UNIVERSITÉ DE MARNE LA VALLÉE

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

Méthodes de Monte Carlo par chaînes de Markov et algorithmes de restauration de données manquantes

Didier CHAUVEAU

Soutenue le 17 décembre 2001 devant le jury composé de :

Rapporteurs :	Eric Moulines	ENST, Paris
	Gareth Roberts	Université de Lancaster, UK
	Bernard Ycart	Université Paris V
Examinateurs :	Marie Duflo	Université de Marne-la-Vallée
	Jean-Pierre Raoult	Université de Marne-la-Vallée
	Christian Robert	Université Paris IX Dauphine
	Tobias Ryden	Université de Lund, Suède

Remerciements

Je remercie vivement Eric Moulines, Gareth Roberts et Bernard Ycart d'avoir accepté d'être rapporteurs et membres du jury de cette habilitation, manifestant ainsi leur intérêt pour mes travaux. Je suis très reconnaissant à Marie Duflo et Christian Robert d'avoir accepté de faire partie de mon jury. Je remercie aussi tout particulièrement Tobias Rydèn qui a accepté de venir de Suède pour assister à ma soutenance.

Jean-Pierre Raoult a guidé mes premiers pas de chercheur, et n'a cessé depuis de me prodiguer conseils et avis, tout en me faisant profiter de sa vaste culture mathématique. Nos relations ont dépassé depuis longtemps le cadre professionnel et je le remercie de sa confiance et de sa grande amitié. Je suis heureux qu'il ait accepté également de faire partie de ce jury.

J'ai fait la connaissance de Jean Diebolt pendant la préparation de ma thèse, et notre collaboration a débuté peu après l'achèvement de celle-ci. Je souhaite qu'elle se poursuive encore longtemps. Sa culture et sa curiosité scientifique, sa disponibilité et sa grande force de travail sont pour moi un formidable moteur. Au-delà de notre collaboration, une véritable amitié nous lie. Pour tout cela un grand merci, Jean.

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 notamment à Frits Ruymgaart, Christian Robert, Gilles Celeux, Florence Muri et à tous les membres du groupe "MC^3" qui se reconnaîtront. Je pense aussi tout particulièrement à Pierre Vandekerkhove pour notre collaboration depuis son arrivée à Marne-la-Vallée.

Je souhaite remercier également toute l'équipe de Mathématique de l'Université de Marne-la-Vallée pour l'ambiance amicale qui y règne. Je pense en particulier à Christiane Cocozza, Michel Roussignol, Sophie Bloch-Mercier, Alain Pajor, Georges Oppenheim et Mireille Morvan, ainsi que les membres des groupes "Fiabilité" et "Algorithmes stochastiques".

Enfin, je voudrais exprimer toute mon affection à mes parents et ma famille, à Sabine et Gabriel qui ont supporté les tensions liées à la préparation de cette habilitation, tout en attendant le petit frère...

Table des matières

Chapitre 1

Introduction

Ce document de synthèse rassemble les travaux de recherche effectués depuis ma thèse. Mon travail s'articule autour de trois thèmes relativement indépendants. J'ai choisi ici de présenter ces thèmes selon l'importance qu'ils représentent dans mon travail plutôt que par ordre chronologique. Le thème principal est axé sur les algorithmes de Monte Carlo par Chaînes de Markov (MCMC), et est au centre de mon activité de recherche depuis plusieurs années maintenant. Ce thème "MCMC" est divisé en deux problématiques distinctes, le *contrôle de convergence* des méthodes MCMC et l'élaboration de nouvelles méthodes adaptatives, qui sont détaillées respectivement dans les chapitres 3 et 4. Le second thème, lié à l'étude des algorithmes de restauration de données manquantes et le troisième, à l'estimation statistique dans le cadre de problèmes mal posés, sont plus anciens et sont éloignés de mon activité principale. Ils sont pour cette raison présentés assez rapidement dans les chapitres 5 et 6.

Mon activité de recherche a considérablement évolué au cours de ces années : initialement motivée par des préoccupations industrielles et centrée sur la statistique et ses applications, elle s'est peu à peu ouverte à des problématiques plus théoriques et probabilistes, telles que le développement de méthodes de Monte Carlo par Chaînes de Markov adaptatives proches des systèmes de particules en interactions, et l'étude de leurs comportements asymptotiques. Mes derniers travaux ne sont cependant pas déconnectés des applications dans la mesure où, lorsque cela a un sens, j'accompagne les méthodes et algorithmes proposés de programmes informatiques de type "boîte noire" utilisables par les praticiens et disponibles en ligne.

Le contenu de ma thèse est brièvement rappelé dans cette introduction mais ne fait pas l'objet d'un chapitre. Vient ensuite une présentation générale de mes travaux ainsi que quelques perspectives de mon activité à venir. J'ai essayé de rendre cette présentation aussi peu technique que possible, les détails des résultats étant contenus dans les chapitres correspondants aux diverses parties. La liste de mes publications, articles soumis et rapports

techniques figure au chapitre 7 et les références de la forme [1] ou [RT1] dans le texte renvoient à cette liste. Une bibliographie figure à la fin du document, et les références à cette bibliographie sont explicites.

1.1 Résumé de la thèse

Ma thèse de doctorat était motivée par un problème de fiabilité en contexte industriel, dont la traduction statistique consistait en l'estimation des paramètres d'un mélange de distributions de durées de vie (voir, e.g., Titterington *et al*, 1985). Cette situation peut naturellement s'interpréter comme un problème d'estimation dans un modèle de données incomplètes, autrement dit en présence de variables *latentes*, ici les indicateurs des sous-populations d'appartenance des observations. Une technique adaptée consiste à utiliser des algorithmes de *restauration* des données manquantes, tels que l'algorithme EM (*Expectation-Maximisation*, Dempster *et al*, 1977), ou sa version stochastique SEM (Celeux et Diebolt, 1983, 1985), initialement motivée par des pathologies de EM telle que sa possible convergence vers des points selle de la vraisemblance (ces algorithmes sont brièvement présentés au chapitre 5).

Dans mon travail ([RT2]), les difficultés étaient de deux ordres : d'une part, les distributions considérées n'appartenaient pas forcément à une famille de lois exponentielle ; d'autre part, l'échantillon observé était soumis à une censure à droite déterministe rendant la structure des variables non observées plus complexe que dans le cas classique. Ces situations peuvent par exemple rendre l'algorithme EM non explicite, donc non utilisable en pratique. J'ai établi la convergence de EM vers un maximum local dans cette situation, étendant ainsi les résultats de Wu (1983), et Redner et Walker (1984). J'ai ensuite proposé des versions stochastiques dans la ligne de SEM, pour les situations où EM ne pouvait pas être utilisé, et étudié la convergence de la chaîne de Markov associée dans un cas simple, étendant ainsi les résultats de Celeux et Diebolt (1992). Ceci a été pour l'essentiel publié dans [1] et [3].

Dans une seconde partie, je me suis intéressé plus spécifiquement au problème industriel, en proposant une méthode d'optimisation mettant à profit les estimateurs issus des algorithmes étudiés dans la première partie. Le versant proprement appliqué de ce travail s'est concrétisé par un logiciel intégrant ces éléments, ainsi que quelques rapports techniques internes (ne figurant pas dans la liste de travaux détaillée au chapitre 7) qui précisent son fonctionnement, à destination des utilisateurs. Enfin, une étude liée à un problème de contrôle de qualité posé par le partenaire industriel, engagée au début de ma thèse mais pas directement liée avec le problème ci-dessus, à fait l'objet du document industriel [RT1].

Les ingrédients essentiels de ma thèse que sont la fiabilité, les modèles

avec données incomplètes, et surtout les techniques d'estimation itérative à base d'algorithmes stochastiques ont motivé les thèmes de recherche sur lesquels je me suis concentré depuis et qui font l'objet de ce document.

1.2 Présentation générale des travaux

Le versant appliqué de mon sujet de thèse (la fiabilité industrielle) m'a tout d'abord conduit à participer au groupe de travail "fiabilité" de l'Université Paris-Sud puis, à partir de 1993, à celui de l'Université de Marne la Vallée. J'y ai collaboré, dans le cadre de contrats entre ces universités et Électricité de France, à l'élaboration des rapports techniques [RT3] et [RT4] qui ne sont pas détaillés ici.

Méthodes de Monte Carlo par Chaînes de Markov

Depuis 1995, mes travaux de recherche sont centrés sur les méthodes de Monte Carlo par Chaînes de Markov (Gilks, Richardson et Spiegelhalter, 1996, ou Robert, 1996). Ces techniques faisant intensivement appel à l'ordinateur et qui entrent dans ce que l'on a coutume d'appeler aujourd'hui en français "le comput statistique" se sont rapidement développées à partir de 1990 (même si elles sont bien plus anciennes). Elles permettent de simuler une chaîne de Markov de loi stationnaire donnée (la *loi cible*) mais inaccessible à l'inférence ou la simulation directe (i.i.d.). Ces méthodes MCMC trouvent une grande part de leurs applications dans l'inférence bayésienne basée sur la loi a posteriori du paramètre d'intérêt. Les deux méthodes les plus utilisées sont l'échantillonneur de Gibbs (Geman et Geman, 1984) et surtout l'algorithme "universel" de Hastings-Metropolis (Hastings, 1970).

Je me suis intéressé naturellement aux algorithmes MCMC car l'algorithme de Gibbs peut être vu dans certaines situations comme une version bayésienne de SEM. Ce changement thématique m'a été facilité par la création, en 1995, d'un groupe de travail "MCMC" animé par Christian Robert (ENSAE-CREST, puis Université de Paris IX Dauphine), et regroupant des chercheurs de différents organismes (Université Paris V, INRIA Rhône Alpes, CNRS Grenoble, INSERM, ENST). Par la suite, la plupart des membres de ce groupe de travail ont été impliqués dans les actions du réseau Européen TMR (Training and Mobility of Researchers) sur le thème *Computational and Statistical methods for the analysis of spatial data*. Le thème MCMC constitue la plus importante partie de mon travail, et est séparé ici en deux problématiques : le contrôle de convergence et l'accélération d'algorithmes par des techniques d'apprentissage. L'une des originalités de ce travail est l'usage intensif, dans les deux thèmes, de la simulation de *chaînes parallèles*, c'est-à-dire de chaînes de Markov de même loi initiale et i.i.d. (ou bien couplées dans les méthodes adaptatives).

Contrôle de convergence des méthodes MCMC

Cette problématique est celle sur laquelle s'est concentré initialement notre groupe de travail MCMC. L'idée en est que même si l'on sait, sous des conditions assez générales, prouver la convergence (ergodicité ou ergodicité géométrique) des chaînes engendrées par les méthodes MCMC employées, ces résultats théoriques ne fournissent pas de *règle d'arrêt* assurant que l'on a effectué suffisamment d'itérations au sens de critères à déterminer. De nombreuses méthodes plus ou moins empiriques ont été proposées, et cette nécessité de disposer de méthodes de contrôle a donné lieu à de nombreuses publications entre 1992 et 2000 (voir Brooks et Roberts, 1998, pour un résumé).

Avec Jean Diebolt, nous avons proposé une méthode fondée théoriquement, et basée sur des critères d'atteinte de la normalité asymptotique pour les chaînes de Markov vérifiant le Théorème de Limite Centrale (TLC). Cette méthode est basée sur l'utilisation d'observations issues de chaînes parallèles, autrement dit i.i.d. de même loi initiale. Elle a tout d'abord fait l'objet d'un chapitre [6] et d'une application en vraie grandeur [7] (modélisation de la séquence d'ADN à partir du travail de Florence Muri) dans l'ouvrage collectif (*Lecture Notes in Statistics*) issu du travail de ce groupe MCMC. Une version plus élaborée a ensuite donné lieu à [RT5] et [9].

Pour être utilisées effectivement par les praticiens, ces méthodes de diagnostic de convergence doivent être *génériques*, i.e. ne pas nécessiter de programmation ou d'implémentation complexe et surtout spécifique des noyaux des algorithmes à contrôler ou des lois cibles. En fait, il semble que seules soient réellement utilisées aujourd'hui les méthodes de type "boîte noire" disponibles en ligne, par exemple dans des librairies telles que `StatLib`[1]. Notre méthode de contrôle par TLC est totalement générique puisque fondée uniquement sur les sorties des algorithmes. Je l'ai donc implémentée sous la forme d'un logiciel disponible en ligne ([L1]). Il a déjà été utilisé dans quelques situations réelles (génome [7], problème de géophysique,...). Ce travail est présenté au début du chapitre 3, § 3.1.

Estimation de la variance limite des chaînes de Markov

Plus récemment, nous avons proposé une méthode d'estimation de la variance limite qui intervient dans le TLC pour les chaînes de Markov, uniquement à partir d'observations issues de chaînes parallèles. Il s'agit d'un problème difficile à cause de la série des covariances provenant de la structure de dépendance. Dans [RT8] nous étudions en préalable les propriétés de stabilité de la *chaîne de Markov produit* dont les composantes sont les chaînes i.i.d, à partir des propriétés de stabilité des composantes. Nous donnons notamment des résultats de transfert à la chaîne produit de *conditions*

[1] `http://lib.stat.cmu.edu`

de drift et de Harris récurrence proposées par Meyn et Tweedie (1993). Ces résultats sont présentés au § 3.2.2. Dans [15], nous donnons la convergence en distribution vers un processus gaussien, d'un "processus variance empirique" issu de la moyenne de fonctions des processus de sommes partielles sur les chaînes i.i.d. Nous montrons que l'estimateur de la variance limite déduit de ce processus est meilleur au sens de sa variance que l'estimateur empirique naturel fondé sur les mêmes observations.

Ce travail théorique n'a pas encore donné lieu à des essais ou simulations. Il a notamment comme champ d'application le contrôle de convergence des algorithmes MCMC, dans la mesure où l'un des prérequis à l'atteinte de la normalité asymptotique est la stabilisation de cette variance limite qui est celle intervenant dans le TLC. Nous avons donc comme perspective de combiner cette méthode avec celle développée dans la boîte à outil de contrôle MCMC basée sur le TLC ([9] et [L1]). Des questions se posent quant à la comparaison des deux techniques et au fait de savoir si leurs conclusions sont en accord. L'objectif final est de proposer le tout dans un outil logiciel générique et complet. Ces travaux sont détaillés au chapitre 3, § 3.2.

Contrôle de la stabilité d'une chaîne de Markov par l'entropie

Avec Pierre Vandekerkhove, nous avons travaillé sur une problématique assez voisine : la caractérisation par des outils statistiques de la stabilité d'une chaîne de Markov, ou encore de sa vitesse de convergence vers la loi cible. L'un des objectifs visés est la comparaison de méthodes MCMC de manière "aveugle", c'est-à-dire uniquement à partir des sorties (observations) des algorithmes, seule information maniable lorsque les noyaux sont trop complexes pour mener une étude théorique. Les outils ici sont complètement différents, et basés sur des critères d'entropie et d'information de Kullback.

Dans [14], nous proposons une technique statistique permettant de s'assurer des propriétés de stabilité d'une chaîne de Markov à partir d'observations successives et aussi parallèles de cette chaîne et d'une connaissance analytique de son noyau. Nous définissons pour cela un estimateur de la "distance" de Kullback entre les lois de chaînes partant de deux positions initiales distinctes, et évoluant avec le même noyau. Cet estimateur, permettant de contrôler la rapidité avec laquelle se réalise l'oubli du point de départ, est construit à partir d'un estimateur original de l'entropie fondé sur une double intégration de Monte Carlo sur les chaînes parallèles. Nous montrons sous des conditions assez générales sa consistance et sa normalité asymptotique. La consistance forte est aussi donnée sous des conditions plus exigeantes. Cette partie figure également au chapitre 3 (§ 3.3) bien qu'il ne s'agisse pas à proprement parler de diagnostic de convergence MCMC.

Ce travail ouvre des perspectives pour les méthodes MCMC. Il s'agit d'utiliser des estimateurs de l'information de Kullback entre la loi d'une

chaîne et la loi cible en fonction du temps afin de comparer les différents algorithmes MCMC utilisables pour un problème donné, et ce encore une fois uniquement à partir des sorties des algorithmes et de certaines informations minimales sur leurs noyaux et sur la loi cible. C'est un problème souvent rencontré par les utilisateurs de méthodes MCMC. En effet l'algorithme de Gibbs et celui de Hastings-Metropolis peuvent être appliqués de nombreuses manières pour résoudre un problème donné, i.e. reconstruire une loi cible. Pour définir un algorithme de Gibbs, plusieurs choix de décomposition de la loi en lois conditionnelles sont possibles. Pour l'algorithme de Hastings-Metropolis, un choix virtuellement infini de *lois instrumentales* s'offre à l'utilisateur (voir § 4.1). Il n'est pas toujours clair de déterminer la meilleure stratégie en terme de vitesse de convergence ou d'exploration des spécificités de la loi cible. Nous avons déjà proposé un estimateur dans le cadre de l'algorithme de Hastings-Metropolis qui se comporte bien expérimentalement, et dont l'étude théorique est en cours.

Algorithmes MCMC adaptatifs

Une autre problématique très vivante actuellement dans le domaine des méthodes MCMC est l'élaboration de nouveaux algorithmes permettant de traiter de manière plus performante certaines situations délicates telles que la reconstruction de lois cible complexes (e.g., multimodales avec éventuellement des modes distants). Les méthodes usuelles donnent des chaînes de faible mélangeance dans ces situations, en raison de la difficulté liée à une bonne exploration des régions d'intérêt du support de la loi cible.

Avec Pierre Vandekerkhove, nous avons proposé des versions adaptatives de l'algorithme de Hastings-Metropolis basées sur l'idée suivante : Dans certains cas, cet algorithme est géométriquement et uniformément ergodique, et la vitesse est d'autant meilleure que la loi instrumentale est proche de la cible. En partant d'une loi instrumentale pratiquement arbitraire, un tel algorithme va donc converger même avec une faible vitesse. Si il est possible d'injecter les lois successives de la chaîne comme lois instrumentales pour les pas suivants, on accélère la convergence puisque les lois instrumentales successives se rapprochent elles-mêmes de la cible. Évidemment ces lois marginales de la chaîne sont inconnues, mais peuvent être estimées à partir de chaînes parallèles. Le problème est que faire cette estimation revient à effectuer un *couplage* des chaînes, qui perdent leur indépendance et leur caractère markovien ce qui rend difficile l'étude théorique de tels processus.

