UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

DOCUMENT DE SYNTHÈSE EN VUE DE L'HABILITATION À DIRIGER DES RECHERCHES

Contribution à l'étude des modèles à données manquantes et apprentissage statistique autour de modèles markoviens

Pierre VANDEKERKHOVE

Soutenue le 10 décembre 2007 devant le jury composé de :

Rapporteurs :	Patrice Bertail	CREST-LS et MODAL'X, Nanterre
	Éric Moulines	ENST, Paris
	Bruce Lindsay	Pennstate University, USA
Examinateurs :	Pierre Del Moral	INRIA, Bordeaux
	Jean-François Delmas	ENPC, Marne-la-Vallée
	Elisabeth Gassiat	Paris XI, Orsay
	Marc Hoffmann	Paris-Est, Marne-la-Vallée
	Damien Lamberton	Paris-Est, Marne-la-Vallée

Remerciements

Je remercie vivement Patrice Bertail, Bruce Lindsay et Éric Moulines d'avoir accepté d'être rapporteurs de cette habilitation à diriger des recherches, manifestant ainsi leur intérêt pour mes travaux. Je suis très reconnaissant à Pierre Del Moral, Jean-François Delmas, Marc Hoffmann et Damien Lamberton d'avoir accepté d'être membre de mon jury. Je remercie tout particulièrement Elisabeth Gassiat qui fut rapporteur de ma thèse de Doctorat et qui me fait l'honneur d'être à nouveau membre du jury de mon habilitation.

J'ai rencontré Didier Chauveau durant mon post-doc, alors que nous faisions partie tous les deux d'un même réseau européen consacré aux méthodes de Monte Carlo par chaînes de Markov. Nous avons commencé à travailler ensemble à mon arrivée à l'université de Marne-la-Vallée en 1998. Je garde de ces années de collaboration le souvenir de séances de travail passionnantes, riches en suspens mathématique, et le goût pour une recherche inspirée par les applications. Nos relations ont depuis longtemps dépassé le cadre professionnel et je tiens à le remercier chaleureusement pour son amitié et la confiance qu'il m'a témoigné durant toutes ces années.

J'ai fait la connaissance de Laurent Bordes par le biais de l'équipe de Fiabilité de Marne-la-Vallée que je tiens à remercier grandement au passage. Notre collaboration a commencé sur un problème transversal à nos domaines de prédilection et s'est poursuivie par sérendipité [1] *à cause d'une formule mathématique laissée un jour dans le coin d'un tableau, et qu'un collègue "qui passait par là" a réinterprété pour nous du point de vue de l'analyse numérique. Sa culture, sa curiosité scientifique, sa générosité et sa grande force de travail ont été pour moi un formidable moteur. Pour toutes les raisons que je viens d'évoquer, merci à toi Laurent.*

Durant ces années j'ai eu la chance de rencontrer de nombreux chercheurs avec qui j'ai eu beaucoup de plaisir à collaborer et que je remercie. Je pense tout particulièrement à Dominique Bakry, Céline Delmas, Paolo Giudici, Xavier Milhaud, Stéphane Mottelet, Tobias Rydén, ainsi que Nadia Oudjane et Pierre Tarrès avec qui j'ai le plaisir de travailler actuellement.

Je souhaite remercier également toute l'équipe de Mathématiques de l'Unviversité de Marne-la-Vallée pour l'ambiance chaleureuse et conviviale qui y règne. Je pense notamment à Thierry Jeantheau, Paul-Marie Samson et Miguel Martinez qui ont partagé mon bureau, Mathieu Meyer pour ses conseils et sa grande disponibilité lors de la rédaction de ce document, et enfin Mireille Morvan pour son aide précieuse durant toutes ces années.

[1] La "serendipity" est un mot inventé en 1754 par le philosophe anglais Sir Horace Walpole, pour qualifier la faculté qu'ont certains de trouver la bonne information par hasard, un peu sans la chercher.

À Laurence.

Table des matières

Chapitre 1

Introduction

Ce document de synthèse rassemble les travaux de recherche que j'ai effectués depuis le début de ma thèse de doctorat. Mon travail s'articule autour de deux grands thèmes : les modèles à données manquantes et les algorithmes stochastiques. Je me suis initié à ces deux sujets durant ma thèse, faite à Montpellier et Toulouse durant les années 1993-97, et le post-doc que j'ai effectué ensuite à l'université de Pavie (Italie) en 1997. Je poursuis actuellement leur étude au sein du laboratoire d'Analyse de Mathématique Appliquées de Marne-la-Vallée, où je suis maître de conférences depuis le mois de septembre 1998. Concernant les modèles à données manquantes, je me suis principalement attaché à la généralisation des modèles de Markov cachés et à l'étude de nouveaux modèles de mélanges semi-paramétriques. Dans la partie algorithmique de mon travail, je me suis essentiellement intéressé à l'optimisation des méthodes de Monte Carlo et à l'étude de deux algorithmes à pas décroissant dans des contextes non standards. Les quatres thèmes de recherche bien distincts que je viens de citer ont en réalité certains points communs qui apparaîtront à la lecture de ce document. En effet deux des questions fondamentales abordées au travers de mes différents travaux concernent l'importance de la "qualité" de l'ergodicité en Statistique, et le rôle que peuvent jouer certaines méthodes non-paramétriques dans des contextes habituellement paramétriques ou bayésiens. La présentation générale des travaux (ainsi que sa traduction anglaise à suivre) décrit de manière aussi peu technique que possible les motivations, résultats et applications de mes différentes contributions aux domaines précédement cités. Dans certaines parties consacrées à la présentation détaillée de mes travaux, j'ai choisi, dans un souci de valorisation, d'illustrer les avancées numériques apportées par nos approches aux moyens de quelques outils graphiques. De même, lorsque nos méthodes ont pu être testées sur des données réelles, j'ai choisi de rappeler le contexte expérimental de l'étude ainsi que les conclusions que nous avons pu apporter. La liste de mes publications, articles soumis ou en préparation figurent au chapitre 6 et les références, respectivement de la forme [A1] ou [B1] dans le texte, renvoient à cette liste. Une bibliographie figure à la fin de ce document et les références à cette bibliographie sont faites en citant le(s) auteur(s) ainsi que l'année de l'article (ou livre) concerné.

1.1 Présentation générale des travaux

Je présente dans ce qui suit les travaux académiques réalisés durant mon parcours de recherche. Dans un souci de clarté, je tiens à préciser que les travaux [A1] et [A2] découle directement de ma thèse de doctorat et que [A4] a été obtenu durant mon post-doc.

Comme il l'a déjà été dit dans le paragraphe précédent, mes travaux s'articulent autour de quatre thèmes principaux : les modèles de Markov cachés ; les modèles de mélange semi-paramétriques ; l'optimisation des méthodes de Monte Carlo (et leur utilisation pour l'étude des chaînes de Markov) ; et enfin les algorithmes stochastiques à pas décroissant.

Modèles de Markov cachés

Dans cette première partie, je m'intéresse à diverses questions en lien avec les modèles de Markov cachés (MMC). La première question concerne l'hypothèse de stationnarité dans l'étude de l'estimateur du maximum de vraisemblance (EMV) pour les chaînes de Markov cachées (CMC). Baum et Petrie (1966) étudient l'EMV pour les CMC en montrant que la log-vraisemblance normalisée (par la taille de l'échantillon) d'une CMC a le même comportement asymptotique que la moyenne empirique du log des probabilités conditionnelles saturées (probabilité conditionnelle de l'observation sachant le passé infini de la chaîne observée). Ces auteurs utilisent alors l'hypothèse de stationnarité pour en déduire que les probabilités conditionnelles saturées constituent une famille de variables aléatoires invariantes en loi sous l'opérateur retard. Cette remarque leur permet en effet d'établir que, pour toute valeur du paramètre, la log-vraisemblance normalisée des CMC vérifie le théorème ergodique et converge presque sûrement vers une quantité (fonction du paramètre) appelée *entropie* et satisfaisant la propriété dite de *contraste*. L'objet du travail, fait durant ma thèse de doctorat en collaboration avec Dominique Bakry et Xavier Milhaud [A1], consiste essentiellement à montrer que l'étude menée par Baum et Petrie peut s'étendre au cas des CMC partant d'une condition initiale déterministe. L'argument clé de la preuve pour ce travail est l'utilisation d'une technique de couplage permettant de comparer efficacement le comportement de la log-vraisemblance d'une CMC sous une condition initiale arbitraire, avec celle d'une CMC démarrant sous le régime stationnaire. Nous montrons aussi dans [A1] la propriété LAN (local asymptotic normality) pour les CMC.

À la suite de ce travail, j'ai eu l'opportunité de travailler (dans le cadre mon post-doc) avec Paolo Giudici et Tobias Rydén [A4], sur des problèmes de test pour les MMC. En utilisant les travaux de Bickel *et al.* (1998) et certains résultats d'identifiabilité pour les familles de mélanges multivariés, nous établissons des tests de type rapport de vraisemblance permettant de traiter des hypothèses ponctuelles ou composites sur les paramètres du modèle. Nous appliquons de plus notre travail sur des donnée réelles de pollution. Les modèles utilisés sont alors des MMC graphiques

sur lesquels il s'agit de tester des zéros dans l'inverse de la matrice de corrélation des vecteurs observés (supposés gaussiens conditionnellement à la chaîne sous-jacente).

Avec Laurent Bordes [A7], je me suis intéressé à un modèle de Markov partiellement caché trouvant un intérêt naturel dans le domaine de la Fiabilité. Il s'agit d'un MMC pour lequel l'information portant sur l'atteinte d'un état fixé de la chaîne sous-jacente est systématiquement connue. Une telle situation se rencontre par exemple lorsqu'un système est en fonctionnement et que l'on souhaite faire de l'inférence sur son modèle de dégradation (supposé markovien) au travers de certaines variables observables (température, niveau vibratoire, etc.). En cas de panne nous pouvons considérer que nous avons atteint de manière sûre le pire état de dégradation du système. Nous montrons pour ce modèle la consistance et la normalité asymptotique de l'EMV sous des conditions plus faibles que celles exigées pour les MMC classiques. Notons en particulier que l'hypothèse d'apériodicité, cruciale dans l'étude de Baum et Petrie (1966) et dans tous les travaux qui ont suivi sur l'inférence des MMC, n'est ici plus recquise.

Je conclus enfin cette partie avec l'introduction et l'étude, faite dans [A8], d'une nouvelle classe de processus à données manquantes. Il s'agit des mélanges Markoviens de processus de Markov dont la définition consiste en la mise bout à bout de trajectoires de processus de Markov mutuellement indépendants, le choix des trajectoires s'opérant au moyen d'une chaîne de Markov non observée. En raison de la grande complexité de la vraisemblance associée à ce modèle et des nombreuses impasses techniques qu'elle engendre, je me suis intéressé à l'estimateur du maximum de vraisemblance des données tronquées (EMVT) introduit par Rydén (1994). Je montre, sous des conditions standards d'identifiabilité, de régularité et de mélangeance des processus, la consistance et la normalité asymptotique de l'EMVT. Une des principales difficultés associées à ce type de modèle étant la paramétrisation des densités des mesures invariantes associées à chaque processus, j'indique une procédure de Monte Carlo permettant de les estimer ponctuellement. Cette étape cruciale permet de calculer en pratique la vraisemblance des données tronquées pour toute valeur du paramètre. Je montre d'autre part que les hypothèses assurant la consistance et la normalité asymptotique de l'EMVT sont satisfaites dans le cadre de mélanges de processus autoregressifs d'ordre 1 gaussiens.

Modèles de mélange semi-paramétriques

Dans cette deuxième partie, je présente diverses contibutions à l'étude des modèles de mélanges semi-paramétriques. Les modèles auxquels nous nous intéressons ont été inspirés par Hall et Zhou (2004). Le premier modèle, étudié en collaboration avec Stéphane Mottelet et Laurent Bordes [A9], est un modèle de mélange à deux composantes symétriques égales à un paramètre de localisation près. Le deuxième (utilisé dans l'analyse des puces ADN), étudié en collaboration avec Céline Delmas et Laurent Bordes [A10], correspond à un mélange à deux composantes dont l'une est connue et l'autre est simplement supposée symétrique autour d'un paramètre de

localisation inconnu. Nous abordons en détail le problème de l'identifiabilité et proposons des méthodes d'estimation des paramètres euclidiens pour ces deux modèles. L'existence de formules d'inversion permettant d'isoler la loi des composantes inconnues, couplée avec l'hypothèse de symétrie, nous permettent d'exhiber des mesures de discrépance sur l'espace des paramètres, induisant ainsi des procédures d'estimation naturelles de type *minimum de contraste*. Une utilisation "plug-in" des formules d'inversion permettent alors de reconstituer les paramètres fonctionnels (fonction de répartition et densité) inconnus des modèles. Nous montrons la consistance de nos procédures d'estimation ainsi que certaines vitesses de convergence presque sûres. Dans un travail en cours [B1], nous montrons pour le deuxième modèle, un théorème central limite (TCL) fonctionnel associé à l'estimateur du vecteur des paramètres constitué par : la proportion du mélange, le paramètre de localisation, et la fonction de répartition de la composante inconnue. Ce dernier résultat nous permet de développer des procédures de test concernant des hypothèses ponctuelles sur les paramètres euclidiens ainsi que l'hypothèse de symétrie sur la composante inconnue. Nous appliquons respectivement les modèles précédents à des données de pluviométrie et à un problème réel de comparaison de gestation chez les bovins issus de l'insémination *artificielle* ou *in vitro*.

À la fin de cette partie, je présente un travail concernant les aspects algorithmiques de l'estimation du premier modèle. Ce travail, réalisé en collaboration avec Didier Chauveau et Laurent Bordes [A12], propose d'adapter l'algorithme EM (Expectation/Maximisation) classique en y ajoutant une étape d'estimation de la densité inconnue. Nous donnons une explication heuristique à cette approche et montrons par des simulations que cette méthodologie donne de très bons résultats avec des temps de calculs beaucoup moins longs que ceux générés par des méthodes d'optimisation classiques.

Méthodes de Monte Carlo

Dans cette troisième partie je présente une série de travaux, tous effectués en collaboration avec Didier Chauveau, concernant l'amélioration de certaines méthodes de Monte Carlo par chaîne de Markov (MCMC). La première approche que nous avons envisagée dans [A6], avait pour but d'accélérer la vitesse de convergence de l'algorithme dit de Hastings-Metropolis (HM). L'algorithme de HM génère, au moyen d'un mécanisme d'acceptation/rejet sur des données simulées au moyen d'une loi instrumentale, une chaîne de Markov dont la loi invariante a pour densité une densité souhaitée. Divers auteurs, comme Menegersen et Tweedie (1996) ou Holden (1998), ont mis en évidence les liens entre la vitesse de convergence de cet algorithme et la proximité entre la loi cible et la loi instrumentale. Étant donné la convergence (même lente) des densités de l'algorithme de HM vers la densité cible, il nous est apparu intéressant d'estimer non-paramétriquement ces densités en générant plusieurs algorithmes de HM en parallèle (i.i.d.) et de les exploiter à leur tour comme densités de nouvelles lois instrumentales (facilement simulables en raison de la structure de mélange des estimateurs à noyaux). Nous montrons, dans le cadre de densités à support

compact, et pour un nombre à priori aussi grand que l'on veut d'algorithmes en parallèle, que notre procédure génère des algorithmes asymptotiquement plus rapides que tout algorithme de HM utilisant une loi instrumentale arbitraire. Nous présentons des simulations réalisées en dimension 1 et 2 illustrant le rendement important de ce type d'approche face à des algorithmes de HM standards même convenablement calibrés.

La deuxième contribution au domaine des MCMC, réalisée dans [B3], , porte sur une méthode destinée à hiérarchiser l'efficacité de diverses approches/stratégies face à un problème de simulation par chaîne de Markov. Considérons par exemple deux algorithmes de MCMC ayant pour but de simuler la même loi. Il est alors intéressant de connaître l'algorithme qui converge le plus rapidement vers sa loi stationnaire. Pour répondre à ce type de question nous devrions être en mesure d'estimer une distance entre la densité des algorithmes et la densité cible, or cette dernière n'est en général connue qu'à une constante de normalisation près (cadre bayésien). On peut cependant remarquer que la différence des distances de Kullback entre la densité des algorithmes et la densité cible est indépendante de cette constante de normalisation ; ainsi l'estimation de ces différences et l'analyse de leur comportement au cours des premières itérations pourrait s'avérer un outil synthétique permettant de mieux appréhender la qualité relative de chaque algorithme. Partant de ce constat, nous avons choisi d'estimer ces différences de distance de Kullback au moyen de plusieurs algorithmes lancés en parallèle. La méthode d'estimation utilisée repose sur l'estimateur de l'entropie introduit par Györfi et Van Der Meulen (1989). La principale difficulté de notre travail a été de montrer que les conditions techniques assurant la consistance des estimateurs étaient satisfaites pour les densités successives de l'algorithme de Hastings-Metropolis au prix de certaines hypothèses (toutes vérifiées dans le cas gaussien).

Nous avons enfin un travail en cours (voir Lavanant, 2007) sur une procédure permettant d'améliorer les performances de la méthode d'*échantillonnage d'importance* (traduction de *importance sampling*). L'échantillonnage d'importance permet de calculer des espérances de fonctions test au sens d'une densité connue à une constante de normalisation près. Le principe de cette méthode utilise un rapport de loi forte des grands nombres portant sur des fonctionnelles d'échantillons instrumentaux judicieusement choisies. Comme pour l'algorithme de HM, il est admis que la qualité de l'estimation (variance) par échantillonnage d'importance est liée à la ressemblance entre la loi instrumentale et la loi cible. Afin d'augmenter cette ressemblance, nous proposons un estimateur à noyau de la loi cible utilisant l'échantillon instrumental, la connaissance de la densité instrumentale, et le numérateur de la densité cible. Nous conjecturons qu'en adaptant le travail de Giné et Guillou (2002), nous devrions pouvoir montrer que notre estimateur est uniformément convergent et préciser une vitesse de convergence presque sûre (dépendante de la dimension du problème). L'étape de re-échantillonnage, au sens de cet estimateur à noyau repondéré, peut s'apparenter à une version lissée de la méthode de re-échantillonnage par poids d'importance (voir, *e.g.* Cappé *et al.*, 2005), couramment utilisée en filtrage

particulaire ou dans les approches bayésiennes.

Nous concluons cette partie consacrée aux méthodes de Monte Carlo, par un travail [A11] concernant l'estimation de l'entropie des chaînes de Markov. Étant donné le noyau de transition d'une chaîne de Markov (supposé simulable), on s'intéresse à la possibilité éventuelle de pouvoir représenter de manière consistante l'évolution dans le temps de la distance de Kullback entre les lois de deux chaînes partant de conditions initiales différentes, ou entre la loi d'une chaîne et sa loi stationnaire, lorsque celle-ci est connue (et simulable). Nous proposons pour cela un estimateur de type double Monte Carlo permettant d'estimer l'entropie de la chaîne contre sa concurrente ou bien sa loi stationnaire. Nous montrons la consistance et la normalité de nos estimateurs sous des conditions faibles de moments. Nous mettons enfin en oeuvre notre méthode pour tester la stabilité de processus autoregressifs d'ordre 1, et pour évaluer la vitesse de convergence (au sens de Kullback) dans le TCL pour des échantillons i.i.d. suivant diverses lois (Student, Uniforme, etc.).

Algorithmes stochastiques à pas décroissant

Cette dernière partie est consacrée à la convergence de deux algorithmes stochastiques à pas décroissant. Dans [A2] je m'intéresse au comportement en temps long de l'algorithme du recuit simulé lorsque le potentiel n'est pas "parfaitement connu", mais peut être approché par une suite d'estimateurs uniformément convergents admettant une vitesse de convergence presque sûre suffisamment rapide. Avec Pierre Tarrès [B2] nous étudions un critère de convergence pour l'algorithme du bandit à deux bras dans le cas où les bras sont simplement supposés ergodiques. Notons que cette extension non triviale du cas i.i.d. laisse entrevoir des applications intéressantes de cet algorithme à la finance (allocation de portefeuilles d'actions).

1.2 General overview of the contributions (English version)

In this chapter, I introduce the organization of my HDR (Habilitation à Diriger des Recherches) report. This work contains four main parts. These parts are respectively dedicated to Hidden Markov Models ; new semiparametric mixture models ; Monte Carlo optimization methods and application to Markov chain behavior analysis ; and the study of two non-decreasing stochastic algorithms.

Hidden Markov models

In this first part, I consider various questions connected with Hidden Markov Models (HMMs). The first question deals with the relevancy of the stationarity assumption in the Maximum Likelihood Estimator (MLE) study for Hidden Markov Chains (HMCs) when the observable process is valued in a finite state-space. Baum and Petrie (1966) investigate the MLE for HMCs by showing that the normalized (by

the sample size) log-likelihood of HMCs behaves asymptotically like the empirical average of its full-conditional log-probabilities. These authors use the stationarity assumption in order to show that the full-conditional probabilities are a law-invariant family of random variables under the shift operator. This last point allows them to establish more specifically that the normalized log-likelihood of a HMC statisfies the ergodic theorem and converges almost surely to a function of the parameter called *entropy* and satisfying the so-called *contrast* property. The aim of the article written in collaboration with Dominique Bakry and Xavier Milhaud [A1] was essentially to prove that the Baum and Petrie results (consistency and asymptotic normality) were still true in the case of non-stationary HMCs. The key idea of the proof was an efficient coupling technique for the log-likelihood comparison between stationary and non-stationary HMCs having the same transitions. We also prove in [A1] the LAN (Local Asymptotic Normality) property for HMCs.

Following this work, I had the opportunity, during my post-doc, to collaborate with Paolo Giudici and Tobias Rydén [A2] on testing problems for HMMs. Using the Central Limit Theorem (CLT) for HMMs established by Bickel, Ritov and Rydén (1998), and standard identifiability results on multivariate mixture density families, we propose consistent likelihood ratio tests to treat simple or composite assumptions on the statistical parameter. We apply our work to graphical HMMs where the natural aim is to test some zeros in the precision matrix of the observed random vectors (assumed to be Gaussian conditionnally on the underlying Markov chain).

With Laurent Bordes [A8], we introduced a Partially Hidden Markov Model (PHMM) that finds natural applications in reliability problems. This model is defined as a standard HMM except for the fact that when the underlying Makov chain reaches a pre-defined state this information is accessible to the practitioner. Such a situation can be found, for example, when the degradation model (assumed to be Markovian and valued in a finite discrete space) of a system is to be estimated through various observable variables (temperature, vibratory level, etc.), the system failure being in that case the only observable state of the degradation process. We establish the consistency and asymptotic normality of the MLE for this model under weaker conditions than those needed for classical HMMs. Consider for example that the aperiodicity assumption usually made on the underlying Markov chain is no longer needed in our work.

