Dossier de candidature à un poste de Maître de conférences

Aurélie Fischer

Qualifiée en section 26 du CNU.

Date : Février 2012
Numéro de qualification : 12226224246

Détails du poste

Sections 26, 25 et 27
Poste n°4073
Profil : Statistiques du risque et Fouilles de données
Institut de Mathématiques de Jussieu et Laboratoire de Probabilités et Modèles Aléatoires
Université Denis Diderot – Paris 7

Table des matières

Thèmes de recherche

Mon domaine d'intérêt majeur est l'apprentissage statistique.
Mes principaux thèmes de recherche sont les suivants :

- Quantification, clustering
- Divergences de Bregman
- Courbes principales
- Grande dimension, projection sur des bases
- Sélection de modèle, choix de paramètres
- Agrégation d'estimateurs

1 Présentation

1.1 Recherche

Après un Master 2 de Probabilités et Statistique à l'Université Paris-Sud, j'ai préparé à l'Université Pierre et Marie Curie, sous la direction de Gérard Biau, une thèse dont le thème principal est l'apprentissage statistique. Une partie des travaux effectués concerne la quantification et le clustering, et l'autre est consacrée aux courbes principales. Je me suis intéressée à des questions théoriques, sans pour autant négliger l'implémentation des méthodes considérées et leur mise en oeuvre sur des applications. J'ai tout d'abord obtenu des résultats sur l'utilisation en quantification et clustering des divergences de Bregman, fonctions de deux variables qui peuvent être non symétriques ou s'appliquer à des fonctions. Ensuite, dans le contexte de l'ingénierie nucléaire, j'ai examiné une méthode de réduction de la dimension par projection pour classer des courbes d'évolution temporelle de certaines quantités physiques. Je me suis également penchée sur un problème de sélection de paramètres pour les courbes principales, courbes passant "au milieu" d'un nuage de points, qui peuvent être vues comme une généralisation non linéaire de l'analyse en composantes principales. Actuellement, je poursuis mon travail sur les courbes principales, et j'étudie aussi les propriétés d'une méthode d'agrégation d'estimateurs, dans la perspective d'applications en médecine génomique.

J'aimerais conserver dans mes recherches à venir les deux composantes, théorie et applications. Je souhaite continuer à explorer des problèmes liés à l'apprentissage statistique ou la statistique non-paramétrique, en m'intéressant à des questions de sélection de modèle et de grande dimension, thématiques présentes au sein du *Laboratoire de Probabilités et Modèles Aléatoires*. Je suis également tout à fait disposée à m'ouvrir à la problématique du risque, ainsi qu'à de nouveaux domaines de recherche représentés dans cette équipe.

1.2 Enseignement

Au cours de ma thèse, j'ai assuré, en tant que monitrice à l'Université Pierre et Marie Curie, des travaux dirigés d'Algèbre et d'Analyse pour le niveau L1, et de Probabilités et Statistique pour le niveau L3. Cette année, je suis Attachée Temporaire d'Enseignement et de Recherche à l'Université Paris Descartes et j'enseigne en Probabilités et Statistique dans le département Statistique et Informatique Décisionnelle (STID) de l'IUT, sous forme de travaux dirigés, travaux pratiques sur ordinateur, cours et encadrement de projets. Mon public est constitué d'étudiants en première année de DUT et en année spéciale, parcours permettant à des étudiants qui se sont réorientés d'obtenir leur diplôme en un an.

En fonction des besoins de l'unité au sein de laquelle j'effectuerai mon service, je suis prête à enseigner aussi bien des Probabilités et Statistique, qu'il s'agisse de modules fondamentaux ou plus spécifiques à mon domaine de recherche, que des Mathématiques générales. Je peux également m'investir dans des enseignements spécialisés plus éloignés de mon champ de compétences.

2 Curriculum Vitae

Aurélie FISCHER

Née le 18 février 1985 à Forbach (Moselle).
Nationalité française.

Laboratoire MAP5, Université Paris Descartes
45 rue des Saints-Pères, 75006 Paris

Département STID, IUT Paris Descartes
143 avenue de Versailles, 75016 Paris

Téléphone : 01 83 94 58 91 ou 06 70 59 28 55
Adresse électronique : `aurelie.fischer@parisdescartes.fr`
Page internet : `http://www.lsta.upmc.fr/doct/fischer`

2.1 Situation actuelle

2011–2012 Attachée Temporaire d'Enseignement et de Recherche (temps complet) à l'Université Paris Descartes – Paris 5.
Membre du Département STID de l'IUT et du Laboratoire Mathématiques Appliquées à Paris 5 (MAP5).

2.2 Parcours universitaire

2008–2011 Allocataire de recherche et monitrice à l'Université Pierre et Marie Curie – Paris 6.

Thèse de doctorat de Mathématiques préparée au Laboratoire de Statistique Théorique et Appliquée (LSTA), intitulée *Apprentissage statistique non supervisé : grande dimension et courbes principales*, soutenue le 9 juin 2011, mention Très honorable.

Directeur de thèse : Gérard BIAU.

Jury : Gérard BIAU (Université Pierre et Marie Curie)
Jérôme DEDECKER (Université Paris Descartes)
Paul DEHEUVELS (Université Pierre et Marie Curie)
Fabrice GAMBOA (*Rapporteur*, Université Paul Sabatier de Toulouse)
Balázs KEGL (Université Paris-Sud)
Gábor LUGOSI (*Rapporteur*, Université Pompeu Fabra de Barcelone)
Pascal MASSART (Université Paris-Sud).

2005–2008 Ecole Normale Supérieure – Formation Interuniversitaire de Mathématiques Fondamentales et Appliquées.

- ◇ **Licence** de Mathématiques.
- ◇ **Maîtrise** de Mathématiques, mention Bien.
  Mémoire encadré par Philippe BIANE, intitulé *Sommes de carrés de polynômes.*
- ◇ **Master 2** de Probabilités et Statistiques – Université Paris-Sud, mention Bien.
  Mémoire encadré par Pascal MASSART, intitulé *Sur quelques méthodes d'analyse de données fonctionnelles.*
- ◇ **Agrégation** de Mathématiques (option Probabilités et Statistiques).

2003–2005 Lycée Fabert à Metz (Moselle) – Classes préparatoires MPSI et MP*.

2003 Lycée Jean Moulin à Forbach (Moselle) – Baccalauréat Scientifique, spécialité Mathématiques, section européenne, mention Très bien.

2.3 Divers

Langues

- Allemand
- Anglais

Compétences informatiques

- Outils bureautiques : LaTeX, Microsoft Word, Excel, PowerPoint
- Calcul et Statistique : R, Matlab, Scilab, SPSS, Maple
- Photo : Corel Paint Shop Pro

Activités culturelles et sportives

- Musique : orgue, chant
- Peinture
- Marche, natation

3 Publications et communications

3.1 Articles publiés ou acceptés dans des revues à comité de lecture

[1] Quantization and clustering with Bregman divergences.
Journal of Multivariate Analysis, **101**, 2207–2221 (2010).

[2] On the number of groups in clustering.
Statistics and Probability Letters, **81**, 1771–1781 (2011).

[3] Parameter selection for principal curves.
Avec Gérard Biau. *IEEE Transactions on Information Theory*, **58**, 1924–1939 (2012).

[4] Projection-based curve clustering.
Avec Benjamin Auder. Accepté en 2011 pour publication dans la revue *Journal of Statistical Computation and Simulation*.

Les articles [1]–[4] sont accessibles sur ma page internet `http://www.lsta.upmc.fr/doct/fischer`

3.2 Communications orales lors de congrès

Mai 2009	Atelier du groupe de recherche MASCOT–NUM, Institut Henri Poincaré, Paris.
Mai 2009	41^{es} Journées de Statistique de la SFdS, Bordeaux.
Sept. 2009	3^{es} Rencontres des Jeunes Statisticiens, Aussois.
Mai 2010	42^{es} Journées de Statistique de la SFdS, Marseille.
Août 2010	Statistique Mathématique et Applications, Fréjus.
Sept. 2011	4^{es} Rencontres des Jeunes Statisticiens, Aussois.