La première solution que nous avons proposée dans [8] et [10], et qui est présentée au chapitre 4, § 4.2, consiste à utiliser un estimateur par histogramme de ces lois successives en certains instants, et à supprimer à chaque fois les chaînes ayant servi à l'estimation. Les chaînes utilisées restent donc i.i.d. mais sont des chaînes de Markov non homogènes. Nous prouvons dans ce cadre, et avec une double asymptotique (en temps et en

nombre de chaînes), qu'une chaîne issue de cet algorithme assure p.s. une vitesse géométrique meilleure que celle induite par la loi instrumentale arbitraire initiale. Cependant, à cause de l'élimination des chaînes aux instants de couplage, cette méthode est gourmande en temps de calcul, et son implémentation est assez lourde. De plus elle ne peut raisonnablement être utilisée en pratique que comme méthode exploratoire en arrêtant les simulations parallèles après quelques apprentissages.

Avec Anas Altaleb, nous avons comparé dans le cadre de l'analyse bayésienne du modèle Logit cette méthode adaptative avec une méthode *ad hoc* : un algorithme de Hastings-Metropolis de type marche aléatoire utilisant une loi instrumentale basée sur une approximation gaussienne calibrée sur les données. Nous montrons dans [11] que la méthode adaptative a un meilleur comportement pour ce type de problème.

Avec Pierre Vandekerkhove, nous avons repris récemment l'idée de base de [10], mais en étudiant directement dans [12] et [16] les processus de Hastings-Metropolis en interaction et non markoviens provenant de l'estimation en certains instants de leur loi commune par un estimateur à noyau sur données dépendantes. Cet estimateur est ensuite utilisé pour construire la loi instrumentale de la dynamique de Hastings-Metropolis de ces même processus jusqu'à l'instant suivant de couplage. Le système obtenu est assez proche des systèmes de particules en interactions utilisés en filtrage (voir, e.g., Del Moral et Miclo 2000). Nous obtenons un résultat théorique similaire à [10], mais bien plus efficace en pratique : il n'y a plus à éliminer de chaînes, et l'apprentissage peut se faire tout au cours du temps. J'ai également écrit un logiciel de type "boîte noire" qui implémente cette méthode de façon générique, et sera bientôt disponible en ligne ([L2]). Ce travail est détaillé au chapitre 4, § 4.3.

Cette étude ouvre de très intéressantes perspectives pour l'utilisation des systèmes de particules en interactions dans les problématiques MCMC. Il suggère aussi l'emploi de méthodes hybrides, entre apprentissage sur les régions d'intérêt déjà découvertes et élargissement de l'exploration du support à l'aide de pas de Hastings-Metropolis de type marche aléatoire, plus aptes à découvrir de nouvelles zones du support contenant de la masse.

Algorithmes de restauration

En prolongement naturel de mon travail de thèse ([1] et [3]), je me suis intéressé aux algorithmes stochastiques de restauration des données manquantes ou des variables latentes, issus de l'algorithme EM. Avec Gilles Celeux et Jean Diebolt, nous avons étudié et comparé dans [5] le comportement des nombreuses versions stochastiques existantes, dans le cadre du problème classique de la reconnaissance d'un mélange de distributions gaussiennes. Nous avons comparé ces algorithmes sur plusieurs exemples simulés et réels plus ou moins difficiles du point de vue de l'estimation des pa-

ramètres (populations imbriquées, populations ne se distinguant que par leurs variances). Nous avons notamment proposé des solutions empiriques au problème de permutation d'étiquetage des composantes du mélange, et montré que, à cause de ce problème, l'estimateur bayésien moyenne a posteriori habituellement utilisé dans ces méthodes n'est pas adapté aux situations multimodales. Ce travail méthodologique fait l'objet du chapitre 5. J'ai également participé dans [13] à la discussion de l'article de Meng et van Dyck (1997) écrit à l'occasion des vingt ans de l'algorithme EM.

Problèmes mal posés en statistique

Je me suis intéressé à des techniques d'estimation statistique dans le contexte de *problèmes mal posés* sous l'impulsion du professeur Frits Ruymgaart, qui était mon responsable scientifique lors de mon post-doctorat à l'Université de Lubbock, Texas, Etats-Unis. Il travaillait alors avec Arnold van Rooij (Université de Nijmegen, Pays-bas) sur ces problèmes qui se ramènent typiquement à de la régularisation d'inverse d'opérateurs. Nous avons étudié dans [2] la construction d'une suite d'inverses régularisés pour la transformée de Laplace, ce qui se ramène à un cas particulier des techniques de déconvolution étudiées par Caroll, van Rooij et Ruymgaart (1991). Le problème statistique associé auquel nous nous sommes intéressés est la reconstruction de la densité de mélange pour un mélange continu de lois exponentielles. Le principe est d'interpréter ce mélange observé comme une transformée de Laplace bruitée et de déterminer son inverse. Dans [4], nous avons proposé une technique de régularisation d'inverse assez similaire, adaptée à la reconstruction des mesures de mélange de lois exponentielles, dans le cas de mesures discrètes signées. Ce travail est présenté au chapitre 6.

Chapitre 2

Introduction (english version)

This chapter is a translation of the introduction (chapter 1). My work can essentially be split into three separate topics. The main topic (in terms of amount of work, publications and current interest) is related to Markov Chain Monte Carlo methods (MCMC), into which I am involved since about 1995. My work in this field can be divided in two subjects : the MCMC convergence assessment problem (presented in chapter 3), and the development of new MCMC adaptive methods for speeding up convergence (presented in chapter 4).

The two other topics are related respectively to the study of stochastic versions of the EM algorithm (started during my PhD), and the study of some estimation techniques through operator inversion related to ill-posed problems (started during my post-doc in 1992). I am not currently working on these fields, so that these are presented more briefly than the MCMC-related subjects in the document, respectively in chapters 5 and 6.

The content of my PhD thesis is briefly summarized in section 2.1. Then section 2.2 gives an overview of my contributions to the fields I have worked on, together with some open issues and prospects for futur work. The list of my publications and technical reports is in chapter 7, and references like, e.g., [1] or [RT1] point to this list. The list of the other references (like Titterington *et al*, 1985) is appended to the document (bibliography section).

2.1 Summary of the PhD thesis

The subject of my thesis was initially motivated by reliability problems in an industrial framework (electronic systems). The mathematical translation of the problem was statistical inference for incomplete data models. The life data were supposed to come from a mixture of distributions (see, e.g.,

Titterington *et al*, 1985), after some censoring process. The application of the EM algorithm (Dempster *et al*, 1977) and its stochastic versions like "SEM" (Celeux and Diebolt, 1983, 1985) seemed appropriate to handle this model.

In this specific situation, there were two difficulties : the distributions of the life data did not always belong to an exponential family, and the incomplete data structure was twofold, due to the censoring process and the missing data coming from the mixture model. These difficulties prevented us to directly use existing results from the literature concerning EM and SEM convergence. Moreover, EM could not be implemented in closed form under some situations. We established the convergence of EM to a local maxima in this situation, and proposed stochastic versions overcoming the difficulty of implementation of EM. We also proved the convergence of the Markov chain associated to this version of SEM in a simple case. This has been published essentially in [1] and [3].

2.2 General overview of the contributions

MCMC methods

Since 1995, my work is essentially related to MCMC simulation methods (see, e.g., Gilks, Richardson and Spiegelhalter, 1996, or Robert, 1996). A MCMC method simulates a Markov chain with some distribution of interest (the target) as its stationary distribution. This target distribution usually comes as the posterior distribution in Bayesian inference, and cannot be simulated with standard i.i.d. Monte-Carlo techniques. The two most-used methods are the Gibbs sampler (Geman and Geman, 1984) and the Hastings-Metropolis algorithm (Hastings, 1970). My interest in MCMC methods originally came from the fact that in certain framework, the Gibbs sampler could be interpreted like a Bayesian version of the stochastic EM algorithm. I started working on MCMC with the working group "MCMC" created and headed by Christian Robert (ENSAE-CREST and Université Paris IX), together with several colleagues from other institutions. Members of this group also became lately members of the TMR network (Training and Mobility of Researchers) on *Computational and Statistical methods for the analysis of spatial data.*

One original approach of my work in both subjects (convergence control and adaptive methods) is the use of *parallel chains*, i.e. of i.i.d. Markov chains with a same initial distribution (some coupling of these chains also occurs in the adaptive methods).

MCMC convergence assessment

Our MCMC working group initially focused on the MCMC convergence assessment problem. The idea is that even if we can prove under general conditions desirable convergence properties of the Markov chain of interest (like ergodicity or Strong Law of Large Numbers), these theoretical results do not provide *stopping rules* for the end user running a MCMC method. There has been a growing concern about convergence assessment methods, and several techniques (sometimes more or less empirical) have been proposed between, say, 1992 and 2000. A survey can be found in, e.g., Brooks and Roberts (1998).

Together with Jean Diebolt, we have proposed a method theoretically valid, grounded on the fact that the normality resulting from the Central Limit Theorem (CLT) for Markov chains is a testable implication of sufficient mixing. A first control tool tests the normality hypothesis for normalized averages of functions of the Markov chain over i.i.d. chains. A second connected tool is based on graphical monitoring of the stabilization of the associated variance. These techniques appeared first in [6], as a chapter of the *Lecture Notes in Statistics* written by our MCMC working group. An actual, real-size MCMC application for the DNA sequence (from the initial work of Florence Muri) was also published in [7]. A more complete and improved version of our method, focusing on its automated aspects, has been published in [9].

An important criterion for convergence assessment methods is the required computer investment : diagnosis requiring problem-specific computer codes for their implementation (e.g., requiring knowledge of the transition kernel of the Markov chain) are far less usable for the end user than diagnosis solely based upon the output of the sampler, since the latter can use available generic code. Actually, it appears that the methods which are used by the practitioners are the generic methods available from online libraries like, e.g., `StatLib`[1]. Our method is completely generic, since it is based only on the realizations from parallel chains, and it works without knowledge on the sampler driving the chain. In addition, the normality diagnosis leads to automated stopping rules. Both tools has been implemented in a software available online ([L1]). This work is presented in chapter 3, § 3.1.

Estimation of the limiting variance for Markov chains

More recently, we have proposed a method for estimating the limiting variance in the Central Limit Theorem (CLT) for Markov chains. It turns out that estimating this variance is not easy, because of the sequence of covariances coming from the dependence structure. The interesting point is that our method uses solely realizations from i.i.d. Markov chains (like

[1] `http ://lib.stat.cmu.edu`

the CLT control method). Hence its usability does not depend on analytical knowledge, or complexity of the kernel, unlike other methods.

First, we study in [RT8] how various forms of stability properties for a single Markov chain transfer to the *m-fold product Markov chain*, i.e. the Markov chain over the product state space resulting from the observation of m i.i.d. copies of the original chain. We give in particular sufficient conditions to carry over *drift conditions* and Harris recurrence properties (as defined in Meyn and Tweedie, 1993) to the product chain. These results, useful for the estimation of the limiting variance, are detailed in § 3.2.2.

In [15], we state the weak convergence to a Gaussian process, of some "empirical variance process" built from the average of functions of partial-sum processes issued from the i.i.d. chains. We use this limiting process to control the fluctuations of the variance, and to compute an estimate better than the intuitive estimate based on the same realizations. This study is presented in chapter 3, § 3.2.

This theoretical study has not yet been tested on simulated data or actual MCMC algorithms. It is related to our MCMC convergence assessment technique monitoring normality. Indeed, the variance appearing in the CLT needs to stabilize before we can assume that functions of the chain have reach an approximate normality. Our objective is to combine the control of fluctuations and estimation of the limiting variance, with the software [L1] in such a way to propose a complete, "black-box type", convergence assessment method.

Control of stability properties of Markov chains through an entropy estimator

With Pierre Vandekerkhove, we have worked on a somehow connected topic : the control or comparison of the stability properties of Markov chains using statistical techniques. One of our objective is to compare the efficiency of several MCMC methods in a "blind" way, i.e. solely from (simulated) realizations from the algorithms. The motivation is that these realizations can be the only manageable information when the kernels are too complex to be studied theoretically. The technique here is rather different than in the previous section, and is based on entropy and Kullback information criteria.

In [14], we propose a way to check stability properties of a Markov chain, on the basis of realizations from parallel chains, provided that the density of the kernel is analytically known. We define an estimate of the Kullback "distance" between de distributions of two chains started from different initial positions, and iterating the same kernel. The idea is that this estimate provides information about how fast the chains forget their starting positions (and if they do forget it). It is constructed from an original estimate of the entropy, grounded on some double Monte Carlo integration over two groups of parallel chains. We show, under mild conditions, its weak and strong

consistency and asymptotic normality. This work is also detailed in chapter 3, § 3.3, even if is not strictly speaking a method for MCMC convergence assessment.

Our futur objective is to extend this technique in order to compare the efficiency of MCMC algorithms. Many different MCMC methods can often be implemented for the same problem : The Gibbs sampler can be defined using different decompositions in full conditionals, and for the Hastings-Metropolis algorithm, an almost unlimited choice of proposal densities is available. Determining which algorithm is the best in term of convergence rate is not clear. We suggest to estimate the Kullback information between the distribution of the chain and its target, and to monitor this estimate against the time (iterations). We have already some good experimental results for the Hastings-Metropolis algorithm, and the theoretical study is in progress.

Adaptive MCMC methods

There is an increasing interest in the current literature for the development of new MCMC methods tailored to delicate situations, such as the reconstruction of complex target densities (e.g., multimodal with distant and/or small modes). Indeed, the classical methods result in slowly mixing chains in these situations, essentially because a good (exhaustive) exploration of the support of the target is difficult to achieve.

With Pierre Vandekerkhove, we have proposed adaptive versions of the Hastings-Metropolis algorithm, based on the following idea : In some situations, this algorithm is geometrically uniformly ergodic, and the convergence rate is driven by the proximity between the target and the proposal density. Hence the successive densities of a chain using an arbitrary proposal density converge to the target. If these successive densities could be used in replacement of the initial proposal density, this would improve the rate of convergence (very rapidly, see fig. 4.1). Unfortunately, these marginals are unknown. Our suggestion is to estimate the successive densities from parallel chains. The main drawback is that performing this estimation leads to a coupling of the chains. These are no longer Markovian and independent, so that the theoretical study is difficult.

The first solution we have suggested in [8] and [10] is to use histogram estimates of these densities at selected instants, and to discard the chains used to perform the estimations, in such a way to preserve the i.i.d. property of the remaining chains. These chains become non homogeneous but their Markov property is also preserved. We prove in this setup, and asymptotically in time and number of chains, that a single chain issued from this strategy is a.s. better than any chain using an arbitrary proposal density. However, this method suffers from two drawbacks in practice : (1) its implementation may be tedious in large dimension (because of the histogram

constructions) ; (2) the number of coupling instants must be kept small, due to the elimination process, so that the method can essentially be used only as an exploratory method to build a good proposal density.

With Anas Altaleb, we have compared this adaptive method against a *ad hoc* MCMC algorithm, in an actual model : the Bayesian analysis of the Logit model. The competing method is a random walk Hastings-Metropolis algorithm using a proposal density grounded on a Gaussian approximation calibrated from the data. We show in [11] that the adaptive method performs better, in terms of exploration of the support of the target and speed of convergence.

With Pierre Vandekerkhove, we have recently proposed in [12] and [16] a new adaptive version based on the same idea. Here, we consider directly parallel "Hastings-Metropolis processes" that are non Markovian and non independent. At selected *coupling times*, their common distribution is estimated using kernel density estimators based on these dependent data. These estimates serve as a basis to build the proposal densities which are used in the Hastings-Metropolis dynamic of all the processes up to the next coupling time. The resulting structure has close connections with the interacting particle systems used in non-linear filtering (see, e.g., Del Moral and Miclo 2000). We prove an asymptotic result similar to [10], but much more efficient in practice : there is no need to discard the chains used at the coupling times, so that adaptation can be performed all over the simulation duration. In addition, the kernel estimator is easier to compute than the histogram in large dimension. A generic, "black-box" type computer code has been written to implement this method, which will be available online ([L2]). This work is detailed in chapter 4, § 4.3.

This technique leads to interesting prospects, such as the use of interacting particle systems (with resampling) in MCMC simulations. It also suggests that hybrid methods should be tried, taking advantage of the adaptation on regions of the support already explored, and of the capacity of the random-walk Hastings-Metropolis dynamic to explore and discover new regions of interest.

Stochastic versions of the EM algorithm

Following the work done during my PhD ([1] and [3]), I studied some stochastic versions of the EM algorithm dedicated to incomplete, or latent data situations. With Gilles Celeux and Jean Diebolt, we have compared in [5] various existing stochastic EM algorithms for recovering mixtures of Gaussian distributions. We have done an intensive simulation study, comparing several methods for different situations (intricate mixtures, mixtures with equal mean and different variances, and a real data case). We have also pointed out the *label switching* difficulty and suggest empirical solutions. This is detailed in chapter 5.

Ill-posed problems in statistics

I have been involved in the study of ill-posed problems during my post-doc with Professor Frits Ruymgaart, at Texas Tech. University (Lubbock, TX). He was working with Professor Arnold van Rooij (University of Nijmegen, NL) on these questions related to regularization of operator inversion. In [2], we have proposed a sequence of regularized inverses for the Laplace transform, by relating it to a particular case of deconvolution studied by Caroll, van Rooij and Ruymgaart (1991). The associated statistical problem is the estimation of the mixing density of a continuous mixture of exponential distributions. Observations from this mixture are interpreted as a noisy Laplace transform, for which inversion is ill-posed. In [4], we have applied a somehow similar technique to recover mixtures of exponential distributions, when the mixing density is a discrete signed measure. This work is presented in chapter 6.

Chapitre 3

Contrôle de convergence des méthodes MCMC

Un algorithme MCMC simule une chaîne de Markov à temps discret $X = (X_t, t \geq 0)$, de loi stationnaire π donnée (la loi cible) et d'espace d'état E. Cette loi est souvent la loi a posteriori d'un modèle bayésien, et un tel algorithme est utilisé lorsque les intégrales de la forme

$$\pi(h) \triangleq \mathbb{E}_\pi(h) = \int_E h(x)\,\pi(dx) \tag{3.1}$$

ne sont pas calculables explicitement, et que π n'est pas simulable directement de façon i.i.d. De bonnes introductions à la très importante littérature statistique sur le sujet sont Gilks, Richardson et Spiegelhalter (1996), Robert (1996), ou encore l'article de Gelfand et Smith (1990). Les méthodes les plus utilisées sont l'algorithme de Hastings-Metropolis (Hastings, 1970) qui sera présenté au chapitre 4, et l'échantillonneur de Gibbs (Geman et Geman, 1984).

