6. The histograms offer concentrated representations of the δ_k's distributions, i.e. the statistical performances of each strategy in all studied configurations. Values near 0 (on the extreme left of the histograms) mean succesful optimizations, whereas right tails stand for the cases of failure (best y value observed far beyond the actual minimum). The *q-q plots* aim at comparing all couples of strategy in plotting the empirical quantiles (i.e. ranked values of the δ_k^i's) of the one against the empirical quantiles of the other. Such kind of graphic allows a far more subtile comparison between strategies

than only scalar indexes like the mean or the median performances. As shown on 9.6, the Kriging Believer strategy does not behave pathologically anymore when using the exponential covariance : it seems in fact to give optimization results with a very good balance between high performances and risk covering.

Even if the three strategies roughly give comparable results within this example, $CL[min]$ and KB appear indeed to provide more often extremely good results (small δ's) than $CL[max]$, which however has thiner tails than $CL[min]$. Note that if the comparaison between strategies is quite stable for small values of δ, this statement doesn't hold for high quantiles since the corresponding fluctuations are too large for samples of 1000 process realizations.

9.3.3 Optimisation parallèle de la fonction de Hartmann (6D)

Nous proposons dans cette section le résumé d'une étude des performances d'EGO et de ses version parallélisées dans le cadre d'une application à une fonction numérique à 6 arguments, la fonction dite « Hartmann 6 », notée $y_{H6} : [0,1]^6 \longrightarrow \mathbb{R}$.

Définition de la fonction Hartmann6

$$\forall x \in [0,1]^6,\ y_{H6}(x) = -\sum_{j=1}^{4} c_j \times \exp\left(-\sum_{i=1}^{6} a_{i,j} \times (x_i - p_{i,j})^2\right)$$

$$a = \begin{pmatrix} 10.00 & 0.05 & 3.00 & 17.00 \\ 3.00 & 10.00 & 3.50 & 8.00 \\ 17.00 & 17.00 & 1.70 & 0.05 \\ 3.50 & 0.10 & 10.00 & 10.00 \\ 1.70 & 8.00 & 17.00 & 0.10 \\ 8.00 & 14.00 & 8.00 & 14.00 \end{pmatrix} \quad p = \begin{pmatrix} 0.1312 & 0.2329 & 0.2348 & 0.4047 \\ 0.1696 & 0.4135 & 0.1451 & 0.8828 \\ 0.5569 & 0.8307 & 0.3522 & 0.8732 \\ 0.0124 & 0.3736 & 0.2883 & 0.5743 \\ 0.8283 & 0.1004 & 0.3047 & 0.1091 \\ 0.5886 & 0.9991 & 0.6650 & 0.0381 \end{pmatrix} \quad c = \begin{pmatrix} 1.0 \\ 1.2 \\ 3.0 \\ 3.2 \end{pmatrix}$$

minimum global $= -\mathbf{3.32}$

minimiseur global $= [0.202, 0.150, 0.477, 0.275, 0.312, 0.657]$

Dans les deux sections à venir, nous considérerons deux plans d'expériences initiaux :

$$\begin{aligned} \mathbf{X}_{20} &= \{\mathbf{x}^1, \ldots, \mathbf{x}^{10}\} \in \left([0,1]^6\right)^{10} \\ \mathbf{X}_{50} &= \{\mathbf{x}^1, \ldots, \mathbf{x}^{10}, \mathbf{x}^{11}, \ldots, \mathbf{x}^{50}\} \in \left([0,1]^6\right)^{50} \end{aligned} \tag{9.22}$$

où les $\mathbf{x}^i$ ($i \in \{1, \ldots, 50\}$) sont les réalisations de variables aléatoires i.i.d. de loi uniforme sur $[0,1]^6$. Voyons dans un premier temps les résultats que donne EGO en partant des deux plans d'expériences $\mathbf{X}_{50}$ (avec un crédit de 20 évaluations, Cf. 9.7) et $\mathbf{X}_{10}$ (avec un crédit de 90 évaluations, Cf. 9.8)).

Application d'EGO avec 2 plans d'expériences initiaux

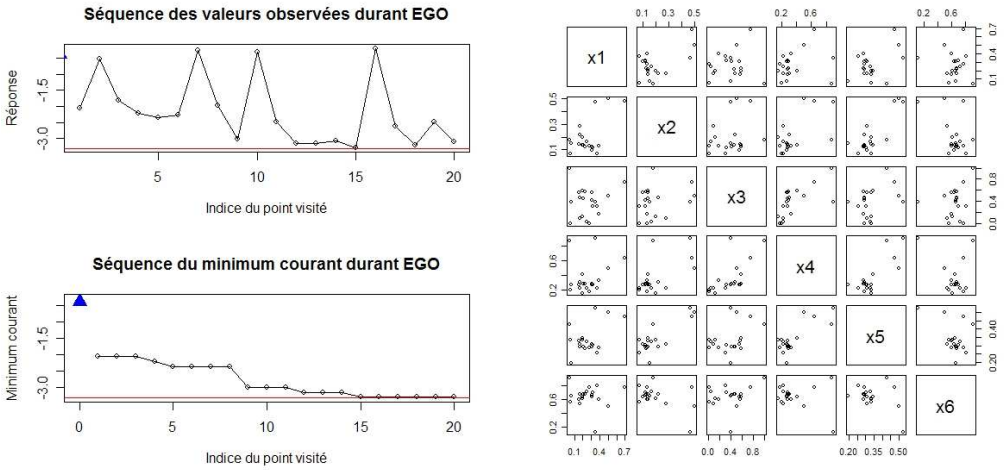


FIG. 9.7 – En haut à gauche : en démarrant avec un plan à 50 points, l'optimum global est atteint en 15 itérations. En bas à gauche : EGO visite séquentiellement la zone du minimum global de Hartmann. Il exploite l'information donnée par le plan initial. A droite : la représentation par projections du plan d'expériences généré par EGO permet de deviner la visite du bassin de minimum, suivie d'une nouvelle phase exploratoire.

L'application d'EGO avec noyau gaussien est concluante dans le cas $\mathbf{X} = \mathbf{X}_{50}$, comme dans le cas $\mathbf{X} = \mathbf{X}_{10}$. Il est intéressant de comparer les nombres totaux d'évaluations nécessités dans un cas comme dans l'autre pour atteindre le minimum global de y_{H6} (avec 1% de tolérance) : $50 + 15 = 65$ évaluations dans le premier cas, contre $10 + 36 = 46$ dans le second. Ceci illustre le fait qu'il est parfois judicieux de se contenter d'un plan d'expériences initial modeste pour investir plus dans la stratégie d'optimisation proprement dite. Cela dit, consacrer 37 unités de temps à l'optimisation n'est pas forcément abordable. Nous allons voir dans la suite de cette étude que l'algorithme CL permet de diminuer très nettement ce délai. Par ailleurs, la figure 9.9 résume une tentative d'optimisation de y_{H6} par Monte-Carlo « pur », se basant sur 500 évaluations de la fonction. Les résultats obtenus illustrent le fait que y_{H6} n'est pas du tout évidente à minimiser.

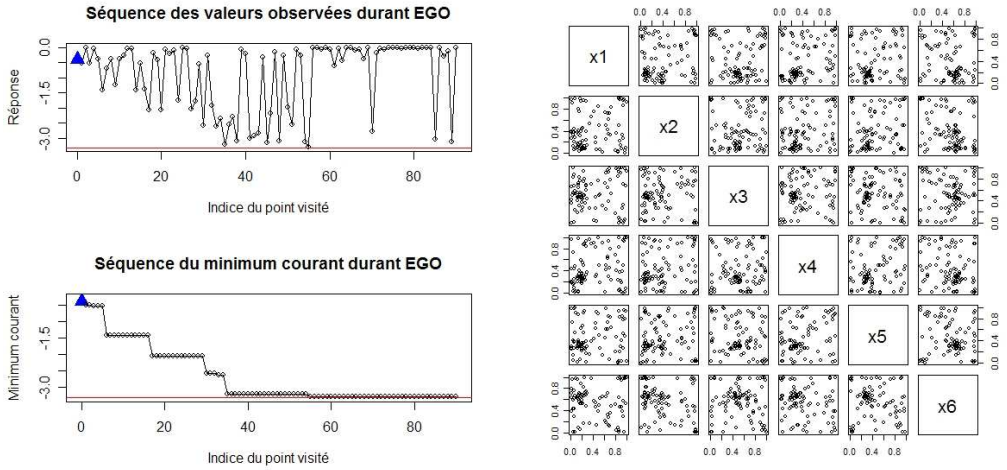


FIG. 9.8 – En haut à gauche : La sequence de points a un comportement nettement plus exploratoire qu'en partant de 50 points. En bas à gauche : EGO trouve le minimum en 36 itérations. C'est peu si l'on considère que le plan initial ne compte que 10 points. A droite : la vue en projections du plan à 90 points généré par EGO illustre le cheminement assez lent des itérés vers le minimiseur global, suivi d'une phase exploratoire.

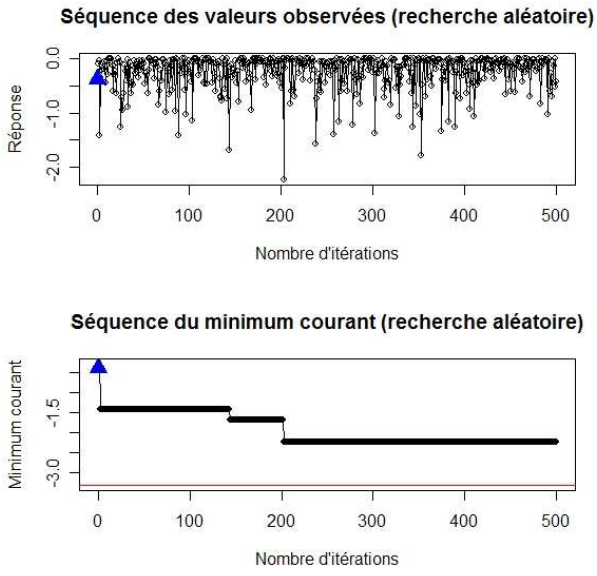


FIG. 9.9 – Résultats de minimisation obtenus par méthode Monte-Carlo. On ne trouve ici pas mieux que -2.2 en 500 itérations, alors que le minimum global est -3.32.

Application à Hartmann6 de différentes parallélisations d'EGO

Nous présentons ici les résultats obtenus en appliquant différentes versions séquentialisées de l'algorithme « Constant Liar » à Hartmann6. L'idée est de permettre une optimisation parallèle synchrone en utilisant CL à chaque itération pour obtenir n_{proc} points d'explorations. Les mensonges du CL sont corrigés à la fin de chaque itération, après les évaluations parallèles du simulateur. $CL[min]$ séquentialisée est appliquée en partant du plan $\mathbf{X}_{50}$ avec $q = 2$ (fig. 9.10), et $q = 10$ (fig. 9.11), en partant du plan $\mathbf{X}_{10}$ avec $q = 10$ (fig. 9.12), et $CL[max]$ en partant du plan $\mathbf{X}_{10}$ avec $q = 10$ (fig. 9.13).

Algorithm 9 Version séquentialisée de l'algorithme du « Constant Liar ».

function CL.STAGES($\mathbf{X}$, $\mathbf{Y}$, y, n_{proc}, n_{it})
  for $i \leftarrow 1, n_{it}$ **do**
    $L = min(\mathbf{Y})$
    for $j \leftarrow 1, n_{proc}$ **do**
      $\mathbf{x}^j_{new} = argmax_{\mathbf{x} \in D} \mathbb{E}I(\mathbf{x})$ ▷ avec $\mathbf{X}$ et $\mathbf{Y}_{CL}$
      $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^j_{new}\}$
      $\mathbf{Y}_{CL} = \mathbf{Y} \bigcup \{L\}$
    end for
    $\mathbf{Y} = \mathbf{Y} \bigcup y(\{\mathbf{x}^1_{new}, \ldots, \mathbf{x}^{n_{proc}}_{new}\})$ ▷ Evaluations du simulateur
    Ré-estimation du Krigeage
  end for
end function

Les résultats présentés sur les figures 9.10, 9.11, 9.12, 9.13 illustrent les gains importants obtenus en termes de *temps effectif* grâce à la parallélisation des heuristiques CL. Si dédoubler la capacité de recherche permet de réduire sensiblement le temps effectif d'étude (de 15 à 10 itérations en doublant le temps CPU, Cf. 9.10), la décupler permet de diviser ce temps par quatre (Cf. 9.11 ; on y voit aussi à la $3^{\text{ème}}$ séquence comment l'algorithme CL proscrit une zone entière après dix evaluations), voire davantage lorsque le plan initial est plus pauvre $\mathbf{X}_{10}$ (Cf. 9.12). $CL[max]$ donne ici à contexte identitique des performances très similaires à celles de $CL[min]$ (Cf. 9.13). D'une manière générale, le fait de procéder à un nombre plus ou moins grand d'évaluations simultanées dans une même zone permet de gagner du temps en « apprenant » en détail un bassin de minimum, ou respectivement en découvrant puis en s'assurant que la zone en question ne présente pas vraiment d'intérêt à être visitée plus longuement. Les compromis possibles entre les valeurs des mensonges L, le nombre de processeurs en parallèle, et le nombre total d'évaluations consituent une des suites potentiellement intéressantes à ce travail.

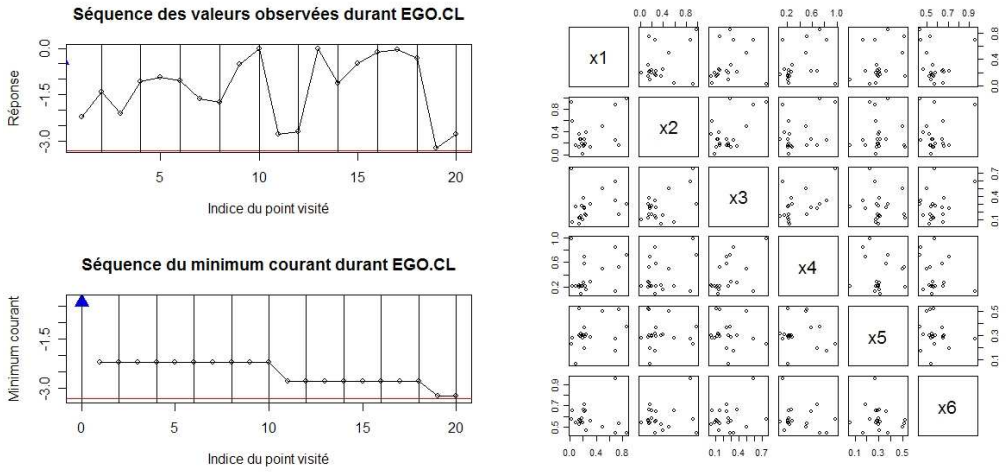


FIG. 9.10 – En haut : En utilisant deux processeurs, on trouve ici le minimum global en 10 unités de temps au lieu de 15 avec un seul processeur. En bas : Le suivi des points en projection permet de constater qu'ils ne sont pas systématiquement en clusters de 2. Il y a un compromis entre visite des zones prometteuses et exploration de l'espace.

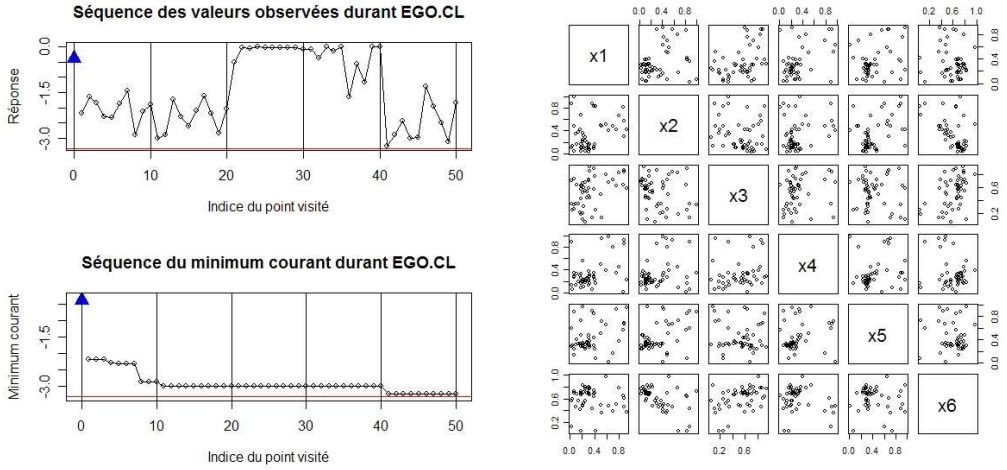


FIG. 9.11 – En haut : On peut encore réduire les délais : avec 10 processeurs en parallèle, le minimum est atteint en 5 unités de temps. En bas : l'algorithme alterne ici entre une première phase d'exploitation (deux premières unités de temps), une phase plus exploratoire ($3^{\text{ème}}$ et $4^{\text{ème}}$ unités de temps), puis une phase finale d'exploitation durant laquelle il trouve le minimum.

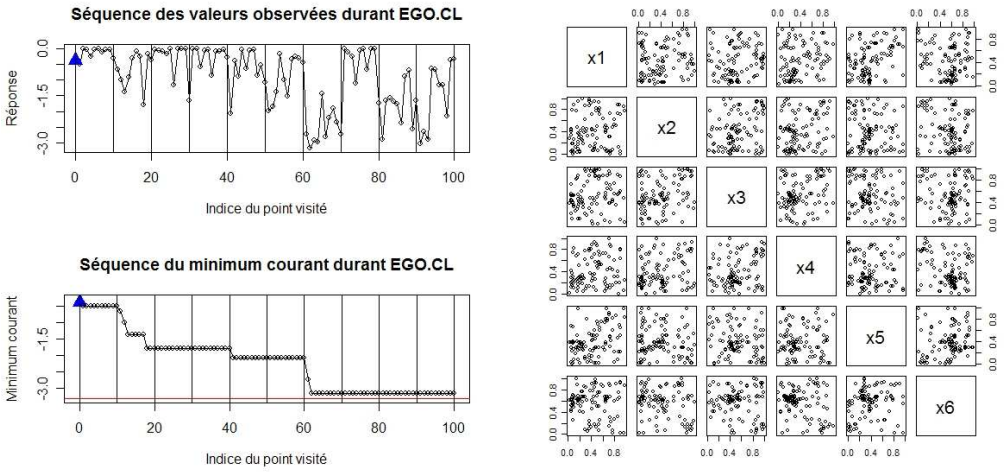


FIG. 9.12 – En haut : en partant d'un plan à 10 points, CL_{min} avec 10 processeurs permet de trouver le minimum en 7 unités de temps. En bas : l'algorithme commence par explorer et condamner des zones peu prometteuses, puis trouve la zone optimale et la visite jusqu'à convergence. Il repart ensuite en exploration.

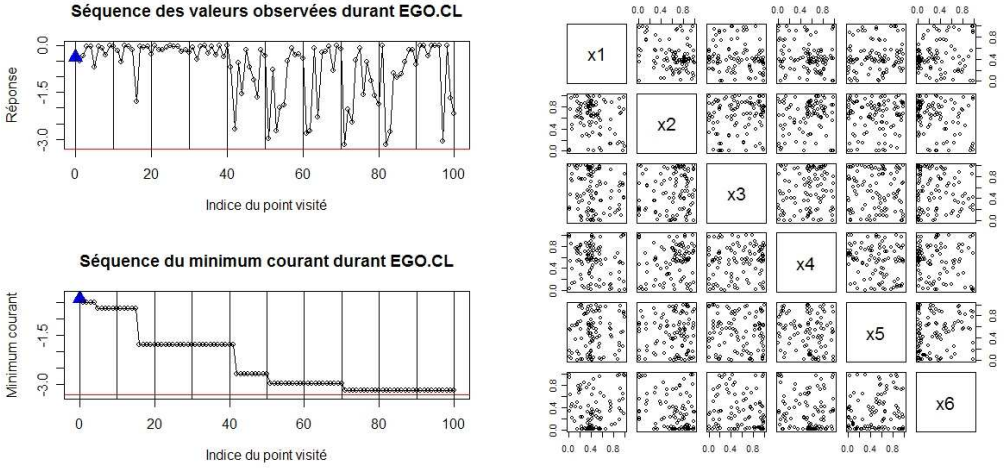


FIG. 9.13 – En haut : CL_{max} donne aussi des résultats satisfaisants sur ce cas. Le réglage de L offre de nombreuses perspectives de recherche. En bas : la répulsion plus forte entre les points visités est apparente. Les plans délivrés par CL_{max} se rapprochent d'avantage de *space filling designs*.

Chapitre 10

Conclusions et perspectives

Les métamodèles probabilistes de type « Krigeage » possèdent une grande versatilité, leur permettant non seulement d'intégrer des tendances déterministes diverses (Cf. Krigeage Universel et son extension au chapitre 6), mais aussi de retranscrire finement des niveaux de régularité souhaités via un choix de noyau *ad hoc* [Abr97], ou encore de respecter certaines invariances algébriques, comme on l'a montré au chapitre 7.

Par ailleurs, les résultats d'équivalence liant le Krigeage avec la régularisation et les splines mis en lumière par Wahba [Wah90], ainsi que les liens ténus existants avec l'approximation linéaire (Cf. chapitre 2 pour ces deux points) donnent à penser que bon nombre de méthodes d'approximation fonctionnelle peuvent s'inscrire dans un même paradigme. La pierre angulaire en est sans doute la notion très générale de noyau, qu'il s'agisse de noyaux reproduisants ou de noyaux de covariance, tous deux étant d'ailleurs omniprésents dans les travaux contemporains d'apprentissage statistique [HTF01, Ver07]. Vu sous cet angle, le Krigeage peut ainsi apparaître comme une instance probabilisée des méthodes classiques d'approximation fonctionnelle, pour laquelle la modélisation en termes de processus gaussiens [RW06] (Cf. chapitre 3) constitue à la fois un socle historique et un cadre pratique très commode.

En particulier, la dérivation de lois conditionnelles gaussiennes a permis l'introduction de critères d'exploration calculables analytiquement tels que l'*amélioration espérée* [JSW98] (Cf. chapitre 4), ayant rendu possibles des implémentations efficaces d'algorithmes tels qu'EGO pour l'optimisation globale de simulateurs numériques coûteux à évaluer. C'est en exploitant le caractère multigaussien des prédictions jointes données par un métamodèle de Krigeage que l'on a pu considérer au chapitre 9 une généralisation multipoints de l'*amélioration espérée*, dédiée au calcul distribué. Les difficultés math-

ématiques inhérentes à l'optimiation d'un tel critère multipoints ont motivé l'introduction de stratégies heuristiques de type « glouton », le *Kriging Believer* puis le *Constant Liar*. Les premiers résultats obetnus sur des cas test ont nettement confirmé l'intérêt de telles méthodes, qui sont aujourd'hui destinées à être enrichies et étudiées sur des problèmes de dimension supérieure.

Ces développement en optimisation sont bien entendu à considérer en gardant à l'esprit que les procédures sont toutes tributaires à la fois du choix d'une famille de Krigeages adaptée (tendances, structures de covariances, Cf. chapitres 6, 7, et 8), et d'une estimation réussie des paramètres de modèle, au premier plan desquels les longueurs de corrélation des noyaux de covariance. Une sous-estimation de ces derniers peut en effet s'avérer dramatique en optimisation, puisque la recherche ne sera alors plus que très locale autour des points du plan d'expériences (Cf. chapitre 14 en annexe pour un début de discussion). L'analyse menée au chapitre 5 au sujet des propriétés de l'estimateur du maximum de vraisemblance avec peu d'observations suggère de prendre des précautions lors de l'utilisation de méthodes automatiques pour l'estimation des paramètres du Krigeage. Approximation parfois grossière de la variance d'estimation, la matrice de dispersion de Fisher semble toutefois constituer un outil intéressant pour guider de nouvelles évaluations et en vue d'améliorer l'estimation.

Pour finir, les mélanges discrets de Krigeages présentés au chapitre 8 apparaîssent comme un moyen intéressant de prendre en compte l'incertitude de modèle, à la fois en prédiction et au sein d'algorithmes d'exploration. La gestion adaptative des poids de modèles permet en effet de transformer le cadre classique de sélection de métamodèle ainsi que les dangers liés à des décisions brutales sur la base d'un plan d'expériences initial en un cadre intégré, dans lequel la pertinence relative des différents métamodèles est ré-évaluée au fil de l'acquisition de nouvelles observations. Au delà du rapport de vraisemblance dont l'applicabilité est illustrée ici, l'usage d'autres procédures de pondération (validation croisée ou autres) dédiées à la nature de l'application traitée reste à investiguer.

Quatrième partie

Annexes

Chapitre 11

Articles

11.1 Choix et estimation d'un modèle de Krigeage

1. Présentation en conférence internationale avec comité de lecture

   - ENBIS-DEINDE « Computer versus Physical Experiments »
     - Lieu et date : Turin (Italie) Avril 2007.
     - Proceedings : oui (Cf. [GDB+07]).

2. Publication de revue scientifique

   - Journal : *Applied Stochastic Models for Business and Industry* (Wiley).
   - Statut : version finale acceptée en Aout 2008, à paraître.
   - pré-print disponible sur HAL.

A note on the choice and the estimation of Kriging models for the analysis of deterministic computer experiments

David Ginsbourger *, Delphine Dupuy*, Anca Badea*, Laurent Carraro*, Olivier Roustant*

April 3, 2008

ABSTRACT

Our goal in the present work is to give an insight on some important questions to be asked when choosing a Kriging model for the analysis of numerical experiments. We are especially concerned about the cases where the size of the design of experiments is small relatively to the algebraic dimension of the inputs. We first fix the notations and recall some basic properties of Kriging. Then we expose two experimental studies on subjects that are often skipped in the field of computer simulation analysis: the lack of reliability of likelihood maximization with few data, and the consequences of a trend misspecification. We finally propose an example from a porous media application, with the introduction of an original Kriging method in which a non-linear additive model is used as external trend.

Keywords: Metamodeling, Kriging, Maximum Likelihood, Determinisitic Drift, Additive Models

1. LINEAR PREDICTORS FOR SPATIAL INTERPOLATION OF NUMERICAL SIMULATORS

We study a deterministic numerical simulator as a function $z : D \subset \mathbb{R}^d \to \mathbb{R}$, where $\mathbf{x} \in D$ is the vector of inputs variables. We denote the set of the design

*Département 3MI , Ecole Nationale Supérieure des Mines, 158 cours Fauriel, 42023 Saint-Etienne (France), tel. +33 04 77 49 97 57, e-mail: ginsbourger@emse.fr

points (or "design") by $\mathbf{X} = \{\mathbf{x^1}, ..., \mathbf{x^n}\}$ and by $\mathbf{Z} = \{z(\mathbf{x^1}), ..., z(\mathbf{x^n})\}$ the set of simulator responses associated with $\mathbf{X}$. Kriging is a class of methods coming from the field of geostatistics [15, 3], known as *linear optimal prediction* in classical statistics. It provides at each point $\mathbf{x} \in D$ a prediction $\hat{Z}(\mathbf{x})$ linearly depending on $\mathbf{Z}$, where the weights depend on the design and on the Kriging model but not on the observations. The way the weights are defined varies as a function of the type of Kriging -Simple (SK), Ordinary (OK), Universal (UK), etc- and many parameters such as the trend functions, the covariance kernel and their own parameters: threshold (or "sill"), scales, nugget, etc... denoted by the r-dimensional vector $\boldsymbol{\psi}$. In the following, we will concentrate on the parameters of sill and scale ($r = 2$), denoted respectively either by $\psi_1,\ \psi_2$ or by $\sigma^2,\ p \in [0, +\infty[$. Most classic Kriging types (including SK, OK, UK, and more) can be interpreted as random process interpolation relying on the assumption that:

$$\forall \mathbf{x} \in D,\ z(\mathbf{x}) = t(\mathbf{x}) + \varepsilon(\mathbf{x}) \tag{1}$$

where t is a numerical deterministic function and $\varepsilon(\mathbf{x})$ is one path of a centered stationary Gaussian Process (GP) with known stationary covariance kernel $k : h \in \mathbb{R}^d \longrightarrow k(h) \in \mathbb{R}$. t is generally known up to a set of parameters or a semi-parametric structure to be estimated within Kriging. Several founder works [19, 7] on the application of Kriging to computer simulations start off with an extremely simplified version of (eq.1). They assume that the trend is an unknown constant (Ordinary Kriging, i.e. $t(x) = \mu \in \mathbb{R}$) and that k is a generalized exponential kernel [20], letting the stochastic part of (eq.1) account for the variability of z. Then the covariance parameters $\boldsymbol{\psi}$ are estimated by maximizing the Gaussian likelihood of the observations $\mathbf{Z}$. On the other hand, recent approaches [8, 14] try to take advantage of more complex trends, from linear and polynomial functions to Fourier series. In other respects, [13] as well as [16] present an application of bayesian analysis to Kriging interpolation of computer codes.

The motivation of this article is to raise some basic questions that should become crucial when applying Kriging techniques with few observations regarding the dimension of inputs, which is quite often the case in numerical simulation. The two coming sections, based on toy experiments, put a focus on the estimation of the covariance parameters $\boldsymbol{\psi}$ and on the choice of the trend t. The two following sections are dedicated at presenting an original combination of additive models and Simple Kriging, with a heuristic fitting methodology. The efficiency of this technique is illustrated on a 3-dimensional example from a porous media simulation test case.

2. FITTING COVARIANCE PARAMETERS BY MLE WITH A SMALL SAMPLE

The Maximum Likelihood (ML) estimation method is widely used in Kriging to choose covariance parameters on the basis of observations. Following the assumptions from (eq. 1), ML estimation relies on the maximization of the density of the observed values $\mathbf{Z}$, seen as a function of the vector $\boldsymbol{\psi}$:

$$L(\boldsymbol{\psi};\mathbf{Z}) := f(\mathbf{Z}|\boldsymbol{\psi}) = (2\pi)^{-\frac{n}{2}}\det(K_{\boldsymbol{\psi}})^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\mathbf{Z}-\mathbf{t})^T K_{\boldsymbol{\psi}}^{-1}(\mathbf{Z}-\mathbf{t})} \tag{2}$$

where $K_{\boldsymbol{\psi}}$ is the covariance matrix of $Z(\mathbf{X}) = \{Z(\mathbf{x^1}), ..., Z(\mathbf{x^n})\}$ provided that $\boldsymbol{\psi}$ is the true vector of covariance parameters, and $\mathbf{t}$ is the vector of values of t at $\mathbf{X}$. The obtained result $\widehat{\boldsymbol{\psi}} = argmax_{\boldsymbol{\psi}}\{L(\boldsymbol{\psi};\mathbf{Z})\}$ is closely depending on $\mathbf{Z}$, i.e. on the observed realization of $Z(\mathbf{X})$. The behaviour of $\widehat{\boldsymbol{\psi}}$ relatively to $\boldsymbol{\psi}$ when the sample of observations fluctuates is a of importance. We recall that $\mathbf{Z}$ is assumed to be one realization of a multivariate Gaussian random vector with given trend, covariance structure, and covariance parameters $\boldsymbol{\psi}$. Then $L(.;\mathbf{Z})$ becomes a random function (fig. 1), and $\widehat{\boldsymbol{\psi}} = argmax_{\boldsymbol{\psi}}\{L(\boldsymbol{\psi};\mathbf{Z})\}$ becomes a random vector as well.

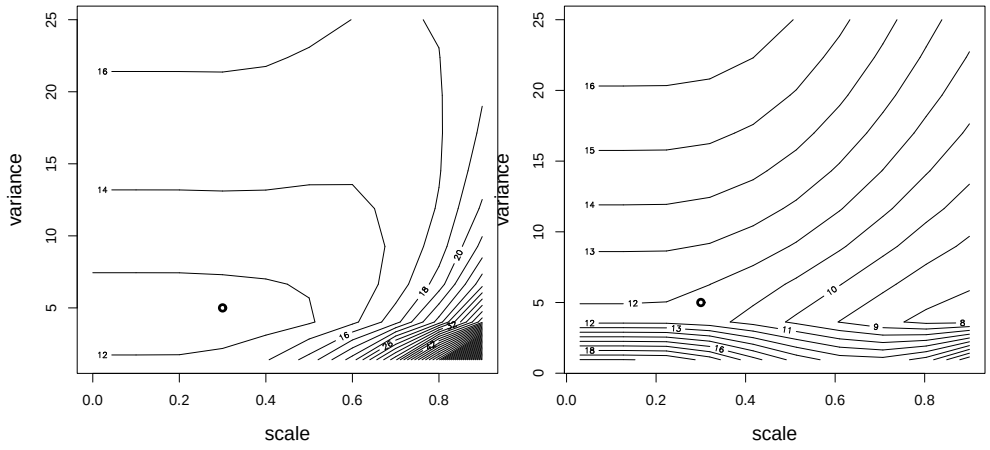


Figure 1. Two realizations of the random function $-2\ln L(.;\mathbf{Z})$ corresponding to two simulated response values $\mathbf{Z}$, both with a Gaussian covariance kernel (c_g defined hereafter) and covariance parameters $\boldsymbol{\psi} = (5,\ 0.3)$. **Left:** ML estimates are close to the actual parameters (bold dot) : $\widehat{\boldsymbol{\psi}} \approx \boldsymbol{\psi}$. **Right:** ML fails to locate the actual parameters: $\widehat{\boldsymbol{\psi}} \neq \boldsymbol{\psi}$.

The distribution of $\widehat{\boldsymbol{\psi}}$ has been studied in detail within the theory of likeli-

hood [23, 12]. A first order Taylor expansion leads to an asymptotic result [2] based on Fisher's Information Matrix $\mathcal{I}(\boldsymbol{\psi})$ (denoted by FIM in the sequel):

$$\begin{cases} \widehat{\boldsymbol{\psi}} \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{\psi}, \mathcal{I}(\boldsymbol{\psi})^{-1}) \\ \mathcal{I}(\boldsymbol{\psi}) = \left(\mathbb{E}\left[\frac{\partial \ln(L(.;\mathbf{Z}))}{\partial \psi_i}(\boldsymbol{\psi}) \frac{\partial \ln(L(.;\mathbf{Z}))}{\partial \psi_j}(\boldsymbol{\psi}) \right] \right)_{i,j \in [1,r]} \end{cases} \tag{3}$$

In many computer experiments, one first picks a covariance kernel from a parametric family: Gaussian, Exponential, Matèrn, etc...(the Gaussian covariance kernel is often chosen for its simplicity and regularity properties) and the associated covariance parameters are then automatically fitted by ML. However, the efficiency and robustness of this estimation method when few data are available are rarely discussed. Our concern is to check in what measure the first order asymptotic results hold with small samples. To do so, we computed the theoretical FIM of $\mathbf{Z}$:

$$\forall i,j \in [1,r] \ (\mathcal{I}(\boldsymbol{\psi}))_{ij} = \frac{1}{2} tr\left(K_{\boldsymbol{\psi}}^{-1} \frac{\partial K_{(.)}}{\partial \psi_i}(\boldsymbol{\psi}) K_{\boldsymbol{\psi}}^{-1} \frac{\partial K_{(.)}}{\partial \psi_j}(\boldsymbol{\psi}) \right) \tag{4}$$

To obtain comparable results for different values of ψ, we introduce a relative inverse FIM: $(\mathcal{J}(\boldsymbol{\psi}))_{ij} = (\mathcal{I}^{-1}(\boldsymbol{\psi}))_{ij}/(\psi_i \psi_j)$. $\mathcal{J}$ is in fact the asymptotical covariance matrix of $\frac{\widehat{\boldsymbol{\psi}}}{\boldsymbol{\psi}}$, where the division is made component by component. We conduct experiments with vectors taken from simulated monodimensional Gaussian Processes to compute empirical means and variances of the ML estimators. For each simulation, we compute covariance parameters estimated by ML and the Integrated Squared Error (ISE) between simulated ($z(\mathbf{x})$) and interpolated ($\widehat{Z}(\mathbf{x})$) data:

$$\text{ISE} = \frac{1}{\text{vol}(D)} \int_D |z(\mathbf{x}) - \widehat{Z}(\mathbf{x})|^2 d\mathbf{x} \tag{5}$$

where vol(D) is Lebesgue's measure of the set D. ISE is approximated by averaging the squared errors on a fine grid (i.e. 200 points). We finally collect the averages and variance matrices of the relative values of the estimated covariance parameters, the averages and variances of ISE (ISE is random since it depends on the realization z), and the covariances between ISE and $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$. The latter two indicators are not presented in the tables. We focus here on GPs with covariance kernels $c_g(h) = \sigma^2 e^{-\frac{h^2}{p^2}}$ (Gaussian) and $c_e(h) = \sigma^2 e^{-\frac{|h|}{p}}$ (Exponential). The covariance parameters reduce to $\boldsymbol{\psi} = (\sigma^2, p) \in]0, +\infty[\times]0, +\infty[$ and the design $\mathbf{X}$ is chosen among uniform subdivisions of $[-1,1]$: $\mathbf{X_n} = \{-1, -1+\frac{2}{n-1}, \ldots, -1+\frac{2(n-2)}{n-1}, 1\}$ ($n \in \mathbb{N}\backslash\{0,1\}$). We

restrict our experiments to the designs $\mathbf{X_5}$ and $\mathbf{X_{10}}$ with both c_g and c_e, and covariance parameters $\psi_1 = \sigma^2 \in \{5, 10\}$, and $\psi_2 = p \in \{0.3, 0.4, 0.5, 0.6\}$.

Table 1. ML and ISE values on 1000 simulated realizations of GPs with Gaussian covariance function, for relative parameters $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$, $i = 1, 2$ and for $\mathbf{X} = \mathbf{X_5}$. The second column shows that the relative ML estimates are almost unbiased even with 5 observations. On the contrary, a comparison between the third and fourth columns illustrates that the ψ_i^{rel} are clearly more dispersed than given by the asymptotical approximation based on the FIM.

$\boldsymbol{\psi}$	$\mathbb{E}\left[\left(\psi_i^{rel}\right)_i\right]$	$Var\left[\left(\psi_i^{rel}\right)_i\right]$	asymptotical $Var\left[\left(\psi_i^{rel}\right)_i\right]$	$\mathbb{E}$[ISE]
$\begin{pmatrix} 5 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.034 \\ 0.018 \end{pmatrix}$	$\begin{pmatrix} 1.105 & 0.277 \\ 0.277 & 1.270 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	4.976
$\begin{pmatrix} 5 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.147 \\ 0.042 \end{pmatrix}$	$\begin{pmatrix} 1.329 & 0.501 \\ 0.501 & 0.976 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	3.287
$\begin{pmatrix} 5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.222 \\ 0.033 \end{pmatrix}$	$\begin{pmatrix} 4.037 & 0.757 \\ 0.757 & 0.679 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	1.947
$\begin{pmatrix} 5 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.187 \\ 0.027 \end{pmatrix}$	$\begin{pmatrix} 2.058 & 0.504 \\ 0.504 & 0.421 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	0.706
$\begin{pmatrix} 10 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.131 \\ 0.006 \end{pmatrix}$	$\begin{pmatrix} 3.334 & 0.867 \\ 0.867 & 1.564 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	10.138
$\begin{pmatrix} 10 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.083 \\ 0.106 \end{pmatrix}$	$\begin{pmatrix} 1.645 & 0.484 \\ 0.484 & 0.862 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	6.398
$\begin{pmatrix} 10 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.166 \\ 0.024 \end{pmatrix}$	$\begin{pmatrix} 1.343 & 0.440 \\ 0.440 & 0.629 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	3.678
$\begin{pmatrix} 10 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.256 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} \mathbf{14.960} & 0.963 \\ 0.963 & 0.392 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	1.459

In the case of a Gaussian covariance, we observe a negative relative bias [1] (-3.4% to -25.6%) in the estimation of $\psi_1 = \sigma^2$. This bias is decreasing with the number of design points $\#\mathbf{X}$ (see table 2 where the negative relative bias varies between -4.2% and -7.5%), which seems in accordance with the asymptotic unbiasedness of MLE. On the other hand, the relative bias of $\hat{\psi}_2$ has a small order of magnitude when $\#\mathbf{X} = 5$ and slightly oscillates around 0 when $\#\mathbf{X} = 10$.

The empirical covariance matrices of the ML estimates offer some surprising results. In particular, the relative variances of $\hat{\psi}_1$ present huge fluctuations: they vary sometimes of an order of more than 10 between two samples of 1000 realizations issued from the same GP; for instance by resimulating a GP with $\psi = (10, 0.4)$ and $\#\mathbf{X} = 5$ we obtain $Var\left[\left(\psi_i^{rel}\right)_i\right] = \begin{pmatrix} 43.555 & 3.242 \\ 3.242 & 0.971 \end{pmatrix}$. Since it is in contradiction with normality and the order of magnitude given by (eq.4),

[1] Mind the fact that by negative relative bias we understood an overestimation of $\boldsymbol{\psi}$.

Table 2. MLE and ISE measures on 1000 simulated GP realizations with Gaussian covariance kernel, for $\mathbf{X} = \mathbf{X_{10}}$. The aproximation based on Fisher's Information Matrix is still underestimating the estimation variances but is less unprecise than in table (1).

$\boldsymbol{\psi}$	$\mathbb{E}\left[\left(\psi_i^{rel}\right)_i\right]$	$Var\left[\left(\psi_i^{rel}\right)_i\right]$	asymptotical $Var\left[\left(\psi_i^{rel}\right)_i\right]$	$\mathbb{E}$[ISE]
$\begin{pmatrix} 5 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.054 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.432 & 0.105 \\ 0.105 & 0.085 \end{pmatrix}$	$\begin{pmatrix} 0.297 & 0.057 \\ 0.057 & 0.033 \end{pmatrix}$	0.177
$\begin{pmatrix} 5 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.042 \\ -0.019 \end{pmatrix}$	$\begin{pmatrix} 0.424 & 0.058 \\ 0.058 & 0.024 \end{pmatrix}$	$\begin{pmatrix} 0.340 & 0.044 \\ 0.044 & 0.014 \end{pmatrix}$	0.009
$\begin{pmatrix} 5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.067 \\ -0.013 \end{pmatrix}$	$\begin{pmatrix} 0.46 & 0.051 \\ 0.051 & 0.013 \end{pmatrix}$	$\begin{pmatrix} 0.362 & 0.036 \\ 0.036 & 0.008 \end{pmatrix}$	0.0004
$\begin{pmatrix} 5 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.075 \\ -0.007 \end{pmatrix}$	$\begin{pmatrix} 0.728 & 0.059 \\ 0.059 & 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.375 & 0.032 \\ 0.032 & 0.005 \end{pmatrix}$	4.e-05

we shall analyze this phenomenon in detail. First, we observe that the extreme values of $Var[\hat{\psi}_1]$ are caused by some outliers, highly perturbating the non-robust estimate of variance. Second, the histogram in (fig.2.) illustrates that the distribution of the $\hat{\psi}_1$'s is rather lognormal than normal. Finally, the comparison with the relative FIM shows that the empirical variance of $\hat{\psi}_1$ is clearly larger than predicted by the second order Fisher approximation, in particular with the smallest designs.

Concerning the relative variances of $\hat{\psi}_2$, the results are much more regular: they decrease monotonically with ψ_2 and with $\#\mathbf{X}$, both for the empirical and theoretical quantities. Once again, the empirical variances tend to match the theoretical variances as $\#\mathbf{X}$ grows, even if the first ones are still typically two times larger than the second ones for a sample of size 10. In other respects, both tables illustrate some fundamental properties of the mean squared error. Obviously decreasing with $\#\mathbf{X}$, the ISE is also decreasing with the range ψ_2 and linearly increasing with the variance ψ_1. Finally, we quantify the linear dependence between the underestimation of both covariance parameters by MLE and the ISE (not in the tables). It is worth noticing that ψ_1 and ψ_2 play drastically different roles here: it seems that a bad estimation of ψ_1 is weakly correlated with the ISE. This result seems natural when considering that the OK predictor is not depending on the process variance, see [3]. Conversely, the correlation between the ISE and the relative MLE error on ψ_2 is significantly positive: it varies between 40.1% and 55.7% when $\#\mathbf{X} = 5$ and between 15% and 62.5% when $\#\mathbf{X} = 10$. This coincides with our previous qualitatives observations of larger ISE when the range is much underestimated.

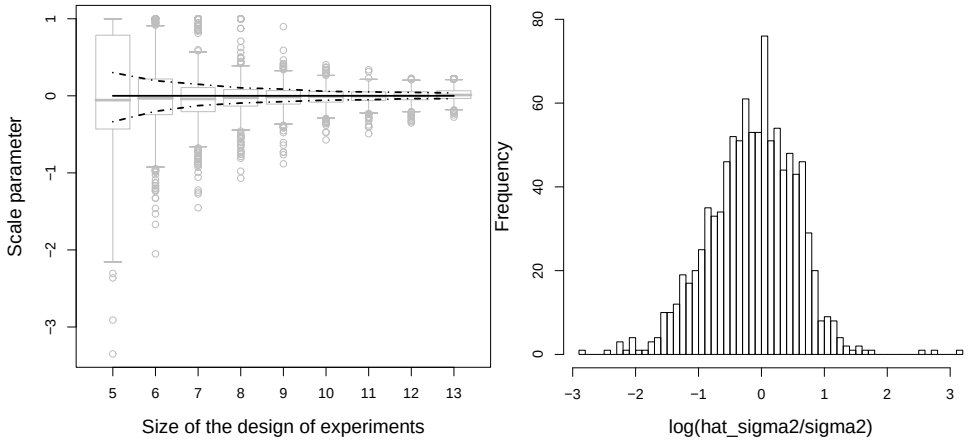


Figure 2. **Left:** Comparison between the experimental law (gray boxplots) and the asymptotic law (black lines) for the scale parameter for increasing size of the design. The boxplots for the experimental laws have been done using 1000 simulations, with Gaussian covariance function of parameters $\boldsymbol{\psi} = (5, 0.5)$. For the asymptotic law the median is represented in continuous line and the first and third quartiles in dashed lines. **Right:** Histogram of the logarithm of relative errors obtained when estimating ψ_1 by ML using 1000 GP realizations. The shape of this histogram suggests that the distribution of relative errors is far closer to a lognormal law than to a Gaussian.

A similar study with exponential covariance function gives very different results both for the bias and the variances of ML estimates (the corresponding tables are not presented here). Indeed, we observe very regular variances of ML estimates while the bias reaches impressive orders of magnitude. However, the behaviour of the ISE and the correlations between ISE and relative MLE errors follow the same sketch as in the Gaussian case.

To sum up this section about ML estimation:

- Fisher's first order asymptotical results must be applied with much care concerning the sample size. More precisely, it has been observed here for $n \leq 5$ that the distribution of the estimated range parameter is asymmetrical with a higher variance than the inverse of Fisher's information, but quickly stabilizes to a Gaussian when n increases (from 5 to 13).

- On the other hand, the distribution of the estimated variance parameter has a very large right tail but its shape is far from being gaussian when n is very small ($n \leq 5$). Furthermore, these results still hold when n increases (from 5 to 13) and it seems that the Gaussian approximation becomes reasonable only for larger values of n.

Estimating covariance parameters by ML with few data appears to produce very dispersed results. Hence, it seems unreasonable to neglect the uncertainty associated with this phase of estimation when performing Kriging. Bayesian techniques are a way to address this issue [16, 4]. In other respects, [1] investigates an extended Kriging variance taking the estimation of parameters into account; a this stage the latter relies on the first order approximation. To finish with, frequentist approaches based on the maximization of penalized likelihood functions seem very promizing since they provide estimators with the same asymptotic properties as ML in addition to a more robust behaviour with few observations [11].

