Pénalités hiérarchiques pour l'intégration de connaissances dans les modèles statistiques

Marie Szafranski

Université de Technologie de Compiègne
HeuDiaSyC

Thèse

présentée pour obtenir le grade de

Docteur de l'Université de Technologie de Compiègne
Spécialité: Technologies de l'Information et des Systèmes

par

Marie Szafranski

Pénalités hiérarchiques pour l'intégration de connaissances dans les modèles statistiques

Thèse soutenue publiquement le 21 novembre 2008

Jury :

Gérard Govaert	Professeur, UTC	*Président*
Samy Bengio	Chercheur, Google Research	*Rapporteur*
Stéphane Canu	Professeur, INSA Rouen	*Rapporteur*
Florence d'Alché-Buc	Professeur, Université d'Évry Val d'Essonne	*Examinatrice*
Francis Bach	Chargé de recherche INRIA, ENS	*Examinateur*
Yves Grandvalet	Chargé de recherche CNRS, UTC	*Directeur*
Pierre Morizet-Mahoudeaux	Professeur, UTC	*Directeur*

Remerciements

Tout d'abord, j'aimerais exprimer mes plus sincères remerciements à Yves Grandvalet. Durant ces trois années, il a su me faire bénéficier de ses nombreuses qualités scientifiques ; il m'a appris à appréhender un problème dans sa globalité et à ne délaisser aucun des aspects théoriques, algorithmiques ou applicatifs. Sa disponibilité et son aide ont incontestablement contribué au bon déroulement de cette recherche. Je remercie aussi chaleureusement Pierre Morizet-Mahoudeaux pour ses conseils toujours avisés, son amitié et son soutien indéfectible, qui remontent à une époque dépassant largement le cadre de cette thèse. Travailler sous leur direction a été une expérience remarquable, et j'espère avoir le plaisir de continuer à collaborer avec eux dans les prochaines années.

Avant de commencer cette thèse, il a fallu parcourir un long chemin. Aussi, j'exprime toute ma gratitude à Georges Szafranski, alias Papa, pour le nombre incalculable d'heures passées à m'expliquer mes cours de mathématiques, et à Bertrand Vachon, qui m'a aidé à en saisir la portée. Je suis très émue que Bertrand et Brigitte aient été à mes côtés le jour de ma soutenance.

Samy Bengio et Stéphane Canu m'ont fait l'honneur de rapporter cette thèse. Florence d'Alché-Buc, Francis Bach et Gérard Govaert ont également eu la gentillesse de s'intéresser à ces travaux. Je leur suis, à tous, reconnaissante d'y avoir consacré du temps, et je les remercie pour leurs remarques constructives et les pistes de recherche suggérées.

Pendant ma dernière année de thèse, j'ai eu le plaisir de travailler avec Alain Rakotomamonjy. Avec lui, nous avons pu donner à ces travaux une dimension applicative. Son expertise m'a été d'une grande aide. Merci Alain ! Je remercie aussi Gangadhar Garipelli de nous avoir autorisés à exploiter ses données, et aidés à interpréter les résultats obtenus.

Christophe Ambroise, Gérard Govaert et Thierry Denœux m'ont fait découvrir les statistiques, et donné envie d'en connaître plus à ce sujet. Gérard m'a également offert l'opportunité d'enseigner dans ce domaine. Pour ces raisons, je les remercie sincèrement. J'ai aussi beaucoup apprécié les discussions avec Benjamin Quost, qui j'espère restera mon « référent-conseil » en matière d'enseignement ! Enfin, je suis reconnaissante à Stéphane Crozat de m'avoir accordé sa confiance. Ses qualités pédagogiques et son enthousiasme auprès des étudiants permettent d'obtenir d'eux des résultats à la hauteur, et même parfois au delà, de nos attentes. Enseigner à ses côtés a été un véritable plaisir.

Je voudrais également remercier Isabelle Boudot, que j'ai tellement embêtée, pour sa gentillesse et son efficacité. Je pense aussi à Corinne Boscolo, Sabine Vidal, Céline Ledent et Nathalie Alexandre que j'ai toujours eu plaisir à croiser.

Liva Ralaivola, Guillaume Stempfel et toute l'équipe « bases de données et apprentissage automatique » du Laboratoire d'Informatique Fondamentale de Marseille m'ont très gentiment accueillie en tant qu'ATER, et m'ont laissé le temps nécessaire pour finir ma thèse. L'ambiance qui règne dans cette équipe m'aide à ne pas (trop !) regretter HeuDiaSyC. Merci à vous.

Je remercie mes « copains de thèse » pour tous les bons moments passés en leur compagnie : Romain, Clément, Olivier, Stéphane, Zoulficar, Saïd et Mohamed, ainsi que Hani et Farid, mes co-bureaux de la première heure, dont la présence m'a manquée cette dernière année. Enfin, je remercie Erik et Benjamin, qui m'ont fait profiter de leur expérience de vieux briscards, professionnellement il est vrai, mais surtout festivement !

Je n'oublie pas la fabuleuse « Red Mount team », au sens large ! Val et Briac, Bruno, Co et No, Fab, Jib et Vaness, Mo, Hervé, Guillaume, Thierry… Malgré la distance, vous savez être présents et me sortir l'esprit de ce monde aussi passionnant qu'envahissant.

Pour terminer, je ne remercierai jamais assez mes parents, Elisabeth et Georges, d'être toujours là, en toutes circonstances. Je remercie également toute ma famille pour les mêmes raisons, avec une pensée particulière pour Maria et Giovanni, Cecilia, Agnese et Jean-Pierre, et Aurélie qui ont embelli mon quotidien Compiègnois. Matthieu, tu as supporté stoïquement mes nombreuses variations d'humeur au cours de ces derniers mois, malgré la rédaction et la soutenance de ta propre thèse. Merci pour ton soutien, merci pour tout.

« Okay… Now what ? »
Mike Slackenerny, [2007]

Table des matières

Contexte

1

1.1 Motivations

Le sujet de cette thèse est issu de problématiques liées au domaine génomique [1]. Le traitement de puces ADN [Soularue et Gidrol, 2002] permet d'étudier en parallèle le comportement de milliers de gènes dans une condition expérimentale donnée : patient malade ou sain, traitement A ou B, etc. Dans chaque situation, on mesure le niveau d'expression de chaque gène. On peut transcrire l'ensemble de ces informations sous la forme d'un tableau de données, où les caractéristiques sont les gènes (cf. tableau 1.1).

mesures	gène 1	…	gène m	…	gène M	conditions
patient 1	x_1^1	…	x_1^m	…	x_1^M	⋮
⋮						malade
patient i	x_i^1	…	x_i^m	…	x_i^M	⋮
patient i'	$x_{i'}^1$	…	$x_{i'}^m$	…	$x_{i'}^M$	⋮
⋮						sain
patient n	x_n^1	…	x_n^m	…	x_n^M	⋮

Table 1.1 – *Exemple de tableau de données associé à une problématique génomique. Dans ce domaine, le nombre d'individus (ici les patients) est généralement très inférieur au nombre de caractéristiques (les gènes) : $n \ll M$.*

Ces mesures peuvent ensuite être utilisées pour étudier des réseaux génétiques, c'est à dire les interactions qui existent entre les différents gènes au sein d'une cellule. Dans ce dernier cas, l'hypothèse admise est que des gènes qui partagent une même fonction biologique sont corégulés.

Les biologistes recherchent alors, sur les nombreuses mesures effectuées, des groupes de gènes qui évoluent de façon similaire, ou au contraire de façon significativement différente. Pour découvrir ces (ir)régularités entre les différents gènes, on peut utiliser des méthodes issues de l'apprentissage statistique.

Afin de pouvoir traiter ce type de problématique, dans lesquelles le nombre d'exemples est très faible devant le nombre de caractéristiques,

1. Plus précisemment, il est issu de notre participation au projet ANR GD2GS : « Genomic Data to Graph Structure, coordonné par Florence d'Alché-Buc. Le site web du projet peut-être consulté à l'adresse : `http://gd2gs.ibisc.univ-evry.fr`.

nous recherchons des méthodes statistiques qui tiennent compte de deux éléments :

1. ***La cohérence des résultats***.
   Pour obtenir des résultats cohérents avec l'état actuel des connaissances, les méthodes statistiques doivent intégrer le savoir des biologistes. Les hiérarchies disponibles dans *Gene Ontology* [Ashburner et coll., 2000] nous permettent de définir préalablement des groupes de gènes partageant un processus biologique ou une activité moléculaire similaire. Ainsi, les solutions fournies par la méthode statistique doivent être contraintes par ces connaissances *a priori*.
2. ***L'interprétabilité des résultats***.
   Si la performance d'un algorithme d'apprentissage statistique est souvent mesurée par sa capacité de généralisation [2], les biologistes sont plus intéressés par une liste de gènes qui permettent d'expliquer en quoi deux conditions expérimentales diffèrent. La taille de cette liste doit rester « raisonnable ». En effet, l'analyse d'un réseau de quelques dizaines de gènes dont les fonctions sont mal connues peut prendre plusieurs mois, voire plusieurs années... De plus, les biologistes font l'hypothèse que peu de processus interviennent dans la différentiation des conditions expérimentales. Ainsi, la méthode statistique doit pouvoir identifier les groupes de caractéristiques associés aux processus pertinents.

Les problématiques liées au domaine génomique sont à l'origine de nos travaux. Néanmoins, ce mémoire ne rapporte pas de résultats relatifs à de telles applications, pour des raisons qui seront abordées dans la dernière section de ce chapitre. On retrouve cependant les éléments de cette problématique dans d'autres domaines applicatifs, comme celui des interfaces cerveau-machine que nous avons traité dans cette thèse.

1.2 Contributions

Nous avons proposé une méthode permettant d'intégrer, dans l'apprentissage d'un modèle statistique, une connaissance issue du domaine d'application considéré. Cette connaissance est relative à la façon dont les caractéristiques d'un problème sont organisées. De manière générale, nous nous intéressons aux problèmes dont les caractéristiques peuvent être structurées de manière hiérarchique. Dans le cadre de ces travaux, ces hiérarchies sont représentées par des arborescences à deux niveaux, comme celle illustrée sur la figure 1.1. Le second niveau contient les différentes caractéristiques associées au problème, alors que le premier niveau permet d'identifier les groupes sur ces caractéristiques.

Pour un problème d'apprentissage statistique donné, notre but est donc de faire émerger les groupes de caractéristiques pertinents, mais également

2. C'est-à-dire sa capacité à prédire correctement un évènement pour des exemples inédits.

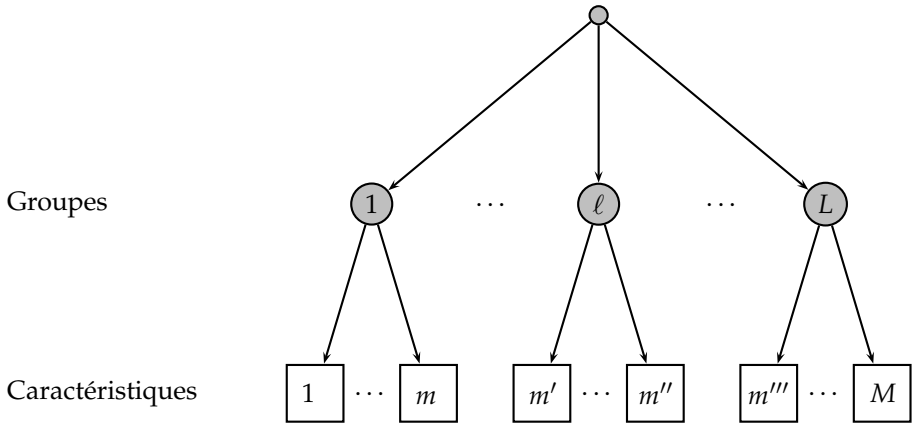


FIGURE 1.1 – *Exemple d'arborescence à deux niveaux sur les caractéristiques. M caractéristiques sont organisées selon L groupes.*

les caractéristiques significatives associées à ces groupes. Pour cela, nous utilisons une formulation variationnelle de type pénalisation adaptative [Grandvalet, 1998]. Nous associons à chaque branche de l'arborescence un facteur de pertinence, et nous imposons une contrainte sur la norme des facteurs d'un même niveau. En fonction du problème considéré, on peut régler sur chaque niveau la sévérité de cette contrainte, et aboutir à des modèles parcimonieux [3] et stables.

Dans un premier temps, nous avons focalisé sur des problèmes de régression paramétrique [Szafranski et coll., 2007, 2008a]. Nous avons ensuite étendu cette première version à des problèmes de classification, dans le cadre de fonctions noyaux [Szafranski et coll., 2008b,c]. Dans cette thèse, nous traçons de nouveaux liens entre ces deux approches.

1.3 Organisation du document

Ce document est organisé en quatre autres chapitres. Le chapitre 2 expose brièvement les concepts de l'apprentissage statistique, ce afin de fixer le vocabulaire et les notations. La régression linéaire et les séparateurs à vaste marge, modèles régulièrement utilisés au fil du document, sont également décrits.

Le chapitre 3 présente un état de l'art relatif aux méthodes parcimonieuses. Nous y répertorions plusieurs critères, que nous regroupons en deux catégories : les régularisations ℓ_p d'une part, et les régularisations structurées d'autre part. Cette dernière catégorie inclut les normes mixtes et celles utilisées pour l'apprentissage de noyaux dans les travaux de Lanckriet et coll. [2004] et de ceux qui ont suivi. Enfin, nous terminons ce chapitre en dressant une typologie des algorithmes d'optimisation associés à ces problèmes, dont nous décrivons les principes généraux.

Dans le chapitre 4, nous développons le processus permettant d'aboutir à la *pénalisation hiérarchique*. Nous établissons les propriétés associées

3. C'est-à-dire des modèles où peu de caractéristiques interviennent.

à ce problème : la relation entre la formulation variationnelle initiale et les normes mixtes, ainsi que les conditions de convexité et de parcimonie. Nous détaillons ensuite les deux algorithmes de résolution que nous avons implémentés. Enfin, nous décrivons plusieurs perspectives visant à étendre le cadre de notre travail à des situations plus complexes.

Le chapitre 5 est consacré à la partie expérimentale. Par le biais de simulations, nous examinons le comportement de la *pénalisation hiérarchique* dans différentes situations. Notre méthode est également appliquée à deux problèmes relatifs aux interfaces cerveau-machine :

1. Dans le premier, nous avons réutilisé les données d'une compétition sur les interfaces cerveau-machine, sur lesquelles Alain Rakotomamonjy [4], professeur à l'université de Rouen, avait déjà eu l'occasion de travailler.
2. Dans le second, Gangadhar Garipelli [5], doctorant à l'Idiap sous la direction de José del R. Millán, nous a transmis ses données. Elles sont issues de ses recherches dans le domaine des interfaces cerveau-machine sur la reconnaissance d'un état reflétant l'anticipation d'un évènement.

Nous détaillons les protocoles de ces deux problèmes, et reportons les résultats associés en terme de performance mais aussi d'interprétation. Bien que l'exemple illustratif relatif à la génomique soit à l'origine de nos motivations, il reste un travail important à effectuer pour adapter notre méthode aux ontologies de gènes [6].

Enfin, nous concluons en récapitulant les différents points importants évoqués dans ce document. Nous mettons également en exergue nos contributions actuelles, et celles en perspective.

4. `http://asi.insa-rouen.fr/enseignants/~arakotom/`
5. `http://www.idiap.ch/~ggaripe/`
6. Ce travail est esquissé dans les perspectives du chapitre 4.

Généralités sur l'apprentissage statistique

2

Sommaire

2.1 Introduction

Ce chapitre a pour but de présenter très brièvement les concepts de l'apprentissage statistique, et d'introduire le vocabulaire et les notations utilisés dans cette thèse. Tout d'abord, nous définissons les notions principales. Nous décrivons ensuite deux modèles, la régression linéaire et les séparateurs à vaste marge, qui seront utilisés par la suite dans nos travaux.

2.2 Formalisme

L'apprentissage statistique regroupe un ensemble de méthodes qui vise à analyser, interpréter, voire prédire un phénomène. Le processus d'apprentissage s'effectue au travers d'objets observés sur un ensemble de M attributs, appelés ***variables explicatives***.

Plus formellement, on représentera ces objets appelés ***observations***, ***exemples*** ou encore ***individus***, sous la forme d'un vecteur de dimension M :

$$\boldsymbol{x} = (x^1, \ldots, x^m, \ldots, x^M) \in \mathcal{X} \ ,$$

où x^m représente l'observation associée à la variable m, et où $\mathcal{X}$ définit l'espace des variables, par exemple $\mathbb{R}^M$. L'apprentissage statistique est divisé en deux catégories : l'apprentissage ***non supervisé*** et l'apprentissage ***supervisé***.

2.2.1 Apprentissage non supervisé

Le but de l'apprentissage non supervisé est de déterminer automatiquement des catégories sur l'ensemble des observations. Dans ce processus de classification automatique, on cherche à regrouper les exemples en fonction de leur ressemblance, dans un espace donné qui peut être une transformation de l'espace initial des variables. L'enjeu consiste à trouver une mesure de ressemblance et un critère de classification satisfaisants.

Dans la mesure où nos travaux se situent dans le cadre de l'apprentissage supervisé, nous nous tiendrons à ce résumé incomplet de la classification automatique. Le lecteur peut se reporter à [Celeux et coll., 1989] pour un état détaillé de cette problématique.

2.2.2 Apprentissage supervisé

En apprentissage supervisé, on dispose d'une information supplémentaire : à chaque observation $\boldsymbol{x}$ est associée une ***étiquette*** y, appelée aussi ***réponse*** ou ***variable expliquée***. Les couples $(\boldsymbol{x}, y)$ sont générés selon une loi de probabilité inconnue $\mathbb{P}(\mathrm{X}, \mathrm{Y})$, où X est le vecteur aléatoire associé aux observations, et Y la variable aléatoire associée aux réponses. Dans cette thèse, nous considérerons que ces couples sont indépendants et identiquement distribués (i.i.d.) selon la loi de probabilité $\mathbb{P}(\mathrm{X}, \mathrm{Y})$.

Définition 2.1 ***Ensemble d'apprentissage*** *— L'ensemble d'apprentissage* $\boldsymbol{S}$ *est défini comme l'ensemble des couples d'observations et de réponses dont on dispose pour analyser un phénomène :* $\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ *, où* $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$*,* $\forall i$*.*

En fonction de l'espace $\mathcal{Y}$, on parlera de :

- régression, si $y \in \mathbb{R}$;
- classification ou discrimination, si $y \in \{1, \ldots, C\}$, avec $C \in \mathbb{N}$;
- classification ou discrimination binaire, si $C = 2$ [1].

Remarque 2.1 *—* On notera $\mathbf{X}$ la matrice associée aux observations de l'ensemble d'apprentissage $\boldsymbol{S}$, avec $\mathbf{X} \in \mathcal{M}^{n \times M}$, où M représente la dimension de $\mathcal{X}$. On notera $\mathbf{y}$ le vecteur associé aux réponses de l'ensemble d'apprentissage $\boldsymbol{S}$, avec $\mathbf{y} \in \mathcal{Y}^n$. ◇

Remarque 2.2 *—* Pour alléger les notations, les variations des indices i et j, représentant les observations de l'ensemble d'apprentissage $\boldsymbol{S}$, seront occultées des sommes, de même que celle des indices m représentant les variables ou les caractéristiques [2]. Le cas échéant, les indices i et j varieront de 1 à n, tandis que l'indice m variera de 1 à M. ◇

2.2.3 Minimisation du risque empirique

À partir de l'ensemble d'apprentissage $\boldsymbol{S}$, on souhaite inférer une relation entre les observations et la réponse, pour tout élément issu de la loi $\mathbb{P}(\mathrm{X}, \mathrm{Y})$. On recherche alors une fonction f capable de caractériser au mieux le lien entre les observations $\boldsymbol{x}_i$ et les réponses y_i. Pour quantifier l'erreur commise par f pour évaluer y, on définit une ***fonction de perte***

$$L \ : \ \mathcal{Y} \times \mathcal{Y} \ \rightarrow \ \mathbb{R}^+ \, .$$

On souhaite minimiser l'espérance de la fonction de perte sur l'ensemble du phénomène. Ainsi, la notion de ***risque*** est définie par

$$R(f) \ = \ \mathbb{E}\left[L\left(\mathrm{Y}, f(\mathrm{X})\right)\right] \ = \ \int L(y, f(\boldsymbol{x}))\, \mathbb{P}(\boldsymbol{x}, y)\, d\boldsymbol{x} dy \, .$$

Cependant, $R(f)$ ne peut pas être calculé puisque la loi de probabilité $\mathbb{P}(\mathrm{X}, \mathrm{Y})$ est inconnue. On définit alors le ***risque empirique*** comme la moyenne de la perte L sur $\boldsymbol{S}$:

$$R_{emp}(f) \ = \ \frac{1}{n} \sum_i L(y_i, f(\boldsymbol{x}_i)) \, .$$

Remarque 2.3 *—* On appelle ***coût*** ou ***critère***, les expressions de type $J(f) = g(L(y, f(\boldsymbol{x})))$. Par abus de langage, on désignera également $J(f)$ par le terme ***fonction de perte***. ◇

Un processus d'apprentissage consistant sous certaines conditions vise à rechercher l'estimateur f qui minimise ce risque empirique. La théorie statistique de l'apprentissage [Vapnik, 1995] étudie les conditions de

1. En particulier, on peut considérer que $y \in \{\pm 1\}$.
2. Le terme « caractéristiques » sera défini en section 2.3.2.

consistance et de convergence de la minimisation du risque empirique. En termes réducteurs, il s'agit de savoir si minimiser R_{emp} est une stratégie acceptable, permettant de minimiser R pour un échantillon de taille infinie. Il faut de plus qu'elle soit applicable, ce qui nécessite qu'une forme de convergence de R_{emp} vers R puisse être établie pour des échantillons de taille finie.

La dimension de Vapnik-Chervonenkis permet de majorer l'écart entre R et R_{emp} en fonction de la perte L, de la taille n de l'ensemble d'apprentissage, mais aussi de l'ensemble de fonctions $\mathcal{H}$ auquel l'estimateur f appartient. L'existence de telles bornes conduit à envisager des stratégies d'apprentissage qui ne sont pas uniquement basées sur la minimisation du risque empirique, mais qui tiennent également compte de la « complexité » du modèle. Cet aspect est développé dans la section suivante.

2.2.4 Contrôle de la complexité

La complexité est définie par l'ensemble des termes qui interviennent dans R et R_{emp} : la loi de (X, Y), la perte L, la taille n de l'échantillon d'apprentissage. Elle dépend également d'une variable qui n'a pas été précisément mentionnée jusqu'à présent : la taille de l'ensemble de fonctions $\mathcal{H}$ dans lequel l'estimateur f est recherché.

Comme la distribution de (X, Y), la perte L et la taille n de l'ensemble d'apprentissage sont des données du problème, le seul point de contrôle est la taille de $\mathcal{H}$. Si $\mathcal{H}$ est un ensemble fini de fonctions, sa taille est le nombre de fonctions essentiellement différentes ; si $\mathcal{H}$ est un espace vectoriel, sa taille peut être mesurée par la dimension.

La complexité fait intervenir la distribution de (X, Y), et n'est pas calculable. Il est donc pratique de choisir une structure d'espaces emboîtés $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_\lambda \subset \cdots \subset \mathcal{H}_p$, pour laquelle on sait que :

1. La complexité est une fonction croissante de la taille, quelle que soit la distribution de (X, Y).
2. Le risque empirique $R_{emp}(f)$ est une fonction décroissante de la taille.

Pour chaque ensemble $\mathcal{H}_\lambda$, la minimisation de R_{emp} fournit un estimateur f_λ. Pour $\lambda < \lambda'$, on a $\mathcal{H}_\lambda \subset \mathcal{H}_{\lambda'}$, ce qui d'après l'hypothèse 2 implique que $R_{emp}(f_\lambda) \geq R_{emp}(f_{\lambda'})$. D'autre part, lorsque $\mathcal{H}_\lambda \subset \mathcal{H}_{\lambda'}$, la borne sur l'écart entre R et R_{emp} augmente. Cette borne s'applique sur toutes les fonctions de $\mathcal{H}_\lambda$, donc en particulier sur f_λ.

La borne sur le risque est donc la somme de deux fonctions de λ, l'une décroissante ($R_{emp}(f_\lambda)$) et l'autre croissante (l'écart maximum sur $R(f_\lambda) - R_{emp}(f_\lambda)$). Trouver un compromis entre ces deux fonctions est donc nécessaire pour minimiser la borne sur le risque. On aboutit ainsi à un schéma d'apprentissage composé de trois étapes :

1. Définition d'une famille de modèle structurée : $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_p$.
2. Estimation de f_λ : $\forall \lambda = \{1, \ldots, p\}$, $f_\lambda = \arg\min_{f \in \mathcal{H}_\lambda} R_{emp}(f)$.

3. Estimation de $R(f_\lambda)$ (ou d'une borne sur $R(f_\lambda)$) et sélection de l'estimateur $f_\lambda^\star$ qui le (la) minimise.

Ces trois étapes forment ce que Vapnik appelle la minimisation du risque structurel.

Dans cette thèse, nous définirons une famille de modèles emboîtés par le biais d'un terme de pénalisation chargé de contraindre les solutions à respecter un *a priori*. Nous développerons alors des algorithmes d'apprentissage spécifiques à la minimisation du risque empirique sur cette famille de modèles. Nos contributions concernent donc les deux premières étapes de la minimisation du risque structurel. Pour la dernière, nous avons eu recours à la validation croisée, une procédure dont le principe est décrit dans tous les manuels traitants de l'apprentissage statistique. Nous ne le rappellerons donc pas dans ce document.

2.3 Modèles de référence

2.3.1 Régression linéaire

Description du problème

À partir de $\mathbf{S}$, on souhaite établir un lien entre les réponses $\mathbf{y} \in \mathbb{R}^n$ et les observations $\mathbf{X}$, à l'aide d'une combinaison ***linéaire*** de variables. On pose le modèle :

$$\begin{aligned} y &= f(\boldsymbol{x}) + \epsilon \\ &= \beta_0 + \sum_m x^m \beta_m + \epsilon \,, \end{aligned}$$

où ϵ est une composante aléatoire représentant l'influence des variables non observées sur les réponses.

Remarque 2.4 — Dans le cadre de cette thèse, nous faisons l'hypothèse que les observations $\boldsymbol{x}_i$ sont centrées et réduites :

$$\forall m,\ \frac{1}{n}\sum_i x_i^m = 0 \,,\ \frac{1}{n}\sum_i (x_i^m)^2 = 1 \,.$$

En effet, les méthodes de régression régularisées, présentées dans le prochain chapitre, ne sont pas invariantes aux échelles associées aux variables. Il paraît donc raisonnable de les normaliser. De plus, nous considérons également que les réponses y_i sont centrées :

$$\frac{1}{n}\sum_i y_i = 0 \,,$$

ce qui permet de considérer que $\beta_0 = 0$. Finalement, le modèle s'écrit

$$y = \boldsymbol{x}\boldsymbol{\beta} + \epsilon \,,$$

où $\boldsymbol{\beta} \in \mathcal{X}$ est le vecteur des composantes appliquées aux variables explicatives des observations $\boldsymbol{x}$. En effet, une observation $\boldsymbol{x}$ est représentée par un vecteur ligne, tandis que $\boldsymbol{\beta}$ est représenté par un vecteur colonne. Ainsi $\boldsymbol{x}\boldsymbol{\beta}$ désigne le produit scalaire entre $\boldsymbol{x}$ et $\boldsymbol{\beta}$. ◇

Formalisation

Le but est de trouver une estimation $\boldsymbol{\beta}$ qui minimise le critère des *moindres carrés* [3], c'est-à-dire la somme des différences au carré entre les réponses réelles et les réponses estimées :

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad J(\boldsymbol{\beta}) &= \sum_i L(y_i, \boldsymbol{x}_i\boldsymbol{\beta}) \\ &= \sum_i (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad . \end{aligned} \tag{2.1}$$

La solution minimisant la perte quadratique (2.1) sous forme matricielle, au sens des *moindres carrés*, consiste à trouver les coefficients du vecteur $\boldsymbol{\beta}$, tel que

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \qquad \Leftrightarrow \qquad \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \quad \text{lorsque } n \geq M\,. \tag{2.2}$$

On peut utiliser la solution obtenue dans un but prédictif ou descriptif. Dans un but prédictif, $\boldsymbol{\beta}$ permet d'obtenir $\hat{y}$, la réponse associée à n'importe quelle observation $\boldsymbol{x}$: $\hat{y} = \boldsymbol{x}\boldsymbol{\beta}$. Dans un but descriptif, $|\beta_m|$ peut-être interprétée comme le degré de pertinence de la variable x^m.

2.3.2 Séparateurs à Vaste Marge

Description du problème

Les Séparateurs à Vaste Marge [4] (SVM) [Vapnik, 1995] font partie des classifieurs dits « à noyaux ». Ils sont communément présentés dans le cadre de la classification binaire : $\forall i,\, y_i \in \{\pm 1\}$, mais peuvent être étendus à des problèmes de plus de deux classes, ou à des problèmes de régression.

Bien souvent, les exemples d'un ensemble d'apprentissage ne peuvent pas être séparés linéairement dans $\mathcal{X}$. Dans les méthodes à noyaux, on considère la transformation de l'espace des variables $\mathcal{X}$ en un espace de ***caractéristiques*** [5], ou d'***hypothèses*** $\mathcal{H}_c$, par une application non linéaire :

$$\begin{aligned} \phi : \quad & \mathcal{X} \to \mathcal{H}_c \\ & \boldsymbol{x} \mapsto \phi(\boldsymbol{x})\,, \end{aligned}$$

où la dimension de $\mathcal{H}_c$ est généralement supérieure à celle de $\mathcal{X}$. De plus, on considère des espaces $\mathcal{H}_c$ dotés d'un produit scalaire $\langle .,. \rangle_{\mathcal{H}_c}$.

Le principe des SVM est de construire une règle de décision permettant de séparer au mieux les exemples positifs des exemples négatifs. Pour un exemple $\boldsymbol{x}$, cette règle de décision est de la forme

$$d(\boldsymbol{x}) = \operatorname{sign}\left(\sum_i \alpha_i \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}_i) \rangle_{\mathcal{H}_c} + b \right)\,, \tag{2.3}$$

3. *Ordinary Least Squares* (OLS), en anglais.
4. Support Vector Machines, en anglais.
5. Feature space, en anglais.

avec $\forall i,\ \boldsymbol{x}_i \in \boldsymbol{S}$, et où $\{\alpha_i\}$ et b sont les paramètres à optimiser. Avant de présenter le problème d'optimisation lié aux SVM, nous allons voir comment construire l'espace des caractéristiques $\mathcal{H}_c$ à partir de fonctions noyaux.

Cadre fonctionnel

On montre ici que certains noyaux permettent de définir implicitement un espace de caractéristiques. Pour une présentation plus détaillée de ce cadre fonctionnel et des SVM, le lecteur peut se référer à [Schölkopf et Smola, 2001]. On commence par introduire les quatre définitions suivantes :

Définition 2.2 ***Noyau*** *— Un noyau est une mesure de similarité entre deux observations :*

$$\begin{aligned} K: \quad & \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ & (\boldsymbol{x}, \boldsymbol{x}') \ \mapsto K(\boldsymbol{x}, \boldsymbol{x}') \ . \end{aligned}$$

Définition 2.3 ***Noyau positif*** *— Un noyau K sur $\mathcal{X}$ est positif s'il est symétrique :*

$$K(\boldsymbol{x}, \boldsymbol{x}') \ = \ K(\boldsymbol{x}', \boldsymbol{x}) \ ,$$

et si pour tout entier positif fini N, il vérifie

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \, K(\boldsymbol{x}_i, \boldsymbol{x}_j) \ \geq \ 0 \ ,$$

où $\forall i = \{1, \ldots, N\}$, $\alpha_i \in \mathbb{R}$ et $\boldsymbol{x}_i \in \mathcal{X}$.

Définition 2.4 ***Espace de Hilbert*** *— Un espace vectoriel $\mathcal{H}$, muni d'un produit scalaire $\langle ., . \rangle_{\mathcal{H}}$, est appelé espace pré-hilbertien. Si de plus, $\mathcal{H}$ muni de la norme $||.||_{\mathcal{H}}$ associée à $\langle ., . \rangle_{\mathcal{H}}$ est complet, $\mathcal{H}$ est un espace de Hilbert.*

Définition 2.5 ***Espace de Hilbert à Noyau Reproduisant (EHNR)*** *— Un espace de Hilbert $\mathcal{H}$ est dit « à noyau reproduisant » s'il existe un noyau K pour lequel, $\forall f \in \mathcal{H}$, et $\forall \boldsymbol{x} \in \mathcal{X}$, $f(\boldsymbol{x})$ peut-être exprimée comme un produit scalaire dans $\mathcal{H}$:*

$$f(\boldsymbol{x}) \ = \ \langle f, K(\boldsymbol{x}, .) \rangle_{\mathcal{H}} \ .$$

Soit N un entier positif fini. Soient $\boldsymbol{x}_i \in \mathcal{X}$, $\forall i = \{1, \ldots, N\}$ et $\boldsymbol{x}'_j \in \mathcal{X}$, $\forall j = \{1, \ldots, N\}$. Soit K, un noyau positif. Soit $\mathcal{H}$, l'ensemble des fonctions définies par les combinaisons linéaires associées au noyau

$$\mathcal{H} \ = \ \left\{ f : \mathcal{X} \to \mathbb{R} \mid f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) \right\} \ ,$$

où $\forall i$, $\alpha_i \in \mathbb{R}$. Soient

$$f \ = \ \sum_{i=1}^{N} \alpha_i \, K(., \boldsymbol{x}_i) \ ,$$

avec $\forall i$, $\alpha_i \in \mathbb{R}$, et $g \in \mathcal{H}$:

$$g \ = \ \sum_{j=1}^{N} \alpha'_j \, K(., \boldsymbol{x}'_j) \ ,$$

avec $\forall j,\ \alpha'_j \in \mathbb{R}$. Pour toutes fonctions f et $g \in \mathcal{H}$, on définit le produit scalaire entre f et g par la forme bilinéaire

$$\langle f, g \rangle_{\mathcal{H}} \ = \ \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha'_j \, K(\boldsymbol{x}_i, \boldsymbol{x}'_j) \ . \tag{2.4}$$

L'espace $\mathcal{H}$ ainsi engendré est un espace pré-hilbertien. Si de plus, on complète $\mathcal{H}$ au sens de $||f||_{\mathcal{H}}$, la norme induite par le produit scalaire, avec $||f||^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}$, alors $\mathcal{H}$ est un espace de Hilbert. On remarque également la propriété suivante :

Propriété 2.1 ***Propriété de reproduction*** *— D'après la définition (2.4) du produit scalaire, on remarque que* $\forall f \in \mathcal{H}$,

$$\langle f, K(\boldsymbol{x}, .) \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{N} \alpha_i \, K(., \boldsymbol{x}_i), K(\boldsymbol{x}, .) \right\rangle_{\mathcal{H}} = f(\boldsymbol{x}) \ .$$

$\mathcal{H}$ est alors appelé Espace de Hilbert à Noyau Reproduisant [6].

En particulier, cette propriété met en évidence la relation

$$\left\langle K(\boldsymbol{x}, .), K(\boldsymbol{x}', .) \right\rangle_{\mathcal{H}} \ = \ K(\boldsymbol{x}, \boldsymbol{x}') \ .$$

Lorsque la transformation ϕ est définie par le biais du noyau K associé à l'EHNR $\mathcal{H}$

$$\begin{aligned} \phi : \ & \mathcal{X} \to \mathcal{H} \\ & \boldsymbol{x} \ \mapsto K(\boldsymbol{x}, .) \ , \end{aligned}$$

on obtient

$$\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}) \rangle_{\mathcal{H}} \ = \ K(\boldsymbol{x}, \boldsymbol{x}') \ .$$

Ainsi, lorsqu'on applique un noyau reproduisant à deux observations issues de l'espace des variables $\mathcal{X}$, on calcule en fait leur produit scalaire dans un espace de caractéristiques défini par l'espace fonctionnel $\mathcal{H}$.

