UNIVERSITE DE POITIERS
UFR Sciences Fondamentales et Appliquées

Document de synthèse en vue de
L'HABILITATION A DIRIGER DES RECHERCHES

Spécialité : Mathématiques appliquées

Sélection de paramètre de lissage des estimateurs récursifs, problème de déconvolution, censure des données, grandes déviations et déviations modérées

par

Yousri Slaoui

Soutenue le 07 octobre 2016 devant le jury composé de :

M. Gérard Biau	Université de Paris VI	Rapporteur
Mme Hermine Biermé	Université de Poitiers	Examinatrice
Mme. Delphine Blanke	Université d'Avignon	Présidente
M. Denis Bosq	Université de Paris VI	Examinateur
M. Hervé Cardot	Université de Bourgogne	Rapporteur
Mme. Aurore Delaigle	Université de Melbourne	Rapportrice
M. Clément Dombry	Université de Franche-Comté	Examinateur
M. Julien Michel	Université de Poitiers	Examinateur
M. Abdelkkader Mokkadem	Université de Versailles	Examinateur

Table des matières

Remerciements

Je tiens en premier lieu à remercier Abdelkader Mokkadem qui a dirigé ma thèse de doctorat et qui m'a fait découvrir le monde des statistiques non-paramétriques. Je remercie également Mariane Pelletier avec qui j'ai eu beaucoup de plaisir à collaborer et à échanger et qui comme Abdelkader a contribué à ouvrir mon imaginaire mathématique.

J'adresse également mes plus sincères remerciements à Gérard Biau, Hervé Cardot et Aurore Delaigle pour l'honneur qu'ils me font en acceptant d'être les rapporteurs de ce mémoire d'habilitation et pour l'intérêt qu'ils ont portés à ce travail.

Je remercie Denis Bosq, Delphine Blanke, Abdelkader Mokkadem, Clément Dombry, Julien Michel et Hermine Biermé qui m'ont honoré en acceptant d'être examinateurs dans le jury mon habilitation.

Je remercie aussi tous mes co-auteurs, avec qui j'ai eu la chance et le plaisir de travailler, pour les nombreux échanges passionnants et enrichissants : Abdelkader Mokkadem, Mariane Pelletier, Grégory Nuel, Vincent Miele, Asma Jmaei, Achour Mohamed Salah, André Garcia, Laurence Watier, Astride de Hauteclocque, Stéphanie Ragot, Elise Gand et Samy Hadjadj.

Je tiens aussi à remercier mes collègues du Laboratoire de Mathématiques et Applications de l'Université de Poitiers, qui m'ont accueilli dans une ambiance agréable et chaleureuse. Je pense en particulier à Pol Vanhaecke et Abderrazak Bouaziz, anciens directeurs du LMA, et à Marc Arnaudon et Julien Michel, anciens responsable de l'équipe de probabilités, statistique et applications, pour leur disponibilité et leurs conseils. Ils ont certainement contribué pour beaucoup au dynamisme de mes débuts en tant que Maître de Conférences. Je pense également à Alessandra Sarti, directrice du LMA et Samuel Boissière, directeur adjoint de l'Ecole doctorale S2IM, pour leur gentillesse et sympathie. Je pense aussi à Pierre-Yves Louis avec lequel, j'ai eu l'occasion de collaborer pour des encadrements d'étudiants à divers niveaux (licence, master et thèse) ainsi pour l'organisation du séminaire de probabilités, statistique et applications.

Merci à Isabelle Amour, Jocelyne Attab, Brigitte Brault, Nathalie Marlet, Benoît Métrot et Nathalie Mongin pour leur disponibilité et les nombreux services rendus au quotidien.

Merci enfin à ma famille, notamment à mes parents, à mon frère et à mes deux sœurs, qui m'ont toujours encouragé, à ma femme Kenza, qui m'a suivi avec enthousiasme dans cette aventure, à ma fille Maram, mon rayon de soleil et à ma belle famille.

À la mémoire de mon grand père

Chapitre 1

Introduction

Ce document présente une synthèse de mes travaux de recherche. Ceux-ci se répartissent sur six thèmes : algorithmes stochastiques, sélection de paramètre de lissage, problème de déconvolution, grandes déviations et déviations modérées, censure des données dans un modèle linéaire mixte, méthode statistique d'analyse d'association. Dans ce mémoire, je décris l'évolution de mes travaux depuis ma thèse tout en mettant en relief le fil conducteur commun.
Mon travail de thèse (2002-2006) au sein du Laboratoire de Mathématiques de Versailles (LMV), sous la direction de Abdelkader Mokkadem et Mariane Pelletier portait sur l'application des méthodes d'approximation stochastiques à l'estimation d'une densité de probabilité et d'une fonction de régression. Dans la première partie, nous avons construit un algorithme stochastique à pas simple qui définit une classe d'estimateurs récursifs à noyau d'une densité de probabilité. Nous avons ensuite étudié les différentes propriétés de cet algorithme. En particulier, nous avons identifié deux classes d'estimateurs ; la première correspond à un choix de pas qui permet d'obtenir un risque minimal, la seconde une variance minimale. Dans la deuxième partie de thèse, nous nous sommes intéressés à l'estimateur proposé par Révész (1973, 1977) pour estimer une fonction de régression $r : x \mapsto \mathbb{E}[Y|X = x]$. Nous avons souligné que l'estimateur de Révész, construit à l'aide d'un algorithme stochastique à pas simple, a un gros inconvénient : les hypothèses sur la densité marginale de X nécessaires pour établir la vitesse de convergence de r_n sont beaucoup plus fortes que celles habituellement requises pour étudier le comportement asymptotique d'un estimateur d'une fonction de régression. Nous avons montré comment l'application du principe de moyennisation des algorithmes stochastiques permet, tout d'abord en généralisant la définition de l'estimateur de Révész, puis en moyennisant cet estimateur généralisé, de construire un estimateur récursif $\bar{r}_n$ qui possède de bonnes propriétés asymptotiques.

J'ai eu par la suite la possiblilité d'effectuer un post doctorale au laboratoire statistique et génome à l'Université d'Evry sous la direction de Bernard Prum et Grégory Nuel, lors de ce séjour post doctorale, j'ai eu l'occasion de découvrir

des méthodes statistiques d'analyse d'association génétiques comme le FBAT (Family Based Association Test) et le TDT (Transmission Disequilibrium Test).

En collaboration avec Grégory Nuel et Vincent Miele, nous avons développé un algorithme qui fait face aux génotypes manquants et aux erreurs de génotypage dans un cadre familial. Dans l'article [C2] nous avons utilisé un algorithme EM (Expectation Maximization) pour estimer à la fois la distribution des vrais génotypes et la fréquence de l'erreur de génotypages. Nous avons ensuite et en collaboration avec Ahmed Rebai illustré notre méthode aux données de microphtalmie dans l'article [C3].

En collaboration avec Grégory Nuel, nous avons développé un algorithme qui corrige le biais engendré par la censure de la variable cible dans un modèle linéaire mixte. Dans l'article [YS9] nous avons considéré les effets aléatoires comme des données manquantes et nous avons utilisé l'échantillonneur de gibbs dans un cadre SEM (Stochastic Expectation Maximization) pour nos multiples imputations des données censurées. Nous avons ensuite et en collaboration avec André Garcia et Omar Gaye illustré notre méthode aux données du paludisme dans l'article [C1].

En collaboration avec André Garcia, Jacqueline Milet, Grégory Nuel, Laurence Watier, David Courtin, Paul Senghor et Florence Migot-Nabias, et en utilisant la statistique FBAT et les modèles linéaires mixtes, nous avons identifiés dans l'article [YS3] trois régions génomiques (6p25.1, 12q22 et 20p11q11) liées à la maladie du paludisme et nous avons detecté un gène associé à l'infection par le paludisme dans la région 5q31-q33.

En collaboration avec Nicolas Brunel et Florence d'alché-Buc, nous avons présenté dans l'article [C4], un algorithme qui permet d'extraire des modules dans un modèle autorégressif avec une application à l'inférence de réseaux de régulation des gènes.

Après avoir eu un poste de maître de conférences à l'Université de Poitiers en Septembre 2011, j'ai eu l'occasion de développer des collaborations avec des biologistes et médecins du CHU de Poitiers, en particulier avec Stéphanie Ragot, Astride de Hauteclocque et Sami Hadjadj. Nous avons utilisé des modèles paramétriques comme les modèles linéaires mixtes et d'autres non-paramétriques comme LOWESS (LOcally WEighted Scatterplot Smoothing), pour décrire des trajectoires des marqueurs biologiques comme la créatinine, l'insuffisance rénale et l'eGFR (estimated Glomerular Filtration Rate), comme des bons marqueurs pronostique de la survenue d'événements cardio-vasculaires ([C5], [C6], [YS6], [YS8], [YS16]).

De plus, je me suis intéressé à établir des principes de grandes déviations et des principes de déviation modérées pour les estimateurs développés dans les articles ([YS1], [YS2]). Ces travaux ont fait l'objet des publications ([YS4], [YS5], [YS11], [YS12]).

Pendant mes travaux de thèse, nous avons souligné la grande influence du paramètre de lissage h_n, appelé aussi fenêtre (bandwidth) sur la performance des estimateurs, je me suis donc ouvert sur la sélection de ce paramètre en utilisant la méthode d'injection (plug-in), dans le cadre d'une densité de probabilité [YS7], dans le cadre d'une fonction de distribution [YS10] et dans le cadre d'une

fonction de régression ([YS13], [YS14]).

Depuis Septembre 2013, je co-encadre avec Julien Michel les travaux de thèse de Asma Jmaei, qui porte sur l'estimation fonctionnelle non-paramétrique avec bord, dans l'article [YS17], nous avons utilisé des algorithmes stochastiques et les polynômes de Bernstein pour réduire le biais de l'estimation sur les bords dans le cas d'une fonction de distribution à support compact, une version courte de ce travail est présenté à la SFDS, 2016 (voir [C10]).

Depuis Septembre 2014, je co-encadre avec Julien Michel les travaux de thèse de Mohamed Salah Achour, qui porte sur l'estimation récursive des quantiles non paramétriques, dans un article en préparation [YS25], nous avons automatisé le choix du paramètre de lissage en utilisant la méthode d'injection, dans un article en préparation [YS26], nous avons utilisé des algorithmes d'approximation stochastiques pour construire des estimateurs récursifs des quantiles, nous avons ensuite comparé notre classe d'estimateurs aux estimateurs non paramétrique existant.

Plus récemment, je me suis intéressé à l'estimation d'une densité de probabilité quand les observations sont entachées d'un bruit gaussien [YS20] ainsi que dans le cas où les observations sont entachées d'un bruit distribué selon une loi de Laplace [YS23]. Je me suis intéressé également à l'estimation d'une fonction de distribution quand les observations sont entachées d'un bruit distribué selon une loi de Laplace [YS15].

Ensuite, je me suis intéressé au problème de l'estimation d'une densité de probabilité quand les observations sont censurées [YS22] et aussi au problème de l'estimation d'une densité de probabilité quand les observations contiennent des données manquantes [YS19].

Revues internationales à comité de lecture

[YS1] Mokkadem, A., Pelletier, M. and Slaoui, Y. (2009), The stochastic approximation method for the estimation of a multivariate probability density. *J. Statist. Plann. Inference*, **139**, 2459–2478.

[YS2] Mokkadem, A., Pelletier, M. and Slaoui, Y. (2009), Revisiting Révész's stochastic approximation method for the estimation of a regression function. *ALEA Lat. Am. J. Probab. Math. Stat.*, **6**, 63–114.

[YS3] Milet, J., Nuel, G., Watier, L., Courtin, D., Slaoui, Y., Senghor, P., Migot-Nabias, F., Gaye, O. and Garcia, A. (2010), Genome Wide Linkage Study, Using a 250K SNP Map, of Plasmodium falciparum Infection and Mild Malaria Attack in a Senegalese Population. *PLoS ONE*, **5** (7) : e11616.

[YS4] Slaoui, Y. (2013), Large and moderate principles for recursive kernel density estimators defined by stochastic approximation method. *Serdica Math. J.*, **39**, 53–82.

[YS5] Slaoui, Y. (2014), Large and moderate deviation principles for kernel distribution estimator. *Int. Math. Forum*, **9**, 871–890.

[YS6] de Hauteclocque, A., Ragot, S., Slaoui, Y., Sosner, P., Halimi, J. M., Rigalleau, V., Roussel, R., Saulnier, P. J,, Hadjadj, S. for the SURDIAGENE Study group (2014), La trajectoire de la créatinine chez les diabétiques de type 2 : un bon marqueur de la survenue d'évènements cardiovasculaires. *Diabetes & Metabolism*, **40**, Supplement 1 doi :10.1016/S1262-3636(14)72186-X.

[YS7] Slaoui, Y. (2014), Bandwidth selection for recursive kernel density estimators defined by stochastic approximation method. *J. Probab. Stat.*, **2014**, ID 739640, doi :10.1155/2014/739640.

[YS8] de Hauteclocque, A., Ragot, S., Slaoui, Y., Grand, E., Miot, A., Sosner, P., Halimi, J. M., Zaoui, P., Rigalleau, V., Roussel, R., Saulnier, P. J., Hadjadj, S. for the SURDIAGENE Study group (2014), The influence of

sex on Renal Function Decline in people with Type 2 Diabetes. *Diabetic Medicine*, **31**, 1121–1128.

[YS9] Slaoui, Y. and Nuel, G. (2014), Parameter estimation in a hierarchical random intercept model with censored response : An approach using a SEM algorithm and Gibbs sampling. *Sankhya B*, **76**, 210–233.

[YS10] Slaoui, Y. (2014), The stochastic approximation method for the estimation of a distribution function. *Math. Methods Statist.*, **23**, 306–325.

[YS11] Slaoui, Y. (2015), Large and Moderate deviation principles for averaged stochastic approximation method for the estimation of a regression function. *Serdica Math. J.*, **41**, 307–328.

[YS12] Slaoui, Y. (2015), Moderate deviation principles for recursive regression estimators defined by stochastic aproximation method. *Int. J. Math. Stat.*, **16**, 51–60.

[YS13] Slaoui, Y. (2015), Plug-In Bandwidth selector for recursive kernel regression estimators defined by stochastic approximation method. *Stat. Neerl.*, **69**, 483–509.

[YS14] Slaoui, Y. (2016), Optimal Bandwidth selection for semi-recursive kernel regression estimators. *Stat. Interface*, **9**, 375–388.

[YS15] Slaoui, Y. (2016), Bandwidth selection in deconvolution kernel distribution estimators defined by stochastic approximation method with Laplace errors. *J. Japan Statist. Soc.*, **46**, 1–26.

[YS16] Ragot, S., Saulnier, J. P., Grand, E., Velho, G., De Hauteclocque, A., Slaoui, Y., Potier, L., Sosner, P., Halimi, J. M., Zaoui, P., Rigalleau, V., Fumeron, F., Roussel, R., Marre, M., Hadjadj, S. on behalf of the SURDIAGENE and DIABHYVAR Study group, (2016), Dynamic changes in renal function are associated with Major Cardiovascular Events in patients with type 2 diabetes. *Diabetes Care*, **39**, 1259–1266.

[YS17] Jmaei, A., Slaoui, Y and Dellagi, D. (2016), Recursive kernel distribution estimators defined by stochastic approximation method using Bernstein polynomials. *en révision.*

[YS18] Slaoui, Y. (2016), On the choice of smoothing parameters for semi-recursive nonparametric hazard estimators. *J. Stat. Theory Pract.*, doi :10.1080/15598608.2016.1214853.

[YS19] Slaoui, Y. (2016), Recursive kernel density estimators under missing data. *Comm. Statist. Theory Methods*, doi :10.1080/03610926.2016.1205618.

[YS20] Slaoui, Y. (2016), Data-driven in deconvolution recursive kernel density estimators defined by stochastic approximation method. *soumis.*

[YS21] Slaoui, Y. (2016), Large and moderate deviation principles for recursive kernel distribution estimators defined by stochastic approximation method. *soumis*.

[YS22] Slaoui, Y. (2016), Smoothing parameters for recursive kernel density estimators under censoring. *soumis*.

[YS23] Slaoui, Y. (2016), Smoothing parameters for deconvolution recursive kernel density estimators defined by stochastic approximation method with Laplace errors. *en révision*.

[YS24] Slaoui, Y. (2016), Data-driven bandwidth selection for recursive kernel density estimators under double truncation. *soumis*.

[YS25] Achour, M. S. and Slaoui, Y. (2016), Optimal bandwidth selection for kernel quantile estimators. *en préparation*.

[YS26] Achour, M. S. and Slaoui, Y. (2016), Bandwidth selection for recursive kernel quantile estimators defined by stochastic approximation method. *en préparation*.

Actes de Conférences

[C1] Slaoui, Y., Garcia, A., Gaye, O. and Nuel, G. (2007), A methodological approach to left censored parasite densities in malaria. Genetics and Mechanisms of susceptibility to infectious diseases. *EMBO, European Molecular Biology Organization, Institut Pasteur, Paris, France.*

[C2] Nuel, G., Slaoui, Y. and Miele, V. (2008), libfbat : a C++ library for family based association testing. *JOBIM, Journées Ouvertes en Biologie, Informatique et Mathématiques, Lille, France*, 119–124.

[C3] Nuel, G., Slaoui, Y., Miele, V. and Rebai, A. (2008), Taking into account missing genotypes and errors in Family Based Association Testing using an Expectation-Maximization framework. *ISB, International Symposium Biotechnology, Sfax, Tunisie*, 508–514.

[C4] Slaoui, Y., Brunel, N. and d'Alché-Buc, F. (2008), Module extraction in autoregressive models : application to gene regulatory networks inference. *MLSB, Machine Learning in Systems Biology, Académie Royale de Belgique, Bruxelles, Belgique.*

[C5] de Hauteclocque, A., Ragot, S., Slaoui, Y., Sosner, P., Halimi, J. M., Rigalleau, V., Roussel, R., Saulnier, P. J., Hadjadj, S. for the SURDIAGENE Study group (2013), *La trajectoire de la créatinine chez les diabétiques de type 2 : un bon marqueur de la survenue d'évènements cardiovasculaires. Journee recherche Tours-Poitiers.* Poitiers, France.

[C6] de Hauteclocque, A., Ragot, S., Slaoui, Y., Sosner, P., Halimi, J. M., Rigalleau, V., Roussel, R., Saulnier, P. J., Hadjadj, S. for the SURDIAGENE Study group (2014), *La trajectoire de la créatinine chez les diabétiques de type 2 : un bon marqueur de la survenue d'évènements cardiovasculaires. Congrès annuel de la société francophone du diabète.* Paris, France.

[C7] Slaoui, Y. (2015), *Large and Moderate Deviation Principles for Nonrecursive and Recursive Estimators of a Regression Function. 9th Annual International Conference on Statistics*, Athens, Greece.

[C8] Slaoui, Y. (2015), *Bandwidth selection in deconvolution recursive kernel density estimators defined by stochastic approximation method. EMS, European Meeting of Statisticians*, Amsterdam, Pays-Bas.

[C9] Slaoui, Y. (2015), *Parameter estimation in a hierarchical random intercept model with censored response : An approach using a SEM algorithm and Gibbs sampling. Biometrics & Biostatistics, San Antonio, USA, 2015. Abstract in Journal of Applied and Computational Mathematics*, **4**, Doi : 10.4172/2168-9679.C1.003.

[C10] Jmaei, A. and Slaoui, Y. (2016), *Recursive kernel distribution estimators defined by stochastic approximation method using Bernstein polynomials. 48 èmes Journées de Statisique de la SFDS*, Montpellier, France.

[C11] Slaoui, Y. (2016), *Smoothing parameters for recursive kernel density estimators under double truncation*, *22nd International Conference on Computational Statistics* (COMPSTAT 2016), Oviedo, Espagne.

Chapitre 2

Algorithme d'approximation stochastique

Sommaire

Mots clés : Estimation d'une densité de probabilité, estimation d'une fonction de distribution, estimation d'une fonction de régression, théorème de la limite centrale, algorithmes d'approximation stochastiques.

Résumé 1. *Dans ce chapitre, nous utilisons les algorithmes stochastiques pour construire des estimateurs récursifs. Le but est d'établir certaines propriétés des estimateurs récursifs à noyau définis par des algorithmes stochastiques afin de comparer leur comportement asymptotique à celui des estimateurs classiques. Dans la première section, nous construisons des estimateurs récursifs à noyau d'une fonction de distribution, nous étudions aussi le cas d'une fonction de distribution à support compact. Dans la deuxième section, nous construisons des estimateurs récursifs à noyau d'une fonction de régression.*

2.1 Introduction

Les algorithmes stochastiques ont été largement utilisés dans de nombreuses applications de recherche, y compris l'identifications des systèmes, contôle adaptatif, système de transmission, détection de changement séquentiel, voir Benveniste et al. [2] pour une liste d'exemple. La forme générale des algorithmes stochastiques est :

$$\theta_n = \theta_{n-1} + \gamma_n \Phi\left(\theta_{n-1}, W_n\right) + \gamma_n^2 \mu_n\left(\theta_{n-1}, W_n\right), \tag{2.1}$$

où (γ_n) est une suite positive qui tend vers zéro, (θ_n) est la séquence à mettre à jour récursivement, (W_n) est une séquence de variables aléatoires représentant les observations en ligne, $\Phi(\theta, W)$ est la fonction qui définit essentiellement comment le paramètre θ est mis à jour en fonction de la nouvelle observation et $\mu_n\left(\theta_{n-1}, W_n\right)$ définit une petite perturbation sur l'algorithme. Le comportement de cet algorithme a été étudié dans Benveniste et al. [2] et dans le cas particulier $(\mu_n \equiv 0)$ dans Delyon [4]. L'algorithme (2.1) coincide avec celui analysé par Kushner [8], Ljung [10] et Ruppert [15] :

$$\theta_n = \theta_{n-1} + \gamma_n\left(\phi\left(\theta_{n-1}\right) - W_n + \beta_n\right), \tag{2.2}$$

où β_n est une variable aléatoire converge vers 0 presque sûrement, ϕ est une fonction mesurable et inconnue. Ils ont montré que (2.2) comprend le processus de rapprochement stochastique de Robbins-Monro [14] et celui de Kiefer-Wolfowitz [7], qui permettent la recherche de zéro θ^* de la fonction ϕ.