I complete this part with the study of a new class of missing data models given in [A8]. I consider the so-called Hidden Markov Mixture of Markov Models (H4M), which is defined as follows : Consider K mutually independent Markov processes, labelled from 1 to K, and a Markov chain valued in this label state space ; the value at time t of the associated H4M coincides with the value of the i-th Markov process if the Markov chain is in state i at time t. Because of the huge complexity of the likelihood for such a model, and the numerous technical difficulties it generates, I proposed to consider instead the Split data Maximum Likelihood Estimator (SMLE) introduced by Rydén (1994). I prove, under weak identifiability and regularity as-

sumptions, and a mixing property on the processes, that the SMLE is consistent and asymptotically normally distributed. One of the main difficulties with this model is the invariant probability density functions parametrization. To solve this problem, I proposed a Monte Carlo pointwise procedure in order to calculate the split data likelihood for all values of the parameter. I check in addition that the technical conditions given to ensure the theoritical results are all satisfied when the (mixed) Markov processes are first-order univariate autoregressive Gaussian models. Various possible applications of this model to biology, ion channels analysis, and EEG signals are also discussed at the end of [A8].

Semiparametric mixture models

In this second part, I present various contributions to semiparametric mixture models. The models we consider are inspired by Hall and Zhou (2003). The first model, studied in collaboration with Stéphane Mottelet and Laurent Bordes [A9], is a mixture of two symmetric probability densities, equal up to a localization parameter, whereas the second model, studied in collaboration with Céline Delmas and Laurent Bordes (coming from microarray problems), is a two-component mixture model where one component is known when the other one is just supposed to be symmetric with respect to an unknown localization parameter. We study carefully the identifiability problems induced by these two models and propose specific methods to estimate the Euclidean part of their parameters. For each model there exists an explicit inversion formula allowing us to isolate the unknown cumulative distribution function, and a slightly different one for the distribution density, under the true value of the parameter. This point, coupled with the symmetry assumption made on the unknown component, then allows us to exhibit discrepancy measures that induce natural minimum contrast estimation procedures. Once the Euclidean part of the parameter is estimated, it is thus reasonable to use the inversion formulae in a "plug-in" way to recover the unknown functional parameters of the models. We prove, under mild conditions, that our estimators are strongly consistent and give an almost sure rate of convergence under additionnal moment conditions. In a work in progress on the second model, in collaboration with Laurent Bordes [B1], we show that there exists an estimator of the functional parameter — the mixing proportion, the localization parameter, and the cumulative distribution function — that satifies a functional CLT towards a Gaussian process whose covariance matrix is given and computable via a tricky Monte Carlo method. As a consequence of this result we propose semiparametric tests to deal with a supposed true value of the parameter, or the symmetry condition on the unknown component. We respectively implement our methods on real data sets coming from rainfall in US cities and bovine gestation mode comparison under *artificial* or *in vitro* insemination.

To end this part, I present a work on algorithmic estimation for semiparametric symmetric two-component mixture models. In this work, done in collaboration with Didier Chaveau and Laurent Bordes [A12], we propose to adapt the standard EM (Expectation/Maximization) algorithm by adding a nonparametric density estima-

tion step. The role of this additionnal step is to provide a "good candidate" able to replace the expression of the unknown mixed density in the EM algorithm. We give a heuristic argument to justify our approach and show, on various simulated examples, that our method provides very good results in practice with computing times much shorter than ones obtained by classical optimization methods.

Monte Carlo methods

This third part is dedicated to Markov Chain Monte Carlo (MCMC) optimization methods, all investigated in collaboration with Didier Chauveau these last few years.

The first approach we considered, in [A6], was motivated by the acceleration of the rate of convergence of the Hastings-Metropolis (HM) algorithm. The HM algorithm generates, by using a rejection/acceptation mechanism on a data set simulated according to a *proposal* distribution, a Markov chain whose invariant distribution admits a density function coinciding with a target density function (known up to normalizing constant). Various authors have exhibited the links between the rate of convergence of this algorithm and the similarity between the density function of its instrumental distribution and the target density function. Given a HM algorithm converging to a target density function, it seemed interesting to us to estimate nonparametrically its successive densities by generating parallel (i.i.d.) HM algorithms, and to use them as new proposal distributions for the next steps of the HM algorithm. When the target density function has a compact support, we show that our procedure converges asymptotically faster than any HM using an arbitrary proposal distribution. We also present simulations in dimensions 1 and 2 that illustrate the efficiency of our method with respect to a standard HM algorithm (even conveniently calibrated).

The second contribution to this topic, given in [B3], concerns a methodological approach to organize into a hierarchy the efficiency of various possible MCMC simulation strategies. Let us consider two different MCMC algorithms having the same invariant distribution. It should be interesting to know which one converges faster to its stationary distribution. To answer this kind of question it would be necessary to estimate a distance between the succesive densities of the algorithms and their common invariant distribution, but unfortunately the latter (posterior Bayesian) distribution is only known up to a normalizing constant. However, we can notice that the difference of the Kullback distances between the algorithm densities and the target density function do not depend on this normalizing constant. From this remark, it should be interesting to estimate these differences and to monitor their sign and magnitude during the firts iterations to detect, in a synthetic way, which algorithm behaves best. To implement this approach, we used parallel algorithms and the Kullback estimation method proposed by Györfy and Van Der Meulen (1989). The main difficulty of our work has been to check that the technical conditions, needed to prove the consistency of our estimator, are satisfied under reasonable conditions on the ingredients of the algorithms (proposal distribution and target density).

We recently initiated (see Lavanant, 2007) very promising work on *importance sampling* method improvements. The importance sampling method allows one to calculate the expectation of functions of random variables whose distribution density function is known up to a normalising constant. This method, based on the strong law of large numbers, uses a ratio of conveniently chosen functions of i.i.d. instrumental random variables. Similarly to the Hastings-Metropolis algorithm, it is commonly assumed that the quality of this method depends on the similarity between the instrumental density function and the target density function. To increase this similarity, we propose a nonparametric estimator of the target density that uses the initial instrumental sample, the knowledge of the instrumental density function, and the analytic form of the target density function. We conjecture that, using work by Giné and Guillou (2002), we will be able to prove, under suitable assumptions, the uniform almost sure convergence of our estimator ; and to specify in addition a uniform almost sure rate of convergence depending on the dimension of the problem. Note that the resampling step necessary for this nonparametric estimate can be viewed as a smoothed version of the weighted resampling scheme (see Cappé *and al.*, 2005) commonly used in particle filtering or Bayesian methods.

We conclude this part dedicated to Monte Carlo Methods with a work [A11] concerning entropy estimation for Markov chains. Given a transition density kernel, assumed to be easy to simulate, we are interested in how to plot, in a consistent way, the evolution in time of the Kullback distance between the laws of two Markov Chains having different initial conditions, or between the law of a Markov chain and its stationary distribution when the latter is explicitly known and easy to simulate. We propose for this purpose a double Monte Carlo type estimator and prove that it is consistent and asymptotically normally distributed under weak moment conditions. We implemented our method in various situations : to check the stability of first-order autoregressive Gaussian models, and to evaluate the rate of convergence in the CLT for i.i.d. samples generated according to various distributions (Student, Uniform, etc.).

Decreasing step stochastic algorithms

In this last part, I consider the convergence of two decreasing step stochastic algorithms. In [A1], I study the asymptotic behaviour of the simulated annealing algorithm when the potential function is not exactly known but can be estimated recursively with a fast enough uniform almost sure rate of convergence. In a work in progress with Pierre Tarrès [B3], we study a minimal criterion ensuring the convergence of the so-called two armed bandit algorithm when the arms are only assumed to be ergodic.

Chapitre 2

Modèles de Markov cachés

Les travaux décrits dans ce chapitre concernent l'inférence statistique pour divers modèles de Markov cachés. Ils ont fait l'objet de collaborations avec Dominique Bakry, Laurent Bordes, Paolo Giudici, Tobias Rydén et Xavier Milhaud.

2.1 Chaînes de Markov Cachées non-stationnaires

Les modèles de Markov cachés (MMC) constituent une vaste classe de processus à temps discret dont la littérature s'est beaucoup développée ces dernières années. Un MMC est constitué généralement de deux parties : (i) une chaîne de Markov $X = (X_n)_{n\geq 0}$ non-observable ; (ii) un processus stochastique $Y = (Y_n)_{n\geq 0}$ seul observable. La structure du processus Y est généralement définie comme suit : sachant X les variables aléatoires consituant Y sont indépendantes et pour tout $n \geq 0$, la loi conditionnelle de Y_n sachant X ne dépend que de X_n. Lorsque X et Y sont tous les deux à valeurs dans des espaces d'état finis, nous parlerons de chaîne de Markov cachée (CMC). Baum et Petrie (1966) sont les premiers à s'intéresser aux CMC, et ils établissent dans une série d'articles les outils fondamentaux permettant l'étude statistique de ces modèles. Ils montrent en particulier la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance (EMV). Baum *et al.* (1970) complètent l'étude des CMC en proposant un algorithme appelé "forward-backward", extension naturelle de l'algorithme EM, permettant d'ajuster les paramètres d'une CMC. Cette procédure d'estimation, dont la mise en oeuvre s'avère d'une très grande simplicité, a permis l'éclosion de nombreuses applications dans des domaines aussi variés que la reconnaissance de la parole (Rabiner, 1989), la neurophysiologie (Fredkin et Rice, 1992), la biologie (Leroux et Puterman, 1992), ou encore la finance (Rydén, Teräsvirta, et Åsbrink, 1998). Près de vingt ans après les derniers travaux de Baum et Petrie sur l'inférence des CMC, Leroux (1992) aborde de nouveau ce sujet en considérant l'étude des MMC dans le cas où la chaîne de Markov X est à valeurs dans un espace d'état fini mais le processus Y est lui à valeurs dans une espace d'état continu (en abrégé CM2C). Leroux montre dans ce travail que, sous des conditions très faibles, l'EMV pour les CM2C est fortement convergent. Notons au passage que la grande nouveauté de son travail est l'emploi du théorème ergodique pour les processus sous-additifs à la log-vraisemblance des

CM2C. À la suite de ce travail, Bickel and Ritov (1996), Bickel *et al.* (1998), Legland et Mevel (2000), montrent la normalité asymptotique de l'EMV sous une série de conditions de moins en moins fortes. Douc et Matias (2001) montrent la consistance et la normalité asymptotique de l'EMV pour des MMC dans le cas où X et Y sont à valeurs dans des espaces d'état continus (en particulier compact pour X). Dans [A1], en collaboration avec Dominique Bakry et Xavier Milhaud, nous généralisons les travaux de Baum et Petrie au cas des CMC non-stationnaires et montrons la propriété LAN dans ce cadre. Je propose de détailler ci-dessous cette première contribution au domaine des MMC.

Notations et définitions. On considère X une chaîne de Markov à valeurs dans un espace d'état fini $E = \{1, \dots, a\}$, et Y une chaîne de Markov cachée (relativement à X) à valeurs dans un espace d'état fini $F = \{1, \dots, b\}$ où $(a, b) \in \{\mathbb{N} \setminus \{0, 1\}\}^2$. On suppose de plus que X est une chaîne Markov irréductible apériodique sur E de matrice de transition notée $\alpha = (\alpha_{i,j})_{1 \leq i,j \leq a}$ à entrées strictement positives. Pour tout $n \geq 0$, la loi conditionnelle de Y_n sachant $\{X_n = j\}$ est donnée par $\gamma_{k,j} = \mathbb{P}(Y_n = k \mid X_n = j)$. On note $Z = (X_n, Y_n)_{n \geq 0}$ la chaîne de Markov à valeurs dans $E \times F$ de transition Π telle que pour tout $n \geq 0$,

$$\Pi_{(j,k),(i,\ell)} = \mathbb{P}(Z_{n+1} = (j, k) \mid Z_n = (i, \ell)) = \alpha_{i,j} \gamma_{k,j}.$$

Par souci de simplicité on identifie la paramètre θ du modèle à $\Pi = \Pi_\theta$ la transition indexée par $(E \times F)^2$ de la chaîne de Markov Z. Nous faisons les deux hypothèses suivantes :

(H1) À l'instant $n = 0$, la loi de Z_0 est une masse de Dirac en $z^{(0)} = (x^{(0)}, y^{(0)}) \in E \times F$.

(H2) L'espace des paramètres Θ est une partie compacte pour la topologie usuelle de l'ensemble des matrices stochastiques à entrées strictement positives indexées par $(E \times F)^2$. La vraie valeur θ_0 du paramètre appartient à l'intérieur supposé non vide de Θ.

Présentation des résultats. L'idée principale dans l'article de Baum et Petrie (1966) consiste à considérer l'existence, d'après le théorème de Kolmogorov, d'une chaîne stationnaire $Z^* = (X^*, Y^*)$ de transition Π_θ indexée sur $\mathbb{Z}$. Nous introduisons τ l'opérateur de retard sur les suites défini par $\tau(y_n) = y_{n+1}$ pour tout $n \in \mathbb{Z}$. Dans le cas stationnaire, la log-vraisemblance associée à une série d'observations $(y_0, \dots, y_n)$ (vue comme partie d'un vecteur infini $\mathbf{y} = (\dots, y_{-2}, y_{-1}, y_0, y_1, y_2, \dots)$) s'écrit alors sous la forme

$$L_n^*(\theta, \mathbf{y}) = \log \mathbb{P}_\theta(Y_0^* = y_0, \dots, Y_n^* = y_n) = \sum_{k=0}^{n} g_k(\theta, \tau^k \mathbf{y}),$$

où par convention $g_0(\theta, \mathbf{y}) = \mathbb{P}_\theta(Y_0^* = y_0)$ et pour tout $k \geq 1$, $g_k(\theta, \mathbf{y}) = \log \mathbb{P}_\theta(Y_0^* = y_0 \mid \mathbf{Y}^{*-1}_{-k} = \mathbf{y}^{-1}_{-k+1})$, avec la notation $\mathbf{y}_\ell^k = (y_\ell, \dots, y_k)^T$, pour tout $k > \ell$.

L'estimateur du maximum de vraisemblance associé à un échantillon $(Y_0^*, Y_1^*, \ldots, Y_n^*)$ peut donc ensuite s'écrire sous la forme

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_n^*(\theta, Y^*).$$

Baum et Petrie définissent alors une sorte d'entropie $\mathcal{H}(\theta)$ dont l'expression est donnée par :

$$\mathcal{H}(\theta) = \mathbb{E}_{\theta_0}\left(\lim_{k\to\infty} g_k(\theta, Y^*)\right),$$

où $\mathbb{E}_{\theta_0}$ désigne l'espérance sous la loi stationnaire de Y^*, et qui a la particularité d'être la limite ergodique de la log-vraisemblance renormalisée par n. Ils montrent de plus que $\mathcal{H}(\theta)$ est $\mathcal{C}^2$ sur Θ et que l'information de Fisher du modèle, que nous noterons $\mathcal{I}(\theta_0)$, est donnée par $-\mathcal{H}^{(2)}(\theta_0)$, où $\mathcal{H}^{(2)}$ désigne la matrice hessienne de $\mathcal{H}$. Nous supposons comme eux les hypothèses suivantes :

(H3) La matrice d'information de Fisher $\mathcal{I}(\theta_0)$ est inversible.

(H4) (Identifiabilité) La fonction $\mathcal{H}$ vérifie : $\mathcal{H}(\theta) = \mathcal{H}(\theta_0)$ si et seulement si $\theta = \theta_0$.

Rappelons que la condition $\theta \neq \theta_0$ entraîne la relation $\mathcal{H}(\theta) < \mathcal{H}(\theta_0)$ (propriété de contraste). Nous montrons alors les théorèmes suivants :

Théorème 1 *Sous les hypothèses* **(H1–3)**, *la structure statistique associée à la chaîne de Markov cachée Y est localement asymptotiquement normale (LAN).*

Rappelons que la propriété LAN énonce le fait que le comportement de la log-vraisemblance sous $\theta_0 + h/\sqrt{n}$ est asymptotiquement voisin de celui de la log-vraisemblance sous θ_0 à une variable aléatoire gaussienne près plus un terme de translation.

Théorème 2 *Sous les hypothèses* **(H1–4)**, *l'EMV $\hat{\theta}_n$ est fortement convergent, asymptotiquement normalement distribué et efficace. Soit,*

$$\hat{\theta}_n \longrightarrow \theta_0 \quad \mathbb{P}_{\theta_0} - p.s., \quad lorsque \quad n \to +\infty,$$
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}), \quad lorsque \quad n \to +\infty.$$

Pour montrer ces résultats nous avons cherché à exploiter pleinement le cadre de Baum et Petrie. Pour cela nous avons prolongé Z sur $\mathbb{Z}$ en considérant une chaîne égale à $z^{(0)}$ pour tout $n < 0$, et à Z_n pour $n \geq 0$. Par abus cette chaîne sera encore notée Z. Les chaînes Z et Z^* étant considérées indépendantes, on montre sans peine en utilisant l'hypothèse **(H2)**, qu'il existe un temps d'arrêt (appelé temps de couplage) presque sûrement fini T défini par :

$$T = \min\left\{k \in \mathbb{N} \ : \ Z_k = Z_k^*\right\}, \tag{2.1}$$

vérifiant

$$\mathbb{P}_\theta(T > n) \leq \rho^n, \quad \text{où} \quad 0 < \rho < 1 \tag{2.2}$$

est une constante indépendante de θ. On désigne alors par $\tilde{Z}$ la chaîne couplée définie par :

$$\tilde{Z}_n = \begin{cases} Z_n, & \text{si} \quad n < T \\ Z_n^*, & \text{si} \quad n \geq T \end{cases}, \qquad n \geq 0. \tag{2.3}$$

D'après la propriété de Markov forte, les chaînes Z et $\tilde{Z}$ ont même loi. Grâce au couplage (2.3) et à l'inégalité (2.2), nous pouvons dire qu'au bout d'un temps presque sûrement fini, la chaîne stationnaire Z^* (avec laquelle Baum et Petrie savent développer leur théorie) se confond avec la chaîne non-stationnaire couplée $\tilde{Z}$ qui a même loi que la chaîne Z que nous souhaitons étudier. Il ne reste plus qu'à vérifier alors que les fonctionnelles du processus $\tilde{Z}$, impliquées dans l'étude de la log-vraisemblance, oublient suffisamment vite ce qui les différencie de leur analogue dans l'étude de Baum et Petrie (i.e. la loi initiale, et le début de la trajectoire avant l'instant de couplage).

Pour être plus précis sur ce dernier point écrivons, dans le cadre non-stationnaire, la log-vraisemblance d'ordre n associée à une trajectoire $\mathbf{y} \in F^{\mathbb{Z}}$:

$$L_n(\theta, \mathbf{y}) = \log \mathbb{P}_\theta(Y_0 = y_0, \ldots, Y_n = y_n) = \sum_{k=0}^{n} \ell_k(\theta, \mathbf{y}),$$

où par convention $\ell_0(\theta, \mathbf{y}) = \mathbb{P}_\theta(Y_0 = y_0)$ et pour tout $k \geq 1$, $\ell_i(\theta, \mathbf{y}) = \log \mathbb{P}_\theta(Y_i = y_i \mid \mathbf{Y}_0^{i-1} = \mathbf{y}_0^{i-1})$. Nous noterons pour toute fonction f de θ, $f^{(d)}$ sa différentielle d'ordre d.

L'étude reposant sur un développement de Taylor de $L_n(\theta, \mathbf{y})$ et de sa dérivée $L_n^{(1)}(\theta, \mathbf{y})$ autour de θ_0, nous avons été amenés à montrer, en utilisant les techniques de Baum et Petrie, que pour tous $\theta \in \Theta$, $(\mathbf{y}, \mathbf{y}^*) \in F^{\mathbb{Z}} \times F^{\mathbb{Z}}$ vérifiant $y_i = y_i^*$ pour $i > t$, $t \geq 1$; on a

$$|\ell_k^{(d)}(\theta, \mathbf{y}) - g_k^{(d)}(\theta, \tau^k(\mathbf{y}^*))| \leq \alpha_d(k - t), \quad k > t,$$

où $|\cdot|$ désigne la norme du max sur les composantes d'une matrice, et $\alpha_d(\ell)$, indépendant de θ et de y, est le terme général d'une série convergente indexée par $\ell \in \mathbb{N}^*$. On montre alors aisément que pour $d = 0, 1, 2$

$$|L_n^{(d)}(\theta, Y) - L_n^{*(d)}(\theta, Y^*)| = O(1) \quad \mathbb{P}_{\theta_0} - p.s.$$

ce qui permet de "recycler" très simplement les résultats de Baum et Petrie dans le cas des CMC non-stationnaires.

2.2 Chaînes de Markov partiellement cachées

Dans [A8], avec Laurent Bordes, nous introduisons une nouvelle classe de modèles à données manquantes, appelés modèles de Markov partiellement cachés (MMPC).

Ces modèles sont en fait très proches des MMC puisqu'ils servent, comme eux, à modéliser des observations dont la loi dépend d'une chaîne de Markov à espace d'état fini, à ceci près qu'il est possible d'obtenir parfois des informations sur la chaîne de Markov latente. De telles situations se rencontrent souvent en Fiabilité, lorsqu'un système en fonctionnement est soumis à des réparations ou à des remplacements. En effet le niveau de dégradation d'un système est une variable latente évoluant dans le temps et à valeurs dans un espace d'état fini, et certains indicateurs du niveau de dégradation peuvent être mesurés (température, niveau vibratoire, etc.). Il peut être alors intéressant de faire de l'inférence sur les quantités mesurées pour obtenir de l'information sur le comportement propre de la dégradation. Clairement dans une telle situation, un MMC standard semble tout à fait adapté. Cependant si l'on dispose d'informations supplémentaires sur la dégradation, telles que les instants de panne du matériel, on observe alors partiellement la chaîne de Markov sous-jacente lorsque celle-ci se trouve dans l'état le plus dégradé (i.e. la panne). Dans de telles situations, il est clair que les MMPC sont plus adaptés que les MMC. Des phénomènes similaires existent dans le domaine biomédical, voir Guihenneuc-Joyaux *et al.* (2000), Jackson et Sharles (2002), lorsque l'on étudie l'évolution d'une maladie via certains marqueurs. L'état visible (absorbant) de la chaîne cachée peut prendre alors la forme du décès du patient. L'asymptotique dans ce type de problème doit être comprise en "nombre d'individus", et non pas "en temps", puisque l'on introduit alors une transition artificielle allant de l'état visible vers un état de sévérité de la maladie (correspondant à l'entrée d'un autre patient dans l'étude).