Formalisation

Dans le cadre des SVM, on recherche dans un EHNR $\mathcal{H}$ la fonction f de norme minimale permettant de séparer au mieux les exemples positifs des exemples négatifs de $\boldsymbol{S}$. L'hyperplan séparateur optimal a pour équation $f(\boldsymbol{x}) + b = 0$, où f et b sont appris à partir de $\boldsymbol{S}$.

Dans le cas séparable, lorsqu'un exemple est bien classé, on impose que l'inégalité $y_i \left(f(\boldsymbol{x}_i) + b \right) \geq 1$ soit vérifiée. Le problème à résoudre est

$$\begin{cases} \min\limits_{f, b} & \dfrac{1}{2} \, \|f\|^2_{\mathcal{H}} \\ \text{s. c.} & y_i \left(f(\boldsymbol{x}_i) + b \right) \ \geq \ 1 \qquad\qquad \forall i \ . \end{cases}$$

6. Reproducing Kernel Hilbert Space (RKHS), en anglais.

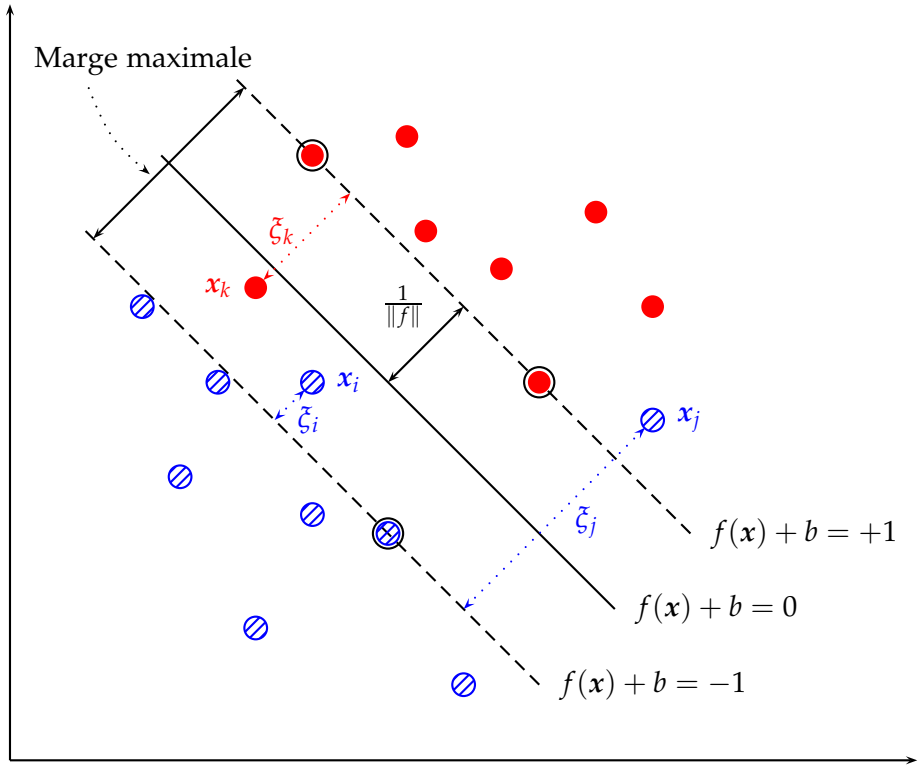


FIGURE 2.1 – *Séparateur à vaste marge dans le cas linéaire non séparable. Les vecteurs supports sont représentés par les exemples encerclés.*

Dans le cas non séparable, on introduit un vecteur de « variables d'écart » $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n) \geq 0$. Le problème d'optimisation devient

$$\begin{cases} \min\limits_{f,b,\boldsymbol{\xi}} & \frac{1}{2}\,\|f\|_{\mathcal{H}}^2 \;+\; C\sum\limits_i \xi_i & & & & (2.5a) \\ \text{s. c.} & y_i\,(f(\boldsymbol{x}_i)+b) \;\geq\; 1-\xi_i & \xi_i \;\geq\; 0 & \forall i\,, & & (2.5b) \end{cases}$$

où $C > 0$ est le paramètre qui règle le compromis entre la largeur de la marge, représentée sur la figure 2.1, et la quantité d'erreur qu'on s'autorise à commettre. On peut résumer la formulation (2.5) par

$$\min_{f,b} \qquad \frac{1}{2}\,\|f\|_{\mathcal{H}}^2 \;+\; C\sum_i [1 - y_i\,(f(\boldsymbol{x}_i)+b)]_+\;,$$

où $[u]_+ = \max(0, u)$. Cette formulation fait apparaître le coût charnière $[1 - y_i\,(f(\boldsymbol{x}_i)+b)]_+$, qui est la fonction de perte associée aux SVM.

En dérivant le Lagrangien associé au problème (2.5), on obtient la forme duale

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} & \sum\limits_i \alpha_i - \frac{1}{2}\sum\limits_i\sum\limits_j \alpha_i\alpha_j\, y_i y_j\, K(\boldsymbol{x}_i, \boldsymbol{x}_j) & \\ \text{s. c.} & \sum\limits_i \alpha_i y_i \;=\; 0 & \\ & 0 \;\leq\; \alpha_i \;\leq\; C & \forall i\,. \end{cases}$$

On peut interpréter cette dernière formulation en fonction des multiplicateurs de Lagrange α_i associés à la contrainte (2.5b) du problème primal :

1. Lorsque $\alpha_i = 0$, la contrainte (2.5b) n'est pas active et la variable d'écart ξ_i est nulle : on a $y_i\,(f(\boldsymbol{x}_i) + b) \geq 1$. Les exemples associés à ces multiplicateurs se situent au delà des hyperplans qui définissent la marge. Ces hyperplans sont représentés en pointillés, sur la figure 2.1, par les droites d'équations $f(\boldsymbol{x}_i) + b = \pm 1$.
2. Lorsque $0 < \alpha_i < C$, la contrainte (2.5b) est active et la variable d'écart ξ_i est nulle : on a $y_i\,(f(\boldsymbol{x}_i) + b) = 1$. Les exemples associés à ces multiplicateurs sont situés sur la frontière des hyperplans définissant la marge. Ce sont les ***vecteurs supports***, représentés entourés sur la figure 2.1.
3. Lorsque $\alpha_i = C$, la contrainte (2.5b) est active et la variable d'écart ξ_i est positive : on a $y_i\,(f(\boldsymbol{x}_i) + b) < 1$. Les exemples associés à ces multiplicateurs se situent du mauvais côté des hyperplans qui définissent la marge. Ils sont soit bien classés à l'intérieur de la marge (exemple $\boldsymbol{x}_i$ sur la figure 2.1), soit mal classés (exemples $\boldsymbol{x}_j$ et $\boldsymbol{x}_k$).

Finalement, seuls les exemples pour lesquels les multiplicateurs α_i ne sont pas nuls, ont une influence dans la solution finale. Les SVM sont parcimonieux dans l'espace des observations.

2.4 Synthèse

Dans ce chapitre, nous avons brièvement introduit les concepts de l'apprentissage supervisé, et précisé le vocabulaire et les notations que nous utiliserons dans la suite de ce manuscrit. Nous avons également décrit la régression linéaire et les SVM. C'est par le biais de ces deux outils que nous allons maintenant présenter un état de l'art sur les méthodes régularisées.

Régularisation et parcimonie 3

Sommaire

3.1 Introduction

Le problème de *sélection de variables* consiste à identifier l'ensemble des variables caractérisant au mieux le phénomène observé, ce dans une perspective double.

1. La première concerne l'amélioration des performances en prédiction. En effet, si certaines variables observées n'ont pas d'influence sur la réponse, le fait de les inclure dans le modèle peut perturber le processus d'estimation.
2. La seconde a trait à l'interprétation du modèle. Depuis plusieurs décennies, les capacités d'acquisition permettent d'observer jusqu'à plusieurs milliers de variables. L'interprétation « humaine » devient alors délicate. Pour faciliter la compréhension du modèle, des procédures « automatiques » deviennent nécessaires.

Il existe différentes façons de sélectionner un ensemble de variables en apprentissage statistique [Guyon et Elisseeff, 2003]. Nous nous intéressons ici aux méthodes régularisées, qui consistent à introduire, dans un problème d'apprentissage statistique, une pénalisation sur la norme des composantes d'un estimateur [1].

Dans un premier temps, nous détaillons les méthodes régularisées par le biais de normes ℓ_p. Nous caractérisons leurs propriétés, notamment en termes de convexité, de parcimonie et de stabilité. Pour cela, nous insistons sur les régularisations « usuelles », à savoir les régularisations ℓ_0 , ℓ_2 et ℓ_1. Nous nous attardons également sur deux variantes pondérées de ces régularisations : l'*adaptive ridge* et l'*adaptive lasso*.

Nous présentons ensuite les méthodes de régularisation structurées. Nous rassemblons dans cette catégorie des méthodes élaborées pour des problèmes dans lesquels une structure sur les variables existe. Ici, cette structure est relative à l'organisation des variables en groupes. L'*elastic net* favorise la découverte et la sélection de groupes de variables corrélées. Lorsque la structure est connue *a priori*, les pénalités composites et les normes mixtes ont pour objectif d'identifier les éléments les plus significatifs de chaque groupe. Enfin, nous décrivons le Multiple Kernel Learning, qui est également apparenté aux méthodes de régularisation structurées. Néanmoins, contrairement à l'ensemble des méthodes précédentes qui sont définies pour des fonctions paramétriques, le MKL considère le cadre fonctionnel des noyaux.

Pour terminer, nous décrivons les principes généraux d'algorithmes d'optimisation permettant de résoudre les problèmes suscités. Nous les regroupons en trois catégories : les algorithmes de seuillage itératif, de contraintes actives, et d'approximation par plans sécants [2]. Enfin, nous décrivons la stratégie dite de « chemin de régularisation », qui permet de parcourir l'ensemble des solutions atteignables quand le paramètre de régularisation varie.

1. Hesterberg et coll. [2008] ont aussi établi une revue sur certaines méthodes régularisées décrites dans ce chapitre, dans le cadre de la régression linéaire.
2. Cutting planes, en anglais.

3.2 Régularisations ℓ_p

3.2.1 Contexte

La notion de problème « bien posé » a été définie par le mathématicien Jacques Hadamard. Selon lui, les modèles mathématiques de phénomènes physiques doivent garantir :

– l'existence de la solution,
– l'unicité et
– la stabilité.

Un problème qui ne vérifie pas une de ces propriétés est dit « mal posé ». Les problèmes inverses sont souvent mal posés. C'est le cas lorsqu'on essaie d'inférer des lois générales à partir de seulement quelques exemples, en particulier lorsque la taille de l'échantillon est inférieure à la dimension du problème ($n < M$).

La régularisation consiste à transformer un problème mal posé en un problème bien posé. Pour ce faire, on peut intégrer une connaissance *a priori* sur la solution par le biais d'un terme de régularisation ou de pénalisation. Le critère minimisé est alors de la forme :

$$\min_{f \in \mathcal{H}} \quad J(f) \; + \; \lambda \, P(f) \; , \tag{3.1}$$

où le paramètre de régularisation $\lambda \in \mathbb{R}^+$ contrôle le compromis entre l'adéquation du modèle aux données $J(f)$ et l'opérateur de pénalisation $P(f)$, et où $\mathcal{H}$ définit l'espace de fonction auquel f appartient.

La formulation (3.1) peut se récrire sous la forme équivalente

$$\begin{cases} \min\limits_{f \in \mathcal{H}} & J(f) & \text{(3.2a)} \\ \text{s. c.} & P(f) \; \leq \; t \; , & \text{(3.2b)} \end{cases}$$

où $t \in \mathbb{R}^+$ est un seuil qui joue un rôle similaire au paramètre de régularisation λ. Lorsque le critère $J(f)$ et l'opérateur de pénalisation $P(f)$ sont convexes, il existe, pour chaque valeur de λ, une valeur de t correspondante, telle que les deux problèmes ont la même solution.

***Remarque* 3.1 —** Le problème (3.1) correspond à la formulation Lagrangienne du problème (3.2). La formulation (3.1) peut également être désignée sous le terme de formulation de Tikhonov, par contraste avec la formulation (3.2) dite d'Ivanov.

◇

Les opérateurs de régularisation auxquels on s'intéresse dans ce document évaluent les normes ℓ_p des estimations. Dans le cadre paramétrique, où $f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$ et où $\boldsymbol{\beta} \in \mathcal{X}$, on définit l'opérateur de régularisation $P(\boldsymbol{\beta})$ par une (quasi-)norme ℓ_p , avec $0 \leq p < \infty$:

$$\|\boldsymbol{\beta}\|_p \quad = \quad \left(\sum_m |\beta_m|^p \right)^{1/p} \; ,$$

pour $p \neq 0$. Lorsque $p = 0$, $P(\boldsymbol{\beta})$ est défini par

$$\|\boldsymbol{\beta}\|_0 \quad = \quad \text{card}\{\beta_m | \beta_m \neq 0\} \ .$$

Enfin, lorsque $p \to \infty$, $P(\boldsymbol{\beta})$ est défini par

$$\|\boldsymbol{\beta}\|_\infty \quad = \quad \max_m \ |\beta_m| \ .$$

***Remarque* 3.2 —** Par abus de langage, les quasi-normes ℓ_p, où $0 \leq p < 1$, seront également appelées normes ℓ_p. ◇

Dans le cadre paramétrique, lorsque $J(\boldsymbol{\beta})$ est une fonction de perte quadratique, le problème (3.1) est connu sous le nom de *bridge regression* [Frank et Friedman, 1993].

3.2.2 Propriétés

Avant de présenter les régularisations les plus courantes, nous allons étudier les propriétés de convexité, de parcimonie et de stabilité des problèmes régularisés par des normes ℓ_p, dans le cadre paramétrique.

***Remarque* 3.3 —** Afin de rendre la lecture de cette section plus fluide, nous raisonnerons uniquement sur la formulation (3.2). Néanmoins, les propositions et les remarques associées s'appliquent de la même façon à la formulation (3.1). ◇

Convexité

Dans un premier temps, introduisons les notions d'ensemble et de fonction convexes [Boyd et Vandenberghe, 2004, chapitre 2 et 3].

Définition 3.1 ***Ensemble convexe* —** *Soit $C \subseteq \mathbb{R}^M$. Un ensemble C est dit convexe si $\forall\, (\boldsymbol{x}_1, \boldsymbol{x}_2) \in C \times C$, et $\theta \in [0, 1]$*

$$\theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \in C \ .$$

Définition 3.2 ***Fonction convexe* —** *Une fonction $f : C \to \mathbb{R}$, définie sur un ensemble convexe C, est dite convexe si pour $\forall\, (\boldsymbol{x}_1, \boldsymbol{x}_2) \in C \times C$, et $\theta \in [0, 1]$*

$$f(\theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2) \ \leq \ \theta f(\boldsymbol{x}_1) + (1 - \theta) f(\boldsymbol{x}_2) \ .$$

***Remarque* 3.4 —** La stricte convexité est obtenue en remplaçant les inégalités par des inégalités strictes, dans les définitions 3.1 et 3.2. ◇

Propriété 3.1 *Le problème (3.2) est convexe si les deux propriétés suivantes sont vérifiées :*

1. *$J(\boldsymbol{\beta})$ est une fonction de perte convexe.*
2. *$P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_p$ est convexe, c'est à dire $p \geq 1$.*

Afin d'illustrer cette propriété, nous représentons sur la figure 3.1 les ensembles admissibles associés à différentes normes, pour un exemple en deux dimensions. On constate que les régions grisées, définies par la contrainte (3.2b), sont strictement convexes pour $p \geq 1$ et concaves sinon.

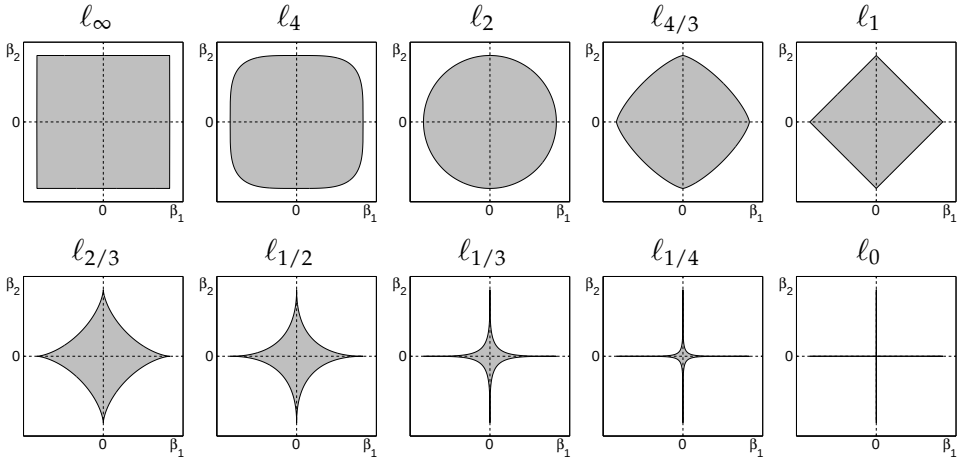


FIGURE 3.1 – *Boules unité pour différentes normes :* $\|\boldsymbol{\beta}\|_p \leq 1$.

Parcimonie

Dans les modèles paramétriques, la valeur absolue des composantes de $\boldsymbol{\beta}$ peut représenter le degré de pertinence d'une variable dans l'interprétation d'un modèle. Plus $|\beta_m|$ est élevée, plus la variable $\boldsymbol{x}^m$ participe à l'explication de la réponse y. Inversement, plus $|\beta_m|$ est proche de 0, moins la variable $\boldsymbol{x}^m$ est pertinente. L'objectif consiste alors à assigner une valeur nulle aux coefficients de la solution associés aux variables non significatives, qui sont ainsi éliminées du modèle.

Définition 3.3 ***Solution parcimonieuse*** *— Une solution $\boldsymbol{\beta}$ est dite parcimonieuse si plusieurs de ses composantes sont nulles.*

Propriété 3.2 *La solution du problème (3.2) est parcimonieuse si les coefficients de l'estimateur sont contraints par le biais d'une norme ℓ_p, où $p \leq 1$. La non-différentiabilté des contours des normes ℓ_p en $\beta_m = 0$ entraîne la parcimonie, lorsque $p \leq 1$ [Nikolova, 2000].*

Les propriétés 3.1 et 3.2 nous indiquent que le problème (3.2) est convexe et la solution associée parcimonieuse, si et seulement si on contraint la norme ℓ_1 de l'estimateur : $P(\beta) = \|\boldsymbol{\beta}\|_1$.

Sur la figure 3.1, les régions définies par le biais d'une contrainte ℓ_p comportent des « coins » situés sur les axes, lorsque $p < 2$. Dans ce cas, les coefficients de la solution contrainte peuvent être annulés. Si la solution des *moindres carrés* est hors du domaine admissible, celle du problème pénalisé (3.2) se situe sur la surface du domaine. Plus exactement, elle se situe sur un point de la surface pour lequel le gradient de $J(\boldsymbol{\beta})$ appartient à la normale de la surface du domaine. Pour $p < 2$, les axes correspondent à des discontinuités de la tangente à la surface, pour lesquels la normale est un cône, alors qu'elle est réduite à un vecteur unique sur les points de continuité. Dès lors, il est clair que pour de nombreuses distributions sur (X, Y), avoir une solution qui annule un ou plusieurs coefficients est un évènement de probabilité non-nulle.

Cette intuition est illustrée sur la figure 3.2, où $\boldsymbol{\beta}^{ls}$ représente la solution non contrainte des *moindres carrés*. Les ellipses représentent les contours de la fonction de perte quadratique autour de l'estimateur $\boldsymbol{\beta}^{ls}$. On montre les solutions obtenues pour le problème (3.2) soumis à une contrainte ℓ_1 (à gauche) et ℓ_2 (à droite), lorsque $\boldsymbol{\beta}^{ls}$ n'appartient pas au domaine admissible. Chaque problème étant strictement convexe, sa solution est unique et se situe sur la frontière de la région définie par la contrainte associée.

***Remarque* 3.5 —** Le problème (3.2) soumis à une contrainte ℓ_1 est en fait strictement convexe si $\mathbf{X}$ est de rang plein, ce qui nécessite d'avoir $n \geq M$. Quand $n < M$, le problème est convexe mais l'unicité de la solution n'est plus assurée [Osborne et coll., 2000a, théorème 2.1]. ◇

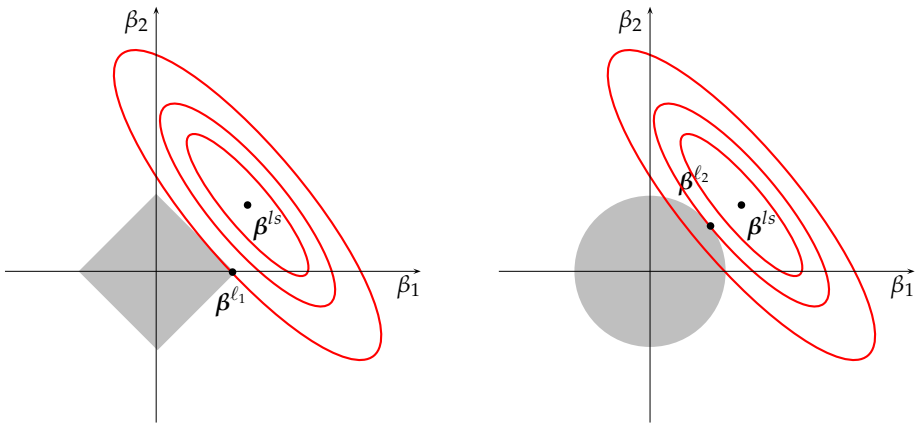


FIGURE 3.2 – *Comparaisons des solutions de problèmes régularisés par une norme ℓ_1 et ℓ_2.*

À gauche de la figure 3.2, $\boldsymbol{\beta}^{\ell_1}$ est la solution du problème (3.2) régularisé par une norme ℓ_1. La deuxième composante de $\boldsymbol{\beta}^{\ell_1}$ est annulée, car l'ellipse atteint la région admissible sur l'angle situé sur l'axe $\beta_2 = 0$. À droite de la figure 3.2, $\boldsymbol{\beta}^{\ell_2}$ est la solution du problème (3.2) régularisé par une norme ℓ_2. La forme circulaire de la région admissible n'incite pas les coefficients à atteindre des valeurs nulles.

Afin de poursuivre cette discussion avec des arguments à la fois simples et formels, on peut donner l'expression d'un coefficient de $\boldsymbol{\beta}^{\ell_1}$ et $\boldsymbol{\beta}^{\ell_2}$, lorsque la matrice $\mathbf{X}$ est orthogonale (ce qui correspond à des contours circulaires pour la fonction de perte quadratique). Pour $\boldsymbol{\beta}^{\ell_2}$, nous avons

$$\beta_m^{\ell_2} \quad = \quad \frac{1}{1+\lambda}\, \beta_m^{ls}\,.$$

Les coefficients subissent un rétrécissement[3] proportionnel par le biais du facteur $1\,/\,(1+\lambda)$. En particulier, $\beta_m^{\ell_2}$ ne peut être nul que si le coefficient β_m^{ls} est lui même exactement nul. Pour $\boldsymbol{\beta}^{\ell_1}$, nous avons

$$\beta_m^{\ell_1} \quad = \quad \text{sign}\left(\beta_m^{ls}\right)\left(|\beta_m^{ls}| - \lambda\right)_+\,,$$

3. Shrinkage, en anglais.

où $[u]_+ = \max(0, u)$. On obtient ainsi un seuillage « doux » : les coefficients de la solution des *moindres carrés* sont rétrécis d'une constante λ lorsque $|\beta_m^{ls}| > \lambda$, et sont annulés sinon.

Stabilité

Définition 3.4 ***Stabilité*** *— Selon Breiman [1996], un problème est instable si pour des ensembles d'apprentissage similaires mais pas identiques (petites perturbations), on obtient des prédictions ou des solutions très différentes (grande perturbation).*

***Remarque* 3.6 —** Bousquet et Elisseeff [2002] ont défini de façon formelle différentes notions de stabilité, basées sur le comportement des estimateurs quand l'échantillon d'apprentissage est perturbé par le retrait ou le remplacement d'un exemple. ◇

Nous avons vu que les régions définies par la contrainte (3.2b) étaient non-convexes pour $p < 1$. L'optimisation de ce type de problème est plus difficile. De plus, les solutions associées ne varient pas de façon continue en t, ce qui occasionne des problèmes d'instabilité [Knight, 2004]. Breiman [1996] a également établi ce constat pour le problème de *sélection de variables*, c'est à dire le problème (3.2) pénalisé par une norme ℓ_0.

3.2.3 Régularisation ℓ_0

La *sélection de variables* ou *sélection de sous-ensembles* [4] consiste, comme son nom l'indique, à sélectionner un sous-ensemble de variables et à éliminer les autres. Présenté par le biais de la formulation (3.2), le problème s'écrit

$$\begin{cases} \min & J(\boldsymbol{\beta}) \\ \text{s. c.} & \|\boldsymbol{\beta}\|_0 \leq t \,, \end{cases}$$

où $\|\boldsymbol{\beta}\|_0 = \text{card}\{\beta_m | \beta_m \neq 0\}$. Ici, t est un entier positif représentant le nombre de variables que l'on souhaite conserver. Ce paramètre t peut soit être fixé avant de lancer la procédure, soit dépendre d'un critère d'arrêt plus « sophistiqué », comme par exemple la stabilisation de la fonction de perte au cours de deux itérations successives [Schuurmans et Southey, 2002]. Si les problèmes régularisés par une norme ℓ_0 conduisent à des modèles parcimonieux, les estimateurs sont quant à eux instables [Breiman, 1996].

3.2.4 Régularisation ℓ_2

Lorsque les variables explicatives sont fortement corrélées, la matrice $\mathbf{X}^T\mathbf{X}$, impliquée dans le calcul de $\boldsymbol{\beta}^{ls}$, la solution des *moindres carrés*, est mal conditionnée : une ou plusieurs valeurs propres sont proches de 0. En conséquence, les coefficients de $\boldsymbol{\beta}^{ls}$ sont susceptibles de prendre des valeurs démesurément élevées, tout comme la variance de l'estimateur.

4. *Subset selection*, en anglais.

Afin de pallier ces inconvénients, Tikhonov et Arsénin [1977] proposent de pénaliser la norme ℓ_2 de l'estimateur f dans la formulation (3.1)

$$\min_{f \in \mathcal{H}} \quad \sum_i (y_i - f(\boldsymbol{x}_i))^2 \; + \; \lambda \, \|f\|^2_{\mathcal{H}} \; .$$

Dans le cadre paramétrique, on parle de *ridge regression* [Hoerl et Kennard, 1970]

$$\min_{\boldsymbol{\beta}} \quad \sum_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 \; + \; \lambda \sum_m \beta_m^2 \; . \tag{3.3}$$

La résolution analytique de ce problème conduit à l'expression

$$\boldsymbol{\beta}^{\ell_2} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y} \; ,$$

où $\mathbf{I} \in \mathcal{M}^{M \times M}$ est la matrice identité. L'ensemble des valeurs propres de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, notamment les plus petites qui reflètent les corrélations, se trouvent décalées d'une constante λ. Imposer une telle contrainte sur les valeurs propres de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ permet de contrôler la magnitude des coefficients, et de réduire la variance de l'estimateur, ce qui peut améliorer les performances en prédiction.

Breiman [1995] propose le *garrot non négatif*[5], qui permet d'introduire l'information apportée par la solution des *moindres carrés* $\boldsymbol{\beta}^{ls}$:

$$\min_{\beta} \quad \sum_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 \; + \; \lambda \sum_m \frac{\beta_m^2}{(\beta_m^{ls})^2} \; .$$

Les coefficients des *moindres carrés* les plus faibles sont ainsi plus sévèrement pénalisés. Géométriquement, tandis que la région définie par une pénalisation ℓ_2 est circulaire, celle ci présente une forme elliptique.

Néanmoins, bien que les solutions obtenues par des problèmes régularisés via une norme ℓ_2 soient uniques et stables, elles ne sont pas parcimonieuses (cf. figure 3.2). Cela présente un inconvénient lorsque l'interprétation est une des finalités du problème d'apprentissage, notamment dans les problèmes de grande dimension.

3.2.5 Régularisation ℓ_1

Afin de combiner sélection de variables et stabilité, Tibshirani [1996] a introduit le critère du *lasso* (*least absolute shrinkage and selection operator*), qui pénalise la norme ℓ_1 d'un estimateur pour une fonction de perte quadratique :

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} \quad \sum\limits_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 & \text{(3.4a)} \\ \text{s. c.} \quad \sum\limits_m |\beta_m| \leq t \; . & \text{(3.4b)} \end{cases}$$

Les solutions $\boldsymbol{\beta}^{\ell_2}$ du problème (3.3) et $\boldsymbol{\beta}^{\ell_1}$ du problème (3.4) voient tous deux leurs coefficients rétrécis. Cependant, nous avons vu dans la section 3.2.2 que la contrainte (3.4b) imposée sur la norme ℓ_1 des estimateurs a pour propriété d'annuler exactement certains coefficients de $\boldsymbol{\beta}^{\ell_1}$.

5. *Non-negative garrote*, en anglais.

***Remarque* 3.7 —** Les problèmes régularisés par le biais d'une norme ℓ_1 se retrouvent également en traitement du signal. En particulier, Chen et coll. [1998] ont proposé une formulation proche du problème (3.4). Le *basis pursuit* défini par

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} & \|\boldsymbol{\beta}\|_1 \\ \text{s. c.} & \mathbf{X}\boldsymbol{\beta} = \mathbf{y} \,, \end{cases}$$

permet de reconstruire le signal $\mathbf{y}$ en utilisant un dictionnaire[6] $\mathbf{X}$ surcomplet ($n < M$). La formulation du *basis pursuit denoising* [Chen et coll., 1998], qui permet de débruiter un signal est quant à elle identique au critère du *lasso*. ◇

Nombre de coefficients non-nuls

Nous avons vu dans les sections précédentes que la solution du *lasso* est de nature parcimonieuse. Il existe également un résultat concernant le nombre de coefficients non-nuls de la solution associée à ce problème :

Théorème 3.1 *Lorsque $n < M$, la solution du lasso comporte au plus n coefficients non-nuls* [Osborne et coll., 2000a, théorème 3.5].

Du point de vue de l'interprétabilité, cette limite peut être trop drastique. Par exemple, dans le domaine génomique, l'ensemble d'apprentissage est composé de seulement quelques dizaines d'expériences sur lesquelles est observé le comportement de plusieurs milliers de gènes. Si l'objectif est de découvrir l'ensemble des gènes significatifs, le *lasso* est alors mal adapté.

Lorsque plusieurs variables explicatives sont fortement corrélées, le *lasso* risque de n'en conserver qu'une. L'interprétation directe des coefficients de $\boldsymbol{\beta}$ peut alors sembler contradictoire avec les connaissances avérées. Si la parcimonie facilite l'interprétation des modèles, il ne faut pas qu'elle masque une partie des phénomènes étudiés.

Consistance en sélection de modèle

Comme le *lasso* combine stabilité et parcimonie, de nombreux auteurs se sont interrogés sur la consistance de ce critère en sélection de modèle.

Définition 3.5 ***Consistance en sélection de modèle*** *— Une procédure de sélection de modèle est dite consistante si la probabilité d'identifier l'ensemble des variables significatives du modèle tend vers* 1*, lorsque la taille de l'échantillon tend vers l'infini.*

En d'autres termes, lorsque le vrai modèle est parcimonieux, les variables sélectionnées par le *lasso* correspondent-elles effectivement à celles du modèle sous-jacent ? Leng et coll. [2004] ont montré que lorsque l'hyperparamètre λ est sélectionné dans l'optique de minimiser l'erreur de prédiction, alors l'estimateur du *lasso* n'aboutit pas à des modèles consistants : de nombreuses variables non significatives sont incluses dans le modèle. Néanmoins, sous certaines conditions, l'estimateur du *lasso* est consistant

6. Par exemple, un dictionnaire d'ondelettes.

[Knight et Fu, 2000 ; Donoho et coll., 2004 ; Meinshausen et Bühlmann, 2006]. On citera également dans ce domaine les travaux de Yuan et Lin [2006] ; Zhao et Yu [2006] ; Zou [2006] ; Bach [2008].

Le problème du *lasso* – néanmoins intéressant pour ses propriétés de stabilité, de convexité et de parcimonie – a inspiré de nombreux travaux ayant pour objectif de produire de estimateurs plus « pertinents » en termes d'interprétabilité et de consistance. Nous en décrirons quelques-uns dans les sections suivantes.

3.2.6 Adaptive ridge

La *pénalisation multiple adaptative*[7] représente, comme le *lasso*, une alternative permettant d'associer les propriétés de stabilité et de parcimonie [Grandvalet, 1998 ; Grandvalet et Canu, 2003]. Elle est définie par

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\boldsymbol{\sigma}} & \sum\limits_i (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \;+\; \sum\limits_m \sigma_m\, \beta_m^2 & & & \text{(3.5a)} \\ \text{s. c.} & \dfrac{\lambda}{M} \sum\limits_m \dfrac{1}{\sigma_m} = 1 & \sigma_m > 0 & \forall m\,, & \text{(3.5b)} \end{cases}$$

où λ est une constante prédéfinie. La contrainte (3.5b) est présente pour éviter les solutions triviales qui consisteraient à affecter un poids nul à chaque paramètre σ_m. Son origine est liée à l'interprétation bayésienne du critère (3.5a)[8]. Chaque coefficient β_m suit une distribution *a priori* gaussienne centrée, de variance proportionnelle à $1/\sigma_m$. La contrainte (3.5b) représente aussi le lien entre les M distributions : elle impose que la variance moyenne des coefficients β_m soit inversement proportionnelle à λ.

Ainsi, les paramètres du vecteur $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_M)$ sont choisis de façon à pénaliser chaque coefficient β_m en fonction de son influence dans le modèle. Les coefficients les moins influents peuvent être annulés. Le problème (3.5) est équivalent au problème (3.4) du *lasso* [Grandvalet, 1998].

3.2.7 Adaptive lasso

Zou [2006] a introduit l'*adaptive lasso* dans le but de construire un estimateur à la fois parcimonieux et consistant. Le problème s'écrit :

$$\min_{\boldsymbol{\beta}} \quad \sum_i (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \;+\; \lambda_n \sum_m \frac{|\beta_m|}{|\sigma_m|^\gamma}\,, \tag{3.6}$$

où la valeur de λ_n varie en fonction de n, la taille de l'ensemble d'apprentissage. La constante $\gamma > 0$ est prédéfinie, et le vecteur $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_M)$ traduit l'influence des variables sur le modèle. On peut par exemple utiliser la solution des *moindres carrés* : $\boldsymbol{\sigma} = \boldsymbol{\beta}^{ls}$.

En fait, pour que le problème (3.6) fournisse un estimateur $\boldsymbol{\beta}$ consistant, le vecteur initial $\boldsymbol{\sigma}$ doit être « zéro-consistant » [Huang et coll., 2007].

7. *Adaptive ridge regression*, en anglais.

8. L'ensemble des problèmes régularisés, notamment (3.3) et (3.4), peuvent aussi être interprétés d'un point de vue bayésien [Hastie et coll., 2001].

