



Spécialité : Mathématiques

**Contributions à la statistique mathématique,
à la statistique pour la biologie,
et à l'étude de quelques processus stochastiques**

Document de synthèse présenté pour l'obtention d'une

Habilitation à Diriger des Recherches

par

Christophe GIRAUD

École Polytechnique

Centre de Mathématiques Appliquées

le 5 décembre 2011

Rapporteurs

Florentina BUNEA Cornell University
Pascal MASSART Université Paris-Sud XI
Jean-Philippe VERT Mines-ParisTech et Institut Curie

Examineurs

Yannick BARAUD Université de Nice
Sylvie MÉLÉARD École Polytechnique
Alexandre TSYBAKOV Université Paris VI et ENSAE
Wendelin WERNER Université Paris-Sud XI

A l'aléa...

...et surtout aux heureux hasards!

Remerciements

Je remercie très chaleureusement Florentina BUNEA, Pascal MASSART et Jean-Philippe VERT pour avoir accepté de prendre un peu de leur temps précieux pour rapporter mon mémoire. C'est un grand honneur que des scientifiques d'une telle qualité se penchent sur mes travaux. Je voudrais aussi dire à quel point j'apprécie leur générosité et leur accessibilité. Je profite de cette occasion pour exprimer ma grande admiration pour les contributions théoriques en statistiques "implémentables" de Florentina BUNEA, pour l'œuvre et le sens de l'intérêt général de Pascal MASSART, et enfin pour le tour de force réalisé par Jean-Philippe VERT, en développant la meilleure des recherches appliquées tout en restant à la pointe des développements les plus théoriques.

Je suis aussi extrêmement heureux de pouvoir compter dans mon jury Yannick BARAUD, dont je reste infiniment reconnaissant pour son amitié et pour avoir partagé sans limite son savoir, Sylvie MÉLÉARD que je remercie vivement pour sa confiance, son accueil chaleureux dans son équipe à l'X et pour m'avoir fait découvrir le monde de l'Écologie scientifique, Alexandre TSYBAKOV dont j'admire beaucoup les contributions statistiques et avec qui j'apprécie de faire cours à l'X, et Wendelin WERNER dont les qualités humaines n'ont d'égales que la profondeur et la finesse de ses mathématiques.

Les travaux présentés dans ce mémoire ne sont pas le fruit du travail d'un individu isolé, mais bien le résultat d'interactions avec divers scientifiques que je tiens à remercier individuellement. Le premier que je tiens à remercier est Jean BERTOIN, mon directeur de thèse, qui par l'élégance de ses mathématiques m'a donné l'envie de faire une thèse puis de la recherche en mathématiques. Je suis aussi extrêmement reconnaissant envers Yannick BARAUD. Bien sûr pour m'avoir fait découvrir les statistiques, mais surtout pour m'avoir soutenu tant scientifiquement que personnellement dans une période difficile. Merci.

Mes principales collaborations et discussions scientifiques sont aujourd'hui avec Sylvie HUET, avec qui j'ai un plaisir permanent à travailler. Et notre boîte à projets déborde encore ! J'espère aussi avoir rapidement de nouvelles occasions de travailler avec Nicolas VERZELEN, dont j'admire l'éthique scientifique et envie secrètement la vélocité mathématique.

Depuis quelques temps, j'ai tourné la majorité de mon activité de recherche vers les statistiques pour la biologie. Je ne conçois cette activité qu'à travers des collaborations étroites avec des collègues biologistes, afin de maximiser la pertinence de mon travail. Cette démarche demande un temps plus long pour produire des résultats, mais cet inconvénient (au regard de nos critères d'évaluation) est largement compensé par la richesse des interactions. Je tiens donc à remercier vivement les collègues biologistes qui ont partagé cette démarche avec moi, en particulier (par ordre alphabétique) Mélisande BLEIN-NICOLAS, Dominique DE VIENNE, Christine DILLMANN, Romain JULLIARD, Emmanuelle PORCHER et Michel ZIVY. J'espère que nos discussions s'intensifieront dans les mois à venir.

Depuis neuf ans, j'ai eu le plaisir de fréquenter trois laboratoires différents : le laboratoire

J.-A. Dieudonné à l'Université de Nice, le laboratoire MIA à l'INRA Jouy-en-Josas et le CMAP à l'École Polytechnique. Dans ces trois lieux, j'y ai apprécié la confiance et la bienveillance de mes collègues, ainsi que l'amitié de certains d'entre eux. Je remercie en particulier celles et ceux qui ont partagé mon bureau, Christophette, Mathieu, Stéphanie et Vincent pour leur bonne humeur. Enfin, je profite de cette occasion pour remercier chaleureusement les gestionnaires de ces trois laboratoires pour leur dévouement, leur gentillesse et leur compétence. Merci à vous toutes.

Je suis aussi profondément redevable envers mes étudiants (encadrés lors de projets, de mémoires ou simplement ayant assisté à mes cours) pour leur fraîcheur, leur avidité d'apprendre et leurs salutaires remises en cause. Si ce mémoire n'est tourné que vers mes activités de recherche, le cœur de mon métier d'enseignement-chercheur reste, de mon point de vue, la structuration, l'épuration et la transmission d'un savoir.

Pour finir j'aimerais remercier trois amis, François, Djalil et Lorenzo, pour nos nombreuses discussions sur le sens et le devenir de notre métier. Plus encore, je remercie de tout mon cœur mes proches pour leur soutien, leur attention et leur patience. En particulier, merci infiniment à *Mademoiselle de Montréal* pour ces belles années passées ensemble, et celles à venir !

Table des matières

Remerciements	iii
Productions scientifiques	2
Projets structurants	4
Parcours scientifique	5
Présentation synthétique	8
1 Statistiques mathématiques	9
1.1 Régression gaussienne à variance inconnue	9
1.2 Modèles Graphiques Gaussiens	14
1.3 Régression multivariée de faible rang	16
1.4 Quelques perspectives	19
2 Statistiques pour la biologie	21
2.1 Inférence de réseaux de régulation	21
2.2 Analyse des déformations anatomiques du cerveau	23
2.3 Délimitation de populations synchrones	24
2.4 Quelques perspectives	26
3 Etude de quelques processus stochastiques	28
3.1 Turbulence de Burgers et particules collantes	28
3.2 Particules collantes en interaction gravitationnelle	33
3.3 Quelques perspectives	34
Bibliographie exogène	39

Productions scientifiques

Prépublications

- [A1] C. GIRAUD, R. JULLIARD, E. PORCHER
Delimiting synchronous populations from monitoring data. Preprint (2011).
- [A2] C. GIRAUD, S. HUET, N. VERZELEN
High-dimensional regression with unknown variance. Preprint (2011).
- [A3] S. ALLASSONNIÈRE, P. JOLIVET, C. GIRAUD
Detecting long distance conditional correlations between anatomical regions using Gaussian Graphical Models. Preprint (2011).
- [A4] Y. BARAUD, C. GIRAUD, S. HUET
Estimator selection in the Gaussian setting. Preprint (2010).
arXiv:1007.2096v2

Articles dans des revues internationales à comité de lecture

- [A5] C. GIRAUD, S. HUET, N. VERZELEN
Graph selection with GGMselect.
A paraître dans *Statistical Applications in Genetics and Molecular Biology*.
- [A6] C. GIRAUD
Low rank multivariate regression.
Electronic Journal of Statistics 5 (2011), 775–799.
- [A7] Y. BARAUD, C. GIRAUD, S. HUET
Gaussian model selection with unknown variance.
Annals of Statistics 37 (2009), no. 2, 630–672.
- [A8] C. GIRAUD
Estimation of Gaussian graph by model selection.
Electronic Journal of Statistics 2 (2008), 542–563.
- [A9] C. GIRAUD
Mixing Least-square estimators when the variance is unknown.
Bernoulli 14, no. 4 (2008) 1089–1107.
- [A10] C. GIRAUD
Gravitational clustering and additive coalescence.
Stoch. Proc. Appl. 115 (2005), no. 8, 1302–1322.
- [A11] C. GIRAUD
On the convex hull of a Brownian excursion with parabolic drift.
Stoch. Proc. Appl. 106 (2003), no. 1, 41–62.
- [A12] C. GIRAUD
On a shock front in Burgers turbulence.
J. Statist. Phys. 111 (2003), no. 1-2, 387–402
- [A13] C. GIRAUD
On regular points in Burgers turbulence with stable noise initial data.
Ann. Inst. H. Poincaré Probab. Statist. 38 (2002), no. 2, 229–251
- [A14] C. GIRAUD
Clustering in a self-gravitating one-dimensional gas at zero temperature.
J. Statist. Phys. 105 (2001), no. 3-4, 585–604.
- [A15] J. BERTOIN, C. GIRAUD, Y. ISOZAKI
Statistics of a flux in Burgers turbulence with one-sided Brownian initial data.
Comm. Math. Phys. 224 (2001), no. 2, 551–564.

- [A16] C. GIRAUD
Genealogy of shocks in Burgers turbulence with white noise initial velocity.
Comm. Math. Phys. 223 (2001), no. 1, 67–86.

Proceeding

- [P1] C. GIRAUD
Some properties of Burgers turbulence with white noise initial conditions.
Probabilistic Methods in Fluids, 161–178, World Scientific Publisher (2003).

Courte note

- [CN1] C. GIRAUD
A pseudo RIP for multivariate regression. Preprint (2011).
arXiv:1106.5599v1

Codes informatiques et documents associés

- [C1] GGMselect. Librairie pour R 2.9.1 dédiée à l'estimation dans les modèles graphiques gaussiens. Optimisé en C et FORTRAN.
Document associé : *An introduction to GGMselect*. (notice détaillée d'utilisation)
<http://cran.r-project.org/web/packages/GGMselect/vignettes/Notice.pdf>
- [C2] KF. Code R 2.10.0 dédié à la régression multivariée.
Document associé : *Notice for KF*. (notice d'utilisation)
- [C3] STV. Code FORTRAN interfacé en R 2.11.0 dédié à la délimitation de populations synchrones.
Le coeur du code est un algorithme d'optimisation Primal-Dual.
Avec Notice d'utilisation.

Thèse de Doctorat

- [T] C. GIRAUD,
Turbulence de Burgers et agrégation de particules lorsque l'état initial est aléatoire.
Thèse de Doctorat, Université Paris VI, 2001.

Livre pédagogique

- [L] C. GIRAUD,
Martingales pour la finance. Cours et exercices corrigés. Preprint.
<http://www.cmap.polytechnique.fr/~giraud/MartingalesFinance.pdf>

Projets structurants

Participation aux projets structurants suivants :

Chaire Modélisation Mathématique et Biodiversité.

École Polytechnique - Muséum d'Histoire Naturelle - Veolia environnement.

Portée par Sylvie Méléard (École Polytechnique).

Membre du comité de pilotage.

SONATA. Projet INRA-INRIA porté par Sébastien Aubourg (INRA-Evry).

Gènes orphelins de *A. thaliana* impliqués dans la réponse aux stress biotiques et abiotiques.

Heteroyeast. Projet ANR porté par Dominique de Vienne (Paris-sud XI).

Analyse et prédiction de l'hétérosis chez la levure avec des visées œnologiques.

IRMgroup. Projet ANR porté par Stéphane Mallat (École Polytechnique).

Représentations invariantes par groupement multirésolution.

Parcimonie. Projet ANR porté par Erwan Le Pennec (INRIA Saclay)

Parcimonie en statistiques.

CBME. Projet ANR porté par Jean-Jacques Daudin (AgroParisTech).

Analyse des données métagénomiques.

Parcours scientifique

Si le hasard est le plus petit dénominateur commun de mon activité de recherche, il en est aussi le fil conducteur. Mon activité scientifique n'est pas celle d'un individu isolé méditant en solitaire, elle reflète au contraire le hasard des rencontres et mon interaction avec les diverses équipes scientifiques que j'ai cotoyé. L'objet de ce court chapitre est de présenter brièvement le contexte scientifique dans lequel s'est développée mon activité de recherche.

Mes premiers travaux sur la turbulence de Burgers, la coalescence et l'agrégation de particules en interaction gravitationnelle, reflètent sans conteste l'influence de mon directeur de thèse Jean Bertoin. Ce fut bien sûr le cas pour les travaux [A16], [A15], [A14] et [A13] réalisés durant mon doctorat à l'Université Paris 6 (1999-2001), mais aussi pour les travaux ultérieurs [A12], [A11], [P1] et [A10] en partie réalisés à l'Université de Nice. Je garde encore aujourd'hui un goût immodéré pour l'élégance et la finesse des mathématiques de Jean Bertoin.

Premier probabiliste recruté à l'Université de Nice (sept. 2002), j'y reçu la formidable mission d'y insuffler une activité de recherche en probabilité. Ma contribution à la formation d'une équipe dynamique en probabilités et statistiques au laboratoire J.A. Dieudonné reste bien-sûr modeste, mais c'est avec plaisir que je vois cette équipe aujourd'hui florissante. Ce succès n'aurait jamais eu lieu sans le soutien indéfectible et inconditionnel du laboratoire, y compris pour toutes les (modestes) actions que j'ai entreprise dans ce sens. Je réalise aujourd'hui pleinement l'importance du soutien apporté par des professeurs de renom, lorsqu'ils venaient régulièrement participer à des groupes de travail / séminaires / journées thématiques que j'organisais sur des sujets probabilistes parfois forts éloignés de leur préoccupations scientifiques.

Les premières années à Nice furent aussi des années de questionnement scientifique. Les premiers projets furent très vite suspendus par la triste disparition de Frédéric Poupaud (dont les qualités humaines et scientifiques restent pour moi un idéal inspirant). Le choix d'un nouveau projet scientifique s'avéra un long processus solitaire semé de doutes et d'indécisions. Finalement, c'est suite à l'arrivée de Yannick Baraud au laboratoire que je découvris les statistiques et fus séduit par ses nouveaux horizons. Le colloque "Mathematical Foundation of Statistical Learning" à Barcelona en juin 2004 acheva de me convaincre d'en faire mon terrain de recherche. Mes premiers travaux en statistique furent avec Yannick Baraud et Sylvie Huet sur la sélection de modèles à variance inconnue [A7]. Lorsque l'objectif est la prédiction, le mélange d'estimateurs est une alternative intéressante à la sélection de modèles. Les mélanges avec des distributions de Gibbs ne pouvant laisser indifférent un ancien probabiliste à coloration "physicienne", j'étendis partiellement les résultats de Leung et Barron [LB06] au cadre de la variance inconnue [A9].

Le laboratoire MIA de l'INRA de Jouy-en-Josas m'offrit l'opportunité de venir passer l'année universitaire 2007-2008 en son sein (en délégation). J'y ai découvert de nouveaux questionnements statistiques liés à l'analyse de données biologiques et ce fut une occasion unique de prendre le temps de discuter avec de nombreux collègues biologistes d'une part de la pertinence de certaines modélisations et d'autre part de leur vision sur des branches émergentes de la biologie, telles que la biologie systémique ou la métagénomique. S'il est inutile de comprendre tous les détails des processus biologiques pour pouvoir développer des analyses pertinentes, il est par contre

indispensable d'avoir une vision claire des enjeux et problèmes étudiés. Ces discussions furent primordiales dans l'élaboration de mon corpus scientifique de statisticien pour la biologie. Au cours de cette année passée à l'INRA, mon activité principale a été l'étude de l'inférence dans les modèles graphiques gaussiens dans l'optique d'inférer des réseaux de régulation génique à partir de données d'expression de gènes. Mon premier travail [A8], de nature théorique, explorait les possibilités d'inférence par sélection de modèles. Avec Sylvie Huet et Nicolas Verzelen, nous en avons extrait une procédure implémentable en pratique [A5], que nous avons mise en oeuvre dans le package R GGMselect [C1]. Cette année passée à Jouy-en-Josas fut aussi l'occasion d'amorcer divers projets au long cours, tels que les projets Heterosyeast (analyse de données peptidiques et phénotypiques pour la construction de marqueurs d'hétérosis chez la levure), CBME (analyse de données metagénomiques) ou SONATA (analyse de gènes orphelins de *A. thaliana* impliqués dans la réaction aux stress biotiques et abiotiques).