La plupart des résultats classiques concernant le processus de Robbins-Monro et Kiefer-Wolfowitz nécessitent l'hypothèse $\mathbb{E}\left[W_n|\mathcal{F}_{n-1}\right] = 0$ où $\mathcal{F}_{n-1}$ est la σ-algèbre des événements qui se produisent avant l'instant $n-1$. Sous des conditions standard sur la fonction ϕ et sur la séquence (γ_n), Duflo [5] et Kushner et Yiu [9] ont montré que θ_n converge vers θ^* presque sûrement.

L'application de la procédure de Robbins-Monro pour construire un algorithme d'approximation stochastique, a été présenté par Révész [12, 13] et étendu par Mokkadem, Pelletier et Slaoui [YS2] pour estimer une fonction de régression, par Tsybakov [16] pour approximer le mode d'une densité de probabilité, par Mokkadem, Pelletier et Slaoui [YS1] pour estimer une densité de probabilité multivariée et récemment par Slaoui [YS10] pour estimer une fonction de distribution et par Slaoui [YS13, YS14] pour estimer une fonction de régression.
Nous définissons maintenant une classe de séquences à variations régulières.

Définition 1. *Soit $\gamma \in \mathbb{R}$ et $(v_n)_{n\geq 1}$ une séquence positive non aléatoire. Nous disons que $(v_n) \in \mathcal{GS}(\gamma)$ si*

$$\lim_{n\to+\infty} n\left[1 - \frac{v_{n-1}}{v_n}\right] = \gamma. \tag{2.3}$$

La Condition (2.3) a été introduite par Galambos et Seneta [6] pour définir des séquences à variations régulières (voir aussi Bojanic et Seneta [3], et par Mokkadem et Pelletier [11] dans le contexte des algorithmes d'approximation stochastique. Notant que l'acronyme $\mathcal{GS}$ représente (Galambos et Seneta). Séquences typiques dans $\mathcal{GS}(\gamma)$ sont, pour $b \in \mathbb{R}$, $n^{\gamma}(\log n)^{b}$, $n^{\gamma}(\log\log n)^{b}$, etc.

2.2 Estimateurs récursifs d'une fonction de distribution

Soit $X_1, \ldots, X_n$ une suite de variable aléatoire indépendantes et de même loi, à valeurs dans $\mathbb{R}$ et soit f et F représentent, respectivement, la densité de probabilité de X_1 et la fonction de distribution de X_1. Pour construire un algorithme stochastique, qui approxime la fonction F à un point donné x, on définit un algorithme de recherche du zéro de la fonction $\phi : y \to F(x) - y$. On procède donc de la façon suivante : (i) nous fixons $F_0(x) \in [0,1]$; (ii) pour tout $n \geq 1$, nous posons

$$F_n(x) = F_{n-1}(x) + \gamma_n T_n(x),$$

où (T_n) est une suite de fonctions $T_n : \mathbb{R} \to \mathbb{R}$ défini par $T_n(x) = \phi(F_{n-1}(x)) - W_n + \beta_n$, en utilisant le fait que $\mathbb{E}[W_n|\mathcal{F}_{n-1}] = 0$, nous avons

$$\begin{aligned} \mathbb{E}[T_n(x)] &= \phi(F_{n-1}(x)) + \beta_n \\ &= F(x) - F_{n-1}(x) + \beta_n. \end{aligned}$$

Pour définir $T_n(x)$, nous suivons l'approche de Révész [12, 13] et de Tsybakov [16] et nous introduisons un noyau K (qui est une fonction telle que $\int_{\mathbb{R}} K(x)dx = 1$), une fonction $\mathcal{K}$ (qui est une fonction défini par $\mathcal{K}(z) = \int_{-\infty}^{z} K(u)\,du$), et une fenêtre (h_n) (qui est une suite de réels positifs qui tend vers zéro). Par ailleurs, il est bien connu que sous certaines conditions de régularité sur la fonction de densité de X_1, nous avons $\mathbb{E}\left[\mathcal{K}\left(h_n^{-1}(x - X_n)\right)\right] = F(x) + \xi_n(x)$, où $\xi_n(x)$ tend vers zéro quand n tend vers l'infini. Par conséquent, nous posons, $T_n(x) = \mathcal{K}\left(h_n^{-1}(x - X_n)\right) - F_{n-1}(x)$. L'algorithme d'approximation stochastique, que nous introduisons pour estimer récursivement la fonction de distribution F en un point x peut donc s'écrire sous la forme suivante :

$$F_n(x) = (1 - \gamma_n) F_{n-1}(x) + \gamma_n \mathcal{K}\left(\frac{x - X_n}{h_n}\right)$$

où la suite de pas $\gamma_n > 0$ vérifie

$$\sum_{n\geq 1} \gamma_n = \infty \quad \text{et} \quad \sum_{n\geq 1} \gamma_n^2 < \infty.$$

Cette propriété récursive est particulièrement intéressante quand la taille de l'échantillon est grande, car nous pouvons facilement faire une mise à jour rapide de F_n avec chaque observation supplémentaire.
Nous supposons que les hypothèses suivantes sont satisfaites.

(A1) $K : \mathbb{R} \to \mathbb{R}$ est une fonction continue bornée, telle que $\int_{\mathbb{R}} K(z)\,dz = 1$, $\int_{\mathbb{R}} zK(z) = 0$ et $\int_{\mathbb{R}} z^2 K(z) < \infty$.

(A2) i) $(\gamma_n) \in \mathcal{GS}(-\alpha)$ avec $\alpha \in]1/2, 1]$.
ii) $(h_n) \in \mathcal{GS}(-a)$ avec $a \in]0, 1[$.
iii) $\lim_{n\to\infty}(n\gamma_n) \in]\min\{2a, (a+\alpha)/2\}, \infty]$.

(A3) f est bornée, différentiable, et f' est bornée.

L'hypothèse (A2)iii) sur la limite de $(n\gamma_n)$ quand n tend vers l'infini est classique dans le cadre des algorithmes d'approximation stochastiques. Les hypothèses (A1) et (A3) sont classiques en estimation non paramétrique. Dans l'article [YS10] nous avons montré le théorème suivant :

Théorème 1 (Convergence ponctuelle faible)**.** *Supposons* $(A1)-(A3)$ *vérifiées, que* f' *est continue en* x *et soit* $\xi = \lim_{n\to\infty}(n\gamma_n)^{-1}$.

1. *S'il existe* $c \geq 0$ *telle que* $\gamma_n^{-1}h_n^3 \to c$, *alors*

$$\sqrt{\gamma_n^{-1}}\,(F_n(x) - F(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{\sqrt{c}}{2(1-2a\xi)}f'(x)\int_{\mathbb{R}} z^2K(z)\,dz, \frac{1}{2-\alpha\xi}F(x)(1-F(x))\right).$$

2. *Si* $\gamma_n^{-1}h_n^3 \to \infty$, *alors*

$$\frac{1}{h_n^2}(F_n(x) - F(x)) \xrightarrow{\mathbb{P}} \frac{1}{2(1-2a\xi)}f'(x)\int_{\mathbb{R}} z^2K(z)\,dz,$$

où $\xrightarrow{\mathcal{L}}$ *désigne la convergence en loi,* $\mathcal{N}$ *la distribution normale et* $\xrightarrow{\mathbb{P}}$ *la convergence en probabilité.*

2.3 Estimateurs d'une fonction de distribution à support compact

Lorsque nous avons une variable aléatoire X avec une fonction de distribution F à support compact $[a, b]$, il est facile de transformer la variable X en Y à support $[0, 1]$ avec le choix $Y = (X - a)/(b - a)$. La transformation tel que $Y = X/(1 + X)$ et $Y = 1/2 + \pi^{-1}\arctan(X)$ peut être utilisé aussi pour couvrir les cas de variables aléatoires à support $\mathbb{R}_+$ et $\mathbb{R}$ respectivement. Par la suite, nous pouvons adapter cette transformation en utilisant les polynômes de Bernstein pour l'approximation d'une fonction de distribution sur l'intervalle $[0, 1]$.
Avec ma première étudiante en thèse, Asma Jmaei, nous avons considéré $X_1, X_2, \ldots, X_n$ une séquence de variables aléatoires i.i.d ayant une distribution commune inconnue F avec une densité associée f à support dans $[0, 1]$. Nous avons ensuite utilisé le schéma de Robbins-Monro (voir [14]), afin de construire un algorithme d'approximation stochastique de la fonction F en un point x. Nous avons considéré l'algorithme de recherche du zéro de la fonction $h : y \mapsto$

$F(x) - y$ comme suivant : (i) on se donne $F_0(x) \in \mathbb{R}$; (ii) pour tout $n \geq 1$, on pose

$$F_n(x) = F_{n-1}(x) + \gamma_n W_n,$$

où W_n est une observation de la fonction h au point $F_{n-1}(x)$ (i.e. $\mathbb{E}[W_n|\mathcal{F}_{n-1}] = 0$, où $\mathcal{F}_{n-1}$ représente la σ-algebre des événements qui se produisent avant l'instant $n-1$). Pour définir W_n, nous avons suivi [17] (voir aussi [1]) et nous avons introduit un polynôme de Bernstein d'ordre $m > 0$ (nous avons supposé que $m = m_n$ dépend de n),

$$b_k(m,x) = C_m^k x^k (1-x)^{m-k} \quad \text{et} \quad W_n = \sum_{k=0}^{m} \mathbb{I}\left\{X_n \leq \frac{k}{m}\right\} b_k(m,x) - F_{n-1}(x).$$

Donc, l'estimateur que nous introduisons pour estimer récursivement une fonction de distribution F au point x peut s'écrire sous la forme suivante

$$F_n(x) = (1-\gamma_n)F_{n-1}(x) + \gamma_n \sum_{k=0}^{m} \mathbb{I}\left\{X_n \leq \frac{k}{m}\right\} b_k(m,x).$$

Nous supposons que les hypothèses suivantes sont satisfaites.

(A1) F est continue, deux fois différentiables et la dérivé seconde de F est bornée.

(A2) $(\gamma_n) \in \mathcal{GS}(-\alpha)$, $\alpha \in (\frac{1}{2}, 1]$.

(A3) $(m_n) \in \mathcal{GS}(a)$, $a \in (0,1)$.

(A4) $\lim_{n\to\infty}(n\gamma_n) \in (\min(a, (2\alpha+a)/4), \infty)$.

Dans l'article [YS17] nous avons prouvé le théorème suivant :

Théorème 2 (Convergence ponctuelle faible)**.** *Supposons (A1)-(A4) vérifiées et soit* $\xi = \lim_{n\to\infty}(n\gamma_n)^{-1}$

1. *Si* $\gamma_n^{-1/2} m_n^{-1} \to c$ *pour une constante* $c \geq 0$*, alors*

$$\gamma_n^{-1/2}(F_n(x) - F(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{cx(1-x)}{2(1-a\xi)} f'(x), \frac{1}{2-\alpha\xi} F(x)(1-F(x))\right).$$

2. *Si* $\gamma_n^{-1/2} m_n^{-1} \to \infty$*, alors*

$$m_n(F_n(x) - F(x)) \xrightarrow{\mathbb{P}} \frac{x(1-x)}{2(1-a\xi)} f'(x).$$

2.4 Estimateurs récursifs d'une fonction de régression

Nous considèrons $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon de la loi du couple (X, Y) et on note f la densité de probabilité de X et g la densité du couple (X, Y).

2.4.1 Estimateurs à noyau dans un cadre équidistant et fixe

Dans [YS13] nous avons considéré le cas où $X_i = i/n$ pour $1 \leq i \leq n$ et nous avons proposé d'estimer la fonction de régression $m : x \mapsto \mathbb{E}[Y|X = x]$ en un point x. Afin de construire un algorithme d'approximation stochastique de la fonction m en un point x, nous avons défini un algorithme de recherche du zéro de la fonction $h : y \to m(x) - y$. Suivant la procédure de Robbins-Monro [14], notre algorithme est définie par : (i) nous fixons $m_0(x) \in \mathbb{R}$, (ii) pour tout $n \geq 1$, nous posons

$$m_n(x) = m_{n-1}(x) + \gamma_n W_n,$$

où $W_n(x)$ est une observation de la fonction h au point $m_{n-1}(x)$, et le pas (γ_n) est une séquence de nombres réels positifs qui tend vers zéro. Pour définir $W_n(x)$, nous suivons l'approche de [12, 13], [16] et de [YS1, YS2] et nous introduisons un noyau K (qui est une fonction telle que $\int_{\mathbb{R}} K(x)dx = 1$), et une fenêtre (h_n) (qui est une suite de réels positifs qui tend vers zéro), et nous posons $W_n(x) = h_n^{-1} Y_n K\left(h_n^{-1}(x - X_n)\right) - m_{n-1}(x)$. L'algorithme d'approximation stochastique, que nous proposons pour estimer récursivement une fonction de régression dans un cadre équidistant et fixe en un point x, peut s'écrire sous la forme suivante :

$$m_n(x) = (1 - \gamma_n) m_{n-1}(x) + \gamma_n h_n^{-1} Y_n K\left(h_n^{-1}[x - X_n]\right).$$

Nous supposons que les hypothèses suivantes sont satisfaites.

(A1) $K : \mathbb{R} \to \mathbb{R}$ est une fonction continue, bornée, telle que $\int_{\mathbb{R}} K(z)\,dz = 1$, $\int_{\mathbb{R}} zK(z) = 0$ et $\int_{\mathbb{R}} z^2 K(z) < \infty$.

(A2) (i) $(\gamma_n) \in \mathcal{GS}(-\alpha)$ avec $\alpha \in (1/2, 1]$.
(ii) $(h_n) \in \mathcal{GS}(-a)$ avec $a \in (0, 1)$.
(iii) $\lim_{n\to\infty}(n\gamma_n) \in (\min\{2a, (\alpha - a)/2\}, \infty]$.

(A3) *(i)* $g(s,t)$ est deux fois continûment différentiable par rapport à s.
(ii) Pour $q \in \{0,1,2\}$, $s \mapsto \int_{\mathbb{R}} t^q g(s,t)\,dt$ est une fonction bornée continue à $s = x$.
Pour $q \in [2,3]$, $s \mapsto \int_{\mathbb{R}} |t|^q g(s,t)\,dt$ est une fonction bornée.
(iii) Pour $q \in \{0,1\}$, $\int_{\mathbb{R}} |t|^q \left|\frac{\partial g}{\partial x}(x,t)\right| dt < \infty$, et $s \mapsto \int_{\mathbb{R}} t^q \frac{\partial^2 g}{\partial s^2}(s,t)\,dt$ est une fonction bornée continue à $s = x$

Nous avons prouvé le théorème suivant dans l'article [YS13] :

Théorème 3 (Convergence ponctuelle faible)**.** *Supposons* $(A1)-(A3)$ *vérifiées, que* $m^{(2)}$ *est continue en* x *et soit* $\xi = \lim_{n\to\infty}(n\gamma_n)^{-1}$.

1. *S'il existe* $c \geq 0$ *telle que* $\gamma_n^{-1} h_n^5 \to c$, *alors*

$$\sqrt{\gamma_n^{-1} h_n}\,(m_n(x) - m(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{\sqrt{c} m^{(2)}(x)}{2(1 - 2a\xi)} \int_{\mathbb{R}} z^2 K(z)\,dz, \frac{f(x)\,\mathbb{E}\left[Y^2|X = x\right]}{2 - (\alpha - a)\xi} \int_{\mathbb{R}} K^2(z)\,dz\right).$$

2. Si $\gamma_n^{-1} h_n^5 \to \infty$, *alors*

$$\frac{1}{h_n^2}\left(m_n(x) - m(x)\right) \xrightarrow{\mathbb{P}} \frac{m^{(2)}(x)}{2(1-2a\xi)} \int_{\mathbb{R}} z^2 K(z)\, dz.$$

2.4.2 Estimateurs à noyau dans un cadre unidimensionnel

Dans [YS14], nous avons considéré l'estimation de la fonction de régression comme rapport de deux fonctions dont nous avons proposé deux estimateurs récursifs dans les parties précédentes. Donc nous avons considéré l'estimateur semi-recursif suivant :

$$r_n(x) = \frac{a_n(x)}{f_n(x)},$$

avec

$$a_n(x) = (1-\beta_n)\, a_{n-1}(x) + \beta_n h_n^{-1} Y_n K\left(h_n^{-1}[x - X_n]\right),$$

et

$$f_n(x) = (1-\gamma_n)\, f_{n-1}(x) + \gamma_n h_n^{-1} K\left(h_n^{-1}[x - X_n]\right).$$

Nous supposons que les hypothèses suivantes sont satisfaites.

(A1) $K : \mathbb{R} \to \mathbb{R}$ est une fonction continue, bornée, telle que $\int_{\mathbb{R}} K(z)\, dz = 1$, $\int_{\mathbb{R}} zK(z) = 0$ et $\int_{\mathbb{R}} z^2 K(z) < \infty$.

(A2) *i)* $(\beta_n) \in \mathcal{GS}(-\beta)$ avec $\beta \in]1/2, 1]$.
ii) $(h_n) \in \mathcal{GS}(-a)$ avec $a \in]0, 1[$.
iii) $\lim_{n\to\infty}(n\beta_n) \in]\min\{2a, (\beta - a)/2\}, \infty]$.

(A3) *(i)* $g(s,t)$ est deux fois continûment différentiable par rapport à s.
(ii) Pour $q \in \{0,1,2\}$, $s \mapsto \int_{\mathbb{R}} t^q g(s,t)\, dt$ est une fonction bornée continue à $s = x$.
Pour $q \in [2,3]$, $s \mapsto \int_{\mathbb{R}} |t|^q\, g(s,t)\, dt$ est une fonction bornée.
(iii) Pour $q \in \{0,1\}$, $\int_{\mathbb{R}} |t|^q \left|\frac{\partial g}{\partial x}(x,t)\right| dt < \infty$, et $s \mapsto \int_{\mathbb{R}} t^q \frac{\partial^2 g}{\partial s^2}(s,t)\, dt$ est une fonction bornée continue à $s = x$.

Nous avons ensuite discuté le choix optimal du couple de pas (γ_n, β_n). En particulier, nous avons montré qu'avec un choix de pas $(\gamma_n) = \left(n^{-1}\right)$ (le choix qui permet d'avoir une erreur quadratique moyenne minimale de l'estimateur récursif d'une densité de probabilité), nous obtenons le théorème suivant :

Théorème 4 (Convergence ponctuelle faible)**.** *Supposons* $(A1) - (A3)$ *vérifiées, et que* $(\gamma_n) = \left(n^{-1}\right)$.

1. S'il existe $c \geq 0$ *telle que* $\beta_n^{-1} h_n^5 \to c$, *alors*

$$\sqrt{\beta_n^{-1} h_n}\left(r_n(x) - r(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\sqrt{c} B_{a,\xi}^{(1)}(x), V_{a,\xi,\beta}^{(1)}(x)\right),$$

où

$$\begin{aligned} B_{a,\xi}^{(1)}(x) &= \frac{1}{2f(x)}\left(\frac{a^{(2)}(x)}{(1-2a\xi)} - \frac{r(x)\,f^{(2)}(x)}{(1-2a)}\right)\mu_2(K) \\ V_{a,\xi,\beta}^{(1)}(x) &= \left\{\frac{\mathbb{E}\left[Y^2|X=x\right]}{(2-(\beta-a)\,\xi)\,f(x)} - \left(\frac{2\xi}{1+a\xi} - \frac{\xi}{1+a}\right)\frac{r^2(x)}{f(x)}\right\}R(K) \end{aligned}$$

2. *Si* $nh_n^5 \to \infty$, *alors*

$$\frac{1}{h_n^2}\left(r_n(x) - r(x)\right) \xrightarrow{\mathbb{P}} B_{a,\xi}^{(1)}(x).$$

De plus, avec un choix particulier de pas $(\gamma_n) = \left((1-a)\,n^{-1}\right)$ le choix qui permet d'avoir une variance minimale de l'estimateur récursif de la densité. Nous avons prouvé le théorème suivant :

Théorème 5 (Convergence ponctuelle faible)**.** *Supposons* $(A1)-(A3)$ *vérifiées, et que* $(\gamma_n) = \left((1-a)\,n^{-1}\right)$.