Notations et définitions. On considère un processus $Z = (X_n, Y_n)_{n\geq 1}$ tel que : (i) $X = (X_n)_{n\geq 1}$ est une chaîne de Markov à valeurs dans un espace d'état fini $E = \{1, \ldots, a\}$; (ii) $Y = (Y_n)_{n\geq 1}$ est une suite de variables indépendantes conditionnellement X et la loi conditionnelle de Y_n à X ne dépend que de X_n. On suppose de plus que pour tout $n \geq 1$, le couple (X_n, Y_n) est observable si $\{X_n = a\}$, sinon seule la variable Y_n est observée. Le processus Y ainsi défini est le MMPC que nous étudions. La chaîne X est supposée irréductible sur E, de transition $\alpha = (\alpha_{i,j})_{1\leq i,j\leq a}$. Les probabilités de transition sont paramétrées par $\phi \in \Phi$, i.e. $\alpha_{i,j} = \alpha_{i,j}(\phi)$ où $\Phi \subset \mathbb{R}^q$. Le processus Y est supposé prendre ses valeurs dans un espace métrique, séparable et complet F, et les distributions conditionnelles de Y_n sachant X_n sont toutes supposées dominées par une mesure σ-finie μ sur $\mathcal{F}$ (la tribu borélienne sur F). On suppose de plus que les densités conditionnelles appartiennent à une famille paramétrique $\mathcal{G} = \{g(\cdot;\theta);\ \theta \in \Theta\}$, et que le paramètre de cette densité est une fonction de X_n et de ϕ ; la densité de Y_n sachant $\{X_n = i\}$ sera alors notée $g(\cdot;\theta_i(\phi))$. Notons que la paramétrisation la plus courante, voir Rydén (1994), est $\phi = (\alpha_{i,j}, 1 \leq i,j \leq a; \theta_1, \ldots, \theta_a)$ où $\alpha_{i,j}(\cdot)$ et $\theta_i(\cdot)$ sont les projections sur les cordonnées de ϕ. Une autre paramétrisation possible consiste par exemple à prendre $\alpha_{i,j}(\cdot)$ et $\theta_i = (i, \sigma_i^2)$ où $g(\cdot;\theta_i)$ est la densité d'une gaussienne centrée en i et de variance σ_i^2, pour $1 \leq i \leq a$.

Nous définissons maintenant les temps succesifs d'entrée et de sortie de l'état a. Soit

$\tau_1 \geq 1$, le premier instant tel que $X_{\tau_1 - 1} \neq a$ et $X_{\tau_1} = a$, puis

$$\tilde{\tau}_1 = \inf \{ n \geq \tau_1 \ : \ X_{n-1} = a, \ X_n < a \},$$

alors pour $p \geq 2$ on définit :

$$\tau_p = \inf \{ n \geq \tilde{\tau}_{p-1} \ : \ X_n = a \} \quad \text{et} \quad \tilde{\tau}_p = \inf \{ n \geq \tau_p \ : \ X_n < a \}.$$

Les suites $(\tau_p)_{p \geq 1}$ et $(\tilde{\tau}_p)_{p \geq 1}$ constituent respectivement les temps d'entrées en $\{a\}$ et $E_a := E \setminus \{a\}$. L'observation de notre MMPC consiste donc en la suite de variables aléatoires

$$((Y_n)_{\tau_1 \leq n \leq \tilde{\tau}_{k+1} - 1}, (\tau_i)_{1 \leq i \leq k+1}, \ (\tilde{\tau}_i)_{1 \leq i \leq k})_{k \geq 1}.$$

L'écriture qui précéde contient le fait qu'entre les instants τ_i et $\tilde{\tau}_i - 1$, la chaîne de Markov X est observée dans l'état a, tandis qu'elle est dans E_a entre les temps $\tilde{\tau}_i$ et $\tau_{i+1} - 1$, pour tout $1 \leq i \leq k$.

Résultats. Dans [A8], nous montrons que le paramètre ϕ est identifiable pour des classes de densités conditionnelles admettant des mélanges identifiables (voir Lindsay 1995), à la condition que la fonction de $[0,1]^{a^2}$ vers $[0,1]^{\mathbb{N}}$ définie par :

$$\alpha \mapsto \left\{ \prod_{i=0}^{n} \alpha_{x_i, x_{i+1}}; \quad (x_0, x_1, \ldots x_{n+1}) \in \{a\} \times E_a \times \{a\}, \quad n \geq 1 \right\}$$

soit injective. Nous montrons au cas par cas que cette hypothèse est vérifiée pour divers modèles (dégradation pure, périodique). Nous montrons ensuite que sous des conditions de régularité minimales (plus faibles que celles rencontrées dans la littérature MMC), l'EMV est fortement convergent et asymptotiquement normal de ϕ_0. Notons le fait que notre MMPC visite un état spécifique de la chaîne de Markov sous-jacente ce qui permet la régénération du processus observé. Ceci autorise en particulier de factoriser la vraisemblance de notre modèle en plusieurs blocs associés à des bouts de trajectoires indépendantes et identiquement distribuées (de longueurs aléatoires), ce qui facilite grandement l'étude du comportement asymptotique de l'EMV par rapport aux approches qui ont été étudiées par Leroux (1992) ou Bickel *et al.* (1998). La frontière technique entre les MMPC et les MMC se sent nettement lorsque l'on a à faire à des chaînes sous-jacentes périodiques. En effet dans ce cas, l'invariance en loi de la chaîne sous l'opérateur retard n'est plus assurée ; or ce dernier point est, nous l'avons vu dans le paragraphe (2.1), un des arguments techniques les plus important de la théorie de Baum et Petrie.

2.3 Test du rapport de vraisemblance pour les CMC

Dans [A4], avec Paolo Giudici et Tobias Rydén, nous nous intéressons à un test de type rapport de vraisemblance pour des MMC multivariés. Nous utilisons pour cela les résultats de Bickel *et al.* (1998) sur la normalité asymptotique de l'EMV

pour les MMC, ainsi que ceux de Yakowitz et Spragins (1968) sur l'identifiabilité de certains mélanges de lois multivariées. Nous montrons que, comme dans le cas i.i.d, la statistique du rapport de vraisemblance pour les MMC est asymptotiquement distribuée suivant une loi du χ^2. Ces problèmes de test sont cruciaux dans la détection de MMC multivariés gaussiens. Un des points intéressants dans la comparaison de tels modèles multivariés est le fait de pouvoir tester des zéros dans les matrices de précision (inverse de la matrice de corrélation) associées aux différentes densités mélangés. En effet si $Y = (Y^1, \ldots, Y^d)$ est un vecteur aléatoire gaussien de matrice de variance-covariance Σ et de matrice de précision $K = (k_{i,j})_{1 \leq i,j \leq d} = \Sigma^{-1}$, la condition $k_{i,j} = 0$ signifie que les variables aléatoires Y^i et Y^j sont indépendantes conditionnellement à l'ensemble des autres composantes du vecteur Y.

Notations et résultats. En utilisant les mêmes notations que dans le paragraphe 2.2, mais dans un cadre purement MMC (c'est à dire que la chaîne de Markov sous-jacente n'est jamais observable), on note π_ϕ (où ϕ est le paramètre global du modèle) la distribution stationnaire associée à la matrice α_ϕ. On rappelle que dans ce cadre la vraisemblance associée à une série d'observations $(y_1, \ldots, y_n)$ du vecteur $(Y_1, \ldots, Y_n)$ s'écrit de manière compacte sous la forme :

$$p_\phi(y_1, \ldots, y_n) = \pi_\phi \left[\prod_{k=1}^{n} G_\phi(y_k) \alpha_\phi \right] \mathbb{1},$$

où pour tout $y \in F$, $G_\phi(y) = \mathrm{diag}[g(y; \theta_i)]$ et $\mathbb{1}$ est le vecteur de taille $a \times 1$ ne contenant que des 1 (a étant le nombre d'états de la chaîne de Markov sous-jacente). Nous noterons comme précédemment la log-vraisemblance sous la forme $L_n(\phi) = \log p_\phi(Y_1, \ldots, Y_n)$.

On s'intéresse d'abord au test du rapport de vraisemblance pour tester une hypothèse ponctuelle du type : $\mathcal{H}_0 \ : \ \phi = \phi_*$ contre $\mathcal{H}_1 \ : \ \phi \neq \phi_*$, où ϕ_* est une valeur fixée du paramètre. Dans ce cas la statistique du rapport de vraisemblance permettant de tester $\mathcal{H}_0$ est une variable aléatoire notée T_n, définie par :

$$T_n = 2(L_n(\hat{\phi}_n) - L_n(\phi_*)),$$

où $\hat{\phi}_n$ désigne l'EMV sur $\Phi \subset \mathbb{R}^q$. Sous des conditions usuelles de régularité, nous montrons que

$$T_n \xrightarrow{\mathcal{L}} \chi^2_q, \quad \text{lorsque } n \to +\infty. \tag{2.4}$$

Ainsi la procédure habituelle consiste à rejeter $\mathcal{H}_0$ au niveau α si $T_n > \chi^2_{q,1-\alpha}$ où $\chi^2_{q,1-\alpha}$ est le quantile de niveau $(1-\alpha)$ d'un Chi-deux à q degrés de liberté.

On propose dans un second temps de tester une hypothèse nulle composite : $\mathcal{H}_0 \ : \ \phi \in \Phi_*$ contre $\mathcal{H}_1 \ : \ \phi \notin \Phi_*$, où Φ_* est un espace de dimension $(q-r)$ déterminé par un ensemble de $r < q$ contraintes d'équation $R_i(\phi) = 0$, $1 \leq i \leq r$,

agissant sur l'espace des paramètres. Dans ce cas, la statistique de test utilisée, encore notée T_n, est définie par :

$$T_n = 2(\sup_{\phi\in\Phi} L_n(\phi) - \sup_{\phi\in\Phi_*} L_n(\phi)).$$

Nous obtenons alors un résultat analogue à (2.4) (en changeant le nombre de degrés de liberté par $(q-r)$) et la procédure de test usuelle associée.

Application à des données de pollution. Dans [A4] nous appliquons le test du rapport de vraisemblance à un jeu de données trivariées composé de mesures journalières de mortalité, température, et pollution dans la ville de Londres. Notre but est de sélectionner le meilleur MMC multivarié Gaussien pour ce jeu de données en utilisant une chaîne sous-jacente à deux états. Nous mettons pour cela en compétition huit modèles correspondants aux différentes configurations possibles de zéros dans la matrice de précision. Au moyen d'une procédure pas à pas, testant successivement des modèles emboités, nous concluons que mortalité et température sont des variables conditionnellement indépendantes sachant la pollution.

2.4 Mélange markovien de processus de Markov

L'étude des mélanges markoviens de processus de Markov (MMPM) que je propose dans [A8] a été inspirée par la littérature sur les modèles autorégressifs à régime markovien. Ces modèles, qui sont une généralisation des MMC vus dans le paragraphe 2.1, peuvent être définis de la manière suivante. On considère un processus X dont l'évolution est donnée par l'équation :

$$X_n = \sum_{i=1}^{d} a_i(U_n)X_{n-i} + \sigma(U_n)\varepsilon_n, \qquad n \geq d+1,$$

où $(\varepsilon)_{n\geq 1}$ est une suite de variables aléatoires i.i.d. , $U = (U_n)_{n\geq 1}$ est une chaîne de Markov à espace d'état discret ou continu, $(a_i(\cdot))_{i=1,\ldots,d}$ et $\sigma(\cdot)$, ce dernier étant appelé terme de *volatilité stochastique*, sont des fonctions définies sur l'espace d'état de U. On constate ici que contrairement aux MMC classiques, le processus X peut dépendre de son propre passé et non pas seulement de U.

Ce modèle a été utilisé par Hamilton (1989) pour modéliser le produit intérieur brut des États Unis (le processus U modélisant les cycles économiques), voir aussi Hamilton et Susmel (1994), Cai (1994), Garcia et Perron (1996), pour des applications et extensions plus récentes de ce modèle. Notons que les processus autoregressifs linéaires à régime markovien ont aussi trouvé leur place dans le domaine de la gestion de la consommation d'électricité, voir Bar-Shalom et Li (1993) pour les problèmes de détection de rupture d'approvisionnement, et Ji *et al.* (1993), ou Kryshnamurty et Rydén (1998) pour le contrôle automatique de la consommation. Comme souvent en matière de MMC, d'importants travaux ont été consacrés d'abord à l'ajustement de ces modèles alors que peu traitaient des propriétés statistiques des estimateurs

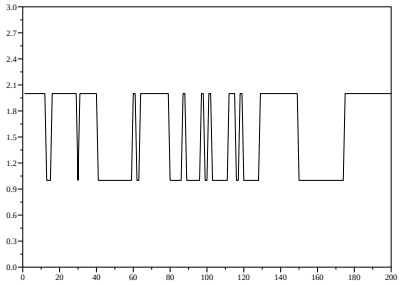
proposés. Citons cependant quelques travaux de référence sur ces questions comme les articles de Krishnamurthy et Rydén (1998), Francq et Roussignol (1998), pour le cas où U est à valeurs dans un espace d'état fini, et Douc *et al.* (2004) pour le cas où U est à valeurs dans un espace d'état continu.

Dans [A8], je définis un nouveau type de processus faisant interagir K processus de Markov à temps discret $X^{[i]} = (X_n^{[i]})_{n\geq 1}$, $1 \leq i \leq K$, indépendants, stationnaires, à valeurs dans un espace d'état mesurable $(E, \mathcal{E})$, de noyau de transition Q^i, $1 \leq i \leq K$, par rapport à une même mesure de référence λ sur $\mathcal{E}$. Le processus $Z = (Z_n)_{n\geq 0}$ auquel je m'intéresse est défini par :

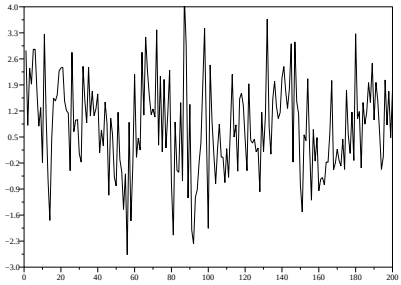
$$Z_n = \sum_{i=1}^{K} \mathbb{1}_{\{U_n=i\}} X_n^{[i]}, \qquad n \geq 1, \tag{2.5}$$

où $U = (U_n)_{n\geq 0}$ est une chaîne de Markov récurrente positive sur $\mathcal{U} = \{1, \ldots, K\}$ non-observable. L'expression (2.5) implique que Z résulte de la mise bout à bout de morceaux de trajectoires issues de processus de Markov indépendants sélectionnés par une chaîne de Markov dont le comportement ne nous est pas accessible. Notons que les MMC sont des sous-modèles des MMPM puisqu'il suffit de considérer pour s'en convaincre le cas où les $X^{[i]}$, $i = 1, \ldots, K$, sont des suites de variables i.i.d. Notons de plus que les MMPM sont capables de modéliser des séries temporelles ayant : (i) des changements abrupts de comportement lorsque, par exemple, la chaîne U change d'état ; (ii) des périodes de grande stabilité dues au maintien de U dans un même état ; (iii) des distributions marginales multimodales en raison de la structure de mélange ; (iv) des effets "feedback" de type phase dans le cas où U quitte brièvement un état, laissant le processus Z dans une zone, puis le retrouvant un peu plus tard dans une zone voisine (si le processus correspondant à l'état de la chaîne de Markov est faiblement mélangeant par exemple).
Afin d'illustrer ces caractéristiques de comportement, nous présentons respectivement au travers des Figures 2.1 et 2.2 une trajectoire de chaîne de Markov (dont seront issus tous les autres processus simulés) et une trajectoire de MMC classique, puis deux MMPM basés sur des AR(1) qui, par construction, ont même loi marginale que la MMC de la Figure 2.1 (voir [A7] pour les détails).

Hypothèses et paramétrisation. Afin d'alléger les notations on considère le cas $K = 2$, et on note $X = X^{[1]}$ et $Y = X^{[2]}$. On considère deux ensembles Φ^i, pour $i = 1, 2$, supposés compacts de $\mathbb{R}^q$. Les noyaux de transition de X et Y sont paramétrés par $\theta \in \Phi^1$ pour Q^1 et $\phi \in \Phi^2$ pour Q^2, et sont supposés appartenir respectivement à des familles paramétriques $\mathcal{K}^1 = \left\{Q_\theta^1(\cdot,\cdot), \theta \in \Phi^1\right\}$, et $\mathcal{K}^2 = \left\{Q_\phi^2(\cdot,\cdot), \phi \in \Phi^2\right\}$. On suppose que pour chaque $\theta \in \Phi^1$ (resp. $\phi \in \Phi^2$) les noyaux de transition Q_θ^1 (resp. Q_ϕ^2) induisent des processus de Markov récurrents positifs admettant une unique mesure invariante de densité q_θ^1 (resp. q_ϕ^2) par rapport à λ. Notons qu'en général la forme analytique de ces densités n'est pas connue explicitement, sauf dans le cas de modèles autorégressifs linéaires d'ordre 1 avec bruit gaussien. Cependant, nous pouvons les caractériser (par définition) comme unique solution d'un problème de

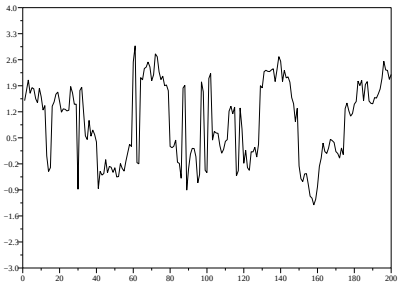


(a) Trajectoire d'une chaîne de Markov obtenue avec $\alpha = \beta = 0.08$.

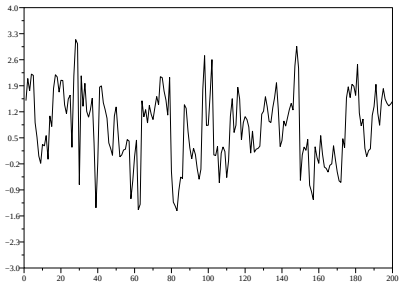


(b) Trajectoire simulée d'un MMC avec lois conditionnelles gaussiennes centrées en 0 et 1.5 et de variance 1.

FIG. 2.1 – Trajectoires d'une chaîne de Markov et d'un MMC associé.



(a) Trajectoire simulée avec $a = 0.9$.



(b) Trajectoire simulée avec $a = 0.7$.

FIG. 2.2 – Deux trajectoires de MMPM construites à partir de la trajectoire (a) : $X \sim AR(1)$ de coefficient $a_1 = a$ et bruit $\mathcal{N}(0, 1 - a^2)$, et $Y \sim AR(1)$ de coefficient $a_2 = a$ et bruit $\mathcal{N}(1.5 * (1 - a), 1 - a^2)$.

point fixe fonctionnel

$$\int_E q_\theta^i(x_1) Q_\theta^i(x_1, \cdot) \lambda(dx_1) = q_\theta^i(\cdot), \qquad i = 1, 2. \tag{2.6}$$

La matrice de transition Π de U est paramétrée par $\gamma = (\alpha, \beta) \in [\delta, 1 - \delta]^2$, avec $0 < \delta < 1$, sous la forme usuelle

$$\Pi_\gamma = \begin{pmatrix} \pi_\gamma(1,1) & \pi_\gamma(1,2) \\ \pi_\gamma(2,1) & \pi_\gamma(2,2) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

Le vecteur de probabilité invariant associé à Π_γ vaut alors $(\pi_\gamma(1), \pi_\gamma(2)) = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$. En conclusion le paramètre à estimer s'écrit donc sous la forme :

$$\vartheta = (\gamma; \theta, \phi) \in \Theta = [\delta, 1 - \delta]^2 \times \Phi^1 \times \Phi^2.$$

Écriture de la vraisemblance. On utilise la notation $\mathbf{Z}_1^n = \{Z_k \ ; \ 1 \leq k \leq n\}$ pour tout processus Z arbitraire. Supposons que U soit obervable, et considérons $u_1^n = (u_1, \ldots, u_n)$ une trajectoire de longueur n de U , et $z_1^n = (z_1, \ldots, z_n)$ une trajectoire de longueur n de Z. La fonction de vraisemblance pour le couple (U, Z) basée sur (u_1^n, z_1^n) peut s'écrire alors $p_\vartheta(u_1^n, z_1^n) = p_\vartheta(z_1^n|u_1^n)p_\vartheta(u_1^n)$, où $p_\vartheta(u_1^n) = \mathbb{P}_\vartheta(\mathbf{U}_1^n = u_1^n)$, et $p_\vartheta(z_1^n|u_1^n)$ désigne la densité conditionnelle de $\mathbf{Z}_1^n$ par rapport à $\{\mathbf{U}_1^n = u_1^n\}$, dont les expressions sont respectivement données par :

$$p_\vartheta(u_1^n) = \pi_\gamma(u_1) \prod_{j=1}^{n-1} \pi_\gamma(u_j, u_{j+1}),$$

et

$$\begin{aligned} &p_\vartheta(z_1^n|u_1^n) = \\ &\int_{E^n} q_\theta^1(x_1) \prod_{j=1}^{n-1} Q_\theta^1(x_j, x_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=2} \lambda(dx_j) \otimes_{j\in\{1,\ldots,n\}/u_j=1} \delta_{z_j}(dx_j) \\ &\times \int_{E^n} q_\phi^2(y_1) \prod_{j=1}^{n-1} Q_\phi^2(y_j, y_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=1} \lambda(dy_j) \otimes_{j\in\{1,\ldots,n\}/u_j=2} \delta_{z_j}(dx_j), \end{aligned}$$

où l'on reconnait la densité jointe des deux vecteurs indépendants $\mathbf{X}_1^n$ and $\mathbf{Y}_1^n$, intégrée composante à composante selon que U n'est pas dans l'état 1, resp. U n'est pas dans l'état 2. Pour obtenir la vraisemblance associée à Z, il suffit de sommer $p_\vartheta(u_1^n, z_1^n)$ sur toutes le valeurs possibles de u_1^n, ce qui donne

$$\begin{aligned} &p_\vartheta(z_1^n) = \\ &\sum_{(u_1,u_2,\ldots,u_n)\in\{1,2\}^n} \pi_\gamma(u_1) \prod_{j=1}^{n-1} \pi_\gamma(u_j, u_{j+1}) \\ &\times \int_{E^n} q_\theta^1(x_1) \prod_{j=1}^{n-1} Q_\theta^1(x_j, x_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=2} \lambda(dx_j) \otimes_{j\in\{1,\ldots,n\}/u_j=1} \delta_{z_j}(dx_j) \\ &\times \int_{E^n} q_\phi^2(y_1) \prod_{j=1}^{n-1} Q_\phi^2(y_j, y_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=1} \lambda(dy_j) \otimes_{j\in\{1,\ldots,n\}/u_j=2} \delta_{z_j}(dy_j). \end{aligned}$$

Remarque. La principale difficulté de ce type de vraisemblance est qu'elle ne peut pas se calculer au moyen d'une formule de récurrence basée sur le filtre de la chaîne U sachant le passé en Z. Or cette technique, qui permet de calculer la vraisemblance déjà compliquée des CMC en des temps linéaires, est à la base de toutes les idées ayant permis l'étude de l'EMV pour les CMC. Ne pouvant disposer d'un tel outil je me suis tourné vers une méthode d'estimation introduite par Rydén (1994), consistant à découper la vraisemblance comme si des paquets successifs de variables de longueur m provenant du processus Z étaient indépendants (ce qui est évidemment faux). Pour un m suffisamment grand (assurant l'identifiabilité du modèle), l'estimateur du maximum de vraisemblance des données tronquées par paquets de longueur

m basé sur $\mathbf{Z}_1^{km}$, pour $k \geq 1$, est défini par

$$\hat{\vartheta}_k = \underset{\vartheta \in \Theta}{\operatorname{argmax}} \prod_{j=1}^{k} p_\vartheta \left(\mathbf{Z}_{(j-1)m+1}^{jm} \right). \tag{2.7}$$

On note ϑ_0 la vraie valeur du paramètre supposée appartenir à l'intérieur non vide de Θ.