L'objectif est donc soit de reconstruire π à partir d'un pseudo-échantillon issu des itérés de la chaîne (détermination des modes et des régions chargées par π), soit d'approcher (3.1) par une moyenne empirique

$$\hat{h}_T = \frac{1}{T}\sum_{t=1}^{T} h(X_t). \tag{3.2}$$

Même si l'ergodicité de la chaîne produite par une méthode MCMC est prouvée sous des conditions assez générales, et que la Loi Forte des Grands Nombres (LFGN) pour les chaînes de Markov assure que $\hat{h}_T \to \mathbb{E}_\pi(h)$ p.s., ces résultats asymptotiques ne donnent pas de critères de contrôle de la chaîne simulée au sens suivant : Si l'on souhaite un échantillon i.i.d. de π, il faut déterminer un instant t_0 de sorte que $X_{t_0+t} \sim \pi$ approximativement, et éventuellement un intervalle entre observations successives afin

d'obtenir des réalisations approximativement indépendantes. Si l'on souhaite utiliser la LFGN, on cherche T assez grand pour assurer une certaine précision dans l'approximation par (3.2) de $\mathbb{E}_\pi(h)$. Déterminer l'instant t_0 à partir duquel on peut raisonnablement admettre que la chaîne est dans son régime stationnaire, et une *règle d'arrêt* T des simulations, est l'objectif du contrôle de convergence des algorithmes MCMC. De nombreuses méthodes ont été proposées (voir, e.g., Brooks et Roberts (1998) pour un panorama de ces méthodes). En consultant cette littérature et celle concernant les applications des méthodes MCMC, on se rend compte que seules sont utilisées en pratique les méthodes de contrôle génériques accompagnées d'un logiciel disponible en ligne et ne nécessitant qu'un minimum d'investissement de programmation. Les méthodes apparemment les plus utilisées sont le *contrôle binaire* de Raftery et Lewis (1992), basée sur une unique chaîne, et la comparaison de variance de Gelman et Rubin (1992), basée sur des chaînes parallèles. Le contrôle binaire est très utilisé car très simple à mettre en œuvre et disponible dans la boîte à outil `CODA` (écrite en `Splus`, voir Best, Cowles et Vines, 1995), mais il est malheureusement peu fondé théoriquement (voir Robert, 1996, chap. 6). De même, la technique de comparaison de variance nécessite une connaissance a priori ou déterminée par des méthodes numériques des modes de π, et repose sur une hypothèse gaussienne souvent fausse en pratique.

Ainsi, il nous a paru opportun de proposer une méthode de contrôle générique, fondée théoriquement, et pouvant donner lieu à un logiciel disponible en ligne et de type "boîte noire".

3.1 Contrôle de convergence par TLC

Nous avons proposé dans [RT5] et [9] une méthodologie de contrôle reposant sur le fait que la normalité asymptotique de fonctions de la chaîne de Markov est un critère que l'on peut tester statistiquement, et qui implique l'atteinte du régime stationnaire. La méthode repose sur la simulation de chaînes parallèles (i.i.d.), ce qui sera le cas de la plupart des techniques présentées dans ce document[1].

Prenons h à valeurs réelles pour simplifier la présentation, et posons

$$S_n(h) = \sum_{t=1}^{n} h(X_t) \quad \text{et} \quad S_n(\bar{h}) = \sum_{t=1}^{n} \left[h(X_t) - \pi(h)\right].$$

Il est naturel, lorsque l'on veut approcher $\mathbb{E}_\pi(h)$ à l'aide de la LFGN, de contrôler la précision de cette approximation à l'aide du Théorème de Limite

[1] J'ai essayé de conserver des notations cohérentes pour ces chaînes i.i.d. tout au long des chapitres suivants, en notant le temps en indice et l'index du numéro de chaîne en exposant. Les articles correspondants ne respectent malheureusement pas forcément ces conventions.

Centrale (TLC) pour les chaînes de Markov, qui assure (lorsqu'il est vérifié) qu'il existe une *variance limite* associée à h,

$$\sigma^2(h) = \lim_{n\to\infty} \frac{1}{n} \mathrm{var}\left[S_n(h)\right], \tag{3.3}$$

telle que $0 < \sigma^2(h) < +\infty$, et

$$\frac{1}{\sqrt{n}} S_n(\bar{h}) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(h)\right). \tag{3.4}$$

Ceci nécessite bien entendu que la chaîne d'intérêt ait des propriétés d'ergodicité suffisantes. Des conditions assurant cela sont données par exemple dans Meyn et Tweedie (1993), chap. 17.

Nous avons proposé de tester le temps n nécessaire pour que les sommes normalisées $S_n(h)/\sqrt{n}$ de certaines fonctions appropriées de X "atteignent" la normalité. Ces tests sont simples à mettre en œuvre si l'on dispose de *chaînes parallèles* : Si l'on simule m chaînes de Markov i.i.d. selon une même loi initiale, que l'on note la ℓ-ième chaîne $X^\ell = (X_t^\ell, t \geq 0)$, pour $\ell = 1, \ldots, m$, et

$$S_n^{(\ell)}(h) = \sum_{t=1}^{n} h(X_t^\ell),$$

la somme associée, il est facile de construire à tout instant n un m-échantillon d'observations de ces sommes normalisées :

$$\left(\frac{S_n^{(1)}(h)}{\sqrt{n}}, \ldots, \frac{S_n^{(m)}(h)}{\sqrt{n}} \right). \tag{3.5}$$

Un choix simple car d'interprétation facile consiste à prendre $h(x) = \mathbb{I}_A(x)$, pour $A \in \mathcal{B}(E)$, de sorte que $S_n(\mathbb{I}_A)$ soit le temps d'occupation de A par une chaîne durant les n premiers sauts. Ainsi, par exemple, dire que l'échantillon (3.5) est approximativement gaussien pour A dans une région modale de π signifie que toutes les chaînes ont visité A approximativement le même nombre de fois, avec certaines fluctuations, ce qui indique une bonne mélangeance. Au contraire, si certaines chaînes sont parties de A et y sont restées bloquées entre $t = 1$ et $t = n$, et si d'autres chaînes sont parties d'ailleurs et n'ont jamais visité A dans le même temps, l'échantillon (3.5) sera bimodal, donc fortement non gaussien, ce qui révélera une mélangeance insuffisante des chaînes, et la nécessité de continuer la simulation au-delà de n.

Notre méthode de contrôle est fondée sur cette constatation, et consiste à appliquer de manière séquentielle des tests de normalité à certaines collections de fonctions telles que $\mathbb{I}_A$ jusqu'à l'atteinte d'une normalité "satisfaisante". Simultanément, nous proposons comme outil complémentaire un contrôle empirique de la stabilisation des variances associées $\sigma^2(\mathbb{I}_A)$. Ce contrôle de la variance est plus satisfaisant théoriquement dans le cas où E est fini, que nous présentons d'abord.

3.1.1 Cas discret fini

Dans le cas où E est fini, $|E| = K$, la transition est une matrice $P = P_{ij}$, $1 \leq i, j \leq K$, et la probabilité invariante est donnée par $\pi = (\pi_i, i \in E)$. On s'intéresse alors aux temps d'occupation des états de E, c'est-à-dire aux fonctions de la chaîne de la forme $\mathbb{I}_i(\cdot)$, et l'on note simplement $S_n(i) = S_n(\mathbb{I}_i)$ ces temps d'occupation.

Contrôle de la normalité des temps d'occupation

Suivant le principe énoncé plus haut, l'algorithme de contrôle consiste à lancer des chaînes i.i.d. suivant une loi initiale dispersée (e.g. uniforme sur E), et à contrôler en des instants prédéterminés n_k, $k = 1, 2, \ldots$, la normalité des échantillons de la forme (3.5). Nous avons choisi pour sa puissance contre une alternative très générale le test de Shapiro-Wilks (Shapiro et Wilks, 1965) avec un niveau α à choisir. Pour $n_0 = 0 < n_1 < n_2 < \cdots$, un premier algorithme simple réservé au cas fini, donné ici pour un $i \in E$ est :

1. `Simuler les` m `chaînes de` n_{k-1} `à` n_k
2. `Mettre à jour l'échantillon` $\left(\frac{S_{n_k}^{(1)}(i)}{\sqrt{n_k}}, \ldots, \frac{S_{n_k}^{(m)}(i)}{\sqrt{n_k}} \right)$
3. `Calculer la statistique de Shapiro-Wilk` $SW(i, n_k)$,
   `Si` H_0 `est rejetée,`
   $k \leftarrow k + 1$ `et aller en 1`
   `sinon fin.`

Cet algorithme retourne donc le premier instant n_k pour lequel H_0 (la normalité) n'est pas rejetée. La statistique SW est à valeur dans $[0, 1]$ et prend des valeurs proches de 1 sous H_0. Il est ainsi possible de suivre graphiquement l'évolution de $n \to SW(i, n)$ jusqu'à son passage au-dessus du seuil correspondant à α. En pratique, on testera simultanément sur les même chaînes une collection d'états $i \in E' \subset E$. Le choix de E' dépend de la taille du problème et est discuté dans [9] : Si K est petit on peut prendre $E = E'$; si K est très grand on se rapproche de la situation ou E est continu et mieux vaut alors utiliser la version adaptée au cas général, décrite au § 3.1.2.

Remarquons que cette détermination d'un instant d'atteinte d'une normalité approximative est liée à la vitesse dans le TLC, et donc au théorème de Berry-Esséen. Nous avons tenté dans [RT5] d'utiliser les bornes de Berry-Esséen pour construire une méthode de contrôle. Malheureusement, la mauvaise qualité de ces bornes (même dans le cadre i.i.d.) rend cette approche inexploitable en pratique.

Contrôle de la stabilisation de la variance

Il est naturel d'associer au contrôle de la normalité un contrôle de la stabilisation de la variance après n pas,

$$\sigma_n^2(h) = \frac{1}{n}\mathrm{var}(S_n(h)),$$

autour de la variance limite $\sigma^2(h)$. Dans le cas discret et pour un état $i \in E$, l'estimateur naturel de $\sigma_n^2(\mathbb{I}_i)$ construit sur m chaînes parallèles observées jusqu'à l'instant n est simplement la variance empirique

$$\hat{\sigma}_n^2(m, \mathbb{I}_i) = \frac{1}{nm}\sum_{\ell=1}^{m}\left(S_n^{(\ell)}(i) - \overline{S_n(i)}\right)^2, \quad \text{où } \overline{S_n(i)} = \frac{1}{m}\sum_{\ell=1}^{m} S_n^{(\ell)}(i).$$

Dans le cas discret, on peut construire d'autre part un estimateur de la variance limite. A partir de l'étude des chaînes finies donnée dans Kemeny et Snell (1960), on vérifie que la variance limite associée à $\sigma^2(\mathbb{I}_i)$, $i \in E$, à une forme simple. Elle s'exprime à partir de la *matrice fondamentale*

$$Z = (I - (P - A))^{-1} = I + \sum_{k=1}^{\infty}(P^k - A), \tag{3.6}$$

où A est la matrice dont toutes les lignes sont égales à π. Si l'on construit la matrice $C = (C_{ij})$ à partir de $Z = (Z_{ij})$ par

$$C_{ij} = \pi_i Z_{ij} + \pi_j Z_{ji} - \pi_i \delta_{ij} - \pi_i \pi_j, \tag{3.7}$$

où $\delta_{ij} = 0$ pour $i \neq j$ et $\delta_{ii} = 1$, on a $\sigma^2(\mathbb{I}_i) = C_{ii}$. On obtient alors facilement le résultat suivant ([9], proposition 2) :

Proposition 1 *Pour toute loi initiale et tout entier n fixé assez grand, on peut construire des estimateurs $\hat{P}$, $\hat{Z}$ et $\hat{C}$ fortement consistants en m des matrices P, Z et C. On déduit alors de (3.6) et (3.7) un estimateur $\hat{\sigma}^2(n, m, \mathbb{I}_i)$ de la variance limite, et*

$$\left.\begin{array}{ll} \hat{\sigma}^2(n, m, \mathbb{I}_i) & \to \sigma^2(\mathbb{I}_i) \\ \hat{\sigma}_n^2(m, \mathbb{I}_i) & \to \sigma_n^2(\mathbb{I}_i) \end{array}\right\} \textit{ p.s. lorsque } m \to \infty.$$

L'algorithme précédent de contrôle par normalité peut calculer séquentiellement ces deux estimateurs de la variance en plus du test de Shapiro-Wilk. On obtient ainsi un indicateur graphique de la stabilisation de la variance, complémentaire de l'indicateur de normalité. Un exemple d'école pour le cas discret fini est donné dans [RT5].

3.1.2 Cas général

L'extension naturelle de l'idée précédente au cas général consiste à choisir une collection de parties $A_r \in \mathcal{B}(E)$, $1 \leq r \leq p$, et à appliquer cette méthode de contrôle aux fonctions $h_r = \mathbb{I}_{A_r}$. Cependant, la technique présentée au § 3.1.1 nécessite quelques aménagements pour être utilisable dans le cas où E est dénombrable ou continu. En effet, il serait illusoire d'espérer atteindre une normalité approchée pour des fonctions telles que $\mathbb{I}_A$ lorsque A est situé dans une queue de π. Il faudrait un très grand nombre d'itérations pour obtenir suffisamment de visites des queues, donc une méthode trop conservative et sans amélioration réelle quant à la précision de l'estimation de $\mathbb{E}_\pi(h)$. D'autre part, toujours dans un souci d'applicabilité, nous avons voulu notre méthode aussi générique que possible, et notamment utilisable "en aveugle" en l'absence de connaissances fines du support de π et de la localisation de ses modes.

Remarquons que cette technique n'est pas limitée aux chaînes de Markov : elle s'étend aux processus ergodiques vérifiant le TLC. Ceci est utile dans la mesure où il sera plus commode en pratique de l'appliquer aux marginales de la chaîne étudiée si celle-ci est de grande dimension, plutôt que de construire son analogue multidimensionnel.

Contrôle automatique par normalité

Le principe que nous avons proposé dans [9] consiste à choisir un compact $\mathcal{A}$ de E appelé "région de contrôle", à réaliser une partition $\mathcal{A} = \bigcup_{r=1}^{p} A_r$ de ce compact, et à appliquer le contrôle par normalité aux fonctions indicatrices $h_r = \mathbb{I}_{A_r}$. Afin d'éviter de contrôler des parties situées dans les queues de π, nous calculons au cours du temps les probabilités empiriques sur les chaînes parallèles,

$$\hat{\mathbb{P}}_n(A_r) = \frac{1}{m} \sum_{\ell=1}^{m} \frac{1}{n} S_n^{(\ell)}(h_r),$$

et éliminons les parties de trop faible probabilité, i.e. telles que $\hat{\mathbb{P}}_n(A_r) < \varepsilon$ où ε est à choisir et peut éventuellement dépendre de n.

Si l'on note $C(n)$ l'ensemble des fonctions $h_r = \mathbb{I}_{A_r}$ pour lesquelles on contrôle la normalité à l'instant n, initialisé par $C(0)$ contenant toutes les fonctions pour $r = 1, \ldots, p$, l'algorithme peut être décrit formellement par :

1. `Simuler les` m `chaînes de` n_{k-1} `à` n_k
2. `Pour` $r \in C(n_{k-1})$ `mettre à jour` $\left(\dfrac{S_{n_k}^{(1)}(h_r)}{\sqrt{n_k}}, \ldots, \dfrac{S_{n_k}^{(m)}(h_r)}{\sqrt{n_k}} \right)$
3. `Pour` $r \in C(n_{k-1})$ `calculer` $\hat{\mathbb{P}}_{n_k}(A_r)$;
   `mettre à jour` $C(n_k) = \left\{ h_r \in C(n_{k-1}) : \hat{\mathbb{P}}_{n_k}(A_r) \geq \varepsilon(n_k) \right\}$

4. Pour $h_r \in C(n_k)$:
   calculer $SW(h_r, n_k)$;
   si H_0 acceptée, $C(n_k) \leftarrow C(n_k) \setminus \{h_r\}$
5. Si $C(n_k) = \emptyset$, fin.
   sinon $k \leftarrow k+1$ et aller en 1.

L'application de cet algorithme nécessite de déterminer les paramètres de réglage $(\mathcal{A}, p, \varepsilon)$ (choix de la région de contrôle, de la finesse de la partition et du seuil d'élimination des parties contrôlées). L'obtention d'un bon choix peut se faire assez facilement par essais successifs, dans la mesure où l'algorithme retourne en sortie la probabilité empirique de la région choisie $\hat{\mathbb{P}}_n(\mathcal{A})$, et celle $\hat{\mathbb{P}}_n(\mathcal{A}_C)$ de la "région effectivement contrôlée", somme des probabilités empiriques des parties A_r sur lesquelles la normalité a été acceptée. Il est souhaitable que ces estimateurs soient proches de 1. En effet, une valeur de $\hat{\mathbb{P}}_n(\mathcal{A})$ trop faible indique un mauvais choix du compact de travail par rapport aux régions d'intérêt de π (des chaînes se sont "échappées" de $\mathcal{A}$ pendant un temps significatif). De même, une valeur de $\hat{\mathbb{P}}_n(\mathcal{A}_C)$ trop faible indique un choix trop élevé de ε qui a conduit à éliminer des parties de probabilité non négligeables. Une fois réglés ces paramètres, l'algorithme retourne le temps nécessaire à l'obtention de la normalité dans les parties contrôlées, et des intervalles de confiance (IC) pour les $\pi(A_r)$ contrôlés construits à partir de cette hypothèse gaussienne. On obtient ainsi un histogramme de la loi stationnaire avec contrôle des fluctuations.

Stabilisation de la variance limite

L'estimateur de la variance après n pas $\sigma_n^2(h)$ à l'aide de la variance empirique reste calculable dans le cas général, mais les calculs algébriques à la base de l'estimation de la variance limite ne sont plus applicables. Nous avons proposé une solution empirique consistant à discrétiser la chaîne de Markov et à calculer l'analogue des matrices définies dans 3.1.1. L'emploi d'une telle discrétisation (déjà utilisée dans le contrôle binaire de Raftery et Lewis,1992) n'est pas valide théoriquement, car le processus discrétisé n'est en général plus une chaîne de Markov. La stabilisation de $\sigma_n^2(h)$ en fonction de n reste donc un indicateur, mais l'estimateur variance limite de la chaîne discrétisée est à considérer avec précaution. Il aurait été possible de discrétiser la chaîne d'une façon théoriquement correcte (Guihenneuc-Jouyaux et Robert, 1998), mais cette technique aurait nécessité la détermination d'*ensembles petits* associés à la chaîne, ce qui aurait compromis l'aspect générique de notre méthode.