En septembre 2008, j'ai été recruté comme professeur chargé de cours résident au département de mathématiques appliquées de l'École Polytechnique. L'ambition était de créer une branche statistique dans l'équipe *Modélisation pour l'Evolution du Vivant* pilotée par Sylvie Méléard. Mon arrivée au CMAP fut une occasion supplémentaire de découvrir de nouveaux horizons. Pour commencer, j'ai découvert les problématiques de l'écologie au travers de mes enseignements et de ma participation aux activités de la chaire MMB. De discussions à la pause café, en encadrement de stage de Master, en expérimentations numériques, nous avons développé avec Emmanuelle Porcher et Romain Julliard une procédure statistique ([A1] et [C3]) visant à délimiter des populations de dynamique synchrone à partir des données amateurs du réseau STOC. Une seconde thématique découverte en arrivant au CMAP est l'analyse d'images. Dans un premier temps nous avons développé avec Stéphanie Allasonnière une procédure d'estimation basée sur les modèles graphiques, dédiée à l'analyse du réseau de connectivité du cerveau [A3]. Dans un second temps, nous nous sommes intéressés avec Stéphane Mallat et Lorenzo Rosasco aux aspects statistiques des transformées de scattering introduites par Stéphane Mallat pour la construction d'invariants en image (dans le cadre du projet IRMgroup). Parallèlement à ces nouveaux horizons, j'ai poursuivi mes collaborations avec les équipes de génétique végétale du Moulon et de génomique végétale de l'INRA d'Evry. Ces projets de longue haleine (conception/expériences/modélisation/analyse des données/interprétation) sont toujours en cours. Enfin, certains questionnements récurrents m'ont conduit vers l'analyse mathématique de divers problèmes statistiques fondamentaux. Dans la poursuite du travail [A5] sur les modèles graphiques gaussiens, nous avons développé avec Yannick Baraud et Sylvie Huet une procédure statistique pour sélectionner un estimateur parmi une famille arbitraire, dans le cadre de la régression gaussienne à variance inconnue [A4]. L'élaboration d'un package R est en cours. Je me suis aussi intéressé à la régression multivariée [A6], [CN1] et [C2], motivé par le problème de relier une famille de phénotypes à des abondances de protéines.

Il est notable que le travail réalisé avec Stéphanie Allasonnière s'inspire de travaux pour la génomique, et qu'à l'inverse l'algorithme pour délimiter des populations synchrones s'inspire de techniques de segmentation d'images. Ce transfert d'idées d'une communauté vers une autre, n'est possible que grâce à la réunion en un même lieu (le CMAP) de mathématiciens appliqués de cultures différentes. C'est une chance de travailler dans un tel lieu !

Présentation synthétique des travaux

Ce chapitre présente de façon synthétique l'essentiel des travaux mentionnés au chapitre *productions scientifiques*. Il est articulé autour de trois grands axes :

1. Statistique mathématique (2005–2011)

- Régression gaussienne à variance inconnue :
 - a) sélection de modèles
 - b) mélange d'estimateurs
 - c) sélection d'estimateurs
- Modèles Graphiques Gaussiens
- Régression multivariée de faible rang

2. Statistique pour la biologie (2008–2011)

- Inférence de réseaux de régulation
- Analyse de déformations anatomiques du cerveau
- Délimitation de populations synchrones

3. Turbulence de Burgers et systèmes de particules en interactions (1999–2004)

- Turbulence de Burgers et particules collantes
- Particules collantes en interaction gravitationnelle

Cette séparation en trois axes dessine des frontières un peu arbitraires entre les différents travaux. En effet, si les travaux probabilistes (troisième axe) sont totalement dissociés à la fois temporellement et thématiquement des travaux de statistiques (deux premiers axes), les travaux de statistique mathématique ne sont pas totalement dissociés de ceux de statistique pour la biologie. Au contraire, ils reflètent en partie des interrogations nées de l'analyse de données biologiques. A l'inverse, les procédures d'estimation développées pour des problèmes biologiques s'inspirent des avancées récentes en statistique mathématique.

Les principales techniques mathématiques mises en oeuvre sont les décompositions de trajectoires, le calcul stochastique, la convergence de processus, les processus empiriques, les inégalités de concentration, l'approximation, les matrices aléatoires et l'optimisation convexe.

1 Statistiques mathématiques

1.1 Régression gaussienne à variance inconnue

La régression gaussienne correspond au cadre où on dispose de n observations

$$Y_i = f_i + \varepsilon_i, \quad i = 1, \dots, n,$$

avec $f = (f_1, \dots, f_n)^T$ un vecteur inconnu de \mathbb{R}^n et les $\varepsilon_1, \dots, \varepsilon_n$ des variables aléatoires indépendantes de loi gaussienne $\mathcal{N}(0, \sigma^2)$. Les formes usuelles pour f sont

1. $f = (F(x_1), \dots, F(x_n))^T$ avec $F : \mathcal{X} \rightarrow \mathbb{R}$ et $x_1, \dots, x_n \in \mathcal{X}$,
2. $f = X\beta$ avec X une matrice réelle de taille $n \times p$ et β un vecteur de \mathbb{R}^p .

Dans le premier cas, les questions classiques sont d'estimer la fonction F , détecter les points de ruptures de F si F est constante par morceaux, etc. Dans le second cas, la dimension p du vecteur β est généralement grande, possiblement plus grande que n , mais le vecteur β est "sparse" dans le sens où peu de coordonnées de β sont non-nulles. L'objectif est alors d'estimer les coordonnées non-nulles de β .

De nombreux estimateurs ont été développés présentant des propriétés d'optimalité pour divers objectifs statistiques et sous diverses hypothèses sur f . Une question générale est de savoir comment obtenir à partir d'une telle collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ d'estimateurs, un estimateur \hat{f} combinant (presque) toutes les bonnes propriétés de chacun des estimateurs \hat{f}_λ . Une façon de poser le problème est d'introduire le risque quadratique d'un estimateur \hat{f}

$$R(\hat{f}) = \mathbb{E} \left[\|f - \hat{f}\|^2 \right]$$

et de chercher un estimateur dont le risque quadratique soit presque aussi petit que le risque oracle

$$R_{\text{oracle}}(\Lambda) = \inf_{\lambda \in \Lambda} R(\hat{f}_\lambda). \quad (1.1)$$

Deux points de vue ont été développés dans la littérature statistique. Le premier consiste à sélectionner un estimateur $\hat{f}_{\hat{\lambda}}$ parmi la collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$. Cette sélection s'effectue typiquement en minimisant un critère $\text{crit}(\lambda) = \hat{R}(\hat{f}_\lambda)$ visant à estimer (éventuellement avec biais) le risque $R(\hat{f}_\lambda)$. Les critères de sélection AIC (Akaike [Ak73]) ou BIC (Schwartz [Sch78]), la sélection de modèles (Birgé et Massart [BM01]) ou la cross-validation [Ge75] entrent dans ce cadre là. Un autre point de vue consiste à prendre une combinaison convexe des estimateurs

$$\hat{f} = \sum_{\lambda \in \Lambda} \omega_\lambda \hat{f}_\lambda, \quad \text{avec } \omega_\lambda \geq 0 \text{ et } \sum_{\lambda \in \Lambda} \omega_\lambda = 1, \quad (1.2)$$

ou plus généralement une combinaison linéaire des \hat{f}_λ (c'est à dire sans contrainte sur les poids ω_λ dans (1.2)). Les estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ et les poids $\{\omega_\lambda, \lambda \in \Lambda\}$ peuvent être construit à partir de deux échantillons différents [Ca04, Ya00, JN00, Ts03, Ya04, BTW07, DT08, Go09] ou être basés sur le même échantillon [LB06, RT11a].

a) Sélection de modèle

Lorsque les estimateurs \hat{f}_λ sont obtenus en maximisant la vraisemblance sur un sous-espace S_λ de \mathbb{R}^n (appelé modèle), le choix d'un estimateur $\hat{f}_{\hat{\lambda}}$ parmi la famille $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ est communément appelé *sélection de modèle*. Dans le cadre où la variance σ^2 des ε_i est connue, la sélection de modèle a été étudiée en détail et d'un point de vue non-asymptotique par Birgé et Massart [BM01].

Lorsque la variance σ^2 est inconnue, de nombreux critères de sélection ont été proposés, les plus classiques étant AIC, BIC ou FPE. Ces critères sont basés sur diverses heuristiques et possèdent certaines qualités d’optimalité asymptotique. Par exemple, si le nombre de modèles S_λ de dimension d croît à vitesse au plus polynomiale en d et si $f \notin \bigcup_{\lambda \in \Lambda} S_\lambda$, les critères AIC et FPE sont asymptotiquement efficaces [Sh81, Li87, Sh97]. Inversement, s’il existe un unique modèle minimal S_{λ_0} contenant f , le critère BIC est consistant dans le sens où $\mathbb{P}(\hat{\lambda}_{\text{BIC}} = \lambda_0) \rightarrow 1$ lorsque n tend vers l’infini [Ni84].

Dans les contextes statistiques actuels où l’espace des paramètres est généralement de très grande dimension, de tels résultats asymptotiques ne peuvent pas nous donner de bonnes intuitions sur le choix de \hat{f}_λ . Par exemple, dans le cadre de la régression linéaire $f = X\beta$, les résultats classiques supposent que la dimension p de β est fixe, alors que n tend vers l’infini. Ce cadre n’est donc pas du tout adapté pour les situations, aujourd’hui courantes, où la dimension p est aussi grande, voir beaucoup plus grande que n . Une approche non-asymptotique permet de prendre en compte tous les paramètres du problème de sélection (taille n de l’échantillon, dimension p de l’espace des paramètres, complexité de la famille de modèle, etc) et permet d’éviter les hypothèses du type $f \notin \bigcup_{\lambda \in \Lambda} S_\lambda$ ou $f \in S_{\lambda_0}$.

Dans l’article [A7], écrit en collaboration avec Y. Baraud et S. Huet, nous analysons d’un point de vue non-asymptotique des critères de la forme log-vraisemblance pénalisée,

$$\text{crit}(\lambda) = \log \left(\|Y - \hat{f}_\lambda\|^2 \right) + \text{pen}(\lambda) \quad \text{ou} \quad \text{crit}'(\lambda) = \|Y - \hat{f}_\lambda\|^2 \left(1 + \frac{\text{pen}'(\lambda)}{n - \dim(S_\lambda)} \right)$$

pour la sélection de modèles lorsque la variance σ^2 est inconnue. Dans un premier temps, nous analysons des critères de sélection classiques avec une pénalité proportionnelle à la dimension du modèle, et donnons des conditions sur la famille $\{S_\lambda, \lambda \in \Lambda\}$ pour que le risque quadratique de l’estimateur sélectionné soit du même ordre de grandeur (ou presque du même ordre de grandeur) que le risque oracle (1.1). Nous soulignons en particulier que dans le cadre de la sélection de variables complète, le critère BIC (et donc a fortiori AIC, FPE, etc) peut conduire à sélectionner un modèle de beaucoup trop grande dimension si la dimension p du paramètre β est du même ordre de grandeur que n . Dans un second temps, nous proposons une pénalité $\text{pen}'(\lambda)$ flexible, s’adaptant à (presque) toute famille de modèle $\{S_\lambda, \lambda \in \Lambda\}$, et pour laquelle le risque quadratique (ou Kullback) de l’estimateur sélectionné est contrôlé (non asymptotiquement) par la borne

$$R(\hat{f}_\lambda) \leq C \inf_{\lambda \in \Lambda} \left\{ \|f - \hat{f}_\lambda\|^2 \left(1 + \frac{\text{pen}'(\lambda)}{n - \dim(S_\lambda)} \right) + \text{pen}'(\lambda)\sigma^2 \right\},$$

où $C > 0$ est une constante numérique universelle.

Dans un dernier temps, nous détaillons comment cette machinerie peut être mise en oeuvre pour détecter les composante non nulles de f , faire de la sélection de variables, détecter des ruptures ou des changements de pente dans un signal, estimer de façon adaptative un signal, etc. Dans chaque cas, nous explicitons la borne de risque obtenue.

b) Mélange d’estimateurs

Les qualités de la sélection de modèle sont doubles. D’une part, elle permet de sélectionner un estimateur dont le risque est contrôlé par le risque oracle $R_{\text{oracle}}(\Lambda)$, d’autre part le modèle sélectionné $S_{\hat{\lambda}}$ donne une information utile sur la structure de f . Le mélange d’estimateurs (1.2) est une alternative intéressante à la sélection de modèle, lorsque la préoccupation principale est d’obtenir un estimateur \hat{f} de risque $R(\hat{f})$ minimal. La plupart des procédures de mélange d’estimateurs [Ca04, Ya00, JN00, Ts03, Ya04, BTW07, BN08, DT08, Go09] nécessitent la connaissance de la variance σ^2 du bruit, et sont basées sur un découpage des données en deux

parties, une partie pour construire les estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ et une partie pour construire les poids $\{\omega_\lambda, \lambda \in \Lambda\}$ dans (1.2).

Dans le cadre de la régression gaussienne avec variance σ^2 connue, Leung et Barron [LB06] (voir aussi [RT11a]) proposent une forme de mélange de Gibbs dont les poids $\{\omega_\lambda, \lambda \in \Lambda\}$ et les estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ sont construits à partir du même jeu de données $Y = (y_1, \dots, y_n)^T$. Les estimateurs considérés sont les estimateurs \hat{f}_λ obtenus en maximisant la vraisemblance sur un sous espace linéaire $S_\lambda \subset \mathbb{R}^n$, c'est-à-dire $\hat{f}_\lambda = \Pi_{S_\lambda} Y$ où Π_{S_λ} est la projection orthogonale sur l'espace S_λ . Leung et Barron [LB06] proposent une analyse très élégante (basée sur le lemme de Stein) du mélange obtenu en prenant comme poids

$$\omega_\lambda = \frac{\pi_\lambda}{\mathcal{Z}} \exp\left(-\beta(\|Y - \hat{f}_\lambda\|^2/\sigma^2 + 2\dim(S_\lambda))\right), \quad \text{où } \mathcal{Z} = \sum_{\lambda \in \Lambda} \pi_\lambda e^{-\beta(\|Y - \hat{f}_\lambda\|^2/\sigma^2 + 2\dim(S_\lambda))},$$

$\{\pi_\lambda, \lambda \in \Lambda\}$ est une mesure de probabilité sur Λ et $\beta \leq 1/4$. Ils obtiennent pour ce mélange (avec $\beta \leq 1/4$) la borne de risque

$$R(\hat{f}) \leq \inf_{\lambda \in \Lambda} \left\{ R(\hat{f}_\lambda) + \frac{1}{\beta} \log \frac{1}{\pi_\lambda} \right\} \quad (1.3)$$

qui possède la qualité remarquable d'avoir une constante égale à 1 devant le risque $R(\hat{f}_\lambda)$. Dans l'article [A9], nous étendons ces résultats au cas où la variance σ^2 est inconnue, avec en tête un cadre fonctionnel où $f_i = F(x_i)$, $i = 1, \dots, n$. Dans ce cadre, les collections de modèles pertinentes possèdent en général une structure d'ordre partiel dont il est possible de tirer partie pour estimer la variance. Nous proposons une forme de mélange ne demandant aucune connaissance sur σ^2 , et démontrons pour l'estimateur \hat{f} obtenu une inégalité de la forme (1.3), mais avec une constante $\left(1 + \frac{1}{2n \log n}\right)$ devant l'infimum. Nous illustrons aussi l'intérêt du mélange d'estimateurs en améliorant cette borne d'un terme $-\frac{\sigma^2}{\beta} \log(k)$, lorsqu'il existe k "bons modèles" pour estimer f .

L'estimateur \hat{f} proposé dans [A9] peut prendre une forme particulièrement simple lorsque la fonction $F(x)$ est recherchée sous la forme

$$F(x) = \sum_{j=1}^p \beta_j \varphi_j(x),$$

avec les φ_j orthogonaux dans $L^2\left(\frac{1}{n} \sum_i \delta_{x_i}\right)$. Pour des modèles de la forme

$$S_\lambda = \{\beta \in \mathbb{R}^p, \beta_j = 0 \text{ si } j \notin \lambda\}, \quad \lambda \subset \{1, \dots, p\}$$

et lorsque $\log(\pi_\lambda)$ est une fonction affine de $\dim(S_\lambda)$, l'estimateur \hat{f} obtenu est un estimateur de "shrinkage" avec une fonction de "shrinkage" correspondant à une version régulière du seuillage dur, voir Figure 1.1. Cette forme est particulièrement appréciable, car son coût de calcul est très léger. Nous montrons aussi pour cette forme particulière, qu'une borne du type (1.3) peut être obtenue pour des paramètres β allant jusqu'à $1/2$, alors qu'un choix $\beta > 1/2$ peut conduire à une explosion de la variance (overfitting). Enfin, à titre d'illustration, nous examinons les performances de cet estimateur pour estimer une fonction à variation bornée à l'aide d'une base de Haar. Dans ce cadre, l'estimateur de "shrinkage" \hat{f} permet d'estimer toute fonction F de variation totale $V(F)$ (inconnue) à la vitesse $(V(F)\sigma^2 \log(n)/n)^{2/3}$. L'estimateur \hat{f} possède donc la qualité de se calculer rapidement, mais il présente une perte de vitesse en $(\log n)^{2/3}$ par rapport à la vitesse minimax. Nous concluons, en montrant qu'un algorithme à la Birgé-Massart [BM00] pour les bases de Haar, permet au mélange d'atteindre la vitesse minimax, au prix cependant d'un coût de calcul un peu plus élevé.