Définition 3.6 **Zéro-consistance** — *Soit $\boldsymbol{\beta}$, le véritable vecteur de paramètres. On définit les ensembles $\mathcal{A} = \{m : \beta_m \neq 0\}$ et $\bar{\mathcal{A}} = \{m : \beta_m = 0\}$.*

Un estimateur $\boldsymbol{\sigma}$ est dit zéro-consistant si :

1. *Lorsque n croît, $\max_{m \in \bar{\mathcal{A}}} |\sigma_m|$ devient en probabilité asymptotiquement négligeable :*
$$\lim_{n \to \infty} \max_{m \in \bar{\mathcal{A}}} |\sigma_m| \xrightarrow{p} 0 \, .$$
2. *Pour n suffisamment grand, il existe une constante $\epsilon > 0$ telle que pour n'importe quelle valeur $\delta > 0$, alors*
$$\mathbb{P}\left(\min_{m \in \mathcal{A}} |\sigma_m| \geq \epsilon \min_{m \in \mathcal{A}} |\beta_m| \right) > 1 - \delta \, .$$

Si le vecteur initial $\boldsymbol{\sigma}$ vérifie cette propriété, alors lorsque n augmente, les facteurs d'échelle $1/|\sigma_m|^\gamma \longrightarrow \infty$, pour $m \in \bar{\mathcal{A}}$. Les coefficients β_m associés sont alors rétrécis vers 0. En revanche, lorsque $m \in \mathcal{A}$, les facteurs d'échelle $1/|\sigma_m|^\gamma \longrightarrow c$, où c est une constante finie [Zou, 2006, remarque 2].

3.3 Régularisations structurées

3.3.1 Elastic-net

L'*elastic-net* a été introduit par Zou et Hastie [2005] dans le but de pallier certaines limites du *lasso* en termes de sélection de modèle, pour des problèmes de grande dimension composés de peu d'exemples ($n \ll M$). En effet, le *lasso* peut sélectionner au plus n caractéristiques (cf. théorème 3.1). De plus, en présence d'un ensemble de variables significatives mais corrélées, il tend à ne sélectionner qu'une seule d'entre elles.

Pour remédier à cela, les auteurs proposent d'associer les propriétés des régularisations ℓ_1 et ℓ_2 en minimisant le critère

$$\min_{\boldsymbol{\beta}} \quad \sum_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 \; + \; \lambda_1 \sum_m |\beta_m| \; + \; \lambda_2 \sum_m \beta_m^2 \, , \tag{3.7}$$

où de façon équivalente

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} & \sum\limits_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 & \text{(3.8a)} \\ \text{s. c.} & \alpha \sum\limits_m |\beta_m| \; + \; (1 - \alpha) \sum\limits_m \beta_m^2 \; \leq \; t \, , & \text{(3.8b)} \end{cases}$$

avec $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$. La contrainte (3.8b) est une combinaison convexe de régularisations ℓ_1 et ℓ_2. Lorsque $\alpha > 0$, le problème (3.8) est strictement convexe. C'est par la stricte convexité que les auteurs expliquent l'« effet groupant ».

de Mol et coll. [2008] ont étudié la consistance de ce problème en sélection de modèle, mais également en estimation. Ils concluent à la « consistance forte universelle [9] » de l'estimateur de l'*elastic-net*.

9. Universal strong consistency, en anglais.

3.3.2 Normes mixtes

Maintenant, nous considérons qu'une structure sur les variables est connue. Les variables sont organisées en groupes disjoints. Par exemple, dans un problème où les variables représentent des catégories, une variable composée de c catégories peut être codée en fonction de c nouvelles variables binaires, comme illustré sur le tableau 3.1. Pour un problème spécifique, une connaissance experte peut permettre de définir des groupes sur les variables. Une fois les groupes de variables définis, l'objectif consiste à identifier les éléments pertinents, en utilisant une pénalité différente sur chaque niveau de la structure.

$\mathbf{X}'$	x' : sexe
x'_1	homme
x'_2	femme

$\mathbf{X}$	x^1 : homme	x^2 : femme
x_1	1	0
x_2	0	1

TABLE 3.1 – *Transformation d'une variable catégorielle ($c = 2$) en variables binaires.*

Soit G_ℓ, l'ensemble regroupant les variables associées au groupe ℓ, $\forall \ell \in \{1, \ldots, L\}$. Soit d_ℓ, le nombre de variables incluses dans le groupe $G_\ell : d_\ell = \text{card}(G_\ell)$, et $\sum_\ell d_\ell = M$.

On définit la norme mixte pondérée $\ell_{(r,s)}$ par

$$\|\boldsymbol{\beta}\|_{(r,s)} \quad = \quad \left(\sum_\ell d_\ell \left(\sum_{m \in G_\ell} |\beta_m|^s \right)^{r/s} \right)^{1/r} . \tag{3.9}$$

Dans chaque groupe, les coefficients sont pénalisés par la norme interne ℓ_s. Soit α_ℓ, le représentant du groupe ℓ :

$$\alpha_\ell \quad = \quad d_\ell \left(\sum_{m \in G_\ell} |\beta_m|^s \right)^{1/s} .$$

La norme ℓ_r externe de l'expression (3.9) pénalise chaque élément de $\boldsymbol{\alpha}$, le vecteur des représentants des groupes. Les groupes sont donc pénalisés par une norme ℓ_r, tandis que les variables au sein d'un groupe sont pénalisées par une norme ℓ_s. En fonction du problème considéré, on peut régler r et s, afin d'obtenir des solutions plus ou moins parcimonieuses sur les différents niveaux de la structure.

Yuan et Lin [2006] ont introduit le *group-lasso* afin de sélectionner l'ensemble des variables associées aux groupes pertinents. Ils utilisent pour cela une norme mixte $\ell_{(1,2)}$ en minimisant le critère

$$\min_{\boldsymbol{\beta}} \quad \sum_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 \; + \; \lambda \sum_\ell \sqrt{d_\ell} \left(\sum_{m \in G_\ell} |\beta_m|^2 \right)^{1/2} . \tag{3.10}$$

Les propriétés de la norme ℓ_1 permettent d'identifier les groupes pertinents et d'éliminer les autres, tandis que la norme ℓ_2 fait participer toutes les variables d'un groupe pertinent de manière proportionnelle.

Lorsque l'objectif est d'identifier les groupes significatifs, mais également les variables significatives au sein des groupes, une norme ℓ_s plus restrictive doit être utilisée. Par exemple, une norme mixte $\ell_{(1,4/3)}$ permet d'atteindre une parcimonie effective sur les groupes, et une parcimonie « numérique » à l'intérieur des groupes [Szafranski et coll., 2008a].

***Remarque* 3.8 —** Dans le chapitre 4, nous verrons que les normes mixtes peuvent être abordées sous un aspect variationnel, où la pénalisation est définie indirectement via la minimisation de facteurs d'échelle associés à chaque niveau de la structure. Nous détaillerons à ce moment leurs propriétés. ◇

3.3.3 Composite Absolute Penalties

Zhao et coll. [à paraître] proposent de moduler les pénalités appliquées au sein des groupes dans les normes mixtes en minimisant le critère suivant :

$$\min_{\boldsymbol{\beta}} \quad \sum_i (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \; + \; \lambda \sum_\ell \left(\sum_{m \in G_\ell} |\beta_m|^{s_\ell} \right)^{r/s_\ell} . \tag{3.11}$$

La norme ℓ_{s_ℓ} interne permet de nuancer la pénalité en fonction de l'*a priori* sur le groupe ℓ. Par exemple, on utilisera une norme ℓ_2 sur un groupe dont on sait que l'ensemble des variables doit jouer un rôle significatif. Au contraire, pour un groupe sur lequel on n'a pas d'*a priori*, on essaiera plutôt d'identifier les variables significatives en utilisant une norme plus restrictive.

***Remarque* 3.9 —** Aucun algorithme de résolution efficace n'est proposé pour résoudre le problème ainsi généralisé. Les auteurs proposent en pratique d'utiliser la norme mixte $\ell_{(1,\infty)}$. ◇

Zhao et coll. proposent également de sélectionner les groupes de façon « hiérarchique », lorsqu'ils ne forment plus une partition mais qu'ils sont imbriqués :

$$G_1 \subset G_2 \subset \cdots \subset G_L \subset I \, ,$$

où $I = \{1, \ldots, M\}$ représente l'ensemble des variables du problème. Cette structure sur les variables permet de sélectionner les groupes dans un ordre prédéfini :

$$\{I \setminus G_L\}, \; \{G_L \setminus G_{L-1}\}, \; \ldots, \; \{G_2 \setminus G_1\}, \; G_1 \, .$$

3.3.4 Multiple Kernel Learning

Le cadre du Multiple Kernel Learning (MKL) a été initié par [Lanckriet et coll., 2004] pour combiner différents noyaux issus de M sources d'information. Le problème consiste à optimiser les paramètres d'un SVM avec comme noyau effectif, la combinaison linéaire des M noyaux : $K(\boldsymbol{x}, \boldsymbol{x}') \; = \; \sum_m K_m(\boldsymbol{x}, \boldsymbol{x}')$.

Pour identifier les éléments significatifs de cette combinaison, Bach et coll. [2004] posent le problème sous la forme suivante :

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi}}} & \frac{1}{2}\left(\sum\limits_m \|f_m\|_{\mathcal{H}_m}\right)^2 + C\sum\limits_i \xi_i & \text{(3.12a)} \\ \text{s. c.} & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i) + b\right) \geq 1-\xi_i \qquad \xi_i \geq 0 \qquad \forall i\ . & \text{(3.12b)} \end{cases}$$

Dans ce problème très similaire au problème SVM original (2.5), la norme mixte associée à chaque EHNR $\mathcal{H}_m$ remplace la norme au carré sur l'EHNR $\mathcal{H}$. La norme ℓ_1 appliquée sur les éléments f_m encourage des solutions parcimonieuses sur les noyaux K_m.

***Remarque* 3.10** — Le problème (3.12) est une généralisation non-paramétrique du *group-lasso*. En effet, si la norme mixte $\ell_{(1,2)}$ utilisée dans *group-lasso* est reformulée en termes de produits scalaires, on obtient :

$$\begin{aligned} \|\boldsymbol{\beta}\|_{(1,2)} &= \sum_\ell \sqrt{d_\ell}\ \langle \boldsymbol{\beta}_\ell, \boldsymbol{\beta}_\ell\rangle_2^{1/2} \\ &= \sum_\ell \sqrt{d_\ell}\ \|\boldsymbol{\beta}_\ell\|_2 \qquad , \end{aligned}$$

où $\boldsymbol{\beta}_\ell$ est le vecteur de coefficients des variables du groupe ℓ. L'élévation au carré de la norme mixte du critère (3.12a) influence la force de la pénalité, mais pas sa forme. ◇

Dans la suite de ce document, nous nous intéresserons plus particulièrement à la définition du problème du MKL donnée par Rakotomamonjy et coll. [2007] :

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi},\boldsymbol{\sigma}}} & \frac{1}{2}\sum\limits_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i & \text{(3.13a)} \\ \text{s. c.} & \sum\limits_m \sigma_m \leq 1 \qquad \sigma_m \geq 0 \qquad \forall m & \text{(3.13b)} \\ & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i) + b\right) \geq 1-\xi_i \qquad \xi_i \geq 0 \qquad \forall i\ , & \text{(3.13c)} \end{cases}$$

où le noyau associé à ce problème est défini par la combinaison linéaire convexe

$$K(\boldsymbol{x},\boldsymbol{x}') = \sum_m \sigma_m K_m(\boldsymbol{x},\boldsymbol{x}')\ .$$

C'est la contrainte ℓ_1 (3.13b) imposée aux coefficients σ_m qui induit la parcimonie sur les éléments f_m et sur les noyaux K_m. En ce sens, les formulations (3.12) et (3.13) sont équivalentes. On montre d'ailleurs, de la même façon que dans [Grandvalet, 1998], que (3.13) est une formulation variationnelle de (3.12), et que les deux problèmes sont équivalents.

***Remarque* 3.11 —** L'approche du MKL peut aussi être utilisée comme alternative à la validation croisée, pour choisir les hyper-paramètres d'une famille de noyaux. Par exemple, dans une famille de noyaux gaussiens où la largeur de bande varie en fonction de m, le MKL sélectionne la combinaison la plus appropriée pour le problème considéré. ◇

3.4 Algorithmes de résolution

Dans cette section, nous décrivons les principes généraux d'algorithmes permettant de résoudre divers problèmes régularisés. Ces algorithmes ont été regroupés en quatre catégories : les algorithmes

- de seuillage itératif ;
- de contraintes actives ;
- d'approximation par plans sécants ;
- de calcul de chemin de régularisation.

***Remarque* 3.12 —** Il existe également une librairie matlab qui implémente différente statégies pour résoudre le problème du lasso [Schmidt]. ◇

3.4.1 Seuillage itératif

Dans cette section, nous nous intéressons aux méthodes itératives basées sur les seuillages obtenus par les conditions d'optimalité du premier ordre du problème général (3.2) :

$$\min_{\boldsymbol{\beta}} \quad \mathcal{L} \;=\; J(\boldsymbol{\beta}) \;+\; \lambda\, P(\boldsymbol{\beta})\;,$$

en particulier pour le problème du *lasso*. Dans ce cadre, ces méthodes de seuillage itératif possèdent de bonnes propriétés de convergence [Fu, 1998 ; Daubechies et coll., 2004]. Néanmoins, nous verrons dans les sections suivantes qu'il existe des méthodes plus efficaces en terme de temps de calcul.

Shooting

Fu [1998] a proposé un algorithme itératif pour le critère du *lasso*. Le seuillage, basé sur les conditions d'optimalité du premier ordre de ce critère, définit chaque coefficient β_m en fonction de l'ensemble des coefficients $\beta_{m'}$, où $m' \neq m$. En effet,

$$\frac{\partial \mathcal{L}}{\partial \beta_m} \;=\; 0 \qquad \Rightarrow \qquad \beta_m \;=\; \frac{-\lambda\, \mathrm{sign}(\beta_m) \;-\; u}{2 \sum\limits_i (x_i^m)^2}\;,$$

où u représente la dérivée de la perte quadratique $J(\boldsymbol{\beta})$ associée aux coefficients $\beta_{m'}$:

$$u \;=\; 2 \sum_i \sum_{m' \neq m} x_i^m\, x_i^{m'} \beta_{m'} \;-\; 2 \sum_i x_i^m y_i\;.$$

L'ensemble des coefficients est initialisé par les estimations des *moindres carrés* : $\forall m$, $\beta_m^0 = \beta_m^{ls}$. Ensuite, $\beta_m^{t+1} = S_\lambda[u]$, avec

$$S_\lambda[u] \;=\; \begin{cases} \dfrac{\lambda - u}{2\sum\limits_i (x_i^m)^2} & u \;> \lambda \\ \dfrac{-\lambda - u}{2\sum\limits_i (x_i^m)^2} & u \;< -\lambda \\ 0 & |u| \leq \lambda \;, \end{cases}$$

où u est ici fonction de $\beta_{m'}^t$, $\forall m' \neq m$.

Yuan et Lin [2006] ont étendu cette approche au critère du *group-lasso*. Cet algorithme converge en quelques itérations vers des solutions stables. Cependant, leurs simulations montrent que le temps de convergence est significativement plus long que ceux des algorithmes basés sur un sous-ensemble de variables (cf. section 3.4.2).

Itérations de Landweber

Daubechies et coll. [2004] ont également proposé une variante de cet algorithme, basé sur les itérations de Landweber, pour le critère du *lasso*. Le principe de cet algorithme de seuillage consiste à utiliser une fonctionnelle de remplacement :

$$\tilde{\mathcal{L}} = \mathcal{L} \;+\; C \sum_m (\beta_m - \alpha_m)^2 \;-\; \sum_i \left(\sum_m x_i^m (\beta_m - \alpha_m) \right)^2 \;,$$

avec $\tilde{\mathcal{L}} = \mathcal{L}$ pour $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Si C majore la plus grande valeur propre de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$: $C > \|\mathbf{X}^{\mathrm{T}}\mathbf{X}\|$, alors $\tilde{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ est strictement convexe en $\boldsymbol{\beta}$ pour $\boldsymbol{\alpha}$ fixé. Les auteurs montrent que la fonctionnelle $\tilde{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ peut se récrire de façon à ce que

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \beta_m} \;=\; 2C\,\beta_m - 2C\,\alpha_m \;-\; 2\sum_i y_i x_i^m \;+\; 2\sum_i \sum_{m'} x_i^m \, x_i^{m'} \alpha_{m'} \;+\; \lambda\,\mathrm{sign}(\beta_m) \;.$$

Ainsi, pour $\boldsymbol{\alpha} = \boldsymbol{\beta}^t$, l'algorithme de seuillage consiste en une descente de gradient à pas fixé, avec :

$$\beta_m^{t+1} \;=\; S_{\lambda/2C} \left[\beta_m^t \;+\; \frac{1}{C} \sum_i y_i \, x_i^m \;-\; \frac{1}{C} \sum_i \sum_{m'} x_i^m \, x_i^{m'} \beta_{m'}^t \right] \;,$$

où $S_\lambda[u]$ est le seuillage défini par les conditions d'optimalité du premier ordre du problème considéré :

$$S_\lambda[u] \;=\; \begin{cases} u - \lambda & u \;> \lambda \\ u + \lambda & u \;< -\lambda \\ 0 & |u| \leq \lambda \;. \end{cases}$$

Ici, le seuillage ne fait intervenir que des produits scalaires entre les vecteurs $\boldsymbol{x}^m$ et des vecteurs $\mathbf{u}$, où $\mathbf{u} \neq \boldsymbol{x}^m$, $\forall m$. Or, ces produits scalaires

peuvent être calculés par des algorithmes de transformations rapides, telles que les transformées en ondelettes ou les transformées de Fourier à court terme. On s'affranchit ainsi du stockage de la matrice $\mathbf{X}$ [10].

Cet algorithme, convergent quelle que soit l'initialisation, peut se révéler assez lent dans certaines situations [Daubechies et coll., à paraître]. Afin de pallier cet inconvénient, les auteurs ont proposé une modification de ce seuillage, par le biais de projections sur la boule engendrée par les contraintes de la pénalité ℓ_1.

3.4.2 Contraintes actives

Les algorithmes décrits dans cette section travaillent avec un sous-ensemble optimal de variables, appelé « ensemble actif ». L'ensemble actif d'un problème, noté $\mathcal{A}$, est défini par l'ensemble des indices des « variables actives », c'est-à-dire celles pour lesquelles les coefficients associés sont non-nuls :

$$\begin{aligned} \mathcal{A} &= \{m \,|\, \beta_m \neq 0\} \ , \\ \gamma &= \{\beta_m\}_{m \in \mathcal{A}} \ . \end{aligned}$$

On note $\bar{\mathcal{A}}$ le complémentaire de l'ensemble actif :

$$\bar{\mathcal{A}} = \{m \,|\, \beta_m = 0\} \ .$$

Ce type d'algorithme itératif, initialement proposé par Osborne et coll. [2000a] pour résoudre le problème du *lasso*, peut se résumer en deux étapes. À chaque itération :

1. Mise à jour de l'ensemble $\mathcal{A}$, avec par exemple, la variable de l'ensemble $\bar{\mathcal{A}}$ qui viole le plus les conditions d'optimalité du problème global (3.1).
2. Résolution du problème d'optimisation sur l'ensemble actif courant $\mathcal{A}$
$$\min_{\gamma} \quad J(\gamma) \ + \ \lambda \, P(\gamma) \ .$$

***Remarque* 3.13 —** Pour résoudre le problème défini à l'étape 2, on peut choisir une méthode d'optimisation adaptée au problème global considéré. Par exemple, Osborne et coll. déterminent une direction de descente $\mathbf{h}$, en résolvant à chaque itération une approximation linéaire

$$\begin{cases} \min\limits_{\mathbf{h}} & J(\gamma + \mathbf{h}) \\ \text{s. c.} & \text{sign}(\gamma)^{\mathrm{T}}(\gamma + \mathbf{h}) \ \leq \ t \ , \end{cases}$$

du problème initial

$$\begin{cases} \min\limits_{\mathbf{h}} & J(\boldsymbol{\beta} + \mathbf{h}) \\ \text{s. c.} & \text{sign}(\boldsymbol{\beta} + \mathbf{h})^{\mathrm{T}}(\boldsymbol{\beta} + \mathbf{h}) \ \leq \ t \ , \end{cases}$$

exprimé sous la forme d'Ivanov. ◇

10. Dans certaines problématiques liées au traitement du signal, les matrices $\mathbf{X}$ sont telles qu'elles ne peuvent être stockées sur des machines récentes disposant d'environ 4 Go de mémoire vive.

L'ensemble actif, vide à l'initialisation pour réduire la quantité de calcul, est mis à jour de façon incrémentale. Chaque itération nécessite la résolution d'un système de taille card$\{\mathcal{A}\}$. Si l'on considère des régularisations parcimonieuses, card$\{\mathcal{A}\}$ reste inférieur à M jusqu'à l'optimalité. Ainsi, cette méthode est moins consommatrice que des approches de type *backward*, où l'ensemble actif de départ est constitué de la totalité des variables du problème, et où les variables sont éliminées une à une [voir par exemple Fu, 1998].

De plus, bien qu'introduit par Osborne et coll. pour un critère quadratique et une régularisation ℓ_1, cette méthode présente l'avantage de pouvoir être adaptée à une fonction de perte et une pénalité génériques. Dans le chapitre 4, nous détaillerons cette approche pour un problème défini par un critère $J(.)$ différentiable et régularisé par une norme mixte $\ell_{(r,s)}$.

[Roth, 2004] a quant à lui utilisé cette approche pour résoudre le problème du *lasso*, avec un critère $J(\boldsymbol{\beta})$ étendu à des modèles linéaires généralisés. Plus récemment, [Roth et Fischer, 2008] ont proposé ce schéma pour le critère du *group-lasso*, toujours dans le cadre de modèles linéaires généralisés. Néanmoins, dans l'étape 1 de l'algorithme associé au critère du *group-lasso*, c'est la totalité des variables appartenant à un groupe qui est ajoutée à l'ensemble actif.

Perkins et coll. [2003] utilisent une méthode similaire, baptisée « grafting », pour résoudre un problème régularisé par une combinaison de pénalités ℓ_0, ℓ_1, et ℓ_2 :

$$\min_{\boldsymbol{\beta}} \quad J(\boldsymbol{\beta}) \;+\; \lambda_0 \sum_m \beta_m^0 \;+\; \lambda_1 \sum_m |\beta_m| \;+\; \lambda_2 \sum_m \beta_m^2 \,,$$

où $0^0 \equiv 0$, et où $J(\boldsymbol{\beta})$ est une log-vraisemblance binomiale négative.

3.4.3 Approximation par plans sécants

Le principe de ces méthodes consiste à déterminer la solution d'un problème régularisé, en utilisant une approximation convexe de la fonction de perte [11] [Hiriart-Urruty et Lemarechal, 1996]. Cette approximation est obtenue par le biais d'une fonction linéaire par morceaux, définie par le maximum d'un ensemble d'hyperplans, appelés plans sécants. Ces derniers sont déterminés par les sous-gradients de la fonction de perte, en un nombre de points spécifiques. La figure 3.3 illustre ce propos.

***Remarque* 3.14 —** Les méthodes d'approximation par plans sécants sont aussi connues sous le nom de méthodes faisceaux [12]. ◇

Utilisation dans le cadre des SVM

Ces méthodes connues depuis plusieurs décennies ont récemment fait l'objet de plusieurs travaux en apprentissage statistique, notamment avec les SVM linéaires [Joachims, 2006 ; Smola et coll., 2008 ;

11. Nous considérons dans cette présentation des fonctions de perte également convexes.

12. Bundle methods, en anglais.

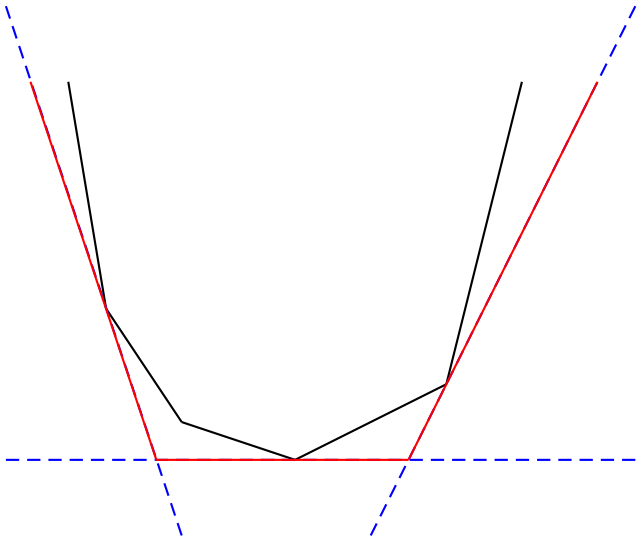


FIGURE 3.3 – *Approximation convexe d'une fonction convexe. En noir, la fonction à approcher ; en pointillés bleus, les hyperplans définis par les sous-gradients ; en rouge, l'approximation convexe de la fonction.*

Franc et Sonnenburg, 2008]. Elles utilisent la formulation Lagrangienne :

$$\min_{\boldsymbol{w}} \qquad \frac{1}{2}\,\|\boldsymbol{w}\|^2 \;+\; C\,R(\boldsymbol{w})\;,$$

où $R(\boldsymbol{w})$ est la fonction de coût charnière $\sum_i [1 - y_i\,(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)]_+$, avec $[u]_+ = \max(0, u)$. C'est donc par le biais de la formulation primale (2.5) que l'on cherche à résoudre le problème SVM.

La résolution s'effectue en deux étapes. Après avoir initialisé à 0 les variables $\boldsymbol{w}^0$, $\mathbf{a}^0$ et δ^0, où le couple $(\mathbf{a}, \delta)$ représente les composantes d'un plan sécant, on alterne :

1. Résolution du problème

$$\boldsymbol{w}^t \;=\; \arg\min_{\boldsymbol{w}} \qquad \frac{1}{2}\,\|\boldsymbol{w}\|^2 \;+\; C\,R^t(\boldsymbol{w})\;,$$

où $R^t(\boldsymbol{w}) = \max_{t' \le t} \left[\langle \mathbf{a}^{t'}, \boldsymbol{w}\rangle + \delta^{t'}\right]$ est l'approximation de $R(\boldsymbol{w})$ formée par l'enveloppe convexe des plans sécants actifs à l'itération t.

2. Ajout d'un plan sécant $\left(\mathbf{a}^{t+1}, \delta^{t+1}\right)$:

$$\begin{aligned} \mathbf{a}^{t+1} &= \partial R(\boldsymbol{w}^t)\;, \\ \delta^{t+1} &= R(\boldsymbol{w}^t) - \langle \mathbf{a}^{t+1}, \boldsymbol{w}^t\rangle\;, \end{aligned}$$

où $\partial R(.)$ représente un sous-gradient de $R(.)$.

Le nombre de plans sécants augmente de façon incrémentale au cours des itérations. En ce sens, ce type d'algorithme peut-être vu comme un algorithme de contraintes actives, où les contraintes sont définies par les plans sécants, et non plus par les variables.

Des résultats théoriques montrent que le nombre d'itérations nécessaires pour la convergence de ces méthodes est indépendante du nombre d'exemples considérés. Ainsi, il est possible de traiter très rapidement des problèmes dans lesquels le nombre d'exemple est important

[Franc et Sonnenburg, 2008, section 4]. Cependant, ce type d'algorithmes peut conduire à des solutions instables, notamment lorsque le nombre de plans sécants est faible, et que l'approximation de la fonction objectif est imprécise [Bonnans et coll., 2006].

Utilisation dans le cadre du MKL

Sonnenburg et coll. [2006] proposent un algorithme de programmation linéaire semi-infinie (SILP) pour résoudre le problème MKL (3.12) sous sa forme duale. Après quelques manipulations, le problème devient :

$$\begin{cases} \max\limits_{\theta, \boldsymbol{\alpha}, \sigma} & \theta \\ \text{s. c.} & \frac{1}{2} \sum\limits_{i,j} \alpha_i \alpha_j \, y_i y_j \, \bar{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \; - \sum\limits_i \alpha_i \; \geq \; \theta \quad \theta \in \mathbb{R} \\ & \sum\limits_i \alpha_i y_i \; = \; 0 \qquad\qquad 0 \leq \; \alpha_i \; \leq C \quad \forall i \\ & \sum\limits_m \sigma_m \; = \; 1 \qquad\qquad \sigma_m \; \geq \; 0 \qquad \forall m \, , \end{cases}$$

où le noyau global du problème SVM est défini par une combinaison linéaire :

$$\bar{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_m \sigma_m K_m(\boldsymbol{x}, \boldsymbol{x}') \, .$$

Néanmoins, l'approche SILP peut également s'interpréter selon le principe des méthodes de plans sécants. L'optimisation de ce problème s'effectue en deux temps. À l'itération t :

1. Estimation des paramètres $\boldsymbol{\alpha}^t$ du sous-problème SVM :

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2} \sum\limits_{i,j} \alpha_i \alpha_j \, y_i y_j \, \bar{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \; - \sum\limits_i \alpha_i \\ \text{s. c.} & \sum\limits_i \alpha_i y_i \; = \; 0 \qquad\qquad 0 \leq \; \alpha_i \; \leq C \quad \forall i \, , \end{cases}$$

   déterminés en considérant les paramètres θ et σ de l'itération précédente.

2. Estimation de θ^t et des paramètres σ^t liés à la sélection des noyaux les plus influents :

$$\begin{cases} \max\limits_{\theta, \sigma} & \theta \\ \text{s. c.} & \frac{1}{2} \sum\limits_{i,j} \alpha_i^{t'} \alpha_j^{t'} \, y_i y_j \, \bar{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \; - \sum\limits_i \alpha_i^{t'} \; \geq \; \theta \quad \theta \in \mathbb{R} \quad \forall t' \leq t \\ & \sum\limits_m \sigma_m \; = \; 1 \qquad\qquad \sigma_m \geq 0 \quad \forall m \, , \end{cases}$$

   déterminés en considérant l'ensemble des paramètres $\boldsymbol{\alpha}$ des itérations précédentes.

À chaque itération, le vecteur $\boldsymbol{\alpha}^t$ est évalué en fonction des paramètres $\boldsymbol{\sigma}^{t-1}$ donnés au noyau $\bar{K}$. L'étape 1 définit en fait les composantes $\{S_m^t\}$ d'un plan sécant :

$$S_m^t \;=\; \frac{1}{2}\sum_{i,j}\alpha_i^t\,\alpha_j^t\,y_i y_j\,K_m(\boldsymbol{x}_i,\boldsymbol{x}_j) \;-\; \sum_i \alpha_i^t\;.$$

Ainsi, à l'étape 2 suivante, une nouvelle contrainte $\sum_m \sigma_m\,S_m^t \geq \theta$ est ajoutée à l'ensemble des contraintes déjà présentes. Au cours des itérations, l'approximation de la fonction objectif devient de plus en plus fine.

3.4.4 Chemin de régularisation

Le ***chemin de régularisation*** est défini par l'ensemble des solutions optimales d'un problème régularisé sous forme Lagrangienne (3.1), en fonction de la variation du paramètre de régularisation :

$$\{\boldsymbol{\beta}(\lambda) : 0 \leq \lambda < \infty\}\;.$$

Pour le critère du *lasso*, lorsque λ varie de façon continue, le chemin de régularisation varie lui aussi continûment. Il est de plus linéaire par morceaux. De manière générale, les problèmes dont la fonction de perte est quadratique par morceaux et le terme de régularisation linéaire par morceaux (en particulier, les régularisations ℓ_1 et ℓ_∞), présentent cette propriété [Rosset et Zhu, 2007]. C'est lorsque ces conditions sur la fonction de perte et le terme de régularisation sont réunies que le chemin de régularisation peut être calculé *exactement*.

Osborne et coll. [2000b] ont introduit le calcul du chemin de régularisation pour le problème du *lasso*. Pour relier les différentes solutions, les auteurs utilisent la notion d'homotopie. C'est cependant avec l'algorithme du LARS (Least Angle Regression) Efron et coll. [2004], basé sur le principe des contraintes actives, que les approches permettant de définir un chemin de régularisation ont été popularisées dans la communauté de l'apprentissage statistique.

Le principe général d'un algorithme définissant un chemin de régularisation consiste en deux étapes principales. À partir de l'ensemble actif $\mathcal{A}$ et de la solution $\boldsymbol{\beta}$ de l'itération courante :

1. Recherche d'un pas γ et d'une direction $\mathbf{h}$, sur le chemin de régularisation, pour lequel l'ensemble actif est modifié.
2. Mise à jour de la solution en fonction du pas γ et de la direction $\mathbf{h}$ déterminés :
$$\boldsymbol{\beta}^{t+1} \leftarrow \boldsymbol{\beta}^t \;+\; \gamma\mathbf{h}\;.$$
L'ensemble actif est modifié en conséquence.

Dans l'algorithme du LARS, on utilise les corrélations des variables au résidu pour déterminer les pas et les variables qui entrent dans (ou sortent de) l'ensemble actif. Une description rudimentaire de l'algorithme est la

suivante. La première variable qui entre dans l'ensemble actif est celle dont la corrélation avec le résidu est la plus importante :

$$\begin{aligned} k &= \arg\max_{m \in \bar{\mathcal{A}}} \left| (\boldsymbol{x}^m)^{\mathrm{T}} (y - \mathbf{X}_{\mathcal{A}} \boldsymbol{\beta}) \right| , \\ \mathcal{A} &\leftarrow k , \end{aligned}$$

où $\boldsymbol{X}_{\mathcal{A}}$ représente la matrice des observations composée des variables de l'ensemble actif. Le coefficient β_k associé voit sa valeur augmentée, dans la direction du signe de sa corrélation avec le résidu, jusqu'à ce qu'une seconde variable k' ait, en valeur absolue, la même corrélation avec le résidu courant. La variable k' est ajoutée à l'ensemble actif courant : $\mathcal{A} \leftarrow \mathcal{A} + k'$. L'algorithme continue de cette façon, jusqu'à ce que toutes les variables soient entrées dans l'ensemble actif. Cette procédure permet d'obtenir l'ensemble du chemin de régularisation [Efron et coll., 2004, lemme 7], pour un coût de calcul équivalent à celui de la solution des *moindres carrés*. Pour une présentation détaillée du LARS, le lecteur peut aussi se référer à Guigue [2005].

D'autres auteurs se sont intéressés aux propriétés et aux algorithmes de chemins de régularisation pour des pénalités ℓ_1. Nous citerons notamment les travaux de Lee et coll. [2006], Park et Hastie [2007] et Rosset et Zhu [2007], qui ont examiné différents critères.

Yuan et Lin [2006] réutilisent le concept du LARS pour définir le chemin de régularisation du *group-lasso* . Ils font intervenir les corrélations entre l'ensemble des variables associées à un groupe ℓ et le résidu :

$$\frac{1}{d_\ell} \| \mathbf{X}_{G_\ell}^{\mathrm{T}} (y - \mathbf{X}_{\mathcal{A}} \boldsymbol{\beta}) \|^2 ,$$

où $\mathbf{X}_{G_\ell}$ représente la matrice des observations composée de l'ensemble des variables du groupe ℓ. Un groupe entre dans l'ensemble actif si la valeur moyenne au carré des corrélations des variables du groupe au résidu est identique à celle des groupes actifs. Park et Hastie [2006] ont dérivé et étendu cette approche en classification. Zhao et coll. [à paraître] proposent également un algorithme pour calculer le chemin d'une régularisation $\ell_{(1,\infty)}$, avec un critère de *moindres carrés*.

3.5 Synthèse

Dans ce chapitre, nous avons présenté différentes formes de régularisations. D'une part, les propriétés des régularisations ℓ_p ont été étudiées, notamment celles de la régularisation ℓ_1. D'autre part, nous avons introduit les régularisations structurées. Enfin, les principes généraux de plusieurs catégories d'algorithmes associés à la résolution de différents problèmes régularisés ont été décrits. Le prochain chapitre nous donne l'occasion d'aborder les normes mixtes sous un nouvel angle.

Pénalisation hiérarchique

4

Sommaire

4.1 Introduction

La *pénalisation hiérarchique* permet d'intégrer une connaissance *a priori* dans l'estimation de modèles statistiques. Cet *a priori* est représenté par une structure sur les variables ou les caractéristiques d'un jeu de données. L'objectif est d'extraire de cette structure les composantes significatives.

Nous décrivons d'abord les différentes étapes de la démarche qui permet d'aboutir au modèle. Nous montrons que la formulation obtenue est le variationnel d'un problème régularisé par une norme mixte. Par le biais de ces deux formulations, nous étudions les propriétés de la *pénalisation hiérarchique* en termes de parcimonie et de convexité, et caractérisons la relation qui lie ces deux notions.