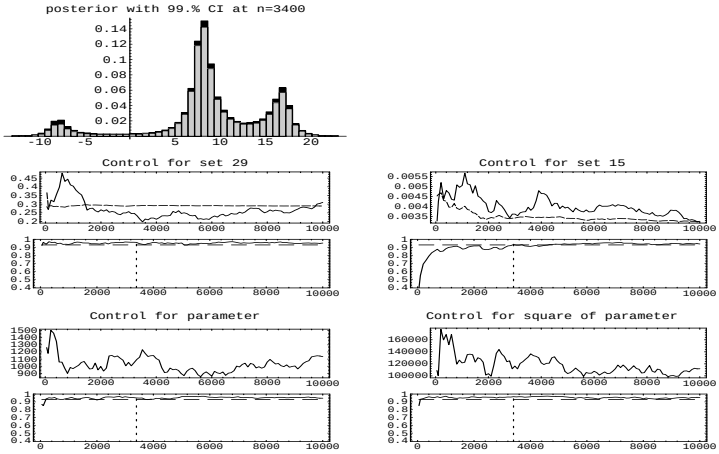
Un exemple

L'algorithme de contrôle dans le cas général est disponible en ligne (voir [L1]). De nombreux exemples d'école et comparaisons avec des méthodes

alternatives dans le cas où E est fini et dans le cas général sont donnés dans [RT5] et [9]. Une application en vraie grandeur figure dans [7], et concerne un algorithme de Gibbs pour un modèle de chaîne de Markov cachée identifiant les régions homogènes de la séquence de l'ADN. Tous ces exemples utilisent la boîte à outil [L1] grâce à son caractère générique. Ils illustrent la simplicité avec laquelle il est possible de déterminer les paramètres de réglage $(\mathcal{A}, p, \varepsilon)$ sans connaissance de la densité cible. Ils montrent aussi la pertinence de la méthode, notamment sa sensibilité aux lois multimodales avec modes distants, donc faiblement mélangeantes et qui demandent plus d'itérations pour parvenir à la normalité approchée.

Nous donnons simplement ici un exemple de sorties de [L1] sur un cas d'école : un échantillonneur de Gibbs tiré de Robert (1996, p.226), concernant l'inférence bayésienne pour le paramètre de localisation d'une loi de Cauchy $\mathcal{C}(\theta, 1)$. L'intérêt est la multimodalité de la loi a posteriori, et la présence d'un mode distant et de faible masse. Nous avons déterminé en quelques essais une région $\mathcal{A}$ convenable de probabilité estimée à 99.7%, et les choix $p = 50$ et $\varepsilon = 0.002$ ont conduit à $\hat{\mathbb{P}}_n(\mathcal{A}_C) = 99\%$ de "masse contrôlée par normalité". Bien entendu, les choix de p et ε sont liés à la précision que l'on souhaite dans la reconstruction de π. La figure 3.1 donne la loi a posteriori empirique avec les IC obtenus à l'instant d'atteinte de la normalité, ainsi que les graphiques de contrôle pour deux des parties contrôlées, et pour les fonctions supplémentaires $h(\theta) = \theta$ et $h(\theta) = \theta^2$. Il est clair que l'atteinte de la normalité a demandé plus de temps (3400 itérations) pour les parties de faible masse situées entre le mode distant et les modes principaux.

FIG. 3.1 – Contrôle par normalité pour le modèle de Cauchy. 1ère ligne, loi a posteriori avec IC (*en noir*). 2ème ligne, graphiques de contrôle pour les $\mathbb{I}_{A_r}$ ayant atteint la normalité le plus rapidement (*gauche*) et le plus lentement (*droite*). 3ème ligne, contrôle pour $h(\theta) = \theta$ et $h(\theta) = \theta^2$. Chaque graphique de contrôle représente la stabilisation de $\sigma_n^2(h)$ et de la variance limite sur la chaîne discrétisée (*haut*), et la statistique SW avec son seuil de rejet (*bas*).



3.2 Estimation de la variance limite des chaînes de Markov

Dans [15], nous développons l'étude théorique d'un estimateur de la variance limite $\sigma^2(h)$ définie en (3.3), uniquement basé comme précédemment sur les réalisations issues de chaînes i.i.d., mais permettant le contrôle des fluctuations de la variance après n pas dans le cas général. Il ne s'agit donc pas à proprement parler d'une méthode de contrôle de convergence MCMC, mais l'un des objectifs visés est de fournir un outil supplémentaire utilisable pour ce contrôle. En effet, le TLC dont nous testons la validité au § 3.1 n'est utilisable que si nous disposons aussi d'un estimateur fiable pour la variance, et les méthodes proposées dans la littérature n'apportent pas — à notre connaissance — une réponse satisfaisante (voir, e.g., Robert 1996).

3.2.1 Processus variance empirique

Nous notons ici encore $X = (X_n, n \geq 0)$ la chaîne de Markov d'intérêt, X^ℓ, $1 \leq \ell \leq m$ les m copies i.i.d. de X, et $S_n^{(\ell)}(h)$ les sommes relatives à une fonction d'intérêt $h : E \to \mathbb{R}$. L'estimateur naturel de $\sigma_n^2(h)$ est la variance empirique du m-échantillon de sommes normalisées, noté

$$\widehat{\sigma}_{n,m}^2(h) = \frac{1}{m} \sum_{\ell=1}^{m} \left(\frac{1}{\sqrt{n}} S_n^{(\ell)}(h) - \frac{1}{\sqrt{n}} \overline{S_n(h)} \right)^2, \quad \overline{S_n(h)} = \frac{1}{m} \sum_{\ell=1}^{m} S_n^{(\ell)}(h).$$

Ici, nous souhaitons notamment établir des bandes de confiance pour les fluctuations de cet estimateur en évaluant, pour $n_1 < n_2$,

$$\mathbb{P}\left[\sup_{n_1\leq n\leq n_2}\left|\widehat{\sigma}^2_{n,m}(h)-\sigma^2(h)\right|\geq u\right], \quad u>0, \tag{3.8}$$

où $\sigma^2(h)$, inconnu, est aussi estimé. Nous sommes donc amenés à étudier le comportement asymptotique d'un processus à temps continu interpolé des sommes partielles $\widehat{\sigma}^2_{[nt],m}(h)$ associées à $\widehat{\sigma}^2_{n,m}(h)$, où $[\cdot]$ désigne la partie entière (voir Billingsley, 1968). Il est naturel alors de considérer plutôt les processus des sommes partielles associés aux $S_n^{(\ell)}(h)$, car ceux-ci vérifient des TLC fonctionnels sous des hypothèses classiques sur X (voir Meyn et Tweedie, 1993).

Ces processus se définissent de la façon suivante : Notons π^i la loi de X à l'instant i, et $\pi^i(h) = \int h\, d\pi^i$. Nous considérons le cas non stationnaire seul réaliste pour les applications, i.e. $X_0 \sim \mu$ loi initiale arbitraire. Alors, le processus des sommes partielles associé à $S_n^{(\ell)}(h)$ correctement centré est

$$\frac{1}{\sqrt{n}}S^{(\ell)}_{[nt]}(\bar{h}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]}\left(h(X_i^\ell)-\pi^i(h)\right), \quad t\in[0,T].$$

Notons $Y_n^{(\ell)}(t)$ son interpolation linéaire :

$$Y_n^{(\ell)}(t) = \frac{1}{\sqrt{n}}\left[S^{(\ell)}_{[nt]}(\bar{h}) + (nt-[nt])\left(S^{(\ell)}_{[nt]+1}(\bar{h}) - S^{(\ell)}_{[nt]}(\bar{h})\right)\right],$$

en omettant la dépendance à la fonction h qui est fixée. Chaque terme de centrage $\pi^i(h)$, inconnu, peut être estimé à partir des chaînes i.i.d. à l'instant i par

$$\widehat{\pi^i}_m(h) = \frac{1}{m}\sum_{\ell=1}^{m} h(X_i^\ell),$$

de sorte que le processus des sommes partielles fonction des observations est

$$\begin{aligned}\frac{1}{\sqrt{n}}\widehat{S}^{(\ell)}_{[nt],m}(\bar{h}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]}\left(h(X_i^\ell)-\widehat{\pi^i}_m(h)\right)\\ &= \frac{1}{\sqrt{n}}S^{(\ell)}_{[nt]}(\bar{h}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]}\left(\pi^i(h)-\widehat{\pi^i}_m(h)\right). \end{aligned}\tag{3.9}$$

Le terme $\widehat{\pi^i}_m(h)$ peut être vu comme une fonction de la réalisation à l'instant i de la *chaîne produit* sur E^m,

$$\mathbf{X}_i = (X_i^1,\ldots,X_i^m).$$

On notera $\widehat{\pi^i}_m(h) = \mathbf{H}_m(\mathbf{X}_i)$ cette fonction dont l'espérance relativement à $(\pi^i)^{\otimes m}$ est $\mathbb{E}[\mathbf{H}_m(\mathbf{X}_i)] = \pi^i(h)$, si bien que le terme de droite dans (3.9) peut être vu comme un processus de sommes partielles sur la chaîne produit,

$$\sum_{i=1}^{[nt]} \Big(\widehat{\pi^i}_m(h) - \pi^i(h)\Big) = \sum_{i=1}^{[nt]} (\mathbf{H}_m(\mathbf{X}_i) - \mathbb{E}[\mathbf{H}_m(\mathbf{X}_i)]) = S_{[nt]}(\bar{\mathbf{H}}_m), \quad (3.10)$$

dont nous notons l'interpolation linéaire $\mathbf{Y}_{n,m}(t)$ (en omettant ici aussi la dépendance à h). L'interpolation du processus des sommes partielles (3.9) est donc $Y_n^{(\ell)}(t) - \mathbf{Y}_{n,m}(t)$. Finalement, nous définissons le *processus variance empirique* comme

$$V_{n,m}(t) = \frac{1}{m} \sum_{\ell=1}^{m} \Big(Y_n^{(\ell)}(t) - \mathbf{Y}_{n,m}(t)\Big)^2. \quad (3.11)$$

Remarquons que les $Y_n^{(\ell)}$ et $\mathbf{Y}_{n,m}$ ne sont pas indépendants. D'autre part, $V_{n,m}(t)$ coïncide bien avec $\widehat{\sigma}^2_{[nt],m}(h)$ aux points $t = i/n$, $i = 1, \ldots, [nT]$, mais est une interpolation non linéaire des sommes partielles associées à la variance empirique après n pas $\widehat{\sigma}^2_{n,m}(h)$.

3.2.2 Propriétés de stabilité pour la chaîne produit

L'étude du processus $V_{n,m}$ nécessite l'usage d'un TLC fonctionnel sur la chaîne de Markov produit $\mathbf{X}$ d'ordre m ; or les hypothèses classiques, par exemple dans le contexte des algorithmes MCMC, portent sur la stabilité de la chaîne simple X. En préalable, nous avons donc étudié dans [RT8] le transfert au produit des conditions usuelles de stabilité d'une chaîne de Markov, telles que les *conditions de dérive* (*drift*) données dans Meyn et Tweedie (1993), la Harris récurrence et l'ergodicité géométrique. Certains auteurs ont déjà eu besoin de ce type de propriétés (par exemple, Roberts et Tweedie 1999, 2001, étudient le transfert au produit d'une condition de dérive géométrique pour $m = 2$ afin de coupler deux chaînes), mais nous n'avons pas trouvé dans la littérature de résultats généraux sur le transfert de conditions de stabilité vers une chaîne produit d'ordre $m > 2$. Cette partie 3.2.2 qui est développée dans [RT8] est donc indépendante de l'étude du processus variance $V_{n,m}$, même si certains résultats seront utilisés dans la suite.

Considérons la chaîne m-produit $\mathbf{X} = (X^1, \ldots, X^m)$ sur $\mathbf{E} = E^m$, de noyau $\mathbf{P}(\mathbf{x}, d\mathbf{y}) = \prod_{i=1}^m P(x_i, dy_i)$. Les conditions de dérive données dans Meyn et Tweedie (1993) utilisent un *ensemble petit* (*small set*) $C \in \mathcal{B}(E)$ et une fonction de dérive (fonction de Lyapounov) $V : E \to [0, \infty]$ de sorte que la chaîne X satisfasse une condition d'attraction vers C, dont par exemple la plus faible est $PV(x) \leq V(x)$ pour $x \in C^c$ (voir Meyn et Tweedie 1993). Il est raisonnable de définir ces objets sur $\mathbf{E}$ uniquement à partir des (C, V)

dont on suppose disposer pour la chaîne initiale. Un choix naturel et maniable est alors

$$\mathbf{C} = C^{\times m}, \qquad \mathbf{V}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} V(x_i).$$

Définissons aussi les ensembles suivants :

$$C(V,r) \triangleq \{x : V(x) \leq r\}, \qquad \mathbf{C}(V,r) = C(V,r)^{\times m}. \tag{3.12}$$

Meyn et Tweedie (1993) utilisent pour ces conditions de stabilité des fonctions V telles que les ensembles $C(V,r)$ soient *petite*, ce qui est un léger affaiblissement des ensembles petits précédents. Or cette propriété ne passe pas au produit ; nous avons donc utilisé la définition un peu plus forte suivante :

Définition 1 *Une fonction mesurable $V : E \to [0,\infty]$ est UOSS*[2] *pour X si V est finie en un point de E et si, pour tout $r < \infty$, $C(V,r)$ est un ensemble petit dès qu'il est non vide.*

L'intérêt de cette propriété est que V UOSS implique $\mathbf{V}$ UOSS.

Transfert des conditions de dérive

Nous étudions tout d'abord le passage au produit des quatre conditions de dérive (V1)–(V4) répertoriées dans Meyn et Tweedie (1993), p.501 (ces conditions donnent ensuite des conditions suffisantes de stabilité de force croissante). Les deux premières ne se transmettent pas à la chaîne produit (voir [RT8]). Les deux plus forte conditions de dérive passent elles au produit pour m quelconque. Nous rappelons ici les résultats obtenus, qui nécessitent un renforcement peu coûteux des conditions de départ (essentiellement le passage de *petite set* à *small set*).

Proposition 2 *Si la chaîne de Markov X vérifie la condition (V3)*

$$PV(x) \leq V(x) - f(x) + b\mathbb{I}_C(x), \quad x \in E,$$

avec $V : E \to [0,\infty]$ et $f : E \to [1,\infty)$ UOSS, V finie en un point de E, $0 < b < \infty$ et C petit, alors $\mathbf{X}$ vérifie la condition (V3) suivante :

$$\mathbf{P}\tilde{\mathbf{V}}(\mathbf{x}) \leq \tilde{\mathbf{V}}(\mathbf{x}) - \mathbf{f}(\mathbf{x}) + \frac{b-\eta}{1-\eta}\mathbb{I}_{\mathbf{C}(f,\, mb/\eta - m + 1)}(\mathbf{x}), \quad \mathbf{x} \in \mathbf{E}, \tag{3.13}$$

où $\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{m} f(x_i)/m$, $0 < \eta < \min\{1, b\}$ et $\tilde{\mathbf{V}} = (1-\eta)^{-1}\mathbf{V}$.

Voici l'analogue pour la condition de dérive géométrique, la plus forte.

[2] "Unbounded off small sets", par analogie avec la terminologie originale *unbounded off petite sets* employée par Meyn et Tweedie (1993).

Proposition 3 *Si la chaîne de Markov X vérifie la condition (V4)*

$$PV(x) \leq \lambda V(x) + b\mathbb{I}_C(x), \quad x \in E,$$

avec $V : E \to [1, \infty]$ UOSS, $0 < \lambda < 1$, $0 \leq b < \infty$ et C petit, alors $\mathbf{X}$ vérifie la condition (V4) suivante :

$$\mathbf{PV}(\mathbf{x}) \leq (\lambda + \eta)\mathbf{V}(\mathbf{x}) + (b - \eta)\mathbb{I}_{\mathbf{C}(V, mb/\eta - m + 1)}(\mathbf{x}), \tag{3.14}$$

où $0 < \eta < \min\{1 - \lambda, b\}$.

Remarquons qu'une dégradation de la mélangeance avec m apparaît au travers de l'ensemble petit d'attraction dans (3.13) et (3.14), sauf dans le cas où $\lambda + b < 1$, car alors $\mathbf{PV} \leq (\lambda + b)\mathbf{V}$ pour tout m.

Transfert de la Harris récurrence et de l'ergodicité

Nous avons déterminé des conditions minimales assurant la Harris récurrence de la chaîne produit. Ces conditions nécessitent au préalable l'irréductibilité de $\mathbf{X}$. Or la ψ-irréductibilité (car E est général) de X ne se transmet pas au produit à cause de la possible dépendance de n à x et A dans la condition $P^n(x, A) > 0$ donnée par Meyn et Tweedie (1993), p. 87. Nous prenons donc comme hypothèse la *ψ-irréductibilité forte* de X, c'est-à-dire l'existence d'un entier n tel que, pour tout x et tout $A \in \mathcal{B}(E)$, $P^n(x, A) > 0$ dès que $\psi(A) > 0$.

Proposition 4 *Si la chaîne de Markov X est fortement ψ-irréductible et vérifie (V3) au sens de la proposition 2, alors $\mathbf{X}$ est Harris récurrente.*

Remarquons que nous pourrons utiliser ce résultat dans l'étude de la variance empirique sans exiger la ψ-irréductibilité forte : en effet dans le contexte des algorithmes MCMC X est toujours positive, ce qui implique notamment que $\mathbf{X}$ est $\pi^{\otimes m}$-irréductible.

En ce qui concerne le transfert au produit des propriétés d'ergodicité de X, il faut d'abord s'assurer que l'apériodicité passe au produit, ce qui est direct ([RT8], lemme 2). Les deux notions d'ergodicité utilisées dans Meyn et Tweedie (1993) sont la f-ergodicité et l'ergodicité géométrique. Toutes deux passent au produit à l'aide des conditions de dérive (V3) et (V4). Nous donnons ici les deux formes les plus simples :

Théorème 1 *Soit X une chaîne de Markov fortement ψ-irreducible et apériodique, et $\mathbf{X}$ la chaîne produit d'ordre m.*

(i) Si X est f-ergodique avec f UOSS, alors X vérifie (V3) et $\mathbf{X}$ est $\mathbf{f}$-ergodique sur un ensemble plein et absorbant, où $\mathbf{f}(\mathbf{x}) = \sum_{i=1}^m f(x_i)/m$.

