Utilisation d'approches probabilistes basées sur les critères entropiques pour la recherche d'information sur supports multimédia

Guilhem Coq

Université de Poitiers

THÈSE

pour obtenir le grade de

Docteur de l'Université de Poitiers

Spécialité : **Mathématiques**

préparée au **Laboratoire de Mathématiques et Applications de Poitiers**

dans le cadre de l'École Doctorale **Sciences Pour l'Ingénieur & Aéronautique SPI&A**

présentée et soutenue publiquement
par

Guilhem Coq

le 5 décembre 2008

Titre:

Utilisation d'approches probabilistes basées sur les critères entropiques pour la recherche d'information sur supports multimédia

Directeurs de thèse: **Marc Arnaudon, Christian Olivier**
Co-directeur de thèse: **Olivier Alata**

Jury

B. LAURENT	Professeur, INSA de Toulouse	Rapporteur
P. SIARRY	Professeur, univ. de Paris XII	Rapporteur
A. EL MATOUAT	Maître de Conférences, univ. du Havre	Examinateur
R. GRIBONVAL	Chargé de recherches INRIA, Rennes	Examinateur
M. ARNAUDON	Professeur, univ. de Poitiers	Directeur de thèse
C. OLIVIER	Professeur, univ. de Poitiers	Directeur de thèse
O. ALATA	Maître de Conférences, univ. de Poitiers	Co-directeur de thèse

Remerciements

C'est avec plaisir que j'entame ce manuscrit en remerciant les personnes qui ont contribué à son existence.

En premier lieu, mes pensées se tournent vers mes encadrants, Marc, Christian et Olivier. Il n'a sûrement pas été facile pour eux, comme pour moi, de mener à bien ce projet de recherche transverse entre deux communautés scientifiques. Leurs qualités d'écoute, de dialogue, d'abnégation, aussi bien que d'intransigeance, ont grandement contribué à me guider sur le chemin a priori glissant qu'est celui de la pluridisciplinarité.

Béatrice Laurent, professeur en statistiques, et Patrick Siarry, professeur dans le domaine du traitement du signal, m'ont fait l'honneur d'accepter d'être les rapporteurs de mon travail. Les échanges que j'ai eus avec eux ont contribué à corriger certaines erreurs ainsi qu'à modeler ce manuscrit pour le mettre en accord avec les exigences habituelles des communautés auxquelles il s'adresse. Ma reconnaissance va également vers Abdelaziz El-Matouat et Rémy Gribonval, une nouvelle fois représentant de leur domaine respectif, qui ont accepté de participer à mon jury.

Je me dois également de remercier certains responsables du laboratoire de mathématiques de Poitiers, Pierre Torasso et Aberrazak Bouaziz pour les principaux, qui m'ont permis d'effectuer mon troisième cycle malgré une traversée de déboires personnels et administratifs (dans l'ordre), dûs seulement à ma forte insouciance revendiquée.

Ma reconnaissance va également vers toutes les personnes qui contribuent à donner à ce laboratoire un charme certain empêchant la lassitude de s'y installer. Parmi celles-là, je cite Frédéric Bosio pour ses histoires drôles, ses chants dans les couloirs et sa moustache surprise, Patrice Tauvel pour sa présence immanquable et ses légendaires humeurs changeantes, une nouvelle fois Abderrazak Bouaziz pour son rire pincé et audible de mon bureau malgré la distance qui le sépare de la cafétéria, Anne Bertrand pour son franc-parler statistique : "sais-tu que boire du jus de carottes réduit de 47% les risques de cancer" a-t-elle lancé un jour, et, puisqu'il me faudra bien terminer cette liste, Alain Miranville pour sa grande générosité tant en termes d'idées sérieuses que de franches rigolades. Dans un style plus discret mais tout aussi important, on trouve entre ces murs Claude Quitté dont l'efficacité mathématique est seulement égalée par sa gentillesse et son caractère, Anthony Phan, vaillant défenseur aussi bien de ses opinions que de l'intérêt d'autrui, ainsi que Pol Vanhaecke dont la gentillesse rend tout problème agréable.

Le laboratoire de mathématiques de Poitiers prépare la future relève des mathématiques française en accueillant bon nombre de doctorants avec lesquels j'ai eu le plaisir de partager mes activités : Ariane, Anne, Céline, Khaoula, Mohamed, Idriss, Grégory, Patience, Sami, Toufic, Pol, Weisam, Armel et Koléhè (sans qui les moeurs politiques et les espaces conpaks n'auraient pas la même saveur) pour les plus anciens ; Caroline, Hélène, Sami, Gang et Willy pour les petits nouveaux. J'ai également une pensée pour Jocelyne Attab, Nathalie Marlet, Benoit Métrot, Nathalie Mongin et Brigitte Brault qui assurent un travail sans faille auprès de personnes qui écrivent beaucoup de choses étranges sur leurs tableaux. Enfin, j'avoue être confus d'oublier ici toutes les autres personnes qui font que, même après plus de cinq ans de fréquentation quotidienne, mes arrivées au laboratoire de mathématiques n'ont jamais été synonymes d'ennui.

Je n'oublierai pas non plus le soutien que m'a apporté mon entourage. Je suis heureux de commencer par Hélène dont l'aptitude à me supporter au quotidien est inespérée et le soutien sans faille. Elle a su remplir ces bientôt trois années de moments uniques (et de hamsters, plantes vertes, chaussures, maillots de bain, géraniums et autres trucs de filles), je lui fais confiance pour faire de même des suivantes.

La ville où j'ai grandi a vu naître des relations durables. Je pense en particulier à Jean-Marie dont les qualités d'ami n'ont jamais été démenties au cours de ces 25 dernières années. Je me remémore les innombrables jeux d'enfants, discussions pré-pubères ou post-bachelière ainsi que les débats enflammés sur la présence, indispensable, de ce coup précis de grosse caisse. Ma reconnaissance va également vers les grands Guillaume et Gaëtan, également musiciens de leurs états ayant laissé (de manière transitoire bien entendu) le taux de saturation de leur ampli sur la position 0. Je remercie également Damien dont le cynisme, l'insouciance apparente et l'intérêt pour (presque) toutes choses m'ont aidé à passer de forts belles années. Enfin, j'ai une pensée pour les amis qui ont compté mais que j'ai perdus de vue sans raison autre que la bête flemme de communiquer : Matthieu, Mickaël, Adrien et Grégory.

J'entend encore aujourd'hui résonner cette phrase : "de toute façon, son bac, il l'aura pas !". Prononcée par maman alors que je jouais aux jeux vidéos à quelques jours seulement de la date fatidique de la disserte de philo, elle est inoubliable. Sans elle, je ne serais certainement pas en train d'écrire ces remerciements, avec le bac en poche qui plus est, merci maman ! Mes pensées se tournent maintenant vers ma soeur Marie-Estelle et sa joyeuse collection de filles, mes nièces pour le coup : Eugénie, Marjolaine, Myriam et Rachel dans leur ordre d'apparition. La joie de vivre, les rires et le gros remue-ménage apporté par ces cinq personnes le temps d'une semaine de vacances, d'un week-end en famille, ou même d'une simple après-midi, suffisent à vous requinquer, à vous fatiguer, et à vous donner envie de recommencer dès le lendemain. J'ai envie de profiter de ces lignes pour encourager mes nièces à continuer à profiter de la vie comme elles le font aujourd'hui ainsi que pour féliciter ma soeur de tout ce qu'elle a accompli.

Je termine sur une note d'espoir en dédiant ce manuscrit à mon père Marcel qui, bien qu'il aurait préfére le tenir entre ses mains, en est certainement très fier.

Résumé

Les problèmes de sélection de modèles se posent couramment dans un grand nombre de domaines applicatifs tels que la compression de données ou le traitement du signal et de l'image. Un des outils les plus utilisés pour résoudre ces problèmes se présente sous la forme d'une quantité réelle à minimiser appelée critère d'information ou critère entropique pénalisé.

La principale motivation de ce travail de thèse est de *justifier* l'utilisation d'un tel critère face à un problème de sélection de modèles typiquement issu d'un contexte de traitement du signal. La justification attendue se doit, elle, d'avoir un solide fondement mathématique.

Nous abordons ainsi le problème classique de la détermination de l'ordre d'une autorégression. La régression gaussienne, permettant de détecter les harmoniques principales d'un signal bruité, est également abordée. Pour ces problèmes, nous donnons un critère dont l'utilisation est justifiée par la minimisation du coût résultant de l'estimation obtenue. Les chaînes de Markov multiples modélisent la plupart des signaux discrets, comme les séquences de lettres ou les niveaux de gris d'une image. Nous nous intéressons au problème de la détermination de l'ordre d'une telle chaîne. Dans la continuité de ce problème nous considérons celui, *a priori* éloigné, de l'estimation d'une densité par un histogramme. Dans ces deux domaines, nous justifions l'utilisation d'un critère par des notions de codage auxquelles nous appliquons une forme simple du principe de *Minimum Description Length*.

Nous nous efforçons également, à travers ces différents domaines d'application, de présenter des méthodes alternatives d'utilisation des critères d'information. Ces méthodes, dites comparatives, présentent une complexité d'utilisation moindre que les méthodes rencontrées habituellement, tout en permettant une description précise du modèle.

Abstract

Model selection problems appear frequently in a wide array of applicative domains such as data compression and signal or image processing. One of the most used tools to solve those problems is a real quantity to be minimized called information criterion or penalized likelihood criterion.

The principal purpose of this thesis is to *justify* the use of such a criterion responding to a given model selection problem, typically set in a signal processing context. The sought justification must have a strong mathematical background.

To this end, we study the classical problem of the determination of the order of an autoregression. We also work on Gaussian regression allowing to extract principal harmonics out of a noised signal. In those two settings we give a criterion the use of which is justified by the minimization of the cost resulting from the estimation. Multiple Markov chains modelize most of discrete signals such as letter sequences or gray scale images. We consider the determination of the order of such a chain. In the continuity we study the problem, *a priori* distant, of the estimation of an unknown density by an histogram. For those two domains, we justify the use of a criterion by coding notions to which we apply a simple form of the "Minimum Description Length" principle.

Throughout those application domains, we present alternative methods of use of information criteria. Those methods, called comparative, present a smaller complexity of use than usual methods but allow nevertheless a precise description of the model.

Présentation générale

Ce travail de recherche se veut situé au centre de l'interaction entre les mathématiques et ses applications en traitement du signal. A cet effet, il a été placé sour la co-direction de trois encadrants :

- Marc Arnaudon, professeur au Laboratoire de Mathématiques et Applications de l'université de Poitiers
- Christian Olivier, professeur au Laboratoire Signal, Images et Communications de l'université de Poitiers
- Olivier Alata, maître de conférence au Laboratoire Signal, Images et Communications de l'université de Poitiers

Les principaux objets d'étude de ce mémoire de thèse sont les critères d'information, outils utilisés dans la résolution de problèmes de sélection de modèles. A la base de ce projet, on trouve la volonté de *"justifier mathématiquement"* l'utilisation de tel ou tel critère devant une situation concrète donnée, typiquement issue d'un problème de traitement de l'information. Il est en effet difficile de décider quel critère sera le mieux adapté à la situation ; d'autant que l'efficacité d'un critère peut grandement varier au sein d'une même expérience.

Ce mémoire de thèse est divisé en 6 chapitres. Dans le chapitre d'introduction, nous présentons brièvement les concepts classiques d'entropie, d'estimation au sens du maximum de vraisemblance et la forme générale des problèmes de sélection de modèles que nous aborderons par la suite. Nous mettons en avant la principale difficulté de ces problèmes : la sur-paramétrisation induite par l'utilisation de la méthode du maximum de vraisemblance.

Le chapitre 2 est consacré à l'introduction des critères d'information dans leur généralité. Nous décrivons l'idée principale qui leur permet de pallier le problème de la surparamétrisation. La forme générale des critères d'information que nous utiliserons dans toute la suite est donnée. Nous introduisons les méthodes comparatives d'utilisation des critères d'information. Ces méthodes, alternatives aux méthodes rencontrées classiquement, présentent une complexité d'utilisation moindre tout en conservant une description précise du modèle ; elles sont présentées dans la partie 2.4.1 du chapitre 2. Les chapitres 3, 4, 5 et 6 sont chacun consacrés à un problème précis de sélection de modèles, paramétrique ou non, sur lequel nous avons particulièrement travaillé pendant ces trois années.

La détermination de l'ordre d'un processus autorégressif est historiquement le premier de ces problèmes puisque Akaike [Aka73, Aka74] y définit son critère. Nous étudions ce contexte dans le chapitre 3. Les méthodes comparatives sont utilisées pour donner une description plus précise du modèle dans le cas unidimensionnel et dans le cas bidimensionnel des images de textures.

Les chaînes de Markov multiples font, au chapitre 4, l'objet de l'élaboration d'un critère dont la justification est donnée en terme de longueur de codage. Ce critère s'inscrit dans la lignée des travaux de Rissanen sur la complexité stochastique [Ris86b] et le principe du *Minimum Description Length* (MDL) [BRY98, GMP05].

Dans le chapitre 5 nous nous intéressons au problème de l'estimation de densité par histogramme. En nous appuyant sur les travaux précédents relatifs aux chaînes de Markov multiples, nous présentons un critère d'information dont la justification repose sur la minimisation d'une longueur de codage et qui répond au problème du choix de l'histogramme.

Nous abordons enfin au chapitre 6 le problème de la régression gaussienne dans le cas où les points de régression sont aléatoires. Ce contexte est appelé "random design" par Baraud et Birgé [Bar02, Bir04]. Nous donnons les propriétés asymptotiques de l'estimation produite par les méthodes comparatives ainsi qu'une étude du comportement du risque de cette estimation sous la forme d'une inégalité oracle.

Quatre annexes complètent ce mémoire. Nous présentons un algorithme de programmation dynamique (annexe A) utilisé dans les problèmes de sélection d'histogramme du chapitre 5. Les annexes B et C contiennent les articles [CAOA] et [COAA] publiés dans des compte-rendus de conférences de traitement du signal. Un article soumis pour l'heure à une nouvelle conférence fait l'objet de la dernière annexe D.

Table des matières

Chapitre 1

Introduction

1.1 Inégalité entropique

Dans cette partie, nous présentons les concepts classiques d'entropie et d'estimation au sens du maximum de vraisemblance.

1.1.1 Entropie

L'entropie est une grandeur clé associée à une distribution de probabilité. Nous en donnons brièvement une définition et le lien qu'elle entretient avec l'information de Kullback.

Dans ce mémoire l'entropie prendra les formes suivantes selon que notre espace Ω sera discret ou de la forme $\mathbb{R}$ muni de la tribu de Borel mesurée par une probabilité à densité f :

$$\left| \begin{array}{rcl} H(P) & = & -\sum_{x\in\Omega} P(x)\log P(x) \\ H(f) & = & -\int_{\mathbb{R}} f(x)\log f(x)dx \end{array} \right. \tag{1.1}$$

On sous-entendra toujours que $0.\log 0 = 0$. La base du logarithme a peu d'importance.

L'entropie est largement utilisée dans des domaines tels que la thermodynamique ou l'astrophysique. Sa principale utilité est de mesurer le désordre apporté par la distribution. Pour illustration, sur un espace discret de cardinal m, la distribution uniforme a la plus forte entropie. Parmi toutes les distributions à densité continue sur $[0,1]$, c'est la distribution uniforme qui a la plus forte entropie. Parmi toutes les distributions à densité continue sur $\mathbb{R}$ de moyenne et variance fixées μ et σ^2, c'est la distribution normale $\mathcal{N}(\mu,\sigma^2)$ qui a la plus grande entropie.

C'est en ce sens que, face à une observation d'une distribution dont il ne connait rien, un utilisateur devrait faire en premier lieu l'hypothèse que cette distribution est d'un de ces types. Dans le cas contraire il diminuerait sans raison le désordre apporté par sa distribution.

1.1.2 Entropie croisée

Considérons maintenant deux distributions définies sur un même espace probabilisé du type de ceux évoqués en (1.1). Nous définissons une grandeur non symétrique entre ces deux distributions, appelée entropie croisée, par

$$\left| \begin{array}{rcl} H(P,Q) & = & -\sum_{x\in\Omega} P(x)\log Q(x) \\ H(f,g) & = & -\int_{\mathbb{R}} f(x)\log g(x)dx \end{array} \right. \tag{1.2}$$

Par commodité de notation, nous ne ferons plus la distinction entre les cadres discret et continu et appellerons nos distributions P et Q. L'inégalité de convexité de Jensen donne le résultat suivant :

Proposition 1.1.1 *L'entropie $H(P)$ et l'entropie croisée $H(P,Q)$ vérifient*

$$H(P) \leq H(P,Q)$$

Ce résultat, également appelé inégalité d'information de Shannon, constitue les fondements de la théorie de l'information dans [Sha48]. Elle permet également de définir l'information non symétrique de Kullback entre P et Q comme

$$K(P,Q) = H(P,Q) - H(P). \tag{1.3}$$

Cette quantité positive mesure donc le défaut d'entropie que présente la distribution Q par rapport à P. C'est l'une des nombreuses divergences permettant de mesurer l'écart entre deux lois de probabilité. Basseville [Bas96] étudie plus généralement la famille des Φ-divergences dont l'information de Kullback fait partie.

Nous reviendrons sur ces notions d'entropie et d'entropie croisée au cours du chapitre 4 dans lequel nous montrerons le lien étroit qu'elles entretiennent avec la théorie du codage.

1.1.3 Estimation au sens du maximum de vraisemblance

Il nous faut ici un espace statistique $(\Omega, \mathcal{A}, P_\theta,\ \theta \in \Theta)$ où les P_θ sont des distributions indexées par l'ensemble Θ. Cet espace peut être discret ou continu ; dans ce dernier cas, on supposera que chaque probabilité P_θ a une densité f_θ par rapport à une mesure fixe et nous confondrons P_θ et f_θ dans les cas où cela ne peut amener d'erreurs.

Les données du statisticien sont le plus souvent un échantillon $x^n = x_1, \ldots, x_n$ de taille n idéalement distribué selon une probabilité inconnue $P^\star$.

L'échantillon étant donné, la fonction de vraisemblance est

$$\begin{array}{rcl} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & P_\theta(x^n) \end{array} \tag{1.4}$$

dans le cas général. Pour nous rapprocher des concepts d'entropie, nous nous intéressons plutôt à l'opposé de la log-vraisemblance, que nous notons l :

$$\begin{array}{rcl} l : \Theta & \to & \mathbb{R} \\ \theta & \mapsto & -\log P_\theta(x^n). \end{array} \tag{1.5}$$

La méthode du maximum de vraisemblance préconise de choisir, lorsqu'il existe et est unique, le paramètre

$$\hat{\theta} = \text{Argmin}(l(\theta),\ \theta \in \Theta) \tag{1.6}$$

et la probabilité associée $P_{\hat{\theta}}$ comme estimation de $P^\star$. Pour plus de détails sur cette méthode, nous renvoyons à des textes classiques comme [Ald97] pour un historique, [vdV98] pour une approche plus théorique et [Kay93] qui s'inscrit dans une optique de traitement du signal.

Lorsque la structure statistique est de type produit construit par indépendance, la vraisemblance devient

$$\begin{array}{rcl} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & \displaystyle\prod_{i=1}^{n} P_\theta(x_i). \end{array} \tag{1.7}$$

et la log-vraisemblance

$$\begin{array}{rcl} l : \Theta & \to & \mathbb{R} \\ \theta & \mapsto & \displaystyle -\sum_{i=1}^{n} \log P_\theta(x_i). \end{array} \tag{1.8}$$

La loi des grands nombres assure que $n^{-1}l(\theta)$ converge vers l'entropie croisée $H(P^\star, P_\theta)$; entropie qui dépasse $H(P^\star)$ conformément à l'inégalité de Shannon.

Par conséquent, au moins asymptotiquement, utiliser la méthode du maximum de vraisemblance c'est choisir une probabilité dont l'entropie croisée avec $P^\star$ se rapproche de l'entropie de $P^\star$. En d'autres termes, c'est choisir une distribution qui se rapproche de $P^\star$ au sens de l'information de Kullback (1.3).

1.2 La sélection de modèles

Nous présentons ici le problème classique de la sélection de modèles dans un cadre assez simple qui conviendra pour la suite de ce mémoire et mettons en avant sa principale difficulté lorsqu'il est traité avec la méthode du maximum de vraisemblance. Considérons toujours notre espace statistique $(\Omega, \mathcal{A}, P_\theta,\ \theta \in \Theta)$ et notre donnée x^n idéalement distribuée selon une probabilité $P^\star$.

Précisons que l'appartenance éventuelle de $P^\star$ à l'ensemble des P_θ n'entre pas dans notre étude. On trouve dans la littérature des références traitant des deux cas ; citons par exemple [Nis84, ZDG01] qui traitent respectivement de régression multiple et chaîne de Markov multiples dans le cas où le modèle contient la vraie probabilité et [Nis88] pour des résultats généraux dans le cas contraire.

Si l'utilisateur dispose *a priori* d'informations assez précises sur ses futures observations, il aura des chances de mettre dans son modèle $P_\theta,\ \theta \in \Theta$ la vraie distribution $P^\star$. Pour illustration, donnons à l'observateur une séquence de 0 et de 1 et précisons lui qu'il s'agit des résultats d'une suite de lancers de pièces ; alors il prendra sûrement pour modèle des lois de Bernoulli, incluant ainsi $P^\star$. Par contre, donnons-lui une séquence de nombre réels dont nous seuls savons qu'elle est un échantillon d'une loi de Student et il prendra pour modèle des lois normales, excluant alors $P^\star$.

Dans la suite de ce mémoire, nous indiquerons l'appartenance de $P^\star$ au modèle $P_\theta, \theta \in \Theta$ dans les situations où elle importe. En l'absence de cette précision, on

peut remplacer $P^\star$ par sa projection, en un sens à préciser selon le cas, sur notre modèle. Selon le contexte, la probabilité $P_{\theta^\star}$ désignera $P^\star$ ou bien cette projection.

1.2.1 Les modèles

Donnons-nous certains sous-ensembles Θ_i, $i \in I$ de Θ. On les appelle sous-modèles ou bien modèles. D'une manière générale, faire de la sélection de modèles c'est déterminer, à partir des données x^n à quel(s) sous-modèle(s) le paramètre $\theta^\star$ appartient.

Pour $i \in I$ fixé nous appellerons, lorsqu'il existe, $\hat{\theta}_i$ le paramètre estimé au sens du maximum de vraisemblance à l'intérieur de Θ_i :

$$\hat{\theta}_i = \text{Argmin}(l(\theta),\, \theta \in \Theta_i). \tag{1.9}$$

Pour l'instant, nous n'avons pas supposé que les modèles sont paramétriques, cadre dans lequel un certain nombre de sous-modèles apparaissent naturellement. Les chapitres 3, 4 et 6 porteront sur de tels modèles tandis que le chapitre 5 traitant de la sélection d'histogrammes se situe dans le cadre non-paramétrique.

Modèles paramétriques naturels

Dans le cas paramétrique où $\Theta = \mathbb{R}^m$ on rencontre souvent les $m+1$ sous-modèles emboités

$$\{0\} = \Theta_0 \subset \Theta_1 \subset \cdots \subset \Theta_m = \mathbb{R}^m \tag{1.10}$$

où $\theta = (\theta_1, \ldots, \theta_m) \in \Theta_j$ si et seulement si $\theta_k = 0$ pour $k \geq j+1$. Ce seront par exemple les modèles utilisés pour la détermination de l'ordre d'une autorégression dans le chapitre 3.

On peut également faire apparaître des sous-modèles disjoints plus fins en fixant un paramètre θ^0 de référence et en considérant les 2^m sous-ensembles indexés par un support S, partie de $[\![1, m]\!]$:

$$\Theta_S = \left\{\theta \in \Theta \text{ tels que } \forall j \in S,\, \theta_j \neq \theta_j^0 \text{ et } \forall j \notin S,\, \theta_j = \theta_j^0\right\}. \tag{1.11}$$

Un paramètre $\theta \in \Theta_S$ est dit de support S. Nous appelons $S^\star$ le support de θ^*. Le cas le plus fréquent est $\theta^0 = 0$, la sélection de modèles revient alors à chercher les composantes non nulles de θ^*. Nous utiliserons souvent ces modèles dans la suite du fait de la précision qu'ils apportent par rapport aux modèles emboités. Un paramètre peut en effet appartenir à Θ_k et avoir également des composantes d'indices $l < k$ nulles. Ces composantes seront estimées si l'on considère seulement les modèles (1.10), alors qu'elles ne le seront pas si on a pu déterminer le support au sens des modèles (1.11).

1.3 La surparamétrisation

Nous abordons ici, à travers des exemples simples, la principale difficulté de la sélection de modèle : la surparamétrisation induite par l'estimation au sens du maximum de vraisemblance.

1.3.1 Modèle de lois normales

Les données x^n sont ici une séquence de nombres réels que l'observateur essaie de modéliser, selon les remarques faites précédemment sur l'entropie, par des réalisations indépendantes d'une loi normale inconnue. En l'absence de plus d'information, la moyenne et la variance de ses lois sont des paramètres libres. L'espace des paramètres est donc $\mathbb{R} \times \mathbb{R}_+^*$, les densités de probabilité en jeu sont

$$f_{\mu,\sigma^2}(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\ \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$$

et la vraisemblance de l'échantillon pour l'une d'entre elles s'écrit

$$\prod_{i=1}^{n} f_{\mu,\sigma^2}(x) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Prenons le paramètre de référence $(0, 1)$, autrement dit la loi normale centrée réduite, et considérons les sous-modèles exprimés par (1.11) :

$$\begin{aligned}
\Theta_\emptyset &: \text{la loi } \mathcal{N}(0,1) \\
\Theta_1 &: \text{les lois } \mathcal{N}(\mu,1) \text{ avec } \mu \neq 0 \\
\Theta_2 &: \text{les lois } \mathcal{N}(0,\sigma^2) \text{ avec } \sigma^2 \neq 1 \\
\Theta_{\{1,2\}} &: \text{les lois } \mathcal{N}(\mu,\sigma^2) \text{ avec } \mu \neq 0,\ \sigma^2 \neq 1.
\end{aligned}$$

Le problème de sélection de modèles se pose donc ainsi : auquel de ces sous-modèles le paramètre $\theta^\star$ appartient-il ?

En essayant de répondre à cette question par la méthode du maximum de vraisemblance, nous allons voir apparaître la raison d'être des critères d'information. En effet les paramètres de $\mathbb{R} \times \mathbb{R}_+^*$ estimés au sens du maximum de vraisemblance sont

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Notons que $\hat{\mu} \neq 0$ et $\hat{\sigma}^2 \neq 1$ avec probabilité 1. De plus pour $S \subset [\![1, 2]\!]$, aucun des paramètres $(\hat{\mu}_S, \hat{\sigma}_S^2)$ estimés au sens de (1.9) dans le modèle respectif Θ_S ne donnera une meilleure vraisemblance que $(\hat{\mu}, \hat{\sigma}^2)$. Par conséquent, du point de vue de la vraisemblance, aucun des modèles en compétition n'est meilleur que le modèle plein $\Theta_{\{1,2\}}$; et ce même si la loi génératrice de l'échantillon est $\mathcal{N}(0, 1)$.

On voit ici que la méthode du maximum de vraisemblance préfère qu'on lui laisse le maximum de liberté sur les paramètres à estimer. Ce phénomène, appelé *surparamétrisation*, est la principale difficulté de la sélection de modèles.

1.3.2 Modèle de dés

La surparamétrisation n'est pas réservée aux variables continues. Supposons cette fois que les données consistent en une séquence de n entiers entre 1 et 6, résultats de

lancers d'un dé. On demande à l'observateur s'il pense que le dé est pipé. Il pourra mettre en confrontation les deux modèles :

$$P_{np}(X = j) = \frac{1}{6},\ \forall j = 1, \dots, 6$$

qui ne contient aucun paramètre libre et

$$P_p(X = j) = \alpha_j,\ \forall j = 1, \dots, 6$$

avec $\sum \alpha_j = 1$ qui contient 5 paramètres libres.

Les paramètres estimés au sens du maximum de vraisemblance pour le deuxième sont $\widehat{\alpha}_j = n_j/n$ avec n_j le nombre d'apparitions du chiffre j dans les données. On note $\widehat{P}_p$ la probabilité correspondante. Par définition de l'esimation au sens du maximum de vraisemblance, on a

$$\prod_{i=1}^{n} \widehat{P}_p(X = x_i) \geq \prod_{i=1}^{n} \frac{1}{6},$$

l'égalité se produisant uniquement si n est multiple de 6 et que chaque n_j vaut $n/6$, auquel cas les deux modèles sont *ex-aequo*.

Encore une fois, au sens du maximum de vraisemblance, le modèle pipé disposant du plus grand nombre de paramètres libres est le meilleur, et ce même si le dé est équilibré.

1.3.3 Modèle d'interpolation polynômiale

Ce modèle se rapproche du problème de régression qui sera traité plus en détails dans le chapitre 6. Il montre bien le phénomène de surparamétrisation.

Les données sont, cette fois, un ensemble de n points du plan (x_i, y_i) où les x_i sont déterministes distincts et les y_i sont donnés par

$$y_i = A^\star(x_i) + \varepsilon_i,\ i = 1, \dots, n$$

Ici $A^\star$ est un polynôme réel inconnu et les ε_i désignent un n-échantillon d'une variable $\varepsilon \sim \mathcal{N}(0, 1)$.

Choisissons $A_0, \dots, A_{n-1}$ une base de $\mathbb{R}_{n-1}[X]$ et décidons qu'un polynôme $B = \sum b_i A_i$ aura pour support $S \subset [\![0, n-1]\!]$ si et seulement si $b_i = 0 \Leftrightarrow i \notin S$. Les modèles inspirés de (1.11) en compétition sont

$$\Theta_S = \{\mathcal{L}(A(x) + \varepsilon),\ \mathrm{supp}(A) = S\},\ S \subset [\![0, n-1]\!].$$

Puisque l'erreur ε est gaussienne, l'opposé de la log-vraisemblance de l'échantillon relativement à un polynôme A fixé est, à des constantes près, la distance

$$\sum_{i=1}^{n} (A(x_i) - y_i)^2 .$$

Par conséquent la vraisemblance maximale au sein d'un modèle Θ_S correspond à l'erreur commise par le polynôme d'interpolation de ce même modèle. Mais le polynôme d'interpolation de Lagrange de degré $n-1$ rend cette erreur nulle. Cela signifie que, encore du point de vue de la vraisemblance, le modèle plein $\Theta_{[\![0,n-1]\!]}$ qui contient le polynôme de Lagrange est le meilleur. Dans tous les cas, la vraisemblance préfère un polynôme de degré $n-1$ au polynôme $A^\star$, qui peut être de degré bien inférieur, voir figure 1.1.

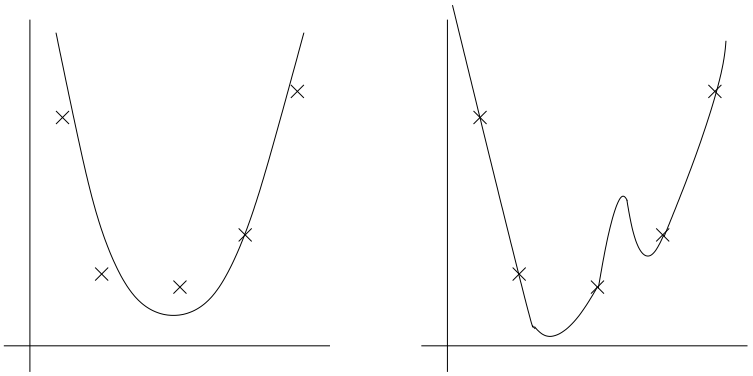


FIG. 1.1 – Le polynôme inconnu et le polynôme d'interpolation surparamétré

1.3.4 Résumé

Dans les trois précédents modèles, nous avons mis en avant la surparamétrisation produite par la méthode du maximum de vraisemblance. Ce phénomène est commun à la plupart des modélisations statistiques de données : la vraisemblance préfère qu'il y ait beaucoup de paramètres à estimer.

Les principaux outils utilisés pour pallier ce problème, également principaux sujets d'étude de ce mémoire, sont les critères d'information.

Dans le chapitre 2 nous les présenterons en détail et dresserons un historique de leurs applications. Nous introduirons également les méthodes comparatives avec lesquelles nous les utilisons.

Les chapitres 3, 4, 5 et 6 sont réservés aux différents résultats théoriques et expérimentaux sur lesquels nous avons travaillé au cours cette thèse.

Chapitre 2

Critères d'information

Les critères d'information peuvent prendre diverses formes, nous renvoyons à la partie 2.1 de ce chapitre pour plus de détails. Rappelons que la motivation principale de ce travail de recherche est la volonté de *justifier mathématiquement* l'utilisation de tel ou tel critère devant une situation concrète donnée.

A l'issue de notre travail nous n'avons pas *justifié* l'utilisation d'un critère-miracle qui serait efficace en toutes circonstances. Qui plus est, à force de parcourir les écrits d'auteurs provenant de milieux variés qui utilisent les critères d'information à des fins toutes aussi variées, il nous est apparu qu'il existe presque autant de ces *justifications* qu'il y a de problèmes de sélection de modèles. Au cours du chapitre 3, nous utiliserons les critères classiques pour une illustration de leur comportement sur un problème usuel de sélection de modèles. Les chapitres 4, 5 et 6 présenteront une justification de l'utilisation de leur critère respectif.

Nous nous intéressons également à différentes méthodes d'utilisation des critères d'information. La méthode la plus couramment utilisée (2.2) présente en effet le défaut de nécessiter une complexité élevée de calculs, cette complexité peut être exponentielle lorsque sont utilisés les modèles fins décrits en (1.11). L'implémentation sur machine est alors coûteuse en temps. Nous présenterons dans la partie 2.4 de ce chapitre des méthodes alternatives moins coûteuses et donc plus facilement exploitables sur machine.

2.1 Principe et expression des critères

Reprenons les notations de la partie 1.2 où les modèles en compétition sont les $\Theta_i \subset \Theta$, $i \in I$. L'estimateur au sens du maximum de vraisemblance à l'intérieur d'un modèle fixé Θ_i produit un paramètre $\widehat{\theta}_i$ comme en (1.9). La log-vraisemblance $l(\widehat{\theta}_i)$ définie en (1.5) peut être vue comme une mesure de l'adéquation du modèle Θ_i aux données. On a vu dans le chapitre précédent que cette mesure est mal adaptée au problème de la sélection de modèles puisqu'elle induit de la surparamétrisation.

La question que l'on peut alors se poser est la suivante : avec un modèle Θ_i on arrive à une certaine adéquation, certes, mais à quel prix ?

L'exemple de l'interpolation polynômiale dans la partie 1.3 montre particulièrement bien l'intérêt de cette question : avec le polynôme de Lagrange, on arrive à une erreur d'interpolation nulle ; en contrepartie le prix à payer est le degré maximal $n-1$ du polynôme.

L'idée des critères d'information est alors la suivante : assigner à chaque modèle en compétition un poids $p(\Theta_i)$ et le mettre en balance avec l'adéquation du modèle aux données mesurée par $l(\widehat{\theta}_i)$. Cette recherche de compromis est effectuée par la minimisation de $l(\widehat{\theta}_i)+p(\Theta_i)$. Nous verrons, par exemple en partie 2.2, que le choix de la forme additive de cette expression apparaît naturellement lors de la justification de l'utilisation de critères.

2.1.1 Historique

Avant de donner de manière plus précise la forme que prendront les critères étudiés par la suite, faisons un tour d'horizon des principaux travaux liés à ce sujet. L'idée d'opposer la vraisemblance au coût nécessaire à son obtention est apparue pour la première fois dans le courant des années 1970 avec les travaux fondateurs de Akaike [Aka73, Aka74] ou de Mallows [Mal73].

Le problème de sélection de l'ordre d'un processus autorégressif se prête bien à l'utilisation des critères d'information. Il est la trame du travail de Akaike, a été étudié par Shibata [Shi76] qui montre des résultats asymptotiques puis par Hannan et Al. qui donnent un critère original dans [HQ79]. Plus récemment Broersen [Bro00] s'intéresse à des échantillons plus réduits tandis que l'utilisation des critères sur les processus autorégressifs bidimensionnels modélisant des images texturées est étudiée par Alata et Al. dans [AO03].

Dans [Nis84], Nishii décrit les réactions asymptotiques des modèles de régression choisis par différents critères d'information. Le cadre non asymptotique de ces problèmes de régression est privilégié aujourd'hui : Baraud en fait une étude dans [Bar00, Bar02] en s'intéressant à l'aléa présent ou non sur les points de régression tandis que Birgé [Bir04] exprime les bornes que l'on peut obtenir selon les différentes mesures du risque choisi.

Les chaînes de Markov multiples et cachées sont d'un intérêt certain dans le domaine du codage de l'information. Les critères d'information permettent de dimensionner ces processus. Ce sujet a été étudié peu après les travaux fondateurs d'Akaike par Tong [Ton75] et son étude se poursuit largement aujourd'hui sur des thèmes plus actuels ; citons les travaux de Olivier et Al. [OPAL97]. Dans les années 1980 Rissanen [Ris86a] relie les chaînes de Markov multiples à la notion de complexité stochastique qu'il précisera dans un cadre général dans [Ris86b, Ris89] ou plus récemment dans [Ris96]. Le lien entre les deux notions sera le principe du *Minimum Description Length* (MDL) pour lequel nous renvoyons à un article fondateur de Rissanen [Ris78] ou aux études plus récentes de Barron et al. [BRY98] et de Grunwald [GMP05]. De ce lien sera dérivé le codage arithmétique présenté par Rissanen en 1976 [Ris76]. Cette technique est utilisée de nos jours dès qu'il s'agit de compresser efficacement l'information, notamment dans la norme JPEG2000 de compression d'images fixes ou dans les normes MPEG4 et H264 de compression vidéo. L'implémentation de ce codage ainsi que les problèmes qui lui sont liés sont traités par Witten et al. dans [WNC87, MNW98]. Nous verrons comment le codage arithmétique est lié à la complexité stochastique et comment regarder cette dernière comme un critère d'information.

L'approximation par un histogramme d'une densité de probabilité, problème non-paramétrique de sélection de modèles, peut également être envisagée par l'utili-

sation de critères d'information. Dans cette optique, Hall présente des résultats basés dans un premier temps sur la minimisation d'une distance du type Kullback-Leibler [Hal87] puis sur un critère du type Akaike [Hal90]. Nous citons aussi les travaux de Rissanen et al. qui relient encore la notion de complexité stochastique à ce problème dans [RSY92]. Plus récemment des résultats non-asymptotiques ont émergé dans ce domaine notamment par les travaux de Castellan [Cas99, Cas00] qui donne des bornes sur le risque de l'estimation. Citons également les travaux de Birgé et Rozenholc [BR06] qui étudient la sélection d'histogramme à l'aide d'un critère d'un point de vue applicatif et les comparent à d'autres méthodes comme le seuillage.

Nous voyons que ces contextes précis – autorégression, regression, dimensionnement des chaînes de Markov multiples et sélection d'histogramme – sont autant de domaines privilégiés d'utilisation de critères d'information. Nous y reviendrons plus en détail dans les chapitres 3, 4, 5 et 6 qui leur sont consacrés.

En dehors de ces domaines, Schwarz [Sch78] justifie un critère devenu très populaire par une approche bayesienne. Nous n'étudions pas ce critère dans ce mémoire et renvoyons à son auteur ou aux travaux plus récents de Lebarbier et Al. [EL06] pour plus de détails le concernant.

Les travaux généraux de Nishii et al. [NBK88, Nis88] étudient les comportements asymptotiques des modèles choisis par un critère. S'appuyant sur ces écrits, El-Matouat et Al. proposent dans [EMH96] le critère φ_β sur lequel nous reviendrons en partie 2.3. Nous utiliserons souvent φ_β au cours de ce mémoire pour sa fléxibilité.

Citons enfin les travaux de Birgé [Bir06] qui s'intéresse à des inégalités sur les risques résultants d'estimation par critères d'information. Ce dernier article, à l'instar des travaux de Massart [Mas07], s'efforce également d'exprimer les problèmes de sélection de modèles avec une approche à la fois mathématique, claire et précise.

2.1.2 Forme générale des critères d'information

Expression

Dans le même cadre que la partie 1.2, nous donnons ici la forme générale des critères d'information que nous étudierons par la suite. Pour $i \in I$, nous notons $|\Theta_i|$ le nombre de paramètres libres à estimer à l'intérieur du modèle Θ_i. Un critère d'information, IC pour *Information Criterion*, se présente sous la forme d'une application $\mathrm{IC} : I \to \mathbb{R}$ avec

$$\mathrm{IC}(i) = 2l(\widehat{\theta}_i) + |\Theta_i|\alpha(n). \tag{2.1}$$

Rappelons que le terme de vraisemblance $l(\widehat{\theta}_i)$ est défini en (1.5) et (1.9). Le facteur 2 devant ce terme est une convention habituelle. Il arrive naturellement lors, par exemple, de la justification du critère d'Akaike que nous donnerons en partie 2.2. Le deuxième terme est appelé *pénalité*, nous y revenons dans quelques lignes.

Dans certains cas, on rencontre une normalisation du critère (2.1) par n^{-1}. Le terme de vraisemblance $n^{-1}l(\widehat{\theta}_i)$ est alors une estimation de l'entropie (1.1) de la distribution inconnue $P_{\theta^\star}$. C'est pour cela que l'on trouve parfois, comme dans [OA08], la terminologie "critère entropique pénalisé" pour désigner un critère d'information. Nous travaillerons la plupart du temps sans cette normalisation qui, de toutes façons, n'entre pas en jeu lors de la sélection d'un modèle.

Interprétation

La quantité IC(i) (2.1) mesure l'adéquation des données à Θ_i pondérée par le poids de ce modèle. Une inégalité du type IC(i) $\leq$ IC(j) signifie que le modèle Θ_i réalise un meilleur compromis entre l'adéquation aux données (mesurée par $2l(\widehat{\theta}_i)$) et le coût de ce modèle (mesuré par $|\Theta_i|\alpha(n)$) que ne le fait Θ_j. En d'autres termes, par IC(i) $\leq$ IC(j) le critère montre que sa préférence va au modèle Θ_i plutôt qu'à Θ_j. Un critère d'information répond donc au problème de sélection de modèle en considérant, par exemple :

$$\widehat{i} = \text{Argmin}\,(\text{IC}(i),\ i \in I)\,. \tag{2.2}$$

On choisit alors $\Theta_{\widehat{i}}$ comme meilleur modèle censé inclure la distribution inconnue.

Cette méthode de sélection du meilleur modèle n'est pas la seule possible. Lorsque le nombre de modèles en compétition est élevé, elle nécessite de calculer beaucoup de critères. Nous présenterons dans la partie 2.4 de ce chapitre des méthodes alternatives qui permettent de diminuer le nombre de calculs nécessaires à la sélection d'un modèle.

Remarques sur la pénalité

Le tableau 2.1 donne un résumé des pénalités usuellement rencontrées ; nous les utiliserons tout au long de ce mémoire.

La pénalité telle qu'exprimée dans (2.1) est linéaire en le nombre de paramètres libres du modèle. Ce choix n'est pas fait au hasard ; nous verrons que les principaux critères abordés dans ce mémoire sont de cette forme. Par conséquent, le seul élément restant à choisir dans l'expression des critères est la fonction de pénalité $\alpha(n)$. Rappelons que la méthode du maximum de vraisemblance a tendance à sur-paramétrer le modèle ; la pénalité étant là pour pallier ce phénomène. Par conséquent, si la pénalité est trop faible, le modèle choisi par (2.2) risque encore d'être sur-paramétré. A l'opposé, si elle est trop forte, on observera le phénomène inverse de sous-paramétrisation. Le choix de la pénalité, et surtout la justification d'un tel choix, est le coeur du problème abordé dans ce mémoire.