3. KRIGING WITH TRENDS: A BLESS OR A CURSE?

Now we wish to examine another difficulty encountered when Kriging based on few data: the selection and the estimation of deterministic trends. In computer experiments, the most commonly used Kriging model seems to be Ordinary Kriging. However, OK reaches one of its limits when the stationarity assumption does not hold any longer, i.e. when non constant trends $t(\mathbf{x})$ are impossible to ignore. In this case, we are back to the general decomposition of (eq.1), where z is assumed to be the sum of a deterministic trend t and one realization of a centered GP ε. At this stage, we may consider several subcases.
If t is known and the parameters of ε have to be estimated, a straightforward solution is to perform Simple Kriging of the residuals $\{z(\mathbf{x}) - t(\mathbf{x})\}_{\mathbf{x}\in D}$.
If t is unknown, it is common to distinguish between a linear and a more general non-linear framework. The case in which t depends linearly on its parameters and ε has a kwown covariance structure has been intensively studied: it is well known as Universal Kriging [14]. When the covariance parameters $\boldsymbol{\psi}$ are known and the trend is a linear combination of some chosen basis functions f_j ($j \in [1, b],\ b \in \mathbb{N}\backslash\{0\}$), the only unknowns are the parameters of the trend ($\forall j \in [1, b],\ \beta_j \in \mathbb{R}$); indeed, if $t(\mathbf{x}) = \sum_{j=1}^{b} \beta_j f_j(\mathbf{x})$, the β_j's can directly be estimated by Generalized Least Squares (GLS):

$$\hat{\beta}(\psi_2) = (\mathbf{F}^T K_{\boldsymbol{\psi}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T K_{\boldsymbol{\psi}}^{-1} \mathbf{Z} = (\mathbf{F}^T R_{\psi_2}^{-1} \mathbf{F})^{-1} \mathbf{F}^T R_{\psi_2}^{-1} \mathbf{Z} \tag{6}$$

where $\mathbf{F}$ denotes the evaluation of $\mathbf{f}(x) = [f_1(x), \ldots, f_b(x)]$ at the n design points and $R_{\psi_2} = (1/\psi_1)K_{\boldsymbol{\psi}}$ (proportionality since the observations are noise-free) is the correlation matrix of $Z(\mathbf{X})$.
In practice, however, one has seldom the value of the covariance parameters at disposal previous to performing UK. So one has to estimate a model with linear trend and unknown covariance parameters $\boldsymbol{\psi}$ (in the following we will also refer to this case as "UK", like many practitioners do). Hence $\boldsymbol{\psi}$ and β have to be estimated within Kriging. At a first sight, this is likely to create a circularity problem: one needs a known trend to work on the residuals and thus estimate $\boldsymbol{\psi}$. On the other hand, estimating t without taking the residuals into account may lead to unadapted trends (the estimation of the trend parameters would rely on Ordinary Least Squares instead of GLS).

Fortunately, ML estimation gives a way to escape this vicious circle. Assuming, like in section 2 that the covariance parameters to be estimated are $\boldsymbol{\psi} = (\sigma^2, p)$, and using MLE (and the same formula (6) for $\hat{\beta}$), one can get a straightforward formula for $\widehat{\sigma^2}$, explicitly depending on ψ_2:

$$\widehat{\sigma^2}(\psi_2) = (1/n)(\mathbf{Z} - \mathbf{F}\hat{\beta}(\psi_2))^T R_{\psi_2}^{-1}(\mathbf{Z} - \mathbf{F}\hat{\beta}(\psi_2)) \tag{7}$$

By injecting (6) and (7) in the expression of the likelihood, one can obtain a *concentrated likelihood* function $L(\psi_2, \widehat{\sigma^2}(\psi_2), \hat{\beta}(\psi_2))$ which clearly depends only on ψ_2 and which has to be maximized to get $\hat{\psi}_2$. The Kriging predictor with plugged-in covariance parameters is then given by:

$$\hat{Z}_{\hat{\psi}_2}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\beta}(\hat{\psi}_2) + r^T(\mathbf{x})R_{\hat{\psi}_2}^{-1}\left(\mathbf{Z} - \mathbf{F}\hat{\beta}(\hat{\psi}_2)\right) \tag{8}$$

where $r(\mathbf{x})$ is a vector of correlation values between Z at an unknown point $\mathbf{x}$ and at the points of the design $\mathbf{X}$. Most of the time (apart in Bayesian Kriging) the variability due to the estimation ψ_2 is not propagated, and one uses the regular UK prediction variance.

UK appears as a very convenient means to incorporate known deterministic trends within Kriging. By the way, we will see in the next section that overcoming the circularity problem is not easy in a more general non-linear framework. Now we would like to go one step deeper in practical considerations and raise a naive but complex question which has to be handled in real-world applications, and particularly in high-dimensional problems: how can one come back to the nature of the trend from raw data? As soon as neither prior information nor obvious graphical clue is available, one has indeed to select a trend on the basis of $(\mathbf{X}, \mathbf{Z})$. What means does he have to do so,

and what risk does he run in case of a bad choice? In order to show that these questions are crucial, let us first perform some toy experiments. The set-up is the following. A realization of a one-dimensional GP with known covariance function and parameters is simulated on a regular grid (401 points on $[-1, 1]$) and an affine trend is added; From this set we choose different subsets of points and perform three types of Kriging : OK, UK with linear trend and UK with quadratic trend.

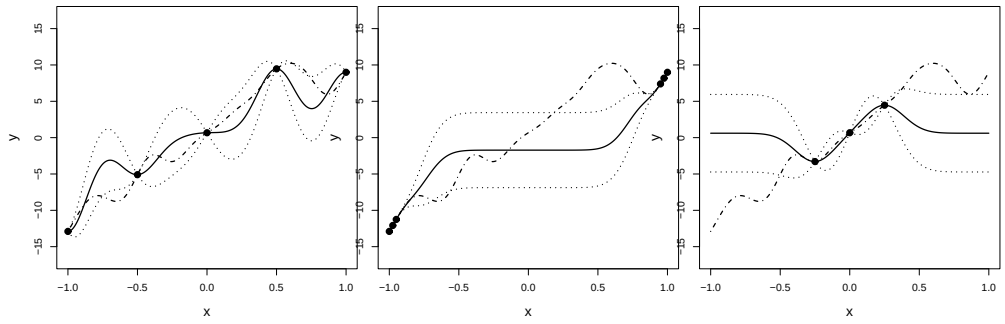


Figure 3. One realization (dashdot line) of a GP with linear trend is interpolated by Ordinary Kriging, based on 3 designs. The OK mean and 95% confidence intervals are represented by bold lines and dotted lines, respectively. The first design (left) is a regular grid; the associated OK prediction seems satisfying, even if the trend model is misspecified. The second design (center) is formed by 6 points concentrated at the boundaries of the domain; The Kriging predictor fails to capture the shape of the realization at the center of the domain. The third design is made of three points clustered at the center of the domain; OK automatically comes back to the mean value outside of the design and dramatically miss the actual trend.

We choose at first a subset of 5 regularly distributed points. Due to the fact that the points are regularly spaced on the grid, all the three kriging give similar good results, even if in two of the three cases the trend is misspecified (fig.3 left for the case of Ordinary Kriging). This may lead to the conclusion that specifying the trend is not very important and we could obtain good results using OK. But if we perform the same Krigings on different designs, where there are few points concentrated either on the boundaries or in the center of the domain, then the results are very bad (due to the ratio between the parameter ψ_2 and the subdivision length) when the trend is misspecified, see (fig. 3, middle and right). The covariance parameters used for the simulated process in (fig. 3) and (fig. 4) are $\psi = (5, 0.2)$. The results are even

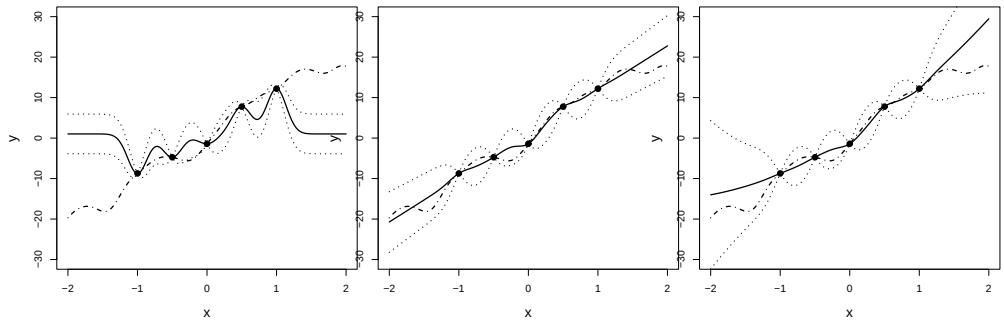


Figure 4. One realization (dashdot line) of a GP with linear trend is interpolated by three different Krigings, based on a regular grid (5 points evenly spaced between -1 and 1). The GP is here represented between -2 and 2, such that all three cases can be referred to as extrapolations. UK with affine trend (center) gives accurate results on the whole domain. On the contrary, both OK (left) and UK with quadratic trend (right) give good results between -1 and 1 but dramatically fail in extrapolation.

worse if we use the Kriging predictor given by OK or by UK with quadratic trend in extrapolation (fig.4, left and right). In the one dimensional case the choice of the trend doesn't seem to be essential while interpolating data which are not very distant one from another with respect to the frequency of variation of the process. On the contrary, when the design is not regular and we are in extrapolation, the performances of Kriging are very sensitive to the adequacy between the real trend of the process and the Kriging trend.

Hence it seems enough to properly fill the space to avoid the risks caused by the choice of trend functions. But what is possible in one or two dimensions becomes unrealistic when the dimension increases: a design with only one point at each vertex of a cubic domain $[0,1]^d$ has 2^d points, i.e. 1024 points in 10 dimensions and more than a billion points in 30 dimensions. As we usually dispose of $10 \times d$ observations per dimension, which is already an optimistic case, choosing a trend based on data only appears as a very difficult task. Let us see nevertheless what would be possible in order to choose a trend starting from a data set $(\mathbf{X},\ \mathbf{Z})$: the classical frame of linear regression offers a panel of diagnostic tools dedicated to validating both assumptions on the trends and on the model of residuals. For instance, commonly used indicators include R^2 (and R^2 adjusted), the F-ratio and the p-values for each estimated regression coefficients, and numerous criteria to check the adequacy of the residuals to the underlying model. In most cases the Gaussian

likelihood of the residuals is considered among the relevant criteria of model selection (some model testing techniques or even based upon it).

Now it seems necessary to recall that the latter measures are exclusively done at the design of experiments, also called "training sample" or "learning sets" in the literature of statistical learning, see [6]. Selecting only on the basis of a R^2 fit would lead for instance to the systematic choice of models interpolating ($\mathbf{X}$, $\mathbf{Z}$). However such models are not meant to be good in prediction outside the design of experiments. This warning leads to the double message:

- Model complexity must be taken into account in selection procedures
- Testing the model at some test points not used in the model fitting could be worth: this is for instance what cross-validation does.

The following experiment is performed in an intent to illustrate the first point. The second point will be illustrated in the next section. Here we investigate on a simple case how trend selection may be misleading when likelihood is the only criterion, without any consideration of model complexity. To do so, we compute, for each trend form of the Kriging model, the optimal parameter $\hat{p}$ by ML, we compare the corresponding values of the likelihoods and we select the kriging model having the highest value of likelihood. In table 3, we compare three Kriging models (OK, UK with linear trend, UK with quadratic trend) for three different functions: one realization of a one-dimensional GP with 11 points and with Gaussian covariance function ($\psi = (5, 0.4)$), the same realization plus a linear trend $0.5 + 5x$, and the same realization plus a quadratic trend $0.5 + 5x + 5x^2$.

Table 3. Comparison of minimum $-2\log(L)$ values obtained by fitting three different Kriging models (OK, UK affine, UK quadratic) to three GP realizations. Each realization is drawn from the GP underlying one of the Kriging models. The design is a regular grid on $[-1, 1]$. The results illustrate that adding degrees of freedom to a Kriging model always lead to a larger value of the maximum likelihood.

	GP		GP +linear $\mathbf{t}$		GP+quadratic $\mathbf{t}$	
kriging type	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$
OK	0.4082	**32.07**	0.4445	36.90	0.4595	38.80
UK, linear $\mathbf{t}$	0.4085	**31.89**	0.4085	31.89	0.4387	35.80
UK, quadratic $\mathbf{t}$	0.4084	31.89	0.4084	31.89	0.4084	31.89

Here it is essential to notice that the likelihood values are necessarily larger when adding more degrees of freedom to a statistical model. This constitutes a misleading incentive to always choose the model with the largest number of parameters within a given family. This happens for instance between Kriging models with first order and second order polynomial trends. As can be observed in table 3, L always increases (i.e. the values of $-2\ln(L(\hat{p}))$ will decrease) with the complexity of nested model. What we should really compare are maximum likelihood values between models with the same number of degrees of freedom. On the last line of table 3, in the cases of the GP without trend and of the GP with linear trend , the estimated values $\hat{\beta}$ are very close to but different from zero. Thus the model obtained by automatically selecting the Kriging with highest likelihood will perform badly in extrapolation because of the higher order terms of the polynomial. The same phenomenon applies in an even more pronounced way with a linear trend in the case of a centered GP (first column, second row).

As a conclusion to this section, we have pointed out that OK and UK may seem to deliver similar results when the design is dense [22], but modeling the trend matters in extrapolation situations [9]. Since working in high-dimensional spaces means that we will practically always be in extrapolation, we need exploratory and visualization tools dedicated at finding trends in multivariate data. Recent methods of data mining and functional analysis may help [6]. We propose now to use additive models within spatial interpolation.

4. USING NON-LINEAR ADDITIVE MODELS AS EXTERNAL DRIFT

Linear models are often used by practitioners of quantitative disciplines since they are simple to interprete and to assess. Additive models (AM) are an extension of linear models. A precise description of these models can be found for instance in the book [5]. The advantage of AM is to conserve the feature of non-interacting predictors, but they allow much more flexible inference for each univariate problem, using kernel smoothers for instance [24]. The generic expression for an additive model is the following:

$$\left\{ \begin{array}{c} Z_i = z(\mathbf{x}) + \varepsilon_i \\ z(\mathbf{x}) = \alpha + \sum_{j=1}^{d} f_j(\mathbf{x}_j) \\ \text{The } \varepsilon_i \text{ are } n.i.i.d. \end{array} \right. \tag{9}$$

and the f_js are arbitrary univariate functions, one for each predictor but

possibly not the same kind of function for each dimension. Hence additive models deal with additive functions observed in a Gaussian noise. $\mathbf{x}$ is in fact assumed here to be a control variable, and not a random variable as in [5]. This model may be used to approximate deterministic computer experiments, provided that the response surface can reasonably be decomposed in an additive way. Once the nature of the f_js is chosen, they can be estimated using a powerful iterative procedure called *backfitting algorithm*, see [6]. Backfitting means that f_1 is estimated on the basis of all data $(\mathbf{X}, \mathbf{Z})$, then f_2 is fitted to the residuals $\mathbf{Z} - f_1(\mathbf{X})$, and so on. Under mild assumptions, the backfitting algorithm converges and finds the unique solution of the additive decomposition of (eq.10). In this section, we propose a combination of additive model and Kriging that offers the great flexibility of *AMs* and yet interpolates the data. It seems very natural to combine both models by using the following decomposition:

$$\begin{cases} z(\mathbf{x}) = t(\mathbf{x}) + \varepsilon_{SK}(\mathbf{x}) \\ t(\mathbf{x}) = \alpha + \sum_{j=1}^{d} f_j(\mathbf{x}_j) \\ \varepsilon_{SK}(\mathbf{x}) \text{ is a GP realization like in (eq.1)} \end{cases} \tag{10}$$

This identity may at first seem similar to the equation of Universal Kriging. However, in this case the non-linear nature of the trend prevents one from solving the estimation globally. Indeed, a likelihood maximization would lead to an optimization problem in infinite dimension:

$$\max_{\boldsymbol{\psi}, (f_j)_{j\in[1,d]}} L(\boldsymbol{\psi}, \mathbf{t}; \mathbf{Z}) \tag{11}$$

To our knowledge such a problem is analytically intractable. On the other hand, the backfitting algorithm is not suited anymore if we take the Kriging part into account. Indeed, kriging the residuals after fitting a smoother in one dimension would lead to an interpolation and thus end the iterative procedure without fitting the additive parts in the other dimensions.

Kriging with external trend [3] seems to constitute a good alternative for solving both the problem of the "general" form of trend and the one of the circularity. Consequently, we consider now a two-step approach (see Alg.1): first, the additive trend $t(\mathbf{x})$ is estimated using the backfitting algorithm, and then Simple Kriging is applied to $(\mathbf{Z} - \mathbf{t})$ with covariance parameters estimated on the basis of those residuals, by likelihood maximization or other.

Unfortunately, there are significant drawbacks in the latter procedure, mainly related to the uncontrolled trade-off between deterministic and stochastic

Algorithm 1 A first two-step approach to fit a Kriging with additive trend

1: Estimate the trend t by backfitting
2: Estimate the covariance parameters and fit a SK model on the basis of the residuals at $\mathbf{X}$

parts. Hence, the whole uncertainty reduces here to the Kriging variance estimated on the residuals; there is indeed no global uncertainty on the trend unless we use only splines in the *AM*. This is likely to cause a large underestimation of the process variance associated with the model. Furthermore, these residuals may be not very well suited to estimate the Gaussian process part: the additive model is constructed to fit z accurately at the design -possibly leading to *overfitting*-, thus the residuals at $\mathbf{X}$ are likely to vary with a smaller magnitude than in prediction. Since we look for a model with reasonable generalization properties, it seems necessary to find an alternative way of estimating the covariance parameters.

We propose here a sequential estimation technique for combined Kriging models like (eq.10). It is based on the idea that when the trend is non-linear, the parameters of the GP model should be estimated on a validation set rather than on the set at which the trend is fitted.

Algorithm 2 An alternative two-step approach to fit a Kriging with additive trend

1: Consider two designs $\mathbf{X}_1$ and $\mathbf{X}_2$ ▷ possibly obtained by splitting $\mathbf{X}$
2: Estimate the trend t by backfitting, based on the data $(\mathbf{X}_1, z(\mathbf{X}_1))$
3: Estimate the SK covariance parameters on the basis of the residuals $\{t(\mathbf{x}) - z(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}_2}$
4: Fit the SK model on the basis of all residuals ▷ with parameters estimated at the previous step

5. A 3-DIMENSIONAL APPLICATION OF KRIGING WITH ADDITIVE TREND (KAT)

The previous approach is applied to a 3-dimensional example from an industrial test case. The data are obtained with a flow simulator and the numerical response z, standing for the outcome of interest, is studied as a function of three physical parameters characterizing the porous media and denoted by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3 \in [-1, 1]$. The response is simulated at 1331 locations corresponding to a 11-level full factorial design, denoted by "F" in

the sequel. Our goal is to provide a surrogate of the simulator on the basis of a poor design of experiments. The metamodel should interpolate the data (to respect the determinism of the underlying simulation) and provide a prediction uncertainty that allows statistical-based exploration, for instance to solve optimization problems. Furthermore, it should take into account a prior knowledge inherited from a previous study: the phenomenon is almost additive in its parameters.

Our initial design, "$\mathbf{X}_1$", is a 20-elements Hammersley sequence. We first perform a graphical analysis (fig. 5) of the response at $\mathbf{X}_1$, discuss the hypothese of additivity, and propose several kinds of linear and additive trends to model the data. Algorithm 1 is tested with the design $\mathbf{X}_1$ (Table 4). Then a second design, "$\mathbf{X}_2$", is used for an intermediate validation of the covariance parameters of the model previously obtained. $\mathbf{X}_2$ is made of 14 points taken from a 40-elements D-optimal design (see fig. 6). Algorithm 2 is then performed by re-estimating the covariance parameters of the previous SK model on the basis of the residuals at $\mathbf{X}_2$. An original estimation method is proposed, which differs from the traditional MLE: the process variance σ^2 is fixed such that the standardized residuals have most of their values between -2 and 2 [7] and the range parameter p is chosen in order to minimize the ISE at the design $\mathbf{X}_2$ (fig. 7). The full factorial design F is finally used for a phase of model validation (fig. 8).

A graphical analysis of the coplots at $\mathbf{X}_1$ does not reject the prior belief of additivity. A first additive decomposition is then estimated using splines in all directions (referred to as "GAM splines" in the following). We observe that we might take a linear trend in the directions of $\mathbf{x}_1$ and $\mathbf{x}_3$, and a non-linear trend in $\mathbf{x}_2$ without loosing much accuracy (see Table 4 for a quantitative validation). Hence we choose to fit an additive model with mixed trends, called "GAM mixed" in the sequel.

Different Krigings with external trend are fitted to the observed data at the design $\mathbf{X}_1$. In all cases, the SK part has a structure of isotropic GP with Gaussian covariance (see section 2). We focus on the two additive trends defined above and on two additional linear trends: a first and a second order regression polynomial. For each model, we fit the trends respectively by OLS and backfitting, and we measure their relevance using indicators computed with the residuals at the design $\mathbf{X}_1$ (residuals deviance and p-values when available). Then we fit a Kriging to the residuals, as explained in Algorithm 1. For each Kriging, we store the maximum reached value of the log-likelihood and the corresponding range and variance values. The results

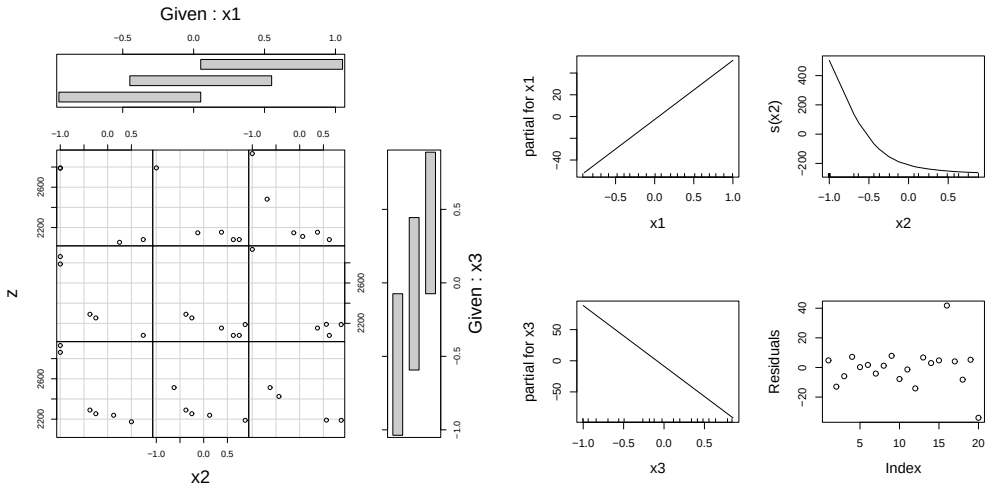


Figure 5. Coplots of z on the Hammersley design $\mathbf{X}_1$ (left) and summary of the additive components and the residuals obtained after application of the backfitting algorithm (right). The additive model is here chosen with a linear function in both directions of $\mathbf{x}_1$ and $\mathbf{x}_3$, and a smoothing spline in $\mathbf{x}_2$'s direction.

are listed in (Table 4).

Table 4. Optimal loglikelihood values and estimated covariance parameters associated with the residuals provided by Algorithm 1 at $\mathbf{X}_1$ with different trend structures. The R^2 values are computed by comparing the residual deviance after fitting the trend only, to the total variance of the response at the design $\mathbf{X}_1$.

Model	Loglikelihood	Range	σ^2	R^2_{adj}	R^2	p-value
1st order Linear + SK	-121.91	1.04	26101.89	0.78	0.82	4.03e-06
2st order Linear + SK	-100.69	0.048	1381.13	0.97	0.98	6.44e-08
GAM splines + SK	-76.01	0.048	117.10	-	0.99	-
GAM mixed +SK	-80.62	0.16	185.71	-	0.99	-

These results support the belief that a general additive trend is adapted for these data: both the variance of residuals and the values of their likelihood (compared to the 2nd order linear model, which uses more degrees of freedom) indicate their good fit to the data.

In practice, however, we care more about the model's abilities to make correct predictions at new points than about its mean squared error at the design.

Hence, model validation should not be blindly supported by the indicator R^2 or the likelihood of the residuals at $\mathbf{X}_1$. First, we should consider the number of degrees of freedom of the model. Second, it may be worth validating the model outside of the design. Indeed, the residuals drawn from (fig. 5) are computed at the same locations as those used to fit the trend.

Concerning the first point, we can compare the degrees of freedom of both "2nd order linear" and "GAM mixed": respectively 10 and 7. Regarding the second point, we conduct a validation test on some additional data, inspired by the cross-validation procedure. Following Algorithm 2, $\mathbf{X}_2$ is used to valid and update the parameters associated with the model fitted at $\mathbf{X}_1$.

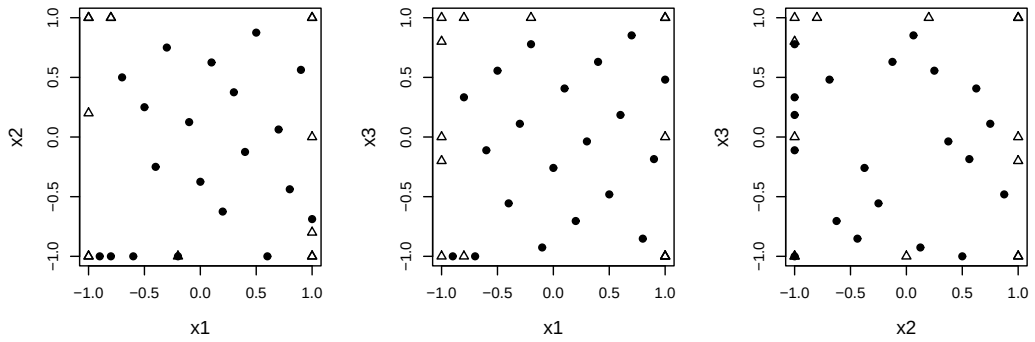


Figure 6. Coplots representing both $\mathbf{X}_1$ (dots) and $\mathbf{X}_2$ (triangles) designs in projection on all pairs of coordinates. The three graphics illustrate the space-filling behaviour of $\mathbf{X}_1$ and the D-optimal nature of $\mathbf{X}_2$. $\mathbf{X}_2$ also appears to be reasonably disconnected from $\mathbf{X}_1$.

The points of $\mathbf{X}_2$ are used to test the validity of the covariance parameters of the model "GAM mixed", previously estimated by ML. Figure 7 shows the associated residuals standardized by the ML variance (left), and the behaviour of the ISE at $\mathbf{X}_2$ as a function of p (right). We recall that the residuals should satisfy the assumption of normality in order to get relevant Kriging variances, insuring correct statistical predictions.
Figure 7 (right) shows that the ISE at the validation design $\mathbf{X}_2$ can be significantly reduced by increasing the range p. Following Algorithm 2, we re-estimate the covariance parameters based on these residuals at $\mathbf{X}_2$. Instead of using ML however , we prefer to directly use the work done hereabove to compute the ISE as a function of the range. It appears indeed that the optimal range to accurately fit the residuals at the validation design is given by

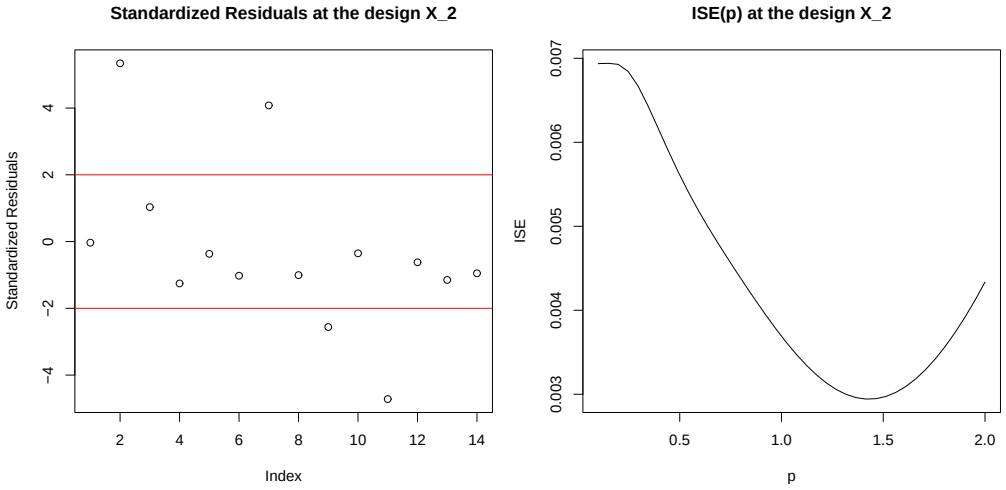


Figure 7. Standardized residuals at the validation design $\mathbf{X}_2$ (left) and variation of the ISE with respect to the covariance parameter p (right). One can observe that the value of p found by ML at $\mathbf{X}_1$ ($p_1 = 0.16$) is clearly suboptimal to get an accurate Kriging predictor when extrapolating to $\mathbf{X}_2$.

$p_2 = 1.4$. Concerning the variance, we observe more satisfying standardized residuals with $\sigma_2^2 = (2 \times \sigma_{ML})^2$. So we keep σ_2^2.

Remark: A ML estimation with the residuals at $\mathbf{X}_2$ delivers $p = 0.97$.

We finally test the model of Algorithm 2 at the design F (fig. 8). The standardized residuals (with the variance σ_2^2) and the ISE as a function of p validate our empirical decisions made on the basis of the intermediate design $\mathbf{X}_2$ (note that ML on $\mathbf{X}_2$ -see remark hereabove- gives also better results than ML on $\mathbf{X}_1$ but the cross-validating strategy minimizing the ISE at $\mathbf{X}_2$ remained the best). To conclude with, the algorithm investigated performed well on this example: Simple Kriging seems to constitute a good complement to additive models in an intent to interpolate data and also possibly explain some non-additive part. The method we use here allows inference of covariance parameters with values suited for a correct quantification of uncertainty. This seems encouraging to develop further "cross-validation-like" methods for the combination *Additive model + Kriging*.

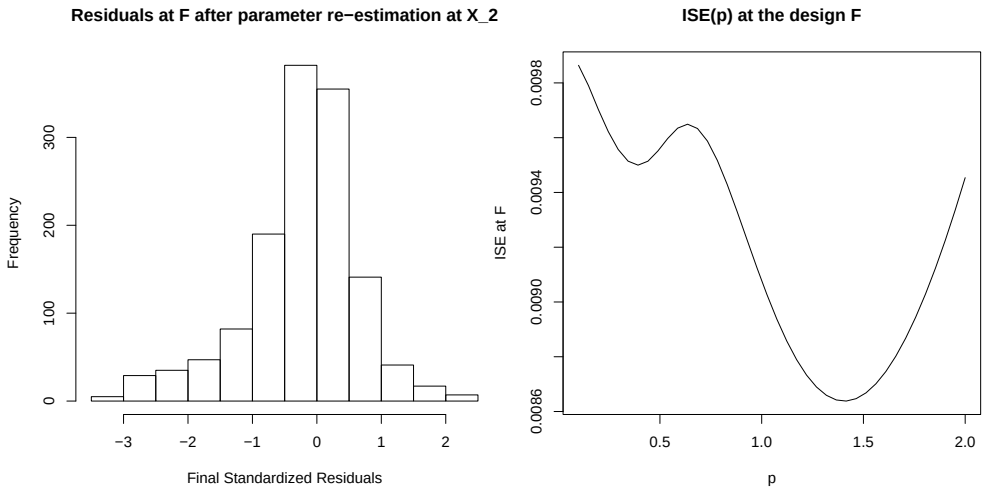


Figure 8. **Left:** Histogram of the standardized residuals at the test design F with the model previously obtained by Algorithm 2 (GAM mixed, $\sigma^2 = \sigma_2^2,\ p = p_2$). **Right:** Variation of the ISE with respect to the covariance parameter p: the value $p_2 = 1.4$ previously chosen at $\mathbf{X}_2$ is almost optimal again.

6. CONCLUSIONS AND PERSPECTIVES

We observed in a one-dimensional frame that MLE could behave very differently from Fisher asymptotical results when n is small. This result should be kept in mind when dealing with higher dimensions, and further studies have to be done in this latter context. Since it relies on the simulation of Gaussian vectors, the experimental approach presented here can easily be transposed in a higher dimensional framework. Perspectives include the empirical comparison of ML and penalized ML [11] when using classical designs.

Further experiments on the topic of trend selection illustrated the fact that the likelihood cannot be considered as only criterion when comparing different functional families. This is suggesting methods penalizing complexity (like in AIC and BIC). But we mainly wish to emphazise on the risks took when predicting with trended Kriging: in higher dimensions, we will allways be in an extrapolation situation. Choosing a trend with the help of a small design then seems very risky. This is an argument to consider Ordinary Kriging in the cases where no prior information on the trend is available.

In other respects, we proposed a model combining an additive model and

Simple Kriging. The application to a simple industrial test case confirmed that directly kriging the residuals by ML gives a poor result. Our attempt to adapt a method inspired by cross-validation with a single test set gave here a Kriging with different features from ML, apparently accounting well for the non-additive part of the response. However, the question of the robustness to a change of design has not been raised yet. This is a subject to be treated in further works.

AKNOWLEDGEMENTS

All the computations have been performed using *R* [17] and the packages *geoR* [18], *RandomFields* [21], *gam* and *gstat*. This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA, and Total S.A.

REFERENCES

[1] Markus Abt. Estimating the prediction mean squared error in gaussian stochastic processes with exponential covariance structure. *Scandinavian Journal of Statistics*, 26:563–578, 1999.

[2] Markus Abt and William J. Welch. Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes. *The Canadian Journal of Statistics*, 26:127–137, 1998.

[3] N.A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics, 1993.

[4] Sarah Goria. *Evaluation d'un projet minier: approche bayésienne et options réelles*. PhD thesis, Ecole des Mines de Paris, 2004.

[5] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1991.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[7] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[8] Astrid Jourdan. Approches statistiques des expériences simulées. *Revue de Statistiques Appliquées*, 50:49–64, 2002.

[9] A. G. Journel and M. E. Rossi. When do we need a trend model in kriging? *Mathematical Geology*, 21(7):715–739, 1989.

[10] J.R. Koehler and A.B. Owen. Computer experiments. Technical report, Department of Statistics, Stanford University, 1996.

[11] R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, (47):111–120, 2005.

[12] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–46, 1984.

[13] J.D. Martin and T. W. Simpson. A monte carlo simulation of the kriging model. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY*, page 4483, 2004.

[14] J.D. Martin and T.W. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43 (4):853–863, 2005.

[15] Georges Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.

[16] A. O'Hagan. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, (91):1290–1300, 2006.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[18] P.J. Ribeiro Jr. and P.J. Diggle. geoR: A package for geostatistical analysis, 2001. ISSN 1609-3631.

[19] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, (4):409–435, 1989.

[20] T.J. Santner, B.J. Williams, and W.J. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[21] M. Schlather. Simulation and analysis of random fields, 2001. URL: http://www2.hsu-hh.de/schlath/R/RandomFields.

[22] Michael L. Stein. *Interpolation of Spatial Data, Some Theory for Kriging*. Springer, 1999.

[23] T.J. Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, 8 (6):1375–1381, 1980.

[24] Grace Wahba. *Spline Models for Observational Data*. S.I.A.M., 1990.

11.2 Mélanges de Krigeage pour l'optimisation

1. Présentations (* : en conférences internationales avec comité de lecture)
   - 7^{th} ENBIS conference*
     - Lieu et date : Dortmund (Allemagne), Septembre 2007.
     - Proceedings : CD.
   - 7^{th} Workshop on Quality Improvement Methods
     - Lieu et date : Bommerholz (Allemagne), 23-24 Mai 2008.
     - Invited talk.

2. Publication de revue scientifique
   - Journal : *Quality and Reliability International* (Wiley).
   - Statut : version finale acceptée en Mai 2008, parue en Aout 2008.
   - pré-print disponible sur HAL.

Discrete Mixtures of Kernels for Kriging-based Optimization

David Ginsbourger, Céline Helbert, Laurent Carraro
Département 3MI, Ecole Nationale Supérieure des Mines
158 cours Fauriel, 42023 Saint-Etienne, France
"last name"@emse.fr

July 3, 2008

Abstract: Kriging-based exploration strategies often rely on a single Kriging model which parametric covariance kernel is selected *a priori* or on the basis of an initial data set. Since choosing an unadapted kernel can radically harm the results, we wish to reduce the risk of model misspecification. Here we consider the simultaneous use of multiple kernels within Kriging. We give the equations of discrete mixtures of Kriging, and derive a multikernel version of the *expected improvement* optimization criterion. We finally provide an illustration of the *Efficient Global Optimization* algorithm with mixed exponential and Gaussian kernels, where the parameters are estimated by *Maximum Likelihood* and the mixing weights are likelihood ratios.

Key words: *Gaussian Processes, Global Optimization, Kernel Selection, Mixtures of Experts*

The global optimization of numerical simulators is a challenging problem as the number of runs is severely limited by computation time. Furthermore, the derivatives are generally not available. For the past decade, Kriging-based derivative-free algorithms such as *Efficient Global Optimization* ([9]) have been developed to address this issue. Kriging metamodels are indeed convenient for building exploration strategies since they provide for every potential input vector, both a mean predicted response value (Kriging mean) and an associated measure of accuracy (Kriging variance). In this paper, the simulator is seen as a determinist numerical black-box function y with d-dimensional input:

$$y : \mathbf{x} \in D \subset \mathbb{R}^d \rightarrow y(\mathbf{x}) \in \mathbb{R} \tag{1}$$

The function y is known at first on the initial Design of Experiments $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^{n_0}\}$, where $n_0 \in \mathbb{N}$ is the number of initial runs. We denote by $\mathbf{Y} = \{y(\mathbf{x}^1), ..., y(\mathbf{x}^{n_0})\}$ the set of observations made by evaluating y at the points of $\mathbf{X}$. In almost all Kriging models, the starting point is to make the assumption that y is one realization of a random process of the following form:

$$Y(x) = \mu(x) + \varepsilon(x) \tag{2}$$

where $\mu(x)$ is a deterministic trend function, $\varepsilon(x)$ is a centered stationary random field with covariance function k, here assumed to belong to a set of positive-definite stationary kernels:

$$\mathcal{K} = \{k_{(r,\sigma^2,\psi)} : h \in D - D \longrightarrow \sigma^2 r(h;\psi),\ r \in \mathcal{R},\ \sigma^2 \in \mathbb{R}^+,\ \psi \in \Psi_r\} \tag{3}$$

$\mathcal{K}$ is indexed by a finite set $\mathcal{R}$ of correlation kernel parametric families, by their respective continous hyperparameters $\psi \in \Psi_r$ (e.g. correlation lengths), and by a positive parameter σ^2

(the process variance). In the following, $\chi = (r, \sigma^2, \psi)$ denotes the tree-structured covariance parameters. In many industrial applications, r is arbitrarily chosen to belong to a parametric family (exponential, Gaussian, Matèrn, etc...), (σ^2, ψ) are then fitted to the data using automatic estimation procedures, and χ is finally plugged in as if it were known. Our particular concern here is to review and extend the EGO Algorithm ([9]) in taking the risk of model into account. After recalling some basics about Gaussian processes and metamodel-based optimization, we propose an adaptation of Ordinary Kriging (OK) with mixed kernels. We then derive an optimization criterion based on a discrete mixture of Kriging, and finally illustrate its efficiency by applying EGO with two simultaneous kernels to a classical test case function.

1 Gaussian Processes and Metamodel-based Optimization

OK is a spatial interpolator developed by G. Matheron and named after the mining engineer D.G. Krige. It provides at each point $\mathbf{x} \in D$ a prediction of Y as a linear combination of the observed values $\mathbf{Y}$. The weights depend on the distance between the prediction point $\mathbf{x}$ and the design of experiments $\mathbf{X}$ through the chosen covariance kernel. Here we give the equations of Kriging for a fixed kernel $k_\chi(.) = \sigma^2 r(.;\psi)$. The Kriging mean m_χ and mean squared error (or variance) s^2_χ at $\mathbf{x}$ are the following functions (see [23] for pointwise derivation):

$$\left\{ \begin{array}{c} m_\chi(\mathbf{x}) = \widehat{\mu}_\chi + \mathbf{k}_\chi(\mathbf{x})^T \mathbf{K}_\chi^{-1} \left(\mathbf{Y} - \widehat{\mu}_\chi \mathbf{1}_n\right) \\ s^2_\chi(x) = \sigma^2 - \mathbf{k}_\chi(\mathbf{x})^T \mathbf{K}_\chi^{-1} \mathbf{k}_\chi(\mathbf{x}) + \left(\mathbf{1}_n^T \mathbf{K}_\chi^{-1} \mathbf{1}_n\right)^{-1} \left(1 - \mathbf{1}_n^T \mathbf{K}_\chi^{-1} \mathbf{k}_\chi(\mathbf{x})\right)^2 \end{array} \right. \tag{4}$$

where we recall that $\chi = (r, \sigma^2, \psi)$. $\mathbf{K}_\chi$ and $\mathbf{k}_\chi(\mathbf{x})$ are the matrices [1] :

$$\mathbf{K}_\chi = \begin{pmatrix} k_\chi(0) & k_\chi(\mathbf{x}_1 - \mathbf{x}_2) & ... & k_\chi(\mathbf{x}_1 - \mathbf{x}_n) \\ k_\chi(\mathbf{x}_2 - \mathbf{x}_1) & k_\chi(0) & ... & k_\chi(\mathbf{x}_2 - \mathbf{x}_n) \\ ... & ... & .... & .... \\ k_\chi(\mathbf{x}_n - \mathbf{x}_1) & ... & .... & k_\chi(0) \end{pmatrix} \text{ and } \mathbf{k}_\chi(\mathbf{x}) = \begin{pmatrix} k_\chi(\mathbf{x} - \mathbf{x}_1) \\ k_\chi(\mathbf{x} - \mathbf{x}_2) \\ ... \\ k_\chi(\mathbf{x} - \mathbf{x}_n) \end{pmatrix}$$

and $\widehat{\mu}_\chi$ is given by: $\widehat{\mu}_\chi = (\mathbf{1}^T \mathbf{K}_\chi^{-1} \mathbf{1})^{-1} \mathbf{1}^T \mathbf{K}_\chi^{-1} \mathbf{Y}$. The classical geostatistical interpretation of Equation(4) is to see $m_\chi(\mathbf{x})$ as the best linear unbiased predictor of $Y(\mathbf{x})$, under the hypothesis that Equation (2) holds with $\mu(\mathbf{x})$ being an unknown constant μ estimated by maximum likelihood. Here we find it more convenient to consider an interpretation of OK in terms of Gaussian processes, in the flavour of ([3]). Assuming that $\varepsilon(\mathbf{x})$ is a centered stationary Gaussian process with known covariance function $k_\chi(.)$ and that μ is an unknown constant with improper uniform ([5]) prior distribution $\mu \sim \mathcal{U}(\mathbb{R})$, we obtain the following conditional distribution for $Y(\mathbf{x})$

$$Y^{OK}_\chi(\mathbf{x}) := [Y(\mathbf{x}) | Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}\left(m_\chi(\mathbf{x}), s^2_\chi(\mathbf{x})\right) \tag{5}$$

This approach allows the analytical calculation of various quantities involving $Y(\mathbf{x})$ knowing the observations, as well as conditional simulations[2] of Y, which ensures for instance that the expectation of any function involving $Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}$ can be estimated by Monte Carlo. In the practice, one chooses a parametric correlation kernel r, then estimates the parameters (σ^2, ψ), and finally plugs in the estimated values in the formulas. It seems however that such a sketch forgets the uncertainty associated with the choice of r and the estimation of (σ^2, ψ) and hence underestimates the modeling uncertainty. Assessing uncertainty with a variance s^2_χ obtained by

[1] In the deterministic case, these equations are often written in an equivalent way using correlation matrices.

[2] A conditional simulation on a fine grid covering D can only be performed when D is low-dimensional (typically up to $d = 3$). However, conditional simulations at a small set of points are affordable whatever the dimension d.

plugging in χ in the Kriging equations entails a hidden risk of trusting too much a "bad" model.

Maximum Likelihood (ML) is one of the standard ways, along with cross-validation, to estimate μ and (σ^2, ψ) on the basis of observations (see [3] for instance). It relies on the hypothesis that $\mathbf{Y}$ is a Gaussian vector with mean μ and covariance matrix K_χ (this can be seen as a direct consequence of the Gaussian process interpretation of Kriging). Considering that r is known, one searches for the parameters $(\widehat{\mu}, \widehat{\psi}, \widehat{\sigma^2})$ that give the largest density value to $\mathbf{Y}$. Noting [3] $R_\chi = \frac{1}{\sigma^2} K_\chi$, MLE then relies on the maximization of the Gaussian likelihood function:

$$L(\sigma^2, \psi, \mu; \mathbf{Y}) = f(\mathbf{Y}|\sigma^2, \psi, \mu) = \frac{1}{(2\pi)^{\frac{d}{2}} (\sigma^2)^{\frac{d}{2}} det(R_\chi)^{\frac{1}{2}}} e^{-\left[\frac{(\mathbf{Y}-\mu\mathbf{1})' R_\chi^{-1} (\mathbf{Y}-\mu\mathbf{1})}{2\sigma^2}\right]} \tag{6}$$

or equivalently on the minimization of $-2 \times log(L(\sigma^2, \psi, \mu; \mathbf{Y}))$. It can be shown that for every fixed ψ, the optimal μ and σ^2 are given by $\mu = \widehat{\mu_\chi}$ [4] , which doesn't depend on σ^2, and:

$$\widehat{\sigma^2}(\psi) = \frac{(\mathbf{Y} - \widehat{\mu_\chi}\mathbf{1})^T R_\chi^{-1} (\mathbf{Y} - \widehat{\mu_\chi}\mathbf{1})}{n} \tag{7}$$

After some direct calculations, ML can be restricted to the p_r-dimensional minimization problem:

$$\min_{\psi \in \Psi_r} \{log(|K_{(r, \widehat{\sigma^2}(\psi), \psi)}|)\} \tag{8}$$

This non-convex optimization problem is generally solved numerically, which adds both computational complexity, dependence on the starting point(s), and randomness in the result in the case of a stochastic optimization algorithm (like genetic algorithms using derivatives, see [29]). Furthermore, there is an inherent, non-reducible uncertainty due to estimating ψ from a limited number of observations. The variability of the parameters obtained by ML on the basis of data actually sampled from a Gaussian process has been studied in detail in the theory of likelihood ([28, 15, 18]), and more recently discussed in this particular framework in ([6]). Alternative estimation procedures dedicated to cope with this variability include restricted maximum likelihood methods ([23]), penalized maximum likelihood estimation ([24]), as well as fully Bayesian approaches ([2, 19]). However, we restrict ourselves here to the classical ML framework.

Once a Kriging interpolator and its associated uncertainty are computed, one dispose of a metamodel that may be used to predict the output at unknown sites, to propagate uncertainties, and last but not least to explore the simulator with a design dedicated to optimize y. One class of methods to derive such a design is by iteratively maximizing a figure of merit based on the Kriging metamodel. ([8]) and ([21]) provide a review of most used Kriging-based optimization criteria.