(ii) Si X vérifie (V4) au sens de la proposition 3 avec $V \geq 1$ UOSS et fini partout, alors $\mathbf{X}$ est géométriquement ergodique.

Enfin, nous avons également établi quelques propriétés formelles de transfert de stabilité d'une chaîne vers un système de particules en interactions, décrit par un noyau de la forme $\mathbf{P}_\Theta(\mathbf{x}, d\mathbf{y}) = \prod_{i=1}^m P_{\theta_i(\mathbf{x})}(x_i, dy_i)$, où $\theta_i(\mathbf{x}) \in \Theta$ pour $\mathbf{x} \in \mathbf{E}$ représente la fonction de couplage des chaînes (voir [RT8]).

3.2.3 Comportement asymptotique du processus variance

Revenons à l'étude du processus variance empirique $V_{n,m}$. Nous étudions tout d'abord la convergence lorsque le temps $n \to \infty$ de chacun des processus $Y_n^{(\ell)}$ et $\mathbf{Y}_{n,m}$ qui interviennent dans (3.11). Les conditions sous lesquelles $Y_n^{(\ell)}$ converge vers un mouvement Brownien sont données dans Meyn et Tweedie (1993) et donnent le (i) ci-dessous. Nos conditions sont en effet similaires, avec le petit renforcement sur la condition de dérive décrit en 3.2.2 qui donne (ii) grâce à la proposition 4, et permet d'obtenir le TLC fonctionnel pour la chaîne produit :

Proposition 5 *Supposons satisfaites les conditions suivantes :*

(C1) La chaîne de Markov X est Harris récurrente positive, et une solution $\widehat{h}$ de l'équation de Poisson $\widehat{h} - P\widehat{h} = h - \pi(h)$ existe avec $\pi(\widehat{h}^2) < \infty$ et $\sigma^2(h) > 0$.

(C2) X vérifie la condition de dérive (V3) de la proposition 2.

Alors, pour tout $m > 0$:

(i) $Y_n^{(\ell)} \xrightarrow{d} \sigma(h)W^{(\ell)}$ lorsque $n \to \infty$, $\ell = 1, \ldots, m$, où $W^{(1)}, \ldots, W^{(m)}$ sont m copies indépendantes du mouvement Brownien.

(ii) La chaîne produit $\mathbf{X}$ est Harris récurrente positive, de probabilité invariante $\underline{\pi} = \pi^{\otimes m}$;

(iii) Pour $\mathbf{H}_m$, il existe une solution $\widehat{\mathbf{H}}$ de l'équation de Poisson

$$\widehat{\mathbf{H}} - \mathbf{P}\widehat{\mathbf{H}} = \mathbf{H}_m - \underline{\pi}(\mathbf{H}_m),$$

et $\underline{\pi}(\widehat{\mathbf{H}}^2) < \infty$;

(iv) $\sigma^2(\mathbf{H}_m) = \sigma^2(h)/m$, et

$$\mathbf{Y}_{n,m} \xrightarrow{d} \frac{1}{\sqrt{m}}\sigma(h)\widetilde{W} \quad \text{lorsque } n \to \infty,$$

où $\widetilde{W}$ est un mouvement Brownien sur $[0, T]$.

En conséquence, $V_{n,m} \xrightarrow{d} Z_m$ lorsque $n \to \infty$, où le processus limite est

$$Z_m(t) = \frac{1}{m}\sum_{\ell=1}^m \left(\sigma(h)W_t^{(\ell)}\right)^2 + \frac{1}{m}\sigma^2(h)\widetilde{W}_t^2 - \frac{2\sigma^2(h)}{m}\widetilde{W}_t B_t,$$

où $B_t = m^{-1/2}\sum_{\ell=1}^m W_t^{(\ell)}$. Le premier terme peut être centré et écrit

$$\frac{1}{\sqrt{m}}\left[\frac{1}{\sqrt{m}}\sum_{\ell=1}^m \left(\left(\sigma(h)W_t^{(\ell)}\right)^2 - \mathbb{E}[(\sigma(h)W_t)^2]\right)\right] + \sigma^2(h)t,$$

car $\mathbb{E}[(\sigma(h)W_t)^2] = \sigma^2(h)t$.

On étudie le comportement de ce terme pour l'asymptotique en nombre de chaînes i.i.d., i.e. lorsque $m \to \infty$. Plus généralement, pour une fonction ψ convenable, nous étudions le comportement d'une somme normalisée de processus de Itô i.i.d.,

$$\xi_m(t) = \frac{1}{\sqrt{m}} \sum_{\ell=1}^{m} \left[\psi(W_t^{(\ell)}) - \mathbb{E}[\psi(W_t)] \right], \quad t \in [0, T].$$

Théorème 2 *Si ψ est deux fois continûment différentiable et vérifie :*

$$\int_{-\infty}^{+\infty} \psi'^2(u) e^{-u^2/2T}\, du < \infty, \quad \textit{et} \quad \int_{-\infty}^{+\infty} \psi''^2(u) e^{-u^2/2T}\, du < \infty,$$

alors

$$\xi_m \xrightarrow{d} G_\psi \quad \textit{lorsque } m \to \infty,$$

où G_ψ est un processus gaussien centré.

La preuve (détaillée dans [15]) se fait en décomposant ξ_m et en étudiant séparément sa partie martingale, dont la tension se montre par le théorème de Rebolledo (voir, e.g., Dacunha-Castelle et Duflo, 1986), et sa partie processus qui utilise un critère classique de tension (Billingsley, 1968).

Dans notre application, $\psi(x) = \sigma^2(h)\, x^2$ et G_ψ admet une représentation sous la forme du processus gaussien W_a de fonction de covariance $(s, t) \mapsto a(s \wedge t)$ avec $a(t) = 2\sigma^4(h)\, t^2$.

3.2.4 Estimation de la variance limite

Le processus limite en n est donc de la forme

$$Z_m(t) = t\sigma^2(h) + \frac{1}{\sqrt{m}} \xi_m(t) + \frac{\sigma^2(h)}{m} \left(\widetilde{W}_t^2 - 2\widetilde{W}_t B_t \right).$$

En appliquant ce qui précède lorsque n et m sont assez grand, et en négligeant le terme en $\mathcal{O}(1/m)$, nous utilisons l'approximation

$$\widehat{\sigma}_h^2(t) = \frac{V_{n,m}(t)}{t} \approx \sigma^2(h) + \frac{1}{\sqrt{m}} \frac{W_{a(t)}}{t}.$$

Notons que la validité de cette approximation (et donc notamment la normalité) doit être vérifiée, au moins par la méthode de contrôle par normalité présentée en 3.1. Par un changement d'échelle, le processus $W_{a(t)}/t$ s'interprète comme un processus de Ornstein-Uhlenbeck, ce qui permet d'utiliser les résultats disponibles dans la littérature sur le supremum de tels processus sur un compact du temps (Delong, 1981) afin de contrôler les fluctuations dans (3.8).

D'autre part, il est possible de construire un estimateur de la variance utilisant les observations pondérées en plusieurs instants $\mathbf{t} = (t_1, \ldots, t_p)$ de ce processus,

$$\widehat{\sigma}_h^2(\mathbf{w}, \mathbf{t}) = \sum_{i=1}^{p} w_i \frac{V_{n,m}(t_i)}{t_i}, \quad \mathbf{w} \in (0,1)^p, \quad \sum_{i=1}^{p} w_i = 1,$$

afin de réduire la variance de l'estimateur. En choisissant bien les instants $t_1, \ldots, t_p$, la variance de l'estimateur à une forme calculable et que nous pouvons optimiser en $\mathbf{w}$. Nous en déduisons quelques schémas simples d'estimation construits sur seulement 7 ou 9 instants qui permettent des réductions de variance relative de l'ordre de 1/2.

3.3 Contrôle de la stabilité d'une chaîne de Markov par l'entropie

Comme nous l'avons dit en introduction, ce travail n'est pas dans sa forme actuelle directement applicable au contrôle de convergence MCMC. Il s'agit ici de contrôler la stabilité d'une chaîne de Markov qui peut être non homogène, à l'aide d'un outil statistique n'utilisant, encore une fois, que les réalisations issues de chaînes i.i.d.

L'hypothèse de base est que le noyau de la chaîne d'intérêt a une densité par rapport à une mesure dominante, et que cette densité est analytiquement accessible. La motivation immédiate de ce travail est l'étude exploratoire de systèmes markoviens de noyaux connus mais trop compliqués pour que l'on puisse déterminer leur propriétés de stabilité par les outils théoriques classiques tels que l'établissement de conditions de dérive avec fonctions de Lyapounov comme au § 3.2.2. Le principe que nous proposons dans [14] est de lancer de deux positions initiales distinctes deux groupes de chaînes i.i.d. évoluant avec le même noyau, et d'estimer au cours du temps l'information de Kullback entre les lois des chaînes de chaque groupe, calculée à l'aide d'une sorte de double intégration de Monte Carlo.

Notons $X = (X_t, t \geq 0)$ et $Y = (Y_t, t \geq 0)$ deux processus de Markov à temps discret de même noyau (non nécessairement homogène) de densité q^t à l'instant t, et de lois initiales différentes p_0^0 et p_1^0 (ou de positions initiales respectives x_0 et x_1). Les densités des lois de chaque processus sont notées respectivement p_0^t et p_1^t, et sont données par

$$p_i^{t+1}(y) = \int_E p_i^t(x) q^t(x,y)\, \nu(dx), \quad i = 0, 1.$$

Nous proposons d'estimer et de représenter graphiquement l'information de Kullback $t \to \mathcal{K}(p_i^t, p_j^t) = \mathbb{E}_{p_i^t}[\log(p_i^t)] - \mathbb{E}_{p_i^t}[\log(p_j^t)]$ pour $i \neq j \in \{0, 1\}$. En effet, la rapidité de la décroissance vers zéro, ou au contraire l'explosion de cet indicateur est représentatif de la stabilité du processus (vitesse d'oubli de la condition initiale). Pour cela, nous introduisons des estimateurs de l'entropie $\mathbb{E}_{p_i^t}[\log(p_i^t)]$ et de l'entropie *externe* $\mathbb{E}_{p_i^t}[\log(p_j^t)]$, dont nous montrons la consistance et la normalité asymptotique. Des estimateurs non paramétriques de l'entropie ont déjà été proposés dans d'autres contextes (voir, e.g., Eggermont et LaRiccia, 1999, ou Tsybakov et Van Der Meulen, 1994), mais la normalité asymptotique n'est alors montrée que dans le cas univarié. D'autre part, aucun estimateur de l'entropie externe n'a été proposé à notre connaissance.

3.3.1 Un estimateur par double Monte Carlo de l'entropie

Le problème étant symétrique en p_0 et p_1, il suffit de traiter par exemple l'estimation de $\mathcal{K}(p_1^t, p_0^t)$, autrement dit des entropies

$$\mathcal{H}_i^t = \mathbb{E}_{p_1^t}[\log(p_i^t)], \quad i = 0, 1.$$

Nous supposons disposer d'observations issues de m copies i.i.d. de X et de Y, notées comme précédemment $(X^1, \ldots, X^m)$ et $(Y^1, \ldots, Y^m)$. Puisque l'expression de q est connue, si la LFGN s'applique (ce que nous supposons), alors pour tout $y \in E$ et $t \geq 0$, l'intégration par Monte Carlo de q construite sur les m v.a. i.i.d. $\mathbf{X}_t = (X_t^1, \ldots, X_t^m)$ de loi p_0^t vérifie

$$\frac{1}{m}\sum_{k=1}^{m} q^t(X_t^k, y) \xrightarrow{\text{p.s.}} \int_E q^t(x, y) p_0^t(x)\, \nu(dx) = p_0^{t+1}(y), \quad m \to \infty. \quad (3.15)$$

Nous pouvons donc espérer que l'intégration par Monte Carlo du logarithme de la partie gauche de (3.15), construite sur les m v.a. i.i.d. de loi p_1^{t+1}, $\mathbf{Y}_{t+1} = (Y_{t+1}^1, \ldots, Y_{t+1}^m)$, approche $\mathcal{H}_0^{t+1}$. Nous introduisons donc l'estimateur de type "double Monte Carlo" construit sur $(\mathbf{X}_t, \mathbf{Y}_{t+1})$ suivant :

$$\widehat{\mathcal{H}}_0^{t+1} = \frac{1}{m}\sum_{\ell=1}^{m} \log\left(\frac{1}{m}\sum_{k=1}^{m} q^t(X_t^k, Y_{t+1}^\ell)\right).$$

Pour définir l'estimateur de $\mathcal{H}_1^{t+1}$ avec la même logique, il se pose un problème car il faudrait utiliser les échantillons $(\mathbf{Y}_t, \mathbf{Y}_{t+1})$. Or ceux-ci ne sont pas indépendants, ce qui est utile pour l'étude asymptotique. Pour conserver la même simplicité et traiter de la même façon $\mathcal{H}_0^{t+1}$ et $\mathcal{H}_1^{t+1}$, nous avons préféré bâtir notre estimateur sur la simulation d'un second m-échantillon de copies de Y indépendant du premier, et que nous notons $\tilde{\mathbf{Y}}_t$ à l'instant t. L'estimateur construit sur $(\mathbf{Y}_t, \tilde{\mathbf{Y}}_{t+1})$ est ainsi

$$\widehat{\mathcal{H}}_1^{t+1} = \frac{1}{m}\sum_{\ell=1}^{m} \log\left(\frac{1}{m}\sum_{k=1}^{m} q^t(Y_t^k, \tilde{Y}_{t+1}^\ell)\right).$$

Nous obtenons sous des conditions de moment le résultat suivant :

Théorème 3 *Si, pour tout $t \geq 0$ et $i = 0, 1$, le noyau normalisé*

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

est non-dégénéré et vérifie

$$\mathbb{E}_{p_i^t \otimes p_1^{t+1}}\left[|r_i^t(X, Y)|^{2+\gamma}\right] < \infty \quad \textit{pour } \gamma > 0, \quad (3.16)$$

et si

$$\mathbb{E}_{p_1^{t+1}}[|\log p_i^{t+1}(Y)|^2] < \infty, \tag{3.17}$$

alors, pour $i = 0, 1$ *:*

$$\begin{aligned} \widehat{\mathcal{H}}_i^{t+1} &\xrightarrow{\mathbb{P}} \mathcal{H}_i^{t+1}, \quad \textit{lorsque } m \to \infty, \\ \sqrt{m}(\widehat{\mathcal{H}}_i^{t+1} - \mathcal{H}_i^{t+1}) &\xrightarrow{d} \mathcal{N}(0, \Sigma_i), \quad \textit{lorsque } m \to \infty, \end{aligned}$$

où $\Sigma_i = var_{p_1^{t+1}}[\log p_i^{t+1}] + var_{p_i^t}[R_i(X)]$*, et* $R_i(x) = \mathbb{E}_{p_1^{t+1}}[r_i^t(x, Y)]$*.*

La preuve utilise une décomposition inspirée de Del Moral et Guionnet (1999). Les techniques utilisées ensuite sont la mise en évidence d'une U-statistique pour traiter l'un des termes (voir, e.g., Serfling, 1980), et l'utilisation d'une inégalité de moyenne déviation due à Fuk et Nagaev (1971, 1976). Sous une condition de moment plus forte, nous montrons la consistance forte à l'aide d'une technique empruntée à Del Moral et Guionnet(1999). Une différence par rapport à ces auteurs est que notre emploi là aussi d'une inégalité de moyenne déviation nous permet de relaxer leur condition de moment de 6 à $4 + \gamma$:

Théorème 4 *Sous les conditions du théorème 3, si l'on remplace (3.16) par :*

$$\mathbb{E}_{p_i^t \otimes p_1^{t+1}}\left[|r_i^t(X, Y)|^{4+\gamma}\right] < \infty \quad \textit{pour } \gamma > 0, \quad i = 0, 1$$

alors $\widehat{\mathcal{H}}_i^{t+1} \xrightarrow{p.s.} \mathcal{H}_i^{t+1}$*, lorsque* $m \to \infty$*.*

3.3.2 Un exemple

Afin d'illustrer la pertinence de notre indicateur de stabilité, nous avons choisi de simuler un AR(1) gaussien

$$X_t = \rho X_{t-1} + \varepsilon_t,$$

où $(\varepsilon_t)_{t \geq 0}$ est une suite i.i.d. de $\mathcal{N}(0, \sigma^2)$, et où nous ne simulerons que des cas pour lesquels $\rho > 0$. L'intérêt de ce modèle est qu'il est facilement calculable et permet d'obtenir des processus stables pour $\rho \in (0, 1)$, instable lorsque $\rho = 1$ et explosifs pour $\rho > 1$. Le noyau est lui-même gaussien, de densité $q(x, y) = \phi_{\sigma^2}(y - \rho x)$, où $\phi_{\sigma^2}(\cdot)$ est la densité de la loi $\mathcal{N}(0, \sigma^2)$, et la loi à l'instant t de chacun des processus est

$$X_t \sim \mathcal{N}\left(\rho^t x_0\,,\, \frac{1 - \rho^{2t}}{1 - \rho^2}\sigma^2\right), \quad Y_t \sim \mathcal{N}\left(\rho^t x_1\,,\, \frac{1 - \rho^{2t}}{1 - \rho^2}\sigma^2\right), \quad \rho \neq 1.$$

La condition (3.17) est facile à vérifier, et on montre par un calcul direct que la condition plus difficile (3.16) est vérifiée pour $\rho \in (0, 1)$ (donc dans le cas stable). Nous avons néanmoins utilisé notre estimateur dans toutes les

configurations, car il demeure un indicateur empirique de l'instabilité ou de l'explosion même lorsque la condition (3.16) n'est pas vérifiée.

Nous avons représenté l'estimation de $\mathcal{K}(p_1^t, p_0^t)$ contre sa vraie valeur calculable ici par intégration numérique via *Mathematica*. Pour tous les modèles la variance a été fixée à $\sigma^2 = 4$. Le cas stable a été testé dans un cas très mélangeant ($\rho = 0.7$) et faiblement mélangeant ($\rho = 0.99$) afin de montrer la pertinence de notre indicateur (fig. 3.2), même avec peu de chaînes parallèles. Le cas instable a été testé avec plus de chaînes (jusqu'à $m = 300$). En effet, les loi respectives p_0^t et p_1^t sont de moyennes fixées mais de variance $(t-1)\sigma^2$ croissante avec t. Il faut donc plus de chaînes pour "couvrir" ce support qui grandit (fig. 3.3). Enfin, dans le cas explosif, les moyennes elles-mêmes s'écartent vers $-\infty$ et $+\infty$ lorsque $t \to \infty$. Notre estimateur est encore correct pour des valeurs de ρ pas trop grandes et suffisamment de chaînes (fig. 3.4). Il indique clairement le caractère explosif du modèle.