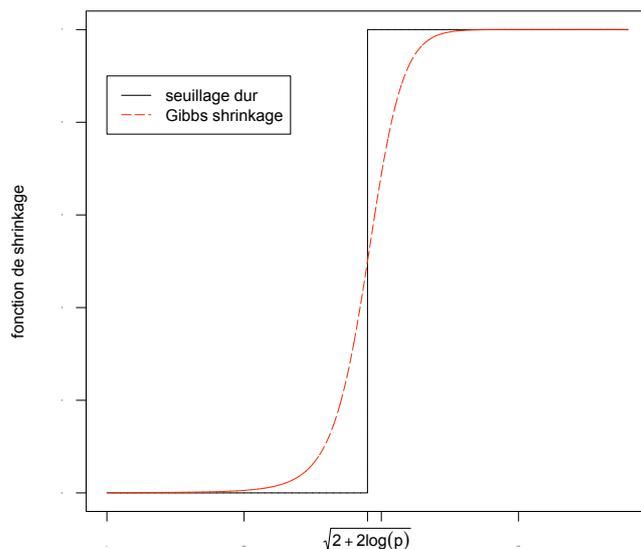


FIG. 1.1 – La fonction de "shrinkage" dans le cas orthogonal.

c) Sélection d'estimateurs

Les procédures de sélection ou de mélange d'estimateurs décrites dans les articles [A7] et [A9] possèdent de bonnes propriétés statistiques (contrôle de leur risque quadratique) mais souffrent de deux rigidités.

La première rigidité, c'est qu'elles sont conçues pour travailler avec une collection *fixe* de modèles $\{S_\lambda, \lambda \in \Lambda\}$. Dans certains cas, comme la sélection de variable complète, la collection de modèles $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ adaptée au problème est extrêmement grande et les procédures de sélection ou mélange ne peuvent pas être mise en oeuvre en pratique à cause des temps de calcul rédhibitoires (à noter cependant que pour la régression linéaire sparse, Rigollet et Tsybakov [RT11a] proposent une méthode MCMC efficace pour le calcul approché du mélange introduit par Leung et Barron [LB06]). Dans ce cas, on aimerait travailler avec une sous-collection de modèles $\{S_\lambda, \lambda \in \hat{\Lambda}\}$ où $\hat{\Lambda} \subset \Lambda$ pourrait dépendre des données. Par exemple, pour la sélection de variables complète, la famille $\{S_\lambda, \lambda \in \hat{\Lambda}\}$ pourrait être générée à l'aide des algorithmes LARS [EHJT04], PLS [Wo66, He90], Random Forest [Br01], etc. L'objectif est alors de sélectionner le meilleur estimateur \hat{f}_λ parmi la collection $\{\hat{f}_\lambda, \lambda \in \hat{\Lambda}\}$.

La seconde rigidité des procédures décrites dans [A7] et [A9], c'est qu'elles ne travaillent qu'avec des estimateurs du type $\hat{f}_\lambda = \Pi_{S_\lambda} Y$. Elles ne sont donc pas utiles pour sélectionner (ou mélanger) parmi des estimateurs tels que le Lasso, les k -plus-proches-voisins, les estimateurs "splines", les estimateurs à noyaux, etc. En particulier, elles ne permettent ni de sélectionner le(s) paramètre(s) d'une procédure comme le λ du Lasso

$$\hat{\beta}_\lambda^{\text{Lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2 + 2\lambda|\beta|_{\ell^1},$$

ni le noyau d'un estimateur à noyau, ni de choisir parmi différentes procédures d'estimation.

La validation-croisée propose une procédure générique pour sélectionner parmi une collection (quasi)-arbitraire d'estimateurs $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ et permet donc de répondre aux deux problèmes

mentionnés ci-dessus. Cependant, la justification des procédures de validation-croisée dans un cadre non-asymptotique est restreinte à quelques cas particuliers, voir Arlot et Celisse [AC10] pour une review récente. Par exemple, aucun résultat non-asymptotique n'existe pour la sélection de variables complète.

Dans l'article [A4], écrit en collaboration avec Y. Baraud et S. Huet, nous proposons une méthode permettant de sélectionner parmi une collection (quasi) arbitraire d'estimateurs $\{\hat{f}_\lambda, \lambda \in \hat{\Lambda}\}$ dans un cadre de régression gaussienne à variance inconnue. Le risque de l'estimateur résultant est contrôlé par une borne non-asymptotique ressemblant à une borne de type oracle. A titre d'exemple, cette procédure permet d'agréger des estimateurs préliminaires et d'étendre au cadre de la variance inconnue les résultats de [Ts03, BTW07] (sous une forme un peu plus faible). Elle permet aussi de sélectionner parmi une famille d'estimateurs linéaires, ou de sélectionner parmi une collection de modèles $\{S_\lambda, \lambda \in \hat{\Lambda}\}$ dépendant des données. La procédure de sélection est basée sur une famille \mathbb{S} (fixée) d'espaces approximants $S \subset \mathbb{R}^n$ et une mesure de complexité $\Delta : \mathbb{S} \rightarrow \mathbb{R}^+$ sur cette famille d'espace. Le critère de sélection, fait alors intervenir l'erreur de reconstruction de Y par $\Pi_S \hat{f}_\lambda$, l'erreur d'approximation de \hat{f}_λ par $\Pi_S \hat{f}_\lambda$ et une pénalité $\text{pen}_\Delta(S)$ liée à la mesure de complexité Δ .

A titre d'exemple, décrivons cette procédure pour sélectionner parmi une famille finie d'estimateurs linéaires (estimateurs à noyaux, ridge, "spline", k -plus-proches-voisins, etc) de la forme $\hat{f}_\lambda = A_\lambda Y$. Notons $\text{Im}(A_\lambda)$ l'image de A_λ et $A_\lambda^+ : \text{Im}(A_\lambda) \rightarrow \text{Im}(A_\lambda^T)$ l'inverse de la bijection induite par la restriction de A_λ à $\text{Im}(A_\lambda^T)$. La restriction de la projection $\Pi_{\text{Im}(A_\lambda^T)}$ à l'espace $\text{Im}(A_\lambda)$ induit une application linéaire notée $\bar{\Pi}_\lambda : \text{Im}(A_\lambda) \rightarrow \text{Im}(A_\lambda^T)$. A chaque estimateur $\hat{f}_\lambda = A_\lambda Y$, associons la collection d'espaces approximant $\{S_\lambda^1, \dots, S_\lambda^{n/2}\}$ où S_λ^k est l'espace engendré par "les" k vecteurs singuliers à droites de $A_\lambda^+ - \bar{\Pi}_\lambda$ associés aux k plus petites valeurs singulières. Remarquons que lorsque la matrice A_λ est symétrique définie positive, l'espace S_λ^k coïncide avec l'espace engendré par "les" k vecteurs propres de A_λ associés aux k plus grandes valeurs propres. Introduisons la fonction $\Delta : \mathbb{S} = \bigcup_{\lambda \in \Lambda} \{S_\lambda^1, \dots, S_\lambda^{n/2}\} \rightarrow \mathbb{R}^+$, définie par

$$\Delta(S) = \beta(1 + \dim(S)), \quad \text{avec } \beta \text{ tel que } \sum_{S \in \mathbb{S}} e^{-\beta(1 + \dim(S))} = 1.$$

L'estimateur $\hat{f}_{\hat{\lambda}}$ est sélectionné en minimisant le critère

$$\text{crit}(\lambda) = \inf_{S \in \{S_\lambda^1, \dots, S_\lambda^{n/2}\}} \left[\|Y - \Pi_S \hat{f}_\lambda\|^2 + \frac{1}{2} \|\hat{f}_\lambda - \Pi_S \hat{f}_\lambda\|^2 + K \text{pen}_\Delta(S) \hat{\sigma}_S^2 \right],$$

où K est une constante plus grande que 1,

$$\hat{\sigma}_S^2 = \frac{\|Y - \Pi_S Y\|^2}{n - \dim(S)},$$

et $\text{pen}_\Delta(S)$ est solution de

$$\mathbb{E} \left[\left(U - \frac{\text{pen}_\Delta(S)}{n - \dim(S)} V \right)_+ \right] = e^{-\Delta(S)},$$

avec x_+ la partie positive de x et U, V deux χ^2 indépendants de degré de liberté respectifs $\dim(S) + 1$ et $n - \dim(S) - 1$. Sous la condition que $A_\lambda^+ - \bar{\Pi}_\lambda$ possède au plus $n/2$ valeurs singulières inférieures à $1/2$, le risque de l'estimateur sélectionné satisfait l'inégalité oracle

$$R(\hat{f}_{\hat{\lambda}}) \leq C\beta \inf_{\lambda \in \Lambda} R(\hat{f}_\lambda)$$

pour une constante numérique $C > 1$.

Il est à noter que lorsque les estimateurs sont des estimateurs par projection $\{\hat{f} = \Pi_{S_\lambda} Y, \lambda \in \Lambda\}$, la procédure de sélection décrite ci-dessus coïncide avec celle du a).

1.2 Modèles Graphiques Gaussiens

Considérons un graphe g non-orienté dont les noeuds sont labellés par $\{1, \dots, p\}$. Si deux noeuds a et b sont joints par une arête dans g , nous noterons $a \stackrel{g}{\sim} b$. La loi d'une variable aléatoire $X = (X_1, \dots, X_p)$ est un modèle graphique par rapport au graphe g si pour tout noeud $a \in \{1, \dots, p\}$:

conditionnellement à $\{X_v, v \stackrel{g}{\sim} a\}$, la variable X_a est indépendante de $\{X_b, b \not\stackrel{g}{\sim} a\}$. (1.4)

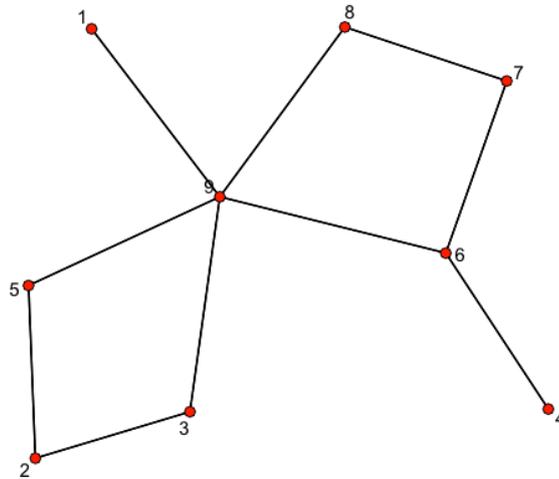


FIG. 1.2 – Exemple : X_6 est indépendante de $\{X_1, X_2, X_3, X_5, X_8\}$ sachant $\{X_4, X_7, X_9\}$.

Lorsque la loi de X possède une densité strictement positive sur \mathbb{R}^p , il existe un unique graphe minimal g_{\min} pour lequel (1.4) est vérifié. Si de plus la variable X est distribuée selon une loi gaussienne $\mathcal{N}(0, \Sigma)$ avec Σ inversible, le graphe g_{\min} est défini par

$$a \stackrel{g_{\min}}{\sim} b \iff \text{cor}(X_a, X_b \mid X_c, c \notin \{a, b\}) \neq 0. \quad (1.5)$$

Le problème d'estimer le graphe g_{\min} d'une loi $\mathcal{N}(0, \Sigma)$ à partir de n réalisations de cette loi a suscité de nombreux travaux récents (par exemple [DP07, WB06, CR06, SPGS00, KB08, HLPL06, YL07, BGA08, FHT08, FFW09, MB06]), avec une attention particulière pour le cas où $p \geq n$, voir $p \gg n$. Au vue de la relation (1.5), une idée naturelle [DP07] est d'estimer g_{\min} en seuillant les corrélations conditionnelles empiriques $\widehat{\text{cor}}(X_a, X_b \mid X_c, c \notin \{a, b\})$. Ces corrélations conditionnelles empiriques ne sont malheureusement pas définie lorsque $p \geq n$ et peuvent être très instables même pour des valeurs modérées de p . Un certain nombre de travaux [WB06, CR06, SPGS00, KB08] ont proposé de travailler à la place avec des corrélations conditionnelles empiriques restreintes $\widehat{\text{cor}}(X_a, X_b \mid X_c, c \in S)$, pour des ensemble S de petite cardinalité.

Les corrélations conditionnelles étant données par la formule

$$\text{cor}(X_a, X_b \mid X_c, c \notin \{a, b\}) = \frac{[\Sigma^{-1}]_{ab}}{\sqrt{[\Sigma^{-1}]_{aa} [\Sigma^{-1}]_{bb}}},$$

le graphe minimal \mathbf{g}_{\min} peut aussi être défini par la relation

$$a \stackrel{\mathbf{g}_{\min}}{\rightsquigarrow} b \iff [\Sigma^{-1}]_{ab} \neq 0. \quad (1.6)$$

Partant de cette caractérisation, un certain nombre de papiers [HLPL06, YL07, BGA08, FHT08, FFW09] proposent de maximiser la log-vraisemblance de Σ^{-1} sous une contrainte ℓ^1 pour obtenir une estimée sparse de Σ^{-1} . Le graphe $\hat{\mathbf{g}}_{\min}$ est alors défini par le squelette de $\hat{\Sigma}^{-1}$, c'est à dire $a \stackrel{\hat{\mathbf{g}}_{\min}}{\rightsquigarrow} b$ si $[\hat{\Sigma}^{-1}]_{ab} \neq 0$. Une alternative à ces approches est de considérer la matrice de régression θ définie par

$$\theta = \operatorname{argmin}_{\theta \in \Theta} \|\Sigma^{1/2}(\theta - I)\|_{p \times p}^2,$$

où Θ est l'ensemble des matrices $p \times p$ à diagonale nulle et $\|\cdot\|_{p \times p}$ est la norme de Hilbert-Schmidt sur les matrices $p \times p$. La relation $\theta_{ab} = -[\Sigma^{-1}]_{ab} / [\Sigma^{-1}]_{aa}$ induit une caractérisation de \mathbf{g}_{\min} en fonction de θ

$$a \stackrel{\mathbf{g}_{\min}}{\rightsquigarrow} b \iff \theta_{ab} \neq 0. \quad (1.7)$$

Cette idée est par exemple exploitée par Meinshausen et Bühlmann [MB06] qui proposent d'estimer la matrice θ à l'aide du Lasso.

Dans l'article [A8], nous proposons d'estimer le graphe \mathbf{g}_{\min} par une approche de sélection de modèles basée sur la matrice θ . A tout graphe \mathbf{g} , associons l'ensemble $\Theta_{\mathbf{g}}$ des matrices de diagonales nulle vérifiant la contrainte $\theta_{a,b} = 0$ s'il n'y a pas d'arête entre a et b dans \mathbf{g} . Si le degré du graphe \mathbf{g} (nombre maximum de voisin d'un noeud de \mathbf{g}) est inférieur à n , on peut alors définir l'estimateur

$$\hat{\theta}_{\mathbf{g}} = \operatorname{argmin}_{\theta \in \Theta_{\mathbf{g}}} \|\mathbf{X}(\theta - I)\|_{n \times p}^2$$

où \mathbf{X} est la matrice $n \times p$ obtenue en rangeant les n réalisations de la loi $\mathcal{N}(0, \Sigma)$ par lignes. Une mesure de risque naturelle associée à l'estimateur $\hat{\theta}_{\mathbf{g}}$ est l'erreur de prédiction

$$R(\hat{\theta}_{\mathbf{g}}) = \mathbb{E} \left[\|\Sigma^{1/2}(\hat{\theta}_{\mathbf{g}} - \theta)\|^2 \right].$$

Dans l'article [A8], nous proposons un critère pour sélectionner un graphe $\hat{\mathbf{g}}$ parmi une collection \mathcal{G} de graphes, et nous montrons que sous la condition

$$\exists 0 < \rho < 1 \text{ tel que : } \deg(\mathbf{g}) \leq \rho \frac{n}{2 \left(1.1 + \sqrt{\log(p)}\right)^2} \text{ pour tout } \mathbf{g} \in \mathcal{G}, \quad (1.8)$$

la borne de risque suivante est vérifiée

$$R(\hat{\theta}_{\hat{\mathbf{g}}}) \leq C \log(p) \inf_{\mathbf{g} \in \mathcal{G}} R(\hat{\theta}_{\mathbf{g}}) + \varepsilon_n, \quad (1.9)$$

où $C > 1$ est une constante numérique et ε_n est un terme résiduel tendant vers 0 à vitesse exponentielle en n . Autrement dit, tant que chaque noeud du graphe \mathbf{g}_{\min} est peu connecté, il est possible d'avoir une estimation raisonnable de θ en terme du risque de prédiction.