1. *S'il existe* $c \geq 0$ *telle que* $\beta_n^{-1}h_n^5 \to c$, *alors*

$$\sqrt{\beta_n^{-1}h_n}\left(r_n(x) - r(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\sqrt{c}B_{a,\xi}^{(2)}(x), V_{a,\xi,\beta}^{(2)}(x)\right),$$

où

$$\begin{aligned} B_{a,\xi}^{(2)}(x) &= \frac{1}{2f(x)}\left(\frac{a^{(2)}(x)}{(1-2a\xi)} - \frac{(1-a)}{(1-3a)}r(x)\,f^{(2)}(x)\right)\mu_2(K) \\ V_{a,\xi,\beta}^{(2)}(x) &= \left\{\frac{\mathbb{E}\left[Y^2|X=x\right]}{(2-(\beta-a)\,\xi)\,f(x)} - (1-a)\,\xi\frac{r^2(x)}{f(x)}\right\}R(K) \end{aligned}$$

2. *Si* $nh_n^5 \to \infty$, *alors*

$$\frac{1}{h_n^2}\left(r_n(x) - r(x)\right) \xrightarrow{\mathbb{P}} B_{a,\xi}^{(2)}(x).$$

Bibliography

[1] Babu, G. J. Canty, A. J. and Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function, *J. Statist. Plann. Inference.*, **105** : 377–392.

[2] Benveniste, A., Métivier, M. and Priouret, P. (1990). Adaptive algorithm and stochastic approximations, *Springer-Verlag.*

[3] Bojanic, R. and Seneta, E. (1973). A unified theory of regularly varying sequences, *Math. Z.* **134** : 91–106.

[4] Delyon, B. (1996). General results on the convergence of stochastic algorithms, *IEEE Trans. Automat. Control.* **41** : 1245–1255.

[5] Duflo, M. (1997). Random iterative models. Collection Applications of Mathematics, *Springer, Berlin.*

[6] Galambos, J. and Seneta, E. (1973). Regularly varying sequences, *Amer. Math. Soc.* **41** : 110–116.

[7] Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression functions, *Ann. Math. Statist.* **23** : 462–466.

[8] Kushner, H. J. (1977). General convergence results for stochastic approximations via week convergence theory, *J. Math. Anal. Appl.* **61** : 490–503.

[9] Kushner, H. J. and Yin, G. G. (2003). Stochastic approximation and recursive algorithms and applications, *volume 35 of Applications of Mathematics. Springer-Verlag, New York.*

[10] Ljung, L. (1978). Strong convergence of a stochastic approximation algorithm, *Ann. Statist.* **6** : 680–696.

[11] Mokkadem, A. and Pelletier, M. (2007). A companion for the Kiefer-Wolfowitz-Blum stochastic approximation algorithm, *Ann. Statist.* **35** : 1749–1772.

[12] Révész, P. (1973). Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes I, *Studia Sci. Math. Hung.* **8** : 391–398.

[13] Révész, P. (1977). How to apply the method of stochastic approximation in the non-parametric estimation of a regression function, *Math. Operationsforsch. Statist., Ser. Statistics.* **8** : 119–126.

[14] Robbins, H. and Monro, S. (1951). A stochastic approximation method, *Ann. Statist.* **22** : 400–407.

[15] Ruppert, D. (1982). Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise, *Ann. Probab.* **10** : 178–187.

[16] Tsybakov, A. B. (1990). Recurrent estimation of the mode of a multidimensional distribution, *Probl. Inf. Transm.* **8** : 119–126.

[17] Vitale, R. A. (1975). A Bernstein polynomial approach to density function estimation, *Stat. Inference Relat. Top.*, **2** : 87–99.

Chapitre 3

Sélection du paramètre de lissage des estimateurs fonctionnels récursifs

Sommaire

Mots clés : Sélection du paramètre de lissage, estimation à noyau d'une densité de probabilité, estimation d'une fonction de distribution, algorithmes d'approximation stochastiques, lissage, ajustement de courbe, méthode de plug-in.

Résumé 2. *Le but dans ce chapitre est d'automatiser le choix du paramètre de lissage dans le cas de l'estimation récursive d'une densité de probabilité ainsi que dans le cas d'une fonction de distribution. Dans la première section, nous proposons un sélecteur du paramètre de lissage d'une densité de probabilité en minimisant l'Erreur Quadratique Pondérée Moyenne Intégrée (EQPMI) en considérant la fonction $f(x)$ comme une fonction de poids. Dans la deuxième section, en se basant sur un selecteur du même type que celui développé dans la première partie mais dans le cadre d'une estimation récursive d'une fonction de distribution, nous prouvons qu'avec un choix particulier de pas de l'algorithme stochastique, l'erreur quadratique moyenne intégrée de l'estimateur récursif peut être plus petite que celle de l'estimateur classique.*

3.1 Introduction

Dans le cadre de l'estimation fonctionnelle par la méthode du noyau, on a affaire à un paramètre dit de lissage ou fenêtre. D'un point de vue théorique, c'est

le comportement asymptotique de ce paramètre qui permet d'atteindre des vitesses de convergence optimales. D'un point de vue pratique, c'est ce paramètre qui va déterminer le degré de lissage de la courbe estimée et donc la qualité de l'estimation. Dans les deux cas, il est nécessaire de choisir un paramètre de lissage qui réalise l'équilibre entre les effets de biais et de dispertion.
Les méthodes de sélection du paramètre de lissage étudiés dans la littérature peuvent être divisé en trois grandes classes : les méthodes de type validation-croisée, les méthodes de type ré-échantillonnage ou bootstrap et les méthodes de type injection ou plug-in.
Les méthodes de type validation-croisée, inspirées des techniques de choix de modèle, consiste à introduire des critères pour estimer l'erreur d'estimation et à prendre ensuite comme paramètre de lissage celui qui minimise un tel critère. Les méthodes de ré-echantillonnage consistent à utiliser les techniques de Bootstrap pour estimer la distribution d'erreur d'estimation. Les méthodes de type plug-in consistent à partir de l'expession de l'erreur d'estimation dans laquelle les constantes sont précisées, d'estimer ensuite ces constantes de manière non paramétrique et finalement d'injecter ces constantes estimées dans l'expression de l'erreur d'estimation afin d'en déduire un paramètre de lissage optimal.
Marron [4] offre une synthèse de ces trois classes de méthodes de sélection du paramètre de lissage. Une comparaison détaillée de ces trois choix du paramètre de lissage peuvent être trouvés dans Delaigle et Gijbels [3]. Ils ont conclu qu'avec un choix approprié de la méthode d'injection et de la méthode de ré-echantillonnage, les deux approches réalisent de meilleurs performances en comparaison avec la méthode de validation croisée, et qu'aucune des deux approches n'a été préférée de manière univoque. Bagkavos et Patil [1] proposent une méthode spécifique d'injection qui est basée sur la minimisation de l'Erreur Quadratique Moyenne Empirique (EQME) et sélectionne le paramètre de lissage qui minimise l'EQME localement au voisinage du point d'estimation. Cheng [2] développe un sélecteur du paramètre de lissage en minimisant l'Erreur quadratique moyenne intégrée (EQMI) en utilisant des ajustements linéaire et cubique. Les estimateurs que nous avons considéré dans ce chapitre sont basés sur la méthode d'injection, qui consiste à la minimisation non pas du EQMI classique, sinon, nous devrions donner un estimateur asymptotiquement sans biais de $\int_{\mathbb{R}} \left(f^{(2)}(x)\right)^2 dx$ (dans le cas de la densité), qui n'est pas très facile à construire en utilisant seulement les donnée observées, plutôt d'utiliser l'EQMI, nous avons utilisé l'Erreur Quadratique Pondérée Moyenne Intégrée (EQPMI), en considérant la fonction $f(x)$ (dans le cas de la densité) comme une fonction de poids.

3.2 Sélection du paramètre de lissage des estimateurs récursifs d'une densité de probabilité

Soit $X_1, \ldots, X_n$ une suite de variable aléatoire indépendante et de même loi, à valeurs dans $\mathbb{R}$ et de densité de pabability f. Nous avons construit un algorithme pour estimer récursivement la densité f au point x :

$$f_n(x) = (1-\gamma_n) f_{n-1}(x) + \gamma_n h_n^{-1} K\left(h_n^{-1}[x - X_n]\right),$$

avec (γ_n) une suite de réels positifs qui tend vers zéro, K un noyau (i.e. une fonction paire telle que $\int_{\mathbb{R}} K(x)\,dx = 1$) et h_n une fenêtre (i.e. une suite déterministe positive qui tend vers zéro). Afin de mesurer la qualité de notre estimation, nous nous sommes basés sur la quantité suivante :

$$EQPMI[f_n] = \mathbb{E}\int_{\mathbb{R}} [f_n(x) - f(x)]^2 f(x)\,dx.$$

Avec un choix particulier du pas (γ_n) dans $\mathcal{GS}(-1)$ et tel que $\lim_{n\to\infty} n\gamma_n = \gamma_0$, nous avons montré que la fenêtre optimale (h_n) doit être égale à

$$\left(2^{-1/5}(\gamma_0 - 2/5)^{1/5}\left\{\frac{R(K)}{\mu_2^2(K)}\frac{I_1}{I_2}\right\}^{1/5} n^{-1/5}\right),$$

et que

$$EQPMI[f_n] = \frac{5}{4}\frac{1}{2^{4/5}}\frac{\gamma_0^2}{(\gamma_0 - 2/5)^{6/5}}\Theta(K) I_1^{4/5} I_2^{1/5} n^{-4/5} + o\left(n^{-4/5}\right).$$

avec $I_1 = \int_{\mathbb{R}} f^2(x)\,dx$, $I_2 = \int_{\mathbb{R}} \left(f^{(2)}(x)\right)^2 f(x)\,dx$, $R(K) = \int_{\mathbb{R}} K^2(z)\,dz$, $\mu_j(K) = \int_{\mathbb{R}} z^j K(z)\,dz$ et $\Theta(K) = R(K)^{4/5}\mu_2(K)^{2/5}$.

Puisque le minimum de $\gamma_0^2(\gamma_0 - 2/5)^{-6/5}$ est atteint à $\gamma_0 = 1$, alors la fenêtre (h_n) doit être égale à

$$\left(\left(\frac{3}{10}\right)^{1/5}\left\{\frac{R(K)}{\mu_2^2(K)}\frac{I_1}{I_2}\right\}^{1/5} n^{-1/5}\right),$$

donc

$$EQPMI[f_n] = \frac{5}{4}2^{2/5}\left(\frac{5}{6}\right)^{6/5}\Theta(K) I_1^{4/5} I_2^{1/5} n^{-4/5} + o\left(n^{-4/5}\right).$$

Afin d'estimer la fenêtre (h_n), nous devrons estimer les I_1 et I_2. Dans [YS7], nous avons proposé les estimateurs suivants :

$$\widehat{I_1} = \frac{Q_n}{n}\sum_{i,k=1}^{n} Q_k^{-1}\beta_k b_k^{-1} K_b\left(\frac{X_i - X_k}{b_k}\right),$$

$$\widehat{I_2} = \frac{Q_n'^2}{n}\sum_{j\neq k} Q_j'^{-1} Q_k'^{-1}\beta_j'\beta_k' b_j'^{-3} b_k'^{-3} K_{b'}^{(2)}\left(\frac{X_i - X_j}{b_j'}\right) K_{b'}^{(2)}\left(\frac{X_i - X_k}{b_k'}\right),$$

où $Q_n = \prod_{i=1}^{n} (1-\beta_i)$, $Q'_n = \prod_{i=1}^{n} (1-\beta'_i)$, K_b est un noyau et b la fenêtre associé et $K_{b'}^{(2)}$ est la dérivée seconde du noyau $K_{b'}$.
Nous devrons maintenant estimer les fenêtres b_n et b'_n et les pas (β_n) et (β'_n). Pour cela, il faut calculer le biais et la variance de $\widehat{I}_1$ et $\widehat{I}_2$. Nous avons montré que

Théorème 6. *Supposons* $(A2) - (A3)$ *vérifiées, et que le noyau* K_b *satisfait l'hypothèse* $(A1)$ *et* $(b_n) \in \mathcal{GS}(-\zeta)$, *avec* $\zeta \in]0,1[$, *et* $\xi = \lim_{n\to\infty} (n\beta_n)^{-1}$ *on a*

$$\begin{aligned} Bias\left[\widehat{I}_1\right] &= \frac{1}{2(1-2a\xi)} b_n^2 \mu_2(K_b) \int_{\mathbb{R}} f^{(2)}(x) f(x)\, dx + o\left(b_n^2\right), \\ Var\left[\widehat{I}_1\right] &= \frac{1}{2-(\alpha-a)\xi} \frac{\beta_n}{nb_n} R(K_b) I_1 + \frac{1}{n}\left(\int_{\mathbb{R}} f^3(x)\, dx - I_1^2\right) + o\left(\frac{1}{n} + \frac{\beta_n}{nb_n}\right). \end{aligned}$$

Un calcul directe, nous a permis de choisir le pas $(\beta_n) = \left(1.36\, n^{-1}\right)$ et la fenêtre $(b_n) \in \mathcal{GS}(-2/5)$ de même après le calcul du biais et de la variance de $\widehat{I}_2$, nous avons obtenu les résultats suivants : $(\beta'_n) = \left(1.48\, n^{-1}\right)$ et $(b'_n) \in \mathcal{GS}(-3/14)$.
En pratique, nous utilisons

$$b_n = n^{-\beta} \min\left\{\widehat{s}, \frac{Q_3 - Q_1}{1.349}\right\}, \quad \beta \in]0,1[$$

(voir [6]) avec $\widehat{s}$ l'écart-type, et Q_1, Q_3 désignant le premier et troisième quartile, respectivement.

3.3 Sélection du paramètre de lissage des estimateurs récursifs d'une fonction de distribution

Soit $X_1, \ldots, X_n$ une suite de variable aléatoire indépendante et de même loi, à valeurs dans $\mathbb{R}$ et de densité de probabilité f et de fonction de distribution F. Nous avons construit dans [YS10] un algorithme pour estimer la fonction F au point x :

$$F_n(x) = (1-\gamma_n) F_{n-1}(x) + \gamma_n \mathcal{K}\left(h_n^{-1}[x - X_n]\right),$$

avec (γ_n) une suite de réels positifs qui tend vers zéro, K un noyau (i.e. une fonction paire telle que $\int_{\mathbb{R}} K(x)\, dx = 1$), $\mathcal{K}$ (une fonction définie par $\mathcal{K}(z) = \int_{-\infty}^{z} K(u)\, du$) et h_n une fenêtre (i.e. une suite déterministe positive qui tend vers zéro).

Afin de mesurer la qualité de notre estimation, nous avons utilisé la quantité suivante :

$$EQPMI[F_n] = \mathbb{E}\int_{\mathbb{R}} [F_n(x) - F(x)]^2 f(x)\, dx.$$

Avec un choix particulier de pas d'algorithme $(\gamma_n) \in \mathcal{GS}(-1)$ tel que $\lim n \to \infty n\gamma_n = \gamma_0$, nous avons montré que la fenêtre optimale (h_n) doit être égale à

$$\left(2^{-1/3}(\gamma_0 - 2/3)^{1/3}\left\{\frac{I_1\phi(K)}{I_2\mu_2^2(K)}\right\}^{1/3} n^{-1/3}\right),$$

donc

$$EQPMI[F_n] = n^{-1}V_F\left\{\frac{\gamma_0^2}{2\gamma_0 - 1} - \frac{3}{4}\frac{1}{2^{4/3}}\frac{\gamma_0^2}{(\gamma_0 - 2/3)^{2/3}}\frac{I_1^{4/3}\Theta(K)}{I_2^{1/3}V_F}n^{-1/3} + o\left(n^{-1/3}\right)\right\},$$

avec $I_1 = \int_{\mathbb{R}} f^2(x)\,dx$, $I_2 = \int_{\mathbb{R}} (f'(x))^2 f(x)\,dx$, $R(K) = \int_{\mathbb{R}} K^2(z)\,dz$, $\mu_j(K) = \int_{\mathbb{R}} z^j K(z)\,dz$, $\phi(K) = 2\int_{\mathbb{R}} zK(z)\mathcal{K}(z)\,dz$ et $\Theta(K) = \phi(K)^{4/3}\mu_2(K)^{2/3}$.

Remarque 1. *Avec un choix de pas* $(\gamma_n) = \left([2/3 + \varepsilon]\,n^{-1}\right)$ *avec* ε *assez proche de zéro, nous avons*

$$EQPMI[F_n] < EQPMI\left[\widetilde{F}_n\right],$$

avec $\widetilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathcal{K}\left(\frac{x - X_i}{h_n}\right)$, *l'estimateur introduit par [5].*

Afin d'estimer la fenêtre (h_n) nous devrons estimer I_2. Dans [YS10], nous avons proposé l'estimateur suivant :

$$\widehat{I}_2 = \frac{\Pi_n'^2}{n}\sum_{j\neq k}\Pi_j'^{-1}\Pi_k'^{-1}\gamma_j'\gamma_k' b_j'^{-3}b_k'^{-3}K_{b'}^{(2)}\left(\frac{X_i - X_j}{b_j'}\right)K_{b'}^{(2)}\left(\frac{X_i - X_k}{b_k'}\right),$$

où $\Pi_n' = \prod_{i=1}^n (1 - \gamma_i')$, K_b est un noyau et b la fenêtre associé et $K_{b'}^{(1)}$ est la dérivée première du noyau $K_{b'}$.
Nous devrons maintenant estimer la fenêtre b_n' et les pas (γ_n) et (γ_n'), pour cela, il faut calculer le biais et la variance de $\widehat{I}_2$, Un calcul directe, nous amène au choix du pas $(\gamma_n) = \left(1.36\,n^{-1}\right)$ et la fenêtre $(b_n) \in \mathcal{GS}(-2/5)$ de même après le calcul du biais et de la variance de $\widehat{I}_2$, nous avons montré que (b_n') doit appartenir à $\mathcal{GS}(-3/10)$.

Bibliography

[1] Bagkavos, D. and Patil, P. N. (2008). Local polynomial fitting in failure rate estimation. *IEEE Trans. Reliab.*, **56** : 126–163.

[2] Cheng, M. Y. (1997). Boundary aware estimators of integrated density derivative products. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **59** : 191–203.

[3] Delaigle, A. and Gijbels, I. (2004). Practical Bandwidth Selection in Deconvolution Kernel Density Estimation. *Comput. Statist. Data Anal.* **45** : 246–267.

[4] Marron, J. S. (1988) Automatic smoothing parameter selction : A survey. *Empir. Econom.* **13** : 187–208.

[5] Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory Prob. Appl.* **9** : 497–500.

[6] Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Chapman & Hall*, London.

Chapitre 4

Estimation fonctionnelle quand les observations sont entachées d'erreurs

Sommaire

Mots clés : Méthodes de déconvolution, sélection du paramètre de lissage, estimation à noyau d'une densité de probabilité, estimation d'une fonction de distribution, algorithmes d'approximation stochastiques, lissage, ajustement de courbe, méthode de plug-in.

Résumé 3. *Dans ce chapitre, nous utilisons les algorithmes stochastiques pour construire des estimateurs récursifs aussi bien dans le cas d'une densité de probabilité que dans le cas d'une fonction de distribution dans le cas où les données sont entachées d'erreurs. Dans la première section, nous construisons des estimateurs récursifs à noyau d'une densité de probabilité quand les observations sont entachées d'erreurs gaussiennes, nous automatisons le choix du paramètre de lissage selon le même critère de selection présenté dans la partie précèdente. Dans la deuxième section, nous construisons des estimateurs récursifs à noyau d'une fonction de distribution dans le cas où les observations sont entachées d'erreurs distribué selon une loi de Laplace.*

4.1 Introduction

On est confronté très souvent au traitement des données entachées d'erreurs. L'origine des erreurs peut provenir, par exemple, de l'appareil de mesure, de la lecture des observations ou de l'échelle utilisée. Notre contribution se situe dans l'estimation fonctionnelle quand les données sont entaché d'erreur additives (c.-à-d. un modèle de déconvolution). Le modèle de déconvolution, c.-à-d. on observe $Z = X + \texttt{erreur}$ existe dans plusieurs domaines. Le modèle de déconvolution peut être recontré en microfluométrie, en biostatistique et en électrophorése. Dans l'étude de la maladie de SIDA, la variable Z peut être considéré comme le temps à partir d'un certain instant jusqu'au moment de l'infection, et la variable X peut être considéré comme le temps entre l'occurence de l'infection jusqu'au moment de l'apparition des symptômes. Notre aport dans ce domaine touche à deux sujets particuliers. D'abord, l'estimation d'une densité de probabilité, ensuite l'estimation d'une fonction de distribution. Dans les deux situations, nous avons explicité l'expression asymptotique de l'erreur quadratiques moyennes intégrés dans le cas où les observations sont entachées par des erreurs qui suivent une loi normale et aussi dans le cas où les erreurs suivent une loi de laplace. Ensuite nous avons données l'expression de la fenêtre optimal en utilisant la méthode de plug-in aussi bien dans le cas récursif que dans le cas non récursif.