Résultats. Dans [A8] je montre, sous des conditions standards de régularité, de mélangeance des processus, et une condition d'identifiabilité (discutée ci-après), que l'estimateur défini en (2.7) est fortement convergent et asymptotiquement normal de ϑ_0. La condition cruciale à vérifier pour ce travail est en réalité la condition d'identifiabilité suivante.

(Identifiabilité) La famille paramétrique $\mathcal{F}^m = \{p_\vartheta(z_1, \ldots, z_m) \ ; \ \vartheta \in \Theta\}$ vérifie : $\forall (\vartheta, \vartheta') \in \Theta^2 \quad$ tels que $\quad p_\vartheta(z_1, \ldots, z_m) = p_{\vartheta'}(z_1, \ldots, z_m) \ \ \lambda^{\otimes m} - p.p., \ \$ on a $\ \vartheta = \vartheta'$.

Cette condition est difficile à vérifier en pratique, car l'écriture des densités $p_\vartheta(z_1, \ldots, z_m)$ suppose la connaissance de la forme analytique des densités des mesures invariantes q^i, $i = 1, 2$. Je montre cependant que cette condition est satisfaite dans le cas de mélanges markoviens de processus autoregressifs d'ordre 1, et donne un moyen pour calculer numériquement la vraisemblance des données tronquées, pour toute valeur ϑ du paramètre, en me passant de la connaissance des q^i, $i = 1, 2$. Ce dernier point utilise une méthode de Monte Carlo pour calculer des intégrales du type (2.6). Je conclus ce travail par une étude bibliographique détaillée concernant les champs d'applications des MMPM, de laquelle il ressort que les MMPM sont des alternatives naturelles aux modèles de type MMC existant en neuroscience (signaux *alpha* et *beta)*, ou en biologie pour l'analyse des canaux d'ions.

Chapitre 3

Modèles de mélange semi-paramétriques

Les travaux décrits dans ce chapitre concernent l'inférence statistique pour des modèles semi-paramétriques de mélanges de lois. Ils ont fait l'objet de collaborations avec Céline Delmas, Laurent Bordes, Didier Chauveau et Stéphane Mottelet.

3.1 Introduction

Les modèles de mélange sont généralement décrits de la manière suivante. Une fonction de répartition (fdr) G est définie sur $\mathbb{R}^p$ ($p \geq 1$) par

$$G(x) = \sum_{i=1}^{k} \lambda_i F_i(x), \qquad x \in \mathbb{R}^p, \tag{3.1}$$

où les paramètres du modèle sont les proportions λ_i du mélange, $1 \leq i \leq k$, vérifiant $\sum_{i=1}^{k} \lambda_i = 1$, les composantes du mélange, c'est à dire les fdr F_i, $1 \leq i \leq k$, et le nombre de composantes k du mélange. Pour estimer les paramètres du modèle (3.1) à partir d'un n-échantillon de vecteurs aléatoires de loi de fdr G, il faut s'affranchir de deux problèmes :

Identifiabilité. Montrer l'unicité de la représentation de G sous la forme du mélange décrit en (3.1).

Identification. Proposer une procédure d'estimation des paramètres inconnus de G.

Pour que ces deux problèmes puissent être résolus, on impose des contraintes sur le modèle et, dans la grande majorité des travaux existants, les composantes F_i du modèle sont supposées appartenir à des familles paramétriques (voir Titterington *et al.*, 1985 ; Mc Lachlan et Peel, 2000). Dans le cas où le nombre de composantes k est connu on parle de modèle *paramétrique*, sinon on parle de modèle *semi-paramétrique*. Hall et Zhou (2003) considèrent pour la première fois le problème de l'estimation des paramètres du modèle (3.1) pour $k = 2$ lorsque F_1 et F_2 n'appartiennent pas à

des familles paramétriques (mais sont les fdr de vecteurs aléatoires de composantes indépendantes). Ces auteurs donnent des conditions générales permettant d'assurer l'identifiabilité et l'identification des paramètres de leur modèle pour $p \geq 3$, considérant le problème ouvert pour $1 \leq p < 3$. Ce travail nous a conduit dans [A9], à considérer l'identifiabilité du modèle (3.1) pour $p = 1$ et $k = 2$ lorsque F_1 et F_2 ne sont pas supposées appartenir à des familles paramétriques. Nous en sommes venus à considérer le modèle :

$$G(x) = \lambda F(x - \mu_1) + (1 - \lambda)F(x - \mu_2), \qquad x \in \mathbb{R}, \tag{3.2}$$

où F est une fdr inconnue admettant une densité f paire, $\lambda \in]0, 1[$ est une proportion de mélange inconnue et $\mu_1 \neq \mu_2$ sont deux paramètres de localisation inconnus.

Sur la base des travaux développés dans [A9], Céline Delmas nous a proposé, à Laurent Bordes et moi-même, d'analyser un modèle proche du modèle (3.2) afin d'étudier la différence d'expression de gènes de puces ADN (*microarrays*). Le modèle est le suivant :

$$G(x) = \lambda F_0(x) + (1 - \lambda)F(x - \mu), \qquad x \in \mathbb{R}, \tag{3.3}$$

où F_0 est une fdr connue et F est la fdr d'une densité paire inconnue, μ est un paramètre de localisation inconnu et $\lambda \in]0, 1[$ est une proportion de mélange inconnue.

3.1.1 Identifiabilité

Modèle (3.2). Dans [A9], en notant $\mathcal{F}$ l'ensemble des fdr de distributions symétriques autour de zéro et $\Delta = \{(x, x)\ ;\ x \in \mathbb{R}\}$, nous montrons que pour

$$((\lambda, \mu_1, \mu_2, F), (\lambda', \mu_1', \mu_2', F')) \in \left(]0, 1/2[\times(\mathbb{R}^2 \setminus \Delta) \times \mathcal{F}\right)^2,$$

la condition

$$\lambda F(x - \mu_1) + (1 - \lambda)F(x - \mu_2) = \lambda' F'(x - \mu_1') + (1 - \lambda')F'(x - \mu_2') \tag{3.4}$$

implique $(\lambda, \mu_1, \mu_2, F), = (\lambda', \mu_1', \mu_2', F')$ ce qui prouve l'identifiabilité du modèle (3.2) sur l'espace des paramètres $]0, 1/2[\times(\mathbb{R}^2 \setminus \Delta) \times \mathcal{F}$. Notons que la preuve de ce résultat passe par une discussion longue et exhaustive de l'équation (3.4) après transformation au sens de Fourier. Parallèlement à nos travaux, Hunter *et al.* (2007) montre le même résultat et donnent une condition suffisante d'identifiabilité pour un modèle de mélange similaire à trois composantes.

Modèle (3.3). En général, ce modèle n'est pas identifiable et nous indiquons dans [A10] quelques contre-exemples à l'identifiabilité. On peut montrer cependant, en utilisant la méthode des moments, que ce modèle est *localement* identifiable, et même *faiblement* identifiable, si l'on impose des contraintes supplémentaires sur les paramètres euclidiens du modèle. Les deux notions citées précédemment en italique seront expliquées un peu plus loin. Parmi les classes de conditions que nous donnons dans [A10], en voici une qui illustre à quel point ce problème est délicat.

Exemple d'identifiabilité. Soit $\mathcal{F}_3$ l'ensemble des densités paires admettant des moments d'ordre 3. On note f_0 la densité associée à la fdr F_0 et on suppose que la fonction caractéristique $\hat{f}_0$ est strictement positive sur $\mathbb{R}$. Le modèle (3.3) exprimé en terme de densités, soit :

$$g(x) = \lambda f_0(x) + (1-\lambda) f(x-\mu), \qquad x \in \mathbb{R}, \tag{3.5}$$

est identifiable si

$$(\lambda, \mu, f) \in]0,1[\times\mathbb{R}^* \times \mathcal{F}_3 \qquad \text{et} \qquad \theta \neq \theta_0 + \frac{k \pm 2}{3k}\mu^2, \qquad k \in \mathbb{N}^*, \tag{3.6}$$

où θ et θ_0 sont respectivement les moments d'ordre 2 de f et f_0.

On constate ici que le modèle (3.3) est identifiable du moment que (μ, θ) est situé sur $\mathbb{R}^*\times]0,+\infty[$ privé d'un ensemble de mesure nulle (d'où la terminologie d'identifiabilité *faible*).

3.1.2 Identification et résultats asymptotiques

Modèle (3.2). Le premier élément clef dans l'identification du modèle (3.2) a été le constat suivant. En notant $\eta = \mu_2 - \mu_1$, il est facile de voir que

$$F(x) = \frac{1}{1-\lambda} G(x+\mu_2) + \frac{-\lambda}{1-\lambda} F(x+\eta), \quad \forall x \in \mathbb{R}, \tag{3.7}$$

et donc en utilisant (3.7) comme formule de récurrence ℓ fois il vient

$$F(x) = \frac{1}{1-\lambda} \sum_{k=0}^{\ell-1} \left(\frac{-\lambda}{1-\lambda}\right)^k G(x+\mu_2+k\eta) + \left(\frac{-\lambda}{1-\lambda}\right)^\ell F(x+\ell\eta), \quad \forall x \in \mathbb{R}. \tag{3.8}$$

Il est alors facile de constater que le dernier terme de (3.8) tend vers 0 lorsque ℓ tend vers l'infini, ce qui nous donne la formule d'inversion suivante :

$$F(x) = \frac{1}{1-\lambda} \sum_{k\geq 0} \left(\frac{-\lambda}{1-\lambda}\right)^k G(x+\mu_2+k\eta), \quad \forall x \in \mathbb{R}. \tag{3.9}$$

Nous introduisons alors les opérateurs linéaires bornés A_θ and A_θ^{-1} définis par

$$A_\theta = \lambda\tau_{\mu_1} + (1-\lambda)\tau_{\mu_2} \quad \text{and} \quad A_\theta^{-1} = \frac{1}{1-\lambda} \sum_{k\geq 0} \left(\frac{-\lambda}{1-\lambda}\right)^k \tau_{-\mu_2-k\eta}, \tag{3.10}$$

où τ_μ ($\mu \in \mathbb{R}$) est l'opérateur de translation défini par $\tau_\mu f = f(\cdot - \mu)$. Avec ces nouvelles notations, l'équation (3.8) devient $G = A_\theta F$ et (3.9) devient $F = A_\theta^{-1} G$.

Le deuxième élément clef dans l'analyse du modèle (3.2) repose sur la remarque suivante. Soit $F_\theta = A_\theta^{-1} G = A_\theta^{-1} A_{\theta_0} F_0$ où $\theta \in \Theta$; clairement si $\theta = \theta_0$ (où θ_0 désigne la vraie valeur du paramètre) nous avons $F_\theta = F_0$, et donc le principe d'invariance $F_0(x) = 1 - F_0(-x)$ d'après l'hypothèse de symétrie sur F_0. On introduit l'opérateur

de symétrisation S_r défini par $S_rF(\cdot) = 1 - F(-\cdot)$. La remarque précédente peut alors être ré-exprimée sous la forme : si $\theta = \theta_0$, alors $A_\theta^{-1}G = S_rA_\theta^{-1}G$, ou de manière équivalente $G = A_\theta S_r A_\theta^{-1}G$, en appliquant A_θ sur le membre de droite de la dernière égalité. Mais que peut-on dire lorsque l'on prend le problème dans l'autre sens ? La réponse est donnée par le théorème suivant.

Théorème 3 *Considérons la modèle (3.2) avec F_0 une fdr associée à une loi symétrique et $\theta_0 \in \Theta$. Si pour $\theta \in \Theta$, nous avons $G = A_\theta S_r A_\theta^{-1}G$, alors $\theta = \theta_0$.*

Une conséquence du théorème précédent est que si G est connue, nous pouvons espérer retrouver la valeur θ_0 de θ en minimisant une mesure de discrépance entre G et $G_\theta = A_\theta S_r A_\theta^{-1}G$. Comme G n'est pas connue, mais peut en revanche être estimée, nous avons choisi de considérer la mesure de discrépance K définie par :

$$K(\theta) \equiv K(\theta; G) = \int_{\mathbb{R}} (G_\theta(x) - G(x))^2 dG(x), \quad \theta \in \Theta. \tag{3.11}$$

Nous montrons que si F est assez régulière et si G est connue, alors K est une fonction de contraste pour le paramètre euclidien, c'est à dire que pour tout $\theta \in \Theta$, $K(\theta; G) \geq 0$ et $K(\theta; G) = 0$ si et seulement si $\theta = \theta_0$.

Ce résultat suggère donc d'estimer θ au moyen de l'approche dite du *minimum de contraste*, soit

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} K(\theta; \hat{G}_n),$$

où $\hat{G}_n$ est un estimateur de la fdr G. Quand les paramètres de localisation μ_1 et μ_2 sont inconnus, afin de pouvoir appliquer des méthodes d'optimisation différentiables, nous sommes contraints de considérer une version régularisée de la fdr empirique (alors que lorsque les paramètres de localisation sont connus la fdr empirique suffit) soit $\tilde{G}_\theta^{(n)} = A_\theta S_r A_\theta^{-1}\tilde{G}_n$, avec $\tilde{G}_n(x) = \displaystyle\int_{-\infty}^{x} \hat{g}_n(y)dy$ et

$$\hat{g}_n(x) = \frac{1}{b_n}\int_{\mathbb{R}} q\left(\frac{x-y}{b_n}\right) d\hat{G}_n(y),$$

où $\hat{G}_n$ désigne la fdr empirique et $b_n \searrow 0$, $nb_n \to +\infty$, et $\sqrt{nb_n} = O(1)$ lorsque $n \to +\infty$. On montre alors le résultat principal suivant :

Théorème 4 *Si la fdr F_0 est strictement croissante et Lipschitzienne sur $\mathbb{R}$, alors $\hat{\theta}_n$ converge presque sûrement vers θ_0 lorsque n tend vers l'infini. Si de plus F_0 est deux fois continuement différentiable avec $F^{(2)}$ dans $L^1(\mathbb{R})$, alors $n^{1/4-\alpha}(\hat{\theta}_n - \theta_0) = o_{a.s.}(1)$, pour tout $\alpha > 0$.*

Une fois le paramètre euclidien estimé, il est alors légitime d'estimer respectivement la fdr F inconnue et sa densité au moyen de la formule d'inversion (3.10) en employant respectivement les estimateurs

$$\hat{F}_n(x) = A_{\hat{\theta}_n}\hat{G}_n \qquad \text{et} \qquad \hat{f}_n(x) = A_{\hat{\theta}_n}\hat{g}_n.$$

Sous des conditions analogues à celles du Théorème 4, on montre alors la convergence presque sûre en norme $L^\infty(\mathbb{R})$ de $\hat{F}_n$ vers F à la vitesse $n^{-1/4+\alpha}$ pour tout $\alpha > 0$, ainsi que la convergence presque sûre de $\hat{f}_n$ vers f.

Modèle (3.3). Les étapes de l'identification du modèle (3.3) sont similaires à celles du modèle (3.2). Nous notons tout d'abord qu'il existe un formule d'inversion simple donnée par :

$$F(x) = \frac{1}{p}\left(G(x+\mu) - (1-p)F_0(x+\mu)\right), \quad \forall x \in \mathbb{R}.$$

Or par hypothèse, F est la fdr d'une loi symétrique, ce qui impose la relation $F(x) = 1 - F(-x)$ pour tout $x \in \mathbb{R}$, d'où l'idée naturelle de comparer les fonctions

$$H_1(x;\theta,G) = \frac{1}{p}G(x+\mu) + \frac{1-p}{p}F_0(x+\mu), \qquad x \in \mathbb{R},$$

et

$$H_2(x;\theta,G) = 1 - \frac{1}{p}G(\mu-x) + \frac{1-p}{p}F_0(\mu-x), \qquad x \in \mathbb{R},$$

qui sous la condition $\theta = \theta_0$ vérifient $H_1(\cdot;\theta,G) = H_2(\cdot;\theta,G)$. Ainsi, si d est une distance entre deux fonctions et si $d(\theta) = d(H_1(\cdot;\theta,G), H_2(\cdot;\theta,G))$, nous avons $d(\theta_0) = 0$, et pouvons estimer θ par

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} d_n(\theta), \tag{3.12}$$

où d_n désigne une estimation empirique de d qui s'obtient en replacant G par $\hat{G}_n$ ou sa version régularisée $\tilde{G}_n$ (voir [A10] et [B1] pour les détails). Dans [A10] la distance choisie provient de la norme $L^q(\mathbb{R})$, alors que dans [B1] la norme considérée est $L^2(G)$ (écart quadratique intégré au sens de g sur $\mathbb{R}$). Dans [A10] nous montrons, sous des conditions minimales, la convergence presque sûre de l'estimateur (3.12) vers la vraie valeur θ_0 du paramètre. Nous établissons dans [B1], sous des conditions un peu plus fortes, la normalité asymptotique de nos estimateurs au sens suivant :

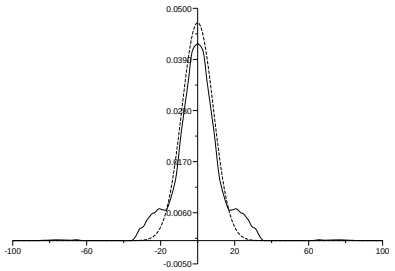
$$\sqrt{n}\left(\mu_n - \mu_0, \hat{p}_n - p_0, \hat{F}_n(\cdot) - F(\cdot)\right)^T \rightsquigarrow \mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)^T, \tag{3.13}$$

dans $\mathbb{R}^2 \times D(\mathbb{R})$ lorsque n tend vers l'infini, $D(\mathbb{R})$ désignant l'espace des fonctions cad-lag sur $\mathbb{R}$ et $\mathcal{G}$ un processus gaussien de matrice de variance-covariance Σ que nous explicitons. Nous proposons de plus un estimateur fortement convergent de la matrice Σ qui nous permet d'utiliser (3.13) dans le cadre de divers problèmes de tests : (i) test du χ^2 pour l'adéquation à un modèle (μ_*, p_*, F_*) fixé ; (ii) test du χ^2 pour la symétrie sur F. Notons que Hunter *et al.* (2007) prouvent aussi la consistance et la normalité asymptotique de leur estimateur au prix de conditions plus abstraites que les notres.

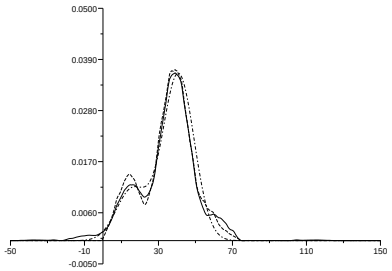
Un projet à court terme est l'obtention de résultats analogues pour le modèle (3.2) et l'étude d'un test du type Kolmogorov.

3.1.3 Applications

Modèle 3.2 : données de pluviométrie. Nous avons étudié les données de précipitation de 70 villes des États-Unis (et Porto Rico) (Statistical Abstracts of the United States, 1975 ; voir McNeil, 1977). Pour cela nous avons ajusté deux modèles. Le premier est le modèle (3.2) pour lequel on note $\hat{\lambda}$, $\hat{\mu}_1$, $\hat{m}_2$, $\hat{f}$ les estimateurs de λ, μ_1, μ_2, et f. Le second est une version paramétrique du modèle (3.2) où l'on suppose que f est la densité de la loi gaussienne centrée et de variance σ^2 (notée $\mathcal{N}(0, \sigma^2)$). Les estimateurs des paramètres inconnus du second modèle sont notés $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{m}_2$, $\tilde{\sigma}^2$, et calculés suivant la méthode du maximum de vraisemblance.



(a) Graphe de $\hat{f}$ (trait plein) et graphe de la densité d'une $\mathcal{N}(0, \tilde{\sigma}^2)$ (traits tirés).



(b) Graphe de $\hat{g}$ (trait plein), graphe de $\hat{\lambda}\tilde{f}(\cdot - \hat{\mu}_1) + (1 - \hat{\lambda})\tilde{f}(\cdot - \hat{\mu}_2)$ (tiré) $\tilde{\lambda}\mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}^2) + (1 - \tilde{\lambda})\mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}^2)$ (tiré-pointillé).

FIG. 3.1 – Paramètres estimés pour le modèle (3.2) : $\hat{\mu}_1 = 13.107$ (3.299), $\hat{\mu}_2 = 39.056$ (1.395), $\hat{\lambda} = 0.171$ (0.078). Paramètres estimés pour le modèle $\lambda * \mathcal{N}(\mu_1, \sigma^2) + (1 - \lambda) * \mathcal{N}(\mu_2, \sigma^2)$: $\tilde{\mu}_1 = 15.715$ (2.220), $\tilde{\mu}_2 = 40.773$ (1.297), $\tilde{\lambda} = 0.235$ (0.060) et $\tilde{\sigma} = 8.504$ (1.187), entre parenthèse figurent les écart-types estimés.