3.3 Exposés dans le cadre d'un séminaire ou groupe de travail

Oct. 2008	Groupe de Travail des doctorants du LSTA, Université Pierre et Marie Curie.
Mars 2009	CIES Jussieu.
Fév. 2010	Séminaire de Statistique, AgroParisTech.
Mars 2010	Groupe de Travail des Thésards et Jeunes Docteurs du MAP5, Université Paris Descartes.
Oct. 2010	Groupe de Travail des doctorants du LSTA, Université Pierre et Marie Curie.
Janv. 2011	Groupe de Travail de Statistique de Jussieu, Université Pierre et Marie Curie.
Oct. 2011	Groupe de Travail de Statistique du MAP5, Université Paris Descartes.
Oct. 2011	Séminaire de Probabilités et Statistique, Montpellier SupAgro.
Nov. 2011	Séminaire de l'équipe Probabilités et Statistiques, Université Paris-Sud.
Nov. 2011	Groupe de Travail en Statistique et Biostatistique, Institut Elie Cartan, Nancy.
Fév. 2012	Séminaire du Laboratoire Hubert Curien, Saint-Etienne.
Fév. 2012	Séminaire de Statistique, Institut de Mathématiques de Toulouse.
Mars 2012	Séminaire de Statistique, AgroParisTech.
Mars 2012	Séminaire de Probabilités et Statistique, Institut Camille Jordan, Lyon.

3.4 Participation à des conférences sans exposé

Déc. 2008	Rencontres de Statistique Mathématique 8 au CIRM à Luminy.
Déc. 2009	Colloque Maths à Venir à Paris.
Déc. 2009	Rencontres de Statistique Mathématique 9 au CIRM à Luminy.
Déc. 2010	Journées pour les 50 ans du LPMA à Paris.
Mai 2011	Journées Etats de la recherche : Théorie de l'Apprentissage, à Paris.

3.5 Participation régulière à des séminaires

Depuis 2007, j'ai fréquenté plusieurs séminaires et groupes de travail :

- Groupe de Travail Apprentissage, puis Séminaire SMILE (Statistical Machine Learning in Paris).
- Séminaire du LSTA et Groupe de Travail de Statistique du LPMA, puis Groupe de Travail de Statistique de Jussieu.
- Séminaire Parisien de Statistique.
- Groupe de Travail de Statistique et Colloquium du MAP5.

4 Activités d'enseignement

Au cours du monitorat, j'ai effectué 64h de travaux dirigés par an (service globalisé sur 3 ans, 2008–2011) à l'UFR de Mathématiques de l'Université Pierre et Marie Curie. En 2011–2012, je dispense en tant qu'ATER 192h d'enseignements, au sein du département STID (Statistique et Informatique Décisionnelle) de l'IUT de l'Université Paris Descartes.

Durant ces années de monitorat et d'ATER, j'ai eu l'occasion d'enseigner aussi bien de l'Algèbre linéaire et de l'Analyse pour le niveau L1, que des Probabilités et Statistique pour le DUT et le niveau L3. Les étudiants de L1 appartenaient soit au parcours MIME (Mathématiques-Informatique-Mécanique-Electronique), soit au parcours PCME (Physique-Chimie-Mécanique-Electronique). Mes enseignements à l'IUT concernent la première année de DUT ainsi que l'Année Spéciale, qui permet à des personnes justifiant au moins d'un niveau L1 d'obtenir le DUT en un an. En outre, j'aurai abordé différentes facettes de l'enseignement, ayant en charge des travaux dirigés, mais aussi des travaux pratiques sur ordinateur, un cours et l'encadrement de projets.

La plupart de ces enseignements m'ont amenée à rédiger des feuilles de travaux dirigés ou travaux pratiques, des sujets de devoirs et des corrigés.

Par ailleurs, j'ai participé à un jury pour le niveau L1 lors du monitorat. Cette année, j'assiste aux pré-soutenances des projets tutorés des étudiants de première année que j'encadre. Plus généralement, je prends part aux réunions du département STID.

Lors de la journée portes ouvertes de l'IUT, j'ai contribué à promouvoir les filières proposées. J'ai également eu l'occasion de présenter le métier d'enseignant-chercheur à des élèves de lycée et classes préparatoires en animant une journée portes ouvertes au lycée Jean Moulin de Forbach (Moselle).

4.1 ATER au département STID de l'IUT Paris Descartes (2011–2012)

- Travaux pratiques de Statistique Descriptive, DUT année 1 (60h).

  → *Utilisation du logiciel SPSS pour la statistique descriptive univariée et bivariée.*

- Travaux dirigés de Probabilités, DUT année 1 (63h).

  → *Dénombrement, probabilités conditionnelles, indépendance, variables aléatoires discrètes et continues, lois de probabilité usuelles.*

- Cours de Modèle linéaire et Analyse de la variance, DUT Année Spéciale (40h).

  → *Régression linéaire simple et multiple, et analyse de la variance à 1 et 2 facteurs ; utilisation du logiciel R pour la partie TP.*

- Projets tutorés, DUT année 1 (20h).

4.2 Monitorat à l'Université Pierre et Marie Curie (2008–2011)

2010–2011 :

- Travaux dirigés de Probabilités, niveau L3 (36h).

  → *Dénombrement, espaces de probabilité, variables et vecteurs aléatoires, indépendance, convergence de suites de variables aléatoires, loi des grands nombres, théorème de la limite centrale.*

2009–2010 :

- Travaux dirigés de Fonctions, niveau L1 (36h).

  → *Fonctions usuelles, fonctions réciproques, étude locale et globale des fonctions, fonctions de plusieurs variables, équations différentielles.*

- Travaux dirigés d'Analyse de données et régression, niveau L3 (36h).

  → *Lois normales multivariées, régression linéaire, analyse en composantes principales, intervalles de confiance et tests statistiques.*

2008–2009 :

- Travaux dirigés de Calcul matriciel, niveau L1 (42h).

  → *Systèmes linéaires, matrices inversibles, familles libres, génératrices, bases de $\mathbb{R}^n$, sous-espace vectoriel, produit scalaire, produit vectoriel, déterminant, diagonalisation.*

- Travaux dirigés de Fonctions, niveau L1 (42h).

Durant ces quatre années, j'ai enseigné devant un public assez varié. En L1, j'ai par exemple eu l'occasion de donner des travaux dirigés de Mathématiques générales en semestre décalé, ce qui entraînait qu'une bonne partie des étudiants avait rejoint cette formation après un premier semestre en classes préparatoires ou en faculté de médecine. Ces étudiants étaient plutôt motivés, tout comme ceux de l'Année Spéciale à l'IUT, qui proviennent d'horizons assez différents et n'ont pas tous le même bagage mathématique : L1 de biologie, première année de classe préparatoire ou de médecine, voire un niveau M1. D'autre part, mes groupes de L3 comprenaient deux ou trois personnes plus âgées, avec des attentes diverses.

Les travaux dirigés de Probabilités étaient intéressants, que ce soit en DUT ou en L3. J'ai aussi pu mesurer les difficultés que peuvent éprouver les étudiants devant le formalisme des Probabilités. Les travaux pratiques de Statistique Descriptive sous SPSS donnés à l'IUT ont constitué ma première expérience de travaux pratiques sur ordinateur. Je me suis rendu compte que, devant l'ordinateur, les étudiants perdent parfois de vue les notions qu'ils ont apprises en cours et ai donc tâché de les convaincre que l'utilisation d'un logiciel ne dispense pas du raisonnement mathématique. J'ai apprécié de retrouver quelques étudiants de première année pour les projets tutorés, dans des circonstances un peu différentes de celles d'une séance de travaux dirigés ou de travaux pratiques. Encadrer ces projets me permet de voir comment les étudiants mettent en application les compétences acquises en travaux pratiques dans ce contexte qui leur confère davantage d'autonomie, mais aussi d'aborder de manière simple quelques notions qu'ils étudieront plus en détail en cours l'année suivante. Enfin, j'ai trouvé le cours de Modèle linéaire et d'Analyse de la variance très plaisant, autant par son contenu que par le fait d'avoir toute latitude quant à l'organisation de mes séances et la répartition entre cours, travaux dirigés et travaux pratiques sous R.

5 Autres activités

Activités administratives et responsabilités collectives

- Relecture d'articles pour les revues *IEEE Transactions on Information Theory*, *Electronic Journal of Statistics* et *Annales de l'Institut Henri Poincaré*.
  Relecture de résumé pour la conférence Compstat 2010.
- Co-responsable du Groupe de Travail des doctorants du Laboratoire de Statistique Théorique et Appliquée (Université Pierre et Marie Curie) en 2009–2011.
- Représentante des étudiants au conseil du Département de Mathématiques et Applications de l'Ecole Normale Supérieure en 2005–2006.

6 Résumé des travaux de recherche

Au cours de ma thèse, réalisée sous la direction de Gérard Biau à l'Université Pierre et Marie Curie, j'ai étudié des problèmes d'apprentissage non supervisé. Une partie des travaux effectués concerne la quantification et le clustering, et l'autre est consacrée aux courbes principales. Actuellement, je continue à m'intéresser aux courbes principales et j'étudie également les propriétés d'une technique d'agrégation d'estimateurs de la régression.