4.2 Estimation d'une densité de probabilité quand les observations sont entachées d'erreurs

Soit $Y_1, \ldots, Y_n$ un échantillon de variables aléatoires i.i.d de densité de probabilité f_Y. Les observations sont de la forme

$$Y_i = X_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

et $X_1, \ldots, X_n$ sont des variables i.i.d, de densité de probabilité f_X. Nous supposons que X et ε sont mutuellement indépendantes et que la densité f_ε des variables i.i.d $\varepsilon_1, \ldots, \varepsilon_n$ est connu. Nous souhaitons utilisé les algorithmes stochastiques pour construire un estimateur récursif de f_X.
Soit ϕ_L le transformer de Fourier d'une fonction ou d'une variable aléatoire L, et nous supposons que $\phi_\varepsilon(t) \neq 0 \ \forall t \in \mathbb{R}$. Le transformée de Fourier inverse :

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-itx)\, \phi_X(t)\, dt, \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-itx)\, \frac{\phi_Y(t)}{\phi_\varepsilon(t)} dt. \end{aligned}$$

De plus, la fonction caractéristique de Y est donnée par

$$\begin{aligned} \phi_Y(t) &= \mathbb{E}[\exp(itY)] \\ &= \int_{\mathbb{R}} \exp(ity)\, f_Y(y)\, dy. \end{aligned}$$

Maintenant, nous estimons ϕ_Y par

$$\widehat{\phi}_{n,Y}(t) = \int_{\mathbb{R}} \exp(ity)\, \widehat{f}_{n,Y}(y)\, dy, \tag{4.1}$$

où

$$\widehat{f}_{n,Y}(y) = (1-\gamma_n)\, \widehat{f}_{n-1,Y}(y) + \gamma_n h_n^{-1} K\left(h_n^{-1}[y - Y_n]\right), \tag{4.2}$$

(γ_n) est un pas de l'algorithme (une suite de nombres positifs décroissante vers 0), K est un noyau (une fonction qui satisfait $\int_{\mathbb{R}} K(x)dx = 1$) et (h_n) est une fenêtre (une suite de nombres positifs décroissante vers 0). Ensuite, nous estimons f_X par

$$\widehat{f}_{n,X}(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-itx)\, \frac{\widehat{\phi}_{n,Y}(t)}{\phi_\varepsilon(t)} dt. \tag{4.3}$$

D'après la combinaison des deux équations (4.1) et (4.2), on a

$$\widehat{\phi}_{n,Y}(t) = (1-\gamma_n)\, \widehat{\phi}_{n-1,Y}(t) + \gamma_n h_n^{-1} \int_{\mathbb{R}} \exp(ity)\, K\left(h_n^{-1}[y - Y_n]\right) dy,$$

et en utilisant l'équation (4.3), nous obtenons

$$\begin{aligned}\widehat{f}_{n,X}(x) &= (1-\gamma_n)\, \widehat{f}_{n-1,X}(x) \\ &\quad + \frac{1}{2\pi}\gamma_n h_n^{-1} \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \exp(ity)\, K\left(h_n^{-1}[y - Y_n]\right) dy\right] \exp(-itx)\, \phi_\varepsilon^{-1}(t)\, dt,\end{aligned}$$

de plus, il est facile de vérifier que

$$\int_{\mathbb{R}} \exp(ity)\, K\left(h_n^{-1}[y - Y_n]\right) dy = h_n \exp(itY_n)\, \phi_K(th_n).$$

Ensuite, on peut écrire que

$$\widehat{f}_{n,X}(x) = (1-\gamma_n)\, \widehat{f}_{n-1,X}(x) + \frac{1}{2\pi}\gamma_n \int_{\mathbb{R}} \exp(-it(x - Y_n))\, \phi_K(th_n)\, \phi_\varepsilon^{-1}(t)\, dt,$$

nous pouvons aussi écrire que

$$\widehat{f}_{n,X}(x) = (1-\gamma_n)\, \widehat{f}_{n-1,X}(x) + \gamma_n h_n^{-1} K^\varepsilon\left(h_n^{-1}[x - Y_n]\right), \tag{4.4}$$

avec

$$K^\varepsilon(u) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itu} \frac{\phi_K(t)}{\phi_\varepsilon\left(\frac{t}{h_n}\right)} dt.$$

Maintenant, nous supposons que $\widehat{f}_{0,X}(x) = 0$, et on pose $\Pi_n = \prod_{j=1}^n (1-\gamma_j)$. Donc, nous pouvons estimer f_X recursivement en un point x avec :

$$\widehat{f}_{n,X}(x) = \Pi_n \sum_{k=1}^n \Pi_k^{-1} \gamma_k h_k^{-1} K^\varepsilon\left(\frac{x - Y_k}{h_k}\right).$$

Nous avons étudié les propriétés des estimateurs de déconvolution récursifs à noyau (4.4), et leurs comparaisons avec l'estimateur de déconvolution non récursif developpé par Carroll et Hall [1] et Stefanski et Carroll [15]

$$\widetilde{f}_{n,X}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K^{\varepsilon}\left(\frac{x-Y_i}{h_n}\right).$$

Les propriétés théoriques de cet estimateur ont été étudiés dans de nombreux documents, y compris Carroll et Hall [1], Stefanski et Carroll [15], Fan [6, 7, 8, 9], Masry [11, 12], Zhang et Karunamuni [18], Meister [13, 14] et Van et Uh [16, 17]. La sélection du paramètre de lissage a été proposé par Stefanski et Carroll [15], Hesse [10] et Delaigle et Gijbels [3, 4, 5].

4.2.1 Estimation d'une densité de probabilité quand les observations sont entachées d'erreurs gaussiennes

Dans cette partie, nous supposons que $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$, avec $\sigma^2 = sh_n^2$ et nous distinguons deux cas, dans le premier cas, nous considèrons que $s \in [0, 1)$ (cas ordinairement lisse) et dans le deuxième cas, nous considèrons que $s \geq 1$ (cas super lisse).

Cas ordinairement lisse
Dans ce cas le choix du noyau n'a pas une grande influence sur la qualité de l'estimation, dans ce cas nous avons utilisé un noyau gaussien. Nous avons ensuite considéré la quantité suivante afin de mesurer la qualité de notre estimateur :

$$EQPMI\left[\widehat{f}_{n,X}\right] = \mathbb{E}\int_{\mathbb{R}}\left[\widehat{f}_{n,X}(x) - f_X(x)\right]^2 f_Y(x)\,dx.$$

Avec un choix particulier du pas (γ_n) dans $\mathcal{GS}(-1)$ et tel que $\lim_{n\to\infty} n\gamma_n = \gamma_0$, nous avons montré que la fenêtre optimale (h_n) doit être égale à

$$\left(2^{-1/5}\left(2\sqrt{\pi}\right)^{-1/5}\left(\gamma_0 - \frac{2}{5}\right)^{1/5}\left\{\frac{I_1}{I_2}\right\}^{1/5} n^{-1/5}\right),$$

et que EQPMI asymptotique

$$AEQPMI\left[\widehat{f}_{n,X}\right] = \frac{5}{4}2^{-4/5}\left(2\sqrt{\pi}\right)^{-4/5}\gamma_0^2\left(\gamma_0 - 2/5\right)^{-6/5} I_1^{4/5} I_2^{1/5} n^{-4/5},$$

avec $I_1 = \int_{\mathbb{R}} f_Y^2(x)\,dx$, $I_2 = \int_{\mathbb{R}}\left(f_X^{(2)}(x)\right)^2 f_Y(x)\,dx$.

De plus, puisque le minimum de $\gamma_0^2\left(\gamma_0 - 2/5\right)^{-6/5}$ est atteint à $\gamma_0 = 1$, alors la fenêtre (h_n) doit être égale à

$$\left(\left(\frac{3}{20\sqrt{\pi}}\right)^{1/5}\left\{\frac{I_1}{I_2}\right\}^{1/5} n^{-1/5}\right),$$

et par conséquent,

$$AEQPMI\left[\widehat{f}_{n,X}\right] = \frac{5}{4}\left(\frac{2}{\pi}\right)^{2/5}\left(\frac{5}{6}\right)^{6/5} I_1^{4/5} I_2^{1/5} n^{-4/5}.$$

Le paramètre de lissage (h_n) obtenu n'est pas directement utilisable puisqu'il dépend des deux quantités inconnues I_1 et I_2. Dans [YS20], nous avons proposé les estimateurs suivants :

$$\begin{aligned}
\widehat{I}_1 &= \frac{\Pi_n}{n}\sum_{i,k=1}^{n} \Pi_k^{-1}\gamma_k b_k^{-1} K_b\left(\frac{Y_i - Y_k}{b_k}\right), \\
\widehat{I}_2 &= \frac{\Pi_n'^2}{n}\sum_{j\neq k} \Pi_j'^{-1}\Pi_k'^{-1}\gamma_j'\gamma_k' b_j'^{-3} b_k'^{-3} K_{b'}^{(2)}\left(\frac{Y_i - Y_j}{b_j'}\right) K_{b'}^{(2)}\left(\frac{Y_i - Y_k}{b_k'}\right).
\end{aligned}$$

où $\Pi_n = \prod_{i=1}^n (1-\gamma_i)$, $\Pi_n' = \prod_{i=1}^n (1-\gamma_i')$, K_b est un noyau et b la fenêtre associé et $K_{b'}^{(2)}$ est la dérivée seconde du noyau $K_{b'}$.

Cas super lisse

Dans ce cas, certaines restrictions doivent être imposées sur la fonction caractéristique ϕ_K du noyau. Dans ce paragraphe nous considérons que $\phi_K(t) = \left(1-t^2\right)^\delta \mathbb{1}_{\{|t|<1\}}$, avec $\delta = 0, 1, 2,$ où 3.

Afin de mesurer la qualité de notre estimateur, nous avons considéré la quantité suivante :

$$EQPMI\left[\widehat{f}_{n,X}\right] = \mathbb{E}\int_{\mathbb{R}}\left[\widehat{f}_{n,X}(x) - f_X(x)\right]^2 f_Y(x)\,dx.$$

Avec un choix particulier du pas (γ_n) dans $\mathcal{GS}(-1)$ et tel que $\lim_{n\to\infty} n\gamma_n = \gamma_0$, nous avons montré que la fenêtre optimale (h_n) doit être égale à

$$\left((2\pi)^{-1/5}(\gamma_0 - 2/5)^{1/5}\left\{\frac{I_1}{I_2}\right\}^{1/5}\frac{(1+2\beta s)^{1/5}}{\mu_2^{2/5}(K)}\exp\left(\frac{\eta s}{5}\right) n^{-1/5}\right),$$

et que

$$\begin{aligned}
AEQPMI\left[\widehat{f}_{n,X}\right] &= \frac{5}{4}(2\pi)^{-4/5}\left(1+\frac{2}{5}\beta s\right)(1+2\beta s)^{-1/5}\exp\left(\frac{4}{5}\eta s\right) \\
&\quad \frac{\gamma_0^2}{(\gamma_0 - 2/5)^{6/5}} I_1^{4/5} I_2^{1/5} \mu_2^{2/5}(K)\, n^{-4/5},
\end{aligned}$$

avec $I_1 = \int_{\mathbb{R}} f_Y^2(x)\,dx$, $I_2 = \int_{\mathbb{R}}\left(f_X^{(2)}(x)\right)^2 f_Y(x)\,dx$ $\eta = s^{-1}\log\left(\int_{-1}^1 \phi_K^2(t)\exp\left(st^2\right)dt\right)$ et $\beta = \left(\int_{-1}^1 t^2\phi_K^2(t)\exp\left(st^2\right)dt\right) / \left(\int_{-1}^1 \phi_K^2(t)\exp\left(st^2\right)dt\right)$.

De plus, puisque le minimum de $\gamma_0^2 \left(\gamma_0 - 2/5\right)^{-6/5}$ est atteint à $\gamma_0 = 1$, alors la fenêtre (h_n) doit être égale à

$$\left(\left(\frac{3}{10\pi} \right)^{1/5} \left\{ \frac{I_1}{I_2} \right\}^{1/5} \frac{(1+2\beta s)^{1/5}}{\mu_2^{2/5}(K)} \exp\left(\frac{\eta s}{5} \right) n^{-1/5} \right),$$

et par conséquent,

$$AEQPMI\left[\widehat{f}_{n,X}\right] = \frac{5}{4}\left(\frac{1}{2\pi}\right)^{4/5}\left(\frac{5}{3}\right)^{6/5}\Theta(K)\, I_1^{4/5} I_2^{1/5} n^{-4/5},$$

avec

$$\Theta(K) = \left(1 + \frac{2}{5}\beta s\right)(1+2\beta s)^{-1/5} \exp\left(\frac{4}{5}\eta s\right) \mu_2^{2/5}(K).$$

Le paramètre de lissage (h_n) obtenu n'est pas directement utilisable puisqu'il dépend des deux quantités inconnues I_1 et I_2. Dans [YS20], nous avons proposé les estimateurs suivants :

$$\begin{aligned}
\widehat{I}_1 &= \frac{Q_n}{n} \sum_{i,k=1}^{n} Q_k^{-1} \beta_k b_k^{-1} K_b\left(\frac{Y_i - Y_k}{b_k}\right), \\
\widehat{I}_2 &= \frac{Q_n'^2}{n} \sum_{j \neq k} Q_j'^{-1} Q_k'^{-1} \beta_j' \beta_k' b_j'^{-3} b_k'^{-3} K_{b'}^{\varepsilon(2)}\left(\frac{Y_i - Y_j}{b_j'}\right) K_{b'}^{\varepsilon(2)}\left(\frac{Y_i - Y_k}{b_k'}\right),
\end{aligned}$$

où $Q_n = \prod_{i=1}^{n}(1-\beta_i)$, $Q_n' = \prod_{i=1}^{n}(1-\beta_i')$, K_b est un noyau et b la fenêtre associé et $K_{b'}^{\varepsilon(2)}$ est la dérivée seconde du noyau de déconvolution $K_{b'}$.

4.3 Estimation d'une fonction de distribution quand les observations sont entachées d'erreurs

Soit $Y_1, \ldots, Y_n$ un échantillon de variables aléatoires i.i.d de densité de probabilité f_Y et de distribution F_Y. Les observations sont de la forme

$$Y_i = X_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

et $X_1, \ldots, X_n$ sont des variables i.i.d, de densité de probabilité f_X et de fonction de distribution F_X. Nous supposons que X et ε sont mutuellement indépendantes et que la densité f_ε des variables i.i.d $\varepsilon_1, \ldots, \varepsilon_n$ est connu. Ce problème est motivé par plusieurs applications pratiques dans différents domaines tels que, par exemple, l'astronomie, la santé publique, et l'économétrie. Dans la littérature, le problème de la déconvolution est généralement abordé sous deux classes d'hypothèses sur les distributions d'erreurs : la distribution lisse ordinaire et la distribution super lisse Fan (voir [7]). Des exemples de distributions lisses ordinaires comprennent Laplacien, gamma, et gamma symétrique ; des exemples de distributions super lisses sont la loi normale, mélange de loi normale et la loi de

Cauchy. D'un point de vue théorique, le taux de convergence ne peut pas être plus rapide qu'une fonction logarithmique dans le cas des erreurs super lisses, alors que pour les erreurs lisses ordinaires, le taux de convergence de F_X est polynomiale.
D'un point de vue pratique, Delaigle et Gijbels [5] ont souligné que les estimateurs de déconvolution avec une erreur de Laplace donne toujours des meilleurs résultats que le cas gaussien, et comme une application, ils ont considéré les données de "second National Health and Nutrition Examination Survey" (NHANES), il s'agit d'une étude de cohorte composée de milliers de femmes qui ont fait l'objet d'une évaluation de leurs comportement alimentaires, afin de rechercher des traces de cancer. La principale variable d'intérêt de l'étude est le logarithme de l'apport quotidien en gras saturé qui a été connu pour être mesuré de manière imprécise, pour plus de détails, voir Stefanski et Caroll [15] et Caroll et al [2]. Dans l'article [YS15], nous avons supposé que ε est distribué selon une loi de Laplace $\varepsilon \sim \mathcal{E}d(\sigma)$, avec σ représente le paramètre d'échelle. En utilisant l'algorithme de Robbins-Monro pour la recherche de zéro de la fonction $h : y \to F_X(x) - y$, nous avons montré qu'afin d'estimer F_X au point x récursivement, nous pouvons utiliser l'algorithme suivant :

$$F_{n,X}(x) = (1-\gamma_n) F_{n-1,X}(x) + \gamma_n \mathcal{K}^{\varepsilon}\left(h_n^{-1}(x - Y_n)\right),$$

avec (γ_n) et (h_n) sont deux suites de réels positifs qui tendent vers zéro, K un noyau tel que $\int_{\mathbb{R}} K(x)\,dx = 1$, et $\mathcal{K}(z) = \int_{-\infty}^{z} K(u)\,du$, et le noyau de déconvolution K^{ε} :

$$K^{\varepsilon}(u) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itu}\frac{\phi_K(t)}{\phi_{\varepsilon}\left(\frac{t}{h_n}\right)}dt,$$

avec ϕ_L le transformer de Fourier d'une fonction ou une variable aléatoire L. La méthode que nous avons développée pour mesurer la qualité de notre estimation est basée sur la quantité suivante :

$$EQPMI[F_{n,X}] = \mathbb{E}\int_{\mathbb{R}} [F_{n,X}(x) - F_X(x)]^2 f_Y(x)\,dx.$$

Avec un choix particulier du pas (γ_n) dans $\mathcal{GS}(-1)$ et tel que $\lim_{n\to\infty} n\gamma_n = \gamma_0$, nous avons montré que la fenêtre optimale (h_n) doit être égale à

$$\left(\left(\frac{3\sigma^4}{8\sqrt{\pi}}\right)^{1/7}(\gamma_0 - 2/7)^{1/7}\left\{\frac{I_1}{I_2}\right\}^{1/7} n^{-1/7}\right),$$

et que

$$AEQPMI[F_{n,X}] = \frac{7}{12}\left(\frac{3\sigma^4}{8\sqrt{\pi}}\right)^{4/7}\frac{\gamma_0^2}{(\gamma_0 - 2/7)^{10/7}} I_1^{4/7} I_2^{3/7} n^{-4/7},$$

avec $I_1 = \int_{\mathbb{R}} f_Y^2(x)\,dx$, $I_2 = \int_{\mathbb{R}} \left(f_X'(x)\right)^2 f_Y(x)\,dx$. Puisque le minimum de $\gamma_0^2\left(\gamma_0 - 2/7\right)^{-10/7}$ est atteint à $\gamma_0 = 1$, alors la fenêtre (h_n) doit être égale à

$$\left(0.7634\sigma^{4/7}\left\{\frac{I_1}{I_2}\right\}^{1/7} n^{-1/7}\right),$$

donc

$$AEQPMI\left[F_{n,X}\right] = 0.3883\,\sigma^{16/7} I_1^{4/7} I_2^{3/7} n^{-4/7}.$$

Afin d'estimer la fenêtre (h_n), nous devrons estimer I_1 et I_2. Dans [YS15], nous avons proposé les estimateurs suivants :

$$\begin{aligned}
\widehat{I}_1 &= \frac{Q_n}{n}\sum_{i,k=1}^{n} Q_k^{-1}\beta_k b_k^{-1} K_b^{\varepsilon}\left(\frac{Y_i - Y_k}{b_k}\right), \\
\widehat{I}_2 &= \frac{Q_n'^2}{n}\sum_{j\neq k}^{n} Q_j'^{-1} Q_k'^{-1}\beta_j\beta_k b_j'^{-2} b_k'^{-2} K_{b'}^{\varepsilon(1)}\left(\frac{Y_i - Y_j}{b_j'}\right) K_{b'}^{\varepsilon(1)}\left(\frac{Y_i - Y_k}{b_k'}\right),
\end{aligned}$$

où $Q_n = \prod_{i=1}^{n}(1-\beta_i)$, $Q_n' = \prod_{i=1}^{n}(1-\beta_i')$, K_b^{ε} est un noyau de déconvolution et b la fenêtre associé et $K_{b'}^{\prime\varepsilon}$ est la dérivée première du noyau de déconvolution $K_{b'}^{\varepsilon}$.
Nous devrons maintenant estimer les fenêtres b_n et b_n' et les pas (β_n) et (β_n'), pour cela, il fallait calculer le biais et la variance de $\widehat{I}_1$ et $\widehat{I}_2$ (pour plus de détails, consulter l'article [YS15]).

Bibliography

[1] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** : 1184–1186

[2] Carroll, R.J., Ruppert, D. & Stefanski, L. (1995). *Measurement Error in Nonlinear Models.* Chapman & Hall, London.

[3] Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** : 869–886.

[4] Delaigle, A. and Gijbels, I. (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.* **56** : 19–47.

[5] Delaigle, A. and Gijbels, I. (2004). Practical Bandwidth Selection in Deconvolution Kernel Density Estimation. *Comput. Statist. Data Anal.* **45** : 246–267.

[6] Fan, J. (1991). Asymptotic normality for deconvolution kernel density estimators. *Sankhya A* **53** : 97–110.

[7] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist* **19** : 1257–1272.

[8] Fan, J. (1991). Global behaviour of deconvolution kernel estimates. *Statist. Sinica* **1** : 541–551.

[9] Fan, J. (1992). Deconvolution with supersmooth distributions. *Canad. J. Statist* **20** : 155–169.

[10] Hesse, C. H. (1999). Data-driven deconvolution. *J. Nonparametr. Stat.* **10** : 343–373.

[11] Masry, A. (1993). Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes. *J. Multivariate Anal.* **44** : 47–68.