The expected improvement (EI) is a broadly used optimization criterion that makes a trade-off between promizing (with low predictions, for minimization) and uncertain zones. Let $y_{min} = \min\{y(\mathbf{x}_1), ..., y(\mathbf{x}_n)\}$ be the minimum of the currently known observation values. Let $\mathbf{x} \in D$ be a candidate point for a next evaluation of y. In the end, evaluating y at $\mathbf{x}$ would bring an improvement of $y_{min} - y(\mathbf{x})$ if $y(\mathbf{x})$ is below y_{min} and no improvement if $y(\mathbf{x})$ is above y_{min}. Of course, this improvement $(y_{min} - y(\mathbf{x}))^+$ cannot be known without evaluating y (else, we could

[3] There is implicitely no "nugget" effect since we work here with deterministic experiments

[4] Directly maximizing the likelihood with respect to μ and (σ^2, ψ) delivers the same value of μ as in the frame of OK, $\widehat{\mu} = \widehat{\mu_\chi}$. Note however that OK includes the variability due to μ's estimation in its prediction variance.

directly find the minimum). But Equation (5) makes it possible to know the statistical distribution of the random variable improvement $(y_{min} - Y(\mathbf{x}))^+$ conditionally on the observations $Y(\mathbf{X}) = \mathbf{Y}$. In particular, the expected improvement is defined as the following function of $\mathbf{x}$:

$$C_\chi^{EI}(\mathbf{x}) = \mathbb{E}\left[\left(y_{min} - Y_\chi^{OK}(\mathbf{x})\right)^+\right] = \mathbb{E}\left[(y_{min} - Y(\mathbf{x}))^+ \,|Y(\mathbf{X}) = \mathbf{Y}\right] \tag{9}$$

Thanks to Equation (5), the expected improvement can be calculated analytically (see [9]):

$$C_\chi^{EI}(\mathbf{x}) = (y_{min} - m_\chi(\mathbf{x}))\,\Phi\left(\frac{y_{min} - m_\chi(\mathbf{x})}{s_\chi(\mathbf{x})}\right) + s_\chi(\mathbf{x})\phi\left(\frac{y_{min} - m_\chi(\mathbf{x})}{s_\chi(\mathbf{x})}\right) \tag{10}$$

where ϕ and Φ are respectively the probability density and the cumulative distribution function of the standard Gaussian law. This expression sheds light on the trade-off between promising and uncertain zones: the first term of the sum enhances local search via the mean prediction $m_\chi(\mathbf{x})$, whereas the second term puts more emphasis on global search via the prediction variance. ([8]) and ([21]) intensively commented the criterion of EI. EGO is an algorithm proposed by ([9]).

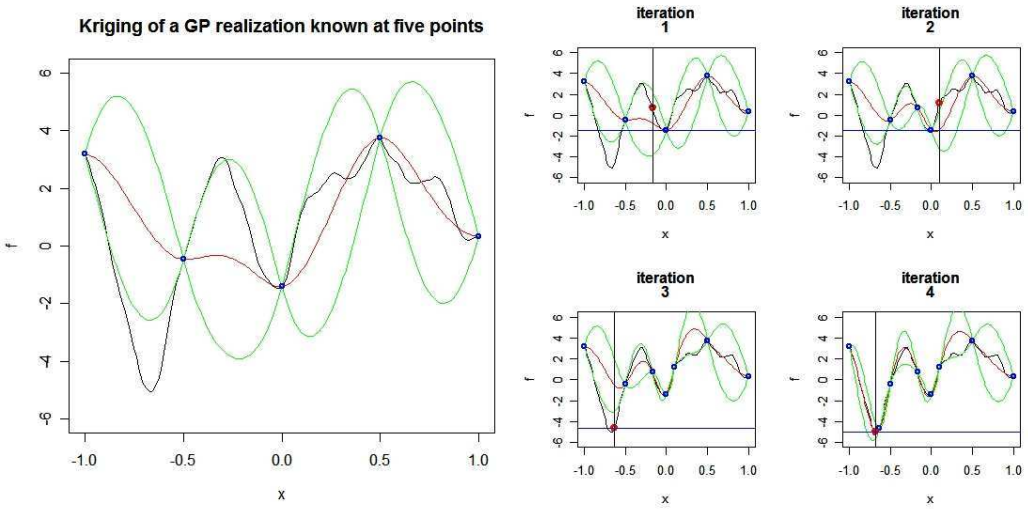


Figure 1: Left: Ordinary Kriging (the smooth interpolator, with light 95% confidence intervals) of y (dark curve), one realization of a Gaussian process with cubic ([17]) covariance function $c(h) = \sigma^2\left(1 - 7\left(\frac{h}{l}\right)^2 + 8.75\left(\frac{h}{l}\right)^3 - 3.5\left(\frac{h}{l}\right)^5 + 0.75\left(\frac{h}{l}\right)^7\right)\mathbf{1}_{[0,l]}(h)$, where $\sigma^2 = 4$ and $l = 0.6$. Right: 4 iterations of EGO applied to y. The first points, regularly spaced, represent the visited sites, the new points (and their associated vertical lines) are the current maximizers of the expected improvement. The horizontal line represents the current y_{min} values. This example illustrates how an EGO sequence explores the objective function without getting trapped in the zones of local optimum.

It relies on a sequential exploration based on OK and on the maximization of the expected improvement criterion (see Alg. 1). In the exact version ([9]), the algorithm loops until $(max_{x\in D}\{C_{\psi^*}^{EI}(x)\} < \delta)$ for a δ fixed by the user, depending on the past evaluations of the objective function (in the original E.G.O., δ is 1% *of the best current function value* [9]). Here we consider the case in which the number of runs $n \in \mathbb{N}$ is fixed in advance. EGO has been found to be a competitive global optimization algorithm, in particular in the fields of automotive and

Algorithm 1 The E.G.O. Algorithm

function EGO($\mathbf{X}$, $\mathbf{Y}$, n)
  for $i \leftarrow 1, n$ **do**
    $(\sigma^{2*}, \psi^*) = argmax_{(\sigma^2,\psi)\in\mathbb{R}^+\times\Psi_r} L(\sigma^2, \psi; Y(\mathbf{X}) = \mathbf{Y})$ ▷ Estimating (σ^2, ψ) by ML
    $\mathbf{x}^* = argmax_{\mathbf{x}\in D}\{C^{EI}_{(r,\sigma^{2*},\psi^*)}(\mathbf{x})\}$ ▷ Maximizing the Expected Improvement
    $\mathbf{X} = \mathbf{X}\bigcup\{\mathbf{x}^*\}$ *and* $\mathbf{Y} = \mathbf{Y}\bigcup\{y(\mathbf{x}^*)\}$ ▷ Updating the Design of Experiments
  end for
end function

aerospace engineering, with objective functions having one to eight inputs (see [8], [21], [11]). In other respects, ([22]) proposes a adaptation of EGO for physical experiments.

2 Mixtures of Kriging for prediction and optimization

Let us focus on the situation in which several kernels are in competition to model a sample of observed data. This is typically the case when different methods are available for the estimation of (σ^2, ψ) (e.g. ML with different initial values, ML and cross-validation, penalized ML with different penalty functions, etc.), or when there is a choice to make between a set of functional forms for the correlation kernel r. Let us assume that the function y has already been evaluated at a finite set of points $\mathbf{X}_{obs}$, and denote by $\mathbf{Y}_{obs}$ the associated responses. We now consider that $M \in \mathbb{N}$ *experts* $\{\mathcal{E}_i,\ i \in [1, M]\}$ are at disposal to estimate χ on the basis of the observations. The $\mathcal{E}_i$'s are functional estimators, providing at the same time a correlation structure and its associated parameters. They are defined as follows:

$$\forall i \in [1, M],\ \mathcal{E}_i : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \in \bigcup_{k=n_0}^{+\infty} \left(D^k \times \mathbb{R}^k\right) \longrightarrow \chi_i = \mathcal{E}_i(\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \in \mathcal{K} \tag{11}$$

For the sake of convenience, we identify $\mathcal{K}$ here with the set of possible χ's (there is an obvious one-to-one mapping between both sets). Given the set of kernels $\{\chi_1, ..., \chi_M\}$ delivered by the experts $\{\mathcal{E}_i, i \in [1, M]\}$ and $\mathcal{W} = \{w_1, ..., w_M\}$ (s.t. $\sum_{i=1}^M w_i = 1$) a set of weights meant to quantify the respective relevance levels of the M experts, we study the idea of replacing the classical approach of kernel selection by a mixture of kernels: instead of keeping the best kernel and dropping off the others, we propose to keep them all and integrate them within OK in probabilizing χ.

Discrete mixture of Gaussian processes: The unknown function y is now seen as one path of a random field associated with an OK model, which underlying kernel is independently chosen at random following a discrete law supported by the set of kernels delivered by the M experts:

$$\begin{cases} [Y^{OK}_{mix}|\chi] = Y^{OK}_{\chi} \\ P(\chi = \chi_i) = w_i \end{cases} \tag{12}$$

Note that the proposed approach is not strictly Bayesian: the *prior distribution* on χ, in the sense of a Bayesian framework, would depend on the data in this case. In this work, we first restrict ourselves to a small set of kernels selected on the basis of the available data, and choose to mix them afterwards: the observations are here used at every step, unlike what could be expected from a straight Bayesian procedure (More detail about that is given at the end of this

section). Following Equation (12), the conditional distribution of $Y(\mathbf{x})$ ($\mathbf{x} \in D$) is a mixture of Gaussians:

$$Y_{mix}^{OK}(\mathbf{x}) := [Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{Y}] \text{ with density function } \sum_{j=1}^{M} w_j p_{\mathcal{N}\left(m_{\chi_j}(\mathbf{x}), s^2_{\chi_j}(\mathbf{x})\right)}(.) \tag{13}$$

Ordinary Kriging with a mixed kernel: Following Equation (13), Y_{mix}^{OK} is a field of Gaussian mixtures. This entails the equations of the mixed mean [5] and variance [6]:

$$m_{mix}(\mathbf{x}) = \mathbb{E}[\mathbb{E}[Y_{mix}^{OK}(\mathbf{x})|\chi]] = \sum_{i=1}^{M} w_i m_{\chi_i}(\mathbf{x}) \tag{14}$$

Hence, the mean of the resulting metamodel is the weighted average of the means associated with the different Krigings (which coincides with the concept of weighted average surrogate model developped in [25]). Furthermore, the corresponding variance is given by

$$\begin{aligned} s^2_{mix}(\mathbf{x}) &= Var[Y_{mix}^{OK}(\mathbf{x})] = \mathbb{E}[Var[Y_{mix}^{OK}(\mathbf{x})|\chi]] + Var[\mathbb{E}[Y_{mix}^{OK}(\mathbf{x})|\chi]] \\ &= \sum_{i=1}^{M} w_i s^2_{\chi_i}(\mathbf{x}) + \sum_{i=1}^{M} w_i[(m_{\chi_i}(\mathbf{x}) - m_{mix}(\mathbf{x}))^2] \end{aligned} \tag{15}$$

The first term is a linear combination of the model variances weigthed by the $w_i's$, whereas the second term reflects the dispersion between the different Kriging means. The latter plays a capital role since it introduces data dependence in the Kriging variance [7] : contrary to the case of regular OK, the variance now depends on the observations $\mathbf{Y}$ through the second term.

Optimization under a mixture of kernels: when several values of χ are possible, finding the next most promizing point with a kriging-based optimization criterion (say C_{χ}^{EI}) becomes a multicriteria decisional problem. Our approach here is to combine all $C_{\chi_i}^{EI}$'s to provide a unified criterion that takes into account both sources of randomness. Using again the so-called law of total expectation, we derive the *mixed* EI:

$$\begin{aligned} C_{mix}^{EI}(\mathbf{x}) : &= \mathbb{E}\left[\left(y_{min} - Y_{mix}^{OK}(\mathbf{x})\right)^+\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(y_{min} - Y_{mix}^{OK}(\mathbf{x})\right)^+ |\chi\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(y_{min} - Y_{\chi}^{OK}(\mathbf{x})\right)^+\right]\right] = \sum_{i=1}^{M} w_i C_{\chi_i}^{EI}(\mathbf{x}) \end{aligned} \tag{16}$$

and hence, the expected improvement function under a mixture of kernels is simply the convex combination of the M expected improvement functions weighted by the $\{w_i, \in [1, M]\}$. Note that any integral criterion under a mixture of Krigings can be calculated in the same manner.

Selecting a benchmark of experts: replacing the step of model selection by a step dedicated to choosing a set of models may seem at first to create more problems than it solves indeed. For instance, if we consider several families of correlation kernels (e.g. a Gaussian, an exponential,

[5] Using the law of total expectation: $\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1|X_2]]$.
[6] Using the law of total variance: $Var[X_1] = \mathbb{E}[Var[X_1|X_2]] + Var[\mathbb{E}[X_1|X_2]]$
[7] Sometimes referred to as *heteroskedasticity of the variance with respect to the data*

and even nonstationary correlation kernels -as in [4]-) and estimate each set of kernel parameters by ML, it naturally increases the computational amount. The price for mixing is in that case to multiply the time needed for model inference by the number of experts. In this flavor, possible approaches would be to consider simultaneously experts relying on the same correlation kernel but with hyperparameters inferred using different methods (e.g. mixing the ML and the Leave-One-Out (LOO) "best" models), or even getting several candidate hyperparameter sets by parametric bootstrap. Ideally, we would like to have all relevant classes of experts represented in a small set. One of the future issues to be addressed seem to be the selection of sets with dissimilar good experts, i.e. of multiple kernels that fit the data well but provide very different results in prediction. In the frame of stationary Gaussian Processes, both regularity and anisotropy properties of the kernels consitute interesting features since they closely condition the behaviour of the corresponding process realizations (see [3] for instance). The example of section 3 presents a mixture bewteen one smooth (Gaussian) and one non-smooth (exponential) kernel.

Setting a probability measure over the set of experts: Once a set of experts is chosen, probability weights have to be defined. The most naïve way of probabilizing the models is to put a uniform distribution on them. This approach may be relevant when mixing models obtained by maximizing different criteria (e.g.: a $50\% - 50\%$ mix of the "best ML model" and the "best LOO model"). On the contrary, it is also possible to consider the density of the mixture of models as a function of both the covariance parameters and parametric weights and then to perform likelihood maximization over all parameters including the $w_i's$. Such problems are typically numerically solved using an *Expectation-Maximization* (EM) algorithm ([10]). We propose a way in between, more informative than a raw uniform distribution and yet computationally cheaper than EM. Since we have a criterion of fit (the Gaussian likelihood), why not use it to weight models? At first, we propose a Kriging mixture with weights based on the likelihood criterion. In what follows, we consider Akaike weights (see [14]):

$$\forall i \in [1, M],\ w_i = \frac{L(\chi_i; \mathbf{Y})}{\sum_{j=1}^{N} L(\chi_j; \mathbf{Y})} \tag{17}$$

The w_i's may be simply interpreted as likelihood[8] profile values divided by a normalization coefficient. Note that they may also be interpreted as conditional probabilities: putting a prior distribution π on χ, an application of Bayes' rule and the total probability formula gives

$$\begin{aligned} P(\chi = \chi_i|\mathbf{Y}) &= \frac{p(\mathbf{Y}|\chi = \chi_i)\pi(\chi_i)}{p(\mathbf{Y})} \\ &= \frac{p(\mathbf{Y}|\chi = \chi_i)\pi(\chi_i)}{\int p(\mathbf{Y}|\chi) d\pi(\chi)} \end{aligned} \tag{18}$$

Let us remark that the weigths w_i of Equation (17) coincide with $P(\chi = \chi_i|\mathbf{Y})$ when $\pi(\chi) = \frac{1}{M}\sum_{j=1}^{M} \delta_{\chi_i}(\chi)$. Recalling that $\chi_i = \mathcal{E}_i(\mathbf{X}_{obs}, \mathbf{Y}_{obs})$, the latter prior is clearly dependent on the observations so that the approach is not really Bayesian. At this stage, we do not know yet if any interpretation with a suitable prior can be found in order to describe this Akaike weighting in a fully Bayesian setting.

[8] The value of μ is here implicitely set to $\widehat{\mu_\chi}$ in the likelihood functions depending on χ.

3 EGO with mixed kernels, applied to Branin's function

The Branin-Hoo function has been intensively studied in the litterature of global optimization of black-box functions ([9]). It is a smooth two-variable function defined by:

$$y(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)cos(x_1) + 10, \ (x_1, x_2) \in [-5, 10] \times [0, 15]$$

y has three global minimizers $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$, and the global minimum is approximately equal to 0.4. We normalized the variables between 0 and 1. Now we wish to illustrate, and compare EGO with different kernels and kernel mixtures.

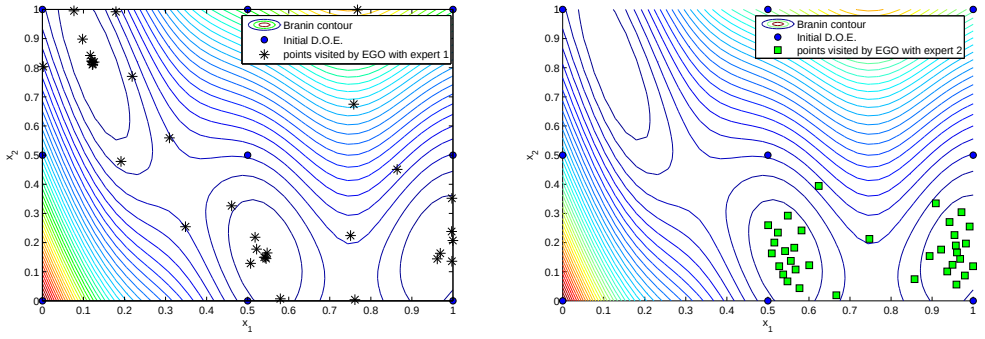


Figure 2: 25 iterations of the EGO algorithm applied to the Branin-Hoo function, with both experts $\mathcal{E}_1$ and $\mathcal{E}_2$ (see eq.(3)) and initial design $\mathbf{X}$ (in dark dots). Left: the path of EGO with expert $\mathcal{E}_1$ is represented by dark thin stars. Right: the path of EGO with expert $\mathcal{E}_2$ is represented by light squares. One of the three zones of minimum (upper left) is not visited.

The experimental set-up is the following: the initial design of experiments is a three-level full factorial design $\mathbf{X} \in ([0, 1] \times [0, 1])^{n_0}$ $(n_0 = 9)$. Two correlation kernels are selected:

Gaussian correlation	Exponential correlation
$r_1(h) = e^{-\frac{\|\|h\|\|^2}{p^2}}$	$r_2(h) = e^{-\frac{\|\|h\|\|}{p}}$

The experts considered here are the two parametric correlation kernels r_1 and r_2 with their respective correlation parameters and associated variances estimated by ML:

$$\begin{cases} \mathcal{E}_1 : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \rightarrow \chi_1 = (r_1, \sigma_1^{2*}, \psi_1^*), \text{ where } (\sigma_1^{2*}, \psi_1^*) = argmax_{\sigma^2,\psi}[L(\sigma^2, \psi; \mathbf{Y}_{obs}, r_1)] \\ \mathcal{E}_2 : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \rightarrow \chi_2 = (r_2, \sigma_2^{2*}, \psi_2^*), \text{ where } (\sigma_2^{2*}, \psi_2^*) = argmax_{\sigma^2,\psi}[L(\sigma^2, \psi; \mathbf{Y}_{obs}, r_2)] \end{cases}$$

All the algorithms and computations are implemented in the frame of the MatLab Simple Kriging free toolbox "Gaussian Processes for Machine Learning" (illustrating the book [3]). In both cases, the kernel hyperparameters intitial values are fixed to $(p, \sigma^2) = (0.1, 10)$.
The results are summarized in Figure (3). The left figure illustrates $n = 25$ iterations of EGO with mixed experts. The pattern of the visited points is close to the trajectory of EGO with Gaussian expert. In particular, the three zones of local optima are visited during the first 25 iterations. This similarities can be understood by looking at the sequences of weights plotted

Algorithm 2 The E.G.O. Algorithm with 2 mixed kernels weighted by their likelihood ratios

function EGO($\mathbf{X}$, $\mathbf{Y}$, n)
  for $i \leftarrow 1, n$ **do**
    for $j \leftarrow 1, 2$ **do**
      $(\sigma_j^{2*}, \psi_j^*) = argmax_{(\sigma_j^2, \psi_j) \in \mathbb{R}^+ \times \Psi_{r_j}} L(\sigma_j^2, \psi_j; Y(\mathbf{X}) = \mathbf{Y}, r_j)$ ▷ MLE
    end for
    for $j \leftarrow 1, 2$ **do**
      $w_j = \frac{L(\sigma_j^{2*}, \psi_j^*; Y(\mathbf{X}) = \mathbf{Y}, r_j)}{\sum_{j=1}^{2} L(\sigma_j^{2*}, \psi_j^*; Y(\mathbf{X}) = \mathbf{Y}, r_j)}$ ▷ Computing the mixing weights
    end for
    $\mathbf{x}^* = argmax_{\mathbf{x} \in D} \sum_{j=1}^{2} w_j C^{EI}_{(r_j, \sigma_j^{2*}, \psi_j^*)}(\mathbf{x})$ ▷ Maximizing the mixed EI
    $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^*\}$ *and* $\mathbf{Y} = \mathbf{Y} \bigcup \{y(\mathbf{x}^*)\}$ ▷ Updating the Design of Experiments
  end for
end function

on the right figure. The two curves on the graphic below represent the log-likelihood associated with both experts as functions of the number of EGO iterations. Note that the likelihood of expert 2 is greater than the likelihood of expert 1 until the number of iterations reaches 6, and then becomes significantly lower than the other one. The likelihood ratios plotted on the graphic above show more precisely how the exponential kernel prevails at the beginning of the EGO algorithm and is later dropped in favour of the Gaussian kernel. This kind of *automatic selection* seems due to an asymptotical steep decrease of the likelihood ratio. Using Akaike weights to mix kernels within a sequential exploration seems here to be a useful means to automatically select an expert without making a decision based on the initial design of experiments only. Hence, the proposed approach appears to be a sound option to increase EGO's robustness to modeling uncertainty.

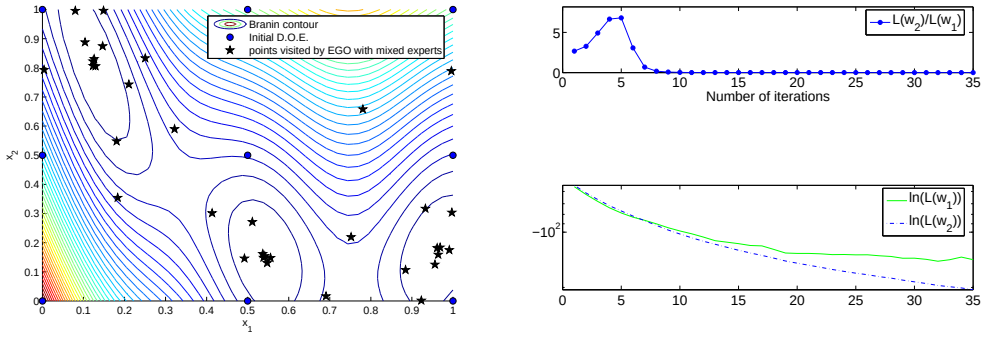


Figure 3: 25 iterations of EGO applied to the Branin-Hoo function, with a mixture of experts $\mathcal{E}_1$ and $\mathcal{E}_2$ weighted by Akaike weights. Left: the path of EGO with the mixture of experts is represented by dark filled stars. Right: evolution of both sequences of log-weights (lower graphic), and the associated series of likelihood ratios (upper graphic).

4 Conclusions

We have derived and discussed an optimization criterion, the *mixed expected improvement*, relying on discrete mixtures of Kriging metamodels with different covariance kernels. The presented framework of multiple experts allows one to handle several parametric correlation structures and/or different parameter estimation techniques within the same Kriging-based procedure. The application of the latter to the optimization of the Branin-Hoo function provided a first example, where mixing appears to be a succesful alternative to model selection based on initial data. It is illustrated that, possibly after an oscillating behaviour for a few iterations, mixing two *experts* within the EGO algorithm (Krigings with Gaussian and Exponential correlation structures, with Akaike weights) may ultimately lead to an automatic selection of a unique metamodel. The issues of selecting parsimonious benchmarks of experts, and of using weighting methods dedicated to different purposes are to be addressed in forthcoming works.

Aknowledgements: This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA, and Total S.A. We wish to thank Raphael T. Haftka and Victor Picheny for their help and useful comments. Special thanks to the R project people for developing such a useful freeware.

References

[1] Journel A. Fundamentals of geostatistics in five lessons. Technical report, Stanford Center for Reservoir Forecasting, 1988.

[2] O'Hagan A. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, 91(91):1290–1300, 2006.

[3] Rasmussen C.E. and Williams K.I. *Gaussian Processes for Machine Learning*. M.I.T. Press, 2006.

[4] Paciorek C.J. *Nonstationary Gaussian Processes for Regression and Spatial Modelling*. PhD thesis, Carnegie Mellon University, 2003.

[5] Robert C.P. *L'analyse statistique bayesienne*. Economica, Paris, 1992.

[6] Ginsbourger D., Dupuy D., Badea A., Roustant O., and Carraro L. On the selection and the estimation of kriging models for deterministic computer experiments. *submitted to ASMBI*, preprint at http://hal.archives-ouvertes.fr/hal-00270173/en, 2008.

[7] R development Core Team. R: A language and environment for statistical computing, 2006.

[8] Jones D.R. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(21):345–383, 2001.

[9] Jones D.R., Schonlau M., and Welch W.J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[10] McLachlan G.J. and Krishnan T. *The EM Algorithm and Extensions*. Wiley series in probability and statistics, 1997.

[11] Kracker H. Methoden zur analyse von computerexperimenten mit anwendung auf die hochdruckblechumformung. Master's thesis, Dortmund University, 2006.

[12] Vert J.P. Double mixture and universal inference. Technical Report DMA-00-15, Ecole Normale Supérieure, 2000.

[13] Koehler J.R. and Owen A.B. Computer experiments. Technical report, Department of Statistics, Stanford University, 1996.

[14] Burnham K.P. and Anderson D.R. *Model Selection and Multimodel Inference*. Springer, 1998.

[15] Mardia K.V. and Marshall R.J. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–46, 1984.

[16] Breiman L. Bagging predictors. Technical Report 421, Department of Statistics, University of California at Berkeley, 1994.

[17] Schlather M. Simulation and analysis of random field. *R News*, 1(2):18–20, 2001.

[18] Abt Markus and Welch William J. Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes. *The Canadian Journal of Statistics*, 26:127–137, 1998.

[19] Jay D. Martin and Timothy W. Simpson. A monte carlo simulation of the kriging model. In *Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization conference, Albany, NY, August 30-September 2, AIAA-2004-4324*, 2004.

[20] Jordan Michael I. and Jacobs Robert A. Hierarchical mixture of experts and the em algorithm. Technical report, Massachusetts Institute of Technology, 1993.

[21] Sasena Michael J., Papalambros Panos, and Goovaerts Pierre. Exploration of metamodeling sampling criteria for constrained global optimization. *Engineering Optimization*, 34 (3):263–278, 2001.

[22] Henkenjohann N., Göbel R., Kleiner M., and Kunert J. An adaptive sequential procedure for efficient optimization of the sheet metal spinning process. *Qual. Reliab. Engng. Int.*, 21:439–455, 2005.

[23] Cressie N.A.C. *Statistics for spatial data*. Wiley series in probability and mathematical statistics, 1993.

[24] Li R. and Sudjianto A. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(47):111–120, 2005.

[25] Goel T., Haftka R., Shyy W., and Queipo N. Simultaneous use of multiple surrogates. In *Proceedings of the 11th AIAA-ISSMO Multidisciplinary Analysis and Optimization, 6-8 September 2006, Portsmouth, Virginia*, 2006.

[26] Hastie T., Tibshirani R., and Friedman J. *The Elements of Statistical Learning*. Springer, 2001.

[27] Santner T.J., Williams B.J., and Notz W.J. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[28] Sweeting T.J. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, 8(6):1375–1381, 1980.

[29] Mebane Jr. Walter R. and Sekhon Jasjeet S. Genetic optimization using derivatives: Theory and application to non-linear models. *Political Analysis*, 7:187–210, 1998.

Authors' biographies:

David Ginsbourger graduated from both the Ecole des Mines de Saint-Etienne (EMSE, France) and Berlin Technical University (Germany) in 2005. He is a PhD candidate in applied mathematics at the EMSE. His research interests are Kriging and global optimization.

Céline Helbert graduated from the EMSE in 2000. She completed her PhD in applied mathematics in 2005 and is now assistant professor at the EMSE. Her research interests are metamodeling and Bayesian inference.

Laurent Carraro has been professor of probability and statistics at the EMSE since 1990. His research interests include functional approximation and random processes.

11.3 Parallélisation d'EGO

1. Présentation en conférence internationale avec comité de lecture

   - NCP07 « Non-Convex Programming »
     - Lieu et date : Rouen (France) Décembre 2007.
     - Proceedings : oui (Cf. [GLRC07]).

2. Publication de revue scientifique

   - Chapitre de livre : *Computational Intelligence in Expensive Optimization Problems* (Springer series Studies in Evolutionary Learning and Optimization).
   - Statut : contribution acceptée en Décembre 2008, à paraître.
   - pré-print disponible sur HAL.

A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes

David Ginsbourger, Rodolphe Le Riche*, Laurent Carraro
Département 3MI
Ecole Nationale Supérieure des Mines
158 cours Fauriel, Saint-Etienne, France
{ginsbourger, leriche, carraro}@emse.fr

March 3, 2008

Abstract

The optimization of expensive-to-evaluate functions generally relies on metamodel-based exploration strategies. Many deterministic global optimization algorithms used in the field of computer experiments are based on Kriging (Gaussian process regression). Starting with a spatial predictor including a measure of uncertainty, they proceed by iteratively choosing the point maximizing a criterion which is a compromise between predicted performance and uncertainty. Distributing the evaluation of such numerically expensive objective functions on many processors is an appealing idea. Here we investigate a multi-points optimization criterion, the *multipoints expected improvement* (q-$\mathbb{E}I$), aimed at choosing several points at the same time. An analytical expression of the q-$\mathbb{E}I$ is given when $q = 2$, and a consistent statistical estimate is given for the general case. We then propose two classes of heuristic strategies meant to approximately optimize the q-$\mathbb{E}I$, and apply them to Gaussian Processes and to the classical Branin-Hoo test-case function. It is finally demonstrated within the covered example that the latter strategies perform as good as the best Latin Hypercubes and Uniform Designs ever found by simulation (2000 designs drawn at random for every $q \in [1, 10]$).

Key words: Kriging, Expected Improvement, EGO, active learning, Monte-Carlo

*C.N.R.S. UMR 5146

1 Introduction

In many engineering applications, such as car crash tests, nuclear criticality safety, reservoir forecasting, the time needed to simulate the physical phenomena is so long that the experimenter can only afford a few simulation runs. It is common to see a deterministic simulator as a numerical black-box function

$$y : \mathbf{x} \in D \subset \mathbb{R}^d \rightarrow y(\mathbf{x}) \in \mathbb{R} \tag{1}$$

y is known at a Design of Experiments $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\} \in (\mathbb{R}^d)^d$, where $n \in \mathbb{N}$ is the number of initial runs or experiments. We denote by $\mathbf{Y} = \{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\}$ the set of observations made by evaluating y at the points of $\mathbf{X}$. The data $(\mathbf{X}, \mathbf{Y})$ provides information that help understanding the function y with an accuracy that depends on n, the geometry of $\mathbf{X}$, and the regularity of y. This partial knowledge of y is needed to build simplified representations of the simulator, also called *surrogate models* or *metamodels*. A metamodel can be used for predicting values of y outside the initial design or visualizing the influence of each variable on y ([9],[13],[18]). It may also guide further sampling decisions for various purposes, such as refining the exploration of the input space in preferential zones or optimizing the function y ([9]).This paper proposes metamodel-based optimization algorithms that are well-suited to parallelization since they yield several points at each iteration. The simulations associated with these points can be distributed on different processors, which helps performing the optimization when the simulations are calculation intensive. The algorithms are derived from a multi-points optimization criterion, named the multi-points expected improvement. Calculations are performed in the framework of Gaussian processes. In particular, the metamodel considered is Ordinary Kriging (see eqs. 3, 4, and 46).

2 Gaussian processes and sequential optimization

2.1 Ordinary Kriging

Probabilistic metamodeling seems to be particularly adapted for the optimization of black-box functions, as analyzed and illustrated in ([7]). Our work follows ([9]), where Ordinary Kriging (OK) is used to derive a sequential optimization strategy (EGO). Kriging is an interpolation method originally developped in geostatistics ([1],[16]). It provides a predictor of spatial phenomena, with a measure of uncertainty quantifying the accuracy of the prediction at each site (A full derivation is proposed in the appendix). Ordinary Kriging is based on the assumption that y is a realization of a stationary Gaussian process Y with unknown constant mean and known covariance structure ([4]). In Kriging-based optimization, one often abusively plugs in maximum likelihood covariance hyperparameters without taking the estimation variance into account ([9]). Here we commit this abuse, and work with the classical Ordinary Kriging equations. This has the advantage of delivering

a Gaussian posterior distribution, even if the uncertainty is slightly underestimated :

$$\forall \mathbf{x} \in D,\ [Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{OK}(\mathbf{x}), s^2_{OK}(\mathbf{x})) \tag{2}$$

where the kriging mean and variance functions are given by the following formulae ([16]):

$$m_{OK}(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \left[c(\mathbf{x}) + \left(\frac{1 - c(\mathbf{x})^T\Sigma^{-1}\mathbb{1}_n}{\mathbb{1}_n^T\Sigma^{-1}\mathbb{1}_n}\right)\mathbb{1}_n\right]^T \Sigma^{-1}\mathbf{Y} \tag{3}$$

$$s^2_{OK}(\mathbf{x}) = Var[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \left[\sigma^2 - c(\mathbf{x})^T\Sigma^{-1}c(\mathbf{x}) + \frac{(1 - \mathbb{1}_n^T\Sigma^{-1}c(\mathbf{x}))^2}{\mathbb{1}_n^T\Sigma^{-1}\mathbb{1}_n}\right] \tag{4}$$

with $c(x) = \left(cov(Y(\mathbf{x}), Y(\mathbf{x}^1)), ..., cov(Y(\mathbf{x}), Y(\mathbf{x}^n))\right)^T$, $\Sigma = \left(cov(Y(\mathbf{x}^i), Y(\mathbf{x}^j)\right)_{i,j\in[1,n]}$, and $\sigma^2 = Var[Y(\mathbf{x})]$ (which is not depending on $\mathbf{x}$ since Y is stationary).

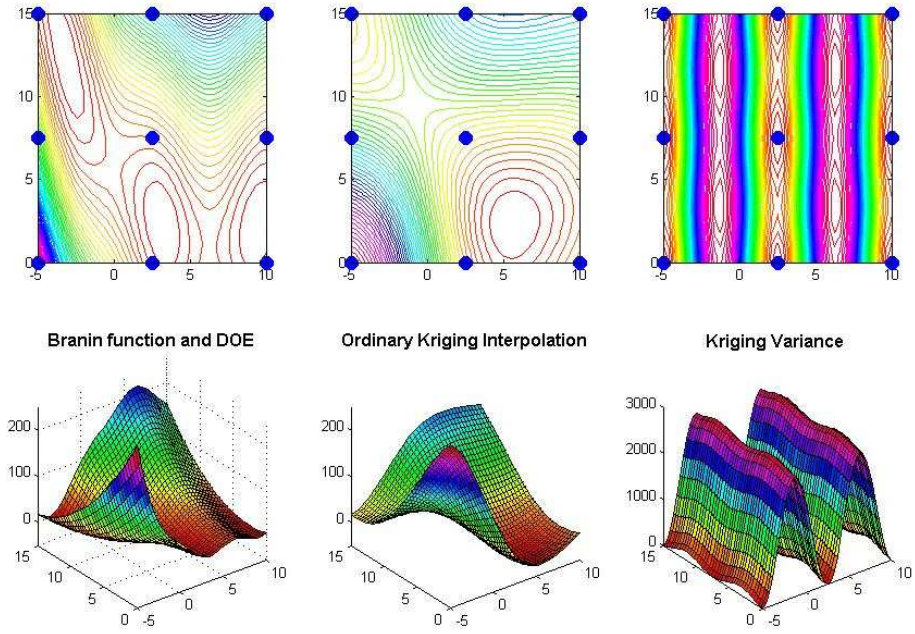


Figure 1: Ordinary Kriging of the Branin-Hoo function (function, Kriging mean value and variance, from left to right). The design of experiments is a 3×3 factorial design. The covariance is an anisotropic squared exponential with parameters estimated by gaussian likelihood maximization ([16]).

In other terms, under the Gaussian process assumptions that have been made, the random variable $Y(\mathbf{x})$ knowing previous observations $\{Y(\mathbf{x^1}), ..., Y(\mathbf{x^n})\}$ follows a normal distribution which mean and variance are $m_{OK}(\mathbf{x})$ and $s^2_{OK}(\mathbf{x})$, respectively.
A full bayesian interpretation can be found in ([18]), or more recently in ([10]). Classical properties of Ordinary Kriging include that $\forall i \in [1, n]$ $m_{OK}(\mathbf{x}^i) = y(\mathbf{x}^i)$ and $s^2_{OK}(\mathbf{x}^i) = 0$, therefore $[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}]$ is interpolating. Note that $[Y(\mathbf{x}^a)/Y(\mathbf{X}) = \mathbf{Y}]$ and $[Y(\mathbf{x}^b)/Y(\mathbf{X}) = \mathbf{Y}]$ are correlated random variables, where $\mathbf{x}^a$ and $\mathbf{x}^b$ are arbitrary points of D (see Appendix C.2).

The OK metamodel of the Branin-Hoo function (see eq. (25)) is plotted on fig. (2.1). The OK interpolation (upper middle) is made only on the basis of the 9 observations (as can be seen in eq. 3). Even if the shape is reasonably respected (lower middle), the contour of the interpolator shows an artificial optimal zone (upper middle, around the point $(6, 2)$). In other respects, the variance is not depending on the observations[1] (see eq. (4)). Note the particular shape of the variance, due to the strong anisotropy of the covariance function estimated by likelihood maximization.

2.2 Kriging-based optimization criteria

Such a Gaussian process regression has been used for optimization (minimization, by default). There is a detailed review of existing optimization methods relying on a metamodel in [7]. It analyzes and illustrates why directly optimizing a deterministic metamodel (like a spline, a polynomial, or the kriging mean) may be dangerous, and does not even necessarily lead to a local optimum. Kriging-based sequential optimization strategies (as developped in [9], and commented in [7]) address the issue of converging to non (locally) optimal points, by taking the kriging variance term into account (hence encouraging the algorithms to explore outside the already visited zones). Such optimization algorithms produce one point at each iteration that maximizes a figure of merit (or criterion) based upon $[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}]$. In essence, the criteria balance kriging mean prediction and uncertainty.

2.2.1 Visiting the point with highest uncertainty: maximizing s_{OK}

The fundamental mistake of minimizing the Kriging mean (m_{OK}) when globally minimizing a function is that no account is done of the uncertainty associated with m_{OK}. At the extreme inverse, it is possible to define the next optimization iterate as the least known point in D,

$$\mathbf{x}' = argmax_{\mathbf{x} \in D} s_{OK}(\mathbf{x}) \tag{5}$$

This procedure defines a series of $\mathbf{x}'$s which will fill the space D (it is dense in D) and, in this sense, it will ultimately locate $\mathbf{x}^*$, a global optimum. Yet, since no use is made of

[1] phenomenon known as homoskedasticity of the Kriging variance with respect to the observations ([16])

previously obtained $\mathbf{Y}$ information (look at formula (4) for s^2_{OK}), there is no bias in favor of high performance regions. Maximizing the uncertainty is inefficient in practice.

2.2.2 Compromizing between m_{OK} and s_{OK}

The most general formulation for compromizing between the exploitation of previous simulations brought by m_{OK} and the exploration based on s_{OK} is the two criteria problem

$$\begin{cases} \min_{\mathbf{x}\in D} m_{OK}(\mathbf{x}) \\ \text{and } \max_{\mathbf{x}\in D} s_{OK}(\mathbf{x}) \end{cases} \tag{6}$$

Let $\mathcal{P}$ denote the Pareto set of solutions [2]. Finding one (or many) elements in $\mathcal{P}$ remains a difficult problem since $\mathcal{P}$ typically contains an infinite number of points. A comparable approach called *direct* ([8])), although not based on Kriging, is described in ([8]) : the metamodel is piecewise constant and the uncertainty measure is an Euclidean distance to already known points. The space D is discretized and the Pareto optimal set defines areas where discretization is refined. The method becomes computationally expensive as the number of iterations and dimensions increase. Note that ([3]) proposes a parallelized version of *direct*.

2.2.3 Maximizing the probability of improvement

Among the numerous criteria presented in [7] and [12], the probability of improving the function beyond the currently known minimum $min(\mathbf{Y}) = \min\{y(\mathbf{x}^1), ..., y(\mathbf{x}^n)\}$ seems to be one of the most fundamental:

$$PI(\mathbf{x}) = P(Y(\mathbf{x}) \leq \min(\mathbf{Y})/Y(\mathbf{X}) = \mathbf{Y}) \tag{7}$$

$$= \mathbb{E}[\mathbb{1}_{Y(\mathbf{x})\leq min(\mathbf{Y})}/Y(\mathbf{X}) = \mathbf{Y}] = \Phi\left(\frac{\min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right) \tag{8}$$

$\min(\mathbf{Y})$ is sometimes replaced by some arbitrary target $T \in \mathbb{R}$. The PI criterion is known to provide a very local search whenever the value of T is close to $min(\mathbf{Y})$. Taking several T's is a remedy proposed by [7] to force global exploration.

2.2.4 Maximizing the expected improvement

An alternative solution is to maximize the *expected improvement*

$$EI(\mathbf{x}) = \mathbb{E}[\max\{0, \min(\mathbf{Y}) - Y(\mathbf{x})\}/Y(\mathbf{X}) = \mathbf{Y}] \tag{9}$$

that additionally takes into account the magnitude of the potential improvement. EI measures how much improvement is expected when sampling at x. *In fine*, the improvement

[2] Definition of the Pareto front of $(s_{OK}, -m_{OK})$: $\forall x \in \mathcal{P}, \nexists\, y \in D : (m_{OK}(y) < m_{OK}(x)$ and $s_{OK}(y) \geq s_{OK}(x))$ or $(m_{OK}(y) \leq m_{OK}(x)$ and $s_{OK}(y) > s_{OK}(x))$

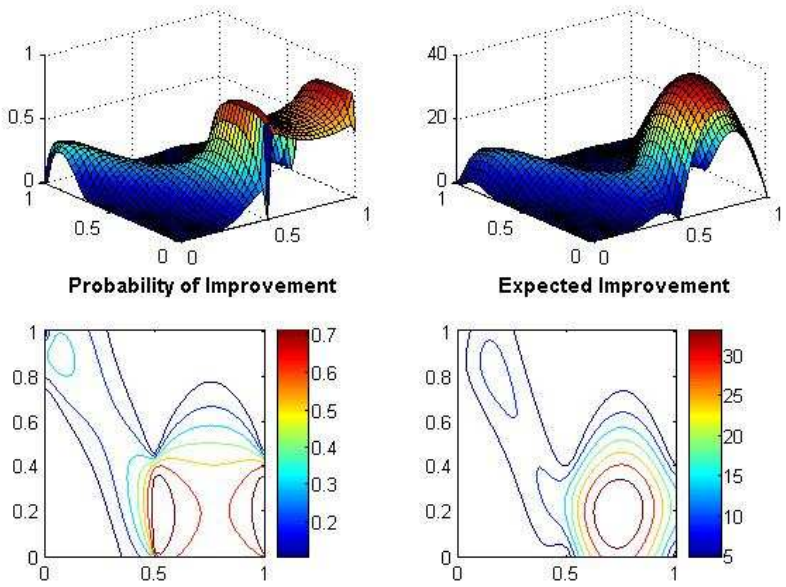


Figure 2: PI and EI surfaces of the Branin-Hoo function (same design of experiments, Kriging model, and covariance parameters as in fig. (2.1)). Maximizing PI leads to sample near the good points (associated with low observations) whereas maximizing EI leads here to sample between the good points. By construction, both criteria are null at the design of experiments, but the probability of improvement is very close to $\frac{1}{2}$ in a neighborhood of the point(s) where the function takes its lower observed value.

will be 0 if the actual $y(\mathbf{x})$ is above $min(\mathbf{Y})$ and $min(\mathbf{Y}) - y(\mathbf{x})$ in the opposite case. Since we know the conditional distribution of $Y(\mathbf{x})$, it is straightforward to calculate EI in closed form (see [9]):

$$\begin{aligned} EI(\mathbf{x}) &= \mathbb{E}[(min(\mathbf{Y}) - Y(\mathbf{x}))\mathbb{1}_{Y(\mathbf{x}) \leq min(\mathbf{Y})} / Y(\mathbf{X}) = \mathbf{Y}] \\ &= (min(\mathbf{Y}) - m_{OK}(\mathbf{x}))\Phi\left(\frac{min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right) + s_{OK}(\mathbf{x})\phi\left(\frac{min(\mathbf{Y}) - m_{OK}(\mathbf{x})}{s_{OK}(\mathbf{x})}\right) \end{aligned} \tag{10}$$

where ϕ and Φ stand for the probability density function and cumulative distribution function of the standard normal law $\mathcal{N}(0,1)$. EI represents a trade-off between promising and uncertain zones. EI has important properties for sequential exploration: it is null at the

already visited sites, and positive everywhere else with a magnitude that is increasing with the Kriging variance and with the decreasing Kriging mean (EI maximizers are indeed part of the Pareto front of $(s_{OK}, -m_{OK})$). Such features are usually demanded from global optimization procedures (see [8] for instance). The expected improvement and the probability of improvement are compared in fig. (2).

2.2.5 The *Stepwise Uncertainty Reduction* strategy

The stepwise uncertainty reduction (SUR) strategy has been introduced in ([5]) and extended to global optimization in ([12]). By looking at possible objective functions as conditional processes, $Y(\mathbf{x})/\mathbf{Y}$, it is possible to define $\mathbf{x}^*/\mathbf{Y}$, the random vector of the location of the minimizer of $Y(\mathbf{x})/\mathbf{Y}$, of density $p_{\mathbf{x}^*/\mathbf{Y}}(\mathbf{x})$. The uncertainty about the location of the optimum of $Y(x)$ is measured as the entropy of $p_{\mathbf{x}^*/\mathbf{Y}}(\mathbf{x})$, $H(\mathbf{x}^*/\mathbf{Y})$. $H(\mathbf{x}^*/\mathbf{Y})$ diminishes as the distribution of $\mathbf{x}^*/\mathbf{Y}$ gets more peaked. Conceptually, the SUR strategy for global optimization chooses as next iterate the point that specifies the most the location of the optimum,

$$\mathbf{x}' = argmin_{\mathbf{x}\in D} H(\mathbf{x}^*/\mathbf{Y}, Y(\mathbf{x})) \tag{11}$$

In practice, $p_{\mathbf{x}^*/\mathbf{Y}}(\mathbf{x})$ is estimated by Monte-Carlo sampling of $Y(\mathbf{x})/\mathbf{Y}$ at a finite number of locations in D, which may become a problem in high dimensional D's as the number of locations must geometrically increase with the number of dimensions to properly fill the space. The SUR criterion is different in nature from the other criteria presented so far in that it does not maximize an immediate (i.e. at the next iteration) payoff defined in terms of Y but rather lays the foundation of a more delayed payoff by gaining a more global knowledge on Y (reduce the entropy of its optima). The multi-points expected improvement criterion introduced in the present article also uses a delayed payoff measure.