Dans cette partie et la suivante, les références du type [1] correspondent à mes publications listées page 5. Toutes les autres références sont regroupées dans la bibliographie générale page 22.

6.1 Quantification et clustering avec des divergences de Bregman

Dans le premier chapitre de ma thèse, je me suis intéressée à la **quantification** (Gersho et Gray (1992), Graf et Luschgy (2000), Linder (2002)) et à la question liée du **clustering** (Duda, Hart et Stork (2000)) en utilisant comme notion de distance des divergences de Bregman (1967). En dimension finie, la **divergence de Bregman** associée à une fonction ϕ strictement convexe et différentiable est donnée par

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

où $\langle\cdot,\cdot\rangle$ désigne le produit scalaire canonique de l'espace euclidien $(\mathbb{R}^d, \|\cdot\|)$, et $\nabla\phi(y)$ le gradient de ϕ au point y. La distance euclidienne standard au carré est par exemple obtenue pour $\phi(x) = \|x\|^2$. Cette définition se généralise à un espace de Banach en écrivant

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

où $D_y\phi(h)$ est la différentielle de ϕ au point y appliquée à h. De nombreuses mesures de dissimilarité fréquemment utilisées en statistique et en théorie de l'information sont des cas particuliers de divergences de Bregman, d'où l'intérêt de considérer cette classe. En outre, certaines de ces divergences s'appliquent à des fonctions ou encore à des mesures de probabilité, ce qui en fait des outils appropriés pour classer des observations de grande dimension ou de nature complexe. Cette caractéristique est très appréciable compte tenu de l'afflux croissant de telles données dans de nombreux domaines.

Mon travail, qui a abouti à une publication dans la revue *Journal of Multivariate Analysis* [1], a consisté à généraliser au cas des divergences de Bregman des résultats d'**existence** d'un quantificateur optimal et de **convergence**. Plus précisément, soit X une variable aléatoire de loi μ à valeurs dans un espace de Banach $\mathcal{X}$, $X_1, \ldots, X_n$ un échantillon de X et d_ϕ une divergence de Bregman. Un quantificateur q envoyant tout $x \in \mathcal{X}$ sur l'un des k éléments $c_1, \ldots, c_k$ de $\mathcal{X}$ est caractérisé par cet ensemble de centres $c_1, \ldots, c_k$ et par la partition de $\mathcal{X}$ en k cellules $S_1, \ldots, S_k$ induite par la relation : $x \in S_j$ si, et seulement si, $q(x) = c_j$. L'erreur résultant de la substitution de $q(X)$ à X est mesurée par la distorsion

$$W(q) = \mathbb{E}[d_\phi(X, q(X))],$$

dont l'équivalent pour la mesure empirique est donné par

$$W_n(q) = \frac{1}{n}\sum_{i=1}^{n} d_\phi(X_i, q(X_i)).$$

Comme dans le cas de la quantification et du clustering k-means avec la distance euclidienne au carré, pour un ensemble de centres donné, la meilleure partition au sens de la distorsion est la partition dite de Voronoi, définie en affectant un élément x à S_j si, et seulement si, x est plus proche de c_j que de tout autre c_ℓ. Cette propriété a pour conséquence qu'il suffit de considérer les quantificateurs associés à la partition de Voronoi, appelés quantificateurs des plus proches voisins. Un tel quantificateur est

donc décrit par l'ensemble de ses centres. En fonction de $\mathbf{c} = (c_1, \ldots, c_k)$, distorsion et distorsion empirique se récrivent

$$W(\mathbf{c}) = \mathbb{E}\Big[\min_{j=1,\ldots,k} d_\phi(X, c_j)\Big], \qquad W_n(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\min_{j=1,\ldots,k} d_\phi(X_i, c_j)\Big].$$

J'ai démontré que, sous des conditions appropriées, il existe un $\mathbf{c}^\star$ optimal au sens de la distorsion $W(\mathbf{c})$, c'est-à-dire $W(\mathbf{c}^\star) = \inf_{\mathbf{c}} W(\mathbf{c})$. Ceci est vrai en particulier pour la mesure empirique μ_n. Une question naturelle consiste alors à se demander si un ensemble de centres empiriques optimaux $\mathbf{c}_n^\star$ basé sur un échantillon $X_1, \ldots, X_n$ constitue, pour n suffisamment grand, une bonne approximation de $\mathbf{c}^\star$. C'est pourquoi je me suis intéressée à la convergence de $W(\mathbf{c}_n^\star)$ vers le minimum de distorsion $W^\star$. En supposant que X reste presque sûrement dans une boule de rayon R et sous certaines conditions, le résultat principal prend ainsi la forme

$$\mathbb{E}\left[W(\mathbf{c}_n^\star)\right] - W^\star \leq \frac{k}{\sqrt{n}} C(\phi, R),$$

où $C(\phi, R) > 0$ est une constante qui ne dépend que du rayon de la boule et de la divergence de Bregman d_ϕ utilisée. Puisque les données peuvent être de dimension élevée voire infinie, l'intérêt de cette borne non-asymptotique est de ne pas faire intervenir la dimension.

J'ai également programmé un algorithme de clustering k-means basé sur les divergences de Bregman, que j'ai appliqué à plusieurs jeux de données simulés pour illustrer les différents comportements des divergences.

6.2 Clustering de courbes dans le cadre de l'industrie nucléaire

Toujours sur le thème du clustering, j'ai eu l'occasion de travailler sur une **application à l'ingénierie nucléaire**, en collaboration avec Benjamin Auder, alors doctorant au CEA de Cadarache. L'objectif de sa thèse était de construire par le biais d'une régression une approximation simple et rapide du code CATHARE (Code Avancé de THermohydraulique pour les Accidents des Réacteurs à Eau), utilisé dans l'industrie nucléaire pour prévoir les risques de rupture d'une cuve. Ce code de calcul donne en sortie les courbes d'évolution temporelle de certains paramètres physiques (pression, température, coefficient d'échanges thermiques). Pour augmenter la précision du modèle, il s'est avéré utile de repérer au préalable des groupes parmi ces courbes, puis d'effectuer la régression sur chacun des groupes séparément. Le clustering d'objets de dimension potentiellement infinie posant problème du point de vue des calculs numériques, l'objet de notre collaboration était d'étudier les propriétés d'une technique de **réduction de la dimension** pour cette étape de clustering de courbes. Ce travail a donné lieu à un article [4], accepté dans la revue *Journal of Statistical Computation and Simulation*.

Sur le plan théorique, nous supposons que les courbes à classer sont des éléments de $L^2([0,1])$ qui se décomposent sur une base hilbertienne avec des coefficients appartenant au sous-ensemble $\mathcal{S}$ de l'espace ℓ^2 des suites de carré sommable donné par

$$\mathcal{S} = \Big\{\mathbf{x} = (x_j)_{j\geq 1} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2\Big\},$$

où $R > 0$ et $(\varphi_j)_{j\geq 1}$ est une suite positive strictement croissante tendant vers l'infini. Il est important de noter que l'ensemble $\mathcal{S}$ est étroitement lié au choix de la base hilbertienne, même si cela n'apparaît pas explicitement dans la définition. Posons

$$W_\infty(\mathbf{c}) = \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|X - c_\ell\|^2\right], \qquad W_d(\mathbf{c}) = \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\right],$$

où Π_d désigne la projection sur $\mathbb{R}^d$. La version empirique de cette distorsion "d-dimensionnelle" est donnée par

$$W_{d,n}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{\ell=1,\dots,k} \|\Pi_d(X_i) - \Pi_d(c_\ell)\|^2.$$

La stratégie de réduction de la dimension que nous avons mise en œuvre consiste à effectuer le clustering dans l'espace de projection de dimension d. Si $\hat{\mathbf{c}}_{d,n}$ désigne un minimiseur de $W_{d,n}(\mathbf{c})$, un contrôle de l'écart entre $W_\infty(\hat{\mathbf{c}}_{d,n})$ et la distorsion minimale $W_\infty^\star$ permet d'évaluer la qualité de la procédure. Notre principal résultat exprime ainsi le fait que la perte en espérance pour le clustering dans l'espace de dimension infinie est bornée par la perte "fini-dimensionnelle" correspondante à laquelle s'ajoute un terme représentant le coût de la projection sur $\mathbb{R}^d$:

$$\mathbb{E}[W_\infty(\hat{\mathbf{c}}_{d,n})] - W_\infty^\star \leq \mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^\star + \frac{8R^2}{\varphi_d}.$$

Cette inégalité permet de discuter le choix de la dimension de projection d en fonction de n, puisque la quantité $\mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^\star$ peut être bornée par un terme de l'ordre de $1/\sqrt{n}$. Par exemple, si $\mathcal{S}$ est un ellipsoïde de Sobolev, avec

$$\varphi_j = \begin{cases} j^{2\beta} & \text{si } j \text{ pair} \\ (j-1)^{2\beta} & \text{si } j \text{ impair,} \end{cases}$$

la dimension d doit être de l'ordre de $n^{1/4\beta}$ pour conserver une vitesse en $1/\sqrt{n}$.