[12] Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Process. Appl.* **47** : 53–74.

[13] Meister, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Canad. J. Statist.* **32** : 439–449.

[14] Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statist. Sinica.* **16** : 195–211.

[15] Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics.* **2** : 169–184.

[16] Van, ES. A. J and Uh, H. W. (2004). Asymptotic normality of kernel type deconvolution estimators : Crossing the Cauchy boundary. *J. Nonparametr. Stat.* **16** : 261–277.

[17] Van, ES. A. J and Uh, H. W. (2005). Asymptotic normality of kernel type deconvolution estimators. *Scand. J. Stat.* **32** : 467–483.

[18] Zhang, S. and Karunamuni, R., (2000). Boundary bias correction for non-parametric deconvolution. *Ann. Inst. Statist. Math.* **52** : 612–629.

Chapitre 5

Grandes déviations et déviations modérées

Sommaire

Mots clés : Grandes déviations et déviations modérées, estimation à noyau d'une densité de probabilité, estimation d'une fonction de distribution, estimation d'une fonction de régression, algorithmes d'approximation stochastiques.

Résumé 4. *Dans ce chapitre, nous étudions le comportement asymptotiques des estimateurs récursifs d'une densité de probabilité, ainsi qu'une fonction de régression. Le but est d'établir certaines propriétés des estimateurs récursifs à noyau défini par des algorithmes stochastiques afin de comparer leur comportement asymptotique à celui des estimateurs classiques. Dans la première section, nous établissons des principes de grandes déviations (PGD) et des principes de déviations modérées (PDM) pour les estimateurs récursifs d'une densité de probabilité. Dans la deuxième section, nous établissons des PGD et des PDM pour*

des estimateurs récursifs à noyau de la régression défini par des algorithmes stochastiques.

5.1 Grandes déviations et déviations modérées pour les estimateurs récursifs à noyau de la densité défini par des algorithmes stochastiques

Soit $X_1, \ldots, X_n$ une suite de variable aléatoire indépendantes et de même loi, à valeurs dans $\mathbb{R}^d$ et de densité de probabilité f.

Les estimateurs récursifs pour estimer la densité f au point x introduit dans [YS1] s'écrivent sous la forme :

$$f_n(x) = (1-\gamma_n) f_{n-1}(x) + \gamma_n h_n^{-d} K\left(h_n^{-1}\left[x - X_n\right]\right),$$

avec (γ_n) une suite de réels positifs qui tend vers zéro, K un noyau (i.e. une fonction paire telle que $\int_{\mathbb{R}} K(x)\, dx = 1$) et h_n une fenêtre (i.e. une suite déterministe positive qui tend vers zéro).
L'objectif de l'article [YS4] était d'établir des principes de grandes déviations (PGD) et de déviations modérées (PDM) ponctuels et uniformes pour f_n.

Rappelons tout d'abord les notions de grandes déviations et de déviations modérées.

Définition 2. *Soit I une fonction définie sur $\mathbb{R}^m$ et à valeurs dans $[0, +\infty]$.*

- *On dit que I est une fonction de taux si ses ensembles de niveau sont fermés ; c'est-à-dire pour tout $\alpha \in \mathbb{R}$, l'ensemble $\{x,\, I(x) \leq \alpha\}$ est fermé.*
- *On dit que I est une bonne fonction de taux si ses ensembles de niveau sont compacts.*

Définition 3. *Une suite de vecteurs $(Z_n)_{n\geq 1}$ de $\mathbb{R}^m$ satisfait un PGD de vitesse (ν_n) et de fonction de taux I si*

1. *(ν_n) est une suite positive telle que $\lim_{n\to\infty} \nu_n = +\infty$;*
2. *Pour tout ouvert U de $\mathbb{R}^m$, et pour tout fermé V de $\mathbb{R}^m$,*

$$\begin{aligned} \liminf_{n\to\infty} \nu_n^{-1} \log \mathbb{P}\left[Z_n \in U\right] &\geq -\inf_{x\in U} I(x) \\ \limsup_{n\to\infty} \nu_n^{-1} \log \mathbb{P}\left[Z_n \in V\right] &\leq -\inf_{x\in V} I(x). \end{aligned}$$

Définition 4. *Soit (v_n) une suite réelle telle que $\lim_{n\to\infty} v_n = \infty$. On dit qu'une suite de vecteurs $(Z_n)_{n\geq 1}$ de $\mathbb{R}^m$ satisfait un PDM si la suite $(v_n Z_n)_{n\geq 1}$ satisfait un PGD.*

5.1.1 PGD ponctuel pour les estimateurs récursifs de la densité défini par les algorithmes stochastiques

Choix du pas (γ_n) qui minimise le risque intégré de f_n

Nos hypothèses sur le noyau K, la fenêtre (h_n) et le pas (γ_n) sont les suivantes :

(L1) $K : \mathbb{R}^d \to \mathbb{R}$ est une fonction bornée et intégrable, $\int_{\mathbb{R}^d} K(z)\,dz = 1$, et $\lim_{\|z\|\to\infty} K(z) = 0$.

(L2) i) $(h_n) = (cn^{-a})$ avec $a \in]0, 1/d[$ et $c > 0$.
ii) $(\gamma_n) = \left(n^{-1}\right)$.

Théorème 7 (PDG ponctuelle pour f_n)**.** *Supposons $(L1) - (L2)$ vérifiées, et supposons que f est continue en x. Alors la suite $(f_n(x) - f(x))$ satisfait un PGD de vitesse $\left(nh_n^d\right)$ et de bonne fonction de taux définie de la façon suivante :*

$$\begin{cases} si \quad f(x) \neq 0, \quad I_{a,x} : t \to f(x)\, I_a\left(1 + \frac{t}{f(x)}\right) \\ si \quad f(x) = 0, \quad I_{a,x}(0) = 0 \quad et \quad I_{a,x}(t) = +\infty \quad for \quad t \neq 0, \end{cases}$$

où

$$I_a(t) = \sup_{u\in\mathbb{R}} \{ut - \psi_a(u)\}$$
$$\psi_a(u) = \int_{[0,1]\times\mathbb{R}^d} s^{-ad}\left(e^{uK(z)} - 1\right) dsdz.$$

Remarque 2. *Sous les hypothèses $(L1) - (L2)$, la fonction de taux obtenue pour le PGD de l'estimateur récursif de la densité est la même fonction de taux obtenue pour le PGD de l'estimateur de Wolverton et Wagner [7] (voir [3]).*

Choix du pas (γ_n) qui minimise la variance de f_n

Nos hypothèses sur la fenêtre (h_n) et le pas (γ_n) sont les suivantes :

(L3) i) $(h_n) \in \mathcal{GS}(-a)$ avec $a \in]0, 1/d[$.
ii) $(\gamma_n) = \left(h_n^d \left(\sum_{k=1}^n h_k^d\right)^{-1}\right)$.

Théorème 8 (PDG ponctuelle pour f_n)**.** *Supposons $(L1)$ et $(L2)$ vérifiées, et que f est continue en x. Alors la suite $(f_n(x) - f(x))$ satisfait un PGD de vitesse $\left(nh_n^d\right)$ et de bonne fonction de taux définie de la façon suivante :*

$$\begin{cases} si \quad f(x) \neq 0, \quad I_x : t \to f(x)\, I\left(1 + \frac{t}{f(x)}\right) \\ si \quad f(x) = 0, \quad I_x(0) = 0 \quad et \quad I_x(t) = +\infty \quad for \quad t \neq 0, \end{cases}$$

où

$$I(t) = \sup_{u\in\mathbb{R}} \{ut - \psi(u)\}$$
$$\psi(u) = \int_{\mathbb{R}^d} \left(e^{uK(z)} - 1\right) dz.$$

Remarque 3. *Sous les hypothèses* $(L1)$ *et* $(L3)$*, la fonction de taux obtenue pour le PGD de l'estimateur récursif de la densité est la même fonction de taux obtenue pour le PGD de l'estimateur de Rosenblatt (voir [2]).*

5.1.2 PDM ponctuel pour les estimateurs récursifs de la densité défini par les algorithmes stochastiques

Soit (v_n) une suite positive. Afin de donner le comportement en déviations modérées des estimateurs récursifs de la densité défini par les algorithmes stochastiques, nous avons besoin des hypothèses suivantes :

(M1) $K : \mathbb{R}^d \to \mathbb{R}$ est continuoue, bornée satisfait $\int_{\mathbb{R}^d} K(z)\,dz = 1$, et pour tous $j \in \{1, \dots d\}$, $\int_{\mathbb{R}} z_j K(z)\,dz_j = 0$ et $\int_{\mathbb{R}^d} z_j^2 |K(z)|dz < \infty$.

(M2) *i)* $(\gamma_n) \in \mathcal{GS}(-\alpha)$ avec $\alpha \in]1/2, 1]$.
ii) $(h_n) \in \mathcal{GS}(-a)$ avec $a \in]0, \alpha/d[$.
iii) $\lim_{n\to\infty}(n\gamma_n) \in]\min\{2a, (\alpha - ad)/2\}, \infty]$.

(M3) f est bornée, deux fois différentiable, et pour tous $i, j \in \{1, \dots d\}$, $\partial^2 f/\partial x_i \partial x_j$ est borné.

(M4) $\lim_{n\to\infty} v_n = \infty$ et $\lim_{n\to\infty} \gamma_n v_n^2/h_n^d = 0$.

Théorème 9 (PDM ponctuelle pour f_n)**.** *Supposons les hypothèses* $(M1) - (M4)$ *vérifiées, et supposons que* f *est continue en* x*. Alors, la suite* $(f_n(x) - f(x))$ *satisfait un MDP de vitesse* $\left(h_n^d / \left(\gamma_n v_n^2\right)\right)$ *et de bonne fonction de taux* $J_{a,\alpha,x}$ *défine par :*

$$\begin{cases} si \quad f(x) \neq 0, \quad J_{a,\alpha,x} : t \to \frac{t^2(2-(\alpha-ad)\xi)}{2f(x)\int_{\mathbb{R}^d} K^2(z)dz} \\ si \quad f(x) = 0, \quad J_{a,\alpha,x}(0) = 0 \quad et \quad J_{a,\alpha,x}(t) = +\infty \quad for \quad t \neq 0. \end{cases}$$

5.2 Grandes déviations et déviations modérées pour les estimateurs récursifs à noyau de la régression défini par des algorithmes stochastiques

Soit $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ une suite de variable aléatoires indépendantes et identiquement distribuées à valeurs dans $\mathbb{R} \times \mathbb{R}$ de densité de probabilité $f(x, y)$ avec $\mathbb{E}|Y| < \infty$. Soient $f(x)$ la densité marginale de X et $r(x) = \mathbb{E}(Y|X = x)$ la régression de Y sur X.

5.2.1 Déviations modérées pour les estimateurs récursifs à noyau de la régression défini par des algorithmes stochastiques

L'objectif de l'article [YS12] était d'établir le comportement en déviations de l'estimateur de Révész généralisé défini par :

$$r_n(x) \quad = \quad \left(1-\gamma_n h_n^{-1} K\left(h_n^{-1}\left[x-X_n\right]\right)\right) r_{n-1}(x)+\gamma_n h_n^{-1} Y_n K\left(h_n^{-1}\left[x-X_n\right]\right).$$

Nous avons montré qu'avec un choix de pas $(\gamma_n) \in \mathcal{GS}(-\alpha)$, avec $\alpha \in]3/4,1]$, et une fenêtre $(h_n) \in \mathcal{GS}(-a)$, $a \in]\frac{1-\alpha}{4}, \frac{\alpha}{3}[$, sous les conditions suivantes sur le pas (γ_n) et la fenêtre (h_n) :

$$\lim_{n\to\infty} v_n = \infty, \qquad \lim_{n\to\infty} \frac{\gamma_n v_n^2}{h_n} = 0 \quad \text{et} \quad \lim_{n\to\infty} v_n h_n^2 = 0,$$

la suite

$$v_n\left(r_n(x)-r(x)\right)$$

satisfait un PGD de vitesse $\left(h_n/\left(\gamma_n v_n^2\right)\right)$ et de bonne fonction de taux $J_{a,\alpha,x}(.)$ définie par

$$\begin{cases} si \quad f(x) \neq 0, \quad J_{a,\alpha,x} : t \to \frac{t^2(2f(x)-(\alpha-a)\xi)}{2f(x)Var[Y|X=x]\int_{\mathbb{R}} K^2(z)dz} \\ si \quad f(x) = 0, \quad J_{a,\alpha,x}(0) = 0 \quad et \quad J_{a,\alpha,x}(t) = +\infty \quad \text{for} \quad t \neq 0. \end{cases}$$

5.2.2 Grandes déviations et déviations modérées pour les estimateurs de Révész moyennisé

L'objectif de l'article [YS11] était d'établir le comportement en déviations de l'estimateur de Révész moyennisé défini par :

$$\overline{r}_n(x) \quad = \quad \frac{1}{\sum_{k=1}^n q_k} \sum_{k=1}^n q_k r_k(x),$$

avec

$$r_k(x) \quad = \quad \left(1-\gamma_k h_k^{-1} K\left(h_k^{-1}\left[x-X_k\right]\right)\right) r_{k-1}(x)+\gamma_k h_k^{-1} Y_k K\left(h_k^{-1}\left[x-X_k\right]\right),$$

et (q_n) une suite positive tel que $\sum q_n = \infty$. Le premier objectif était d'établir le PGD de l'estimateur de Révész moyennisé. Il se trouve que la fonction de taux dépend du choix de la fenêtre (h_n) et de la suite de poids (q_n).

Nous avons montré qu'avec un choix particulier de la fenêtre $(h_n) \equiv (cn^{-a})$ avec $c > 0$ et $a \in (1-\alpha, (4\alpha-3)/2)$ (avec $\alpha \in (\frac{3}{4}, 1]$), et la suite de poids $(q_n) = (c'n^{-q})$ avec $c' > 0$ et $q < \min\{1-2a, (1+a)/2\}$, la suite $(\overline{r}_n(x) - r(x))$ satisfait un PGD de vitesse (nh_n) et de bonne fonction de taux définie par

$$I_{a,q,x}(t) = \sup_{u\in\mathbb{R}} \{ut - \psi_{a,q,x}(u)\},$$

qui est la transformation de Fenchel-Legendre de la fonction $\psi_{a,q,x}$ définie par :

$$\psi_{a,q,x}(u) = (1-q) \int_{[0,1]\times\mathbb{R}^2} s^{-a} \left(e^{us^{a-q}K(z)\frac{(y-r(x))}{f(x)}} - 1\right) g(x,y)\, dsdzdy.$$

Nous avons souligné que, dans le cas speciale $(q_n) = (h_n)$, qui représente le cas où la suite de poids (q_n) minimise la variance asymptotique de $\overline{r}_n$ (voir [YS2]), nous obtenons la même fonction de taux dans le cas du PGD ponctuelle que celle obtenue pour l'estimateur de Nadaraya-Watson (voir [1]).

Notre deuxième objectif était d'établir le comportement en déviations modérées des estimateurs de Révész moyennisé.

Pour toute suite positive (v_n) telle que

$$\lim_{n\to\infty} v_n = \infty, \qquad \lim_{n\to\infty} \frac{v_n^2}{nh_n} = 0 \quad \text{et} \quad \lim_{n\to\infty} v_n h_n^2 = 0 \tag{5.1}$$

et une fenêtre $(h_n) \in \mathcal{GS}(-a)$, nous avons montré que la suite

$$v_n \left(\overline{r}_n(x) - r(x)\right)$$

satisfait une PGD de vitesse $\left(nh_n/v_n^2\right)$ et de bonne fonction de taux $J_{a,q,x} : \mathbb{R} \to \mathbb{R}$ définie par

$$J_{a,q,x}(t) = \frac{1+a-2q}{(1-q)^2} \frac{f(x)}{Var[Y|X=x] \int_{\mathbb{R}} K^2(z)\,dz} \frac{t^2}{2}.$$

Rappelons que lorsque la suite (q_n) est à variation régulière d'ordre $(-a)$, qui représente le cas où la suite (q_n) minimise la variance asymptotique de $\overline{r}_n$ (voir [YS2]), le facteur $(1+a-2q)/(1-q)^2$ peut être réduit à $1/(1-a)$, donc nous pouvons écrire

$$J_{a,x}(t) = \frac{1}{(1-a)} \frac{f(x)}{Var[Y|X=x] \int_{\mathbb{R}} K^2(z)\,dz} \frac{t^2}{2}. \tag{5.2}$$

Louani [1] a établit le comportement en déviations modérées de l'estimateur de Nadaraya-Watson ([5], [6]) define par

$$\widehat{r}_n(x) = \begin{cases} \frac{\widehat{m}_n(x)}{\widehat{f}_n(x)} & si \quad \widehat{f}_n(x) \neq 0 \\ 0 & sinon, \end{cases} \tag{5.3}$$

où

$$\widehat{m}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h_n}\right) \quad \text{et} \quad \widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right).$$

Il a montré que pour toute suite positive (v_n) qui satisfait (5.1), la suite $v_n\left(\widehat{r}_n(x) - r(x)\right)$ satisfait un PGD avec une vitesse $\left(nh_n/v_n^2\right)$ et une bonne fonction de taux $\widehat{J}_x : \mathbb{R} \to \mathbb{R}$ définie par

$$\widehat{J}_x(t) = \frac{f(x)}{Var[Y|X=x] \int_{\mathbb{R}} K^2(z)\,dz} \frac{t^2}{2}. \tag{5.4}$$

Mokkadem et al. [4] ont établif le comportement en déviations modérées de la version semi-recursive de l'estimateur de Nadaraya-Watson définie par

$$\tilde{r}_n(x) = \begin{cases} \frac{\tilde{m}_n(x)}{\tilde{f}_n(x)} & si \quad \tilde{f}_n(x) \neq 0 \\ 0 & sinon, \end{cases} \tag{5.5}$$

où

$$\tilde{m}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{h_i} K\left(\frac{x - X_i}{h_i}\right) \quad \text{et} \quad \tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right).$$

Ils ont montré que, pour toute suite positive (v_n) telle que (5.1), la suite $v_n(\tilde{r}_n(x) - r(x))$ satisfait un PGD de vitesse $\left(nh_n/v_n^2\right)$ et une bonne fonction de taux $\tilde{J}_{a,x} : \mathbb{R} \to \mathbb{R}$ définie par

$$\tilde{J}_{a,x}(t) = (1+a)\frac{f(x)}{Var[Y|X=x]\int_{\mathbb{R}} K^2(z)\,dz}\frac{t^2}{2}. \tag{5.6}$$

Nous pouvons constater à partir de (5.2), (5.4) et (5.6) que la fonction de taux qui apparaît dans les PDM pour l'estimateur de Révész moyennisé défini avec le suite de poids (q_n) qui minimise la variance asymptotique de $\overline{r}_n$ est plus grande que celle obtenue pour l'estimateur semi-recursif (5.5) qui est plus grande que celle obtenue pour l'estimateur de Nadaraya-Watson (5.3) ; ce qui signifie que l'estimateur de Révész moyennisé $\overline{r}_n(x)$ défini avec le choix de $(q_n) \in \mathcal{GS}(-a)$, est plus concentré autour de $r(x)$ que les deux autres estimateurs de la régression.

Bibliography

[1] Louani, D. (1999). Some large deviations limit theorems in conditionnal nonparametric statistics, *Statistics*. **33** : 171–196.

[2] Mokkadem, A., Pelletier, M. and Worms, J. (2005). Large and moderate deviations principles for kernel estimation of a multivariate density and its partial derivatives, *Austral. J. Statist.* **4** : 489–502.

[3] Mokkadem, A., Pelletier, M. and Thiam, B. (2006). Large and moderate deviations principles for recursive kernel estimation of a multivariate density and its partial derivatives, *Serdica Math. J.* **32** : 323–354.

[4] Mokkadem, A., Pelletier, M. and Thiam, B. (2008). Large and moderate deviations principles for kernel estimators of the multivariate regression, *Math. Methods Statist*, **17** : 1–27.

[5] Nadaraya, E. A. (1964). On estimating regression, *Theory Probab. Appl*, **10** : 186–190.

[6] Watson, G. S. (1990). Smooth regression analysis, *Sankhya Ser. A*, **26** : 359–372.

[7] Wolverton, C. T. and Wagner, T. J. (1969). Asymptotically optimal discriminant functions for pattern classification. *IEEE Trans. Inform. Theory* **15** : 258–265.

Chapitre 6

Censures des données dans un modèle linéaire mixte

Sommaire

Mots clés : Echantillonneur de Gibbs, Algorithme EM, Algorithme SEM, Modèle linéaire mixte, censure des données, distribution tronquée.

Résumé 5. *Dans ce chapitre, nous proposons une approche basée sur l'algorithme SEM (Stochastique Espérance Maximisation) qui fait appel à l'échantillonneur de Gibbs pour faire face au problème de la censure des données dans la réponse d'un modèle linéaire mixte. Nous avons comparé notre approche avec des méthodes existantes en considérant à la fois des données simulées et des données réelles. Les résultats obtenus montrent que notre approche est plus performante que les autres approches en terme de précision et d'efficacité.*

6.1 Introduction

Les modèles linéaires à effets mixtes (LEM) ([5], [10]) ont été largement utilisés dans de nombreuses applications de recherche, notamment en économie,

en sociologie, en assurance, en agronomie et en génétique. Il y a eu une discussion intensive dans la littérature sur les méthodes d'estimation de ces modèles ([11] ; [14] ; [12]).