2.2.6 The *Efficient Global Optimization* (EGO) algorithm

The EGO algorithm ([9]) relies on the EI criterion. Starting with an initial Design $\mathbf{X}$ (typically a Latin Hypercube), EGO sequentially visits the current global maximizer of EI (say the first visited one if there is more than one global maximizer) and updates the Kriging metamodel at each iteration, including hyperparameters re-estimation:

1. `Evaluate` y `at` $\mathbf{X}$`, set` $\mathbf{Y} = y(\mathbf{X})$ `and estimate covariance parameters of` Y `by MLE (Maximum Likelihood Estimation)`
2. `While stopping criterion not met`
   (a) `Compute` $\mathbf{x}' = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$`, set` $\mathbf{X} = \mathbf{X} \cup \{\mathbf{x}'\}$ `and` $\mathbf{Y} = \mathbf{Y} \cup \{y(\mathbf{x}')\}$
   (b) `Re-estimate covariance parameters by MLE`

After having been developed and applied in [15], EGO has been considered as a reference and has inspired contemporary works in optimization of expensive-to-evaluate functions. For instance, ([11]) exposes some EGO-based methods for the optimization of noisy black-box functions. ([14]) proposes an adaptation of EGO to multi-objective optimization.
EGO does not allow parallel evaluations of y, which is desirable for costly simulators (for instance, a crash-test simulation run typically lasts 24 hours). Here we present a criterion meant to choose an arbitrary number of points without intermediate evaluations of y.

3 The multi-points expected improvement

The main objective of this article is to propose and analyze a global optimization criterion, the multi-points expected improvement or q-points EI, that yields many (say q) points. Since the q-points EI is an extension of the expected improvement, all derivations are performed within the framework of Ordinary Kriging. Such criterion is the first step towards a parallelized version of the EGO algorithm [9]. It also departs, like the SUR criterion, from other criteria that look for an immediate payoff.

The q-points EI criterion (as already defined but not developed in ([15]) under the name "q-step EI") is the expectation of the improvement brought by the q considered points:

$$\begin{aligned} EI(\mathbf{x}^{n+1},...,\mathbf{x}^{n+q}) &= \mathbb{E}\left[\max\left\{(min(\mathbf{Y}) - Y(\mathbf{x}^{n+1}))^+,...,(min(\mathbf{Y}) - Y(\mathbf{x}^{n+q}))^+\right\}/Y(\mathbf{X}) = \mathbf{Y}\right] \\ &= \mathbb{E}\left[\left(\min\left(\mathbf{Y}\right) - \min\left(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q})\right)\right)^+/Y(\mathbf{X}) = \mathbf{Y}\right] \end{aligned} \tag{12}$$

Hence, the q-points EI may be seen as the regular EI applied to the random variable $\min(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q}))$. We have to deal with a minimum of dependent random variables. Fortunately, classical results of multivariate statistics[3] provide us with the exact joint distribution of the q unknown responses conditionally on the observations:

$$[(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+q}))/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}((m_{OK}(\mathbf{x}^{n+1}),...,m_{OK}(\mathbf{x}^{n+q})), S_q) \tag{13}$$

where the elements of the conditional covariance matrix S_q are:

$$\begin{aligned} (S_q)_{i,j} =& c(\mathbf{x}^{n+i} - \mathbf{x}^{n+j}) - \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}) \\ &+ \sigma^2\left[\frac{(1 - \mathbb{1}_n^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+i}))(1 - \mathbb{1}_n^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}))}{\mathbb{1}_n^T\Sigma^{-1}\mathbb{1}_n}\right] \end{aligned} \tag{14}$$

A full derivation of the joint Simple and Ordinary Kriging predictors and some overall considerations about the minimum of dependent random variables are presented respectively in the Appendix C.2.-C.4. and A.1.

[3] Cochran's theorem for the projection of Gaussian vectors

3.1 Analytical calculation of 2-EI

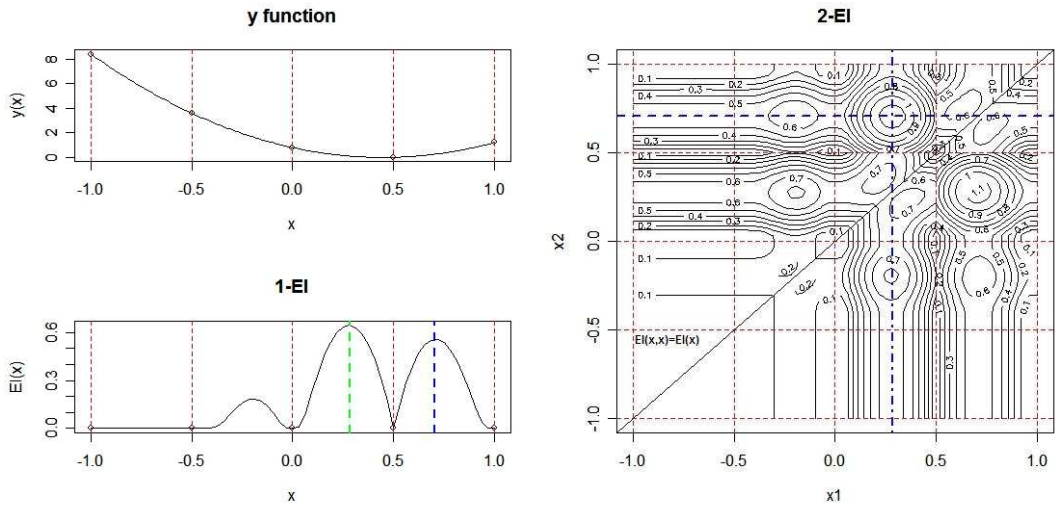


Figure 3: 1-point EI (lower left) and 2-points EI (right) functions associated with a monodimensional quadratic function ($y(x) = 4 \times (x - 0.45)^2$ known at $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. The ordinary kriging has here a cubic covariance with parameters $\sigma^2 = 10,\ \text{scale} = 0.9$).

2-EI can be derived as an expression depending on the mono- and bi-dimensional Gaussian cdf's. Using the following decomposition

$$
\begin{aligned}
&EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2}) \\
&= \mathbb{E}[(\min(\mathbf{Y}) - min(Y(\mathbf{x}^{n+1}), Y(\mathbf{x}^{n+2})))\mathbb{1}_{min(Y(\mathbf{x}^{n+1}),Y(\mathbf{x}^{n+2}))\leq \min(\mathbf{Y})}/Y(\mathbf{X}) = \mathbf{Y}] \\
&= \mathbb{E}[(\min(\mathbf{Y}) - Y(\mathbf{x}^{n+1}))\mathbb{1}_{Y(\mathbf{x}^{n+1})\leq \min(\mathbf{Y})}\mathbb{1}_{Y(\mathbf{x}^{n+1})\leq Y(\mathbf{x}^{n+2})}/Y(\mathbf{X}) = \mathbf{Y}] \\
&+ \mathbb{E}[(\min(\mathbf{Y}) - Y(\mathbf{x}^{n+2}))\mathbb{1}_{Y(\mathbf{x}^{n+2})\leq \min(\mathbf{Y})}\mathbb{1}_{Y(\mathbf{x}^{n+2})\leq Y(\mathbf{x}^{n+1})}/Y(\mathbf{X}) = \mathbf{Y}] \\
&= EI(\mathbf{x}^{n+1}) + EI(\mathbf{x}^{n+2}) \\
&- \mathbb{E}[(\min(\mathbf{Y}) - Y(\mathbf{x}^{n+1}))\mathbb{1}_{Y(\mathbf{x}^{n+1})\leq \min(\mathbf{Y})}\mathbb{1}_{Y(\mathbf{x}^{n+1})\geq Y(\mathbf{x}^{n+2})}/Y(\mathbf{X}) = \mathbf{Y}] \\
&- \mathbb{E}[(\min(\mathbf{Y}) - Y(\mathbf{x}^{n+2}))\mathbb{1}_{Y(\mathbf{x}^{n+2})\leq \min(\mathbf{Y})}\mathbb{1}_{Y(\mathbf{x}^{n+2})\geq Y(\mathbf{x}^{n+1})}/Y(\mathbf{X}) = \mathbf{Y}]
\end{aligned}
$$

one can analytically calculate $EI(\mathbf{x}^{n+1}, \mathbf{x}^{n+2})$. A complete derivation of the 2-points EI and some basic properties are proposed in the appendix A.2. and A.3.
fig. (3.1) represents the 1-EI and the 2-EI contour plots associated with a deterministic polynomial function known at 5 points. The 1-point EI advises here to sample between the "good points" of the initial design. The 2-points EI contour illustrates some general

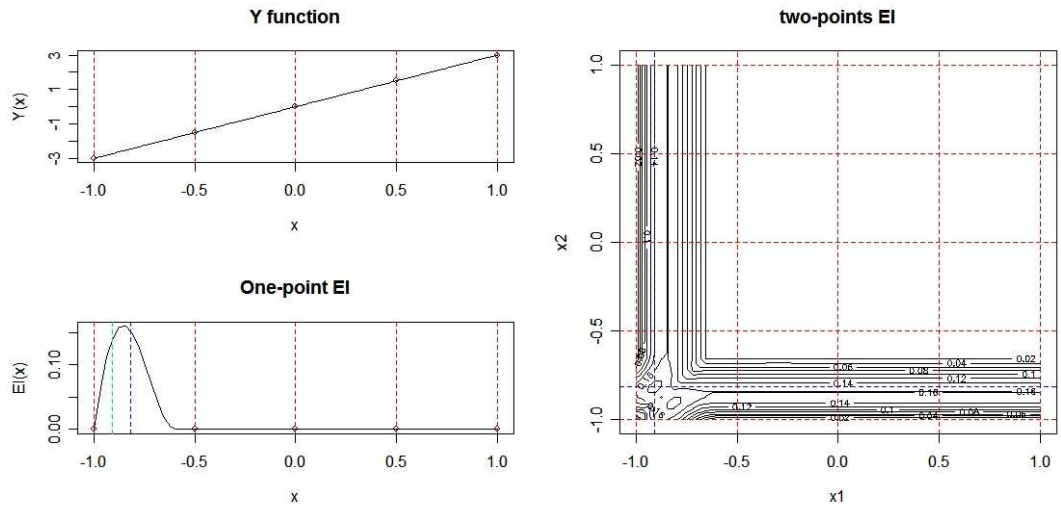


Figure 4: 1-point EI (lower left) and 2-points EI (right) functions associated with a monodimensional linear function ($y(x) = 3 \times x$) known at $\mathbf{X} = \{-1, -0.5, 0, 0.5, 1\}$. The ordinary kriging has here a cubic covariance with parameters $\sigma^2 = 10,\ \text{scale} = 1.4$).

properties: 2-EI is symmetric and its diagonal equals the 1-point EI, which can be easily seen by coming back to the definitions. Roughly said, 2-EI is high whenever the 2 points have high 1-EI and are reasonably distant from another (precisely, in the sense of the metric used in kriging). Additionally, maximizing 2-EI selects here the two best local optima of 1-EI ($x_1 = 0.3$ and $x_2 = 0.7$). This is not a general fact. Other examples illustrate for instance how 2-EI maximization can yield two points located around (but different from) 1-EI's global optimum whenever 1-EI has one single peak of great magnitude (see fig. (4)).

3.2 q-EI computation by Monte Carlo Simulations

Extrapolating the calculation of the 2-EI to the general case gives a complex expression depending on q-dimensional Gaussian cdf's. Hence, it seems that the direct computation of q-EI when q grows large would have to rely on numerical multivariate integral approximation techniques anyway. Therefore, directly evaluating q-EI by Monte-Carlo Simulation then makes sense. Thanks to Eqs. (13) and (14), the random vector $[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))/\mathbf{Y}]$ can easily be simulated using the Mahalanobis decomposition of Gaussian vectors:

$$\forall k \in [1, n_{sim}],\ M_k = (m_{OK}(\mathbf{x}^{n+1}), ..., m_{OK}(\mathbf{x}^{n+q})) + [S_q^{\frac{1}{2}} N_k]^T, N_k \sim \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q)\ i.i.d. \quad (15)$$

Computing the integral of any function (not necessarily linearly) depending on the vector $[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))/\mathbf{Y}]$ can then be done in averaging the images of the simulated vectors by the considered function:

function Q-$\mathbb{E}$I($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X}^{new}$)
  $L = \text{chol}(Var[Y(\mathbf{X}^{new})/Y(\mathbf{X}) = \mathbf{Y}])$ ▷ Cholesky decomposition of S_q
  for $i \leftarrow 1, n_{sim}$ **do**
    $N \sim \mathcal{N}(0, I_q)$ ▷ Drawing a vector N at random
    $M_i = m_{OK}(\mathbf{X}^{new}) + LN$ ▷ Simulating Y at $\mathbf{X}^{new}$
    $qI_{sim}(i)$=$[\min(\mathbf{Y}) - \min(M_i)]^+$ ▷ Simulating the improvement at $\mathbf{X}^{new}$
  end for
  $qEI_{sim} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} qI_{sim}(i)$ ▷ Empirical Expected Improvement
end function

A straightforward application of the Law of Large Numbers yields indeed

$$qEI_{sim} = \sum_{i=1}^{n_{sim}} \frac{[\min(\mathbf{Y}) - \min(M_i)]^+}{n_{sim}} \xrightarrow[n_{sim} \to +\infty]{} EI(\mathbf{x}^1, ..., \mathbf{x}^q) \text{ a.s.} \tag{16}$$

The Central Limit Theorem can finally be used to control the precision of the Monte Carlo approximation as a function of n_{sim} (see [2] for details concerning the variance estimation):

$$\sqrt{n_{sim}} \left(\frac{qEI_{sim} - EI(\mathbf{x}^1, ..., \mathbf{x}^q)}{\sqrt{Var[I(\mathbf{x}^1, ..., \mathbf{x}^q)]}} \right) \xrightarrow[n_{sim} \to +\infty]{} \mathcal{N}(0, 1) \text{ in law} \tag{17}$$

4 Approximated q-EI maximization

In the last section, we presented a multi-points criterion meant to deliver a design of experiments in one step through the optimization problem

$$(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}' \in D^q}[EI(\mathbf{X}')] \tag{18}$$

However, the computation of q-EI becomes intensive as q increases. Moreover, the optimization problem (18) is of dimension $d \times q$. Here we try to find pseudo-sequential strategies that approach the result of problem (18) while avoiding its numerical cost. Let us first come back to the notations. In the following, we will use the shortcut

$$EI[Y(\mathbf{Z}) = z](\mathbf{x}) = \mathbb{E}[(min(\mathbf{Y}, Y(\mathbf{Z})) - Y(\mathbf{x}))^+ / Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{Z}) = z] \tag{19}$$

where $\mathbf{Z}$ stands for a set of points in D and z is a vector of (true or assumed) images of $\mathbf{Z}$ by y. For instance, expressing q iterations of EGO (without hyperparameter updating) in

this formalism yields

$$\left\{ \begin{array}{c} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}) = argmax_{\mathbf{x}\in D} EI[\,](\mathbf{x}) \\ \forall j \in [1, q-1],\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = y(\mathbf{x}^{n+j}), \\ \qquad\qquad ..., Y(\mathbf{x}^{n+1}) = y(\mathbf{x}^{n+1})](x) \end{array} \right. \tag{20}$$

Note that this formalism holds when the event "$Y(\mathbf{Z}) = z$" is replaced by an event of the form "$Y(\mathbf{Z})$". E.g. if $Y(\mathbf{Z})$ is random, $EI[Y(\mathbf{Z})](\mathbf{x}) = \mathbb{E}[(min(\mathbf{Y}, Y(\mathbf{Z})) - Y(\mathbf{x}))^+ / Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{Z})]$ then becomes a random variable too, depending on the random variable $Y(\mathbf{Z})$. This is the basis of the following strategies.

4.1 A q-points design built with the 1-point expected improvement

Instead of searching for the globally optimal vector $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q})$, an intuitive way of replacing it by a sequential approach is the following: first look for the best single point $\mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$, then feed the model and look for $\mathbf{x}^{n+2} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+1})](x)$, and so on. Of course, the value $y(\mathbf{x}^{n+1})$ is not known at the second step (else we would be in a real sequential algorithm, like EGO). Nevertheless, we dispose of two pieces of information: the site $\mathbf{x}^{n+1}$ has already been visited, and $[Y(\mathbf{x}^{n+1})/\mathbf{Y} = Y(\mathbf{X})]$ is a random variable with known distribution. More precisely, the latter is $[Y(\mathbf{x}^{n+1})/\mathbf{Y} = Y(\mathbf{X})] \sim \mathcal{N}(m_{OK}(\mathbf{x}^{n+1}), s^2_{OK}(\mathbf{x}^{n+1}))$. Hence, the second site $\mathbf{x}^{n+2}$ can be computed as:

$$\mathbf{x}^{n+2} = argmax_{\mathbf{x}\in D} \mathbb{E}\left[EI[Y(\mathbf{x}^{n+1})](\mathbf{x}) / Y(\mathbf{X}) = \mathbf{Y}\right] \tag{21}$$

The same procedure can be applied iteratively to deliver q points, computing $\forall j \in [1, q-1]$:

$$\begin{aligned} &\mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} \mathbb{E}\left[EI[Y(\mathbf{x}^{n+j}), ..., Y(\mathbf{x}^{n+1})](x) / Y(\mathbf{X}) = \mathbf{Y}\right] \\ &= argmax_{\mathbf{x}\in D} \int_{\mathbf{u}\in\mathbb{R}^j} \left[EI[(Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+j-1})) = \mathbf{u}](\mathbf{x})\right] f_{(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+j}))/Y(\mathbf{X})=\mathbf{Y}}(\mathbf{u}) d\mathbf{u} \end{aligned} \tag{22}$$

where $f_{(Y(\mathbf{x}^{n+1}),...,Y(\mathbf{x}^{n+j}))/Y(\mathbf{X})=\mathbf{Y}}(.)$ is the multi-Gaussian density of the joint kriging predictor at $(\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+j})$. Although Eq. (22) is a sequentialized version of the q-points expected improvement maximization, it doesn't completely fulfill our objectives. There is still a multi-Gaussian density to integrate, which seems to be a typical curse in such problems dealing with dependent random vectors. We now present two classes of heuristic strategies meant to circumvent the computational complexity encountered in eq. (22).

4.2 Constant Liar and Kriging Believer strategies

Lying to escape intractable calculations

We propose to weaken the conditional knowledge taken into account at each iteration. This idea inspired two heuristic strategies that we expose and test in the next two subsections: the *Kriging Believer* and the *Constant Liar*.

4.2.1 The "kriging believer" heuristic

The *Kriging Believer* strategy replaces the conditional knowledge about the responses at the sites chosen within the last iterations by deterministic values equal to the expectation of the kriging predictor. Keeping the same notations as previously, the strategy can be summed up as follows:

$$\left\{\begin{array}{c} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}),\ m^n_{OK}(\mathbf{x}^{n+1}) = \mathbb{E}[Y(\mathbf{x}^{n+1})/Y(\mathbf{X}) = \mathbf{Y}] \text{ and } \forall j \in [1, q-1] : \\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = m^{n+j-1}_{OK}(\mathbf{x}^{n+j}), ..., Y(\mathbf{x}^{n+1}) = m^n_{OK}(\mathbf{x}^{n+1})](x) \\ m^{n+j}_{OK}(\mathbf{x}^{n+j+1}) = \mathbb{E}[Y(\mathbf{x}^{n+j+1})/Y(\mathbf{X}) = \mathbf{Y}, Y(\mathbf{x}^{n+j}) = m^{n+j-1}_{OK}(\mathbf{x}^{n+j}), \\ ..., Y(\mathbf{x}^{n+1}) = m^n_{OK}(\mathbf{x}^{n+1})] \end{array}\right. \tag{23}$$

Algorithm 1: The Kriging Believer algorithm: a first approximate solution of the multi-points problem $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}'\in D^q}[EI(\mathbf{X}')]$

1: **function** KB($\mathbf{X}$, $\mathbf{Y}$, q)
2: **for** $i \leftarrow 1, q$ **do**
3: $\mathbf{x}^{n+i} = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$
4: $m_{OK}(\mathbf{x}^{n+i}) = \mathbb{E}[Y(\mathbf{x}^{n+i})/Y(\mathbf{X}) = \mathbf{Y}]$
5: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
6: $\mathbf{Y} = \mathbf{Y} \bigcup \{m_{OK}(\mathbf{x}^{n+i})\}$
7: **end for**
8: **end function**

This sequential strategy delivers a q-points design and is computationally affordable since it relies on the analytically known EI, optimized in d dimensions. However, there is a risk of failure, since believing a kriging surface that overshoots the observed data may lead to a sequence that gets trapped in a non-optimal region for many iterations (see 4.3). We now propose a second strategy that reduces this risk.

4.2.2 The "constant liar" heuristic:

Now consider a sequential strategy in which the model is actualized at each iteration with a value exogenously fixed by the user, and not necessarily connected with the Kriging predictor. The strategy referred to as the *constant liar* consists in lying with the same

value L for every iteration: maximize the expected improvement (find x_{n+1}), actualize the model as if $y(x_{n+1}) = L$, and so on always with the same $L \in R$:

$$\left\{ \begin{array}{c} \mathbf{x}^{n+1} = argmax_{\mathbf{x}\in D} EI(\mathbf{x}) \text{ and } \forall j \in [1, q-1] : \\ \mathbf{x}^{n+j+1} = argmax_{\mathbf{x}\in D} EI[Y(\mathbf{x}^{n+j}) = L, ..., Y(\mathbf{x}^{n+1}) = L](\mathbf{x}) \end{array} \right. \tag{24}$$

Algorithm 2: The Constant Liar algorithm: another approximate solution of the multi-points problem $(\mathbf{x}'^{n+1}, \mathbf{x}'^{n+2}, ..., \mathbf{x}'^{n+q}) = argmax_{\mathbf{X}'\in D^q}[EI(\mathbf{X}')]$

1: **function** CL($\mathbf{X}$, $\mathbf{Y}$, L, q)
2: **for** $i \leftarrow 1, q$ **do**
3: $\mathbf{x}^{n+i} = argmax_{\mathbf{x}\in D} EI(\mathbf{x})$
4: $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
5: $\mathbf{Y} = \mathbf{Y} \bigcup \{L\}$
6: **end for**
7: **end function**

The effect of L on the performance of the resulting optimizer is investigated in the next section. L should logically be determined on the basis of the values taken by y at the initial design. Three values, $min\{\mathbf{Y}\}$, $mean\{\mathbf{Y}\}$, and $max\{\mathbf{Y}\}$ are considered here. The larger L is, the more explorative the algorithm will be, and vice versa.

5 Empirical comparisons

5.1 Application to the Branin-Hoo function

The four optimization strategies presented in the last section are now compared on the the Branin-Hoo function which is a classical test-case in global optimization ([9],[15],[17]).

$$\left\{ \begin{array}{c} y_{BH}(x_1, x_2) = (x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1 - \frac{1}{8\pi})cos(x_1) + 10 \\ x_1 \in [-5, 10],\ x_2 \in [0, 15] \end{array} \right. \tag{25}$$

y_{BH} has three global minimizers $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$, and the global minimum is approximately equal to 0.4. The variables are normalized by the transformation $x_1^{'} = \frac{x_1+5}{15}$ and $x_2^{'} = \frac{x_2}{15}$. The initial design of experiments is a 3×3 complete factorial design $\mathbf{X}_9$ (see fig. (5)), thus $\mathbf{Y} = y_{BH}(\mathbf{X}_9)$. Ordinary Kriging is applied with a stationary, anisotropic, Gaussian covariance function

$$\forall h = (h_1, h_2) \in \mathbb{R}^2,\ C(h_1, h_2) = \sigma^2 e^{-\theta_1 h_1^2 - \theta_2 h_2^2} \tag{26}$$

where the parameters (θ_1, θ_2) are fixed to their Maximum Likelihood Estimate $(5.27, 0.26)$, and σ^2 is estimated within kriging, as an implicit function of (θ_1, θ_2) (like in [9]). We

build a 10-points optimization design with each strategy. We additionally estimated by Monte Carlo simulations ($n_{sim} = 10^4$) the probability of improvement and the expected improvement brought by the q first points of each strategy (here $q \in \{2, 6, 10\}$). The results are gathered in Table 1.

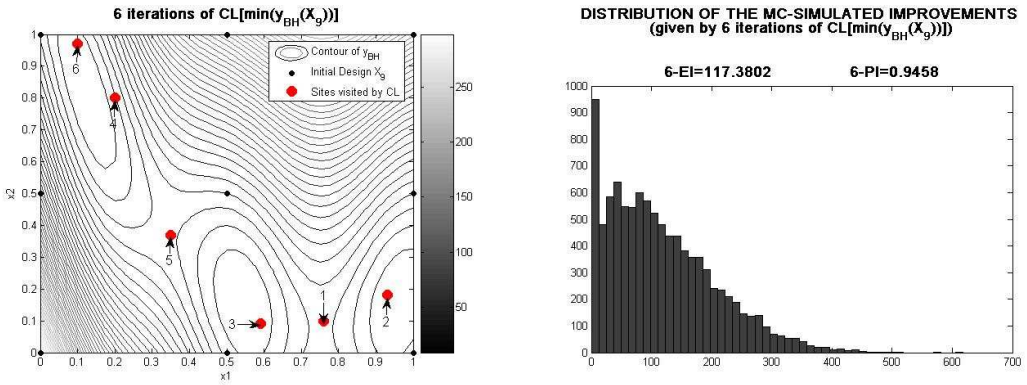


Figure 5: (Left) contour of the Branin-Hoo function with the initial design $\mathbf{X}_9$ (small black points) and the 6 first points given by the heuristic strategy CL[min($f_{BH}(\mathbf{X}_9)$)] (large bullets). (Right) Histogram of 10 000 Monte Carlo simulated values of the improvement brought by the 6-points CL[min($f_{BH}(\mathbf{X}_9)$)] strategy. The corresponding estimations of the 6-points PI and EI are given above.

The four strategies (KB and the three variants of CL) gave clearly different designs and optimization performances. In the first case, *Constant Liar* (CL) sequences behaved as if the already visited points generated a repulsion, with a magnitude increasing with L. The tested values $L = \max(\mathbf{Y})$ and $L = mean(\mathbf{Y})$ forced the exploration designs to fill the space by avoiding $\mathbf{X_9}$. Both strategies provided space-filling, exploratory designs with high probabilities of improvement (10-PI near 100%) and promising q-EI values (see Table 1). *In fine*, they brought respective actual improvements of 7.86 and 6.25.

Of all the tested strategies, CL[min($\mathbf{Y}$)] gave here the best results. In 6 iterations, it visited the three locally optimal zones of y_{BH}. In 10 iterations, it gave the best actual improvement among the considered strategies, which is furthermore in agreement with the 10-points EI values simulated by Monte-Carlo. It seems in fact that the soft repulsion when $L = \min(\mathbf{Y})$ is the right tuning for the optimization of the Branin-Hoo function, with the initial design $\mathbf{X}_9$.

In the second case, the *Kriging Believer* (KB) has yielded here disappointing results. All the points (except one) were clustered around the first visited point $\mathbf{x}^{n+1}$ (the same as in CL, by construction). This can be explained by the exaggeratedly low prediction given by Kriging at this very point: the mean predictor overshoots the data (because of

	CL[$min(\mathbf{Y})$]	CL[$mean(\mathbf{Y})$]	CL[$max(\mathbf{Y})$]	KB
PI (first 2 points)	87.7%	87%	88.9%	65%
EI (first 2 points)	114.3	114	113.5	82.9
PI (first 6 points)	94.6%	95.5%	92.7%	65.5%
EI (first 6 points)	117.4	115.6	115.1	85.2
PI (first 10 points)	99.8%	99.9%	99.9%	66.5%
EI (first 10 points)	122.6	118.4	117	85.86
Improvement (first 6 points)	7.4	6.25	7.86	0
Improvement (first 10 points)	8.37	6.25	7.86	0

Table 1: Multipoints PI, EI, and actual improvements for the 2, 6, and 10 first iterations of the heuristic strategies CL[min($\mathbf{Y}$)], CL[$mean(\mathbf{Y})$], CL[max($\mathbf{Y}$)], and Kriging Believer (here $\min(\mathbf{Y}) = \min(y_{BH}(\mathbf{X}_9))$). $q-PI$ and $q-EI$ are evaluated by Monte-Carlo simulations (Eq. (16), $n_{sim} = 10^4$).

the Gaussian covariance), and the expected improvement becomes abusively large in the neighborhood of $\mathbf{x}^{n+1}$. Then $\mathbf{x}^{n+2}$ is then chosen near $\mathbf{x}^{n+1}$, and so on. The algorithm gets temporarily trapped at the first visited point. KB behaves in the same way as CL would do with a constant L below $\min(\mathbf{Y})$. As can be seen in Table 1 (last column), the phenomenon is visible on both the q-PI and q-EI criteria: they remain almost constant when q increases. This illustrates in particular how q-points criteria can help in rejecting unappropriate strategies.

The results shown in Table 1 highlight a major drawback of the q-points PI criterion. When q increases, the PI associated with all 3 CL strategies quickly converges to 100%, such that it is not possible to discriminate between the good and the very good designs. The q-points EI is a more selective measure thanks to taking the magnitude of possible improvements into account. Nevertheless, the q-EI criterion overevaluates the improvement associated with all designs considered here. This effect (already pointed out in [15]) can be explained by considering both the high value of σ^2 estimated from $\mathbf{Y}$ and the small difference between the minimal value reached at $\mathbf{X}_9$ (9.5) and the actual minimum of y_{BH} (0.4).

We now compare CL[min], CL[max], latin hypercubes (LHS) and uniform random designs (UNIF) in terms of q-EI values, with $q \in [1, 10]$.For every $q \in [1, 10]$, we sampled 2000 q-elements designs of each type (LHS and UNIF) and compared the empirical Expected Improvement distributions to the Expected Improvement estimates associated with the q first points of both CL strategies. As can be seen on fig. (6), CL[max] (light bullets) and CL[min] (dark squares) offer very good q-EI results compared to random designs, especially for small values of q. By definition, the two of them start with the 1-EI global maximizer, which ensures a q-EI at least equal to 83 for all $q \geq 1$. Both associated q-EI series then seem to converge to threshold values, almost reached for $q \geq 2$ by CL[max] (which dominates CL[min] when $q = 2$ and $q = 3$) and for $q \geq 4$ by CL[min] (which dominates CL[max]

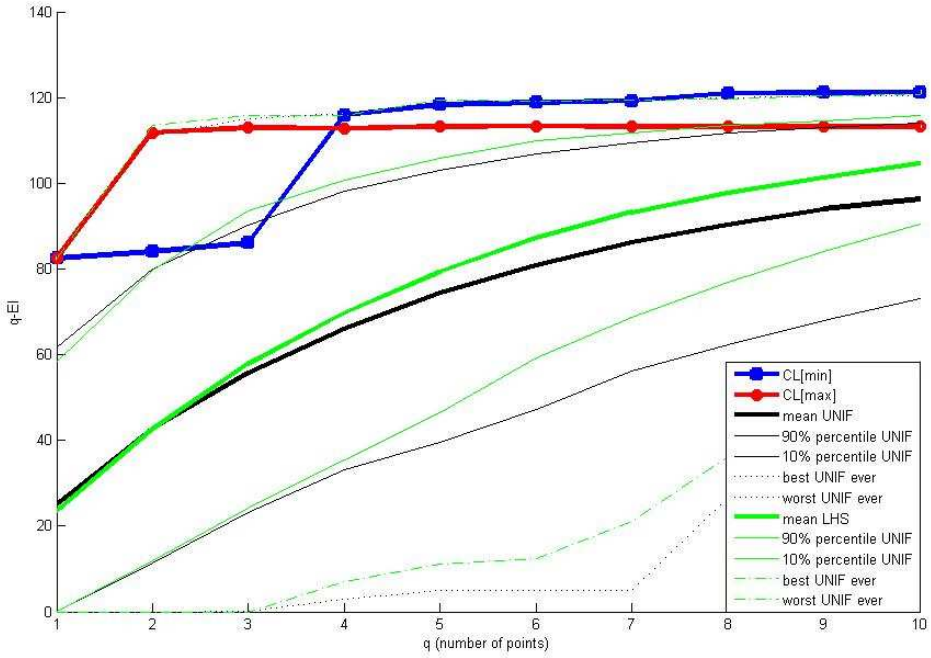


Figure 6: Comparaison of the q-EI associated with the q first points ($q \in [1, 10]$) given by the constant liar strategies (min and max), 2000 q-points designs taken uniformly at random for every q, and 2000 q-points LHS designs taken at random for every q.

for all $4 \leq q \leq 10$). Thr random designs have less promizing q-EI expected values. Their q-EI distributions are quite dispersed, which can be seen for instance by looking at the $10\% - 90\%$ interpercentiles represented on fig. (6) by thin full lines (respectively dark and light for UNIF and LHS designs). Note in particular that the q-EI distribution of the LHS designs seem globally better than the one of the uniform designs. Interestingly, the best designs ever found among the UNIF designs (dark dotted lines) and among the LHS designs (light dotted lines) almost match with CL[max] when $q \in \{2, 3\}$ and CL[min] when $4 \leq q \leq 10$. We haven't yet observed a design sampled at random that clearly provides better q-EI values than the heuristic strategies.

5.2 Kriging-based optimization of gaussian process realizations

With the intent to produce general results, we chose to study and compare the 3 heuristics KB, CL[min $\mathbf{Y}$], and CL[max $\mathbf{Y}$] presented in 2.3 in applying them to random functions.

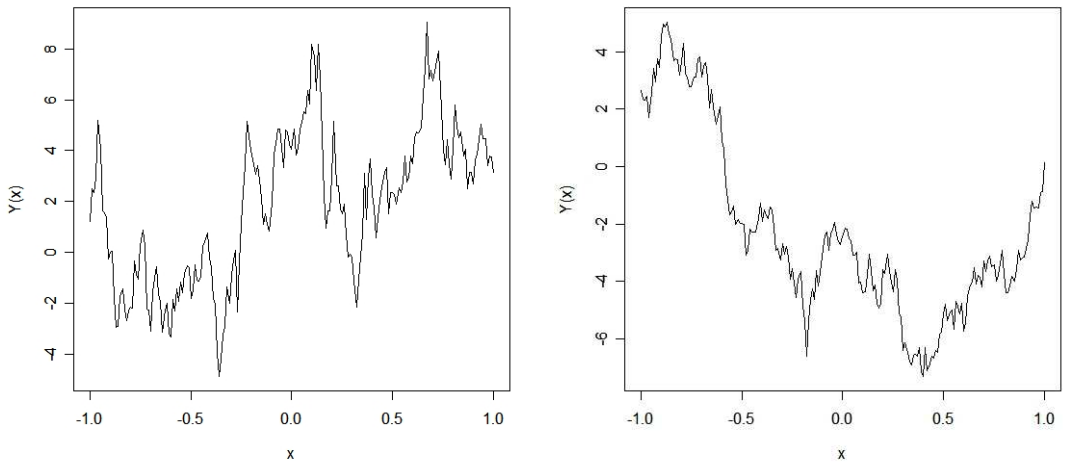


Figure 7: Two stationary Gaussian Process paths (both centered, with variance 10 and exponential covariance structure with respective correlation lengths 0.2 and 0.7). This family of Gaussian Process is often referred to as *Ornstein-Uhlenbeck* Process ([4])

Gaussian Process simulation is a handy way to work with such functions.[4]. We considered four experimental configurations (denoted by $k \in [1,4]$) involving Gaussian Processes $Y^k(x)$ and 1000 realizations $\{y_i^k(x), i \in [1,1000]\}$ of them for each configuration. For all configurations, the outputs varied between -1 and 1 ($D = [-1,1]$), and the initial design of experiments was fixed to the 3-elements set $\mathbf{X} = \{-1,0,1\}$ (see fig. (8)). The other experimental parameters varied accordingly to the values specified in Table (5.2).

k	covariance	correlation length	variance	N_k
1	Exponential	0.3	40	2
2	Exponential	1	40	2
3	Exponential	0.3	40	10
4	Exponential	1	40	10

Table 2: Design of experiments for a comparison between the 3 heuristics

Formally, each heuristic strategy $\mathcal{S}$ (here $\mathcal{S} \in \{KB, CL[min], CL[max]\}$) provides a sequence of points $X^{k,1}(\mathcal{S}), ..., X^{k,N_k}(\mathcal{S})$. These points are random variables since they closely

[4]simulating mono- or multi-dimensional Gaussian processes on a grid (having m elements) is theorically (but not always numerically) straightforward, the cost being the inversion of an $m \times m$ covariance matrix.

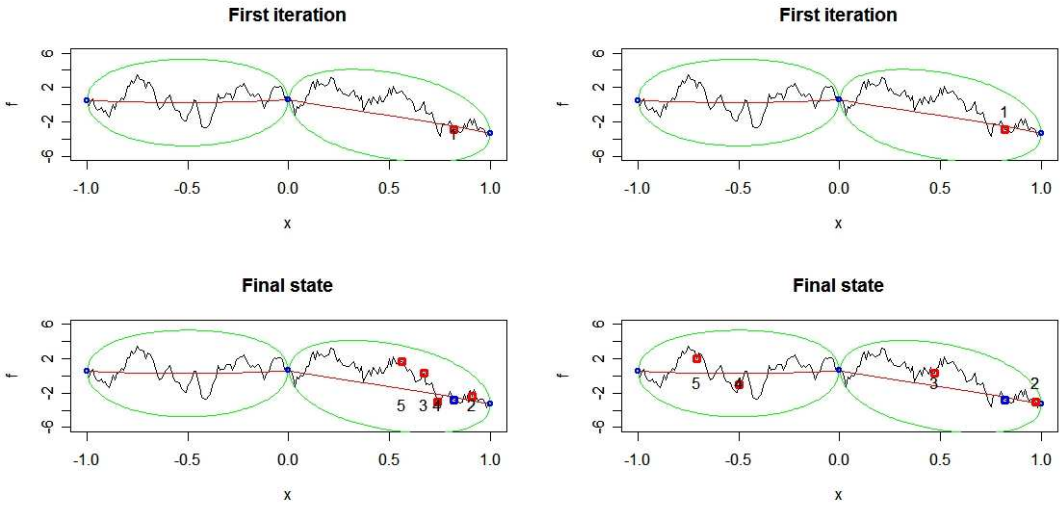


Figure 8: 5 iterations of CL[min(**Y**)] (left) and CL[max(**Y**)] (right) to one Gaussian Process path (scale 0.7, variance 10, exponential covariance). This example clearly illustrates how the CL strategy priveleges local search whenever $L = \min(\mathbf{Y})$, and possesses a more space-filling behaviour when $L = \max(\mathbf{Y})$.

depend on the process Y^k which is itself random. Here we wish to study the performances of each strategy (given a configuration) by looking at the behavior of the random variable

$$\Delta_k(\mathcal{S}) = \min\{Y^k(\mathbf{X}), Y^k(X^{k,1}(\mathcal{S})), ..., Y^k(X^{k,N_k}(\mathcal{S}))\} - \min_{x\in D}[Y^k(x)] \geq 0 \qquad (27)$$

which measures how far we are from having perfectly optimized the process $Y^k(x)$ after having ran N_k iterates of the strategy $\mathcal{S}$. Hence, the closer the realizations of $\Delta_k(\mathcal{S})$ are to 0, the better $\mathcal{S}$ fullfils its goals as optimizer.
We studied the experimental performances of the three algorithms applied to the 1000 realizations ran for every configuration k. We considered the realizations of $\Delta_k(\mathcal{S})$,

$$\delta_k^i(\mathcal{S}) = \min\{y_i^k(\mathbf{X}), y_i^k(x_i^{k,1}(\mathcal{S})), ..., y_i^k(x_i^{k,N_k}(\mathcal{S}))\} - \min_{x\in D}[y_i^k(x)] \geq 0 \qquad (28)$$

where the $x_i^{k,j}(\mathcal{S})$'s stand for the 1000 realizations of the $X^{k,j}(\mathcal{S})$'s. The results are summarized in figures (10) and (11). The histograms offer concentrated representations of the δ_k's distributions, i.e. the statistical performances of each strategy in all studied configurations. Values near 0 (on the extreme left of the histograms) mean succesful optimizations, whereas right tails stand for the cases of failure (best y value observed far beyond the

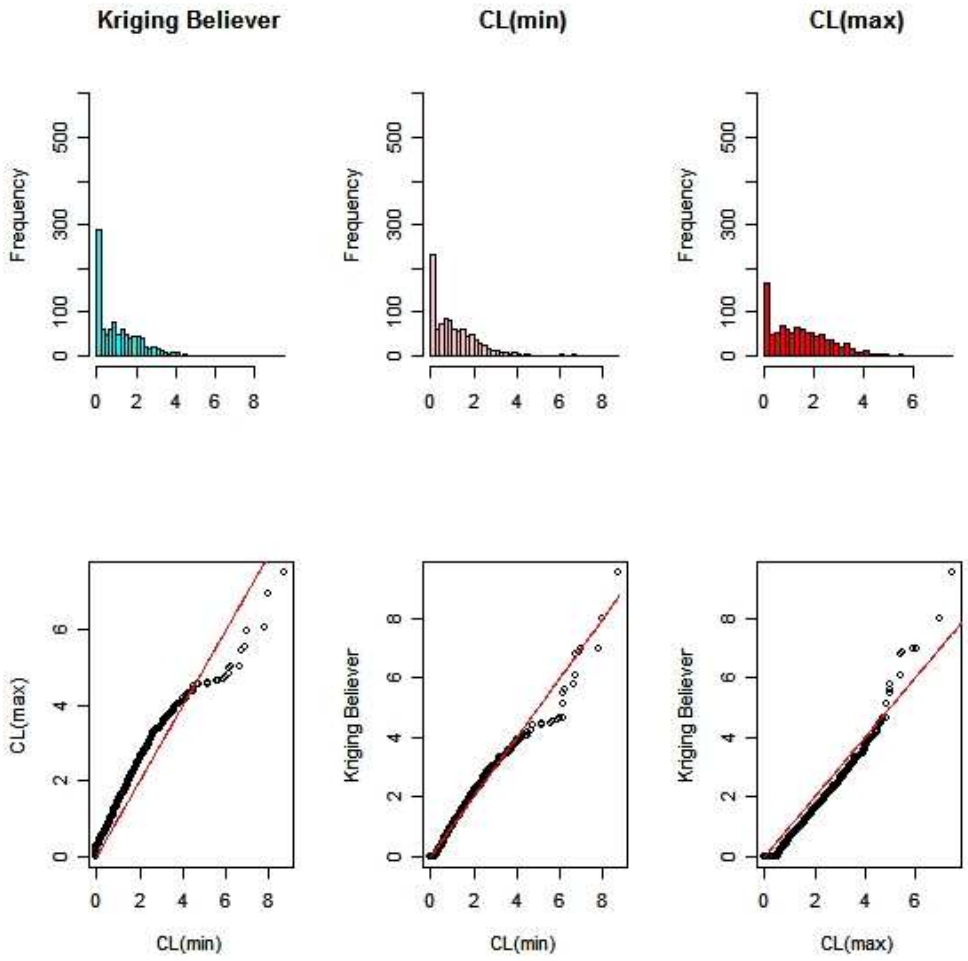


Figure 9: Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configurations 4. $CL[max]$ and KB keep their positions of best performers, respectively at the right and left extremes. Note the particular shape of the qq-plot between $CL[min]$ and $CL[max]$. The first one is statistically more likely to perform very well, but also more likely to fail dramatically. Conversely to configuration 1 (see appendix D), $CL[max]$ is here a good challenger for a risk averse user.

actual minimum). The *q-q plots* aim at comparing all couples of strategy in plotting the empirical quantiles (i.e. ranked values of the δ_k^i's) of the one against the empirical quantiles of the other. Such kind of graphic allows a far more subtile comparison between strategies than only scalar indexes like the mean or the median performance.

As shown on fig. (11), the Kriging Believer strategy does not behave pathologically anymore when using the exponential covariance: it seems in fact to give optimization results with a very good balance between high performances and risk covering. Even if the three strategies roughly give comparable results within this example, $CL[min]$ and KB appear indeed to provide more often extremely good results (small δ's) than $CL[max]$, which however has thiner tails than $CL[min]$. Note that if the comparaison between strategies is quite stable for small values of δ, this statement doesn't hold for high quantiles since the corresponding fluctuations are too large for samples of 1000 process realizations.

6 Conclusions

Gaussian Process regression is very convenient for metamodel-based optimization. Its probabilistic frame allows to build explicit criteria accounting for the exploration/exploitation trade-off, like the *expected improvement*, EI. The q-EI criterion developed here makes it possible to get an evaluation of the "optimization potential" given by a set of q experiments. It can be used to analytically derive EI-optimal singletons and couples. Monte-Carlo simulations offer the opportunity to evaluate the q-EI associated with any given design of experiment, whatever its size. Four heuristic strategies, the "Kriging Believer" and three "Constant Liars" have been proposed and compared that aim at maximizing q-EI while being numerically tractable. It has been verified that they provide higher q-EI's than Latin Hypercubes and random uniform designs of experiments.

Aknowledgements: This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA, and Total S.A. We wish to thank Xavier Bay, Raphael T. Haftka, Ben Smarslok, Yann Richet, Olivier Roustant, and Victor Picheny for their help and rich comments. Special thanks to the R project people ([6]) for developing and spreading such a useful freeware.

A More details on the 2-points Expected Improvement

A.1 Minimum of two random variables

Let us consider two real random variables U and V defined on the same probability space. In the case where U and V are independent, the cumulative distribution of the couple (U, V) is well known:

$$\forall x, y \in \mathbb{R},\ P(U \leq x, V \leq y) = P(U \leq x) \times P(V \leq y) = F_U(x) \times F_V(y)$$

This is the key to the distribution of $m = min(U, V)$ since $\forall x \in \mathbb{R},\ P(m \leq x) = 1 - P(m > x) = 1 - P(U > x, V > x) = 1 - P(U > x) \times P(V > x) = 1 - (1 - P(U \leq x)) \times (1 - P(V \leq x)) = P(U \leq x) + P(V \leq x) - P(U \leq x) \times P(V \leq x)$. For instance, in the case where $N_1, N_2 \sim \mathcal{N}(\mu, \sigma^2)$ independently, the distribution of their minimum is given by:

$$\forall x \in \mathbb{R},\ P(min(N_1, N_2) \leq x) = 2\left[\Phi(\frac{x-\mu}{\sigma})\right] - \left[\Phi(\frac{x-\mu}{\sigma})\right]^2 \tag{29}$$

where Φ is the gaussian cumulative distribution function. Now we consider the case of two dependent random variables U and V. Since the last multiplicativity property doesn't hold anymore, all we can say is that $\forall x \in \mathbb{R},\ P(m \leq x) = P(U \leq x, V \leq x)$ which is the cumulative distribution function of the random vector (U, V) evaluated at (x, x). In the case where (U, V) is a dependent multigaussian vector $(N_1, N_2) \sim \mathcal{N}(\mu, \Sigma)$, we have the more general expression:

$$\forall x \in \mathbb{R},\ P(min(N_1, N_2) \leq x) = P(N_1 \leq x, N_2 \leq x) = CDF(\mu, \Sigma)(x, x) \tag{30}$$

where CDF stands for the bi-gaussian cumulative distribution function. This is the integral of the bi-gaussian density function (corresponding to the distribution $\mathcal{N}(\mu, \Sigma)$) over the surface of the southwestern quadrant delimited by (x, x). This expression is so forth considered as analytically intractable and must be numerically approximated.