Nous présentons ensuite deux algorithmes pour résoudre ce problème. La première approche est développée pour des fonctions paramétriques de l'espace des variables explicatives. Elle exploite l'expression régularisée par une norme mixte et utilise le principe d'« ensemble actif » de variables. La seconde approche quant à elle s'appuie sur la formulation variationnelle, qui nous permet d'étendre notre méthode au cadre des fonctions noyaux associées à l'espace des caractéristiques, en utilisant un algorithme de type « wrapper [1] », basé sur une méthode de gradient. Nous montrons également comment relier ces deux approches, afin de pouvoir les comparer par la suite.

Enfin, nous terminons en discutant des extensions envisagées pour étendre ce formalisme à des situations plus complexes.

4.2 Formalisation du modèle

4.2.1 Cadre général

Nous nous intéressons aux problèmes où les variables explicatives (resp. les caractéristiques) d'un jeu de données peuvent être structurées dans une arborescence. Dans cette structure, les variables (resp. les caractéristiques) sont situées sur les feuilles de l'arbre, c'est-à-dire au niveau le plus profond. Les niveaux intermédiaires identifient les sous-groupes auxquels appartiennent les variables (resp. les caractéristiques). Ces groupes sont reliés par une racine commune, associée au niveau 0 de l'arbre.

***Remarque* 4.1** — Dans la section 4.2, nous considérons les arborescences associées aux variables explicatives, afin d'alléger la présentation du formalisme. Néanmoins, ce dernier s'étend de façon immédiate à l'espace des caractéristiques. ◇

Dans cette représentation arborescente, supprimer des variables consiste à élaguer des branches de l'arbre. Si une feuille est élaguée, la variable correspondante est supprimée du modèle. Si un sous-arbre est élagué, c'est le sous-groupe entier qui est éliminé.

1. Un algorithme où deux problèmes d'optimisation sont emboîtés.

4.2.2 Arborescences à deux niveaux

Nous formalisons maintenant le problème pour des arborescences à deux niveaux [2]. Nous considérons, sans perte de généralité, des arbres dont toutes les feuilles sont au même niveau. Si nécessaire, un pré-traitement introduit des nœuds intermédiaires, comme illustré en figure 4.1.

Les groupes de variables, symbolisés par les nœuds de hauteur 1, sont indicés par les valeurs $\{1, \dots, \ell, \dots, L\}$. L'ensemble G_ℓ désigne les fils du nœud ℓ, c'est-à-dire les variables appartenant au groupe ℓ. Cet ensemble est de cardinalité d_ℓ.

Les branches allant de la racine au nœud ℓ sont étiquetées par $\sigma_{1,\ell}$, qui représente le facteur d'échelle associé au groupe ℓ. Les branches allant du nœud ℓ à la feuille m sont étiquetées par $\sigma_{2,m}$, qui représente le facteur d'échelle associé à variable m. Ces différentes notations sont illustrées sur la figure 4.1.

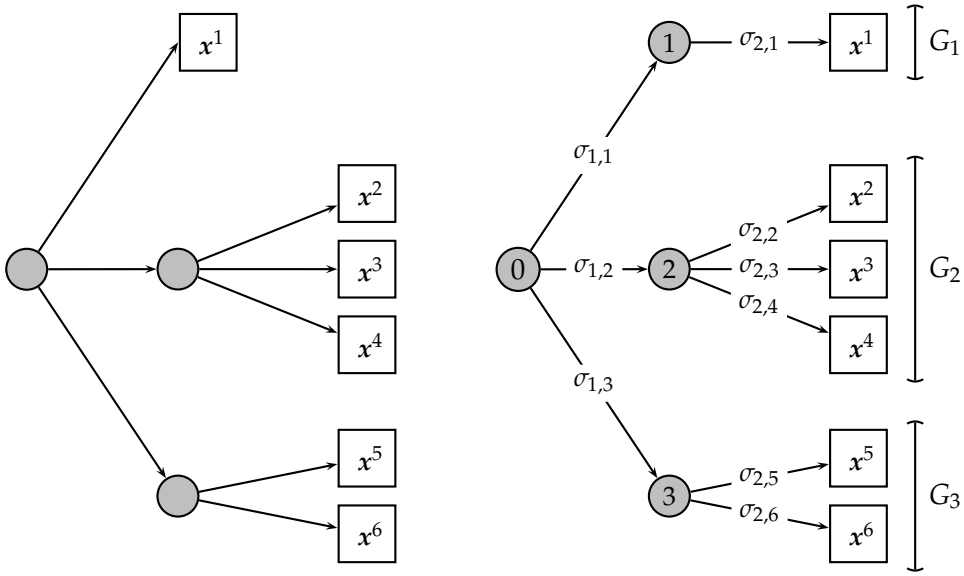


FIGURE 4.1 – *Exemple d'arborescence à deux niveaux associée aux variables explicatives d'un jeu de données.*

***Remarque* 4.2 —** Dorénavant, pour alléger les notations, les variations des indices dans les sommes seront occultées. Le cas échéant, les indices i et j, représentant les observations, varieront de 1 à n ; l'indice m, représentant les variables, variera de 1 à M ; l'indice ℓ, représentant les groupes de variables, variera de 1 à L. ◇

4.2.3 Sélection « exacte » de variables

Éliminer les variables du modèle consiste à élaguer les branches de l'arbre. Dans ce but, on commence par considérer les facteurs $\boldsymbol{\sigma}_1$ et $\boldsymbol{\sigma}_2$ comme des poids binaires. Supprimer la variable m consiste à affecter la valeur 0 au

2. La hauteur de l'arborescence est donc $H = 2$.

poids $\sigma_{2,m}$. De la même façon, supprimer le groupe ℓ consiste à annuler le poids $\sigma_{1,\ell}$. Cette dernière action a pour effet d'annuler l'ensemble des poids associés aux variables contenues dans G_ℓ.

Si le problème consiste uniquement à sélectionner des variables, $\sigma_{1,\ell}$ et $\sigma_{2,m}$ sont redondants. Néanmoins, nous souhaitons ici favoriser des solutions parcimonieuses sur les deux niveaux de la hiérarchie. Dans cette optique, en se référant à la figure 4.1, il est préférable d'annuler le poids $\sigma_{1,3}$ (et de fait, les poids $\sigma_{2,5}$ et $\sigma_{2,6}$) plutôt que d'annuler $\sigma_{2,4}$ et $\sigma_{2,5}$. Si au final les deux options permettent d'éliminer le même nombre de variables, seule la première permet de s'affranchir d'un groupe.

Ainsi, pour intégrer le processus de sélection à chaque niveau, nous considérons le problème de minimisation de l'erreur quadratique suivant

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\boldsymbol{\sigma}} & \sum\limits_i (y_i - \boldsymbol{x}_i(\boldsymbol{\sigma} \star \boldsymbol{\beta}))^2 & & & \text{(4.1a)} \\ \text{s. c.} & \sigma_m = \sigma_{1,\ell}\,\sigma_{2,m} & & \forall \ell,\ m \in G_\ell & \text{(4.1b)} \\ & \sum\limits_\ell d_\ell\,\sigma_{1,\ell} \leq s_1 & \sigma_{1,\ell} \in \{0,1\} & \forall \ell & \text{(4.1c)} \\ & \sum\limits_m \sigma_{2,m} \leq s_2 & \sigma_{2,m} \in \{0,1\} & \forall m\ , & \text{(4.1d)} \end{cases}$$

où $\mathbf{u} \star \mathbf{v}$ représente le produit terme à terme des vecteurs $\mathbf{u}$ et $\mathbf{v}$, et où s_1 et s_2 désignent le nombre de feuilles que l'on souhaite conserver après élagage à chaque niveau de l'arborescence. Les coefficients du vecteur $\boldsymbol{\beta}$ sont soumis à des contraintes de parcimonie :

- au niveau des groupes, par le biais de la contrainte ℓ_0 (4.1c) associée aux facteurs σ_1,
- au niveau des variables, par le biais de la contrainte ℓ_0 (4.1d), associée aux facteurs σ_2.

Bien que le système (4.1) exprime de façon exacte les différents aspects liés à notre problématique, il présente plusieurs inconvénients. D'une part, ce problème est de nature combinatoire. D'un point de vue calculatoire, trouver un minimum global est donc délicat. D'autre part, Breiman [1996] montre que résoudre un problème combinatoire n'est pas nécessairement approprié. En effet, la résolution naïve d'un tel problème requiert d'abord l'évaluation d'un nombre combinatoire de solutions, puis la sélection de la configuration la mieux adaptée. Lorsque cette procédure est supporté par peu d'exemples, elle peut être très instable, contrairement à des approches régularisées de type ℓ_2 ou ℓ_1.

4.2.4 Sélection « douce » de variables

Nous venons de voir qu'au delà de la motivation combinatoire, la décision de relâcher les contraintes du problème (4.1) se base également sur un fondement statistique.

Ainsi, les poids binaires des vecteurs σ_1 et σ_2 sont remplacés par des variables continues positives. Nous remplaçons également par 1 les variables s_1 et s_2 des contraintes (4.1c) et (4.1d), afin de ne pas multiplier le nombre

d'hyper-paramètres à régler. Enfin, les contraintes de taille sur les vecteurs $\boldsymbol{\sigma}_1$ et $\boldsymbol{\sigma}_2$ n'auront d'effet que si les coefficients du vecteur $\boldsymbol{\beta}$ sont également contraints. Le problème devient

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\,\boldsymbol{\sigma}} & \sum\limits_i \left(y_i - \boldsymbol{x}_i(\boldsymbol{\sigma} \star \boldsymbol{\beta})\right)^2 + \lambda \sum\limits_m \beta_m^2 & & & \text{(4.2a)} \\ \text{s. c.} & \sigma_m = \sigma_{1,\ell}\,\sigma_{2,m} & & \forall \ell,\ m \in G_\ell & \text{(4.2b)} \\ & \sum\limits_\ell d_\ell\,\sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell & \text{(4.2c)} \\ & \sum\limits_m \sigma_{2,m} \leq 1 & \sigma_{2,m} \geq 0 & \forall m\ , & \text{(4.2d)} \end{cases}$$

où le paramètre de régularisation λ est le seul hyper-paramètre à régler.

Nous introduisons une modification supplémentaire. Dans l'équation (4.2a), le terme de pénalisation sur le vecteur $\boldsymbol{\beta}$ – induit par les facteurs $\boldsymbol{\sigma}_1$ et $\boldsymbol{\sigma}_2$ – est dissocié de la fonction de perte et intégré dans le terme de régularisation. Cela permet d'une part d'introduire une fonction de perte générique et de traiter indifféremment des problèmes de régression ou classification. D'autre part, cela permet également d'assurer en partie la convexité du problème d'optimisation. Pour que cette dernière devienne effective, nous introduisons une racine sur les éléments $\sigma_{1,\ell}$ et $\sigma_{2,m}$. Le problème précédent est transformé de la façon suivante

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\,\boldsymbol{\sigma}} & J(\boldsymbol{\beta}) + \lambda \sum\limits_m \dfrac{\beta_m^2}{\sigma_m} & & & \text{(4.3a)} \\ \text{s. c.} & \sigma_m = \sqrt{\sigma_{1,\ell}\,\sigma_{2,m}} & & \forall \ell,\ m \in G_\ell & \text{(4.3b)} \\ & \sum\limits_\ell d_\ell\,\sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell & \text{(4.3c)} \\ & \sum\limits_m \sigma_{2,m} \leq 1 & \sigma_{2,m} \geq 0 & \forall m\ . & \text{(4.3d)} \end{cases}$$

***Remarque* 4.3 —** Pour effectivement assurer la convexité, il convient de considérer le problème d'optimisation (4.3) en fonction des variables $(\boldsymbol{\beta}, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$, c'est-à-dire :

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\,\boldsymbol{\sigma}_1,\,\boldsymbol{\sigma}_2} & J(\boldsymbol{\beta}) + \lambda \sum\limits_\ell \sum\limits_{m \in G_\ell} \dfrac{\beta_m^2}{\sqrt{\sigma_{1,\ell}\,\sigma_{2,m}}} & & \\ & \sum\limits_\ell d_\ell\,\sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell \\ & \sum\limits_m \sigma_{2,m} \leq 1 & \sigma_{2,m} \geq 0 & \forall m\ . \end{cases}$$

Néanmoins, pour des questions de lisibilité, nous préférons continuer à utiliser la formulation en fonction de $(\boldsymbol{\beta}, \boldsymbol{\sigma})$. ◇

***Remarque* 4.4 —** Par continuité en 0, u/v est défini comme $u/0 = \infty$ si $u \neq 0$ et $0/0 = 0$. ◇

Nous obtenons donc un problème où l'on encourage la parcimonie à chaque niveau par le biais des contraintes (4.3c) et (4.3d) sur les normes

ℓ_1 des vecteurs σ_1 et σ_2. Cette parcimonie est ensuite répercutée dans le modèle via l'ajustement des coefficients du vecteur β par σ dans le terme de régularisation de (4.3a).

Si la formulation du problème (4.3) semble désormais satisfaisante, elle peut être davantage généralisée. En effet, afin d'ajuster le degré de pénalité sur les variables et les groupes de variables, nous allons introduire un exposant p sur les éléments de σ_1 et un exposant q sur ceux de σ_2.

$$\begin{cases} \min\limits_{\beta,\sigma} & J(\beta) + \lambda \sum\limits_m \dfrac{\beta_m^2}{\sigma_m} & & & \text{(4.4a)} \\ \text{s. c.} & \sigma_m = \sigma_{1,\ell}^p \, \sigma_{2,m}^q & & \forall \ell,\ m \in G_\ell & \text{(4.4b)} \\ & \sum\limits_\ell d_\ell \, \sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell & \text{(4.4c)} \\ & \sum\limits_m \sigma_{2,m} \leq 1 & \sigma_{2,m} \geq 0 & \forall m\ . & \text{(4.4d)} \end{cases}$$

Intuitivement, plus p (resp. q) est grand, plus le poids $\sigma_{1,\ell}$ (resp. $\sigma_{2,m}$) se comporte comme un poids binaire (cf. figure 4.2).

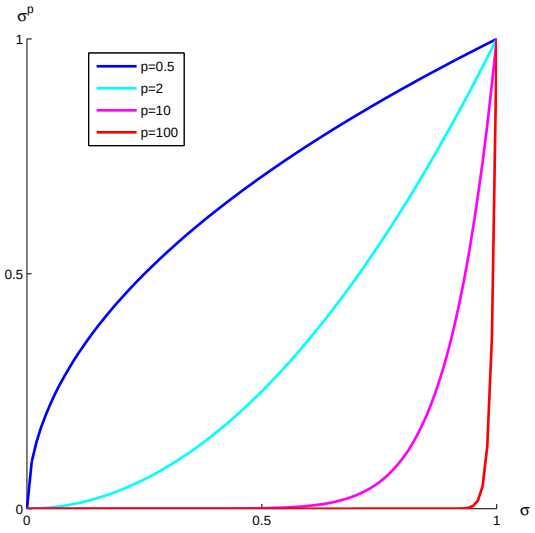


FIGURE 4.2 – *Évolution de $f(\sigma) = \sigma^p$ en fonction de p.*

On peut ainsi régler ces exposants et donner plus ou moins d'influence à la structure de groupe et/ou aux variables, en fonction du problème considéré.

4.2.5 Propriétés

Le problème maintenant formalisé, nous pouvons nous intéresser à quelques propriétés. Nous montrons dans un premier temps la relation qui existe entre (4.4) et les normes mixtes, ce qui nous permettra d'établir les conditions de convexité de ce problème. Nous visualiserons également le comportement de différentes normes sur un exemple à trois variables explicatives.

Relation avec les normes mixtes

Proposition 4.1 *Le problème (4.4) équivaut à minimiser*

$$\min_{\boldsymbol{\beta}} \quad J(\boldsymbol{\beta}) \;+\; \lambda \left(\sum_{\ell} d_{\ell}^{t} \left(\sum_{m \in G_{\ell}} |\beta_m|^{s} \right)^{r/s} \right)^{2/r} , \tag{4.5}$$

avec $s = \dfrac{2}{q+1}$, $r = \dfrac{2}{p+q+1}$ *et* $t = 1 - \dfrac{r}{s}$.

Démonstration. La démonstration de la proposition 4.1 est reportée en annexe A.1. □

La pénalité associée au problème (4.5) est analogue aux pénalités mixtes, présentées en section 3.3.2. Notons que l'exposant 2 externe sur la norme de l'équation (4.5) n'a d'effet que sur la force de la pénalité, mais qu'elle n'influence pas son type.

On peut donc reformuler notre problème initial en s'affranchissant des facteurs σ_1 et σ_2. Ainsi, les contraintes associées à ces facteurs dans le problème (4.4) consistent en fait à pénaliser la norme mixte $\ell_{(r,s)}$ (au carré) des coefficients de l'estimateur $\boldsymbol{\beta}$.

Nous pouvons d'ores et déjà dégager deux cas particuliers :

- $p = 0$ et $q = 1$ correspond à minimiser un problème de type *lasso* ;
- $p = 1$ et $q = 0$ correspond à minimiser un problème de type *group-lasso*.

Convexité

En suivant la remarque 4.3, on considère que le problème d'optimisation (4.4) s'exprime en fonction des variables $(\boldsymbol{\beta}, \sigma_1, \sigma_2)$.

Proposition 4.2 ***Conditions nécessaires de convexité*** *— Le problème (4.4) n'est pas convexe si* $|q| > 1$ *ou si* $|p + q| > 1$.

Démonstration. La proposition 4.1 établit l'équivalence entre les problèmes (4.4) et (4.5). Ce dernier est convexe lorsque $r \geq 1$ et $s \geq 1$ [Zhao et coll., à paraître]. Les conditions nécessaires de convexité sur p et q sont déduites des relations établies dans la proposition 4.1 entre p, q, r, et s :

$$\begin{aligned} s = \frac{2}{q+1} \geq 1 \quad &\Rightarrow \quad q \geq -1 \quad \text{et} \quad q \leq 1 \\ r = \frac{2}{p+q+1} \geq 1 \quad &\Rightarrow \quad p+q \geq -1 \quad \text{et} \quad p+q \leq 1 \ . \end{aligned}$$

Si une des ces conditions n'est pas vérifiée, le problème (4.5) ne peut être convexe. □

Nous avons vu dans la section 3.2 que la convexité d'une norme ℓ_p était induite pour $p \geq 1$. Il n'est donc pas surprenant de retrouver une propriété identique sur chaque niveau d'une norme mixte, par le biais des conditions sur r et s.

Proposition 4.3 ***Conditions suffisantes de convexité*** *— Le problème (4.4) est convexe si* $0 \leq q \leq 1$ *et si* $p + q = 1$.

Démonstration. Un problème qui minimise un critère convexe sur un ensemble convexe est convexe.

1. À condition que $J(\boldsymbol{\beta})$ soit une fonction de perte convexe, et puisque λ est positif, le critère (4.4a) est convexe si lorsque $q = 1 - p$, $f(x, y, z) = \dfrac{x^2}{y^p z^{(1-p)}}$ est convexe pour y et z positifs. Pour cela, nous montrons que $\nabla^2{}_f$, la matrice Hessienne relative à $f(x, y, z)$ peut-être décomposée en somme de deux matrices définie-positives :

$$y^p z^{(1-p)} \nabla^2{}_f = \begin{bmatrix} 2 & -\frac{2xp}{y} & -\frac{2x(1-p)}{z} \\ -\frac{2xp}{y} & \frac{x^2(p+1)}{y^2} & \frac{x^2 p(1-p)}{yz} \\ -\frac{2x(1-p)}{z} & \frac{x^2 p(1-p)}{yz} & \frac{x^2(1-p)^2 + x^2(1-p)}{z^2} \end{bmatrix}$$
$$= 2 \begin{bmatrix} 1 \\ -\frac{xp}{y} \\ \frac{x(p-1)}{z} \end{bmatrix} \begin{bmatrix} 1 \\ -\frac{xp}{y} \\ \frac{x(p-1)}{z} \end{bmatrix}^{\mathrm{t}} + x^2(1-p) \begin{bmatrix} 0 \\ \sqrt{\frac{p}{y}} \\ -\frac{\sqrt{p}}{z} \end{bmatrix} \begin{bmatrix} 0 \\ \sqrt{\frac{p}{y}} \\ -\frac{\sqrt{p}}{z} \end{bmatrix}^{\mathrm{t}}.$$

2. Les contraintes (4.4c) et (4.4d) de positivité et d'inégalité sur les éléments de σ_1 et σ_2 définissent des demi-espaces qui forment des ensembles convexes. Puisque l'intersection d'ensembles convexes forme un ensemble convexe, les contraintes définissent un ensemble admissible convexe. □

Propriété 4.1 ***Convexité et parcimonie*** *— Le problème (4.5) ne peut à la fois être convexe et strictement parcimonieux sur chaque niveau, lorsque la structure de groupe est prise en compte.*

Nous avons également vu dans la section 3.2 que la parcimonie n'était effective que pour des normes ℓ_p où $p \leq 1$.

Les conditions de convexité du problème (4.5) ($r \geq 1$ et $s \geq 1$, cf. proposition 4.2) mettent donc en exergue le fait que la parcimonie sur deux niveaux ne peut être atteinte, avec une norme mixte, que pour r et $s \leq 1$. Dans cette situation, le seul cas convexe correspond au *lasso* [3], avec $r = s = 1$. De fait, les autres configurations entraînent la minimisation d'un problème non-convexe.

Parcimonie

La figure 4.3 présente différentes configurations de normes mixtes. Les deux axes horizontaux représentent le plan (β_1, β_2) associé à un premier groupe, tandis que l'axe vertical (β_3) est associé à un second groupe. Les

3. Nous verrons dans le paragraphe suivant pourquoi le *lasso* ne tient pas compte de la structure de groupe.

ensembles admissibles sont définis avec la norme $\ell_{(r,s)}$ contrainte et pondérée par la taille des groupes :

$$\ell_{(r,s)} = \sum_{\ell} d_\ell^t \left(\sum_{m \in G_\ell} |\beta_m|^s \right)^{r/s} \leq 1 \ .$$

Le tableau 4.1 contient les équivalences entre les normes mixtes $\ell_{(r,s)}$ présentées en figure 4.3, et les valeurs de p et q correspondantes.

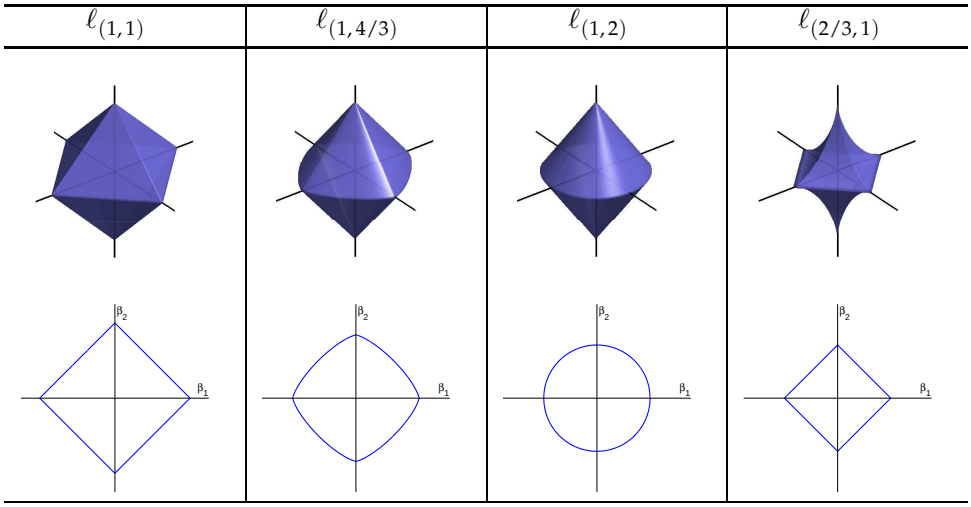


FIGURE 4.3 – *Boules unité pour différentes normes mixtes.*

$\ell_{(r,s)}$	$\ell_{(1,1)}$	$\ell_{(1,4/3)}$	$\ell_{(1,2)}$	$\ell_{(2/3,1)}$
p	0	1/2	1	1
q	1	1/2	0	1

TABLE 4.1 – *Correspondance entre différentes normes mixtes $\ell_{(r,s)}$ et les valeurs de p et q.*

La norme mixte $\ell_{(1,1)}$ (*lasso*) est parcimonieuse aux deux niveaux, mais ne tient pas compte de la structure de groupe : l'ensemble des coefficients est pénalisé de façon identique.

La norme mixte $\ell_{(1,2)}$ (*group-lasso*) est parcimonieuse sur les groupes de variables. Cependant, au sein d'un groupe, la norme ℓ_2 applique un rétrécissement proportionnel sur les variables, et ne favorise pas la parcimonie à ce niveau.

La norme mixte $\ell_{(1,4/3)}$ représente une alternative au *lasso* et au *group-lasso*. D'une part, elle tient compte de la structure de groupe, et est parcimonieuse à ce niveau. De plus, la pénalité $\ell_{4/3}$ appliquée sur les variables d'un groupe est de nature plus restrictive que la norme ℓ_2 du *group-lasso*. Bien que la frontière de l'ensemble admissible soit différentiable dans l'hyperplan (β_1, β_2) associé aux groupes, en $\beta_1 = 0$ et $\beta_2 = 0$, sa courbure est très importante. Cela permet d'obtenir, au sein des groupes, des solutions avec peu de coefficients significatifs.

Lorsque la motivation principale consiste à interpréter un modèle, on peut favoriser des solutions strictement parcimonieuses, et utiliser une pénalité

non-convexe. La norme mixte $\ell_{(2/3,1)}$, particulièrement restrictive au niveau des groupes, illustre ce cas.

4.2.6 Deux approches

Nous avons montré que notre problème s'exprime de deux manières, soit par le biais d'un problème variationnel (4.4), soit par le biais d'un problème régularisé par une norme mixte (4.5).

Nous allons maintenant présenter, pour chaque approche, une façon de résoudre ces problèmes. Dans un premier temps, nous verrons comment optimiser le problème (4.5), dans un cadre d'apprentissage général (régression ou classification). Nous nous concentrerons ensuite sur le problème variationnel (4.4), dans le cadre de la classification et des méthodes à noyaux.

4.3 Approche régularisée via une norme mixte

4.3.1 Formulation dans un espace de fonctions paramétriques

Nous nous situons dans le cadre fonctionnel décrit en section 4.2, utilisé pour formaliser le problème de la pénalisation hiérarchique, à savoir les espaces de fonctions paramétriques. Nous présentons ici un méthode permettant de résoudre une expression simplifiée équivalente au problème (4.5)

$$\min_{\boldsymbol{\beta}} \quad J(\boldsymbol{\beta}) \; + \; \nu \sum_{\ell} d_\ell^t \left(\sum_{m \in G_\ell} |\beta_m|^s \right)^{r/s} \quad . \tag{4.6}$$

En fonction de la perte $J(\boldsymbol{\beta})$ utilisée, on pourra considérer des problèmes de régression ou de classification.

4.3.2 Principe de résolution

Le principe de résolution du problème (4.6) se base sur une approche proposée par Osborne et coll. [2000a] pour résoudre le problème du *lasso*. Avant d'en décrire le principe, nous rappelons le concept d'ensemble actif.

Définition 4.1 *L'ensemble actif courant $\mathcal{A}$ est défini par l'ensemble des variables non-nulles, appelées variables « actives », auxquelles on associe le vecteur $\boldsymbol{\gamma}$:*

$$\begin{aligned} \mathcal{A} &= \{m \,|\, \beta_m \neq 0\} \ , \\ \boldsymbol{\gamma} &= \{\beta_m\}_{m \in \mathcal{A}} \ . \end{aligned}$$

On définit $\forall \ell$ le sous-ensemble des coefficients de $\boldsymbol{\gamma}$ associé au groupe ℓ :

$$J_\ell = \{G_\ell \cap \mathcal{A}\} \ .$$

Notre algorithme comporte deux étapes :

1. À chaque itération, on résout un problème d'optimisation avec $\mathcal{A}$, l'ensemble actif courant

$$\min_{\gamma} \mathcal{L}(\gamma) = J(\gamma) + \nu \sum_{\ell} d_\ell^t \left(\sum_{m \in J_\ell} |\gamma_m|^s \right)^{r/s} , \tag{4.7}$$

   en utilisant la procédure décrite aux étapes A et B de l'algorithme 1.
2. Ensuite, l'ensemble actif des variables est mis à jour de façon incrémentale. On utilise pour cela les étapes C et D de l'algorithme 1.

4.3.3 Conditions d'optimalité

Avant de proposer l'algorithme associé au problème (4.6), nous allons énumérer les conditions d'optimalité du premier ordre sur $\boldsymbol{\beta}$.

1. Lorsque $\beta_m = 0$, $m \in G_\ell$, et $\sum_{m \in G_\ell} |\beta_m| = 0$

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_m} + \nu\, r\, d_\ell^t\, v_j = 0 \ ,$$

   avec $v_j \in [-1, 1]$.
2. Lorsque $\beta_m = 0$, $m \in G_\ell$, et $\sum_{m \in G_\ell} |\beta_m| \neq 0$

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_m} = 0 \ ,$$

   pour $r \geq 1$ et $s \geq 1$.

   ***Remarque* 4.5 —** Lorsque $r < 1$ ou $s < 1$, la forme de la condition d'optimalité est indéterminée. Dans l'approche présentée dans cette section, on se limitera donc aux cas convexes où $r \geq 1$ et $s \geq 1$. ◇

3. Lorsque $\beta_m \neq 0$, $m \in G_\ell$

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_m} + \nu\, r\, d_\ell^t\ \mathrm{sign}(\beta_m)\ |\beta_m|^{r-1} \left(1 + \frac{1}{|\beta_m|^s} \sum_{\substack{m' \in G_\ell \\ m' \neq m}} |\beta_{m'}|^s \right)^{(r-s)/s} = 0 \ .$$

Ces conditions, notamment les deux premières, seront utilisées dans l'algorithme 1 pour tester l'optimalité d'une solution.

4.3.4 Algorithme

L'algorithme est initialisé avec $\mathcal{A} = \emptyset$. La première variable active est sélectionnée selon le processus décrit à l'étape D de l'algorithme 1, page 53.

***Remarque* 4.6 —** Une stratégie différente de celle décrite à l'étape D peut-être utilisée pour sélectionner la variable à intégrer à l'ensemble actif. Par exemple, on peut calculer les paramètres de Lagrange associés aux contraintes, et prendre la variable pour laquelle ce paramètre est le plus petit. ◇

Algorithme 1 : Approche par contraintes actives

A ***Recherche d'une mise à jour candidate à partir de la solution admissible courante*** γ

résoudre $\min_{\mathbf{h}} \mathcal{L}(\gamma + \mathbf{h})$, `// cf. remarque 4.7`

avec $\mathbf{h} \in \mathbb{R}^{|\mathcal{A}|}$ `// direction de descente`

B ***Recherche d'une nouvelle solution admissible*** γ^+

$\gamma^+ = \gamma + \mathbf{h}$

si $\forall m$, $\text{sign}(\gamma_m^+) = \text{sign}(\gamma_m)$

alors

aller à l'étape C

sinon

B1 $\mathcal{S} = \{j : \text{sign}(\gamma_j^+) \neq \text{sign}(\gamma_j)\}$ `// cf. remarque 4.8`

$\forall j \in S$, $\mu_j = -\dfrac{\gamma_j}{h_j}$

$k = \arg\min_{j \in \mathcal{S}} \mu_j$

$\gamma = \gamma + \mu_k \mathbf{h}$ `// annule` γ_k

$\text{sign}(\gamma_k) = -\text{sign}(\gamma_k)$

B2 calculer $\mathbf{h}$ `// cf. étape A`

$\gamma^+ = \gamma + \mathbf{h}$

si $\text{sign}(\gamma_k^+) \neq \text{sign}(\gamma_k)$

alors

$\mathcal{A} \leftarrow \mathcal{A} - \{k\}$

aller à l'étape A.

C ***Test sur l'optimalité de*** γ

si $\forall \beta_k \notin \mathcal{A}$

C1 pour $k \in G_\ell$, et $\sum_{m \in G_\ell} |\beta_m| = 0$, $\left| \dfrac{\partial J(\boldsymbol{\beta})}{\partial \beta_k} \right| \leq \nu\, r\, d_\ell^t$

C2 pour $k \in G_\ell$, et $\sum_{m \in G_\ell} |\beta_m| \neq 0$, $\dfrac{\partial J(\boldsymbol{\beta})}{\partial \beta_k} = 0$

alors

γ est une solution

sinon

aller à l'étape D

D ***Sélection de la variable à ajouter à l'ensemble actif***

D1 $k = \arg\max_{m \notin \mathcal{A}} \dfrac{1}{r\, d_\ell^t} \left| \dfrac{\partial J(\boldsymbol{\beta})}{\partial \beta_m} \right|$ `// ℓ, groupe de la variable k`

D2 $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$

$\gamma = [\gamma, 0]^{\mathrm{T}}$ `// mise à jour de la solution courante`

`// avec` $-\text{sign}\left(\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_k}\right)$, `le signe de la nouvelle composante`

D3 aller à l'étape A

***Remarque* 4.7 —** Dans l'étape A, la dérivée du problème (4.7) est discontinue à cause des valeurs absolues. Nous contournons ce problème en remplaçant $|\gamma_m + h_m|$ par $\text{sign}(\gamma_m)(\gamma_m + h_m)$. Nous pouvons ainsi utiliser des outils d'optimisation continue, basés sur des méthodes de Newton, de quasi-Newton, ou encore de gradient conjugué, selon la taille du problème. ◇

***Remarque* 4.8 —** L'étape B1 consiste à déterminer μ, le plus grand pas dans la direction $\mathbf{h}$, permettant d'obtenir $\forall j \in \mathcal{S}$

$$\text{sign}(\gamma_j + \mu h_j) = \text{sign}(\gamma_j) \ .$$

Seule la variable $k = \arg\min_j -\frac{\gamma_j}{h_j}$ est annulée : $\gamma_k + \mu h_k = 0$. ◇

4.4 Approche variationnelle

Dans cette section, on montre comment enrichir le problème du MKL, présenté en section 3.3.4, pour sélectionner les différentes composantes d'une structure de noyaux. Notre approche est fondée sur la formulation (3.13) [Rakotomamonjy et coll., 2007]. Ce dernier problème peut facilement s'étendre sur deux niveaux, grâce à une adaptation de la formulation variationnelle (4.4).

4.4.1 Contexte

Dans la section 3.3.4, nous avons décrit deux usages du MKL : le premier en sélection de variables, lorsque les données proviennent de différentes sources et qu'un noyau est associé à chaque source ; le second comme alternative à la validation croisée pour choisir l'hyper-paramètre d'une famille de noyaux.

Néanmoins, le MKL ne traite pas les problèmes où, par exemple, plusieurs noyaux se rapportent à une même variable d'entrée. En effet, la norme ℓ_1 appliquée sur les éléments associés aux EHNR ne favorise pas des solutions éliminant tous les noyaux associés à une variable non significative.

4.4.2 Formulation dans un ensemble d'EHNR

Pour prendre en compte de telles situations, nous devons donc définir une structure de groupe sur les noyaux, afin de guider le processus de sélection. Pour cela, nous généralisons la formulation (4.4) en considérant des pénalités dans des EHNR, et non plus dans des espaces fonctionnels

paramétriques. On obtient

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi},\boldsymbol{\sigma}}} & \frac{1}{2}\sum\limits_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i & & & & (4.8a) \\ \text{s. c.} & \sigma_m = \sigma_{1,\ell}^p\,\sigma_{2,m}^q & & & \forall \ell,\ m \in G_\ell & (4.8b) \\ & \sum\limits_\ell d_\ell\,\sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell & & (4.8c) \\ & \sum\limits_m \sigma_{2,m} \leq 1 & \sigma_{2,m} \geq 0 & \forall m & & (4.8d) \\ & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i & \xi_i \geq 0 & \forall i\ . & & (4.8e) \end{cases}$$

Néanmoins, ce problème formulé sur deux niveaux est difficile à optimiser. Nous allons le simplifier en le reformulant uniquement en fonction de $\boldsymbol{\sigma}$, tout en conservant l'information sur la structure de groupe.