Notons ici que dans certains travaux datant de ces 20 dernières années, on trouve des pénalités qui ne présentent plus cette linéarité. Plus précisément on trouve dans les travaux de Castellan, Barron et Al., Baraud, ou Birgé [Cas99, BBM99, Bar00, Bar02, Bir04] des termes de pénalités de la forme

$$\alpha(n) = K(1 + \sqrt{2L_i})^2 \tag{2.3}$$

où $K > 1$ et $L_i \geq 0$ est un nouveau poids associé au modèle Θ_i. C'est ce nouveau poids qui peut faire perdre la linéarité de la pénalité. Ces critères sont justifiés par des arguments de minimisation d'un risque de manière non-asymptotique. A l'opposé, les critères "classiques" élaborés avant 1990 reposent sur des considérations asymptotiques. La différence majeure entre ces deux types de critères repose sur la dépendance de la fonction $\alpha(n)$ en n. Si l'on veut un bon comportement asymptotique, il faut que $\alpha(n)$ croisse avec n de manière à ce que la pénalité ne soit pas écrasée par la vraisemblance dans (2.1). Au contraire, dans un cadre non asymptotique, la pénalité peut rester constante en n comme dans (2.3). A noter

TAB. 2.1 – Un rapide tour d'horizon des critères usuels

Critère	Pénalité $\alpha(n)$	Références	Commentaires
AIC	2	[Aka73, Aka74]	Utilisé historiquement pour la détermination de l'ordre d'une autorégression.
BIC	$\log n$	[Sch78],[EL06]	Justification bayésienne non étudiée dans ce mémoire.
φ	$\log\log n$	[HQ79]	Dans le cadre de l'autorégression cette pénalité assure une convergence presque-sûre.
BIC/RIC	$\log n$	[Ris86b, Ris86a] [BRY98]	Basé sur les notions de complexité stochastique, principe MDL. Adapté à la sélection de l'ordre d'une chaîne de Markov multiple.
φ_β	$n^\beta \log\log n$, $0 < \beta < 1$	[EMH96]	Ces pénalités assurent une convergence presque-sûre.
IC	constante ou dépendante des modèles	[Cas99]	Basé sur la minimisation de risques non asymptotiques pour la sélection d'histogrammes.
IC	constante ou dépendante des modèles	[Bar00, Bar02]	Basé sur la minimisation de risques non asymptotiques pour les problèmes de régression.

que le critère d'Akaike ($\alpha(n) = 2$) fait office d'exception puisqu'il fait partie des critères "classiques" élaborés asymptotiquement mais que sa pénalité ressemble à celles (2.3) dérivées actuellement dans un cadre non-asymptotique. Les auteurs cités précédemment présentent d'ailleurs les pénalités (2.3) comme des généralisations de celle d'Akaike.

2.2 Justification d'un critère usuel : celui d'Akaike

A titre d'exemple, nous donnons ici la justification de l'utilisation du critère d'Akaike. Il est l'un des premiers critères d'information apparu dans la littérature [Aka74]. Une démarche visant à minimiser un risque justifie le choix de la fonction de pénalité $\alpha(n)$ dans (2.1).

Nous adoptons la notation à densité f_θ, $\theta \in \Theta$ pour désigner les distributions. Plaçons-nous dans un cadre paramétrique où $\Theta = \mathbb{R}^m$ et où les modèles Θ_i, $i \in I$ en compétition sont des sous-espaces affines de dimension $|\Theta_i|$, comme c'est le cas pour les sous-modèles définis par (1.11). Supposons également que la structure statistique soit de type produit si bien que la vraisemblance des données x^n pour une distribution f_θ vaut $\prod_{i=1}^n f_\theta(x_i)$.

Sous des conditions classiques [Kay93, vdV98] de régularité sur les fonctions $\theta \mapsto f_\theta(x)$, la matrice d'information de Fisher en $\theta^\star$ définie par

$$I(\theta^\star)_{r,s} = -\mathbb{E}_{\theta^\star}\left[\frac{\partial^2 \log f_\theta(.)}{\partial\theta_r\ \partial\theta_s}_{|\theta=\theta^*}\right]$$

définit un produit scalaire $\langle .,. \rangle$ sur l'espace des paramêtres de norme associée $\|.\|$.

Supposons que l'on dispose d'un estimateur $\widehat{\theta}$ de θ^*, dont on omet la dépendance en n pour la clarté, qui soit non biaisé et tel que $\sqrt{n}(\widehat{\theta} - \theta^*)$ soit asymptotiquement normal. Nous notons θ_i^* et $\widehat{\theta}_i$ les projections orthogonales respectives de θ^* et $\widehat{\theta}$ sur un modèle donné Θ_i. On obtient alors les lois asymptotiques suivantes :

$$\begin{aligned} n\left\|\widehat{\theta} - \theta^*\right\|^2 &\sim \chi^2(m) \\ n\left\|\widehat{\theta}_i - \theta_i^*\right\|^2 &\sim \chi^2(|\Theta_i|) \\ n\left\|(\theta^* - \theta_i^*) - \left(\widehat{\theta} - \widehat{\theta}_i\right)\right\|^2 &\sim \chi^2(m - |\Theta_i|) \end{aligned} \tag{2.4}$$

De plus on a

$$\begin{aligned} \left\|\theta^* - \widehat{\theta}_i\right\|^2 &= \|\theta^* - \theta_i^*\|^2 + \left\|\theta_i^* - \widehat{\theta}_i\right\|^2 \\ \left\|\widehat{\theta} - \widehat{\theta}_i\right\|^2 &= \|\theta^* - \theta_i^*\|^2 + \left\|(\theta^* - \theta_i^*) - \left(\widehat{\theta} - \widehat{\theta}_i\right)\right\|^2 \\ &\quad -2\left\langle \widehat{\theta} - \theta^*, \theta_i^* - \theta^* \right\rangle. \end{aligned} \tag{2.5}$$

Notons que ce dernier produit scalaire converge vers 0 au sens L^2 par (2.4) et Schwarz :

$$\mathbb{E}\left[\left\langle \widehat{\theta} - \theta^*, \theta_i^* - \theta^* \right\rangle^2\right] \leq \frac{1}{n}\|\theta_i^* - \theta^*\|^2 \,\mathbb{E}\left[n\left\|\widehat{\theta} - \theta^*\right\|^2\right] \underset{n\to\infty}{\to} 0.$$

Rappelons que l'information (non symétrique) de Kullback définie en (1.3) permet de mesurer l'écart entre deux densités par

$$K(\theta, \theta') = -\int f_\theta(x) \log \frac{f_{\theta'}(x)}{f_\theta(x)} dx \geq 0.$$

Cette mesure fait partie de la famille des Φ-divergences dont on trouve une étude détaillée par exemple dans [Bas89, Bas96]. Nous aurons simplement besoin du lien suivant entre l'information de Kullback et la norme définie par la matrice d'information de Fisher. Pour θ au voisinage de $\theta^\star$, on a :

$$2K(\theta^*, \theta) = \|\theta^* - \theta\|^2 (1 + o(1)). \tag{2.6}$$

L'argument du critère d'Akaike

La norme $\|.\|$ héritée de la métrique de Fisher, reliée à l'information de Kullback par (2.6), nous donne un moyen de mesurer le risque de notre procédure de sélection de modèle. Pour $i \in I$ fixé, le risque résultant du choix de Θ_i comme espace d'estimation de θ^* est

$$\mathbb{E}\left[\left\|\widehat{\theta}_i - \theta^*\right\|^2\right]. \tag{2.7}$$

La démarche de Akaike consiste alors à choisir la pénalité du critère (2.1) de telle sorte que ce risque soit minimal.

Le critère

Le risque (2.7) est inaccessible à l'utilisateur. Nous voyons maintenant comment l'estimer pour obtenir le critère d'Akaike.

Utilisons les égalités (2.5) dans lesquelles on néglige le terme mixte à cause de sa convergence L^2 vers 0 pour estimer $\mathbb{E}\left[n\left\|\widehat{\theta}_i - \theta^*\right\|^2\right]$ par

$$\mathbb{E}\left[n\left\|\theta_i^* - \widehat{\theta}_i\right\|^2\right] + \mathbb{E}\left[n\left\|\widehat{\theta} - \widehat{\theta}_i\right\|^2\right] - \mathbb{E}\left[n\left\|(\theta^* - \theta_i^*) - \left(\widehat{\theta} - \widehat{\theta}_i\right)\right\|^2\right]$$

Etant données les lois asymptotiques (2.4) le premier terme s'estime par $|\Theta_i|$ et le dernier par $m - |\Theta_i|$.

Pour le terme restant $\mathbb{E}\left[n\left\|\widehat{\theta} - \widehat{\theta}_i\right\|^2\right]$ utilisons le lien (2.6) et estimons

$$\mathbb{E}\left[2nK(\widehat{\theta}, \widehat{\theta}_i)\right] = \mathbb{E}\left[2n\mathbb{E}_{\widehat{\theta}}\left[-\log\frac{f_{\widehat{\theta}_i}(.)}{f_{\widehat{\theta}}(.)}\right]\right]$$

par

$$-2\sum_{j=1}^{n}\log\frac{f_{\widehat{\theta}_i}(x_j)}{f_{\widehat{\theta}}(x_j)}. \tag{2.8}$$

Ainsi une estimation à partir de l'échantillon du risque (2.7) à minimiser est :

$$-\frac{2}{n}\sum_{j=1}^{n}\log\frac{f_{\widehat{\theta}_i}(x_j)}{f_{\widehat{\theta}}(x_j)} + \frac{2|\Theta_i| - m}{n}$$

La minimisation se faisant par rapport à Θ_i on peut oublier tous les termes dans cette expression qui n'en dépendent pas pour obtenir le critère AIC (Akaike Information Criterion) sous la forme générale annoncée en (2.1) :

$$\mathrm{AIC}(k) = -2\sum_{j=1}^{n}\log f_{\widehat{\theta}_i}(x_j) + 2|\Theta_i|. \tag{2.9}$$

L'argument de Akaike préconise donc de choisir $\alpha(n) = 2$ pour fonction de pénalité.

2.3 Le résultat de R. Nishii

R. Nishii donne dans [Nis88] une étude des propriétés asymptotiques des modèles choisis par critères d'information. Lorsque les modèles en compétition sont de la forme (1.11) et lorsque $S^\star$ désigne le support de θ^*, il montre le résultat suivant :

Théorème 2.3.1 *Désignons par $\widehat{S}$ le support choisi par minimisation du critère d'information :*

$$\widehat{S} = \mathit{Argmin}(\mathit{IC}(S),\ S \subset [\![1, m]\!]).$$

Si la pénalité $\alpha(n)$ satisfait $\alpha(n) \to \infty$ et $\alpha(n) = o(n)$ alors $\widehat{S}$ converge en probabilité vers $S^\star$.

Si elle satisfait $\log\log n = o(\alpha(n))$ et $\alpha(n) = o(n)$ alors $\widehat{S}$ converge presque sûrement vers $S^\star$.

Ce résultat permet de justifier l'utilisation d'une multitude de critères ; le choix de la pénalité n'ayant d'autre signification que le mode de convergence que l'on souhaite obtenir. Par exemple, choisir

$$\alpha(n) = (\log\log\log n)^\alpha,\ \alpha > 0$$

suffit à assurer la convergence en probabilité de notre estimateur du support. Le choix

$$\alpha(n) = (\log n)^\gamma,\ \gamma > 0$$

donnera, quant à lui, une convergence presque-sûre.

Le critère φ_β

Dans cette optique, El-Matouat et Al. [EMH96] proposent d'utiliser le critère φ_β dont la pénalité est donnée pour $0 < \beta < 1$ par

$$\alpha(n) = n^\beta \log\log n. \tag{2.10}$$

Le choix restant sur la valeur de β donne une paramétrisation des pénalités assurant, via le théorème 2.3.1, la convergence presque-sûre du support estimé vers le vrai support. Elle permet de plus de retrouver, pour $\beta = 0$, le critère φ de Hannan et Quinn [HQ79]. Cette paramétrisation n'est bien sûr pas la seule possible.

A n fixé, un ajustement de la valeur de β permet de retrouver la pénalité des autres critères usuels pour lesquels nous renvoyons au tableau 2.1. Pour cette raison nous utiliserons souvent le critère φ_β dans lequel nous ferons varier β. Les valeurs particulières de β sont :

$$\begin{aligned} \beta_{\mathrm{AIC}} &= (\log 2 - \log\log\log n)/\log n \\ \beta_{\mathrm{BIC}} &= (\log\log n - \log\log\log n)/\log n \end{aligned} \tag{2.11}$$

Notons que $\beta_{\mathrm{AIC}} < 0$ dès que $n \geq 1619$. Dans [OJM99], les auteurs du critère φ_β introduisent aussi les valeurs particulières de β suivantes :

$$\beta_{\min} = \frac{\log\log n}{\log n} = 1 - \beta_{\max} \tag{2.12}$$

et conseillent d'utiliser le critère φ_β pour $\beta_{\min} < \beta < \beta_{\max}$. Nous constaterons par la pratique qu'un choix proche de $\beta_{\min}$ donne souvent les meilleurs résultats en situation de simulation.

Notons que $\beta_{\mathrm{BIC}} < \beta_{\min}$, en d'autres termes, le critère $\varphi_{\beta_{\min}}$ pénalise plus que le critère BIC.

2.4 Complexité d'utilisation

Outre la justification de l'utilisation des critères, nous nous intéressons dans ce mémoire à la complexité de cette utilisation. Par complexité, nous entendons simplement le nombre de critères qu'il est nécessaire de calculer pour achever la

procédure de sélection de modèles. Dans le cas des modèles fins décrits en (1.11) la méthode (2.2), que nous qualifierons de globale, et qui s'écrit ici :

$$\widehat{S} = \text{Argmin}\left(\text{IC}(S),\ S \subset [\![1,m]\!]\right) \tag{2.13}$$

a une complexité valant 2^m. Cette complexité exponentielle est souvent trop élevée pour envisager une exécution rapide sur machine. Nous décrivons ici des méthodes d'utilisation des critères d'information alternatives à la méthode globale.

2.4.1 Méthode comparative

Cette méthode est proposée par Nishii dans [Nis88]. Fixons la valeur du critère lorsque tous les paramètres sont libres comme référence et estimons le support $S^\star$ par $\widehat{S}$ défini comme suit :

$$\left\{\begin{array}{rcl} \text{IC}_{\text{ref}} & = & \text{IC}([\![1,m]\!]) \\ \widehat{S} & = & \left\{j \in [\![1,m]\!],\ \text{IC}_{\text{ref}} \leq \text{IC}\left([\![1,m]\!] \setminus \{j\}\right)\right\}. \end{array}\right. \tag{2.14}$$

Ainsi $\widehat{S}$ contient les composantes que le critère juge importantes via $\text{IC}_{\text{ref}} \leq \text{IC}([\![1,m]\!] \setminus \{j\})$. Cette méthode nécessite le calcul de $m+1$ critères.

2.4.2 Méthode comparative inversée

La méthode comparative inversée prend, cette fois, comme référence la valeur du critère lorsque tous les paramètres sont gelés et estime $S^\star$ par

$$\left\{\begin{array}{rcl} \text{IC}_{\text{ref}} & = & \text{IC}(\emptyset) \\ \widehat{S} & = & \left\{j \in [\![1,m]\!],\ \text{IC}(\{j\}) \leq \text{IC}_{\text{ref}}\right\}. \end{array}\right. \tag{2.15}$$

Ce support contient une nouvelle fois les composantes que le critère juge importantes. Cette méthode nécessite également le calcul de $m+1$ critères.

Nous verrons au chapitre 6 que la méthode comparative inversée est moins robuste que la précédente méthode. Cela tient globalement au fait que les quantités IC_{ref}, $\text{IC}(\{j\})$, $j = 1, \ldots, m$, sur lesquelles nous nous basons pour l'estimation du support contiennent moins d'informations que celles utilisées par la méthode (2.14).

2.4.3 Méthode comparative descendante

Cette méthode, à l'instar de (2.14), élimine les composantes jugées non utiles par le critère. La procédure se déroule cette fois en un nombre aléatoire d'étapes. Pour commencer, fixons

$$\begin{array}{rcl} S^{(0)} & = & [\![1,m]\!] \\ \text{IC}_{\text{ref}}^{(0)} & = & \text{IC}(S^{(0)}). \end{array}$$

A la première étape, nous sélectionnons les candidats à l'élimination par

$$C^{(1)} = \left\{j \in S^{(0)},\ \text{IC}\left(S^{(0)} \setminus \{j\}\right) \leq \text{IC}_{\text{ref}}^{(0)}\right\}.$$

TAB. 2.2 – Les méthodes comparatives et leur complexité.

Méthode	Complexité
Globale (2.13)	2^m
Comparative (2.14)	$m+1$
Comparative inversée (2.15)	$m+1$
Comparative descendante, partie 2.4.3	$\leq m(m+1)/2$

Ecartons ensuite la composante $J^{(1)}$ jugée la moins utile par le critère et actualisons notre référence en vue de la prochaine étape :

$$\begin{aligned} J^{(1)} &= \text{Argmin}\left(\text{IC}\left(S^{(0)} \setminus \{j\}\right),\ j \in C^{(1)}\right) \\ S^{(1)} &= S^{(0)} \setminus \left\{J^{(1)}\right\} \\ \text{IC}_{\text{ref}}^{(1)} &= \text{IC}(S^{(1)}). \end{aligned}$$

L'étape $k \geq 1$ ayant été effectuée, on procède à l'étape $k+1$:

$$\begin{aligned} C^{(k+1)} &= \left\{j \in S^{(k)},\ \text{IC}\left(S^{(k)} \setminus \{j\}\right) \leq \text{IC}_{\text{ref}}^{(k)}\right\} \\ J^{(k+1)} &= \text{Argmin}\left(\text{IC}\left(S^{(k)} \setminus \{j\}\right),\ j \in C^{(k+1)}\right) \\ S^{(k+1)} &= S^{(k)} \setminus \left\{J^{(k+1)}\right\} \\ \text{IC}_{\text{ref}}^{(k+1)} &= \text{IC}(S^{(k+1)}). \end{aligned}$$

A une certaine étape $k_f + 1$, on obtiendra $C^{(k_f+1)} = \emptyset$. Cela signifie que le critère juge qu'il n'y a plus de composantes inutiles dans $S^{(k_f)}$. Nous arrêtons donc la procédure et choisissons ce dernier support comme estimation de $S^\star$.

Notons que toutes les quantités $C^{(.)}, J^{(.)}, S^{(.)}, \text{IC}_{\text{ref}}^{(.)}, k_f$ produites par la méthode comparative descendante sont aléatoires. La complexité de cette méthode est donc elle aussi aléatoire, mais elle est bornée par $m(m+1)/2$.

2.4.4 Complexités des méthodes

Le tableau 2.2 résume les complexités des méthodes décrites. Les méthodes comparatives présentent toutes une complexité polynômiale en le nombre total de paramètres choisis pour la sélection de modèles. Cela assure que leur utilisation sera plus rapide après implémentation sur machine. De plus, leur utilisation permet la sélection de n'importe quel support contenu dans $[\![1, m]\!]$ et fournit donc une description du modèle aussi fine que la méthode globale.

Résumé

Au cours de ce chapitre, nous avons présenté l'idée générale des critères d'information et dressé un historique de leurs apparitions et utilisations. Nous avons rencontré les premières justifications de certains critères, AIC (2.9) et φ_β (2.10), basées sur les considérations respectives de minimisation de risques et de stabilité asymptotique.

Des méthodes comparatives sont nouvellement introduites en partie 2.4. Elles peuvent être utilisées avec n'importe quel critère et répondent à un souci de complexité d'utilisation. Au cours de ce mémoire, nous nous efforcerons de les utiliser et de montrer que leur efficacité égale celle des méthodes classiques, bien plus coûteuse en temps de calcul.

Chapitre 3

Les modèles auto-régressifs gaussiens

On dit d'une suite de variables aléatoires réelles $X = (X_t)_{t\in\mathbb{N}}$ que c'est un processus auto-régressif unidimensionnel si pour tout $t \in \mathbb{N}$ elle vérifie la relation

$$X_t = -\sum_{i\geq 1} a_i X_{t-i} + E_t, \tag{3.1}$$

où $(a_i)_{i\geq 1}$ est une suite déterministe de réels presque nulle et le processus $E = (E_t)_{t\in\mathbb{N}}$ est un bruit blanc gaussien de variance σ^2 appelé excitation ou innovation. Par convention, on pose $X_t = 0$ pour $t < 0$.

Le dernier rang k pour lequel la suite (a) ne s'annule pas est appelé l'ordre d'autorégression, ainsi X_t dépend linéairement de $X_{t-1}, \ldots, X_{t-k}$ mais pas de X_{t-k-1}. Les a_i, $i = 1, \ldots, k$ sont appelés les coefficients d'autorégression.

Plus précisément, on pourra s'intéresser au support d'autorégression, *i.e.* l'ensemble

$$S = \{j \in \mathbb{N}^\star,\ a_j \neq 0\} \subset [\![1, k]\!]. \tag{3.2}$$

La donnée de ce support est plus précise que la simple donnée de l'ordre puisqu'elle permet de savoir exactement de quelles variables passées X_t dépend linéairement.

Dans la suite nous fixons trois modèles autorégressifs (AR), tous stables au sens Entrée-Bornée Sortie-Bornée, notion pour laquelle nous renvoyons à l'ouvrage général de Benidir et Al. [BB99], et présentant une excitation de variance $\sigma^2 = 1$:

$$\begin{cases} \text{Un AR d'ordre 2 de coefficients (-0.5 -0.05)} \\ \text{Un AR d'ordre 8 de coefficients (0.2 0.2 -0.8 0 0 0.5 0.2 0.2)} \\ \text{Un AR d'ordre 15 de coefficients (0.5, 0.06, 0, \ldots, 0, 0.45)} \end{cases} \tag{3.3}$$

Ces modèles sont typiquement choisis dans la littérature pour leur nature ambiguë. Le modèle AR2 peut facilement être confondu avec un modèle d'ordre 1 à cause de la faible valeur de son deuxième paramètre. AR15 pourrait quant à lui être confondu avec un modèle d'ordre 1 ou 2 étant donné qu'on trouve 12 coefficients nuls après les deux premiers non-nuls.

On peut essayer de modéliser une séquence de données réelles x^n par un processus auto-régressif. Dans un premier temps la détermination de l'ordre de ce processus est le problème de sélection de modèles que nous abordons à l'aide des critères d'information. Dans un deuxième temps, nous essaierons de déterminer le support de régression.

3.1 Détermination de l'ordre

Il est naturel de travailler ici avec les modèles emboités $\Theta_0 \subset \Theta_1 \subset \cdots \subset \Theta_m$ comme définis en (1.10) où m est un ordre maximal fixé par l'utilisateur. Dans notre cadre le modèle auto-régressif Θ_k d'ordre k contient les suites réelles presque nulles dont le dernier rang de non-annulation est k auxquelles on adjoint le paramètre de variance σ^2. L'appartenance $\theta \in \Theta_k$ doit se comprendre comme $\theta \in \Theta_k$ et $\theta \notin \Theta_{k-1}$, un paramètre $\theta \in \Theta_k$ se résume alors à

$$\theta = (a_1, \ldots, a_k, \sigma^2),\ a_k \neq 0$$

et a donc $k+1$ composantes libres.

A partir des données, la résolution du système d'équations de Yule-Walker fournit une estimation

$$\widehat{\theta}(k) = (\widehat{a}_1(k), \ldots, \widehat{a}_k(k), \widehat{\sigma}^2(k))$$

des paramètres au sens du maximum de vraisemblance. C'est donc le modèle autorégressif d'ordre k ayant ces paramètres qui donne la meilleure adéquation aux données x^n. De plus l'opposé de la log-vraisemblance de x^n vis-à-vis de ce modèle Θ_k est

$$2l(\widehat{\theta}(k)) = n \log \widehat{\sigma}^2(k).$$

Pour plus de détails sur ces résultats classiques, nous renvoyons à Akaike [Aka69]. Par conséquent la forme générale des critères d'information (2.1) devient

$$\mathrm{IC}(\Theta_k) = n \log \widehat{\sigma}^2(k) + (k+1)\alpha(n) \tag{3.4}$$

et la sélection de l'ordre d'autorégression se fait par

$$\widehat{k} = \mathrm{Argmin}(\mathrm{IC}(\Theta_k),\ k = 0, \ldots, m) \tag{3.5}$$

3.1.1 Simulations

Nous générons une réalisation x^n des processus détaillés en (3.3) et testons l'efficacité du critère (3.4) pour la reconnaissance de l'ordre du modèle. Les pénalités de ces critères se trouvent dans le tableau 2.1. Pour comparer les critères d'information à la méthode du maximum de vraisemblance, nous appelons MV le terme de vraisemblance dans (3.4) :

$$\mathrm{MV}(\Theta_k) = n \log \widehat{\sigma}^2(k).$$

Nous verrons que cette quantité induit de la surparamétrisation.

Les figures 3.1, 3.2, 3.3 et 3.4 présentent l'allure des critères en fonction de k sur les différents modèles considérés (3.3). Chaque critère sélectionne un ordre selon (3.5), en d'autres termes l'ordre sélectionné est l'endroit où la courbe atteint son minimum.

Comme attendu, la méthode du maximum de vraisemblance (courbe MV) choisit toujours un ordre trop élevé. Le critère BIC quant à lui montre un bon comportement. Les critères φ_β sont également efficaces à condition que la valeur de β soit bien choisie. Si elle est trop faible, le terme de pénalité parvient à peine à redresser

la surparamétrisation induite par la vraisemblance tandis que si elle est trop forte c'est la pénalité qui l'emporte et on observe une sous-paramétrisation.

C'est par exemple le cas sur la figure 3.1 avec $\beta = 0.5$: le critère choisit l'ordre $\widehat{k} = 3$ au lieu de 8. Cela tient également à la forme des coefficients de AR8 (3.3) qui peuvent laisser croire à une autorégression d'ordre 3, ou même à un ordre 6 puisque les 7ème et 8ème coefficients sont faibles. Toutes les courbes de cette figure reflètent ce fait.

Le modèle AR15 est également délicat puisqu'il pourrait se confondre avec un ordre 1 ou 2. Sur la figure 3.2, le critère $\varphi_{0.3}$ préfère de peu l'ordre 15 à l'ordre 2 tandis que $\varphi_{0.4}$ choisit $\widehat{k} = 1$.

La figure 3.3 montre une faible variation des critères, la décision des critères est donc à considérer avec précaution. Même si les coefficients de AR2 peuvent laisser croire à un ordre 1, AIC et $\varphi_{0.1}$ parviennent à déterminer le bon ordre. Les autres sous-paramétrisent tandis que MV continue à sur-paramétriser. En augmentant substantiellement le nombre d'échantillons jusqu'à $n = 10.000$ sur la figure 3.4, on voit apparaître une décision plus nette des critères ; AIC restant à la limite de la sur-paramétrisation de MV et $\varphi_{0.3}$ proche de la sous-paramétrisation.

Notons que sur les figures 3.1, 3.2 et 3.3, on a $n = 500$, $\beta_{\min} \approx 0,3$ et $\beta_{\max} \approx 0,7$ par l'équation (2.12). Le critère $\varphi_{\beta_{\min}}$ est donc similaire à la courbe $\varphi_{0,3}$. Le critère $\varphi_{\beta_{\max}}$ présente quant à lui une forte sous-paramétrisation. Sur la figure 3.4, on a choisi $n = 10.000$, ce qui amène $\beta_{\min} \approx 0,24$; le critère correspondant est donc encore efficace tandis que $\varphi_{\beta_{\max}}$ sous-paramétrise.

Ces simulations mettent bien en avant la sur-paramétrisation induite par la méthode du maximum de vraisemblance. De plus AIC est toujours proche d'avoir ce défaut pour n assez grand du fait que sa pénalité reste constante et s'efface devant le terme de vraisemblance. Le critère BIC, quant à lui, présente souvent un bon comportement. Le critère φ_β est plus délicat à utiliser car le réglage du paramètre β est critique. Il fait l'objet des figures 3.5 et 3.6.

Nous travaillons dans un premier temps avec le processus AR8. Comptons une réussite si le critère choisit $\widehat{k} = 8$ par (3.5). La figure 3.5 présente le taux de réussite sur 100 expériences en fonction de β. On y trouve les valeurs particulières de β définies en (2.11) et (2.12). On constate que le critère AIC est toujours moins efficace que BIC. A mesure que n augmente, ce dernier, ainsi que $\varphi_{\beta_{\min}}$, donne 100% de réussite comme le théorème 2.3.1 le prédisait. Par contre, la valeur $\beta_{\max}$ donne de mauvais résultats alors qu'elle satisfait les conditions de ce même théorème. La vitesse de convergence est donc lente et nous n'avons pas réussi à mettre en évidence un bon comportement de $\varphi_{\beta_{\max}}$ même pour des tailles très élevées d'échantillon. La plage de valeurs de β qui donnent un taux de réussite de 100% dépend apparemment de certains facteurs, dont le modèle autorégressif lui-même. Pour comparaison, nous effectuons le même travail avec AR15 et présentons les résultats sur la figure 3.6.

3.2 Détermination du support

On s'intéresse maintenant au support du processus autorégressif modélisant les données x^n. Fixons $[\![1, m]\!]$ comme support d'autorégression maximal. Pour $S \subset [\![1, m]\!]$ un support fixé, une simple adaptation de la résolution des équations de Yule-Walker permet de déterminer les paramètres $\widehat{\theta}(S)$ au sens du maximum de

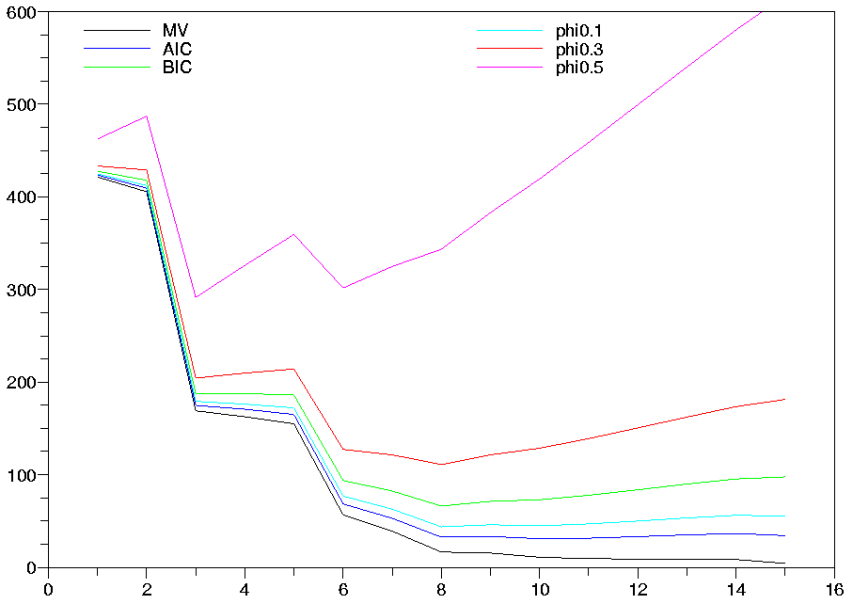


FIG. 3.1 – Allure des critères sur une réalisation de AR8 en fonction de k, n=500. Ici, $\beta_{\min} \approx 0,3$ et $\beta_{\max} \approx 0,7$.

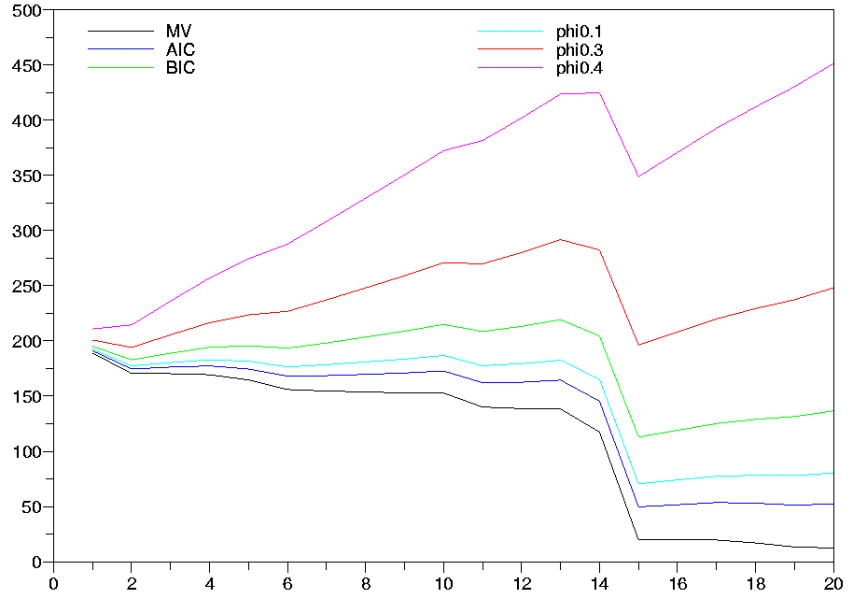


FIG. 3.2 – Allure des critères sur une réalisation de AR15 en fonction de k, n=500. Ici, $\beta_{\min} \approx 0,3$ et $\beta_{\max} \approx 0,7$.

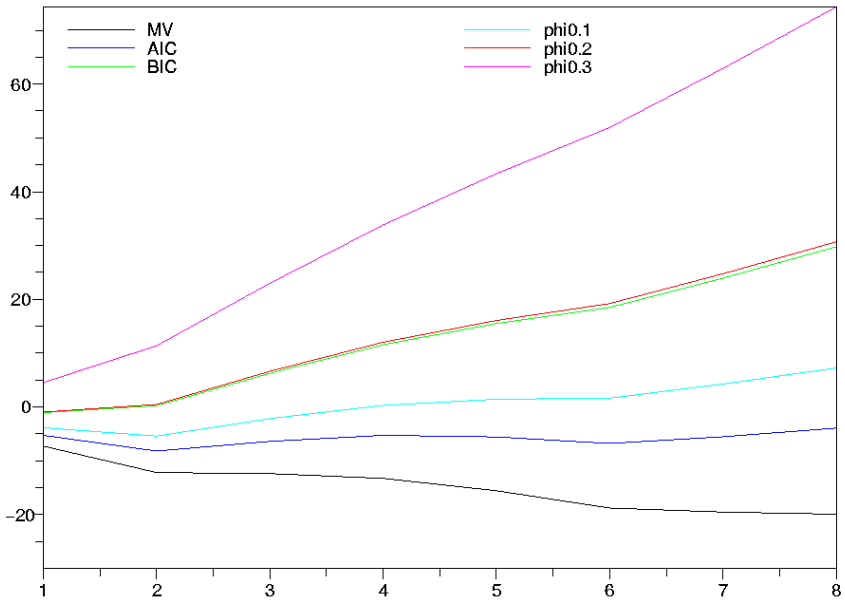


FIG. 3.3 – Allure des critères sur une réalisation de AR2 en fonction de k, n=500. Ici, $\beta_{\min} \approx 0,3$ et $\beta_{\max} \approx 0,7$.

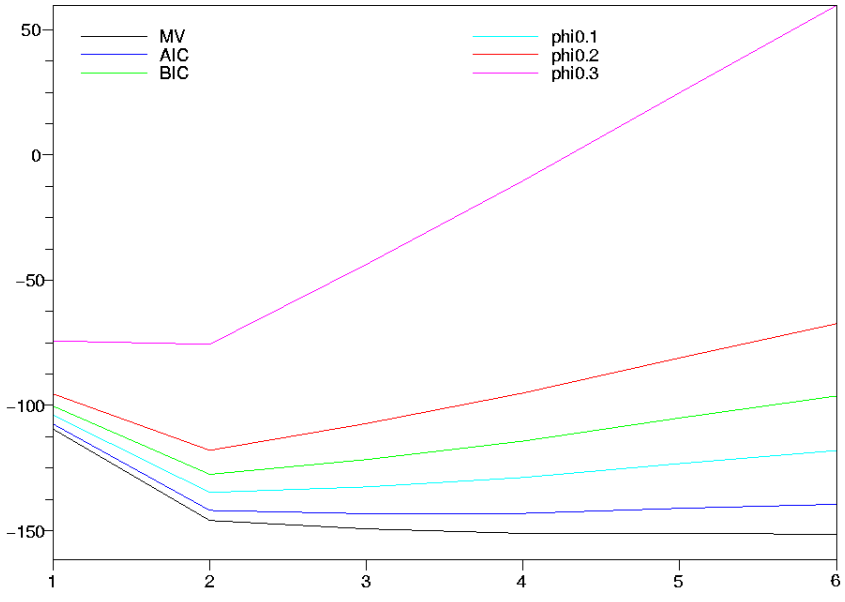
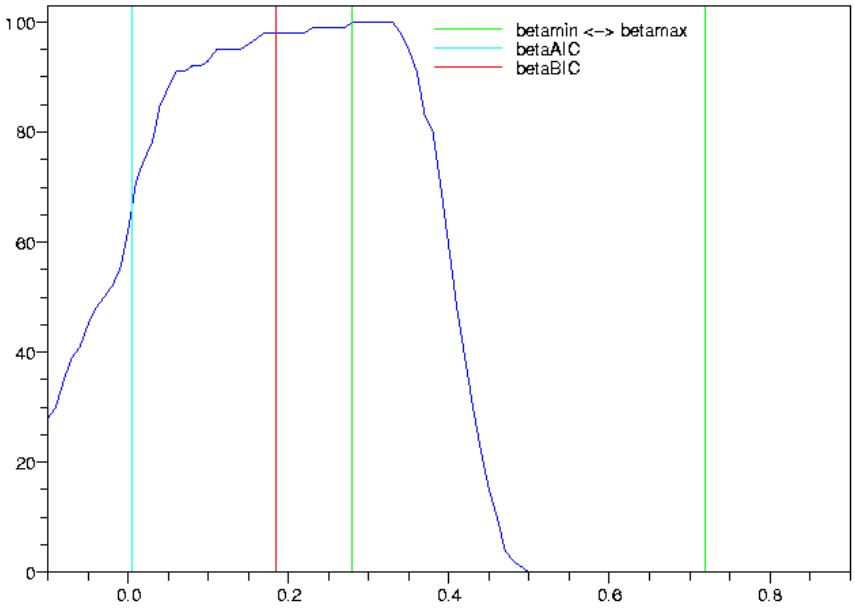
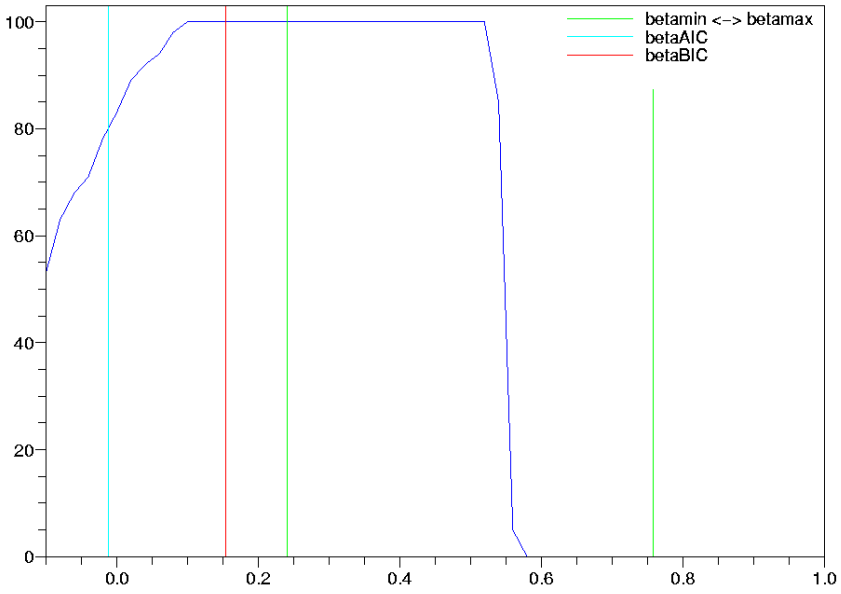


FIG. 3.4 – Allure des critères sur une réalisation de AR2 en fonction de k, n=10.000. Ici, $\beta_{\min} \approx 0,24$ et $\beta_{\max} \approx 0,76$.

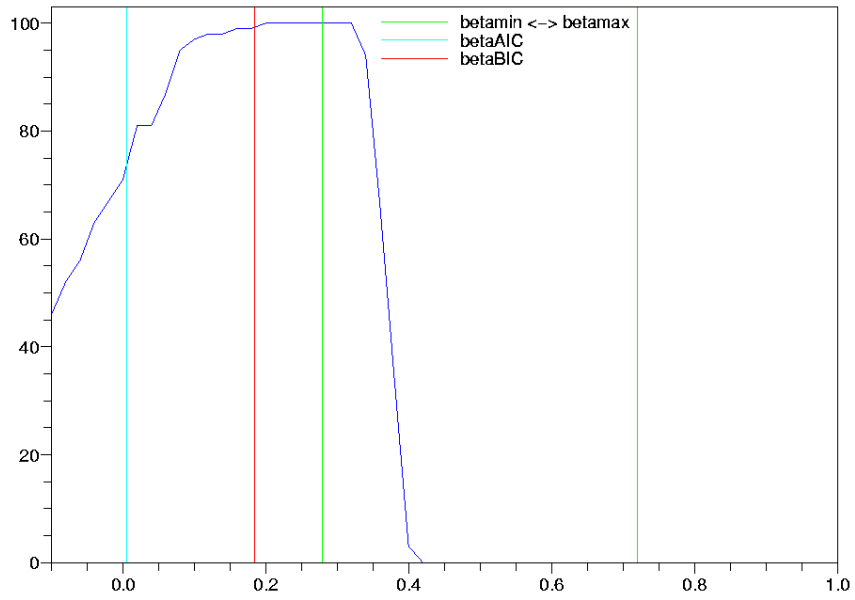


$n = 1000$

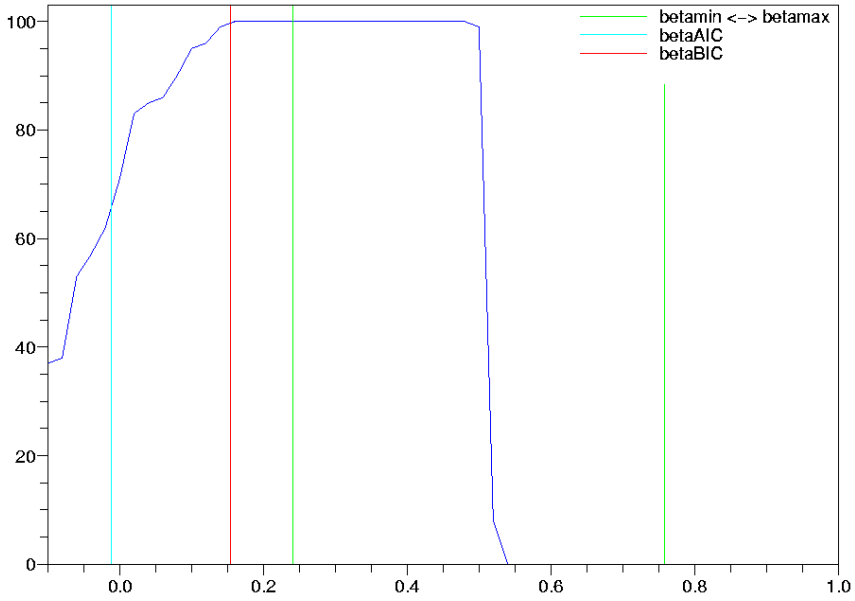


$n = 10.000$

FIG. 3.5 – AR8 : pourcentage de sélection du bon ordre par (3.5) avec le critère φ_β (2.10) en fonction de β.



$n = 1000$



$n = 10.000$

FIG. 3.6 – AR15 : pourcentage de sélection du bon ordre par (3.5) avec le critère φ_β (2.10) en fonction de β.

vraisemblance :

$$\widehat{\theta}(S) = (\widehat{a}_j,\ j \in S,\ \widehat{\sigma}^2(S)).$$

Les critères s'écrivent alors

$$\mathrm{IC}(S) = n \log \widehat{\sigma}^2(S) + |S+1|\alpha(n).$$

On peut utiliser ici les différentes méthodes abordées dans la partie 2.4 pour la détermination du support. Nous étudions les résultats donnés par la méthode globale (2.13) et la méthode comparative (2.14).

3.2.1 Simulations

Le modèle utilisé est AR8 pour n =1000 et 100.000, il a pour support $S^\star = \{1,2,3,6,7,8\}$. Nous générons 100 réalisations de ce processus et appliquons à chacune d'elles le critère φ_β pour β variant dans $]0,1[$ avec les deux méthodes, globale (2.13) et comparative (2.14). Un succès est compté si $\widehat{S} = S^\star$. Le nombre de succès pour chaque méthode est donné en fonction de β par la figure 3.7.

Rappelons que la méthode comparative nécessite le calcul de seulement $m+1$ critères quand la méthode globale en demande 2^m. La méthode comparative est moins efficace pour des échantillons modérés (n =1000) ; les taux de succès de BIC et $\varphi_{\beta_{\min}}$ étant tout de même de l'ordre de 75%. Pour des échantillons plus étendus, les taux de succès sont comparables.