La Figure 3.1(a) montre $\tilde{f}$, l'estimateur de f, ainsi que la densité de la loi $\mathcal{N}(0, \tilde{\sigma}^2)$. La Figure 3.1(b) montre $\hat{g}$, l'estimateur à noyaux de g avec les reconstructions de g obtenues à partir des deux modèles. Les écarts-types sont estimés par Jacknife. On peut voir que le modèle (3.2) offre une meilleure adéquation aux données pour ce qui concerne la reconstruction de g et que la densité f ne semble pas appartenir à une famillle paramétrique usuelle.

Modèle 3.3 : comparaison de gestation des bovins. Dans [A10] nous étudions un jeu de données issu de la technologie des puces ADN afin de détecter les gènes statistiquement différement exprimés suivant que leur porteur est issu d'un mode d'insémination *artificiel* ou *in vitro*. Les données se présentent sous la forme de réalisations de variables aléatoires S_i, $i = 1, \ldots, 10214$, i.i.d. et distribuées suivant (3.3), où F_0 est la fdr d'une loi de Student à 18 degrés de liberté. Les paramètres sont estimés en recherchant l'argument du minimum de d_n sur une discrétisation fine de l'espace paramétrique restreint à $[0.01, 0.1] \times [0.5, 1.5]$ (voir la Figure 3.2 pour l'allure de d_n). Les estimations obtenues dans ce modèle sont $\hat{p} = 0.037$ et $\hat{\mu} = 1.05$ et une procédure à la Benjamini-Hochberg permet de détecter environ 370

gènes potentiellement différemment exprimés.

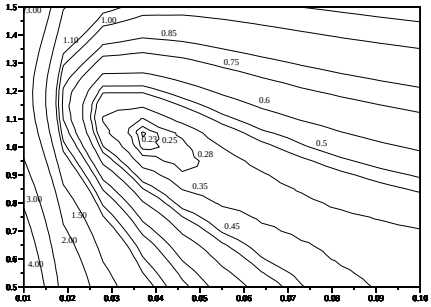
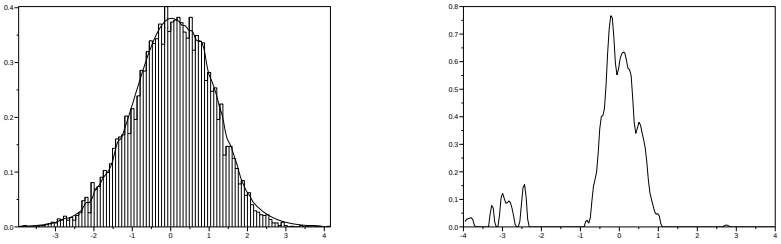


FIG. 3.2 – Courbes de niveau de $(p, \mu) \mapsto d_n(p, \mu)$ pour le jeu de données sur la gestation des bovins avec $(p, \mu) \in [0.01, 0.1] \times [0.5, 1.5]$.



(a) Histogramme des données comparé avec $(1 - \hat{p})f_0(\cdot) + \hat{p}f(\cdot - \hat{\mu})$ pour $\hat{p} = 0.037$ et $\hat{\mu} = 1.05$.

(b) Estimation de la densité inconnue f.

FIG. 3.3 – Reconstruction du mélange à partir de l'estimation de (p, μ, f) et de l'estimateur de la densité de f.

La figure 3.3 indique clairement la robustesse de notre méthode face à l'hypothèse technique de symétrie sur la composante inconnue. On constate en effet que le contraste n'est pas tout à fait nul (0.23) pour la valeur estimée du paramètre mais que notre méthode détecte, pour cette valeur, une densité presque symétrique permettant de très bien ajuster le modèle, comme l'indique 3.3(a).

3.2 Algorithme EM semi-paramétrique

La méthode présentée dans cette section est l'objet d'un travail [A12] fait en collaboration avec Laurent Bordes et Didier Chauveau. Notons qu'une approche sensiblement voisine a été considérée récemment par Robin *et al.* (2007) dans le cadre du modèle (3.6) en lien avec l'analyse des puces ADN.

3.2.1 Analyse de l'algorithme EM

On considère le modèle de mélange semi-paramétrique

$$g_\varphi(x) = g(x|\varphi) = \sum_{j=1}^{m} \lambda_j f(x-\mu_j), \qquad x \in \mathbb{R}, \tag{3.14}$$

où $\varphi = (\theta, f) = ((\lambda_j, \mu_j)_{j=1,\dots,m}, f) \in \Phi = \Theta \times \mathcal{F}$ désigne le paramètre inconnu du modèle avec

$$\Theta = \left\{ (\lambda_j, \mu_j)_{j=1,\dots,m} \in \{]0,1[\times\mathbb{R}\}^m \ ; \quad \sum_{j=1}^{m} \lambda_j = 1, \quad \mu_i \neq \mu_j,\ 1 \leq i < j \leq m \right\}.$$

et $\mathcal{F}$ désigne l'ensemble des densités paires sur $\mathbb{R}$. Avant de présenter l'algorithme EM (Expectation/Maximisation) pour le modèle (3.14), nous proposons de rappeler son pendant dans le cadre d'un modèle de mélange paramétrique.

Algorithme EM paramétrique. Le but de l'algorithme EM est d'approcher l'estimateur du maximum de vraisemblance pour les modèles à données manquantes au moyen d'une procédure itérative simple. Soit $\mathbf{x} = (x_1, \dots, x_n)$ la réalisation d'un n-échantillon i.i.d. suivant une densité de probabilité $g(\cdot|\theta)$ définie par

$$g_\theta(x) = g(x|\theta) = \sum_{j=1}^{m} \lambda_j\, f(x|\xi_j), \tag{3.15}$$

où $\theta = (\lambda_j, \xi_j)_{j=1,\dots,m}$ est le paramètre euclidien du modèle et $f(\,\cdot\,|\xi)$ appartient à une famille paramétrique notée $\mathcal{F}_p$. Dans ce type de modèle, la principale difficulté provient du fait que l'on ne connait pas les provenances des x_i, c'est à dire suivant quelle composante ils ont été tirés. Supposons pourtant que l'on connaisse pour chaque x_i la composante $z_i \in \{1, \dots, m\}$ du mélange dont il est issu, i.e.

$$X_i|Z_i = j \sim f(\cdot|\xi_j) \ \text{ et } \ \mathbb{P}(Z_i = j) = \lambda_j, \quad j = 1, \dots, m.$$

Dans ce cas, la densité pour une observation complètes $y = (x, z)$ s'écrit

$$h(y|\theta) = h((x,y)|\theta) = \lambda_z f(x|\xi_z),$$

et donc pour une série d'observations complète $\mathbf{y} = (y_1, \dots, y_n)$, la log-vraisemblance s'écrit $\log \mathbf{h}(\mathbf{y}|\theta) = \sum_{i=1}^{n} \log h(y_i|\theta)$. Notons que l'EMV pour le modèle (3.15), basé sur la maximisation de $\log \mathbf{h}(\mathbf{y}|\theta)$, est en général explicite du moment que la

famille paramétrique $\mathcal{F}_p$ l'autorise (famille exponentielle, gaussienne, etc.). Au lieu de maximiser la log-vraisemblance des données observées, l'algorithme EM maximise de manière itérative l'opérateur

$$Q(\theta|\theta^t) = \mathbb{E}[\log \mathbf{h}(\mathbf{y}|\theta)|\mathbf{x}, \theta^t],$$

où θ^t est la valeur courante du paramètre à l'étape t (voir Wu, 1983, pour une présentation générale de l'algorithme EM ainsi que la preuve de la convergence).

L'itération $\theta^t \to \theta^{t+1}$ définie dans la cadre précédent est donnée par

1. Étape E : `calculer` $Q(\theta|\theta^t)$.
2. Étape M : `prendre` $\theta^{t+1} = \text{argmax}_{\theta\in\Theta} Q(\theta|\theta^t)$.

On note que l'opérateur $Q(\cdot|\theta^t)$ est une espérance au sens de la loi $\mathbf{k}(\mathbf{y}|\mathbf{x}, \theta^t)$ c'est à dire de $\mathbf{y}$ sachant $\mathbf{x}$, pour la valeur θ^t du paramètre. Dans un modèle de mélange, nous avons en particulier

$$\mathbf{k}(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i=1}^{n} k(y_i|x_i, \theta) = \prod_{i=1}^{n} k(z_i|x_i, \theta),$$

puisque $(z_i|x_i), i = 1, \ldots, n$, sont indépendants. Les z_i étant discrets, leur loi est donnée par la formule de Bayes

$$k(j|x, \theta^t) = \mathbb{P}(Z = j|x, \theta^t) = \frac{\lambda_j^t f(x|\xi_j^t)}{\sum_{\ell=1}^m \lambda_\ell^t f(x|\xi_\ell^t)}, \qquad j = 1, \ldots, m. \tag{3.16}$$

Dans le cas où les densités $f(\cdot|\mu)$ impliquées dans le modèle (3.15) sont de la forme $f(\cdot - \mu)$, nous obtenons simplement

$$k(j|x, \theta) = \frac{\lambda_j^t f(x - \mu_j^t)}{\sum_{j=1}^m \lambda_j^t f(x - \mu_j^t)}, \qquad j = 1, \ldots, m.$$

Lorsque les données $(\mathbf{x}, \mathbf{z})$ sont entièrement observées, l'EMV pour les proportions du mélange est indépendant de f et vaut composante à composante

$$\hat{\lambda}_j = \frac{\sum_{i=1}^n \mathbb{I}_{z_i=j}}{n}, \quad j = 1, \ldots, m, \tag{3.17}$$

(où $\mathbb{I}_{z_i=j}$ vaut 1 si $z_i = j$). Considérons enfin le cas ou les ξ_j's sont des paramètres d'espérance, i.e. $\mathbb{E}(X|Z = j) = \xi_j$ (ce qui n'impose pas le fait que $f(\cdot|\xi)$ soit paire). Dans ce cas des estimateurs sans biais naturels, fortement convergents des μ_j sont les moyennes empiriques calculées dans chacune des sous-populations

$$\hat{\mu}_j = \frac{\sum_{i=1}^n x_i \mathbb{I}_{z_i=j}}{\sum_{i=1}^n \mathbb{I}_{z_i=j}}, \quad j = 1, \ldots, m. \tag{3.18}$$

Lorsque les données $\mathbf{z}$ sont manquantes, l'EMV pour les données complètes est remplacé par l'estimateur EM paramétrique. L'implémentation de l'étape M pour l'itération conduisant de $\theta^t \to \theta^{t+1}$ se fait de manière tout à fait standard par une

maximisation directe de $Q(\cdot|\theta^t)$, voir Redner et Walker (1984), ce qui se voit au travers des équations (3.17) et (3.18), où chaque indicatrice inconnue $\mathbb{I}_{z_i=j}$ est remplacée par son espérance conditionnelle aux données observées x_i, et pour la valeur courante du paramètre, soit $\mathbb{E}(\mathbb{I}_{Z_i=j}|x_i,\theta^t) = k(j|x_i,\theta^t)$ donnée par (3.16).

1. Étape E : `pour` $i = 1,\ldots,n$ `et` $j = 1,\ldots,m$, `calculer` $k(j|x_i,\theta^t)$.
2. Étape M : `réactualiser` θ^{t+1} `par :`

$$\lambda_j^{t+1} = \frac{1}{n}\sum_{i=1}^{n} k(j|x_i,\theta^t) \tag{3.19}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{n} k(j|x_i,\theta^t)\, x_i}{\sum_{i=1}^{n} k(j|x_i,\theta^t)}, \quad j = 1,\ldots,m. \tag{3.20}$$

Algorithme EM formel en semi-paramétrique. Revenons au modèle (3.14) pour lequel nous pouvons encore définir les densités des données observées et complètes

$$\begin{aligned} g_\varphi(x) &= g(x|\varphi) = \sum_{j=1}^{m} \lambda_j f(x-\mu_j), \\ h(y|\varphi) &= h((x,z)|\varphi) = \lambda_z f(x-\mu_z). \end{aligned}$$

Formellement, la log-vraisemblance associée aux $\mathbf{x}$ pour la valeur φ du paramètre peut s'écrire

$$L_{\mathbf{x}}(\varphi) = \sum_{i=1}^{n} \log g(x_i|\varphi).$$

De la sorte, nous sommes quasiment en mesure de proposer un algorithme de type EM, à ceci près qu'il nous faut définir une valeur courante du paramètre $\varphi^t = (\theta^t, f^t)$ à l'itération t, et l'opérateur

$$Q(\varphi|\varphi^t) = \mathbb{E}[\log \mathbf{h}(\mathbf{y}|\varphi)|\mathbf{x},\varphi^t].$$

Comme dans le cas paramétrique, l'espérance est prise suivant la loi de $\mathbf{y}$ sachant $\mathbf{x}$, pour la valeur φ^t du paramètre :

$$\mathbf{k}(\mathbf{y}|\mathbf{x},\varphi^t) = \prod_{i=1}^{n} k(y_i|x_i,\varphi^t) = \prod_{i=1}^{n} k(z_i|x_i,\varphi^t),$$

où

$$k(j|x,\varphi^t) = \mathbb{P}(Z=j|x,\varphi^t) = \frac{\lambda_j^t f^t(x-\mu_j^t)}{\sum_{\ell=1}^{m} \lambda_\ell^t f^t(x-\mu_j^t)}, \quad j = 1,\ldots,m. \tag{3.21}$$

Ainsi $Q(\varphi|\varphi^t)$ est donné par

$$Q(\varphi|\varphi^t) = \sum_{i=1}^{n}\sum_{j=1}^{m} k(j|x_i,\varphi^t)\big[\log(\lambda_j) + \log f(x_i-\mu_j)\big].$$

Pour une initialisation $\varphi^0 = (\theta^0, f^0)$, un algorithme formel pour estimer φ est donc

1. Étape E : `calculer` $Q(\varphi|\varphi^t)$ `en utilisant (3.21) et (3.22).`
2. Étape M : `choisir` φ^{t+1} `qui maximise` $Q(\varphi|\varphi^t)$.

La principale difficulté de l'algorithme précédent est de trouver f^{t+1} tel que φ^{t+1} $\varphi^{t+1} = (\theta^{t+1}, f^{t+1})$ maximise $Q(\cdot|\varphi^t)$. Dans le paragraphe suivant nous indiquons une solution heuristique à ce problème dans le cadre du modèle (3.14).

Méthodologie pour l'EM semi-paramétrique. L'idée heuristique permettant l'implémentation d'un algorithme EM dans notre cas vient du fait que les paramètres des composantes du mélange sont en réalité des espérances. L'idée est alors de procéder itérativement de la manière suivante :

1. calculer un estimateur f^{t+1} de f, en utilisant θ^t ;
2. remplacer f^{t+1} dans l'étape M de l'algorithme EM (3.19)–(3.20) pour calculer θ^{t+1}.

Le principe d'estimation de f que nous avons retenu est le suivant : étant donné le paramètre euclidien θ (ou un estimateur θ^t), et le fait que les composantes du mélange sont égales à un paramètre de localisation près, nous conjecturons, qu'en pratique, il est raisonnable d'estimer f au moyen d'un estimateur à noyau basé sur les données convenablement recentrées (la justification de ce point sera détaillée plus tard). Dans la suite, nous noterons $\tilde{x}_i$ la i-ème observation recentrée et par $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ le vecteur correspondant.

Afin de décrire l'esprit de la méthode, considérons la *situation idéale* où nous aurions accès à $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ et où θ serait connu. Un estimateur consistant de f s'obtiendrait alors en effectuant les étapes suivantes :

1. calculer $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_n)$, où $\tilde{x}_i = x_i - \mu_{z_i}$, $i = 1, \ldots, n$;
2. calculer l'estimateur à noyau de la densité f au moyen d'un noyau K et d'une largeur de fenêtre h_n,

$$\hat{f}_{\tilde{\mathbf{x}}}(u) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{u - \tilde{x}_i}{h_n}\right).$$

Supposons maintenant que les $\mathbf{z}$ soient inconnus, mais que la vraie valeur de φ soit connue. La difficulté est alors de retrouver un échantillon distribué suivant f étant donné un échantillon distribué suivant g_φ. La solution que nous proposons, et dont nous prouvons la validité dans le Lemme 1 de [A12], est la suivante. On simule la i-ème allocation d'après la loi à postériori $(k(j|x_i, \varphi), j = 1, \ldots, m)$, et l'on effectue un recentrage *ad hoc* comme suit :

S-1 : `Pour` $i = 1, \ldots, n$`, on simule` $Z(x_i, \varphi) \sim \mathcal{M}(1; k(j|x_i, \varphi), j = 1, \ldots, m)$.

S-2 : `On prend` $\tilde{x}_i = x_i - \mu_{Z(x_i,\varphi)}$`, où` $Z(x, \varphi) \in \{1, \ldots, m\}$ `et` $\mu_{Z(x,\varphi)} = \mu_j$ `quand` $Z(x, \varphi) = j$.

L'idée forte derrière ce dernier point est que si φ^t est proche de la vraie valeur du paramètre φ, alors l'échantillon $\tilde{\mathbf{X}}^{t+1}$ sera presque distribué suivant f, ce qui rendra l'estimateur à noyau à l'étape $(t+1)$ fiable (proche de f), et permettra de se rapprocher du comportement d'un algorithme EM classique avec f connue.

3.2.2 Procédure semi-paramétrique de type EM

L'analyse du paragraphe précédent nous pousse donc à considérer l'algorithme EM semi-paramétrique (EM-SP) suivant, dont le passage de $\varphi^t \to \varphi^{t+1}$ s'opère en effectuant :

1. Étape E : `calculer` $k(j|x_i,\varphi^t)$, $i=1,\ldots,n$, $j=1,\ldots,m$ `utilisant` (3.21).
2. Étape S :
   - `pour` $i=1,\ldots,n$, `générer` $Z^{t+1}(x_i,\varphi^t) \sim \mathcal{M}(1;k(j|x_i,\varphi^t),j=1,\ldots,m)$;
   - `prendre` $\tilde{x}_i^{t+1} = x_i - \mu^t_{Z^{t+1}(x_i,\varphi^t)}$.
3. Étape non-paramétrique (réactualisation du paramètre fonctionnel)

   - `estimateur à noyaux de la densité`

   $$\hat{f}_{\tilde{\mathbf{x}}^{t+1}}(u) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{u-\tilde{x}_i^{t+1}}{h_n}\right);$$

   - `symétrisation`

   $$f^{t+1}(u) = \frac{\hat{f}_{\tilde{\mathbf{x}}^{t+1}}(u) + \hat{f}_{\tilde{\mathbf{x}}^{t+1}}(-u)}{2}.$$

4. Étape M : (stratégie EM paramétrique pour réactualiser le paramètre euclidien)

$$\begin{aligned}
\lambda_j^{t+1} &= \frac{1}{n}\sum_{i=1}^{n} k(j|x_i,\varphi^t);\\
\mu_j^{t+1} &= \frac{\sum_{i=1}^{n} k(j|x_i,\varphi^t)\, x_i}{\sum_{i=1}^{n} k(j|x_i,\varphi^t)}, \quad j=1,\ldots,m.
\end{aligned}$$

Notons que l'étape M peut être changée dans l'esprit de l'algorithme SEM (Stochastique EM), i.e. en utilisant le fait que les données complètes ont été préalablement simulées pour calculer l'EMV du paramètre euclidien, soit :

$$\begin{aligned}
\lambda_j^{t+1} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i^{t+1}(x_i,\varphi^t)=j\}};\\
\mu_j^{t+1} &= \frac{\sum_{i=1}^{n} x_i\, \mathbb{I}_{\{Z_i^{t+1}(x_i,\varphi^t)=j\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{Z_i^{t+1}(x_i,\varphi^t)=j\}}}, \quad j=1,\ldots,m.
\end{aligned}$$

L'avantage de cette deuxième version est qu'elle génère une chaîne de Markov plus simple à étudier que celle engendrée par l'étape M de l'algorithme précédent. Ayant constaté que les estimateurs à noyaux des données recentrées étaient assez proches de la vraie densité (paramétrique) mélangée, il peut être intéressant, dans un premier temps, d'étudier la robustesse de l'algorithme EM face à la famille paramétrique de lois considérée. ll demeure néanmoins que l'étude de cet algorithme est d'une incroyable complexité, et que nous devrons sans doute nous doter de moyen techniques supplémentaires (voir Robin *et al.*, 2007) pour le contrôler et montrer, un jour peut être, sa convergence (en un sens qui reste à définir).

Chapitre 4

Méthodes de Monte Carlo

Dans ce chapitre nous présentons quelques contributions au domaine des méthodes de Monte Carlo par chaîne de Markov (MCMC) adaptatives. Le but des MCMC est de reconstruire des lois d'intérêt qui peuvent s'avérer complexes, telles que les lois multimodales en grande dimension avec des modes très éloignés. Dans ce type de situations, les méthodes classiques comme Hastings-Metropolis ou l'échantillonneur de Gibbs peuvent mettre un temps très long avant de découvrir leurs différentes régions modales. Le principe de base des méthodes adaptatives consiste à utiliser l'information sur la loi cible, obtenue lors des itérations passées de la chaîne, afin d'explorer le plus efficacement possible son support dans le but d'améliorer la vitesse de convergence. Les travaux décrits dans ce chapitre sont en collaboration avec Didier Chauveau.