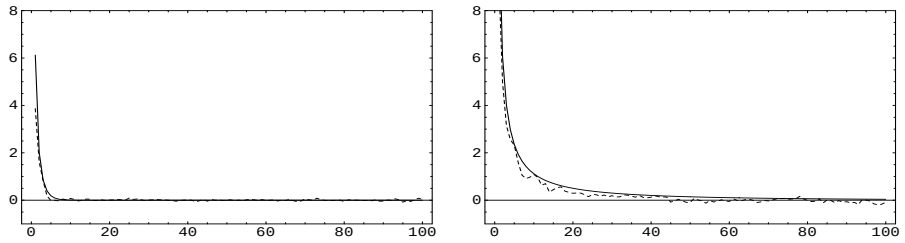


FIG. 3.2 – Vraie $\mathcal{K}(p_1^t, p_0^t)$ (trait plein) et estimation. Cas stable, conditions initiales $x_0 = -5$ et $x_1 = 5$. *Gauche* : $\rho = 0.7$ et $m = 30$; *Droite* : $\rho = 0.99$ et $m = 50$.

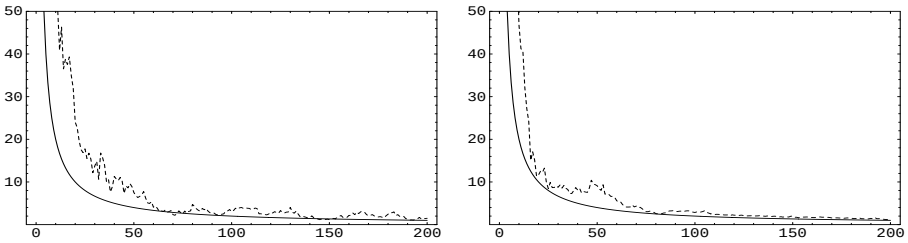


FIG. 3.3 – Cas instable $\rho = 1$ avec les conditions initiales $x_0 = -20$ et $x_1 = 20$. *Gauche* : $m = 50$; *Droite* : $m = 300$.

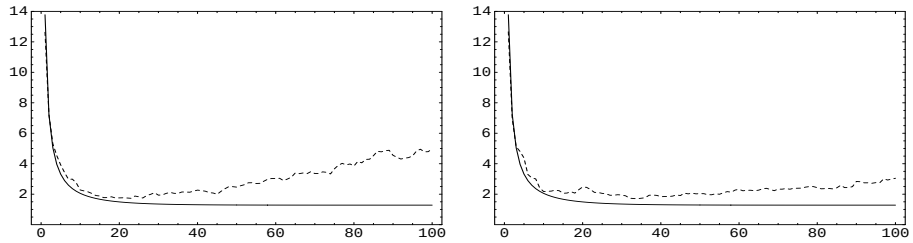


FIG. 3.4 – Cas explosif $\rho = 1.05$ avec les conditions initiales $x_0 = -5$ et $x_1 = 5$. *Gauche* : $m = 300$; *Droite* : $m = 600$.

Chapitre 4

Algorithmes MCMC adaptatifs

Ce chapitre est consacré au développement de méthodes MCMC adaptatives utilisant la dynamique de l'algorithme de Hastings-Metropolis. La motivation de ce travail est la reconstruction de lois cibles complexes responsables de chaînes faiblement mélangeantes, telles que par exemple les lois multimodales avec des modes éloignés et de grandes régions de peu de masse. En effet, dans de tels cas, les méthodes classiques basées sur l'algorithme de Hastings-Metropolis ou l'échantillonneur de Gibbs peuvent mettre énormément de temps à découvrir et explorer correctement toutes les régions modales (nous allons préciser pourquoi ci-dessous). L'idée de base des méthodes adaptatives est d'utiliser l'information sur la loi cible déjà glanée au cours des itérations précédentes de la chaîne, afin de diriger l'exploration future dans le but d'améliorer la vitesse de convergence.

4.1 L'algorithme de Hastings-Metropolis

Nous présentons d'abord l'algorithme de Hastings-Metropolis (HM), dont la dynamique sera à la base de nos méthodes adaptatives. Cet algorithme très général nécessite la connaissance de la densité f de la loi cible π à une constante multiplicative près, et est donc tout à fait adapté à la reconstruction des lois a posteriori des modèles bayésiens. Chaque déplacement est basé sur la génération, à partir de la position courante x, d'une valeur *candidate* y à partir d'une *loi instrumentale* qui est une densité conditionnelle $q(y|x)$ "presque quelconque" (mais facile à simuler). Un mécanisme d'acceptation-rejet fait que la chaîne se déplace ou reste sur place avec probabilité positive. Le pas $x_n \to x_{n+1}$ est donné par :

1. `simuler` $y \sim q(\cdot\,|x_n)$
2. `calculer` $\alpha(x_n, y) = \min\left\{1, \dfrac{f(y)q(x_n|y)}{f(x_n)q(y|x_n)}\right\}$

3. `prendre` $x_{n+1} = \begin{cases} y & \texttt{avec probabilité } \alpha(x_n, y), \\ x_n & \texttt{avec probabilité } 1 - \alpha(x_n, y). \end{cases}$

Algorithme de Hastings-Metropolis à marche aléatoire

C'est la version de l'algorithme de HM probablement la plus employée en raison de sa simplicité de mise en œuvre. La chaîne de Markov associée à q est une marche aléatoire, i.e. $y_{n+1} = x_n + \varepsilon_{n+1}$, ε étant une perturbation aléatoire de loi g, indépendante de x_n, 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 (la symétrie fait que le taux d'acceptation se réduit alors à $\alpha(x, y) = \min\{1, f(y)/f(x)\}$).

Dans tous les cas, il est nécessaire de calibrer l'algorithme en choisissant le paramètre d'échelle de la loi utilisée. En effet, de ce paramètre dépendra crucialement la vitesse d'exploration du support de π par l'algorithme, donc finalement la vitesse de convergence (voir Gilks *et al.*, 1996). En effet, une loi q générant de trop petits sauts donnera un fort taux d'acceptation, mais restera longtemps (ou pour toujours) dans le bassin d'attraction d'un seul mode si les autres sont trop éloignés. Dans l'autre sens, une loi q générant de trop grands sauts tentera très souvent de visiter les queues de π, et entraînera un trop faible taux d'acceptation. Des auteurs conseillent donc de calibrer la variance de sorte que le taux d'acceptation empirique ne soit ni trop petit ni trop grand (la valeur 0.23 a même été recommandée). Nous verrons que cette recommandation ne s'applique pas forcément, notamment dans le cas de modes distants, car il faut de toutes façon que la chaîne tente de visiter de larges régions pour espérer découvrir les modes. Le taux d'acceptation induit par le "bon calibrage", disponible dans le cas de simulations, est alors inévitablement faible.

Le point capital est en fait que dans le cas de lois complexes, ce bon calibrage est difficile à déterminer, et nécessite en fait déjà une bonne connaissance du paysage induit par la loi π cherchée. L'une de nos motivations est de donner une réponse à ce type de problème par une méthode "aveugle".

Algorithme de Hastings-Metropolis indépendant

C'est l'autre version de l'algorithme de HM très utilisée. Ici, la loi instrumentale est indépendante de la position courante de l'algorithme, $q(y|x) = q(y)$, ce qui autorise des déplacements très libres par rapport à la version marche aléatoire. En revanche, la performance de cet algorithme est liée à la bonne qualité de la loi instrumentale q, qui doit proposer de visiter souvent les régions d'intérêt de π.

Dans ce cadre, Mengersen et Tweedie (1996) donnent un résultat d'ergodicité géométrique uniforme (équivalent à la condition de Doeblin) : Si $q(\cdot) > 0$ sur le support de π, et que il existe $a \in (0, 1)$ tel que $q(x) > af(x)$

pour tout $x \in E$, alors

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

Ce résultat montre en particulier que plus q "ressemble" à π, (donc a proche de 1), plus la convergence est rapide. Ce résultat a été amélioré par Holden (1998), qui utilise la norme relative

$$\|p^n - f\|_f \triangleq \sup_{x\in\Omega} \left| \frac{p^n(x) - f(x)}{f(x)} \right|. \tag{4.1}$$

Holden montre sous la même condition de minoration $q \geq af$ la convergence de la densité p^n de l'algorithme à l'instant n au sens de cette norme :

$$\|p^n - f\|_f \leq D(1-a)^n, \quad \text{où } D = ||p^0 - f||_f. \tag{4.2}$$

Méthodes adaptatives

Nos méthodes adaptatives utilisent fortement ce résultat, et sont fondées sur l'idée de base suivante : Dans le cadre de l'algorithme de HM indépendant, supposons que nous disposons d'une loi instrumentale q^0 vérifiant $q^0 \geq a_0 f$ et assurant donc la convergence géométrique (4.2) avec une vitesse déterminée par a_0. Si l'on veut améliorer la vitesse, il faut améliorer a_0, et donc q^0. Il est alors naturel de chercher à exploiter la connaissance de f dont on dispose au travers de la densité p^n de l'algorithme à l'instant n, grâce au fait que $p^n \to f$ au sens de (4.2). Si l'on pouvait, par exemple, remplacer directement au cours du temps q^0 par les densités successives p^n, on obtiendrait le schéma d'apprentissage idéal décrit dans la fig. 4.1, avec les améliorations extrêmement rapides des constantes de minorations $a_1, a_2, \ldots$, associées à la vitesse dans (4.2).

Les densités p^n sont évidemment inconnues, mais si l'on dispose de m chaînes i.i.d. lancées suivant l'algorithme de HM indépendant de loi q^0, il est possible d'estimer p^n (de façon non-paramétrique) à partir des réalisations à l'instant n des m chaînes. Malheureusement, dès le premier instant d'apprentissage n_1, la construction de l'estimateur de p^{n_1} crée un *couplage* des chaînes, qui perdent leur indépendance et leur caractère markovien, ce qui rend difficile l'étude théorique de ces processus. Au § 4.2, nous contournons cette difficulté en faisant en sorte de ne travailler que sur des chaînes i.i.d. grâce à un artifice de simulation. Au § 4.3 nous reprenons l'idée de base avec une étude directe de la structure de dépendance des processus couplés.

D'autres auteurs ont proposé des méthodes MCMC adaptatives dans la littérature. Par exemple, Gelfland et Sahu (1994) ont suggéré d'utiliser une phase adaptative durant une période de temps fini, puis de lancer un algorithme MCMC usuel . Gilks et Roberts dans Gilks *et al.* (1996, chap. 6), proposent d'étendre l'*adaptive direction sampling* à la dynamique de HM.

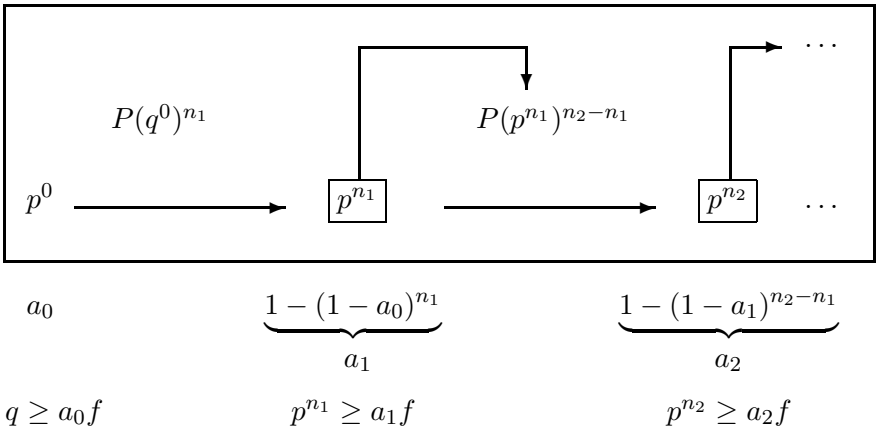


FIG. 4.1 – Schéma idéal d'apprentissage en ligne aux instants $n_1, n_2, \ldots$, où $P(q)^n$ est le n-itéré du noyau de HM de loi instrumentale q. *2ème ligne* : constantes de minoration associées à l'emploi de la densité de l'algorithme comme loi instrumentale. *3ème ligne* : conditions de minorations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

Cependant, à notre connaissance, il n'a pas été établi de résultats asymptotiques prouvant que ces méthodes faisaient mieux que leurs contreparties classiques. Les algorithmes que nous présentons sont construits sur des chaînes parallèles, et peuvent être envisagés dans un cadre asymptotique (en temps et en nombre de chaînes). Ce cadre permet de donner des résultats théoriques satisfaisants, desquels on peut espérer le bon comportement à distance finie des algorithmes utilisés en pratique.

4.2 Un algorithme de Hastings-Metropolis avec apprentissage

Dans [8] et [10], nous proposons une solution consistant à utiliser un estimateur par histogramme de la densité de l'algorithme. La convergence est obtenue au prix d'un schéma théorique consistant à éliminer, aux instants d'apprentissage, les chaînes utilisées pour construire les mises-à-jour des lois instrumentales qui sont injectées dans les autres chaînes. Ceci permet de préserver l'indépendance et le caractère markovien (non-homogène) des chaînes restantes, et ainsi d'utiliser des résultats classiques telles que des inégalités exponentielles sur l'histogramme dans le cadre i.i.d. Les inconvénients majeurs de cette technique sont une implémentation difficile (histogramme en grandes dimensions et méthode parallèle non standard à cause des chaînes à éliminer), et un temps de calcul pouvant être important en raisons du grand nombre de chaînes à lancer au départ à cause du procédé d'élimination.

Nous supposons ici que la densité cible f est C-Lipschitzienne à support compact, et minorée. Ceci est restrictif mais indispensable pour utiliser des

résultats de convergence p.s. uniforme sur E. Dans la pratique ce n'est pas très gênant, cette méthode servant essentiellement à construire une "bonne" loi instrumentale (i.e. une loi qui localise bien la masse) sur un compact aussi grand que nécessaire, comprenant les régions d'intérêt de π. Soit donc f la densité cible strictement positive sur $E \subset \mathbb{R}^s$ et minorée par une constante $\alpha > 0$. On note aussi $A = \sup_{x \in E} f(x)$.

Pour l'étude asymptotique, nous supposons disposer d'une infinité de copies i.i.d. d'un processus de Hastings-Metropolis inhomogène défini pour une suite de lois instrumentales de densités q^n. Afin d'alléger les notations, nous considérons que l'apprentissage se fait à tout instant (ce ne sera pas le cas en pratique). L'apprentissage à l'instant n utilise $m = m(n)$ copies qui sont empruntées à cet ensemble infini, et sont ensuite éliminées. La densité p^n est estimée par l'histogramme H_m construit sur les réalisations de cet ensemble de $m(n)$ chaînes (voir, e.g., Bosq et Lecoutre, 1987, chap. 6 pour la définition et les propriétés de l'histogramme). Afin d'assurer la consistance de l'estimateur, nous exigerons $m(n) \to \infty$ avec n pour un régime qui sera précisé. La loi instrumentale q^n est soit H_m, soit une modification légère de H_m, consistant à rendre toutes les classes strictement positives, de sorte que la condition de minoration $q^n \geq a_n f$ soit satisfaite pour une constante $a_n \in (0, 1)$.

4.2.1 Convergence et amélioration apportée par l'algorithme

Nous montrons tout d'abord dans ce cadre la convergence des marginales au sens de la norme $||p^n - f||_f$ ([10], proposition 2), par une application de la technique empruntée à Holden (1998) dans notre situation non-homogène. Nous montrons ensuite sous des conditions techniques précisées ci-dessous une inégalité exponentielle à distance finie pour l'histogramme H_m basé sur m réalisations i.i.d. de p^n. Ce résultat utilise notamment une inégalité exponentielle pour la loi multinomiale (Bosq et Lecoutre 1987, p. 174), qui exige dans notre situation que la fenêtre h_m ne tende pas trop vite vers 0 (condition (4.3)), et que m et n soient assez grands.

Proposition 6 *Soit $H_m = H_{m(n)}$ l'histogramme de p^n, h_m sa fenêtre, et $\varepsilon > 0$. Posons $\delta_{m,n} = 2A\left(1 - 1/(Amh_m^s)\right)^n ||p^0 - f||_f + \sqrt{s} h_m C$.*
Si $h_m \to 0$, $mh_m^s \to \infty$ lorsque $n \to \infty$, $mh_m^s = o(n)$, et

$$mh_m^{3s} \geq (20/(\varepsilon - \delta_{m,n})^2) \quad \textit{pour } m > m_0, n > n_0, \tag{4.3}$$

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

$$\mathbb{P}\left(\sup_{x \in E} |H_m(x) - p^n(x)| > \varepsilon\right) \leq 3 \exp\left(-mh_m^{2s}(\varepsilon - \delta_{m,n})^2/25\right). \tag{4.4}$$

Nous montrons enfin que l'algorithme avec apprentissage converge plus rapidement vers f, en n, que tout algorithme de HM homogène usuel utilisant une loi instrumentale arbitraire q^0 satisfaisant $q^0 \geq a_0 f$.

Le résultat ci-dessous exprime le fait que l'on n'utilisera pas infiniment souvent une loi instrumentale "moins bonne" que q^0, c'est à dire associée à une condition de minoration $q^n \geq a_n f$ avec $a_n < a_0$. Il conduit à calibrer le régime $m(n)$ afin de montrer par un lemme de Borel-Cantelli fondé sur (4.4) que les événements "indésirables" ne peuvent survenir un nombre infini de fois. Une manière plus concise d'exprimer ce résultat est d'introduire l'instant associé à la suite de constantes de minoration (aléatoires) (a_n),

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

instant après lequel tout algorithme de HM indépendant utilisant la loi instrumentale q^t pour $t > T(a_0)$ est plus rapide que l'algorithme initial.

Théorème 5 *Si $m(n)$ et $h_{m(n)}$ vérifient les conditions de la proposition 6, et*

$$m(n) h_{m(n)}^{2s} \geq c \log(n), \tag{4.5}$$

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

Nous avons également proposé dans [10] une extension de ce résultat au cas général $E = \mathbb{R}^s$, en appliquant notre méthode sur un compact et en supposant disposer d'une approximation fine des queues de f hors de ce compact.