La condition (1.8) sur le degré est assez naturelle. Soit \mathcal{G}_{ρ} l'ensemble de tous les graphes à p sommets de degré vérifiant (1.8). Le Lemme 1 de [A8] assure que pour tout $\delta \in]\sqrt{\rho}, 1[$, les inégalités

$$1 - \delta \leq \sup_{\theta \in \bigcup_{\mathbf{g} \in \mathcal{G}_{\rho}} \Theta_{\mathbf{g}}} \frac{\frac{1}{\sqrt{n}} \|\mathbf{X}(\theta - I)\|_{n \times p}^2}{\|\Sigma^{1/2}(\theta - I)\|_{p \times p}^2} \leq 1 + \delta$$

ont lieu avec une probabilité supérieure à $1 - 2 \exp(-n(\delta - \sqrt{\rho})^2/2)$. Inversement, il découle de [BDDW08, CDD09] qu'il existe une constante $c(\delta)$ telle que les inégalités précédentes ne

peuvent pas être vérifiées si le degré de \mathbf{g} est supérieur à $c(\delta)n/(1 + \log(p/n))$. Ces arguments suggèrent donc que les estimateurs $\hat{\theta}_g$ sont fiables tant que (1.8) est vérifié. Les résultats minimaux récents de N. Verzelen [Verz10] pour le problème lié de la régression gaussienne à design aléatoire, confirment que l'ordre de grandeur " $n/2 \log(p)$ " est celui à partir duquel il devient quasi-impossible de faire de l'estimation.

La procédure décrite dans l'article [A8] possède de bonnes propriétés statistiques, mais elle est extrêmement coûteuse en temps de calcul. En pratique, elle ne peut pas être mise en oeuvre lorsque la dimension p de X dépasse quelques petites dizaines. Partant du constat que l'ensemble \mathcal{G} des graphes de degré inférieur à $n/(2 \log p)$ est trop vaste, nous proposons dans l'article [A5] (écrit en collaboration avec N. Verzelen et S. Huet) d'implémenter la procédure décrite dans [A8] uniquement sur une sous famille $\hat{\mathcal{G}}$ de ces graphes, la famille $\hat{\mathcal{G}}$ étant obtenue à l'aide de procédures numériquement efficaces. Cette famille $\hat{\mathcal{G}}$ est de cardinalité beaucoup plus petite que \mathcal{G} , mais elle dépend des données \mathbf{X} . Nous montrons une inégalité ressemblant à (1.9) pour la procédure résultante, ainsi qu'un résultat de consistance dans un cadre $p \gg n$. En résumé, le critère de sélection introduit dans [A8] peut être utilisé pour

1. sélectionner les paramètres dont dépendent les multiples procédures proposées dans la littérature,
2. sélectionner parmi toutes ces procédures la "meilleure" sur le jeu de données.

1.3 Régression multivariée de faible rang

Considérons le modèle de régression multivariée

$$y_i = A^T x_i + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

où l'observation y_i est un vecteur de \mathbb{R}^T , la matrice A est de taille $p \times T$ et la covariable x_i est un vecteur de \mathbb{R}^p . En notant Y , X et E les matrices de lignes respectives y_i^T , x_i^T et ε_i^T , le modèle de régression multivarié se met sous la forme matricielle

$$Y = XA + \sigma E.$$

Un objectif classique est de chercher à estimer la matrice A avec l'a priori que son rang est faible [An51, Iz75, RV98]. Remarquons que l'on a

$$\begin{aligned} Y_{ia} &= \langle A^T x_i, e_a \rangle + \sigma E_{ia} \\ &= \underbrace{\langle x_i e_a^T, A \rangle}_{=Z_{ia}} + \sigma E_{ia} \end{aligned}$$

donc le modèle de régression multivarié est un cas particulier du modèle de régression trace

$$w_j = \langle Z_j, A \rangle + \sigma \xi_j, \quad j = 1, \dots, N, \tag{1.10}$$

étudiés en détails dans [Ba08, NW09, RT11b, KLT10] (voir aussi [LMY10, YELM07]). L'estimateur de type " ℓ^1 " proposé pour la régression trace (1.10) s'écrit dans le cas particulier de la régression multivariée sous la forme

$$\hat{A}_\lambda = \operatorname{argmin}_A \left\{ \|Y - XA\|^2 + 2\lambda \sum_k \sigma_k(A) \right\}$$

où $\|\cdot\|$ est la norme de Hilbert-Schmidt et $\sigma_1(A) \geq \sigma_2(A) \geq \dots$ est la suite des valeurs singulières de A . Les bornes obtenues (par exemple dans [KLT10]) pour la régression trace offrent donc

un contrôle sur l'estimateur \widehat{A}_λ . Cependant, ces résultats sont basés sur une condition "RIP" qui dans le cas multivarié requiert que $X^T X$ est inversible et donc $n \geq p$. Dans la courte note [CN1] nous montrons que cette condition "RIP" peut être relâchée dans le cas de la régression multivariée en la condition suivante (compatible avec $n < p$) :

pseudo-RIP :

$$1 \leq \frac{\sigma_1(X)}{\sigma_q(X)} \leq \eta < +\infty, \quad \text{avec } q = \text{rang}(X).$$

Sous cette condition, une adaptation immédiate des résultats de Koltchinskii *et al.* [KLT10] garantit le résultat suivant. Si les entrées de la matrice E sont i.i.d. de loi gaussienne $\mathcal{N}(0, \sigma^2)$, pour le choix $\lambda = K\sigma_{\max}(X)(\sqrt{T} + \sqrt{q})\sigma$ avec $K > 1$, on a

$$\begin{aligned} \|X\widehat{A}_\lambda - XA\|^2 &\leq \inf_{B \in \mathbb{R}^{p \times T}} \left\{ \|XB - XA\|^2 + 6K^2\eta^2 \left(\sqrt{T} + \sqrt{q}\right)^2 \text{rang}(B) \sigma^2 \right\} \\ &= \min_{r \leq \min(T, q)} \left\{ \sum_{k \geq r+1} \sigma_k(XA)^2 + 6K^2\eta^2 \left(\sqrt{T} + \sqrt{q}\right)^2 r \sigma^2 \right\}, \end{aligned}$$

avec probabilité supérieure à $1 - e^{-(K-1)^2(T+q)/2}$ (voir [CN1]). Le premier terme du membre de droite de l'inégalité correspond au biais minimal lorsque l'on cherche à approximer XA par XB avec B de rang inférieur à r . Le second terme est du même ordre de grandeur que la vitesse minimax pour estimer XA avec rang de A inférieur à r . Cette borne possède donc les bons ordres de grandeurs tant que le rapport $\eta = \sigma_1(X)/\sigma_q(X)$ reste de taille raisonnable.

Dans le cadre de la régression multivariée, il est possible de s'affranchir de l'estimateur \widehat{A}_λ et de travailler directement avec les estimateurs

$$\widehat{A}_r = \underset{A : \text{rang}(A) \leq r}{\text{argmin}} \|Y - XA\|^2, \quad r = 1, \dots, \min(q, T).$$

En effet, ceux-ci peuvent se calculer numériquement à l'aide d'une simple décomposition en valeurs singulières [RV98]. Dans un cadre non-asymptotique à variance connue, le problème de la sélection du rang r en minimisant un critère de la forme

$$\text{crit}_{\sigma^2}(r) = \|Y - X\widehat{A}_r\|^2 + \text{pen}(r)\sigma^2, \quad (1.11)$$

a été analysé récemment par Bunea *et al.* [BSW11]. Dans le cadre où les entrées de la matrice E sont i.i.d. de loi gaussienne $\mathcal{N}(0, \sigma^2)$, Bunea *et al.* proposent une pénalité de la forme $\text{pen}(r) = \lambda \left(\sqrt{q} + \sqrt{T}\right)^2 r$, avec $\lambda > 1$. Pour cette pénalité, les auteurs démontrent une inégalité oracle pour le risque $R(\widehat{A}) = \mathbb{E} \left[\|X\widehat{A} - XA\|^2 \right]$ et ils obtiennent des bornes sur la probabilité d'estimer correctement le rang avec \widehat{r} . L'une des qualités de ces résultats, c'est qu'ils ne nécessitent *aucune hypothèse* sur le design X . Dans le cas où la variance σ^2 est inconnue et le rang de X est inférieur à n , Bunea *et al.* proposent de remplacer σ^2 dans (1.11) par

$$\widehat{\sigma}^2 = \frac{\|Y - \Pi_X Y\|^2}{T(n - q)},$$

où Π_X est la projecteur orthogonal sur l'image de X . Le cas de figure où le rang de X est inférieur à n est hélas peu probable lorsque le nombre p de covariables est supérieur à la taille n de l'échantillon.

Dans l'article [A6], nous explorons en détail le problème de la sélection du rang r dans un cadre gaussien à variance inconnue, sans hypothèse sur le rang de X . Nous considérons un critère de sélection de la forme

$$\text{crit}(r) = \log \left(\|Y - X\hat{A}_r\|^2 \right) + \text{pen}(r), \quad (1.12)$$

pour lequel nous calculons une forme de pénalité minimale. Soit $q = \text{rang}(X)$ et $G_{q \times T}$ une matrice aléatoire $q \times T$ d'entrées i.i.d. de loi normale $\mathcal{N}(0, 1)$. En notant $\mathcal{S}_{q \times T}(r)$ l'espérance de la $(2, r)$ -norme de Ky-Fan de $G_{q \times T}$, c'est à dire

$$\mathcal{S}_{q \times T}(r) = \mathbb{E} [\|G_{q \times T}\|_{(2,r)}], \quad \text{avec} \quad \|G_{q \times T}\|_{(2,r)}^2 = \sum_{k=1}^r \sigma_k(G_{q \times T})^2,$$

nous montrons dans l'article [A6] que la pénalité

$$\text{pen}(r) = -\log \left(1 - \lambda \frac{\mathcal{S}_{q \times T}(r)^2}{nT - 1} \right)$$

avec $\lambda = 1$ est "minimale" pour le critère (1.12) dans le sens suivant. Pour $\lambda > 1$ l'estimateur sélectionné $\hat{A}_{\hat{r}}$ satisfait une borne de type oracle, alors que pour $\lambda < 1$ le rang \hat{r} sélectionné est du même ordre de grandeur que $\min(q, T)$ lorsque $A = 0$ (et donc $\text{rang}(A) = 0$). Nous montrons aussi que dans le cadre à variance connue, la pénalité

$$\text{pen}(r) = \lambda \mathcal{S}_{q \times T}(r)^2$$

avec $\lambda = 1$ est minimale pour le critère (1.11). Les pénalités $\text{pen}(r) = (\sqrt{T} + \sqrt{q})^2 r$ et $\text{pen}(r) = \mathcal{S}_{q \times T}(r)^2$ sont représentées Figure 1.3.

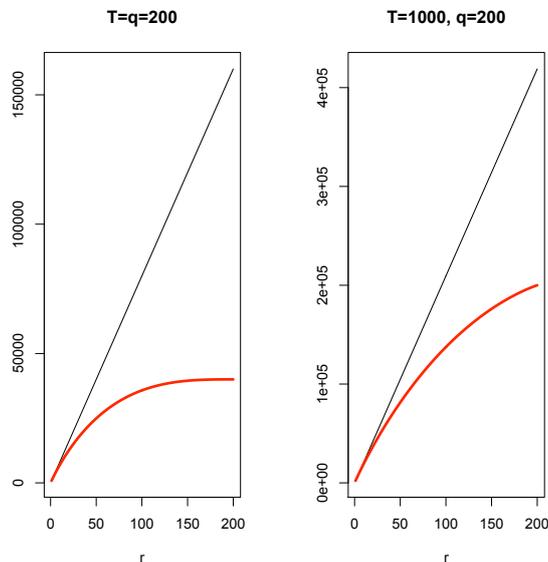


FIG. 1.3 – Pénalité $\text{pen}(r) = (\sqrt{T} + \sqrt{q})^2 r$ (en noir) et $\text{pen}(r) = \mathcal{S}_{q \times T}(r)^2$ (en rouge épais).

D'un point de vue pratique, la quantité $\mathcal{S}_{q \times T}(r)$ peut être évaluée efficacement par Monte Carlo lorsque q ou T est petit. Lorsque q et T sont grands, $\mathcal{S}_{q \times T}(r)$ peut être approximé précisément à l'aide de la loi de Marchenko-Pastur [MP67]. La procédure de sélection d'estimateur est mise en oeuvre dans le code [C2].

Enfin, lorsque les entrées de la matrice d'erreur E sont i.i.d. de loi \mathbf{P}_E sous-gaussienne, les mêmes résultats ont lieu en remplaçant $\mathcal{S}_{q \times T}(r)$ par

$$\mathcal{S}_X(r) = \mathbb{E} \left[\|\Pi_X G\|_{(2,r)} \right],$$

où Π_X est le projecteur sur l'image de X et G est une matrice aléatoire $n \times T$ d'entrées i.i.d. de loi \mathbf{P}_E .

1.4 Quelques perspectives

Parmi les sujets abordés dans ce chapitre, deux points mériteraient une plus grande attention. Le premier est l'analyse de modèles de régression multivariée plus complexes, le second est le mélange d'estimateurs arbitraires dans un esprit similaire à [A4].

Régression multivariée

Le modèle de régression multivariée analysé dans [A6] suppose que les composantes du bruit ε_i sont indépendantes et de même variance. Un modèle plus réaliste en pratique, serait de supposer que les ε_i sont i.i.d. de loi gaussienne $\mathcal{N}(0, \Sigma)$. Lorsque la matrice Σ est connue, l'analyse du modèle ne pose pas de difficultés. Par contre, l'analyse est beaucoup plus complexe dans le cas où Σ est inconnue. Un premier pas serait d'analyser le cas où la matrice Σ est diagonale (et inconnue). Dans ce cas, on ne sait pas calculer efficacement l'estimateur \hat{A}_r obtenu en maximisant la vraisemblance conjointement en Σ et A avec la contrainte $\text{rang}(A) \leq r$. On peut chercher à relâcher ce problème d'optimisation en le convexifiant selon le schéma suivant. La première étape est de reparamétriser le problème en $D = \Sigma^{-1/2}$ et $B = A\Sigma^{-1/2}$ afin de rendre la log-vraisemblance concave. La seconde étape est de convexifier la contrainte $\text{rang}(A) \leq r$ en imposant à la place une contrainte sur la norme nucléaire de B . Le résultat de cette procédure est le problème d'optimisation convexe

$$(\hat{B}_\lambda, \hat{D}_\lambda) \in \operatorname{argmin}_{B,D} \left\{ -n \log(|D|) + \frac{1}{2} \|YD - XB\|^2 + \lambda \sum_k \sigma_k(B) \right\}.$$

Il serait intéressant d'analyser l'estimateur résultant $\hat{A}_\lambda = \hat{B}_\lambda \hat{D}_\lambda^{-1}$ avec les outils développés pour la régression trace. La difficulté réside dans le contrôle de \hat{D}_λ .

Une autre classe de modèles de régression multivariée mériterait une plus grande attention. Dans un cadre où on chercherait à prédire des phénotypes à partir de niveaux d'expression de gènes et / ou d'abondances de protéines, une hypothèse naturelle est de supposer que la matrice A est à la fois "creuse" (beaucoup de zéros) et de faible rang. Une analyse de la complexité de la collection de modèles utile pour estimer une matrice A creuse et de faible rang (sans connaître ni le rang, ni les zéros de A), laisse présupposer qu'il est possible dans certains cas d'estimer A à une vitesse plus rapide que si on suppose seulement "A creuse" ou "A de faible rang". La collection de modèles nécessaire étant de très grande taille, il n'est pas envisageable de mettre en oeuvre une telle approche en pratique. Une nouvelle fois, une possibilité est de convexifier le problème en résolvant

$$\hat{A}_\lambda \in \operatorname{argmin}_A \left\{ \|Y - XA\|^2 + \lambda |A|_{\ell^1} + \mu \sum_k \sigma_k(A) \right\}. \quad (1.13)$$

Gaiffas et Lecué [GL10] se sont intéressés à des estimateurs de ce type pour la régression trace. Leurs résultats (difficiles) ne fournissent que des vitesses "lentes". Pour l'estimateur (1.13), il est

possible d'obtenir des vitesses plus rapides, mais celles-ci ne montrent pas le bénéfice de la double pénalisation ℓ^1 et nucléaire. Beaucoup de chemin reste à parcourir pour mieux comprendre ce cadre.

Mélange d'estimateurs arbitraires

De nombreux résultats ont été obtenus pour mélanger des estimateurs arbitraires. Ces procédures sont basées sur deux jeux de données : un jeu pour calculer les estimateurs \hat{f}_λ , un autre jeu pour calculer les poids ω_λ . Quelques travaux [LB06, RT11a] et [A9] réalisent un mélange en se basant sur le même jeu de données à la fois pour calculer les estimateurs \hat{f}_λ et les poids ω_λ . Cependant, les seuls estimateurs considérés dans ces articles sont les estimateurs par projection $\hat{f}_\lambda = \Pi_{S_\lambda} Y$ (à noter néanmoins l'existence de résultats dans la thèse de G. Leung pour des estimateurs de James-Stein). Lorsque l'objectif est la prédiction, il serait intéressant de développer des procédures performantes de mélange

- (i) ne demandant pas la connaissance de la variance σ^2 ,
- (ii) permettant de mélanger des estimateurs arbitraires,

dans un esprit semblable à [A4].