3.3 Le cas bidimensionnel

La modélisation d'images en niveau de gris peut être envisagée à l'aide de processus autorégressifs bidimensionnels : AR2D.

Une image en niveau de gris est un ensemble de pixels, ou sites, repérés par leur coordonnées (t,u) dans le réseau $\mathbb{Z}^2$ à chacun desquels est associé un niveau de gris, nombre entre 0 et 1. Le niveau de gris 0 signifie que le pixel est noir, le niveau 1 désigne un pixel blanc. Chaque réel de $[0,1]$ ne correspond pas à un niveau de gris, cependant on confondra $x \in [0,1]$ avec le niveau de gris qui lui est le plus proche.

Les processus AR2D se présentent sous la forme d'une famille $X_{t,u}$, $(t,u) \in \mathbb{Z}^2$ de variables à valeurs dans $[0,1]$ liée par une relation du type

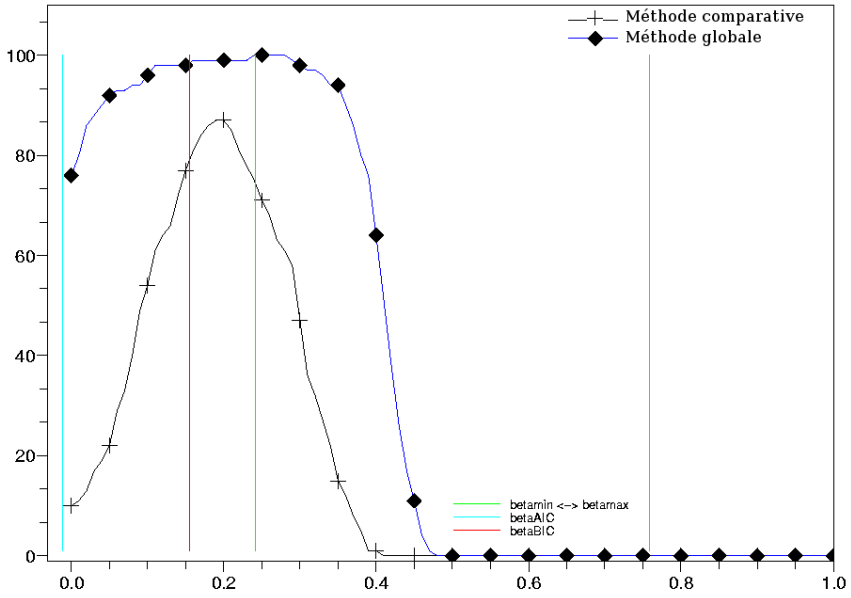
$$X_{t,u} = -\sum_{(t',u') \in S} a_{t',u'} X_{t-t',u-u'} + E_{t,u} \tag{3.6}$$

où E est un bruit blanc gaussien de variance σ^2 et S est appelé le support d'autorégression.

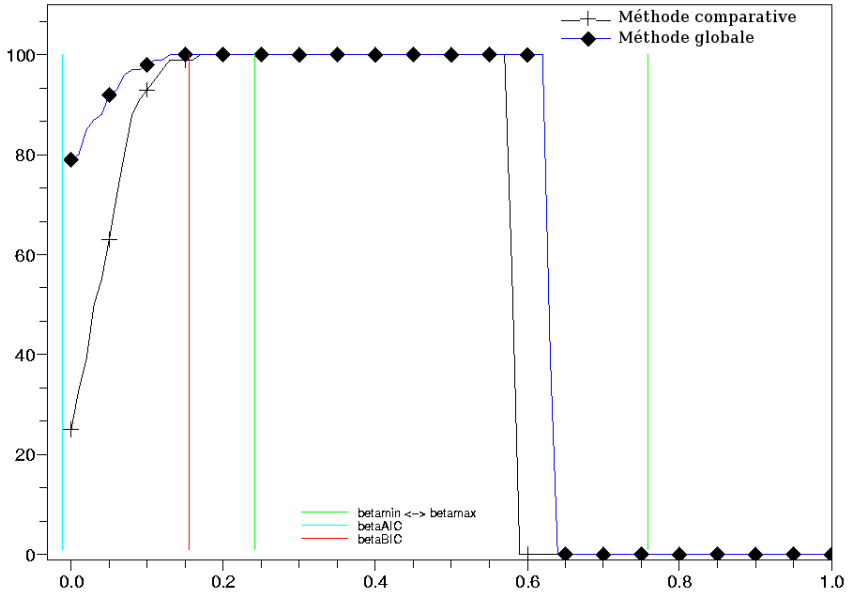
Il nous faut préciser ce que l'on entend par support S. Choisissons pour support d'autorégression maximal la partie $[\![0,m]\!] \times [\![0,m]\!] \setminus (0,0)$ de $\mathbb{Z}^2$. Un support sera pour nous de la forme

$$S = \big\{(t_j,u_j) \in [\![0,m]\!]^2,\ j = 1,\ldots,|S|\big\} \setminus (0,0). \tag{3.7}$$

Quand nous travaillerons sur des images, nous négligerons les problèmes de bord en supposant que pour un site (t,u) et un support S fixé, le site $(t-t',u-u')$



$n = 1000$



$n = 100.000$

FIG. 3.7 – AR8 : Pourcentage de succès des deux méthodes (2.13) et (2.14) en fonction de β.

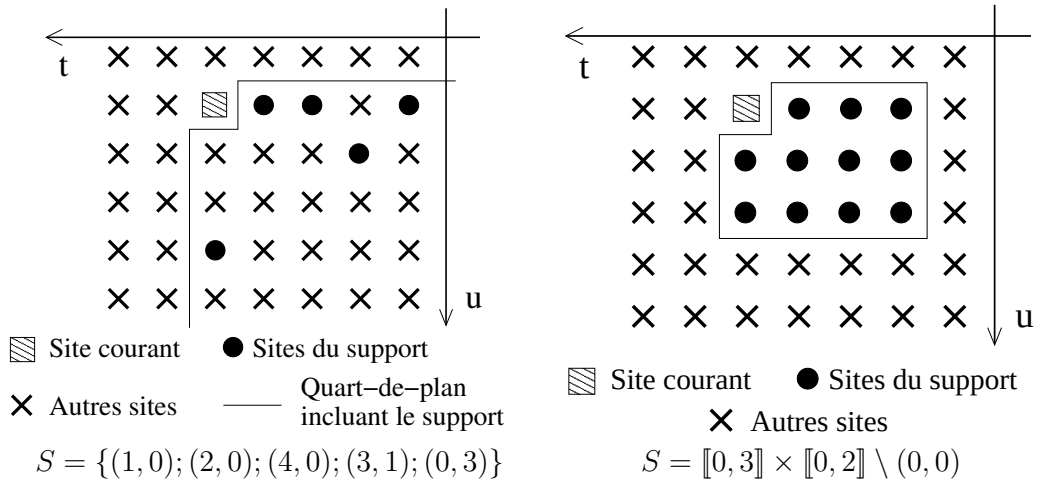


FIG. 3.8 – Exemples de supports AR2D.

appartient à l'image pour tout $(t', u') \in S$. Si l'image est assez grande, les sites pour lesquels cela n'est pas réalisé sont très minoritaires. La valeur $X_{t,u}$ d'un processus AR2D en le site (t, u) dépend donc linéairement des valeurs de ce processus en des sites situés dans un quart de plan dont le site courant (t, u) est le sommet. Notons que dans, (3.7), nous interdisons l'appartenance de $(0, 0)$ à un support S pour que la valeur en un site ne dépende pas d'elle-même. Si nous choisissons S de la forme $[\![0, k_t]\!] \times [\![0, k_u]\!] \setminus (0, 0)$ avec $k_t, k_u \leq m$, le support d'autorégression est un rectangle. Un tel support est à mettre en relation avec la notion d'ordre d'une autorégression unidimensionnelle. Les supports généraux (3.7) correspondent à la notion également nommée "support" dans le cas unidimensionnel. La figure 3.8 clarifie ces notations.

On trouve dans la littérature [RJ85, AR05] d'autres géométries pour les supports d'autorégressions bidimensionnelles : *causal Non-Symetrical Half Plane* (NSHP), *semi-causal* ou *Non-Causal* (NC). Notre choix s'est porté vers les supports de type QP (pour *causal Quarter Plane*) car ils sont les plus simples à utiliser.

3.3.1 Illustration sur des données réelles

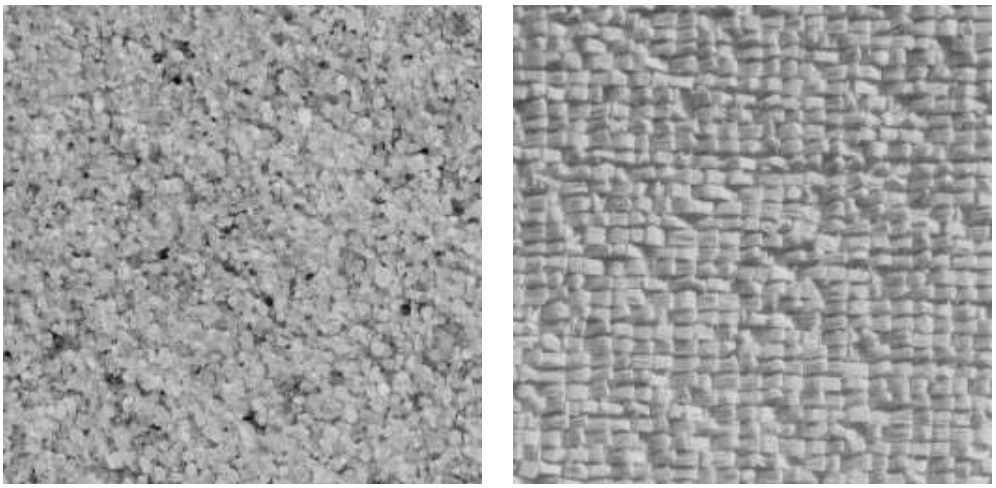
Nous considérons les images de textures d29 et d84 issues de l'album de Brodatz [Bro66] présentées en figure 3.9. Nous modélisons ces images par des processus autorégressifs bidimensionnels. La détermination du support de ces processus est le problème de sélection de modèle que nous envisageons avec le critère φ_β pour $\beta = \beta_{\min}$ défini en (2.12). Fixons comme support d'autorégression maximal $S_m = [\![0, m]\!] \times [\![0, m]\!] \setminus (0, 0)$ avec $m = 18$.

Pour $S \subset S_m$ un support fixé, l'estimation des paramètres $\widehat{a}_{t',u'}(S)$, $(t', u') \in S$, $\widehat{\sigma}^2(S)$ du modèle (3.6) au sens du maximum de vraisemblance se fait à nouveau à l'aide de la méthode de Yule-Walker et le critère utilisé s'écrit

$$\varphi_\beta(S) = n \log \widehat{\sigma}^2(S) + (|S| + 1) n^\beta \log \log n.$$

Nous effectuons la sélection de supports rectangulaires par une méthode globale :

$$\widehat{S}_r = \text{Argmin} \left(\varphi_\beta(S),\ S = [\![0, k_t]\!] \times [\![0, k_u]\!] \setminus (0, 0),\ k_t, k_u \in [\![0, m]\!] \right). \tag{3.8}$$



texture d29 texture d84

FIG. 3.9 – Les deux images de textures considérées.

Ensuite, nous envisageons la sélection de supports quelconques par une méthode de type comparative (2.14) qui s'écrit ici :

$$\begin{aligned} \varphi_{\beta,\text{ref}} &= \varphi_\beta(S_m) \\ \widehat{S}_c &= \{(t,u) \in S_m,\ \varphi_\beta(S_m) \leq \varphi_\beta\left(S_m \setminus (t,u)\right)\} \end{aligned} \tag{3.9}$$

Notons que les travaux présentés ici font suite à ceux de Alata et Al. [AR05] sur la segmentation d'images texturées, la sélection d'un support d'autorégression plus précis via (3.9) n'étant pas abordée dans ces derniers.

Les supports ainsi sélectionnés sont présentés sur la figure 3.10. Lorsque l'on se restreint aux supports rectangulaires, on conserve des sites qui sont éliminés par la méthode comparative puisque cette dernière permet une géométrie plus souple des supports. A l'inverse, certains sites conservés par la méthode comparative ne sont pas inclus dans le support rectangulaire puisque ce dernier présenterait trop de paramètres libres s'il voulait les englober. Le support sélectionné par la méthode comparative est donc plus précis.

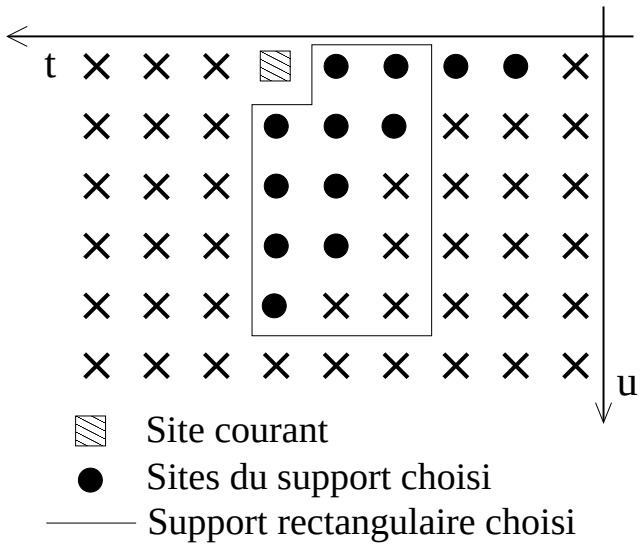
Résumé

Le problème de la description d'une autorégression est le terrain privilégié pour l'utilisation des critères d'information, c'est pourquoi nous l'avons traité dans ce premier chapitre applicatif.

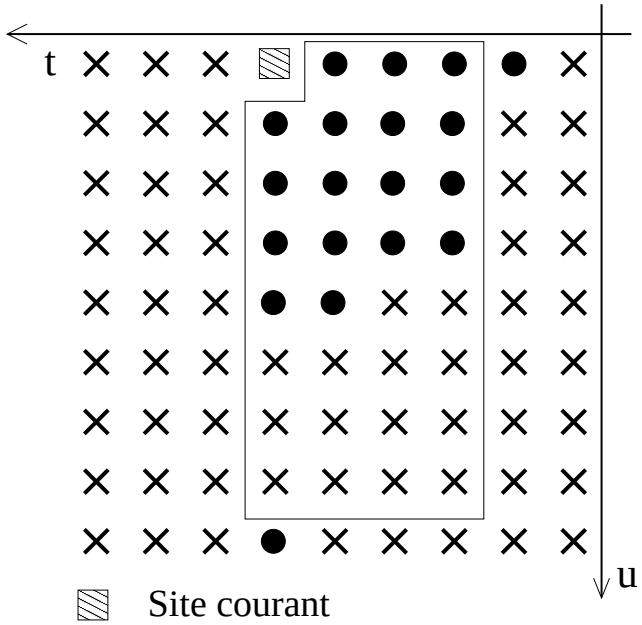
Outre la détermination usuelle de l'ordre d'autorégression, nous nous intéressons à la description, plus précise, de son support (3.2). A cette fin, nous exploitons le critère φ_β (2.10), malléable grâce à sa pénalité et peu utilisé dans la littérature, ainsi que les méthodes comparatives introduites en section 2.4.

Ces considérations, transposées aux images ou textures comme en section 3.3, peuvent ouvrir la voie à de nouvelles méthodes de traitement. La forme du support

d'autorégression choisi par les critères, présentée par exemple en figure 3.10, constitue en effet un facteur discriminant entre plusieurs images de textures. Ce facteur pourrait être utilisé, par exemple, pour résoudre des problèmes de reconnaissance de forme.



Texture d29



Texture d84

FIG. 3.10 – Supports rectangulaires et quelconques d'autorégression 2D choisis par les méthodes (3.8) et (3.9) pour les textures d29 et d84.

Chapitre 4

Les chaînes de Markov multiples

Dans ce chapitre, nous abordons le problème de la détermination de l'ordre d'une chaîne de Markov multiple à l'aide des critères d'information (2.1).

Pour cela, nous rappelons dans un premier temps le concept de codage et les liens connus qu'il entretient avec les probabilités ; ces liens sont plus développés, par exemple, par Berstel et Al. [BP85]. Nous nous intéressons ensuite à la notion de complexité stochastique, introduite de manière générale par Rissanen [Ris86b, Ris86a, Ris89], que nous interprétons dans notre cadre comme un codage *adaptatif*. Différentes inégalités entropiques sont interprétées en terme de codage au cours des parties 4.3, 4.4 et 4.5.

L'application du principe de *Minimum Description Length* [Ris78, GMP05] à ces codages nous amènera, en partie 4.6, à la justification de l'utilisation du critère BIC (voir tableau 2.1) pour le problème posé.

Le logarithme est en base 2 tout au long de ce chapitre.

4.1 Codes et probabilités

Considérons E un ensemble fini de $m \geq 2$ symboles. Un message est une séquence $x^n = x_1, \ldots, x_n$ d'éléments de E qu'un utilisateur souhaite transmettre à un autre utilisateur. Puisque les moyens actuels pour effectuer une telle tâche sont principalement de nature informatique, il vient à l'idée de coder le message en une séquence de bits (0 ou 1). Le code devra être convenu à l'avance par les deux utilisateurs et la séquence de bits, une fois reçue, devra être décodable. Donnons quelques définitions permettant la modélisation d'une telle expérience.

4.1.1 Les codes préfixes

Définition 4.1.1 *Un code sur E est une application injective*

$$\begin{array}{rcl} C : E & \longrightarrow & \bigcup\limits_{i \in \mathbb{N}^*} \{0,1\}^i \\ x & \mapsto & C(x). \end{array}$$

La séquence de bits $C(x)$ est appelée le code de x ; le nombre de bits qu'elle contient est appelé la longueur de code de x et noté $L(x)$.

Une fois une telle application choisie, l'idée vient de coder le message x^n en concaténant les codes de chacun des x_i, $i = 1, \ldots, n$. L'exemple suivant montre que, pour que cette méthode soit efficace, il nous faudra imposer une autre condition sur le code C. Prenons $E = \{a, b, c\}$ et

$$C(a) = 0,\ C(b) = 1,\ C(c) = 01. \tag{4.1}$$

Codons maintenant les messages "abc" et "cc" :

$$C(abc) = 0101,\ C(cc) = 0101.$$

Les codes de ces messages sont identiques ; le décodage n'est pas possible. Pour pallier ce problème, introduisons la notion de code préfixe.

Définition 4.1.2 *Soit C un code sur E.*

- *On dit qu'un mot de code $C(x)$ est le préfixe d'un autre mot $C(y)$ si $L(x) \leq L(y)$ et si la séquence de bits $C(x)$ se trouve au début de la séquence $C(y)$.*
- *On dira que C est un code préfixe si aucun des mots de code qui le compose n'est le préfixe d'un autre mot de code.*

Voici un code préfixe pour l'exemple précédent :

$$C(a) = 00,\ C(b) = 1,\ C(c) = 01. \tag{4.2}$$

On obtient

$$C(abc) = 00101,\ C(cc) = 0101.$$

Pour effectuer le décodage, on parcourt la séquence de bits dans l'ordre chronologique en écrivant, à chaque fois qu'un mot de code est reconnu, le symbole correspondant. Le choix d'un code préfixe assure donc que la concaténation des codes des différents symboles du message est décodable.

A travers notre exemple nous voyons que le code préfixe (4.2) nécessite plus de bits que le code non préfixe (4.1), ce constat est renforcé par le résultat usuel suivant, pour lequel nous renvoyons à [BP85].

Proposition 4.1.3 *Soit L une fonction sur E à valeurs dans $\mathbb{N}^*$. On a équivalence entre :*

(i) La fonction L vérifie l'inégalité de Kraft :

$$\sum_{x \in E} 2^{-L(x)} \leq 1 \tag{4.3}$$

(ii) Il existe un code préfixe C sur E tel que pour tout $x \in E$ la longueur de la séquence de bits $C(x)$ soit $L(x)$.

On voit donc que la longueur d'un code préfixe, contrainte à vérifier (4.3), ne saurait être trop courte. Cette inégalité établit de plus un parallèle entre code préfixe sur E et distribution de probabilité sur E comme le montre la partie suivante.

4.1.2 Le lien avec les probabilités

Soit maintenant P une probabilité sur E muni de la tribu de toutes ses parties. On suppose que P charge au moins deux points distincts. Posons

$$L = \lceil -\log P \rceil \tag{4.4}$$

où $\lceil . \rceil$ désigne la partie entière supérieure. Alors L est à valeurs dans $\mathbb{N}^\star$ et

$$\sum_{x \in E} 2^{-L(x)} \leq \sum_{x \in E} 2^{\log P(x)} = 1.$$

En d'autres termes, L vérifie l'inégalité de Kraft (4.3) ; elle est par conséquent la longueur d'un certain code préfixe C. Partant d'une probabilité P, on aboutit donc à un code préfixe, pas nécessairement unique, de longueur L, le lien entre ces deux objets étant donné par (4.4). L'inégalité de Kraft (4.3) donne donc un lien étroit entre probabilités et codes préfixes. Dans la suite nous confondrons souvent une probabilité avec "le" code qui lui est associé. La quantité à laquelle nous nous intéresserons principalement est la longueur du code L, qui est unique et donnée par (4.4), plutôt que le code en lui-même qui ne l'est pas. On dira par conséquent que l'on code des messages "avec la probabilité P" ou "avec un P-code" pour signifier que l'on utilise un code préfixe de longueur $L = \lceil -\log P \rceil$. A travers cette correspondance, un symbole de forte probabilité se voit attribuer un code court.

Lorsque cela sera commode, on omettra la partie entière dans l'expression (4.4). Dans ce cas la fonction L n'est plus à valeurs entières et ne saurait être la longueur d'un code au sens de la définition 4.1.1. On parle alors de longueurs de codes idéalisées. La différence entre la longueur d'un code et sa longueur idéalisée n'excède jamais 1 bit.

4.1.3 Inégalités entropiques et codages

La notion d'entropie (1.1), présentée au début de ce mémoire, a une interprétation importante dans le cadre du codage tel qu'on l'aborde ici. Rappelons que l'entropie d'une probabilité P sur E est définie par

$$H(P) = -\sum_{x \in E} P(x) \log P(x) \tag{4.5}$$

tandis que l'entropie croisée entre deux distributions P et Q s'écrit

$$H(P, Q) = -\sum_{x \in E} P(x) \log Q(x). \tag{4.6}$$

Supposons que les symboles constituant le message à encoder soient des réalisations indépendantes de la loi P. L'entropie croisée $H(P, Q)$ est alors la moyenne de la longueur (idéalisée) du Q-code d'un symbole. L'entropie $H(P)$ représente quant à elle la longueur moyenne obtenue en codant avec la probabilité P. Alors l'inégalité de convexité de Jensen :

$$H(P) \leq H(P, Q) \tag{4.7}$$

affirme qu'aucun code préfixe sur E ne donne une longueur moyenne de codage plus faible que celui associé à la distribution P elle-même. Shannon est le premier a

donner dans [Sha48] cette interprétation de l'inégalité (4.7), c'est pourquoi on parle dans le domaine du traitement du signal d'entropie de Shannon et d'inégalité de Shannon.

Plus l'entropie d'une distribution est forte, plus il est coûteux en terme de bits de coder une séquence de symboles qu'elle régit. Le cas extrême est donné par la distribution uniforme P_u sur $E = \{1, \dots, m\}$. Prenons m de la forme 2^d pour clarifier la situation. L'entropie de P_u vaut :

$$H(P_u) = -\sum_{i=1}^{m} \frac{1}{m} \log \frac{1}{m} = \log m = d.$$

Le code C qui associe à chaque entier i l'écriture binaire de $i-1$ avec d bits vérifie la propriété du préfixe. Par exemple, pour $m = 8 = 2^3$, on a $C(3) = 010$. L'inégalité de Jensen affirme qu'il est inutile d'essayer de coder des données distribuées de manière uniforme avec un autre code préfixe.

4.1.4 Codage de Huffman

Le but de cette section n'est pas de présenter l'algorithme de codage de Huffman pour lequel nous renvoyons à l'auteur lui-même [Huf52]. Notons que cet algorithme est dérivé de la preuve, constructive, de la proposition 4.1.3. Nous essayons ici de faire une distinction entre l'aspect théorique du codage et son aspect pratique, ce dernier étant le codage de Huffman.

Soit P une probabilité inconnue régissant les données à encoder. Au regard de l'inégalité (4.7), l'utilisateur qui veut utiliser le moins de bits possibles pour le codage doit trouver une probabilité Q dont l'entropie croisée $H(P, Q)$ se rapproche au maximum de $H(P)$. En d'autres termes, il doit trouver Q minimisant l'information de Kullback $K(P, Q)$ définie en (1.3). A cet effet, la probabilité empirique $\widehat{P}$ estimée au sens du maximum de vraisemblance est un bon candidat. C'est là ce que nous appelons l'aspect théorique du codage.

Maintenant, en pratique, la difficulté pour l'utilisateur est de construire un $\widehat{P}$-code, *i.e.* un code préfixe dont la longueur L est donnée par $L = \left\lceil -\log \widehat{P} \right\rceil$. C'est l'algorithme de Huffman qui permet une telle construction.

Nous voyons donc le codage de Huffman comme un outil pratique permettant d'implémenter l'idée théorique selon laquelle la probabilité empirique $\widehat{P}$ se rapproche de P au sens entropique. Nous retrouverons cette même distinction dans la suite entre la notion, théorique, de complexité stochastique et celle, pratique, de codage arithmétique. Le cadre qui se prête le mieux à l'introduction de ces notions est celui des modèles de chaînes de Markov multiples que nous présentons maintenant.

4.2 Les chaînes de Markov multiples

Définition et interprétation en termes de codage

Un processus $(X_n)_{n \in \mathbb{N}^*}$ est une chaîne de Markov multiple d'ordre $k \in \mathbb{N}$ (abrégée en CMM(k)) si k est le plus petit entier satisfaisant l'égalité en lois

$$\mathbb{P}(X_n | X_{n-1}, \dots, X_1) = \mathbb{P}(X_n | X_{n-1}, \dots, X_{n-k})$$

pour tout n. Nous nous placons de plus dans le cas homogène où cette loi ne dépend pas de n. Une CMM(0) est une suite de variables indépendantes.

Nous supposons que les k premières variables $X_1, \ldots, X_k$ d'une CMM(k) sont indépendantes et identiquement distribuées de loi uniforme sur E. Pour $i \in E$ un état et $j \in E^k$ un état multiple, on désigne par $\theta(i|j)$ la probabilité de voir apparaître i après j. La donnée des $(m-1)m^k$ probabilités $\theta(i|j)$ pour $j \in E^k$ et i parcourant E privé d'un élément suffit donc à décrire l'évolution de la chaîne. Nous désignons par θ de telles données.

Vraisemblance

Si $x^n = x_1, \ldots, x_n$ est une séquence donnée d'éléments de E, la vraisemblance de x^n relativement à la CMM(k) décrite par un paramètre θ s'écrit

$$\mathbb{P}_\theta(x^n) = \frac{1}{m^k} \prod_{j \in E^k} \prod_{i \in E} \theta(i|j)^{n(i|j)} \tag{4.8}$$

où $n(i|j)$ est le nombre d'occurrence de l'état i après l'état j dans x^n. Par conséquent, un ordre k étant choisi, le paramètre $\widehat{\theta}_k$ de la CMM(k) estimé au sens du maximum de vraisemblance à partir des données x^n est

$$\widehat{\theta}_k(i|j) = \frac{n(i|j)}{n(j)}, \ i \in E, \ j \in E^k, \tag{4.9}$$

où $n(j)$ est le nombre d'occurrences de l'état multiple j dans x^n. Remarquons que $n(j)$ ne doit pas comptabiliser une éventuelle apparition de j à la fin de la séquence x^n. Dans le cas $k = 0$, les états j disparaissent, on remplace alors $n(i|j)$ par $n(i)$ et $n(j)$ par n. Notons que nous n'estimons pas les paramètres relatifs aux k premières valeurs $x_1, \ldots, x_k$ puisqu'elles sont supposées distribuées uniformément.

Le paramètre (4.9) permet d'atteindre une vraisemblance maximale dont l'opposé du logarithme est donné par la fonction l, au sens de (1.5), qui vaut ici

$$l\left(\widehat{\theta}_k\right) = -\sum_{j \in E^k} \sum_{i \in E} n(i|j) \log \frac{n(i|j)}{n(j)} + k \log m. \tag{4.10}$$

Entropie des CMM

Nous avons défini en (4.5) l'entropie d'une distribution sur un ensemble fini quelconque et donné son interprétation en terme de longueur de codage. Posant $F = E^n$, ensemble fini de taille m^n, une CMM(k), $k \geq 0$, décrite par un paramètre θ devient une distribution sur F. On peut définir son entropie d'une manière similaire :

$$H_n(\theta) = -\sum_{x^n \in E^n} \mathbb{P}_\theta(x^n) \log \mathbb{P}_\theta(x^n) \tag{4.11}$$

Cette quantité s'interprète comme la longueur moyenne du code d'un "symbole" $x^n \in F = E^n$ régi par la distribution θ donnée par (4.8), ce code étant défini sur F et associé à θ lui même. Dans cette optique, nous ne codons plus les éléments de x^n un par un mais nous effectuons le codage de la chaîne globalement. L'entropie H_n est indicée par n pour rappeler ce fait.

Considérons maintenant une distribution sur E^n notée $\mathbb{P}_{\theta'}$. On définit l'entropie croisée entre θ et θ' de manière similaire à (4.6) par

$$H_n(\theta, \theta') = -\sum_{x^n \in E^n} \mathbb{P}_\theta(x^n) \log \mathbb{P}_{\theta'}(x^n). \tag{4.12}$$

Comme précédemment, cette quantité est la longueur moyenne du $P_{\theta'}$-code d'un "symbole" x^n régi par P_θ et l'inégalité de convexité de Jensen/Shannon donne

$$H_n(\theta) \leq H_n(\theta, \theta'). \tag{4.13}$$

Notons que quand $k = k' = 0$, ces notions d'entropie et leurs interprétations en terme de codage sont similaires à celles exposées dans la partie 4.1 à un facteur n près. Dans ce dernier cas, nous avions expliqué au paragraphe 4.1.4 la position du codage de Huffman par rapport à l'inégalité entropique (4.7). Nous présentons maintenant le codage arithmétique dans la même optique.

4.3 Codage arithmétique

4.3.1 Situation

Supposons disposer d'une réalisation x^n d'une CMM d'ordre $k^\star$ de paramètre inconnu θ. Nous souhaitons coder le message x^n globalement, *i.e.* sans concaténer les codes de ses différents symboles. Au regard de l'inégalité de Shannon (4.13), il nous faut, pour coder le message avec un nombre réduit de bits, choisir un code sur E^n, donc une distribution sur E^n, dont l'entropie croisée avec $P(\theta)$ soit la plus proche possible de $H(\theta)$.

Désignons par k une estimation de l'ordre $k^\star$ de la CMM. Nous reviendrons plus en détail dans la partie 4.6 sur la façon de construire une telle estimation ; ce point étant en fait le coeur du problème abordé dans ce chapitre. La distribution $\widehat{\theta}_k$ estimée au sens du maximum de vraisemblance comme en (4.9) est un bon candidat. C'est encore une fois ce que nous appelons l'aspect théorique du codage.

Pour l'aspect pratique, il faut maintenant définir une procédure qui encode effectivement x^n en une chaîne de bits de longueur

$$L(x^n) = \left\lceil -\log P_{\widehat{\theta}_k} \right\rceil = \left\lceil l\left(\widehat{\theta}_k\right) \right\rceil \tag{4.14}$$

où l est définie en (4.10). Cette procédure, introduite par Rissanen dès 1976 [Ris76] est appelée codage arithmétique. Nous présentons maintenant son algorithme.

4.3.2 Algorithme

Contrairement au codage de Huffman, nous donnons ici une description de l'algorithme de codage arithmétique. Cette description se veut succincte et théorique en ce sens qu'elle ne tient pas compte des difficultés liées à l'implémentation sur machine de l'algorithme. Pour plus de détails nous renvoyons aux écrits de Witten et Al. [WNC87] ou de Moffat et Al. [MNW98] qui abordent le codage arithmétique d'un point de vue plus informatique.

TAB. 4.1 – Codage arithmétique à l'ordre 1 de la chaîne *abaa*.

t	x^t	$I_c(t)$	Partitionnement
0	$\emptyset$	$[0,1]$	$[0,1/2,1]$
1	a	$[0,1/2]$	$[0,1/4,1/2]$
2	ab	$[1/4,1/2]$	$[1/4,1/2,1/2]$
3	aba	$[1/4,1/2]$	$[1/4,3/8,1/2]$
4	$abaa$	$[1/4,3/8]$	non utilisé
$\lceil -\log(3/8-1/4)\rceil = 3$ Code : 011 ; prédécesseur : 010 1/4+1/8 et 1/4 appartiennent à $I_c(4)$			

Codage

Soit donc $x^n = x_1, \ldots, x_n$ une séquence d'éléments de E à coder et $\widehat{\theta}_k$ (4.9) l'estimateur au sens du maximum de vraisemblance des paramètres d'une CMM(k) dont elle serait une réalisation.

Pour $1 \leq t \leq n$ nous noterons $j(t) = x_{t-k}, \ldots, x_{t-1}$ l'état présent de la chaîne et $i(t) = x_t$ le prochain symbole à coder. Si $t \leq k$ alors l'état présent $j(t)$ est mal défini puisqu'il contient un indice négatif ; cependant X_t est distribué selon la loi uniforme sur E et nous considérons que

$$\widehat{\theta}_k(i(t)|j(t)) = \frac{1}{m}.$$

Le codage arithmétique de x^n à l'ordre k se fait de façon itérative sur $t = 0, \ldots, n$, $t = 0$ signifiant que le codage n'a pas débuté, en créant à chaque étape un nouvel intervalle courant $I_c(t) \subset [0,1]$. Le premier intervalle courant est fixé à $I_c(0) = [0,1]$. Supposons que les $t \geq 0$ premières étapes aient été effectuées. L'étape $t+1$ consiste à partitionner l'intervalle courant $I_c(t)$ en $m = |E|$ sous-intervalles de longueurs proportionnelles à $\widehat{\theta}_k(i|j(t+1))$, $i \in E$. De cette façon, nous associons à chaque future valeur possible $i \in E$ un intervalle dont la longueur est proportionnelle à la probabilité avec laquelle on l'attend. Le symbole x_{t+1} est traité en choisissant pour nouvel intervalle courant $I_c(t+1)$ celui qui lui correspond.

Une fois le dernier symbole x_n ainsi traité, nous obtenons un intervalle $I_c(n) =$ [inf,sup]. A l'intérieur de cet intervalle, il existe deux nombres dyadiques dont les parties fractionnaires présentent exactement $\lceil -\log(\text{sup-inf})\rceil$ bits et sont consécutives. Le code arithmétique de x^n est choisi comme étant la partie fractionnaire du plus grand de ces nombres.

Pour illustration, dans la table 4.1, nous prenons $m = 2,\ E = \{a,b\}$ et codons $x^4 = abaa$ à l'ordre $k = 1$. Pour cette chaîne, on a

$$\begin{array}{ll} \widehat{\theta}_1(a|a) = \frac{1}{2}, & \widehat{\theta}_1(b|a) = \frac{1}{2}, \\ \widehat{\theta}_1(a|b) = 1, & \widehat{\theta}_1(b|b) = 0. \end{array}$$

Lors des partitionnements, nous allouons l'intervalle de gauche au symbole a.

Longueur du code

Par construction, la longueur du dernier intervalle courant $I_c(n)$ vérifie

$$\lceil -\log(\text{sup -inf}) \rceil = \left\lceil -\log\left(\frac{1}{m^k}\prod_{t=k+1}^{n}\widehat{\theta}_k(i(t)|j(t))\right)\right\rceil = \left\lceil l(\widehat{\theta}_k)\right\rceil$$

comme en (4.10). Cette quantité égale aussi la longueur du codage arithmétique de x^n qui atteint donc l'objectif fixé en (4.14).

L'inégalité entropique (4.13) peut ici se réécrire

$$H_n(\theta) \leq H_n(\theta, \widehat{\theta}_k). \tag{4.15}$$

En terme de codage, cette inégalité s'interprète comme précédemment : le membre de droite correspond à la longueur moyenne du codage arithmétique à l'ordre k d'un message x^n régi par une CMM décrite par θ. Cette longueur moyenne dépasse nécessairement l'entropie de la CMM.

Décodage

L'opération inverse du codage arithmétique s'effectue facilement à condition toutefois que codeurs et décodeurs se soient accordés au préalable sur l'ordre k auquel le codage est effectué ainsi que sur le paramètre $\widehat{\theta}_k$ utilisé. Nous ne décrivons pas la procédure de décodage qui est relativement directe mais renvoyons une nouvelle fois à [WNC87, MNW98] pour des détails plus techniques.

L'accord préalable sur $\widehat{\theta}_k$ pose problème. En effet, pour le codage arithmétique tel que nous venons de le présenter, l'encodeur détermine le paramètre $\widehat{\theta}_k$ à partir des données. Ce paramètre n'est donc pas accessible au décodeur ; de plus il est renouvelé avec chaque nouveau message à transmettre. Inclure ce paramètre en préambule du codage semble être une solution peu efficace : il peut avoir une grande taille et est constitué de nombres réels qui doivent être connus avec une grande précision pour éviter les erreurs de décodage. Le paramètre d'ordre k, simple entier quant à lui, peut raisonnablement être inclu en préambule.

Pour pallier le problème restant sur $\widehat{\theta}_k$, nous présentons maintenant le codage arithmétique *adaptatif* qui nous permettra de faire le lien avec la notion théorique de complexité stochastique.

4.4 Codage arithmétique adaptatif

Ce codage est conçu pour éviter le problème soulevé en fin de partie 4.3.2. L'algorithme ressemble à celui du codage arithmétique classique (non-adaptatif) présenté dans cette même partie, sauf que l'on ne calcule pas le paramètre $\widehat{\theta}$ avant le codage mais au cours du codage.

4.4.1 Algorithme

Le seul paramètre que nous ayions besoin de fixer au préalable est un ordre de codage k. Soit à nouveau $x^n = x_1, \ldots, x_n$ un message à coder et $I_c(0) = [0, 1]$.

TAB. 4.2 – Codage arithmétique *adaptatif* à l'ordre 1 de *abaa*.

t	x^t	$I_c(t)$	$\widehat{\theta}_1^{(t)}(.\|.)$	Partitionnement
0	$\emptyset$	$[0,1]$	$(a\|a)=1/2$ $(a\|b)=1/2$	$[0,\frac{1}{2},1]$
1	a	$[0,\frac{1}{2}]$	$(a\|a)=1/2$ $(a\|b)=1/2$	$[0,\frac{1}{4},\frac{1}{2}]$
2	ab	$[\frac{1}{4},\frac{1}{2}]$	$(a\|a)=1/3$ $(a\|b)=1/2$	$[\frac{1}{4},\frac{3}{8},\frac{1}{2}]$
3	aba	$[\frac{1}{4},\frac{3}{8}]$	$(a\|a)=1/3$ $(a\|b)=2/3$	$[\frac{1}{4},\frac{7}{24},\frac{3}{8}]$
4	$abaa$	$[\frac{1}{4},\frac{7}{24}]$	$(a\|a)=$ non utilisé $(a\|b)=$ non utilisé	non utilisé
$\lceil -\log(1/4-7/24)\rceil = 5$ Code : 01001 ; prédécesseur : 01000 1/4+1/32 et 1/4 appartiennent à $I_c(4)$				

Codage

Supposons que les $t \geq 0$ symboles aient été traités ; $t = 0$ signifiant que le codage n'a pas commencé. Pour traiter x_{t+1}, nous commençons par actualiser les probabilités de transition comme suit :

$$\widehat{\theta}_k^{(t)}(i|j) = \frac{n^{(t)}(i|j)+1}{n^{(t)}(j)+m} \tag{4.16}$$

où $i \in E$, $j \in E^k$, $n^{(t)}(i|j)$ et $n^{(t)}(j)$ désignent respectivement le nombre d'occurrences de i après j et de j dans la chaîne $x^t = x_1, \ldots, x_t$; rappelons que $n^{(t)}(j)$ ne doit pas comptabiliser une éventuelle occurrence de j à la fin de cette chaîne. Ces probabilités reflètent ce que nous savons de la chaîne à l'instant t du codage ; elles sont l'aspect *adaptatif* de l'algorithme. Notons que pour $0 \leq t \leq k$ on a $\widehat{\theta}_k^{(t)}(i|j) = 1/m$ pour tout i, j conformément à la distribution uniforme des k premières variables d'une CMM(k). Considérant $j(t+1) = x_{t-k+1}, \ldots, x_t$, on découpe comme précédemment l'intervalle courant $I_c(t)$ en m intervalles de longueurs proportionnelles à $\widehat{\theta}_k^{(t)}(i|j(t+1))$, $i \in E$ et on traite x_{t+1} en choisissant pour nouvel intervalle courant celui qui lui correspond. Le code de x^n est construit à partir de $I_c(n)$ comme en partie 4.3.2.

Pour illustration, dans la table 4.2, nous considérons à nouveau $m = 2$, $E = \{a, b\}$ et codons $x^4 = abaa$ de manière adaptative à l'ordre $k = 1$.

Commentaires

Remarquons sur cet exemple que la longueur du code adaptatif, 5 bits, est plus importante que celle du code non-adaptatif (table 4.1), 3 bits. Cela provient du fait que, dans le cas adaptatif, le paramètre utilisé pour le codage évolue à chaque itération. Il n'est donc pas optimal dès le début du codage. Nous reviendrons plus en détail sur ce fait lorsque nous présenterons l'inégalité de Rissanen en partie 4.5.

Cet exemple montre aussi le fait général suivant concernant le codage arithmétique : plus la chaîne comporte de transitions d'ordre k inattendues, plus la longueur du codage augmente. Par exemple, entre les lignes $t = 3$ et $t = 4$ du tableau précédent, l'état courant est a. On attend donc le symbole b avec probabilité $2/3$ et le symbole a avec probabilité $1/3$. Le symbole a, qualifié d'inattendu, se présente. Par conséquent, il nous faut choisir à l'étape $t = 4$ le plus petit des deux intervalles $[1/4, 7/24]$ et $[7/24, 3/8]$ du partitionnement. L'intervalle final $I_c(n)$ a donc une longueur plus faible que si b s'était présenté et la longueur du code s'en trouve augmentée à 5 bits. Pour comparaison, le message *abab* se code quant à lui en 4 bits 0110.

Notons enfin que les termes 1 et m au numérateur et dénominateur de (4.16) assurent que ces probabilités de transition ne s'annulent jamais. Ainsi, lors du codage arithmétique adaptatif, les partitions ne contiennent pas d'intervalles réduits à un point ; cela assure que $I_c(n)$ n'est pas d'intérieur vide et que le codage peut bien se terminer. Dans le cas non-adaptatif (par exemple dans la table 4.1), on peut trouver des intervalles réduits à un point, mais alors c'est nécessairement un autre intervalle qui sera choisi puisque le paramètre $\widehat{\theta}$ est déterminé à partir des données.

Décodage

Seul l'entier k est nécessaire au décodeur pour retrouver le message x^n. Cet entier peut aisément être donné en préambule.

Longueur du code

La longueur du code arithmétique adaptatif à l'ordre k de x^n, notée $L_{a,k}(x^n)$ vérifie

$$L_{a,k}(x^n) = \left\lceil -\log\left(\prod_{t=1}^{n} \widehat{\theta}_k^{(t-1)}(i(t)|j(t))\right)\right\rceil = \left\lceil -\sum_{t=1}^{n} \log \widehat{\theta}_k^{(t-1)}(x_t|x_{t-1},\ldots,x_{t-k})\right\rceil. \tag{4.17}$$

Cette formule reflète bien l'idée de l'adaptativité du codage : on code chaque élément x_t avec une probabilité $\widehat{\theta}_k^{(t-1)}$ estimée à partir des observations passées du processus $x_{t-1},\ldots,x_1$. Cette longueur nous amène à la notion de complexité stochastique.

4.4.2 Complexité stochastique

Rissanen introduit cette notion par exemple dans [Ris86b, Ris86a]. Notons que ces travaux portent sur un cadre plus général que celui des chaînes de Markov multiples. Ces dernières restent cependant, à notre sens, l'exemple le plus parlant où cette notion s'applique en terme de codage comme nous venons de l'expliquer.