Mise en œuvre

Il peut être difficile en pratique de construire l'approximation sur les queues utilisée pour traiter le cas général. Nous avons donc plutôt fondé l'utilisation de notre méthode sur la détermination d'un compact assez grand contenant les régions d'intérêt de π, de sorte que la masse hors de ce compact soit négligeable du point de vue de l'inférence faite sur π. La détermination d'un tel compact est bien plus simple que celle des positions des modes (nécessaire pour calibrer l'algorithme de HM à marche aléatoire), et peut se faire par essais successifs à l'aide de méthodes telles que l'outil de contrôle MCMC présenté au § 3.1.2 qui informe l'utilisateur de la pertinence du "compact de travail $\mathcal{A}$".

D'autre part, il est impossible en pratique d'effectuer des apprentissages tout au long de la simulation, ceci à cause du procédé d'élimination qui est le vrai inconvénient de cette méthode. Une implémentation de cet algorithme approchant la situation théorique est possible en construisant l'apprentissage sur un nombre a priori fixé de chaînes i.i.d., divisé en k paquets. A l'instant n_i, $i = 1, \ldots, k$, on utilise le i-ème paquet de $m(n_i)$ chaînes pour construire $H_{m(n_i)}$, puis l'on n'utilise plus ces chaînes afin de préserver l'indépendance

sur les chaînes restantes. L'algorithme inhomogène utilise donc des lois instrumentales qui apprennent f de mieux en mieux, et il devient homogène de loi instrumentale q^{n_k} après n_k. Un autre paquet (éventuellement réduit à une seule chaîne) ayant subi toutes les mutations de lois instrumentales mais n'ayant jamais servi à la construction de ces lois peut alors être exécuté jusqu'à la fin de la simulation. Les chaînes de ce paquet ont en principe une dynamique de HM plus rapide que celle associée à q^0.

J'ai ici aussi écrit un logiciel générique implémentant la méthode, et nous l'avons testé sur des exemples simulés ([10]), et sur un vrai modèle bayésien (§ 4.2.2 ci-dessous). Cet algorithme semble particulièrement adapté aux situations dans lesquelles la loi π est multimodale, car les lois instrumentales proposées favorisent rapidement les sauts entre modes déjà "découverts", accélérant ainsi l'exploration du domaine d'intérêt.

4.2.2 Application à l'analyse bayésienne du modèle Logit

Dans [11], nous avons comparé l'algorithme de HM avec apprentissage ci-dessus contre un algorithme de HM à marche aléatoire avec une loi instrumentale calibrée par une méthode *ad hoc*, dans le cadre de l'analyse bayésienne d'un modèle de régression non linéaire généralisée, le modèle Logit. Il s'agit donc de l'étude par simulation d'une situation réelle dans le cas où le paramètre est à valeur dans $\mathbb{R}^2$.

Pour la version de HM à marche aléatoire, le principe est de déterminer une "bonne" loi instrumentale à partir d'une approximation gaussienne bidimensionnelle construite sur les données (voir Altaleb, 1999). Il s'agit d'un algorithme à chaîne unique. Pour la version adaptative, un compact englobant largement la région d'intérêt est facile à trouver, la loi a posteriori étant assez simple ici (unimodale). Nous avons utilisé une phase d'apprentissage assez courte, et donné à la marche aléatoire une phase d'échauffement (*burn-in*) du même nombre total d'itérations, afin de rendre les simulations comparables. Nous avons enfin effectué les comparaisons à l'aide du logiciel publique de diagnostic de convergence MCMC `CODA` (Best *et al.*, 1995).

Dans cette situation réelle mais assez simple, on constate une bien meilleure capacité d'exploration du support pour la chaîne produite par l'algorithme adaptatif, à l'issue des étapes d'apprentissage. Ceci indique une meilleure mélangeance que pour la chaîne issue de l'autre stratégie, et est confirmé par cinq diagnostics de convergence (voir [11]), dont nous ne donnons ici que le plus intuitif : le déplacement sur la loi a posteriori (fig. 4.2).

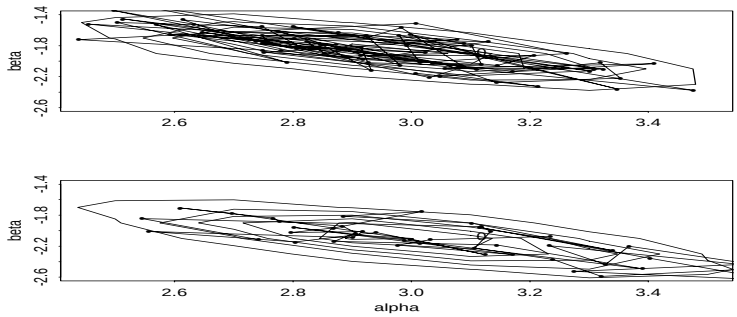


FIG. 4.2 – Déplacement sur les contours de la loi a posteriori pour le modèle Logit (100 itérations). HM adaptatif (*haut*) ; HM avec approximation gaussienne (*bas*).

4.3 Algorithmes de Hastings-Metropolis en interaction

Récemment, nous avons proposé dans [12] et [16] une amélioration de la méthode adaptative présentée au § 4.2. L'apprentissage est fondé sur la même idée, mais la méthode est bien plus élégante car correspondant à ce que l'on souhaite faire en pratique. Surtout, elle ne présente pas les inconvénients précédents : l'apprentissage peut très bien se faire tout au long de la simulation ; la mise en œuvre pratique est ainsi plus proche du cadre asymptotique ; enfin, l'implémentation est plus simple (et générique).

Un seul ensemble de m chaînes est utilisé au cours du temps, et les densités successives sont "apprises" par des estimateurs à noyau (plus faciles à implémenter). Le système observé est donc composé de m processus qui ne sont plus ni markoviens ni indépendants, puisque à chaque instant de mutation une loi instrumentale est construite, qui dépend de toutes les autres réalisations, et dont dépendent les pas de HM suivants. Le principe général est illustré fig. 4.3. Une propriété intéressante pour les perspectives qu'elle ouvre est que ce système présente des analogies avec les systèmes de particules en interactions (voir, e.g., Del Moral et Miclo, 2000).

Nous étudions la structure de dépendance du système et donnons son comportement asymptotique lorsque le temps (n) et le nombre de processus en parallèle (m) tendent vers l'infini de façon contrôlée. Nous supposons ici encore que la densité cible f est à support compact, et minorée. Comme nous l'avons dit précédemment, ce n'est pas très gênant en pratique pourvu que l'on puisse déterminer un compact aussi grand que nécessaire.

Soit $E \subset \mathbb{R}^s$, et f la densité cible vérifiant $f(x) \geq \alpha > 0$ sur E. Nous notons $n_1 < \cdots < n_k < \cdots$ les instants d'apprentissage, que nous supposons suivre un schéma $n_{k+1} = n_k + d$, où $d = d(n_k)$ sera précisé. Le processus vectoriel global est noté $\mathbf{X} = (\mathbf{X}_n, n \geq 0)$, et ses m composantes à l'instant n

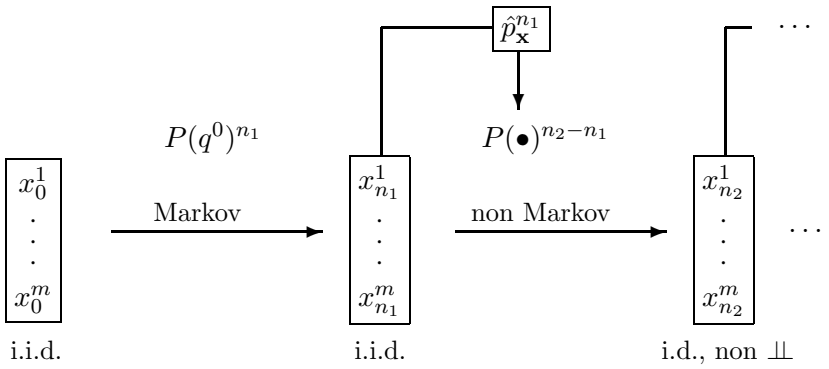


FIG. 4.3 – Principe général des algorithmes de HM en interactions : Les lois instrumentales sont construites à partir d'estimateurs $\hat{p}_{\mathbf{x}}^{n_k}$ des lois successives p^{n_k}.

sont notées $\mathbf{X}_n = (X_n^1, \ldots, X_n^m)$ comme au chapitre 3[1]. De même, nous notons $\mathbf{x}_n = (x_n^1, \ldots, x_n^m)$ le vecteur des m observations de ce processus à l'instant n. Nous aurons besoin aussi du vecteur des $(m-1)$ observations parallèles x_n^j, pour $j \neq i$ à l'instant n, que nous noterons $\mathbf{x}_n^{\neq i}$. Enfin, la loi initiale commune à toutes les chaînes est π^0.

L'algorithme global "IHM" (*Interacting Hastings-Metropolis*) est décrit par les étapes S_1–S_3 ci-dessous qui donnent l'évolution de la i-ème composante ($i = 1, ..., m$). Pour une raison technique, il est indispensable que chaque composante i n'utilise que les réalisations des autres processus $j \neq i$ pour construire ses lois instrumentales aux instants n_k. Cette condition permet de conserver la propriété de π-invariance des noyaux de HM, car chaque processus utilise alors un noyau de ce type dépendant (fonctionnellement) de variables *externes* (voir [RT10], § 2.2 et lemme 1).

Algorithme IHM

(S_1) Pour une valeur initiale $x_0^i \sim \pi^0$, et pour $0 \leq n < n_1$, le pas $x_n^i \to x_{n+1}^i$ est une étape de HM indépendant de loi instrumentale q^0.

(S_2) A l'instant n_k, $k \geq 1$, une loi instrumentale est construite à partir des observations $\mathbf{x}_{n_k}^{\neq i}$ de $\mathbf{X}^{\neq i}$ (i.e. sans l'observation $x_{n_k}^i$ elle-même) ; on la note $q_{\mathbf{x}_{n_k}^{\neq i}}$.

(S_3) Pour $n_k \leq n < n_{k+1}$, les itérations $x_n^i \to x_{n+1}^i$ sont des pas de HM indépendant de loi instrumentale $q_{\mathbf{x}_{n_k}^{\neq i}}$, donnés par :

1. `simuler` $y^i \sim q_{\mathbf{x}_{n_k}^{\neq i}}(\cdot)$

[1]Les notations sont inversées par rapport à [16] et [RT10]. J'ai fait ce choix afin de préserver des notations consistantes tout au long du présent document.

$$\texttt{2. calculer } \alpha_{\mathbf{x}_{n_k}^{\neq i}}(x_n^i, y^i) = \min\left\{1, \frac{f(y^i) q_{\mathbf{x}_{n_k}^{\neq i}}(x_n^i)}{f(x_n^i) q_{\mathbf{x}_{n_k}^{\neq i}}(y^i)}\right\}$$

$$\texttt{3. prendre } x_{n+1}^i = \begin{cases} y^i & \texttt{avec probabilité } \alpha_{\mathbf{x}_{n_k}^{\neq i}}(x_n^i, y^i), \\ x_n^i & \texttt{avec probabilité } 1 - \alpha_{\mathbf{x}_{n_k}^{\neq i}}(x_n^i, y^i). \end{cases}$$

Remarques :

(a) les composantes X^i, $i = 1, \ldots, m$ entre n_k et n_{k+1} sont markoviennes homogènes et indépendantes conditionnellement à $\mathbf{X}_{n_k} = \mathbf{x}_{n_k}$;

(b) si $(X_0^1, \ldots, X_0^m)$ sont indépendants et que les lois instrumentales $q_{\mathbf{x}_{n_k}^{\neq i}}(\cdot)$ sont symétriques en les $(\mathbf{x}_{n_k}^{\neq i})$, alors il est facile de vérifier que, pour tout $n \geq 1$, $\mathbf{X}_n$ est un vecteur aléatoire *échangeable*.

4.3.1 Convergence des processus marginaux

Nous montrons tout d'abord la convergence géométrique des marginales vers la loi cible, au sens de la norme $||\cdot||_f$ définie en (4.1), sous une condition de minoration uniforme pour les lois instrumentales construites aux instants d'apprentissage. Cette condition peut sembler restrictive, mais on verra qu'elle est satisfaite pour les lois que nous construisons.

Proposition 7 *Supposons qu'il existe une constante $a^* \in (0,1)$ telle que*

$$\forall k \geq 0, \quad \forall \mathbf{x}_{n_k}^{\neq i} \in E^{m-1}, \quad \forall x \in E, \quad q_{\mathbf{x}_{n_k}^{\neq i}}(x) \geq a^* f(x). \tag{4.6}$$

Alors la densité de la loi d'un processus marginal issu de (S_1–S_3) vérifie,

$$\left\|p^{n_k+r} - f\right\|_f \leq C_0\, \rho^{n_k+r}, \qquad k \geq 1, \quad 1 \leq r \leq n_{k+1} - n_k, \tag{4.7}$$

où $\rho = (1 - a^*)$ et $C_0 = \|p^0 - f\|_f$.

Ce résultat est basé sur le fait que, conditionnellement à l'événement $\{\mathbf{X}_{n_k}^{\neq i} = \mathbf{x}_{n_k}^{\neq i}\}$, le i-ème processus itère un noyau de HM homogène dépendant de variables *externes* grâce à la suppression de l'observation de la marginale i lors de la construction de la loi instrumentale. Il est alors possible d'appliquer (4.2) à la densité conditionnelle de la loi de cette marginale, puis de déconditionner. Il suffit ensuite d'itérer cette procédure jusqu'à n_1.

4.3.2 Structure de dépendance et estimateur à noyau

Pour utiliser un estimateur à noyau sur données dépendantes, nous devons préciser la structure de dépendance des v.a. échangeables $X_n^1, \ldots, X_n^m$. Ce processus n'est pas mélangeant au sens classique puisque il n'y a pas de

notion de futur ou de passé en m à tout instant n fixé. L'analogue du *coefficient de dépendance forte* (voir, e.g., Bosq, 1996) dont nous aurons besoin, devient ici

$$\begin{aligned}\alpha(X^i, \mathbf{X}^{\neq i}) &= \sup_{\substack{B\in\mathcal{B}(E)\\ C\in\mathcal{B}(E^{m-1})}} \left|\mathbb{P}(X^i \in B \cap \mathbf{X}^{\neq i} \in C) - \mathbb{P}(X^i \in B)\mathbb{P}(\mathbf{X}^{\neq i} \in C)\right| \\ &= \alpha(\mathbf{X}).\end{aligned}$$

Nous le notons simplement $\alpha(\mathbf{X})$ car il ne dépend pas de i par l'échangeabilité. La proposition suivante précise "l'oubli" des processus en fonction du temps passé depuis le dernier apprentissage. Ce résultat est une conséquence de la convergence géométrique des marginales, qui induit de bonnes propriétés de mélangeance après les instants de couplage.

Proposition 8 *Si l'hypothèse (4.6) est satisfaite, alors pour $k \geq 1$ et $1 \leq r \leq n_{k+1} - n_k$,*

$$\alpha(\mathbf{X}^{n_k+r}) \leq C\rho^r, \quad \textit{avec } C \textit{ constante positive.}$$

Nous considérons ensuite l'estimation par noyau d'une densité "générique" p à partir d'observations $\mathbf{X} = (X^1, \ldots, X^m)$ dont la dépendance est décrite par $\alpha(\mathbf{X})$, en oubliant pour l'instant l'aspect temporel. L'estimateur à noyau classique est

$$p_m(x) = \frac{1}{mh_m^s} \sum_{i=1}^{m} K\left(\frac{x - X^i}{h_m}\right), \quad x \in \mathbb{R}^s, \tag{4.8}$$

où h_m est la fenêtre et K un noyau borné symétrique et Lipschitzien, vérifiant les conditions $\lim_{||x||\to\infty} ||x||^s K(x) = 0$, et $\int ||u||^2 K(x)\,dx < \infty$. En adaptant l'approche de Bosq (1996), théorème 1, à notre situation de dépendance, nous obtenons tout d'abord une inégalité exponentielle pour la somme de v.a. échangeables centrées à partir du lemme de Bradley utilisant le coefficient $\alpha(\mathbf{X})$. En appliquant cette inégalité aux v.a. $Y^i = h_m^{-s} K((x - X^i)/h_m)$ convenablement centrées, et en utilisant une technique de chaînage (Bosq, 1996, p. 48), on déduit l'inégalité exponentielle suivante :

Proposition 9 *Soit $\mathbf{X} = (X^1, ..., X^m)$ m v.a. échangeables de loi p, continue sur E. Pour tout $\varepsilon > 0$, il existe $m_0(\varepsilon)$ tel que, pour $m > m_0(\varepsilon)$:*

$$\begin{aligned}\mathbb{P}\left(||p_m - \mathbb{E}(p_m)|| > \varepsilon\right) &\leq c_3 h_m^{-s(s+1)} (\log m)^s \\ &\quad \times \left[2\exp(-c_1 m h_m^{2s}) + c_2 m h_m^{-s/2} \alpha(\mathbf{X})\right],\end{aligned}$$

où $||p|| = \sup_{x\in E} |p(x)|$, et $c_1 = c_1(\varepsilon, K)$, $c_2 = c_2(\varepsilon, K)$, $c_3 = c_3(K, E)$ sont des constantes positives.