4.1 Algorithme de Hastings-Metropolis avec apprentissage séquentiel

Nous présentons tout d'abord le travail effectué en [A6] qui concerne une des premières avancées en matière de MCMC adaptatives. L'algorithme de Hastings-Metropolis (HM) permet de simuler une loi π lorsque l'on ne connaît que la forme analytique de sa densité à une constante multiplicative près. Ce problème survient par exemple lorsque l'on doit reconstruire la loi à postériori d'un modèle bayésien. Le principe de l'algorithme de HM est le suivant : étant donné la position courante x de l'algorithme, on propose une valeur *candidate* y au moyen d'une loi *instrumentale* pouvant dépendre de x notée $q(y|x)$ facilement simulable, puis on applique un mécanisme d'acceptation-rejet au candidat y pour constituer la valeur courante de l'algorithme à l'étape suivante. Le pas de l'algorithme conduisant de x_n à x_{n+1} est décrit ci-dessous :

1. `simuler` $y \sim q(\,\cdot\,|x_n)$;
2. `calculer` $\alpha(y, x_n) = \min\left\{1, \dfrac{f(y)q(x_n|y)}{f(x^{(n)})q(y|x_n)}\right\}$;
3. `prendre` $x_{n+1} = \begin{cases} y & \texttt{avec probabilité } \alpha(y, x_n), \\ x_n & \texttt{avec probabilité } 1 - \alpha(y, x_n). \end{cases}$

Il est courant d'utiliser l'algorithme de HM avec une règle de proposition de type marche aléatoire, i.e. $y_n = x_n + \varepsilon_{n+1}$, où ε_{n+1} est une perturbation aléatoire de densité g indépendante de la position courante x, et $q(y|x) = g(y - x)$. Les implémentations les plus courantes utilisent pour g une loi symétrique telle que la gaussienne $\mathcal{N}(0, \sigma^2)$ en dimension 1, car la symétrie réduit le taux d'acceptation à $\alpha(x, y) = \min\{1, f(y)/f(x)\}$. La grande difficulté dans ce type d'approche est le choix du paramètre d'échelle (voir Gilks *et al.*, 1996). Typiquement, si la densité q génère des sauts trop petits, l'algorithme aura un taux d'acceptation fort, mais pourra rester piégé dans le bassin d'attraction d'un mode sans pouvoir visiter les autres modes, alors que si la densité q génère des sauts trop grands, l'algorithme de HM aura tendance à visiter les queues de distribution de π, où la masse est très faible, ce qui entraînera un taux d'acceptation trop petit (algorithme figé). Certains auteurs ont proposé de calibrer la variance de la marche aléatoire, de manière à obtenir un taux d'acceptation ni trop grand ni trop petit (la valeur 0.23 a même été proposée). Or il est aisé de voir que cette recommandation n'est pas adaptée à toutes les situations. Prenons par exemple une densité à trois modes avec un mode très éloigné des deux autres. Il se peut que, dans une telle situation, nous puissions ajuster une variance permettant de visiter aisément les zones situées sous les modes les plus rapprochés (avec un α ni trop grand ni trop petit durant la période ou l'algorithme ne visite que cette zone), mais que cet ajustement nous empêche de visiter le mode le plus éloigné, créant ainsi un biais important dans les estimations futures. Le problème du calibrage de la variance nécessite en réalité un bonne connaissance du paysage induit par la loi π que l'on cherche à reconstituer. L'un de nos objectifs est de proposer une réponse à ce type de problème par une méthode "aveugle". D'autres auteurs ont cherché des procédures adaptatives permettant de mieux explorer le support de la loi cible comme Haario *et al.* (2001), Atchadé *et al.* (2005) ; citons aussi les travaux de Andrieu et Moulines (2006) sur l'ergodicité de certains algorithmes MCMC adaptatifs.

Une autre version de l'algorithme de HM consiste à considérer une loi instrumentale indépendante de la position courante x, i.e. $q(y|x) = q(y)$. Cette deuxième approche autorise en réalité des déplacements beaucoup plus libres que ceux de la version marche aléatoire. Cependant la performance de cet algorithme est liée à la qualité de la loi instrumentale q comparativement à π. Dans ce cadre, Mengersen et Tweedie (1996) montrent l'ergodicité géométrique uniforme de l'algorithme de HM sous des conditions éclairant ce dernier point. Si $q(\,\cdot\,) > 0$ sur le support Ω de π et qu'il existe $a \in]0, 1[$ tel que $q(x) > af(x)$ pour tout $x \in \Omega$, alors pour tout $n \geq 1$,

$$\|\mathbb{P}^n(x, \cdot\,) - \pi\|_{VT} \leq (1 - a)^n,$$

où $\mathbb{P}^n(x, \cdot\,)$ désigne la loi de l'algorithme de HM à l'instant n partant de x et $\|\cdot\|_{VT}$ la norme en variation totale. Ce résultat montre en effet, que plus q est proche de f (densité de π), c'est à dire "a proche de 1", plus la convergence est rapide. Ce résultat a été amélioré par Holden (1998), qui montre que, sous la condition $q \geq af$, la densité p^n de l'algorithme de HM à l'étape n vérifie :

$$\sup_{x \in \Omega} \left| \frac{p^n(x) - f(x)}{f(x)} \right| \leq D(1 - a)^n, \qquad \text{où} \quad D = \sup_{x \in \Omega} \left| \frac{p^0(x) - f(x)}{f(x)} \right|. \tag{4.1}$$

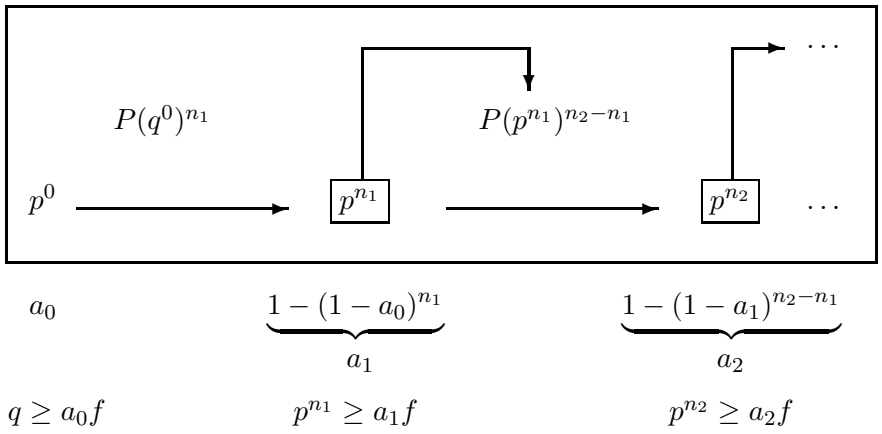


FIG. 4.1 – Schéma idéal d'apprentissage en ligne aux instants $n_1, n_2, \ldots$, où $P(q)^n$ est le n-itéré du noyau de HM de loi instrumentale q. *2ème ligne* : constantes de minoration associées à l'emploi de la densité de l'algorithme comme loi instrumentale. *3ème ligne* : conditions de minorations associées.

Méthode adaptative. Notre but est d'améliorer la convergence d'un algorithme de HM indépendant ayant une loi instrumentale q^0 vérifiant $q^0 \geq a_0 f$, et assurant donc la convergence géométrique (4.1) déterminée par a_0. Un des leviers pour améliorer cette vitesse est de chercher une loi instrumentale q admettant une constante a se rapprochant davantage de 1. A cette fin, il apparaît naturel d'utiliser la connaissance de f dont on dispose au travers de la densité p^n, puisque celle-ci, pour n assez grand, se trouve arbitrairement uniformément proche de f. Imaginons par exemple que l'on puisse remplacer au cours du temps la densité q^0 par les densités successives p^n ; nous obtiendrions alors le schéma idéal décrit dans la Figure 4.1, avec des améliorations extrêmement rapides des constantes de minorations $a_1,\ a_2, \ldots$, associées à la vitesse donnée dans (4.1). Les densités p^n étant inconnues, on peut se doter de m chaînes i.i.d. lancées suivant l'algorithme de HM indépendant de loi initiale et instrumentale q^0, et estimer non-paramétriquement p^n à partir des états des m chaînes à l'étape n. Malheureusement, dès le premier instant d'apprentissage n_1, la construction de l'estimateur de p^{n_1} entraîne un couplage des chaînes qui perdent alors leur indépendance et leur caractère Markovien, rendant difficile l'étude théorique de ces processus. Nous contournons cette difficulté en faisant en sorte de ne travailler que sur des chaînes i.i.d. grâce à un artifice de simulation.

Algorithme de HM adaptatif et résultat d'optimalité asymptotique. Dans [A6], pour estimer la densité de l'algorithme, nous avons choisi de considérer, en raison de sa simplicité, la méthode de l'histogramme. Afin de lever les difficultés techniques discutées dans le paragraphe précédent, nous considérons un schéma théorique consistant à éliminer, aux instants d'apprentissage, les chaînes ayant permis la mise à jour de la dernière loi instrumentale (injectée ensuite dans les chaînes restantes). Ce procédé permet de conserver l'indépendance et le caractère Markovien (non-homogène) des chaînes restantes, et d'utiliser des résultats classiques sur l'histogramme tels que les inégalités exponentielles de déviation sur les classes dans le cadre i.i.d. Nous supposons dans notre travail que la densité cible f est

C-Lipschitzienne à support compact $\Omega \subset \mathbb{R}^s$ et minorée par une constante $\alpha > 0$. Cette condition est restrictive mais nécessaire pour pouvoir utiliser les résultats de convergence presque sûre uniformes sur Ω. Dans la pratique, il n'est pas nécessaire que ces conditions soient strictement vérifiées, puisque cette méthode sert essentiellement à reconstruire une loi instrumentale localisant bien les zones chargées par π sur un compact aussi grand que nécessaire. Pour l'étude asymptotique de notre algorithme, nous supposons disposer d'une infinité de copies i.i.d. d'un processus de HM inhomogène défini pour une suite de lois instrumentales q^n. Afin d'alléger les notations, on suppose que l'apprentissage se fait à tous les instants (ce qui n'est pas le cas en pratique). L'apprentissage à l'instant n utilise $m(n)$ copies empruntées à cet ensemble infini, qui sont ensuite éliminées. La densité p^n est estimée par l'histogramme $H_{m(n)}$ construit sur les réalisations de ces $m(n)$ chaînes (voir Bosq et Lecoutre, 1987, pour la définition de l'histogramme et ses propriétés). Afin d'assurer la consistance des estimateurs, nous exigerons que $m(n)$ tende vers l'infini avec n à un régime que nous préciserons. La loi instrumentale q^n est soit $H_{m(n)}$, soit une légère modification de $H_{m(n)}$ (lorsque surviennent certaines classes vides) de manière à assurer la condition de minoration $q^n \geq a_n f$, pour tout $n \geq 1$, avec $a_n \in]0,1[$. Dans ce cadre, nous montrons tout d'abord une convergence du type (4.1) pour les lois marginales des chaînes encore en vie à l'étape n. Nous montrons ensuite, sous des conditions techniques rappelées ci-dessous, une inégalité exponentielle à distance finie (4.3) pour l'histogramme $H_{m(n)}$ basé sur $m(n)$ réalisations i.i.d. suivant p^n. Ce résultat découle d'une inégalité exponentielle pour la loi multinomiale (Bosq et Lecoutre 1987), et exige dans notre situation que la largeur des classes $h_{m(n)}$ ne tende pas trop vite vers 0 (condition (4.2))

Proposition 1 *Soit $H_m := H_{m(n)}$ l'histogramme de p^n, h_m la largeur de ses classes, et $\varepsilon > 0$. Posons*

$$\delta_{m,n} = 2A\left(1 - \frac{1}{Amh_m^s}\right)^n \sup_{x\in\Omega}\left|\frac{p^0(x) - f(x)}{f(x)}\right| + \sqrt{s}h_m C.$$

Si $h_m \to 0$, $mh_m^s \to +\infty$ lorsque $n \to +\infty$, $mh_m^s = o(n)$ et

$$mh_m^{3s} \geq \left(20/(\varepsilon - \delta_{m,n})^2\right) \quad \text{for } m > m_0, n > n_0, \tag{4.2}$$

où n_0 et m_0 vérifient $(\varepsilon - \delta_{m_0,n_0}) > 0$ et $(\varepsilon - \delta_{m_0,n_0})h_{m_0}^s \leq 1$, nous avons alors, pour $n > n_0$ et $m > m_0$:

$$\mathbb{P}\left(\sup_{x\in\Omega}|H_m(x) - p^n(x)| > \varepsilon\right) \leq 3\exp\left(-mh_m^{2s}(\varepsilon - \delta_{m,n})^2/25\right). \tag{4.3}$$

Nous montrons enfin que l'algorithme avec apprentissage que nous proposons converge plus rapidement vers f que tout algorithme de HM homogène usuel utilisant une loi arbitraire q^0 satisfaisant $q^0 \geq a_0 f$. Le résultat donné ci-dessous exprime le fait que l'algorithme n'utilisera pas infiniment souvent une loi instrumentale "moins bonne" que q^0, c'est à dire vérifiant une condition de minoration du type $q^n \geq af$ avec $a_n < a_0$. Il faut pour cela calibrer le régime $m(n)$ afin de montrer, par un lemme de

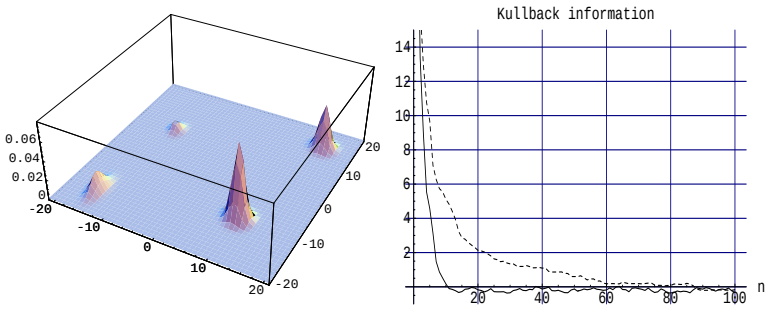


FIG. 4.2 – *Gauche : Vraie densité. Droite : estimation de la distance de Kullback entre p^n et f pour la chaîne adaptative (trait plein) et la chaîne homogène (trait pointillé).*

Borel-Cantelli utilisant (4.3), que les évènements "indésirables" ne peuvent survenir qu'un nombre fini de fois. Une façon plus rigoureuse d'exprimer ce résultat consiste à introduire

$$T(a_0) = \inf\{t \in \mathbb{N} \ : \ \forall n \geq t, \quad a_n > a_0\},$$

l'instant aléatoire au-delà duquel tout algorithme de HM indépendant, utilisant la loi instrumentale q^n pour $n > T(a_0)$, est plus rapide que l'algorithme initial (au sens où l'on sait démontrer une vitesse plus rapide).

Théorème 5 *Si $m(n)$ vérifie les conditions de la Proposition (1), et*

$$m(n)h^{2s}_{m(n)} \geq c\log(n), \tag{4.4}$$

où $c = c(\alpha)$ est une constante explicite, alors $\mathbb{P}(T(a_0) < \infty) = 1$.

Simulation bi-dimensionnelle. Afin d'illustrer le comportement de notre méthode dans un cadre multivarié, nous proposons de la confronter à une loi de mélange à quatre composantes gaussiennes bi-dimensionnelles. Les paramètres du modèles sont

$$\begin{aligned}
\text{poids} \quad &: \quad \alpha_1 = 0.5, \quad \alpha_2 = 0.3, \quad \alpha_3 = 0.15, \quad \alpha_4 = 0.05, \\
\text{moyennes} \quad &: \quad \mu_1 = \begin{pmatrix} +10 \\ -10 \end{pmatrix}, \mu_2 = \begin{pmatrix} 15 \\ 15 \end{pmatrix}, \mu_3 = \begin{pmatrix} -15 \\ -15 \end{pmatrix}, \mu_4 = \begin{pmatrix} -12 \\ +07 \end{pmatrix}, \\
\text{variances} \quad &: \quad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.5 & 0 \\ 0 & 3 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}.
\end{aligned}$$

La Figure 4.2 donne le graphe de la loi cible, ainsi que l'allure de la convergence, au sens de Kullback, des algorithmes de HM adaptatif et homogène. La méthode d'estimation pour ce deuxième graphique est détaillée dans le paragraphe 4.2.

Nous avons utilisé, pour notre algorithme adaptatif, le schéma d'apprentissage suivant : il a d'abord été initialisé avec 231 chaînes en parallèle, et a effectué 4 mutations aux instants $I = (1, 3, 5, 7)$, avec des ensembles de chaînes en parallèle de

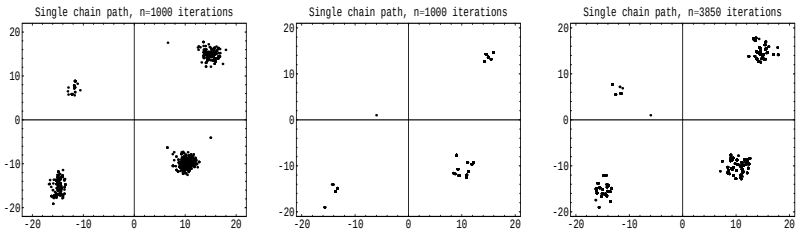


FIG. 4.3 – *Visites d'une seule chaîne . Gauche : pour $n = 1000$ étapes de la chaîne adaptative ; Milieu : pour $n = 1000$ étapes de la chaîne homogène ; Droite : pour $n = 3850$ étapes de la chaîne homogène.*

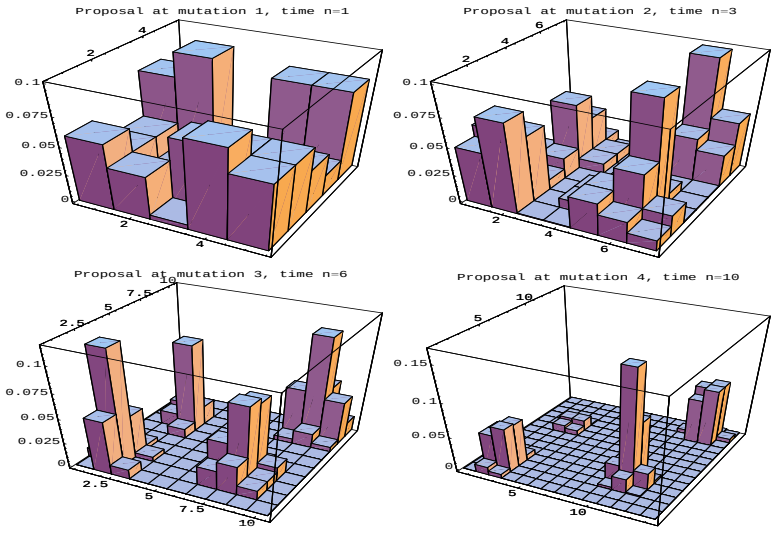


FIG. 4.4 – *Lois intrumentales successives à chaque instant d'apprentissage.*

taille $N(\cdot) = (40, 50, 60, 80)$. La chaîne finale qui utilise q_7 a éffectué $n = 2000$ sauts. Cet algorithme adaptatif a finalement effectué 3050 sauts, dont 1050 ont été utilisés pour construire les quatres densités instrumentales. Nous constatons aux travers des Figure 4.2 et 4.3 que la richesse d'exploration de la chaîne adaptative lui permet, à nombre égal de sauts, de faire beaucoup mieux que la chaîne homogène. La Figure 4.4 montre enfin avec quelle rapidité l'apprentissage de la loi cible s'effectue.

Algorithme de Hastings-Metropolis en intéraction. Afin d'éviter le gaspillage de chaînes qui, une fois utilisées pour l'apprentissage de la loi cible, étaient mise de coté, nous avons proposé dans [A5] un système de couplage d'algorithmes en parallèle (non indépendants) pour lequel nous pensions que la convergence des marginales vers la loi cible était plus rapide et sans biais. Or un contre-exemple dans un cas simple nous a montré que cette intuition était erronée. Depuis, ce problème a été reconsidéré avec succès par Del Moral et Doucet (2003) dans une approche à rebours. Notons toutefois que la procédure proposée dans [A5] donne de très bons résultats en pratique, et qu'une fois le système de apprentissage arrêté, les chaînes résultantes

convergent exponentiellement vite vers la loi cible.

4.2 Comparaison de MCMC via l'entropie

Les méthodes de Monte Carlo par chaînes de Markov (MCMC) génèrent une chaîne de Markov $(X^{(n)})_{n\geq 0}$ dont la loi stationnaire admet une densité de probabilité f sur un espace d'état $\Omega \subseteq \mathbb{R}^s$. Dans les situations où l'on ne sait pas simuler suivant f, ou lorsqu'il n'existe pas de calcul exact pour des intégrales de la forme $\mathbb{E}_f[h] = \int h(x)f(x)dx$, les méthodes MCMC sont très utiles, puisque pour T assez grand $X^{(T)}$ est presque distribuée suivant f, et $\mathbb{E}_f[h]$ peut être approchée par la moyenne ergodique de la chaîne. Les méthodes les plus couramment utilisées sont l'algorithme de Hastings-Metropolis (Hastings, 1970) et l'échantillonneur de Gibbs (Geman et Geman, 1984). Nous avons évoqué dans le paragraphe 4.1 les difficultés inhérentes au choix d'une bonne stratégie de simulation et les efforts menés dans ce domaine pour proposer des méthodes adaptatives efficaces. Nous proposons dans ce travail une approche méthodologique permettant de hiérarchiser diverses stratégies de simulation au moyen d'un critère basé sur l'entropie. Supposons que l'on souhaite comparer deux algorithmes MCMC partant d'une même loi initiale $p_1^0 = p_2^0$ dont les densités à l'itération n sont p_1^n et p_2^n. Un indicateur naturel de la qualité de l'algorithme est l'évolution dans le temps de la divergence de Kullback-Leibler entre p_i^n, $i = 1, 2$ et f donnée par

$$\mathcal{K}(p_i^n, f) = \int \log\left(\frac{p_i^n(x)}{f(x)}\right) p_i^n(x)dx = \mathcal{H}(p_i^n) - \mathbb{E}_{p_i^n}[\log f],$$

où pour toute densité p sur Ω, $\mathcal{H}(p) = \int \log(p(x))p(x)dx$ désigne l'entropie relative de p. Lorsque f est analytiquement connue, un estimateur fortement convergent de $\mathbb{E}_{p_i^n}[\log f]$ s'obtient au moyen d'une intégration de type Monte Carlo utilisant par exemple N chaînes i.i.d. de l'algorithme à l'étape n. Malheureusement comme f est en général la densité à postériori d'un modèle bayésien, on ne la connait en réalité qu'à une constante multiplicative près, i.e $f(\cdot) = C\varphi(\cdot)$ où la constante de normalisation C n'est pas accessible. Cependant, si l'on souhaite comparer les comportements en entropie de deux stratégies, la connaissance de C n'est pas nécessaire pour estimer la différence de leur divergence par rapport à f, puisque

$$\begin{aligned} D(p_1^n, p_2^n, f) &= \mathcal{K}(p_1^n, f) - \mathcal{K}(p_2^n, f) \\ &= \mathcal{H}(p_1^n) - \mathcal{H}(p_2^n) + \mathbb{E}_{p_1^n}[\log\varphi] - \mathbb{E}_{p_2^n}[\log\varphi]. \end{aligned} \tag{4.5}$$

La distance de Kullback est la seule divergence assurant cette propriété, ce qui en fait un outil particulièrement appréciable pour l'étude des algorithmes MCMC. Notons que nous aurions pu chercher à estimer d'autres distances comme la distance L^1 ou L^2, mais l'estimation de ces distances requiert des conditions techniques aussi difficiles à vérifier que celles exigées pour l'estimation de l'entropie. Nous aurions du de plus nous confronter au problème de l'estimation de C par d'autres techniques, ce qui totalement inenvisageable dans un contexte MCMC non-trivial. D'autres auteurs, comme Douc *et al.* (2006), se sont servi de la distance de Kullback pour ajuster

des stratégies de simulation. Pour estimer les entropies relatives $\mathcal{H}(p_i^n, f)$, $i = 1, 2$, impliquées dans (4.5), nous utilisons une méthode proposée par Györfi et Van Der Meulen (1989), qui décompose un échantillon $\mathbf{X}_N$ i.i.d. de taille N, distribué suivant une densité générique p, en deux sous-échantillons $\mathbf{Y}_N = (Y_i)$ et $\mathbf{Z}_N = (Z_i)$ définis par :

$$\begin{aligned} Y_i &= X_{2i}, \quad \text{pour } i = 1, \ldots, [N/2], \\ Z_i &= X_{2i-1}, \quad \text{pour } i = 1, \ldots, [(N+1)/2], \end{aligned}$$

où $[\cdot]$ désigne la partie entière. Soit $\hat{p}_N(x) = \hat{p}_N(x, \mathbf{Z}_N)$ l'estimateur à noyau de la densité p défini par

$$\hat{p}_N(x) = \frac{1}{h_N^s (N+1)/2} \sum_{i=1}^{[(N+1)/2]} K\left(\frac{x - Z_i}{h_N}\right), \qquad x \in \mathbb{R}^s,$$

où le noyau K est une densité de probabilité sur $\mathbb{R}^s$, $h_N > 0$ avec $h_N \to 0$ et $N h_N^s \to +\infty$ lorsque $N \to +\infty$. L'estimateur de l'entropie introduit par Györfi et Van Der Meulen (1989) s'écrit alors sous la forme

$$\mathcal{H}_N(p) = \frac{1}{[N/2]} \sum_{i=1}^{[N/2]} \log \hat{p}_N(Y_i) \mathbb{I}_{p_N(Y_i) \geq a_N}, \tag{4.6}$$

où $0 < a_N < 1$ avec $a_N \to 0$ lorsque $N \to +\infty$. L'une des grandes difficultés de notre travail a été de montrer que les hypothèses de régularité (caractère Lipschitz, décroissance des queues, etc.) exigées pour la consistance de l'estimateur (4.6) étaient vérifiées sur les densités successives de certains algorithmes MCMC. Nous donnons en particulier deux séries d'hypothèses (voir Propositions 2 et 3 et le Théorème 1 dans [B1]) assurant la validité de notre approche pour l'algorithme de Hastings-Metropolis indépendant vu dans le paragraphe 4.1. Une illustration de cette méthode est donnée dans la Figure 4.2 (dans un cadre où la constante C de normalisation de f est connue), montrant clairement la sensibilité de l'entropie aux étapes d'apprentissage.