Dans un premier temps, il est raisonnable de considérer le cas simple des estimateurs linéaires $\hat{f}_\lambda = A_\lambda Y$ avec $A_\lambda = \sum_i \sigma_i(\lambda) v_i v_i^T$ et les $\sigma_i(\lambda) \geq 0$. Lorsque la variance σ^2 est connue, une analyse très similaire à [LB06] permet d'obtenir une borne de risque fine pour l'estimateur $\hat{f} = \sum_{\lambda \in \Lambda} \omega_\lambda \hat{f}_\lambda$ avec $\omega_\lambda \propto \pi_\lambda e^{-\hat{r}_\lambda / (8\sigma^2)}$ où

$$\hat{r}_\lambda = \|Y - \hat{f}_\lambda\|^2 + 2\sigma^2 \text{Tr}(A_\lambda), \quad \text{et } \pi_\lambda = e^{-\alpha \text{Tr}(A_\lambda^T A_\lambda)}, \quad \text{avec } \alpha > 0 \text{ tel que } \sum_{\lambda \in \Lambda} e^{-\alpha \text{Tr}(A_\lambda^T A_\lambda)} = 1.$$

Cet estimateur \hat{f} satisfait la borne de risque

$$\begin{aligned} \mathbb{E} \left[\|\hat{f} - f\|^2 \right] &\leq \min_{\lambda \in \Lambda} \left[\mathbb{E} \left[\|\hat{f}_\lambda - f\|^2 \right] + 8\alpha \text{Tr}(A_\lambda^T A_\lambda) \sigma^2 \right] \\ &\leq (1 + 8\alpha) \min_{\lambda \in \Lambda} \mathbb{E} \left[\|\hat{f}_\lambda - f\|^2 \right]. \end{aligned}$$

Dans le cadre d'estimateurs linéaires quelconques, il n'est pas possible de mettre en oeuvre une stratégie similaire à [A9] pour le cas où la variance est inconnue. Des bornes de risque peuvent être obtenues pour des formes de poids un peu différentes (avec σ^2 inconnue), mais ces bornes donnent des contrôles moins bons que la sélection d'estimateurs avec [A4].

2 Statistiques pour la biologie

2.1 Inférence de réseaux de régulation

La biologie des systèmes est une branche émergente de la biologie, dont l'objet est l'étude d'une entité biologique (cellule, organe, organisme) dans son ensemble, avec une attention spéciale pour ses mécanismes de régulation. Il s'agit moins d'identifier précisément comment s'opèrent les régulations (mécanismes biochimiques complexes), que de chercher à comprendre les propriétés émergentes du système (stabilité, résilience, etc).

L'une des grandes difficultés à laquelle se heurte la biologie des systèmes est l'identification des liens de régulation existant au sein du génome / protéome / métabolome, y compris pour les organismes modèles tels *E. coli*, *B. subtilis*, *A. thaliana*, etc. En effet, il est matériellement impossible de tester expérimentalement tous les liens de régulation possibles entre les 25 à 30 000 gènes de *A. thaliana*. L'analyse statistique offre une approche intéressante pour identifier des liens "possibles" ou "probables" entre différents gènes. Les techniques actuelles des biopuces ou de protéomique LC-MS/MS permettent de mesurer simultanément l'abondance d'un grand nombre d'ARNm ou de peptides. À partir de ces mesures de niveau d'expression des gènes ou d'abondance de protéines, il est possible de conduire deux types d'analyse statistique.

La première est une analyse *différentielle* des niveaux d'abondances mesurés dans deux conditions expérimentales différentes. Ce type d'analyse permet d'identifier de façon fiable des gènes / protéines intervenant dans la réponse à un stress (physiologique, chimique, biotique, etc). Cependant, elle demande la réalisation coûteuse d'un très grand nombre d'expériences et ne permet d'identifier qu'un petit nombre de liens à la fois. Une seconde approche consiste à effectuer une analyse des dépendances statistiques *globales* entre les niveaux d'abondances mesurés. Le principe est que le réseau des dépendances conditionnelles entre les abondances reflèterait, au moins en partie, le réseau de régulation biologique. Dans cette optique diverses approches ont été développées. Certaines approches [MV08, Vert10] se basent sur la connaissance de certains liens pour chercher à identifier les autres : elles se situent donc dans le domaine de l'apprentissage *supervisé*. D'autres approches basées sur les modèles graphiques sont *non-supervisées*.

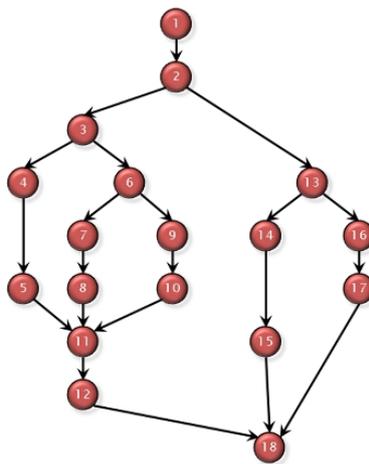


FIG. 1.4 – Exemple de graphe acyclique orienté.

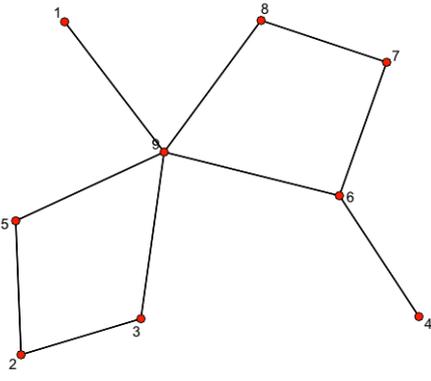
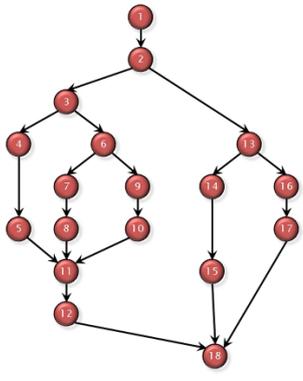
Supposons qu'un *graphe acyclique orienté* \vec{g} , comme à la Figure 1.4, représente le réseau de régulation entre p variables X_1, \dots, X_p . Les *modèles graphiques orientés* fournissent alors un cadre statistique naturel pour modéliser la loi du vecteur $X = (X_1, \dots, X_p)$. En effet, dans un tel

modèle, une variable X_a est indépendante des variables "non-descendantes" conditionnellement aux variables "parents directs". Par exemple dans le cas de la Figure 1.4, conditionnellement à $\{X_5, X_8, X_{10}\}$ la variable X_{11} est indépendante des variables autres que X_{12} et X_{18} . Si on modélise les variables observées par un modèle graphique orienté, le problème statistique devient alors d'estimer à partir de n réalisations du vecteur X le graphe \vec{g} . Ce problème est malheureusement mal posé. En effet, il existe en général plusieurs graphes orientés minimaux compatibles avec la loi de X . Par exemple, dans le simple modèle autorégressif d'ordre 1, AR(1), les deux graphes

$$1 \rightarrow 2 \rightarrow \dots \rightarrow p \quad \text{et} \quad 1 \leftarrow 2 \leftarrow \dots \leftarrow p,$$

sont des graphes minimaux compatibles avec la loi de X . Maathuis *et al.* [MKB09, KMCMB11] proposent de contourner le problème en estimant les différents graphes orientés compatibles avec la loi de X , puis en donnant une mesure d'importance à chaque arrête correspondant à l'effet d'intervention minimal parmi tous les graphes obtenus.

Une alternative aux modèles graphiques orientés est la *modélisation graphique non-orientée*, présentée section 1.2. Dans ce cas, le graphe minimal est unique et le problème est bien posé. En plus, il n'y a pas d'hypothèse d'acyclicité du réseau de régulation, évitant d'exclure tous les phénomènes de rétroaction. Mentionnons que si la loi de X est un modèle graphique orienté par rapport au graphe acyclique \vec{g} , alors c'est aussi un modèle graphique non-orienté par rapport au graphe moral de \vec{g} (graphe non orienté obtenu en retirant les "flèches" aux arêtes de \vec{g} et en joignant les noeuds de \vec{g} pointant vers un même noeud).

Modèles non-orientés	Modèles orientés
 <p data-bbox="255 1624 774 1713">X_a indépendant de $\{X_b : b \overset{\vec{g}}{\approx} a\}$ sachant $\{X_b : b \overset{\vec{g}}{\approx} a\}$</p>	 <p data-bbox="821 1624 1332 1713">X_a indépendant de $\{X_b : a \overset{\vec{g}}{\rightarrow} \dots \overset{\vec{g}}{\rightarrow} b\}$ sachant $\{X_b : b \overset{\vec{g}}{\rightarrow} a\}$</p>

En collaboration avec Sylvie Huet et Nicolas Verzelen, nous avons développé une procédure statistique [A5] et un package R [C1], dédiés à l'estimation dans le cadre des modèles graphiques gaussiens non-orientés. A côté des aspects théoriques évoqués section 1.2, une grosse partie du travail a été de tester sur des jeux de données simulées un grand nombre de méthodes pour

explorer l'espace des graphes. Nous avons retenu les plus performantes et nous décrivons dans [A5] leur comportement typique lors de nos expérimentations. Enfin, en collaboration avec Annie Bouvier nous avons optimisé le code [C1] en C et FORTRAN là où c'était nécessaire. De nouvelles fonctionnalités (actuellement uniquement disponibles sur demande) seront intégrées lors de la prochaine mise à jour du package.

A titre d'illustration, la Figure 1.5 donne le graphe sélectionné par GGMselect à partir des données d'expression de gènes de Hess *et al.* [He06]. L'analyse est effectuée sur 26 gènes présentant un fort pouvoir prédictif pour la guérison du cancer du sein après chimiothérapie. La population étudiée est celle des 99 patients non guéris (non-PCR) après traitement.

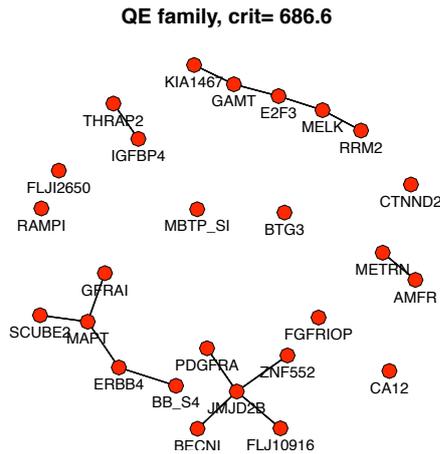


FIG. 1.5 – Graphe sélectionné par GGMselect pour 26 gènes de patients non-PCR.

2.2 Analyse des déformations anatomiques du cerveau

Les structures de connectivité du cerveau portent une information importante sur son fonctionnement et sur les relations anatomiques (existence de fibres neuronales) entre les différentes zones fonctionnelles. Il est en particulier intéressant de chercher des régions qui ne sont pas reliées directement par une fibre mais dont les comportements sont corrélés. L'existence ou le manque de certains de ces liens peut s'avérer être un marqueur d'une pathologie. La structure de dépendance statistique entre les déformations de diverses régions anatomiques du cerveau par rapport à un cerveau type reflèterait une partie de ces structures de connectivité. Selon ce paradigme, l'analyse statistique permet donc d'explorer les grandes structures du réseau de connexion du cerveau [Ki10, LCKKL06].

Les modèles graphiques gaussiens fournissent un cadre naturel pour explorer les dépendances conditionnelles entre divers points du cerveau. Cependant, contrairement au cadre de la section 1.2, le graphe \mathbf{g}_{\min} sous-jacent n'a aucune raison d'être de faible degré car chaque point est fortement connecté à son entourage immédiat. Nous avons donc *a priori* qu'il existe d'une part de multiples connexions à faible distance (qui ne nous intéressent pas) et d'autre part quelques connexions longue distance que l'on cherche à estimer. Dans l'article [A3], écrit en collaboration avec Stéphanie Allasonnière, nous développons une procédure d'estimation dédiée à ce contexte. Nous partons de *a priori* que le graphe \mathbf{g}_{\min} contient un graphe \mathbf{g}_0 liant chaque point à ses voisins et nous explorons l'existence d'arêtes en dehors de \mathbf{g}_0 . L'exploration de ces

arêtes est basée sur l'algorithme Elastic Net [ZH05] et la sélection finale des arêtes s'effectue à l'aide d'une version modifiée du critère développé dans [A8].

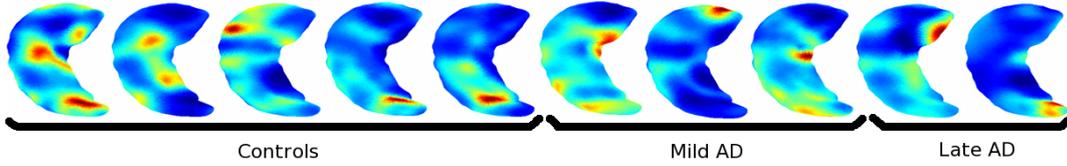


FIG. 1.6 – Représentation visuelle des log-jacobiens de déformation.

Cette procédure a été mise en oeuvre pour explorer des déformations de la surface de l'hippocampe. Les données, gracieusement transmises par l'Université John Hopkins [Mi09], fournissent le log du Jacobien de la déformation de l'hippocampe de 101 sujets par rapport à un hippocampe "modèle", voir Figure 1.6. Parmi les 101 sujets, 44 souffrent de la maladie d'Alzheimer et 57 sont des patients sains (contrôle). Nous constatons que les patients atteints par la maladie d'Alzheimer présentent une connectivité statistique plus faible que les patients sains, mais il faudra une plus grande base de données pour éventuellement confirmer cette tendance.

2.3 Délimitation de populations synchrones

La dynamique des populations d'oiseaux est très dépendante du climat : des populations soumises aux mêmes contraintes climatiques auront tendance à fluctuer de manière synchrone. Ce forçage climatique peut s'opérer à des échelles spatiales très variées selon les composantes du climat et le cycle de vie. D'autres forçages peuvent aussi s'opérer (fluctuation des ressources, synchronisation par échanges d'individus, etc), mais ces phénomènes sont probablement de faible ampleur chez les oiseaux. Délimiter les populations synchrones apporterait donc beaucoup à la compréhension des échelles spatiales auxquelles s'opère le forçage climatique et contribuerait à mieux comprendre les liens que l'on peut faire entre climat et dynamique de populations.

Le Muséum National d'Histoire Naturelle dispose d'une vaste base de données STOC recueillies par des ornithologues amateurs selon un processus très strict. Ces données ont été collectées une fois par an depuis 2001, sur environ 1700 sites. Les comptages sont ainsi disponibles pour de nombreux sites d'observation, mais les séries temporelles sont courtes (au plus 9 années) et très incomplètes. En effet, sur la période 2001-2009, seuls 361 sites disposent d'au moins 7 années d'observations.

Dans l'article [A1], écrit en collaboration avec Emmanuelle Porcher et Romain Julliard, nous nous intéressons au problème de délimiter des populations synchrones à partir des données STOC. Nous avons en tête que chaque site s possède sa propre abondance caractéristique mais qu'il existe une synchronie au niveau des variations temporelles de ces abondances. Autrement dit, on s'attend à ce qu'il existe une partition des sites en régions R_1, R_2, \dots , telle que tous les sites d'une même région R ont une dynamique temporelle (quasi) identique $t \rightarrow \rho_R(t)$. En notant x_s la localisation du site s et Z_{st} l'observation au site s l'année t , on cherche donc à ajuster le modèle

$$Z_{st} \sim \text{Poisson}(\exp(\theta_s + f(x_s, t))) \quad \text{avec } f(x, t) = \sum_R \rho_R(t) \mathbf{1}_R(x) \quad \text{et } f(., 1) = 0, \quad (1.14)$$

les régions R et les paramètres $\theta_s, \rho_R(t)$ étant inconnus. Chaque région R délimite une population synchrone dont la dynamique est donnée par $t \rightarrow \rho_R(t)$. Le paramètre $\exp(\theta_s)$ est proportionnel à l'abondance sur le site s (qui dépend des caractéristiques du site), θ_s représente donc un "effet

site". Il n'est pas possible d'ajuster directement le modèle (1.14) par maximum de vraisemblance car la complexité de l'optimisation en les régions R_1, \dots, R_k est beaucoup trop importante. Nous choisissons donc de relâcher le problème d'optimisation en le convexifiant. Nous introduisons sur l'espace des fonctions $f : \mathcal{D} \times \{1, \dots, T\} \rightarrow \mathbb{R}$ une norme $\text{STV}(f)$ qui favorise les fonctions de la forme $f(x, t) = \sum_R \rho_R(t) \mathbf{1}_R(x)$. Cette norme est définie pour les fonctions f dans $L^1_{\text{loc}}(\mathcal{D} \times \{1, \dots, T\})$ par

$$\text{STV}(f) = \sup \left\{ -\sum_t \int_{\mathcal{D}} f(x, t) \text{div}_x(\phi(x, t)) dx : \phi(\cdot, t) \in C_c^\infty(\mathcal{D}, \mathbb{R}^d) \text{ and } \left\| \sum_t \|\phi(\cdot, t)\| \right\|_\infty \leq 1 \right\}.$$

Lorsque la fonction f est C^1 en la variable x la norme $\text{STV}(f)$ prend la forme simple

$$\text{STV}(f) = \int_{\mathcal{D}} \max_t \|\nabla_x f(x, t)\| dx.$$

L'estimation s'effectue alors en maximisant la log-vraisemblance moins une pénalité, cette pénalité étant proportionnelle à une version discrétisée de $\text{STV}(f)$. Nous obtenons ainsi un problème d'optimisation convexe qui peut être attaqué numériquement à l'aide d'un algorithme primal-dual, dont la convergence est assurée par le Théorème 3.14 de Chambolle *et al.* [Ch10]. Cet algorithme [C3], programmé en FORTRAN avec une interface en R, a été mis à disposition du Muséum National d'Histoire Naturelle.