Définition 4.4.1 *Soit $x^n \in E^n$ un message constitué de n symboles. La complexité stochastique de x^n relativement au modèle des chaînes de Markov multiples d'ordre $k \in \mathbb{N}$ est donnée par*

$$C_k(x^n) = -\sum_{t=1}^{n} \log \widehat{\theta}_k^{(t-1)}(x_t|x_{t-1},\ldots,x_{t-k}). \tag{4.18}$$

Rappelons que, pour $0 \leq t \leq n-1$, *on a*

$$\widehat{\theta}_k^{(t-1)}(i|j) = \frac{n^{(t-1)}(i|j)+1}{n^{(t-1)}(j)+m}$$

avec $i \in E$, $j \in E^k$, $n^{(t-1)}(i|j)$ *et* $n^{(t-1)}(j)$ *le nombre d'occurrences respectives de* i *après* j *et de* j *dans la chaîne* $x^{t-1} = x_1, \ldots, x_{t-1}$; $n^{(t-1)}(j)$ *ne devant pas comptabiliser une occurrence de* j *à la fin de cette chaîne. Dans le cas* $k = 0$, *les états multiples* j *disparaissent et on remplace* $n^t(j)$ *par* t.

D'après (4.17), la complexité stochastique $C_k(x^n)$ coïncide avec la longueur du codage arithmétique adaptatif de x^n à l'ordre k. La prochaine section est consacrée à l'interprétation en terme de codage de l'inégalité de Rissanen.

4.5 L'inégalité entropique de Rissanen

4.5.1 Théorème

Pour $k \in \mathbb{N}$, notons $\Theta_k \subset [0,1]^{(m-1)m^k}$ l'ensemble des paramètres θ décrivant une CMM(k) à valeurs dans E, ensemble à m éléments. C'est un sous-ensemble compact d'intérieur non-vide de l'espace euclidien de dimension $(m-1)m^k$. On munit ce dernier de la mesure de Lebesgue par rapport à laquelle l'expression "pour presque tout $\theta \in \Theta_k$" se référera. Pour $\theta \in \Theta_k$ et n donnés, on note $\mathbb{P}_\theta$ la probabilité sur E^n associée à θ comme en (4.8) et $\mathbb{E}_\theta$ l'espérance en découlant. Dans [Ris86b], Rissanen montre l'inégalité suivante :

Théorème 4.5.1 *Soit* $k^\star, k \in \mathbb{N}$ *et* $\varepsilon > 0$. *Pour presque tout* $\theta \in \Theta_{k^\star}$ *et pour* n *assez grand, on a*

$$H_n(\theta) + (1-\varepsilon)\left((m-1)m^{k^\star}\right)\frac{\log n}{2} \leq \mathbb{E}_\theta\left[C_k(.)\right] \tag{4.19}$$

où $H_n(\theta)$ *est l'entropie associée à la distribution* P_θ *(4.11).*

4.5.2 Interprétation

Soit $k^\star \in \mathbb{N}$ et $\theta \in \Theta_{k^\star}$ un paramètre décrivant l'évolution d'une CMM d'ordre $k^\star$. Rappelons que pour tout $x^n \in E^n$, $k \in \mathbb{N}$, la complexité stochastique $C_k(x^n)$ coïncide, à une partie entière supérieure près, à la longueur $L_{a,k}(x^n)$ (4.17) du codage arithmétique adaptatif à l'ordre k du message x^n. L'inégalité de Rissanen (4.19) affirme alors que la longueur moyenne du codage arithmétique adaptatif d'un message régi par la CMM décrite par θ dépasse nécessairement l'entropie de cette dernière, augmentée d'un terme d'obstruction $2^{-1}(m-1)m^{k^\star}\log n$.

Il faut noter que, dans le cas non-adaptatif, on obtient l'inégalité entropique classique (4.15) dans laquelle cette obstruction n'apparaît pas. Elle est donc créée par le caractère *adaptatif* du codage. L'adaptation aux données, exprimée par l'actualisation des probabilités de transition au cours du codage (équation (4.16)), prend un certain temps pendant lequel le codage n'est pas optimal. Cependant, c'est cette actualisation qui résout le problème de décodabilité soulevé en partie 4.3.2.

Ce temps d'adaptation, et donc l'obstruction dans l'inégalité entropique, disparaît dans le cas non-adaptatif où le paramètre optimal $\widehat{\theta}_k$ est calculé *avant* le codage. Toutefois c'est ce même paramètre, qui crée le problème de décodabilité.

En résumé, effectuer le codage de manière *adaptative* assure la décodabilité du message mais a pour contrepartie de créer une obstruction dans l'inégalité entropique $2^{-1}(m-1)m^{k^\star}\log n$.

L'inégalité (4.19) de ce théorème est donc à mettre en parallèle avec l'inégalité entropique (4.7) classique portant seulement sur E : pour toute distribution P et Q sur E on a

$$H(P) \leq H(P,Q).$$

Dans ce dernier cadre, il est également classique que, pour toute distribution P sur E, il existe une distribution Q satisfaisant

$$H(P,Q) \leq H(P) + 1.$$

Cette dernière inégalité affirme que la borne de Shannon est approchable à 1 bit près par une distribution adéquate. Rissanen montre également dans [Ris86b] un parallèle de cette inégalité que nous énonçons maintenant.

4.5.3 La seconde inégalité de Rissanen

Les notations sont identiques à celles du theorème 4.5.1.

Théorème 4.5.2 *Pour $k \in \mathbb{N}$, $\theta \in \Theta_k$ et n assez grand, on a :*

$$\mathbb{E}_\theta\left[C_k(.)\right] \leq H_n(\theta) + \left((m-1)m^k\right)\frac{\log n}{2} + o(\log n). \tag{4.20}$$

L'inégalité énoncée affirme que le codage arithmétique adaptatif à l'ordre k, dont la longueur coïncide avec la complexité stochastique $C_k(x^n)$, approche asymptotiquement et en moyenne la borne "entropie + obstruction" présentée dans l'inégalité entropique (4.19). Notons que le noyau du travail de Rissanen dans [Ris86b] reste l'inégalité entropique (4.19).

4.6 Détermination de l'ordre d'une CMM

Cette section est consacrée à la détermination de l'ordre d'une CMM à partir d'une de ses réalisations. Notre propos est de justifier, en nous appuyant sur les concepts de codage présentés précédemment ainsi que sur le principe du *Minimum Description Length*, l'utilisation du critère d'information BIC pour la résolution de ce problème.

4.6.1 Position du problème

On dispose d'une réalisation $x^n = x_1, \ldots, x_n$ de cette CMM dont les paramètres d'ordre $k^\star$ et de transition $\theta \in \Theta_{k^\star}$ sont inconnus. On se propose d'estimer, à partir des données x^n, l'ordre $k^\star$ de la chaîne. Une fois que cet ordre est estimé, il est aisé d'estimer θ au sens du maximum de vraisemblance par l'expression (4.9). Nous sommes en présence d'un problème de sélection de modèles. Nous justifions maintenant pourquoi l'utilisation du critère d'information BIC (4.22) est adapté à sa résolution.

4.6.2 Le principe du *Minimum Description Length* (MDL)

Nous utilisons ici le principe MDL sous sa forme la plus simple : disposant d'un message x^n à encoder et de plusieurs techniques pour ce faire, il nous faut choisir celle qui donnera le code le plus court. Pour plus de détails concernant ce principe, nous renvoyons aux travaux fondateurs de Rissanen [Ris78] ou plus récemment à Rissanen ou Grunwald et Al. [BRY98, GMP05].

Enoncé de cette manière, le principe MDL s'applique parfaitement à notre problème de sélection de modèles. Nous disposons en effet d'un message x^n que nous pouvons coder à l'aide des différentes techniques de codages arithmétiques présentées dans les parties 4.3 et 4.4. Nous avons vu (partie 4.3.2) que le codage arithmétique non-adaptatif présente le défaut majeur de ne pas être décodable, nous le laisserons donc de côté pour l'instant. Le codage arithmétique adaptatif, quant à lui, nécessite le choix préalable d'un ordre k de codage et produit un code dont la longueur $L_{a,k}(x^n)$ est donnée par l'équation (4.17). Le principe MDL nous pousse donc à choisir comme ordre de codage celui qui réalise

$$\widetilde{k} = \text{Argmin}\left(L_{a,k}(x^n),\ k \in \mathbb{N}\right). \tag{4.21}$$

4.6.3 Lien avec la détermination de l'ordre

Rappelons qu'ici le message est régi par une CMM d'ordre $k^\star$; en d'autres termes la valeur d'un symbole x_t est dictée par son passé $x_{t-1}, \ldots, x_{t-k^\star}$. Par conséquent, pour $k < k^\star$ le message va certainement contenir des transitions d'ordre k inattendues puisqu'on essaye de prédire la valeur d'un symbole x_t en ne regardant pas assez loin dans le passé. De la même façon, pour $k > k^\star$, on prend en compte trop d'informations pour la prédiction du prochain symbole et on observe à nouveau des transitions d'ordre k inattendues.

Pour k fixé, nous avons expliqué précédemment (partie 4.4.1) comment ces transitions d'ordre k inattendues présentes dans le message font augmenter la longueur de son code arithmétique adaptatif à l'ordre k. On s'attend donc à ce que son codage à l'ordre $k^\star$ soit le plus efficace. De cette manière, on établit un lien entre le codage arithmétique adaptatif et le problème de sélection de modèle abordé ici. On préconise donc d'estimer l'ordre $k^\star$ inconnu de la CMM régissant le message par le $\widetilde{k}$ (4.21) minimisant sa longueur de code.

Simulation

Pour illustration, nous choisissons $m = 2$, $k^\star = 5$ ainsi qu'un paramètre $\theta \in \Theta_5$ que nous n'explicitons pas ici et en générons un message x^n de longueur $n = 1000$. Nous effectuons ensuite les codages arithmétiques adaptatifs de ce message aux ordres $k = 0, \ldots, 10$. Les longueurs de codes résultantes sont présentées en fonction de k sur la figure 4.1.

Comme attendu, c'est à l'ordre $k = k^\star$ que le codage est le plus efficace. Cependant, effectuer tous ces codages est coûteux en terme de calculs. Nous voyons maintenant comment estimer cette longueur de codage à partir des données. L'estimation à laquelle nous aboutirons n'est autre que le critère d'information BIC. En ce sens, le codage arithmétique adaptatif est lui aussi un critère d'information.

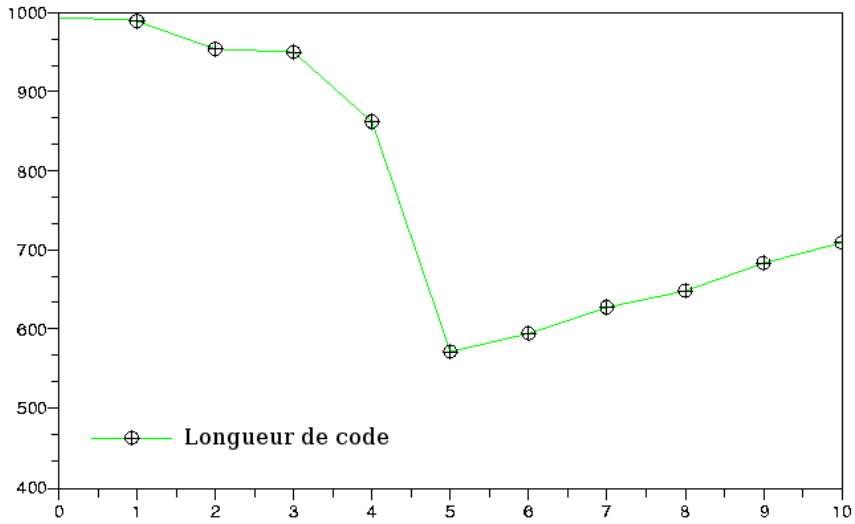


FIG. 4.1 – Longueurs des codes arithmétiques adaptatifs de x^n pour $k = 0, \ldots, 10$.

4.6.4 Le critère BIC

A k fixé les inégalités (4.19) et (4.20) préconisent de prendre pour estimation de la complexité stochastique $C_k(x^n)$ la quantité

$$H_n(\theta) + \big((m-1)m^k\big)\,\frac{\log n}{2}.$$

Le terme d'entropie, inconnu de l'utilisateur, s'estime à l'ordre k à partir des données par $l(\widehat{\theta}_k)$ défini en (4.10). L'estimation que nous faisons de la complexité stochastique de x^n relativement au modèle d'ordre k, et donc de la longueur du codage arithmétique adaptatif à l'ordre k de x^n, est :

$$\begin{aligned} \mathrm{BIC}(k) &= l(\widehat{\theta}_k) + \big((m-1)m^k\big)\,\frac{\log n}{2} \\ &= -\sum_{j\in E^k}\sum_{i\in E} n(i|j)\log\frac{n(i|j)}{n(j)} + k\log m \\ &\quad + \big((m-1)m^k\big)\,\frac{\log n}{2}. \end{aligned} \tag{4.22}$$

Par conséquent, le principe MDL utilisé précédemment nous pousse à estimer l'ordre $k^\star$ de la CMM inconnue par

$$\widehat{k} = \mathrm{Argmin}(\mathrm{BIC}(k),\ k \in \mathbb{N}). \tag{4.23}$$

La quantité $\mathrm{BIC}(k)$ est un critère d'information au sens où nous l'entendons dans l'équation (2.1). Le terme de vraisemblance maximale relativement au modèle d'ordre k est $l(\widehat{\theta}_k)$ tandis que la pénalité est $\big((m-1)m^k\big)\,\frac{\log n}{2}$. Le coefficient $(m-1)m^k$ de cette pénalité correspond au nombre de paramètres libres d'une CMM d'ordre k pouvant prendre m valeurs. Notons que nous n'avons pas gardé dans l'écriture (4.22) du critère le coefficient 2 adopté dans (2.1) ; ce coefficient ne ferait

plus correspondre le critère à une estimation de la longueur du code de x^n à l'ordre k. La pénalité de BIC correspond donc à $\alpha(n) = \log n$.

Dans notre cas, il faudrait plutôt appeler ce critère RIC, pour *Rissanen Information Criterion*. Il est similaire dans sa forme au critère BIC (*Bayesian Information Criterion*) donné par Schwarz [Sch78]. La différence se situe au niveau de la méthode de justification d'utilisation de ce critère. Nous venons d'établir une justification basée sur une interprétation en terme de codage tandis que Schwarz se base sur une méthode bayésienne d'estimation que nous n'abordons pas ici.

4.6.5 Retour au codage arithmétique non-adaptatif

Dans la partie 4.3, nous avons montré que la longueur $L_{na,k}(x^n)$ de code résultant du codage arithmétique non-adaptatif à l'ordre k d'un message x^n satisfait par construction

$$L_{na,k}(x^n) = \left\lceil l(\widehat{\theta}_k) \right\rceil \tag{4.24}$$

où $\widehat{\theta}_k$ est l'estimateur au sens du maximum de vraisemblance des paramètres de la CMM(k) régissant x^n comme en (4.9).

De la même manière que le critère BIC(k) (4.22) s'interprète comme la longueur du code arithmétique adaptatif à l'ordre k de x^n, la quantité $\mathrm{MV}(k) := l(\widehat{\theta}_k)$ correspond donc à la longueur du code arithmétique *non*-adaptatif à l'ordre k. Cette quantité, en tant que maximum de vraisemblance, est inadaptée au problème de sélection de modèle abordé ici : une méthode de sélection du type $\widehat{k} = \mathrm{Argmin}(\mathrm{MV}(k))$ amènerait à surestimer l'ordre $k^\star$ de la CMM régissant le message. La pénalisation des critères d'information tels qu'exprimés en (2.1) permet de pallier ce problème de sur-paramétrisation. En transposant ces considérations en termes de codages, il vient que c'est la manière *adaptative* d'effectuer le codage arithmétique qui crée la pénalité et permet donc d'éviter la sur-paramétrisation.

4.6.6 Simulations

Nous choisissons à nouveau $m = 2$, $k^\star = 5$ ainsi qu'un paramètre $\theta \in \Theta_5$ qu'il est inutile d'expliciter ici. Nous générons un message x^n de longueur $n = 2500$. Nous effectuons ensuite les codages arithmétiques adaptatifs et non-adaptatifs de ce message aux ordres $k = 0, \ldots, 10$, ainsi que le calcul du critère BIC(k) (4.22) pour les mêmes ordres. Les résultats sont présentés, divisés par n, en fonction de k sur la figure 4.2. Notons que, pour cette simulation, les codages arithmétiques ont nécessité 6,25 secondes de calcul tandis que le critère BIC en nécessite seulement 1,85 sur la même machine.

Nous retrouvons ici l'allure classique des critères d'information. Le critère du maximum de vraisemblance (MV), correspondant à la longueur du codage non-adaptatif par (4.24), présente le phénomène habituel de sur-paramétrisation tandis que le critère BIC atteint son minimum en $\widehat{k} = k^\star = 5$. Il sélectionne donc le bon ordre, de la même manière que le fait la longueur du code adaptatif dont il est une estimation.

On observe que le critère BIC a tendance, pour les ordres grands, à croître rapidement. Cela est dû au nombre de paramètres libres $(m-1)m^k$ présent dans sa pénalité qui explose quand l'ordre k grandit. Pour notre exemple, avec $m = 2$ et

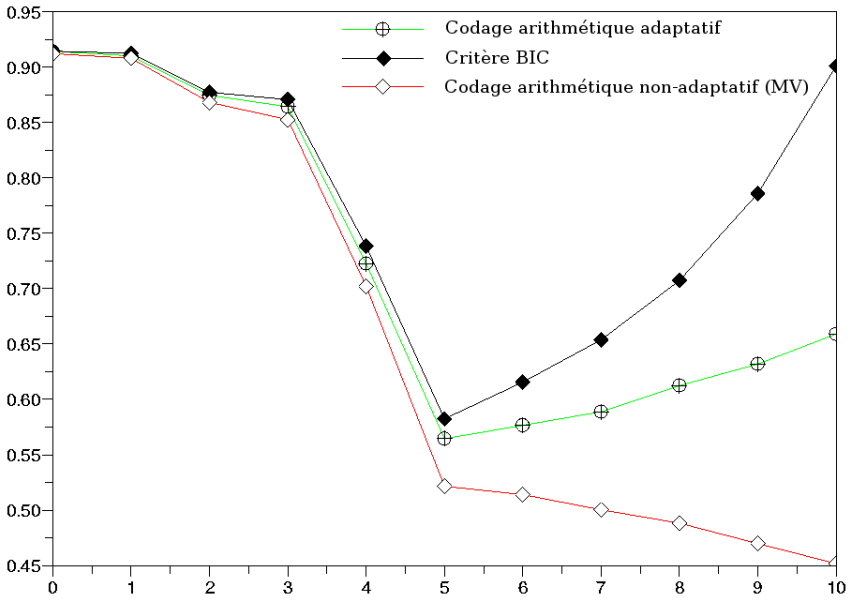


FIG. 4.2 – Superposition des longueurs de codage arithmétique adaptatif, non-adaptatif et du critère BIC en fonction de l'ordre k.

k	0	1	2	3	4	5	6	7	8	9	10
$(m-1)m^k$	1	2	4	8	16	32	64	128	256	512	1024
Estimation	1	2	4	8	16	28	46	78	139	245	467

TAB. 4.3 – Le nombre réel de paramètres libres et le nombre de transitions observées dans le message.

pour le dernier ordre testé $k = 10$, on obtient 1024 paramètres libres alors que la longueur de notre message est seulement de 2500 symboles. Nous avons donc modifié le critère BIC en remplaçant, pour chaque k, le terme $(m-1)m^k$ de sa pénalité par une estimation du nombre de paramètres effectivement libres, cette estimation étant égale aux nombre de transitions observées au moins $m = 2$ fois dans le message. Le tableau 4.3 présente cette estimation en fonction de k pour notre message. Il faudrait choisir n plus grand de manière à observer toutes les transitions dans le message pour éviter d'avoir à effectuer cette estimation et garder un critère BIC raisonnable pour les grands ordres.

Résumé

Dans ce chapitre, nous présentons les notions classiques de codages et le lien qu'elles entretiennent avec la théorie des probabilités. Dans ce cadre, l'inégalité de convexité de Jensen (4.7), appelée inégalité d'information de Shannon en théorie de l'information, exprime l'existence d'une borne inférieure pour la longueur de codage de symboles issus d'une source. La technique de codage de Huffman est vue ici comme un moyen pratique d'approcher cette borne.

Nous rappelons ensuite comment Rissanen développe, entre autre dans [Ris86b], la notion de complexité stochastique et démontre l'inégalité (4.19). Cette dernière est vue ici comme une extension de l'inégalité d'information dans le cas où les données sont codées globalement et de manière adaptative. Elle présente à nouveau l'existence d'une borne inférieure pour la longueur d'un tel codage, cette borne étant celle de Shannon augmentée d'une obstruction. L'outil pratique permettant d'approcher la borne est, cette fois, le codage arithmétique adaptatif que nous décrivons en détail.

Une application basique du principe du *Minimum Description Length* nous amène enfin à la résolution du problème étudié dans ce chapitre : la détermination de l'ordre d'une chaîne de Markov multiple à partir de l'une de ses réalisations. Le MDL nous pousse en effet à minimiser la quantité BIC(k) (4.22) entrant dans la classe des critères d'information (2.1) définie au début de ce mémoire. Nous avons donc justifié, dans notre cadre et en nous basant sur des considérations de type longueur de codage, l'utilisation d'un critère d'information pour le problème de sélection de modèle posé.

En perspective, notons qu'il est moins rapide d'effectuer le codage arithmétique adaptatif que de calculer le critère BIC qui estime sa longueur. Un utilisateur désirant choisir un ordre de codage pour son message pourra donc calculer les critères BIC(k) dans un premier temps, pour ensuite coder à l'ordre minimisant ce critère.

Chapitre 5

La sélection d'histogramme

Ce chapitre a pour but de justifier, en terme de longueur de codage, l'utilisation du critère (5.6) pour le problème de la sélection d'un histogramme estimant une densité inconnue. Pour cela, nous nous appuyons sur les résultats du chapitre précédent concernant le critère BIC (4.22) et ses liens avec le codage arithmétique adaptatif.

Le logarithme est également exprimé en base 2 au cours de ce chapitre.

5.1 Position du problème

Soit f une densité de probabilité inconnue définie sur un intervalle I de $\mathbb{R}$ à estimer à partir d'un de ses échantillons $x^n = x_1, \ldots, x_n$. Pour $P = (I_j)_{j=1,\ldots,m}$ une partition de I à $m \in \mathbb{N}^\star$ intervalles il est aisé d'estimer f à partir de x^n par son estimateur au sens du maximum de vraisemblance au sein de la classe des densités constantes par morceaux sur P. L'estimateur obtenu est

$$\widehat{f}_P = \sum_{j=1}^{m} \frac{n_j}{n|I_j|} \mathbb{1}_{I_j} \tag{5.1}$$

où n_j est le nombre de données tombant dans l'intervalle I_j et $|I_j|$ sa longueur.

Le problème est maintenant de décider, à partir des données uniquement, quelle partition P est la mieux adaptée à une telle estimation. Pour cela, nous élaborons, à partir d'une partition, une technique de codage sans perte des données x^n. Une application du principe MDL, tel qu'exprimé en section 4.6.2, conduira au choix d'une partition.

5.2 Le codage sans perte des données

Choisissons un réel $r > 0$ et discrétisons l'intervalle I en $R = |I|/r$ intervalles. On peut imaginer que le réel r représente la précision en dessous de laquelle la machine avec laquelle nous travaillons ne peut plus distinguer deux réels, l'intervalle I contenant en fait R réels distinguables. Cette approche permet également de traiter le cas où les données sont des réalisations d'une variable discrète. Appelons $P_{\max}$ la partition de I à R intervalles, tous de longueur r.

Nous expliquons maintenant comment, à partir d'une sous-partition P de $P_{\max}$ quelconque, un utilisateur peut effectuer un codage sans perte de ses données x^n. Ce codage se fait en deux étapes.

5.2.1 Première étape : le codage arithmétique

La sous-partition $P = (I_j)_{j=1,\dots,m}$ de $P_{\max}$ présente en premier lieu l'avantage de réduire les données à une nouvelle chaîne y^n à valeurs dans $[\![1, m]\!]$ comme suit :

$$\forall i \in [\![1, n]\!], \quad y_i = \sum_{j=1}^{m} j \mathbb{1}_{I_j}(x_i) \tag{5.2}$$

Clairement y_i est le numéro de l'intervalle de P dans lequel x_i tombe. Cette définition montre également que les y_i sont des observations indépendantes de la variable $Y = \sum j \mathbb{1}_{I_j}(X)$ où X a pour densité f.

Effectuons le codage arithmétique adaptatif de cette chaîne y^n tel qu'expliqué en section 4.4. Vu l'indépendance des y_i, le codage à l'ordre 0 sera moins coûteux en terme de bits qu'un codage à un ordre supérieur ; la longueur obtenue par un tel codage est désormais notée $L(y^n|P)$.

En première approche, puisque m est la taille de l'alphabet avec lequel le message y^n est écrit, la longueur $L(y^n|P)$ augmente avec m. Il se peut cependant qu'un découpage judicieux de l'intervalle crée de la régularité, ou plutôt une chute d'entropie, dans la chaîne y^n faisant ainsi diminuer cette longueur.

5.2.2 Deuxième étape : le codage de précision

Une fois la chaîne y^n encodée, nous avons un aperçu global des données x^n : chaque x_i est dans l'intervalle I_{y_i} de P. Puisque nous souhaitons coder les données sans perte, il nous faut décrire plus précisément où se situe x_i. Chaque intervalle I_j présente exactement $|I_j|/r$ réels distinguables. Pour préciser la valeur de x_i parmi ces réels distinguables, nous effectuons un codage à longueur fixe sur l'intervalle d'entiers $[\![1, |I_j|/r]\!]$. La longueur idéalisée de ce codage est $\log(|I_j|/r)$. Ainsi le nombre de bits nécessaires à la précision de tous les x_i est :

$$L(x^n|y^n) = \sum_{j=1}^{m} n_j \log(|I_j|/r) = \sum_{j=1}^{k} n_j \log |I_j| - n \log r. \tag{5.3}$$

Globalement, cette longueur a tendance à diminuer lorsque m augmente puisque les intervalles de P deviennent plus petits et la précision des x_i d'autant plus aisée.

5.2.3 Application du principe MDL

Commençons par noter que, pour que le codage que nous venons de présenter soit entièrement décodable, il faudrait également encoder la partition P à partir de laquelle il est effectué. Un tel encodage de partition est par exemple traité dans [RSY92]. En pratique, la longueur du code nécessaire à la partition est négligeable devant la longueur du code des données elles-mêmes, c'est pour cette raison que nous la négligeons.

Au final, la longueur $L(x^n|P)$ du code sans perte des données x^n à partir de la sous-partition P de $P_{\max}$ est donnée par

$$L(x^n|P) = L(y^n|P) + L(x^n|y^n). \tag{5.4}$$

Nous disposons ici d'autant de techniques de codages de nos données que $P_{\max}$ présente de sous-partitions. Le principe MDL tel qu'exprimé dans la section 4.6.2 nous pousse donc à choisir comme partition celle qui donnera le codage le plus efficace. Concernant la minimisation de $L(x^n|P)$, nous avons vu qu'à mesure que le nombre d'intervalles m de P augmente, le premier terme de (5.4) augmente également tandis que le second diminue. C'est ce comportement inverse de $L(x^n|y^n)$ par rapport à $L(y^n|P)$ qui peut permettre de décider si la partition P est acceptable.

La quantité $L(x^n|P)$ présentée ici n'est pas un critère d'information au sens où nous les entendons dans ce mémoire. Une estimation des longueurs de codage entrant en jeu dans (5.4) grâce aux notions développées dans le chapitre précédent va maintenant nous permettre de retrouver un tel critère.

5.3 Estimation de la longueur de codage

Le terme $L(y^n|P)$ présent dans (5.4) est la longueur du codage arithmétique adaptatif à l'ordre 0 de la chaîne y^n. Par les considérations du chapitre précédent portant sur ce codage et ses liens avec le critère BIC (4.22) (section 4.6.4), ce terme s'estime par

$$L(y^n|P) \approx \mathrm{BIC}(0) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n} + (m-1)\frac{\log n}{2}. \tag{5.5}$$

Cette estimation, couplée à (5.3), nous permet d'estimer la longueur (5.4) du codage sans perte des données par

$$\begin{aligned} \mathrm{CRIT}(P) &= \mathrm{BIC}(0) + L(x^n|y^n) \\ &= -\sum_{j=1}^{m} n_j \log \frac{n_j}{n|I_j|} + (m-1)\frac{\log n}{2} - n\log r. \end{aligned} \tag{5.6}$$

Via le principe MDL, c'est donc la partition qui réalise

$$\widehat{P} = \mathrm{Argmin}(\mathrm{CRIT}(P),\ P \text{ sous-partition de } P_{\max}) \tag{5.7}$$

qu'il nous faut choisir pour l'estimation de f par $\widehat{f}_{\widehat{P}}$ comme en (5.1).

Modulo le terme $-n\log r$, la quantité $\mathrm{CRIT}(P)$ est un critère d'information au sens de (2.1). En effet, son premier terme est l'opposé du logarithme de la vraisemblance maximale des données relativement à la classe des histogrammes construits sur P :

$$-\log\left(\prod_{i=1}^{n} \widehat{f}_P(x_i)\right) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n|I_j|}.$$

Le second terme est, quant à lui, la pénalisation correspondante à l'estimation par un histogramme à m intervalles. La fonction de pénalité $\alpha(m)$ de (2.1) vaut ici $\log n$.

En ce sens, le critère CRIT peut être vu comme le critère BIC (tableau 2.1) appliqué au problème non-paramétrique de la sélection d'histogramme.

Nous choisissons de conserver le terme $-n \log r$ dans CRIT (5.6) puisqu'il permet de conserver l'aspect longueur du codage du critère. Sa présence n'influe pas sur le choix de $\widehat{P}$ par (5.7).

Plus spécifiquement, la précision r est isolée dans le CRIT. Par conséquent, le choix de $\widehat{P}$ par (5.7) est donc indépendant de cette précision. Cette remarque nous permet d'étendre la procédure de sélection (5.7) à d'autres classes de partitions sans perdre l'interprétation de CRIT en terme de longueur de codage. Considérons en effet une classe $\mathcal{P}$ de partitions qui soient toutes des sous-partitions d'une certaine Q, partition régulière de I. Choisir r comme la longueur des intervalles de Q donne le sens voulu à la procédure

$$\widehat{P} = \text{Argmin}(\text{CRIT}(P),\ P \in \mathcal{P}).$$

En particulier, la classe des partitions de I à m intervalles tous de longueurs $|I|/m$ pour $m \in [\![1, M]\!]$ avec $M \in \mathbb{N}^\star$ fixé peut être utilisée de la sorte. Dans la suite, un histogramme choisi parmi une telle classe de partitions régulières sera dit *régulier*.

Par opposition, lorsque nous travaillerons avec une précision r fixée, une partition $P_{\max}$ associée et la méthode (5.7), nous parlerons d'histogrammes *dynamiques*. Cette terminologie est emprunté à l'algorithme de programmation *dynamique* présenté par Rissanen et Al. dans [RSY92]. Cet algorithme permet de déterminer la partition $\widehat{P}$ vérifiant (5.7) en un nombre d'opérations de l'ordre de R^2 alors que, *a priori*, il est nécessaire de calculer 2^{R-1} critères, un pour chaque sous-partition de $P_{\max}$. Nous détaillons cet algorithme en annexe A.

5.4 Simulations

Dans cette section nous choisissons pour densité inconnue celle d'une loi β de paramètres $(1/2, 1/2)$,

$$\frac{1}{\Gamma(1/2)^2} x^{-1/2}(1-x)^{-1/2},$$

et en générons un échantillon de taille $n = 1000$.

5.4.1 Histogramme régulier

Nous travaillons dans un premier temps avec des partitions régulières P_m de $]0, 1[$ à m intervalles pour $m = 1, \ldots, 50$. La figure 5.1(a) présente le critère CRIT calculé sur l'échantillon en fonction de m. On trouve également en figure 5.1(b) la longueur du codage de précision $L(x^n|y^n)$ (5.3) ainsi que l'estimation BIC(0) (5.5) de la longueur du codage arithmétique $L(y^n|P)$; la somme de ces deux quantités étant CRIT (5.6). La partition réalisant le minimum de CRIT est obtenue pour $\widehat{m} = 17$ et l'histogramme associé est représenté, superposé à la loi $\beta(1/2, 1/2)$, en figure 5.2.

La divergence de Kullback entre deux densités de probabilités f et g est définie comme la demi-somme des informations de Kullback (1.3) $K(f, g)$ et $K(g, f)$. Elle vaut donc :

$$D(f, g) = \frac{1}{2} \int (g - f) \log \frac{g}{f}\, d\lambda. \tag{5.8}$$

Nous avons calculé cette divergence entre la loi $\beta(1/2, 1/2)$ et l'histogramme régulier obtenu désigné par h_r :

$$D(\beta, h_r) \simeq 1,20. \tag{5.9}$$

Ce résultat est à comparer avec celui (5.10) de la partie suivante.

5.4.2 Histogramme dynamique

Cette fois nous fixons pour précision $r = 2.10^{-2}$ et déterminons l'histogramme dynamique réalisant (5.7) grâce à l'algorithme dynamique présenté en annexe A. L'algorithme dynamique utilisé permet difficilement de garder une trace des estimations des longueurs de codages partielles $L(x^n|y^n)$ (5.3) et $L(y^n|P)$ (5.5) pour P une sous-partition de $P_{\max}$. Par contre il permet, pour chaque $m = 1, \ldots, R$, d'obtenir la valeur $\mathrm{CRIT}(\widehat{P}_m)$ où $\widehat{P}_m$ réalise le minimum de CRIT parmi les sous-partitions de $P_{\max}$ à m intervalles. Ce sont ces quantités qui sont représentées sur la figure 5.3, en fonction de $m = 1, \ldots, R$. La partition $\widehat{P}$ satisfaisant (5.7) vérifie aussi

$$\widehat{P} = \mathrm{Argmin}(\mathrm{CRIT}(\widehat{P}_m),\ m = 1, \ldots, R).$$

Dans notre cas ce minimum est obtenu pour $m = 7$ et l'histogramme construit sur $\widehat{P}$ est donné en figure 5.4, superposé à la loi $\beta(1/2, 1/2)$.

Ici, la divergence de Kullback (5.8) entre la loi $\beta(1/2, 1/2)$ et l'histogramme dynamique h_d obtenu vaut

$$D(\beta, h_d) \simeq 0,89. \tag{5.10}$$

Comparant ce résultat avec (5.9), on observe que l'histogramme dynamique, même s'il présente moins d'intervalles que le régulier, permet une meilleure approche de la loi inconnue.

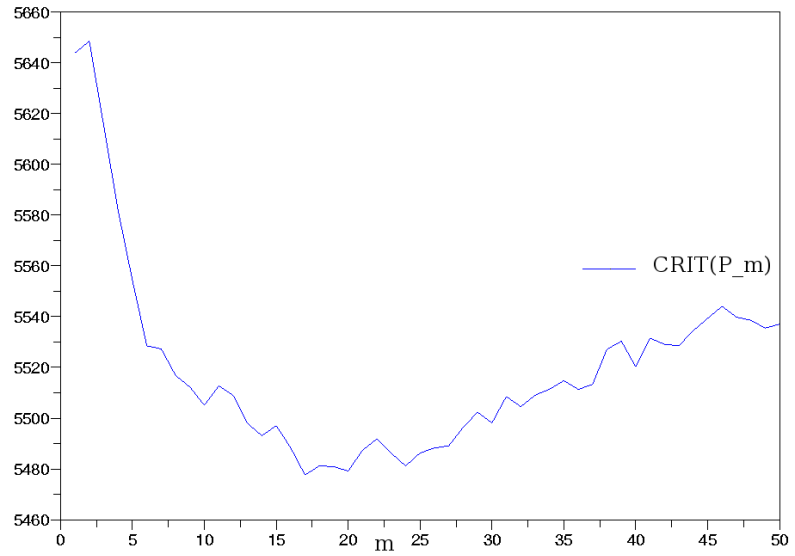
5.5 Application à la reconnaissance de loi

Ce problème de reconnaissance de loi nous a été posé dans le cadre d'une collaboration entre le groupe Systèmes de Communications du laboratoire XLIM-SIC de Poitiers et l'équipe France Télécom de Brest. Les méthodes que nous allons présenter pour le traiter sont aisément transposables à tout autre problème similaire. Nous reprenons ici certains résultats extraits d'un travail soumis à cette date à une conférence internationale. Cette soumission est présentée en annexe D.

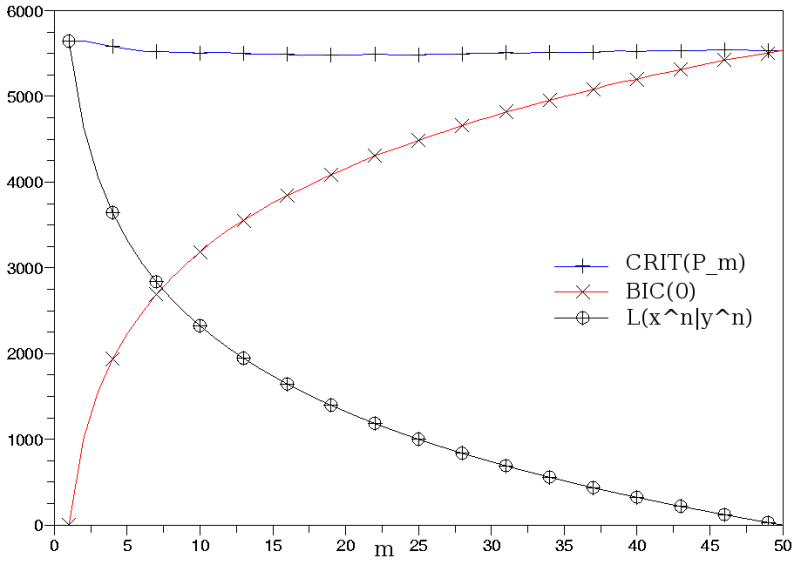
Disposant d'un échantillon de taille n d'une loi de Rayleigh, nous nous intéressons à la reconnaissance de cette loi parmi d'autres lois candidates pour lesquelles nous avons choisi les distributions de Rice et de Weibull. Ces lois sont habituellement utilisées pour la modélisation de canaux de transmission sans-fil à propagation multitrajectorielle. Pour plus de détails concernant ces lois et leurs applications dans l'étude des canaux de propagation, on peut consulter le livre de Patzold [Pat01].

La méthode classique par Kolmogorov-Smirnov

La méthode classique pour la communauté scientifique traitant de transmission numérique consiste à calculer la distance de Kolmogorov-Smirnov (KS) entre la fonction de répartition empirique de l'échantillon et les fonctions de répartitions



(a)



(b)

FIG. 5.1 – Le critère (5.6) d'un échantillon de la loi $\beta(1/2, 1/2)$ sur les partitions régulières P_m (a) avec ses deux composantes BIC(0) et $L(x^n|y^n)$ (b). Le minimum est atteint pour $\widehat{m} = 17$.

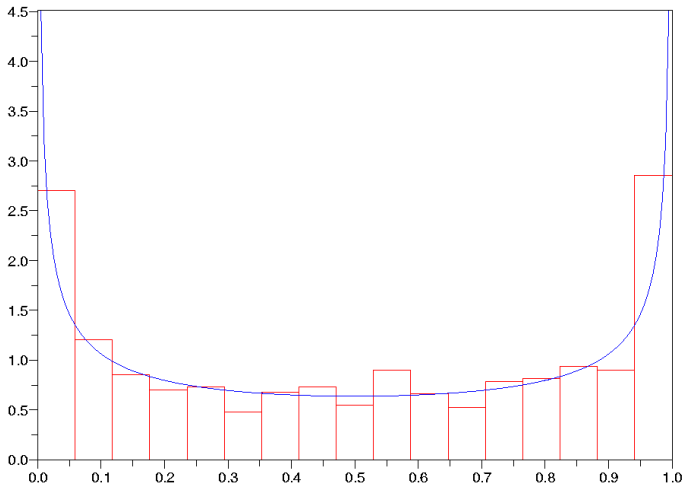


FIG. 5.2 – L'histogramme régulier à 17 classes déterminé par le critère (5.6) superposé à la densité inconnue.

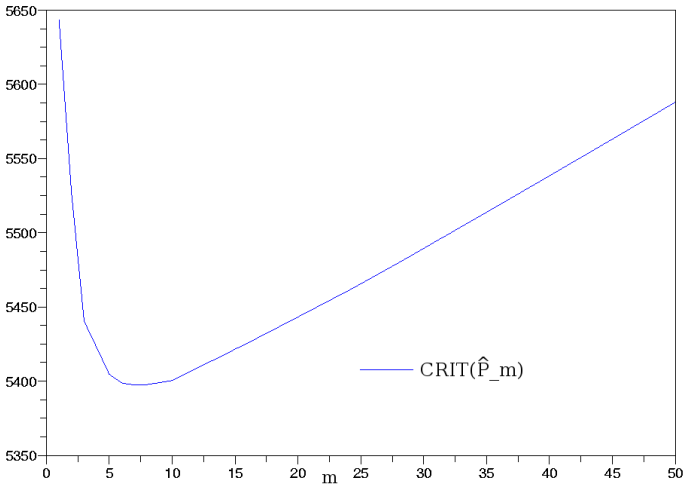


FIG. 5.3 – Les critères minimaux CRIT($\widehat{P}_m$) sur un échantillon d'une loi $\beta(1/2, 1/2)$.

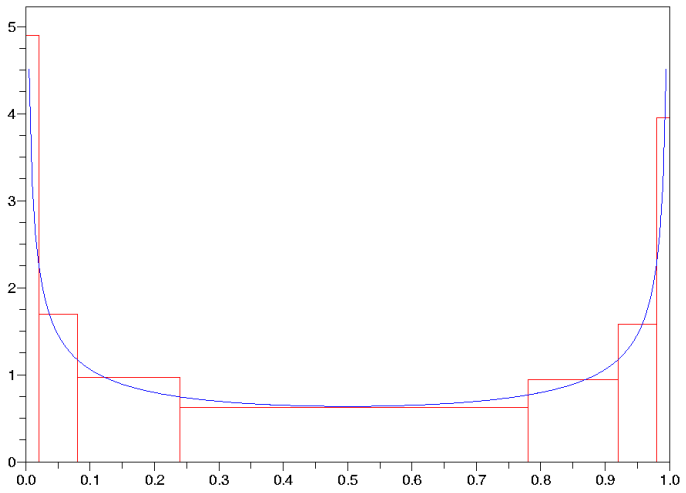


FIG. 5.4 – L'histogramme dynamique à 7 classes déterminé par le critère (5.6) superposé à la densité inconnue.

théoriques des lois en compétition. On choisit alors comme loi celle qui minimise ces distances.

La méthode à histogramme

Nous proposons ici d'utiliser une méthode à histogramme. Pour cela, déterminons l'histogramme (régulier ou dynamique, ce dernier nécessitant le choix préalable d'une précision) associé à l'échantillon par le critère CRIT (5.6) puis calculons la divergence de Kullback-Leibler (5.8) entre cet histogramme et les lois en compétition. Nous choisissons comme loi celle qui minimise ces distances.

Notons que le choix d'une autre Φ-divergence comme celle de Hellinger (cf Basseville [Bas89, Bas96]) conduirait à des résultats similaires.

Résultats

La figure 5.5 présente les taux de reconnaissance de la méthode utilisant la distance KS ainsi que des méthodes à histogrammes réguliers ou dynamiques en fonction de la taille n de l'échantillon. La méthode à histogramme dynamique atteint les 100% de réussite le plus rapidement. La méthode à histogramme régulier est, quant à elle, comparable à la méthode KS.

Notons aussi que les histogrammes, réguliers ou dynamiques, déterminés par le critère CRIT (5.6) sur nos échantillons n'ont pas plus de 50 classes pour $n \leq 10.000$. Il est donc intéressant de remarquer que l'histogramme, résumé de l'information apportée par l'échantillon à une cinquantaine de paramètres au maximum, permet une aussi bonne, voire meilleure, reconnaissance de la loi de base que la fonction de répartition empirique qui conserve, elle, toute l'information de l'échantillon.