Considérons d'abord le changement de variable permettant de transformer $\boldsymbol{\sigma}_2$ en $\boldsymbol{\sigma}$. Lorsque $q \neq 0$, cette transformation est bijective, à condition que $\sigma_{1,\ell} \neq 0$. De plus, si pour le problème (4.8), $\sigma^\star_{1,\ell}$ et $\sigma^\star_{2,m}$ sont des solutions optimales associées à $\sigma_{1,\ell}$ et $\sigma_{2,m}$, alors $\sigma^\star_{1,\ell} = 0 \Rightarrow \sigma^\star_{2,m} = 0$. Ainsi, le problème (4.8) est équivalent à

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\sigma}_1}} & \frac{1}{2}\sum\limits_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i & & & & (4.9a) \\ \text{s. c.} & \sum\limits_\ell d_\ell\,\sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell & & (4.9b) \\ & \sum\limits_\ell \sigma_{1,\ell}^{-p/q} \sum\limits_{m\in G_\ell} \sigma_m^{1/q} \leq 1 & \sigma_m \geq 0 & \forall m & & (4.9c) \\ & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i & \xi_i \geq 0 & \forall i\ . & & (4.9d) \end{cases}$$

Ce nouveau problème est davantage simplifié, en éliminant $\boldsymbol{\sigma}_1$ du processus d'optimisation. On aboutit finalement à

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi},\boldsymbol{\sigma}}} & \frac{1}{2}\sum\limits_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i & & & & (4.10a) \\ \text{s. c.} & \sum\limits_\ell \left(d_\ell^p \left(\sum\limits_{m\in G_\ell} \sigma_m^{1/q}\right)^q\right)^{1/(p+q)} \leq 1 & \sigma_m \geq 0 & \forall m & & (4.10b) \\ & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i & \xi_i \geq 0 & \forall i\ . & & (4.10c) \end{cases}$$

C'est cette dernière formulation que nous utiliserons pour adapter l'algorithme du MKL de Rakotomamonjy et coll. [2007]. On constate que le problème (4.10) ne diffère de la formulation initiale du MKL (3.13) qu'au niveau de la contrainte (4.10b) sur les paramètres $\boldsymbol{\sigma}$, qui s'exprime désormais sous la forme d'une norme mixte.

Proposition 4.4 ***Conditions de convexité —*** *Le problème (4.10) est convexe si et seulement si* $0 \leq q \leq 1$ *et* $0 \leq p+q \leq 1$.

Démonstration. Un problème qui minimise un critère convexe sur un ensemble convexe est convexe.

1. La fonction objectif du problème (4.10) est convexe [cf. Boyd et Vandenberghe, 2004, p. 89].
2. Les contraintes de positivité sur les éléments de $\boldsymbol{\sigma}$ (4.10b) et $\boldsymbol{\xi}$ (4.10c) définissent des ensembles convexe, tout comme la partie gauche de la contrainte (4.10c). La partie gauche de la contrainte (4.10b) définit également un ensemble convexe si

   i) $\left(\sum_{m \in G_\ell} \sigma_m^{1/q} \right)^q$ est une norme, c'est-à-dire si $0 \leq q \leq 1$;

   ii) $\sum_\ell t_\ell^{1/(p+q)}$ est convexe en t_ℓ, c'est-à-dire si $0 \leq p+q \leq 1$. □

Comme dans le cadre des fonctions paramétriques, on peut reformuler le problème (4.10) pour aboutir à la minimisation d'une norme mixte sur les éléments des EHNR.

Proposition 4.5 *Le problème (4.10) est équivalent à*

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi}}} & \frac{1}{2}\left(\sum_\ell d_\ell^t \left(\sum_{m \in G_\ell} \|f_m\|_{\mathcal{H}_m}^s \right)^{r/s} \right)^{2/r} + C \sum_i \xi_i & \text{(4.11a)} \\ \text{s. c.} & y_i \left(\sum_m f_m(\boldsymbol{x}_i) + b \right) \geq 1 - \xi_i \qquad \xi_i \geq 0 \qquad \forall i \ , & \text{(4.11b)} \end{cases}$$

avec $s = \dfrac{2}{q+1}$, $r = \dfrac{2}{p+q+1}$ *et* $t = 1 - \dfrac{r}{s}$.

Démonstration. La démonstration de la proposition 4.5 est reportée en annexe A.2. □

4.4.3 Principe de résolution

Pour résoudre le problème (4.10), nous utilisons un algorithme de type wrapper. Le schéma consiste à considérer deux problèmes imbriqués :

$$\begin{cases} \min\limits_{\boldsymbol{\sigma}} & J(\boldsymbol{\sigma}) & \text{(4.12a)} \\ \text{s. c.} & \sum_\ell \left(d_\ell^p \left(\sum_m \sigma_m^{1/q} \right)^q \right)^{1/(p+q)} \leq 1 \qquad \sigma_m \geq 0 \qquad \forall m \ , & \text{(4.12b)} \end{cases}$$

où $J(\sigma)$ est défini comme la valeur de la fonction objectif, pour le problème

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi}}} & \frac{1}{2}\sum\limits_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} \;+\; C\sum\limits_i \xi_i & \text{(4.13a)} \\ \text{s. c.} & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i)+b\right) \geq 1-\xi_i \qquad \xi_i \geq 0 \qquad \forall i \;. & \text{(4.13b)} \end{cases}$$

Dans une boucle interne, on optimise les paramètres $\{f_m\}$, b et $\boldsymbol{\xi}$ du problème (4.13), en considérant que les coefficients du vecteur $\boldsymbol{\sigma}$ sont fixés. Dans une boucle externe, on optimise les coefficients de $\boldsymbol{\sigma}$ de façon à diminuer la valeur du critère $J(\boldsymbol{\sigma})$, avec les paramètres $\{f_m\}$, b et $\boldsymbol{\xi}$ calculés précédemment.

4.4.4 Conditions d'optimalité

Le Lagrangien associé au problème (4.10) s'écrit en fonction des multiplicateurs $\lambda \geq 0$ et $\boldsymbol{\mu} = (\mu_1,\dots,\mu_M) \geq 0$ associés aux contraintes (4.10b), ainsi que $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_n) \geq 0\ \forall n$, et $\boldsymbol{\eta} = (\eta_1,\dots,\eta_n) \geq 0\ \forall n$, associés aux contraintes (4.10c) :

$$\begin{aligned} \mathcal{L} \;=\; & \frac{1}{2}\sum_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} \;+\; C\sum_i \xi_i \\ & +\; \lambda\left[\sum_\ell \left(d_\ell^p\Big(\sum_{m\in G_\ell}\sigma_m^{1/q}\Big)^q\right)^{1/(p+q)} - 1\right] \;-\; \sum_m \mu_m\sigma_m \\ & -\; \sum_i \alpha_i\left[y_i\left(\sum_m f_m(\boldsymbol{x}_i)+b\right)+\xi_i-1\right] \;-\; \sum_i \eta_i\xi_i \;. \end{aligned}$$

Conditions d'optimalité sur f_m, b et ξ_i

La forme duale du problème (4.13) permet de calculer facilement le gradient $\nabla J(\boldsymbol{\sigma})$. Les dérivées directionnelles de $\mathcal{L}$ en fonction de f_m, b et ξ_i s'annulent lorsque

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial f_m} = \frac{f_m(.)}{\sigma_m} - \sum_i \alpha_i y_i K_m(.,\boldsymbol{x}_i) = 0 \qquad &\Leftrightarrow \qquad f_m(.) = \sigma_m\sum_i \alpha_i y_i K_m(.,\boldsymbol{x}_i) \;, \\ \frac{\partial\mathcal{L}}{\partial b} = -\sum_i \alpha_i y_i = 0 \qquad &\Leftrightarrow \qquad \sum_i \alpha_i y_i = 0 \;, \\ \frac{\partial\mathcal{L}}{\partial \xi_i} = C-\alpha_i-\eta_i = 0 \qquad &\Leftrightarrow \qquad 0 \leq \alpha_i \leq C \;. \end{aligned}$$

En reportant les résultats dans le problème primal (4.13), on obtient une forme duale identique à celle d'un problème SVM

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} & -\frac{1}{2}\sum\limits_{i,j}\alpha_i\alpha_j\, y_i y_j\, \overline{K}_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j) \;+\; \sum\limits_i \alpha_i & \text{(4.14a)} \\ \text{s. c.} & \sum\limits_i \alpha_i y_i = 0 & \text{(4.14b)} \\ & C \geq \alpha_i \geq 0 \qquad \forall i \;, & \text{(4.14c)} \end{cases}$$

où $\overline{K}_\sigma(\boldsymbol{x}, \boldsymbol{x}') = \sum_m \sigma_m K_m(\boldsymbol{x}, \boldsymbol{x}')$.

D'après les propriétés de dualité [Boyd et Vandenberghe, 2004, chap. 5], la valeur optimale de la fonction objective associée au problème primal (4.13) est également la valeur de la fonction objectif duale :

$$J(\sigma) \;=\; -\frac{1}{2} \sum_{i,j} \alpha_i^\star \alpha_j^\star \, y_i y_j \, \overline{K}_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) \;+\; \sum_i \alpha_i^\star \;, \tag{4.15}$$

où les coefficients $\alpha_i^\star$ sont les solutions du problème (4.14).

***Remarque* 4.9 —** L'existence et le calcul des dérivées de $J(\cdot)$ découlent de résultats généraux sur l'analyse des valeurs optimales. Dans Bonnans et Shapiro [1998], le théorème 4.1 montre que l'unicité de la solution $\boldsymbol{\alpha}^\star$ et la différentiabilité de (4.15) assurent la différentiabilité de $J(\sigma)$ [4]. ◇

De plus, les dérivées de $J(\sigma)$ peuvent être calculées comme si les coefficients du vecteur $\boldsymbol{\alpha}^\star$ n'étaient pas dépendants de σ. Les composantes du gradient $\nabla J(\sigma)$ sont donc

$$\frac{\partial J(\sigma)}{\partial \sigma_m} \;=\; -\frac{1}{2} \sum_{i,j} \alpha_i^\star \alpha_j^\star \, y_i y_j \, K_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \;. \tag{4.16}$$

Conditions d'optimalité sur σ

La dérivée directionnelle de $\mathcal{L}$ en fonction de σ_m est

$$\frac{\partial \mathcal{L}}{\partial \sigma_m} \quad = \quad -\frac{1}{2} \frac{\|f_m\|_{\mathcal{H}m}^2}{\sigma_m^2} \;+\; \frac{\lambda}{p+q} \, \sigma_m^{(1-q)/q} \left(d_\ell^{-1} \sum_{m \in G_\ell} \sigma_m^{1/q} \right)^{-p/(p+q)} \;-\; \mu_m \;.$$

En suivant le processus décrit en annexe, A.2, on obtient

$$\sigma_m = \left(\sum_\ell \left(d_\ell^p \, s_\ell^{q+1} \right)^{\frac{1}{p+q+1}} \right)^{-(p+q)} \left(d_\ell^{-1} \, s_\ell \right)^{\frac{p}{p+q+1}} \left(\|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}} \right)^q \;, \tag{4.17}$$

où $s_\ell = \sum_{m \in G_\ell} \|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}$.

D'après le principe de résolution décrit en section 4.4.3, on pourrait utiliser dans la boucle externe de l'algorithme l'équation (4.17) pour mettre à jour le vecteur σ, au lieu d'optimiser le problème (4.12). Néanmoins, cette approche n'assure pas la convergence de l'algorithme et peut engendrer des problèmes numériques, notamment lorsque des éléments de σ sont proches de 0.

Ainsi, comme dans Rakotomamonjy et coll. [2007], nous utilisons le fait que le résultat de la fonction objectif $J(\sigma)$ est optimal pour le problème SVM (4.13). Le vecteur σ est ensuite mis à jour grâce à une descente de gradient projeté.

4. L'unicité de $\boldsymbol{\alpha}^\star$ est garantie, à condition que la matrice de Gram associée au noyau $\overline{K}_\sigma$ soit définie-positive. Dans le cas contraire, on peut forcer cette propriété en ajoutant une perturbation ϵ sur les diagonales des noyaux K_m.

4.4.5 Algorithme

Nous disposons maintenant de tous les éléments pour adapter l'algorithme du MKL de Rakotomamonjy et coll. [2007], selon le procédé décrit en section 4.4.3.

Algorithme 2 : Approche basée sur une méthode de gradient

initialisation de σ
calcul de la solution du problème SVM $\rightarrow J(\sigma)$
répéter
  calcul de la direction $d = -\nabla J(\sigma)$
  répéter
    calcul de d', la projection de d sur l'hyperplan tangent à la surface de l'ensemble admissible, définie par (4.10b)
    calcul du plus petit pas qui annule un coefficient de σ
      $\mathcal{S} = \left\{ j : d'_j < 0 \text{ and } \sigma_j \neq 0 \right\}$
      $\mu = \min_{j \in \mathcal{S}} -\frac{\sigma_j}{d'_j}$
      $k = \arg\min_{j \in \mathcal{S}} -\frac{\sigma_j}{d'_j}$
      $d_k = 0$
    $\sigma^\dagger = \sigma + \mu\, d'$
    projection de $\sigma^\dagger$ sur la surface de l'ensemble admissible
    calcul de la solution du problème SVM $\rightarrow J(\sigma^\dagger)$
    si $J(\sigma^\dagger) < J(\sigma)$
    alors
      $\sigma = \sigma^\dagger$
  tant que $J(\sigma^\dagger) \geq J(\sigma)$
  $\mu^\star = \arg\min_\mu J(\sigma + \mu\, d)$
  $\sigma = \sigma + \mu^\star\, d$
jusqu'à *la convergence*

4.5 Parallèle entre les deux approches

Dans les deux sections précédentes, nous avons introduit deux algorithmes permettant de résoudre les problèmes (4.6) et (4.10). Nous allons maintenant présenter un moyen d'unifier ces deux approches. Dans un premier temps, nous verrons comment dériver la formulation (4.6) de l'algorithme 1, afin de pourvoir utiliser des fonctions noyaux. Ensuite, nous montrerons qu'il est possible d'utiliser l'algorithme 2 avec une fonction de coût quadratique.

4.5.1 Extension de l'approche régularisée via une norme mixte pour la sélection de noyaux

Pour intégrer une pénalité sur une combinaison linéaire de fonctions noyaux, nous posons le problème de régression suivant

$$y_i = \sum_j \beta_j K_m(\boldsymbol{x}_i, \boldsymbol{x}_j) + \epsilon_i \ ,$$

où K_m est le noyau de paramètre m, et ϵ_i l'erreur résiduelle. Afin de combiner m noyaux, le critère général (4.6) s'écrit

$$\min_{\{\boldsymbol{\beta}_m\}_{m=1}^M} \sum_i \left(y_i - \sum_m \sum_j \beta_{m,j} K_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^2 + \nu \sum_m n^t \left(\sum_j |\beta_{m,j}|^s \right)^{r/s} \ . \quad (4.18)$$

On retrouve ainsi un problème similaire au MKL (cf. section 3.3.4). La parcimonie induite sur les noyaux correspond ici à la pénalité ℓ_r appliquée sur la norme des coefficients qui représentent les groupes. La parcimonie induite sur les individus par l'utilisation des SVM, dans le problème du MKL, se retrouve dans quant à elle avec pénalité ℓ_s appliquée sur la norme les individus.

***Remarque* 4.10 —** Cette extension ne permet que de sélectionner des noyaux, sans pouvoir effectuer de regroupement sur ces derniers. En effet, dans l'expression (4.18), m indice les noyaux, tandis que j indice les observations. Afin de considérer des groupes de noyaux, il faudrait ajouter un niveau supplémentaire à l'arborescence. L'expression de la pénalisation devrait être de la forme :

$$\left(\sum_\ell \left(\sum_{m \in G_\ell} \left(\sum_j |\beta_{m,j}|^s \right)^{r/s} \right)^{t/r} \right)^{1/t} \ .$$

L'équivalence avec la formulation variationnelle reste à déterminer. ◇

4.5.2 Extension de l'approche variationnelle à une fonction de coût quadratique

Nous avons vu en section 4.4 que le problème variationnel pouvait être décomposé en deux sous problèmes, correspondants à deux phases « indépendantes » de l'algorithme. Ainsi, le problème (4.13, 4.12) peut être étendu à d'autres fonctions objectifs $J(\sigma)$, dans la mesure où l'on peut évaluer leur valeur et leur gradient. En particulier, on peut considérer un problème de régression en remplaçant la fonction objectif $J(\sigma)$ du problème (4.13) par la minimisation du coût quadratique :

$$\min_{f_1,\ldots,f_M} \quad \sum_i \left(y_i - \sum_m f_m(\boldsymbol{x}_i) \right)^2 + \lambda \sum_m \frac{\|f_m\|^2}{\sigma_m} \ . \quad (4.19)$$

Ici $f_m(\boldsymbol{x}_i)$ peut être une fonction paramétrique définie par $f_m(\boldsymbol{x}_i) = x_i^m \beta_m$. On peut également envisager d'utiliser des fonctions noyaux, comme en section 4.5.1, où $f_m(\boldsymbol{x}_i) = \sum_j \beta_{m,j} K_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Cependant, pour comparer les deux approches, nous nous limiterons aux fonctions paramétriques. Ainsi, pour un vecteur de paramètres σ fixé, la solution paramétrique du problème (4.19) est

$$\boldsymbol{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \ ,$$

où $\mathbf{D}$ est une matrice diagonale de terme général $d_m = \lambda/\sigma_m$. Les termes σ_m sont quant à eux optimisés de la même façon que dans la boucle externe de l'algorithme 2.

4.6 Perspectives

Nous montrons ici comment formuler la *pénalisation hiérarchique* sur des arborescences de hauteur arbitraire, et discutons la possibilité de considérer des graphes acycliques dirigés plutôt que des arborescences.

4.6.1 Arborescences de hauteur arbitraire

À partir des notations précédentes, nous définissons les variables suivantes :

- H est la hauteur de l'arborescence ;
- h indice la hauteur dans une arborescence :
  $h = 0$ correspond à la racine ;
  $h = H$ correspond aux feuilles ;
- m_h indice les nœuds de hauteur h ;
- G_{h,m_h} correspond à l'ensemble des fils du nœud m_h ;
- d_{h,m_h} est le cardinal de l'ensemble G_{h,m_h} , c'est-à-dire :
  $d_{h,m_h} = \sum_{m_{h+1}\in G_{h,m_h}} d_{h+1,m_{h+1}}$;
  $d_{H-1,m_{H-1}} = \mathrm{card}\left(G_{H-1,m_{H-1}}\right)$.

La généralisation du problème variationnel (4.4) est alors

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\boldsymbol{\sigma}} & J(\boldsymbol{\beta}) + \lambda \sum\limits_{m_1} \sigma_{1,m_1}^{-p_1} \sum\limits_{m_2\in G_{1,m_1}} \sigma_{2,m_2}^{-p_2} \quad \cdots \sum\limits_{m_H\in G_{H-1,m_{H-1}}} \sigma_{H,m_H}^{-p_H}\, \beta_{m_H}^2 \\ \text{s. c.} & \sum\limits_{m_1} d_{1,m_1}\, \sigma_{1,m_1} = 1 \\ & \sum\limits_{m_1}\sum\limits_{m_2\in G_{1,m_1}} d_{2,m_2}\, \sigma_{2,m_2} = 1 \\ & \vdots \\ & \sum\limits_{m_1}\sum\limits_{m_2\in G_{1,m_1}} \cdots \sum\limits_{m_H\in G_{H-1,m_{H-1}}} \sigma_{H,m_H} = 1 \\ & \sigma_{h,m_h} \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad m_h \in G_{h-1,m_{h-1}}\,, \forall h \quad . \end{cases}$$

Nous ne faisons ici qu'esquisser une perspective de travail. Les propriétés de cette formulation, en terme de convexité, de parcimonie et de résolution, doivent être étudiées plus en avant. Néanmoins, nous mettons cette perspective en exergue, car c'est par ce biais que nous pourrons accéder

à de nouvelles applications. En effet, dans la problématique concernant la génomique évoquée au chapitre 1, l'ontologie qui permet de définir une structure de groupes sur les gènes comporte plusieurs niveaux.

4.6.2 Graphes acycliques dirigés

Toujours dans le cadre d'applications génomiques, l'ontologie permettant de construire la structure de groupe sur les gènes utilise des graphes acycliques dirigés. Un gène peut être relié à plusieurs processus biologiques. De la même façon, un processus biologique intermédiaire peut lui aussi être issu de processus biologiques différents. Ainsi, pour considérer ce type d'application, il convient de ne pas se limiter à des structures arborescentes qui définissent des partitions exactes sur les groupes, mais d'étendre ce cadre de travail à des groupes non disjoints.

On peut considérer différentes possibilités d'extensions, selon le type d'appartenance d'une variable à plusieurs groupes. Celle-ci peut-être :

– dure : la variable appartient à l'intersection de deux ensembles ;
– floue : la variable appartient à des degrés divers à l'un ou l'autre des ensembles associés.

Dans le second cas, on pourra encore distinguer deux problèmes : ceux où les degrés d'appartenance sont connus et définis *a priori*, et ceux où ils doivent être estimés dans l'apprentissage.

4.7 Synthèse

Dans ce chapitre, nous avons détaillé le cadre théorique de la *pénalisation hiérarchique* pour des structures arborescentes de deux niveaux. En particulier, nous avons montré la relation entre la formulation variationnelle initiale, et les normes mixtes.

Nous avons aussi examiné les propriétés de ce modèle en terme de convexité et de parcimonie. Cela nous a permis de dégager une caractéristique importante, qui spécifie que la parcimonie ne peut être atteinte sur chaque niveau que si l'on considère un problème non-convexe (propriété 4.1). Ainsi, en fonction de l'application considérée, il faut envisager de sacrifier la convexité au profit de l'interprétabilité. Nous illustrerons ce cas dans le chapitre 5, sur les expériences relatives aux interfaces cerveau-machine.

Nous avons également proposé deux méthodes de résolution, où la première utilise l'expression de la norme mixte dans un algorithme de type « contraintes actives », et où la seconde utilise la formulation variationnelle dans un algorithme de type « wrapper ». Nous avons établi le parallèle entre ces deux approches afin de pouvoir les comparer.

Enfin, nous avons discuté de l'extension du formalisme de la *pénalisation hiérarchique* pour des arborescences de hauteurs arbitraires et des graphes acycliques dirigés, ce qui nous permettrait de pouvoir traiter de nouvelles applications, en particulier des applications génomiques. Il est maintenant temps de voir la *pénalisation hiérarchique* en action.

Applications

5

Sommaire

5.1 Introduction

Dans ce chapitre, nous présentons les résultats obtenus par la *pénalisation hiérarchique* sur différents jeux de données.

Tout d'abord, nous comparons les algorithmes 1 et 2, respectivement basés sur la méthode de contraintes actives et la méthode de gradient, sur deux jeux de données simulés, pour une fonction de perte quadratique régularisée par la norme convexe $\ell_{(1,4/3)}$. Dans la première simulation, les variables sont catégorielles tandis que dans la seconde, elles sont continues. Nous comparons les performances en apprentissage et en test, l'influence des variables et des groupes sélectionnés, ainsi que les temps de calculs obtenus avec chaque algorithme.

Nous décrivons ensuite les protocoles de deux problèmes relatifs aux interfaces cervau-machine. Sur chaque problème, les performances en prédiction mais également en sélection seront détaillées. Ces expériences mettent en évidence le fait que la régularisation d'un problème par une norme mixte non-convexe permet d'obtenir des solutions plus interprétables qu'avec des méthodes convexes, pour des performances similaires.

Pour terminer, nous discuterons du bien fondé d'utiliser ou non un seuillage avec l'algorithme de gradient, afin d'éliminer des variables dont les poids sont très faibles. Nous illustrerons cela sur l'un des problèmes d'interfaces cervau-machine.

5.2 Comparaison des deux algorithmes

5.2.1 Cadre de comparaison

Cette section a pour but de définir un cadre pour comparer les deux approches, en fonction d'un paramètre régularisation. En effet, les critères minimisés par les deux algorithmes sont similaires, mais la norme mixte utilisée par l'algorithme 2, basé sur la méthode de gradient, est élevée au carré.

Nous rappelons dans un premier temps les deux problèmes considérés. Dans l'algorithme 1, le critère minimisé est :

$$C_1(\nu, \boldsymbol{\beta}) \;=\; \min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \;+\; \nu\, P(\boldsymbol{\beta})\,,$$

où $J(\boldsymbol{\beta})$ est la fonction de perte quadratique et

$$P(\boldsymbol{\beta}) \;=\; \sum_{\ell} d_\ell^{1/4} \left(\sum_{m \in G_\ell} |\beta_m|^{4/3} \right)^{3/4} .$$

Sous sa forme duale, le critère minimisé par l'algorithme 2 est régularisé par une norme mixte et, d'après la proposition 4.1, s'écrit :

$$C_2(\lambda, \boldsymbol{\beta}) \;=\; \min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \;+\; \lambda\, P(\boldsymbol{\beta})^2\,.$$

Dans la mesure où la pénalité du critère de l'algorithme 2 est élevée au carré, il convient de trouver une relation entre ν et λ. Ainsi, nous pourrons

comparer les solutions obtenues par les deux algorithmes lorsque le paramètre de régularisation varie. Pour le critère $C_2(\lambda, \boldsymbol{\beta})$, nous proposons d'utiliser :

$$\lambda^\star \;=\; \frac{1}{2}\,\frac{\nu}{P(\boldsymbol{\beta}^\star)}\,, \tag{5.1}$$

où $\boldsymbol{\beta}^\star = \arg\min_\beta C_1(\nu, \boldsymbol{\beta})$. La justification de ce choix est reportée en annexe A.3.

***Remarque* 5.1** — L'expression de $\lambda^\star$ repose sur la condition que l'algorithme 1 utilisé pour déterminer $\boldsymbol{\beta}^\star$ retourne une valeur optimale. Nous constaterons dans la section suivante que les valeurs de ν permettant de minimiser l'erreur de test pour les deux algorithmes peuvent être légèrement différentes. ◇

5.2.2 Modèle I

Protocole expérimental

Pour construire cette simulation, nous reprenons le protocole de génération de données utilisé par Yuan et Lin [2006] pour le *group-lasso*. Dans un premier temps, 15 variables aléatoires ont été simulées selon une loi normale multivariée : $(X_1, \ldots, X_{15}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, où la covariance entre la variable m et la variable m' est définie par la valeur $(1/2)^{|m-m'|}$. Chaque variable a ensuite été divisée en trois catégories, en fonction des fractiles de la loi normale. Ainsi, la catégorie d'une observation vaut :

$$\begin{cases} 1 & \text{si } x_i^m < \phi^{-1}(1/3) \\ 2 & \text{si } x_i^m > \phi^{-1}(2/3) \\ 3 & \text{si } x_i^m \in [\phi^{-1}(1/3),\, \phi^{-1}(2/3)]. \end{cases}$$

Les données sont ensuite transformées en données binaires, selon le procédé décrit page 29, par le biais du tableau 3.1. Notre matrice d'observations contient donc 45 variables, réparties sur 15 groupes. Le vecteur de réponse a ensuite été défini par :

$$\begin{aligned} \mathbf{y} \;=\; & -\,1.2\, I(\boldsymbol{x}^1 = 1) \;+\; 1.8\, I(\boldsymbol{x}^1 = 2) \\ & +\; 0.5\, I(\boldsymbol{x}^3 = 1) \;+\; \quad I(\boldsymbol{x}^3 = 2) \\ & +\; \quad I(\boldsymbol{x}^5 = 1) \;+\; \quad I(\boldsymbol{x}^5 = 2) \;+\; \boldsymbol{\epsilon}\,, \end{aligned}$$

où I représente la fonction indicatrice, et où $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{var}(\boldsymbol{\epsilon}))$ est un aléa dont la variance est choisie de façon à ce que le rapport signal sur bruit soit égal à 1.8 dB :

$$10 \log_{10}\left(\frac{\mathrm{var}(\mathbf{y})}{\mathrm{var}(\boldsymbol{\epsilon})}\right) \;=\; 1.8\,.$$

Nous avons généré 600 observations selon ce schéma. L'ensemble d'apprentissage est composé de 400 observations choisies aléatoirement. Ainsi, l'ensemble de test contient les 200 observations restantes. Le processus a été répété 100 fois.

Résultats

Nous nous sommes intéressés aux résultats obtenus pour les valeurs de ν permettant de minimiser l'erreur moyenne de test sur les 100 répétitions, à savoir $\nu = 15$ pour l'algorithme 1 et $\nu = 30$ pour l'algorithme 2. Nous avons également regardé le comportement des algorithmes pour des valeurs de correspondant à un sous-apprentissage ($\nu = 0.1$ et $\nu = 5$), et à un sur-apprentissage ($\nu = 100$ et $\nu = 150$).

Les résultats moyens concernant les erreurs d'apprentissage et de test, ainsi que les valeurs des pénalités et des critères sont reportés sur tableau 5.1 avec

$$\begin{aligned}
\bar{J}_t(\boldsymbol{\beta}) &= \frac{1}{100 \times 200}\ J(\mathbf{X}_t, \boldsymbol{\beta}) \\
\bar{J}_a(\boldsymbol{\beta}) &= \frac{1}{100 \times 400}\ J(\mathbf{X}_a, \boldsymbol{\beta}) \\
\bar{P}(\boldsymbol{\beta}) &= \frac{1}{100}\ P(\boldsymbol{\beta}) \\
\bar{C}(\boldsymbol{\beta}) &= \frac{400}{100}\ \bar{J}_a(\boldsymbol{\beta}) + \nu\, \bar{P}(\boldsymbol{\beta})\ ,
\end{aligned}$$

où $\mathbf{X}_t$ représente la matrice d'observations de l'ensemble de test et $\mathbf{X}_a$ la matrice d'observations de l'ensemble d'apprentissage. On constate sur tableau 5.1 que l'algorithme 1 de contraintes actives parvient à mieux minimiser le critère.

	ν	$\bar{J}_t(\boldsymbol{\beta})$	$\bar{J}_a(\boldsymbol{\beta})$	$\bar{P}(\boldsymbol{\beta})$	$10^{-2}\times \bar{C}(\boldsymbol{\beta})$
Alg. 1	0.1	4.53 ± 2.45	3.90 ± 2.05	13.46 ± 2.37	7.81 ± 4.11
Alg. 2		4.53 ± 2.45	3.90 ± 2.05	13.44 ± 2.37	7.81 ± 4.11
Alg. 1	5	4.37 ± 2.40	3.96 ± 2.06	8.22 ± 1.81	8.33 ± 4.18
Alg. 2		4.41 ± 2.42	3.93 ± 2.05	9.62 ± 2.16	8.34 ± 4.19
Alg. 1	15	$\mathbf{4.30 \pm 2.36}$	$\mathbf{4.12 \pm 2.10}$	$\mathbf{4.52 \pm 0.82}$	$\mathbf{8.92 \pm 4.27}$
Alg. 2		4.32 ± 2.38	4.06 ± 2.08	5.56 ± 1.13	8.95 ± 4.27
Alg. 1	30	4.32 ± 2.34	4.25 ± 2.13	3.19 ± 0.37	9.47 ± 4.31
Alg. 2		$\mathbf{4.30 \pm 2.35}$	$\mathbf{4.17 \pm 2.11}$	$\mathbf{3.88 \pm 0.53}$	$\mathbf{9.51 \pm 4.32}$
Alg. 1	100	4.70 ± 2.32	4.70 ± 2.15	1.71 ± 0.25	11.11 ± 4.34
Alg. 2		4.49 ± 2.32	4.47 ± 2.15	2.28 ± 0.27	11.22 ± 4.34
Alg. 1	150	5.22 ± 2.31	5.24 ± 2.13	0.84 ± 0.26	11.74 ± 4.38
Alg. 2		4.93 ± 2.31	4.93 ± 2.13	1.31 ± 0.36	11.83 ± 4.39

TABLE 5.1 – *Résultats moyen et écart-types obtenus sur les* 100 *répétitions pour le modèle I. En gras, les valeurs relatives au paramètre de régularisation qui minimise l'erreur de test.*

Les résultats en sélection de groupes et de variables sont présentés sur le tableau 5.2, ainsi que sur les figures 5.1 et 5.2 de la page 69, sur lesquelles la boite de gauche correspond aux résultats de l'algorithme 1, et celle de droite aux résultats de l'algorithme 2. On y voit que l'algorithme 1 tend à être plus sévère en ce qui concerne la sélection de variables et de groupes.

Plus le paramètre de régularisation augmente, et plus la différence entre les nombres de groupes et de variables sélectionnés par les deux algorithmes est marquée. On observe également ces écarts sur les valeurs

	ν	# Groupes	# Variables	$\bar{P}(\boldsymbol{\beta})$
Alg. 1	0.1	14.99 ± 0.10	40.50 ± 3.01	13.46 ± 2.37
Alg. 2		14.98 ± 0.14	42.90 ± 1.74	13.44 ± 2.37
Alg. 1	5	12.28 ± 1.70	27.72 ± 6.03	8.22 ± 1.81
Alg. 2		11.87 ± 1.97	31.46 ± 5.72	9.62 ± 2.16
Alg. 1	15	$\mathbf{5.54 \pm 2.01}$	$\mathbf{13.53 \pm 4.44}$	$\mathbf{4.52 \pm 0.82}$
Alg. 2		6.75 ± 2.07	17.00 ± 5.50	5.56 ± 1.13
Alg. 1	30	2.13 ± 0.94	6.23 ± 2.60	3.19 ± 0.37
Alg. 2		$\mathbf{5.73 \pm 3.68}$	$\mathbf{15.59 \pm 11.03}$	$\mathbf{3.88 \pm 0.53}$
Alg. 1	100	1.00 ± 0.00	3.00 ± 0.00	1.71 ± 0.25
Alg. 2		3.37 ± 4.72	9.72 ± 13.79	2.28 ± 0.27
Alg. 1	150	1.00 ± 0.00	3.00 ± 0.00	0.84 ± 0.26
Alg. 2		3.69 ± 5.06	10.59 ± 14.77	1.31 ± 0.36

TABLE 5.2 – *Résultats moyens et écart-types en sélection sur les* 100 *répétitions pour le modèle I. En gras, les valeurs relatives au paramètre de régularisation qui minimise l'erreur de test.*

des pénalités $\bar{P}(\boldsymbol{\beta})$, rappelées sur le tableau 5.2. Pour l'algorithme 1, les nombres de groupes et de variables sélectionnés décroissent avec $\bar{P}(\boldsymbol{\beta})$. Pour l'algorithme 2, de nombreux coefficients très proches de 0 sont conservés, en particulier pour de petites valeurs de $\bar{P}(\boldsymbol{\beta})$. Si l'on compare maintenant $\bar{P}(\boldsymbol{\beta})$ pour $\nu = 15$ pour l'algorithme 1, et $\nu = 30$ pour l'algorithme 2, on voit que, pour une pénalité moyenne donnée, l'algorithme 1 est plus parcimonieux. Toutes ces observations suggèrent d'explorer une stratégie de seuillage pour l'algorithme 2, ce que nous feront en section 5.4.