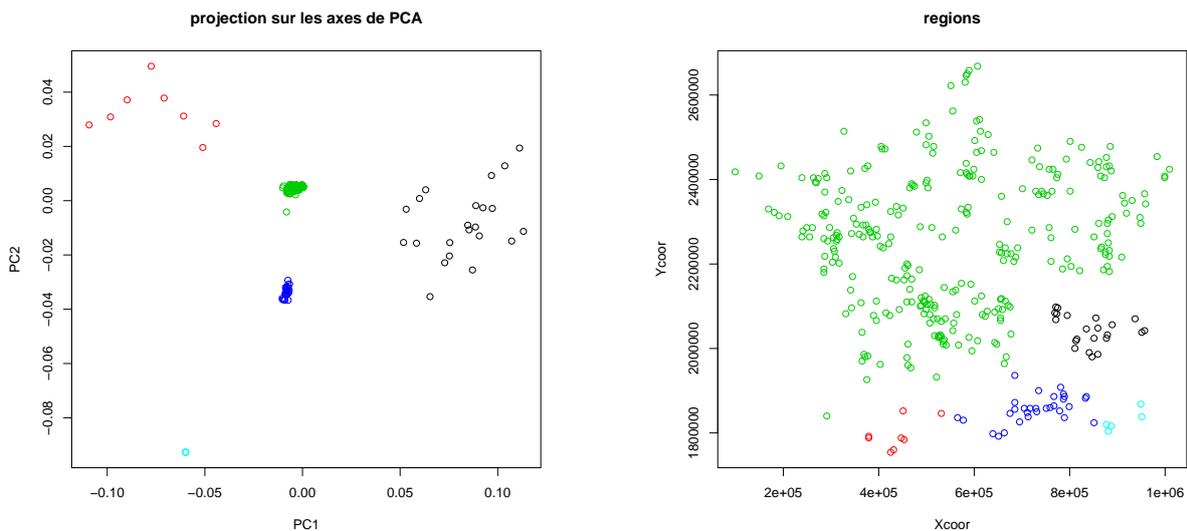


FIG. 1.7 – Gauche : ACP de \hat{f} et segmentation par *kmeans*. Droite : régions obtenues.

La Figure 1.7 représente les régions obtenues en réalisant une segmentation par *kmeans* du \hat{f} estimé pour les données STOC. Les 5 régions étant délimitées, la dynamique de chacune des régions est estimée par maximum de vraisemblance. Ces dynamiques sont représentées Figure 1.8. Les régions estimées ont du sens d'un point de vue écologique, mais nous restons prudents dans l'interprétation des résultats. En effet, les sites sont distribués de façon très hétérogène et les frontières entre les régions estimées ont tendance à passer à travers les "déserts" d'observation. Il est donc difficile de conclure si nous apprenons des limites de populations synchrones ou simplement la topologie des sites d'observation.

Cela nous a conduit à explorer la fiabilité de la procédure d'estimation sur des données simulées. Il ressort que l'algorithme détecte plutôt bien les régions périphériques, mais a du mal

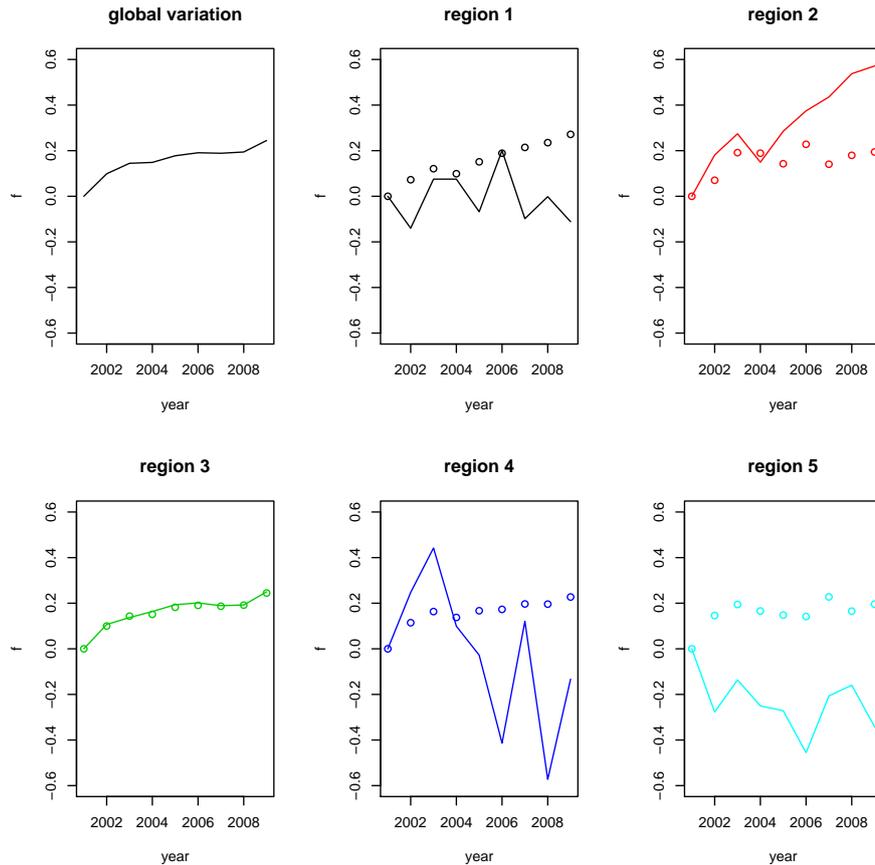


FIG. 1.8 – Dynamique temporelle pour chaque région. Continu : modèle (1.14) ajusté avec les régions R de la Figure 1.7. Pointillés : $mean(\hat{f}[R == r,])$.

à détecter les régions "centrales" (régions ayant des limites ne touchant pas les frontières du domaine). Cette tendance s'explique facilement par le fait que la norme STV pénalise proportionnellement au périmètre d'une région. Il est donc beaucoup plus coûteux en terme de la norme STV d'ajouter une région centrale qu'une région périphérique.

2.4 Quelques perspectives

Le champ des problèmes statistiques posés par le développement soutenu des biotechnologies est extrêmement vaste. Les caractéristiques usuelles des données récoltées actuellement sont

- des jeux de données massifs mais présentant relativement peu de répétitions,
- un bruit biologique et/ou technique important et une proportion non négligeable de données manquantes,
- des données hétérogènes apportant des informations de nature différentes.

Parallèlement, les questionnements biologiques restent bien souvent de nature assez qualitative. Par exemple, l'intérêt se portera en général plus sur la compréhension de la structure des interactions entre différentes entités biologiques, que sur la prédiction d'une (ou plusieurs) quantité(s) d'intérêt. Les problèmes statistiques sont donc plus souvent de la catégorie des problèmes inverses, que de celle des problèmes de prédiction, à l'exception notable des problèmes de diagnostic médical. Cet ensemble de caractéristiques engendre de réelles difficultés d'un point de vue statistique. Il apparaît clairement que (dans le meilleur des cas) seules des procédures statis-

tiques performantes et exploitant pleinement les structures du problème permettent d'apporter des éléments de réponses aux questions biologiques actuelles. Pour éviter une inflation démesurée de cette section, nous allons nous circonscrire à quelques perspectives liées aux travaux évoqués précédemment. En particulier, nous n'allons pas évoquer toutes les perspectives liées aux projets structurants auxquels nous participons.

Concernant l'estimation non-supervisée de réseaux de gènes, de nombreuses approches ont été proposées, mais leur pertinence reste à démontrer. En effet, nous sommes actuellement à un stade où il faudrait pouvoir tester les différentes méthodes développées sur des données relatives à un ensemble de gènes dont on connaîtrait *parfaitement* la structure du réseau de régulation (les seuls tests effectués aujourd'hui sont sur des données synthétiques). De telles données pourraient être accessibles très prochainement sur *B. subtilis*. Un tel retour expérimental permettra de mieux cerner les forces et faiblesses de chacune des approches et les prochains axes à développer.

Concernant la délimitation de populations synchrones, deux points sont à approfondir. Le premier point est de parvenir à contourner l'hétérogénéité spatiale des sites d'observation. Une approche possible serait de déplacer virtuellement les sites afin d'obtenir une répartition plus homogène. Bien sûr, il faut parvenir à préserver un maximum de propriétés topologiques et l'idéal serait d'avoir des déformations très locales de l'espace. Des approches par particules en interaction, ou par transport optimal seraient à envisager. Le second point à approfondir concerne la sélection du nombre de régions à conserver. Le chemin de régularisation complet peut apporter une information écologiquement intéressante, il faut cependant veiller à ce que les motifs obtenus reflètent autre chose que du bruit. Une analyse théorique du modèle semble difficile, cependant une approche empirique par des tests d'hypothèses ou par validation-croisée est envisageable.

Enfin, signalons une famille de problèmes méritant une attention pratique et théorique. Dans de nombreux cas, l'objectif est de comprendre le lien entre un nombre modéré de mesures et un grand nombre de covariables, en ayant à disposition un échantillon de taille relativement limitée. A titre d'exemple, il peut s'agir de relier des phénotypes à des abondances de protéines, des niveaux d'expressions de gènes à des *DNA copy number*, des scores cognitifs à des images IRMf, etc. Les modèles de régression multivariée peuvent alors s'avérer utiles pour explorer ces liens. Selon le problème considéré, diverses hypothèses peuvent être envisagées : régression linéaire avec matrice "creuse" et de faible rang, modèles à effets mixtes, inverse creux de la matrice de covariance du bruit, régression non linéaire sparse, etc. Développer des procédures statistiques efficaces et des critères de sélection performants dans ces contextes apporterait des outils très utiles pour l'analyse de données biologiques.

3 Etude de quelques processus stochastiques

3.1 Turbulence de Burgers et particules collantes

La turbulence de Burgers est un modèle simplifié de turbulence hydrodynamique introduit par Burgers [Bu48, Bu50, Bu74]. Ce modèle ne reflète que très partiellement la turbulence hydrodynamique, voir Kraichnan [Kr68] pour une analyse des similitudes et différences. Ce modèle apparaît cependant dans de nombreux domaines de physique mathématique, par exemple pour décrire la formation des grandes structures de l'univers [VDFN94], des phénomènes de sédimentation ou en acoustique [Wo98]. En dehors de la modélisation physique, les propriétés statistiques remarquables de la turbulence de Burgers suscitent une attention particulière. Par exemple, Bertoin [Be00] a mis à jour des liens étroits entre la turbulence de Burgers et la coalescence additive.

Contexte

L'équation de Burgers non visqueuse (aussi appelée équation de Riemann)

$$\partial_t u + u \partial_x u = 0, \quad x \in \mathbb{R}, \quad t \geq 0 \quad (1.15)$$

admet une unique solution faible entropique $u(x, t)$ pouvant être obtenue comme limite lorsque $\varepsilon \rightarrow 0$ de l'unique solution forte u^ε de l'équation avec viscosité

$$\partial_t u + u \partial_x u = \varepsilon \partial_{xx}^2 u.$$

Dès que la condition initiale $u(\cdot, 0)$ vérifie la condition

$$W(x) = \int_0^x u(z, 0) dz \stackrel{\pm\infty}{=} o(x^2) \quad (1.16)$$

la solution entropique $u(x, t)$ de (1.15) peut se mettre sous la forme

$$u(x, t) = \frac{x - a(x, t)}{t} \quad (1.17)$$

où $a(\cdot, t)$ est une fonction croissante. Les discontinuités de $u(\cdot, t)$ coïncident donc avec les sauts (positifs) de $a(\cdot, t)$ et sont appelées "chocs" de la solution.

Les solutions de l'équation de Burgers (1.15) sont intimement liées au modèle de particules collantes proposé par Zeldovich [Ze70]. Les particules collantes sont des particules ponctuelles évoluant librement entre deux collisions. Lorsque plusieurs particules se rencontrent, elles fusionnent en une seule particule avec conservation de la masse et du moment (et donc avec dissipation d'énergie). Cette évolution est une solution faible particulière du système

$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0 \\ \partial_t(\rho v) + \partial_x(\rho v^2) = 0 \end{cases} \quad (1.18)$$

où $\rho(\cdot, t)$ représente la densité de masse au temps t et $v(\cdot, t)$ le champ de vitesse. La dynamique de particules collantes infinitésimales réparties au temps initial selon la mesure de Lebesgue sur \mathbb{R} est liée à l'équation de Burgers comme suit. Si on note $u(x, t) = (x - a(x, t))/t$ la solution entropique de (1.15) avec condition initiale $u(\cdot, 0) = v_0$ vérifiant (1.16), alors le couple

$$v(x, t) = \frac{u(x+, t) + u(x-, t)}{2} \quad \text{et} \quad \rho = \partial_x a(dx, t),$$

$(\partial_x a(dx, t))$ étant la dérivée de Stieljes de $a(., t)$ est solution faible de (1.18) avec condition initiale $\rho(dx, 0) = dx$ et $v(., 0) = v_0$. La répartition au temps t des particules collantes est donc totalement décrite par la dérivée spatiale (au sens de Stieljes) de $a(x, t) = x - tu(x, t)$. En particulier, la présence d'un amas macroscopique dans le modèle de particules collantes correspond à la présence d'un choc dans la solution $u(x, t)$ de (1.15) : la vitesse et la position du choc coïncident avec celle de l'amas, et l'amplitude du choc est égale à la masse de l'amas divisé par t . Inversement, un point régulier de $u(., t)$ correspond à une particule restée isolée au temps t .

L'un des aspects remarquables de l'équation de Burgers non visqueuse (1.15) est l'existence d'une formule explicite pour la solution $u(x, t)$. En effet, dès que (1.16) est vérifiée, la solution u de (1.15) est de la forme (1.17) avec $a(x, t)$ donné par la célèbre formule de Hopf-Cole [Ho50, Co51]

$$a(x, t) = \operatorname{argmax}_a^+ \left\{ W(a) + \frac{1}{2t}(a - x)^2 \right\}$$

où argmax^+ représente le sup des a pour lequel le maximum est atteint, voir Figure 1.9. La

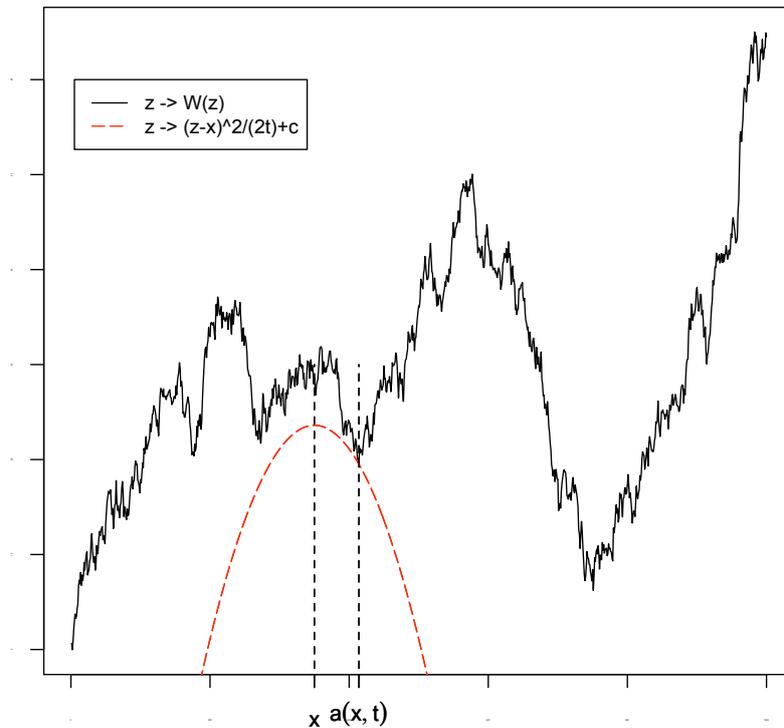


FIG. 1.9 – La fonction $x \rightarrow a(x, t)$.

solution $u(x, t)$ est alors entièrement caractérisée par "l'enveloppe t^{-1} -parabolique" de W . Cette enveloppe est définie comme suit. Rappelons que l'enveloppe convexe f_c d'une fonction f est donnée par

$$f_c(x) = \sup\{g(x) : g \text{ est linéaire et } g \leq f\}.$$

L'enveloppe α -parabolique de f est définie de la même manière mais en remplaçant la contrainte "g est linéaire" par "g(x) = $-\frac{\alpha}{2}(x - a)^2 + b$, avec $a, b \in \mathbb{R}$ ". La présence d'un choc en x dans la solution $u(\cdot, t)$ correspond alors à la présence dans l'enveloppe t^{-1} -parabolique de W d'une parabole $g(x) = -\frac{1}{2t}(x - a)^2 + b$ entrant en contact en au moins deux points avec W . La masse et la vitesse de l'amas de particules associé s'interprète géométriquement selon le schéma de la Figure 1.10.