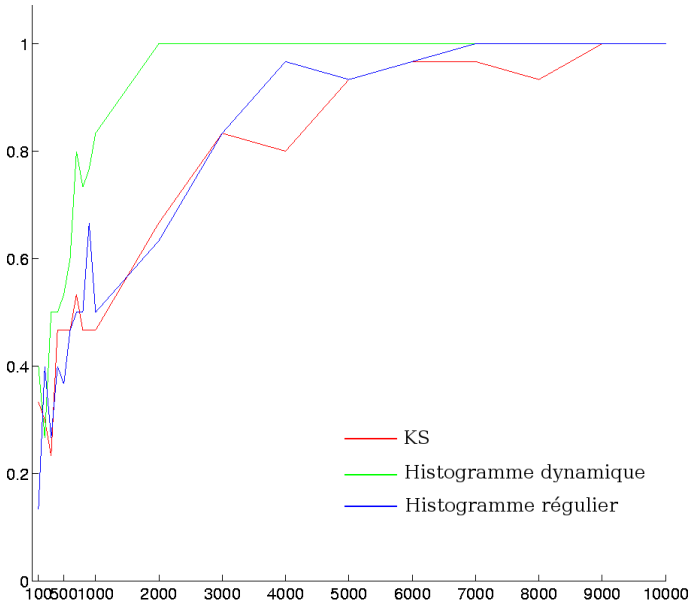


FIG. 5.5 – Taux de reconnaissance de la loi de Rayleigh par la méthode de Kolmogorov-Smirnov (KS) et par les méthodes à histogrammes.

5.6 Applications au traitement de l'image

Nous reprenons ici les résultats principaux d'applications au domaine du traitement de l'image donnés dans [CAOA] et [COAA]. Ces publications sont également présentées dans les annexes B et C de ce mémoire.

5.6.1 Quantification d'une image en niveaux de gris

Une image en niveaux de gris, ensemble de n pixels auxquels sont attribués un niveau entier entre 0 et 255, peut être vue comme un échantillon d'une loi discrète à valeurs dans $[\![0, 255]\!]$. Nous pouvons donc lui appliquer le critère CRIT (5.6) en choisissant pour $P_{\max}$ la partition à 256 intervalles de longueurs 1 de l'intervalle $I = [-1/2, 255 + 1/2]$. Chaque intervalle de cette partition représente un niveau de gris particulier.

Nous travaillons ici avec l'image Lena présentée en figure 5.7(a). L'histogramme de cette image construit sur la partition $P_{\max}$ est donné en figure 5.6 (a). En figure 5.6 (b) on trouve l'histogramme de cette même image construit cette fois sur la partition $\widehat{P}$ déterminée par le critère CRIT (5.6) et par (5.7).

La détermination d'une telle partition permet la quantification de l'image. Notons $\widehat{P} = (I_j)_{j=1,\dots,m}$. La quantification consiste à réassigner à chaque pixel appartenant à un certain I_j le niveau de gris égal à l'entier le plus proche de

$$\frac{1}{n_j} \sum_{i \,|\, x_i \in I_j} x_i.$$

L'image quantifiée est alors une image dont chaque pixel ne peut prendre que m

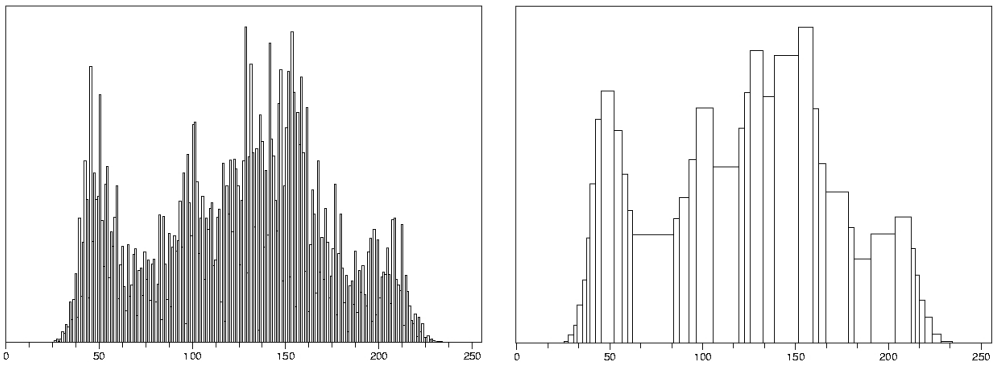


FIG. 5.6 – Les histogrammes de Lena sur la partition $P_{\max}$ (a) et sur la partition $\widehat{P}$ (b) à $\widehat{m} = 39$ intervalles déterminée par (5.6) et (5.7).

valeurs, au lieu des 256 habituelles. La figure 5.7 présente l'image Lena de base et sa quantification sur la partition $\widehat{P}$ à 39 intervalles présentée en figure 5.6.

Pour évaluer la distorsion entre ces deux images, nous utilisons la mesure habituelle PSNR (*Peak Signal to Noise Ratio*). Bien que sujette à discussion, elle est utilisée en compression d'images faute de critère psycho-visuel reconnu. Exprimée en décibels, cette quantité est définie entre deux images en niveaux de gris de même taille $l \times c$. Si ces images sont notées $x_{l,c}$ et $x'_{l,c}$, leur PSNR est donné par

$$\mathrm{PSNR}(x, x') = 10 \log_{10} \left(\frac{255^2}{\sum_{l,c} \left| x_{l,c} - x'_{l,c} \right|^2} \right). \tag{5.11}$$

Le PSNR entre deux images identiques est infini. Plus il est fort, plus les images se ressemblent. Habituellement, on considère que deux images ont le même rendu pour l'oeil humain si leur PSNR dépasse les 30 dB, avec les réserves précédentes. Le PSNR entre l'image Lena de base et sa quantification vaut ici 38,52 dB.

5.6.2 Sur l'hypothèse d'indépendance pour les images

Dans cette section, nous mettons en avant le fait que l'hypothèse d'indépendance des niveaux de gris d'une image est évidemment à remettre en cause. En effet, l'image est interprétable par le cerveau humain car ses pixels présentent une certaine cohérence qui ne saurait être modélisée par un échantillon indépendant d'une certaine loi. Notons que dans la partie précédente, si nous avions "mélangé" l'image Lena, *i.e.* appliqué une permutation sur ses pixels, la partition choisie par le critère (5.6) (figure 5.6) et la quantification résultante (figure 5.7) auraient été strictement identiques. Nous voyons donc que considérer les pixels comme indépendants revient à oublier toute l'information spatiale apportée par l'image.

Le critère à un ordre $k \geq 0$

Rappelons brièvement, que dans la partie 5.2, nous avons introduit un codage sans perte des données x^n à partir d'une partition P de l'intervalle dans lequel elles

(a) (b)

FIG. 5.7 – L'image Lena de base (a) et sa quantification sur la partition $\widehat{P}$ (figure 5.6(b)) à 39 intervalles. PSNR = 38,52 dB.

vivent. Ce codage est estimé par le critère CRIT (5.6) qui présente deux termes

1. le terme BIC(0) qui estime la longueur du codage arithmétique adaptatif à l'ordre 0 de la chaîne auxiliaire y^n,
2. le terme $L(x^n|y^n)$ qui représente le codage de précision nécessaire pour retrouver les données x^n.

Nous nous contentons ici d'essayer d'utiliser l'information spatiale apportée par l'image en augmentant l'ordre du codage arithmétique de y^n. On peut en effet penser que, si les pixels x^n sont corrélés, une CMM d'ordre $k > 0$ les modélisera mieux qu'une CMM(0) et que cette meilleure modélisation se transmettra à la chaîne y^n définie par (5.2).

Dans un premier temps, il nous faut choisir un moyen de linéariser l'image pour la transformer en un vecteur de données. Cette linéarisation se devant de conserver au mieux l'adjacence des pixels, nous choisissons le parcours de Hilbert illustré en figure 5.8. Notons que, dans le domaine de la compression d'images, on utilise la plupart du temps un parcours de type zig-zag dans des domaines transformés.

Une fois cette opération effectuée, nous disposons de la chaîne de données $x^n = x_1, \ldots, x_n$ où n est le nombre de pixels de l'image. Soit P une partition de $I = [-1/2, 255 + 1/2]$ dont le nombre d'intervalles est noté m et la chaîne y^n (5.2) associée. Nous définissons, pour $k \geq 0$

$$\text{CRIT}(k, P) \quad = \quad \text{BIC}(k) + L(x^n|y^n) \tag{5.12}$$

où BIC(k) est défini en (4.22) et se rapporte à la chaîne y^n à valeurs dans $E = [\![1, m]\!]$. Ce critère est une estimation de la longueur résultante du codage sans perte des données x^n à partir de la partition P ; ce codage étant effectué en deux temps :

1. codage arithmétique adaptatif à l'ordre k de y^n
2. codage de précision de x^n.

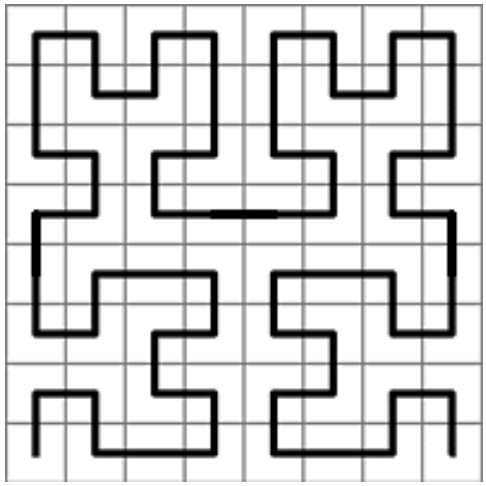


FIG. 5.8 – Le parcours de Hilbert

La quantité $n^{-1}\text{CRIT}(k, P)$ est une estimation du débit, en bits par pixel (bpp), obtenu par cette technique de codage sans perte.

Encore une fois, le principe MDL (partie 4.6.2) nous pousse à choisir pour partition P et pour ordre de codage k ceux qui minimisent (5.12).

Résultats

Nous travaillons encore avec l'image Lena présentée en figure 5.7(a). Pour des raisons de simplicité et de temps de calcul, nous considérons ici uniquement les partitions régulières de I à m intervalles que nous notons P_m pour $m = 1, \ldots, 256$. Remarquons que le nombre de degrés de liberté $(m-1)m^k$ présent dans la partie $\text{BIC}(k)$ (4.22) de notre critère $\text{CRIT}(k, P_m)$ (5.12) explose très rapidement avec m et k. C'est pour cette raison que nous présentons seulement des résultats pour $k = 0, 1$, $m = 1, \ldots, 256$.

La figure 5.9 présente les valeurs des débits estimés $n^{-1}\text{CRIT}(k, P_m)$ pour les valeurs de k et de m citées précédemment.

Comme attendu, effectuer le codage arithmétique adaptatif de y^n à l'ordre 1 diminue le débit, et donc le critère. Le principe MDL préconise ici de choisir l'ordre 1 et la partition P_{50} pour effectuer le codage de l'image Lena le plus efficacement possible. Le codage à l'ordre $k = 2$, non représenté ici, ne semble par apporter d'amélioration par rapport à l'ordre 1, montrant ainsi que les pixels de notre image sont liés seulement à leur voisin le plus proche au sens du parcours de Hilbert.

Remarquons aussi que, sur la figure 5.9, la valeur des débits pour $m = 1$ correspond aux formats .bmp ou .pgm de codage des images en niveaux de gris. Dans ces formats simples, 8 bits sont alloués à chaque pixel pour préciser la valeur de son niveau parmi les $256 = 2^8$ possibles. Par opposition, pour $m = 256$, les chaînes y^n et x^n coïncident et on observe en fait le débit obtenu par le codage arithmétique adaptatif direct de l'image. Par conséquent, le fait que les courbes de débit présentent un minimum montre qu'un mélange de ces deux codages est plus efficace que chacun d'eux séparément.

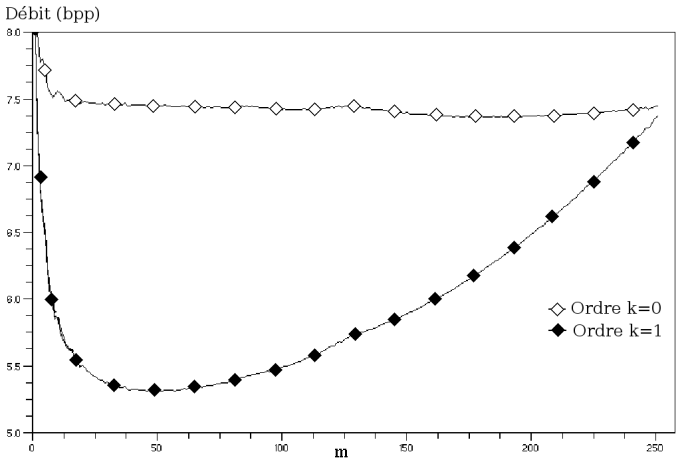


FIG. 5.9 – Les débits estimés $n^{-1}\mathrm{CRIT}(k, P_m)$ (5.12) sur l'image Lena.

Notons enfin que nous ne présentons pas ici de résultats efficaces de compression. Le meilleur débit que nous obtenons, environ 5 bpp pour un codage sans perte, est en effet très éloigné des performances des normes actuelles JPEG et JPEG2000 qui atteignent facilement les 0,3 ou 0,2 bpp avec une perte invisible à l'oeil. Pour obtenir des résultats en terme de compression, il serait intéressant d'insérer les techniques de codage présentées ici à la suite de la transformation en ondelettes de la norme JPEG2000 ou de la transformation Discrete Cosinus Transform (DCT) de la norme JPEG. On déterminerait alors, via un critère similaire à (5.12) appliqué aux coefficients transformés de l'image, l'ordre de corrélation résiduelle de ces coefficients ainsi que la partition qui réaliserait leur codage de manière optimale.

Résultats concernant les images quantifiées

Chaque partition régulière P_m pour $m = 1, \ldots, 256$ permet d'effectuer une quantification de l'image (voir partie 5.6.1). L'image quantifiée est encodée à l'aide seulement de la chaîne auxiliaire y^n définie en (5.2) ; rappelons que cette chaîne dépend de P_m et donc de m. La longueur du code résultante est estimée par le critère $\mathrm{BIC}(k)$ (4.22) avec $E = [\![1, m]\!]$ où k est l'ordre du codage arithmétique adaptatif effectué. La quantité $n^{-1}\mathrm{BIC}(k)$ est donc une estimation du débit nécessaire à l'encodage de l'image quantifiée sur la partition P_m.

Cette image quantifiée présente une certaine distorsion par rapport à l'image de base que nous mesurons par le PSNR (5.11). Sur la figure 5.10, on représente cette distorsion tracée contre le débit correspondant pour les ordres $k = 0$ et $k = 1$.

Pour illustration, nous présentons en figure 5.11 les deux images Lena quantifiées sur les partitions P_m pour $m = 3$ et $m = 13$ niveaux ainsi que leurs PSNR respectifs. Nous donnons également le débit atteint par les codages arithmétiques adaptatifs aux ordres $k = 0$ et $k = 1$ sur chacune de ces images. On observe par exemple que, pour un débit imposé d'environ 1,4 bpp, le codage à l'ordre 1 autorise un PSNR de 33,15 dB tandis que l'ordre 0 ne donne que 22,11 dB.

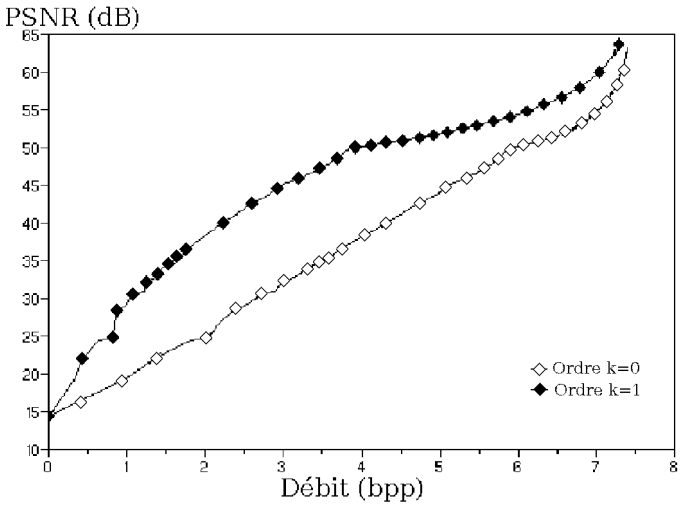


FIG. 5.10 – Les courbes débits-distorsion estimées de l'image Lena quantifiée sur les partitions régulières et encodée aux ordres $k = 0, 1$.

$m = 3$ niveaux : 22,11 dB
ordre 0 : 1,36 bpp
ordre 1 : 0,43 bpp

$m = 13$ niveaux : 33,15 dB
ordre 0 : 3,18 bpp
ordre 1 : 1,39 bpp

FIG. 5.11 – Débits et distorsion estimés de l'image Lena quantifiée sur $m = 3$ et $m = 13$ niveaux pour les codages aux ordres $k = 0$ et $k = 1$.

Résumé

Au cours de ce chapitre, le problème de l'estimation d'une densité inconnue par un histogramme est abordé. Nous élaborons une technique de codage en deux temps de l'échantillon à notre disposition. Cette technique repose sur le codage arithmétique adaptatif présenté dans le chapitre précédent. Une nouvelle application du principe *Minimum Description Length* nous pousse alors à minimiser la quantité CRIT (5.6) qui entre dans la catégorie des critères d'information (2.1) au sens entendu dans ce mémoire. Nous avons donc à nouveau justifié, en nous basant sur des considérations de type longueur de codage, l'utilisation d'un critère d'information pour le problème de la sélection d'histogramme.

Nous appliquons ensuite ce critère à quelques uns des domaines dans lesquels la construction d'un histogramme est requise. Le traitement d'une image, plus précisément sa quantification, est ainsi abordée. Dans ce cadre, le problème de l'indépendance des observations est également traité. Ces résultats font l'objet de publications plus détaillées présentées en annexe B et C. Nous nous intéressons enfin au problème de reconnaissance d'une loi parmi une certaine famille de lois candidate. Ici encore, l'approximation de la loi par un histogramme apporte une information non-négligeable pour la reconnaissance. Ces résultats font l'objet d'une soumission à une conférence internationale du groupe Institute of Electrical and Electronics Engineers IEEE présentée en annexe D.

Chapitre 6

Gaussian regression

This chapter is mainly an article entitled *Alternative utilizations of Information Criteria for Gaussian regression on a random design* which have been submitted to the journal European Series in Applied and Industrial Mathematics ESAIM : Probability and Statistics. Here, it is adapted to follow the continuity of this manuscript.

6.1 Motivation

We study the problem of Gaussian regression in the case called "random design" by Baraud [Bar02] or Birgé [Bir04]. In this setting, more precisely described in part 6.2.3, we consider a set of abscisses $x^n = (x_1, \ldots, x_n)$ that is a sample from a specified density w on an interval I of $\mathbb{R}$. The ordinates $y^n = (y_1, \ldots, y_n)$ are the images of those abscisses by an unknown function $f^\star$ deteriorated by a Gaussian white noise. By opposition, the more common "fixed design" studied for instance by Baraud [Bar00] consider the same problem except the abscisses are deterministic.

We will provide methods solving this model selection problem using Information Criteria. One of our main concern is that those method are both “efficient” and “fast”. The efficiency will be studied in terms of the behavior of the selected model as well as in terms of risks. As for the rapidity, we will require the methods to present an algorithm that does not require too many computations in view of an implementation on machines.

Firstly, we write the form the Information Criterion (2.1) takes in this setting. Then, in part 6.3.5, we recall the different methods under study and add a new one (part 6.3.5) specific to the regression setting. The asymptotic study, section 6.4, shows why those methods are efficient by giving the asymptotic behavior of the selected model. Result of this section are inspired by the work of Nishii [Nis88]. We give in section 6.5 some simulations results illustrating the convergence theorems given earlier. Section 6.6 is devoted to the computation of risks in the random design case. We stress the main differences between the fixed design setting. Finally, in section 6.7, we give an oracle inequality regarding the risk of the estimation of $f^\star$ resulting from a fast model selection method presented earlier.

6.2 Notations

6.2.1 Regression space

Let I be an interval of $\mathbb{R}$ endowed with the Lebesgue measure λ and w be a given nonnegative function with integral 1 assumed to vanish only on a set of I of Lebesgue measure 0. The following defines a scalar product on $L^2 := L^2(I, wd\lambda)$

$$\langle f, g \rangle_w = \int_I fgwd\lambda. \tag{6.1}$$

whose associated norm is denoted by $\|.\|_w$

We choose $F = \text{Vect}(f_1, \ldots, f_d)$ a d-dimensional subspace of L^2 with $d \leq n$. Let us stress that, contrarily to work of Baraud [Bar00, Bar02], our d is not allowed to depend on n. We denote by M_w the Gram matrix of the f_j's

$$M_w(j,k) = \langle f_j, f_k \rangle_w, \ j,k = 1, \ldots, d.$$

For any support $S \subset [\![1, d]\!]$ we denote by F_S the $|S|$-dimensional subspace of F

$$F_S = \text{Vect}\{f_j, \ j \in S\} \tag{6.2}$$

and by $M_{w,S}$ the associated Gram matrix $M_{w,S}(j,k) = \langle f_j, f_k \rangle_w, \ j,k \in S$. In F, the orthogonal projector on F_S is denoted by Π_S^F.

In the case where $f^\star$ actually lives in F, we write $f^\star = \sum a_j^\star f_j$ and call $S^\star$ its support

$$S^\star = \{j \text{ such that } a_j^\star \neq 0\}. \tag{6.3}$$

In the sequel it will be convenient to identify F to $\mathbb{R}^d$ via $F \to \mathbb{R}^d$, $f = \sum_{k=1}^d a_k f_k \mapsto a = (a_1, \ldots, a_d)^T$. This way, the squared norm of a function may be written as $\|f\|_w^2 = f^T M_w f$.

We will also need to apply central limit theorem several times in the sequel. In order to ensure that all considered variables have a variance, we assume from now on that there exists a $\eta > 0$ such that

$$f^\star, f_j \in L^{4+\eta}(w), \ j = 1, \ldots, d. \tag{6.4}$$

Note that we are constrained to suppose that $f^\star$ presents this integrability only in the case where it does not belong to the space of regression F.

6.2.2 Observations space

We endow the observations space $\mathbb{R}^n$ with the canonical scalar product $\langle ., . \rangle$ and its associated norm $\|.\|$. When $x \in I^n$ and a function f are given, we denote by $f(x)$ the vertical vector of $\mathbb{R}^n$ $f(x) = (f(x_1), \ldots, f(x_n))^T$. For any support S we define E_S as the subspace

$$E_S = \text{Vect}\{f_j(x), \ j \in S\} \tag{6.5}$$

and shorten $E_{[\![1,d]\!]}$ to E. We assume in the sequel that the family $(f_1(x), \ldots, f_d(x))$ is linearily independant, which makes each E_S a $|S|$-dimensional space. In $\mathbb{R}^n$, the orthogonal projector on E_S is denoted by Π_S^E.

6.2.3 Modelisation

Let $X = X_1, \ldots, X_n$ be independent variables with density w on I. Let also G be a n-dimensional gaussian white noise $G \sim \mathcal{N}(0, \sigma^2 \mathrm{id}_n)$ independent of X. We assume that the variance σ^2 is known to the user. If this is not the case, Baraud suggests in [Bar00] to replace it by a suitable estimation built solely on the data; then he shows that this procedure gives results as good as if σ^2 was known. We do not consider this case here.

We modelize the abscisses X and ordinates Y of a set of n points in the plane as follows :

$$Y = f^\star(X) + G \tag{6.6}$$

Namely, the abscisses are given by a sample of the law w and the ordinates by the images of those abscisses by the unknown function $f^\star$ deteriorated by the noise. Note that all the E_S (6.5) and their projection Π_S^E become random.

For any support $S \subset [\![1, d]\!]$, let us set the following model Θ_S :

$$\Theta_S = \{\mathcal{L}(f(X) + G) |\ f \in F_S\}\,. \tag{6.7}$$

Those are the law of n-dimensional variables which write as $f(X)+G$ where $f \in F_S$. If $f^\star$ lives in F and has support $S^\star$, the law of Y belongs to the model $\Theta_{S^\star}$.

6.3 Information criteria and their use

6.3.1 Maximum likelihood

Relatively to a function $f \in F$, a realization (x, y) of (X, Y) has the following likelihood :

$$L(f) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} w(x_i) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i))^2\right).$$

In the sequel we shall rather work with $l = -2\sigma^2 \ln(L(f))$. Since we are only concerned in minimization of l with respect to f, we drop terms not depending on it and l may be seen as

$$\begin{aligned} l : F \sim \mathbb{R}^d &\to \mathbb{R} \\ f &\mapsto \sum_{i=1}^{n} (y_i - f(x_i))^2 \\ a &\mapsto \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{d} a_k f_k(x_i)\right)^2. \end{aligned}$$

Consequently, for any support S, the maximization of the likelihood of y within the model Θ_S amounts to the minimization of the quadratic error l on F_S. Then the (opposite of the) maximum (log-)likelihood of y with respect to the model Θ_S writes as :

$$\min_{f \in F_S} l(f) = d^2(y, E_S) =: l(\widehat{f}_S). \tag{6.8}$$

6.3.2 Estimators $\widehat{f}_S$, $\widetilde{f}_S$

In order to express asymptotics of this likelihood as well as to give a matricial expression of the function $\widehat{f}_S \in F_S$ that realizes the minimum in (6.8), it will be convenient to use the Vandermonde-type $n \times d$ matrix V depending only of the x_i's :

$$V = \begin{pmatrix} f_1(x_1) & \dots & f_d(x_1) \\ \vdots & \vdots & \vdots \\ f_1(x_n) & \dots & f_d(x_n) \end{pmatrix}$$

as well as the $d \times d$ Gram matrix $M = V^T V$. The passage from a function f to the vector $f(x)$ is then a simple multiplication $f(x) = Vf$. Also note that, in the case where $f^\star \in F$, we may write $y = Vf^\star + g$ and the Taylor expansion of l at any function f writes as

$$\begin{aligned} l(h) - l(f) &= (\text{grad } l_f)^T (h - f) + \frac{1}{2}(h-f)^T \text{Hess } l_f (h-f) \\ &= -2\left(M(f - f^\star) - V^T g\right)^T (h - f) + (h-f)^T M (h-f). \end{aligned} \tag{6.9}$$

Now let us set D_S the $d \times |S|$ matrix of zeros and ones such that $V_S := VD_S$ contains columns j of V for $j \in S$ only. We also set the Gram matrix $M_S = V_S^T V_S$. Consequently, the likelihood (6.8) writes as

$$L(y|\Theta_S) = d^2(y, E_S) = \frac{\text{Gram}\left(\{y, f_j(x),\, j \in S\}\right)}{\det\left(M_S\right)}. \tag{6.10}$$

where $\text{Gram}(u_1, \dots, u_k)$ denotes the Gram determinant of those vectors. Moreover the matrix of the orthogonal projection Π_S^E of $\mathbb{R}^n$ onto E_S is

$$\text{Mat}(\Pi_S^E) = V_S M_S^{-1} V_S^T \tag{6.11}$$

which gives the following expression for $\widehat{f}_S$ realizing the minimum in (6.8) :

$$\widehat{f}_S = D_S M_S^{-1} V_S^T y \in F_S. \tag{6.12}$$

It is important to note that, unlike the fixed points case, it may happen that $\left\|\widehat{f}_S\right\|_w^2$ is not integrable. To handle this problem we will use at some points in the sequel a truncated estimator

$$\widetilde{f}_S = \widehat{f}_S . \mathbb{1}_{\{\|nM_S^{-1}\| < C\}} \tag{6.13}$$

where $\|.\|$ is any norm on matrices and C is a constant satisfying $C > \|M_{w,S}^{-1}\|$. Note that, from the law of large numbers, M_S/n converges to $M_{w,S}$ a.s. so that $\mathbb{1}_{\{\|nM_S^{-1}\| < C\}} \to 1$ a.s. Now, from (6.12), the squared norm of $\widetilde{f}_S$ writes as

$$\left\|\widetilde{f}_S\right\|_w^2 = \mathbb{1}_{\{\|nM_S^{-1}\| < C\}} \left(D_S M_S^{-1} V_S^T y\right)^T M_{w,S} \left(D_S M_S^{-1} V_S^T y\right)$$

with $y = f^\star(x) + g$. In that expression, matrices D_S and $M_{w,S}$ are deterministic and the indicator function ensures the boundedness of M_S^{-1}. Moreover, the integrability

we have requested on $f^\star$ and on the f_j's in (6.4) along with the fact that G is Gaussian ensure that

$$\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right] < \infty \text{ for any } p \text{ such that } 2p \leq 4+\eta. \tag{6.14}$$

Furthermore, for an event A, Hölder's inequality gives

$$\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2}\mathbb{1}_A\right] \leq \mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right]^{1/p}\mathbb{P}(A)^{1/q}. \tag{6.15}$$

for any $p>1$ such that $2p \leq 4+\eta$ and $1/p+1/q=1$.

Now, since all f_j's belong to $L^4(w)$, any variable of the form $f_j(X)f_k(X)$ has a variance smaller than

$$V_{max} = \max\left\{V\left(f^\star(X)^2\right), V\left(G_1^2\right), V\left(f_j(X)f_k(X)\right),\ j,k=1,\ldots,d\right\}. \tag{6.16}$$

Consequently we may apply central limit theorem to obtain a control given by, for any $\varepsilon>0$,

$$\mathbb{P}\left(\|M_s/n - M_{w,S}\| > \varepsilon\right) = O\left(\exp\left(\frac{-n\varepsilon^2}{4V_{max}}\right)\right).$$

Since the event $\{\|nM_S^{-1}\| < C\}^c$ is included in an event of the form $\|M_s/n - M_{w,S}\| > \varepsilon_C$ for a certain $\varepsilon_C>0$ depending on C, we obtain that

$$\mathbb{E}\left[\mathbb{1}_{\{\|nM_S^{-1}\|<C\}^c}\right] = O\left(\exp\left(\frac{-n\varepsilon_C^2}{4V_{max}}\right)\right). \tag{6.17}$$

6.3.3 Asymptotics of the maximum likelihood

Because all the f_j's belong to $L^2(w)$, the law of large numbers ensures that each entry $M_S(i,j)$ of M_S satisfies :

$$\frac{1}{n}M_S(i,j) \to M_{w,S}(i,j),\ a.s. \tag{6.18}$$

Moreover, since $y = f^\star(x)+g$ and by independence between X and G, the entry (1,1) of the first Gram matrix in (6.10) satisfies

$$\frac{1}{n}\left\langle f^\star(x)+g, f^\star(x)+g\right\rangle \to \left\langle f^\star, f^\star\right\rangle_w + \sigma^2\ a.s. \tag{6.19}$$

Any other entry converges to the corresponding one of the Gram matrix of functions $f^\star, f_j, j\in S$. Consequently we get :

$$\frac{1}{n}l(\widehat{f}_S) \to \sigma^2 + d^2(f^\star, F_S) \text{ a.s.} \tag{6.20}$$

Now we need to control the speed of the convergence in (6.20). Every variables for whom the law of large numbers has been used in (6.18) and (6.19) have a variance because of (6.4) and because the error G is Gaussian. All those variances are smaller than V_{max} defined earlier in (6.16). Then we may apply central limit theorem to obtain that convergences in (6.18) and (6.19) have a speed given by $O\left(\exp(-n\varepsilon^2/4V_{max})\right)$, for any $\varepsilon>0$. That speed passes to the determinant to give

$$\mathbb{P}\left(\left|\frac{1}{n}l(\widehat{f}_S) - \sigma^2 - d^2(f^\star, F_S)\right| > \varepsilon\right) = O\left(\exp\left(\frac{-n\varepsilon^2}{4V_{max}}\right)\right). \tag{6.21}$$

6.3.4 The criterion

From part 6.3.1, the general information criterion (2.1) should become here

$$\mathrm{IC}(S) = \frac{l(\widehat{f}_S)}{\sigma^2} + |S|\beta(n). \tag{6.22}$$

However, we have assumed that σ^2 is known to the user and are only interested in minimization of IC relatively to S. Consequently, we may set

$$\mathrm{IC}(S) := l(\widehat{f}_S) + |S|\alpha(n). \tag{6.23}$$

where $\alpha = \sigma^2\beta$ is the penalty function.

Note that, in the case where the regression error G in (6.6) is not Gaussian, the likelihood term in (6.23) does not equal $d^2(y, E_S)$. Consequently, in this case, the quantity $d^2(y, E_S) + |S|\alpha(n)$ may not be called an information criterion in the general sense given by (2.1).

We will use this criterion along with fast methods presented in part 2.4.

6.3.5 Methods

As stressed in part 2.4, the methods we use allow to select a support that may take any value in $\mathcal{P}([\![1, d]\!])$. In this sense, our basis $(f_j)_j$ is not required to present a natural order as, for instance, basis of polynomials or wavelets would do. Our methods will always select a set of functions to be used for regression, without favouring any of them and regardless of which kind of functions are mixed in the basis.

We briefly recall those methods here and add a new one, the adapted reversed comparative method, part 6.3.5, specific to our setting. We will always denote the selected support by $\widehat{S}$ but in the sequel the context will allow to determine which method is used.

Global method

The estimated support is chosen as

$$\widehat{S} = \mathrm{Argmin}\left\{\mathrm{IC}(S), S \in \mathcal{P}([\![1, d]\!])\right\}. \tag{6.24}$$

Comparative method

The expression "$-j$" denotes the support $[\![1, d]\!] \setminus \{j\}$. The estimated support is chosen via

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}([\![1, d]\!]) \\ \widehat{S} &= \left\{j \in [\![1, d]\!] \text{ such that } \mathrm{IC}_{\mathrm{ref}} \leq \mathrm{IC}(-j)\right\} \end{aligned} \tag{6.25}$$

Reversed comparative method

We estimate the support by

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}(\emptyset) \\ \widehat{S} &= \left\{ j \in [\![1,d]\!] \text{ such that } \mathrm{IC}(\{j\}) \leq \mathrm{IC}_{\mathrm{ref}} \right\} \end{aligned} \tag{6.26}$$

The asymptotic flaws of this method will be precisely discussed in part 6.4.2. However, we may already give some comments about it. Suppose the basis $(f_j)_j$ is orthogonal relatively to the scalar product (6.1), that $f^\star$ belongs to F and writes $f^\star = \sum_j a_j f_j$. Then from the asymptotic behaviour of likelihood terms in our criterion given in (6.20), we obtain that the difference $n^{-1}(\mathrm{IC}(\{j\}) - \mathrm{IC}(\emptyset))$ behaves as $n^{-1}\alpha(n) - a_j^2$. Then the decision of the reversed method regarding the function f_j is related only to the coefficient a_j of $f^\star$ which vanishes if and only if f_j is to be rejected. Note that, in the same setting, the regular comparative method (6.25) is interested in the difference $n^{-1}(\mathrm{IC}_{[\![1,d]\!]} - \mathrm{IC}(-j))$ that also behaves as $n^{-1}\alpha(n) - a_j^2$ because of (6.20). Those two methods, in the orthogonal case, are asymptotically equivalent.

Now for the non-orthogonal case, still from (6.20), the difference $n^{-1}(\mathrm{IC}(\{j\}) - \mathrm{IC}(\emptyset))$ only behaves as $n^{-1}\alpha(n) + d^2(f^\star, F_j) - \|f^\star\|_w^2$. Here, even in the case where f_j is to be rejected, the term $d^2(f^\star, F_j) - \|f^\star\|_w^2$ remains negative, requiring a stronger penalization for the reversed method to actually reject f_j. In fact, it will be shown in the asymptotic study that this need of a large penalization may almost not be fulfilled without rejecting every functions. By opposition, still in the non-orthogonal case, for the regular comparative method (6.25) the difference $n^{-1}(\mathrm{IC}_{[\![1,d]\!]} - \mathrm{IC}(-j))$ behaves as $n^{-1}\alpha(n) - d^2(f^\star, F_{-j})$ where the term $d^2(f^\star, F_{-j})$ vanishes when f_j is to be rejected. In this sense, the regular comparative method is not affected by the orthogonality of the basis.

To sum up those comments, let us just say that in the orthogonal case, reversed and regular comparative methods have the same asymptotic behaviour whereas in the non-orthogonal case the reversed method suffers flaws that will be more precisely described in the part 6.4.2 of the asymptotic study.

Finally, let us say that the reversed method might be transposed to the case where the space of regression has infinite dimension. Indeed, even though the basis (in the classical or Hilbert sense) of F was infinite, using the reversed method would never require to compute an IC with an infinite number of free parameters. By opposition the reference of the regular comparative method, as well as all others criteria needed to select $\widehat{S}$ in (6.25), are not computable in the infinite dimensional case.

Adapted reversed comparative method

We give here an adaptation of the reversed comparative method that will be shown to avoid issues occuring with the previous one. To this end we need to define new functions f_j^N ; the superscript N stands for normal. Indeed, each function f_j^N is chosen to be normal, relatively to the scalar product (6.1), to the hyperplane F_{-j} defined in (6.2). The orientation of f_j^N as well as its norm does not matter in the sequel.

We may then define new alternative models as in (6.7) :

$$\Theta_j^N = \left\{ \mathcal{L}(f(X)+G) |\ f = a_j^N f_j^N \right\},\ a_j^N \neq 0.$$

For any j, the family $\{f_k, k \neq j\} \cup \{f_j^N\}$ is a basis of F and a function does not live in F_{-j} if and only if it has a component along f_j^N. The idea comes that, instead of determining wether $f^\star$ should have a non-vanishing component along f_j, we will determine if it should have one along f_j^N by following the so called adapted reversed comparative method :

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}(\emptyset) \\ \widehat{S} &= \left\{ j \in [\![1,d]\!] \text{ such that } l(\widehat{f}_j^N) + \alpha(n) \leq \mathrm{IC}_{\mathrm{ref}} \right\}. \end{aligned} \tag{6.27}$$

Note that the basis $(f_j^N)_j$ is orthogonal if and only $(f_j)_j$ is. In this case, f_j^N is colinear to f_j which makes (6.26) and (6.27) equivalent methods of selection of $\widehat{S}$.

In the sequel, it will be shown that under some assumptions on the penalty, this method satisfies a convergence theorem 6.4.3 . Let us stress that, even though this theorem applies, this method suffers the same flaws as the reversed comparative method (6.26) since it remains a *reversed* method. Indeed, as in comments made before, when f_j is to be rejected, the adapted reversed comparative method is strongly affected by the orthogonality of f_j^N and $f^\star$. However, calculations are not made with those function but rather with $f_j^N(X)$ and $f^\star(X)$ that are orthogonal only asymptotically. Consequently, for a fixed n, we also expect this method to require a larger penalization in order to avoid overparametrization.

Descending comparative method

The descending comparative method is designed especially in the aim of using results from Baraud in [Bar00, Bar02]. It requires a random computation of ICs. Firstly let us set

$$\begin{aligned} S^{(0)} &= [\![1,d]\!] \\ \mathrm{IC}_{\mathrm{ref}}^{(0)} &= \mathrm{IC}(S^{(0)}). \end{aligned}$$

The first step of the descending method produces new quantities that have superscript (1) as follows

$$\begin{aligned} C^{(1)} &= \left\{ j \in S^{(0)},\ \mathrm{IC}\left(S^{(0)} \setminus \{j\}\right) \leq \mathrm{IC}_{\mathrm{ref}}^{(0)} \right\} \\ J^{(1)} &= \mathrm{Argmin}\left\{ \mathrm{IC}\left(S^{(0)} \setminus \{j\}\right),\ j \in C^{(1)} \right\}. \end{aligned} \tag{6.28}$$

This way, among the functions of $C^{(1)}$ found useless by the criterion, $J^{(1)}$ is the worst one. This is consequently the function we should remove in priority. This is what we do now by refreshing our reference with superscript (1) :

$$\begin{aligned} S^{(1)} &= S^{(0)} \setminus \{J^{(1)}\} \\ \mathrm{IC}_{\mathrm{ref}}^{(1)} &= \mathrm{IC}(S^{(1)}). \end{aligned} \tag{6.29}$$

From there, we start a second step by computing useless functions and the worst one by

$$\begin{aligned} C^{(2)} &= \left\{ j \in S^{(1)},\ \mathrm{IC}\left(S^{(1)} \setminus \{j\}\right) \leq \mathrm{IC}^{(1)}_{\mathrm{ref}} \right\} \\ J^{(2)} &= \mathrm{Argmin}\left\{ \mathrm{IC}\left(S^{(1)} \setminus \{j\}\right),\ j \in C^{(2)} \right\} \end{aligned}$$

and refresh again our reference by adding 1 to all superscripts in (6.29).

This process is repeated until the random step $k_f + 1$ where $C^{(k_f+1)} = \emptyset$. This means that the criterion does not reject functions anymore and that the current support $S^{(k_f)}$ should be our estimator $\widehat{S}$. We say that the procedure stops at step k_f for "k final".

6.3.6 Methods complexities

Table 2.2 gives the complexities of the comparative methods used here. The adapted reversed comparative method, specific to the regression problem, present a complexity that equals $d + 1$. As stressed earlier, the interest of the comparative methods is that they require less computations in order to determine a support that is as precise as the one produced by the global methods.

The remainder of the chapter is devoted to explain in which ways the comparative methods are also "efficient".

6.4 Asymptotic study

In all our asymptotic study, we assume that $f^\star$ lives in F and has support $S^\star$ as in (6.3). Our concern is then to determine conditions on our criterion (6.23), more precisely on its penalty term, to obtain some convergence of $\widehat{S}$ to $S^\star$ as n grows. The main results are given in theorems 6.4.1, 6.4.2, 6.4.3, 6.4.4.

6.4.1 Asymptotics of the comparative method

Here we work with the comparative method (6.25) and criterion (6.23). The main result is as follows.

Theorem 6.4.1 *If the penalty $\alpha(n)$ of the criterion (6.23) satisfies*

(i) $\alpha(n) = o(n)$

(ii) $\ln \ln n = o(\alpha(n))$

then the method (6.25) is strongly consistent in the sense that $\widehat{S}$ converges (stationnarily) to $S^\star$ almost surely.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ and $\widehat{S} \subset S^\star$ a.s. above a certain rank.

Remark. The proof actually shows that we might relax condition (i) to

$$\alpha(n) = c\,n \text{ where } c < \min\left\{ d^2(f^\star, F_{-j}),\ j \in S^\star \right\}$$

but that min is not available to the user.

Proof : Let us split it into two parts.

$\star$ *First part* : the case $j \in S^\star$. Here, let us use asymptotics of the likelihood term given in (6.20) to write :

$$\frac{1}{n}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) - \alpha(n)\right) = -d^2(f^\star, F_{-j}) + o(1) \text{ a.s.}, \tag{6.30}$$

where that latter distance $D^2 := d^2(f^\star, F_{-j})$ does not vanish since $j \in S^\star$.

Again because $j \in S^\star$, one would like to select j as a part of $\widehat{S}$; in other terms one would like $\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j)$ to be nonpositive. It suffices to choose $\alpha(n)$ such that $\alpha(n) = o(n)$ to ensure that fact a.s. for n large enough.

Taking a not too large penalty (of order $o(n)$) thus ensures that we do not reject basis functions f_j that actually appear in the unknown function $f^\star$, this is the first statement of the theorem.

In this case, let us take n large enough to make $0 \leq \alpha(n)/n < D^2/2$ to get the following control :

$$\begin{aligned}
\mathbb{P}(j \notin \widehat{S}) &= \mathbb{P}\left(n^{-1}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j)\right) \geq 0\right) \\
&\leq \mathbb{P}\left(n^{-1} l(\widehat{f}_{[\![1,d]\!]}) - \left(n^{-1} l(\widehat{f}_{-j}) - D^2\right) > D^2/2\right) \\
&\leq \mathbb{P}\left(\left|n^{-1} l(\widehat{f}_{[\![1,d]\!]}) - \sigma^2\right| > D^2/4\right) \\
&\quad + \mathbb{P}\left(\left|n^{-1} l(\widehat{f}_{-j}) - \sigma^2 - D^2\right| > D^2/4\right).
\end{aligned}$$

Hence, from (6.21) :

$$\mathbb{P}(j \notin \widehat{S}) = O\left(\exp\left(-\frac{nD^4}{64 V_{max}}\right)\right), \ j \in S^\star \tag{6.31}$$

which will be useful in studies of risks to come later.