4.3 Échantillonnage d'importance : f-correction de la loi instrumentale

Dans un travail en cours [B4], en collaboration avec Didier Chauveau et avec l'aide de Cécilia Lavanant, étudiante de Master, nous proposons une méthode de réduction de la variance pour la méthode dite d'*échantillonnage d'importance* (traduction de *importance sampling*). Le principe de cette méthode est le suivant : supposons que, comme dans les paragraphes 4.1 et 4.2, nous souhaitions approcher des intégrales du type $\mathbb{E}_f[h]$, où f est une densité sur $\Omega \subset \mathbb{R}^s$ de la forme $f(\cdot) = C\varphi(\cdot)$ où la constante $C = \int \varphi(x)dx$ de normalisation est inconnue. L'idée consiste à simuler un échantillon i.i.d. $X_1, \ldots, X_n$ suivant une loi instrumentale de densité g et d'estimer $E_f[h]$ au moyen de la loi forte des grands nombres qui, sous des conditions

ad hoc de moments, donne :

$$T_g = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{h\varphi}{g}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} \frac{\varphi}{g}(X_i)} \xrightarrow{p.s.} \int \frac{hf}{g}(x)g(x)dx = \mathbb{E}_f[h], \quad \text{lorsque } n \to +\infty. \tag{4.7}$$

Comme pour l'algorithme de HM la qualité de l'estimation décrite précédemment (voir Robert et Casella, 2004) dépend fortement de la ressemblance entre la densité instrumentale g et la densité cible f, l'idéal étant de pouvoir simuler directement sous f. Nous proposons dans ce travail un moyen d'améliorer la qualité de la densité instrumentale g en rectifiant son allure aux endroits où g est loin de f. Une solution est par exemple de créer des modes aux endroits où f en possède mais pas g, et à contrario d'abaisser g aux endroits où f est faible et g élevée. Douc *et al.* (2006) s'intéressent à ce problème et proposent d'ajuster des lois de mélange (avec un nombre de composantes fixé) afin de mimer au mieux le relief de la densité cible f. Nous proposons pour notre part un estimateur consistant de f basé sur un estimateur à noyau repondéré utilisant l'échantillon instrumental. Cet estimateur de f est défini de la manière suivante :

$$\hat{f}_n(x) = \frac{\frac{1}{nh_n^s}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\frac{\varphi}{g}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} \frac{\varphi}{g}(X_i)}, \qquad x \in \mathbb{R}^s, \tag{4.8}$$

où le noyau K est une densité de probabilité sur $\mathbb{R}^s$, $h_N > 0$ avec $h_N \to 0$ et $Nh_N^s \to +\infty$ lorsque $n \to +\infty$. Nous conjecturons que sous certaines conditions la famille de fonctions

$$\mathcal{F} = \left\{ K\left(\frac{x - \cdot}{h}\right)\frac{\varphi}{g}(\cdot),\ x \in \mathbb{R}^s,\ h > 0 \right\}$$

forme une classe de Vapnik-Cervonĕnkis bornée de fonctions mesurables. Si nous parvenons à montrer un résultat de ce type, avec des conditions raisonnables sur φ, g et K, nous aurons alors, en utilisant les travaux de Giné et Guillou (2002), une vitesse de convergence presque sûre de $||\hat{f}_n - f||_\infty$ vers 0 lorsque $n \to +\infty$. Notons que d'un point de vue méthodologique la densité (4.8) est facilement simulable puisqu'elle correspond à un mélange à n composantes $1/h_n^s K\left(\frac{x-X_i}{h_n}\right)$, où $i = 1, \ldots, n$, dont les poids du mélange sont $\frac{\varphi}{g}(X_i)/\sum_{i=1}^{n}\frac{\varphi}{g}(X_i)$. On estime alors $\mathbb{E}_f[h]$ au moyen de l'estimateur $T_{\hat{f}_n}$ défini en (4.7). L'étape de re-échantillonnage au sens de cet estimateur à noyau repondéré, peut s'apparenter à une version lissée de la méthode de re-échantillonnage par poids d'importance (voir, *e.g.*, Cappé *et al.*, 2005) couramment utilisée en filtrage particulaire (Del Moral, 2004) ou dans les approches bayésiennes (Robert et Casella, 2004). Un projet intéressant concerne la normalité asymptotique de $T_{\hat{f}_n}$, lorsque le nombre de variables instrumentales tirées

tend vers l'infini. Il est à noter, comme toujours dans ce type de problème, que le contrôle des queues de distribution de $\hat{f}_n$ (ou d'une version modifiée) jouera une rôle prépondérant.

4.4 Simulation de la convergence en entropie de chaînes de Markov

Dans [A11] nous proposons, en collaboration avec Didier Chauveau, de contrôler le comportement en entropie de certaines chaînes de Markov (stabilité, convergence vers la loi stationnaire, vitesse dans le TCL) au moyen d'un outil statistique utilisant des réalisations i.i.d. de ces chaînes. On suppose simplement que le noyau de transition de la chaîne d'intérêt admet une densité par rapport à une mesure dominante, que l'on connait analytiquement cette densité, et que l'on sait la simuler. Nous introduisons pour cela un estimateur de l'entropie *relative* $\mathcal{H}(p^t) = \mathbb{E}_{p^t}(\log p^t)$ associée à la densité marginale p^t de la chaîne de Markov d'intérêt à l'instant $t \geq 1$. Cette estimation sera ensuite comparée à une entropie *externe* $\mathcal{H}(p_1^t, p^t) = \mathbb{E}_{p_1^t}(\log p^t)$, où les p_1^t sont les densités successives d'une autre chaîne de Markov bien choisie.

Estimation par double Monte Carlo. Soit $X = (X^t)_{t\geq 0}$ une chaîne de Markov inhomogène à temps discret, à valeurs dans une espace d'état mesurable $(E, \mathcal{E})$. On suppose que pour tout $t \geq 0$, la densité de transition q^t par rapport à une mesure de référence ν σ-finie est analytiquement connue. On sait que les densités marginales de la chaîne de Markov X sont données par sa densité marginale p^0 et la formule de récurrence :

$$p^{t+1}(y) = \int p^t(x) q^t(x, y) \nu(dx), \qquad t \geq 0.$$

L'idée principale de notre travail consiste à remarquer que si nous disposions d'un échantillon

$$\mathbf{X}^t = (X_1^t, X_2^t, \ldots, X_N^t), \qquad \text{i.i.d.} \sim p^t,$$

nous serions en mesure d'estimer ponctuellement p^{t+1} au moyen de la loi forte des grands nombres

$$\frac{1}{N}\sum_{k=1}^{N} q^t(X_k^t, y) \xrightarrow{p.s.} \int q^t(x, y) p^t(x) \nu(dx) = p^{t+1}(y), \quad \text{lorsque } n \to +\infty. \tag{4.9}$$

Ainsi nous pouvons attendre qu'en intégrant le log du membre de droite de (4.9) au sens de Monte Carlo, au moyen d'un deuxième échantillon i.i.d.

$$\tilde{\mathbf{X}}^{t+1} = (\tilde{X}_1^{t+1}, \tilde{X}_2^{t+1}, \ldots, \tilde{X}_N^{t+1}), \qquad \text{i.i.d.} \sim p^{t+1},$$

indépendant de $\mathbf{X}^t$, qu'il y ait convergence vers $\mathcal{H}(p^{t+1})$. Il est donc naturel d'introduire

$$\hat{\mathcal{H}}(p^{t+1}) = \frac{1}{N}\sum_{\ell=1}^{N} \log\left(\sum_{k=1}^{N} q^t(X_k^t, \tilde{X}_\ell^{t+1})\right). \tag{4.10}$$

On notera, dans le même esprit, que la quantité $\hat{\mathcal{H}}(p_1^{t+1}, p^{t+1})$ obtenue en remplacant $\tilde{\mathbf{X}}^t$ dans (4.10) par un échantillon i.i.d. $\mathbf{Y}^t = (Y_1^t, Y_2^t, \ldots, Y_N^t) \sim p_1^{t+1}$ devrait converger vers l'entropie externe $\mathcal{H}(p_1^t, p^t)$. Nous montrons, sous des conditions de moments, les résultats suivants :

Théorème 6 *Si pour tout $t \geq 0$, le noyau de transition re-normalisé*

$$r^t(x, y) = \frac{q^t(x, y)}{p^{t+1}(y)}$$

est non-dégénéré et vérifie :

$$\mathbb{E}_{p^t \otimes p_1^{t+1}} \left[|r^t(X, Y)|^{2+\gamma} \right] < \infty, \qquad \textit{pour } \gamma > 0,$$

et

$$\mathbb{E}_{p_1^{t+1}} \left[|\log p^{t+1}(X, Y)|^2 \right] < \infty,$$

alors

$$\hat{\mathcal{H}}(p_1^{t+1}, p^{t+1}) \xrightarrow{\mathbb{P}} \mathcal{H}(p_1^{t+1}, p^{t+1}), \qquad \textit{lorsque } n \to +\infty,$$

et

$$\sqrt{N} \left(\hat{\mathcal{H}}(p_1^{t+1}, p^{t+1}) - \mathcal{H}(p_1^{t+1}, p^{t+1}) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^t), \qquad \textit{lorsque } n \to +\infty,$$

où $\Sigma^t = var_{p_1^{t+1}}[\log p^{t+1}] + var_{p^t}[R(X)]$ et $R(x) = \mathbb{E}_{p_1^{t+1}}[(x, Y)]$.

La preuve de ce théorème utilise une décomposition inspirée de Del Moral et Guionnet (1999). Les techniques mise en oeuvre sont liées à la théorie des U-statistiques (voir Serfling, 1980) et à une inégalité de moyenne déviation due à Fuk et Nagaev (1971, 1976). Sous une condition de moment plus forte, et en utilisant de nouveau une technique inspirée de Del Moral et Guionnet (1999), nous montrons la convergence forte de nos estimateurs. Une différence par rapport à ces auteurs est que l'emploi d'une inégalité de moyenne déviation nous permet de relaxer leur condition de moment de 6 à $4 + \gamma$ pour tout $\gamma > 0$.

Théorème 7 *Sous les conditions du Théorème 6, si l'on remplace la condition (4.11) par :*

$$\mathbb{E}_{p^t \otimes p_1^{t+1}} \left[|r^t(X, Y)|^{4+\gamma} \right] < \infty, \qquad \textit{pour } \gamma > 0,$$

alors

$$\hat{\mathcal{H}}(p_1^{t+1}, p^{t+1}) \xrightarrow{p.s.} \mathcal{H}(p_1^{t+1}, p^{t+1}), \qquad \textit{lorsque } n \to +\infty.$$

Applications. Nous avons mis en oeuvre notre méthode pour étudier la stabilité d'un processus AR(1) avec bruit gaussien, ainsi que la vitesse de convergence dans le théorème central limite. Pour cette deuxième application, nous avons utilisé le fait que les variables du type

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \tag{4.11}$$

où $X_1, X_2, \ldots$ sont des variables aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$, de densité commune f par rapport à la mesure de Lebesgue, forment une chaîne de Markov. Il suffit d'étudier pour s'en convaincre, la structure autoregressive inhomogène du modèle

$$Y_{n+1} = \sqrt{\frac{n}{n+1}} Y_n + \frac{1}{\sqrt{n+1}} X_{n+1}, \quad n \geq 1,$$

le noyau de transition de cette chaîne de Markov s'écrivant alors très simplement sous la forme :

$$q^n(x, y) = (n+1)^{d/2} f(\sqrt{n+1}y - \sqrt{n}x).$$

En supposant que $\mathbb{E}[X] = 0$, et sous une condition de moment d'ordre 2, nous savons que

$$Y_n \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Sigma), \qquad \text{lorsque } \ n \to +\infty,$$

où Σ est la matrice de variance-covariance de X et $\mathcal{N}_d$ désigne la loi gaussienne en dimension d. Ce constat fait, il nous est apparu intéressant de représenter l'évolution dans le temps de la distance de Kullback entre p^n, la densité de Y_n, et ϕ_Σ, la densité de la loi $\mathcal{N}_d(0, \Sigma)$, soit

$$\mathcal{K}(p^n, \phi_\Sigma) = \mathbb{E}_{p^n}(\log p^n) - \mathbb{E}_{p^n}(\log \phi_\Sigma). \tag{4.12}$$

Nous constatons que nous pouvons estimer le premier terme du membre de droite de 4.12 en utilisant (4.10), et que le deuxième terme s'estime plus simplement encore en considérant $N^{-1} \sum_{\ell=1}^{N} \log(\phi_\Sigma(Y_{n,\ell}))$, où $Y_{n,1}, Y_{n,2}, \ldots$ sont i.i.d. suivant p^n. La

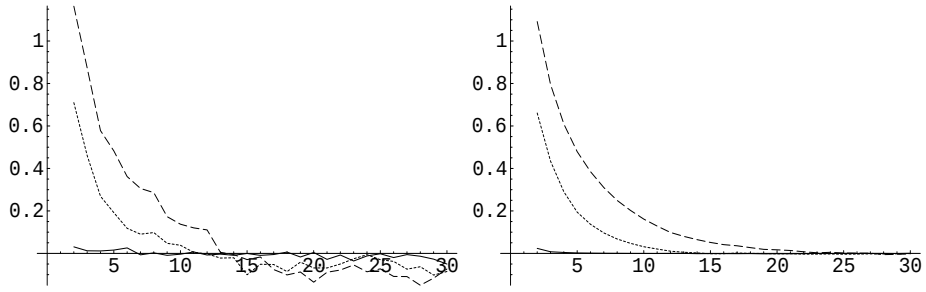


FIG. 4.5 – *Modèle $\mathcal{U}_{[-20;20]}$ (trait plein), M_1 (trait pointillé) et M_2 (trait discontinu) pour $N = 200$ (gauche), et $N = 5000$ (droite).*

Figure 4.5 représente le comportement de la distance de Kullback, estimée avec 200

puis 5000 chaînes, entre la loi de Y_n définie en (4.11) et la loi $\mathcal{N}(0,1)$ lorsque les X_i sont respectivement i.i.d. suivant une loi uniforme $\mathcal{U}_{[-20;20]}$ et les deux modèles de mélange M_1 et M_2 à composantes gaussiennes dont les paramètres sont donnés ci-dessous :

$$\begin{aligned} M_1 \quad &: \quad \alpha = \frac{1}{2}, \quad \mu_1 = -8, \quad \sigma_1^2 = 1, \quad \mu_2 = 8, \quad \sigma_2^2 = 4, \\ M_2 \quad &: \quad \alpha = \frac{1}{2}, \quad \mu_1 = -20, \quad \sigma_1^2 = 1, \quad \mu_2 = 20, \quad \sigma_2^2 = 16. \end{aligned}$$

Nous constatons clairement aux travers de la figure 4.5 les disparités importantes qui peuvent exister dans l'approximation gaussienne évoquée plus haut et situer aussi, pour la première fois, des ordres de grandeurs raisonnables sur la taille d'échantillon nécessaire en pratique pour sa prise en compte. En particulier, au travers du graphe associé au modèle M_2, lorsque la loi des X_i a des queues de distributions plus lourdes celles de la loi $\mathcal{N}(0,1)$, nous pouvons observer que la loi de Y_n met plus de temps à converger vers la loi normale centrée réduite.

Chapitre 5

Algorithmes stochastiques

Ce chapitre est consacré à l'étude de deux algorithmes stochastiques à pas décroissant que j'ai choisi d'étudier en marge de mes travaux en Statistique. Le travail (en cours de rédaction) sur l'algorithme du bandit à deux bras [B2] est le fruit d'une collaboration avec Pierre Tarrès.

5.1 Recuit simulé avec un estimateur séquentiel de l'énergie

L'objectif premier était de se doter d'un algorithme d'optimisation stochastique pouvant gérer l'arrivée de séquences d'information sur une fonction cible à minimiser. Un champs d'application naturel pour ce type d'approche est l'optimisation des fonctions de contraste ou de régression intervenant en Statistique.

De manière plus précise, on se donne $(H_n)_{n\geq 0}$ une suite d'estimateurs fonctionnels d'une certaine fonction H (appelée *potentiel* ou *energie*) définie sur un espace mesuré $(\Theta, \mathcal{B}, \lambda)$, telle que les hypothèses suivantes soient vérifiées :

(H1) Il existe une suite déterministe $(\delta_n)_{n\geq 0}$, décroissante vers 0, telle que

$$\|H_n - H\|_\infty = O(\delta_n) \quad p.s.$$

où $\|\cdot\|_\infty$ désigne la norme uniforme sur Θ.

(H2) La fonction H est de classe $\mathcal{C}^2$ sur Θ et admet un ensemble fini de minima globaux (noté H^*).

Nous rappelons qu'il est possible de définir pour la fonction H une mesure de Gibbs $G_{\beta,H}$ à la température $T = 1/\beta$, avec $\beta > 0$, de densité relativement à λ de la forme

$$\frac{1}{Z_\beta} e^{-\beta H}, \quad \text{où} \quad Z_\beta = \int_\Theta e^{-\beta H}.$$

Si H vérifie **(H2)**, cette mesure de Gibbs a la propriété de charger, lorsque la température T tend vers 0, uniquement les minima globaux de la fonction H, i.e.

$$G_{\beta,H} \longrightarrow \frac{1}{Z} \sum_{\theta \in H^*} k_\theta \delta_\theta, \quad \text{lorsque} \quad \beta \to +\infty, \tag{5.1}$$

où $k_\theta = (\det \nabla^2 H(\theta))^{-1/2}$ et $Z = \sum_{\theta \in H^*} k_\theta$. On suppose dans cette étude que la fonction H est inconnue, mais peut être approchée au sens de (**H1**). La méthode décrite ci-dessous a pour but d'optimiser (pas à pas) les estimateurs H_n dans l'espoir d'optimiser asymptotiquement le potentiel H. A cette fin, on définit pour tout $n \geq 0$ une mesure de Gibbs G_{β_n,H_n} adaptée à H_n, où β_n est un paramètre donné dépendant de n.

Nous montrons dans un premier théorème que, sous les hypothèses **(H1-2)**, les mesure de Gibbs G_{β_n,H_n} et $G_{\beta_n,H}$ ont le même comportement asymptotique, au sens de (5.1), si $\beta_n \delta_n \to 0$ lorsque $n \to +\infty$.

Ce premier point établi, on propose de simuler le comportement asymptotique du processus à valeurs mesures $(G_{\beta_n,H_n})_{n\geq 0}$ au moyen d'une chaîne de Markov inhomogène $(\theta_n)_{n\geq 0}$, dont le noyau de transition dit de *Metropolis* s'écrit pour chaque $n \geq 0$ sous la forme :

$$N_{\beta_n,H_n}(\theta, d\xi) = \mathbb{I}_{H_n(\xi)\leq H_n(\theta)} N(\theta, d\xi) + \mathbb{I}_{H_n(\xi)>H_n(\theta)} N(\theta, d\xi) + Q_n \delta_\theta(d\xi),$$

où N est un noyau markovien, homogène, vérifiant la condition de Harris sur Θ (voir Meyn et Tweedie, 2003), et la propriété de réversibilité par rapport à la mesure λ. Le terme Q_n sert à avoir $\int_\Theta N_{\beta_n,H_n}(\theta, d\xi) = 1$. La réversibilité permet de montrer que la mesure G_{β_n,H_n} est réversible pour N_{β_n,H_n}. On montre alors le théorème suivant :

Théorème 8 *Sous les hypothèses précèdentes, et pour δ_n de la forme $n^{-\alpha}$, avec $\alpha > 0$, il existe une constante γ_0 explicite telle que pour tout $\gamma < \gamma_0$ et $\beta_n = \gamma \log n$, on ait :*

$$\|\mathbb{P}_{\theta_n} - G_{\beta_n,H_n}\|_{VT} \longrightarrow 0 \quad p.s., \quad \textit{lorsque} \quad n \to +\infty,$$

où $\mathbb{P}_{\theta_n}$ désigne la mesure de probabilité de la chaîne de Markov θ_n à l'instant $n \geq 0$.

En conclusion, ce théorème exprime le fait que la loi de la chaîne de Markov $(\theta_n)_{n\geq 0}$ se concentrera, comme voulu, presque sûrement sur les minima globaux de la fonction H. Des simulations de cet algorithme, appliqué à l'optimisation de la log vraisemblance des données tronquées pour les CMC, se trouvent à la fin de ma thèse de doctorat.