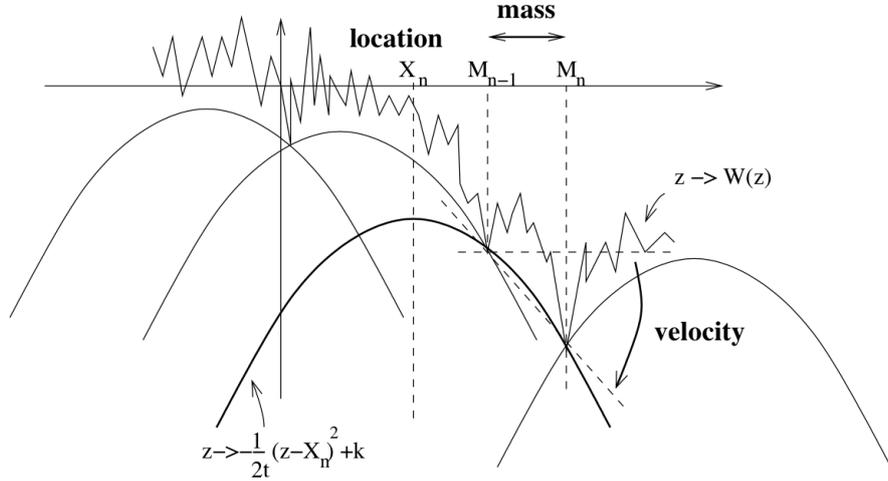


FIG. 1.10 – Interprétation géométrique de la masse et la vitesse d'un amas.

Condition initiale bruit blanc sur \mathbb{R}

L'étude des propriétés statistiques de la solution $u(x, t)$ de l'équation de Burgers (1.15) avec condition initiale bruit blanc a suscité de nombreux travaux [AE95, Av95, Bu74, Ry98a, Ry98b, SAF92]. Dans ce cas, $(W(x), x \in \mathbb{R})$ est un mouvement brownien sur \mathbb{R} et les propriétés de la solution $u(x, t)$ sont liées aux propriétés statistiques de l'enveloppe t^{-1} -parabolique d'un mouvement brownien, schématisée Figure 1.10. D'un point de vue qualitatif, la fonction $a(\cdot, t)$ est p.s. croissante en escalier, sans accumulation de sauts [Gr89, Av95]. Du point de vue des particules collantes, cela signifie qu'à tout instant $t > 0$ toutes les particules se sont agrégées en amas macroscopiques dont les positions forment une suite discrète de \mathbb{R} . La loi du système à un temps fixe $t > 0$ est intégralement décrite par la loi du processus de saut pur $\{a(x, t) = x - tu(x, t) : x \in \mathbb{R}\}$. Cette loi a été décrite indépendamment par Groeneboom [Gr89] et Frachebourg et Martin [FM00].

Dans les articles [A16] et [A11], nous nous sommes intéressés à l'évolution temporelle du système de particules lorsque le temps croît. La question principale est de comprendre quelle est la loi du système à un temps $s < t$, sachant l'état du système au temps t . L'évolution du système est régie par la dynamique totalement déterministe des particules collantes. Cette dynamique engendre une perte d'information au cours du temps : la filtration $\mathcal{F}_t = \sigma(a(x, t), x \in \mathbb{R})$ est strictement décroissante au sens de l'inclusion. Lorsque l'on retourne le sens d'écoulement du temps, on obtient un processus stochastique de fragmentation schématisé Figure 1.11. L'objet principal des articles [A16] et [A11] est de caractériser ce processus de fragmentation et donner une description du système à un temps fixe $s < t$ sachant l'état au temps t . Il faut donc comprendre comment évolue la fragmentation de l'enveloppe t^{-1} -parabolique lorsque t décroît. La première clef pour analyser ce processus est de remarquer que

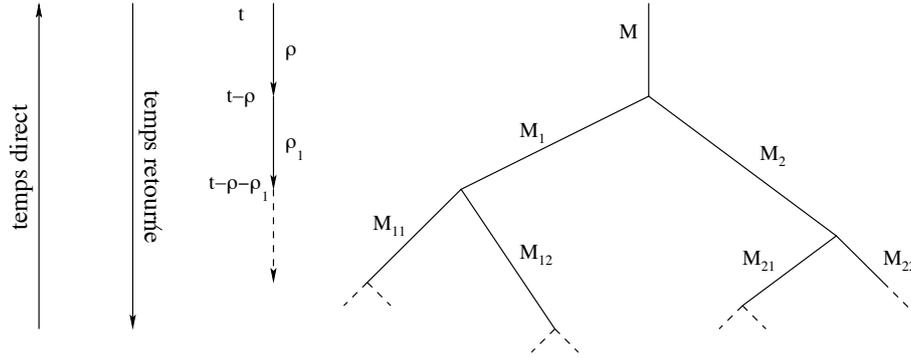


FIG. 1.11 – Fragmentation des amas.

- (i) les points de contact du brownien avec son enveloppe t^{-1} -parabolique sont des splitting times [Ge79] pour le mouvement brownien,
- (ii) les excursions du mouvement brownien au-dessus de son enveloppe parabolique sont indépendantes sachant \mathcal{F}_t , de loi conditionnelle décrite à partir de la loi d'une excursion brownienne normalisée conditionnée à rester au dessus d'une parabole $x \rightarrow \frac{1}{2t}x(1-x)$.

La seconde clef pour décrire ce processus de fragmentation est de calculer la loi du couple de variable aléatoire

$$\begin{cases} \sigma &= \min \left\{ \frac{2}{s(1-s)} e_s : s \in]0, 1[\right\} \\ \eta &= \operatorname{argmin}_s^+ \left\{ \frac{2}{s(1-s)} e_s : s \in]0, 1[\right\}, \end{cases}$$

où $(e_s : s \in [0, 1])$ est une excursion brownienne normalisée, voir Figure 1.12. La densité de cette

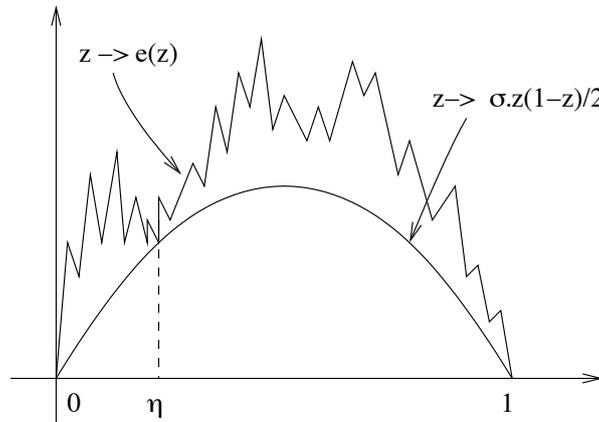


FIG. 1.12 – Les variables σ et η .

loi est calculée dans [A11] et vaut

$$\mathbb{P}(\sigma \in da, \eta \in dx) = \frac{e^{-a^2/24}}{\sqrt{8\pi x(1-x)}} C(ax^{3/2}) C(a(1-x)^{3/2}) dadx$$

où

$$C(\lambda) := \mathbb{E} \left(\exp \left(- \int_0^1 e_s ds \right) \right) = \lambda \sqrt{2\pi} \sum_{n=1}^{\infty} \exp \left(-2^{-1/3} \omega_n \lambda^{2/3} \right), \quad \text{pour } \lambda > 0,$$

avec $0 < \omega_1 < \omega_2 < \dots$ la suite des zéros de la première fonction de Airy Ai (cf [AS64] p. 446). A noter la remarquable expression pour la loi de σ (voir [A16]) :

$$\mathbb{P}(\sigma \geq a) = e^{-a^2/24} C(a).$$

A partir de ces lois, l'article [A11] décrit le mécanisme de fragmentation d'un amas, et donne une description complète de la loi de l'arbre généalogique d'un amas.

Condition initiale bruit blanc asymétrique

Lorsque $u(\cdot, 0)$ est un bruit blanc asymétrique, c'est à dire

$$W(x) = \begin{cases} 0 & \text{si } x < 0 \\ \text{mouvement brownien} & \text{si } x \geq 0, \end{cases}$$

le chaos présent à droite de l'origine se propage au cours du temps vers la gauche. Nous nous sommes intéressé dans ce modèle à deux phénomènes. Le premier est le flux de masse qui traverse l'origine de la droite vers la gauche. Le second est la propagation du front de choc vers la gauche.

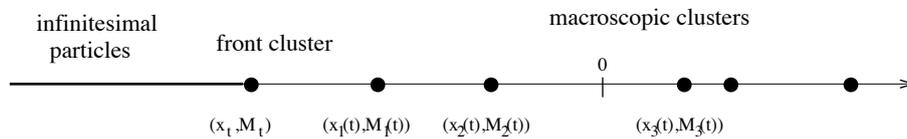


FIG. 1.13 – Etat du système à un temps $t > 0$.

L'article [A15], écrit en collaboration avec J. Bertoin et Y. Iozaki, décrit complètement la loi du flux de masse à travers l'origine à la fois à un temps fixe et d'un point de vue dynamique. Ces lois s'expriment de nouveau à partir de fonctions liées à la fonction d'Airy Ai . Des résultats similaires sont obtenus pour le cas où $(W(x), x > 0)$ est l'intégrale d'un mouvement brownien.

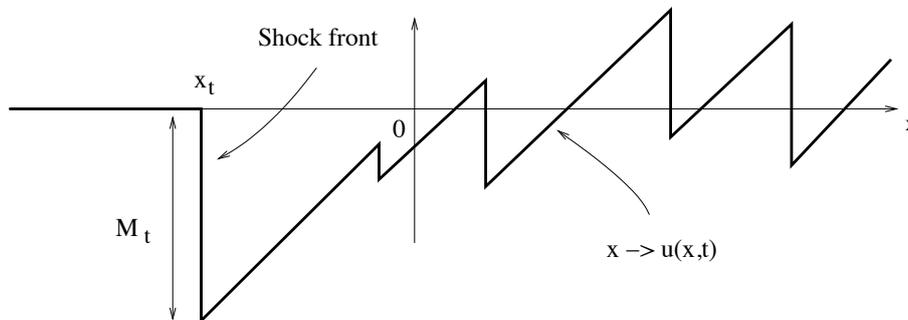


FIG. 1.14 – Front de choc progressant vers la gauche.

L'article [A12] est dédié à l'étude de la propagation du front de choc vers la gauche, voir Figures 1.13 et 1.14. Ce front se propage globalement à une vitesse sub-balistique. La loi de la position x_t et de l'amplitude M_t du front s'expriment de nouveau en terme des fonctions d'Airy Ai et Bi . De plus le comportement asymptotique de x_t pour $t \rightarrow 0$ ou $t \rightarrow \infty$ est obtenu grâce à une connexion avec le comportement asymptotique des premiers temps de passage du mouvement brownien à travers un niveau. Enfin, le calcul du générateur infinitésimal du processus $(M_t, t \geq 0)$ donne une description complète de la dynamique du front grâce à la relation $x_t = -\frac{1}{2} \int_0^t M_s \frac{ds}{s}$.

Condition initiale bruit stable

Lorsque la condition initiale $u(., 0)$ est un processus α -stable, Bertoin [Be99] montre que la solution $u(x, t)$ présente des points réguliers dans deux cas de figure : lorsque $u(., 0)$ est un bruit de Cauchy et lorsque $u(., 0)$ est un bruit stable de Lévy non-complètement asymétrique d'indice $\alpha \in]1/2, 1[$.

Dans l'article [A13] nous étudions ces points réguliers. dans le cas bruit stable de Lévy non-complètement asymétrique d'indice $\alpha \in]1/2, 1[$, nous montrons que l'ensemble des points réguliers est p.s. discret et régénératif. De plus, les points réguliers ont une vitesse nulle et l'évolution de la turbulence de part et d'autre d'un point régulier est indépendante. Ce dernier résultat est physiquement intuitif. Un point régulier correspond à une particule restée isolée et de vitesse nulle. Les particules présentes à sa droite et à sa gauche n'ont donc pas pu interagir entre elles.

Dans le cas où $u(., 0)$ est un bruit de Cauchy, nous montrons que l'ensemble des points réguliers est p.s. non dénombrable mais avec une dimension de Minkowsky nulle.

3.2 Particules collantes en interaction gravitationnelle

La distribution actuelle des grandes structures de l'univers reflèteraient une petite fluctuation de densité par rapport à la densité uniforme datant de l'époque du découplément du baryon et du photon [VDFN94]. Cette hypothèse a suscité quelques travaux explorant la distribution de masse obtenue dans un système de particules en interaction gravitationnelle issu d'une distribution initiale uniforme [MP96, LS05, KL06, Vy08, Za08].

Dans les articles [A14] et [A10], nous avons étudié diverses propriétés statistiques d'un tel modèle en dimension 1. Nous avons considéré un modèle de particules collantes dont la dynamique entre les chocs est régie par l'hamiltonien

$$H = \sum_i \frac{p_i^2}{2m_i} + \gamma \sum_{i \neq j} m_i m_j |x_i - x_j|,$$

où x_i, m_i, p_i représentent la position, la masse et le moment de la particule i au temps t et où γ est une constante gravitationnelle. L'accélération subie par une particule est proportionnelle à la différence des masses entre sa gauche et sa droite. Cette dynamique est une solution faible du système

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0 \\ \partial_t(\rho u) + \partial_x(\rho u^2) = -\rho \partial_x \phi \\ \partial_{xx}^2 \phi = \gamma \rho \end{cases}$$

où ρ, u, ϕ représentent la densité massique, la vitesse et le potentiel gravitationnel en x au temps t , voir [ERS96]. La dynamique du modèle est donc totalement déterministe et l'aléa n'est présent qu'au niveau de la distribution initiale des particules.

L'article [A14] est consacré à l'étude du système lorsqu'au temps initial les particules sont inertes et réparties selon la loi uniforme sur un intervalle. Nous avons étudié la distribution de la taille des amas (plus gros amas, amas "typique", etc..) ainsi que les échelles de temps auxquelles apparaissent différents types d'amas. Les résultats obtenus sont assez spécifiques des conditions initiales considérées. En effet, divers travaux récents [LS05, KL06, Vy06, Vy08, Za08] ont montrés tout un panel de propriétés possibles selon la nature des conditions initiales considérées.

L'article [A10] exhibe quant à lui des liens entre ce modèle gravitationnel et la coalescence/coagulation additive. Nous montrons que lorsque les particules sont réparties initialement selon un processus de Poisson avec des vitesses v_i vérifiant

$$v_{i+1} - v_i = -\lambda \frac{m_{i+1} + m_i}{2}, \quad \text{avec } \lambda \geq 0,$$

la dynamique de coalescence des particules est alors celle d'un coalescent additif [EP98, AP98] à un changement de temps stochastique près. Dans un second temps nous étudions une limite hydrodynamique du système et nous montrons que la distribution de masses peut être décrite à partir de l'enveloppe parabolique de l'intégrale d'un pont (respectivement mouvement) brownien lorsque la condition initiale est "uniforme" (resp. "poissonnienne"). Cette connexion entre la dynamique du système et d'une part la coalescence additive et d'autre part l'intégrale d'un pont brownien, induit une connexion entre le coalescent additif standard [AP98] et l'enveloppe parabolique d'un pont brownien. Plus précisément, si b est un pont brownien rendu périodique sur \mathbb{R} , l'enveloppe α -parabolique de $\int_0^x b$ est constituée de petits morceaux de paraboles. Si on oublie la position de ces morceaux, la dynamique selon laquelle fusionnent les morceaux de l'enveloppe e^{-t} -parabolique de $\int_0^x b$ lorsque t croît, est celle du coalescent additif standard.