A.2 Analytical calculation of the 2-points Expected Improvement

Some classical results of conditional calculus allow us to precise this dependance and fix the notations. Let us first give shortened notations for the means, standard deviations, and covariance of the random variables $(Y_{OK}(\mathbf{x}^i) = [Y(\mathbf{x}^i)/Y(\mathbf{X}) = \mathbf{Y}], i \in \{1, 2\})$:

$$\begin{aligned} &m_i = \mathbb{E}[Y(\mathbf{x}^i)/\mathbf{Y}] = m_{OK}(\mathbf{x}^i), \sigma_i = \sqrt{Var[Y(\mathbf{x}^i)/\mathbf{Y}]} = s_{OK}(\mathbf{x}^i), \\ &c_{1,2} = cov[Y_{OK}(\mathbf{x}^1), Y_{OK}(\mathbf{x}^2)/\mathbf{Y}] = C_{12} = \rho_{1,2}\sigma_1\sigma_2 \end{aligned}$$

Well-kwown results from linear regression (for instance) then give us conditional means and variances of one response knowing the other:

$$m_{2/1} = E[Y(\mathbf{x}^2)/\mathbf{Y}, Y_{OK}(\mathbf{x}^1)] = m_2 + \frac{c_{1,2}}{\sigma_1^2}(Y_{OK}(\mathbf{x}^1) - m_1),\ \sigma_{2/1}^2 = \sigma_2^2 - \frac{c_{1,2}^2}{\sigma_1^2} = \sigma_2^2(1 - \rho_{12}^2) \tag{31}$$

$$m_{1/2} = E[Y(\mathbf{x}^1)/\mathbf{Y}, Y_{OK}(\mathbf{x}^2)]m_1 + \frac{c_{1,2}}{\sigma_2^2}(Y_{OK}(\mathbf{x}^2) - m_2),\ \sigma_{1/2}^2 = \sigma_1^2 - \frac{c_{1,2}^2}{\sigma_2^2} = \sigma_1^2(1 - \rho_{12}^2) \tag{32}$$

At this stage we are in position to compute $EI(\mathbf{x}^1, \mathbf{x}^2)$. Starting here, we replace the complete notation $Y_{OK}(\mathbf{x}^i)$ by Y_i and forget the conditioning on $\mathbf{Y}$ for the sake of clarity.

Phase 1

$$\begin{aligned} EI(\mathbf{x}^1, \mathbf{x}^2) &= E[I(\mathbf{x}^1, \mathbf{x}^2)] = E[max(0, min(\mathbf{Y}) - min(Y_1, Y_2))] \\ &= E[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbb{1}_{min(Y_1,Y_2) \leq min(\mathbf{Y})}] \\ &= E[(min(\mathbf{Y}) - min(Y_1, Y_2))\mathbb{1}_{min(Y_1,Y_2) \leq min(\mathbf{Y})}(\mathbb{1}_{Y_1 \leq Y_2} + \mathbb{1}_{Y_2 \leq Y_1})] \\ &= E[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_1 \leq Y_2}] + E[(min(\mathbf{Y}) - Y_2)\mathbb{1}_{Y_2 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}] \end{aligned}$$

Since both terms of the last sum are similar (up to a permutation between $\mathbf{x}^1$ and $\mathbf{x}^2$), we will at first restrict our attention to the first one. Using $\mathbb{1}_{Y_1 \leq Y_2} = 1 - \mathbb{1}_{Y_2 \leq Y_1}$ [5] , we get:

$$\begin{aligned} E[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_1 \leq Y_2}] &= E[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}(1 - \mathbb{1}_{Y_2 \leq Y_1})] \\ &= EI(\mathbf{x}^1) - E[(min(\mathbf{Y}) - Y_1)\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}] \\ &= EI(\mathbf{x}^1) + B(\mathbf{x}^1, \mathbf{x}^2) \end{aligned}$$

where $B(\mathbf{x}^1, \mathbf{x}^2) = E[(Y_1 - min(\mathbf{Y}))\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}]$. Informally, $B(\mathbf{x}^1, \mathbf{x}^2)$ is the opposite of the improvement brought by Y_1 when $Y_2 \leq Y_1$ and hence that doesn't contribute to the 2-step expected improvement. Our aim in the next phases will be to give an explicit expression for $B(\mathbf{x}^1, \mathbf{x}^2)$.

Phase 2

$$B(\mathbf{x}^1, \mathbf{x}^2) = E[Y_1\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}] - min(\mathbf{Y})E[\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}]$$

At this point, it is worth noticing that $Y_1 = m_1 + \sigma_1 N_1$ with $N_1 \sim \mathcal{N}(0, 1)$. Substituing this decomposition in the last expression of $B(\mathbf{x}^1, \mathbf{x}^2)$ leads to:

$$B(\mathbf{x}^1, \mathbf{x}^2) = \sigma_1 E[N_1\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}] + (m_1 - min(\mathbf{Y}))E[\mathbb{1}_{Y_1 \leq min(\mathbf{Y})}\mathbb{1}_{Y_2 \leq Y_1}]$$

The two terms of this sum require some attention. We compute both of them in detail respectively in phase 3 and phase 4.

Phase 3

Using a classical property of conditional calculus [6], we have that:

$$E[N_1 1_{Y_1 \leq min(\mathbf{Y})} 1_{Y_2 \leq Y_1}] = E[N_1 1_{Y_1 \leq min(\mathbf{Y})} E[1_{Y_2 \leq Y_1}/Y_1]]$$

Using the fact that $Y_2/Y_1 \sim \mathcal{N}(m_{2/1}(Y_1), s^2_{2/1}(Y_1))$, we obtain the following:

$$E[1_{Y_2 \leq Y_1}/Y_1] = \Phi(\frac{Y_1 - m_{2/1}}{s_{2/1}}) = \Phi(\frac{Y_1 - m_2 - \frac{c_{1,2}}{\sigma_1^2}(Y_1 - m_1)}{\sigma_2\sqrt{1 - \rho_{12}^2}})$$

Back to the main term and using again the normal decomposition of Y_1, we get:

$$E[N_1 1_{Y_1 \leq min(\mathbf{Y})} 1_{Y_2 \leq Y_1}] = [N_1 1_{N_1 \leq \frac{min(\mathbf{Y}) - m_1}{\sigma_1}} \Phi(\frac{m_1 - m_2 + (\sigma_1 - \rho_{12}\sigma_2)N_1}{\sigma_2\sqrt{1 - \rho_{12}^2}})] = E[N_1 1_{N_1 \leq \gamma_1}\Phi(\alpha_1 N_1 + \beta_1)]$$

$$\text{where } \gamma_1 = \frac{min(\mathbf{Y}) - m_1}{\sigma_1},\ \beta_1 = \frac{m_1 - m_2}{\sigma_2\sqrt{1 - \rho_{12}^2}} \text{ and } \alpha_1 = \frac{\sigma_1 - \rho_{12}\sigma_2}{\sigma_2\sqrt{1 - \rho_{12}^2}}$$

Finally, $E[N_1 1_{N_1 \leq \gamma_1}\Phi(\alpha_1 N_1 + \beta_1)]$ can be computed applying an integration by parts:

$$\begin{aligned} \int_{-\infty}^{\gamma_1} u\phi(u)\Phi(\alpha_1 u + \beta_1)du &= [-\phi(u)\Phi(\alpha_1 u + \beta_1)]_{-\infty}^{\gamma} + \int_{-\infty}^{\gamma_1} \alpha_1\phi(u)\phi(\alpha_1 u + \beta_1)du \\ &= -\phi(\gamma_1)\Phi(\alpha_1\gamma_1 + \beta_1) + \frac{\alpha_1}{2\pi}\int_{-\infty}^{\gamma_1} e^{\frac{-u^2 - (\alpha_1 u + \beta_1)^2}{2}}du \end{aligned}$$

[5] This expression should be rigorously noted $1 - \mathbb{1}_{Y_2 < Y_1}$. Since we work here with (continous) gaussian random variables, it sufficies however that their correlation is different from 1 for the expression to be exact ($\{Y_1 = Y_2\}$) is then neglectable). We implicitely do this assumption here and in the following.

[6] For all function ϕ in $L^2(\Omega)$, $E[X\phi(Y)] = E[E[X/Y]\phi(Y)]$

Since $u^2 + (\alpha_1 u + \beta_1)^2 = \left(\sqrt{(1+\alpha_1^2)}u + \frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)^2 + \frac{\beta_1^2}{1+\alpha_1^2}$, the last integral reduces to:

$$\sqrt{2\pi}\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)\int_{-\infty}^{\gamma_1} e^{\frac{-\left(\sqrt{(1+\alpha_1^2)}u+\frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)^2}{2}}du = \frac{2\pi\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)}{\sqrt{(1+\alpha_1^2)}}\int_{-\infty}^{\sqrt{(1+\alpha_1^2)}\gamma_1+\frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}}\frac{e^{\frac{-v^2}{2}}}{\sqrt{2\pi}}dv$$

We conclude in using the definition of the cumulative distribution function:

$$E[N_1 1_{Y_1\leq min(\mathbf{Y})}1_{Y_2\leq Y_1}] = -\phi(\gamma_1)\Phi(\alpha_1\gamma_1+\beta_1) + \frac{\alpha_1\phi\left(\sqrt{\frac{\beta_1^2}{1+\alpha_1^2}}\right)}{\sqrt{(1+\alpha_1^2)}}\Phi\left(\sqrt{(1+\alpha_1^2)}\gamma_1 + \frac{\alpha_1\beta_1}{\sqrt{1+\alpha_1^2}}\right)$$

Phase 4

We then finally compute the term:

$$E[1_{Y_1\leq min(\mathbf{Y})}1_{Y_2\leq Y_1}] = E[1_{X\leq min(\mathbf{Y})}1_{Z\leq 0}]$$

where $(X, Z) = (Y_1, Y_2 - Y_1)$ is following a two-dimensional normal distribution of mean $M = (m_1, m_2 - m_1)$, and variance matrix $\Gamma = \begin{pmatrix} \sigma_1^2 & c_{1,2} - \sigma_1^2 \\ c_{1,2} - \sigma_1^2 & \sigma_2^2 + \sigma_1^2 - 2c_{1,2} \end{pmatrix}$. The final results rely on the fact that:

$$E[1_{X\leq min(\mathbf{Y})}1_{Z\leq 0}] = CDF(M,\Gamma)(min(\mathbf{Y}), 0)$$

where CDF stands for the bi-gaussian cumulative distribution function.

Proposition:

$$EI(\mathbf{x}^1, \mathbf{x}^2) = EI(\mathbf{x}^1) + EI(\mathbf{x}^2) + B(\mathbf{x}^1, \mathbf{x}^2) + B(\mathbf{x}^2, \mathbf{x}^1) \tag{33}$$

$$\text{with } B(\mathbf{x}^1, \mathbf{x}^2) = (m_{OK}(\mathbf{x}^1) - min(\mathbf{Y}))\delta(\mathbf{x}^1, \mathbf{x}^2) + \sigma_{OK}(\mathbf{x}^1)\epsilon(\mathbf{x}^1, \mathbf{x}^2)$$

$$\epsilon(\mathbf{x}^1, \mathbf{x}^2) = \alpha_1\phi(\frac{|\beta_1|}{\sqrt{(1+\alpha_1^2)}})\Phi(\frac{\gamma + \frac{\alpha_1\beta_1}{1+\alpha_1^2}}{(1+\alpha_1^2)^{-\frac{1}{2}}}) - \phi(\gamma)\Phi(\alpha_1\gamma+\beta_1),\ \delta(\mathbf{x}^1, \mathbf{x}^2) = CDF(\Gamma)\begin{pmatrix} min(\mathbf{Y}) - m_1 \\ m_1 - m_2 \end{pmatrix}$$

B Generalities about the q-points expected improvement

B.1 An alternative definition

After the definition of the bivariate expected improvement, it seems natural to define the multivariate expected improvement as:

$$EI(\mathbf{x}^1, ..., \mathbf{x}^q) = E[max(min(\mathbf{Y}) - min\{Y_{OK}(\mathbf{x}^1), ..., Y_{OK}(\mathbf{x}^q)\}, 0)]$$

Shortening again the notations, we have the equivalent definition:

$$EI(\mathbf{x}^1, ..., \mathbf{x}^q) = \sum_{i=1}^{q} E[(min(\mathbf{Y}) - Y_i)\mathbb{1}_{Y_i\leq min(\mathbf{Y})}(\Pi_{j\neq i}\mathbb{1}_{Y_i\leq Y_j})] \tag{34}$$

Proof of 34: like in phase 1, we use the property $1 = \sum_{i=1}^{q}(\Pi_{j\neq i}\mathbb{1}_{Y_i\leq Y_j})$ which only means that the smallest Y_i is among the Y_is!

B.2 First bounds on q-EI

$$EI(\mathbf{x}^1, ..., \mathbf{x}^q) \leq \sum_{i=1}^{q} EI(\mathbf{x}^i) \tag{35}$$

Proof of 35: $\forall i \in [1, q]$, $(\Pi_{j \neq i} \mathbb{1}_{Y_i \leq Y_j}) \leq 1$ and hence $(min(\mathbf{Y}) - Y_i)\mathbb{1}_{Y_i \leq min(\mathbf{Y})}(\Pi_{j \neq i} \mathbb{1}_{Y_i \leq Y_j}) \leq (min(\mathbf{Y}) - Y_i)\mathbb{1}_{Y_i \leq min(\mathbf{Y})}$. The property follows from 34.

$$EI(\mathbf{x}^1, ..., \mathbf{x}^q) \geq \max_{J \subsetneq [1,q]} EI(\{\mathbf{x}^i, i \in J\}) \geq \max_{1 \leq i \leq q} EI(\mathbf{x}^i) \tag{36}$$

Proof of 36: Be $J \subsetneq [1, n]$. Both statements directly come from the inequality: $min(Y_i, i \in J) \geq min(Y_i, i \in [1, n])$.

$$\forall \sigma \in \Sigma_n,\ EI(\mathbf{x}^{\sigma(1)}, ..., \mathbf{x}^{\sigma(q)}) = EI(\mathbf{x}^1, ..., \mathbf{x}^n) \tag{37}$$

Proof of 37: This follows the invariance of *min* by permutation.

C Joint predictions using Simple and Ordinary Kriging

Here we give some details about the calculation of the joint distribution obtained when simultaneously predicting at different points in the cases of Simple and Ordinary Kriging (SK and OK in the following). Let us first recall some basics about Kriging and Gaussian Processes.

C.1 Gaussian Processes for Machine Learning

A real (L^2) random process $(Y(\mathbf{x}))_{\mathbf{x} \in D}$ is defined as a *Gaussian Process* (GP) whenever all its finite-dimensional distributions are Gaussian. Consequently, for all $n \in \mathbb{N}$ and for all set $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ of n points of D, there exists a vector $\mathbf{m_X} \in \mathbf{R}^n$ and a symmetric positive semi-definite matrix $\Sigma_{\mathbf{X}} \in \mathcal{M}_n(\mathbb{R})$ such that $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n))$ is a Gaussian Vector, following a multigaussian probability distribution $\mathcal{N}(\mathbf{m_X}, \Sigma_{\mathbf{X}})$. More specifically, for all $i \in [1, n]$, $Y(\mathbf{x}^i) \sim \mathcal{N}(\mathbb{E}[Y(\mathbf{x}^i)], Var[Y(\mathbf{x}^i)])$ where $\mathbb{E}[Y(\mathbf{x}^i)]$ is the ith coordinate of $\mathbf{m_X}$ and $Var[Y(\mathbf{x}^i)]$ is the ith diagonal term of $\Sigma_{\mathbf{X}}$. Furthermore, all couples $(Y(\mathbf{x}^i), Y(\mathbf{x}^j))$ $i, j \in [1, n], i \neq j$ are multigaussian with a covariance $Cov[Y(\mathbf{x}^i), Y(\mathbf{x}^j)]$ equal to the non-diagonal term of $\Sigma_{\mathbf{X}}$ indexed by i and j.

A Random Process Y is said to be *first order stationary* if its mean is a constant, i.e. if $\forall \mathbf{x} \in D$, $\mathbb{E}[Y(\mathbf{x})] = \mu$ where $\mu \in \mathbb{R}$. Y is said to be *second order stationary* if there exists a positive semidefinite function $c : D - D \longrightarrow \mathbb{R}$ such that for all pairs $(\mathbf{x}, \mathbf{x}') \in D^2$, $Cov[Y(\mathbf{x}), Y(\mathbf{x}')] = c(\mathbf{x} - \mathbf{x}')$. We then have the following expression for the covariance matrix of the observations at $\mathbf{X}$:

$$\Sigma_{\mathbf{X}} = (Cov[Y(\mathbf{x}_i), Y(\mathbf{x}_j)])_{i,j \in [1,n]} = (c(\mathbf{x}_i - \mathbf{x}_j))_{i,j \in [1,n]} = \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1 - \mathbf{x}_2) & ... & c(\mathbf{x}_1 - \mathbf{x}_n) \\ c(\mathbf{x}_2 - \mathbf{x}_1) & \sigma^2 & ... & c(\mathbf{x}_2 - \mathbf{x}_n) \\ ... & ... & ... & ... \\ c(\mathbf{x}_n - \mathbf{x}_1) & c(\mathbf{x}_n - \mathbf{x}_2) & ... & \sigma^2 \end{pmatrix} \tag{38}$$

where $\sigma^2 := c(0)$. If Y is first and second order stationary, it is said *weakly stationary*. A major feature of Gaussian Processes is that their *weak stationarity* is equivalent to *strong stationarity*: if Y is a weakly stationary GP, the law of probability of the random variable $Y(\mathbf{x})$ doesn't depend on $\mathbf{x}$, and the joint distribution of $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n))$ is the same as the distribution of $(Y(\mathbf{x}^1 + \mathbf{h}), ..., Y(\mathbf{x}^n + \mathbf{h})$ whatever the set of points $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \in D^n$ and the vector $\mathbf{h} \in \mathbb{R}^n$ such that $\{\mathbf{x}^1 + \mathbf{h}, ..., \mathbf{x}^n + \mathbf{h}\} \in D^n$. To sum up, a stationary GP is entirely defined by its mean μ and its covariance function $c(.)$. The classical framework of Kriging for Computer Experiments is to make predictions of a costly simulator y at a new set of sites $\mathbf{X}_{new} = \{\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+q}\}$ (most of the time, $q = 1$), on the basis of the collected observations at the initial

design $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$, and under the assumption that y is one realization of a stationary GP Y with known covariance function (in theory) Simple Kriging (SK) assumes a known mean, $\mu \in \mathbb{R}$. In Ordinary Kriging (OK), μ is estimated.

C.2 Conditioning Gaussian Vectors

Let us consider a centered Gaussian vector $V = (V_1, V_2)$ with covariance matrix

$$\Sigma_V = \mathbb{E}[VV^T] = \begin{pmatrix} \Sigma_{V_1} & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{V_2} \end{pmatrix} \tag{39}$$

Key properties of Gaussian vectors include that the orthogonal projection of a Gaussian vector is still a Gaussian vector, and that the orthogonality of two subvectors V_1, V_2 of a Gaussian vector V (i.e. $\Sigma_{cross} = \mathbb{E}[V_2 V_1^T] = 0$) is equivalent to their independance. We now express the conditional expectation $\mathbb{E}[V_1/V_2]$. $\mathbb{E}[V_1/V_2]$ is by definition such that $V_1 - \mathbb{E}[V_1/V_2]$ is independent of V_2. $\mathbb{E}[V_1/V_2]$ is thus fully characterized as orthogonal projection on the vector space spanned by V_2, solving the equation:

$$\mathbb{E}[(V_1/ - \mathbb{E}[V_1/V_2])V_2^T] = 0 \tag{40}$$

Assuming linearity of $\mathbb{E}[V_1/V_2]$ in V_2, i.e. $\mathbb{E}[V_1/V_2] = AV_2$ ($A \in \mathcal{M}_n(\mathbb{R})$), a straightforward development of (eq.40) gives the matrix equation $\Sigma_{cross}^T = A\Sigma_{V_2}$, and hence $\Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2$ is a suitable solution provided Σ_{V_2} is full ranked[7]. We conclude that

$$\mathbb{E}[V_1/V_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2 \tag{41}$$

by uniqueness of the orthogonal projection in a Hilbert space. Using the independence between $(V_1 - \mathbb{E}[V_1/V_2])$ and V_2, we calculate the conditional covariance matrix Σ_{V_1/V_2}:

$$\begin{aligned} \Sigma_{V_1/V_2} &= \mathbb{E}[(V_1 - \mathbb{E}[V_1/V_2])(V_1 - \mathbb{E}[V_1/V_2])^T/V_2] \\ &= \mathbb{E}[(V_1 - AV_2)(V_1 - AV_2)^T] \\ &= \Sigma_{V_1} - A\Sigma_{cross} - \Sigma_{cross}^T A^T + A\Sigma_{V_2} A^T \\ &= \Sigma_{V_1} - \Sigma_{cross}^T \Sigma_{V_2}^{-1} \Sigma_{cross} \end{aligned} \tag{42}$$

Now consider the case of a non-centered random vector $V = (V_1, V_2)$ with mean $m = (m_1, m_2)$. The conditional distribution V_1/V_2 can be obtained by coming back to the centered random vector $V - m$. We then find that $\mathbb{E}[V_1 - m_1/V_2 - m_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$ and hence $\mathbb{E}[V_1/V_2] = m_1 + \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$.

C.3 Simple Kriging Equations

Let us come back to our metamodeling problem and assume that y is one realization of a Gaussian Process Y, defined as follows:

$$\begin{cases} Y(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}) \\ \varepsilon(\mathbf{x}) \text{ centered stationary GP with covariance function } c(.) \end{cases} \tag{43}$$

where $\mu \in \mathbb{R}$ is a known scalar. Now say that Y has already been observed at n locations $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ ($Y(\mathbf{X} = \mathbf{Y})$) and that we wish to predict Y a q new locations $\mathbf{X}_{new} = \{\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+q}\}$.
Since $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n), Y(\mathbf{x}^{n+1}), ..., Y(\mathbf{x}^{n+q}))$ is a Gaussian Vector with mean $\mu \mathbb{1}_{n+q}$ and covariance matrix

$$\Sigma_{tot} = \begin{pmatrix} \Sigma & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{new} \end{pmatrix} = \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1 - \mathbf{x}_2) & ... & c(\mathbf{x}_1 - \mathbf{x}_{n+q}) \\ c(\mathbf{x}_2 - \mathbf{x}_1) & \sigma^2 & ... & c(\mathbf{x}_2 - \mathbf{x}_{n+q}) \\ ... & ... & ... & ... \\ c(\mathbf{x}_{n+q} - \mathbf{x}_1) & c(\mathbf{x}_{n+q} - \mathbf{x}_2) & ... & \sigma^2 \end{pmatrix} \tag{44}$$

[7] If Σ_{V_2} is not invertible, the equation holds in replacing $\Sigma_{V_2}^{-1}$ by the pseudo-inverse $\Sigma_{V_2}^{\dagger}$.

We can directly apply eq. (41) and eq. (42) to derive the Simple Kriging Equations:

$$[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}(m_{SK}(\mathbf{X}_{new}), \Sigma_{SK}(\mathbf{X}_{new})) \tag{45}$$

with $m_{SK}(\mathbf{X}_{new}) = \mathbb{E}[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}] = \mu\mathbb{1}_q + \Sigma_{cross}^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbb{1}_q)$ and $\Sigma_{SK}(\mathbf{X}_{new}) = \Sigma_{new} - \Sigma_{cross}^T\Sigma^{-1}\Sigma_{cross}$. When $q = 1$, $\Sigma_{cross} = \mathbf{c}(\mathbf{x}^{n+1}) = Cov[Y(\mathbf{x}^{n+1}), Y(\mathbf{X})]$ and the covariance matrix reduces to $s_{SK}^2(\mathbf{x}) = \sigma^2 - \mathbf{c}(\mathbf{x}^{n+1})^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+1})$, which is called the *Kriging Variance*.
When μ is constant but not known in advance, it is not mathematically correct to sequentially estimate μ and plug in the estimate in the Simple Kriging equations. Ordinary Kriging addresses this issue.

C.4 Ordinary Kriging Equations

Compared to Simple Kriging, Ordinary Kriging (OK) is used when the mean of the underlying random process is constant and unknown. We give here a derivation of OK in a Bayesian framework, assuming that μ has an improper uniform prior distribution $\mu \sim \mathcal{U}(\mathbb{R})$. y is thus seen as a realization of a random process Y, defined as the sum of μ and a centered GP [8]:

$$\begin{cases} Y(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}) \\ \varepsilon(\mathbf{x}) \text{ centered stationary GP with covariance function } c(.) \\ \mu \sim \mathcal{U}(\mathbb{R}) \text{ (prior)} \end{cases} \tag{46}$$

Note that conditioning with respect to μ actually provides SK equations. Letting μ vary, we aim to find the law of $[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}]$. Starting with $[Y(\mathbf{X}) = \mathbf{Y}/\mu] \sim \mathcal{N}(\mu\mathbb{1}_n, \Sigma)$, we get μ's posterior distribution:

$$[\mu/Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}\left(\hat{\mu}, \sigma_\mu^2\right) = \mathcal{N}\left(\frac{\mathbb{1}^T\Sigma^{-1}\mathbf{Y}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}, \frac{1}{\mathbb{1}_q^T\Sigma^{-1}\mathbb{1}_q}\right) \text{ (posterior)} \tag{47}$$

We can re-write the SK equations $[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}, \mu] \sim \mathcal{N}(m_{SK}(\mathbf{X}_{new}), \Sigma_{SK}(\mathbf{X}_{new}))$. Now it is very useful to notice that the conditional random vector $[(Y(\mathbf{X}_{new}), \mu)/Y(\mathbf{X}) = \mathbf{Y}]$ is Gaussian [9]. It follows that $[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}]$ is Gaussian, and its mean and covariance matrix can finally be calculated with the help of classical conditional calculus results. Hence using $m_{OK}(\mathbf{X}_{new}) = \mathbb{E}[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}] = \mathbb{E}_\mu\left[\mathbb{E}[Y(\mathbf{X}_{new})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right]$, we find that $m_{OK}(\mathbf{X}_{new}) = \hat{\mu} + \Sigma_{cross}^T\Sigma^{-1}(\mathbf{Y} - \hat{\mu}\mathbb{1}_n)$. Similarly, $\Sigma_{OK}(\mathbf{X}_{new})$ can be obtained using that $Cov[A, B] = Cov[\mathbb{E}[A/C], \mathbb{E}[B/C]] + \mathbb{E}[Cov[A, B/C]]$ for all random variables A,B, C such that all terms exist. We get for all couples of points $(\mathbf{x}^{n+i}, \mathbf{x}^{n+j})$ $(i, j \in [1, q])$:

$$\begin{aligned} &Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}] \\ =&\mathbb{E}\left[Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] + Cov\left[\mathbb{E}[Y(\mathbf{x}^{n+i})/Y(\mathbf{X}) = \mathbf{Y}, \mu], \mathbb{E}[Y(\mathbf{x}^{n+j})/, Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] \end{aligned} \tag{48}$$

The left term $Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/, Y(\mathbf{X}) = \mathbf{Y}, \mu]$ is the conditional covariance under the Simple Kriging Model. The right term is the covariance between $\mu + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbb{1}_q)$ and $\mu + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}(\mathbf{Y} - \mu\mathbb{1}_q)$ conditionally to the observations $Y(\mathbf{X}) = \mathbf{Y}$. Using eq. (47), we finally obtain:

$$\begin{aligned} &Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}] \\ =&\mathbb{E}\left[Cov[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}, \mu]\right] \\ &+Cov[\mathbf{E}[Y(\mathbf{x}^{n+i})/Y(\mathbf{X}) = \mathbf{Y}, \mu], \mathbf{E}[Y(\mathbf{x}^{n+j})/, Y(\mathbf{X}) = \mathbf{Y}, \mu]] \\ =&Cov_{SK}[Y(\mathbf{x}^{n+i}), Y(\mathbf{x}^{n+j})/Y(\mathbf{X}) = \mathbf{Y}] \\ &+Cov[\mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}(\mathbf{Y}) + \mu(1 + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbb{1}_q), \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}(\mathbf{Y}) + \mu(1 + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}\mathbb{1}_q)] \\ =&c(\mathbf{x}^{n+i} - \mathbf{x}^{n+j}) - \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbf{c}(\mathbf{x}^{n+j}) + \frac{(1 + \mathbf{c}(\mathbf{x}^{n+i})^T\Sigma^{-1}\mathbb{1}_q)(1 + \mathbf{c}(\mathbf{x}^{n+j})^T\Sigma^{-1}\mathbb{1}_q)}{\mathbb{1}_q^T\Sigma^{-1}\mathbb{1}_q} \end{aligned} \tag{49}$$

[8] The resulting random process Y is not Gaussian

[9] which can be proved by considering its Fourier transform

And the Ordinary Kriging Variance now appears as a particular case. For all $\mathbf{x} \in D$, we have indeed:

$$s^2_{OK}(\mathbf{x}) = Var[Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] = \sigma^2 - \mathbf{c}(\mathbf{x})^T \Sigma^{-1} \mathbf{c}(\mathbf{x}) + \frac{(1 - \mathbb{1}_n^T \Sigma^{-1} \mathbf{c}(\mathbf{x}))^2}{\mathbb{1}_n^T \Sigma^{-1} \mathbb{1}_n} \tag{50}$$

D More graphics to compare the KB and CL strategies

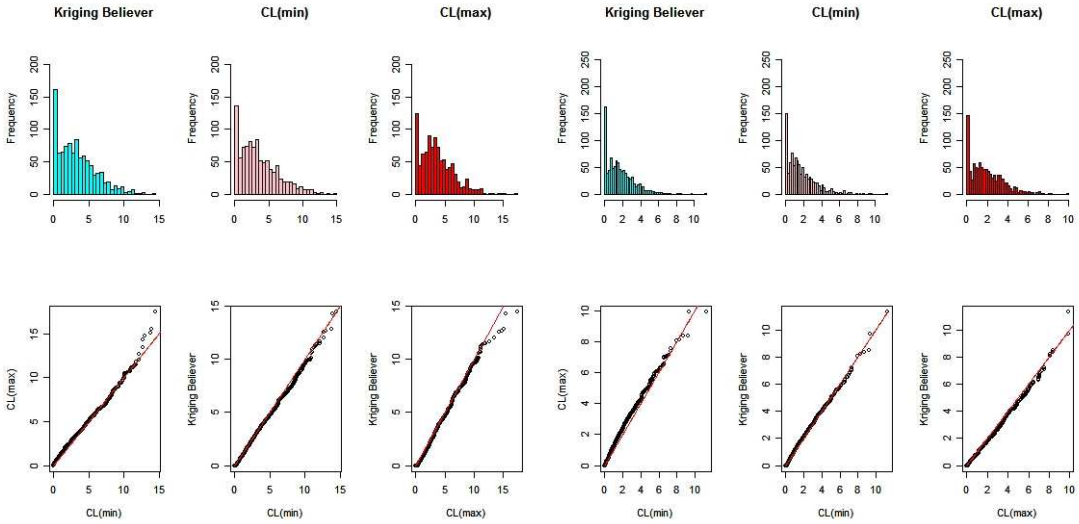


Figure 10: Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configurations 1 (left) and 3 (right).

This figure focuses on the δ_k^i's associated with 2 iterations of the three strategies applied to Gaussian processes with exponential covariance and respective scales 0.3 and 1 (see 5.2). In the first case ($k = 1$), the KB and $CL[min]$ show very similar results. $CL[max]$ has a slightly different right tail, and both first and third qq-plots illustrate how it is dominated in terms of extreme risk. This effect doesn't hold when the scale is 1 (right side of fig.(10)). All the strategies then behave almost equally. Note that the performances are uniformly better than with the previous configuration. This is because the realizations are more regular, and are as such easier to optimize starting with only 3 points.
We now look at (11), where 10 iterations are considered with the same sets of covariance parameters as previously (see 5.2). This time, clearer dissimilarities appear between the strategies. In configuration 2 (scale 0.3), $CL[max]$ shows better *right-tail* performances than both other strategies, which almost match. Note the dominance of KB in terms of extreme performance (near 0). These effect are amplified in configuration 4 where $CL[max]$ and KB keep their positions of best performers, respectively at the right and left extremes. Note the particular shape of the *qq*-plot between $CL[min]$ and $CL[max]$. The first one is statistically more likely to perform very well, but also more likely to fail dramatically. Conversely to configuration 1, $CL[max]$ is here a good challenger for a risk averse user.

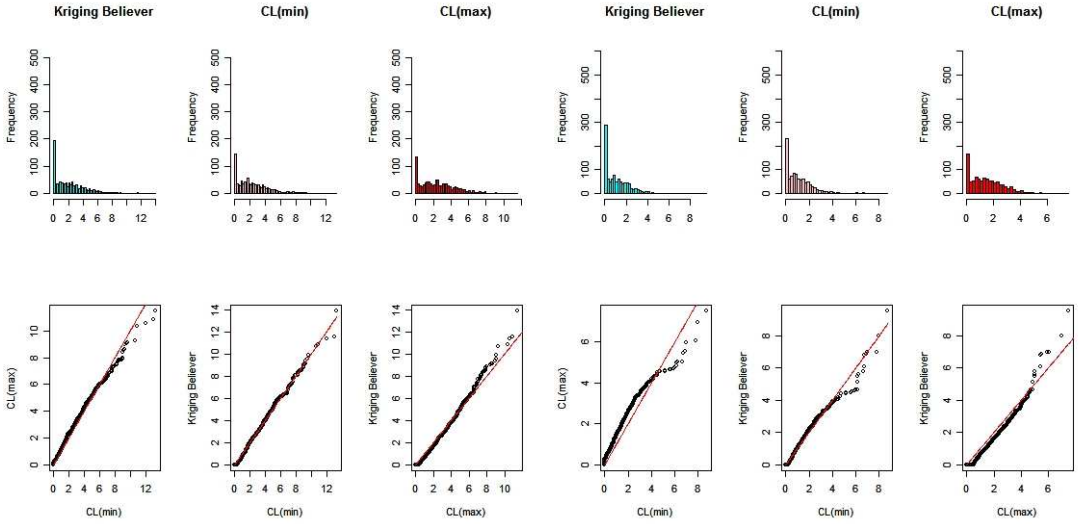


Figure 11: Comparison of the heuristic strategies $CL[min]$, $CL[max]$, KB applied to 1000 Gaussian process realizations with configurations 2 (left) and 4 (right).

References

[1] Journel A. Fundamentals of geostatistics in five lessons. Technical report, Stanford Center for Reservoir Forecasting, 1988.

[2] Ripley B.D. *Stochastic Simulation*. John Wiley and Sons, New York, 1987.

[3] Baker C.A., Watson L. T., Grossman B., Mason W. H., and Haftka R. T. Parallel global aircraft configuration design space exploration. *Practical parallel computing*, pages 79–96, 2001.

[4] Rasmussen C.E. and Williams K.I. *Gaussian Processes for Machine Learning*. M.I.T. Press, 2006.

[5] Geman D. and Jedynak B. An active testing model for tracking roads in satellite images. Technical report, Institut National de Recherches en Informatique et Automatique (INRIA), December 1995.

[6] R development Core Team. R: A language and environment for statistical computing, 2006.

[7] Jones D.R. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, (21):345–383, 2001.

[8] Jones D.R., Pertunen C.D., and Stuckman B.E. Lipshitzian optimization without the lipshitz constant. *Journal of Optimization Theory and Application*, (79), October 1993.

[9] Jones D.R., Schonlau M., and Welch W.J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[10] Sarah Goria. *Evaluation d'un projet minier: approche bayésienne et options réelles.* PhD thesis, Ecole des Mines de Paris, 2004.

[11] D. Huang, T.T. Allen, W. Notz, and N. Zheng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, to appear.

[12] Villemonteix J., Vazquez E., and Walter E. An informational approach to the global optimization of expensive-to-evaluate functions. *Elsevier science direct*, 2006.

[13] Koehler J.R. and Owen A.B. Computer experiments. Technical report, Department of Statistics, Stanford University, 1996.

[14] Joshua Knowles. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE transactions on evolutionnary computation*, 2005.

[15] Schonlau M. *Computer Experiments and Global Optimization.* PhD thesis, University of Waterloo, 1997.

[16] Cressie N.A.C. *Statistics for spatial data.* Wiley series in probability and mathematical statistics, 1993.

[17] Queipo N.V., Verde A., Pintos S., and Haftka R.T. Assessing the value of another cycle in surrogate-based optimization. In *11th Multidisciplinary Analysis and Optimization Conference.* AIAA, 2006.

[18] Santner T.J., Williams B.J., and Notz W.J. *The Design and Analysis of Computer Experiments.* Springer, 2003.

11.4 Symétries

1. Présentations en conférences internationales avec comité de lecture

   - Congrès joint de la Société de Statistique Canadienne et de la Société Française de Statistique
     - Lieu et date : Ottawa (Canada), du 25 au 29 Mai 2008.
     - Proceedings : non.
   - 8^{th} ENBIS plenary conference
     - Lieu et date : Athènes (Grèce), du 21 au 24 Septembre 2008.
     - Proceedings : non.

2. Publication de revue scientifique

   - Journal : *Comptes Rendus de l'Académie des Sciences* (section Maths).
   - Statut : en révision mineure.

Noyaux de covariance pour le krigeage de fonctions symétriques

David Ginsbourger [a], Xavier Bay [a] Laurent Carraro [a]

[a] *Ecole Nationale Supérieure des Mines, 158 cours Fauriel, 42023 Saint-Etienne, France*

Reçu le *****; accepté après révision le +++++

Présenté par

Résumé

L'apprentissage statistique d'une fonction déterministe par un processus gaussien nécessite de sélectionner un noyau de covariance. Lorsque l'on dispose a priori d'informations sur les symétries de la fonction que l'on souhaite approximer, il est fort dommageable de ne pas les utiliser à l'étape du choix du noyau. Nous proposons une caractérisation des noyaux de covariance dont les processus gaussiens associés possèdent des réalisations invariantes par l'action d'un groupe fini de transformations. Nous donnons ensuite un exemple de tels processus symétriques, construits sur la base de processus gaussiens stationnaires, et jouissant d'intéressantes propriétés de régularité. *Pour citer cet article : A. Nom1, A. Nom2, C. R. Acad. Sci. Paris, Ser. ? (2008).*

Abstract

Covariance kernels for the kriging of symmetrical functions. Learning a deterministic function using a gaussian process relies on the selection of a covariance kernel. When some prior information is available concerning symmetries of the function to be approximated, it is clearly unreasonable not to use it in the choice of the kernel. We propose a characterization of the kernels which associated gaussian processes have all paths invariant under the action of a finite group of transformations. We then give an example of such symmetrical processes, built on the basis of stationary gaussian processes, and having interesting regularity properties. *To cite this article: A. Nom1, A. Nom2, C. R. Acad. Sci. Paris, Ser. ? (2008).*

L'interpolation optimale par processus gaussiens -souvent appelée krigeage en sciences de la terre- permet de modéliser une fonction numérique déterministe multivariable $y : D \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ $(d \in \mathbb{N}^*)$ sur la base d'observations faites en un plan d'expériences $\mathbf{X} \in D^k$ $(k \in \mathbb{N}^*)$. Le choix du noyau de covariance permet de prendre en compte la connaissance a priori que l'on a de la fonction y à approximer, en particulier en ce qui concerne sa régularité et ses propriétés spectrales. Le lien entre l'invariance du noyau de covariance par différentes actions de groupe et les propriétés des réalisations du processus gaussien correspondant ont fait l'objet de nombreuses études (la stationnarité à l'ordre 2 découle par exemple de l'invariance par translation du noyau de covariance). Nous nous intéressons ici aux processus

Email addresses: ginsbourger@emse.fr (David Ginsbourger), bay@emse.fr (Xavier Bay), carraro@emse.fr (Laurent Carraro).

aléatoires centrés dont les réalisations sont sûrement invariantes par l'action d'un groupe fini donné de transformations de $\mathbb{R}^d$, et plus particulièrement aux propriétés que doit satisfaire un noyau défini positif pour engendrer un tel processus. Dans ce qui suit, nous considérons $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $(\Omega, \mathcal{A}, P)$ un espace probabilisé, $\{(Y_x)_{x\in\mathbb{R}},\ \forall x \in \mathbb{R}\ Y_x \in L^2(P)\}$ un processus aléatoire réel centré de noyau de covariance $k_Y : D \times D \longrightarrow \mathbb{R}$, et G un groupe fini d'ordre $n \in \mathbb{N}^*$ agissant sur E via l'action $\Phi : (g, x) \in G \times E \longrightarrow \Phi(g, x) := g.x \in E$. Nous supposerons de plus que D est stable par Φ.

Définition. On dit que Y a toutes ses réalisations invariantes sous l'action Φ lorsque

$$\forall \omega \in \Omega,\ \forall x \in D,\ \forall g \in G,\ Y_x(\omega) = Y_{g.x}(\omega) \tag{1}$$

Propriété. Si Y *a toutes ses réalisations invariantes sous l'action* Φ*, alors* k_Y *satisfait*

$$\forall x, x' \in E,\ k_Y(x, x') = \frac{1}{n^2} \sum_{(g,g')\in G^2} k_Y(g.x, g'.x') \tag{2}$$

Démonstration. Par définition du noyau k_Y et par invariance de Y_x sous l'action Φ de G sur E, on a que $\sum_{(g,g')\in G^2} k_Y(g.x, g'.x') = \sum_{(g,g')\in G^2} Cov[Y_{g.x}, Y_{g'.x'}] = \sum_{(g,g')\in G^2} Cov[Y_x, Y_{x'}] = n^2 k_Y(x, x')$.

Remarque 1 Cette décomposition de k_Y *en double somme indexée par les éléments de* G *est loin d'être unique. Soit en effet* $\mathcal{R}$ *la relation d'équivalence sur* D *qui lie deux éléments appartenant à une même orbite de* ϕ *et* $\pi : D \longrightarrow D/\mathcal{R}$ *la surjection canonique. Soit* $A \subset D$ *un système de représentants des orbites de* Φ *; Y s'écrit alors comme le symétrisé d'un processus Z :*

$$Y_x = \sum_{g\in G} Y_x \frac{\mathbb{1}_{g.A}(x)}{\#Stab_\Phi(x)} = \sum_{g\in G} Y_{g^{-1}.x} \frac{\mathbb{1}_A(g^{-1}.x)}{\#Stab_\Phi(g^{-1}.x)} = \sum_{g\in G} Z_{g.x} \tag{3}$$

où $Stab_\Phi(x) := \{g \in G :\ g.x = x\}$ *est le stabilisateur de* x *par* Φ *(sous-groupe de* G*, d'ordre* $\#Stab_\Phi(x)$*) et* $Z_x := Y_x \frac{\mathbb{1}_A(x)}{\#Stab_\Phi(x)}$*. Il vient en particulier en notant* k_Z *le noyau de covariance de* Z*, que*

$$\forall x, x' \in E,\ k_Y(x, x') = Cov\left[\sum_{g\in G}^{m} Z_{g.x}, \sum_{g'\in G}^{m} Z_{g'.x'}\right] = \sum_{(g,g')\in G^2} k_Z(g.x, g'.x') \tag{4}$$

Corollaire. Y *a toutes ses réalisations invariantes sous l'action* Φ *seulement s'il existe un noyau défini postitif* $\bar{k}_Y : (D/\mathcal{R})^2 \longrightarrow \mathbb{R}$ *tel que* k_Y *s'écrive :*

$$\forall x, x' \in D,\ k_Y(x, x') = \bar{k}_Y(\pi(x), \pi(x')) \tag{5}$$

Démonstration. En notant $R_A : D/\mathcal{R} \longrightarrow A$ la bijection entre les orbites de Φ et le système de représentants A, on introduit $\bar{k}_Y : (o_1, o_2) \in (D/\mathcal{R})^2 \longrightarrow k_Y(R_A(o_1), R_A(o_2))$. Il est clair d'une part que $\bar{k}_Y$ est défini positif sur $(D/\mathcal{R})^2$, puisque toute matrice $\left(\bar{k}_Y(o_i, o_j)\right)_{1\le i,j\le k}$ associée à un vecteur $(o_1, ..., o_k) \in (D/\mathcal{R})^k$ est par construction égale à une matrice définie positive $(k_Y(R_A(o_i), R_A(o_j)))_{1\le i,j\le k}$. D'autre part, il existe $g, g' \in G$ tels que $R_A(\pi(x)) = g.x$ et $R_A(\pi(x')) = g'.x$. Il vient finalement que $k_Y(x, x') = k_Y(g.x, g'.x') = k_Y(R_A(\pi(x)), R_A(\pi(x'))) = \bar{k}_Y(\pi(x), \pi(x'))$.

Exemple 1 Soit X *le processus Gaussien réel centré de covariance* $k_X : x, x' \in \mathbb{R} \longrightarrow k_X(x, x') = e^{-|x-x'|} \in \mathbb{R}$ *(Ornstein-Uhlenbeck), et* $\Phi : (g, x) \in (\mathbb{Z}/2\mathbb{Z}) \times \mathbb{R} \longrightarrow \mathbb{R}$ *l'action qui a* $(\bar{1}, x)$ *associe* $-x$*. Le processus* Y *obtenu par symétrisation de la restriction de* X *à* $A = [0, +\infty[$*, défini par* $Y_x = \frac{1}{1+\mathbb{1}_{\{0\}}(x)} X_x \mathbb{1}_{[0,+\infty[}(x) + \frac{1}{1+\mathbb{1}_{\{0\}}(x)} X_x \mathbb{1}_{[0,+\infty[}(-x)$*, a toutes ses réalisations invariantes sous l'action* Φ*. Son noyau de covariance est donné par* $\forall x, x' \in \mathbb{R},\ k_Y(x, x') = e^{-\left||x|-|x'|\right|}$*. Remarquons que le processus Y, symétrisé du processus stationnaire X, n'est évidemment pas stationnaire.*

Exemple 2 Soit X le processus Gaussien réel centré de covariance $k_X : x, x' \in \mathbb{R}^2 \longrightarrow k_X(x,x') = e^{-||x-x'||^2} \in \mathbb{R}$, et $\Phi : (g,x) \in (\mathbb{Z}/2\mathbb{Z}) \times \mathbb{R} \longrightarrow \mathbb{R}$ l'action qui a $(\bar{1},x)$ associe $s(x) = (x_2,x_1)$, son symétrique par rapport à la première bissectrice. Le processus Y obtenu par symétrisation de la restriction de X à $A = \{x \in \mathbb{R} : x_1 \leq x_2\}$ est défini par $Y_x = \frac{1}{1+\mathbb{1}_{\{x\in\mathbb{R}^2 : s(x)=x\}}(x)} X_x \mathbb{1}_A(x) + \frac{1}{1+\mathbb{1}_{\{x\in\mathbb{R}^2 : s(x)=x\}}(x)} X_x \mathbb{1}_A(s(x))$. Remarquons en particulier que Y conserve la régularité de X en dehors de l'ensemble $\{x \in \mathbb{R}^2 : s(x) = x\}$.

Supposons maintenant qu'un processus gaussien réel centré possède un noyau de covariance s'écrivant, à l'instar de k_Y dans l'eq. (4), comme la somme sur $G \times G$ d'un noyau défini positif quelconque. Il est alors possible de remonter aux propriétés d'invariance par Φ des réalisations du processus considéré.