Cependant, l'estimation du LEM souffre souvent du problème de la censure des données. Plusieurs stratégies ont été utilisé pour surmonter ce problème. Tout d'abord, des approches naïves qui consistent soit à ignorer les données censurées, soit de les remplacer par des imputations numériques. Le principal intérêt d'une telle approche est qu'une méthode d'estimation classique peut être utilisé sur l'ensemble des données. L'inconvénient est qu'un biais peut être introduit dans le processus puisque les propriétés statistiques de l'estimateur résultant ne sont par claires ([15] et [6]).

Des approches plus performantes consistent à prendre en compte la censure des données dans le calcul de la vraisemblance. Pour ce faire, Hughes [8] suggère de considérer les effets aléatoires et les données censurées comme des données manquantes ainsi il utilise un algorithme de Monte Carlo EM (MCEM). Une approche alternative introduite par Jacqmin-Gadda et al. [9] consiste à introduire un modèle d'erreur autorégressif et de maximiser la vraisemblance en utilisant les algorithmes du Simplexe et de Marquardt.

Le problème de la censure des données a été également considéré dans le cas des modèles non linéaire à effets mixtes (NLEM). Ces approches consistent à résoudre le problème en utilisant une extension de l'algorithme EM : MCEM [19], SAEM [16], HEM [17], toutes ces méthodes sont statistiquement significatives et consistantes, elles aboutissent à des algorithmes sophistiqués qui ne sont pas faciles à comprendre ou à mettre en œuvre.

Une alternative intéressante pour surmonter le problème de la censure des données dans un modèle linéaire mixte est basée sur l'imputation multiple (MI). Fondamentalement, MI est une technique de Monte Carlo, dans laquelle les valeurs manquantes sont remplacées par des valeurs imputées et les résultats obtenus sont combinés pour produire des estimations afin d'incorporer les données manquantes, Rubin [15] en LMM et Wu et Wu [18] et Fitzgerald et al. [6] en NLMM, ces derniers ont montré que leur méthode d'imputation multiple à un biais inférieur à 5% et une couverture raisonnable (88% pour les intervalles de confiance à 95%) par rapport à Hughes [8]. Il est néanmoins important de souligner que la performance d'une méthode MI dépend fortement de la pertinence du modèle utilisé pour les imputations. En collaboration avec Grégory Nuel nous avons montré dans l'article [YS9] que l'utilisation de la distribution conditionnelle pour imputer les données censurées, nous permet d'obtenir un algorithme SEM classique, dont les propriétés de convergence sont bien connues ([13]). Les avantages de la méthode proposée sont à la fois d'être statistiquement fondée et facilement implémentable en utilisant des logiciels existants.

6.2 Méthode et notations

6.2.1 Modéle à effets aléatoires hiérarchiques

Dans les études longitudinales, les sujets sont mesurés plusieurs fois au cours du temps en plus ils sont souvent emboîtés dans des groupes. Les études des données à plusieurs niveaux ont conduit à l'introduction de nombreuses méthodes statistiques, mentionné sous un certain nombre de noms, les modèles linéaires hiérarchiques (MLH, [1]), la modélisation à plusieurs niveaux ([7]). La forme générale des modèles à effet aléatoire hiérarchique avec p effets fixes et q effets aléatoires hiérarchiques.

$$\underbrace{Y_{i_1,\ldots,i_q,k}}_{\substack{\text{k-ième mesure de la variable réponse}}} = \underbrace{\beta_0 + \beta_1 X^1_{i_1,\ldots,i_q,k} + \ldots + \beta_p X^p_{i_1,\ldots,i_q,k}}_{\text{effets fixes } F_{i_1,\ldots,i_q,k}} + \underbrace{b_{i_1} + \ldots + b_{i_1,\ldots,i_q}}_{\text{effets aléatoires } B_{i_1,\ldots,i_q}} + \underbrace{\varepsilon_{i_1,\ldots,i_q,k}}_{\text{erreur}}.$$

On note $\mathcal{I}$ l'ensemble de tous les $(i_1, \ldots, i_q, k)$, et par $\mathbf{Y} = \{Y_{i_1,\ldots,i_q,k}, (i_1, \ldots, i_q, k) \in \mathcal{I}\}$ le vecteur de toutes les valeurs de la variable réponse. Nous supposons que les effets aléatoires et l'erreur du modèle sont indépendants et que $b_{i_1} \sim \mathcal{N}(0, \sigma_1^2)$, $\ldots$, $b_{i_1,\ldots,i_q} \sim \mathcal{N}(0, \sigma_q^2)$, et $\varepsilon_{i_1,\ldots,i_q,k} \sim \mathcal{N}(0, \sigma^2)$. On note $\boldsymbol{\Theta} = \left(\beta_0, \beta_1, \ldots, \beta_p, \sigma_1^2, \ldots, \sigma_q^2, \sigma^2\right)$ l'ensemble des paramètres du modèle.

6.2.2 Censure des données

Nous avons considéré seulement le cas de la censure à gauche des données, le cas de la censure à droite ainsi que le cas de la censure à droite et à gauche peut être déduit. Soit t le niveau de la censure, $\mathcal{T} = \left\{(i_1, \ldots, i_q, k) \in \mathcal{I}, Y_{i_1,\ldots,i_q,k} < t\right\}$ l'ensemble des indices des données censurées. On note $\mathbf{Y} = (\mathbf{Y}^{\mathrm{O}}, \mathbf{Y}^{\mathrm{C}})$ avec $\mathbf{Y}^{\mathrm{O}} = \left\{Y_{i_1,\ldots,i_q,k}, (i_1, \ldots, i_q, k) \notin \mathcal{T}\right\}$ le vecteur des valeurs observées de la variable réponse, $\mathbf{Y}^{\mathrm{C}} = \left\{Y_{i_1,\ldots,i_q,k}, (i_1, \ldots, i_q, k) \in \mathcal{T}\right\}$ le vecteur des valeurs censurées de la variable réponse.

6.3 Algorithme EM et sa version stochastique

L'algorithme EM permet d'obtenir les estimateurs du maximum de vraisemblance dans le cadre des modèles paramétriques quand les données observées peuvent être vu comme des données incomplètes. Notons par X les variables observées, par S les variables latentes et par $\mathbb{P}_\Theta$ la distribution associée aux paramètres Θ du modèle. Notre objectif est d'estimer les paramètres du modèle :

$$\widehat{\Theta} = \arg\max_{\Theta} \mathbb{P}_\Theta\left(X\right),$$

avec $\mathbb{P}_\Theta(X)$ représente la vraisemblance des données incomplètes :

$$\mathbb{P}_\Theta(X) = \int_S \mathbb{P}_\Theta(X,S)\, dS,$$

où $\mathbb{P}_\Theta(X,S)$ représente la vraisemblance des données complètes. Nous pouvons remarquer que :

$$\log\left[\mathbb{P}_\Theta(X)\right] = \log\left[\mathbb{P}_\Theta(X,S)\right] - \log\left[\mathbb{P}_\Theta(S|X)\right],$$

et par passage à l'espérance conditionnelle, nous obtenons :

$$\log\left[\mathbb{P}_\Theta(X)\right] = \mathbb{E}\left\{\log\left[\mathbb{P}_\Theta(X,S)\right]|X\right\} - \mathbb{E}\left\{\log\left[\mathbb{P}_\Theta(S|X)\right]|X\right\}.$$

L'algorithme EM consiste en l'optimisation indirecte de la log-vraisemblance des données observées à travers l'optimisation itérative de l'espérance conditionnelle de la log-vraisemblance des données complètes.

Algorithme EM (Dempster et al., [5]) :

- Choix arbitraire de $\Theta^{(0)}$
- Pour $i = 1, \ldots, N$:
  - Étape Espérance : calcule de

$$Q\left(\Theta|\Theta^{(i-1)}\right) = \int_S P_{\Theta^{(i-1)}}(S|X) \log \mathbb{P}_\Theta(X,S)\, dS.$$

  - Étape Maximisation : Actualisation des paramètres avec

$$\Theta^{(i)} = \arg\max_{\Theta^{(i-1)}} Q\left(\Theta|\Theta^{(i-1)}\right).$$

Les étapes Espérance et Maximisation sont itérées jusqu'à la convergence de l'algorithme.

La propriété fondamentale de l'algorithme EM, est que la log-vraisemblance des données observées augmente d'une itération à une autre :

$$\log\left[\mathbb{P}_\Theta(X)|\Theta^{(i)}\right] \geq \log\left[\mathbb{P}_\Theta(X)|\Theta^{(i-1)}\right].$$

En effet, l'étape Maximisation garantit

$$\mathbb{E}\left\{\log\left[\mathbb{P}_{\Theta^{(i)}}(X,S)\right]|X\right\} \geq \mathbb{E}\left\{\log\left[\mathbb{P}_{\Theta^{(i-1)}}(X,S)\right]|X\right\},$$

et l'application de l'inégalité de Jensen permet d'avoir

$$\mathbb{E}\left\{\log\left[\mathbb{P}_{\Theta^{(i)}}(S|X)\right]|X\right\} \leq \mathbb{E}\left\{\log\left[\mathbb{P}_{\Theta^{(i-1)}}(S|X)\right]|X\right\}.$$

Ainsi, la quantité $\mathbb{E}\left\{\log\left[\mathbb{P}_{\Theta^{(i)}}(S|X)\right]|X\right\}$ diminue à chaque itération, ce qui assure que $\log\left[\mathbb{P}_\Theta(X)|\Theta\right]$ tend vers un maximum local si la log-vraisemblance des données observées est majorée. Cependant l'algorithme EM souffre de certaines limitations telles que la dépendance aux valeurs initiales, la convergence vers un

point selle ou un optimum local, la vitesse de convergence peut être très lente et le calcule de la quantité $Q\left(\Theta|\Theta^{(i-1)}\right)$ peut être une tâche très lourde.
Afin de répondre à ces limitations, Celeux et Diebolt [2] ont introduit l'algorithme SEM, qui a pour principe de maximiser la log-vraisemblance des données complètes grâce à une évaluation numérique à travers l'introduction d'une étape stochastique entre les étapes Espérance et Maximisation. L'étape stochastique consiste à générer un échantillon pseudo complet en générons les données manquantes à partir de la densité conditionnellement aux données observées.
Algorithme SEM (Celeux et Diebolt [2])

- générer $s_1, \ldots, s_M$ à partir de $P_{\Theta^{(i-1)}}(S|X)$
- $\Theta^{(i)} = \arg\max_\Theta \frac{1}{M}\sum_{j=1}^{M} \log \mathbb{P}_\Theta(X, S = s_j)$.

De plus le théorème de la loi des grands nombre assure que

$$\lim_{M\to\infty} \frac{1}{M}\sum_{j=1}^{M} \log \mathbb{P}_\Theta(X, S = s_j) = \underbrace{\int_S P_{\Theta^{(i-1)}}(S|X) \log \mathbb{P}_\Theta(X, S)\, dS}_{Q\left(\Theta|\Theta^{(i-1)}\right)}.$$

Dans notre cadre de la censure des données dans la réponse d'un modèle linéaire mixte, nous avons proposé dans l'article [YS9] l'algorithme suivant :
Algorithme SEM (Slaoui et Nuel [YS9]) :

- Choix arbitraire de $\Theta^{(0)}$
- Pour $i = 1, \ldots, N$:
  - générer Y^c à partir de $P_{\Theta^{(i-1)}}(Y^c|Y^o)$, `en utilisant l'échantillonneur de Gibbs`
  - calculer $\Theta^{(i)} = \arg\max_\Theta \mathbb{P}_\Theta(Y^o, Y^c)$, `en utilisant des estimateurs standards des modèles linéaires mixtes`.

Nous avons donc besoin de générer des échantillons à partir de la loi conditionnelle. Pour ce faire nous avons utilisé l'algorithme de Gibbs sampling.

Notons qu'une loi jointe est caractérisée par l'ensemble de ces lois conditionnelles. En dimension deux, si la densié jointe $f(x, y)$ a des lois conditionnelles notées $f(x|y)$ et $f(y|x)$ alors le théorème de Hammersley et Clifford assure que

$$f(x, y) = \frac{f(y|x)}{\int \frac{f(y|x)}{f(x|y)} dy}.$$

En dimension supérieure ou égale à trois, il est nécessaire d'utiliser des algorithmes itératifs, tel que l'échantillonneur de Gibbs qui permet de simuler une distribution jointe à partir des simulations successives des lois conditionnelles.
Déroulement de l'algorithme de Gibbs Sampling :
Nous considérons trois variables aléatoires X, Y et Z. Le processus d'échantillonnage

se déroule comme suit

$$\begin{aligned} X^{(j+1)} &\sim f\left(X|Y^{(j)}, Z^{(j)}\right) \\ Y^{(j+1)} &\sim f\left(Y|X^{(j+1)}, Z^{(j)}\right) \\ Z^{(j+1)} &\sim f\left(Z|X^{(j+1)}, Y^{(j+1)}\right) \end{aligned}$$

ce qui implique que, $\left(X^{(t)}, Y^{(t)}, Z^{(t)}\right)$ défini une chaîne de Markov qui converge vers la loi de (X, Y, Z), (la loi stationnaire de cette chaîne de Markov).
Illustration de l'algorithme de Gibbs Sampling :
Nous considérons que $\varepsilon \sim \mathcal{N}(0, 2)$, $Y \sim \mathcal{N}(0, 1)$ et $Z \sim \mathcal{N}(0, 1)$, sous les contraintes suivantes :

$$\left\{ \begin{array}{rcl} Y + \varepsilon & < & t, \\ Z + \varepsilon & < & t. \end{array} \right.$$

avec $t = 0.4$.
Pour cela, nous commençons par la génération d'une valeur initiale $\varepsilon^{(0)}$ à partir de $\mathcal{N}(0, 2, \texttt{lower} = t)$ qui représente une distribution normale tronquée à gauche en t. Donc, pour $j = 0, 1, \ldots, M-1$, nous générons

$$\begin{array}{ll} Y^{(j+1)} & \text{à partir de } \mathcal{N}\left(0, 1, \texttt{upper} = t - \varepsilon^{(j)}\right) \\ Z^{(j+1)} & \text{à partir de } \mathcal{N}\left(0, 1, \texttt{upper} = t - \varepsilon^{(j)}\right) \\ \varepsilon^{(j+1)} & \text{à partir de } \mathcal{N}\left(0, 2, \texttt{upper} = \min\left(t - Y^{(j+1)}, t - Z^{(j+1)}\right)\right). \end{array}$$

Une distribution normale tronquée est définie pour toutes x, tel que $-\infty \leqslant a \leqslant x \leqslant b \leqslant +\infty$, par :

$$f(x; \mu, \sigma, \texttt{lower} = a, \texttt{upper} = b) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}{\int_a^b \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right) du}.$$

Dans notre cadre nous avons proposé l'algorithme suivant :
Algorithme de Gibbs Sampling (Slaoui et Nuel [YS9])

- Choix initial arbitraire de $(\mathbf{b}_1^{(0)}, \ldots, \mathbf{b}_q^{(0)}, \mathbf{e}^{(0)})$
- Pour $g = 1, \ldots, G$
  - Générer $\varepsilon_{i_1,\ldots,i_q,k}^{(g)}$ à partir de $\mathcal{N}\left(0, \sigma^2, \texttt{upper} = t - F_{i_1,\ldots,i_q,k} - B_{i_1,\ldots,i_q}^{(g-1)}\right)$;
  - Pour $r = 1, \ldots, q$
  - Générer $b_{i_1,\ldots,i_r}$ à partir $\mathcal{N}\left(0, \sigma_r^2, \texttt{upper} = a_{i_1,\ldots,i_r}\right)$ où

$$\begin{aligned} a_{i_1,\ldots,i_r} = \min\{ & t - F_{i_1,\ldots,i_q,k} - \left(B_{i_1,\ldots,i_q}^{(g-1)} - b_{i_1,\ldots,i_r}^{(g-1)}\right) - \varepsilon_{i_1,\ldots,i_q,k}^{(g)}, \\ & \texttt{pour } (i_{r+1}, \ldots, i_q, k) \texttt{ tel que } (i_1, \ldots, i_q, k) \in \mathcal{T} \end{aligned}$$

- Output : $\mathbf{Y}^{\mathbf{c}}$ obtenu à partir de $(\mathbf{b}_1^{(G)}, \ldots, \mathbf{b}_q^{(G)}, \mathbf{e}^{(G)})$

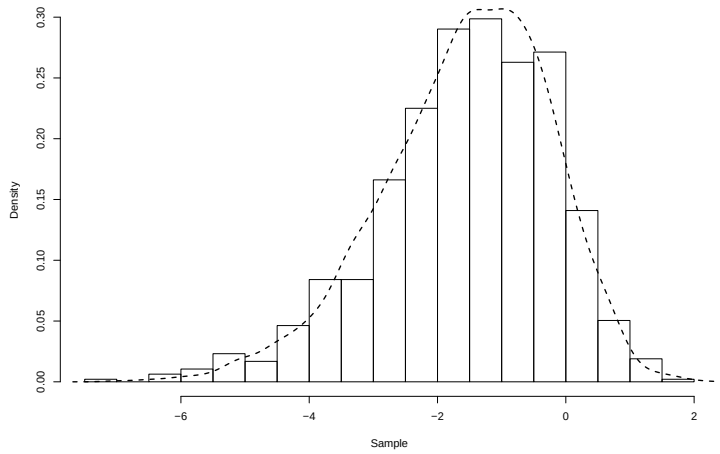

FIGURE 6.1 – Nous comparons les densités de ε et celle de $\varepsilon^{(M)}$, avec $M = 10000$. Cet exemple simple illustre le chéma de l'échantillonneur de Gibbs pour la génération d'une variable aléatoire lorsque la loi jointe n'est pas facilement calculable.

6.3.1 Exemple :

Nous considèrons un modèle à effet aléatoire hiérarchique avec p effets fixes et $q = 2$ effets aléatoires hiérarchiques.

$$Y_{i_1,i_2,k} = F_{i_1,i_2,k} + b_{i_1} + b_{i_1,i_2} + \varepsilon_{i_1,i_2,k},$$

avec $Y_{i_1,i_2,k}$ est la $k$ième mesure de l'enfant i_2 de la famille i_1. Nous supposons que nous avons un totale de 9 observations et que 5 d'entre eux ont été censurés :

$$\begin{cases} Y_{1,1,1} &= F_{1,1,1} + b_1 + b_{1,1} + \varepsilon_{1,1,1} &< t \\ Y_{1,1,2} &= F_{1,1,2} + b_1 + b_{1,1} + \varepsilon_{1,1,2} & \\ Y_{1,2,1} &= F_{1,2,1} + b_1 + b_{1,2} + \varepsilon_{1,2,1} & \\ Y_{2,1,1} &= F_{2,1,1} + b_2 + b_{2,1} + \varepsilon_{2,1,1} &< t \\ Y_{2,2,1} &= F_{2,2,1} + b_2 + b_{2,2} + \varepsilon_{2,2,1} &< t \\ Y_{2,2,2} &= F_{2,2,2} + b_2 + b_{2,2} + \varepsilon_{2,2,2} & \\ Y_{2,2,3} &= F_{2,2,3} + b_2 + b_{2,2} + \varepsilon_{2,2,3} & \\ Y_{3,1,1} &= F_{3,1,1} + b_3 + b_{3,1} + \varepsilon_{3,1,1} &< t \\ Y_{3,1,2} &= F_{3,1,2} + b_3 + b_{3,1} + \varepsilon_{3,1,2} &< t \end{cases}$$

Dans cet exemple, $\mathcal{T} = \{(1,1,1),(2,1,1),(2,2,1),(3,1,1),(3,1,2)\}$ représente l'ensemble des indices pour lesquelles nous avons des censures, $\mathcal{T}_1 = \{(1),(2),(3)\}$ représente l'ensemble des familles pour lesquelles nous avons des censures et $\mathcal{T}_2 = \{(1,1),(2,1),(2,2),(3,1)\}$ représente l'ensemble des enfants des familles pour lesquelles nous avons des censures. De plus, $\mathbf{e} = \{\varepsilon_{1,1,1}, \varepsilon_{2,1,1}, \varepsilon_{2,2,1}, \varepsilon_{3,1,1}, \varepsilon_{3,1,2}\}$, $\mathbf{b}_1 = \{b_1, b_2, b_3\}$, et $\mathbf{b}_2 = \{b_{1,1}, b_{2,1}, b_{2,2}, b_{3,1}\}$. Notre algorithme de Gibbs sam-

pling à l'itération g s'écrit sous la forme suivante :

$$
\begin{aligned}
\varepsilon_{1,1,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma, \texttt{upper} = t - b_{1,1}^{(g-1)} - b_{1}^{(g-1)} - F_{1,1,1}\right), \\
b_{1}^{(g)} &\sim \mathcal{N}\left(0, \sigma_1, \texttt{upper} = t - \varepsilon_{1,1,1}^{(g)} - b_{1,1}^{(g-1)} - F_{1,1,1}\right), \\
b_{1,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma_2, \texttt{upper} = t - \varepsilon_{1,1,1}^{(g)} - b_{1}^{(g)} - F_{1,1,1}\right), \\
\varepsilon_{2,1,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma, \texttt{upper} = t - b_{2,1}^{(g-1)} - b_{2}^{(g-1)} - F_{2,1,1}\right), \\
\varepsilon_{2,2,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma, \texttt{upper} = t - b_{2,2}^{(g-1)} - b_{2}^{(g-1)} - F_{2,2,1}\right), \\
b_{2}^{(g)} &\sim \mathcal{N}\left(0, \sigma_1, \texttt{upper} = \min\left\{t - \varepsilon_{2,1,1}^{(g)} - b_{2,1}^{(g-1)} - F_{2,1,1}; t - \varepsilon_{2,2,1}^{(g)} - b_{2,2}^{(g-1)} - F_{2,2,1}\right\}\right), \\
b_{2,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma_2, \texttt{upper} = t - \varepsilon_{2,1,1}^{(g)} - b_{2}^{(g)} - F_{2,1,1}\right), \\
b_{2,2}^{(g)} &\sim \mathcal{N}\left(0, \sigma_2, \texttt{upper} = t - \varepsilon_{2,2,1}^{(g)} - b_{2}^{(g)} - F_{2,2,1}\right), \\
\varepsilon_{3,1,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma, \texttt{upper} = t - b_{3,1}^{(g-1)} - b_{3}^{(g-1)} - F_{3,1,1}\right), \\
\varepsilon_{3,1,2}^{(g)} &\sim \mathcal{N}\left(0, \sigma, \texttt{upper} = t - b_{3,1}^{(g-1)} - b_{3}^{(g-1)} - F_{3,1,2}\right), \\
b_{3}^{(g)} &\sim \mathcal{N}\left(0, \sigma_1, \texttt{upper} = \min\left\{t - \varepsilon_{3,1,1}^{(g)} - b_{3,1}^{(g-1)} - F_{3,1,1}; t - \varepsilon_{3,1,2}^{(g)} - b_{3,1}^{(g-1)} - F_{3,1,2}\right\}\right), \\
b_{3,1}^{(g)} &\sim \mathcal{N}\left(0, \sigma_2, \texttt{upper} = \min\left\{t - \varepsilon_{3,1,1}^{(g)} - b_{3}^{(g)} - F_{3,1,1}; t - \varepsilon_{3,1,2}^{(g)} - b_{3}^{(g)} - F_{3,1,2}\right\}\right).
\end{aligned}
$$

Ensuite, nous gardons les effets aléatoires et les erreurs collectées lors de la dernière étape.