Enfin, lorsque les masses et positions initiales des particules sont données par un processus ponctuel de Poisson, nous établissons dans [A10] un lien entre la dynamique du système de particules et l'équation de coagulation de Smoluchowsky, ce lien étant similaire à celui exhibé par Bertoin dans [Be02].

3.3 Quelques perspectives

D'un point de vue physique, l'étude de l'équation de Burgers et de la dynamique d'agrégation ballistique en dimension supérieure serait plus intéressante. Cette analyse est cependant beaucoup plus difficile et seules quelques propriétés sont connues, voir Poupaud [Po02], Frisch *et al.* [FBV01], Bec et Khanin [BK07].

Plus proche des travaux de la section 3.1, mentionnons que les propriétés asymptotiques de l'estimateur du maximum de vraisemblance d'une densité monotone sont reliées à l'enveloppe convexe d'un mouvement brownien avec drift polynomial

$$\left(\frac{1}{p} |x|^p - W(x), x \in \mathbb{R} \right), \quad \text{avec } p \geq 2. \quad (1.19)$$

Parallèlement, cette enveloppe convexe est intimement liée aux solutions de l'équation de Lax

$$\partial_t u + \frac{1}{q} \partial_x |u|^q = 0, \quad \text{avec } \frac{1}{p} + \frac{1}{q} = 1.$$

Menon et Srinivasan [MS10] proposent une analyse formelle de cette équation avec condition initiale aléatoire. Cette analyse, proposant une équation dont serait solution le générateur de $u(x, t)$, offre une caractérisation de l'enveloppe convexe de (1.19) pour $p \geq 3$.

Bibliographie exogène

- [AS64] M. Abramowitz, I.A. Stegun, Handbook of mathematical functions. Nat. Bur. Stand., Washington, 1964.
- [Ak73] H. Akaike, Information theory and an extension of the maximum likelihood principal. In B.N. Petrov and F. Csaki ed., 2nd International Symposium on Information Theory, 267–281, 1973.
- [AP98] D.J. Aldous and J. Pitman. The standard additive coalescent. *Ann. Probab.* 26 (1998), 1703–1726.
- [An51] T.W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distribution. *Annals of Mathematical Statistics* 22 (1951), 327–351.
- [AC10] S. Arlot and A. Celisse, A survey of cross-validation procedures for model selection. *Statistics Surveys* 4 (2010), 40–79.
- [Av95] M. Avellaneda, 1 Statistical Properties of Shocks in Burgers Turbulence II : Tail Probabilities for Velocities, Shock-strengths and Rarefaction Intervals. *Comm. Math. Phys.* 169 (1995), 45–59.
- [AE95] M. Avellaneda, Weinan E, Statistical Properties of Shocks in Burgers Turbulence. *Comm. Math. Phys.* 172 (1995), 13–38.
- [Ba08] F. Bach. Consistency of trace norm minimization, *Journal of Machine Learning Research*, 9 (2008), 1019–1048.
- [BGA08] O. Banerjee, L.E. Ghaoui and A. d’Aspremont. Model selection through sparse maximum likelihood estimation. *J. Machine Learning Research* 9 (2008), 485–516.
- [BDDW08] R. Baraniuk, M. Davenport, R. De Vore and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* 28 (2008), no. 3, 253–263.
- [BK07] J. Bec and K. Khanin. Burgers turbulence. *Phys. Rep.* 447 (2007), 1–66.
- [Be98] J. Bertoin, Large Deviation Estimates in Burgers Turbulence with Stable Noise Initial Data. *J. Stat. Phys.* 91 no 3/4 (1998), 655–667.
- [Be99] J. Bertoin, Structure of Shocks in Burgers Turbulence with Stable Noise Initial Data. *Comm. Math. Phys.* 203 (1999), 729–741.
- [Be00] J. Bertoin, Clustering Statistics for Sticky Particles with Brownian Initial Velocity. *J. Math. Pures Appl.* 79 no 2 (2000), 173–194.
- [Be02] J. Bertoin. Self-attracting Poisson clouds in an expanding univers. *Comm. Math. Phys.* 232 (2002), no.1, 59–81.
- [BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.* 3 (2001), 203–268.
- [BM00] L. Birgé and P. Massart, An adaptive compression algorithm in Besov spaces. *Constr. Approx.* 16 (2000), no. 1, 1–36.

- [BMPZ98] J.C. Bonvin, Ph.A. Martin, J. Piasecki, X. Zotos, Statistics of Mass Aggregation in a Self-Gravitating One-Dimensional Gas. *J. Stat. Phys.* 91 no 1/2 (1998), 177–197.
- [Br01] L. Breiman, Random forests. *Machine Learning* 45 (2001), 5–32.
- [BN08] F. Bunea and A. Nobel. Sequential Procedures for Aggregating Arbitrary Estimators of a Conditional Mean. *IEEE Transactions in Information Theory* 54 (2008), no. 4, 1725–1735.
- [BSW11] F. Bunea, Y. She and M. Wegkamp. Optimal selection of reduced rank estimation of high-dimensional matrices. *To appear in the Ann. Statist.*
- [BTW07] F. Bunea, A.B. Tsybakov and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.* 35 (2007), no 4, 1674–1697
- [Bu48] J.M. Burgers, *Adv. Appl. Mech.* 1 (1948), p. 171.
- [Bu50] J.M. Burgers, *Proc. Acad. Sci. Amst.* 53 (1950), pp 247, 393, 718, 732.
- [Bu74] J.M. Burgers, *The Nonlinear Diffusion Equation*. Dordrecht, Reidel 1974.
- [CR06] R. Castelo and A. Roverato. A robust procedure for Gaussian graphical model search from microarray data with p larger than n *J. Mach. Learn. Res.*, 7 (2006), 2621–2650.
- [Ca04] O. Catoni. Statistical learning theory and stochastic optimization. In *Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001*. Springer-Verlag, Berlin, 2004.
- [Ch10] A. Chambolle and V. Caselles and M. Novaga and D. Cremers and T. Pock. An introduction to Total Variation for Image Analysis. *Radon Series Comp. Appl. Math.* 9 (2010), 1–78.
- [CDD09] A. Cohen, W. Dahmen and R. De Vore. Compressed sensing and the best k -term approximation. *J. Amer. Math. Soc.* 22 (2009), 211–231.
- [Co51] J.D. Cole, On a Quasi Linear Parabolic Equation Occuring in Aerodynamics. *Quart. Appl. Math.* 9 (1951), 225–236.
- [DT08] A. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72 (2008), 39–61.
- [DP07] M. Drton and M. Perlman. Multiple testing and error control in Gaussian Graphical model selection. *Statist. Sci.* 22 (2007) no. 3, 430–449.
- [ERS96] Weinan E, Ya.G. Rykov, Ya.G. Sinai, Generalized Variational Principles, Global Weak Solutions and Behavior with Random Initial Data for Systems of Conservation Laws Arising in Adhesion Particle Dynamics. *Comm. Math. Phys.* 177 (1996), 349–380.
- [EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression. *Ann. Statist.* 32 (2004), no 4, 407–499.
- [EP98] S. Evans and J. Pitman (1998). Construction of Markovian coalescents. *Ann. Inst. H. Poincaré* 34 (1998), 339–383.
- [FFW09] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat.* 3 (2009), 521–541.
- [FM00] L. Frachebourg, Ph.A. Martin, Exact Statistical Properties of the Burgers Equation. *J. Fluid. Mech.* 417 (2000), 323–349.
- [FMP00] L. Frachebourg, Ph.A. Martin, J. Piasecki, Ballistic Aggregation : a Solvable Model of Irreversible many Particles Dynamics. *Physica A* 279 (2000), 69–99.
- [FHT08] J. Friedman, T. Hastie, R. Tibshirani. Sparse inverse covariance estimation with the lasso. *Biostatistics* 9 (2008) no. 3, 432–441.

- [FBV01] U. Frisch, J. Bec and B. Villone. Singularities and the distribution of density in the Burgers / adhesion model. *Physica D* 152-153 (2001), 620–635,
- [Ge75] S. Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* 70 (1975), 320–328.
- [GL10] S. Gaïffas and G. Lecué. Sharp oracle inequalities for the prediction of a high-dimensional matrix. arXiv :1008.4886v1 (2010).
- [Ge79] R. K. Gettoor Splitting times and shift functionals. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 47 (1979), 69–81.
- [Go09] A. Goldenshluger. A universal procedure for aggregating estimators. *Ann. Statist.* 37 (2009), no. 1, 542–568.
- [Gr89] P. Groeneboom, Brownian Motion with a Parabolic Drift and Airy Functions. *Proba. Th. Rel. Fields* 81 (1989), 79–109.
- [JN00] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.* 28 (2000) no 3, 681–712.
- [He90] I. Helland, PLS regression and statistical models. *Scandinavian Journal of Statistics*, 17 (1990), 97–114.
- [He06] K.R. Hess *et al.* Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology* 24 (2006), 4236–4244.
- [Ho50] E. Hopf, The Partial Differential Equation $u_t + uu_x = \mu u_{xx}$. *Comm. Pure Appl. Math.* 3 (1950), 201–230.
- [HLPL06] J.Z. Huang, N. Liu, M. Pourahmadi and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* 93 no 1, (2006), 85–98.
- [Iz75] A.J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate analysis* 5 (1975), 248–262.
- [KB08] M. Kalisch and P. Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *J. Comput. Graph. Statist.* 17 (2008), 773–789.
- [KMCMB11] M. Kalisch, M. Machler, D. Colombo, M.H. Maathuis and P. Bühlmann. Causal inference using graphical models with the R package pcalg. A paraître dans *J. Statist. Software*.
- [Ki10] S.G. Kim, M. Chung, J. Hanson, B. Avants, J. Gee, R. Davidson, S. Pollak. Structural connectivity via the tensor-based morphometry. *Biannual Meeting of Korean Society of Human Brain Mapping* (2010).
- [KLT10] V. Koltchinskii, K. Lounici and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. (2010) arXiv :1011.6256v2
- [Kr68] R.H. Kraichnan, Lagrangian History Statistical Theory for Burgers’ Equation. *Phys. Fluids* 11 (1968), 265–277.
- [KL06] L.V. Kuoza and M.A. Lifshits, Aggregation in a one-dimensional gas model with stable initial data. *J. Math. Sci.* 133 (2006), no. 3, 1298–1307
- [LCKKL06] H. Lee, M. Chung, H. Kang, B.N. Kim, D.S. Lee. Discriminative persistent homology of brain networks (2011).
- [LB06] G. Leung and A.R. Barron, Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* 52 (2006) no 8, 3396–3410.
- [Li87] K.-C. Li. Asymptotic optimality for C_p , CL , cross-validation and generalized cross-validation : discrete index set. *Ann. Statist.* 15 (1987), no. 3, 958–975.

- [LS05] M. Lifshits and Z. Shi. Aggregation rates in one-dimensional stochastic systems with adhesion and gravitation. *Ann. Probab.* 33 (2005), no. 1, 53–81.
- [LMY10] Z. Lu, R. Monteiro and M. Yuan. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming* 123 (2010), no. 2.
- [MKB09] M.H. Maathuis, M. Kalisch and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics* 37 (2009), 3133–3164.
- [MP67] V. A. Marchenko, L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb. (N.S.)*, 72 (1967), no 4, 507–536.
- [MP96] Ph.A. Martin, J. Piasecki, Aggregation Dynamics in a Self-Gravitating One-Dimensional Gas. *J. Stat. Phys.* 84 no 3/4 (1996), 837–857.
- [MB06] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics* 34 (2006), 1436–1462.
- [MS10] G. Menon and R. Srinivasan. Kinetic Theory and Lax Equations for Shock Clustering and Burgers Turbulence. *J. Stat. Phys.* 140 (2010), 1195–1223.
- [Mi78] P.W. Millar, A Path Decomposition for Markov processes. *Ann. of Proba.* 6 (1978), 345–348.
- [Mi09] M. Miller *et. al.* Morphometry birn. collaborative computational anatomy : An MRI morphometry study of the human brain via diffeomorphic metric mapping. *Human Brain Mapping* 30 (2009), no 7, 2132–2141.
- [MV08] F. Mordelet and J.-P. Vert. SIRENE : Supervised Inference of REgulatory NEtworks. *Bioinformatics* 24 (2008), no16, 76–82.
- [NW09] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. (2009) arXiv :0912.5100v1
- [Ni84] R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* 12 (1984), no 2, 758–765.
- [Po02] F. Poupaud. Diagonal Defect Measures, Adhesion Dynamics and Euler Equation. *Methods Appl. Anal.* 9(2002), no 4, 533–562.
- [RV98] G.C. Reinsel and R.P. Velu. *Multivariate Reduced-Rank Regression : Theory and Applications.* Lecture Notes in Statist. 136. Springer, New York, 1998.
- [RT11a] P. Rigollet, A.B. Tsybakov. Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.* 39 (2011), 731–771
- [RT11b] A. Rohde, A.B. Tsybakov. Estimation of High-Dimensional Low-Rank Matrices. *Ann. Statist.* 39, (2011), no 2, 887–930.
- [Ry98a] R. Ryan, Large-deviation Analysis of Burgers Turbulence with White-noise Initial Data. *Comm. Pure Appl. Math.* 51 (1998), 47–75.
- [Ry98b] R. Ryan, The Statistics of Burgers Turbulence Initialized with White-noise Initial Data. *Comm. Math. Phys.* 191 (1998), 71–86.
- [Sch78] G. Schwartz, Estimating the dimension of a model. *Ann. Statist.* 6 (1978), 461–464
- [SZ89] S.F. Shandarin, Ya.B. Zeldovich, The Large-scale Structures of the Universe : Turbulence, Intermittency, Structures in a Self-gravitating Medium. *Rev. Mod. Phys.* 61 (1989), 185–220.
- [Sh97] J. Shao. An asymptotic theory for linear model selection. *Statist. Sinica* 7 (1997), no 2, 221–264.

- [SAF92] Z.S. She, E. Aurell, U. Frisch, The Inviscid Burgers Equation with Data of Brownian Type. *Comm. Math. Phys.* 148 (1992), 632–641.
- [Sh81] R. Shibata. An optimal selection of regression variables. *Biometrika* 68 (1981), no 1, 45–54.
- [SPGS00] P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search. *Adaptive Computation and Machine Learning*, Cambridge, MIT Press, 2000.
- [Ts03] A.B. Tsybakov. Optimal rates of aggregation. *COLT* (2003), 303–313.
- [VDFN94] M. Vergassola, B. Dubrulle, U. Frisch, A. Noullez, Burgers’ Equation, Devil’s Staircases and the Mass Distribution Function for Large-scale Structures. *Astron. Astrophys.* 289 (1994), 325–356.
- [Vert10] J.-P. Vert. Reconstruction of biological networks by supervised machine learning approaches. in H. Lodhi and S. Muggleton (Eds.), *Elements of Computational Systems Biology*, Wiley, p.189-212, 2010.
- [Verz10] N. Verzelen. Minimax risks for sparse regressions : Ultra-high-dimensional phenomena. (2010) arXiv :1008.0526
- [Vy06] V. Vysitsky. On energy and clusters in stochastic systems of sticky gravitating particles. *Theory Probab. Appl.* 50 (2006) no. 2, pp. 265D283
- [Vy08] V. Vysotsky. Clustering in a stochastic model of one-dimensional gas. *Ann. Appl. Probab.* 18 (2008), no. 3, 1026–1058.
- [WB06] A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.* 5 (2006).
- [Wo66] H. Wold, Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiah (Ed.). *Multivariate Analysis.* (pp. 391D420) New York, Academic Press, 1966.
- [Wo98] W.A. Woyczyński, Göttingen Lectures on Burgers-KPZ turbulence. *Lecture Notes in Math.* 1700, Springer 1998.
- [Ya00] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.* 74 (2000), no. 1, 135–161.
- [Ya04] Y. Yang. Combining forecasting procedures : some theoretical results. *Econometric Theory* 20 (2004), no. 1, 176–222.
- [YELM07] M. Yuan, A. Ekici, Z. Lu and R. Monteiro. Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression. *Journal of the Royal Statistical Society, Series B*, 69 (2007), 329–346.
- [YL07] M. Yuan and Y. Lin Model selection and estimation in the Gaussian graphical model. *Biometrika* 94 (2007), 19–35.
- [Za08] V.F. Zakharova, Aggregation in a one-dimensional stochastic gas model with finite power moments of particle velocities. *J. Math. Sci.* 152 (2008), no. 6, 885–896
- [Ze70] Ya.B. Zeldovich, Gravitational Instability : an Approximate Theory for Large Density Perturbations. *Astron. Astrophys.* 5 (1970), pp 84-89.
- [ZH05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67 (2005), no. 2, 301–320.