5.2 Problème du bandit à deux bras dans un cadre ergodique

L'algorithme du bandit à deux bras est une procédure d'apprentissage statistique permettant de détecter entre deux sources de bénéfices, bras d'une machine à sous

par exemple, laquelle est la plus profitable. Supposons qu'à tout instant $n \in \mathbb{N}$, nous ayons le choix de jouer avec un bras A ou un bras B, et que ces bras puissent, indépendemment du passé, nous faire gagner respectivement un euro avec probabilité p_A (resp. p_B) et ne rien perdre avec probabilité $1-p_A$ (resp. $1-p_B$). Dans une telle situation, le bras A est dit plus profitable que la bras B si $p_A > p_B$. L'algorithme du bandit à deux bras a été introduit par Norman (1968) en psychologie mathématique, puis par Shapiro et Narendra (1969) en ingéniérie comme automate d'apprentissage linéaire (voir le "survey" et le livre de Narendra et Thathachar, 1974, 1989). Cette procédure permet de jouer aléatoirement avec les bras A et B, avec la propriété de sélectionner presque sûrement le bras le plus favorable lorsque que le nombre d'essai tend vers l'infini. Une application naturelle de ce type d'algorithme a été developpée par Niang (1999) dans le cadre de l'allocation de confiance pour les marchés financiers. Décrivons plus en détail le fonctionnement de cet algorithme. Pour tout $n \in \mathbb{N}$, on note X_n la probabilité de choisir le bras A à l'étape n, et on fixe $X_0 = x \in]0,1[$. Les probabilités X_n évoluent récursivement selon la procédure

$$\forall n \geq 0, \quad X_{n+1} = \begin{cases} X_n + \gamma_n(1 - X_n) & \text{si } U_{n+1} = A, \text{ et } \eta_{A,n+1} = 1 \ , \\ (1-\gamma_n)X_n & \text{si } U_{n+1} = B, \text{ et } \eta_{B,n+1} = 1 \ , \\ X_n & \text{sinon.} \end{cases} \tag{5.2}$$

où $\gamma_0 = c$ et $\gamma_n = c/(c+n)$, avec $c \in]0,+\infty[$, U_{n+1} est une variable aléatoire correspondant au label du bras choisi à l'instant $n+1$, et $\eta_{A,n+1}$ (resp. $\eta_{B,n+1}$) correspondent à la performance, à valeurs dans $\{0,1\}$ (1 pour succès, 0 pour échec), du bras A (resp. du bras B) à l'instant $n+1$. La formule (5.2) dit explicitement ceci : si on joue A et que l'on gagne, on renforce la probabilité de le jouer à l'étape suivante à hauteur d'un terme additif valant $\gamma_n(1-X_n)$, si on joue B et que l'on gagne on affaiblit X_n d'un facteur $0 < (1-\gamma_n) < 1$, et dans les autres cas on ne change rien.

En considérant $(I_n)_{n\geq 1}$ une suite i.i.d. de variables aléatoires uniformes sur $[0,1]$, le label du bras joué à l'instant $n+1$ est modélisé par $U_{n+1} = A\mathbb{1}_{I_{n+1}\leq X_n} + B\mathbb{1}_{I_{n+1}>X_n}$. Comme nous l'avons évoqué précédemment, on considère en général que les suites de variables aléatoires $(\eta_{A,n})_{n\geq 1}$ and $(\eta_{B,n})_{n\geq 1}$ sont i.i.d., mutuellement indépendantes, et respectivement distribuées suivant une loi de Bernoulli de paramètre p_A et p_B. Paradoxalement à la grande simplicité de l'algorithme (5.2), il a fallu attendre une trentaine d'années pour que le bon critère de convergence voit le jour. Ce résultat a été obtenu par Tarrès (2001), puis généralisé dans Lamberton *et al.* (2004). Notons que Benaïm et Ben Arous (2003) étudient aussi un algorithme de type bandit à deux bras dans un contexte de théorie des jeux. Récemment Lamberton et Pagès (2005a–b) ont établi des vitesses de convergence pour l'algorithme du bandit à deux bras ainsi qu'une version pénalisée de cet algorithme.

Dans [B2], nous proposons avec Pierre Tarrès, l'étude asymptotique de l'algorithme du bandit à deux bras lorsque l'hypothèse d'indépendance est levée sur les suites $(\eta_{A,n})_{n\geq 1}$ and $(\eta_{B,n})_{n\geq 1}$. Nous supposons désormais la condition d'ergodicité suivante :

(E) (Ergodicité). Les réalisations des bras A et B vérifient

$$\frac{1}{n}\sum_{k=1}^{n}\eta_{A,k}\underset{n\to\infty}{\longrightarrow}\theta_A \;\; p.s. \;\text{ et }\; \frac{1}{n}\sum_{k=1}^{n}\eta_{B,k}\underset{n\to\infty}{\longrightarrow}\theta_B \;\; p.s., \tag{5.3}$$

où $1 \geq \theta_A > \theta_B \geq 0$, ce qui signifie que A rapporte asymptotiquement plus que B. Notons qu'une telle condition permet de considérer des situations plus réalistes en pratique. Par exemple, dans le contexte de la sélection du meilleur trader citée dans Lamberton *et al.* (2004), les hypothèses d'indépendance et d'homogéneité non réalistes pour qualifier les "performances humaines" des traders, peuvent ainsi être allégées. Nous avons voulu mettre en évidence dans ce travail le fait que l'algorithme du bandit à deux bras se joue des structures stochastiques locales associées aux performances des bras A et B, mais valorise surtout leur aptitude au succès sur de très grandes plages de temps (du moment qu'un principe de supériorité ergodique du type (5.3) existe). Notre résultat principal repose sur une nouvelle décomposition de l'algorithme (voir Lemme 1) faisant apparaître la différence des moyennes ergodiques de chaque bras. Nous définissons pour tout $n \geq 1$, $\Gamma_n := \sum_{k=0}^{n}(\eta_{A,k+1} - \eta_{B,k+1} - (\theta_A - \theta_B))$, avec $\Gamma_0 = 0$, ainsi que la fonction $f(x) = x(1-x) \geq 0$, pour tout $x \in [0,1]$. Les σ-algèbres d'intérêt sont :

$$\mathcal{G} = \sigma(\eta_{A,1}, \ldots, \eta_{A,n}; \eta_{B,1}, \ldots, \eta_{B,n}; n \in \mathbb{N}).$$

et

$$\text{pour } \; n \geq 1, \;\; \mathcal{F}_n = \mathcal{G} \cup \sigma(I_1, \ldots, I_n).$$

Lemme 1 *Il existe une martingale $(M_n)_{n\geq 0}$, définie pour tout $n \geq 0$ par*

$$M_{n+1} = \sum_{k=0}^{n}\gamma_k\varepsilon_{k+1},$$

avec

$$\forall\; k \geq 0, \quad \mathbb{E}_x(\varepsilon_{k+1}|\mathcal{F}_k) = 0, \quad \mathbb{E}_x((\varepsilon_{k+1})^2|\mathcal{F}_k) \leq CX_k,$$

où $C = 3(2+c^2)$ et telle que, pour tout $n \geq 2$:

$$X_{n+1} = x + M_n + (\theta_A - \theta_B)\sum_{k=1}^{n}\gamma_k(1+W_k)f(X_k) + \Gamma_{n+1}\gamma_n f(X_n), \tag{5.4}$$

où

$$|W_k| \leq D\left|\frac{\Gamma_{k+1}}{c+k+1}\right| \to 0 \;\; \mathbb{P}_x - a.s., \; \textit{lorsque } k \to +\infty,$$

avec $D = \frac{1+2c}{\theta_A - \theta_B}$.

En utilisant une technique de preuve inspirée par celle de Tarrès (2001) dans le cadre i.i.d. nous établissons le théorème suivant :

Théorème 9 *Sous l'hypothèse d'ergodicité* **(E)**, *nous avons*
(i) Pour tout $c \in]0, +\infty[$, *le processus* $(X_n)_{n\geq 0}$ *défini en (5.2) converge* $\mathbb{P}_x$*-presque sûrement vers* 0 *ou* 1.
(ii) Pour tout $c \in]0, 1/\theta_B[$, *le processus* $(X_n)_{n\geq 0}$ *défini en (5.2) converge* $\mathbb{P}_x$*-presque sûrement vers* 1.

Le Lemme 1 permet de montrer, étant donnée la convergence asymptotique de la martingale $(M_n)_{n\geq 0}$, la convergence presque sûre de $\Gamma_k\gamma_k$ vers 0 due à **(E)**, et à l'encadrement $0 \leq X_n \leq 1$, que

$$\sum_{k=1}^{n} \gamma_k X_k(1 - X_k) < \infty \quad \mathbb{P}_x - p.s. \tag{5.5}$$

La condition (5.5) implique que X_n tend presque sûrement vers 0 ou 1 lorsque n tend vers l'infini. Nous montrons aussi un principe de "frein" pour la descente de l'algorithme vers 0.

Lemme 2 *Sous la condition d'ergodicité* **(E)**, *pour tout* $\varepsilon > 0$, *il existe presque sûrement un rang* $n_0(\varepsilon)$ *(aléatoire) tel que pour tout* $n \geq n_0(\varepsilon)$, $X_n \geq n^{-c\theta_B(1+\varepsilon)}$.

Sous la condition $0 < c < 1/\theta_B$, nous montrons en utilisant les Lemmes 1 et 2, et l'inégalité de Doob pour les martingales, que "moralement" si l'algorithme descend près de 0, la variance conditionnelle du bruit ne lui permettra pas de passer en dessous d'une fraction fixée de sa valeur courante. Ceci entraîne que $\liminf X_n > 0$ pour chaque trajectoire, et donc que l'algorithme converge obligatoirement vers 1. Nous étudions actuellement la possibilité d'accélérer la décroissance vers 0 de γ_n sous une condition faible de vitesse associée au comportement ergodique **(E)**.

Chapitre 6

Liste des travaux

[A1] Bakry, D., Milhaud, X. et Vandekerkhove, P. (1997). Statistics of Hidden Markov chains with finite state space. The nonstationary case. *C. R. Acad. Sci. Paris*, Série. I, 203–206.

[A2] Vandekerkhove, P. (1998). Simulated annealing with a sequential estimator of the energy. *C. R. Acad. Sci. Paris*, Série I, 1003–1006.

[A3] Chauveau, D. et Vandekerkhove, P. (1999). Un algorithme de Hastings-Metropolis avec apprentissage séquentiel. *C. R. Acad. Sci. Paris*, Série I, 173–176.

[A4] Giudici, P., Rydén, T. et Vandekerkhove, P. (2000). Likelihood-Ratio Tests for Hidden Markov Models. *Biometrics*, **56**, 742-747.

[A5] Chauveau, D. et Vandekerkhove, P. (2001). Algorithmes de Hastings Metropolis en intéraction. *C. R. Acad. Sci. Paris*, Série I, 881–884.

[A6] Chauveau, D. et Vandekerkhove, P. (2002). Improving convergence of the Hastings-Metropolis Algorithm with a learning proposal. *Scand. J. Statist*, **28**, 13–29.

[A7] Bordes, L. et Vandekerkhove, P. (2005). Statistical inference for Partially Hidden Markov Models. *Communications in Statistics*, **34**, 1081–1104.

[A8] Vandekerkhove, P. (2005). Consistent et asymptotically normal estimates for hidden Markov mixtures of Markov models. *Bernoulli*, **11**, 103–129.

[A9] Bordes, L., Mottelet, S. et Vandekerkhove, P. (2006). Semiparametric estimation of a two component mixture model. *Ann. Statist.*, **34**, 1204–1232.

[A10] Bordes, L., Delmas, C. et P. Vandekerkhove. (2006). Semiparametric estimation of a two-component mixture model where a component is known. *Scand. J. Statist.*, **33**, 733–752.

[A11] Chauveau, D. et Vandekerkhove, P. (2007). A Monte Carlo estimation of the entropy for Markov chains. *Methodology et Computing in Applied Probability*, **9**, 133–149.

[A12] Bordes, L., Chauveau, D. et Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, **51**, 5429–5443.

Thèse de doctorat

Identification de l'ordre des processus ARMA stables–Contribution à l'étude statistique des chaînes de Markov cachés. Thèse soutenue à l'université Montpellier II, le 19 septembre 1997.

Liste des articles soumis ou en préparation

[B1] Bordes, L. et Vandekerkhove, P. (2007). Semiparametric two-component mixture model with a known component : a class of asymptotically normal estimators. *Soumis à Annals of Statistics.*

[B2] Tarrès, P. et Vandekerkhove, P. (2007). On the ergodic two-armed bandit algorithm. *En préparation.*

[B3] Chauveau, D. et Vandekerkhove, P. (2007). How to compare MCMC simulation strategies ? *En révision pour Statistics and Computing.*

[B4] Chauveau, D. et Vandekerkhove, P. (2007). Variance reduction method for importance sampling schemes via f-driven proposals. *En préparation.*

Bibliographie

[1] Andrieu, C. et Moulines, E. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, **16**, 1462–1505.

[2] Atchadé, Y.F. et Rosenthal, J.S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* , **11**, 815–828.

[3] Bakry, D., Milhaud, X. et Vandekerkhove, P. (1997). Statistique de chaînes de Markov cachées à espace d'états fini. Le cas non stationnaire. *C. R. Acad. Sci. Paris*, Série I, **325**, 203–206.

[4] Bar-Shalom, Y. et Li, X. R. (1993). *Estimation et Tracking : Principles, Technics, and Software.* Artech House, Boston, London.

[5] Baum, L.E. et Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

[6] Baum, L.E., Petrie, T., Soules, G. et Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

[7] Benaïm, M. et Ben Arous, G. (2003) A Two-Armed Type Bandit Problem. *International Journal of Game Theory*, **32**, 3–16.

[8] Bickel, P.J. et Ritov, Y. (1996). Inference in Hidden Markov models I : LAN in the stationary case. *Bernoulli*, **2**, 199–228.

[9] Bickel, P.J., Ritov, Y. et Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

[10] Bosq, D. et Lecoutre, J.P. (1987). Théorie de l'estimation fonctionnelle. *Economica*, Paris.

[11] Bordes, L., Chauveau, D. et Vandekerkhove, P. (2007). Semiparametric EM algorithm for a two-component mixture model. *Computational Statistics and Data Analysis*, **51**, 5429–5443.

[12] Bordes, L., Delmas, C. et Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where a component is known. *Scand. J. Statist.*, **33**, 733–752.

[13] Bordes, L., Mottelet, S. et Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, **34**, 1204–1232.

[14] Bordes, L. et Vandekerkhove, P. (2005). Statistical inference for Partially Hidden Markov Models. *Communications in Statistics*, **34**, 1081–1104.

[15] Cappé, O., Moulines, E. et Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York.

[16] Cai, J. (1994). A Markov unconditional variance in ARCH. *J. Business Econom. Statist.*, **12**, 309–316.

[17] Chauveau, D. et Vandekerkhove, P. (1999). Un algorithme de Hastings-Metropolis avec apprentissage séquentiel. *C. R. Acad. Sci. Paris*, Série I, 173–176.

[18] Chauveau, D. et Vandekerkhove, P. (2001). Algorithmes de Hastings Metropolis en intéraction. *C. R. Acad. Sci. Paris*, Série I, 881–884.

[19] Chauveau, D. et Vandekerkhove, P. (2002). Improving convergence of the Hastings-Metropolis Algorithm with a learning proposal. *Scand. J. Statist.*, **28**, 13–29.

[20] Chauveau, D. et Vandekerkhove, P. (2007). A Monte Carlo estimation of the entropy for Markov chains. *Methodology and Computing in Applied Probability*, **9**, 133–149.

[21] Del Moral, P. (2004) *Feynman-Kac formulae. Genealogical and interacting particle systems with applications.* Probability and its Applications (New York). Springer-Verlag, New York.

[22] Del Moral, P. et Doucet, A. (2003). On a class of genealogical and interacting Metropolis Models. *Séminaire de Probabilités XXXVII*, Ed. J. Azéma and M. Emery and M. Ledoux and M. Yor, Lecture Notes in Mathematics 1832, Springer-Verlag Berlin, 415–446.

[23] Del Moral, P. et Guionnet, A. (1999). Central Limit Theorem for Nonlinear Filtering and Interacting Particle Systems. *Ann. Appl. Probab.*, **9**, 275–297.

[24] Douc, R. et Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, **3**, 381–420.

[25] Douc, R., Guillin, A., Marin, J.M. et Robert, C.P. (2007) Convergence of adaptive sampling schemes. *Ann. Statist.*, **35**, 420–448.

[26] Douc, R., Moulines, E. et Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *Ann. Statist.*, **32**, 2254–2304.

[27] Francq, C. et Roussignol, M. (1998). Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator. *Statistics,* **32**, 151–173.

[28] Fredkin, D.R. et Rice, J.A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Royal Soc. Lond. B*, **249**, 125–132.

[29] Fuk, D.K., et Nagaev, S.V. (1971-76). Probability Inequalities for Sums of Independent Random Variables. *Th. Probab. Appl.,* **16**, **21**, 643–660, 875.

[30] Garcia, R. et Perron, P. (1996). An analysis of the real interest rate under regime shift. *The review of Economics and Statistics.*

[31] Geman, S. et Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.,* **6**, 721–741.

[32] Gilks, W.R., Roberts, G.O. et Sahu, S.K. (1998). Adaptive Markov chain Monte carlo through regeneration. *JASA*, **93**, 1045–1054.

[33] Giné, E. et Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.,* **38**, 907–921.

[34] Giudici, P., Rydén, T. et Vandekerkhove, P. (2000). Likelihood-Ratio Tests for hidden Markov models. *Biometrics*, **56**, 742–747.

[35] Guihenneuc-Jouyaux, C., Richardson, S. et Longini, I. M. (2000). Modelling markers of disease progression by a hidden Markov process : application to characterizing CD4 cell decline. *Biometrics*, **56**, 733–741.

[36] Györfi, L. et Van Der Meulen, E. C. (1989). An entropy estimate based on a kernel density estimation, *Colloquia Mathematica societatis János Bolyai 57. Limit Theorems in Probability and Statistics Pécs (Hungary)*, 229–240.

[37] Haario, H., Saksman, E. et Tamminen, J. (2001). An adaptive Metropolis Algorithm. *Bernoulli*, **7**, 223–242.

[38] Hall, P. et Zhou, X-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, **31**, 201–224.

[39] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica,* **57**, 357–384.

[40] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, **57**, 97–109.

[41] Hamilton, J.D. et Susmel. R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, **64**, 307–333.

[42] Holden, L. (1998). Geometric Convergence of the Metropolis-Hastings Simulation Algorithm. *Statist. Prob. Letters*, **39**, 371–377.

[43] Hunter, D.R., Wang, S. et Hettmansperger T.P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.*, **35**, 224–251.

[44] Jackson, C.H. et Sharples, L.D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, **21**, 113–128.

[45] Ji, C., Snapp, R. et Psaltis, D. (1990). Generalizing smoothness constraints from discrete samples. *Neural Computation,* **2**, 188–197.

[46] Krishnamurthy, V. et Rydén, T. (1998). Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of Times Series Analysis*, **19**, 291–307.

[47] Lamberton, D. et Pagès, G. (2004). When can the two-armed bandit algorithm be trusted ? *Ann. Appl. Probab.*, **14**, 1424–1454.

[48] Lamberton, D. et Pagès, G. (2005). How fast is the two armed-bandit algorithm ? *Preprint*, LPMA n^o 1018.

[49] Lamberton, D. et Pagès, G. (2005). A penalized bandit algorithm. *Preprint*, LPMA n^o 1019.

[50] Lavanant, C. (2007). Réduction de la variance pour l'importance sampling au moyen de lois instrumentales corrigées non-paramétriquement par la loi cible. *Mémoire de Master 1*, Université de Marne-la -Vallée.

[51] LeGland, F. et Mevel, L. (2000). Exponential forgetting and geometric ergodicity in Hidden Markov Models. *Math. Control Signals Syst.,* **13**, 63–93.

[52] Leroux, B.G., et Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48**, 545–558.

[53] Leroux, B.G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.*, **40**, 127–143.

[54] Lindsay, B.G. (1995). *Mixture Models : Theory, Geometry and Applications.* NSF-CBMS Regional Conference Series in Probability and statistics.

[55] McLachlan, G.J. et Peel, D. (2000). *Finite mixture models*. Wiley, New York.

[56] McNeil, D.R. (1977). *Interactive Data Analysis*. Wiley, New York.

[57] Mengersen, K.L. et Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.,* **24**, 101–121.

[58] Meyn, S.P. et Tweedie, R.L. (2003). *Markov chains and Stochastic Stability*. Springer-Verlag.

[59] Narendra, K.S. et Thathachar, M.A.L. (1974) Learning automata–a survey. *IEEE. Trans. Systems Man. Cybernetics*, SMC-4, 323–334.

[60] Narendra, K.S. et Thathachar, M.A.L. (1989) *Learning automata–An Introduction.* Prentice-Hall, Englewood Cliffs.

[61] Niang, M. (1999). Algorithme de Narendra et application à l'allocation d'actifs. *Rapport de stage de DEA*. Olympia Capital Management et Univ. Marne-la-Vallée, France.

[62] Norman, M.F. (1968). On linear models with two absorbing barriers. *J. Math. Psych.*, **5**, 225-241.

[63] Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–284.

[64] Robert, C. et Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

[65] Robin, S., Avner, B.H., Daudin, J.J. et Pierre, L. (2007) A semi-parametric approach for mixture models : Application to false discovery rate estimation. *Comput. Statist. Data Analysis*, **51**, 5483–5493.

[66] Redner, R.A. et Walker, H.F. (1984). Mixtures densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195–249.

[67] Rydén, T. (1994). Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.* **22**, 1884–1895.

[68] Rydén, T., Teräsvirta, T. et Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model of absolute returns. *Journal of Applied Econometrics*. **13**, 217–244.

[69] Serfling, R.J. (1980). *Approximation theorems of Mathematical Statistics*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

[70] Titterington, D.M., Smith, A.F.M. et Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

[71] Tarrès, P. (2001). Pièges des Algorithmes stochastiques et marches aléatoires renforcées par sommets. *Thèse de doctorat*, ENS Cachan, France.

[72] Vandekerkhove, P. (1998). Simulated annealing with a sequential estimator of the energy. *C.R. Acad. Sci. Paris*, Série I, 1003–1006.

[73] Vandekerkhove, P. (2005). Consistent and asymptotical normal parameter estimates for hidden Markov mixtures of Markov models. *Bernoulli*, **11**, 103–129.

[74] Wu, C.F. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.

[75] Yakowitz, S. J. et Spragins, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.*, **39**, 209–214.