Théorème. *Soit Y un processus gaussien réel centré dont le noyau de covariance est de la forme $k_Y(x,x') = \sum_{(g,g')\in G^2} k_Z(g.x, g'.x')$, où $k_Z : E \times E \longrightarrow \mathbb{R}$ est un noyau défini positif. Y est alors équivalent à un processus Z^Φ ("version" de Y, Cf. [?]) de réalisations invariantes par Φ. Il existe de plus une modification de Y dont toutes les réalisations sont invariantes par Φ.*

Démonstration. Soit Z un processus gaussien réel centré de noyau de covariance k_Z, et Z^Φ le processus défini par $\forall x \in D,\ Z^\Phi_x = \sum_{g\in G} Z_{g.x}$. Il est clair que Z^Φ aussi un processus gaussien centré, et que ses réalisations sont toutes invariantes par l'action de Φ car $\forall h \in G, \forall x \in E,\ Z^\Phi_{g.x} = \sum_{g\in G} Z_{h.(g.x)} = \sum_{g\in G} Z_{(hg).x} = \sum_{k\in G} Z_{k.x}$ puisque $\forall h \in G,\ hG = G$ par propriété de groupe. On a par ailleurs que $\forall x, x' \in E,\ Cov[Z^\Phi_x, Z^\Phi_{x'}] = Cov[\sum_{g\in G} Z_{g.x}, \sum_{g'\in G} Z_{g'.x}] = \sum_{(g,g')\in G^2} k_Z(g.x, g'.x') = k_Y(x,x')$. Comme Y et Z^Φ sont deux processus gaussiens réels centrés de même noyau d'autocovariance, ils sont de même loi, et on a donc en particulier que $P(Y_{x_1} \in B_1, ..., Y_{x_m} \in B_m) = P(Z^\Phi_{x_1} \in B_1, ..., Z^\Phi_{x_m} \in B_m)$ quelque soit $m \in \mathbb{N}$ et $\{B_i \in \mathcal{B}(\mathbb{R}), i \in [\![1,m]\!]\}$. En reprenant les notations de la démonstration précédente, introduisons maintenant le processus $\tilde{Y}$ défini comme suit : $\forall x \in D,\ \tilde{Y}_x = Y_{R_A(\pi(x))}$. Par construction, $\tilde{Y}$ a toutes ses réalisations invariantes par Φ. Enfin, pour tout $x \in D$, il existe $g \in G$ tel que $R_A(\pi(x)) = g.x$, et il vient que $Var[Y_x - \tilde{Y}_x] = Var[Y_x - Y_{g.x}] = k_Y(x,x) + k_Y(g.x, g.x) - 2k_Y(x, g.x) = 0$. On a ainsi que $\forall x \in D,\ P(Y_x = \tilde{Y}_x) = 1$ et $\tilde{Y}$ est bien une modification de Y.

Remarque 2 En pratique, les réalisations de Y sont souvent simulées sur une partie finie $S = \{x_1, ..., x_m\}$ d'éléments de E en se basant sur une décomposition (Cholesky, Mahalanobis) de la matrice de covariance $K = (k_Y(x_i, x_j))_{1\leq i,j\leq m}$. L' invariance par Φ des vecteurs ainsi obtenus est donc certaine (i.e. $\forall \omega \in \Omega$).

L'exemple suivant illustre le fait qu'un processus Gaussien Y et une modification de Y dont toutes les réalisations sont invariantes par Φ (le cas échéant) ne sont pas nécessairement indistinguables.

Exemple 3 Soit $\Omega =]0,1[,\ \mathcal{A} = \mathcal{B}(]0,1[),\ P$ la mesure de Lebesgue sur Ω , $E = \mathbb{R},\ \mathcal{E} = \mathcal{B}(\mathbb{R}),\ F : x \in \mathbb{R} \longrightarrow \int_{-\infty}^{x} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} du \in]0,1[,\ \varepsilon : \omega \in \Omega \longrightarrow \varepsilon(\omega) = F^{-1}(\omega) \in \mathbb{R},\ G = \{e, s_0\}$ (s_0 symétrie par rapport à 0) et $Y : (x,\omega) \in E \times \Omega \longrightarrow Y_x(\omega) = |x|\varepsilon(\omega)\mathbb{1}_{x\neq\varepsilon(\omega)}$. Le processus défini par $\tilde{Y}_x(\omega) = |x|\varepsilon(\omega)$ a clairement toutes ses réalisations invariantes par la symétrie s_0, et on a bien que $\tilde{Y}$ est une modification de Y puisque pour tout $x \in D$, $P(Y_x = \tilde{Y}_x) = P(\varepsilon \neq 1) = 1$. En revanche, $\left\{\omega \in \Omega \ /\ (\forall x \in D,\ Y_x(\omega) = \tilde{Y}_x(\omega))\right\} = \left\{\frac{1}{2}\right\}$ est négligeable, et les deux processus ne sont donc pas indistinguables.

Pour finir, l'exemple suivant illustre la possibilité de construire une classe particulière de processus invariants en symétrisant des processus stationnaires.

Exemple 4 Reprenons les notations de l'exemple 2. On peut construire un processus de réalisations invariantes par Φ sur la base du processus X (lui-même stationnaire, i.e. invariant par l'action des translations de $D = E = \mathbb{R}^2$) en posant $\forall x \in D,\ X_x^\Phi = \frac{1}{2}(X_x + X_{s(x)}) = \frac{1}{2}(X_{(x_1,x_2)} + X_{(x_2,x_1)})$. Le noyau de covariance du nouveau processus X^Φ est donné par

$$\begin{aligned} k_{X^\Phi}(x,x') &= \frac{1}{4}[k_X(x-x') + k_X(s(x)-x') + k_X(x-s(x')) + k_X(s(x)-s(x'))] \\ &= \frac{1}{4}\left[e^{-||(x_1-x'_1,x_2-x'_2)||^2} + e^{-||(x_1-x'_1,x'_2-x_2)||^2} + e^{-||(x'_1-x_1,x_2-x'_2)||^2} + e^{-||(x'_1-x_1,x'_2-x_2)||^2}\right] \end{aligned} \tag{6}$$

Remarquons que X^Φ conserve la régularité de X, y compris sur l'ensemble $\{x \in \mathbb{R}^2 : s(x) = x\}$.

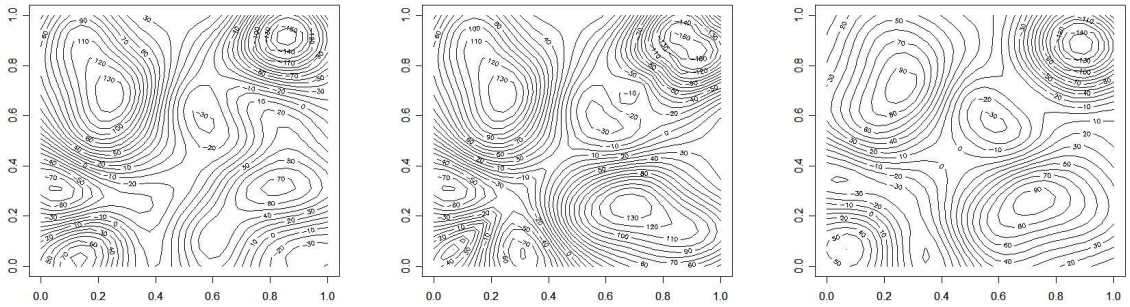


FIG. 1. A gauche : une réalisation simulée sur une grille 30×30 d'un processus gaussien réel centré X de covariance $k_X(x,x') = \sigma^2 e^{-\theta_1(x_1-x'_1)^2-\theta_2(x_2-x'_2)^2}$ (ici $\sigma^2 = 10,\ \theta_1 = \theta_2 = 10$). Au centre : la symétrisée de la restriction à $\{x \in [0,1]^2 : x_1 \leq x_2\}$ de la réalisation de gauche (Exemple 2). A droite : une réalisation simulée (en utilisant le même aléa que précédemment, i.e. en transformant la même réalisation d'un vecteur gaussien centré réduit) d'un processus gaussien stationnaire symétrisé Y (Exemple 4), de covariance construite à partir de k_X selon le procédé de l'eq. (6).

Remerciements : Nous souhaitons remercier Anestis Antoniadis pour une question décisive sur la non nécessaire symétrie des réalisations d'un processus égal en loi à un processus sûrement symétrique, et Yann Richet (IRSN) pour avoir soulevé le problème dans son contexte applicatif.

Chapitre 12

Quelques propriétés des vecteurs gaussiens

12.1 Introduction aux Vecteurs gaussiens

12.1.1 Préambule : la loi $\mathcal{N}(\mathbf{0}, I_d)$

Définition

On dit que $\mathbf{X} = (X_1, ..., X_d)^T \sim \mathcal{N}(\mathbf{0}, I_d)$ lorsque les X_j ($j \in [1, d]$) suivent indépendamment des lois gaussiennes centrées réduites $\mathcal{N}(0, 1)$.

Propriété (densité)

$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_d)$ admet une densité $f_{\mathbf{X}}$ définie par

$$\forall \mathbf{x} \in \mathbb{R}^d \ f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\sum_{j=1}^{d} x_j^2} = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}} \tag{12.1}$$

Preuve : $f_{\mathbf{X}}(\mathbf{x}) = f_{(X_1,...,X_d)}(x_1, ..., x_d) = \Pi_{j=1}^{d} f_{X_j}(x_j)$ par indépendance des X_j. Il suffit alors d'injecter les densités univariées $f_{X_j}(x_j) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x_j^2}$ (les X_j étant gaussiennes centrées réduites par définition).

Propriété (fonction caractéristique)

$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_d)$ admet pour fonction caractéristique $\Phi_{\mathbf{X}}$, définie par

$$\forall \mathbf{u} \in \mathbb{R}^d \ \Phi_{\mathbf{X}}(\mathbf{u}) = e^{-\sum_{j=1}^{d} \frac{1}{2}u_j^2} \tag{12.2}$$

Preuve : $\Phi_{\mathbf{X}}(\mathbf{u}) = \mathbb{E}[e^{i<u,\mathbf{X}>}] = \mathbb{E}[e^{i\sum_{j=1}^{d} u_j X_j}] = \Pi_{j=1}^{d}\mathbb{E}[e^{iu_j X_j}]$ par indépendance des X_j. On utilise alors le résultat déja connu $\mathbb{E}[e^{iu_j X_j}] = e^{-\frac{1}{2}u_j^2}$ (fonction caractéristique d'une v.a.r. de loi $\mathcal{N}(0,1)$).

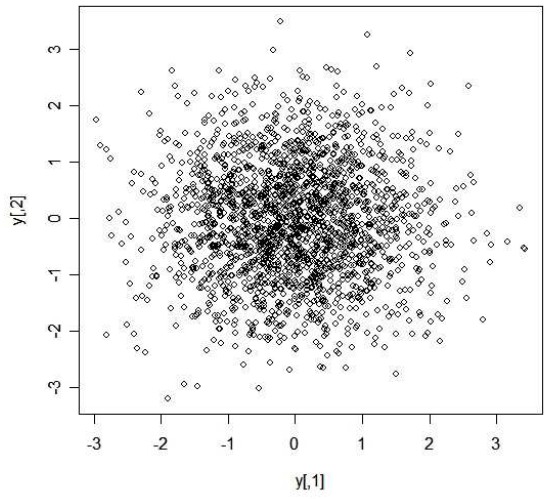


FIG. 12.1 – Un échantillon de 2000 points tirés selon la loi bigaussienne $\mathcal{N}(0, I_2)$.

12.1.2 Trois définitions des vecteurs gaussiens — cas général.

Définition 1 (par projections univariées)

$\mathbf{X}$ est un vecteur gaussien d-dimensionnel si toute combinaison linéaire de ses composantes $< a, \mathbf{X} >_{\mathbb{R}^d} = a^T\mathbf{X} = \sum_{j=1}^{d} a_j X_j$ ($a \in \mathbb{R}^d$) suit une loi gaussienne.

Définition 2 (par fonctions caractéristiques)

$\mathbf{X}$ est un vecteur gaussien d-dimensionnel si sa fonction caractéristique $\Phi_{\mathbf{X}}$ s'écrit sous la forme

$$\forall \mathbf{u} \in \mathbb{R}^d,\ \Phi_{\mathbf{X}}(\mathbf{u}) = e^{i<\mathbf{m},\mathbf{u}>_{\mathbb{R}^d} - \frac{1}{2}<\mathbf{u},S\mathbf{u}>_{\mathbb{R}^d}} \tag{12.3}$$

avec $\mathbf{m} \in \mathbf{R}^d$ et $S \in \mathcal{M}_d(\mathbb{R})$ une matrice symétrique positive.

Définition 3 (par transformation affine)

$\mathbf{X}$ est un vecteur gaussien d-dimensionnel s'il existe un vecteur $\mathbf{m} \in \mathbf{R}^d$ et une matrice $A \in \mathcal{M}_d(\mathbb{R})$ tels que $X = \mathbf{m} + AN$ (*égalité en loi*) où $N \sim \mathcal{N}(\mathbf{0}, I_d)$.

Preuve de l'équivalence des trois définitions

(1) $\Rightarrow$ (2)

Considérons $\mathbf{X}$ satisfaisant la définition (1) et notons μ sa moyenne et Σ sa matrice de covariance. Calculons sa fonction caractéristique. Soit $\mathbf{u} \in \mathbb{R}^d$. On a par définition $\Phi_X(\mathbf{u}) = \mathbb{E}[e^{i<\mathbf{u},\mathbf{X}>_{\mathbb{R}^d}}]$. Par ailleurs, (1) permet d'affirmer que $Y =< \mathbf{u}, \mathbf{X} >_{\mathbb{R}^d}$ est gaussienne. On a alors (après un court calcul de $\mathbb{E}[Y]$ et $Var[Y]$) $Y \sim \mathcal{N}(< \mathbf{u}, \mu >_{\mathbb{R}^d}, < \mathbf{u}, \Sigma\mathbf{u} >_{\mathbb{R}^d})$. Il suffit finalement de remarquer que $\Phi_X(\mathbf{u}) = \Phi_Y(1) = e^{i\mathbb{E}[Y]-\frac{1}{2}(Var[Y])^2} = e^{i<\mathbf{u},\mu>_{\mathbb{R}^d}-\frac{1}{2}<\mathbf{u},\Sigma\mathbf{u}>_{\mathbb{R}^d}}$.

(2) $\Rightarrow$ (3)

Analyse : Soit $A \in \mathcal{M}(\mathbb{R}^d)$. On note $\mathbf{Z} = \mathbf{m}\mathbb{1}_\Omega + AN$. $\forall \mathbf{u} \in \mathbb{R}^d$, $\Phi_{\mathbf{Z}}(\mathbf{u}) = \mathbb{E}[e^{i<\mathbf{u},\mathbf{Z}>}] = e^{i<\mathbf{m},\mathbf{u}>}\mathbb{E}[e^{i<\mathbf{u},AN>}] = e^{i<\mathbf{m},\mathbf{u}>}\mathbb{E}[e^{i<A^T\mathbf{u},N>}] = e^{i<\mathbf{m},\mathbf{u}>}\Phi_N(A^T\mathbf{u})$. Il reste à utiliser le fait que $\Phi_N(A^T\mathbf{u}) = e^{-\frac{1}{2}<A^T\mathbf{u},A^T\mathbf{u}>} = e^{-\frac{1}{2}<\mathbf{u},AA^T\mathbf{u}>}$ pour voir que $\Phi_{\mathbf{Z}}(\mathbf{u}) = \Phi_{\mathcal{N}(\mathbf{m},AA^T)}(\mathbf{u})$.
Synthèse : Soit $\mathbf{X}$ satisfaisant (2).Comme S est symétrique positive, il existe $A \in \mathcal{M}_d(\mathbb{R})$ telle que $S = AA^T$ (une telle décomposition n'est pas nécessairement unique). Soit $N_1 \sim \mathcal{N}(\mathbf{0}, I_d)$. D'après notre analyse, $\mathbf{m} + \mathbf{A}N_1$ a $\Phi_{\mathbf{X}}$ comme fonction caractéristique. Il suit que $\mathbf{m} + \mathbf{A}N_1$ et $\mathbf{X}$ sont égaux en loi.

(3) $\Rightarrow$ (1)

Soit $\mathbf{a} \in \mathbb{R}^d$ et $\mathbf{X} = \mathbf{m} + \mathbf{A}N$ où $N \sim \mathcal{N}(\mathbf{0}_{1\times d}, I_d)$. On a directement que $< \mathbf{a}, \mathbf{m} + \mathbf{A}N >_{\mathbb{R}^d} = (\mathbf{a}^T\mathbf{m}) + \mathbf{a}^T\mathbf{A}N \sim \mathcal{N}(\mathbf{a}^T\mathbf{m}, \mathbf{a}^T\mathbf{A}\mathbf{A}^T\mathbf{a})$. Toute combinaison linéaire des composantes de $\mathbf{X}$ est donc bien gaussienne.

12.1.3 Propriétés élémentaires

Les composantes d'un vecteur gaussien sont gaussiennes...

Soit $\mathbf{X} = (X_1, ..., X_d)^T$ un vecteur gaussien d-dimensionnel. On a par la définition (1) que $\sum_{j=1}^d a_j X_j$ est gaussien pour tout d-uplet de réels $(a_1, ..., a_d)$. Il suffit alors de considérer les $d-$uplets de la forme $(0, ..., 0, a_k = 1, 0, ..., 0)$ $(k \in [1, d])$.

...mais un vecteur de composantes gaussiennes peut ne pas être gaussien

Contre-exemple : Soit $N_1 \sim \mathcal{N}(0,1)$ et N_2 une v.a.r. définie comme suit :

$$\begin{cases} N_2 = -N_1 \text{ si } N_1 \in [-a,a] \ (a \in]0,+\infty[\text{ fixé}) \\ N_2 = N_1 \text{ sinon} \end{cases}$$

Montrons que N_2 est une variable gaussienne. Soit $x \in \mathbb{R}$ quelconque.

$$\begin{aligned} F_{N_2}(x) = \mathbb{P}(N_2 < x) &= \mathbb{P}\left((N_2 < x) \cap (\{N_2 \in [-a,a]\} \cup \{N_2 \notin [-a,a]\})\right) \\ &= \mathbb{P}\left((N_2 < x) \cap \{N_2 \in [-a,a]\}\right) + \mathbb{P}\left((N_2 < x) \cap \{N_2 \notin [-a,a]\}\right) \\ &= \mathbb{P}\left((N_1 < x) \cap \{N_1 \in [-a,a]\}\right) + \mathbb{P}\left((-N_1 < x) \cap \{N_1 \notin [-a,a]\}\right) \\ &= \mathbb{P}\left((N_1 < x) \cap \{N_1 \in [-a,a]\}\right) + \mathbb{P}\left((N_1 < x) \cap \{N_1 \notin [-a,a]\}\right) \\ &= \mathbb{P}\left(N_1 < x\right) \end{aligned}$$

En revanche, le couple (N_1, N_2) **n'est pas** un vecteur gaussien. En effet, la somme $S = N_1 + N_2$ (qui est une combinaison linéaire de N_1 et N_2 avec $(a_1, a_2) = (1,1)$) est telle que $0 < \mathbb{P}(N_1 + N_2 = 0) < 1$. S ne peut ainsi ni être une variable continue ni une constante. Elle n'est donc *a fortiori* pas gaussienne.

La somme de 2 vecteurs gaussiens indépendants est un vecteur gaussien

Cela découle directement de la définition (1).

Dépendance $\iff$ corrélation

Il est clair que tout V.a.r. $\mathbf{X} = (X_1, ..., X_d)$ de composantes indépendantes a une matrice de covariance diagonale. On a en effet déjà vu que les termes $Cov[X_i, X_j]$ $(i \neq j)$ sont nuls par application de l'eq. (**??**). Montrons que la réciproque est vraie dans le cas d'un vecteur gaussien.

Soit $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Gamma)$ où $\Gamma = (\gamma_{ij})_{i,j \in [1,d]}$ est une matrice diagonale (positive). On a en vertu de la définition (2) la fonction caractéristique de $\mathbf{X}$:

$$\begin{aligned} \Phi_{\mathbf{X}}(\mathbf{u}) = \mathbb{E}[e^{i<\mathbf{u},\mathbf{X}>}] &= e^{i<\mathbf{u},\mathbf{m}> - \frac{1}{2}<\mathbf{u},\Gamma\mathbf{u}>} \\ &= e^{i\sum_{j=1}^d u_j m_j - \frac{1}{2}\sum_{j=1}^d \gamma_j u_j^2} = \Pi_{j=1}^d e^{i m_j u_j - \frac{1}{2}\gamma_j u_j^2} \\ &= \Pi_{j=1}^d \Phi_{X_j}(u_j) \end{aligned}$$

d'où l'on conclut que les X_j ($j \in [1, d]$) sont indépendants.

Remarque : la transformation de Mahalanobis $\mathbf{Y} = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \mathbf{m})$ d'un vecteur gaussien $\mathbf{X} \in \mathcal{N}(\mathbf{m}, \Sigma)$ fait donc plus que de le décorréler. Les composantes de Y sont en effet indépendantes (ce qui est clair par définition de $\mathcal{N}(\mathbf{0}, I_d)$).

12.1.4 Représentation de la loi multigaussienne

On dit qu'un vecteur gaussien d-dimensionnel $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ est dégénéré lorsque sa matrice de covariance Σ est de rang inférieur à d (i.e. Σ n'est pas inversible). Nous allons voir dans cette section que $\mathbf{X}$ admet dans le cas non-dégénéré une densité, appelée *multigaussienne*, et dont l'expression peut être déduite de la densité $f_{\mathcal{N}(\mathbf{0},I)}$ par changement de variable affine. On traitera ensuite du cas des vecteurs gaussiens dégénérés. La fin du chapître est consacrée à des illustrations de la loi bigaussienne (multigaussienne en dimension 2).

Cas non-dégénéré : fondement de la densité multigaussienne

Propriété : si $\mathbf{X} = A\mathbf{Y} + \mathbf{b}$, où $\mathbf{Y}$ est un V.a.r.c. d-dimensionnel de densité $f_{\mathbf{Y}}$ et $A \in \mathcal{GL}_d(\mathbb{R})$, alors $\mathbf{X}$ est continu et

$$\forall \mathbf{x} \in \mathbb{R}^d,\ f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(A^{-1}(\mathbf{x} - \mathbf{b})) \times |det(A^{-1})| = \frac{f_{\mathbf{Y}}(A^{-1}(\mathbf{x} - \mathbf{b}))}{|det(A)|} \tag{12.4}$$

Ce résultat est une application directe de la formule de changement de variables avec comme difféomorphisme $h(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$.

Considérons maintenant un V.a.r. gaussien non-dégénéré $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$. On sait déjà par transformation de Mahalanobis que $M = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \mathbf{m}) \sim \mathcal{N}(\mathbf{0}, I)$ et on peut alors écrire $\mathbf{X}$ sous la forme

$$\mathbf{X} = \mathbf{m} + \Sigma^{\frac{1}{2}} M =: h(M) \tag{12.5}$$

La propriété précédente permet alors d'affirmer que $\mathbf{X}$ admet une densité, et on a en vertu de l'eq. (12.4) :

$$\begin{aligned} \forall \mathbf{x} \in \mathbb{R}^d,\ f_{\mathbf{X}}(\mathbf{x}) &= (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}[\Sigma^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})]^T[\Sigma^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})]} det(\Sigma^{-\frac{1}{2}}) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}(det(\Sigma))^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\Sigma^{-1}(\mathbf{x}-\mathbf{m})} \end{aligned} \tag{12.6}$$

Cas dégénéré

Dans le cas où Σ n'est pas inversible, il est clair que la fonction h de l'eq. (12.5) n'est pas bijective. On ne peut donc pas appliquer le changement de variables qui nous a permis d'obtenir la densité multigaussienne. Les V.a.r. gaussiennes dégénérés n'ont en fait pas de densité ; on a toutefois le résultat suivant :

Propriété : Soit $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ et $rg(\Sigma) = k < p$. Il existe alors un sous-espace vectoriel $H \subset \mathbb{R}^d$ de dimension $d - k$ telle que $\mathbf{a}^T\mathbf{X}$ soit une v.a. constante pour tout $\mathbf{a} \in H$.

Preuve : On sait que $\mathbf{X} = \mathbf{m} + \Sigma^{\frac{1}{2}}M$ où $M \sim \mathcal{N}(\mathbf{0}, I)$ et $rg(\Sigma^{\frac{1}{2}}) = k$. Soit $H = Ker((\Sigma^{\frac{1}{2}})^T)$, de dimension $dim(Ker(\Sigma^{\frac{1}{2}})) = d - dim(Im(\Sigma^{\frac{1}{2}})) = d - k$. Si $\mathbf{a} \in H$, on a $(\Sigma^{\frac{1}{2}})^T\mathbf{a} = 0$ et ainsi $\Sigma\mathbf{a} = 0$. Considérons la fonction caractéristique de la v.a.r. $\mathbf{a}^T\mathbf{X}$:

$$\begin{aligned} \forall u \in \mathbb{R},\ \Phi_{\mathbf{a}^T\mathbf{X}}(u) &= \mathbb{E}[e^{i(\mathbf{a}^T\mathbf{X})u}] = \mathbb{E}[e^{i(u\mathbf{a}^T)\mathbf{X}}] \\ &= e^{i(u\mathbf{a})^T\mathbf{m} - \frac{1}{2}u\mathbf{a})^T\Sigma u\mathbf{a})} = e^{i(u\mathbf{a})^T\mathbf{m}} \end{aligned} \tag{12.7}$$

et on conclut que $\mathbf{a}^T\mathbf{X}$ est un constante de valeur $\mathbf{a}^T\mathbf{m}$.

Illustration : étude de la densité bigaussienne

Soit $\mathbf{X} = (X_1, X_2) \sim \mathcal{N}(\mathbf{m}, \Sigma_\mathbf{X})$ un vecteur gaussien bidimensionnel (ou *couple bigaussien*). Supposons que $\sigma_{X_1} > 0$ et $\sigma_{X_2} > 0$. En notant $\rho = \frac{Cov[X_1, X_2]}{\sigma_{X_1}\sigma_{X_2}}$ le coefficient de corrélation de X_1 et X_2, on peut écrire

$$\Sigma_\mathbf{X} = \begin{pmatrix} \sigma^2_{X_1} & Cov[X_1, X_2] \\ Cov[X_1, X_2] & \sigma^2_{X_2} \end{pmatrix} = \begin{pmatrix} \sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{pmatrix} \tag{12.8}$$

et on calcule sans peine le déterminant de $\Sigma_\mathbf{X}$

$$det(\Sigma_\mathbf{X}) = \sigma^2_{X_1}\sigma^2_{X_2}(1 - \rho^2) \geq 0 \text{ puisque } \rho \in [-1, 1] \tag{12.9}$$

Dans le cas où $\rho = 1$, il vient que $rg(\Sigma_\mathbf{X}) = 1$. On sait alors que $\mathbf{X}$ n'admet pas de densité, et la propriété énoncée plus haut au sujet des vecteurs gaussiens dégénérés permet d'affirmer qu'il existe une droite vectorielle H telle que $\mathbf{a}^T\mathbf{X}$ soit une constante pour tout $\mathbf{a} \in H$. En remarquant la relation de colinéarité $\sigma_{X_2} \times \Sigma_\mathbf{X}(., 1) = \sigma_{X_1} \times \Sigma_\mathbf{X}(., 2)$, on trouve que $H = Vect_{L^2(\mathbb{P})}\{(\sigma_{X_2} \quad -\sigma_{X_1})\}$. Le cas $\rho = -1$ peut être traité de la même manière. Intéressons-nous maintenant au cas où $\rho \in]-1, 1[$. Puisque $det(\Sigma_\mathbf{X}) > 0$, $\Sigma_\mathbf{X}$ est inversible et $\mathbf{X}$ admet la densité

$$\forall (x_1, x_2) \in \mathbb{R}^2,\ f_\mathbf{X}(x_1, x_2) = \frac{1}{2\pi \times \sqrt{det(\Sigma_\mathbf{X})}} e^{-\frac{1}{2}\left[(x_1 - m_1 \quad x_2 - m_2)\Sigma_\mathbf{X}^{-1}(x_1 - m_1 \quad x_2 - m_2)^T\right]}$$

En explicitant $\Sigma_{\mathbf{X}}^{-1}$ (à l'aide de la fameuse formule de la comatrice), il vient

$$\Sigma_{\mathbf{X}}^{-1} = \frac{1}{\sigma_{X_1}^2 \sigma_{X_2}^2 (1-\rho^2)} \begin{pmatrix} \sigma_{X_2}^2 & -\rho\sigma_{X_1}\sigma_{X_2} \\ -\rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_1}^2 \end{pmatrix} \tag{12.10}$$

et la densité de $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma_X)$ peut alors s'écrire

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-m_1)^2}{\sigma_{X_1}^2} - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_{X_1}\sigma_{X_2}} + \frac{(x_2-m_2)^2}{\sigma_{X_2}^2}\right]} \tag{12.11}$$

Remarquons que l'on retrouve bien la densité de l'eq. (12.1) en prenant $\rho = 0$, $\mathbf{m} = \mathbf{0}$, et $\sigma_{X_1} = \sigma_{X_2} = 1$ comme valeurs des paramètres.

Considérons maintenant, pour $C \in \mathbb{R}_*^+$ quelconque, l'ensemble $\mathcal{E}_C$ des points de $\mathbb{R}^2$ en lesquels la densité $f_{\mathbf{X}}$ est constante égale à C (*courbe d'isodensité*) :

$$\mathcal{E}_C = \{(x_1, x_2) \in \mathbb{R}^2 : \ f_{\mathbf{X}}(x_1, x_2) = C\} \tag{12.12}$$

Un passage au logarithme dans l'eq. (12.11) va nous donner accès à une caractérisation géométrique simple de $\mathcal{E}_C$. Posons pour ce faire les notations suivantes

$$\begin{cases} Q(x_1, x_2) = \frac{(x_1-m_1)^2}{\sigma_{X_1}^2(1-\rho^2)} - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_{X_1}\sigma_{X_2}(1-\rho^2)} + \frac{(x_2-m_2)^2}{\sigma_{X_2}^2(1-\rho^2)} \\ K = -2ln\left(2\pi det(\Sigma_{\mathbf{X}})C\right) = -2ln\left(2\pi\sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)C\right) \end{cases} \tag{12.13}$$

On arrive alors sans difficulté au constat que $\mathcal{E}_C$ est une ellipse d'équation

$$\mathcal{E}_C = \left\{(x_1, x_2) \in \mathbb{R}^2 : \ Q(x_1, x_2) = K\right\} \tag{12.14}$$

Les $\mathcal{E}_C$ sont aussi appelés ellipses de concentration. L'ensemble de ces ellipses permet de prolonger la notion d'intervalle de confiance au cas bidimensionnel (et au cas d-dimensionnel, où l'on parle d'*ellipsoïdes de concentration*). Notons $\mathcal{E}_C^+ = \{(x_1, x_2) \in \mathbb{R}^2 : \ f_{\mathbf{X}}(x_1, x_2) \geq C\}$. Pour un niveau de confiance $\alpha \in [0, 1[$ donné, on peut en effet trouver C_α tel que $\mathbb{P}\left(\mathbf{X} \in \mathcal{E}_{C_\alpha}^+\right) = \alpha$. On appelle alors $\mathcal{E}_{C_\alpha}$ l'*ellipse de confiance de niveau* α du vecteur gaussien $\mathbf{X}$. On peut montrer (c.f. exercice d'entraînement 6)) que la constante C_α s'exprime en fonction du quantile à $\alpha\%$ de la loi du χ_2^2, noté $q_{\chi_2^2}(\alpha)$:

$$C_\alpha = \frac{e^{-\frac{1}{2}q_{\chi_2^2}(\alpha)}}{2\pi\sqrt{det(\Sigma_{\mathbf{X}})}} \tag{12.15}$$

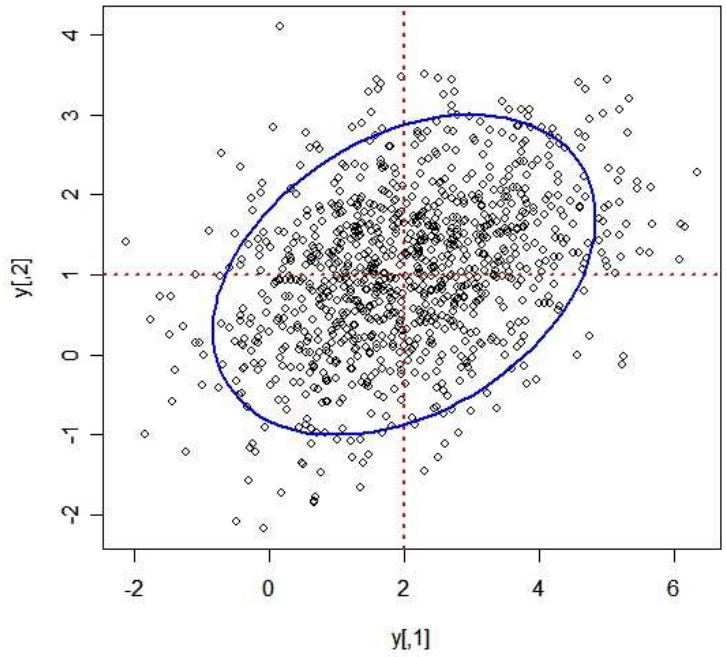


FIG. 12.2 – Un échantillon de 1000 points tirés selon la loi $\mathcal{N}(\mathbf{m}, \Sigma)$, avec $\mathbf{m} = (2,1)^T$, $\sigma_1^2 = 2,\ \sigma_2^2 = 1$, et $\rho\sigma_1\sigma_2 = 0.5$. En bleu : l'ellipse de confiance à 95% de la loi $\mathcal{N}(\mathbf{m}, \Sigma)$.

12.2 More Gaussian vectors ?

12.2.1 Vecteurs gaussiens et conditionnement

A real random process $(Y(\mathbf{x}))_{\mathbf{x}\in D}$ is defined as a *Gaussian Process* (GP) whenever all its finite-dimensional distributions are gaussian. Consequently, for all $n \in \mathbb{N}$ and for all set $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ of n points of D, there exists a vector $\mathbf{m} \in \mathbf{R}^n$ and a symmetric positive semi-definite matrix $\Sigma \in \mathcal{M}_n(\mathbb{R})$ such that $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n))$ is a gaussian Vector, following a multigaussian probability distribution $\mathcal{N}(\mathbf{m}, \Sigma)$. More specifically, for all $i \in [1, n]$, $Y(\mathbf{x}^i) \sim \mathcal{N}(\mathbb{E}[Y(\mathbf{x}^i)], Var[Y(\mathbf{x}^i)])$ where $\mathbb{E}[Y(\mathbf{x}^i)]$ is the ith coordinate of $\mathbf{m}$ and $Var[Y(\mathbf{x}^i)]$ is the ith diagonal term of Σ. Furthermore, all couples $(Y(\mathbf{x}^i), Y(\mathbf{x}^j))$ $i, j \in [1, n], i \neq j$ are multigaussian with a covariance $Cov[Y(\mathbf{x}^i), Y(\mathbf{x}^j)]$ equal to the non-diagonal term of Σ indexed by i and j.

A Random Process Y is said to be *first order stationary* if its mean is a constant, i.e. if

$\exists \mu \in \mathbb{R} | \ \forall \mathbf{x} \in D, \ \mathbb{E}[Y(\mathbf{x})] = \mu$. Y is said to be *second order stationary* if it is first order stationary and if there exists furthermore a function of positive type, $c : D - D \longrightarrow \mathbb{R}$, such that for all pairs $(\mathbf{x}, \mathbf{x}') \in D^2$, $Cov[Y(\mathbf{x}), Y(\mathbf{x}')] = c(\mathbf{x} - \mathbf{x}')$. We then have the following expression for the covariance matrix of the observations at $\mathbf{X}$:

$$\Sigma := \begin{pmatrix} \sigma^2 & c(\mathbf{x}_1 - \mathbf{x}_2) & ... & c(\mathbf{x}_1 - \mathbf{x}_n) \\ c(\mathbf{x}_2 - \mathbf{x}_1) & \sigma^2 & ... & c(\mathbf{x}_2 - \mathbf{x}_n) \\ ... & ... & ... & ... \\ c(\mathbf{x}_n - \mathbf{x}_1) & c(\mathbf{x}_n - \mathbf{x}_2) & ... & \sigma^2 \end{pmatrix} \tag{12.16}$$

where $\sigma^2 := c(0)$. Second order stationary processes are sometimes called *weakly stationary*. A major feature of GPs is that their *weak stationarity* is equivalent to *strong stationarity* : if Y is a weakly stationary GP, the law of probability of the random variable $Y(\mathbf{x})$ doesn't depend on $\mathbf{x}$, and the joint distribution of $(Y(\mathbf{x}^1), ..., Y(\mathbf{x}^n))$ is the same as the distribution of $(Y(\mathbf{x}^1 + \mathbf{h}), ..., Y(\mathbf{x}^n + \mathbf{h}))$ whatever the set of points $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \in D^n$ and the vector $\mathbf{h} \in \mathbb{R}^n$ such that $\{\mathbf{x}^1 + \mathbf{h}, ..., \mathbf{x}^n + \mathbf{h}\} \in D^n$.

To sum up, a stationary GP is entirely defined by its mean μ and its covariance function $c(.)$. The classical framework of Kriging for Computer Experiments is to make predictions of a costly simulator y at a new set of sites $\mathbf{X}_{new} = \{\mathbf{x}^{n+1}, ..., \mathbf{x}^{n+q}\}$ (most of the time, $q = 1$), on the basis of the collected observations at the initial design $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$, and under the assumption that y is one realization of a stationary GP Y with known covariance function c (in theory). Simple Kriging (SK) assumes a known mean, $\mu \in \mathbb{R}$. In Ordinary Kriging (OK), μ is estimated.

12.2.2 Conditioning Gaussian Vectors

Let us consider a centered Gaussian vector $V = (V_1, V_2)$ with covariance matrix

$$\Sigma_V = \mathbb{E}[VV^T] = \begin{pmatrix} \Sigma_{V_1} & \Sigma_{cross}^T \\ \Sigma_{cross} & \Sigma_{V_2} \end{pmatrix} \tag{12.17}$$

Key properties of Gaussian vectors include that the orthogonal projection of a Gaussian vector onto a linear subspace is still a Gaussian vector, and that the orthogonality of two subvectors V_1, V_2 of a Gaussian vector V (i.e. $\Sigma_{cross} = \mathbb{E}[V_2 V_1^T] = 0$) is equivalent to their independence. We now express the conditional expectation $\mathbb{E}[V_1|V_2]$. $\mathbb{E}[V_1|V_2]$ is by definition such that $V_1 - \mathbb{E}[V_1|V_2]$ is independent of V_2. $\mathbb{E}[V_1|V_2]$ is thus fully characterized as orthogonal projection on the vector space spanned by the components of V_2, solving

the so called *normal equations* :

$$\mathbb{E}[(V_1 - \mathbb{E}[V_1|V_2])V_2^T] = 0 \tag{12.18}$$

Assuming linearity of $\mathbb{E}[V_1|V_2]$ in V_2, i.e. $\mathbb{E}[V_1|V_2] = AV_2$ ($A \in \mathcal{M}_n(\mathbb{R})$), a straightforward development of eq. 12.18 gives the matrix equation $\Sigma_{cross}^T = A\Sigma_{V_2}$, and hence $\Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2$ is a suitable solution provided Σ_{V_2} is full ranked[1]. We conclude that

$$\mathbb{E}[V_1|V_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1} V_2 \tag{12.19}$$

by uniqueness of the orthogonal projection onto a closed linear subspace in a Hilbert space. Using the independence between $(V_1 - \mathbb{E}[V_1|V_2])$ and V_2, one can calculate the conditional covariance matrix $\Sigma_{V_1|V_2}$:

$$\begin{aligned}\Sigma_{V_1|V_2} &= \mathbb{E}[(V_1 - \mathbb{E}[V_1|V_2])(V_1 - \mathbb{E}[V_1|V_2])^T|V_2] = \mathbb{E}[(V_1 - AV_2)(V_1 - AV_2)^T] \\ &= \Sigma_{V_1} - A\Sigma_{cross} - \Sigma_{cross}^T A^T + A\Sigma_{V_2}A^T = \Sigma_{V_1} - \Sigma_{cross}^T \Sigma_{V_2}^{-1} \Sigma_{cross}\end{aligned} \tag{12.20}$$

Now consider the case of a non-centered random vector $V = (V_1, V_2)$ with mean $m = (m_1, m_2)$. The conditional distribution $V_1|V_2$ can be obtained by coming back to the centered random vector $V - m$. We then find that $\mathbb{E}[V_1 - m_1|V_2 - m_2] = \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$ and hence $\mathbb{E}[V_1|V_2] = m_1 + \Sigma_{cross}^T \Sigma_{V_2}^{-1}(V_2 - m_2)$.

12.2.3 Preuve du lemme 3.95

Soient n et m les tailles respectives des V.a.r. X_1 et X_2. On veut montrer que la densité du vecteur (X_1, X_2) est multi-gaussienne. Il suffit de montrer que la fonction caractéristique $\Phi_{(X_1,X_2)} : (T_n, T_m) \in \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{E}[e^{i(<T_n,X_1>+<T_m,X_2>)}]$ est de la forme :

$$\Phi_{(X_1,X_2)}(T_n, T_m) = e^{\mathbf{i}\left(\begin{bmatrix} T_n & T_m \end{bmatrix}\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}\right) - \frac{1}{2}\left(\begin{bmatrix} T_n & T_m \end{bmatrix}\begin{bmatrix} \Sigma_1 & \Sigma_{12}^T \\ \Sigma_{12} & \Sigma_2 \end{bmatrix}\begin{bmatrix} T_n \\ T_m \end{bmatrix}\right)}$$

En utilisant la définition de $\Phi_{(X_1,X_2)}$ et la formule de l'espérance totale, on a :

$$\begin{aligned}\Phi_{(X_1,X_2)}(T_n, T_m) &= \mathbb{E}[e^{\mathbf{i}(<T_n,X_1>+<T_m,X_2>)}] = \mathbb{E}\left[\mathbb{E}[e^{\mathbf{i}(<T_n,X_1>+<T_m,X_2>)}/X_1]\right] \\ &= \mathbb{E}\left[e^{\mathbf{i}(<T_n,X_1>)}\mathbb{E}[e^{\mathbf{i}(<T_m,X_2>)}/X_1]\right] = \mathbb{E}\left[e^{\mathbf{i}(<T_n,X_1>)}\Phi_{X_2/X_1}(T_m)\right] = (*)\end{aligned}$$

On utilise alors l'hypothèse de gaussianité de la variable conditionnelle X_2/X_1 :

$$(*) = \mathbb{E}\left[e^{\mathbf{i}(T_n^T X_1)} e^{\mathbf{i}(T_m^T(AX_1+b)) - \frac{1}{2}(T_m^T \Sigma T_m)}\right] = e^{\mathbf{i}(T_m^T b) - \frac{1}{2}(T_m^T \Sigma T_m)}\mathbb{E}\left[e^{\mathbf{i}((T_n^T + T_m^T A)X_1)}\right] = (**)$$

[1] If Σ_{V_2} is not invertible, the equation holds in replacing $\Sigma_{V_2}^{-1}$ by the pseudo-inverse $\Sigma_{V_2}^{\dagger}$.

Puis l'hypothèse de gaussianité de la variable X_1 :

$$(**) = e^{\mathbf{i}(T_m^T b)-\frac{1}{2}(T_m^T\Sigma T_m)}e^{\mathbf{i}((T_n^T+T_m^T A)m_1)-\frac{1}{2}((T_n^T+T_m^T A)\Sigma_1(T_n^T+T_m^T A)^T)}$$
$$= e^{\mathbf{i}(T_m^T b+(T_n^T+T_m^T A)m_1)}e^{-\frac{1}{2}(T_m^T\Sigma T_m+T_n^T\Sigma_1 T_n+T_m^T A\Sigma_1 A^T T_m+2T_n^T\Sigma_1 A^T T_m)} = (***)$$

Et on conclut en écrivant $(***)$ matriciellement :

$$(***) = e^{\mathbf{i}\left(\begin{bmatrix} T_n & T_m \end{bmatrix}\begin{bmatrix} m_1 \\ Am_1+b \end{bmatrix}\right)-\frac{1}{2}\left(\begin{bmatrix} T_n & T_m \end{bmatrix}\begin{bmatrix} \Sigma_1 & \Sigma_1 A^T \\ A\Sigma_1^T & \Sigma+A\Sigma_1^T\Sigma_1^{-1}\Sigma_1 A^T \end{bmatrix}\begin{bmatrix} T_n \\ T_m \end{bmatrix}\right)}$$

Par identification, on trouve finalement que $\Sigma_{12} = A\Sigma_1^T$ d'où $A = \Sigma_{12}\Sigma_1^{-1}$. On remarque aussi que $\Sigma_2 = \Sigma + \Sigma_{12}\Sigma_1^{-1}\Sigma_{12}^T$ soit $\Sigma_{2/1} = \Sigma = \Sigma_2 - \Sigma_{12}\Sigma_1^{-1}\Sigma_{12}^T$ ce qui est conforme aux résultats bien connus de conditionnement des vecteurs gaussiens. Remarquons que le résultat ci-dessus avait en fait déjà été établi dans [AALF97].

12.2.4 divergence de Küllback-Leibler entre deux multigaussiennes

On donne ci-dessous quelques résultats classiques sur l'entropie et la divergence de Küllback-Leibler. Après avoir rappelé les définitions et propriétés de base, on exposera en détail le calcul de la divergence de K-L entre deux vecteurs gaussiens quelconques.

Entropie

Soit X une variable aléatoire à valeurs dans un espace mesuré $(\mathbb{X}, \tau, \mu)$, admettant f pour densité de probabilité par rapport à μ. On appelle *entropie* H de la v.a. X (ou indifférement de la densité f) le réel positif défini comme suit :

$$\mathbb{H}[X] = \mathbb{H}[f(\mathbf{x})] = \mathbb{E}[-ln(f(X))] = -\int_{\mathbb{X}} ln(f(\mathbf{x}))f(\mathbf{x})d\mu(\mathbf{x}) \tag{12.21}$$

L'entropie mesure en quelque sorte l'incertitude associée à une variable aléatoire X : $\mathbb{H}$ elle est d'autant plus forte qu'il y a d'imprévisibilité dans le comportement de X.

Exemple : si $(\mathbb{X}, \tau, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ et $X \sim N(0,1)$, on a

$$\mathbb{H}[X] = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} ln\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)dx = -ln\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{1}{2}Var[X] = \frac{1}{2}ln(2\pi e)$$

De même, dans le cas d'une gaussienne de paramètres (μ, σ^2), on trouve que l'entropie vaut $\mathbb{H} = \frac{1}{2}ln(2\pi e\sigma^2)$. Il apparaît ainsi que l'entropie d'une gaussienne croît de manière logarithmique avec la variance.

Divergence de Küllback-Leibler

On considère cette fois-ci deux variables aléatoires X et Y à valeurs dans $(\mathbb{X}, \tau, \mu)$, avec pour densités respectives f et g par rapport à μ. La divergence de Küllback-Leibler (ou entropie relative) de Y à X est définie par :

$$KL(X||Y) = KL(f||g) = \int_{\mathbb{X}} ln\left[\frac{f(\mathbf{x})}{g(\mathbf{x})}\right] f(\mathbf{x}) d\mu(\mathbf{x}) \tag{12.22}$$

On peut montrer que $KL(f||g)$ est positive ou nulle, et nulle ssi f=g μ-presque partout. On remarque également que KL n'est pas symétrique en ses arguments.

Interprétation de la divergence KL dans le cas de distributions paramétriques

Lorsque f et g sont des densités issues d'une même famille de loi (exponentielles, gaussiennes, Bernouilli, etc...) mais possédant différentes valeurs de paramètres, la divergence de KL peut-être vue sous l'angle de la théorie de la vraisemblance. On note ici θ_1 et θ_2 deux occurences du vecteur des paramètres associé à une famille de distributions (par exemple, $\theta = (\mu, \sigma^2)$ pour la famille des gaussiennes monodimensionnelles). On suppose ainsi que $f_{\theta_1}(.) = f(.|\theta_1)$ et $f_{\theta_2}(.) = f(.|\theta_2)$ sont de telles distributions. La divergence KL de f_2 à f_1 peut alors être écrite comme :

$$KL(f_1||f_2) = \mathbb{E}_{\theta_1}[ln(f_{\theta_1}(X)) - ln(f_{\theta_2}(X))] = \mathbb{E}_{\theta_1}\left[ln\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}\right)\right] \tag{12.23}$$

Cette quantité mesure ainsi l'écart moyen entre les log-vraisemblance associées au paramètres θ_1 et θ_2 lorsque l'échantillon considéré est généré par la loi de densité $f(.|\theta_1)$.