5.2.3 Modèle II

Protocole expérimental

Pour cette seconde simulation, nous reprenons le protocole de génération de données de [Yuan et Lin, 2006], cette fois ci sur des variables continues. Dans un premier temps, 16 variables aléatoires ont été indépendamment simulées selon une loi normale standardisée : $(X_1, \ldots, X_{17}) \sim \mathcal{N}(0, 1)$. Chaque variable de la matrice d'observation a ensuite été définie par $\boldsymbol{x}^m = \left(\boldsymbol{x}^m + \boldsymbol{x}^{17}\right) / \sqrt{2}$, pour $m = \{1, \ldots, 16\}$. Le vecteur de réponse associé à cette matrice d'observation est :

$$\begin{aligned} \mathbf{y} \quad = \quad & \boldsymbol{x}^3 \; + \; \left(\boldsymbol{x}^3\right)^2 \; + \quad \left(\boldsymbol{x}^3\right)^3 \; + \\ & \frac{2}{3}\, \boldsymbol{x}^6 \; - \; \left(\boldsymbol{x}^6\right)^2 \; + \; \frac{1}{3}\, \left(\boldsymbol{x}^6\right)^3 \; + \; \boldsymbol{\epsilon} \; , \end{aligned}$$

où $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 4)$. Les monômes d'ordre 1, 2 et 3 associés à la variable m forment un groupe. Ainsi, nous disposons de 48 variables, réparties sur 16 groupes.

De la même façon que dans la simulation précédente, nous avons généré 600 observations selon ce schéma, puis 400 observations ont été utilisées pour l'apprentissage, et 200 pour mesurer l'erreur de test. Le processus a été répété 100 fois.

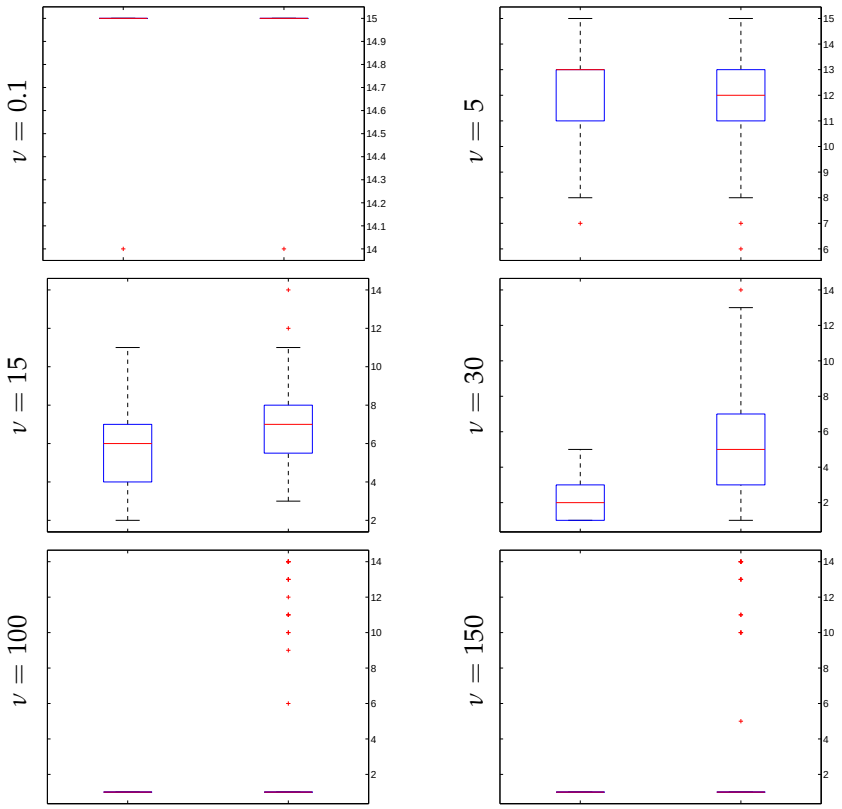


FIGURE 5.1 – *Répartition du nombre de groupes sélectionnés en fonction de ν pour le modèle I.*

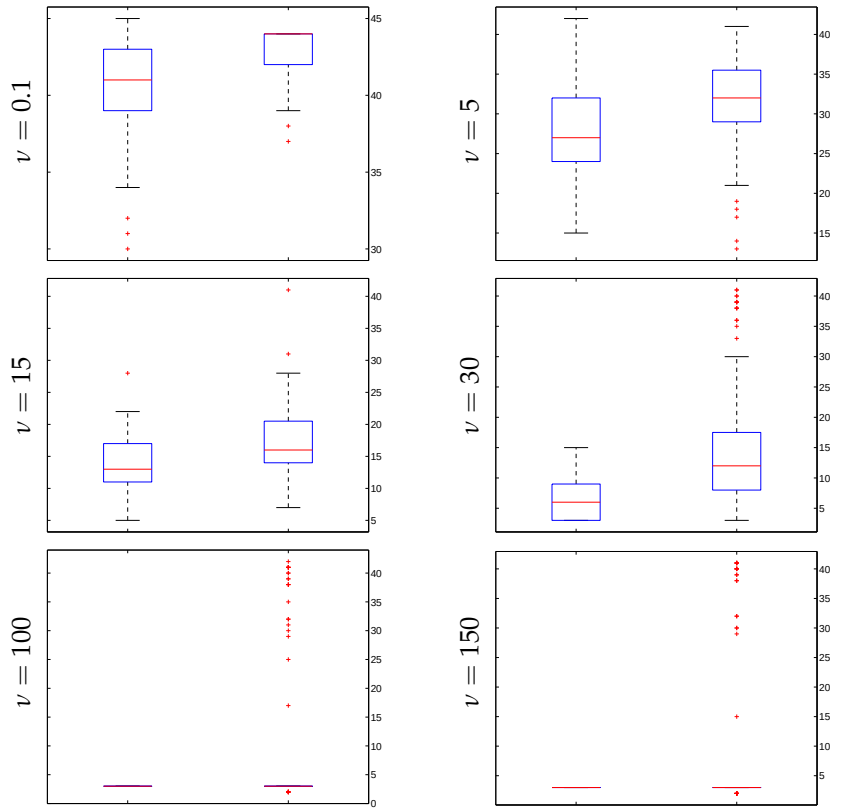


FIGURE 5.2 – *Répartition du nombre de variables sélectionnées en fonction de ν pour le modèle I.*

Résultats

Dans le tableau 5.3, nous reportons les erreurs moyennes en apprentissage et en test, ainsi que que les valeurs moyennes des pénalités et des critères. Les valeurs de ν permettant d'obtenir les erreurs moyennes de test les plus faibles sont $\nu = 10^{3/2}$ pour l'algorithme 1 et $\nu = 10^{7/4}$ pour l'algorithme 2. Les résultats associés aux valeurs de paramètres correspondant à un sous-apprentissage ($\nu = 10^{-2}$ et $\nu = 1$), et à un sur-apprentissage ($\nu = 10^{5/2}$ et $\nu = 10^{3}$) sont également indiqués. Les indicateurs $\bar{J}_{\mathrm{t}}(\boldsymbol{\beta})$, $\bar{J}_{\mathrm{a}}(\boldsymbol{\beta})$, $\bar{P}(\boldsymbol{\beta})$ et $\bar{C}(\boldsymbol{\beta})$ sont définis comme à la section précédente. Le tableau 5.4 présente les résultats obtenus en sélection.

On peut constater que les deux algorithmes se comportent comme lors de la simulation précédente. En effet, à mesure que ν croît, l'algorithme 1 parvient à mieux minimiser le critère.

Ici aussi, l'algorithme 1 pénalise plus sévèrement les coefficients que l'algorithme 2 lorsque la pénalisation augmente. En conséquence, pour des valeurs de ν importantes, les nombres de groupes et de variables sélectionnés sont en moyenne légèrement supérieurs pour l'algorithme 2, ce qui révèle ici la présence de nombreux coefficients proches de 0.

	ν	$\bar{J}_{\mathrm{t}}(\boldsymbol{\beta})$	$\bar{J}_{\mathrm{a}}(\boldsymbol{\beta})$	$\bar{P}(\boldsymbol{\beta})$	$10^{-2}\times\ \bar{C}(\boldsymbol{\beta})$
Alg. 1	10^{-2}	4.60 ± 0.51	3.49 ± 0.27	16.05 ± 1.35	6.98 ± 0.54
Alg. 2		4.60 ± 0.51	3.49 ± 0.27	16.05 ± 1.35	6.98 ± 0.54
Alg. 1	1	4.55 ± 0.50	3.49 ± 0.27	15.54 ± 1.33	7.14 ± 0.54
Alg. 2		4.54 ± 0.51	3.50 ± 0.27	15.30 ± 1.41	7.14 ± 0.54
Alg. 1	$10^{6/4}$	$\mathbf{4.14 \pm 0.47}$	$\mathbf{3.78 \pm 0.27}$	$\mathbf{10.17 \pm 1.06}$	$\mathbf{10.78 \pm 0.65}$
Alg. 2		4.17 ± 0.47	3.72 ± 0.27	10.79 ± 1.05	10.84 ± 0.65
Alg. 1	$10^{7/4}$	4.18 ± 0.50	3.93 ± 0.28	9.44 ± 1.00	13.17 ± 0.81
Alg. 2		$\mathbf{4.13 \pm 0.47}$	$\mathbf{3.85 \pm 0.28}$	$\mathbf{9.88 \pm 1.00}$	$\mathbf{13.25 \pm 0.81}$
Alg. 1	$10^{5/2}$	6.68 ± 1.44	6.28 ± 0.41	6.73 ± 0.94	33.84 ± 3.06
Alg. 2		5.83 ± 1.38	5.46 ± 0.87	7.46 ± 1.02	34.52 ± 3.23
Alg. 1	10^{3}	18.28 ± 5.48	18.06 ± 1.24	2.93 ± 0.83	65.41 ± 8.91
Alg. 2		13.59 ± 4.36	13.04 ± 1.86	4.12 ± 0.98	67.25 ± 9.57

TABLE 5.3 – *Résultats moyen et écart-types obtenus sur les* 100 *répétitions pour le modèle II. En gras, les valeurs relatives au paramètre de régularisation qui minimise l'erreur de test.*

5.2.4 Temps de calcul

Les temps de calcul moyens sont reportés sur le tableau 5.5. Dans l'algorithme 1 de contraintes actives, une variable est ajoutée à chaque itération. En règle générale, plus le paramètre de régularisation augmente, moins l'algorithme 1 sélectionne de variables, et plus il se termine rapidement. À l'inverse, dans l'algorithme 2 de gradient, les variables sont éliminées une à une. Plus le paramètre de régularisation augmente, plus il y a de variables éliminées. Cependant, l'algorithme 2 nécessite plus d'itérations pour converger pour des valeurs de pénalisation légèrement inférieures à celles permettant de minimiser l'erreur de test.

	ν	# Groupes	# Variables	$\bar{P}(\boldsymbol{\beta})$
Alg. 1	10^{-2}	16.00 ± 0.00	47.99 ± 0.10	16.05 ± 1.35
Alg. 2		16.00 ± 0.00	48.00 ± 0.00	16.05 ± 1.35
Alg. 1	1	16.00 ± 0.00	47.90 ± 0.33	15.54 ± 1.33
Alg. 2		15.58 ± 0.54	43.40 ± 2.57	15.30 ± 1.41
Alg. 1	$10^{6/4}$	$\mathbf{7.90 \pm 1.87}$	$\mathbf{18.86 \pm 4.64}$	$\mathbf{10.17 \pm 1.06}$
Alg. 2		7.00 ± 2.16	16.49 ± 5.19	10.79 ± 1.05
Alg. 1	$10^{7/4}$	4.06 ± 1.29	11.53 ± 3.59	9.44 ± 1.00
Alg. 2		$\mathbf{4.34 \pm 2.50}$	$\mathbf{10.81 \pm 7.05}$	$\mathbf{9.88 \pm 1.00}$
Alg. 1	$10^{5/2}$	2.00 ± 0.00	5.99 ± 0.10	6.73 ± 0.94
Alg. 2		2.08 ± 0.71	6.23 ± 2.11	7.46 ± 1.02
Alg. 1	10^{3}	1.02 ± 0.14	3.06 ± 0.42	2.93 ± 0.83
Alg. 2		1.64 ± 0.59	4.87 ± 1.78	4.12 ± 0.98

TABLE 5.4 – *Résultats moyens et écart-types en sélection sur les* 100 *répétitions pour le modèle II. En gras, les valeurs relatives au paramètre de régularisation qui minimise l'erreur de test.*

Pour les paramètres de régularisation correspondant aux erreurs de tests les plus faibles, et sur ces expériences comportant peu de variables, l'algorithme 2 semble converger plus rapidement que l'algorithme 1. Sur le modèle I, l'algorithme 2 a l'avantage, bien que le nombre de variables sélectionnées par les deux algorithmes soit équivalent. Cela est confirmé avec le modèle II, où l'algorithme 1 (qui ajoute en moyenne 18 variables) produit des solutions moins parcimonieuses que l'algorithme 2 (qui n'en conserve que 10 en moyenne, et en élimine donc environ 38).

ν	Modèle I	
	Alg. 1	Alg. 2
0.1	0.23 ± 0.30	0.11 ± 0.08
5	0.50 ± 0.23	0.22 ± 0.04
15	$\mathbf{0.38 \pm 0.15}$	0.17 ± 0.03
30	0.17 ± 0.10	$\mathbf{0.11 \pm 0.03}$
100	0.10 ± 0.21	0.09 ± 0.02
150	0.09 ± 0.08	0.09 ± 0.03

ν	Modèle II	
	Alg. 1	Alg. 2
10^{-2}	0.16 ± 0.00	0.01 ± 0.00
1	0.26 ± 0.24	0.22 ± 0.11
$10^{3/2}$	$\mathbf{0.37 \pm 0.15}$	0.22 ± 0.03
$10^{7/4}$	0.18 ± 0.13	$\mathbf{0.16 \pm 0.04}$
$10^{5/2}$	0.03 ± 0.01	0.13 ± 0.09
10^{3}	0.07 ± 0.05	0.10 ± 0.00

TABLE 5.5 – *Temps de calcul moyens pour les modèles I et II. En gras, les temps de calculs obtenus pour le paramètre de régularisation qui minimise l'erreur de test.*

Bien que l'algorithme 1, basé sur les contraintes actives, soit plus stable numériquement et plus parcimonieux que l'algorithme 2, basé sur la méthode de gradient, nous n'avons donc pas d'argument définitif permettant de privilégier une de ces deux approches.

***Remarque* 5.2 —** Dans l'algorithme basé sur la méthode de gradient, estimer $\boldsymbol{\beta}$ nécessite l'inversion de la matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D})$, où le terme général de la matrice diagonale $\mathbf{D}$ est $d_m = \nu/\sigma_m$ (cf. section 4.5). La stratégie utilisée pour inverser cette matrice influe sur la qualité de l'estimation, et donc de la prédiction. C'est pourquoi cet algorithme est moins stable que celui basé sur les contraintes actives. ◇

5.3 Interfaces Cerveau-Machine

Pour les deux problèmes considérés ici, nous utilisons l'approche variationnelle associée à l'algorithme 2. Si au regard des simulations effectuées, l'algorithme 1, basé sur les contraintes actives, possède des qualités certaines, il ne peut être utilisé que pour une pénalité convexe (cf. remarque 4.5). De plus, dans le cadre de méthodes non paramétriques, on ne peut que sélectionner des noyaux, sans effectuer de regroupement sur ces derniers (cf. remarque 4.10).

Nous désignerons désormais l'algorithme 2 par l'acronyme CKL, pour Composite Kernel Learning [Szafranski et coll., 2008b]. Ainsi, $\text{CKL}_{1/2}$ fait référence à l'algorithme dans le cas où $p = q = 1/2$, c'est-à-dire pour une régularisation $\ell_{(1,4/3)}$. Cette version est convexe, contrairement au CKL_1 qui fait référence à l'algorithme lorsque $p = q = 1$, c'est-à-dire pour une régularisation $\ell_{(2/3,1)}$. Nous comparons ces deux versions au MKL, implémenté lui aussi par l'algorithme CKL, avec $p = 0$ et $q = 1$. Les noyaux utilisés ici consistent en une transformation linéaire : $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$.

5.3.1 Contexte

Dans les problèmes liés aux Interfaces Cerveau-Machine [1] (ICM), on mesure, avec des capteurs, l'activité cérébrale d'un sujet en présence d'un stimulus particulier. Nous nous intéressons ici aux interfaces qui utilisent des capteurs non intrusifs. Dans les deux cas étudiés, des signaux électroencéphalogrammes (EEG) sont mesurés au moyen de 64 électrodes appelées également canaux, réparties sur le cerveau. Cette répartition est illustrée sur la figure 5.3.

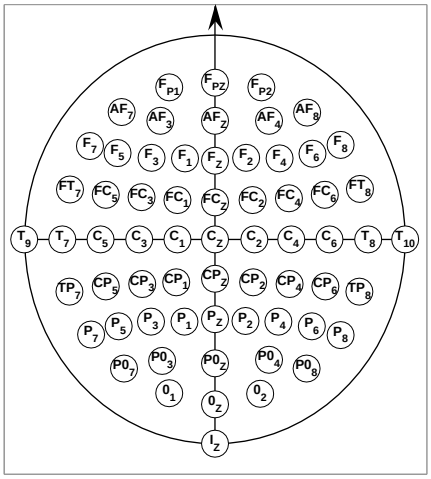


Figure 5.3 – *Répartition des électrodes sur le cerveau pour les problèmes d'interfaces cerveau-machine considérés. La flèche représente la direction frontale.*

1. Brain-Computer Interfaces (BCI), en anglais.

5.3.2 Problème I

Protocole expérimental

L'interface « BCI P300 speller », a été développée par Farwell et Donchin [1998] pour épeler des mots. Elle est basée sur l'apparition de potentiels évoqués dans les EEG, en réponse à un stimulus visuel.

Dans le protocole associé au BCI P300 speller, un patient doit regarder un tableau composé de 6 lignes et de 6 colonnes de caractères. Ce tableau est représenté sur la figure 5.4. Après que le patient ait choisi un caractère, les lignes ou les colonnes du tableau sont illuminées aléatoirement 12 fois. Lorsqu'une ligne ou une colonne illuminée contient le caractère choisi, le patient doit compter. Le fait de compter génère un potentiel évoqué. Ce potentiel appelé P300 apparaît 300 ms après le stimulus visuel. Le processus est répété 15 fois par caractère, car le rapport signal sur bruit de ce type de signaux est faible.

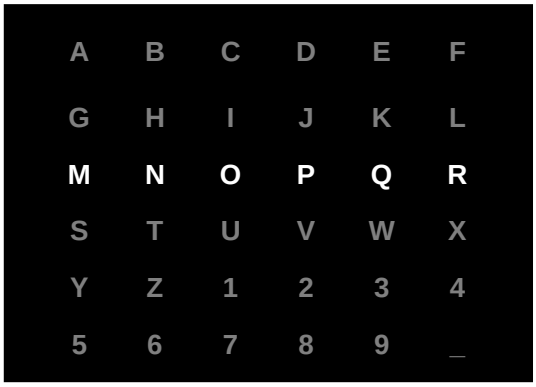


FIGURE 5.4 – *Tableau d'épellation des mots.*

Nous disposons de 7560 signaux, issus d'une base de données d'une compétition sur les interfaces cerveau-machine [Blankertz et coll., 2004], pré-traités de la même façon que dans Rakotomamonjy et Guigue [2008]. Puisque le potentiel apparaît autour de 300 ms après le stimulus visuel, la fenêtre temporelle considérée s'étend jusque 666 ms après le stimulus. Pour réduire le rapport signal sur bruit, chaque canal a été filtré par un passe-bande d'ordre 8 pour des fréquences de coupure de 0.1 Hz et 10 Hz. Le signal est ensuite décimé en accord avec la fréquence de coupure haute. Chaque signal est ainsi composé de 14 fenêtres temporelles. Les 6^e et 7^e fenêtres sont respectivement centrées sur 300 et 350 ms environ.

Finalement, nous disposons de 7560 signaux, de dimension 896 = 14 fenêtres temporelles $\times$ 64 électrodes. L'arborescence associée à ce schéma est représentée sur la figure 5.5. À chaque signal est associé une réponse, qui vaut $+1$ si le signal contient un P300, et -1 sinon.

Notre objectif consiste à reconnaître la présence ou l'absence d'un P300, parmi les signaux mesurés De plus, nous souhaitons identifier les électrodes significatives d'une part, et les fenêtres temporelles significatives d'autre part, qui permettent d'effectuer cette distinction.

Parmi les 7560 signaux, 567 exemples d'apprentissage ont été choisis aléa-

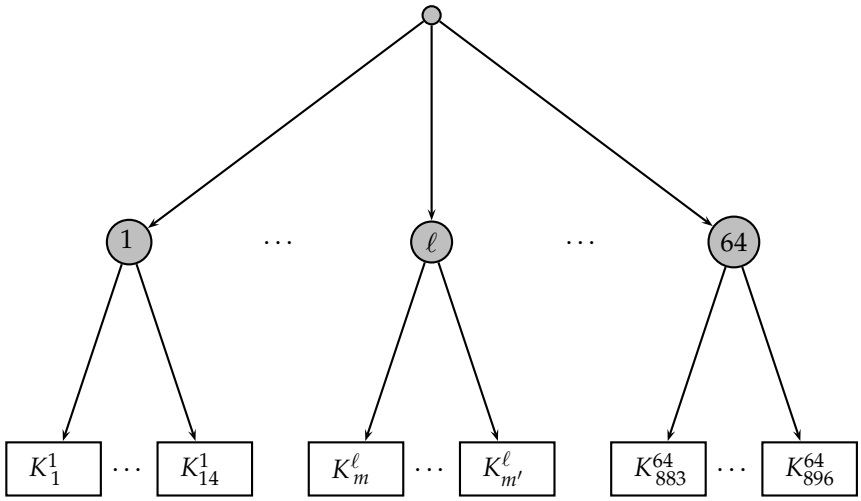


FIGURE 5.5 – *Arborescence associée au problème I.*

toirement. La base de test est composée des 6993 signaux restants. Le paramètre de régularisation C a été choisi par validation croisée sur 5 blocs. Les performances du classifieur obtenu ont été mesurées par l'aire sous la courbe ROC (AUC). Cette procédure a été répétée 10 fois.

Résultats

Le tableau 5.6 résume les performances moyennes des trois méthodes : le MKL, le $\text{CKL}_{1/2}$ et le CKL_1. Le nombre de canaux et de noyaux sélectionnés sont également reportés.

Méthode	AUC	# Canaux	# Noyaux
$\text{CKL}_{1/2}$	85.61 ± 1.21	64.00 ± 0.00	886.20 ± 04.13
CKL_1	86.16 ± 0.76	33.40 ± 6.26	130.30 ± 37.57
MKL	85.81 ± 0.52	60.10 ± 2.08	228.10 ± 32.53

TABLE 5.6 – *Résultats moyens pour le problème I.*

Les performances moyennes des trois méthodes sont très similaires. On remarque néanmoins une différence significative en terme de sélection de canaux et de noyaux. En ce qui concerne les noyaux, le $\text{CKL}_{1/2}$ est bien moins parcimonieux que le MKL, qui lui même conserve presque deux fois plus de noyaux que le CKL_1. Cette différence est encore plus notable en terme de sélection de canaux : le $\text{CKL}_{1/2}$, qui conserve l'ensemble des électrodes, a un comportement assez proche du MKL, qui en élimine quatre en moyenne. Le CKL_1 quant à lui élimine jusqu'à la moitié des électrodes.

***Remarque* 5.3 —** Le manque de parcimonie pour le $\text{CKL}_{1/2}$ s'explique par la nature de la pénalisation $\ell_{(1,4/3)}$ appliquée. En effet, nous savons que ces signaux EEG comportent une activité significative autour des fenêtres temporelles 7 et 8. Ainsi, lorsqu'un ou plusieurs éléments significatifs d'un

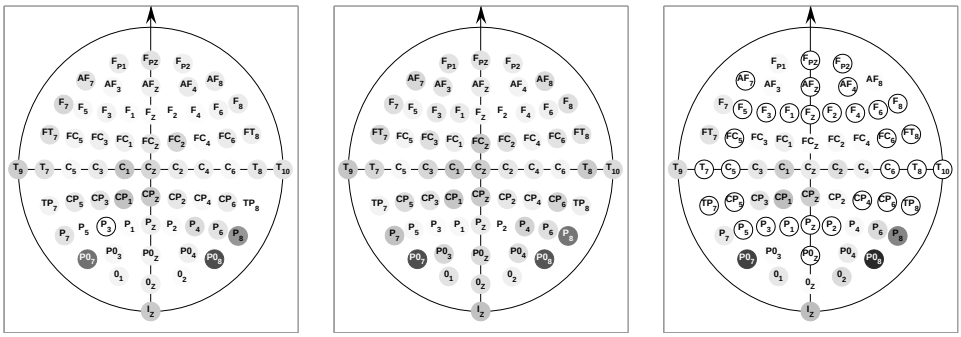
canal entrent dans le modèle, c'est l'ensemble des coefficients associés au canal qui est intégré. Les éléments les moins significatifs ont cependant des valeurs proches de 0, comme le confirme la figure 5.7. ◇

La figure 5.6 représente le degré de pertinence des électrodes sur les 10 essais. Elle montre en particulier les électrodes sélectionnées, et celles éliminées pour les trois méthodes. Pour un essai, le degré de pertinence de l'électrode ℓ est calculée par le représentant d'un groupe associé à la norme mixte $\ell_{(r,s)}$ de l'expression (4.11) :

$$\frac{d_\ell^t}{Z} \left(\sum_{m \in G_\ell} \|f_m^\star\|_{\mathcal{H}_m}^s \right)^{1/s} , \tag{5.2}$$

où Z est un facteur qui normalise la somme des représentants des groupes à 1, et où

$$\|f_m^\star\|_{\mathcal{H}_m}^2 \;=\; (\sigma_m^\star)^2 \sum_{i,j} \alpha_i^\star \alpha_j^\star \, y_i y_j \, K_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \; .$$



MKL $\quad$ $\mathrm{CKL}_{1/2}$ $\quad$ CKL_1

FIGURE 5.6 – *Pertinence médiane des électrodes pour le problème I. Plus les couleurs sont foncées, plus la pertinence est élevée. Les électrodes blanches entourées d'un cercle noir sont celles pour lesquelles les degrés de pertinence valent* 0.

Les résultats du CKL_1 sont particulièrement nets. Les degrés de pertinence des électrodes situées dans le cortex visuel, notamment ceux des électrodes PO_7 and PO_8, sont importants. On remarque également que l'électrode CP_Z, ainsi que d'autres électrodes associées à l'aire motrice primaire et au cortex somatosensoriel, jouent un rôle [2]. Les résultats du MKL et du $\mathrm{CKL}_{1/2}$ montrent l'influence des mêmes électrodes. Cependant, contrairement au CKL_1, elles mettent en valeur de nombreuses électrodes frontales, qui ne semblent pas être significatives pour le paradigme du BCI P300 Speller.

Enfin, nous pouvons observer sur la figure 5.7 que ce sont essentiellement les coefficients associés aux fenêtres temporelles 7 et 8 qui permettent d'identifier les électrodes pertinentes. Cependant, si le $\mathrm{CKL}_{1/2}$ et le CKL_1 focalisent quasiment exclusivement sur ces deux fenêtres, le MKL met en relief quelques électrodes sur l'ensemble des fenêtres, notamment les quatre premières. Les valeurs représentées sur les électrodes correspondent à $\|f_m^\star\|^s$.

2. Le site `http://lecerveau.mcgill.ca/` présente, entre autre, les fonctions associées aux différentes aires du cerveau.

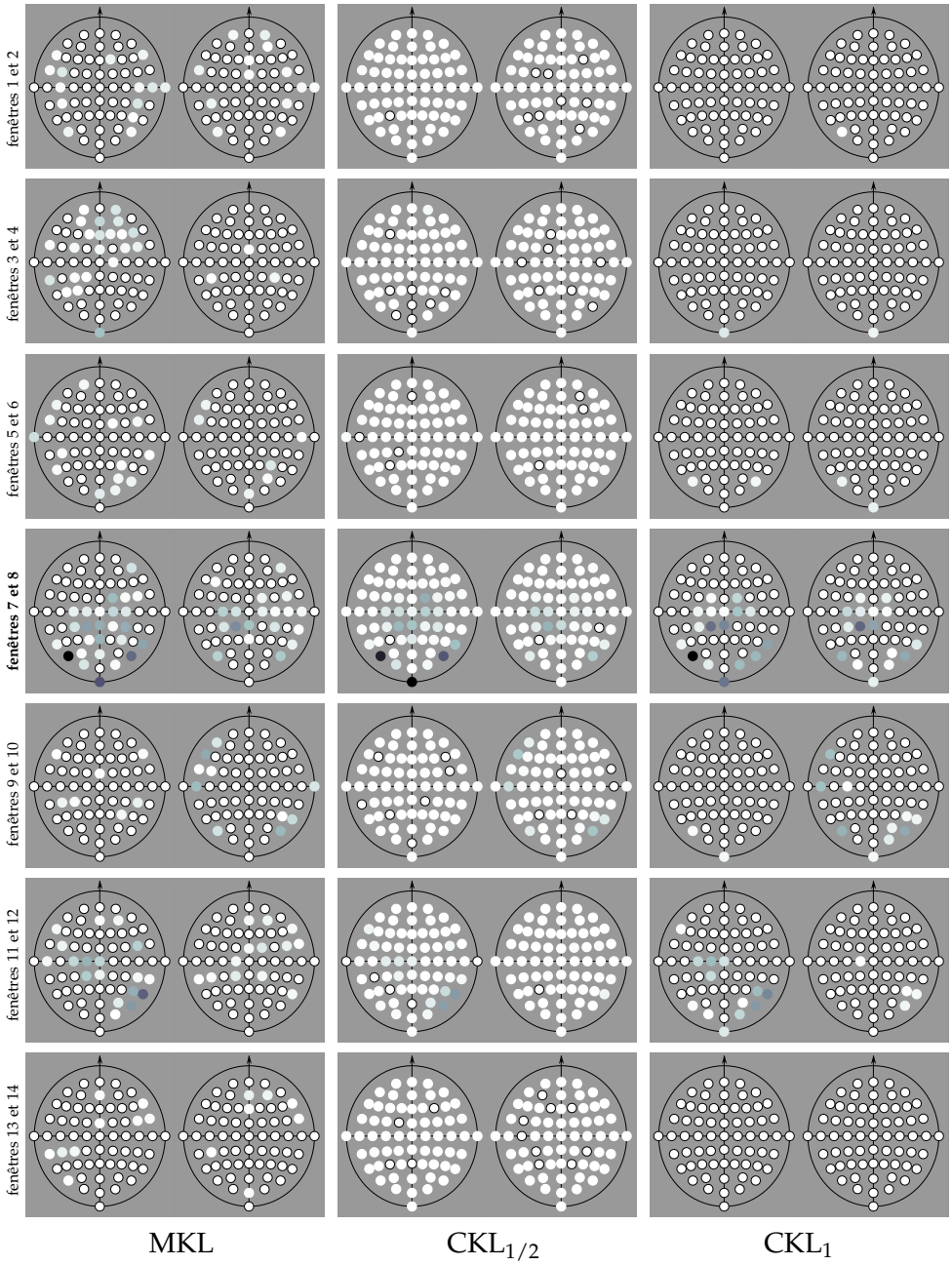


FIGURE 5.7 – *Évolution temporelle de la pertinence des électrodes sur un essai pour le problème I. Plus les couleurs sont foncées, plus la pertinence est élevée. Les électrodes blanches entourées d'un cercle noir sont celles pour lesquelles les degrés de pertinence valent* 0.

5.3.3 Problème II

Les données utilisées dans ce problème sont issues des recherches de Gangadhar Garipelli, doctorant à l'Idiap sous la direction de José del R. Millán.

Protocole expérimental

L'objectif est ici de détecter les zones actives du cerveau pendant la phase d'anticipation d'un évènement. Dans cette expérience, on s'intéresse donc à la fenêtre temporelle située juste avant un stimulus particulier.

Dans le protocole associé à cette expérience, un sujet face à un écran est soumis à deux types d'évènements :

1. Dans les évènements « GO », un point vert est affiché au milieu de l'écran. Au bout de 4 secondes, le point change de couleur et devient rouge. On demande alors au sujet de réagir.
2. Dans les évènements « NOGO », un point jaune est affiché au milieu de l'écran. Au bout de 4 secondes, le point change de couleur et devient rouge, on demande au sujet de ne pas réagir.

Dans la phase d'apprentissage, on demande au sujet de réagir en pressant un bouton. On dispose de deux jeux de données correspondant à deux individus. Une phase d'apprentissage est constituée de 20 sessions, et chaque session de 10 essais. Pour chaque sujet, l'ensemble d'apprentissage contient 200 signaux.

Pour chacune des 64 électrodes, le signal est échantillonné selon 21 coefficients, entre 0 et 3.25 secondes[3]. La dimension globale du problème est donc de $64 \times 21 = 1344$. L'arborescence associée à ce schéma est représentée sur la figure 5.8.

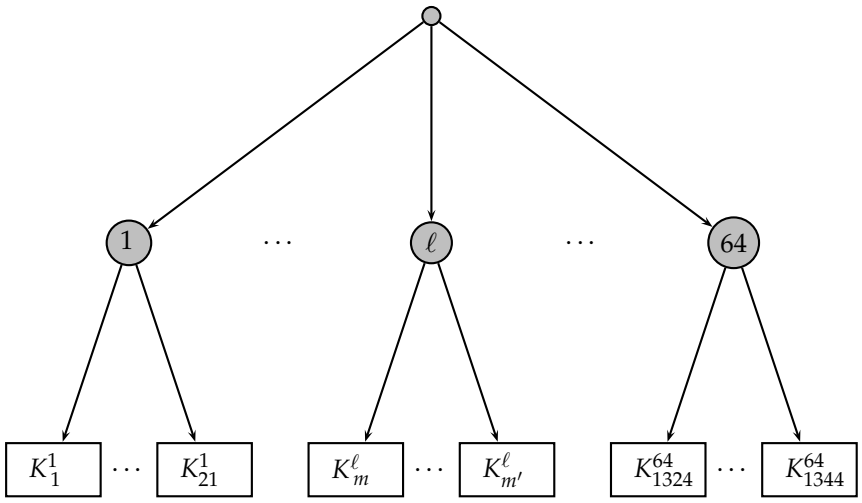


FIGURE 5.8 – *Arborescence associée au problème II.*

3. Le signal est donc tronqué avant l'apparition de l'évènement, qui a lieu au bout de 4 secondes.

Les connaissances sur ce type de problème identifient l'électrode centrale C_Z comme l'une des plus pertinentes. De façon générale, les électrodes situées dans la zone centrale du cerveau doivent intervenir dans ce problème, contrairement aux électrodes situées en périphérie.

À partir de cette connaissance, les auteurs de ces données ont construit un classifieur basé sur l'analyse linéaire discriminante, en tenant compte uniquement de l'électrode C_Z. Le pourcentage d'erreur de classification en test se situe autour 30% pour le sujet 1 et 33% pour le sujet 2.

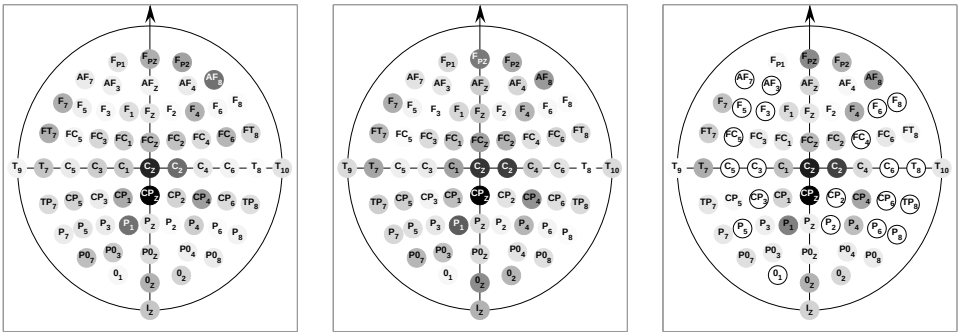
Résultats

Nous allons maintenant comparer ces connaissances et ces premiers résultats à ceux retournés par nos trois algorithmes. L'erreur de test est évaluée par une procédure de double validation croisée (cf. annexe A.5). Pour l'interprétation, on estime les coefficients σ_m sur la totalité de l'ensemble d'apprentissage, en sélectionnant le paramètre de régularisation C par validation croisée sur 10 blocs. Les résultats obtenus par le $CKL_{1/2}$, le CKL_1 et le MKL sont reportés sur les tableaux 5.7 et 5.8.