$\star\star$ *Second part* : the case $j \notin S^\star$. Recall the definition of the (random) matrix M in part 6.3.2. Since each f_j is in L^2 the law of large numbers gives

$$M/n \longrightarrow M_w \text{ a.s.}$$

that latter limit being invertible and positive definite. Consequently each entry of nM^{-1} is bounded a.s. at least above a certain rank, which we denote by

$$M^{-1} = O\left(\frac{1}{n}\right) \text{ a.s.} \tag{6.32}$$

More precisely, if m_j is the j-th diagonal coefficient of M^{-1}, then $n.m_j$ goes a.s. to the j-th diagonal coefficient of M_w^{-1} which is positive and

$$\frac{1}{m_j} = O(n) \text{ a.s.} \tag{6.33}$$

We call here for brievety $\widehat{f}$ and $\widehat{f^j}$ the functions in F and F_{-j} respectively maximizing the likelihood on those spaces as in (6.8). From (6.9), the quantities we are interested in are :

$$l\left(\widehat{f^j}\right) - l\left(\widehat{f}\right) = \left(\widehat{f^j} - \widehat{f}\right)^T M \left(\widehat{f^j} - \widehat{f}\right) \tag{6.34}$$

since $\widehat{f}$ satisfies grad $l\big(\widehat{f}\big) = 0$.

Now for $\widehat{f^j}$, it satisfies

$$\Big(\text{grad } l(\widehat{f^j})\Big)^T = M\widehat{f^j} - M\widehat{f} = (0, \ldots, 0, \lambda_j, 0, \ldots, 0)^T \tag{6.35}$$

where λ_j is a Lagrange coefficient set at the j-th place in the latter vector which necessarily satisfies

$$\lambda_j = -\frac{\widehat{f}_j}{m_j} \tag{6.36}$$

since $\widehat{f}^j_j$, the j-th coefficient of $\widehat{f^j}$, must vanish.

Plugging (6.33), (6.35) and (6.36) in expansion (6.34) gives

$$l\big(\widehat{f^j}\big) - l\big(\widehat{f}\big) = \Big(\widehat{f}_j\Big)^2 O(n) \tag{6.37}$$

Now note that $\widehat{f}$ is defined by $\widehat{f} = f^\star + M^{-1}V^T g$ which yields $\mathbb{E}[\widehat{f}] = f^\star$ keeping in mind that matrices V and M only depend on X which is independent of G. Consequently, in our case $j \notin S^\star$, we get

$$\widehat{f}_j = 0 + \big(M^{-1}V^T g\big)_j .$$

Any entry in the vector $V^T g$ is of the form $\sum_{i=1}^n f_k(x_i)g_i$ which is the sum of n independent variables with mean 0, finite variance from (6.4), and thus is $O(\sqrt{n \ln\ln n})$ a.s. by the law of iterated logarithm. The order of M^{-1} given in (6.32) makes $\widehat{f}_j = O\left(\sqrt{\frac{\ln\ln n}{n}}\right)$ a.s.

Plugging in (6.37), we finally get

$$l\big(\widehat{f^j}\big) - l\big(\widehat{f}\big) = O(\ln\ln n) \text{ a.s.}$$

Now in our case $j \notin S^\star$, one would like to reject j; in other terms, on would like $\text{IC}_{\text{ref}} - \text{IC}(-j)$ to be nonnegative. Write

$$\text{IC}_{\text{ref}} - \text{IC}(-j) = l\big(\widehat{f}\big) - l\big(\widehat{f^j}\big) + \alpha(n) = \alpha(n) + O\left(\ln\ln n\right) \text{ a.s.}$$

so that

$$\frac{1}{\ln\ln n}\left(\text{IC}_{\text{ref}} - \text{IC}(-j) - \alpha(n)\right) = O(1) \text{ a.s.} \tag{6.38}$$

Now it suffices to choose $\alpha(n)$ such that $\ln\ln n = o(\alpha(n))$ to ensure that a.s. and for n large enough, $\text{IC}_{\text{ref}} - \text{IC}(-j) > 0$.

We have shown that taking a large enough penalty ensures that we reject basis functions f_j which do not appear in $f^\star$, this is the second statement of the theorem. □

6.4.2 Asymptotics of the reversed comparative method

We work here with the method (6.26).

Asymptotics of the likelihood difference

The function achieving the maximum likelihood for $\mathrm{IC}_{\mathrm{ref}}$ obviously vanishes and the function with support $\{j\}$ achieving it for $\mathrm{IC}(\{j\})$ is the new function $\widehat{f^j}$ whose component along f_j is :

$$\widehat{f_j^j} = \frac{\langle f_j(x), y\rangle}{\langle f_j(x), f_j(x)\rangle}$$

Now the difference of the log-likelihoods is easier to compute than with the regular comparative method :

$$\begin{aligned} l(0) - l(\widehat{f^j}) &= \sum_{i=1}^{n} \left(y_i^2 - \left(y_i - \frac{\langle f_j(x), y\rangle}{\langle f_j(x), f_j(x)\rangle} f_j(x_i) \right)^2 \right) \\ &= \frac{\langle f_j(x), y\rangle^2}{\langle f_j(x), f_j(x)\rangle} \end{aligned}$$

In order to control the asymptotics, we firstly use the law of that large numbers to write $\langle f_j(x), f_j(x)\rangle^{-1} = O(1/n)$ a.s. Moreover, two applications of the law of iterated logarithm give, a.s. :

$$\langle f_j(x), y\rangle = \langle f_j(x), f^\star(x)\rangle + \langle f_j(x), g\rangle = n \langle f_j, f^\star\rangle_w + O\left(\sqrt{n \ln\ln n}\right).$$

Consequently, a.s. :

$$l(0) - l(\widehat{f^j}) = n \langle f_j, f^\star\rangle_w^2 + \langle f_j, f^\star\rangle_w O\left(\sqrt{n \ln\ln n}\right) + O(\ln\ln n). \tag{6.39}$$

Non-orthonormal case

Here appears the main problem about the reversed method (6.26). Suppose we are in a case where $\langle f_j, f^\star\rangle_w = \sum_{k\in S^\star} a_k \langle f_j, f_k\rangle_w$ vanishes for no index j. This happens most of the times if the basis $(f_k, k \in [\![1, d]\!])$ is not orthonormal even though $j \notin S^\star$. Then formula (6.39) gives

$$\frac{1}{n}\left(\mathrm{IC}_{\mathrm{ref}}^r - \mathrm{IC}(\{j\}) + \alpha(n)\right) = \langle f_j, f^\star\rangle_w^2 + o(1). \tag{6.40}$$

Now assume $\alpha(n) = o(n)$, then $\mathrm{IC}_{\mathrm{ref}}^r - \mathrm{IC}(j) > 0$ a.s. above a certain rank and this for all j. This means we keep every indices $j \in [\![1, d]\!]$. Here the condition $\alpha(n) = o(n)$ which ensured we kept good indices with the regular comparative method (see theorem 6.4.1) turns out to make us keep every indices.

One should thus think about taking a penalty a bit larger. However, formula (6.40) also implies that if

$$\alpha(n) = Cn \text{ where } C > \max\left\{\langle f_j, f^\star\rangle_w^2,\ j \in [\![1, d]\!]\right\}$$

then $\mathrm{IC}_{\mathrm{ref}}^r - \mathrm{IC}(j) < 0$ a.s. above a certain rank and this for all j. This means we reject every indices $j \in [\![1, d]\!]$.

Therefore, in order to obtain a result similar to theorems 6.4.1 one should firstly choose a penalty of order not smaller than n and not greater than Cn to ensure that we do not have a criterion that accepts or rejects systematically every indices. In

certain cases, there exists a good order for the penalty in the few place left between n and Cn as follows.

$$\alpha(n) = kn \text{ where } \min\left\{\langle f_j, f^\star\rangle_w^2,\ j \in S^\star\right\} > k > \max\left\{\langle f_j, f^\star\rangle_w^2,\ j \notin S^\star\right\}.$$

However, that k might not exist if its bounds are not ordered the correct way; moreover, it is unavailable to the user and thus not of a practical use.

In the case where the basis (f_j) is orthonormal, those issues disappear and we establish in the next part a convergence theorem for the reversed comparative method.

Orthonormal case

We assume here that the basis $(f_j, j \in [\![1, d]\!])$ is orthonormal relatively to the scalar product (6.1). Note that this is often the case; let us cite for instance the situation where one needs to decompose a measured signal on a wavelet or a Fourier basis (in this case, w is the uniform density on the corresponding interval).

Recall that $f^\star = \sum_{j\in S^\star} a_j f_j$, consequently $a_j = \langle f_j, f^\star\rangle_w = 0 \Leftrightarrow j \notin S^\star$ and (6.39) yields

$$\mathrm{IC}^r_{\mathrm{ref}} - \mathrm{IC}(j) + \alpha(n) = na_j^2 + a_j O\left(\sqrt{n\ln\ln n}\right) + O(\ln\ln n) \text{ a.s.}$$

This formula is to be related to equations (6.30) and (6.38) concerning the regular comparative method to see that, in the orthonormal case, the reversed method behaves asymptotically as the regular method. The following theorem is derived from considerations similar to the proof of theorem 6.4.1.

Theorem 6.4.2 *In the orthonormal case and under the following assumptions on the penalty :*

(i) $\alpha(n) = o(n)$

(ii) $\ln\ln n = o(\alpha(n))$

the selection of indexes done by the reversed comparative method (6.26) is strongly consistent.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ *and* $\widehat{S} \subset S^\star$ *a.s. above a certain rank.*

The orthonormal case as treated here gives the idea of considering the adapted reversed comparative method (6.27). In the next part, we show that this method is consistent without the orthonormality hypothesis.

6.4.3 Asymptotics of the adapted reversed comparative method

We work here with the method (6.27) and fix $j \in [\![1, d]\!]$. Arguments in part 6.4.2 may be transposed here simply by replacing f_j with f_j^N and noting that $\langle f_j^N, f^\star\rangle_w = a_j^N$. We get a formula similar to (6.39) :

$$\mathrm{IC}^N_{\mathrm{ref}} - \mathrm{IC}^N(j) + \alpha(n) = n\left(a_j^N\right)^2 + a_j^N O\left(\sqrt{n\ln\ln n}\right) + O(\ln\ln n).$$

Now recalling that $a_j^N \neq 0 \Leftrightarrow j \in S^\star$, we obtain a result similar to theorem 6.4.2 for the adapted reversed comparative method without the orthonormality hypothesis.

Theorem 6.4.3 *Under the following assumptions on the penalty :*

(i) $\alpha(n) = o(n)$

(ii) $\ln \ln n = o(\alpha(n))$,

the selection of indexes by the adapted reversed comparative method (6.27) is strongly consistent.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ *and* $\widehat{S} \subset S^\star$ *a.s. above a certain rank.*

6.4.4 Asymptotics of the descending comparative method

We work here with the method described in part 6.3.5 and show the following theorem

Theorem 6.4.4 *Under the following assumptions on the penalty :*

(i) $\alpha(n) = o(n)$

(ii) $\ln \ln n = o(\alpha(n))$,

the selection of indexes by the descending comparative method in part (6.3.5) is strongly consistent.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ *and* $\widehat{S} \subset S^\star$ *a.s. above a certain rank.*

Proof : Under assumptions of that theorem, we may apply theorem 6.4.1 concerning the regular comparative method. Then, with probability 1 and for n large enough, the indices $C^{(1)}$ selected in the first step (6.28) are exactly $S^{\star c}$ so that $J^{(1)} \notin S^\star$ and $S^\star \subset S^{(1)}$.

Then, applying again theorem 6.4.1 gives that, with n possibly larger, $J^{(2)} \notin S^\star$. Once that process has been iterated enough times to eliminate all indices out of $S^\star$, theorem 6.4.1 once again ensures that the following choice of $C^{(k)}$ will lead to the empty set. This completes the proof. □

6.5 Simulations

In this section, we present simulation results illustrating theorems 6.4.1, 6.4.2, 6.4.3, 6.4.4 regarding convergence of the chosen supports $\widehat{S}$ selected by our alternative methods toward the true support $S^\star$.

6.5.1 Setting

The unknown function we consider is

$$\begin{aligned} f^\star : [-\pi, \pi] &\rightarrow \mathbb{R} \\ x &\mapsto f^\star(x) = -x + \cos(2x) - \sin(2x). \end{aligned} \tag{6.41}$$

The distribution of the abscisses is given by the density

$$w = \frac{1}{2\pi}\mathbb{1}_{[-\pi,\pi]} \tag{6.42}$$

and the model considered is a 6-dimensional space $F = \mathrm{Vect}(f_1, \ldots, f_6)$ where

$$\begin{array}{lll} f_1(x) = x, & f_3(x) = \cos(x), & f_5(x) = \sin(x), \\ f_2(x) = x^2, & f_4(x) = \cos(2x), & f_6(x) = \sin(2x). \end{array} \tag{6.43}$$

Note that $f^\star \in F$.

6.5.2 Results

We generate 100 set of observations of the couple (X, Y) linked by the relation $Y = f^\star(X) + G$. On each of those observations we apply the φ_β criterion (2.10), for β ranging from 0 to 1 by step 0.05, along with the following methods :

- ⋆ Global method (6.24).
- ⋆ Comparative method (6.25)
- ⋆ Reversed comparative method (6.26)
- ⋆ Adapted reversed comparative method (6.27), shortened to "adapted"
- ⋆ Descending method (part 6.3.5).

Recall that the first one has exponential complexity while the others have polynomial complexities which make them much faster to use. In our setting, the global method took 270 seconds to provide results while any of the 4 comparative methods needed about 13 seconds.

We count a success if the selected support is exactly the one of $f^\star$ in the setting given by (6.41), (6.42) and (6.43) ; that is $S^\star = \{1, 4, 6\}$. Note that, since the basis of our setting is not orthonormal relatively to (6.1), the reversed comparative method should not give good results as seen in part 6.4.2. Actually, this method always give a percentage succes of 0 in all our simulations, this is why it does not appear in our results.

Figure 6.1 presents the percentage of succes of the different methods plugged against the value of β in (2.10) for $n = 20, 50, 200$ and 1000. The four vertical lines correspond, from the left to the right, to the values β_{AIC}, β_{BIC}, β_{min} and β_{max} given in (2.11) and (2.12). Note that most of the time, descending and global methods give the same percentage.

6.5.3 Comments

Firstly let us say that when β is too low, the penalization of the criterion is too weak and overparametrization occurs : $\widehat{S}$ contains too many functions, thus the failure. By opposition, when β is too large, underparametrization occurs and $\widehat{S}$ does not contain $S^\star$.

Now, as n grows, we observe an increasing rate of succes for any fixed value of β as convergence theorems of the previous section announced.

However, the AIC criterion (corresponding to the first vertical line) does not fullfill those theorems requirements and thus present a quite low percentage of succes. The AIC criterion is known for its lack of penalization yielding overparametrization ; this is what we observe here.

The BIC criterion (corresponding to the second vertical line) does not give here the fastest increasing rate of success. It seems it also lacks a little more penalization to reach the 100% of success sooner. Our previous use of the φ_β criterion in other model selection problems (such as autoregression order determination) also resulted in the same conclusion regarding the BIC criterion.

As announced when the adapted reversed comparative method (6.27) was introduced (part 6.3.5), it requires a bit more penalization than regular comparative method in order to avoid overparametrization. This fact appears on figure 6.1.

Finally let us stress that the regression problem as studied here is the one, to our knowledge, that allows the biggest penalization before underparametrization occurs. Indeed, in the others model selection problems we studied with the φ_β criterion, we obtained results similar to figure 6.1 except that the success rate of any method fell back to 0 for smaller value of β regarding the value of n. See for instance figures 3.5, 3.6 or 3.7 in the autoregression context.

6.5.4 A brief word on future applications

Choosing $I = [-\pi, \pi]$, w the uniform density on I, $f^\star(x) = x$ and a basis consisting of $\cos(ax)$, $\sin(bx)$ where $a, b \in [\![0, 5]\!]$, we observed that comparative methods select supports containing only sinus functions and produced the following estimation of $f^\star$:

$$2.02\sin(x) - 1.02\sin(2x) + 0.71\sin(3x) - 0.47\sin(4x) + 0.46\sin(5x)$$

whereas the Fourier series of $f^\star$ starts by

$$2\sin(x) - \sin(2x) + \frac{2}{3}\sin(3x) - \frac{1}{2}\sin(4x) + \frac{2}{5}\sin(5x) + ...$$

This observations enlights the fact that linear regression helps finding the most important harmonics contained in a noised signal. We currently work on finding those harmonics on real signals, e.g. heartbeat signals measured before or after physical efforts. The same work might also be done with wavelets basis.

6.6 Study of the risks

6.6.1 Expression of the risks

The aim of this part is to show non-asymptotic differences between our setting and the fixed points case in terms of risk for our estimators. Because of the remark following (6.12), we are constrained here to make the following assumption :

$$\mathbb{E}\left[M_S^{-1}\right] < +\infty. \tag{6.44}$$

Note that, from (6.12), this assumption also implies that $\mathbb{E}\left[\left\|\widehat{f}_S\right\|_w^2\right] < +\infty$ and thus gives sense to following computations about the risk of $\widehat{f}_S$ as an estimator of $f^\star$. Here, we are not interested in conditions ensuring (6.44) but in their consequences on expressions of the bias and risk of $\widehat{f}_S$.

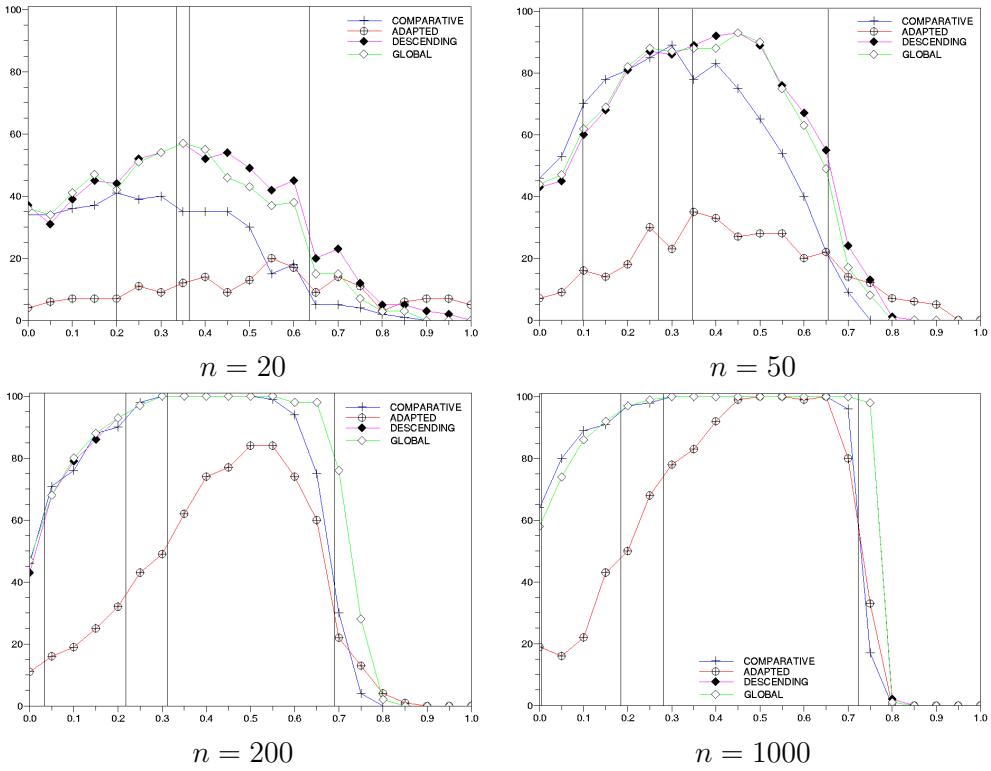


FIG. 6.1 – Percentage of succes of the different methods against the value of β in the φ_β criterion (2.10). Vertical lines correspond respectively to β_{AIC}, β_{BIC}, β_{min} and β_{max}, equations (2.11) and (2.12).

We work here with a fixed support S that belongs to either of the families $\mathcal{F}_1$ and $\mathcal{F}_2$ defined by

$$\mathcal{F}_1 = \{S \subset [\![1,d]\!] \,|\, S^\star \not\subseteq S\} \text{ and } \mathcal{F}_2 = \{S \subset [\![1,d]\!] \,|\, S^\star \subseteq S\}\,. \tag{6.45}$$

In the case where $f^\star$ does not live in F, we simply set $\mathcal{F}_2 = \emptyset$ and $\mathcal{F}_1 = \mathcal{P}([\![1,d]\!])$.

Recall briefly that, in the fixed points setting studied for instance in [Bar00], computations of losses are done directly in the space of observations $\mathbb{R}^n$ endowed with the normalized canonical norm $n^{-1}\|.\|_n$ that is chosen to satisfy $n^{-1}\|t\|_n^2 = n^{-1}\sum_i t(x_i)^2$ for all $t \in F$. We obtain :

$$\begin{aligned} \mathbb{E}\widehat{f}_S &= \Pi_S^F f^\star \\ \mathbb{E}\left[\|\widehat{f}_S - f^\star\|_n^2\right] &= \|f^\star - \Pi_S^F f^\star\|_n^2 + \sigma^2|S|/n. \end{aligned} \tag{6.46}$$

The case $S \in \mathcal{F}_2$

Remark that the multiplication $D_S D_S^T f$ sets to 0 the components of f along each $f_j,\ j \notin S$. Since $f^\star$ has support $S^\star$ we get $D_S D_S^T f^\star = f^\star$ and formula (6.12) gives

$$\widehat{f}_S = D_S M_S^{-1} V_S^T (V D_S D_S^T f^\star + g) = f^\star + D_S M_S^{-1} V_S^T g. \tag{6.47}$$

By independence between X and G we obtain $\mathbb{E}\widehat{f}_S = f^\star = \Pi_S^F f^\star$, that is $\widehat{f}_S$ is an unbiased estimator of $\Pi_S^F f^\star$. In order to compute the risk let us remark that

$$\left\|\widehat{f}_S - f^\star\right\|_w^2 = g^T V_S M_S^{-1} D_S^T M_w D_S M_S^{-1} V_S^T g =: g^T A_S\, g. \tag{6.48}$$

One of the main difference between the deterministic points case and our case appears in that matrix A_S. Indeed, in the former case, the computation of losses is done in the space of observations $\mathbb{R}^n$ rather than in F. In other words, the matrix $M_{w,S}$ in A_S is replaced by M_S which reduces A_S to the matrix of Π_S^E as in (6.11) and gives an exact formula for the variance. In our settings we are not able to derive such a formula and may only write :

$$\begin{aligned} \mathbb{E}\widehat{f}_S &= f^\star = \Pi_S^F f^\star \\ \mathbb{E}\left[\left\|\widehat{f}_S - f^\star\right\|_w^2\right] &= \sigma^2 \mathbb{E}\left[\mathrm{Tr}(A_S)\right] = \sigma^2 \mathbb{E}\left[\mathrm{Tr}(M_{w,S} M_S^{-1})\right]. \end{aligned} \tag{6.49}$$

In this case $S \in \mathcal{F}_2$, the result (6.49) is quite close to (6.46).

The case $S \in \mathcal{F}_1$

Here we may only write $y = V f^\star + g$ and (6.12) gives

$$\mathbb{E}\widehat{f}_S = \mathbb{E}\left[D_S M_S^{-1} D_S^T M f^\star\right] =: \mathbb{E}\left[B_S f^\star\right].$$

Note that $B_S \to D_S M_{w,S}^{-1} D_S^T M_w$ a.s., that latter being the matrix of the orthogonal projector Π_S^F. However $\mathbb{E}B_S \neq D_S M_{w,S}^{-1} D_S^T M_w$. Consequently, as an estimator of $\Pi_S^F f^\star$, $\widehat{f}_S$ has a bias.

Now for the risk we write

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{f}_S - f^\star\right\|_w^2\right] &= \mathbb{E}\left[\left\|f^\star - \Pi_S^F f^\star + \Pi_S^F f^\star - \mathbb{E}\widehat{f}_S + \mathbb{E}\widehat{f}_S - \widehat{f}_S\right\|_w^2\right] \\ &= \left\|f^\star - \Pi_S^F f^\star\right\|_w^2 + \left\|\Pi_S^F f^\star - \mathbb{E}\widehat{f}_S\right\|_w^2 \\ &\quad + \mathbb{E}\left[\left\|\mathbb{E}\widehat{f}_S - \widehat{f}_S\right\|_w^2\right],\end{aligned}$$

moreover

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{f}_S - \mathbb{E}\widehat{f}_S\right\|_w^2\right] &= \mathbb{E}\left[\left\|B_S f^\star + D_S M_S^{-1} V_S^T g - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2\right] \\ &= \mathbb{E}\left[\left\|B_S f^\star - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2\right] + \mathbb{E}\left[\left\|D_S M_S^{-1} V_S^T g\right\|_w^2\right],\end{aligned}$$

the expectation of the scalar product vanishing since X and G are independent. Finally, non asymptotically :

$$\begin{aligned}\mathbb{E}\widehat{f}_S &= \mathbb{E}\left[B_S f^\star\right] \\ \mathbb{E}\left[\left\|\widehat{f}_S - f^\star\right\|_w^2\right] &= \left\|f^\star - \Pi_S^F f^\star\right\|_w^2 + \sigma^2 \mathbb{E}\left[\mathrm{Tr}(M_{w,S} M_S^{-1})\right] \\ &\quad + \left\|\Pi_S^F f^\star - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2 + \mathbb{E}\left[\left\|B_S f^\star - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2\right] \\ &= \left\|f^\star - \Pi_S^F f^\star\right\|_w^2 + \sigma^2 \mathbb{E}\left[\mathrm{Tr}(M_{w,S} M_S^{-1})\right] \\ &\quad + \mathbb{E}\left[\left\|B_S f^\star - \Pi_S^F f^\star\right\|_w^2\right]\end{aligned} \tag{6.50}$$

Comparing this result to (6.46), we get new bias and variance terms. Those are created by the randomness on the X_i's since, in the fixed setting, the expression $B_S f^\star - \Pi_S^F f^\star$ vanishes.

6.6.2 Asymptotics of the risks

We no longer suppose (6.44). Our aim is now to derive asymptotic results similar to those given in [Nis84], except we handle the random points case. More precisely, we prove that assumptions of theorems 6.4.1, 6.4.2, 6.4.3, 6.4.4 also ensure an asymptotic risk equivalent to an oracle risk. Recall that the remark following (6.12) prevents us from computing risks in a general case. We handle this issue by using the truncated estimator $\widetilde{f}_S$ defined in (6.13).

The ideal case

Assume for a moment that the user knows the support $S^\star$. Then he will estimate $f^\star$ by $\widetilde{f}_{S^\star}$ and get an oracle risk $\mathcal{O}R(n, S^\star) = \mathbb{E}\left[\left\|\widetilde{f}_{S^\star} - f^\star\right\|_w^2\right]$.

Note that the event $\{\|M_{S^\star}^{-1}\| < C\}$ appearing in (6.13) is independent of the noise g. Therefore, following computations similar to part 6.6.1 we get

$$\mathcal{O}R(n, S^\star) = \|f^\star\|_w^2 \, \mathbb{E}\left[\mathbb{1}_{\{\|nM_{S^\star}^{-1}\| < C\}^c}\right] + \sigma^2 \mathbb{E}\left[\mathbb{1}_{\{\|nM_{S^\star}^{-1}\| < C\}} \mathrm{Tr}(M_{w,S^\star} M_{S^\star}^{-1})\right].$$

The first term is handled by (6.17). For the second, the indicator function gives a dominated convergence allowing to write

$$\mathcal{O}R(n,S^\star) \sim \frac{\sigma^2|S^\star|}{n} \tag{6.51}$$

Risk of our procedures

We assume now that the support $\widehat{S}$ has been selected by an IC (6.23) whose penalty satisfies

(i) $\alpha(n) = o(n)$

(ii) $\ln\ln n = o(\alpha(n))$,

and using either of the following method :

- ⋆ comparative method (6.25)
- ⋆ Reversed comparative method in the orthonormal case (6.26)
- ⋆ Adapted reversed comparative method (6.27)
- ⋆ Descending comparative method (part 6.3.5)

Recall the notations of part 6.6.1. Theorems of the previous section as well as control (6.31), easily transposable to other selection procedures listed above, ensure that in any of those cases we get

$$\begin{cases} \widehat{S} \to S^\star & \text{a.s.} \\ \mathbb{P}(\widehat{S}=S) \to 0 & \text{for any} \quad S \in \mathcal{F}_2 \setminus \{S^\star\} \\ \mathbb{P}(\widehat{S}=S) = O\left(\exp(-cn)\right) & \text{for any} \quad S \in \mathcal{F}_1 \end{cases} \tag{6.52}$$

where c is a positive constant.

Our estimation procedure of $f^\star$ by $\widetilde{f} = \widetilde{f}_{\widehat{S}}$ has a risk $R(n)$ given by

$$R(n) = \mathbb{E}\left[\left\|f^\star - \widetilde{f}\right\|_w^2\right] = \sum_{S\subset[\![1,d]\!]} \mathbb{E}\left[\left\|f^\star - \widetilde{f}_S\right\|_w^2 \mathbb{1}_{\widehat{S}=S}\right] =: \sum_{S\subset[\![1,d]\!]} R(n,S). \tag{6.53}$$

Now distinguish between two cases.

Asymptotics of the case $S \in \mathcal{F}_2$

Computations similar to (6.47) and (6.48) where $\widehat{f}_S$ is replaced by $\widetilde{f}_S$ reduce $R(n,S)$ to

$$\begin{aligned} R(n,S) &= \|f^\star\|_w^2\, \mathbb{E}\left[\mathbb{1}_{\{\|nM_S^{-1}\|<C\}^c \cap \{\widehat{S}=S\}}\right] \\ &\quad + \mathbb{E}\left[g^T A_S g\, \mathbb{1}_{\{\|nM_S^{-1}\|<C\}\cap\{\widehat{S}=S\}}\right]. \end{aligned}$$

Remark that

$$\begin{aligned} Z_n &:= ng^T A_S g\, \mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \\ &= \frac{1}{n}\sum_{i,j=1}^n g_i g_j \left(V_S^T nM_S^{-1} M_{w,S} nM_S^{-1} V_S\right)_{i,j} \mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \end{aligned} \tag{6.54}$$

has an $L^2(\Omega)$ norm satisfying $\mathbb{E}\, Z_n^2 = O(1)$.

Let us begin by $S = S^\star$. Then (6.52) yields $\lim \mathbb{E}\left[Z_n \mathbb{1}_{\{\widehat{S}=S^\star\}}\right] = \lim \mathbb{E} Z_n$ which is $\sigma^2|S^\star|$ by dominated convergence. This, with (6.17), gives

$$R(n, S^\star) \sim \frac{\sigma^2|S^\star|}{n}. \tag{6.55}$$

Now for $S \in \mathcal{F}_2 \setminus \{S^\star\}$ we use Schwarz's inequality to write

$$\mathbb{E}\left[Z_n \mathbb{1}_{\{\widehat{S}=S\}}\right] \leq \left(\mathbb{E}\, Z_n^2\, \mathbb{P}(\widehat{S} = S)\right)^{1/2} \to 0$$

because of (6.54) and (6.52). Consequently,

$$nR(n, S) \to 0, \text{ for any } S \in \mathcal{F}_2 \setminus \{S^\star\}. \tag{6.56}$$

Asymptotics of the case $S \in \mathcal{F}_1$

Here we simply write

$$\begin{aligned} R(n, S) &= \mathbb{E}\left[\left\|f^\star - \widetilde{f}_S\right\|_w^2 \mathbb{1}_{\widehat{S}=S}\right] \\ &\leq 2\left\|f^\star\right\|_w^2 \mathbb{P}(\widehat{S} = S) + 2\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right]^{1/p} \mathbb{P}(\widehat{S} = S)^{1/q} \end{aligned}$$

where p, q are chosen as in (6.15). Recall (6.14) and (6.52) in our present case $S \in \mathcal{F}_1$ to obtain

$$nR(n, S) \to 0, \text{ for any } S \in \mathcal{F}_1. \tag{6.57}$$

Summary

Plugging (6.53), (6.55), (6.56) and (6.57) together we get the following

Theorem 6.6.1 *Assume that the penalty of our IC (6.23) satisfies*

(*i*) $\alpha(n) = o(n)$

(*ii*) $\ln \ln n = o(\alpha(n))$,

and that either of the following method has been used to determine $\widehat{S}$.

- ⋆ *comparative method (6.25)*
- ⋆ *Reversed comparative method in the orthonormal case (6.26)*
- ⋆ *Adapted reversed comparative method (6.27)*
- ⋆ *Descending comparative method (part 6.3.5)*

Then the estimation of $f^\star$ by $\widetilde{f} = \widetilde{f}_{\widehat{S}}$ defined in (6.13) presents a risk $R(n)$ equivalent to the oracle risk $\mathcal{O}R(n, S^\star)$ (6.51) in the sense that

$$R(n) \sim \frac{\sigma^2|S^\star|}{n}$$

6.7 An oracle inequality for the risk of the comparative descending method

In this section, we no longer assume that $f^\star$ lives in F. Our main purpose is to give theorem 6.7.1. This theorem presents an oracle inequality on the risk achieved by the estimator of $f^\star$ resulting from the use of an information criterion of the form (6.23) along with the (fast) descending comparative method (part 6.3.5).

6.7.1 Preliminary result

For S a support, we define the "risk"

$$R^\star(S) = d^2(f^\star, F_S) + \sigma^2|S|/n. \tag{6.58}$$

This quantity $R^\star(S)$ does not represent the risk resulting from the estimation of $f^\star$ within F_S which are expressed in (6.49) and (6.50). Actually, $R^\star(S)$ is the risk resulting from such an estimation in the case where the x_i's are deterministic as in (6.46). In the sequel, as Baraud in [Bar02], we still work with quantities $R^\star(S)$.

Let $\mathcal{F}$ be a family of supports. We associate with it an oracle risk by

$$\mathcal{O}_{\mathcal{F}}(f^\star) = \min_{S\in\mathcal{F}} \{R^\star(S)\} \tag{6.59}$$

The oracle (6.59) is the minimum risk the user could achieve by selecting the support $S_{\mathcal{O}}$ in $\mathcal{F}$ that realizes the minimum. However, as the name oracle implies, that quantity as well as $S_{\mathcal{O}}$ is unavalaible to the user. Also note that, even though $f^\star$ would live in F, there is no reason why $S_{\mathcal{O}}$ would equal $S^\star$.

From now on, we choose a penalty of the form

$$\alpha(n) = (1+\theta)\sigma^2|S|,\ \theta > 0. \tag{6.60}$$

Let us assume briefly that this user has chosen the global method (6.24). The only thing he knows is what his criterion has found is the best support, namely

$$\widehat{S} = \text{Argmin}\left\{\text{IC}(S),\ S \in \mathcal{F}\right\}.$$

Baraud shows in [Bar02] that in our setting and by using the penalty (6.60) in his criterion (6.23), the user did not take too much risks in the sense that

$$\mathbb{E}\left[\left\|f^\star - \widetilde{f}_{\widehat{S}}\right\|_w^2\right] \leq C\mathcal{O}_{\mathcal{F}}(f^\star). \tag{6.61}$$

where C is a constant depending on θ appearing in the penalty (6.60) but neither on n nor on $f^\star$.

Now, as stressed in part 6.3.6, the global method has exponential complexity. Here, if we wanted to be able to select any support, we would have computed 2^d criteria. Our aim in the sequel is to show that the descending method, that has polynomial complexity, also gives an oracle inequality of the type (6.61).

6.7.2 A family of nested deterministic supports

We define here a sequence of decreasing unknown supports $S^{\star(k)}$, $k = 0, \ldots, d$ all with cardinality $d - k$. Firstly we set $S^{\star(0)} = [\![1, d]\!]$ then, when $S^{\star(k)}$ is defined, we set

$$S^{\star(k+1)} = \text{Argmin}\left(R^\star(S),\ S \subset S^{\star(k)},\ |S| = d - (k+1)\right) \tag{6.62}$$

where the function $R^\star$ is defined in (6.58).

We thus obtain a sequence of risks $R^\star(S^{\star(k)})$, $k = 0, \ldots, d$. Each of those represents the minimum risk achieved by removing a single function in the previous support.

In the sequel, we are constrained to make the following assumption :

$$R^\star(S^{\star(k)}) \neq R^\star(S^{\star(k+1)}),\ k = 0, \ldots, d-1. \tag{6.63}$$

This holds most of the time regarding the expression of R^* (6.58). Indeed, it is rather unlikely that the potential increase of the first term in R^* is exactly compensated by the decrease σ^2/n of the second. However, if that happens, it suffices to add one or several points to the sample to fix the problem. Assumption (6.63) ensures that the first index $1 \leq k^\star \leq d-1$ such that

$$R^\star(S^{\star(k^\star-1)}) > R^\star(S^{\star(k^\star)}) \text{ and } R^\star(S^{\star(k^\star+1)}) > R^\star(S^{\star(k^\star)}) \tag{6.64}$$

is correctly defined. In the case where the sequence $R^\star(S^{\star(k)})$, $k = 0, \ldots, d$ is always decreasing, we set $k^\star = d$. In the case where $R^\star(S^{\star(1)}) > R^\star(S^{\star(0)})$, we set $k^\star = 0$.

This way, $S^{\star(k^\star)}$ is the first support that does not immediatly include a support achieving a smaller risk. The quantity $R^\star(S^{\star(k^\star)})$ is an oracle risk, not among any risks possible as in (6.59) with $\mathcal{F} = \mathcal{P}([\![1, d]\!])$, but among a smaller, nested, family of risks.

This deterministic family of supports $S^{\star(k)}$ (6.62) is related to the random family $S^{(k)}$ produced by the descending comparative method in part 6.3.5. Recall that this method stops at a random step k_f and thus produces only supports $S^{(k)}$, $k = 0, \ldots, k_f$. Ideally, one would like the method to choose good supports and stop at the right step in the sense that

$$S^{(0)} = S^{\star(0)},\ S^{(1)} = S^{\star(1)}, \ldots, S^{(k^\star)} = S^{\star(k^\star)}, \text{ and } k_f = k^\star.$$

Even though the forecoming theorem 6.7.1 deals with an oracle inequality, its proof also shows that this happens except on a event the probability of which decreases exponentially fast with n.

6.7.3 The oracle inequality

Let us give the main result.

Theorem 6.7.1 *Consider an information criterion of the form (6.23) whose penalty term writes as*

$$\alpha(n) = (1 + \theta)\sigma^2|S|$$

with $\theta > 0$. Using this criterion along with the descending comparative method described in part 6.3.5, one produces $\widetilde{f}_{S^{(k_f)}}$ as an estimation of the unknown function $f^\star$. The risk of such an estimation satisfies

$$\mathbb{E}\left[\left\|f^\star - \widetilde{f}_{S^{(k_f)}}\right\|_w^2\right] \leq C.R^\star(S^{\star(k^\star)}) + r_n \tag{6.65}$$

where C is a constant depending on θ but neither on n nor $f^\star$, $R^\star(S^{\star(k^\star)})$ is the nested oracle risk defined in (6.64) and r_n is a deterministic term satisfying $r_n = O(\exp(-an))$ with $a > 0$.

Proof : It is split into two parts.

$\star$ *First part* : some probability controls. Our aim here is to show controls (6.72) that will be of use in the second part. Let us define the decreasing sequence of events $(A_k)_{k=1,\dots,d}$ by

$$A_k = \left\{S^{(k-1)} = S^{\star(k-1)}, \dots, S^{(0)} = S^{\star(0)}\right\}. \tag{6.66}$$

Note that A_k is implicitely included in the event $\{k_f \geq k-1\}$.

Let us choose $1 \leq k \leq k^\star$. We are interested in the probability that the descending comparative method selects good supports up to step $k-1$ and fails to select $S^{\star(k)}$ at step k.

$$\mathbb{P}\left(S^{(k)} \neq S^{\star(k)}\,,\, A_k\right).$$

Write

$$\begin{aligned}
\mathbb{P}\left(S^{(k)} \neq S^{\star(k)}\,,\, A_k\right) &\leq \sum_S \mathbb{P}\left(\frac{1}{n}\mathrm{IC}(S) - \sigma^2 \leq \frac{1}{n}\mathrm{IC}(S^{\star(k)}) - \sigma^2\,,\, A_k\right) \\
&\leq \sum_S \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S) - \sigma^2 - R^\star(S)\right| > \varepsilon_k\,,\, A_k\right) \\
&\quad + \sum_S \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S^{\star(k)}) - \sigma^2 - R^\star(S^{\star(k)})\right| > \varepsilon_k\,,\, A_k\right)
\end{aligned} \tag{6.67}$$

where the sums are extended to all supports $S \subset S^{\star(k-1)}$ with cardinal $|S| = d-k$ except $S^{\star(k)}$ and

$$\varepsilon_k = \frac{1}{2}\min_S(R^\star(S) - R^\star(S^{\star(k)})) > 0,$$

the minimun being taken among the same set of supports. Let us denote by $\varepsilon > 0$ the smallest of those ε_k's, $k = 1, \dots, k^\star$. Because of the form of the penalty term (6.60), expressions appearing in the probabilities of equation (6.67) simplify to

$$\frac{1}{n}\mathrm{IC}(S) - \sigma^2 - R^\star(S) = \frac{1}{n}d^2(y, E_S) - \sigma^2 - d^2(f^\star, F_S) + \frac{\theta\sigma^2|S|}{n}.$$

It remains to choose n large enough to make $\theta\sigma^2|S|/n < \varepsilon/4$ and apply control (6.21) to have

$$\mathbb{P}\left(S^{(k)} \neq S^{\star(k)}\,,\, A_k\right) = O\left(\exp\left(-\frac{\varepsilon^2 n}{64 V_{max}}\right)\right),\ k = 1, \dots, k^\star. \tag{6.68}$$

Now let us stress that, conditionnaly to the distribution X of the abscisses, the information criterion (6.23) has a distribution of the type non-central χ^2 ; the degrees

of freedom as well as the mean parameter being related to which support S the IC is calculated on. Now, an event of the form A_k (6.66) may be expressed in term of some IC(S) constrained to belong to some, random but non-empty, intervals of $\mathbb{R}$. Therefore, via the positivity of the densities of the joint laws of those IC, we get the positivity of $\mathbb{P}(A_k)$ for any k. Moreover, iterating control (6.68) also shows that $\mathbb{P}(A_k^c)$ behaves as $O(\exp(-an))$ for some $a > 0$ and $k = 1, \ldots, k^\star$. Finally we get the existence of a positive constant c satisfying for all n :

$$\mathbb{P}(A_k) > c > 0, \ k = 1, \ldots, k^\star. \tag{6.69}$$

We are now interested in

$$\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right).$$

This is the probability that the method goes well up to step $k^\star - 1$ but stops here. Write :

$$\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right) = \mathbb{P}\left(\bigcap_S \left(\frac{1}{n}\mathrm{IC}(S) - \sigma^2 \geq \frac{1}{n}\mathrm{IC}(S^{\star(k^\star-1)}) - \sigma^2\right)\,,\, A_{k^\star}\right)$$

where the intersection is extended to all supports S of cardinal $d - k^\star$ included in $S^{\star(k^\star-1)}$. In particular, choosing $S = S^{\star(k^\star)}$:

$$\begin{aligned}
\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right) &\leq \mathbb{P}\left(\frac{1}{n}\mathrm{IC}(S^{\star(k^\star)}) - \sigma^2 \geq \frac{1}{n}\mathrm{IC}(S^{\star(k^\star-1)}) - \sigma^2\,,\, A_{k^\star}\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S^{\star(k^\star-1)}) - \sigma^2 - R^\star(S^{\star(k^\star-1)})\right| > \varepsilon\,,\, A_{k^\star}\right) \\
&\quad + \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S^{\star(k^\star)}) - \sigma^2 - R^\star(S^{\star(k^\star)})\right| > \varepsilon\,,\, A_{k^\star}\right)
\end{aligned}$$

where

$$\varepsilon = \frac{1}{2}\left(R^\star(S^{\star(k^\star-1)}) - R^\star(S^{\star(k^\star)})\right) > 0.$$

Again because of the form (6.60) of the penalty, expressions in those probabilities reduce to

$$\frac{1}{n}\mathrm{IC}(T) - \sigma^2 - R^\star(T) = \frac{1}{n}d^2(y, E_T) - \sigma^2 - d^2(f^\star, F_T) + \frac{\theta\sigma^2|T|}{n}.$$

Choosing n large enough to make $\theta\sigma^2|T|/n < \varepsilon/4$ and applying control (6.21) yields

$$\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right) = O\left(\exp\left(-\frac{\varepsilon^2 n}{64 V_{max}}\right)\right). \tag{6.70}$$

The last probability we need to control is the following :

$$\mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star}\right).$$

Let us write

$$\mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star}\right) \leq \mathbb{P}\left(k_f > k^\star\,,\, S^{(k^\star)} \neq S^{\star(k^\star)}\,,\, A_{k^\star}\right) + \mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star+1}\right).$$

The first probability has already been dealt with when we obtained control (6.68). For the second one, it is handled by arguments similar to the ones we developped to justify (6.70). Consequently we get, for some $a > 0$:

$$\mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star}\right) = O\left(\exp(-an)\right). \tag{6.71}$$

We will conclude this serie of probability controls by putting together (6.68), (6.70), (6.71) with (6.69) to write, for some $a > 0$:

$$\begin{aligned} \mathbb{P}\left(S^{(k)} \neq S^{\star(k)} \,|\, A_k\right) &= O\left(\exp(-an)\right),\ k = 1, \ldots, k^\star, \\ \mathbb{P}\left(k_f = k^\star - 1 \,|\, A_{k^\star}\right) &= O\left(\exp(-an)\right), \\ \mathbb{P}\left(k_f > k^\star \,|\, A_{k^\star}\right) &= O\left(\exp(-an)\right). \end{aligned} \tag{6.72}$$

$\star\star$ *Second part* : the oracle inequality. It is time to apply Baraud's result. Recall that the descending comparative method produces $\widehat{f}_{S^{(k_f)}}$ as an estimation of $f^\star$. The loss is measured by $\left\|f^\star - \widetilde{f}_{S^{(k_f)}}\right\|_w^2$ which we shorten to d_w^2.