Application au cas de deux multigaussiennes

Entropie d'un vecteur aléatoire gaussien continu

Soit $X \sim \mathcal{N}(m_X, \Sigma_X)$ un vecteur aléatoire gaussien continu. Son entropie $\mathbb{H}[X]$ peut s'écrire $\mathbb{E}[-ln(f(X))]$, soit $\mathbb{E}[ln((2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma_X)}) + \frac{1}{2}(X - m_X)'\Sigma_X^{-1}(X - m_X)]$. Le premier terme est déterministe et le second fait apparaître l'espérance d'une forme quadratique définie sur un vecteur gaussien :

$$\mathbb{E}[(X - m_X)'\Sigma_X^{-1}(X - m_X)] = \mathbb{E}[Z'Z] = \mathbb{E}[Z_1^2 + ... + Z_d^2],$$

où $Z = \Sigma_X^{-\frac{1}{2}}(X - m_X)$ est un vecteur gaussien centré réduit. Comme $E[Z'Z] = d$ en tant qu'espérance d'un χ_d^2, on trouve finalement :

$$\mathbb{H}[X] = \frac{1}{2} ln(\det(\Sigma_X)) + \frac{d}{2}(ln(2\pi e)) \tag{12.24}$$

Et on observe ainsi que l'entropie du vecteur gaussien X est croissante en le déterminant de Σ_X, i.e. en le volume des ellipsoïdes de niveau de la densité multigaussienne.

Divergence KL entre deux vecteurs gaussiens

Soient maintenant X et Y deux vecteurs gaussiens à valeurs dans $\mathbb{R}^d$, de lois respectives $\mathcal{N}(m_X, \Sigma_X)$ et $\mathcal{N}(m_Y, \Sigma_Y)$. On notera $\theta_X = (m_X, \Sigma_X)$ et $\theta_Y = (m_Y, \Sigma_Y)$ et $f_{\theta_X}, f_{\theta_Y}$ les densités associées au deux distributions multigaussiennes étudiées. Nous proposons ici le calcul détaillé de la divergence KL de X et Y :

On part de l'écriture de la divergence en termes de vraisemblances comparées :

$$KL(X||Y) = \mathbb{E}_{\theta_X}[ln(f_{\theta_X}(U)) - ln(f_{\theta_Y}(U))]$$

Comme les termes en $ln((2\pi)^{\frac{d}{2}})$ se neutralisent, il reste la somme d'une expression déterministe, et de l'espérance d'une différence de formes quadratiques :

$$KL(X||Y) = \underbrace{\frac{1}{2} ln(det(\Sigma_Y \Sigma_X^{-1}))}_{(*)} + \underbrace{\frac{1}{2} E_{\theta_X}[(U-\mu_Y)' \Sigma_Y^{-1}(U-\mu_Y) - (U-\mu_X)' \Sigma_X^{-1}(U-\mu_X)]}_{(**)}$$

Le terme (**) peut lui aussi être alors simplifié. On peut en effet observer de nouveau que $\mathbb{E}_{\theta_X}[(U-\mu_X)' \Sigma_X^{-1}(U-\mu_X)]$ est l'espérance d'un χ^2_d, et est à ce titre égale à d. Développons enfin l'expression $\mathbb{E}_{\theta_X}[(U-\mu_Y)' \Sigma_Y^{-1}(U-\mu_Y)]$:

En écrivant cette dernière comme $\mathbb{E}_{\theta_X}[(U-\mu_X+\mu_X-\mu_Y)' \Sigma_Y^{-1}(U-\mu_X+\mu_X-\mu_Y)]$, on arrive à la somme d'un terme déterministe $(\mu_X-\mu_Y)' \Sigma_Y^{-1}(\mu_X-\mu_Y)$ et de l'espérance $\mathbb{E}_{\theta_X}[(U-\mu_X)' \Sigma_Y^{-1}(U-\mu_X)]$. Il suffit alors d'exprimer U sous la forme $m_X + \Sigma_X^{\frac{1}{2}} N$, où N est un vecteur gaussien centré réduit. L'espérance en question se réduit alors à $\mathbb{E}[N'AN]$ avec $A = \Sigma_X^{\frac{1}{2}} \Sigma_Y^{-1} \Sigma_X^{\frac{1}{2}}$, ce qui permet de conclure en remarquant que :

$$\mathbb{E}[N'AN] = \mathbb{E}\left[\sum_{1 \le i,j \le d} a_{ij} N_i N_j\right] = \sum_{1 \le i,j \le d} a_{ij} \mathbb{E}[N_i N_j] = \sum_{i=1} a_{ii} = tr(A) = tr(\Sigma_Y^{-1} \Sigma_X)$$

Au final, la divergence KL recherchée est ainsi égale à :

$$KL(X||Y) = \frac{1}{2}\left[ln(\det(\Sigma_Y \Sigma_X^{-1})) - d + (\mu_X-\mu_Y)' \Sigma_Y^{-1}(\mu_X-\mu_Y) + tr(\Sigma_Y \Sigma_X^{-1})\right] \tag{12.25}$$

12.2.5 Préliminaire à un calcul de matrice d'information de Fisher : différentielle de l'application *déterminant*

On rappelle que le déterminant d'une matrice carrée $A = (a_{ij})_{i,j=1,..,n} \in \mathcal{M}_n(\mathbb{R})$ peut être défini de la manière suivante :

$$\det(A) = \sum_{\sigma \in S_n} \varepsilon(\sigma) \Pi_{i=1}^n a_{i,\sigma(i)} \tag{12.26}$$

où $\varepsilon : S_n \longrightarrow \{-1, 1\}$ est l'application *signature*, mesure de la parité du nombre d'échanges d'une permutation. On remarque que l'application déterminant est une fonction polynômiale de plusieurs variables. Elle est donc à ce titre indéfiniment différentiable. On s'intéresse ici à l'application différentielle de det, c'est à dire l'application

$$d(\det) : \mathcal{M}_n(\mathbb{R}) \longrightarrow L(\mathcal{M}_n(\mathbb{R}), \mathbb{R}),$$

qui à chaque élément $A \in \mathcal{M}_n(\mathbb{R})$ associe une forme linéaire $d(\det)_A : \mathcal{M}_n(\mathbb{R}) \longrightarrow \mathbb{R}$ telle que

$$\lim_{||H|| \to 0} \left| \frac{\det(A+H) - \det(A) - d(\det)_A(H)}{||H||} \right| = 0, \tag{12.27}$$

où $||H|| := sup_{||v||_{\mathbb{R}^n} \leq 1} \frac{||Hv||_{\mathbb{R}^n}}{||v||_{\mathbb{R}^n}}$ est la norme "triple" usuelle de H. Nous proposons ci-dessous le calcul de $d(\det)_{\cdot}$ en deux étapes : nous calculons dans un premier temps la différentielle en l'identité, $d(\det)_I$, puis nous étendons le résultat obtenu au cas général.

Soit $h \in \mathbb{R}^+$ et $U \in \mathcal{M}_n(\mathbb{R})$ quelconque de norme unité. On a le développement suivant :

$$\begin{aligned}
\det(I + hU) &= \sum_{\sigma \in S_n} \varepsilon(\sigma) \Pi_{i=1}^n (h u_{i,\sigma(i)} + \delta_{i,\sigma_i}) \\
&= \Pi_{i=1}^n (h u_{i,i} + 1) + \sum_{\sigma \in S_n, \sigma \neq id} \varepsilon(\sigma) \Pi_{i=1}^n (h u_{i,\sigma(i)} + \delta_{i,\sigma_i}) \\
&= 1 + h \sum_{i=1}^n u_{ii} + O(h^2) + \sum_{\sigma \in S_n, \sigma \neq id} \varepsilon(\sigma) \Pi_{i=1}^n (h u_{i,\sigma(i)} + \delta_{i,\sigma_i}) \\
&= \det(I) + h \times \mathrm{tr}(U) + O(h^2)
\end{aligned} \tag{12.28}$$

Et comme *det* est $\mathcal{C}^\infty$, le fait que :

$$\lim_{h \to 0} \left| \frac{\det(I + hU) - \det(I)}{||hU||} - \mathrm{tr}(hU) \right| = 0 \tag{12.29}$$

suffit à établir que $\forall H \in \mathcal{M}_n(\mathbb{R}),\ d(\det)_I(H) = tr(H)$.

Soit maintenant $A \in \mathcal{GL}_n(\mathbb{R})$ et $H \in \mathcal{M}_n(\mathbb{R})$ quelconques. On peut alors affirmer que $\forall H \in \mathcal{M}_n(\mathbb{R})$:

$$d(\det)_A(H) = \det(A)\mathrm{tr}(A^{-1}H) \tag{12.30}$$

On a en effet en vertu du précédent résultat :

$$\begin{aligned}\left|\frac{\det(A+H)-\det(A)-\det(A)\mathrm{tr}(A^{-1}H)}{||H||}\right| &= |\det(A)|\left|\frac{\det(I+A^{-1}H)-\det(I)-\mathrm{tr}(A^{-1}H)}{||H||}\right| \\ &\leq \frac{|\det(A)|}{||A^{-1}||}\left|\frac{\det(I+A^{-1}H)-\det(I)-\mathrm{tr}(A^{-1}H)}{||A^{-1}H||}\right| \\ &\longrightarrow 0 \text{ lorsque } ||H|| \to 0\end{aligned} \tag{12.31}$$

On peut finalement montrer par prolongement analytique que 12.30 vaut sur tout $\mathcal{M}_n(\mathbb{R})$.

Chapitre 13

Eléments d'optimisation non-contrainte

Lorsque l'on veut optimiser la réponse d'un simulateur numérique y, le coût (en temps de calcul) de chaque évaluation rend bien souvent impossible l'application des routines d'optimisation usuelles (gradients, simplexe, recuit, etc...) directement sur la fonction y. On peut alors être tenté d'estimer un métamodèle déterministe m et de s'en servir pour faciliter l'optimisation. La manière la plus directe est d'optimiser la fonction m. On se ramène alors à un classique problème d'optimisation.

Nous évoquons dans la section 13.1 quelques généralités sur ce pan largement étudié des mathématiques appliquées qu'est l'optimisation numérique multivariables. Les méthodes à employer doivent bien sûr être adaptées en fonction de la nature du métamodèle choisi. Nous présentons dans la section 3.2 quelques cas particuliers de couplages entre métamodèles déterministes et méthodes d'optimisations.

La suite logique de cette annexe est le chapitre 4, voué en particulier à mettre en garde le lecteur contre certains dangers de l'optimisation sur base de métamodèles déterministes. On y aborde en effet certains pièges classiques liés au remplacement de y par m lors de l'étape d'optimisation, et pointons en quoi les qualités attendues d'un métamodèle doivent être très différentes selon que l'on souhaite obtenir une approximation fidèle (au sens des critères usuels) ou que l'on cherche à se donner une fonction de remplacement pour l'optimisation.

13.1 Optimisation de fonctions déterministes

Cette section donne un rapide aperçu des techniques classiques d'optimisation numérique des fonctions. Le lecteur désireux d'étudier plus en détail et/ou de mettre en pratique ces différentes méthodes pourra consulter des ouvrages tels que ([Cia98], [GMW81], [Cul94]).

13.1.1 Généralités et outils de base de l'optimisation numérique

On considére le problème de *minimisation*[1] suivant :

$$\min_{\mathbf{x}\in A} m(\mathbf{x}) \tag{13.1}$$

où $A \subseteq D$ est un sous-ensemble de valeurs *admissibles* du paramètre $\mathbf{x}$ et m est une fonction connue analytiquement, ce qui est le cas pour les métamodèles abordés au cours du chapitre 2. Dans l'éq. 13.1 et dans la suite de cette section, nous supposons implicitement que des conditions suffisantes d'existence du minimum de m sur A sont réunies (par exemple la continuité de m et la compacité de A). Ainsi, si $\mathbf{x}^* = \arg\min_{\mathbf{x}\in A} m(\mathbf{x})$, $\mathbf{x}^*$ et $m(\mathbf{x}^*)$ sont respectivement appelés *minimiseur* et *minimum* de m sur A (attention, il ne peut exister qu'un unique minimum mais il peut exister plusieurs minimiseurs !). On dit lorsque $A = D$ que le problème est *non-contraint*. Les méthodes possibles de résolution du problème 13.1 sont nombreuses. Rappelons avant de présenter une selection d'algorithmes quelques notions et propriétés indispensables.

Local v.s. global

On appelle *minimum global* de m —noté $\min(m)$— la plus petite valeur que peut atteindre m sur tout son domaine de définition D (cas non-contraint), autrement dit lorsque $\min(m)$ est atteint et que pour tout $\mathbf{x} \in D$, $m(\mathbf{x}) \geq \min(m)$. Les valeurs de $\mathbf{x}$ telles que le minimum global soit atteint sont appelées *minimiseurs globaux* de m (il se peut qu'il n'y en ait qu'un seul, mais cela n'est pas nécessairement le cas : penser à une sinusoïde). Seules quelques rares classes de problèmes (e.g. convexes) possèdent un minimiseur global accessible sans faire appel à des méthodes très coûteuses en temps de calcul et au succès incertain. On doit donc souvent se contenter de résultats *sub-optimaux*, tels que ceux fournis par des méthodes dites « locales ». On appelle *minimiseur local* tout $\mathbf{x}_0 \in D$ pour lequel il existe une boule B centrée en $\mathbf{x}_0$ et de rayon positif telle que $\forall \mathbf{x} \in B,\ m(\mathbf{x}_0) \leq m(\mathbf{x})$. Autrement dit, les minimiseurs locaux sont le fond des "bassins" (potentiellement nombreux) que possède m. La valeur prise par m en un minimiseur local est appelée *minimum local*.

[1] la maximisation de m étant formellement équivalente à la minimisation de la fonction $-m$.

Pour revenir brièvement sur le fait qu'un problème d'optimisation tel que 13.1 n'admet pas forcément de solution, nous en donnons ici deux illustrations élémentaires en une dimension. Il est clair que $\min_{x\in A}(-x^2)$ n'a pas de minimum (ni local, ni global) : pour tout $x \in \mathbb{R}$, il existe $y \in \mathbb{R}$ (e.g. $y = x + 1$ dans le cas où $x > 0$) tel que $m(y) < m(x)$. De même, cette fonction n'admet pas de minimum sur l'intervalle ouvert $]0, 1[$.

Outils différentiels

La plupart des méthodes d'optimisation locale se basent sur des outils différentiels tel que le *gradient* ou la matrice *Hessienne*, pour peu que la régularité de la fonction étudiée le permette. Lorsqu'il existe, le gradient de m en $\mathbf{x} \in D$ —noté $\nabla m(\mathbf{x})$— est défini comme le vecteur des dérivées partielles de m :

$$\nabla m(\mathbf{x}) = \begin{pmatrix} \frac{\partial m}{\partial x_1}(\mathbf{x}) \\ ... \\ \frac{\partial m}{\partial x_d}(\mathbf{x}) \end{pmatrix} \tag{13.2}$$

Le développement de Taylor de m à l'ordre 1 au voisinage de $\mathbf{x}$ va nous permettre de faire apparaître une des idées fondatrices des méthodes locales dites *d'ordre* 1. On a

$$m(\mathbf{x} + t\mathbf{h}) = m(\mathbf{x}) + t\langle \nabla m(\mathbf{x}), \mathbf{h}\rangle + o(t||\mathbf{h}||) \tag{13.3}$$

et le gradient $\nabla m(\mathbf{x})$ nous permet ainsi de construire la meilleure approximation linéaire de m au voisinage de $\mathbf{x}$. On va chercher à tirer parti de l'eq. 13.3 pour trouver des *directions de descente*, i.e. $\mathbf{d} \in \mathbb{R}^d$ tel que $m(\mathbf{x} + t\mathbf{d}) < m(\mathbf{x})$ pour $t > 0$ choisi suffisamment petit. La direction $\mathbf{d} = -\nabla m(\mathbf{x})$ apparaît être idéale puisque $\langle \nabla m(\mathbf{x}), -\nabla m(\mathbf{x})\rangle = -||\nabla m(\mathbf{x})||^2 < 0$ et que l'on a donc en vertu de 13.3 que $m(\mathbf{x} - t\nabla m(\mathbf{x})) < m(\mathbf{x})$ pour t suffisamment petit. Notons que toutes les directions $\mathbf{d}$ telles que $\langle \nabla m(\mathbf{x})), \mathbf{d}\rangle < 0$ conviennent (i.e. faisant un angle concave avec le gradient), et certaines sont plus efficaces que la direction opposée au gradient. Ce constat est à l'origine de nombreuses méthodes de *descente*, dont on présentera quelques variantes dans la prochaine section.

La connaissance —s'il y a lieu— de la différentielle d'ordre 2 de la fonction à optimiser est un atout de taille pour mener une optimisation locale. Lorsqu'elle existe, la Hessienne de m en $\mathbf{x}$ est définie comme la matrice des dérivées partielles secondes de m :

$$\nabla^2 m(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 m}{\partial x_1^2}(\mathbf{x}) & ... & \frac{\partial^2 m}{\partial x_1 x_d}(\mathbf{x}) \\ ... & ... & ... \\ \frac{\partial^2 m}{\partial x_1 x_d}(\mathbf{x}) & ... & \frac{\partial^2 m}{\partial x_d^2}(\mathbf{x}) \end{pmatrix} \tag{13.4}$$

Un développement de Taylor à l'ordre 2 offre en chaque $\mathbf{x}$ la meilleure approximation locale de m par une forme quadratique (i.e. la partie polynômiale du DL) :

$$m(\mathbf{x}+t\mathbf{h}) = m(\mathbf{x}) + t\langle \nabla m(\mathbf{x}), \mathbf{h}\rangle + \frac{t^2}{2}\langle \nabla^2 m(\mathbf{x})\mathbf{h}, \mathbf{h}\rangle + o(t^2||\mathbf{h}||^2) \tag{13.5}$$

Les propriétés algébriques de la matrice Hessienne $\nabla^2 m(\mathbf{x})$ permettent ainsi de préciser le comportement de m au voisinage de tout $\mathbf{x}$, et même comme nous allons le voir ci-dessous de caractériser les minima locaux. L'exploitation de 13.5 donne lieu à un certain nombre de méthodes locales, dites *d'ordre* 2 (Cf. section suivante).

Eléments de caractérisation des optima locaux (cas non-contraint)

Lorsque la fonction à optimiser est de classe $\mathcal{C}^1$: si $\mathbf{x}$ est un minimiseur local de m alors on a nécessairement $\nabla m(\mathbf{x}) = 0$ (condition nécessaire d'optimalité). Ainsi les minimiseurs de m sont à chercher parmi les $\mathbf{x}$ qui annulent le gradient, appelés *points critiques*.

Lorsque la fonction à optimiser est de classe $\mathcal{C}^2$: si $\mathbf{x}$ est un minimiseur local de m alors on a nécessairement (en plus de la condition sur le gradient) que la matrice Hessienne $\nabla^2 m(\mathbf{x})$ est **semi**-définie positive (i.e. $\forall \mathbf{h} \in \mathbb{R}^d - \{\mathbf{0}\},\ \mathbf{h}^T\nabla^2 m\mathbf{h} \geq 0$). Cela revient à dire que m est nécessairement *localement convexe* en tout minimiseur. Cette condition nécessaire est en fait presque une équivalence puisqu'on a la condition suffisante suivante : si $\nabla m(\mathbf{x}) = 0$ et si la matrice Hessienne $\nabla^2 m(\mathbf{x})$ est définie positive (i.e. $\forall \mathbf{h} \in \mathbb{R}^d - \{\mathbf{0}\},\ \mathbf{h}^T\nabla^2\mathbf{h} > 0$) alors $\mathbf{x}$ est un minimiseur local de m. Cela revient à dire qu'il est suffisant que m soit *localement strictement convexe* en $\mathbf{x}$ pour que $\mathbf{x}$ soit un minimiseur local de m.

Lorsque la fonction à optimiser est strictement convexe sur $\mathbb{R}^d$: le minimiseur existe et est unique. C'est donc là un cas particulier très favorable à l'optimisation globale !

13.1.2 Méthodes locales non-contraintes

Maintenant que nous avons fait quelques rappels sur les outils de base de l'optimisation numérique, nous proposons de passer en revue une sélection de méthodes d'optimisation locale non-contrainte parmi les plus utilisées et/ou représentatives de leur catégorie. Nous les avons rangées par ordre décroissant du nombre de dérivées utilisées car certaines méthodes d'ordre 1 découlent de méthodes d'ordre 2.

Les méthodes présentées ici sont itératives, en ce sens qu'elles partent d'un point $\mathbf{x}_0 \in D$ arbitrairement choisi et selectionnent ensuite une suite finie de points $\{\mathbf{x}_k \in D\}_{k\in[1,N]}$ ($N \in \mathbb{N}$) successivement dans le temps et de manière automatique, en utilisant à chaque étape k l'information disponible sur la fonction m et ses dérivées au point courant $\mathbf{x}_k$ pour trouver le prochain point à visiter $\mathbf{x}_{k+1}$.

Méthodes d'ordre 2

Méthode de Newton :

La méthode de Newton consiste à chaque itération à chercher le minimum de la forme quadratique osculatrice à m. Considérons à $\mathbf{x}$ fixé l'approximation quadratique de m

$$m(\mathbf{x}) \approx Q_k(\mathbf{x}) := m(\mathbf{x}_k) + \nabla m(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T\nabla^2 m(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) \quad (13.6)$$

$$\text{et le gradient de } Q_k : \nabla Q_k(\mathbf{x}) = \nabla m(\mathbf{x}_k) + \nabla^2 m(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) \quad (13.7)$$

La direction de Newton à l'étape $k+1$ est par définition comme le point critique de Q_k :

$$(\nabla Q_k(\mathbf{x}_{k+1}) = 0) \Rightarrow \left(\mathbf{x}_{k+1} = \mathbf{x}_k - [\nabla^2 m(\mathbf{x}_k)]^{-1}\nabla m(\mathbf{x}_k)\right) \quad (13.8)$$

et on obtient ainsi la suite des itérés de la méthode de Newton au prix du calcul puis de l'inversion d'une matrice hessienne à chaque étape. Le coût computationnel important se justifie par une grande efficacité dans certaines conditions. On peut en effet montrer que la méthode de Newton a une vitesse de convergence quadratique lorsque l'on part d'un point où la Hessienne est définie positive (i.e. au voisinage d'un minimum local). En revanche, la méthode est nettement moins intéressante voire inefficace (car succeptible de diverger) lorsque la fonction n'est pas localement strictement convexe au point d'initialisation. Différentes variantes ont été développées pour pallier ce problème. On présente ci-après deux méthodes parmi les plus reconnues pour rendre plus robuste la méthode de Newton.

Méthodes de Levenberg-Marquardt et des régions de confiance :

La méthode de Levenberg-Marquardt part du constat que l'on peut rendre $\nabla^2 m(\mathbf{x}_k)$ définie postive en lui ajoutant λI_d pour $\lambda > 0$ suffisamment grand. On obtient alors des itérés modifiés à partir de la méthode de Newton

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\nabla^2 m(\mathbf{x}_k) + \lambda I_d]^{-1}\nabla m(\mathbf{x}_k) \quad (13.9)$$

Le choix du coefficient λ joue un rôle majeur dans le déroulement de la méthode. Lorsque λ est grand, la matrice identité écrase l'influence de la Hessienne et l'on se retrouve dans la direction du gradient, avec un pas dont la norme décroît en $\frac{1}{\lambda}$. Lorsque λ est très petit, on retombe sur la méthode de Newton. Les situations intermédiares fournissent des directions composites entre direction de Newton et gradient, qui peuvent donner des résultats de qualité très variable. La méthode de Levenberg-Marquardt repose sur une adaptation de λ à chaque étape k en fonction du résultat de l'étape précedente. Si $m(\mathbf{x}_k) < m(\mathbf{x}_{k-1})$ on se rapproche de la direction de Newton en diminuant λ ; sinon, on augmente λ de manière à prendre une direction plus proche de la direction classique de descente (opposée au gradient). Dans ce dernier cas, la diminution de la longueur du pas (i.e. la norme $||m(\mathbf{x}_{k+1}) - m(\mathbf{x}_k)||$) ralentit considérablement l'algorithme.

La méthode des régions de confiance (*trust regions*) est une autre variante de la méthode de Newton. La prise en charge des itérations où la Hessienne n'est pas définie positive est assurée d'une manière alternative à celle de Levenberg-Marquardt. Au lieu de "biaiser" la Hessienne et de continuer à chercher le point critique d'une certaine forme quadratique (qui n'est pas vraiment Q_k), on s'intéresse à la minimisation de Q_k sous contrainte de rester dans un certain voisinage de $\mathbf{x}_k$:

$$\begin{cases} \mathbf{x}_{k+1} = \arg\min Q_k(\mathbf{x}) \\ \quad s.c.\ ||\mathbf{x} - \mathbf{x}_k|| \leq \Delta_k \end{cases} \tag{13.10}$$

La contrainte Δ_k (si elle est bien reglée) n'agit pas lorsque l'on est effectivement au voisinage d'un minimum local. Si en revanche la Hessienne n'est pas définie positive, la solution de l'eq. 13.10 est sur la contrainte ($||\mathbf{x}_{k+1} - \mathbf{x}_k|| = \Delta_k$) et on force ainsi $\mathbf{x}_{k+1}$ à s'éloigner des zones non-prometteuses (d'autant plus que Δ_k est grand).

Méthodes d'ordre 1

Deepest descent :

C'est la point de départ des méthodes de gradient. Comme expliqué plus haut, la direction opposée au gradient est systématiquement une direction de descente. La méthode de la plus profonde descente (*deepest descent*) repose ainsi sur la construction d'une suite d'itérés selon la relation

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla m(\mathbf{x}_k) \tag{13.11}$$

où $\alpha_k > 0$ est un coefficient à déterminer à chaque itération. La subtilité de la méthode repose sur le calcul des $\alpha_k > 0$ optimaux, i.e. permettant à chaque itération de descendre

le plus possible sur la droite de descente dirigée par $-\nabla m(\mathbf{x}_k)$. On obtient les α_k en résolvant le problème monodimensionnel suivant

$$\alpha_k = \arg\min_{\alpha>0} \{m(\mathbf{x}_k - \alpha \nabla m(\mathbf{x}_k))\} \tag{13.12}$$

Malgré la garantie qu'une amélioration est réalisée à chaque itération lorsque $\nabla m(\mathbf{x}_k) \neq 0$, cet algorithme n'est aujourd'hui guère employé en pratique. La descente (vers un point critique et pas nécessairement un minimum) peut devenir extrêmement lente lorsque l'on explore des régions de l'espace où la Hessienne est mal conditionnée. On peut en effet montrer que les gradients successifs de l'algorithme sont orthogonaux, ce qui peut occasionner des phénomènes de "zigzags" entre itérés successifs.

Méthodes de gradient conjugué :

Les méthodes de gradient conjugué consistent à rechercher des directions de descente qui tiennent à la fois compte du gradient et des directions précédemment empruntées. Après une première itération de plus profonde descente, on va chercher chaque direction $\mathbf{d}_{k+1}$ ($k \geq 1$) comme combinaison linéaire du gradient $\nabla m(\mathbf{x}_{k+1})$ et de la direction $\mathbf{d}_k$

$$\mathbf{d}_{k+1} = -\nabla m(\mathbf{x}_{k+1}) + \beta_k \mathbf{d}_k \tag{13.13}$$

Le déplacement vers le point suivant se fait alors dans la direction $\mathbf{d}_{k+1}$ avec un pas α_{k+1} déterminé en résolvant un problème d'optimisation monodimensionnelle similaire à 13.12. Le choix des β_k a fait l'objet de nombreux développements, dont les algorithmes de *Fletcher-Reeves* (1964) et de *Polak-Ribière* (1971) sont les aboutissements les plus connus. Les directions y sont respectivement déterminées par

$$\beta_k^{FR} = \frac{||\nabla m(\mathbf{x}_{k+1})||^2}{||\nabla m(\mathbf{x}_k)||^2} \text{ et } \beta_k^{PR} = \frac{\langle \nabla m(\mathbf{x}_{k+1}) - \nabla m(\mathbf{x}_k), \nabla m(\mathbf{x}_{k+1}) \rangle}{||\nabla m(\mathbf{x}_k)||^2} \tag{13.14}$$

La seconde est réputée d'autant meilleure que la fonction m est différente d'une fonction quadratique. Dans un cas comme dans l'autre, il est recommandé de s'assurer régulièrement au cours des algorithmes que la norme des gradients successifs reste suffisamment grande pour éviter les problèmes numériques (Cf. dénominateur de l'eq. 13.14).

Méthodes de quasi-Newton :

Ces méthodes ont pour vocation d'approcher la méthode de Newton tout en faisant l'économie des calculs associés au calcul et l'inversion des matrices hessiennes successives

$\nabla^2 m(\mathbf{x}_k)$ ($k \in [1, N]$). La clef des algorithmes DFP et BFGS (aujourd'hui implémentés dans la plupart des logiciels de calcul scientifique) est de se baser sur l'approximation

$$\mathbf{x}_k - \mathbf{x}_{k-1} \approx [\nabla^2 m(\mathbf{x}_k)]^{-1}(\nabla m(\mathbf{x}_k) - \nabla m(\mathbf{x}_{k-1})) \tag{13.15}$$

pour construire une suite de matrices symétriques S_k qui puissent jouer le rôle de l'inverse de la Hessienne dans l'éq. (13.8). On leur impose ainsi la *condition de quasi-Newton*

$$\mathbf{x}_k - \mathbf{x}_{k-1} = S_k(\nabla m(\mathbf{x}_k) - \nabla m(\mathbf{x}_{k-1})) \tag{13.16}$$

et on construit la suite des $\mathbf{x}_k$ par recherche linéaire dans les directions $\mathbf{d}_k = -S_k \nabla m(\mathbf{x}_k)$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k S_k \nabla m(\mathbf{x}_k) \tag{13.17}$$

où les α_k sont optimisés tout comme dans l'eq. 13.12. Remarquons que la condition de de quasi-Newton 13.16 ne permet pas de définir la suite des S_k de manière univoque. Le principe des fameux algorithmes DFP et BFGS est de choisir une point de départ $\mathbf{x}_0$ et une matrice d'initialisation définie positive (notée S_0 pour DFP et $\hat{S}_0$ pour BFGS), puis de mettre à jour les S_k (resp. $\hat{S}_k$) à chaque itération.

L'algorithme DFP (*Davidon-Fletcher-Powell*, 1971) repose sur la mise à jour suivante

$$S_{k+1} = S_k + \frac{\delta_k \delta_k^T}{\langle \delta_k, g_k \rangle} - \frac{S_k g_k g_k^T S_k}{\langle S_k g_k, g_k \rangle} \tag{13.18}$$

$$\text{où } \begin{cases} \delta_k = \mathbf{x}_{k+1} - \mathbf{x}_k \\ g_k = \nabla m(\mathbf{x}_{k+1}) - \nabla m(\mathbf{x}_k) \end{cases} \tag{13.19}$$

L'algorithme BFGS (*Broyden-Fletcher-Goldfarb-Shanno*, 1970)

$$\hat{S}_{k+1} = \hat{S}_k + \left[1 + \frac{\langle \hat{S}_k g_k, g_k \rangle}{\delta_k, g_k}\right] \frac{\delta_k \delta_k^T}{\langle \delta_k, g_k \rangle} - \frac{\delta_k g_k^T \hat{S}_k + \hat{S}_k g_k \delta_k^T}{\langle \delta_k, g_k \rangle} \tag{13.20}$$

où l'on a conservé les notations de l'eq. 13.19. BFGS fait toujours aujourd'hui partie des algorithmes jugés les plus efficaces. Il est employé par défaut dans la résolution de nombreux problèmes d'optimisation locale non-contrainte, et on trouve même parfois des versions de BFGS d'ordre 0 où les gradients sont approchés par différences finies.

Méthodes d'ordre 0

Simplexe de Nelder-Mead :

Appelée aussi *méthode du polytope*, la méthode du simplexe de Nelder-Mead (1965)[2] consiste à faire évoluer dans l'espace des variables (de dimension d) une figure formée de $d+1$ points $\{x_1, ..., x_{d+1}\}$ (*simplexe* ou "hyper-triangle", i.e. enveloppe convexe des $d+1$ points) en cheminant vers le minimum de la fonction m. Elle se base exclusivement sur des évaluations de m en certains sommets des simplexes successifs. La transition du simplexe courant vers sa mise à jour se fait à chaque itération via diverses opérations géométriques. La méthode nécessite la spécifications de quatre paramètres : les coefficients de *réflection* ($\rho > 0$), d'*expansion* ($\chi > 1$)[3], de *contraction* ($\gamma \in]0, 1[$), et de *réduction* ($\sigma \in]0, 1[$). Le déroulement d'une itération se fait de la manière suivante :

1. **Ordonner les** $d+1$ **points** de manière à avoir $y(x_1) \leq ... \leq y(x_{d+1})$
2. **Calculer le point de réflection** $x_r = \bar{x} + \rho(\bar{x} - x_{d+1})$, où $\bar{x} = \frac{1}{d}\sum_{j=1}^{d} x_j$ est l'isobarycentre des d meilleurs points. Calculer $y_r = y(x_r)$. Si $y(x_1) \leq y_r < y(x_d)$, accepter le point de réflection x_r et terminer l'itération.
3. **Calculer le point d'expansion**. Si $y_r < y(x_1)$, calculer $x_e = \bar{x} + \chi(x_r - \bar{x})$ et évaluer $y_e = y(x_e)$. Garder entre x_e et x_r celui dont l'image par y (y_e ou y_r) est la plus basse. Terminer l'itération.
4. **Calculer un point de contraction**. Si $y_r \geq y(x_d)$, faire une contraction entre $\bar{x}$ et le meilleur de x_r et x_{d+1} selon la régle suivante :
   (a) *contraction extérieure :* si $y(x_d) \leq y_r < y(x_{d+1})$, calculer $x_c = \bar{x} + \gamma(x_r - \bar{x})$ et évaluer $y_c = y(x_c)$. Si $y_c \leq y_r$, accepter x_c et terminer l'itération. Sinon, aller en 5.
   (b) *contraction intérieure :* si $y_r \geq y(x_{d+1})$, calculer $x_{cc} = \bar{x} - \gamma(\bar{x} - x_{d+1})$ et évaluer $y_{cc} = y(x_{cc})$. Si $y_{cc} < y_{d+1}$, accepter x_{cc} et terminer l'itération. Sinon, aller en 5.
5. **Réduire le simplexe.** Evaluer y aux d points $v_i = x_1 + \sigma(x_i - x_1)$ $(i = 2, ..., d+1)$. Les sommets (non-ordonnés) du simplexe suivant sont $\{x_1,\ v_2,\ ...,\ v_{d+1}\}$.

Cette méthode est codée par défaut comme algorithme d'ordre 0 dans bon nombre de logiciels de calcul (c.f. la célèbre fonction *fminsearch* de MatLab ™).

[2] à ne pas confondre avec la célèbre *méthode du simplexe*, développée par Dantzig pour la résolution de problèmes de programmation linéaire.

[3] On a de plus la condition $\chi > \rho$. On choisit généralement les valeurs $\rho = 1$, $\chi = 2$, $\gamma = \frac{1}{2}$, $\sigma = \frac{1}{2}$.

13.1.3 Méthodes globales non-contraintes

Algorithmes stochastiques (*meta-heuristiques*)

Recherche aléatoire :

Une possibilité lorsque l'on veut approcher le minimum d'une fonction m dont on ne connaît pas les gradients (voire une fonction qui n'en possède pas) est d'évaluer m en un grand nombre de points choisis au hasard (le plus souvent suivant une loi uniforme sur le domaine de définition de m) et de retenir celui d'entre eux dont l'image par m est la plus basse. On appellera ici *population* l'ensemble des points choisis. Il est bien évident que le résultat d'un tel mode opératoire dépend du bon remplissage de l'espace par la population, et par là même de la taille de cette dernière. Or, lorsque la dimension du problème ($d \in \mathbb{N}$) augmente, le nombre de points nécessaire à un remplissage de l'espace de qualité constante croît exponentiellement en d (on peut penser à l'exemple —non aléatoire— du passage d'une grille en $2D$ à une grille en $3D$). La recherche aléatoire exposée ci-dessus a l'avantage d'être extrêmement simple (!). Cela dit, elle n'est généralement utilisée que "faute de mieux" et on essayera dès que possible de raffiner cette approche.

Recuit simulé :

L'algorithme du recuit simulé (Metropolis) consiste à construire une suite de points (dits "*états*") dont l'image par m ("l'*enérgie*") s'approche du minimum recherché. Le terme de "recuit simulé" provient d'un problème de métallurgie : la configuration d'énergie minimale d'un métal ne peut être obtenue qu'au prix d'un refroidissement très lent, avec d'éventuelles phases de léger recuit. Un refroidissement trop rapide peut empêcher les atomes de s'agencer dans une configuration idéale, et ainsi entraîner la production d'un métal en état métastable (i.e. un minimum local d'énergie). L'idée clef du recuit simulé est de remonter de temps en temps la température afin de faciliter la mobilité des atomes. Elle a été adaptée en termes mathématiques et a permis la résolution de nombreux problèmes concrets d'optimisation globale. A chaque itération, on fait une recherche aléatoire (la population est réduite à un point) dans le voisinage du point précédent $\mathbf{x}_n$. Si le nouveau point $\mathbf{x}_{new}$ est meilleur que le précédent, il devient le nouveau point courant $\mathbf{x}_{n+1}$. Dans le cas contraire, on prend une décision aléatoire : on fait un tirage d'une variable aléatoire ζ uniforme sur $[0,1]$. Si $\zeta \leq e^{-\frac{y(x_{new})-y(x_n)}{T_n}}$ on pose quand même $\mathbf{x}_{n+1} = \mathbf{x}_{new}$, où T_n est un paramètre (appelé "*température*"). Si $\zeta > e^{-\frac{y(\mathbf{x}_{new})-y(\mathbf{x}_n)}{T_n}}$, on stationne en posant $\mathbf{x}_{n+1} = \mathbf{x}_n$. Les règles d'évolution de la

température pour obtenir des résultats de convergence les plus rapides ont fait l'objet de nombreuses recherches en mathématiques dans les années 1980 et 1990. Nous renvoyons aux travaux d'Olivier Catoni et/ou de Christian Mazza pour plus de détails.

Algorithmes génétiques :

Les algorithmes génétiques sont des stratégies d'optimisation inspirées par les mécanismes naturels décrits dans la théorie de l'évolution de Darwin. Ils consistent à générer puis à faire évoluer une population de points $\mathbb{P} = \{\mathbf{p}^1, ..., \mathbf{p}^g\}$ de l'espace des entrées ($D \subset \mathbb{R}^d$) en leur appliquant à chaque itération différents opérateurs probabilistes (Cf. ci-dessous) destinés à faire apparaître dans la population des éléments dont l'image par m est aussi proche de l'optimum que possible. Les opérateurs de diversification classiques sont

- des mutations : on bruite les $\mathbf{p}^j$ ($j \in [1, g]$) (à des fins d'exploration)
- des croisements : on mélange les composantes (appelées *gènes*) de certains $\mathbf{p}^j$
- des selections : une fois la population augmentée d'éléments mutés et croisées, on en sélectionne une partie pour former la polulation de l'itération suivante.

La taille de la population, le nombre total de générations (ou le critère d'arrêt), les lois de probabilité des opérateurs de croisement et de mutation sont autant de réglages qui permettent d'obtenir des comportements différents pour l'algorithme considéré.
Les *stratégies évolutionnaires* ES(λ, μ) constituent une classe d'algorithmes génétiques de base, permettant d'illustrer simplement les notions de population et d'opérateurs de diversification. Le principe des algorithmes ES(λ, μ) est de passer au cours de chaque itération de λ points *parents* à $\mu \geq \lambda$ points *enfants* par des mutations probabilistes (souvent des perturbations gaussiennes). On selectionne alors les λ meilleurs parmi les μ enfants, et ils deviennent les parents de l'itération suivante.
Plus récemment (Novembre 2005), Nikolaus Hansen a proposé l'algorithme "CMA-ES" (*Covariance Matrix Adaptation Evolution Strategy*). CMA-ES est un type d'ES(λ, μ) avec mutation probabiliste multigaussienne, dans laquelle la matrice de covariance est estimée dynamiquement de manière à converger vers l'inverse de la matrice Hessienne (au point courant) de la fonction considérée. Ce dernier fait aujourd'hui partie des algorithmes d'optimisation évolutionnaires jugés prometteurs par la littérature spécialisée.

13.1.4 Récapitulatif de quelques fonctions d'optimisation en R

Nous donnons dans le tableau ci-dessous un résumé des principales fonctions et packages d'optimisation en langage R. Ce récapitulatif est à compléter et surtout à actualiser dans la mesure où les packages R existant évoluent, et que de nouveaux packages et fonctions

sont édités régulièrement.

Fonction {Package}	Contraintes	Méthode	Gradient	Comment.
Optimisation non-contrainte				
`optim`	-	Nelder-Mead	non	-
	-	BFGS	oui/non	-
	-	Gradient conj.	oui/non	-
	$a \leq x \leq b$	L-BFGS-B	oui/non	-
	-	Recuit simulé	non	opt. glob.
`nlm`				
`nlminb`				
`trust` {`trust` }	-	Régions de confiance	oui	$\nabla^2 m$
Optimisation contrainte				
`solve.QP`{`quadprog`}	$Ax \geq b$	Goldfarb- Idnani	non	m quad.
`constrOptim`	$Ax \geq b$	Nelder-Mead	non	-
	$Ax \geq b$	BFGS	oui	-
	$Ax \geq b$	Gradient conj.	oui	-
	$Ax \geq b$	Recuit simulé	oui	opt. glob.
`genoud` {`rgenoud`}	$g(x) \geq 0$	EAs & quasi-Newton	oui/non	opt. glob.
Pour information				
`optimize`	$a \leq x \leq b$	section dorée	non	uniq. $1D$

TAB. 13.1 – Une sélection de packages d'optimisation en langage R

13.2 Exemples d'optimisation sur métamodèle déterministe

On propose ici quelques illustrations de couplages entre métamodèles déterministes et méthodes d'optimisation classiques telles que celles présentées plus haut dans ce chapitre. Ces exemples sont loin d'être exhaustifs —on se restreint ici à trois cas particuliers— mais ils permettent de soulever certaines questions de portée générale : le danger d'une optimisation directe sur une surface de réponse polynômiale, la séparabilité des problèmes d'optimisation de fonctions additives, et une vision *duale* du krigeage (permettant une ré-interprétation de cette méthode en termes d'approximation par fonctions de base).

13.2.1 Optimisation de la partie déterministe d'une régression

En pratique, la régression linéaire (ou encore l'approximation linéaire, jouant ici un rôle équivalent) est souvent utilisée comme surface de réponse lorsque l'on a pas la possibilité ou simplement pas de bonne raison de faire autrement. Il est alors naturel de vouloir prolonger son utilisation à des fins d'optimisation. Par ailleurs, il ne faut pas oublier

que les méthodes d'optimisation usuelles remplacent localement la fonction objectif par un polynôme de degré 2. Cela n'est pas sans danger si la fonction est irrégulière ou multimodale comme on va le voir dans ce qui suit.

modèle linéaire de degré 2 sans interaction

Utiliser un modèle de régression linéaire de degré 2 sans intéraction pour approcher y revient à rechercher m sous la forme

$$m(\mathbf{x}) =< \mathbf{b}, \mathbf{x} > + \frac{1}{2} < \mathbf{x}, D\mathbf{x} > \text{, avec } D = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & \\ 0 & 0 & ... & \lambda_d \end{pmatrix} \tag{13.21}$$

où $\mathbf{b}$ et les λ_j ($j \in [1, d]$) sont choisis par moindres carrés. L'optimisation de ce type de métamodèle se révèle particulièrement simple grâce à la particularité suivante : m peut s'écrire sous forme d'une somme de fonctions monovariables. On a en effet :

$$m(\mathbf{x}) = \sum_{j=1}^{d} b_j x_j + \sum_{j=1}^{d} \lambda_j x_j^2 = \sum_{j=1}^{d} \left[b_j x_j + \lambda_j x_j^2 \right] := \sum_{j=1}^{d} \left[m_j(x_j) \right] \tag{13.22}$$

Minimiser m revient alors à minimiser une à une les d fonctions monovariables m_j, ce qui est extrêmement simplificateur (d optimisations en dimension 1 ont un coût computationnel nettement moins important qu'une optimisation en d dimensions). On parle dans un tel cas de *problème séparable*.

Nous proposons ci-dessous deux exemples : optimisation d'une spline en 1D (pour deux plans d'expériences : subdivisions de l'intervalle d'étude à 5 puis 20 points) et optimisation de la fonction de Branin-Hoo 2D. Pour chaque exemple, les approximations polynômiales de degré 2 sans interaction ont été réalisées avec la fonction *lm*, et les optimisations en utilisant la fonction *optim* (qui fait appel par défaut à l'algorithme du simplexe de Nelder-Mead).

On peut remarquer sur la fig. 13.1 que le minimiseur de l'approximation quadratique apprise en 5 points (à gauche) n'est ni un minimum global ni un local de la vraie fonction étudiée. De même en augmentant la taille du plan à 20 expériences, minimiser l'approximation quadratique ne donne pas une meilleure approximation de l'optimum réel. Nous examinons dans le chapitre 4 l'effet de réitérer en tenant compte à chaque étape des résultats d'optimisation obtenus aux étapes précédentes.

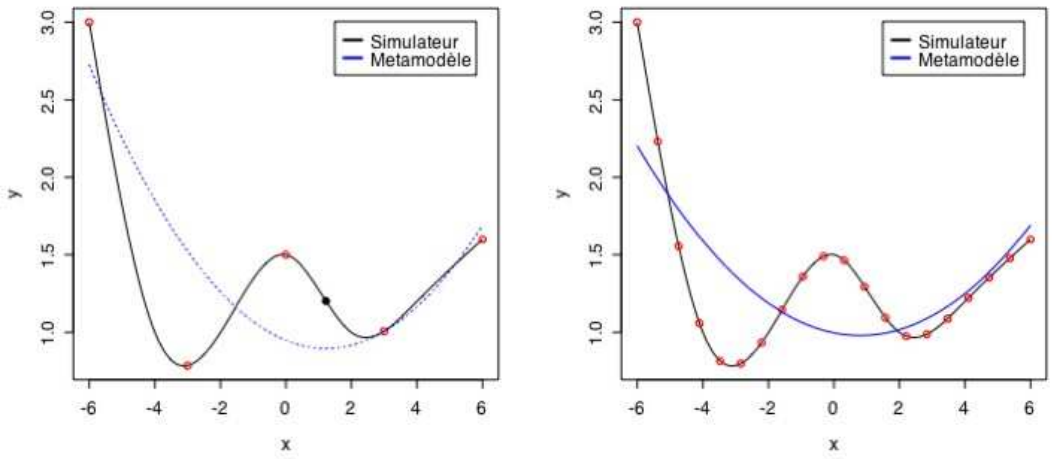


FIG. 13.1 – Optimisation de l'approximation quadratique (en bleu) d'une spline (en noir), en considérant comme plan d'expériences (en rouge) une subdivision régulière à 5 points (graphique de gauche) puis à 20 points (graphique de droite) de l'intervalle d'étude $[-6, 6]$. Le point noir représente le minimiseur de l'approximation quadratique.