Pour le sujet 1, les trois méthodes permettent en moyenne d'améliorer les résultats initiaux. Néanmoins, les écart-types reportés suggèrent une variabilité importante des résultats en prédiction, en fonction des sessions utilisées en validation et en test (cf. annexe A.5). Dans cette expérience, le CKL_1 fait intervenir nettement moins d'électrodes que le $CKL_{1/2}$ et le MKL.

Sujet 1	Erreur (%)	# Canaux	# Noyaux
$CKL_{1/2}$	22.00 ± 4.81	64.00 ± 0.00 (64)	1323.00 ± 2.24 (1324)
CKL_1	23.00 ± 6.47	16.40 ± 10.90 (43)	177.60 ± 67.66 (326)
MKL	21.50 ± 9.62	52.00 ± 22.94 (64)	267.60 ± 114.27 (343)

TABLE 5.7 – *Résultats moyens sur le sujet* 1 *pour le problème II. Les valeurs entre parenthèses représentent le nombre de canaux et de noyaux sélectionnés sur la totalité de l'ensemble d'apprentissage.*



MKL $CKL_{1/2}$ CKL_1

FIGURE 5.9 – *Pertinence des électrodes chez le sujet* 1 *pour le problème II. Plus les couleurs sont foncées, plus la pertinence est élevée. Les électrodes blanches entourées d'un cercle noir sont celles pour lesquelles les degrés de pertinence valent* 0*.*

La figure 5.9 rend compte de l'influence des électrodes, en utilisant l'expression (5.2) du problème précédent. Les trois méthodes mettent en évi-

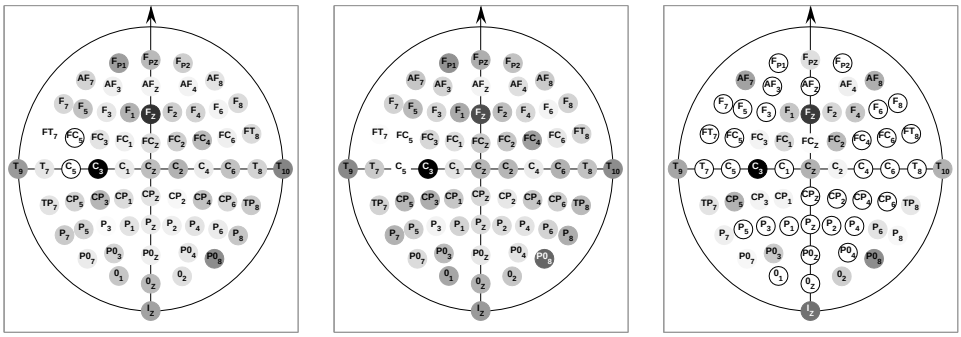
dence la zone centrale du cerveau. Cependant, le CKL_1 annule de nombreux coefficients relatifs aux électrodes périphériques, sans toutefois parvenir à tous les éliminer. Bien que le MKL soit plus parcimonieux en termes de noyaux que le $CKL_{1/2}$, ces deux méthodes localisent de nombreuses électrodes autour de la zone centrale.

En ce qui concerne le sujet 2, les performances des trois méthodes sont en moyenne similaires à celle du classifieur initial. Encore une fois, le $CKL_{1/2}$ et le MKL font intervenir la quasi-totalité des électrodes, alors que le CKL_1 en élimine plus de la moitié.

Sujet 2	Erreur (%)	# Canaux	# Noyaux
$CKL_{1/2}$	33.00 ± 8.37	64.00 ± 0.00 (64)	1321.60 ± 0.89 (1324)
CKL_1	33.00 ± 6.23	34.60 ± 21.62 (30)	261.20 ± 157.52 (185)
MKL	27.50 ± 3.54	60.40 ± 1.67 (62)	331.40 ± 48.42 (358)

TABLE 5.8 – *Résultats moyens sur le sujet* 2 *pour le problème II. Les valeurs entre parenthèses représentent le nombre de canaux et de noyaux sélectionnés sur la totalité de l'ensemble d'apprentissage.*

La figure 5.10 montre l'influence des électrodes, toujours en utilisant l'expression (5.2). Les résultats sont moins nets pour ce sujet. En effet, les pertinences se répartissent sur l'ensemble des zones du cerveau pour les trois méthodes. Le CKL_1 élimine toutefois beaucoup plus d'électrodes en périphérie que le $CKL_{1/2}$ et le MKL. De façon surprenante, l'électrode C_Z n'a ici qu'une contribution moyenne.



MKL $CKL_{1/2}$ CKL_1

FIGURE 5.10 – *Pertinence des électrodes chez le sujet* 2 *pour le problème II. Plus les couleurs sont foncées, plus la pertinence est élevée. Les électrodes blanches entourées d'un cercle noir sont celles pour lesquelles les degrés de pertinence valent* 0.

5.4 Seuillage pour l'algorithme 2

5.4.1 Définition d'un seuillage

Dans les précédentes sections, nous avons observé, notamment pour la pénalité $\ell_{(1,4/3)}$, que les solutions retournées par l'algorithme 2 sont peu parcimonieuses. En effet, il se peut que tous les coefficients σ_m appartenant à un groupe soient très faibles. Ainsi, un groupe peut être sélectionné même si sa pertinence est quasi-nulle.

Nous nous sommes donc interrogés sur une façon adéquate de seuiller ces petits coefficients. Pour cela, nous avons utilisé les conditions d'optimalité en σ du problème (4.10). On obtient la condition :

$$\frac{\partial J(\sigma)}{\partial \sigma_m} > \sum_{\ell} \sum_{m \in G_\ell} \sigma_m \frac{\partial J(\sigma)}{\partial \sigma_m} \left(d_\ell^{-1} \sum_{m \in G_\ell} \sigma_m^{1/q} \right)^{-p/(p+q)} \sigma_m^{(1-q)/q} \quad \Rightarrow \quad \sigma_m = 0 \,,$$

où $\frac{\partial J(\sigma)}{\partial \sigma_m}$ est défini par (4.16). Les détails de la dérivation de cette condition sont reportés en annexe A.4.

5.4.2 Application du seuillage au problème II

Nous avons utilisé la condition énoncée ci-dessus pour seuiller les solutions retournées par l'algorithme CKL, dans les trois configurations. Ici également, les erreurs de test sont évaluées par le biais d'une double validation croisée. Pour l'interprétation, les coefficients σ_m sont estimés sur la totalité de l'ensemble d'apprentissage, en sélectionnant le paramètre de régularisation C par validation croisée sur 10 blocs. Les résultats obtenus par le $\text{CKL}_{1/2}$, le CKL_1 et le MKL sont reportés sur les tableaux 5.9 et 5.10.

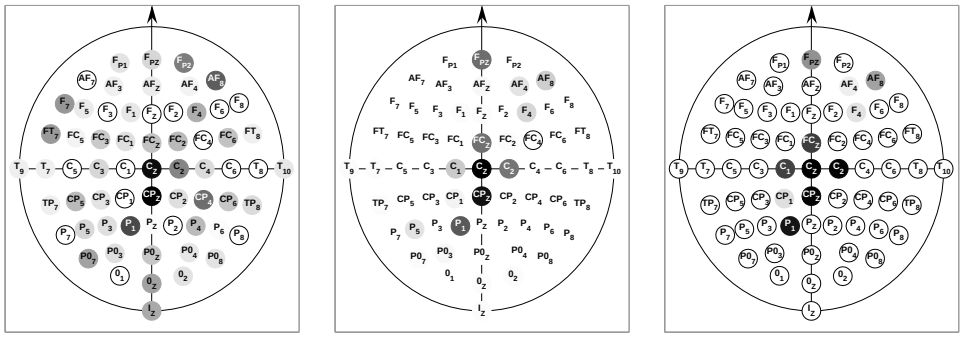
Sujet 1

Pour le sujet 1, on constate, notamment avec le $\text{CKL}_{1/2}$, que le seuillage permet d'éliminer un nombre important de noyaux, qui correspondent en fait aux coefficients les moins influents dans chaque groupe. Cependant, l'ensemble des électrodes est conservé. Le seuillage affecte également le MKL qui sélectionne en moyenne moitié moins de noyaux qu'initialement. Le CKL_1 est également plus parcimonieux, avec environ 10 électrodes conservées. Cependant, si pour les trois méthodes les erreurs moyennes restent inférieures à celles du classifieur basé sur l'électrode C_Z, on remarque que les performances sont dégradées par rapport à celles obtenues lorsque les solutions ne sont pas seuillées.

Sujet 1	Erreur (%)	# Canaux	# Noyaux
$\text{CKL}_{1/2}$	24.50 ± 7.70	60.60 ± 1.67 (63)	434.60 ± 52.26 (471)
CKL_1	25.50 ± 5.70	10.20 ± 2.59 (11)	45.00 ± 4.90 (51)
MKL	24.50 ± 5.70	25.00 ± 19.16 (48)	65.40 ± 47.42 (103)

TABLE 5.9 – *Résultats obtenus en seuillant la solution sur le sujet* 1 *pour le problème II. Les valeurs entre parenthèses représentent le nombre de canaux et de noyaux sélectionnés sur la totalité de l'ensemble d'apprentissage.*

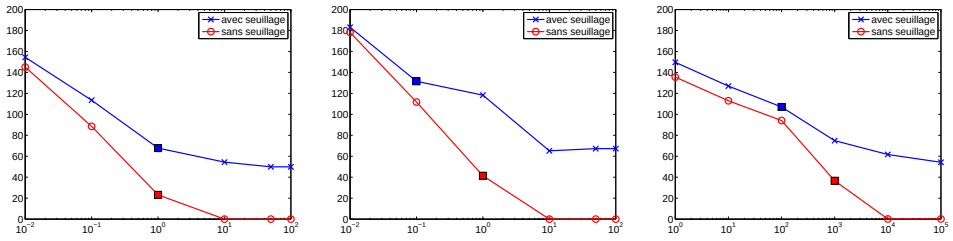
La figure 5.11 montre la pertinence des électrodes. Ici, le CKL_1 met clairement en évidence la zone centrale du cerveau. Le $\text{CKL}_{1/2}$ présente un comportement similaire au CKL_1 : bien qu'il conserve l'essentiel des électrodes, en accord avec la nature de la pénalisation $\ell_{(1,4/3)}$, celles situées en périphérie de la zone centrale sont très peu influentes. Le MKL plus parcimonieux en termes de canaux et de noyaux que le $\text{CKL}_{1/2}$ localise cependant de nombreuses électrodes hors de la zone centrale.



MKL $\quad$ $\mathrm{CKL}_{1/2}$ $\quad$ CKL_1

FIGURE 5.11 – *Pertinence des électrodes chez le sujet* 1 *pour le problème II. Plus les couleurs sont foncées, plus la pertinence est élevée. Les électrodes blanches entourées d'un cercle noir sont celles pour lesquelles les degrés de pertinence valent* 0.

Nous avons également reporté sur la figure 5.12 les valeurs des coûts charnière en apprentissage pour les solutions seuillées (en bleu) et non seuillées (en rouge). Pour le MKL et le CKL_1, ce coût varie de façon cohérente, à savoir décroissante, en fonction du paramètre de régularisation. Pour le $\mathrm{CKL}_{1/2}$, ce coût remonte sensiblement, lorsque $C > 10$.



MKL $\quad$ $\mathrm{CKL}_{1/2}$ $\quad$ CKL_1

FIGURE 5.12 – *Évolution du coût charnière sur le sujet* 1 *pour le problème II. En abscisse, les valeurs du paramètre de régularisation C ; en ordonnée, les valeurs du coût charnière. Les carrés présents sur les courbes correspondent au paramètre de régularisation choisi par validation croisée.*

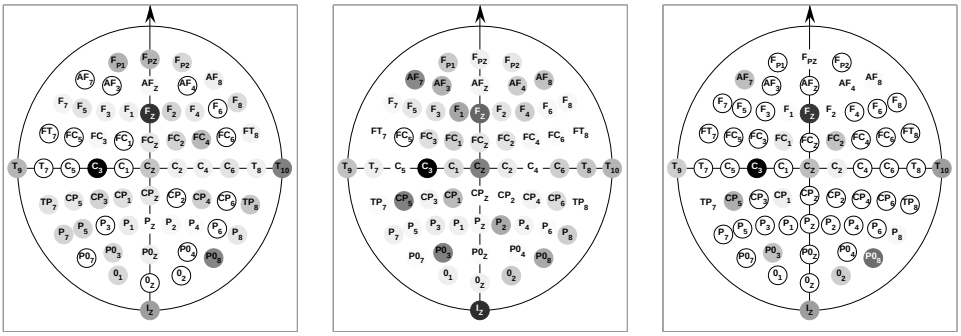
Sujet 2

En ce qui concerne le sujet 2, seule l'erreur du CKL_1 reste similaire aux résultats initiaux. Les erreurs du $\mathrm{CKL}_{1/2}$ et du MKL sont en moyenne nettement supérieures à celles obtenues en n'utilisant pas de seuillage. On observe, comme chez le sujet 1, une influence du seuillage sur la sélection de noyaux et de canaux.

Méthode	Erreur (%)	# Canaux		# Noyaux	
$\mathrm{CKL}_{1/2}$	40.00 ± 4.68	62.00 ± 3.39	(63)	659.00 ± 219.56	(770)
CKL_1	33.00 ± 4.81	34.80 ± 11.39	(22)	119.20 ± 52.37	(45)
MKL	32.00 ± 7.37	35.20 ± 16.72	(45)	148.60 ± 119.78	(97)

TABLE 5.10 – *Résultats obtenus en seuillant la solution sur le sujet* 2 *pour le problème II. Les valeurs entre parenthèses représentent le nombre de canaux et de noyaux sélectionnés sur la totalité de l'ensemble d'apprentissage.*

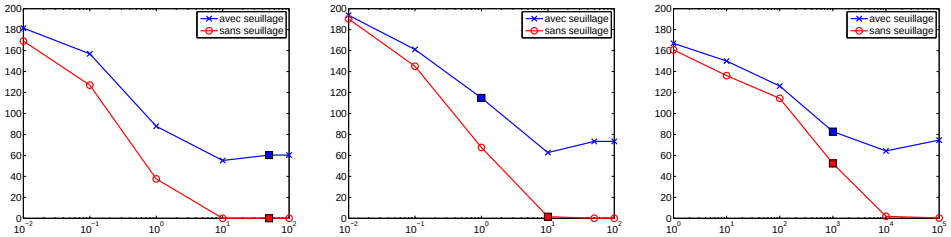
La figure 5.13 représente l'influence des électrodes. Pour ce sujet, les résultats restent flous et moins conformes aux attentes, même si le CKL_1 élimine plus d'électrodes en périphérie de la zone centrale qu'avec une solution non seuillée.



MKL $CKL_{1/2}$ CKL_1

FIGURE 5.13 – *Pertinence des électrodes chez le sujet* 2 *pour le problème II. Plus les couleurs sont foncées, plus la pertinence est élevée. Les électrodes blanches entourées d'un cercle noir sont celles pour lesquelles les degrés de pertinence valent* 0*.*

Finalement, on constate des comportements incohérents en apprentissage. La figure 5.14 montre que les coûts charnières relatifs aux trois méthodes remontent pour des pénalisations importantes, lorsque les solutions sont seuillées. Notons que si l'on réitère l'apprentissage en utilisant la solution seuillée à l'initialisation, le problème persiste.



MKL $CKL_{1/2}$ CKL_1

FIGURE 5.14 – *Évolution du coût charnière sur le sujet* 2 *pour le problème II. En abscisse, les valeurs du paramètre de régularisation ; en ordonnée, les valeurs du coût charnière. Les carrés présents sur les courbes correspondent au paramètre de régularisation choisi par validation croisée.*

5.5 SYNTHÈSE

Dans ce chapitre, nous avons comparé, au travers de simulations, les comportements des deux algorithmes, à savoir l'algorithme basé sur le principe des contraintes actives, et celui basé sur la méthode de gradient. Nous avons pour cela utilisé la norme mixte $\ell_{(1,4/3)}$.

L'algorithme basé sur la méthode de gradient présente l'avantage de pouvoir utiliser une pénalisation non-convexe et de considérer des regroupements sur différents noyaux. Nous l'avons appliqué sur des problèmes relatifs aux interfaces cerveau-machine, pour lesquels l'utilisation d'une pénalité non-convexe facilite l'interprétation des résultats, tout en atteignant des performances en prédiction similaires à celles de méthodes convexes.

Nous avons également discuté et testé le seuillage des solutions retournées par l'algorithme de gradient, afin d'obtenir des résultats plus interprétables, notamment pour la pénalité convexe $\ell_{(1,4/3)}$. Nous avons dérivé une condition de seuillage, qui ne semble pas apporter une solution satisfaisante dans tous les cas de figure. En particulier, le fait d'utiliser ce seuillage peut introduire de l'instabilité et dégrader les performances du classifieur.

Cette dernière problématique laisse place à un certain nombre de questions. En effet, on peut envisager d'autres façons de définir un seuillage optimal. On peut aussi s'interroger sur la façon d'intégrer ce processus au sein de l'algorithme, que ce soit par le biais de l'initialisation, en partant de solutions parcimonieuses, ou que ce soit par la modification du critère d'arrêt.

Conclusion

6

6.1 Synthèse et contributions

La *pénalisation hiérarchique* permet d'intégrer un *a priori* spécifique dans l'estimation de modèles statistiques. Cette connaissance porte sur :

1. L'organisation des caractéristiques d'un problème en une structure hiérarchique.
2. Un principe de parcimonie sur les composantes de cette structure.

Dans cette thèse, nous représentons cette structure par une arborescence à deux niveaux, ce qui permet de constituer des groupes distincts de caractéristiques. Nous avons utilisé des approches régularisées pour extraire les composantes significatives de chaque niveau.

Dans un premier temps, nous avons présenté un état de l'art des techniques de régularisation en apprentissage supervisé. Ce préalable nous a permis de dresser les propriétés de parcimonie et de convexité de ce type de méthodes, et d'esquisser les questions relatives à la stabilité et à la consistance en sélection de modèles des solutions qu'elles engendrent. Nous avons également décrit les principes généraux des algorithmes de résolution associés.

Nous avons ensuite déroulé les étapes successives permettant d'aboutir à la *pénalisation hiérarchique*. Dans cette formulation, des facteurs de pertinence sont associés à chaque branche de l'arborescence, et soumis à des contraintes de parcimonie. En fonction du problème considéré, le degré de parcimonie peut être réglé par deux paramètres distincts : l'un est associé aux facteurs relatifs aux groupes et l'autre aux facteurs relatifs aux caractéristiques. Nous avons montré que minimiser ce problème est une forme variationnelle d'un problème régularisé par une norme mixte. La mise en relation de ces deux approches offre deux points de vues pour étudier les propriétés de convexité et de parcimonie de ce modèle.

Dans la chronologie de ces trois années de thèse, nous avons d'abord étudié un modèle de régression paramétrique, qui a conduit à minimiser un problème régularisé par la norme mixte convexe $\ell_{(1,4/3)}$. Un algorithme de type « contraintes actives » a été développé pour résoudre ce problème particulier [Szafranski et coll., 2007, 2008a]. Nous avons ensuite étendu cette approche pour des normes mixtes $\ell_{(r,s)}$ génériques et potentiellement non-convexes, dans le cadre de fonctions noyaux. Nous nous sommes alors

concentrés sur les problèmes de classification binaire. C'est avec un algorithme de type « wrapper », basé sur une méthode de gradient, que nous avons proposé de résoudre ce problème [Szafranski et coll., 2008b,c].

Dans ce document, nous avons noué de nouveaux liens entre ces deux approches. D'une part, nous avons dérivé les résultats de l'approche paramétrique régularisée par la norme mixte $\ell_{(1,4/3)}$ pour aboutir à une norme mixte générique $\ell_{(r,s)}$, et adapté l'algorithme de contraintes actives en conséquence. Pour que la solution retournée par cet algorithme soit définie, la norme $\ell_{(r,s)}$ doit toutefois être convexe. D'autre part, nous avons montré comment implémenter le wrapper avec une fonction de perte quadratique, ce qui permet de considérer des problèmes de régression.

Nous avons comparé les deux approches sur des simulations, pour un critère quadratique régularisé par une norme mixte convexe $\ell_{(1,4/3)}$. Pour l'algorithme de contraintes actives, les valeurs des critères minimisés sont inférieures et les solutions sont stables numériquement. De plus, elles sont plus parcimonieuses alors que les solutions retournées par l'algorithme basé sur la méthode de gradient comportent de nombreux coefficients presque nuls. Ce constat nous a mené à explorer une stratégie de seuillage visant à améliorer l'interprétabilité des solutions dans le cas convexe. Cependant, ce seuillage induit des problèmes de stabilité, qui doivent être examinés plus en profondeur. En revanche, sur les expériences considérées, où le nombre de variables est restreint, le wrapper converge plus rapidement.

Pour illustrer le comportement de la *pénalisation hiérarchique* utilisant différentes normes mixtes, nous avons appliqué le wrapper à des problèmes d'interfaces cerveau-machine. Nous avons mis en évidence, sur les deux problèmes étudiés, qu'une pénalisation non-convexe permettait d'obtenir des solutions plus parcimonieuses et donc plus facilement interprétables, tout en atteignant des performances en prédiction similaires à celles de méthodes convexes.

6.2 Perspectives

Dans la continuité directe de nos travaux de thèse, nous envisageons différentes perspectives. Elles concerne notamment des aspects applicatifs, algorithmiques, ainsi que d'autres aspects liés aux principes d'optimisation.

Applications en signal / image

Nous travaillons actuellement sur un problème de détection en image, qui se traduit par un problème de classification binaire sur chaque pixel. Dans cette application, on utilise des filtres définis par les paramètres d'une gaussienne :

1. L'échelle est analogue à une largeur de bande ou un écart-type ;
2. L'ordre de dérivation de la gaussienne définit une notion de régularité ;

3. L'orientation désigne la direction dans laquelle s'effectue la dérivation.

Le coût lié au calcul de la réponse d'un filtre sur tous les pixels de l'image est important. Les filtres orientables [1] permettent de calculer des réponses sur toutes les orientations à partir d'une combinaison linéaire de peu de filtres. Une fois qu'on a choisi une échelle et un ordre, les réponses des filtres pour toutes les orientations viennent à un coût calculatoire négligeable. Le coût calculatoire est donc lié à la présence de un ou plusieurs filtres d'un ordre et d'une échelle donnés. Notre objectif consiste à minimiser le nombre d'erreurs à coût calculatoire borné. Ce coût calculatoire peut être approché par une norme mixte.

Applications en génomique

Dans le chapitre 4, nous avons décrit le formalisme de la *pénalisation hiérarchique* pour des hiérarchies de profondeurs arbitraires, et discuté de la généralisation aux graphes acycliques dirigés. Si la formulation concernant les arborescences de hauteurs arbitraires est déjà définie, l'étude des propriétés et la mise en œuvre de cette extension doivent être réalisées.

De plus, pour être fidèle à la sémantique des hiérarchies de *Gene Ontology*, il faudra choisir comment représenter ces graphes acycliques dirigés. En particulier, il faudra déterminer si les parentés multiples correspondent à des appartenances dures ou floues, et dans le dernier cas, connues ou inconnues. Ainsi, l'ensemble de ces travaux est envisagé à plus long terme.

Interprétabilité

Pour l'algorithme de type wrapper, nous souhaitons investiguer deux pistes pour obtenir des solutions plus intéprétables, notamment pour la pénalité $\ell_{(1,4/3)}$. D'une part, il est possible de jouer sur l'initialisation de la solution. D'autre part, des stratégies de seuillage restent à développer, que ce soit au sein même du processus de descente de gradient, ou sur la définition du critère d'arrêt. Dans tous les cas, l'intégration de ce seuillage devra garantir la stabilité des solutions.

Optimisation de problèmes non-convexes

Nos résultats expérimentaux illustrent l'intérêt des normes mixtes non-convexes. La non-convexité pose toutefois des problèmes théoriques et algorithmiques. Pour répondre à ces problèmes, nous pouvons considérer différentes approches.

La première consiste à partir d'une pénalisation convexe, qui évolue doucement pour finalement atteindre la pénalité non-convexe souhaitée. Cette évolution serait contrôlée par un ou plusieurs paramètres, par exemple les puissances d'une norme mixte. Le suivi des points d'équilibre permettrait d'assurer la convergence vers le minimum global. Cependant, le nombre

1. Steerable filters, en anglais.

de bifurcations croît potentiellement de manière exponentielle, ce qui peut rendre l'approche inutilisable quand le nombre de variables dépasse la dizaine.

Nous pouvons aussi envisager une approche gloutonne, en résolvant successivement une série de problèmes convexes. Par exemple, une première itération consiste à résoudre le problème régularisé par une norme mixte. Si les variables de la matrice d'observations sont ensuite mises à l'échelle des coefficients de la solution retournée, et que la même procédure est répétée avec cette nouvelle matrice d'observations, la deuxième solution correspond alors à un minimum local d'un problème régularisé par une norme mixte plus restrictive, peut-être déjà non-convexe, que nous pouvons déterminer. Après une ou plusieurs itérations de ce type, la solution retournée correspondrait à un minimum local d'un problème régularisé par une norme mixte non-convexe. Le fait de minimiser une succession de problèmes convexes pourrait conserver les propriétés de stabilité des solutions.

Chemin de régularisation

Nous avons vu au chapitre 3 que des travaux concernant l'étude des chemins de régularisation pour des normes mixtes ont été initiés. Ces recherches sont néanmoins focalisées sur certaines normes mixtes bien particulières, et se concentrent sur des pénalités permettant d'obtenir un chemin de régularisation linéaire par morceau [Zhao et coll., à paraître] .

Lorsque le chemin de régularisation ne peut être calculé de façon exacte, il peut être échantillonné. En effet, dans les algorithmes de contraintes actives, les solutions sont de plus en plus parcimonieuses, à mesure que le paramètre de régularisation croît. Pour un paramètre de régularisation $\lambda < \lambda'$, la solution peut donc être initialisée avec celle obtenue par l'optimisation du critère pour le paramètre λ'. On parle d'initialisation à chaud [2]. Le problème consiste alors à établir la séquence de paramètres induisant un changement dans l'ensemble actif.

Pour des normes mixtes génériques impliquant d'autres pénalités que des normes ℓ_1 ou ℓ_∞, le chemin n'est pas linéaire par morceaux. Caractériser la trajectoire entre deux mises à jour de l'ensemble actif devrait permettre d'améliorer l'efficacité des algorithmes d'optimisation, si le chemin est constitué de solutions optimales.

Pour une norme mixte non-convexe, il paraît difficile d'envisager ce genre de perspectives. En effet, assurer l'optimalité de la solution demande la résolution d'un problème combinatoire. De plus, le chemin n'est plus nécessairement continu. Au delà des difficultés techniques occasionnées, on peut s'interroger sur l'intérêt pratique du suivi d'un chemin discontinu. En revanche, pour des normes mixtes convexes, l'analyse des propriétés des chemins de régularisation pourrait aboutir à des gains calculatoires substantiels.

2. Warm start, en anglais.

Annexes

A

Sommaire

A.1 Éléments de preuve de la proposition 4.1

La preuve de la proposition 4.1 s'effectue en quatre temps. Dans un premier temps, on déduit du Lagrangien associé au problème (4.4) les conditions d'optimalité associées à $\sigma_{1,\ell}$ et $\sigma_{2,m}$. Puis, on établit une relation entre les facteurs $\sigma_{1,\ell}$ et les coefficients du vecteur $\boldsymbol{\beta}$. La même procédure est ensuite appliquée aux facteurs $\sigma_{2,m}$. Enfin, on reporte ces résultats, qui ne dépendent plus que des coefficients du vecteur $\boldsymbol{\beta}$, dans le problème initial.

A.1.1 Définition des conditions d'optimalité

Commençons par rappeler le problème initial :

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\,\sigma} & J(\boldsymbol{\beta}) + \lambda \sum\limits_{\ell} \sigma_{1,\ell}^{-p} \sum\limits_{m \in G_\ell} \sigma_{2,m}^{-q}\, \beta_m^2 & & & \text{(A.1a)} \\ \text{s. c.} & \sum\limits_{\ell} d_\ell\, \sigma_{1,\ell} \leq 1 & \sigma_{1,\ell} \geq 0 & \forall \ell & \text{(A.1b)} \\ & \sum\limits_{m} \sigma_{2,m} \leq 1 & \sigma_{2,m} \geq 0 & \forall m\ . & \text{(A.1c)} \end{cases}$$

Le Lagrangien associé à ce problème est

$$\begin{aligned} \mathcal{L} = J(\boldsymbol{\beta}) \;+\; & \lambda \sum_{\ell} \sigma_{1,\ell}^{-p} \sum_{m \in G_\ell} \sigma_{2,m}^{-q}\, \beta_m^2 \\ & + \nu_1 \left(\sum_{\ell} d_\ell\, \sigma_{1,\ell} - 1 \right) \;+\; \nu_2 \left(\sum_{\ell} \sum_{m \in G_\ell} \sigma_{2,m} - 1 \right) \\ & - \sum_{\ell} \eta_{1,\ell}\, \sigma_{1,\ell} \;-\; \sum_{m} \eta_{2,m}\, \sigma_{2,m}\ . \end{aligned}$$

Les conditions d'optimalité pour $\sigma_{2,m}$ et $\sigma_{1,\ell}$ sont

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \sigma_{1,\ell}} &= -\frac{\lambda\, p}{\sigma_{1,\ell}^{p+1}} \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_{2,m}^{q}} + \nu_1\, d_\ell - \eta_{1,\ell} = 0\ , \\ \frac{\partial \mathcal{L}}{\partial \sigma_{2,m}} &= -\frac{\lambda\, q}{\sigma_{2,m}^{q+1}} \frac{\beta_m^2}{\sigma_{1,\ell}^{p}} + \nu_2 - \eta_{2,m} = 0\ , \end{aligned}$$

dont ont déduit les expressions de $\sigma_{1,\ell}$ et $\sigma_{2,m}$

$$\sigma_{1,\ell} = 0 \qquad \text{ou} \qquad \sigma_{1,\ell} = \left(\frac{\lambda\, p}{\nu_1} \right)^{\frac{1}{p+1}} \left(\frac{1}{d_\ell} \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_{2,m}^{q}} \right)^{\frac{1}{p+1}}\ , \tag{A.2}$$

$$\sigma_{2,m} = 0 \qquad \text{ou} \qquad \sigma_{2,m} = \left(\frac{\lambda\, q}{\nu_2} \right)^{\frac{1}{q+1}} \left(\frac{\beta_m^2}{\sigma_{1,\ell}^{p}} \right)^{\frac{1}{q+1}}\ . \tag{A.3}$$

***Remarque* A.1 —** Les multiplicateurs de Lagrange $\eta_{1,\ell}$ (respectivement $\eta_{2,m}$) sont nuls lorsque les contraintes associées ne sont pas saturées, c'est à dire si $\sigma_{1,\ell} > 0$ (respectivement $\sigma_{2,m} > 0$). C'est pourquoi elles n'apparaissent pas dans les expressions (A.2) et (A.3). ◇

A.1.2 Expression de $\sigma_{1,\ell}$ en fonction de β_m

En reportant (A.3) dans (A.2), on trouve

$$\sigma_{1,\ell} = (\sigma_{1,\ell})^{\frac{pq}{(q+1)(p+1)}} \left(\frac{\lambda\, p}{\nu_1}\right)^{\frac{1}{p+1}} \left(\frac{\lambda\, q}{\nu_2}\right)^{\frac{1}{(p+1)(q+1)}} \left(d_\ell^{-1} \sum_{m\in G_\ell} |\beta_m|^{\frac{2}{q+1}}\right)^{\frac{1}{p+1}} .$$

On en déduit l'expression

$$\sigma_{1,\ell} \quad = \quad c_1 \left(d_\ell^{-1}\, s_\ell\right)^{\frac{q+1}{p+q+1}} , \tag{A.4}$$

où $s_\ell = \sum_{m\in G_\ell} |\beta_m|^{\frac{2}{q+1}}$, et où $c_1 = \left[\left(\frac{\lambda}{p\,\nu_1}\right)^{\frac{1}{p+1}} \left(\frac{\lambda}{q\,\nu_2}\right)^{\frac{1}{(p+1)(q+1)}}\right]^{\frac{p+q+1}{(p+1)(q+1)}}$ est une constante qui ne dépend pas de l'élément $\sigma_{1,\ell}$ considéré.

On détermine maintenant c_1 en reportant l'expression (A.4) dans la contrainte (A.2). À l'optimalité

$$c_1 \sum_\ell d_\ell \left(d_\ell^{-1}\, s_\ell\right)^{\frac{q+1}{p+q+1}} \quad = \quad 1 ,$$

dont on déduit

$$c_1 \quad = \quad \left[\sum_\ell d_\ell^{\frac{p}{p+q+1}}\ (s_\ell)^{\frac{q+1}{p+q+1}}\right]^{-1} .$$

Ainsi

$$\sigma_{1,\ell} \quad = \quad \frac{\left(d_\ell^{-1}\, s_\ell\right)^{\frac{q+1}{p+q+1}}}{\sum_\ell d_\ell^{\frac{p}{p+q+1}}\ (s_\ell)^{\frac{q+1}{p+q+1}}} . \tag{A.5}$$

A.1.3 Expression de $\sigma_{2,m}$ en fonction de β_m

En reportant maintenant (A.5) dans (A.3), nous obtenons

$$\sigma_{2,m} \quad = \quad c_2\, \frac{|\beta_m|^{\frac{2}{q+1}}}{\left(d_\ell^{-1}\, s_\ell\right)^{\frac{p}{p+q+1}}} , \tag{A.6}$$

où $c_2 = (c_1)^{\frac{-p}{q+1}} \left(\frac{\lambda}{q\,\nu_2}\right)^{\frac{1}{q+1}}$ est une constante qui ne dépend pas de l'élément $\sigma_{2,m}$ considéré. On détermine c_2 en reportant l'expression (A.6) dans la contrainte (A.3). À l'optimalité

$$c_2 \sum_\ell (d_\ell)^{\frac{p}{p+q+1}}\ (s_\ell)^{\frac{-p}{p+q+1}} \sum_{m\in G_\ell} |\beta_m|^{\frac{2}{q+1}} \quad = \quad 1 ,$$

dont on déduit

$$c_2 \quad = \quad \left[\sum_\ell d_\ell^{\frac{p}{p+q+1}}\ (s_\ell)^{\frac{q+1}{p+q+1}}\right]^{-1} \quad = \quad c_1 .$$

Ainsi

$$\sigma_{2,m} \quad = \quad \frac{|\beta_m|^{\frac{2}{q+1}} \left(d_\ell^{-1} s_\ell\right)^{\frac{-p}{p+q+1}}}{\sum_\ell d_\ell^{\frac{p}{p+q+1}} (s_\ell)^{\frac{q+1}{p+q+1}}} \ . \tag{A.7}$$

A.1.4 Expression du problème initial en fonction de β_m

Nous devons maintenant déterminer la valeur du produit $\sigma_{1,\ell}^p \, \sigma_{2,m}^q$

$$\sigma_{1,\ell}^p \, \sigma_{2,m}^q \quad = \quad \frac{|\beta_m|^{\frac{2q}{q+1}} \left(d_\ell^{-1} s_\ell\right)^{\frac{p}{p+q+1}}}{\left(\sum_\ell d_\ell^{\frac{p}{p+q+1}} (s_\ell)^{\frac{q+1}{p+q+1}}\right)^{p+q}} \ ,$$

et l'intégrer dans la somme correspondant à la partie régularisée de l'équation (A.1a) problème initial, afin d'obtenir

$$\sum_\ell \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_{1,\ell}^p \, \sigma_{2,m}^q} \quad = \quad \left(\sum_\ell d_\ell^{\frac{p}{p+q+1}} s_\ell^{\frac{q+1}{p+q+1}}\right)^{p+q+1} \ .$$

En remplaçant s_ℓ par sa valeur, le problème final s'exprime uniquement en fonction des coefficients β_m

$$\min_{\boldsymbol{\beta}} \quad J(\boldsymbol{\beta}) \quad + \quad \lambda \left(\sum_\ell d_\ell^{\frac{p}{p+q+1}} \left(\sum_{m \in G_\ell} |\beta_m|^{\frac{2}{q+1}}\right)^{\frac{q+1}{p+q+1}}\right)^{p+q+1} \ .$$

C'est bien cette dernière expression qui correspond à la formulation (4.5) de la proposition 4.1. □

A.2 Éléments de preuve de la proposition 4.5

La preuve de la proposition 4.5 s'effectue en trois temps, dans le même esprit que celle présentée en annexe A.1. Dans un premier temps, on déduit du Lagrangien associé au problème (4.10) les conditions d'optimalité associées à σ_m. Puis, on établit une relation entre les paramètres σ_m et les éléments des EHNR f_m. Enfin, on reporte ces résultats, qui ne dépendent plus que des éléments f_m, dans le problème initial.