We condition by the event $\{k_f = k^\star\} \cap A_{k^\star}$ which means that the method has chosen good supports up to step $k^\star - 1$ and will stop at the good step $k^\star$. Moreover, the comparative descending method (part 6.3.5) used here ensures that we are going to choose a support by minimization of our criterion among the family of supports of cardinal $d - k^\star$ included in $S^{\star(k^\star-1)}$. The oracle risk associated with this family as in (6.59) is precisely $R^\star(S^{\star(k^\star)})$. Consequently, the result of Baraud (6.61), more precisely equation (15) following theorem 1.1 in [Bar02], ensures that

$$\mathbb{E}\left[d_w^2 \middle|\, k_f = k^\star,\, A_{k^\star}\right] \leq C.R^\star(S^{\star(k^\star)}),$$

where C depends on θ but neither on n nor on $f^\star$. We need to remove the conditionning in order to obtain the oracle inequality (6.65) :

$$\mathbb{E}\left[d_w^2 \middle|\, A_{k^\star}\right] \leq \mathbb{E}\left[d_w^2 \middle|\, k_f = k^\star,\, A_{k^\star}\right] + \mathbb{E}\left[d_w^2 \mathbb{1}_{k_f = k^\star - 1} \middle|\, A_{k^\star}\right] + \mathbb{E}\left[d_w^2 \mathbb{1}_{k_f > k^\star} \middle|\, A_{k^\star}\right]$$

handles the event $k_f = k^\star$. Moreover :

$$\mathbb{E}\left[d_w^2 \middle|\, A_{k^\star - 1}\right] \leq \mathbb{E}\left[d_w^2 \middle|\, A_{k^\star}\right] + \mathbb{E}\left[d_w^2 \mathbb{1}_{S^{(k^\star)} \neq S^{\star(k^\star)}} \middle|\, A_{k^\star - 1}\right]$$

handles the event $S^{(k^\star-1)} = S^{\star(k^\star-1)}$ in $A_{k^\star}$. Iterating that latter argument we get

$$\begin{aligned} \mathbb{E}\left[d_w^2 | A_1\right] = \mathbb{E}\left[d_w^2\right] \leq\ & C.R^\star(S^{\star(k^\star)}) + \\ & \mathbb{E}\left[d_w^2 \mathbb{1}_{k_f = k^\star - 1} \middle|\, A_{k^\star}\right] + \mathbb{E}\left[d_w^2 \mathbb{1}_{k_f > k^\star} \middle|\, A_{k^\star}\right] + \\ & \sum_{k=1}^{k^\star - 1} \mathbb{E}\left[d_w^2 \mathbb{1}_{S^{(k)} \neq S^{\star(k)}} \middle|\, A_k\right]. \end{aligned}$$

It suffices to apply Hölder's inequality (6.15) along with controls (6.72) to conclude the proof.

□

Appendices

Annexe A

Un algorithme de programmation dynamique

L'agorithme de programmation dynamique que nous présentons dans cette annexe permet de trouver la partition $\widehat{P}$ qui réalise (5.7) sans avoir à calculer la valeur de CRIT(P) sur toutes les sous-partitions de $P_{\max}$.

A.1 Notations et propriétés

Rappelons que nous disposons de données $x = x_1, \ldots, x_n$ vivant dans un intervalle $[a, b]$ subdivisés en une partition maximale $P_{\max}$ régulière à R intervalles de pas r :

$$P_{\max} : a < a + r < a + 2r < \cdots < a + (R-1)r < a + Rr = b.$$

Cette partition présente 2^{R-1} sous-partitions parmi lesquelles nous souhaitons minimiser la quantité CRIT(x, P) (5.6) :

$$\text{CRIT}(x, P) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n|I_j|} + (m-1)\frac{\log n}{2} - n \log r.$$

où $m \leq R$, $P = (I_j)_{j=1,\ldots,m}$ et n_j est le nombre de données tombant dans l'intervalle I_j de longueur $|I_j| = \sup I_j - \inf I_j$.

Pour tout $\tau \in [\![1, R]\!]$, définissons

$$I(\tau) = [a, a + \tau r],\ x(\tau) = \{x_i \text{ tels que } x_i \in I_\tau\},\ n(\tau) = \text{Card}(x(\tau)).$$

En d'autres termes, $x(\tau)$ représente les données vivant dans l'intervalle tronqué $I(\tau)$ et $n(\tau)$ représente leur nombre. Nous appelons également $P_{\max}(\tau)$ la partition maximale tronquée à ce même intervalle :

$$P_{\max}(\tau) : a < a + r < a + 2r < \cdots < a + (\tau - 1)r < a + \tau r.$$

Pour $m \leq \tau$ et $P(\tau) = (I_j)_{j=1,\ldots,m}$ une sous-partition de $P_{\max}(\tau)$, on définit

$$\text{CRIT}(x(\tau), P(\tau)) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n(\tau)|I_j|} + (m-1)\frac{\log n(\tau)}{2} - n(\tau) \log r.$$

C'est la valeur que prendrait le critère si on oubliait les données hors de l'intervalle $I(\tau)$. Notons que les n_j sont indépendants de la troncature de l'intervalle.

Soit maintenant $\tau \in [\![1, R-1]\!]$ fixé, $P(\tau)$ une sous-partition de $P_{\max}(\tau)$ à $m \leq \tau$ intervalles et $\sigma \in [\![\tau, R]\!]$. On note $P(\tau, \sigma)$ la partition à $m+1$ intervalles de $I(\sigma)$ constituée de $P(\tau)$ et de l'intervalle $I_{m+1} = [a + \tau r, a + \sigma r]$. On a alors

$$\begin{aligned} \mathrm{CRIT}(x(\tau), P(\tau)) &= -\sum_{j=1}^{m} n_j \log \frac{n_j}{n(\tau)|I_j|} + (m-1)\frac{\log n(\tau)}{2} - n(\tau)\log r \\ \mathrm{CRIT}(x(\sigma), P(\tau,\sigma)) &= -\sum_{j=1}^{m+1} n_j \log \frac{n_j}{n(\sigma)|I_j|} + m\frac{\log n(\sigma)}{2} - n(\sigma)\log r. \end{aligned}$$

et la différence entre ces deux critères s'évalue par

$$\begin{aligned} \mathrm{CRIT}(x(\sigma), P(\tau,\sigma)) - \mathrm{CRIT}(x(\tau), P(\tau)) &= (n(\sigma) - n(\tau)) \log \frac{n(\sigma) - n(\tau)}{n(\sigma)(\sigma - \tau)} \\ &\quad -n(\tau)\log\frac{n(\tau)}{n(\sigma)} \\ &\quad +(m-1)\frac{\log n(\tau)}{2} - m\frac{\log n(\sigma)}{2}. \end{aligned} \tag{A.1}$$

Cette quantité est l'évolution que subit le critère lorsque l'on passe de l'intervalle $I(\tau)$ partitionné par $P(\tau)$ à l'intervalle $I(\sigma)$ partitionné par $P(\tau, \sigma)$. A m fixé, elle ne dépend pas de $P(\tau)$. Plus précisément, si $P'(\tau)$ est une autre partition de $I(\tau)$ à m intervalles, alors

$$\mathrm{CRIT}(x(\sigma), P(\tau,\sigma)) - \mathrm{CRIT}(x(\tau), P(\tau)) = \mathrm{CRIT}(x(\sigma), P'(\tau,\sigma)) - \mathrm{CRIT}(x(\tau), P'(\tau)).$$

C'est cette égalité qui assure que l'algorithme dynamique 1 présenté dans la suite choisit bien la parition $\widehat{P}$ vérifiant (5.7).

A.2 Description de l'algorithme

A.2.1 Les sorties de l'algorithme

L'algorithme de programmation dynamique produit deux tableaux, appelés dans la suite Tableau et TableauCorrespondance, de taille $R \times R$ dont les lignes seront indexées par $m = 1, \ldots, R$ et les colonnes par $\sigma = 1, \ldots, R$.

Une fois l'algorithme dynamique 1 achevé, Tableau(m, σ) contiendra la valeur minimale du critère sur les données $x(\sigma)$ parmi les sous-partitions de $P_{\max}(\sigma)$ à m intervalles. Par conséquent, la moitié du Tableau pour $m > \sigma$ n'est pas utilisée.

Le TableauCorrespondance stockera quant à lui des nombres entiers servant à reconstituer la partition permettant d'atteindre le minimum. Cette reconstitution se fera à l'aide de l'algorithme de reconstruction 2. La première ligne du TableauCorrespondance ainsi que sa moitié inférieure ($m > \sigma$) ne sont pas utilisées.

A.2.2 L'algorithme dynamique

Nous présentons maintenant l'algorithme dynamique 1 avec les notations précédentes dans un langage de programmation générique. Nous supposons que les quantités $n(\sigma),\ \sigma = 1, \ldots, R$ sont déjà calculées et stockées dans un vecteur encore noté n.

Algorithm 1 L'algorithme dynamique

Require: $n(\sigma),\ \sigma = 1, \ldots, R$
Ensure: Tableau, TableauCorrespondance
for $\sigma = 1$ to R **do**
  Tableau(1,σ) = $n(\sigma)\log(\sigma)$
end for
for $m = 2$ to R **do**
  for $\sigma = m$ to R **do**
    Temp = Vecteur($\sigma - m + 1$)
    for $\tau = m - 1$ to $\sigma - 1$ **do**
      Temp($\tau - m + 2$) = Tableau($m - 1, \tau$) + (A.1)
    end for
    Tableau(m, σ) = min(Temp)
    TableauCorrespondance(m, σ) = Argmin(Temp) $+m - 2$
  end for
end for

La valeur minimale du critère est obtenue en minimisant les valeurs présentes sur la dernière colonne du Tableau ($\sigma = R$). Nous notons $\widehat{m}$ l'indice de la ligne correspondant à ce minimum ; c'est également le nombre d'intervalles de la partition minimisant le critère.

A.2.3 L'algorithme de reconstruction

Pour reconstruire la partition $\widehat{P}$ vérifiant (5.7), nous aurons besoin de $\widehat{m}$, du TableauCorrespondance construit dans l'algorithme dynamique et de l'algorithme 2 de reconstruction présenté maintenant.

Algorithm 2 L'algorithme de reconstruction

Require: $\widehat{m}$, TableauCorrespondance
Ensure: Partition
Partition = Vecteur($\widehat{m}$)
Partition($\widehat{m}$) = R
for $i = \widehat{m}$ to 2 by -1 **do**
  Partition($i - 1$) = TableauCorrespondance(i, R)
  R = Partition($i - 1$)
end for

Notons le vecteur à $\widehat{m}$ composantes Partition, sortie de l'algorithme 2 de reconstruction, ainsi :

$$\text{Partition} = (\sigma_1, \ldots, \sigma_{\widehat{m}}).$$

La partition $\widehat{P}$ vérifiant (5.7) recherchée est alors

$$\widehat{P} : a + 0.r = a < a + \sigma_1 r < \cdots < a + \sigma_{\widehat{m}} r = b$$

A.2.4 Le nombre d'itérations

L'algorithme dynamique 1 présente trois boucles FOR imbriquées qui génèrent la quasi-totalité des calculs nécessaires à son exécution. Le nombre d'itérations correspondant à ces boucles vaut

$$\sum_{m=2}^{R} \sum_{\sigma=m}^{R} \sum_{\tau=m-1}^{\sigma-1} 1 = \frac{1}{6}(R^3 - R).$$

Cette quantité est bien moindre que le nombre de sous-partitions de $P_{\max}$ qui est 2^{R-1}.

Nous avions cependant annoncé précédemment que l'algorithme dynamique se terminait en un nombre d'opérations de l'ordre de R^2. Cela est vrai si nous nous limitons aux sous-partitions de $P_{\max}$ présentant un nombre d'intervalles plus petits qu'un certain I_{m} fixé. Dans ce cas, il suffit d'adapter l'algorithme 1 en remplaçant R par I_{m} à la ligne 4 et le nombre d'itérations devient

$$\sum_{m=2}^{I_{\mathrm{m}}} \sum_{\sigma=m}^{R} \sum_{\tau=m-1}^{\sigma-1} 1 = \frac{1}{2}R^2(I_{\mathrm{m}} - 1) + R(-\frac{1}{2}I_{\mathrm{m}}^2 + I_{\mathrm{m}} - \frac{1}{2}) + \frac{1}{6}I_{\mathrm{m}}^3 - \frac{1}{2}I_{\mathrm{m}}^2 + \frac{1}{3}I_{\mathrm{m}}.$$

Cette limitation sur le nombre maximal d'intervalles que pourra contenir la partition recherchée est peu restrictive en pratique. En effet, le nombre R est souvent très élevé et il importe peu à l'utilisateur de choisir une partition contenant un aussi grand nombre d'intervalles. De plus, une première utilisation de l'algorithme pour $I_{\mathrm{m}} = R$ pourra permettre à l'utilisateur de situer approximativement le nombre optimal de classes de l'histogramme pour le type de données qu'il utilise et donc de choisir un I_{m} adéquat pour ses prochaines expériences.

Annexe B

Codage arithmétique pour la description d'une distribution

Cette annexe présente l'article [CAOA] paru dans les comptes-rendus de la conférence Traitement et Analyse de l'Information : Méthodes et Applications TAIMA'07.

Codage arithmétique pour la description d'une distribution

Guilhem Coq[1], Olivier Alata[2], Christian Olivier[2], et Marc Arnaudon[1]

[1] Laboratoire de Mathématiques et Applications, UMR CNRS 6086
BP 30179 - 86962 Futuroscope Chasseneuil Cedex France
Tél : 05 49 49 68 97 Fax : 05 49 49 69 01
coq,arnaudon@math.univ-poitiers.fr

[2] Laboratoire Signal Image et Communication, EA 4103
BP 30179 - 86962 Futuroscope Chasseneuil Cedex France
Tél. : 05 49 49 65 67 Fax : 05 49 49 65 70
alata,olivier@sic.sp2mi.univ-poitiers.fr

Résumé Partant du codage arithmétique adaptatif et utilisant le principe du Minimum Description Length, nous arrivons à un outil efficace pour la sélection de modèles : le critère d'information RIC. Nous présentons ensuite une extension de ces techniques de codage à l'estimation non-paramétrique et l'illustrons sur l'histogramme des niveaux de gris d'une image.

Mots clés Critères d'informations, MDL, sélection de modèles, estimation non-paramétrique, histogrammes.

1 Introduction

Le codage arithmétique, présenté par Rissanen [7], est optimal en terme d'entropie. Une version simple de ce codage, pour laquelle nous renvoyons à [6], est utilisée dans JPEG2000 où plusieurs modèles de référence sont utilisés. Nous présentons une version adaptative, utilisée notamment dans le codeur d'images médicales CALIC, qui est un outil efficace pour la sélection de modèles. Sa longueur entre en effet dans le cadre plus général des critères d'informations ou d'entropie pénalisée, introduits par exemple dans [1,10] et dont les domaines d'applications sont nombreux, citons [2,4]. Nous présentons ensuite une procédure de sélection de la partition, non nécessairement régulière, d'un histogramme basée sur ces techniques de codage.

2 Codage entropique et arithmétique adaptatif

2.1 Codage entropique

Soit E un ensemble de m symboles. Un code binaire sur E est une application injective $C : E \to \cup_{i\in\mathbb{N}^*}\{0,1\}^i$. La longueur de $C(x)$ est notée $L(x)$. On code ainsi chaque symbole par une chaîne. Si L vérifie l'inégalité de Kraft [5], on sait qu'elle est la longueur d'un certain code qui satisfait la condition du préfixe, indispensable au décodage. Prenant P une probabilité sur E et $L = \lceil -\log P \rceil$, où log est le logarithme à base 2, L vérifie cette inégalité et est donc la longueur d'un code que nous confondrons avec P. Ainsi, si $P(x)$ est grand, $L(x)$ est faible.

Rappelons l'inégalité de convexité de Jensen : si P et Q sont deux probabilités sur E, en notant $\mathbb{E}_P$ l'espérance sous P, on a :

$$H(P) := \mathbb{E}_P[-\log P] \leq \mathbb{E}_P[-\log Q] =: H(P,Q) \tag{1}$$

Sur des données provenant de P inconnue, l'objectif est donc de trouver un codage Q dont l'entropie croisée $H(P,Q)$ se rapproche de $H(P)$. A cet effet, le codage de Huffman est optimal. Cependant le codage arithmétique, consistant à coder plusieurs symboles simultanément, donne de meilleurs résultats.

2.2 Chaînes de Markov multiples

Les chaînes de Markov multiples (CMM) sont le cadre naturel du codage arithmétique. Un processus $(X_n)_{n\in\mathbb{N}^*}$ à valeurs dans E est une CMM d'ordre $k \in \mathbb{N}$ si k est le plus petit entier vérifiant l'égalité $P(X_n|X_{n-1},\ldots,X_0) = P(X_n|X_{n-1},\ldots,X_{n-k})$ pour tout n. Nous nous placerons toujours dans le cas où cette loi conditionnelle ne dépend pas de n ; la chaîne est alors dite homogène. Une CMM d'ordre 0 est une suite de variables aléatoires indépendantes.

Prenons les k premières variables d'une CMM d'ordre k indépendantes et de distribution uniforme sur E. Notons $i \in E$ un état, $j \in E^k$ un état composé et $\theta(i|j)$ la probabilité de voir apparaitre i après j. La donnée des $(m-1)m^k$ réels $\theta(i|j)$, pour j parcourant E^k et i parcourant $m-1$ états de E, suffit à décrire l'évolution de X. Pour θ un tel paramètre et $x^n = x_1,\ldots,x_n$ une chaîne d'éléments de E, la vraisemblance de x^n relativement à θ s'écrit :

$$P(x^n|\theta) = \frac{1}{m^k} \prod_{j\in E^k} \prod_{i\in E} \theta(i|j)^{n(i|j)} \tag{2}$$

avec $n(i|j)$ le nombre d'occurences de i après j dans x^n.

2.3 Codage arithmétique adaptatif

Soit l'intervalle courant $I_c = [0,1[$. Pour coder $x^n \in E^n$ à l'ordre k choisi au préalable, on procède par itération. Supposons traités les t premiers symboles, $t \geq 0$. Pour traiter le $(t+1)-$ième, on actualise les probabilités de transitions comme suit :

$$\hat{\theta}^{(t)}(i|j) = \frac{n^{(t)}(i|j)+1}{n^{(t)}(j)+m}$$

où $i \in E$, $j \in E^k$, $n^{(t)}(i|j)$ et $n^{(t)}(j)$ sont les nombres d'occurences respectifs de i après j et de j dans la chaîne tronquée après le $t-$ième symbole ; $n^{(t)}(j)$ ne devant pas compter une apparition de j à la fin de cette chaîne. Ces probabilités ne sont jamais nulles. On pose $j = x_{t-k+1},\ldots,x_t$ l'état actuel et on découpe I_c selon les probabilités $\hat{\theta}^{(t)}(i|j)$, un intervalle correspondant à un état i de E. On choisit comme nouvel I_c celui correspondant à x_{t+1}.

Une fois le dernier symbole traité et notant $I_c = [a,b[$, il existe deux nombres dyadiques de longueur $\lceil -\log(b-a)\rceil$ consécutifs dans I_c. On prend pour code de x^n la partie fractionnaire du plus grand de ces nombres. Pour illustration, prenons $E = \{a,b\}$ et codons $abaa$ à l'ordre 1 :

Initialement	$[0 \quad , \quad 1[$	$\hat{\theta}^{(0)}(a\|a) = 1/2,\ \hat{\theta}^{(0)}(a\|b) = 1/2$
Découpage	$[0 \quad , \ 1/2 \ , \quad 1[$	
Premier symbole : a	$[0 \quad , \quad 1/2[$	$\hat{\theta}^{(1)}(a\|a) = 1/2,\ \hat{\theta}^{(1)}(a\|b) = 1/2$
Découpage	$[0 \quad , \ 1/4 \ , \quad 1/2[$	
Deuxième symbole : b	$[1/4 \quad , \quad 1/2[$	$\hat{\theta}^{(2)}(a\|a) = 1/3,\ \hat{\theta}^{(2)}(a\|b) = 1/2$
Découpage	$[1/4 \ , \ 3/8 \ , \quad 1/2[$	
Troisième symbole : a	$[1/4 \quad , \quad 3/8[$	$\hat{\theta}^{(3)}(a\|a) = 1/3,\ \hat{\theta}^{(3)}(a\|b) = 2/3$
Découpage	$[1/4 \ , 7/24 \ , \quad 3/8[$	
Quatrième symbole : a	$[1/4 \quad , \quad 7/24[$	

Le code 01001 convient puisque $1/4 \leq 2^{-2} + 2^{-5} < 7/24$ et $1/4 \leq 2^{-2} < 7/24$. Notons que, contrairement au cas simple, seul l'ordre k est nécessaire au codage et au décodage.

Remarque 1. Lors de ce codage, nous apprenons les régularités d'ordre k de la chaîne à mesure que nous la découvrons. Par conséquent, plus la chaîne est régulière à cet ordre, plus nous choisirons des intervalles "grands", plus la longueur de codage sera faible.

3 Sélection de modèles par Minimum Description Length (MDL)

Considérons le problème de sélection de modèles suivant : étant donné une chaîne x^n, sélectionner l'ordre $\hat{k}$ d'une CMM dont x^n serait une réalisation. Pour $k \in \mathbb{N}$, notons Θ_k le modèle d'ordre k constitué des paramètres décrivant les transitions d'une CMM d'ordre k et Θ la réunion des Θ_k. Le nombre de composantes libres d'un $\theta \in \Theta_k$ est $|\Theta_k| = (m-1)m^k$.

Appelons complexité stochastique de x^n relativement au modèle d'ordre k la longueur du code arithmétique adaptatif de x^n à l'ordre k, notée $C_k(x^n)$. Suivant la remarque 1, si x^n est une réalisation d'une CMM d'ordre $k^\star$, alors $k^\star$ minimise $C_k(x^n)$, et donc son espérance. Le MDL préconise donc de choisir pour $\hat{k}$ l'ordre minimisant $C_k(x^n)$ ou $\mathbb{E}\,[C_k(x^n)]$.

3.1 Estimation de la complexité stochastique

Le calcul des $C_k(x^n)$ étant complexe, Rissanen effectue dans [8] une étude détaillée de $\mathbb{E}\,[C_k(x^n)]$ dont le résultat essentiel est : pour $k \in \mathbb{N}$, $\varepsilon > 0$, presque-tout $\theta_k \in \Theta_k$ et n assez grand on a :

$$nH(\theta_k) + (1-\varepsilon)\frac{|\Theta_k|}{2}\log n \leq \mathbb{E}_{\theta_k}\,[C_k(x^n)] \leq nH(\theta_k) + \frac{|\Theta_k|}{2}\log n + o(\log n). \qquad (3)$$

La première inégalité de (3) peut-être vue comme un raffinement de (1) : sur des données provenant de θ_k inconnu, le codage adaptatif à l'ordre k donne en moyenne un nombre de bits par symbole, $\mathbb{E}_{\theta_k}[C_k(x^n)]/n$, plus élevé que $H(\theta_k) + |\Theta_k|\log n/2n$. Le facteur $|\Theta_k|\log n/2n$ empêche l'entropie du codage adaptatif de se rapprocher de l'entropie théorique $H(\theta_k)$.

A partir de x^n, nous estimons $H(\theta_k)$ par $-1/n \log P(x^n|\hat{\theta}_k)$, où $\hat{\theta}_k$ est l'estimateur au sens du maximum de vraisemblance de θ_k au sein du modèle Θ_k. Les inégalités (3) suggèrent d'estimer $\mathbb{E}_{\theta_k}[C_k(x^n)]$ par :

$$\mathrm{RIC}(x^n, k) = -\log P(x^n|\hat{\theta}_k) + \frac{|\Theta_k|}{2}\log n, \qquad (4)$$

et le principe du MDL répond au problème de sélection de modèles posé par :

$$\hat{k} = \text{Argmin}\{\text{RIC}(x^n, k) \mid k \in \mathbb{N}\}. \tag{5}$$

Ce critère RIC (Rissanen Information Criterion) prend la même forme que BIC (Bayesian Information Criterion) proposé par Schwarz [10] et étudié dans le cadre des CMM par Zhao et al. [11]. On sait que la longueur du codage arithmétique simple [6] de x^n à l'ordre k avec le paramètre $\hat{\theta}_k$ est $\lceil -\log P(x^n|\hat{\theta}_k) \rceil$; ainsi, minimiser la longueur de codage simple revient à maximiser la vraisemblance. En termes de critères d'informations, c'est donc le fait de coder de manière adaptative qui crée la pénalité $\frac{|\Theta_k|}{2} \log n$.

3.2 Comparaison des codages et critères sur simulation d'une CMM

Nous générons une réalisation x^n d'une CMM d'ordre $k^\star = 5$ à 2 états avec $n = 2000$. L'entropie du paramètre θ_5 utilisé, non explicité ici, vaut $H(\theta_5) = 0.527$. Sur cette chaîne, pour $k = 0, \ldots, 7$, nous effectuons un codage arithmétique simple à l'ordre k avec le paramètre $\hat{\theta}_k$, un codage arithmétique adaptatif à l'ordre k, le calcul de $\text{MV}(x^n, k) = -\log P(x^n|\hat{\theta}_k)$, et enfin le calcul de $\text{RIC}(x^n, k)$. Les résultats divisés par n sont présentés en figure 1. Les courbes de codage adaptatif et RIC présentent nettement un minimum en $k^\star$; cela s'explique par la remarque 1. Le critère RIC choisit le bon ordre : $\hat{k} = k^\star$.

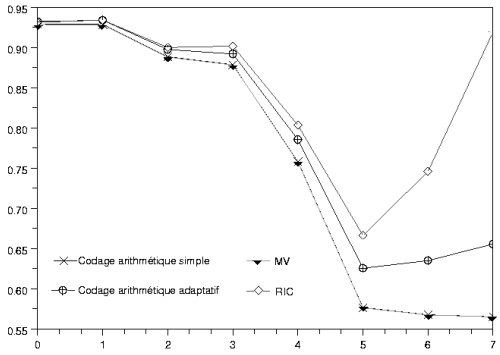


Fig. 1. Comparaison des différentes longueurs de codages et du critère étudié

4 Application à la description d'une distribution par histogramme

Nous présentons ici une application du codage arithmétique adaptatif au cadre non-paramétrique de l'estimation de densité par histogramme. Soit une densité f inconnue sur I et x^n un échantillon de cette distribution. Soit Π une partition de I à m intervalles $(I_j)_{j \in [\![1,m]\!]}$. Nous allons coder sans perte les données x^n à l'aide de Π.

Pour $i \in [\![1, n]\!]$, on note $y_i \in E = [\![1, m]\!]$ le numéro de l'intervalle dans lequel tombe x_i. Par indépendance des x_i, le codage arithmétique adaptatif de y^n à l'ordre $k = 0$ sera le

meilleur. Notons $L(y^n|\Pi)$ la longueur de ce codage. A i fixé, nous pouvons retrouver le x_i correspondant à un y_i en effectuant, à l'intérieur de l'intervalle I_{y_i}, un codage à longueur fixe $\log l_{y_i}/r$ où l_{y_i} est la longueur de I_{y_i} et r la précision de la machine. La longueur du code nécessaire pour retrouver x^n à partir de y^n est alors $L(x^n|y^n) = \sum_{j=1}^{m} n_j \log l_j - n \log r$ où n_j est le nombre de x_i tombant dans I_j.

La longueur du code sans perte de x^n est $L(x^n|\Pi) = L(y^n|\Pi) + L(x^n|y^n)$. Il faut, pour décoder, connaitre la partition utilisée ; la longueur nécessaire à son codage étant faible devant $L(x^n|\Pi)$, nous l'omettons. Estimons $L(y^n|\Pi)$ par $\mathrm{RIC}(y^n, 0)$ (4) et définissons le critère :

$$\mathrm{Crit}(x^n|\Pi) = \mathrm{RIC}(y^n, 0) + L(x^n|y^n) = -\log P(y^n|\hat{\theta}_0) + \frac{m-1}{2}\log n + \sum_{j=1}^{m} n_j \log l_j - n \log r$$

où $\hat{\theta}_0(j) = n_j/n$ est estimé au sens du maximum de vraisemblance. Utilisant (2) il vient

$$\mathrm{Crit}\ (x^n|\Pi) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n l_j} + \frac{m-1}{2}\log n - n \log r \tag{6}$$

qui entre dans le cadre général des critères utilisés par exemple par Birgé [3] pour la sélection d'un histogramme.

Le principe du MDL préconise de choisir pour partition celle qui minimise ce critère. Le nombre de partitions de I étant trop élevé, on se restreint à la classe des sous-partitions d'une partition Π_{max} donnée. Si Π_{max} a R intervalles, il y a 2^{R-1} telles sous-partitions. Rissanen et al. [9] présentent une méthode de programmation dynamique qui permet de ramener en $O(R^2)$ le nombre d'opérations à effectuer pour trouver la sous-partition optimale.

Pour une densité Laplacienne $e^{-|x|}/2$ (par exemple une distribution de coefficients DCT dans JPEG), avec $I = [-5, 5]$ et Π_{max} la partition régulière de pas 2.10^{-2}, on obtient la partition présentée en figure 2.(a). Sur l'histogramme des 256 niveaux de gris de l'image Léna, avec $I = [0, 255]$, la partition choisie 2.(b) a 39 intervalles. Dans les deux cas, le critère choisit plus d'intervalles aux endroits où la densité présente de fortes variations.

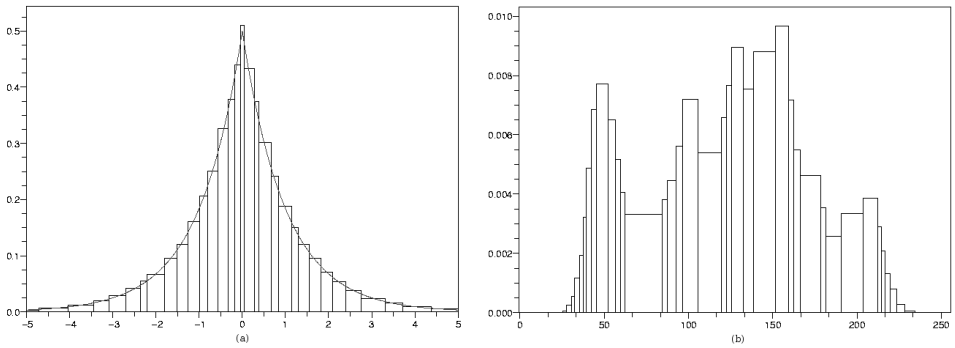


Fig. 2. (a) : Histogramme choisi par le critère sur une densité Laplacienne
(b) : Histogramme approché de l'image Léna en 39 niveaux de gris

Fig. 3. L'image originale (a) et l'image reconstruite (b) sur 39 niveaux de gris pour un PSNR de 38,52 dB

La reconstruction de l'image Léna sur les 39 niveaux de gris choisis est donnée en figure 3. Cette image possède un PSNR de 38,52 dB par rapport à l'image d'origine en 256 niveaux de gris.

En conclusion, nous avons présenté, à partir du MDL, un procédé de description d'une distribution par un histogramme. L'obtention d'un tel histogramme à partir des données numériques peut-être exploitée dans un contexte de reconnaissance de formes. De plus, l'utilisation du critère présenté peut aussi être d'un intérêt certain dans la chaîne de codage source d'une image ou d'une vidéo.

Références

1. H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.
2. Olivier Alata and Christian Olivier. Choice of a 2-d causal autoregressive texture model using information criteria. *Pattern Recognition Letters*, 24(9-10) :1191–1201, 2003.
3. Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10 :24–45 (electronic), 2006.
4. Abdelaziz El Matouat, Christian Olivier, and Frédéric Jouzel. Choix du nombre de composantes d'un modèle de mélange gaussien par critères d'informations. *12ème congrès RFIA, Paris*, 2000.
5. Peter D. Grunwald, In Jae Myung, and Mark A. Pitt. *Advances in Minimum Description Length : Theory and Applications (Neural Information Processing)*. The MIT Press, 2005.
6. Paul G. Howard and Jeffery Scott Vitter. Arithmetic coding for data compression. Technical Report Technical report DUKE–TR–1994–09, 1994.
7. Jorma Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20(3) :198–203, 1976.
8. Jorma Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3) :1080–1100, 1986.
9. Jorma Rissanen, Terry P. Speed, and Bin Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2) :315–323, 1992.
10. Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.
11. L. C. Zhao, C. C. Y. Dorea, and C. R. Goncalves. On determination of the order of a Markov chain. *Statistical Inference for Stochastoc Processes*, 4(3) :273–282, 2001.

Annexe C

Information Criteria and arithmetic codings : an illustration on raw images

Cette appendice présente l'article [COAA] publiée dans les comptes-rendus de la 15th European Signal Processing Conference EUSIPCO 2007.

INFORMATION CRITERIA AND ARITHMETIC CODINGS : AN ILLUSTRATION ON RAW IMAGES

Guilhem Coq[1] , *Christian Olivier*[2] , *Olivier Alata*[2] , *Marc Arnaudon*[1]

[1] Laboratoire de Mathématiques et Applications
Université de Poitiers
Téléport 2 - BP 30179 86962 Chasseneuil FRANCE
Phone: +(33) 5 49 49 68 97 Fax: +(33) 5 49 49 69 01
email: coq,arnaudon@math.univ-poitiers.fr

[2] Laboratoire Signal Image et Communications
Université de Poitiers
Téléport 2 - BP 30179 86962 Chasseneuil FRANCE
Phone: +(33) 5 49 49 65 67 Fax: +(33) 5 49 49 65 70
email: olivier,alata@sic.sp2mi.univ-poitiers.fr

ABSTRACT

In this paper we give a short theoretical description of the general predictive adaptive arithmetic coding technique. The links between this technique and the works of J. Rissanen in the 80's, in particular the BIC information criterion used in parametrical model selection problems, are established. We also design lossless and lossy coding techniques of images. The lossless technique uses a mix between fixed-length coding and arithmetic coding and provides better compression results than those separate methods. That technique is also seen to have an interesting application in the domain of statistics since it gives a data-driven procedure for the non-parametrical histogram selection problem. The lossy technique uses only predictive adaptive arithmetic codes and shows how a good choice of the order of prediction might lead to better results in terms of compression. We illustrate those coding techniques on a raw grayscale image.

1. INTRODUCTION

Arithmetic Coding (AC) is an efficient binary coding technique. We use it here in one of its most general form : the *predictive* and *adaptive* one. Even though those aspects of AC are known, it is quite hard to find literature dealing with both of them ; as well as to determine which aspects are actually used in image coding norms such as JPEG and JPEG2000. We try here to answer the first issue but could not collect useful informations about the second. This paper does not seek compression efficiency but wants to show how different AC processes may be used in both parametrical (§3) and non-parametrical (§4) model selection problems. This explains why we choose to work on raw images.

After a description of AC algorithm in §2, we take a closer look at the resulting codelength. To this end, we use works of J. Rissanen in [6, 7] and especially [8]. The main conclusion of §3 is that the codelength enters the family of information criteria, a widely used tool in the vast problem of model selection. We aim at showing that the *adaptive* aspect of the AC used here is an essential feature.

Next, we design in §4 a new lossless coding technique. It uses a mix between AC, which is compression efficient, and fixed-length coding, which is not. It is shown in §4.2 that correctly mixing those two methods gives better compression efficiency than using only AC. The most important parameter to be adjusted in order to get that "correct" mix is the order of prediction. Moreover, that method is shown in §4.3 to have a direct application in the histogram selection problem.

Finally we design in §5 a lossy coding technique which, once again, shows the importance of the order of prediction.

2. GENERALITIES ON ARITHMETIC CODING

2.1 Multiple Markov Chain

The notion of Multiple Markov Chain (MMC) leads to arithmetic coding. Let $E = \{a_1, \dots, a_m\}$ be a finite set with m elements. An E-valued process $(X_n)_{n\in\mathbb{N}^*}$ is an order k MMC if $k \in \mathbb{N}$ is the smallest integer satisfying the law equality $\mathbb{P}(X_n|X_{n-1}, \dots, X_1) = \mathbb{P}(X_n|X_{n-1}, \dots, X_{n-k})$ for all n. We will always work in the case where that law does not depend on n ; the chain is said homogeneous. An order 0 MMC is a sequence of independent random variables.

If X is an order k MMC, we will suppose that $X_1, \dots, X_k$ are independent and uniformly distributed on E. For $i \in E$ a state and $j \in E^k$ a multiple state, we denote by $\theta(i|j)$ the probability to see i after j. Consequently, choosing $(m-1)m^k$ real numbers $\theta(i|j)$ for $j \in E^k$ and $i \in \{a_1, \dots, a_{m-1}\}$ is enough for describing the evolution of X. Let θ denote such a parameter and $x^n = x_1, \dots, x_n$ be a sequence of elements of E, the likelihood of x^n relatively to θ writes as :

$$\mathbb{P}(x^n|\theta) = \frac{1}{m^k} \prod_{j\in E^k} \prod_{i\in E} \theta(i|j)^{n(i|j)} \tag{1}$$

where $n(i|j)$ is the number of occurences of i after j in x^n.

2.2 Predictive adaptive arithmetic coding : PAAC

We deal here with a general AC which is both k-predictive and adaptive ; we shorten it to k-PAAC. *Predictive* means we code using orders k that may be greater than 1, hence a prediction of the future state of the chain from the current state. *Adaptive* means we do not need any prior knowledge on the chain, except its order ; we learn how to predict the future step by step. Both notions have been formally introduced and studied by Rissanen [6, 7, 8]. For a more concrete description of arithmetic coding, we refer to [11] ; note that this paper does not mention the predictive aspect. Let us now give a theoretical description of the general k-PAAC algorithm.

Let $x^n = x_1, \dots, x_n$ be a chain of elements of E to be encoded and I_c be the current interval firstly set to $I_c = [0, 1)$. For $n \geq t \geq 1$ we note $x^t = x_1, \dots, x_t$. The only prior we need is an order of coding $k \geq 0$, then the algorithm works as follows.

Suppose that the $t \geq 0$ first symbols are dealt with ; $t = 0$ means we have not started the coding yet. To deal with the $(t+1)$-th symbol we actualize transition probabilities as follows :

$$\hat{\theta}^{(t)}(i|j) = \frac{n^{(t)}(i|j) + 1}{n^{(t)}(j) + m}$$

where $i \in E$, $j \in E^k$, $n^{(t)}(i|j)$ and $n^{(t)}(j)$ denote the respective number of occurences of i after j and of j in the chain x^t ; $n^{(t)}(j)$ must not count an occurence of j at the very end of that chain. If $k = 0$, the multiple states j vanish and we set $n^{(t)}(j) = t$. Those probabilities reflect what we know of the chain at the time t of the coding process ; they are the *adaptive* aspect. We then set $j = x_{t-k+1}, \ldots, x_t$ the current state and split the current interval I_c in m smaller intervals according to the probabilities $\hat{\theta}^{(t)}(i|j)$, $i \in E$. This way, we associate to each possible future state $i \in E$ an interval whose length is proportional to the probability with which we expect it. The $(t+1)$-th symbol is dealt with by choosing for new I_c the interval corresponding to $i = x_{t+1}$.

Once the last symbol x_n has been dealt with, we are left with an interval $I_c = [\text{low,high})$. Let $\lceil . \rceil$ denote the superior integer part, there exists two consecutive dyadic numbers with length $\lceil -\log(\text{high-low}) \rceil$ in I_c. We take as the arithmetic code of x^n the sequence of bits given by the fractionnal part of the biggest one. If encoder and decoder agree on the order k of coding, that sequence of bits is decodable, we refer again to [11].

For illustration in table 1, we take $m = 2$, $E = \{a, b\}$ and encode $x^4 = abaa$ at order $k = 1$. In the splits, we allow the left interval to a.

TAB. 1 – Order 1 PAAC of the chain *abaa*.

t	x^t	I_c	$\hat{\theta}^{(t)}(.\|.)$	Split
0	$\emptyset$	$[0,1)$	$(a\|a) = 1/2$ $(a\|b) = 1/2$	$[0, \frac{1}{2}, 1)$
1	a	$[0, \frac{1}{2})$	$(a\|a) = 1/2$ $(a\|b) = 1/2$	$[0, \frac{1}{4}, \frac{1}{2})$
2	ab	$[\frac{1}{4}, \frac{1}{2})$	$(a\|a) = 1/3$ $(a\|b) = 1/2$	$[\frac{1}{4}, \frac{3}{8}, \frac{1}{2})$
3	aba	$[\frac{1}{4}, \frac{3}{8})$	$(a\|a) = 1/3$ $(a\|b) = 2/3$	$[\frac{1}{4}, \frac{7}{24}, \frac{3}{8})$
4	$abaa$	$[\frac{1}{4}, \frac{7}{24})$	$(a\|a) =$ not used $(a\|b) =$ not used	not used
$\lceil -\log(1/4 - 7/24) \rceil = 5$ Code : 01001 ; predecessor : 01000 Both 1/4+1/32 and 1/4 belong to I_c				

This example shows the following general fact about k-PAAC : the more unexpected behaviours occur in the chain, the smaller is the last I_c, the longer is the code. For instance at step $t = 4$ we expected b with probability 2/3, and observed a. This caused us to choose the small interval $I_c = [1/4, 7/24)$. For comparison, if b had occured the code would have been 0110 which is 1 bit shorter. This leads us to the notion of information criteria (IC).

3. INFORMATION CRITERIA

Let us show how the PAAC may be used to solve a model selection problem being : if x^n is a realisation of an unknown MMC (§2.1), which is its order ? More precisely, we will see how the *adaptive* aspect of the PAAC is involved.

3.1 Coding approach of the model selection problem

As mentionned earlier the k-PAAC length of x^n, say $L(x^n|k)$, is ruled by the unexpected events in x^n : the more unexpected events, the longer the code. Consequently, if x^n is ruled by an unknown order $k^\star$ MMC and we try to k-PAAC it at an order $k \neq k^\star$, many unexpected events might occur : either because $k < k^\star$ and we do not look far enough in the past, or because $k > k^\star$ and we take into account informations relative to a too far away past which has actually no influence on the future. Thus the minimization of $L(x^n|k)$ is an appropriate tool for seeking $k^\star$.

The works of Rissanen will confirm that idea and establish a link with Information Criteria (IC).

3.2 Rissanen's result

In [8] it is shown that $L(x^n|k)$ asymptotically behaves as :

$$\text{BIC}(x^n|k) = -\log \mathbb{P}(x^n|\hat{\theta}_k) + \frac{(m-1)m^k}{2} \log n \qquad (2)$$

where $\hat{\theta}_k$ is the maximum likelihood (ML) estimator of order k for x^n, *i.e.* the parameter that maximizes (1).

BIC stands for Bayesian Information Criterion and enters the formalism of IC first introduced by Akaike [1] ; let us mention [10, 5, 3] in addition to [1, 8] as important steps in the theory of IC.