Comme mentionné plus haut, la forme de l'ensemble $\mathcal{S}$ dépend de la base hilbertienne considérée. Il est donc essentiel de choisir une base appropriée. D'un point de vue pratique, notre approche est basée sur un algorithme permettant de **construire une base** à l'aide de paquets d'ondelettes suivant la méthode de Coifman et Wickerhauser (1992), en incluant une comparaison avec les bases de Haar, de Fourier et de Karhunen-Loève.

6.3 Courbes principales et sélection de paramètres

Parallèlement à ce travail appliqué à l'industrie nucléaire, j'ai commencé à m'intéresser aux **courbes principales**. Ces courbes passant "au milieu" d'une loi de probabilité ou d'un nuage de points peuvent être vues comme une généralisation non linéaire de la première composante principale. Selon la définition originelle de Hastie et Stuetzle (1989), une courbe principale pour un vecteur aléatoire $\mathbf{X}$ de $\mathbb{R}^d$ est une courbe paramétrée $\mathbf{f} = (f_1, \dots, f_d)$ vérifiant la propriété d'auto-consistance, c'est-à-dire

$$\mathbf{f}(t) = \mathbb{E}[\mathbf{X} | t_{\mathbf{f}}(\mathbf{X}) = t],$$

où l'indice de projection $t_{\mathbf{f}}$ est défini par

$$t_{\mathbf{f}}(\mathbf{x}) = \sup\{t, \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\|\}.$$

La propriété d'auto-consistance s'interprète en disant que chaque point de la courbe est la moyenne des observations qui se projettent sur la courbe autour de ce point. D'autres définitions ont été proposées ensuite. Le premier chapitre de la seconde partie de la thèse présente ces diverses définitions et quelques applications des courbes principales. Dans la définition de Kégl, Krzyżak, Linder et Zeger (2000), plus facile à manipuler, une courbe principale de longueur L pour $\mathbf{X}$ est une courbe paramétrée minimisant le critère

$$\Delta(\mathbf{f}) = \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] = \mathbb{E}\left[\inf_t \|\mathbf{X} - \mathbf{f}(t)\|^2\right]$$

parmi toutes les courbes de longueur au plus L, tandis que Sandilya et Kulkarni (2002) remplacent la contrainte de longueur par une contrainte sur la courbure. Afin de déterminer une courbe principale

à partir d'un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbf{X}$, on cherche à minimiser le critère empirique

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \Delta(\mathbf{f}, \mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^{n} \inf_t \|\mathbf{X}_i - \mathbf{f}(t)\|^2.$$

Adoptant cette définition de courbe principale sous forme de problème des moindres carrés, je me suis penchée sur la sélection de certains paramètres de la courbe, tels que la longueur ou la courbure. En effet, le but est d'obtenir une courbe rendant compte correctement de la forme des données sans pour autant interpoler. J'ai choisi de considérer cette question sous l'angle de la **sélection de modèle par pénalisation** (Birgé et Massart (1997), Barron, Birgé et Massart (1999)).

Contexte gaussien

En premier lieu, j'ai supposé que l'on observe des points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbb{R}^d$ tels que

$$\mathbf{X}_i = \mathbf{x}_i^\star + \sigma \boldsymbol{\xi}_i, \quad i = 1, \ldots, n,$$

où les $\mathbf{x}_i^\star$ sont inconnus, les $\boldsymbol{\xi}_i$ sont des vecteurs gaussiens standards indépendants de $\mathbb{R}^d$ et $\sigma > 0$, et que l'on cherche une courbe principale correspondant à ces observations. Un cadre similaire est utilisé par Caillerie et Michel (2011), qui étudient des questions de sélection de modèle pour des graphes appelés complexes simpliciaux.

Considérant des courbes à extrémités fixées F et G, et ayant introduit une collection dénombrable de modèles $\{\mathcal{F}_\ell\}_{\ell \in \mathcal{L}}$ de courbes de longueur ℓ, je me suis donné comme but de choisir la longueur adéquate. Pour chaque longueur ℓ, la minimisation du risque empirique $\Delta_n(\mathbf{f})$ mène à une certaine courbe $\hat{\mathbf{f}}_\ell$. L'idée consiste alors à sélectionner la longueur $\hat{\ell}$ en minimisant un critère pénalisé. En notant $\tilde{\mathbf{x}}_i = \hat{\mathbf{x}}_{i\hat{\ell}}$ les projections des observations sur la courbe résultante $\hat{\mathbf{f}}_{\hat{\ell}}$ et $\|\cdot\|$ la norme euclidienne normalisée de $\mathbb{R}^d$, la qualité de l'estimation peut être mesurée au moyen de la perte

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\tilde{\mathbf{x}}_i - \mathbf{x}_i^\star\|^2.$$

Dans ce contexte gaussien, le résultat principal s'énonce ainsi : étant donné une famille de poids $\{w_\ell\}_{\ell \in \mathcal{L}}$ vérifiant $\sum_{\ell \in \mathcal{L}} e^{-w_\ell} = \Sigma < +\infty$, si le niveau de bruit σ n'est pas trop grand, il existe des constantes c_1, c_2 telles que pour tout $\eta > 1$, si

$$\mathrm{pen}(\ell) \geq \eta\sigma^2 \left[c_1 \ln\left(\frac{\ell^{1/d} \lambda^{1-1/d}}{\sigma} \right) + c_2 + \frac{4w_\ell}{nd} \right],$$

où $\lambda = \sqrt{\ell^2 - FG^2}$, alors, presque sûrement, il existe un minimiseur $\hat{\ell}$ du critère pénalisé

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \hat{\mathbf{x}}_{i\ell}\|^2 + \mathrm{pen}(\ell).$$

En outre,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\tilde{\mathbf{x}}_i - \mathbf{x}_i^\star\|^2 \leq c(\eta)\Big[\inf_{\ell \in \mathcal{L}} \{ d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) + \mathrm{pen}(\ell) \} + \frac{\sigma^2}{nd}(\Sigma + 1) \Big], \tag{6.1}$$

où $\mathcal{C}_\ell \subset \mathbb{R}^{nd}$ dépend de la classe $\mathcal{F}_\ell$ et $d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) = \inf_{\overrightarrow{\mathbf{y}} \in \mathcal{C}_\ell} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i^\star\|^2$, avec $\overrightarrow{\mathbf{y}}$ le vecteur de $\mathbb{R}^{nd}$ constitué de tous les $\mathbf{y}_i$.

Cette approche fait l'objet d'un article en cours de rédaction, qui sera soumis pour publication prochainement.

Contexte borné

Dans la suite de ce chapitre de thèse a été développé un second point de vue, qui a donné lieu à un article écrit en collaboration avec mon directeur de thèse Gérard Biau, publié dans la revue *IEEE Transactions on Information Theory* [3]. Nous cherchons à établir un résultat similaire, mais avec une inégalité portant cette fois-ci non plus sur les estimateurs des points échantillonnés sur la courbe, mais sur l'estimateur de la courbe principale elle-même. Dans cette perspective, nous sommes amenés à changer de cadre de travail et à supposer $\mathbf{X}$ presque sûrement bornée. Considérant alors des modèles de lignes polygonales, nous examinons successivement le cas des courbes principales de longueur bornée de Kégl, Krzyżak, Linder et Zeger (2000) et de courbure intégrale bornée de Sandilya et Kulkarni (2002).