6.4 Application : Paludisme

Nous avons considéré un jeu de données de 176 familles Sénégalaise, 505 enfants âgés entre 2 et 19 ans de deux villages de Niakhar (Diohine et Toucar). Le nombre d'observations est 6986. Nous avons mesuré la charge parasitaire dans le sang `Plasmodium falciparum` pendant deux saisons différentes et sur une période d'observation de trois ans (2001-2003), le nombre de mesures par enfant vari de 1 à 15, pour plus de détails, voir ([YS3]).

Nous avons sélectionné les variables suivantes :
- `identification de la famille` : un facteur à 176 niveaux.
- `identification de l'enfant` : un facteur à 505 niveaux.
- `CP` : Charge parasitaire.
- `infection` : un facteur à deux niveaux (infecté : 1 ou non infecté : 0).
- `année` : un facteur à trois niveaux (0 pour 2001, 1 pour 2002 et 2 pour 2003).
- `nombre de mesures par enfant` : un facteur à 15 niveaux.
- `âge` : Âge de l'enfant entre 2 et 19 ans.

– `saison` : un facteur à deux niveaux (Juillet-Octobre et Octobre-Mars).
– `village` : un facteur à deux niveaux (Diohine et Toucar).

Notons que, le jeu de données considéré est complet, donc, le calcul de la vraisemblance peut se faire comme dans Laird et Ware [10]. Nous avons effectué une sélection de modèle en utilisant le critère `BIC` (Bayesian Information Criterion, voir Chabanet et Pineau [3]), ce qui nous a permis de choisir le modèle suivant :

$$\begin{aligned}\log(\texttt{CP})_{i_1,i_2,k} &= \texttt{A1} + \texttt{A2} * \hat{\texttt{age}}_{i_1,i_2,k} + \texttt{A3} * \texttt{saison}_{i_1,i_2,k} + \texttt{A4} * \texttt{infection}_{i_1,i_2,k} \\ &\quad + \texttt{A5} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=1} + \texttt{A6} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=2} + b_{i_1} + b_{i_1,i_2} + \varepsilon_{i_1,i_2,k},\end{aligned}$$

ce qui signifie que nous considérons les cinq effets fixes : $\hat{\texttt{age}}$, `saison`, `infection`, la deuxième `année` (2002) et la troisième `année` (2003) (en outre l'ordonnée à l'origine) aussi bien qu'un effet aléatoire lié à chaque famille, $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$ et un effet aléatoire lié à chaque enfant de chaque famille, $b_{i_1,i_2} \sim \mathcal{N}\left(0, \sigma_2^2\right)$ en plus de l'erreur $\varepsilon_{i_1,i_2,k} \sim \mathcal{N}\left(0, \sigma^2\right)$. Nous avons supposé que les coefficients aléatoires $(b_{i_1}, b_{i_1,i_2}, \varepsilon_{i_1,i_2,k})$ sont indépendants les uns des autres.

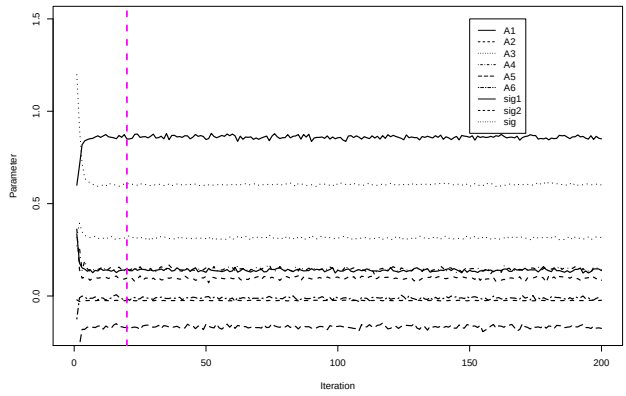

FIGURE 6.2 – Valeur des estimations des paramètres à chaque itération de l'algorithme `SEM` pour le modèle $\log(\texttt{CP})_{i_1,i_2,k} = \texttt{A1}+\texttt{A2}*\hat{\texttt{age}}_{i_1,i_2,k}+\texttt{A3}*\texttt{saison}_{i_1,i_2,k}+ \texttt{A4} * \texttt{infection}_{i_1,i_2,k} + \texttt{A5} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=1} + \texttt{A6} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=2} + b_{i_1} + b_{i_1,i_2} + \varepsilon_{i_1,i_2,k}$, avec $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$, $b_{i_1,i_2} \sim \mathcal{N}\left(0, \sigma_2^2\right)$ et $\varepsilon_{i_1,i_2,k} \sim \mathcal{N}\left(0, \sigma^2\right)$. La période de rodage $M = 20$ est indiqué par une ligne verticale. Le niveau de censure t a été choisi de telle sorte que 30% des données sont censurés.

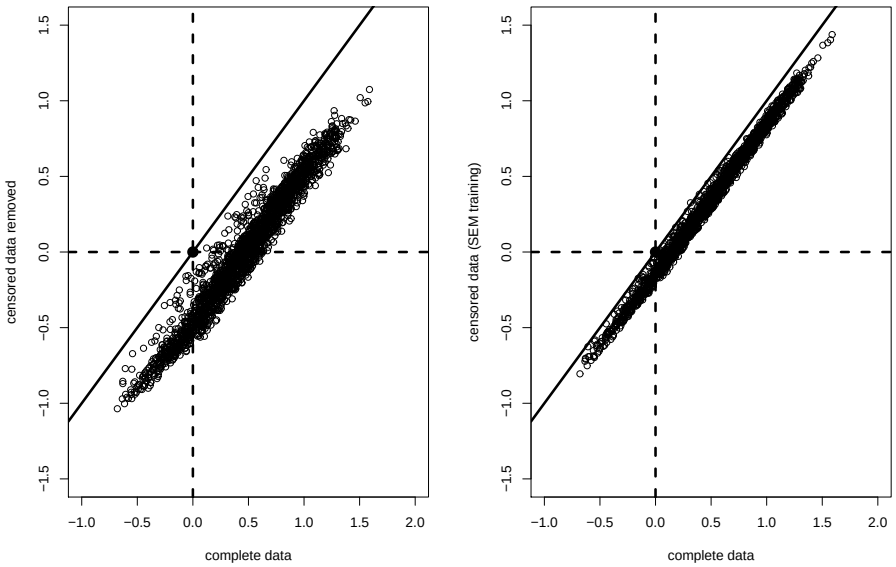

FIGURE 6.3 – Résidus normalisés obtenus en utilisant les données complètes (pas de censure) contre ceux obtenus avec deux méthodes : soit en supprimant les données censurées de l'ensemble des données (dans le panneau de gauche) ou en utilisant notre algorithme `SEM` (dans le panneau de droite). Le niveau de censure t a été choisi de telle sorte que 30% des données sont censurées. Le modèle est $\log\left(\texttt{CP}\right)_{i_1,i_2,k} = \texttt{A1} + \texttt{A2} * \texttt{âge}_{i_1,i_2,k} + \texttt{A3} * \texttt{saison}_{i_1,i_2,k} + \texttt{A4} * \texttt{infection}_{i_1,i_2,k} + \texttt{A5} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=1} + \texttt{A6} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=2} + b_{i_1} + b_{i_1,i_2} + \varepsilon_{i_1,i_2,k}$, avec $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$, $b_{i_1,i_2} \sim \mathcal{N}\left(0, \sigma_2^2\right)$ et $\varepsilon_{i_1,i_2,k} \sim \mathcal{N}\left(0, \sigma^2\right)$.

6.5 Comparaison

En utilisant un jeu de donnée complet, nous avons simulé différents niveaux de censures, les résidus obtenus ont été comparés qualitativement par des visualisation graphiques des résidus normalisés, et quantitativement en calculant l'erreur moyenne, l'erreur robuste moyenne relative, la corrélation de rang de kendall et la corrélation linéaire.

Nous indiquons par r^* les résidus de référence, i.e. les résidus obtenus en utilisant les données complètes, et par r_i les résidus d'essai, i.e. les résidus obtenus en utilisant les approches testées en utilisant des données censurées. Ensuite, nous avons claculé les mesures suivantes : (n désignant le nombre d'individus) : Erreur Moyenne ($\texttt{EM} = n^{-1}\sum_i |r_i - r_i^*|$), Erreur Robuste Moyenne Relative ($\texttt{ERMR} = \frac{1}{n}\sum_{i,|r_i|>\varepsilon}\left|\frac{r_i}{r_i^*} - 1\right|$), (qui est simplement l'erreur relative moyenne obtenue en supprimant les résidus proches de zéro, quand l'estimation est connue pour être peu fiable dans notre cadre), la corrélation de rang de kendall $\texttt{RCor} = \frac{4P}{n(n-1)} - 1$, où P est le nombre de paires concordantes, et enfin,

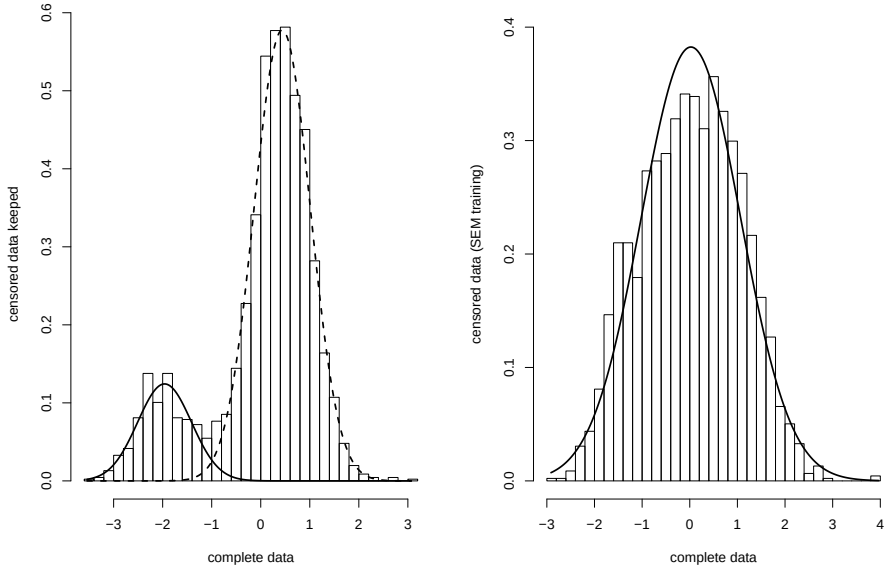

FIGURE 6.4 – Répartition des résidus normalisés obtenus avec : soit en gardant les valeurs censurées (dans le panneau de gauche) ou en utilisant notre algorithme `SEM` (dans le panneau de droite). Le niveau de censure t a été choisie de telle sorte que 30% des données sont censurées.

la corrélation linéaire ($\texttt{LCor} = \mathbb{C}ov\left(r_i, r_i^*\right) \sigma\left(r_i\right)^{-1} \sigma\left(r_i^*\right)^{-1}$).

6.5.1 Comparaison avec l'algorithme MCEM : package `censure3`

Nous avons simulé des données selon le modèle linéaire à effet mixte suivant :

$$Y_{i_1,k} = \texttt{A1} + \texttt{A2} * X_{i_1,k} + b_{i_1} + \varepsilon_{i_1,k}$$

avec l'hypothèse que $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$ et $\varepsilon_{i_1,k} \sim \mathcal{N}\left(0, \sigma^2\right)$. Nous avons supposé que les coefficients aléatoires $(b_{i_1}, \varepsilon_{i_1,k})$ sont indépendants les uns des autres. Les valeurs des paramètres ont été choisi pour être similaires à celles obtenues à partir de la cohorte Aquitaine $\texttt{HIV} - 1$ (human immunodeficiency virus = virus de l'immunodéficience humaine)-1 des patients infectés. Les paramètres étaient $\texttt{A1} = 4$, $\texttt{A2} = -0.5$, $\sigma_1^2 = 0.25$ et $\sigma^2 = 1$. Nous avons simulé des échantillons avec trois niveaux de censure (10%, 20% et 30%), 100 sujets de 200 échantillons. Le nombre des mesures répétées pour chaque sujet a été distribué d'une façon aléatoire entre 2 et 7 (moyenne 4) et les temps des mesures ont été réparties uniformément entre 0 et 6.

	0%	10%				30%			
		RV	DL	HDL	SEM	RV	DL	HDL	SEM
A1	**0.715**	0.919	0.787	0.814	**0.760**	1.123	0.931	0.896	**0.850**
A2	**-0.030**	-0.021	-0.026	-0.024	**-0.027**	-0.014	-0.019	-0.021	**-0.023**
A3	**0.350**	0.297	0.322	0.310	**0.339**	0.250	0.263	0.280	**0.316**
A4	**0.242**	0.090	0.184	0.164	**0.207**	-0.015	0.096	0.116	**0.140**
A5	**-0.140**	-0.169	-0.163	**-0.159**	-0.170	-0.123	**-0.137**	-0.147	-0.175
A6	**0.038**	-0.039	-0.001	-0.009	**0.001**	-0.013	-0.011	-0.011	**-0.005**
σ_1^2	**0.158**	0.121	0.146	0.140	**0.159**	0.061	0.104	0.116	**0.137**
σ_2^2	**0.121**	0.094	0.108	0.102	**0.116**	**0.089**	0.083	0.087	0.085
σ^2	**0.869**	0.607	0.711	0.660	**0.766**	0.407	0.484	0.531	**0.599**
EM	**0**	0.209	0.068	0.100	**0.040**	0.463	0.237	0.194	**0.146**
ERMR	**0**	0.182	0.056	0.085	**0.033**	0.433	0.183	0.160	**0.123**
RCor	**1**	0.944	0.969	0.978	**0.981**	0.870	0.956	0.941	**0.948**
LCor	**1**	0.996	0.998	**0.999**	**0.999**	0.976	0.996	0.994	**0.996**

TABLE 6.1 – Comparaison quantitative de quatre méthodes différentes et à deux niveaux de censure (10% et 30%) avec le modèle $\log(\texttt{CP})_{i_1,i_2,k} = \texttt{A1} + \texttt{A2} * \hat{\texttt{age}}_{i_1,i_2,k} + \texttt{A3} * \texttt{saison}_{i_1,i_2,k} + \texttt{A4} * \texttt{infection}_{i_1,i_2,k} + \texttt{A5} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=1} + \texttt{A6} * \mathbb{1}_{\texttt{T}_{i_1,i_2,k}=2} + b_{i_1} + b_{i_1,i_2} + \varepsilon_{i_1,i_2,k}$, with $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$, $b_{i_1,i_2} \sim \mathcal{N}\left(0, \sigma_2^2\right)$ et $\varepsilon_{i_1,i_2,k} \sim \mathcal{N}\left(0, \sigma^2\right)$. Les quatre dernières lignes correspondent respectivement à l'erreur moyenne (EM), erreur robuste moyenne relative (ERMR), la corrélation de rang de kendall (RCor), la corrélation linéaire (LCor) entre les résidus estimés normalisés et les résidus de référence obtenus avec les données complètes. La colonne 0% pourrait être utilisée comme une référence pour tous les critères. Le temps de fonctionnement était d'environ 150s sur le poste de travail de l'auteur avec $G = 5$ nombre d'itération de l'échantillonneur de Gibbs et $N = 25$ nombre d'itération de l'algorithme SEM.

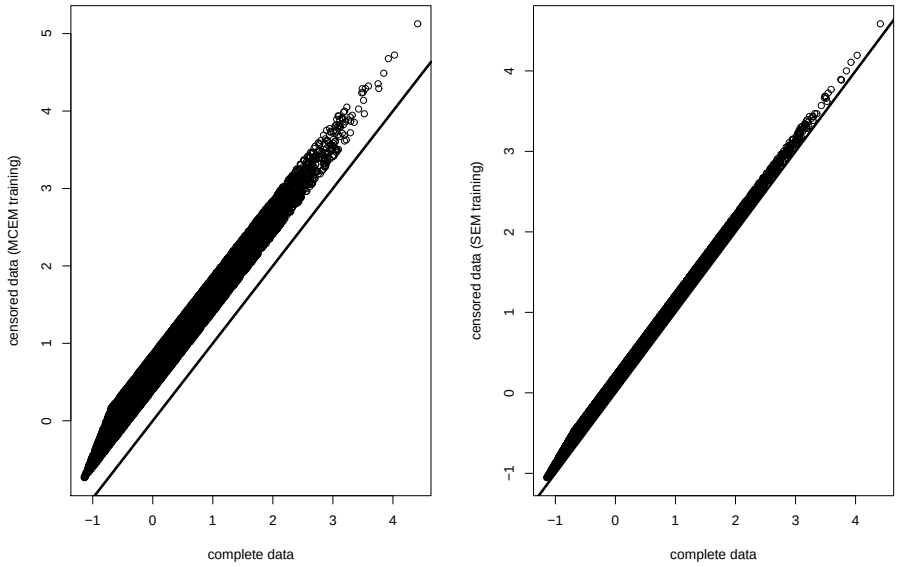

FIGURE 6.5 – Résidus normalisés obtenus en utilisant les données complètes (pas de censure) contre ceux obtenus avec deux méthodes : soit en utilisant l'algorithme `MCEM` développé par Hughes [8] (dans le panneau de gauche) ou en utilisant notre algorithme `SEM` (dans le panneau de droite). Le niveau de censure t a été choisie de telle sorte que 20% des données sont censurées. Le modèle est $Y_{i_1,k} = \texttt{A1} + \texttt{A2} * X_{i_1,k} + b_{i_1} + \varepsilon_{i_1,k}$, with $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$ et $\varepsilon_{i_1,k} \sim \mathcal{N}\left(0, \sigma^2\right)$.

	0%	10%		20%		30%	
		MCEM	SEM	MCEM	SEM	MCEM	SEM
A1	**4.0015**	3.8739	**3.9892**	3.6028	**3.9160**	3.1721	**3.7697**
A2	**-0.4663**	-0.6664	**-0.5190**	-0.9262	**-0.5942**	-1.1591	**-0.7465**
σ_1^2	**0.2211**	0.2693	**0.2453**	**0.2728**	0.28872	**0.2704**	0.4006
σ^2	**1.0008**	1.3414	**1.2055**	1.6938	**1.5792**	**1.5636**	2.6195
EM	**0**	0.2260	**0.0381**	0.6218	**0.1475**	1.1621	**0.3663**
ERMR	**0**	0.1348	**0.0236**	0.4883	**0.1171**	1.0598	**0.3368**
RCor	**1**	0.9596	**0.9893**	0.8973	**0.9709**	0.8345	**0.9301**
LCor	**1**	0.9933	**0.9995**	0.9741	**0.9971**	0.9518	**0.9910**

TABLE 6.2 – Comparaison quantitative de notre méthode à l'algorithme SEM développé par Hughes [8] à trois niveaux de censure (10%, 20% et 30%) avec le modèle $Y_{i_1,k} = \texttt{A1} + \texttt{A2} * X_{i_1,k} + b_{i_1} + \varepsilon_{i_1,k}$, avec $b_{i_1} \sim \mathcal{N}\left(0, \sigma_1^2\right)$ et $\varepsilon_{i_1,k} \sim \mathcal{N}\left(0, \sigma^2\right)$. Les quatres dernières lignes correspondent respectivement à l'erreur moyenne (EM), erreur robuste moyenne relative (ERMR), la corrélation de rang de kendall (RCor), la la corrélation linéaire (LCor) entre les résidus estimés normalisés et les résidus de référence obtenus avec les données complètes. La colonne 0% pourrait être utilisée comme une référence pour tous les critères. Le temps de fonctionnement était d'environ 25s sur le poste de travail de l'auteur avec $G = 5$ nombre d'itération de l'échantillonneur de Gibbs et $N = 25$ nombre d'itération de l'algorithme SEM.