Here is the idea behind IC : the first term of the criterion (2), referred to as the ML term, decreases as k grows. This is mainly because the ML estimator $\hat{\theta}_k$ fits the datas more accurately if we let him look far away in the past. This phenomena is known as *overparametrization* and is the major problem to be solved in model selection, it appears on figure 1. On the other hand, the second term, the penalty, increases as k grows due to $(m-1)m^k$ which is the number of free parameters in the MMCs model of order k. Therefore, the minimization of IC over k realizes a balance between the data fitting, measured by the ML term, and the complexity of the model needed to obtain such a fitting, measured by the penalty.

The quantity $\text{BIC}(x^n|k)$ is much faster to compute than $L(x^n|k)$; the encoder should use BIC before encoding to find which order will achieve the minimum codelength.

One can design a *non-adaptive* order k-predictive arithmetic coding process whose codelength would be exactly $\lceil -\log \mathbb{P}(x^n|\hat{\theta}_k) \rceil = \lceil \text{ML} \rceil$. However, this process requires to send the parameter $\hat{\theta}_k$ for decodability and, especially, it no longer answers the problem of order selection since ML suffers the overparametrization issue. In terms of IC, the *adaptive* aspect of the process creates the penalty term which avoids overparametrization, see again figure 1.

3.3 Comparison of actual codings with criterion

We generate a realization x^n of an order $k^\star = 5$ MMC with $m = 2$ and $n = 25000$. For $k = 0, \ldots, 10$ we encode it with k-PAAC process. We also compute the criterion $\text{BIC}(x^n|k)$ and the quantity $\text{ML} = -\log \mathbb{P}(x^n|\hat{\theta}_k)$. Results are presented on figure 1 divided by n to express them as a bitrate.

As expected, BIC and k-PAAC curves present a minimum at $k = k^\star$ while the ML method overparametrizes at $k = 9$.

Note that, when computing BIC, it is desirable to have enough observations compared to the number of free parameters, empirically :

$$n \approx \alpha(m-1)m^k \text{ with } \alpha \geq 20 \qquad (3)$$

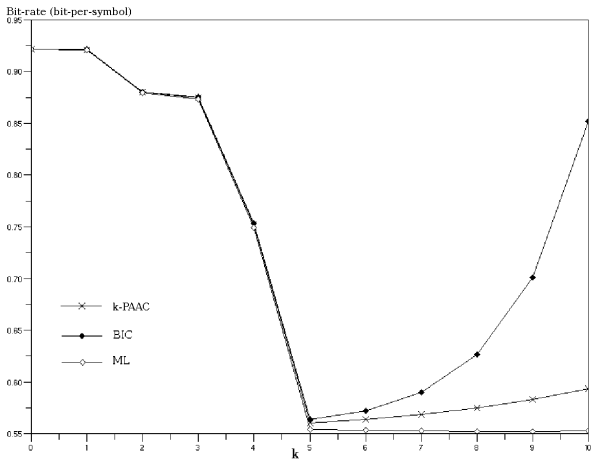


FIG. 1 – Superposition of codelengths and criteria.

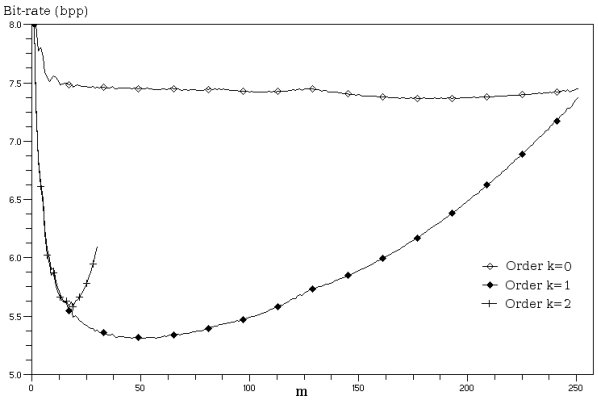


FIG. 2 – Lossless estimated bit-rates of Lena at order 0,1,2.

would be good. If n is too small behind the number of transition probabilities to be estimated, those transitions do not occur often in the chain and their estimation is weak, resulting in the penalty to dominate the ML term. An alternative would be to compute the number of transitions actually observed in the chain and plug them in (2) instead of $(m-1)m^k$.

4. LOSSLESS CODING OF RAW IMAGES

Let $[\![p,q]\!]$ be the set of integers from p to q. Let us choose an $r \times c$ greyscale image and set $n = rc$. Firstly, the image has to be turned into a vector $x^n \in I^n$. For order $k \geq 1$ codings, the way this linearization is done does matter since one does not want to lose proximity information on the pixels. We have chosen the "zigzag" linearization used in 8×8 blocks of DCT transform in JPEG norm [12]. Other transformations have been tested and results are quite similar. Let us now describe our lossless coding method.

4.1 Lossless coding method

It is a two-part coding technique. In first, choose a partition P of $I = [0,255]$; that is a set of m disjoined intervals $(I_j)_{j \in [\![1,m]\!]}$ whose union is I. Then, from x^n, form a new chain y^n as follows :

$$\forall i \in [\![1,n]\!],\ y_i = \sum_{j=1}^{m} j \mathbb{1}_{I_j}(x_i). \tag{4}$$

That is, each y_i denotes the number of the interval of P in which x_i falls. The chain y^n has values in $E = [\![1,m]\!]$. For k an order, we denote by $L(y^n|k,P)$ its k-PAAC codelength. If $m = 1$, we set $L(y^n|k,P)$ to 0.

Secondly, we denote by A_j the number of integers in I_j. Once $y_i = j$ is known one needs, in order to recover $x_i \in I_j$, to specify which one of those integers x_i actually is. This is done for each $x_i \in I_j$ by a simple code with fixed length $\lceil \log A_j \rceil$. Therefore, the number of bits required to recover x^n from y^n is $L(x^n|y^n) = \sum_{j=1}^{m} n_j \lceil \log A_j \rceil$.

For decodability, one should also send the partition chosen to encode. We do not take this into account here since the codelength required to this end is very small compared to the quantities $L(y^n|k,P)$ and $L(x^n|y^n)$ we work on.

Let us note $L(x^n|k,P) := L(y^n|k,P) + L(x^n|y^n)$ the total lossless codelength of x^n with help of the partition P.

4.2 Choice of partition and order of prediction

As m grows $L(y^n|k,P)$ also grows because y^n has values in $[\![1,m]\!]$. By opposition $L(x^n|y^n)$ decreases since the intervals I_j get smaller. Consequently, there should exist a partition P which balances those two phenomena by minimizing the codelentgh $L(x^n|k,P)$. This argument takes place in the theory of Minimum Description Length (MDL) introduced by Rissanen and for which we refer to Grunwald and al. [4].

We estimate $L(y^n|k,P)$ by $\mathrm{BIC}(y^n|k)$, see §3. We then define the following criterion as an estimation of the lossless order k coding of x^n with the partition P :

$$\mathrm{CRIT}(x^n|k,P) = \mathrm{BIC}(y^n|k) + L(x^n|y^n). \tag{5}$$

We restrict ourselves to regular partitions ; *i.e.* partitions $P(m)$ whose intervals all have length $256/m$. We work with the 512×512 greyscale Lena image.

Figure 2 presents, for m ranging from 1 to 256 the estimated bit-rate $\mathrm{CRIT}(x^n|k,P(m))/n$ for $k = 0,1,2$. For $k = 1$, the condition (3) is satisfied for m up to 115 but we still give the $k = 1$ curve up to $m = 256$ for completeness. The algorithm complexity increases considerably with the order k and computations for $k \geq 2$ shows no significant improvements ; in the case $k = 2$ we went up to $m = 30$ which makes α about 10.

Note that our coding technique with $P(1)$ is equivalent to the pgm format[1]. In the other extreme case, with $P(256)$ we get $y^n = x^n$ and $L(x^n|y^n) = 0$; this means we directly encode the chain x^n with the k-PAAC process. Considering this, figure 2 shows how a mix of those two methods leads to better bit-rates. The minimization of the criterion (5) tells us which partition is to be chosen in order to get the correct mix.

More important, 1-PAAC is clearly seen to reaches better bit-rates than 0-PAAC : roughly 7 bpp with huge $P(200)$ partition for 0-PAAC against 5.4 bpp with $P(50)$ for 1-PAAC. Note that the order k chosen for the coding process only affects the first term $\mathrm{BIC}(y^n|k)$ of the criterion (5), hence we

[1] http ://www.imagemagick.org/script/formats.php

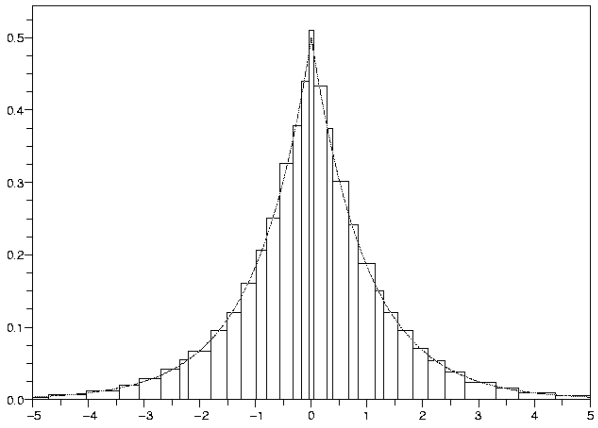


FIG. 3 – Laplace distribution and histogram chosen by (6).

may also give the following interpretation of the curves in figure 2 : no matter how we quantize them via a partition, the grey scales in our image should not be considered independent but rather of order 1. Unsurprisingsly, that dependance of a pixel greyscale on its neighboors may be shown this way on most of common images which content is comprehensible by the human brain.

4.3 Histogram selection statistical problem

It is interesting to note that the criterion (5) may be directly extended to the histogram selection statistical problem : if f is an unknown density on an interval I and x^n is a sample from this density, which partition of I is to be chosen for building an histogram estimator of f ?

For such a partition P, by independence of x^n and formula (4), it is readily seen that the y_i's are independent so that the 0-PAAC of y^n will be the best. Let us denote by L_j the length of I_j and suppose that each I_j contains a number of real numbers proportional to L_j. Then, up to terms which do not depend on P and after little calculations, the estimated lossless order 0 codelength of x^n using P is :

$$\begin{aligned} \mathrm{CRIT}(x^n|0,P) &= \mathrm{BIC}(y^n,0)+L(x^n|y^n). \\ \mathrm{CRIT}(x^n|0,P) &= -\sum_{j=1}^{m} n_j \log\frac{n_j}{nL_j}+\frac{m-1}{2}\log n. \quad (6) \end{aligned}$$

This criterion is in shape really similar to the one used by Birgé and al. in [2] except it has a coding background which justifies its use. Moreover it is not restricted to regular partitions of I. If I is supposed to contain R real numbers, there could be 2^{R-1} partitions to be tested, which is huge. Rissanen and al. presented in [9] a dynamic programing method which shrinks to $O(R^2)$ the number of computations required to find which one of the 2^{R-1} partitions achieves the minimum of (6). For illustration, we present in figure 3 the partition chosen on a 2000-sample from the Laplace distribution used to represent DCT coefficients in the JPEG norm. We assume that $I=[-5,5]$ and $R=200$.

5. LOSSY CODING OF RAW IMAGES

We keep the same linearization as in §4 to turn an image into a vector x^n and now describe our lossy coding method.

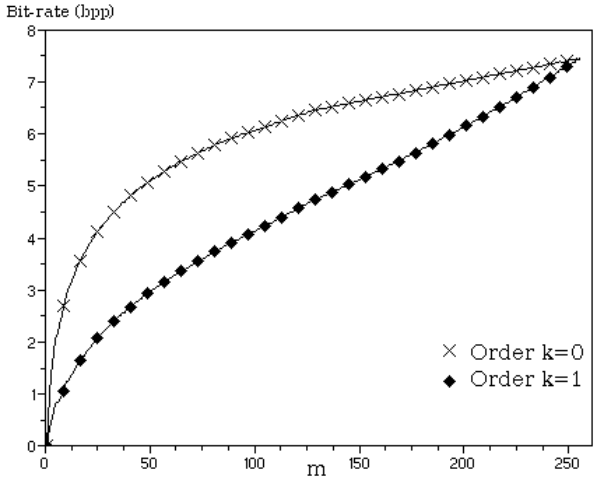


FIG. 4 – Estimated Lena's bit-rates for 0-PAAC and 1-PAAC.

5.1 Lossy coding method

For P a partition of $[0,255]$ in m intervals, we define the $[\![1,m]\!]$-valued chain y^n as in (4). Next, we quantize the datas x^n on P at their barycenter. That is, for each $j\in[\![1,m]\!]$, we consider all x_i's falling into I_j, compute their barycenter, round it to the closest integer B_j and finally set all those x_i's to B_j. This gives a new image with only m grey levels, this is where the loss occurs. Moreover, that quantization creates an injective map :

$$\begin{array}{rccc} B: & [\![1,m]\!] & \longrightarrow & [\![0,255]\!] \\ & j & \longmapsto & B_j \end{array}$$

With the help that map, the decoder is able to reconstruct the quantized image from only the chain y^n ; therefore B is to be sent. However, the coding of such a map is very short compared to the codelength of the chain y^n, so we drop it.

Now we are left to encode y^n with the k-PAAC process, hence the estimation of the lossy codelength of our image by the BIC criterion (2) :

$$\mathrm{BIC}(y^n|k)=-\log\mathbb{P}(y^n|\hat{\theta}_k)+\frac{(m-1)m^k}{2}\log n.$$

5.2 Influence of the order on bit-rates

We still restrict ourselves to regular partition $P(m)$ and work with Lena. Figure 4 presents the estimated bit-rates $\mathrm{BIC}(y^n|k)/n$ for m ranging from 1 to 256 and orders $k=0,1$. For any m, the fact that the $k=1$ curve is under the $k=0$ curve means, as in §4 and via IC interpretation, that the chain y^n is of order 1 rather than order 0.

5.3 Comparison involving distortion

Each value of m brings a certain quantization, thus a certain distortion. We measure this distortion by the Peak Signal to Noise Ratio (PSNR) and plot it against the corresponding bit-rate of 0-PAAC and 1-PAAC in figure 5. For illustration, we present in figure 6 the two quantized Lena images obtained for $m=3$ and $m=13$ with their respective PSNR. We also give bit-rates achieved by 0-PAAC and 1-PAAC on

each of those image. For instance, this shows that at an imposed rate of about 1.4 bpp, the 1-PAAC allows to encode Lena with a PSNR of 33.15 dB while the 0-PAAC only gives 22.11 dB.

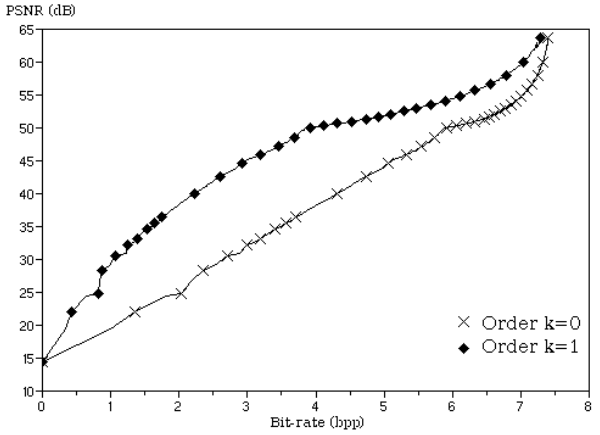


FIG. 5 – Estimated Lena's bit-rates/PSNR for 0 and 1-PAAC.

$m = 3$ levels : 22.11 dB
0-PAAC : 1.36 bpp
1-PAAC : 0.43 bpp

$m = 13$ levels : 33.15 dB
0-PAAC : 3.18 bpp
1-PAAC : 1.39 bpp

FIG. 6 – Estimated PSNR and bit-rates on Lena quantized at $m = 3$ and $m = 13$ levels for 0-PAAC and 1-PAAC.

6. PERSPECTIVES

As mentionned in the introduction we did not provide efficient compression results by intentionally working on raw images. Therefore it would be interesting to insert the discussed binary coding methods after, for instance, the wavelet transform block of the JPEG2000 norm. In order to compress, one should in first determine with the BIC criterion (2) the order of the sequence of wavelet coefficients and then use the criterion (5) to determine the partition which allows to encode those coefficients efficiently.

ACKNOWLEDGMENTS

The authors would like to thank the PIMHAI, INTERREG IIIB "Arc Atlantique" project for its support in the writing of this paper.

REFERENCES

[1] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[2] Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10 :24–45 (electronic), 2006.

[3] Abdelaziz El Matouat and Marc Hallin. Order selection, stochastic complexity and Kullback-Leibler information. In *Athens Conference on Applied Probability and Time Series Analysis, Vol. II (1995)*, volume 115 of *Lecture Notes in Statist.*, pages 291–299. Springer, New York, 1996.

[4] Peter D. Grunwald, In Jae Myung, and Mark A. Pitt. *Advances in Minimum Description Length : Theory and Applications (Neural Information Processing)*. The MIT Press, 2005.

[5] R. Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.*, 27(2) :392–403, 1988.

[6] Jorma Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20(3) :198–203, 1976.

[7] Jorma Rissanen. Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory*, 32(4) :526–532, 1986.

[8] Jorma Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3) :1080–1100, 1986.

[9] Jorma Rissanen, Terry P. Speed, and Bin Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2) :315–323, 1992.

[10] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[11] Ian H. Witten, Radford M. Neal, and John G. Cleary. Arithmetic coding for data compression. *Commun. ACM*, 30(6) :520–540, 1987.

[12] Azza Ouled Zaid, Christian Olivier, Olivier Alata, and Francois Marmoiton. Transform image coding with global thresholding : application to baseline jpeg. *Pattern Recognition Letters*, 24(7) :959–964, 2003.

Annexe D

Law recognition via histogram-based estimation

Cette appendice présente le travail intitulé *Law recognition via histogram-based estimation* et soumis à la conférence International Conference on Acoustics, Speech, and Signal Processing ICASSP 2009 du groupe Institute of Electrical and Electronics Engineers IEEE.

LAW RECOGNITION VIA HISTOGRAM-BASED ESTIMATION

G. Coq

Laboratoire de Mathématiques et Applications
Université de Poitiers
TELEPORT 2 - BP 30179, 86962 Chasseneuil, France
Phone : (+33) 5 49 49 68 97 Fax : (+33) 5 49 49 69 01
email : coq@math.univ-poitiers.fr

O. Alata, Y. Pousset, X. Li, C. Olivier

Laboratoire Signal, Images et Communications
Université de Poitiers
TELEPORT 2 - BP 30179, 86962 Chasseneuil, France
Phone : (+33) 5 49 49 65 67 Fax : (+33) 5 49 49 65 70
email : alata,pousset,li,olivier@sic.sp2mi.univ-poitiers.fr

ABSTRACT

In this paper, we study the problem of recognizing an unknown probability density function from one of its sample which is of interest in signal and image processing or telecommunication applications. By opposition with the classical Kolmogorov-Smirnov method based on empirical cumulative functions, we consider histogram estimators of the density itself built from our data. Those histograms are generated via model selection, more specifically via a codelength-based Information Criterion. From the histograms, we may compute a Kullback-Leibler distance to any theoretical law which is used to complete the recognition. We apply this histogram-based method for law recognition in a theoretical setup where the true density is known as well as in a real setup where data come from radio channel propagation experimentation.

Index Terms— HF radio propagation, probabilty, information criteria, histograms, law recognition

1. INTRODUCTION

Law recognition is a problem of great interest in various domains such as image processing, shape recognition or telecommunication applications. The most widely used tool to solve this problem is the Kolmogorov-Smirnov (KS) test that will be more precisely presented in the sequel. This test is based on a data-based estimation of the cumulative function of our unknown density. Here, we choose to estimate the density itself by an histogram. This estimation of the density will allow computation of Kullback-Leibler type distance on which the law recognition will be based.

In order to estimate the unknown density by an histogram, one needs to use non-parametrical model selection. More precisely Information Criteria (IC), also called penalized likelihood criteria, will be used. Birgé [1] and Birgé and Al. [2] address this problem in recent works. The authors suggest to use an IC in order to determine from the data which histogram is the most suitable for the estimation. The justification of the use of such a criterion is based on the minimization of the risk of the resulting estimation. From another point of view, Rissanen develops in [3] the notion of Minimum Description Length (MDL) and in [4], that of stochastic complexity, strongly related to the theory of coding as expressed in [5]. From those notions, one may, as in [6], construct an IC suitable for our present concern of histogram selection. The use of that latter criterion is thus justified via coding arguments rather than risk-minimizing arguments as were former criteria from Birgé and Al. [1, 2].

In this paper, we work in the continuity of Rissanen and Al. [6] by developping a two-steps coding technique of our data from which is derived a codelength-based Information Criterion. Both notions and their connections are described in section 2. This criterion allows to select, from the data, an histogram estimating the unknown density. Then, in section 3, we present the method of law recognition based on the previous histogram and on Kullback-Leibler distance We also recall the classical Kolmogorov-Smirnov method of law recognition. Part 4 and 5 are dedicated to applications in theoretical and experimental backgrounds.

2. THE HISTOGRAM SELECTION CRITERION

The main setup is as follows : f is an unknown density defined on an interval $I = [a, b]$ of $\mathbb{R}$ and $x^n = x_1, \ldots, x_n$ is a sample from f. Given $P = (I_j)_{j=1,\ldots,m}$ a partition of I into m intervals, one constructs an histogram estimator of f by

$$\widehat{f}_P = \sum_{j=1}^{m} \frac{n_j}{nL_j} \mathbb{1}_{I_j},$$

where $\mathbb{1}_X$ denotes the indicator function of a set X, n_j the number of data x_i falling into I_j and L_j the length of I_j. The main problem of histogram selection is to determine which partition P is to be chosen in order to estimate f by $\widehat{f}_P$.

For the sake of simplicity, we choose $r > 0$ a precision, usually small, and denote by P_{max} the partition of I consisting of R intervals all with length r. Then we restrict ourselves to the 2^{R-1} partitions which intervals are unions of adjacent

intervals of $P_{\max}$. Those are called sub-partitions of $P_{\max}$ and their set is denoted by $\mathcal{SP}$. Note that this restriction allows to handle the case where data live in a discrete space.

Previous work [5] allows to design an IC derived from a data-coding technique answering the partition selection problem. Here, we present the coding technique and the IC resulting in our setting.

2.1. Two steps coding

The main idea is to choose a partition $P \in \mathcal{SP}$ and, with help of it, to encode our data x^n. We consider here the natural idea of coding: transforming our data x^n in a sequence of bits that is decodable if encoder and decoder agree. Then, via the principle of Minimum Description Length [7], the partition to be chosen is the one that realizes the best encoding of our datas. That encoding is now described ; it is lossless up to the precision r and presents two steps.

2.1.1. First step : arithmetic coding

In the first step, data x^n are transformed into y^n as follows

$$y_i = \sum_{j=1}^{m} j.\mathbb{1}_{\{x_i \in I_j\}},\ i = 1, \ldots, n.$$

In other words, y_i denotes the number of the interval of P in which x_i falls. In order to encode y^n, we use a version of the arithmetic coding technique presented in [5] as Predictive Adaptive Arithmetic Coding of order 0, 0-PAAC for short. In the sequel, $L(y^n|P)$ denotes the length of the binary string resulting from such a coding. As also discussed in the previous reference, the PAAC is strongly related to the work of Rissanen [4] in the sense that $L(y^n|P)$ is asymptotically estimated by the so-called stochastic complexity of y^n:

$$L(y^n|P) \approx -\sum_{j=1}^{n} n_j \log \frac{n_j}{n} + \frac{m-1}{2}\log(n). \qquad (1)$$

As m grows, that quantity tends to grow as well since encoding symbols y_i, that may take m different values, gets harder.

2.1.2. Second step : fixed length coding

For a fixed $i = 1, \ldots, n$, the information $y_i = j$ alone does not allow to recover x_i. In order to do this, one needs to precise where x_i is located inside the interval I_j. Up to the precision r, there are L_j/r real numbers in this interval. Consequently, the precision of x_i may be done with an encoding of (ideal) fixed length equal to $\log L_j/r$.

Now, the total number of bits required to precise all the x_i's equals

$$L(x^n|y^n) = \sum_{j=1}^{m} n_j \log \frac{L_j}{r}. \qquad (2)$$

By opposition to (1), this quantity tends to decreases as m increases. Indeed, the larger m, the smaller the intervals of P, the easier it is to precise where each x_i is.

2.1.3. The criterion

Via our two-steps encoding method, the estimated total lossless codelenght of the data x^n with help of the partition P writes as the sum of (1) and (2), that is

$$\text{CRIT}(x^n, P) = -\sum_{j=1}^{n} n_j \log \frac{r n_j}{n L_j} + \frac{m-1}{2}\log(n). \qquad (3)$$

This quantity enters the formalism of Information Criteria (IC), widely used tools in model selection problem for which one may for instance refer to [8, 9, 2].

The MDL principle thus suggests to choose $\widehat{P}$ as

$$\widehat{P} = \text{Argmin}\,\{\text{CRIT}(x^n, P),\ P \in \mathcal{SP}\} \qquad (4)$$

and consider $\widehat{f}_{\widehat{P}}$ as an estimator of the unknown density f. Note that this minimization does not depend on the chosen precision r.

The opposite behaviors of (1) and (2) described earlier reflect the usual fact that this minimization of the IC (3) realizes the best compromise between the complexity of the partition and how well it fits the data.

The resulting histogram is referred to in the sequel as dynamic histogram. This word is inherited from the dynamic programming method introduced by Rissanen [6] that allows to determine $\widehat{P}$ in (4) in a number of operations of the order R^2 instead of having to compute the 2^{R-1} values of the criterion for each $P \in \mathcal{SP}$. We may also restrict ourselves to the class of regular histograms on I. Those are built on partitions P_m that have m intervals all of length $(b-a)/m$ for $m = 1, \ldots, M$. Those M partitions may be seen as sub-partitions of $P_{\max}$ when $r = (b-a)/lcm(1, \ldots, M)$ where lcm denotes the least common multiple. The resulting estimation is referred to as regular histogram. We use the term optimal histograms to refer to either dynamic or regular ones.

3. METHODS

Let $\mathcal{F}$ be a family of density functions on I ; they are the laws in competition.

3.1. Optimal histogram method

Once the optimal histogram estimator $\widehat{f}$ is selected via (3) and (4), we may compute the Kullback-Leibler distance from it to any $f \in \mathcal{F}$ as follows :

$$\text{KL}(\widehat{f}, f) = \frac{1}{2}\int_I (\widehat{f} - f)\log\frac{\widehat{f}}{f} d\mu. \qquad (5)$$

Then, the law recognition is done via

$$f_{\text{KL},opt} = \text{Argmin}(\text{KL}(\widehat{f}, f),\ f \in \mathcal{F}). \qquad (6)$$

3.2. Empirical histogram method

From our set of data x^n, it is usual to build an empirical histogram describing the distribution. Classically this histogram $\widehat{f}_{emp}$ is built on the regular partition of I that counts $\lfloor 2\sqrt{n}-1 \rfloor$ intervals where $\lfloor . \rfloor$ denotes the inferior integer part. From it, one may solve the law recognition problem by

$$f_{\mathrm{KL},emp} = \mathrm{Argmin}(\mathrm{KL}(\widehat{f}_{emp}, f),\ f \in \mathcal{F}). \tag{7}$$

3.3. Kolmogorov-Smirnov method

The Kolmogorov-Smirnov (KS) method is the tool used classically for law recognition. From the data x^n, we construct the empirical cumulative function denoted by $\widehat{F}$. Let us denote by $\mathcal{F}_c$ the set of cumulative functions of the laws $f \in \mathcal{F}$. The KS distance between $\widehat{F}$ and $F \in \mathcal{F}_c$ is

$$\mathrm{KS}(\widehat{F}, F) = \sup_{t \in I} \left| \widehat{F}(t) - F(t) \right|. \tag{8}$$

From that distance, the law recognition is done via

$$F_{\mathrm{KS}} = \mathrm{Argmin}(\mathrm{KS}(\widehat{F}, F),\ F \in \mathcal{F}_c). \tag{9}$$

This method does not require to estimate the density itself.

3.4. The model

Here, our model contains three densities in competition

$$\begin{aligned}
\text{Rayleigh} \quad &: \quad f_R(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\
\text{Nakagami} \quad &: \quad f_N(x) = \frac{2\mu^\mu x^{2\mu-1}}{(\mu-1)!\Omega^\mu} \exp\left(-\frac{\mu x^2}{\Omega}\right) \\
\text{Weibull} \quad &: \quad f_W(x) = \frac{k x^{k-1}}{\lambda^k} \exp\left(-\frac{x^k}{\lambda^k}\right)
\end{aligned} \tag{10}$$

This choice of model is usually done for radio channel propagation modelization. All coefficients $\sigma, \mu, \Omega, k, \lambda$ are shape parameters to be described later. Note that choosing $\mu = 1$ and $k = 2$ in Nakagami and Weibull laws make them similar to a Rayleigh law.

4. APPLICATION IN A THEORETICAL SETUP

Our aim in this part is to show that the recognition method defined in (5) and (6) is efficient and to compare it with the usual KS method ((8) and (9)) and the empirical histogram method (7).

The shape parameters $\sigma, \mu, \Omega, k, \lambda$ in (10) are all set to obtain a mean of 73 and a standard deviation of 1.2. We generate 30 samples of sizes n ranging from 100 to 3000 of the laws in model (10). On each of those sample, we apply the three recognition methods discussed earlier: optimal histograms (6), empirical histogram (7) and KS distance (9). Since the generating law is known, we may compute successfull recognition rates (RR) and plot them in figure 1.

We choose to show only RR of the Rayleigh law. Results are similar for other generating distributions. We see that using optimal histograms, especially dynamic one, yields a lightly better RR than using the usual KS method. This is remarkable since, in our setting, optimal histograms never contain more than 50 bins. This means that resuming the data information to about 50 classes allows a better recognition of the underlying law than conserving all the information in the n steps of the empirical cumulative function used in the KS method. Moreover, if one wants to avoid KS method for law recognition, optimal histograms should be used since the empirical one gives poor results.

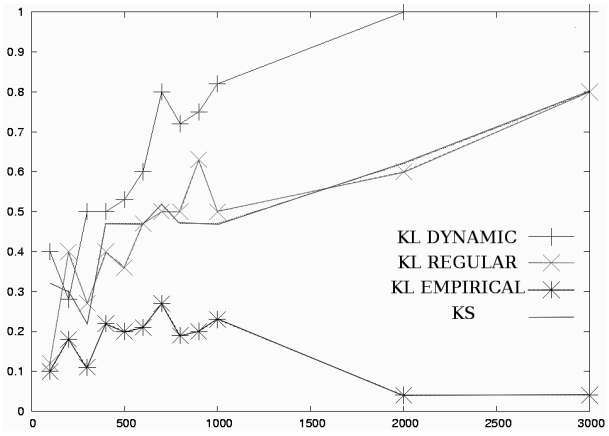


Fig. 1. Rayleigh RR using optimal histograms, empirical histogram and KS methods.

5. APPLICATION IN A REAL SETUP

Laboratory SIC-XLIM developped a software allowing to simulate the fast fading behavior of a radio propagation channel in various environments, see [10]. From this software, we collected $n = 700$ data representing the attenuation (dB) of the signal in both Light Of Sight (LOS) and Non Light Of Sight (NLOS) configurations, see figure 2. We choose to modelize the radio channel by either a Rayleigh, Weibull, or Nakagami distribution from the model (10). In order to determine which of those laws best suits the radio channel, we apply recognition method (6) with optimal histograms.

5.1. LOS configuration

In this experiment, the computed average attenuation and its standard deviation are respectively 73 dB and 1.2 dB. Shape parameters in (10) are set in consequence and shown in Table 1(a). Here, the Weibull distribution suits the best the fast

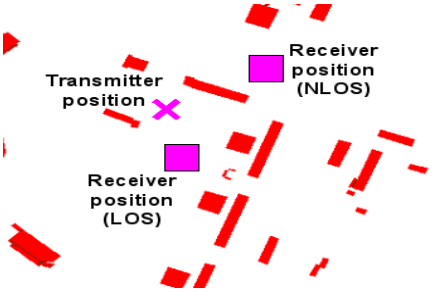


Fig. 2. The environment of the simulation.

fading behavior of the propagation channel in a LOS configuration. This is the usual conclusion as in [11].

5.2. NLOS configuration

Here, average and standard deviation equal respectively 99.8 dB and 3.5 dB. Shape parameters in (10) are set in consequence and shown in Table 1(b). It is not obvious to decide which law suits the best the data but it does not matter since shape parameters $\mu \approx 1$ and $k \approx 2$ for Nakagami and Weibull laws actually correspond to a Rayleigh law. As in [12], the Rayleigh modelization appears as the best in the NLOS configuration.

Histo. \ Law	Rayleigh $\sigma = 1.58$	Nakagami $\mu = 0.68$ $\Omega = 4.99$	Weibull $k = 1.53$ $\lambda = 2.05$
Dynamic	0.12	0.05	0.04
Regular	0.12	0.08	0.06

(a)

Histo. \ Law	Rayleigh $\sigma = 4.09$	Nakagami $\mu = 0.99$ $\Omega = 33.43$	Weibull $k = 2.01$ $\lambda = 5.79$
Dynamic	0.033	0.034	0.035
Regular	0.114	0.115	0.115

(b)

Table 1. KL distances from optimal histograms to the laws in competition in LOS(a) and NLOS(b) cases.

6. CONCLUSION

In this paper, we developped an information-theoretic criteria (3) allowing to estimate an unknown probability law by an histogram. This histogram, summing our data to a few number of parameters and used along with Kullback-Leibler distance, is shown to allow a rate of succesfull law recognition as good as or even better than the usual Kolmogorov-Smirnov method. It is then applied to a realistic environment where it matches usual results of modelization.

7. REFERENCES

[1] Lucien Birgé, "Statistical estimation with model selection," *Indag. Math. (N.S.)*, vol. 17, no. 4, pp. 497–537, 2006.

[2] Lucien Birgé and Yves Rozenholc, "How many bins should be put in a regular histogram," *ESAIM Probab. Stat.*, vol. 10, pp. 24–45 (electronic), 2006.

[3] Jorma Rissanen, "Modeling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[4] Jorma Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[5] Guilhem Coq, Olivier Alata, Christian Olivier, and Marc Arnaudon, "Information criteria and arithmetic codings : an illustration on raw image.," in *15th European Signal Processing Conference proceedings*, pp. 634–638.

[6] Jorma Rissanen, Terry P. Speed, and Bin Yu, "Density estimation by stochastic complexity.," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 315–323, 1992.

[7] Peter D. Grunwald, In Jae Myung, and Mark A. Pitt, *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*, The MIT Press, 2005.

[8] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281. Akadémiai Kiadó, Budapest, 1973.

[9] Peter Hall, "Akaike's information criterion and Kullback-Leibler loss for histogram density estimation," *Probab. Theory Related Fields*, vol. 85, no. 4, pp. 449–467, 1990.

[10] F. Escarieu, Y. Pousset, R. Vauzelle, and L. Aveneau, "Outdoor and indoor channel characterization by a 3d simulation software.," Septembre 2001, PIMRC '2001, San diego, USA.

[11] N. C. Sagias and G. K. Karagiannidis, "Gaussian class multivariate Weibull distributions: Theory and applications in fading channels," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3608–3619, Oct 2005.

[12] T.K. Sarkar, Zhong Ji, Kyungjung Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication," *Antennas and Propagation Magazine, IEEE*, vol. 45, no. 3, pp. 51–82, 2003.

Bibliographie

[Aka69] H. Akaike. Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, 21 :243–247, 1969.

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[Aka74] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[Ald97] J. Aldrich. R. A. Fisher and the making of maximum likelihood 1912–1922. *Statist. Sci.*, 12(3) :162–176, 1997.

[AO03] O. Alata and C. Olivier. Choice of a 2-d causal autoregressive texture model using information criteria. *Pattern Recognition Letters*, 24(9-10) :1191–1201, 2003.

[AR05] O. Alata and C. Ramananjarasoa. Unsupervised textured image segmentation using 2-d quarter plane autoregressive model with four prediction supports. *Pattern Recogn. Lett.*, 26(8), 2005.

[Bar00] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.

[Bar02] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 :127–146 (electronic), 2002.

[Bas89] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing. An Interdisciplinary Journal*, 18(4) :349–369, 1989.

[Bas96] M. Basseville. Information : entropies, divergences et moyennes. *Institut de Recherche en Informatique et Systèmes Aléatoires IRISA*, 1996.

[BB99] M. Benidir and M. Barret. *Stabilité des filtres et des systèmes linéaires.* Dunod, 1999.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

[Bir04] L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6) :1039–1051, 2004.

[Bir06] L. Birgé. Statistical estimation with model selection. *Indag. Math. (N.S.)*, 17(4) :497–537, 2006.

[BP85] J. Berstel and D. Perrin. *Theory of codes*, volume 117 of *Pure and Applied Mathematics.* Academic Press Inc., Orlando, FL, 1985.

[BR06] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10 :24–45 (electronic), 2006.

[Bro66] P. Brodatz. *Texture : a Photographic Album for Artists and Designers.* New York, Dover, 1966.

[Bro00] P. M. T. Broersen. Finite sample criteria for autoregressive order selection. *IEEE Transaction Signal Processing*, 48(12) :3550–3558, 2000.

[BRY98] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6) :2743–2760, 1998.

[CAOA] G. Coq, O. Alata, C. Olivier, and M. Arnaudon. Codage arithmétique pour la description d'une distribution. In *Conférence TAIMA'07*, pages 65–71.

[Cas99] G. Castellan. Modified akaike's criterion for histogram density estimation. *Technical Report 99.61, Université de Paris-Sud*, 1999.

[Cas00] G. Castellan. Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(8) :729–732, 2000.

[COAA] G. Coq, C. Olivier, O. Alata, and M. Arnaudon. Information criteria and arithmetic codings : an illustration on raw image. In *15th European Signal Processing Conference proceedings*, pages 634–638.

[EL06] T. Mary-Huard E. Lebarbier. Un introduction au critère bic : fondements théoriques et applications. *Journal de la Société française de Statistiques*, 147(1) :39–57, 2006.

[EMH96] A. El-Matouat and M. Hallin. Order selection, stochastic complexity and Kullback-Leibler information. 115 :291–299, 1996.

[GMP05] P. D. Grunwald, In Jae Myung, and M. A. Pitt. *Advances in Minimum Description Length : Theory and Applications (Neural Information Processing).* The MIT Press, 2005.

[Hal87] P. Hall. On Kullback-Leibler loss and density estimation. *Ann. Statist.*, 15(4) :1491–1519, 1987.

[Hal90] P. Hall. Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theory Related Fields*, 85(4) :449–467, 1990.

[HQ79] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B*, 41(2) :190–195, 1979.

[Huf52] D. A. Huffman. A method for the construction of minimum redundancy codes. *Proceedings of the Institute of Radio Engineers*, 40(9) :1098–1101, Sep 1952.

[Kay93] S. M. Kay. *Fundamentals of statistical signal processing : estimation theory.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[Mal73] C. Mallows. Some comments on cp. *Technometrics*, 15(661-675), 1973.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics.* Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[MNW98] A. Moffat, R. M. Neal, and I. H. Witten. Arithmetic coding revisited. *ACM Trans. Inf. Syst.*, 16(3) :256–294, 1998.

[NBK88] R. Nishii, Z. D. Bai, and P. R. Krishnaiah. Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, 18(3) :451–462, 1988.

[Nis84] R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, 12(2) :758–765, 1984.

[Nis88] R. Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.*, 27(2) :392–403, 1988.

[OA08] C. Olivier and O. Alata. *Optimization in Signal and Image Processing*, chapter 4, Information Criteria : examples of applications for the processing of signals and images. ISTE, Wiley, June 2008.

[OJM99] C. Oliver, F. Jouzel, and A. El Matouat. Choice of the number of component clusters in mixture models by information criteria. *Proc. Vision Interface*, pages 74–81, May 1999.

[OPAL97] C. Olivier, T. Paquet, M. Avila, and Y. Lecourtier. Optimal order of markov models applied to bankchecks. *IJPRAI*, 11(5) :789–800, 1997.

[Pat01] M. Patzold. *Mobile Fading Channels : Modeling, Analysis and Simulation*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[Ris76] J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM J. Res. Develop.*, 20(3) :198–203, 1976.

[Ris78] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14 :465–471, 1978.

[Ris86a] J. Rissanen. Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory*, 32(4) :526–532, 1986.

[Ris86b] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3) :1080–1100, 1986.

[Ris89] J. Rissanen. *Stochastic complexity in statistical inquiry*, volume 15 of *World Scientific Series in Computer Science*. World Scientific Publishing Co. Inc., Teaneck, NJ, 1989.

[Ris96] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42(1) :40–47, 1996.

[RJ85] S. Ranganath and A. K. Jain. Two-Dimensional Linear Prediction Models - part I : Spectral Factorization and Realization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33(1) :280–299, February 1985.

[RSY92] J. Rissanen, T. P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2) :315–323, 1992.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27 :379–423, 623–656, 1948.

[Shi76] R. Shibata. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63(1) :117–126, 1976.

[Ton75] H. Tong. Determination of the order of a Markov chain by Akaike's information criterion. *J. Appl. Probability*, 12(3) :488–497, 1975.

[vdV98] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[WNC87] I. H. Witten, R. M. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Commun. ACM*, 30(6) :520–540, 1987.

[ZDG01] L. C. Zhao, C. C. Y. Dorea, and C. R. Goncalves. On determination of the order of a Markov chain. *Statistical Inference for Stochastic Processes*, 4(3) :273–282, 2001.

Titre

Utilisation d'approches probabilistes basées sur les critères entropiques pour la recherche d'information sur supports multimédia

Résumé

Les problèmes de sélection de modèles se posent couramment dans un grand nombre de domaines applicatifs tels que la compression de données ou le traitement du signal et de l'image. Un des outils les plus utilisés pour résoudre ces problèmes se présente sous la forme d'une quantité réelle à minimiser appelée critère d'information ou critère entropique pénalisé.

La principale motivation de ce travail de thèse est de *justifier* l'utilisation d'un tel critère face à un problème de sélection de modèles typiquement issu d'un contexte de traitement du signal. La justification attendue se doit, elle, d'avoir un solide fondement mathématique.

Nous abordons ainsi le problème classique de la détermination de l'ordre d'une autorégression. La régression gaussienne, permettant de détecter les harmoniques principales d'un signal bruité, est également abordée. Pour ces problèmes, nous donnons un critère dont l'utilisation est justifiée par la minimisation du coût résultant de l'estimation obtenue. Les chaînes de Markov multiples modélisent la plupart des signaux discrets, comme les séquences de lettres ou les niveaux de gris d'une image. Nous nous intéressons au problème de la détermination de l'ordre d'une telle chaîne. Dans la continuité de ce problème nous considérons celui, *a priori* éloigné, de l'estimation d'une densité par un histogramme. Dans ces deux domaines, nous justifions l'utilisation d'un critère par des notions de codage auxquelles nous appliquons une forme simple du principe de *Minimum Description Length*.

Nous nous efforçons également, à travers ces différents domaines d'application, de présenter des méthodes alternatives d'utilisation des critères d'information. Ces méthodes, dites comparatives, présentent une complexité d'utilisation moindre que les méthodes rencontrées habituellement, tout en permettant une description précise du modèle.

Mot-clefs

Statistiques, traitement du signal, sélection de modèles, critères d'information, critères entropiques pénalisés, autoregression, modèles de Markov, histogrammes, regression gaussienne.

Abstract

Model selection problems appear frequently in a wide array of applicative domains such as data compression and signal or image processing. One of the most used tools to solve those problems is a real quantity to be minimized called information criterion or penalized likelihood criterion.

The principal purpose of this thesis is to *justify* the use of such a criterion responding to a given model selection problem, typically set in a signal processing context. The sought justification must have a strong mathematical background.

To this end, we study the classical problem of the determination of the order of an autoregression. We also work on Gaussian regression allowing to extract principal harmonics out of a noised signal. In those two settings we give a criterion the use of which is justified by the minimization of the cost resulting from the estimation. Multiple Markov chains modelize most of discrete signals such as letter sequences or gray scale images. We consider the determination of the order of such a chain. In the continuity we study the problem, *a priori* distant, of the estimation of an unknown density by an histogram. For those two domains, we justify the use of a criterion by coding notions to which we apply a simple form of the "Minimum Description Length" principle.

Throughout those application domains, we present alternative methods of use of information criteria. Those methods, called comparative, present a smaller complexity of use than usual methods but allow nevertheless a precise description of the model.

Keywords

Statistics, signal processing, model selection, information criteria, penalized entropy criteria, autoregression, Markov models, histograms, Gaussian regression.