Dans le premier cas, soit

$$\mathbf{f}^\star \in \underset{\mathbf{f}, \mathscr{L}(\mathbf{f}) \leq L}{\arg\min}\, \Delta(\mathbf{f}),$$

où $\mathscr{L}(\mathbf{f})$ désigne la longueur de la courbe $\mathbf{f}$. Pour $k \geq 1$ et $\ell \in \mathcal{L} \subset]0, L]$, le modèle $\mathcal{F}_{k,\ell}$ est défini comme la classe des lignes polygonales à k segments et de longueur au plus ℓ. A chaque modèle $\mathcal{F}_{k,\ell}$ correspond une courbe $\hat{\mathbf{f}}_{k,\ell}$ minimisant le critère empirique $\Delta_n(\mathbf{f})$ sur tous les éléments de ce modèle et notre objectif est de choisir la meilleure courbe principale parmi les estimateurs de la collection $\{\hat{\mathbf{f}}_{k,\ell}\}_{k\geq 1, \ell \in \mathcal{L}}$. Pour ce faire, nous construisons une fonction de pénalité $\mathrm{pen}(k,\ell)$, et les paramètres $(\hat{k}, \hat{\ell})$ retenus sont ceux qui minimisent le critère $\Delta_n(\mathbf{f}_{k,\ell})$ pénalisé par $\mathrm{pen}(k,\ell)$. La qualité de l'estimateur $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$ sélectionné est évaluée en contrôlant la perte $\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})]$. Si $\{w_{k,\ell}\}_{k\geq 1, \ell \in \mathcal{L}}$ désigne une famille de poids positifs telle que $\sum_{k\geq 1, \ell \in \mathcal{L}} e^{-w_{k,\ell}} = \Sigma$, pour une pénalité de la forme

$$\mathrm{pen}(k,\ell) \geq \frac{1}{\sqrt{n}} \left[c_1 \sqrt{k} + c_2 \max\left(\frac{\ell}{\sqrt{k}}, \sqrt{k \ln \frac{\ell}{k}} \right) + c_0 \right] + \delta^2 \sqrt{\frac{w_{k,\ell}}{2n}},$$

où les constantes c_i ne dépendent que de d et δ, la courbe $\tilde{\mathbf{f}}$ obtenue en minimisant le critère $\Delta_n(\mathbf{f})$ pénalisé vérifie l'inégalité

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1, \ell \in \mathcal{L}} \left(\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) + \mathrm{pen}(k,\ell) \right) + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) = \inf_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$. Observons que la pénalité proposée tend vers 0 à la vitesse $1/\sqrt{n}$.

Grâce à un résultat géométrique reliant la courbure intégrale d'une courbe paramétrée à sa longueur (Alexandrov et Reshetnyak (1989)), il est possible de mettre en œuvre une approche similaire dans le contexte des courbes principales de Sandilya et Kulkarni (2002), où la contrainte est mise sur la courbure. Les modèles $\mathcal{F}_{k,\kappa}$ sont ici indexés par le nombre de segments $k \geq 1$ et la courbure intégrale $\kappa \in \mathcal{K}$. De même que la longueur d'une ligne polygonale est la somme des segments qui la composent, la courbure intégrale est la somme des angles aux sommets. Les courbes $\mathbf{f}^\star$ et $\tilde{\mathbf{f}}$, ainsi que la famille $\{w_{k,\kappa}\}_{k\geq 1, \kappa \in \mathcal{K}}$, sont définies dans ce cadre par analogie avec le cas précédent. Alors, si

$$\mathrm{pen}(k,\kappa) \geq \frac{\delta^2}{\sqrt{n}} \left[c_1 \sqrt{k} + + c_2 \max\left(\frac{\zeta(\kappa)}{\sqrt{k}}, \sqrt{k \ln \frac{\zeta(\kappa)}{k}} \right) + c_0 + \sqrt{\frac{w_{k,\kappa}}{2}} \right],$$

où ζ est une fonction croissante de κ et les c_i ne dépendent que de d, nous obtenons

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1, \kappa \in \mathcal{K}} \left(\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\kappa}) + \mathrm{pen}(k,\kappa) \right) + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\kappa}) = \inf_{\mathbf{f} \in \mathcal{F}_{k,\kappa}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

Finalement, un algorithme permettant de calculer une courbe principale d'un point de vue pratique a été implémenté. Pour évaluer les constantes intervenant dans les fonctions de pénalité, nous avons choisi d'employer l'**heuristique de pente**, méthode de calibration de pénalités introduite par

Birgé et Massart (2007). L'algorithme, qui estime à la fois $\hat{k}$ et $\hat{\ell}$, est basé sur une version bivariée de cette heuristique. Après avoir appliqué la méthode à des données simulées, nous étudions des applications à la **reconnaissance de caractères**, puis à la reconstruction de **limites de plaques lithosphériques** à l'aide de données d'impacts de tremblements de terre. Comme certaines failles sont peu repérables, par exemple parce qu'elles se trouvent dans une zone densément recouverte de végétation, mais qu'il convient de les surveiller de près dans le cadre de la prévention des risques sismiques, les courbes principales peuvent s'avérer très utiles en permettant de les localiser.

Poursuite des travaux sur les courbes principales durant mon année d'ATER

Courbes principales d'ordre supérieur

Si l'on entend par "courbe principale" une généralisation de la notion de première composante principale, alors la question des courbes principales d'ordre supérieur se pose naturellement, en particulier du point de vue des applications. Dans une analyse en composantes principales, on extrait les composantes principales successives, chacune expliquant une certaine proportion décroissante de la variance. De même, une loi de probabilité ou un nuage de points peut présenter **plusieurs courbes principales**. Supposons que l'on dispose d'une courbe paramétrée $\mathbf{f}$ minimisant sur une certaine classe $\mathcal{M}$ le critère $\Delta(\mathbf{f})$ ou son équivalent empirique $\Delta_n(\mathbf{f})$. On cherche, par analogie avec l'analyse en composantes principales, à définir une courbe principale $\mathbf{g}$ d'ordre 2, comme une courbe paramétrée minimisant le critère sous une contrainte d'orthogonalité entre $\mathbf{f}$ et $\mathbf{g}$. Si l'on considère des courbes paramétrées sur $[0,1]$ étendues sur une base orthonormale $(\varphi_k)_k$ sous la forme

$$\mathbf{f}(t) = \sum_{k=0}^{+\infty} \mathbb{a}_k \varphi_k(t) = \left(\sum_{k=0}^{+\infty} a_k^1 \varphi_k(t), \ldots, \sum_{k=0}^{+\infty} a_k^d \varphi_k(t) \right),$$

la norme sur ces courbes définie par $\|\mathbf{f}\|_{\mathcal{C}}^2 = \int_0^1 \|\mathbf{f}(t)\|^2 \mathrm{d}t$ se récrit $\|\mathbf{f}\|_{\mathcal{C}}^2 = \sum_{j=1}^d \sum_{k=0}^{+\infty} (a_k^j)^2$ et la condition d'orthogonalité peut s'exprimer comme la nullité du produit scalaire correspondant. Avec Avner Bar-Hen, professeur à l'Université Paris Descartes, nous nous intéressons aux conséquences de cette définition, aussi bien sur un plan théorique que pratique.

Pénalisation par la somme des angles

D'autre part, nous commençons à explorer un cadre de **sélection de paramètres** pour les courbes principales un peu différent de celui de ma thèse. Considérant toujours des lignes polygonales, nous supposons que les extrémités des courbes sont fixées. Au lieu de chercher à estimer directement le nombre de segments, nous le fixons initialement aussi, potentiellement grand, et étudions les propriétés de la courbe principale obtenue en minimisant le critère empirique pénalisé par la somme des angles. En effet, cette méthode pourrait s'avérer intéressante pour construire une courbe principale si elle permet de contrôler correctement les angles, en mettant certains d'entre eux à 0, à la manière du Lasso (Least absolute shrinkage and selection operator, Tibshirani (1996)).

Application à l'étude du climat

Nous avons aussi pour objectif, partant du commentaire de Smith (2010) d'un article de Li, Nychka et Ammann (2010), une application à la **reconstruction de températures** à l'aide de données de **dendrochronologie**. La dendrochronologie est l'étude des cernes de croissance des arbres, dont la largeur est influencée entre autres par la température. Or, on possède des relevés de température à la surface de la terre à partir de 1850 environ. Dans le cadre de l'analyse de l'évolution du climat, il est essentiel de connaître les températures des siècles antérieurs. Nous disposons de données correspondant à 70 arbres, de sorte que le nombre de variables explicatives approche le nombre d'années pour lesquelles la température est connue. De plus, ces données relatives à différents arbres sont corrélées,

et les années sont également corrélées entre elles. Les courbes principales pourraient constituer une méthode non-paramétrique de réduction de la dimension appropriée dans ce contexte. La technique générale consiste à approcher l'espace des prédicteurs par une courbe passant au milieu de ces points en dimension 70, puis à effectuer une régression des températures sur les indices de projection sur la courbe. Combiner une analyse en composantes principales préalable et une courbe principale peut être utile également. Par ailleurs, si une seule courbe ne contient pas suffisamment d'information, la construction proposée plus haut pourra être exploitée pour obtenir une courbe principale d'ordre supérieur.