L'approximation quadratique sans interaction de la fonction de Branin-Hoo présentée sur la fig. (13.2, à gauche) illustre sans détour le fait qu'un métamodèle déterministe mal spécifié peut conduire à une optimisation insatisfaisante : en recherchant à exploiter une tendance trop grossière, on peut négliger à tort les variations locales de la fonction et explorer trop hâtivement les bords du domaine d'étude. Notons que nous n'avons jusqu'à présent considéré une surface de réponse sans interaction. Voyons ci-dessous si la prise en compte des interactions permet d'améliorer significativement l'optimisation sur base d'approximation quadratique.

modèles linéaires de degré 2 avec interactions

On considère maintenant une approximation quadratique avec interactions. Le métamodèle peut alors s'écrire sous la forme :

$$m(\mathbf{x}) = < \mathbf{b}, \mathbf{x} > + \frac{1}{2} < \mathbf{x}, A\mathbf{x} > \tag{13.23}$$

où $A \in \mathcal{M}_d(\mathbb{R})$ est une matrice carrée à coefficients réels, non-nécessairement symétrique. Le problème de minimisation de m n'est alors plus séparable, et on a affaire à une *optimisation quadratique*. Le graphique ci-dessus (13.2, à droite) illustre à nouveau en

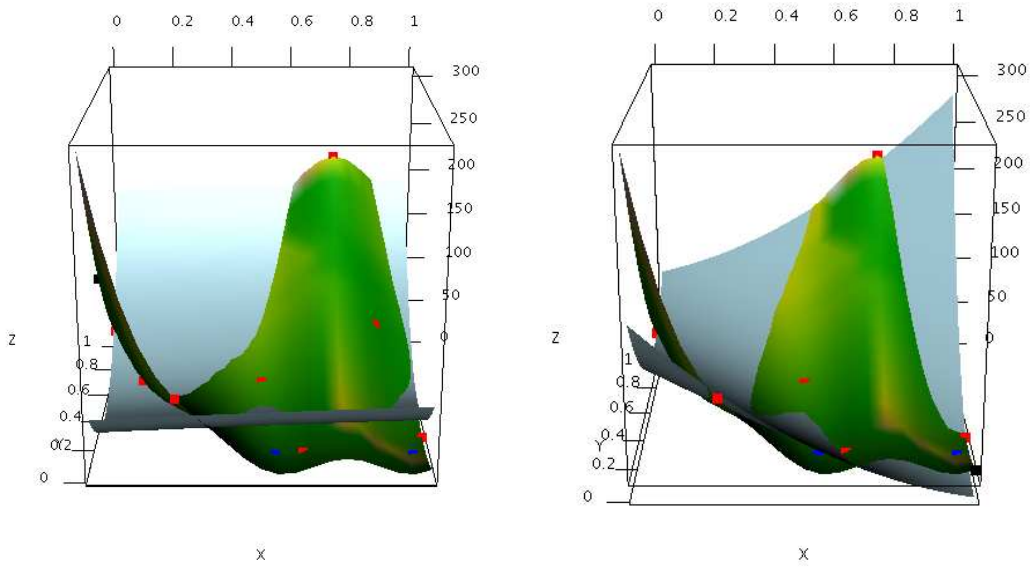


FIG. 13.2 – Optimisation d'approximations quadratiques (en bleu) de la fonction de Branin-Hoo (en noir), en considérant comme plan d'expériences (en rouge) un plan factoriel tourné à 9 points. Les graphiques de gauche et de droite correspondent respectivement à des approximations sans et avec interactions. Les point noirs représentent les minimiseurs des approximations, et les points bleus ceux de la fonction étudiée.

quoi l'optimisation sur la base d'un métamodèle déterministe est risquée et ici particulièrement sensible au choix de modèle : la prise en compte des interactions a fait passer l'optimiseur à un bord opposé du domaine. Le nouveau point candidat est meilleur que précédemment, mais le modèle ne semble pas encore assez précis pour que l'on trouve directement un vrai optimum local de la fonction.

13.2.2 Optimisation sur base de modèles additifs

Optimiser sur la base d'un métamodèle additif tel que

$$m(\mathbf{x}) = \sum_{j=1}^{d} m_j(x_j) \tag{13.24}$$

revient (à l'instar du cas d'un modèle linéaire de degré 2 sans interaction) à séparer l'optimisation multidimensionnelle en plusieurs problèmes de dimension 1, ou de petites dimensions dans le cas plus général où m s'exprime comme somme de fonctions dépendant de paquets de variables disjoints. Ces considérations sont potentiellement très utiles pour traiter les cas d'étude des chapitres 6 et 7.

13.2.3 Calcul du gradient d'une surface de réponse

On considère ici la moyenne de krigeage ordinaire comme surface de réponse :

$$m_{OK}(\mathbf{x}) = \bar{\mu} + \mathbf{c}(\mathbf{x})^T \Sigma^{-1}(\mathbf{Y} - \bar{\mu}\mathbb{1}_n) \tag{13.25}$$

où $\bar{\mu} = \frac{\mathbb{1}^T\Sigma^{-1}\mathbf{Y}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}$. A première vue, différencier cette expression peut sembler ardu. On se convainc aisément du contraire en remarquant que la seule source de dépendance en $\mathbf{x}$ provient en fait du vecteur $\mathbf{c}(\mathbf{x})$ des covariances entre la valeur recherchée et les observations faites au plan d'expériences. On peut faire apparaître ceci encore plus distinctement en notant $m_{OK}(\mathbf{x}) = \bar{\mu} + \mathbf{c}(\mathbf{x})^T\alpha = \bar{\mu} + \alpha^T\mathbf{c}(\mathbf{x})$ avec $\alpha = \Sigma^{-1}(\mathbf{Y} - \bar{\mu}\mathbb{1}_n)$ un vecteur de coefficients indépendants de $\mathbf{x}$. Remarquons au passage que cette dernière expression permet de faire le lien entre le krigeage et ce qui est communément appelé *interpolation par fonctions radiales de base*. Pour revenir au gradient, il vient que

$$\nabla m_{OK}(\mathbf{x}) = \nabla\mathbf{c}(\mathbf{x})^T\alpha = (\mathbf{Y} - \bar{\mu}\mathbb{1}_n)^T\Sigma^{-1}\nabla\mathbf{c}(\mathbf{x}) \tag{13.26}$$

Précisons ici que ce résultat n'est pas une invitation à utiliser directement des méthodes de gradient sur la moyenne de krigeage. Comme cela est illustré en détail dans ([Jon01]), cette méthode donnerait de forts mauvais résultats dans beaucoup de configurations (et mènerait bien souvent à retrouver comme solution un des points du plan d'expériences initial). Nous discutons dans les chapitres 4, 8, et 9 de cette thèse de différentes manières plus indiquées d'utiliser le krigeage pour l'optimisation.

Chapitre 14

Une pré-étude sur l'application d'EGO à des processus gaussiens

14.1 Protocole d'étude

Variante considérée dans cette étude

Le fil rouge de toute notre étude est la quantification des performances de l'algorithme EGO appliqué à des réalisations de processus (c'est à dire dans son contexte originel de conception, bien qu'il ne soit pas celui d'application). Il nous faut toutefois préciser deux points qui font que l'objet de notre étude n'est pas l'algorithme EGO *stricto sensu*, mais une version d'EGO sensiblement déconstruite.

- Tout d'abord, nous introduisons lors de l'étape 2 un contrôle supplémentaire du modélisateur : il est désormais possible de choisir les paramètres du variogramme selon différentes méthodes (vraisemblance, mais aussi variographie empirique ou autres), voire même de considérer des fonctions de corrélation différentes de celle de [JSW98] (on lui préfèrera d'ailleurs ici le corrélogramme cubique). La possibilité de contrôler les valeurs données au paramètre θ est primordial pour la suite de l'étude.
- La fin de l'algorithme est imposée par l'utilisateur (nombre d'itérations fixe).

Voici le résultat d'une application de 4 itérations de cet algorithme EGO à une réalisation de PSG de covariance cubique, de variance unité et de portée 0.3. Il est important de préciser que l'on a pris également une portée de 0.3 dans EGO.

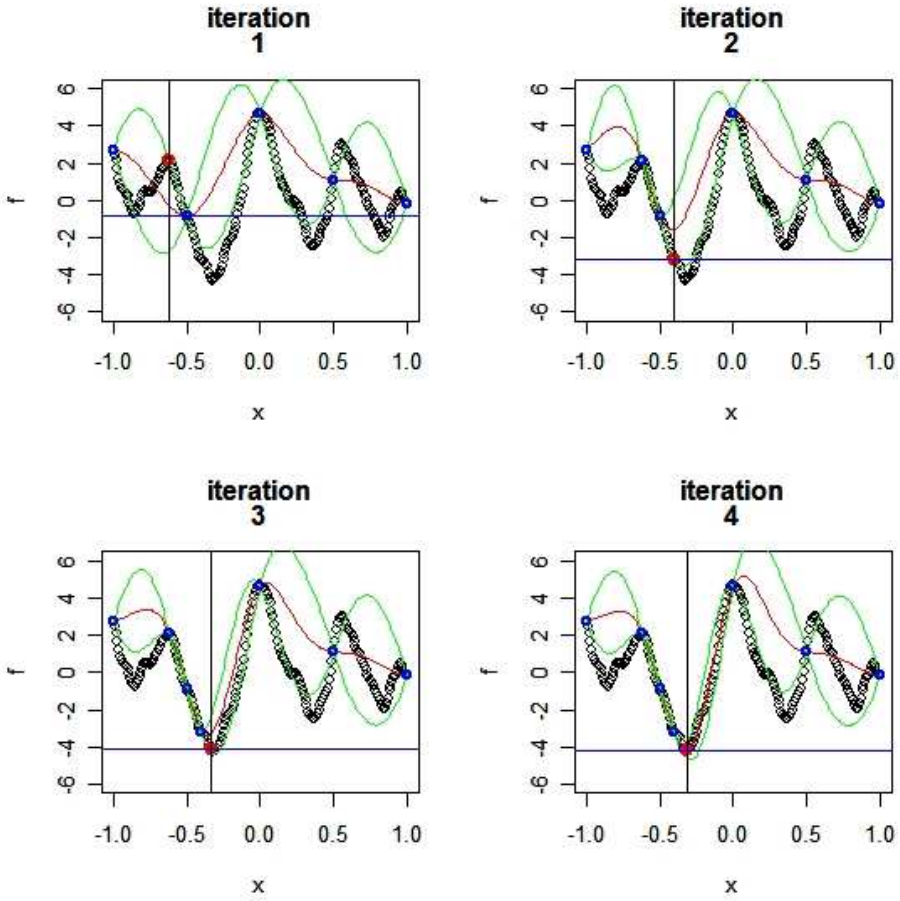


FIG. 14.1 – 4 itérations d'EGO à portée 0.3 connue

Comme on peut le constater sur la figure 14.1, la première itération de EGO illustre bien le compromis entre prédiction prometteuse et prédiction incertaine : le critère d'amélioration espérée nous entraîne sur un site dans le voisinage du minimiseur connu, mais suffisamment éloigné pour que l'incertitude de krigeage se fasse bien sentir.

Cadre d'étude des performances de l'algorithme (un peu de modélisation) : on cherche ici à optimiser en x une trajectoire du processus, $[Z_{\theta_{real}}(x)](\omega)$, où l'indice θ_{real} représente la portée du variogramme du processus considéré. $Z_{\theta_{real}}$ est gaussien stationnaire de moyenne nulle, de variance unitaire, et de variogramme cubique sans effet de pépite. Pour chaque état ω, EGO rend une suite de points $\{x^1_{\theta_{optim}}, ..., x^{n_{optim}}_{\theta_{optim}}\}(\omega)$, où l'in-

dice θ_{optim} représente la valeur de la portée prise par l'algorithme lors de l'application du Krigeage Ordinaire. Comme nous l'avons vu dans la section précédente, nous relâchons la contrainte $\theta_{optim} = \theta^*_{MV}$ de l'algorithme originel, et autorisons ce paramètre à prendre tout valeur souhaitée. Voici la notation que nous avons adoptée pour décrire l'ensemble des positions visitées par EGO (à paramètre θ_{optim}), appliqué à la réalisation $[Z_{\theta_{real}}(x)](\omega)$:

$$D_{\theta_{optim}}(Z_{\theta_{real}}(x)(\omega)) = \{x^1_{\theta_{optim}}, ..., x^{n_{optim}}_{\theta_{optim}}\}(\omega)$$

Notre but est d'étudier les performances de l'algorithme EGO sur des réalisations de processus stochastiques gaussiens, à la fois lorsque la portée est connue et lorsqu'elle est choisie de manière exogène. Nous définissons donc une variable aléatoire de performance indicée par deux variables (portée réelle θ_{real},et portée estimée (ou fixée a priori par l'utilisateur) θ_{optim}), simplement basée sur la distance entre le vrai minimum global de la fonction étudiée et le minimum global observé sur $X \cup D_{\theta_{optim}}(Z_{\theta_{real}}(x)(\omega))$:

$$\forall \omega \in \Omega \; [\Delta_{f^*}(\theta_{real}, \theta_{optim})](\omega) = \min_{x \in D_{\theta_{optim}}} (Z_{\theta_{real}}(x)(\omega)) - \min_{x \in I}(Z_{\theta_{real}}(x)(\omega))$$

Le critère déterministe que nous avons retenu pour rendre compte de la performance de l'algorithme appliqué au processus de portée θ_{real} avec une portée connue de θ_{optim} est l'espérance de Δ_{f^*}, c'est à dire la distance moyenne du résultat d'EGO au vrai optimum :

$$g(\theta_{real}, \theta_{optim}) = \mathbb{E}\left[\Delta_{f^*}(\theta_{real}, \theta_{optim})\right] = \mathbb{E}\left[\min_{x \in D_{\theta_{optim}}} (Z_{\theta_{real}}(x))\right] - \mathbb{E}\left[\min_{x \in I}(Z_{\theta_{real}}(x))\right]$$

Numériquement, nous estimeons g via la performance moyenne sur n_{simu} simulations :

$$\hat{g}(\theta_{real}, \theta_{optim}) = \frac{1}{n_{simu}} \sum_{s=1}^{n_{simu}} [\min_{x \in D_{\theta_{optim}}} (z^s_{\theta_{real}}(x)) - \min_{x \in I}(z^s_{\theta_{real}}(x))]$$

Notre objectif dans la section suivante est d'étudier succintement la fonction $g(\theta, \theta)$ en fonction de θ et du nombre d'itérations de EGO.

14.2 Résultats expérimentaux

Nous proposons d'étudier les performances d'EGO en fonction du nombre d'itérations sur l'exemple de deux PG de variogramme cubique. La portée —qui est ici la même pour

la simulation et pour l'optimisation— est prise successivement de valeur 0.02 puis 0.1.

mode opératoire : pour chaque portée ainsi que pour chaque nombre d'itérations (n_{init}) considéré, nous avons réalisé $n_{optim} = 100$ simulations de PG de caractéristiques voulues. Insistons sur le fait que les simulations utilisées d'un nombre d'itérations à un autre ne sont pas les mêmes : cela explique les non-monotonies entre box-plots (Cf. figure suivante), qui sont en fait imputables à des fluctuations d'échantillonnage.

Faute de résultats théoriques sur la vitesse de convergence d'EGO, ces résultats empiriques donnent un premier aperçu de ses performances. Il faut bien garder à l'esprit pour l'interprétation de ces résultats que l'aléa joue sur Δ_{f^*} à deux niveau : sur la forme de la réalisation à optimiser, mais aussi sur l'écart à l'optimum que l'on peut déjà observer sur le plan d'expériences initial. Penser au cas extrême où l'optimum est atteint sur D.

Interprétation : Chaque boxplot permet de se faire une idée de la distribution empirique de Δ_{f^*} pour une portée et un nombre d'itérations donnés. Il apparaît clairement que ces distributions positives on tendance à se décaler vers 0 avec l'augmentation du nombre d'itérations : cela s'explique par le regain d'information occasionné par des itérations supplémentaires. Par ailleurs il y a une nette augmentation des performances de l'algorithme avec l'augmentation de la portée : c'est une conséquence triviale de l'applatissement des réalisations avec la portée. Il est en particulier à retenir la grande variabilité des performances de l'algorithme : en-dessous d'un certain "n_{init} seuil", la variabilité des résultats autour de la performance médiane semble peu diminuée par l'augmentation de n_{init}. Seule une tendance décroissante de la médiane apparaît distinctement, quoiqu'il paraisse difficile de conclure sur la forme analytique de sa décroissance sur la seule base de ces figures.

Pour une portée de 0.1, le graphique montre bien comment, pour un nombre d'itérations supérieur à 18, la médiane de l'écart à l'optimum est inférieure ou à peine supérieure à 0 : on observe que dans environ 50% des cas (voire plus), l'algorithme EGO atteint l'optimum global lorsque le nombre d'itérations dépasse le seuil de 18. Ce seuil recule considérablement sur les deux figures suivantes : en toute logique, des fonctions plus plates sont faciles à optimiser en un nombre moins conséquent d'itérations. A l'extrême, on obtient une fonction constante, optimisable à coup sûr en une itération !

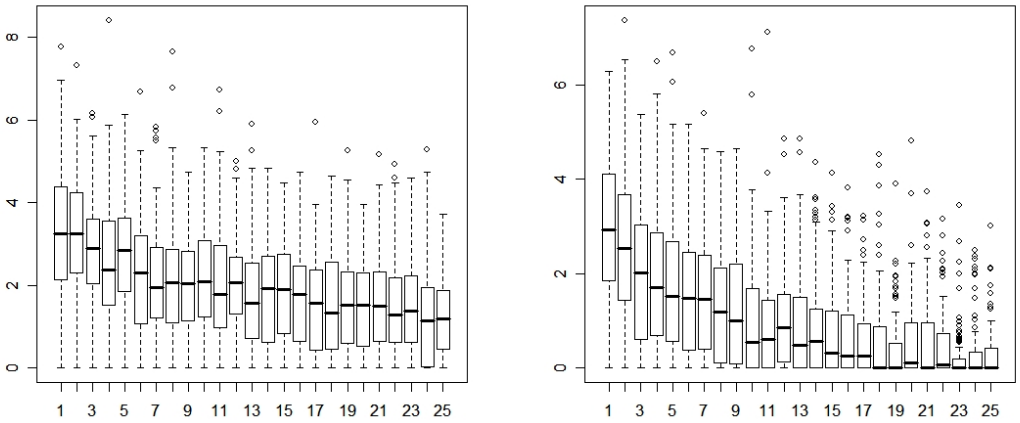


FIG. 14.2 – distribution de Δ_{f^*} en fonction de n_{init} pour des portées de 0.02 et 0.1

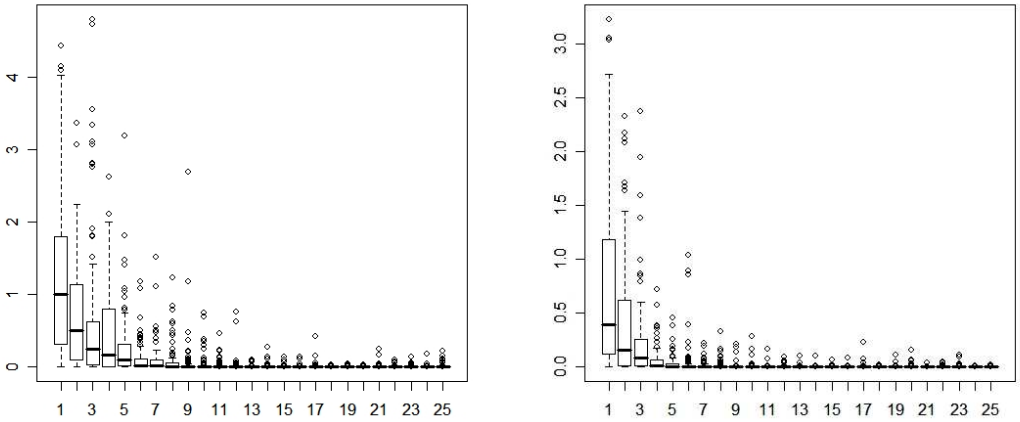


FIG. 14.3 – distribution de Δ_{f^*} en fonction de n_{init} pour des portées de 0.5 et 0.9

14.3 Sensibilité aux erreurs d'estimation de la covariance

Notre étude numérique s'articule autour de l'estimation de $g(\theta_{real}, \theta_{optim})$ pour trois valeurs de θ_{real} : 0.05, 0.1, 0.2. Le processus simulé prend ses valeurs dans l'intervalle [-1,1], et le plan d'expériences est une simple grille régulière monodimensionnel à 10 éléments. 6 Itérations d'EGO sont faites pour chaque simulation. Entre 100 et 1000 simulations ont été réalisées pour chaque valeur de θ_{real} et θ_{optim}.

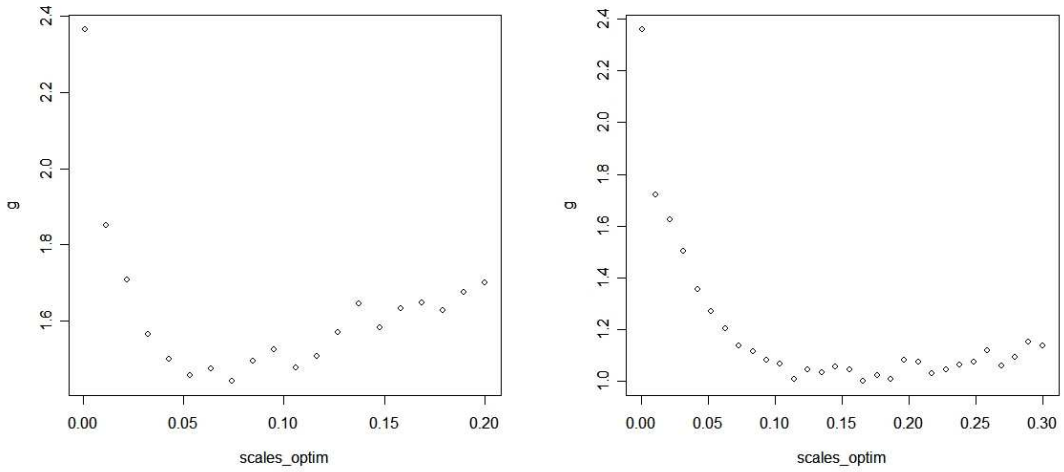


FIG. 14.4 – $\hat{g}(\theta_{real}, .)$ en fonction de θ_{optim} pour $\theta_{real} = 0.05$ *and* $\theta_{real} = 0.1$

Chacun de ces graphiques représente une coup de la fonction $\hat{g}(\theta_{real}, \theta_{optim})$ pour un valeur donnée de θ_{real}.

Le premier graphique a une tendance convexe, conforme à notre première intuition : les performances sont bonnes lorsque θ_{optim} prend des valeurs proches de θ_{real}, et deviennent d'autant moins bonnes que θ_{optim} s'éloigne de θ_{real}, inférieurement et supérieurement. Il est toutefois remarquable que cet effet est asymétrique : une sous-estimation de θ_{optim} (c'est à dire lorsque $\theta_{optim} \leq \theta_{real}$) a de bien pires conséquences sur les performances qu'une sur-estimation du même ordre de grandeur. Par ailleurs, la suite de points d'abscisses inférieures θ_{real} décroît de manière strictement monotone, alors que la suite de points au-dessus de θ_{real} oscille légèrement autour de sa tendance. Nous ne savons pas encore si ce curieux effet est du à l'inférence statistique (ici sur 1000 simulations) ou

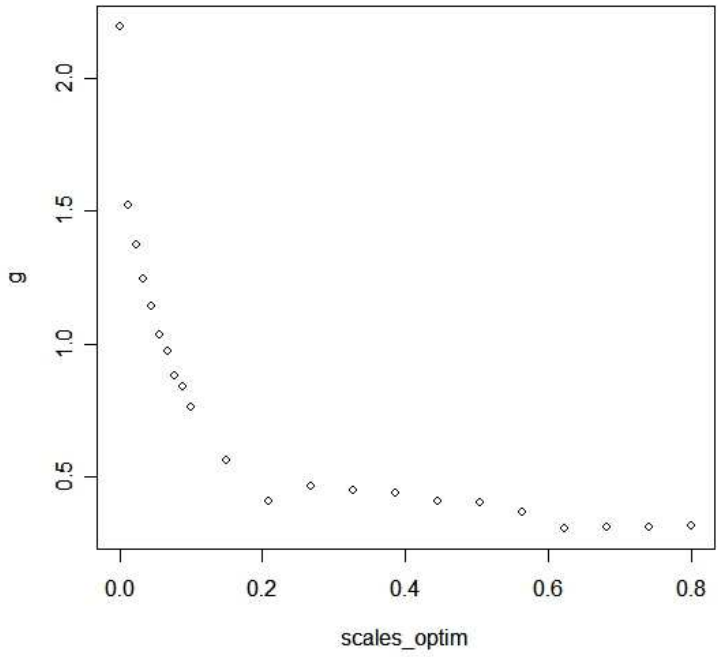


FIG. 14.5 – $\hat{g}(\theta_{real}, .)$ en fonction de θ_{optim} pour $\theta_{real} = 0.2$

bien reflète un phénomène déterministe.

Le second graphique a de quoi surprendre. Nous retrouvons la sequence de points nettement décroissante, pour $\theta_{optim} \leq \theta_{real}$. En revanche, le critère $\hat{g}$ continue pour $\theta_{optim} \geq \theta_{real}$ à décroître un temps, puis recroît très timidement lorsque θ_{optim} grandit. Ainsi, une surestimation de θ_{optim} semble-t-elle presque sans effet sur les performances de l'optimisation lorsque $\theta = 0.1$. Ce phénomène sera naturellement à prendre en considération dans nos futures estimation de portée à des fins d'optimisation.

Alors que les deux derniers graphiques sont issus des résultats de 1000 simulations, le troisième est le fruit de seulement 100 expériences numériques. Cela explique probablement au moins partiellement son allure chaotique. Nous observons sur ce graphique une amplification des effets observés sur le graphique à $\theta = 0.1$: g semble avoir tendance à diminuer lorsque $\theta_{optim} \geq \theta_{real}$ croît, particulièrement pour les valeurs extrêmes, i.e. $\theta_{optim} \cong 1$. Ces résultats méritent être confirmés par des expériences supplémentaires.

Bibliographie

[AALF97] R. Ardanuy Albajar and J.F. López Fidalgo, *Characterizing the general multivariate normal distribution through the conditional distributions*, Extracta Mathematicae **12** (1997), 15–18.

[AAP06] A. Antoniadis, U. Amato, and M. Pensky, *Wavelet kernel penalized estimation for non-equispaced design regression*, Statistics and Computing **16** (2006), 37–56.

[ABC92] A. Antoniadis, J. Berruyer, and R. Carmona, *Régression non linéaire et applications*, Economica, Paris, 1992.

[Abr97] P. Abrahamsen, *A review of gaussian random fields and correlation functions, second edition*, Tech. report, Norwegian Computing Center, 1997.

[Abt99] M. Abt, *Estimating the prediction mean squared error in gaussian stochastic processes with exponential covariance structure*, Scandinavian Journal of Statistics **26** (1999), 563–578.

[AF04] M. Avalos Fernandez, *Modèles additifs parcimonieux*, Ph.D. thesis, Université de Technologie Compiègne, 2004.

[AN00] S. Amari and H. Nagaoka, *Transactions of mathematical monographs : methods of information geometry*, vol. 191, Oxford University Press, 2000.

[Aro50] N. Aronszajn, *Studies in partial differential equations, report* 11 *: theory of reproducing kernels*, Tech. report, Harvard University, Division of Engineering Sciences, 1950.

[AW98] M. Abt and W. J. Welch, *Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes*, The Canadian Journal of Statistics **26** (1998), 127–137.

[BA98] K.P. Burnham and D.R. Anderson, *Model selection and multimodel inference*, Springer, 1998.

[Bai05] S. Baillargeon, *Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations*, Master's thesis, Université Laval, 2005.

[Bel01] J. Bellissard, *Mécanique statistique des systèmes hors d'équilibre, cours de master 2*, 2001.

[Bre94] L. Breiman, *Bagging predictors*, Tech. Report 421, Department of Statistics, University of California at Berkeley, 1994.

[BTW07] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Aggregation for gaussian regression*, Annals of Statistics **Vol. 35** (2007), no. 4, 1674–1697.

[BWG+01] C.A. Baker, L. T. Watson, B. Grossman, W. H. Mason, and R. T. Haftka, *Parallel global aircraft configuration design space exploration*, Practical parallel computing (2001), 79–96.

[Car] L. Carraro, *Notes sur le cours "interpolation et approximation", master de mathématiques appliquées de l'université jean monnet, saint-etienne.*

[Car03a] P.C. Caragea, *Approximate likelihoods for spatial processes*, Ph.D. thesis, Chapel Hill, 2003.

[Car03b] L. Carraro, *Introduction à la régression*, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2003.

[CG47] H. Cartan and R. Godement, *Théorie de la dualité et analyse harmonique dans les groupes abéliens localement compacts*, Annales Scientifiques de l'Ecole Normale Supérieure **64** (1947), 79–99.

[Chi04] J.-P. Chilès, *La modelisation geostatistique de la variabilite spatiale et ses applications*, Ph.D. thesis, Université Pierre et Marie Curie, 2004.

[Cia98] P.G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Dunod, 1998.

[Cre93] N.A.C. Cressie, *Statistics for spatial data*, Wiley series in probability and mathematical statistics, 1993.

[Cul94] J.-C. Culioli, *Introduction à l'optimisation*, Ellipses, 1994.

[Cur05] F.C. Curriero, *On the use of non-euclidean isotropy in geostatistics*, Tech. report, Johns Hopkins University, 2005.

[dCT06] R development Core Team, *R : A language and environment for statistical computing*, 2006.

[DHB07] D. Dupuy, C. Helbert, and A. Badea, *Livrable dice : Métamodèles, retour d'expérience.*, Tech. report, Ecole des Mines de Saint-Etienne, 2007.

[DM02] G. Dreyfus and J.-M. Martinez, *Réseaux de neurones*, Eyrolles, 2002.

[EGN06] M. Emmerich, K. Giannakoglou, and B. Naujoks, *Single-and multiobjective optimization assisted by gaussian random field metamodels*, IEEE Transactions on Evolutionary Computation **10 (4)** (2006).

[Eld92] John F. Elder, *Global rd optimization when probes are expensive : the grope algorithm*, IEEE International Conference on Systems, Man, and Cybernetics, 1992.

[ET93] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall/CRC, 1993.

[Fra08] J. Franco, *Planification d'expériences numériques en phase exploratoire pour la simulation de phénomènes complexes*, Ph.D. thesis, Ecole Nationale Supérieure de Mines de Saint-Etienne, 2008.

[GA] M. Gomez-Aparicio, *Compléments d'analyse : analyse hibertienne et analyse fonctionnelle*, Cours de Bachelor, Institut de Mathématiques de l'Université de Neuchâtel.

[GDB+07] D. Ginsbourger, D. Dupuy, A. Badea, O. Roustant, and L. Carraro, *A note on the choice and the estimation of kriging models for deterministic computer experiments*, Proceedings of the ENBIS-DEINDE workshop "Computer experiments versus Physical Experiments", Torino, 2007.

[GDB+09] ———, *A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments*, Applied Stochastic Models for Business and Industry **25** (2009), no. 2, 115–131.

[Gen01] M.G. Genton, *Classes of kernels for machine learning : A statistics perspective*, Journal of Machine Learning Research **2** (2001), 299–312.

[GG08] C. Gaetan and X. Guyon, *Modélisation et statistique spatiales*, Springer, 2008.

[GHC08] D. Ginsbourger, C. Helbert, and L. Carraro, *Discrete mixtures of kernels for kriging-based optimization*, Quality and Reliability Eng. Int. **24** (2008), no. 6, 681–691.

[GHSQ06] T. Goel, R. Haftka, W. Shyy, and N. Queipo, *Simultaneous use of multiple surrogates*, AIAA Journal (2006).

[GJ95] D. Geman and B. Jedynak, *An active testing model for tracking roads in satellite images*, Tech. report, Institut National de Recherches en Informatique et Automatique (INRIA), December 1995.

[GLRC07] D. Ginsbourger, R. Le Riche, and L. Carraro, *A multipoints criterion for parallel global optimization of deterministic computer experiments*, Non-Convex Programming 07, 2007.

[GMW81] P.E. Gill, W. Murray, and M.H. Wright, *Practical optimization*, Academic Press, 1981.

[Gor04] S. Goria, *Evaluation d'un projet minier : approche bayésienne et options réelles*, Ph.D. thesis, Ecole des Mines de Paris, 2004.

[GS92] P. Guttorp and P.D. Sampson, *Methods for estimating heterogeneous spatial covariance functions with environmental applications*, Tech. report, Department of Statistics, University of Washington, 1992.

[GWB98] P.W. Goldberg, C.K.I. Williams, and C.M. Bishop, *Regression with input-dependent noise : A gaussian process treatment*, Advances in Neural Information Processing Systems (Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, eds.), vol. 10, The MIT Press, 1998.

[HANM06] D. Huang, T.T. Allen, W. Notz, and R.A. Miller, *Sequential kriging optimization using multiple fidelity evaluations*, Sructural and Multidisciplinary Optimization **32** (2006), pp. 369–382 (14).

[HANZ06] D. Huang, T.T. Allen, W. Notz, and N. Zheng, *Global optimization of stochastic black-box systems via sequential kriging meta-models*, Journal of Global Optimization **34** (2006), 441–466.

[HGKK05] N. Henkenjohann, R. Göbel, M. Kleiner, and J. Kunert, *An adaptive sequential procedure for efficient optimization of the sheet metal spinning process*, Qual. Reliab. Engng. Int. **21** (2005), 439–455.

[HMM03] J. Hall, K. McKinnon, and T. Mayer, *Efficient global optimization : testing, reliability, and efficiency*, Tech. report, School of Mathematics, University of Edinburgh, 2003.

[HT91] T. Hastie and R. Tibshirani, *Generalized additive models*, Chapman and Hall, 1991.

[HT95] _____, *Generalized additive models*, Encyclopedia of Statistical Sciences (1995).

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, 2001.

[JJ93] M.I. Jordan and R.A. Jacobs, *Hierarchical mixture of experts and the em algorithm*, Tech. report, Massachusetts Institute of Technology, 1993.

[Jon01] D.R. Jones, *A taxonomy of global optimization methods based on response surfaces*, Journal of Global Optimization **21** (2001), no. 21, 345–383.

[Jou88] A. Journel, *Fundamentals of geostatistics in five lessons*, Tech. report, Stanford Center for Reservoir Forecasting, 1988.

[Jou02] A. Jourdan, *Approches statistiques des expériences simulées*, Revue de Statistiques Appliquées **50** (2002), 49–64.

[JPS93] D.R. Jones, C.D. Pertunen, and B.E. Stuckman, *Lipshitzian optimization without the lipshitz constant*, Journal of Optimization Theory and Application (1993), no. 79.

[JR89] A. G. Journel and M. E. Rossi, *When do we need a trend model in kriging ?*, Mathematical Geology **21** (1989), no. 7, 715–739.

[JSW98] D.R. Jones, M. Schonlau, and W.J. Welch, *Efficient global optimization of expensive black-box functions*, Journal of Global Optimization **13** (1998), 455–492.

[Kno05] Joshua Knowles, *Parego : A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems*, IEEE transactions on evolutionnary computation (2005).

[KO96] J.R. Koehler and A.B. Owen, *Computer experiments*, Tech. report, Department of Statistics, Stanford University, 1996.

[KPPB07] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, *Most likly heteroskedastic gaussian process regression*, Proceedings of the 24^{th} International Conference on Machine Learning, Corvallis, OR, 2007.

[Kra06] H. Kracker, *Methoden zur analyse von computerexperimenten mit anwendung auf die hochdruckblechumformung*, Master's thesis, Dortmund University, 2006.

[Kri51] D.G. Krige, *A statistical approach to some basic mine valuation problems on the witwatersrand*, J. of the Chem., Metal. and Mining Soc. of South Africa **52 (6)** (1951), 119139.

[KVB05] J.P.C. Kleijnen and W.C.M. Van Beers, *Robustness of kriging when interpolating in random simulation with heterogeneous variances : Some experiments*, European Journal of Operational Research **165** (2005), 826–834.

[Lé65] P. Lévy, *Processus stochastiques et mouvement brownien*, 1965.

[Lan65] S. Lang, *Algebra*, Addison-Wesley, Reading, Mass., 1965.

[Lar08] A. Largillier, *Introduction aux espaces de hilbert à noyau reproduisant*, Tech. report, Groupe de travail RKHS de l'université Jean Monnet et de l'Ecole des Mines de Saint-Etienne, 2008.

[LG06] J.F. Le Gall, *Intégration, probabilités et processus aléatoires*, Ecole Normale Supérieure de Paris, September 2006.

[Lin96] J.K. Lindsey, *Parametric Statistical Inference*, Oxford Science Publications, 1996.

[LLR04] Marco A. Luersen and Rodolphe Le Riche, *Globalized neldermead method for engineering optimization*, Computers and Structures **82** (2004), 2251–2260.

[LS05] R. Li and A. Sudjianto, *Analysis of computer experiments using penalized likelihood in gaussian kriging models*, Technometrics **47** (2005), no. 47, 111–120.

[LSC05] Q. V. Le, A. J. Smola, and S. Canu, *Heteroskedastic gaussian process regression*, Proceedings of the 22^{nd} International Conference on Machine Learning, Bonn, Germany, 2005.

[Ma05] Chunseng Ma, *Linear combinations of space-time covariance functions and variograms*, IEEE transactions on signal processing **53** (2005).

[Mal99] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1999.

[Mat63] G. Matheron, *Principles of geostatistics*, Economic Geology **58** (1963), 1246–1266.

[Mat69] ———, *Le krigeage universel*, Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau **1** (1969).

[Mat70] ———, *La théorie des variables régionalisées et ses applications*, Tech. report, Centre de Morphologie Mathématique de Fontainebleau, Ecole Nationale Supérieure des Mines de Paris, 1970.

[MdBMV94] A.M.J. Meijerink, H. de Brouwer, C.M. Mannaerts, and C.R. Valenzuela, *Introduction to the use of geographic information systems for practical hydrology*, ITC, 1994.

[MIvDV08] A. Marrel, B. Iooss, F. van Dorpe, and E. Volkova, *An efficient methodology for modeling complex computer codes with gaussian processes*, Computational Statistics and Data Analysis (2008).

[MK97] G.J. McLachlan and T. Krishnan, *The em algorithm and extensions*, Wiley series in probability and statistics, 1997.

[MM84] K.V. Mardia and R.J. Marshall, *Maximum likelihood estimation of models for residual covariance in spatial regression*, Biometrika **71** (1984), 135–46.

[MR07] J. Miss and Y. Richet, *La simulation Monte Carlo ; de la propagation des neutrons appliquée à la criticité*, Rapport scientifique et technique IRSN (2007), 158–164.

[MS04a] J.D. Martin and T.W. Simpson, *A monte carlo simulation of the kriging model*, 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY, August 30 - September 2, AIAA, AIAA-2004-4483., 2004.

[MS04b] ———, *A monte carlo simulation of the kriging model*, AIAA Journal (2004).

[MS05] ———, *Use of kriging models to approximate deterministic computer models*, AIAA Journal **43 (4)** (2005), 853–863.

[MS08] W.R. Jr. Mebane and J.S. Sekhon, *Genetic optimization using derivatives : The rgenoud package for r*, Journal of Statistical Software **to appear** (2008).

[MT86] R. Mneimné and F. Testard, *Groupes de lie classiques*, Hermann, 1986.

[O'H06] A. O'Hagan, *Bayesian analysis of computer code outputs : a tutorial*, Reliability Engineering and System Safety **91** (2006), no. 91, 1290–1300.

[Pac03] C.J. Paciorek, *Nonstationary gaussian processes for regression and spatial modelling*, Ph.D. thesis, Carnegie Mellon University, 2003.

[QHS+05] N. V. Queipo, R.T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P.K. Tucker, *Surrogate-based analysis and optimization*, Progress in Aerospace Sciences **41** (2005), 1–28.

[QVPH06] N.V. Queipo, A. Verde, S. Pintos, and R.T. Haftka, *Assessing the value of another cycle in surrogate-based optimization*, 11th Multidisciplinary Analysis and Optimization Conference, AIAA, 2006.

[RBCG08] A. Rakotomamonjy, F.R. Bach, S. Canu, and Y. Grandvalet, *Simplemkl*, Journal of Machine Learning Research **9** (2008), 2491–2521.

[Rip87] B.D. Ripley, *Stochastic simulation*, John Wiley and Sons, New York, 1987.

[RMC07] P. Renard, G. Mariethoz, and A. Communian, *Formation et modélisation des reservoirs*, Université de Neuchâtel, Mars 2007.

[Rob92] C.P. Robert, *L'analyse statistique bayesienne*, Economica, Paris, 1992.

[Rob96] D.J.S. Robinson, *A course in the theory of groups*, Springer Graduate texts in Mathematics, 1996.

[Rud77] W. Rudin, *Analyse relle et complexe*, 1975-1977.

[RW06] C.E. Rasmussen and K.I. Williams, *Gaussian processes for machine learning*, M.I.T. Press, 2006.

[RY91] D. Revuz and M. Yor, *Continuous martingales and brownian motion*, Springer-Verlag, 1991.

[Sas02] M.J. Sasena, *Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations*, Ph.D. thesis, University of Michigan, 2002.

[Sch97] M. Schonlau, *Computer experiments and global optimization*, Ph.D. thesis, University of Waterloo, 1997.

[Sch98] L. Schwartz, *Mathématiques pour la licence. algèbre.*, Dunod, 1998.

[Sch00] B. Schölkopf, *The kernel trick for distances*, Neural Information Processing Systems, 2000.

[Sch01] M. Schlather, *Simulation and analysis of random fields*, 2001, pp. 18–20.

[SPG02] M. Sasena, P. Papalambros, and P. Goovaerts, *Global optimization of problems with disconnected feasible regions via surrogate modeling*, AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimi, 2002.

[ST07] A. Samarov and A. Tsybakov, *Advances in statistical modeling and inference : Essays in honor of Kjell A. Doksum*, ch. 12 : aggregation of density estimators and dimension reduction, pp. 233–251, World Scientific, 2007.

[Ste99] M.L. Stein, *Interpolation of spatial data, some theory for kriging*, Springer, 1999.

[Swe80] T.J. Sweeting, *Uniform asymptotic normality of the maximum likelihood estimator*, The Annals of Statistics **8** (1980), no. 6, 1375–1381.

[SWJ97] M. Schonlau, W.J. Welch, and D.R. Jones, *A data-analytic approach to bayesian global optimization*, Proceedings of the A.S.A., 1997.

[SWMW89] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, *Design and analysis of computer experiments*, Statistical Science (1989), no. 4, 409–435.

[SWN03] T.J. Santner, B.J. Williams, and W.J. Notz, *The design and analysis of computer experiments*, Springer, 2003.

[Vap98] V.N. Vapnik, *Statistical learning theory*, Wiley-Interscience, 1998.

[Vaz05] E. Vazquez, *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*, Ph.D. thesis, Université Paris XI Orsay, 2005.

[VBK08] W.C.M. Van Beers and J.P.C. Kleijnen, *Customized sequential designs for random simulation experiments : Kriging metamodeling and bootstrapping*, European Journal of Operational Research **186** (2008), 1099–1113.

[VdV98] A.W. Van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[Ver00] J.-P. Vert, *Double mixture and universal inference*, Tech. Report DMA-00-15, Ecole Normale Superieure, 2000.

[Ver07] ———, *Kernel methods*, Centre for Computational Biology, Ecole des Mines de Paris, 2007.

[VVW09] J. Villemonteix, E. Vazquez, and E. Walter, *An informational approach to the global optimization of expensive-to-evaluate functions*, Journal of Global Optimization **44** (2009), no. 4, 509–534.

[Wac04] H. Wackernagel, *Géostatistique et assimilation séquentielle de données*, Ph.D. thesis, Université Pierre et Marie Curie, 2004.

[Wah90] G. Wahba, *Spline models for observational data*, Siam, 1990.

[Yao98] T. Yao, *Automatic covariance modeling and conditional spectral suimulation using Fast Fourier Transform*, Ph.D. thesis, Stanford Center for Reservoir Forecasting, 1998.

École Nationale Supérieure des Mines
de Saint-Étienne

N° d□ordre : **519 MA**

David GINSBOURGER

MULTIPLE METAMODELS FOR THE APPROXIMATION AND THE OPTIMIZATION OF MULTIVARIATE NUMERICAL FUNCTIONS

Speciality: Applied Mathematics

Keywords: Gaussian Processes, Global Optimization, Kernel Methods, Maximum Likelihood, Kriging, Mixtures of Experts, Synchronous Distributed Computing.

Abstract:

This dissertation takes place in the framework of design and analysis of computer experiments. More precisely, its main focus is on optimization strategies based on surrogate models of the objective function, or metamodels. Its principal motivation is to expose and strengthen existing works on Kriging-based optimization. Some relationships between different classical metamodels are adressed, and some light is shed on the versatility of Kriging and its suitability for sequential and parallel optimization.

After a detailed introduction to Kriging (end of part I), several tracks for the enrichment of this metamodel are proposed in part II. Part III is dedicated to some novelties in Kriging-based optimization, in particular concerning the integration of a mixture of metamodels or the parallelisation of evaluations for synchronous distributed computing.

École Nationale Supérieure des Mines
de Saint-Étienne

N° d'ordre : **519 MA**

David GINSBOURGER

MULTIPLES MÉTAMODÈLES POUR L'APPROXIMATION ET L'OPTIMISATION DE FONCTIONS NUMÉRIQUES MULTIVARIABLES

Spécialité: Mathématiques Appliquéees

Mots clefs : Processus Aléatoires Gaussiens, Optimisation Globale, Mélanges d'Experts, Maximum de Vraisemblance, Noyaux de Covariance, Krigeage, Calcul Distribué Synchrone.

Résumé :

Cette thèse s'inscrit dans la thématique de planification d'expériences numériques. Elle porte plus précisément sur l'optimisation de simulateurs numériques coûteux à évaluer, par des stratégies d'échantillonnage basées sur des représentations simplifiées du simulateur, *les metamodèles*. Une fois choisi un metamodèle parmi les familles existantes (polynômes, splines, modèles additifs, Krigeage, réseaux de neurones), on estime les paramètres du metamodèle. On dispose alors d'une représentation simplifiée du simulateur, que l'on pourra faire évoluer en fonction des informations apportées par de nouvelles évaluations. Etant donné qu'il est difficile de savoir a priori quel sera le type de metamodèle capable de guider au mieux un algorithme d'optimisation, une des motivations de ce travail est d'examiner comment une construction ad hoc de la structure du metamodèle, voire la prise en compte de plusieurs metamodèles, peuvent améliorer les méthodes d'approximation et les stratégies d'optimisation globale actuellement employées.

Cela soulève à la fois des questions mathématiques et statistiques de sélection de modèle (quelles familles de métamodèles considérer ? Comment estimer les termes de covariance et/ou de tendance d'un métamodèle de Krigeage, et selon quels critères les évaluer ? Comment prendre en compte certaines formes d'instationnarité dans la covariance de Krigeage que sont les symétries et la présence de bruits d'observation hétérogènes ?), de combinaison de modèles (Une fois un ensemble de metamodèles choisis, comment agrège-t-on les pseudo-informations qu'ils nous apportent ?), et de définition de critères décisionnels pour guider les évaluations au sein d'algorithmes d'optimisation (Comment paralléliser EGO ou des procédures similaires d'exploration sur base de Krigeage ?).