Bibliography

[1] Bryk, A., and Raudenbush, S. W. (1992). Hierarchical Linear Models for Social and Behavioral Research : Applications and Data Analysis Methods. *Newbury Park*, CA : Sage.

[2] Celeux, G. and Diebolt, J. (1985). The SEM Algorithm : a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Comput. Statist. Quater.*, **2** : 73-82.

[3] Chabanet, C. and Pineau, N. (2006). Using linear mixed models to handle variablility of consumer's liking. *Food Quality and Preference*, **17** : 658-668.

[4] Dempster, A. P., Laird and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **39** : 1-38.

[5] Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *J. Amer. Statist. Assoc.*, **76** (374) : 341-353.

[6] Fitzgerald, A. P., DeGruttola, V. G., and Vaida, F. (2002). Modeling HIV viral rebound using non-linear mixed effects models. *Stat. Med.*, **21** : 2093-2108.

[7] Goldstein, H. (1995). Multilevel statistical models. London, *E*dward Arnold.

[8] Hughes, J. P. (1999). Mixed effects models with censored data with application to HIV RNA Levels. *Biometrics*, **55** : 625-629.

[9] Jacqmin-Gadda, H., Chene G., Thiebaut, R., and Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostat.*, **1** (4) : 355-368.

[10] Laird, N. M. and Ware, J. H., (1982). Random effects models for longitudinal data. *Biostat.*, **38** : 963–974.

[11] Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.*, **83** (404) : 1014–1021.

[12] McLean, R. A., Sanders, W. L. and Stroup, W. W. (1991). A Unified Approach to Mixed Linear Models. *J. Amer. Statist. Assoc.*, **45** (1) : 54–64.

[13] Nielsen, S. F. (2000). The stochastic EM algorithm : estimation and asymptotic results. *Bernoulli*, **6** (3) : 457-489.

[14] Robinson, G. K. (1991). That BLUP is a Good Thing : The Estimation of Random Effects. *Statist. Sci.*, **6** (1) : 15-32.

[15] Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *J. Amer. Statist. Assoc.*, **91** : 473-520.

[16] Samson, A., Lavielle, M., and Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model : Application to HIV dynamics model. *Comput. Statist. Data Anal.*, **51** (3) : 1562-1574.

[17] Vaida, F., Fitzgerald, A. P., and Deruttola, V. (2007). Efficient hybrid EM for linear and nonlinear mixed effects models with censored response. *Comput. Statist. Data Anal.*, **51** (12) : 5718-5730.

[18] Wu, H. and Wu, L. (2000). A multiple imputation method for missing covariates in non-linear mixed-effects models with application to hiv dynamics. *Stat. Med.*, **20** (12) : 1755-1769.

[19] Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies. *J. Amer. Statist. Assoc.*, **97** (460) : 955-964.

Chapitre 7

Test d'associations familiales

Sommaire

Mots clés : Épidémiologie génétique, erreurs de génotypage, génotypes manquants, Algorithme EM.

Résumé 6. *Dans ce chapitre, nous proposons une approche basée sur l'algorithme EM (Espérance Maximisation) pour faire face aux erreurs de génotypages en utilisant un modèle d'erreur de génotypage explicite dans des statistiques d'associations familiales.*

7.1 Introduction

En épidémiologie génétique, il est fréquent de considérer les familles des personnes pour lesquelles nous avons des génotypes (ex : valeur pour un bi- allélique Single Nucleotide Polymorphism - SNP) et phénotypes (ex : statut de la maladie/infection, de la valeur d'un caractère quantitatif phénotypique). Le défi statistique relevé consiste alors de trouver les marqueurs génotypiques qui sont associés de façon significative aux phénotypes étudiés .

Une approche classique pour répondre à ce genre de question, consiste à utiliser le cadre générale des tests d'associations familiales (FBAT pour Family

Based Association Test) [1, 2, 3], qui est largement utilisé dans une forme ou une autre. L'idée de FBAT est de combiner les génotypes et les phénotypes dans une statistique (en utilisant une fonction de codage pour les génotypes) et puis de tester l'association en comparant la valeur observée à la distribution sous l'hypothèse nulle où les génotypes sont distribués conditionnellement à leurs proches ancestraux.

En collaboration avec Gégory Nuel et Vincent Miele nous avons proposé dans l'article [C2] une approche qui fait face à deux difficultés : 1-Comment traiter les erreurs de génotypage ? 2- Comment traiter les génotypes manquants ? Pour la première question, une approche commune consiste à détecter les incohérences mendélienne avec un logiciel comme [4] puis de considérer ceux qui correspondent aux génotypes manquants. Tous les génotypes manquants (à la fois ceux qui viennent d'incohérences mendélienne et les vrais) sont ensuite déduites dans FBAT en utilisant un cadre bayésien avec un prior uniforme.
Dans cette section, nous considérons un modèle de données incomplètes où les vrais (inobservé) génotypes produisent ceux observés à travers un modèle d'erreur explicite de génotypage. Nous proposons un algorithme Expectation-Maximisation [5] permettant d'estimer les paramètres de notre modèle ainsi de produire la distribution à posteriori des vrais génotypes (voir [2]). Cette distribution peut être utile pour détecter les erreurs de génotypage (voir [4]). Cependant, notre objectif est plutôt d'utiliser cette distribution pour proposer une mise en œuvre de FBAT à la fois robuste face aux erreurs de génotypage et aux génotypes manquants (see [3]).

7.2 Estimation des paramètres

Structure Pedigree Soit $\mathcal{I} = \{1, \ldots, n\}$ l'ensemble d'individus. On note par F_i (resp. M_i) le père (resp. mère) de l'individu $i \in \mathcal{I}$, et par ? les individus avec un génotype inconnu. Soit $F_1 \cup \ldots \cup F_k$ une partition de $\mathcal{I}$ dans k familles disjoints, un individu $i \in \mathcal{I}$ tel que $(F_i, M_i) =?$ est appelé fondateur. Nous introduisons ensuite l'ensemble des parents $\mathcal{P}_i$ de l'individu $i \in \mathcal{I}$ qui est défini de manière récursive par $\mathcal{P}_i = \{i\} \cup \mathcal{P}_{F_i} \cup \mathcal{P}_{M_i}$ (avec la convention que $\mathcal{P}_{\texttt{fondateur}} = \emptyset$). Deux individus $i, j \in \mathcal{I}$ puisse appartenir à la même famille si et uniquement si $\mathcal{P}_i \cap \mathcal{P}_j \neq \emptyset$.

Génotypes Notons par $g_{s,i} \in \mathcal{G}$ le génotype de l'individu i du marqueur $s \in \mathcal{S} = \{1, \ldots, N\}$, où $\mathcal{G}$ est l'ensemble des génotypes possibles. À partir de maintenant, et par souci de simplicité, nous considérons seulement le cas biallélique $\mathcal{G} = \{aa, aA, AA\}$ (mais nous ne sommes pas limités à ce cas particulier).

7.3 Modèle

Vrais génotypes Soit $G^*_{s,i} \in \mathcal{G}$ pour tous $s \in \mathcal{S}$ et $i \in \mathcal{I}$ représentent les

vrais génotypes et $G_{s,i}$ (avec $i \in \mathcal{I}$ tel que $g_{s,i} \neq ?$) représentent les génotypes observées. Nous supposons que les vrais génotypes sont indépendants d'une famille à une autre pour un marqueur s. Cependant, au sein de chaque famille, les vrais génotypes des fondateurs sont censés être indépendantes et les descendants sont distribués conditionnellement à leurs parents. Pour des raisons de simplification, nous supposons que les fondateurs vérifient l'équilibre de Hardy-Weinberg (HWE) pour les vrais génotypes.

Pour les fondateurs

$$\mathbb{P}\left(G^*_{s,i} = g\right) = D_s\left(g\right) \quad \forall i \in \mathcal{I}, s \in \mathcal{S}$$

où D_s représente la fonction de distribution de probabilité du génotype du marqueur $s \in \mathcal{S}$ et elle est donnée par :

$$D_s\left(g\right) = \begin{cases} p_s^2 & \text{si} \quad g = aa \\ 2\left(1-p_s\right)p_s & \text{si} \quad g = aA \\ \left(1-p_s\right)^2 & \text{si} \quad g = AA \end{cases}$$

avec p_s représente la probabilité de l'allèle "a" pour le marqueur s.

Pour les non foundateurs

$$\mathbb{P}\left(G^*_{s,i} = g | G^*_{s,F_i} = f, G^*_{s,M_i} = m\right) = \texttt{conditionnel}\left(f, m, g\right) \quad \forall i \in \mathcal{I}, s \in \mathcal{S}$$

où $\texttt{conditionnel}\left(f, m, g\right)$ est la probabilité conditionnelle pour les parents avec de vrais génotypes f et m d'avoir un enfant avec le génotype g (*e.g.* $\texttt{conditionnel}\left(aa, aa, .\right) = (1; 0; 0)$, $\texttt{conditionnel}\left(aa, aA, .\right) = (1/2; 1/2; 0)$, $\texttt{conditionnel}\left(aa, AA, .\right) = (0; 1; 0)$, $\texttt{conditionnel}\left(aA, aa, .\right) = (1/2; 1/2; 0)$, $\texttt{conditionnel}\left(aA, aA, .\right) = (1/4; 1/2; 1/4)$, $\texttt{conditionnel}\left(aA, AA, .\right) = (0; 1/2; 1/2)$, $\texttt{conditionnel}\left(AA, aa, .\right) = (0; 1; 0)$, $\texttt{conditionnel}\left(AA, aA, .\right) = (0; 1/2; 1/2)$, $\texttt{conditionnel}\left(AA, AA, .\right) = (0; 0; 1)$).

Donc,

$$\mathbb{P}\left(G^*_{s,i} = g\right) = \sum_{f,m \in \mathcal{G}} \mathbb{P}\left(G^*_{s,F_i} = f\right) \mathbb{P}\left(G^*_{s,M_i} = m\right) \texttt{conditional}\left(f, m, g\right) \quad \forall i \in \mathcal{I}, s \in \mathcal{S}, g \in \mathcal{G}.$$

Génotypes observés Pour tous $i \in \mathcal{I}$ et $s \in \mathcal{S}$, on a :

$$\mathbb{P}\left(G_{s,i} = g_{s,i} | G^*_{s,i} = g^*_{s,i}\right) = \left(1-\varepsilon\right)\mathbb{I}_{\left\{g_{s,i} = g^*_{s,i}\right\}} + \frac{\varepsilon}{2}\mathbb{I}_{\left\{g_{s,i} \neq g^*_{s,i}\right\}}$$

où $\varepsilon \in [0, 1]$ est la probabilité de l'erreur de génotypage. Autres modèles d'erreurs plus complexes peuvent être défini (*ex.* taux d'erreur en fonction de g^*) mais le modèle que nous considérons dans ce travail est assez illustratif pour présenter notre méthode.

7.3.1 Algorithme EM

Pour tout marqueur $s \in \mathcal{S}$ et pour toute famille $\mathcal{F}_j$ nous définissons :

$$\begin{aligned}\mathbb{P}(\mathcal{F}_j) &= \mathbb{P}\left(G^*_{s,i} = g^*_{s,i}, G_{s,i} = g_{s,i}\right) \\ &= \mathbb{P}\left(G^*_{s,i} = g^*_{s,i}\right) \prod_{i \in \mathcal{F}_j} \mathbb{P}\left(G_{s,i} = g_{s,i} | G^*_{s,i} = g^*_{s,i}\right).\end{aligned}$$

Ensuite, nous obtenons le schéma suivant, pour tout $i \in \mathcal{F}_j$ et pour tout $s \in \mathcal{S}$, si $(F_i, M_i) = \texttt{foundateur}$, nous avons :

$$\mathbb{P}\left(G^*_{s,i} = g^*_{s,i}\right) = \sum_{f,m \in \mathcal{G}} D_s(f) D_s(m) \texttt{conditionnel}\left(f, m, g^*_{s,i}\right),$$

et si $(F_i, M_i) \neq \texttt{foundateur}$, nous avons :

$$\mathbb{P}\left(G^*_{s,i} = g^*_{s,i}\right) = \sum_{f,m \in \mathcal{G}} \mathbb{P}\left(G^*_{s,F_i} = f\right) \mathbb{P}\left(G^*_{s,M_i} = m\right) \texttt{conditionnel}\left(f, m, g^*_{s,i}\right).$$

Il en résulte que, dans le cas simple quand $(F_i, M_i) = \texttt{foundateur}$,

$$\mathbb{P}(\mathcal{F}_j) = \sum_{f,m \in \mathcal{G}} D_s(f) D_s(m) \texttt{conditionnel}\left(f, m, g^*_{s,i}\right) \prod_{i \in \mathcal{F}_j} \left[(1-\varepsilon) \mathbb{I}_{\left\{g_{s,i} = g^*_{s,i}\right\}} + \frac{\varepsilon}{2} \mathbb{I}_{\left\{g_{s,i} \neq g^*_{s,i}\right\}}\right]$$

et, quand $(F_i, M_i) \neq \texttt{foundateur}$,

$$\begin{aligned}\mathbb{P}(\mathcal{F}_j) &= \sum_{f,m \in \mathcal{G}} \mathbb{P}\left(G^*_{s,F_i} = f\right) \mathbb{P}\left(G^*_{s,M_i} = m\right) \texttt{conditionnel}\left(f, m, g^*_{s,i}\right) \\ &\quad \times \prod_{i \in \mathcal{F}_j} \left[(1-\varepsilon) \mathbb{I}_{\left\{g_{s,i} = g^*_{s,i}\right\}} + \frac{\varepsilon}{2} \mathbb{I}_{\left\{g_{s,i} \neq g^*_{s,i}\right\}}\right]\end{aligned}$$

et en utilisant le schéma donné ci-dessus, nous pouvons calculer la probabilité souhaitée.

Estimation EM des fréquences p_s pour tout $s \in \mathcal{S}$.

- Initialisation aléatoire de p_s pour tout $s \in \mathcal{S}$
- paramètre n'a pas encore convergé
  - $s \in \mathcal{S}$
  - nallele=0
  - pour $j = 1, \ldots, k$
    - nlocalallele=0, et normalisation=0
    - pour tous les valeurs possibles de $g^*_{s,i}$ pour $i \in \mathcal{F}_j$
      - calcule du proba = $\mathbb{P}(\mathcal{F}_j)$
      - normalisation+=proba
      - nlocalallele+=proba $\times \sum_{i \in \texttt{fondateur}} \left(2 \times \mathbb{I}\left\{g^*_{s,i} = aa\right\} + \mathbb{I}\left\{g^*_{s,i} = aA\right\}\right)$
    - nallele+=nlocalallele/normalisation
  - $p_s = nallele/(4 \times |\{(F_i, M_i) = \texttt{fondateur}, i \in \mathcal{I}\}|)$

7.4 Statistique FBAT

Notons par $\phi_i \in \mathbb{R}$ le phénotype de l'individu $i \in \mathcal{I}$ et $X : \mathcal{G}^h \to \mathbb{R}^d$ une fonction de codage correspondant à un mode phénotypique donné (ex : additif, génotypique, récessif, etc). Nous supposons d'abord que $h = 1$ (seulement un marqueur à la fois est considéré) et $d = 1$ ensuite nous allons discuter l'extension de nos résultats avec des fonctions de codage plus complexes. Par exemple, on peut considérer la fonction de codage additif suivante :

$$X(aa) = 2 \quad X(aA) = 1 \quad X(AA) = 0.$$

Nous définissons la statistique de FBAT, $t(s)$ du marqueur s par :

$$t(s) = \sum_{g_s^*} \left[\sum_{i \in \mathcal{I}'} (\phi_i - \texttt{offset}) \times X\left(g_{s,i}^*\right) \right] \mathbb{P}\left(G_s^* = g_s^* | G_s = g_s\right)$$

où $\texttt{offset} \in \mathbb{R}$ est une constante (ex $\texttt{offset} = 0$), $\mathcal{I}' = \{i \in \mathcal{I}, \phi_i \neq ?\}$, $g_s^* = \left\{g_{s,i}^*, i \in \mathcal{I}\right\}$ et $g_s = \{g_{s,i}, i \in \mathcal{I}'\}$.

Pour utiliser cette statistique dans un cadre de tests, nous voulons comparer $t(s)$ à la distribution de

$$T(s) = \sum_{g_s^*} \left[\sum_{i \in \mathcal{I}'} (\phi_i - \texttt{offset}) \times X\left(\Gamma_{s,i}^*\right) \right] \mathbb{P}\left(G_s^* = g_s^* | G_s = g_s\right)$$

où toutes les $\Gamma_{s,i}^*$ sont indépendantes et distribuées selon :

$$\mathbb{P}\left(\Gamma_{s,i}^* = \gamma\right) = \begin{cases} \texttt{offspring}\left(g_{s,F_i}^*, g_{s,M_i}^*, \gamma\right) & \text{si } F_i \neq ? \text{ et } M_i \neq ? \\ D_s(\gamma) & \text{si } F_i = ? \text{ et } M_i = ? \end{cases}$$

Nous définissons alors le $\mathrm{Z} - \mathrm{score}$ normalisé z_s par :

$$z_s = \frac{t(s) - \mathbb{E}\left[T(s)\right]}{\sqrt{\mathbb{V}\left[T(s)\right]}}$$

qui peut être calculé avec l'algorithme suivant :

Calcul de la statistique FBAT.
- stat=0, esperance=0, et variance=0
- pour $j = 1, \ldots, k$
  - localstat=0, localesperance=0 et localvariance=0
  - normalisation=0
  - pour tous les valeurs possibles de $g_{s,i}^*$ pour $i \in \mathcal{F}_j$
    - calcul proba = $\mathbb{P}(\mathcal{F}_j)$
    - normalisation+=proba
    - localstat+=proba $\times \left[\sum_{i \in \mathcal{I}' \bigcap \mathcal{F}_j} (\phi_i - \texttt{offset}) \times X\left(g_{s,i}^*\right)\right]$

- localesperance+=proba $\times \left(\sum_{i \in \mathcal{I}' \bigcap \mathcal{F}_j} (\phi_i - \texttt{offset}) \times \mathbb{E}\left[X\left(\Gamma^*_{s,i}\right)\right] \right)$
- localesperance+=proba $\times \left(\sum_{i \in \mathcal{I}' \bigcap \mathcal{F}_j} (\phi_i - \texttt{offset}) \times \mathbb{V}\left[X\left(\Gamma^*_{s,i}\right)\right] \right)$

– stat+=localstat/normalisation
– esperance+=localesperance/normalisation
– variance+=localvariance/($normalisation \times normalisation$).

Ce résultat s'étend naturellement en utilisant des fonctions de codage plus complexe ($h > 1$ or $d > 1$) d'une statistique FBAT multidimensionnelle. Dans un tel cas, il est alors nécessaire de calculer à la fois l'espérance et la matrice de covariance de la statistique multidimensionnelle afin d'effectuer une normalisation du chi-deux classique au lieu de la normalisation gaussienne qui est faite ci-dessus.

7.4.1 Conclusion

La méthode que nous proposons ici permet à la fois de détecter les erreurs de génotypage et de produire des statistiques de FBAT robustes aux erreurs et aux génotypes manquants. En dépit du fait qu'il ne soit pas son but principal, notre algorithme semble afficher des performances similaires à GMCheck [6] pour le problème de détection et de correction des erreurs de génotypage.

Ajoutons que nos méthodes sont disponibles dans une bibliothèque de programmation appelé libfbat qui est écrite en ANSI C ++ et développé sur les systèmes x86 GNU / Linux avec GCC 4.1.3. Compilation et installation sont conformes à la procédure standard GNU. La bibliothèque est gratuite et disponible sur le web. libfbat est sous licence GNU General.

Bibliography

[1] Laird, N., Horvath, S. and Xu, X., (2000). Implementing a unified approach to family based tests of association. *Genet. Epidemiol.*, **19**(Suppl 1) : 921-927.

[2] Lange, C., Silverman, E. K., Xu, X., Weiss, S. T. and Laird, N. M., (2003). A multivariate family-based association test using generalized estimating equations : FBAT-GEE. *Biostatistics*, **4**(2) : 195-206.

[3] Xu, X., Rakovski, C., Xu, X. and Laird, N., (2006). An efficient family-based association test using multiple markers. *Genet. Epidemiol.*, **30** : 620-626.

[4] O'Connell, J. R. and Weeks, D. E., (1998). PedCheck : a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **63**(1) : 259-266.

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **39**(1) : 1-38.

[6] Thomas, A., (2005). GMCheck : Bayesian error checking for pedigreegenotypes and phenotypes. *Bioinformatics*, **21**(14) : 3187-3188.