6.4 Nombre de groupes en clustering

Dans mes travaux sur le clustering avec des divergences de Bregman et sur la classification de courbes dans l'industrie nucléaire, le nombre de groupes k est supposé connu. Cependant, le **choix du nombre de classes** à spécifier constitue un problème majeur en clustering. Pour obtenir un résultat pertinent, en évitant autant que possible de couper artificiellement une classe ou de fusionner plusieurs groupes, il est en effet indispensable de déterminer correctement k. Chercher à sélectionner automatiquement ce paramètre m'a donc intéressée tout naturellement.

Or, travaillant sur la quantification et le clustering d'une part, et les courbes principales d'autre part, j'ai peu à peu réalisé à quel point le **lien entre courbes principales et quantification** dans $\mathbb{R}^d$ euclidien est étroit, un quantificateur optimal et une courbe principale pouvant être définis comme des éléments minimisant des critères très similaires. En particulier, cette relation m'a permis d'employer la même approche que pour les paramètres des courbes principales, pour obtenir un résultat sur le choix du nombre de groupes en clustering. Il en résulte une note, publiée dans *Statistics and Probability Letters* [2]. Dans $(\mathbb{R}^d, \|\cdot\|)$ euclidien, et en supposant la variable aléatoire $\mathbf{X}$ sous-jacente presque sûrement bornée, je démontre que pour une fonction de pénalité $\mathrm{pen}(k)$ de l'ordre de $\sqrt{k/n}$, l'ensemble de centres $\tilde{\mathbf{c}}$ obtenu par minimisation de la distorsion empirique pénalisée vérifie

$$\mathbb{E}[W(\tilde{\mathbf{c}})] \leq \inf_{1\leq k\leq n} (W_k^\star + \mathrm{pen}(k)) + r_n,$$

où $W_k^\star$ est la distorsion optimale pour k centres et r_n tend vers 0 lorsque n tend vers l'infini. De plus, $\mathbb{E}[W(\tilde{\mathbf{c}})]$ décroît vers 0 à la vitesse $n^{-2/(d+4)}$.

La mise en œuvre pratique de la méthode repose à nouveau sur l'heuristique de pente. Après avoir comparé les résultats sur des simulations avec ceux de la *Gap Statistic* de Tibshirani, Walther et Hastie (2001), j'ai appliqué le procédé à plusieurs types de données réelles. A l'aide d'un certain nombre d'informations concernant différentes **espèces animales**, je retrouve ainsi le nombre de classes présentes : mammifère, poisson, invertébré, oiseau, insecte. De même, je recherche le nombre de classes d'âge au sein d'un groupe d'**ormeaux**, dont les caractéristiques sont données dans Nash, Sellers, Talbot, Cawthorn et Ford (1994). Enfin, j'ai également utilisé des données provenant d'une étude sur la **dyslexie** effectuée dans le Laboratoire de Sciences Cognitives et Psycholinguistique, situé dans le Département d'Etudes Cognitives de l'Ecole Normale Supérieure. Pour mieux comprendre ce trouble, différentes hypothèses doivent être testées en comparant les performances d'adultes dyslexiques avec celles d'adultes témoins, dont j'ai moi-même fait partie. Il est intéressant de voir que le nombre de groupes trouvé est 4, correspondant aux personnes dyslexiques, au groupe témoin, plus quelques faux positifs et faux négatifs.

6.5 Une méthode d'agrégation d'estimateurs de la régression

Cette partie, indépendante des précédentes, concerne un travail en cours, en collaboration avec Gérard Biau, Benjamin Guedj (Université Pierre et Marie Curie) et James Malley (National Institutes of Health, Bethesda, Etats-Unis), avec pour finalité des applications en **génomique** et dans le **domaine biomédical**. Il s'agit de l'étude d'une méthode d'**agrégation d'estimateurs de la régression**.

L'un des objectifs est d'étudier le lien entre le génome et certaines maladies, telles que le trouble du déficit de l'attention ou la schizophrénie infantile. Dans ce contexte, la réponse est binaire, et la liste des variables explicatives, arbitraire, est constituée de quelques centaines de milliers de SNP (*Single Nucleotide Polymorphism*), variations de l'ADN correspondant à une différence sur une seule paire de bases de nucléotides, ce qui constitue la forme la plus abondante de variabilité dans le génome humain. Bien qu'il s'agisse d'un problème de classification, l'équipe est amenée à utiliser des techniques de régression, étant préoccupée avant tout par l'estimation des probabilités d'être atteint d'une maladie. Pour cela, plusieurs techniques d'estimation de la régression sont employées, comme par exemple les méthodes de forêts aléatoires et des k plus proches voisins, donnant lieu à différentes probabilités. L'idée est de combiner les estimateurs pour améliorer la performance de l'estimation de cette probabilité. Une autre application vise la détection des sites de début de transcription d'un génome entier, comme celui de la mouche du fruit *Drosophila Melanogaster*, qui est constitué d'environ 165 millions de bases. Dans ce cas, une probabilité est affectée à chaque base à l'aide de forêts aléatoires. Les résultats devraient ici aussi être améliorés en combinant plusieurs méthodes.

Les techniques d'agrégation, dont le développement est justement motivé par l'existence d'un grand nombre de méthodes d'estimation, consistent en général à construire une combinaison linéaire ou une combinaison convexe de plusieurs estimateurs (voir par exemple Bunea, Tsybakov et Wegkamp (2007)). Nous proposons une approche un peu différente, basée sur une idée de Mojirsheibani (1999) en classification. Soit $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ un échantillon d'un couple de variables aléatoires $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$. On le divise en deux sous-échantillons indépendants $\mathcal{D}_\ell = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_\ell, Y_\ell)\}$ et $\mathcal{D}_k = \{(\mathbf{X}_{\ell+1}, Y_{\ell+1}), \ldots, (\mathbf{X}_n, Y_n)\}$, où $k = n - \ell \geq 1$. Supposons donnés M estimateurs de la régression $r_{k,1}, \ldots, r_{k,M}$ construits à l'aide de $\mathcal{D}_k$. Soit $\varepsilon > 0$ et soit $\mathbf{x}$ une nouvelle donnée. La stratégie proposée débute par la recherche des observations $(\mathbf{X}_i)_{i \in I}$, $I \subset \{1, \ldots, \ell\}$, pour lesquelles $|r_{k,m}(\mathbf{X}_i) - r_{k,m}(\mathbf{x})| < \varepsilon$ pour tout $m = 1, \ldots, M$, puis la prévision $\hat{y}$ pour $\mathbf{x}$ est obtenue en prenant la moyenne des $(Y_i)_{i \in I}$. Plus formellement, l'estimateur combiné ρ_n est défini par

$$\rho_n(r_{k,1}(\mathbf{x}), \ldots, r_{k,M}(\mathbf{x})) = \sum_{i=1}^{\ell} W_{n,i}(\mathbf{x}) Y_i,$$

où les poids $W_{n,i}$ sont donnés par

$$W_{n,i}(\mathbf{x}) := \frac{\mathbf{1}_{\bigcap_{m=1}^{M} \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_i)| \leq \varepsilon\}}}{\sum_{j=1}^{\ell} \mathbf{1}_{\bigcap_{m=1}^{M} \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_j)| \leq \varepsilon\}}}.$$

Nous étudions les **propriétés asymptotiques** de cet estimateur, avec un seuil ε_n dépendant de n. Si $\lim_{n\to\infty} \varepsilon_n = 0$ et $\lim_{n\to\infty} n\varepsilon_n^M = +\infty$, ρ_n vérifie, pour toute loi de $(\mathbf{X}, Y)$,

$$\lim_{n\to\infty} \mathbb{E}\left[(\rho_n(r_{k,1}(\mathbf{X}), \ldots, r_{k,M}(\mathbf{X})) - \mathbb{E}[Y | r_{k,1}(\mathbf{X}), \ldots, r_{k,M}(\mathbf{X})])^2\right] = 0.$$

De plus, ρ_n est asymptotiquement au moins aussi bon que n'importe lequel des estimateurs initiaux, et, si l'un de ces estimateurs individuels est universellement consistant, c'est-à-dire, quelle que soit la loi de $(\mathbf{X}, Y)$, $\lim_{k\to\infty} \mathbb{E}\left[(r_{k,m}(\mathbf{X}) - \mathbb{E}[Y|\mathbf{X}])^2\right] = 0$, alors ρ_n l'est également :

$$\lim_{n\to\infty} \mathbb{E}\left[(\rho_n(r_{k,1}(\mathbf{X}), \ldots, r_{k,M}(\mathbf{X})) - \mathbb{E}[Y|\mathbf{X}])^2\right] = 0.$$

7 Projet de recherche

Au cours de mes activités futures, j'aimerais continuer à explorer des problématiques liées à l'apprentissage statistique et à la statistique non-paramétrique, étudier des questions de sélection de modèle et de grande dimension, en associant théorie et applications.