A.2.1 Définition des conditions d'optimalité

Commençons par rappeler le problème initial :

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\boldsymbol{\xi},\boldsymbol{\sigma}}} & \frac{1}{2}\sum\limits_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i & \text{(A.8a)} \\ \text{s. c.} & \sum\limits_\ell \left(d_\ell^p \Big(\sum\limits_{m\in G_\ell} \sigma_m^{1/q} \Big)^q \right)^{1/(p+q)} \leq 1 \qquad \sigma_m \geq 0 \qquad \forall m & \text{(A.8b)} \\ & y_i \left(\sum\limits_m f_m(\boldsymbol{x}_i) + b \right) \geq 1 - \xi_i \qquad \xi_i \geq 0 \qquad \forall i \ . & \text{(A.8c)} \end{cases}$$

Le Lagrangien associé à ce problème est

$$\begin{aligned} \mathcal{L} \quad = \quad & \frac{1}{2}\sum_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \xi_i \\ & + \quad \lambda \left[\sum_\ell \left(d_\ell^p \Big(\sum_{m\in G_\ell} \sigma_m^{1/q} \Big)^q \right)^{1/(p+q)} - 1 \right] \quad - \quad \sum_m \mu_m \sigma_m \\ & - \quad \sum_i \alpha_i \left[y_i \left(\sum_m f_m(\boldsymbol{x}_i) + b \right) + \xi_i - 1 \right] \quad - \quad \sum_i \eta_i \xi_i \ . \end{aligned}$$

Les conditions d'optimalité pour les coefficients σ_m sont

$$\frac{\partial \mathcal{L}}{\partial \sigma_m} = -\frac{1}{2}\frac{\|f_m\|^2_{\mathcal{H}_m}}{\sigma_m^2} + \frac{\lambda}{p+q}\sigma_m^{(1-q)/q} \left(d_\ell^{-1} \sum_{m\in G_\ell} \sigma_m^{1/q} \right)^{-p/(p+q)} - \mu_m = 0 , \tag{A.9}$$

dont on déduit l'expression de σ_m

$$\begin{aligned} & \sigma_m = 0 \qquad \text{ou} \\ & \sigma_m = \left(\frac{2\lambda}{p+q} \right)^{-q/(q+1)} \left(d_\ell^{-1} \sum_{m\in G_\ell} \sigma_m^{1/q} \right)^{\frac{pq}{(p+q)(q+1)}} \left(\|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}} \right)^q \ . \end{aligned} \tag{A.10}$$

A.2.2 Expression de σ_m en fonction de f_m

Pour trouver l'expression de σ_m en fonction de f_m, transforme (A.10) de la façon suivante :

$$\sigma_m^{1/q} \;=\; \left(\frac{2\,\lambda}{p+q}\right)^{-1/(q+1)} \left(d_\ell^{-1} \sum_{m\in G_\ell} \sigma_m^{1/q}\right)^{\frac{p}{(p+q)(q+1)}} \left(\|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}\right) .$$

On en déduit

$$\sum_{m\in G_\ell} \sigma_m^{1/q} \;=\; \left(\frac{2\,\lambda}{p+q}\right)^{-1/(q+1)} \left(d_\ell^{-1} \sum_{m\in G_\ell} \sigma_m^{1/q}\right)^{\frac{p}{(p+q)(q+1)}} \sum_{m\in G_\ell} \left(\|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}\right) ,$$

puis

$$\left(\sum_{m\in G_\ell} \sigma_m^{1/q}\right)^{\frac{q(p+q+1)}{(p+q)(q+1)}} \;=\; \left(\frac{2\,\lambda}{p+q}\right)^{-1/(q+1)} \left(d_\ell^{-1}\right)^{\frac{p}{(p+q)(q+1)}} \sum_{m\in G_\ell} \left(\|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}\right) .$$

Finalement, on obtient

$$\sum_{m\in G_\ell} \sigma_m^{1/q} \;=\; \left(c^{-(p+q)}\; d_\ell^{-p}\; s_\ell^{(p+q)(q+1)}\right)^{\frac{1}{q(p+q+1)}} \quad , \tag{A.11}$$

où $s_\ell = \sum_{m\in G_\ell} \|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}$, et où $c = \dfrac{2\,\lambda}{p+q}$ est une constante qui ne dépend pas de l'élément σ_m considéré. La constante c est déterminée en reportant l'expression (A.11) dans la partie gauche de la contrainte (A.8b), qui à l'optimalité vaut 1. On trouve

$$c \quad = \quad \left[\sum_\ell d_\ell^{\frac{p}{p+q+1}}\; (s_\ell)^{\frac{q+1}{p+q+1}}\right]^{(p+q+1)} \quad . \tag{A.12}$$

En reportant (A.11) et (A.12) dans (A.10), on obtient

$$\sigma_m = \left(\sum_\ell \left(d_\ell^p\; s_\ell^{(q+1)}\right)^{\frac{1}{p+q+1}}\right)^{-(p+q)} \left(d_\ell^{-1}\; s_\ell\right)^{\frac{p}{p+q+1}} \left(\|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}\right)^q \tag{A.13}$$

A.2.3 Expression du problème initial en fonction de f_m

En intégrant (A.13) dans (A.8a), et en remplaçant s_ℓ par sa valeur, le problème final s'exprime uniquement en fonction des éléments f_m

$$\begin{cases} \min\limits_{\substack{f_1,\dots,f_M,\\ b,\xi}} & \dfrac{1}{2}\left(\sum\limits_\ell d_\ell^{\frac{p}{p+q+1}} \left(\sum\limits_{m\in G_\ell} \|f_m\|_{\mathcal{H}_m}^{\frac{2}{q+1}}\right)^{\frac{q+1}{p+q+1}}\right)^{p+q+1} \quad + \quad C\sum\limits_i \xi_i \\ \text{s. c.} & y_i\left(\sum\limits_m f_m(\boldsymbol{x}_i) + b\right) \geq 1-\xi_i \qquad\qquad \xi_i \geq 0 \quad \forall i \ , \end{cases}$$

Cette dernière expression correspond à la formulation (4.11) de la proposition 4.5. □

A.3 Éléments de comparaison des algorithmes

Soit

$$C_1(\nu, \boldsymbol{\beta}) \;=\; \min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \;+\; \nu\, P(\boldsymbol{\beta})\;,$$

et

$$C_2(\lambda, \boldsymbol{\beta}) \;=\; \min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \;+\; \lambda\, P(\boldsymbol{\beta})^{2/r}\;.$$

Les conditions d'optimalités pour ces deux problèmes sont :

$$\begin{aligned}
\frac{\partial C_1(\nu, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \nu\, \frac{\partial P(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &&= 0 \\
\frac{\partial C_2(\lambda, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \frac{2\lambda}{r}\, P(\boldsymbol{\beta})^{(2-r)/r}\, \frac{\partial P(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &&= 0\,, \qquad \text{(A.14)}
\end{aligned}$$

et nous permettent de définir λ en fonction de ν. Soient :

$$\begin{aligned}
\boldsymbol{\beta}^\star &= \arg\min_{\boldsymbol{\beta}}\; C_1(\nu, \boldsymbol{\beta}) \\
\lambda^\star &= \frac{\nu\, r}{2}\, P(\boldsymbol{\beta}^\star)^{(r-2)/r} \quad .
\end{aligned}$$

L'expression (A.14) devient :

$$\frac{\partial C_2(\lambda^\star, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \;=\; \frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \;+\; \nu\, P(\boldsymbol{\beta}^\star)^{(r-2)/r}\, P(\boldsymbol{\beta})^{(2-r)/r}\, \frac{\partial P(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \;=\; 0\,.$$

Or pour $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$, on a $\dfrac{\partial C_2(\lambda^\star, \boldsymbol{\beta}^\star)}{\boldsymbol{\beta}} \;=\; \dfrac{\partial C_1(\nu, \boldsymbol{\beta}^\star)}{\partial \boldsymbol{\beta}}$.
Donc pour $\boldsymbol{\beta}^\star = \arg\min_{\boldsymbol{\beta}}\; C_1(\nu, \boldsymbol{\beta})$, on a $\boldsymbol{\beta}^\star = \arg\min_{\boldsymbol{\beta}}\; C_2(\lambda^\star, \boldsymbol{\beta})$.

A.4 Définition d'un seuillage pour l'algorithme 2

A.4.1 Cadre paramétrique

Rappelons le critère minimisé par l'algorithme 2 sous sa forme variationnelle :

$$\begin{cases} \min\limits_{\beta,\sigma} & \sum\limits_i \left(y_i - \sum\limits_m x_i^m \beta_m \right)^2 + \lambda \sum\limits_m \dfrac{\beta_m^2}{\sigma_m} \\ \text{s. c.} & \sum\limits_\ell \left(d_\ell^p \Big(\sum\limits_{m \in G_\ell} \sigma_m^{1/q} \Big)^q \right)^{1/(p+q)} \leq 1 \qquad \sigma_m \geq 0 \qquad \forall m \,. \end{cases}$$

Le Lagrangien associé est

$$\begin{aligned} \mathcal{L} \;=\; & \sum_i \left(y_i - \sum_m x_i^m \beta_m \right)^2 \;+\; \lambda \sum_m \frac{\beta_m^2}{\sigma_m} \;+ \\ & \nu \left[\sum_\ell \left(d_\ell^p \Big(\sum_{m \in G_\ell} \sigma_m^{1/q} \Big)^q \right)^{1/(p+q)} - 1 \right] \;-\; \sum_m \mu_m \sigma_m \,, \end{aligned}$$

dont on déduit la condition d'optimalité

$$\frac{\partial \mathcal{L}}{\partial \sigma_m} \;=\; \frac{\nu}{p+q} \, \frac{d_\ell^{p/(p+q)} \, \sigma_m^{(1-q)/q}}{\left(\sum\limits_{m \in G_\ell} \sigma_m^{1/q} \right)^{p/(p+q)}} \;-\; \frac{\lambda}{2} \frac{\beta_m^2}{\sigma_m^2} \;-\; \mu_m \;=\; 0 \,. \quad \text{(A.15)}$$

***Remarque* A.2 —** Les multiplicateurs de Lagrange μ_m sont nuls lorsque les contraintes associées ne sont pas saturées, c'est à dire lorsque $\sigma_m > 0$. À l'optimum, soit $\sigma_m > 0$ et $\mu_m = 0$, soit $\sigma_m = 0$ et $\mu_m > 0$. Ainsi, $\mu_m \, \sigma_m = 0$, $\forall m$. Multiplier (A.15) par σ_m annule donc la dernière partie de l'expression. ◇

En multipliant (A.15) par σ_m, on obtient

$$\begin{aligned} 0 \;&=\; \frac{\nu}{p+q} \, d_\ell^{p/(p+q)} \left(\sum_{m \in G_\ell} \sigma_m^{1/q} \right)^{-p/(p+q)} \sigma_m^{1/q} \;-\; \frac{\lambda}{2} \frac{\beta_m^2}{\sigma_m} \\ &=\; \frac{\nu}{p+q} \, d_\ell^{p/(p+q)} \left(\sum_{m \in G_\ell} \sigma_m^{1/q} \right)^{-p/(p+q)} \sum_{m \in G_\ell} \sigma_m^{1/q} \;-\; \frac{\lambda}{2} \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_m} \\ &=\; \frac{\nu}{p+q} \, d_\ell^{p/(p+q)} \left(\sum_{m \in G_\ell} \sigma_m^{1/q} \right)^{q/(p+q)} \;-\; \frac{\lambda}{2} \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_m} \\ &=\; \frac{\nu}{p+q} \sum_\ell \left(d_\ell^p \left(\sum_{m \in G_\ell} \sigma_m^{1/q} \right)^q \right)^{1/(p+q)} \;-\; \frac{\lambda}{2} \sum_\ell \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_m} \\ \Rightarrow \quad \nu \;&=\; (p+q) \, \frac{\lambda}{2} \sum_\ell \sum_{m \in G_\ell} \frac{\beta_m^2}{\sigma_m} \,. \end{aligned} \quad \text{(A.16)}$$

Ainsi, en reportant (A.16) dans (A.15), on obtient :

$$\frac{\lambda}{2}\sum_{\ell}\sum_{m\in G_\ell}\frac{\beta_m^2}{\sigma_m}\; d_\ell^{p/(p+q)}\left(\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{-p/(p+q)}\sigma_m^{(1-q)/q} - \frac{\lambda}{2}\frac{\beta_m^2}{\sigma_m^2} - \mu_m \;=\; 0\,,$$

et pour $\mu_m > 0$,

$$\begin{aligned}\frac{\beta_m^2}{\sigma_m} &< \sum_{\ell}\sum_{m\in G_\ell}\frac{\beta_m^2}{\sigma_m^2}\;\left(d_\ell^{-1}\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{-p/(p+q)}\sigma_m^{(1-q)/q}\\ \Rightarrow\quad \sigma_m &= 0\,.\end{aligned}$$

A.4.2 Cadre non paramétrique

De la même façon qu'à la section précédente, nous pouvons énoncer une condition de seuillage dans le cadre non paramétrique. Pour le problème initial (A.8), on peut récrire la condition d'optimalité (A.9) en utilisant le gradient de $J(\sigma)$ défini par l'expression (4.16) :

$$\frac{\partial\,\mathcal{L}}{\partial\,\sigma_m} \;=\; \frac{\partial J(\sigma)}{\partial\,\sigma_m} + \frac{\lambda\,\sigma_m^{(1-q)/q}}{p+q}\left(d_\ell^{-1}\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{-p/(p+q)} - \mu_m \;=\; 0\,, \quad \text{(A.17)}$$

À l'optimum, on a $\mu_m\,\sigma_m = 0$, $\forall m$ (cf. remarque A.2). Ainsi, $\forall\ \sigma_m \neq 0$, on déduit de (A.17) l'expression de λ :

$$\begin{aligned}0 &= \frac{\lambda\,d_\ell^{p/(p+q)}}{p+q}\left(\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{-p/(p+q)}\sigma_m^{1/q} + \sigma_m\frac{\partial J(\sigma)}{\partial\,\sigma_m}\\ &= \frac{\lambda\,d_\ell^{p/(p+q)}}{p+q}\left(\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{-p/(p+q)}\sum_{m\in G_\ell}\sigma_m^{1/q} + \sum_{m\in G_\ell}\sigma_m\frac{\partial J(\sigma)}{\partial\,\sigma_m}\\ &= \frac{\lambda\,d_\ell^{p/(p+q)}}{p+q}\left(\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{q/(p+q)} + \sum_{m\in G_\ell}\sigma_m\frac{\partial J(\sigma)}{\partial\,\sigma_m}\\ &= \frac{\lambda}{p+q}\sum_{\ell}\left(d_\ell^{p}\left(\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{q}\right)^{1/(p+q)} + \sum_{\ell}\sum_{m\in G_\ell}\sigma_m\frac{\partial J(\sigma)}{\partial\,\sigma_m}\\ \Rightarrow\quad \lambda &= -(p+q)\sum_{\ell}\sum_{m\in G_\ell}\sigma_m\frac{\partial J(\sigma)}{\partial\,\sigma_m}\,. \qquad \text{(A.18)}\end{aligned}$$

Ainsi, en reportant (A.18) dans (A.17), on a pour $\mu_m > 0$

$$\begin{aligned}\frac{\partial J(\sigma)}{\partial\sigma_m} &> \sum_{\ell}\sum_{m\in G_\ell}\sigma_m\frac{\partial J(\sigma)}{\partial\,\sigma_m}\left(d_\ell^{-1}\sum_{m\in G_\ell}\sigma_m^{1/q}\right)^{-p/(p+q)}\sigma_m^{(1-q)/q}\\ \Rightarrow\quad \sigma_m &= 0\,.\end{aligned}$$

***Remarque* A.3 —** Dans le cadre non paramétrique, le sens de l'inégalité du seuillage est différent de celui obtenu dans le cadre paramétrique. Cela peut sembler contre intuitif. Rappellons donc que l'expression du gradient $\dfrac{\partial J(\sigma)}{\partial\,\sigma_m}$, définie par (4.16), est une valeur négative. ◇

A.5 Double validation croisée

La double validation croisée consiste à évaluer l'erreur de test sur un ensemble l'apprentisage, par une procédure de validation croisée. Les hyperparamètres de régularisation optimaux pour estimer cette erreur de test sont choisis, au sein de cette procédure, par validation croisée également.

Ainsi, on évalue par validation croisée :
– le paramètre de régularisation optimal, dans une boucle interne ;
– l'erreur de test, dans une boucle externe.

Algorithme 3 : Double validation croisée

Entrées :
$(\mathbf{X}, \mathbf{y})$: la matrice des observations et le vecteur des réponses associé
$\mathbf{C}$: le vecteur d'hyperparamètres
K_{ext} : le nombre de blocs utilisé pour évaluer l'erreur de test
K_{int} : le nombre de blocs utilisé pour choisir l'hyperparamètre

pour i allant de 1 à K_{ext} **faire**
- $train_{ext}[i] \leftarrow (K_{ext} - 1)$ blocs de $\mathbf{X}$
- $valid_{ext}[i] \leftarrow 1$ bloc de $\mathbf{X}$
- **pour** j allant de 1 à K_{int} **faire**
  - $train_{int}[j] \leftarrow (K_{int} - 1)$ blocs de $train_{ext}[i]$
  - $valid_{int}[j] \leftarrow 1$ bloc de $train_{ext}[i]$
  - **pour** k allant de 1 à taille($\mathbf{C}$) **faire**
    - $model_{int}[j, k] \leftarrow$ apprentissage($train_{int}[j]$, $C[k]$)
    - $erreur_{int}[j, k] \leftarrow$ erreur($model_{int}[j, k]$, $valid_{int}[j]$)
- $erreur_{int}[i] \leftarrow$ moyenne $_j(erreur_{int}[j, k])$
- $C_{opt}[i] \leftarrow$ min($erreur_{int}[i]$)
- $model_{ext}[i] \leftarrow$ apprentissage($train_{ext}[i]$, $C_{opt}[i]$)
- $erreur_{ext}[i] \leftarrow$ erreur($model_{ext}[i]$, $valid_{ext}[i]$)

$erreur_{test} \leftarrow$ moyenne $_i(erreur_{ext}[i])$

Bibliographie

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin et G. Sherlock,
« Gene Ontology : Tool for the unification of biology. The Gene Ontology Consortium. »,
Nature Genetics, vol. 25, n° 1, p. 25–29, 2000.
(Cité page 2.)

F. R. Bach, G. R. G. Lanckriet et M. I. Jordan,
« Multiple kernel learning, conic duality, and the SMO algorithm »,
Dans *Proceedings of the 21th Annual International Conference on Machine Learning (ICML 2004)*, vol. ACM International Conference Proceeding Series, p. 41–48, ACM, 2004.
(Cité page 31.)

F. Bach,
« Bolasso : model consistent lasso estimation through the bootstrap »,
Dans *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, A. McCallum et S. Roweis (coordinateurs), p. 33–40, Omnipress, 2008.
(Cité page 27.)

B. Blankertz, K.-R. Mueller, G. Curio, T. Vaughan, G. Schalk, J. Wolpaw, A. Schloegl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schroeder et N. Birbaumer,
« The BCI competition 2003 : progress and perspectives in detection and discrimination of EEG single trials »,
IEEE Transactions on Biomedical Engineering, vol. 51, n° 6, p. 1044–1051, 2004.
(Cité page 73.)

J. Bonnans et A. Shapiro,
« Optimization problems with pertubation : A guided tour »,
SIAM Review, vol. 40, n° 2, p. 228–264, 1998.
(Cité page 58.)

J. Bonnans, J. Gilbert, C. Lemaréchal et C. Sagastizábal,
Numerical Optimization – Theoretical and Practical Aspects,
vol. Universitext, Springer Verlag, Berlin, Berlin, Germany, second édition, 2006.
(Cité page 37.)

O. Bousquet et A. Elisseeff,
« Stability and generalization »,
Journal of Machine Learning Research, vol. 2, p. 499–526, 2002.
(Cité page 24.)

S. Boyd et L. Vandenberghe,
Convex optimization,
Cambridge University Press, New York, NY, USA, 2004.
(Cité pages 21, 56 et 58.)

L. Breiman,
« Heuristics of instability and stabilization in model selection »,
Annals of Statistics, vol. 24, n° 6, p. 2350–2383, 1996.
(Cité pages 24 et 45.)

L. Breiman,
« Better subset regression using the nonnegative garrote »,
Technometrics, vol. 37, n° 4, p. 373–384, 1995.
(Cité page 25.)

G. Celeux, E. Diday, G. Govaert, Y. Lechevallier et H. Ralambondrainy,
Classification automatique des données : environnement statistique et informatique,
Dunod, Paris, France, 1989.
(Cité page 7.)

J. Cham,
PhD Comics, vol. 3,
Piled Higher & Depper Publishing, Los Angeles, California, USA,
adresse : `http://www.phdcomics.com/comics/archive.php?comicid=844`,
2007.
(Cité page v.)

S. S. Chen, D. L. Donoho et M. A. Saunders,
« Atomic decomposition by Basis Pursuit »,
SIAM Journal on Scientific Computing, vol. 20, p. 33–61, 1998.
(Cité page 26.)

I. Daubechies, M. Defrise et C. D. Mol,
« An iterative thresholding algorithm for linear inverse problems with a sparsity constraint »,
Communications on Pure and Applied Mathematics, vol. 57, n° 11, p. 1413–1457, 2004.
(Cité pages 32 et 33.)

I. Daubechies, M. Fornasier et I. Loris,
« Accelerated projected gradient method for linear inverse problems with sparsity constraints »,
Journal of Fourier Analysis and Applications, à paraître.
(Cité page 34.)

D. L. Donoho, M. Elad et V. Temlyakov,

« Stable recovery of sparse overcomplete representations in the presence of noise »,
IEEE Transactions on Information Theory, vol. 52, nº 1, p. 6–18, 2004.
(Cité page 27.)

B. Efron, T. Hastie, I. Johnstone et R. Tibshirani,
« Least Angle Regression »,
Annals of Statistics, vol. 32, nº 2, p. 407–499, 2004.
(Cité pages 38 et 39.)

A. Farwell et E. Donchin,
« Talking off the top of your head : toward a mental prosthesis utilizing event-related brain potentials »,
Electroencephalography and Clinical Neurophysiology, vol. 70, nº 6, p. 510–523, 1998.
(Cité page 73.)

V. Franc et S. Sonnenburg,
« Optimized cutting plane algorithm for support vector machines »,
Dans *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, A. McCallum et S. Roweis (coordinateurs), p. 320–327, Omnipress, 2008.
(Cité pages 36 et 37.)

I. E. Frank et J. H. Friedman,
« A statistical view of some chemometrics regression tools »,
Technometrics, vol. 35, nº 2, p. 109–148, 1993.
(Cité page 21.)

W. J. Fu,
« Penalized regressions : The bridge versus the lasso »,
Journal of Computational and Graphical Statistic, vol. 7, nº 3, p. 397–416, 1998.
(Cité pages 32 et 35.)

Y. Grandvalet,
« Least absolute shrinkage is equivalent to quadratic penalization »,
Dans *Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN 1998)*, p. 201–206, Springer, 1998.
(Cité pages 3, 27 et 31.)

Y. Grandvalet et S. Canu,
« Adaptive scaling for feature selection in SVMs »,
Dans *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun et K. Obermayer (coordinateurs), p. 553–560, MIT Press, 2003.
(Cité page 27.)

V. Guigue,
Méthodes à noyaux pour la représentation et la discrimination de signaux non-stationnaires,
Thèse de doctorat, Institut National des Sciences Appliquées de Rouen, 2005.
(Cité page 39.)

I. Guyon et A. Elisseeff,
« An introduction to variable and feature selection »,
Journal of Machine Learning Research, vol. 3, p. 1157–1182, 2003.
(Cité page 19.)

T. Hastie, R. Tibshirani et J. H. Friedman,
The Elements of Statistical Learning,
Springer, 2001.
(Cité page 27.)

T. Hesterberg, N. H. Choi, M. L. et C. Fraley,
« Least angle and ℓ_1 penalized regression : a review »,
Statistics Surveys, vol. 2, p. 61–93, 2008.
(Cité page 19.)

J.-B. Hiriart-Urruty et C. Lemarechal,
Convex Analysis and Minimization Algorithms,
Springer-Verlag New York, Inc., New York, NY, USA, second édition, 1996.
(Cité page 35.)

A. E. Hoerl et R. W. Kennard,
« Ridge regression : Biased estimation for nonorthogonal problems »,
Technometrics, vol. 12, n° 1, p. 55–67, 1970.
(Cité page 25.)

J. Huang, S. Ma et C.-H. Zhang,
« Adaptive lasso for sparse high-dimensional regression models »,
Rapport technique, University of Iowa, 2007.
(Cité page 27.)

T. Joachims,
« Training linear svms in linear time »,
Dans *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD 2006)*, T. Eliassi-Rad, L. H. Ungar, M. Craven et D. Gunopulos (coordinateurs), p. 217–226, ACM, 2006.
(Cité page 35.)

K. Knight,
« Discussion on Least Angle Regression »,
Annals of Statistics, vol. 32, n° 2, p. 458–460, 2004.
(Cité page 24.)

K. Knight et W. Fu,
« Asymptotics for lasso-type estimators »,
Annals of Statistics, vol. 28, n° 5, p. 1356–1378, 2000.
(Cité page 27.)

G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui et M. I. Jordan,
« Learning the kernel matrix with semi-definite programming »,
Journal of Machine Learning Research, vol. 5, p. 27–72, 2004.
(Cité pages 3 et 30.)

S.-I. Lee, H. Lee, P. Abbeel et A. Ng,
« Efficient L_1 regularized logistic regression »,
Dans *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI 2006)*, p. 1–9, 2006.
(Cité page 39.)

C. Leng, Y. Lin et G. Wahba,
« A note on lasso and related procedures in model selection »,
Statistica Sinica, vol. 16, n° 4, p. 1273–1284, 2004.
(Cité page 26.)

N. Meinshausen et P. Bühlmann,
« High-dimensional graphs and variable selection with the Lasso »,
Annals of Statistics, vol. 34, n° 3, p. 1436–1462, 2006.
(Cité page 27.)

C. de Mol, E. de Vito et L. Rosasco,
« Elastic-net regularization in learning theory »,
2008, adresse : `http://arxiv.org/abs/0807.3423`.
(Cité page 28.)

M. Nikolova,
« Local strong homogeneity of a regularized estimator »,
SIAM Journal on Applied Mathematics, vol. 61, n° 2, p. 633–658, 2000.
(Cité page 22.)

M. R. Osborne, B. Presnell et B. A. Turlach,
« On the lasso and its dual »,
Journal of Computational and Graphical Statistics, vol. 9, n° 2, p. 319–337, 2000a.
(Cité pages 23, 26, 34, 35 et 51.)

M. Osborne, B. Presnell et B. Turlach,
« A new approach to variable selection in least squares problems »,
IMA Journal of Numerical Analysis, vol. 20, n° 3, p. 389–403, 2000b.
(Cité page 38.)

M. Y. Park et T. Hastie,
« Regularization path algorithms for detecting gene interactions »,
Rapport technique, Stanford University, 2006.
(Cité page 39.)

M. Y. Park et T. Hastie,
« L_1-regularization path algorithm for generalized linear models »,
Journal of the Royal Statistical Society, Series B, vol. 69, n° 4, p. 659–677, 2007.
(Cité page 39.)

S. Perkins, K. Lacker et J. Theiler,
« Grafting : Fast, incremental feature selection by gradient descent in function space »,
Journal of Machine Learning Research, vol. 3, p. 1333–1356, 2003.
(Cité page 35.)

A. Rakotomamonjy et V. Guigue,
« BCI competition 3 : Dataset 2 - ensemble of SVM for BCI P300 speller »,
IEEE Transactions on Biomedical Engineering, vol. 55, n° 3, p. 1147–1154, 2008.
(Cité page 73.)

A. Rakotomamonjy, F. Bach, S. Canu et Y. Grandvalet,
« More efficiency in Multiple Kernel Learning »,
Dans *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Z. Ghahramani (coordinateur), p. 775–782, Omnipress, 2007.
(Cité pages 31, 54, 55, 58 et 59.)

S. Rosset et J. Zhu,
« Piecewise linear regularized solution paths »,
Annals of Statistics, vol. 35, n° 3, p. 1012–1030, 2007.
(Cité pages 38 et 39.)

V. Roth et B. Fischer,
« The group-lasso for generalized linear models : uniqueness of solutions and efficient algorithms »,
Dans *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, A. McCallum et S. Roweis (coordinateurs), p. 848–855, Omnipress, 2008.
(Cité page 35.)

V. Roth,
« The generalized lasso »,
IEEE Transactions on Neural Networks, vol. 15, n° 1, p. 16–28, 2004.
(Cité page 35.)

M. Schmidt, adresse : `http://www.cs.ubc.ca/~schmidtm/Software/lasso.html`.
(Cité page 32.)

D. Schuurmans et F. Southey,
« Metric-based methods for adaptive model selection and regularization »,
Machine Learning, vol. 48, n° 1-3, p. 51–84, 2002,
Special issue on new methods for model selection and model combination.
(Cité page 24.)

B. Schölkopf et A. J. Smola,
Learning with kernels : Support Vector Machines, regularization, optimization, and beyond,
MIT Press, Cambridge, MA, USA, 2001.
(Cité page 12.)

A. Smola, S. V. N. Vishwanathan et Q. Le,
« Bundle methods for machine learning »,
Dans *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer et S. Roweis (coordinateurs), p. 1377–1384, MIT Press, 2008.
(Cité page 35.)

S. Sonnenburg, G. Rätsch, C. Schäfer et B. Schölkopf,
« Large scale Multiple Kernel Learning »,
Journal of Machine Learning Research, vol. 7, p. 1531–1565, 2006.
(Cité page 37.)

P. Soularue et X. Gidrol,
« Puces à ADN »,
Techniques de l'ingénieur, vol. RE 6, 2002.
(Cité page 1.)

M. Szafranski, Y. Grandvalet et P. Morizet-Mahoudeaux,
« Hierarchical penalization »,
Dans *Conférence d'Apprentissage (CAp)*, Cépaduès, 2007.
(Cité pages 3 et 85.)

M. Szafranski, Y. Grandvalet et P. Morizet-Mahoudeaux,
« Hierarchical penalization »,
Dans *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer et S. Roweis (coordinateurs), p. 1457–1464, MIT Press, 2008a.
(Cité pages 3, 30 et 85.)

M. Szafranski, Y. Grandvalet et A. Rakotomamonjy,
« Composite Kernel Learning »,
Dans *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, A. McCallum et S. Roweis (coordinateurs), p. 1040–1047, Omnipress, 2008b.
(Cité pages 3, 72 et 86.)

M. Szafranski, Y. Grandvalet et A. Rakotomamonjy,
« Learning with groups of kernels »,
Dans *Conférence d'Apprentissage (CAp)*, Cépaduès, 2008c.
(Cité pages 3 et 86.)

R. Tibshirani,
« Regression shrinkage and selection via the lasso »,
Journal of the Royal Statistical Society, Series B, vol. 58, n° 1, p. 267–288, 1996.
(Cité page 25.)

A. N. Tikhonov et V. Y. Arsénin,
Solution of ill-posed problems,
W. H. Wilson, Washington, D. C, USA, 1977.
(Cité page 25.)

V. N. Vapnik,
The nature of statistical learning theory,
Springer-Verlag New York, Inc., New York, NY, USA, 1995.
(Cité pages 8 et 11.)

M. Yuan et Y. Lin,
« Model selection and estimation in regression with grouped variables »,
Journal of the Royal Statistical Society, Series B, vol. 68, n° 1, p. 49–67, 2006.
(Cité pages 27, 29, 33, 39, 66 et 68.)

P. Zhao, G. Rocha et B. Yu,
« The Composite Absolute Penalties family for grouped and hierarchical variable selection »,
Annals of Statistics, à paraître.
(Cité pages 30, 39, 48 et 88.)

P. Zhao et B. Yu,
« On model selection consistency of lasso »,
Journal of Machine Learning Research, vol. 7, p. 2541–2563, 2006.
(Cité page 27.)

H. Zou,
« The adaptive lasso and its oracle properties »,
Journal of the American Statistical Association, vol. 101, p. 1418–1429, 2006.
(Cité pages 27 et 28.)

H. Zou et T. Hastie,
« Regularization and variable selection via the elastic net »,
Journal of The Royal Statistical Society, Series B, vol. 67, n° 2, p. 301–320, 2005.
(Cité page 28.)

Ce document a été préparé à l'aide de l'éditeur de texte GNU Emacs et du logiciel de composition typographique LaTeX 2ε. Le style du document est basé sur le modèle de thèse élaboré par Julien Chiquet.

Titre Pénalités hiérarchiques pour l'intégration de connaissances dans les modèles statistiques

Résumé L'apprentissage statistique vise à prédire, mais aussi analyser ou interpréter un phénomène. Dans cette thèse, nous proposons de guider le processus d'apprentissage en intégrant une connaissance relative à la façon dont les caractéristiques d'un problème sont organisées. Cette connaissance est représentée par une structure arborescente à deux niveaux, ce qui permet de constituer des groupes distincts de caractéristiques. Nous faisons également l'hypothèse que peu de (groupes de) caractéristiques interviennent pour discriminer les observations. L'objectif est donc de faire émerger les groupes de caractéristiques pertinents, mais également les caractéristiques significatives associées à ces groupes. Pour cela, nous utilisons une formulation variationnelle de type pénalisation adaptative. Nous montrons que cette formulation conduit à minimiser un problème régularisé par une norme mixte. La mise en relation de ces deux approches offre deux points de vues pour étudier les propriétés de convexité et de parcimonie de cette méthode. Ces travaux ont été menés dans le cadre d'espaces de fonctions paramétriques et non paramétriques. L'intérêt de cette méthode est illustré sur des problèmes d'interfaces cerveaux-machines.

Mots-clés Apprentissage statistique supervisé ; parcimonie ; régularisation ; lasso ; normes mixtes ; sélection de variables et de caractéristiques ; Séparateurs à Vaste Marge (SVM) ; apprentissage de noyaux.

Title Hierarchical penalties for integrating prior knowledge in statistical models

Abstract Supervised learning aims at predicting, but also analyzing or interpreting an observed phenomenon. Hierarchical penalization is a generic framework for integrating prior information in the fitting of statistical models. This prior information represents the relations shared by the characteristics of a given studied problem. In this thesis, the characteristics are organized in a two-levels tree structure, which defines distinct groups. The assumption is that few (groups of) characteristics are involved to discriminate between observations. Thus, for a learning problem, the goal is to identify relevant groups of characteristics, and at the same time, the significant characteristics within these groups. An adaptive penalization formulation is used to extract the significant components of each level. We show that the solution to this problem is equivalent to minimize a problem regularized by a mixed norm. These two approaches have been used to study the convexity and sparseness properties of the method. The latter is derived in parametric and non parametric function spaces. Experiences on brain-computer interfaces problems support our approach.

Keywords Supervized learning ; sparseness ; regularization ; lasso ; mixed norms ; variable and feature selection ; SVM ; kernel learning.