Les deux premières sections ci-dessous décrivent des pistes de recherche en **apprentissage non supervisé**, se situant dans la continuité de ma thèse et se donnant pour but de répondre à des questions soulevées durant celle-ci. Ces parties reprennent ainsi les deux grands thèmes de la thèse, **quantification et clustering** d'une part, et **courbes principales** d'autre part. Bien que j'aie commencé à travailler indépendamment sur chacune de ces deux notions, j'ai ensuite réalisé qu'elles étaient étroitement liées, ce qui m'a paru une caractéristique très importante. Puisque ces deux domaines s'enrichissent mutuellement, une étude effectuée dans l'un des deux cadres pourra suggérer d'intéressantes idées conduisant à des résultats dans le second. La dernière section de cette partie concerne l'**apprentissage supervisé** et traite de quelques prolongements relatifs à la méthode d'**agrégation d'estimateurs** de la régression que j'étudie dans un travail en cours.

Ces axes de recherche seront amenés à évoluer en fonction de l'équipe qui m'accueillera et je suis prête à m'investir dans d'autres projets suivant ses souhaits et ses besoins.

7.1 Quantification et clustering

Clustering et divergences

Les divergences de Bregman forment une riche classe de mesures de distorsion, englobant comme cas particuliers des fonctions de proximité effectivement employées en statistique et en théorie de l'information. Elles s'appliquent à des objets variés, comme des fonctions ou des mesures de probabilité, ce qui peut être très utile devant l'afflux croissant de données complexes et de grande dimension dans différents domaines. Mon travail sur la quantification et le clustering avec ces divergences [1] se comprend dans une optique de généralisation ou d'unification. En particulier, il s'agissait d'étendre les propriétés établies pour une norme hilbertienne (Biau, Devroye et Lugosi (2008)).

Cependant, je ne me suis pas focalisée dans la thèse sur la question du **choix de la divergence de Bregman**. C'est un problème essentiel en pratique, qui mérite de l'attention dans deux directions. D'une part, l'existence d'une relation entre divergences de Bregman en dimension finie et lois de la famille exponentielle suggère d'explorer plus avant le cas de la dimension infinie afin de décrire clairement le lien avec les notions appropriées de famille exponentielle. D'autre part, dans le contexte du clustering, il convient de réfléchir au choix de la divergence en lui-même. Il serait intéressant de dégager quelques éléments pouvant guider ce choix alors que nous ne connaissons pas la loi sous-jacente des observations.

A propos du nombre de groupes

Dans la thèse et dans la note [2], je me suis penchée sur le **choix du nombre de groupes** dans ce contexte de quantification et de clustering, différent de celui des modèles de mélange, dans lequel on dispose d'informations supplémentaires. Les résultats obtenus pourraient être généralisés dans différentes directions. Tout d'abord, on peut se demander s'il est possible d'affaiblir les hypothèses, en supposant par exemple que la variable aléatoire sous-jacente, au lieu d'être presque sûrement bornée, satisfait seulement une **condition de moment**. Il s'agirait de s'affranchir de l'hypothèse de bornitude dans les majorations conduisant à la pénalité, en exploitant des outils de concentration récents. Des résultats comme celui de Bentkus (2008), qui démontre une inégalité de type Hoeffding pour des variables non bornées, pourraient constituer une piste prometteuse. Notons que la borne

non asymptotique de type lemme de Pierce, utilisée pour déduire de l'inégalité oracle la vitesse $\mathcal{O}(n^{-2/d+4})$, ne nécessite qu'une condition de moment (voir Luschgy et Pagès (2008)). En outre, il serait intéressant d'examiner le cas où les données n'appartiennent pas nécessairement à $\mathbb{R}^d$ ainsi que l'utilisation d'**autres distances** que la distance euclidienne. Par ailleurs, il devrait être possible d'établir des résultats similaires avec une forme de **pénalité dépendant des données**. En effet, le théorème de sélection de modèle employé dans [2] peut aussi s'appliquer dans un tel contexte, dès lors que la borne inférieure sur la pénalité est vérifiée avec grande probabilité (Massart (2007)).

Enfin, une autre direction de recherche qui me semble digne d'attention consiste à essayer de fournir des **garanties théoriques** pour des **heuristiques existantes** de choix de nombre de groupes. A titre d'exemple, l'article de Kim, Park et Park (2001) suggère de pénaliser la distorsion empirique par un terme de l'ordre de $k/d_{\min}$, où $d_{\min}$ représente la distance minimale entre les centres des classes. Cette quantité est effectivement croissante en le nombre de groupes k et aura tendance à exploser pour de grandes valeurs de k puisque $d_{\min}$ s'approche alors de 0.

Clustering de courbes et projections

L'article [4], fruit d'une collaboration avec Benjamin Auder, traite d'une technique de **clustering de courbes** reposant sur une méthode de projection dans le contexte de l'industrie nucléaire, et présente une discussion sur le lien entre le nombre d'observations et la dimension de projection. Pour aller plus loin, il serait utile à ce niveau de développer une méthode automatique de **choix de la dimension de projection**. Les techniques étudiées pour choisir le nombre de groupes pourront fournir un point de départ à la réflexion.

D'autre part, bien que la méthode, qui inclut la programmation du choix de la base, soit déjà relativement complexe d'un point de vue algorithmique, il me paraît aussi intéressant d'envisager de remplacer les projections simples, que nous avons utilisées, par des **projections aléatoires** basées sur le lemme de Johnson-Lindenstrauss (1984), en reprenant l'idée de Biau, Devroye et Lugosi (2008).

7.2 Courbes principales

Dans le cadre des courbes principales, je souhaiterais en premier lieu mener à son terme le travail initié avec Avner Bar-Hen, concernant les courbes principales d'ordre supérieur et le point de vue pénalisation par la somme des angles, ainsi que l'application à l'étude du climat. Je présente ci-dessous trois autres grandes directions de recherche envisageables. La première concerne des améliorations possibles des méthodes de sélection de paramètres considérées dans ma thèse, la seconde les définitions alternatives, divers algorithmes et applications des courbes principales, et la dernière a pour objet la méthode de la pente.

Quelques améliorations pour les méthodes de sélection de paramètres

Plusieurs extensions des techniques de sélection de paramètres pour les courbes principales développées dans la deuxième partie de la thèse peuvent être envisagées.

Le premier résultat, relatif à la sélection de la longueur d'une courbe principale dont les extrémités sont supposées fixées, est basé sur un **modèle gaussien**

$$\mathbf{X}_i = \mathbf{x}_i^\star + \sigma \boldsymbol{\xi}_i, \quad i = 1, \ldots, n,$$

où les $\mathbf{x}_i^\star$ sont inconnus. On remarque que dans ce cas, la pénalité obtenue ne tend pas vers 0 lorsque n tend vers $+\infty$. Il en résulte en particulier que le terme de droite dans l'inégalité de type oracle associée ne converge pas vers 0. Ce point est intrinsèquement lié à la géométrie du problème. Ainsi posé, ce dernier n'est pas simplifié par l'augmentation du nombre d'observations, car rien n'est spécifié sur la répartition des éléments $\mathbf{x}_i^\star$. En imaginant un autre modèle, dans lequel ces derniers sont **répartis selon une loi uniforme** sur la courbe (type de modèle considéré par Genovese, Perone-Pacifico,

Verdinelli et Wasserman (2010)), notre problème pourrait se replacer dans le cadre de la sélection de modèle en estimation de densité. Ce point de vue met en jeu des calculs d'entropie à crochets de classes de densités de mélanges gaussiens continus en dimension d. Même s'ils ne s'appliquent pas directement dans notre contexte, les résultats d'entropie de Genovese et Wasserman (2000) et Ghosal et van der Vaart (2001) fourniraient probablement des pistes utiles.

L'article [3] étudie le choix de paramètres de longueur, courbure et nombre de segments pour des **courbes principales polygonales** à extrémités non fixées dans le cas où $\mathbf{X}$ est presque sûrement bornée. L'inégalité permettant d'évaluer la qualité de la courbe sélectionnée s'applique cette fois-ci à la courbe elle-même et non aux projections des observations sur la courbe comme dans le cadre gaussien. De plus, la pénalité tend vers 0 à la vitesse $1/\sqrt{n}$. Néanmoins, il me semble que cette approche pourrait gagner en précision, notamment par le biais d'**arguments de concentration** permettant de contrôler le comportement local d'un processus empirique. Bien que la transposition au cadre des courbes principales risque d'être difficile en raison de la présence de la minimisation en t dans le critère

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \inf_t \|\mathbf{X}_i - \mathbf{f}(t)\|^2,$$

je voudrais m'inspirer du cas de la régression examiné par Massart et Nédélec (2006). Ce minimum en t, en plus du caractère multidimensionnel, est justement ce qui distingue l'estimation des courbes principales de la régression. Notons aussi que des **pénalités dépendant des données** pourraient certainement être considérées, comme mentionné pour le choix du nombre de groupes en clustering.

Enfin, puisque l'on sait calculer l'entropie métrique de tels objets, une piste supplémentaire consisterait à exploiter le fait que les courbes de $\mathbb{R}^d$ de longueur bornée correspondent aux **fonctions lipschitziennes** à valeurs dans $\mathbb{R}^d$.

D'une manière générale, comme continuation logique des résultats de sélection de paramètres de ma thèse, il serait appréciable d'obtenir des résultats d'**adaptation** et de faire un lien avec la **théorie de l'approximation**. Les améliorations possibles et les nouveaux points de vue évoqués m'intéressent tout particulièrement dans cette perspective.

Autres définitions, algorithmes et applications

Plusieurs définitions de courbe principale ont été introduites depuis la fin des années 1980, plus ou moins éloignées de la définition de Hastie et Stuetzle (1989), basée sur la propriété d'auto-consistance. Dans les travaux et pistes de recherche mentionnés jusqu'ici, j'ai utilisé la définition de Kégl, Krzyżak, Linder et Zeger (2000), qui peut être vue comme une version de la définition originelle exprimée sous forme de problème de minimisation de moindres carrés. Cependant, d'**autres définitions** pourraient également être exploitées. Par exemple, la définition d'Ozertem et Erdogmus (2011) est basée sur les **lignes de crête** des densités de probabilités. Il me paraît intéressant d'analyser rigoureusement les propriétés d'**existence** et de **convergence** de ces courbes principales.

En outre, il existe des **algorithmes** destinés à fournir une approximation de courbe principale reposant sur différentes heuristiques, incluant notamment l'utilisation de **pénalités locales**, pour lesquels il pourrait être utile d'établir des garanties théoriques.

D'une manière générale, je suis aussi très intéressée par les **applications** des courbes principales, qui se révèlent extrêmement variées, allant de la reconnaissance de caractères au domaine médical en passant par la géographie ou encore l'écologie.

Autour de la méthode de la pente

Enfin, je place ici un projet correspondant à une question qui s'est posée dans le cadre des courbes principales, mais me paraît importante en général. La méthode de sélection de paramètres obtenue pour les courbes principales a été illustrée dans la thèse et l'article [3] en s'inspirant de manière heuristique de la méthode de la pente, méthode de calibration des pénalités introduite par Birgé et

Massart (2007). Cependant, réfléchir aux **justifications théoriques** de cette technique dans de tels contextes, où l'on peut notamment avoir à choisir **plusieurs paramètres**, constitue une direction de recherche intéressante.

7.3 Sur l'estimateur combiné de la régression

Concernant l'estimateur combiné de la régression défini dans la Section 6.5, mes objectifs dans un futur proche sont d'une part la mise en œuvre pratique sur les **problèmes de biologie** avec l'équipe de James Malley, et d'autre part quelques extensions d'ordre plus théoriques. Il me paraît ainsi intéressant de chercher à obtenir une **vitesse de convergence**, ou encore de se demander si l'on peut s'affranchir de couper l'échantillon en deux, ce qui implique de manipuler des sommes de variables aléatoires qui ne sont plus indépendantes. En outre, le **choix du seuil** ε est une question importante.

A plus long terme, on pourrait aussi imaginer des définitions d'estimateur plus complexes, par exemple en conservant l'ensemble de tous les $\mathbf{X}_i$, $i = 1, \ldots, n$, et en estimant la réponse $\hat{y}$ pour une nouvelle observation $\mathbf{x}$ grâce à une certaine moyenne pondérée des Y_i. La difficulté réside ici dans la construction des poids. Ceux-ci devraient faire intervenir les distances $|r_{k,m}(\mathbf{X}_i) - r_{k,m}(\mathbf{x})|$, $m = 1, \ldots, M$, et dépendre en particulier des quantités $\varepsilon_i = \max_{m=1,\ldots,M} |r_{k,m}(\mathbf{X}_i) - r_{k,m}(\mathbf{x})|$. Notre estimateur initial serait alors un cas particulier de cette nouvelle définition, pour lequel tous les poids valent 0 ou 1.

8 Bibliographie générale

A.D. Alexandrov et Y.G. Reshetnyak : *General Theory of Irregular Curves*. Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht, 1989.

A. Barron, L. Birgé et P. Massart : Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

V. Bentkus : An extension of the Hoeffding inequality to unbounded random variables. *Lithuanian Mathematical Journal*, 48:137–157, 2008.

G. Biau, L. Devroye et G. Lugosi : On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.

L. Birgé et P. Massart : From model selection to adaptive estimation. *In* D. Pollard, E. Torgersen et G. Yang, éditeurs : *Festschrift for Lucien Le Cam : Research Papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

L. Birgé et P. Massart : Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

L.M. Bregman : The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

F. Bunea, A.B. Tsybakov et M.H. Wegkamp : Aggregation for Gaussian Regression. *The Annals of Statistics*, 35:1674–1697, 2007.

C. Caillerie et B. Michel : Model selection for simplicial approximation. *Foundations of Computational Mathematics*, 11:707–731, 2011.

R.R. Coifman et M.V. Wickerhauser : Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.

R.O. Duda, P.E. Hart et D.G. Stork : *Pattern Classification*. Wiley-Interscience, New York, 2000.

C.R. Genovese, M. Perone-Pacifico, I. Verdinelli et L. Wasserman : The geometry of nonparametric filament estimation. 2010. `http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.5536v2.pdf`.

C.R. Genovese et L. Wasserman : Rates of Convergence for the Gaussian Mixture Sieve. *The Annals of Statistics*, 28:1105–1127, 2000.

A. Gersho et R.M. Gray : *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, 1992.

S. Ghosal et A.W. van der Vaart : Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29:1233–1263, 2001.

S. Graf et H. Luschgy : *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, Heidelberg, 2000.

T. Hastie et W. Stuetzle : Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

W. Johnson et J. Lindenstrauss : Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

B. Kégl, A. Krzyżak, T. Linder et K. Zeger : Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:281–297, 2000.

D.J. Kim, Y.W. Park et D.J. Park : A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and System*, E84D:281–285, 2001.

B. Li, D.W. Nychka et C.M. Ammann : The Value of Multi-Proxy Reconstruction of Past Climate. *Journal of the American Statistical Association* 105:883–895, 2010.

T. Linder : Learning-theoretic methods in vector quantization. *In* L. Györfi, éditeur : *Principles of Nonparametric Learning*. Springer-Verlag, Wien, 2002.

H. Luschgy et G. Pagès : Functional quantization rate and mean regularity of processes with an application to Levy processes. *The Annals of Applied Probability*, 18:427–469, 2008.

P. Massart : *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

P. Massart et E. Nédélec : Risk bounds for statistical learning. *The Annals of Statistics*, 34:2326–2366, 2006.

M. Mojirsheibani : Combining Classifiers via Discretization. *Journal of the American Statistical Association*, 94:600–609, 1999.

W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn et W.B Ford : The population biology of Abalone (Haliotis species) in Tasmania : Blacklip Abalone (H. rubra) from the north coast and islands of Bass Strait. Rapport technique 48, Sea Fisheries Division, 1994.

U. Ozertem et D. Erdogmus : Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286, 2011.

S. Sandilya et S.R. Kulkarni : Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48:2789–2793, 2002.

R. Smith : Commentaire de l'article "The value of multi-proxy reconstruction of past climate" de Li, Nychka et Ammann (2010). *Journal of the American Statistical Association* 105:905–910, 2010.

R. Tibshirani : Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society* 58:267–288, 1996.

R. Tibshirani, G. Walther et T. Hastie : Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.

9 Annexes

Personnes à contacter :

- Gérard Biau, Université Pierre et Marie Curie
  `gerard.biau@upmc.fr`
- Fabienne Comte, Université Paris Descartes
  `fabienne.comte@parisdescartes.fr`
- Thierry Levy, Université Pierre et Marie Curie
  `thierry.levy@upmc.fr`
- Pascal Massart, Université Paris-Sud
  `pascal.massart@math.u-psud.fr`

Voici la liste des documents joints au dossier :

- Déclaration de candidature.
- Rapport de soutenance de thèse.
- Rapports préalables à la soutenance de thèse, rédigés par Fabrice GAMBOA et Gábor LUGOSI.
- Attestation de diplôme.
- Photocopie de la carte d'identité.