4.3.3 Amélioration de la vitesse de convergence géométrique

Nous supposons que l'on observe $m+1$ processus, afin de construire les estimateurs sur m observations et simplifier ainsi les notations. Aux instants d'apprentissage n_k, la densité p^{n_k} est estimée par $p_m^{n_k}$ donné par (4.8) et construite à l'aide des m observations $\mathbf{X}_{n_k}^{\neq i}$ pour le i-ème processus. Comme $p_m^{n_k}$ ne satisfait pas directement la condition (4.6), on prend pour loi instrumentale son mélange avec une loi uniforme sur E de faible poids ($\lambda \approx 1$),

$$q_{\mathbf{x}_{n_k}^{\neq i}}(x) = \lambda p_m^{n_k}(x) + (1-\lambda)\frac{1}{|E|}, \quad x \in E. \tag{4.9}$$

Alors il existe $b \in (0,1)$ tel que $|E|^{-1} \geq bf(x)$, $\forall x \in E$, et la condition (4.6) est satisfaite avec $a^* = (1-\lambda)b$. L'étude asymptotique nécessite que m et n tendent tous deux vers l'infini avec un régime $m = m(n)$ à calibrer, afin d'utiliser la consistance de $p_m^{n_k}$ et la convergence de p^n vers f. Dans ce cadre, la meilleure vitesse possible asymptotiquement (lorsque $p_m^{n_k} = f$) est associée à la constante de minoration

$$a_{opt} = \lambda + (1-\lambda)b.$$

Soit alors, comme dans la méthode précédente (§ 4.2.1)

$$T(a_0) = \inf\{n_k, k \geq 1 : \forall n \geq n_k, a_n > a_0\},$$

l'instant après lequel une marginale est meilleure que tout algorithme de HM arbitraire associé à a_0. Pour un choix déterministe de constantes a_{n_k} vérifiant $a_{n_k} > a_0$ pour n assez grand, on montre que les événements indésirables du type "la condition de minoration n'est pas vérifiée pour a_{n_k}" ne peuvent survenir qu'un nombre fini de fois. Ces événements s'expriment comme des déviations contrôlables grâce à la proposition 9 et à la convergence géométrique de p^n vers f. Les vitesses (4.10) et (4.11) précisées ci-dessous permettent d'appliquer un lemme de Borel-Cantelli d'où le résultat :

Théorème 6 *Sous les hypothèses précédentes, et si $m(n)$, $h_{m(n)}$ et $d(n)$ vérifient les conditions*

$$c_1 m h_m^{2s} - \log\left(c_2 h_m^{-s(s+1)} (\log m)^s\right) \geq (1+\gamma)\log(n), \tag{4.10}$$

$$d(n) \geq c_3 \log\left[\frac{(1+\gamma)nm(\log m)^s}{h_m^{s(s+3/2)}}\right] \tag{4.11}$$

où c_1, c_2, c_3 constantes positives et $\gamma > 0$, alors pour $a_0 < a_{opt}$,

$$\mathbb{P}(T(a_0) < \infty) = 1.$$

Remarquons qu'il est facile de donner des schémas de simulation de sorte que $m(n)$, $h_{m(n)}$ et $d(n)$ satisfassent les conditions (4.10) et (4.11). Une première condition est que $m(n)h_{m(n)}^{2s} \to \infty$ lorsque $n \to \infty$, ce qui est assuré par exemple par le choix $h_m = m^{(\beta-1)/2s}$ pour $\beta \in (0,1)$. Le choix $m(n) = (\log n)^{(1/\delta)}$ avec $0 < \delta < \beta < 1$ assure (4.10) pour n assez grand. Enfin, il suffit de prendre $d(n) = c\log[(1+\gamma)n]$ avec $c = c(c_3, s) > 1$ pour vérifier (4.11) pour n assez grand.

4.3.4 Mise en œuvre et exemple

Comme pour l'algorithme de HM avec apprentissage (§ 4.2), il faut déterminer un compact assez grand sur lequel appliquer la méthode, et ceci se fait de la même façon. Mais ici, il est possible de conserver les m processus parallèles tout le long de la simulation, puisque le schéma ne nécessite plus d'éliminer de chaînes. La méthode est générique et le schéma adaptatif est implémenté sous forme d'une boîte noire fonction des paramètres de réglages $m(n)$, $h_{m(n)}$, $d(n)$ et du compact de travail. Cet outil sera disponible en ligne ([L2]). L'algorithme IHM est aussi plus simple que la méthode précédente car les estimateurs à noyau sont faciles à construire quelle que soit la dimension. D'autre part, les lois instrumentales sont des mélanges d'une loi uniforme et d'estimateurs à noyau, qui sont eux-même des mélanges de gaussiennes en pratique. Ces lois sont donc faciles à évaluer et à simuler.

La seule différence avec le schéma asymptotique est que le nombre m de processus parallèles est fixé au début et n'évolue pas avec n. On choisit en pratique la durée totale n de la simulation et en déduit $m(n)$ assurant (4.10). La suite de lois instrumentales se stabilise donc au bout d'un certain temps autour d'une position moyenne et l'on n'observe plus ensuite que des fluctuations. On peut alors décider d'arrêter les simulations parallèles, mais il se pose le problème du choix de la dernière loi instrumentale qui sera utilisée pour la suite. Une solution raisonnable semble être de prendre une moyenne des derniers apprentissages. Il est aussi envisageable de faire croître m avec n en échantillonnant des particules (chaînes) supplémentaires suivant les lois construites aux instants d'apprentissage. Cette perspective nous rapproche encore plus des schémas de systèmes de particules en interactions mais demande une nouvelle étude théorique.

Nous avons testé l'algorithme IHM sur un exemple en dimension 2 reproduisant une situation délicate car peu mélangeante. La loi cible choisie, un mélange de quatre gaussiennes, est multimodales avec des modes éloignés et un mode de faible poids (fig. 4.4). Nous avons utilisé IHM en arrêtant les simulations parallèles après une durée T choisie empiriquement, au-delà de laquelle seule une marginale est simulée, avec la dernière loi instrumentale apprise. A fin d'illustration, un exemple des lois instrumentales obtenues est donné fig. 4.5. Nous avons comparé IHM avec un algorithme de HM à marche aléatoire (RW) avec calibration de la variance σ^2, et un algorithme

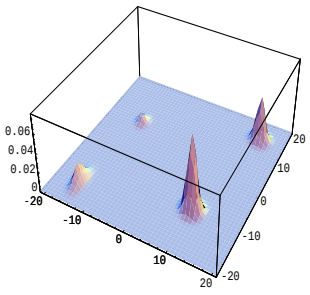


FIG. 4.4 – Loi cible π mélange de quatre gaussiennes bidimensionnelles.

RW	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$	$\sigma = 17$	$\sigma = 30$
	19.7	7.4	2.2	1.1	0.6
IS	2.4				
IHM	$m = 50$	$m = 100$	$m = 200$	$m = 300$	$m = 500$
	50.2	49.2	54.9	55.1	60.4

TAB. 4.1 – Taux d'acceptation empiriques $\hat{\alpha}$ (en %) calculés sur une chaîne de $n = 10,000$ itérations issue de chacune des trois stratégies avec différents réglages.

de HM indépendant (IS) utilisant comme loi instrumentale la loi uniforme sur le compact de travail. Ce dernier algorithme peut être considéré comme "presque" géométrique puisque la condition de minoration est satisfaite sur ce compact, et que la masse à l'extérieur est négligeable. Nous donnons ici quelques éléments de comparaison : les taux d'acceptation empiriques et la reconstruction des lois marginales (d'autres éléments figurent dans [RT10]).

Tout d'abord, le "bon" calibrage de la méthode RW ici est $\sigma \approx 17$, en raison des positions relatives des modes de π. La table 3.1 montre que pour RW, le taux d'acceptation correspondant au bon calibrage est très faible (1%), et que les recommendations usuelles (citées au § 4.1) ne s'appliquent donc pas. Pour IHM, ce taux d'acceptation est très important, et croît logiquement avec m puisque l'approximation de π par les lois instrumentales s'améliore. Ainsi, l'exploration du support par la chaîne issue de l'algorithme IHM est meilleure que celle des autres méthodes.

Les histogrammes marginaux construits sur des chaînes de même longueur montrent aussi une meilleure reconstruction par IHM des marginales (fig. 4.6). Notons que pour RW, d'autres choix que $\sigma = 17$ donnent des reconstructions encore moins bonnes que celles données ici. Enfin, des essais pour RW avec un nombre total d'itérations comparable à celui utilisé par IHM pour construire ses lois instrumentales donnent des résultats similaires.

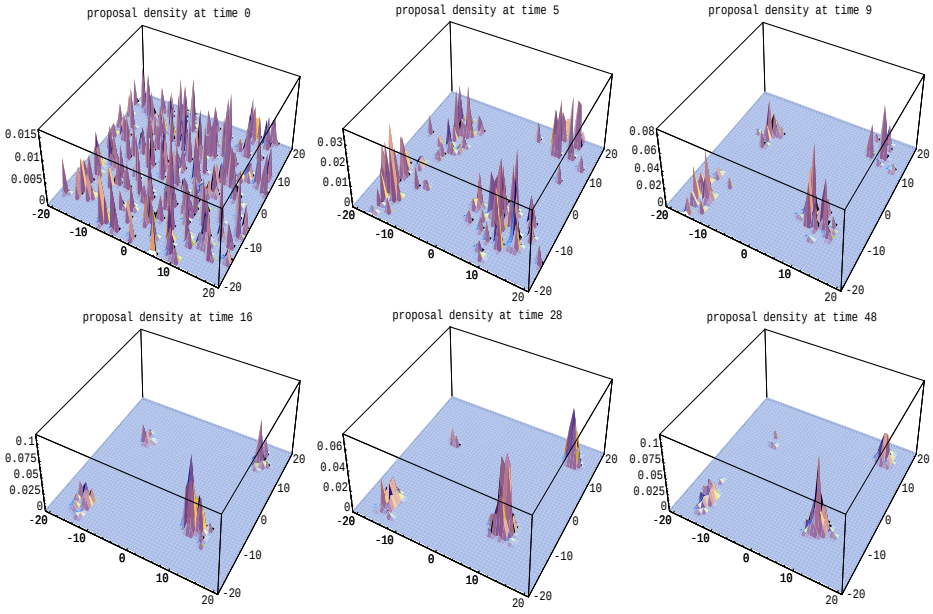


FIG. 4.5 – Suite de lois instrumentales $q_{\mathbf{x}_{n_k}^{\neq i}}$ construites par IHM, $k = 0, \ldots, 5$.

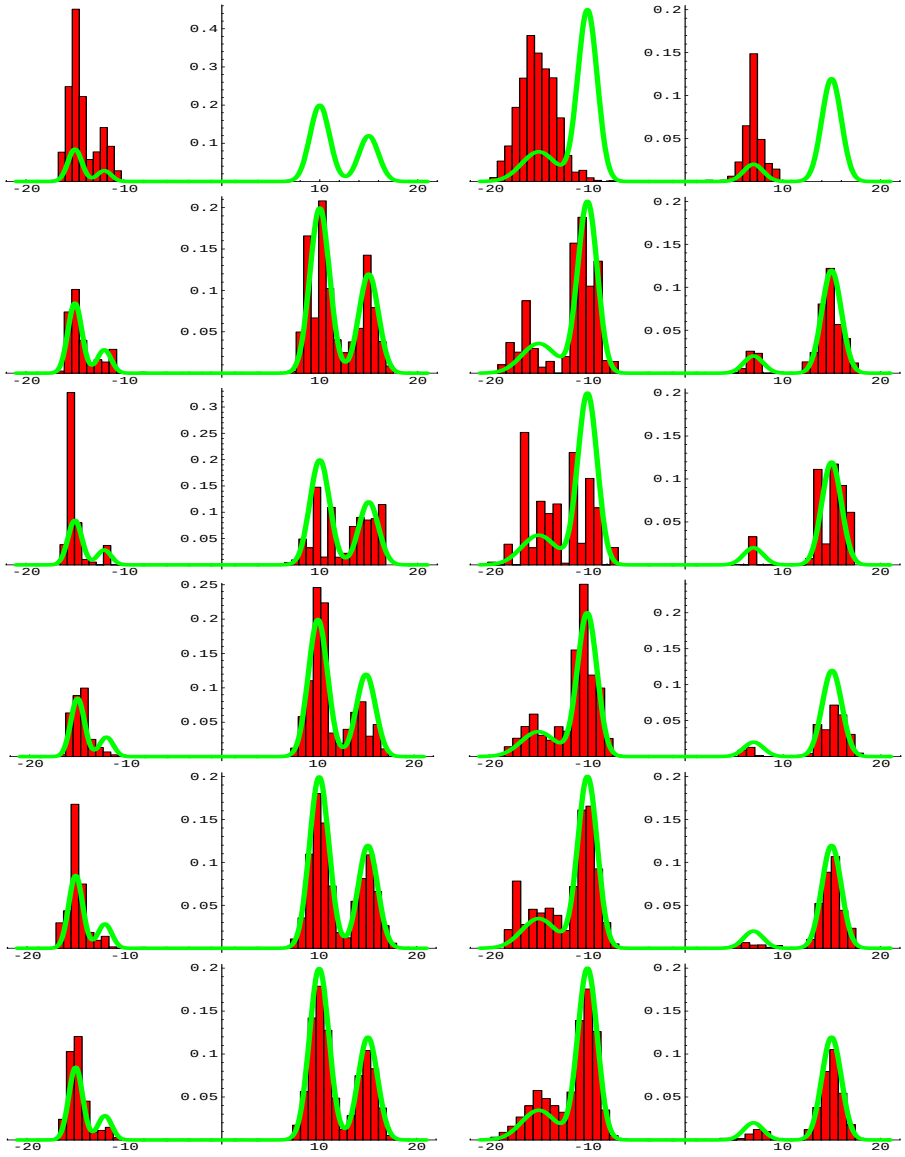


FIG. 4.6 – Densités marginales vraies et empiriques sur une chaîne de $n = 10,000$ itérations issue de chacune des stratégies. De haut en bas : RW $\sigma = 5$; RW $\sigma = 17$; RW $\sigma = 30$; IS ; IHM $m = 100 \times T = 50$; IHM $m = 500 \times T = 18$.

Chapitre 5

Algorithmes de restauration

De nombreux contextes conduisent à une certaine perte d'information sur les données observées. Les situations classiques sont la *censure* d'une partie des données, ou l'*agrégation* de données. Mais il existe d'autres situations (mélanges de lois, chaînes de Markov cachées,...) pour lesquelles les données disponibles s'interprètent naturellement comme des données manquantes par rapport à un certain niveau d'observation qui serait plus satisfaisant du point de vue statistique. On parle donc de données incomplètes chaque fois que l'on établit un modèle statistique jugé satisfaisant relativement à un certain niveau d'information sur les phénomènes aléatoires considérés, et que l'on ne dispose en fait que d'une part de cette information. Dans ce cas, il est classique de pouvoir assez facilement appliquer des méthodes d'estimation par maximum de vraisemblance sur le modèle initial – dit complet – alors que la forme analytique du modèle observé – dit incomplet – rend cet objectif inatteignable. On est alors amené à mettre en œuvre les algorithmes de nature probabiliste considérés dans ce chapitre.

On note $\mathbf{y}$ la donnée complète, dont la loi admet la densité $g(\mathbf{y}|\theta)$ par rapport à une mesure ν. Seule la donnée incomplète $\mathbf{x} = \pi(\mathbf{y})$ est observée, où π est une application (surjective) "perte d'information". La loi de $\mathbf{x}$ admet la densité $f(\mathbf{x}|\theta)$ par rapport à une mesure μ, donnée par $f(\mathbf{x}|\theta) = \int_{\pi^{-1}(\mathbf{x})} g(\mathbf{y}|\theta)\, d\nu(\mathbf{y})$. Dans le cas traité ici du mélange de lois, on peut écrire $\mathbf{y} = (\mathbf{x}, \mathbf{z})$. La perte d'information est une projection et une version de f est $f(\mathbf{x}|\theta) = \int_{\mathcal{Z}} g((\mathbf{x}, \mathbf{z})|\theta)\, d\mathbf{z}$.

L'objectif est l'estimation par maximum de vraisemblance du paramètre θ dans le modèle observé, i.e. la détermination de

$$\hat{\theta}_{\mathbf{x}} = \operatorname*{argmax}_{\theta \in \Theta} L(\theta), \quad \text{où } L(\theta) = \log f(\mathbf{x}|\theta).$$

Il est donc fréquent que $\hat{\theta}_{\mathbf{x}}$ soit difficile à calculer, mais que la nature analytique du modèle complet rende facile le calcul de $\hat{\theta}_{\mathbf{y}} = \operatorname{argmax}_{\theta \in \Theta} \log g(\mathbf{y}|\theta)$.

5.1 L'algorithme EM et ses versions stochastiques

L'idée de l'algorithme EM (*Expectation-Maximisation*, Dempster *et al.*, 1977) est de remplacer la maximisation de la log-vraisemblance des données complètes par la maximisation de son espérance conditionnellement à l'observation $\mathbf{x}$ et pour une valeur θ_0 du paramètre a priori arbitraire, puis d'itérer cette procédure. Si on note $Q(\theta|\theta') = \mathbb{E}[\log g(\mathbf{y}|\theta) \mid \mathbf{x}, \theta']$, l'itération $\theta_{t+1} = EM(\theta_t)$ est :

$$\begin{aligned} \texttt{étape E} &: \texttt{calcul de } Q(\cdot|\theta_t) \\ \texttt{étape M} &: \theta_{t+1} = \underset{\theta\in\Theta}{\texttt{argmax}}\, Q(\theta|\theta_t). \end{aligned}$$

Cet algorithme assure que $L(\theta_{t+1}) \geq L(\theta_t)$. La convergence de EM est étudiée dans Wu (1983) et dans Redner et Walker (1984) pour le cas des mélanges (un résumé figure dans [RT2]). La sensibilité de EM à toute sortes de pathologies telles que la convergence et l'immobilisation dans un maximum local de L, ou pire la stabilisation sur un point selle est bien connue. C'est initialement la raison principale pour laquelle sont apparues des versions stochastiques de EM, qui sont les algorithmes de restauration de données manquantes.

Dans [5], nous avons comparé EM avec les différentes versions stochastiques utilisées alors, dans le cas des mélanges de lois. Comme l'un des objectifs de cet article était de donner un état de l'art sur ce problème, nous avons précisé les formes explicites de ces algorithmes pour les mélanges. Ils sont simplement rappelés ici dans le cadre général.

SEM Le principe de l'algorithme SEM (*Stochastic EM*, Celeux et Diebolt 1985, 1992) consiste à restaurer la fraction manquante des données afin de pouvoir appliquer la procédure de maximum de vraisemblance sur les données complètes. La motivation initiale était de pallier aux inconvénients de EM en "bruitant" la suite (θ_t), mais comme nous l'avons rappelé au § 1.1, SEM peut aussi résoudre les difficultés d'implémentation qui peuvent survenir, notamment lorsque le calcul ou la maximisation de $Q(\cdot|\theta)$ n'est pas explicite (voir [1] et [3]). Si on note $k(\mathbf{y}|\mathbf{x}, \theta) = g(\mathbf{y}|\theta)/f(\mathbf{x}|\theta)$ la densité de $\mathbf{y}$ conditionnellement à $\mathbf{x}$ et pour la valeur θ du paramètre, l'itération $\theta_{t+1} = SEM(\theta_t)$ de SEM est :

$$\begin{aligned} \texttt{Restauration} &: \texttt{simuler } \mathbf{y}_{t+1} \sim k(\cdot|\mathbf{x}, \theta_t) \\ \texttt{Maximisation} &: \theta_{t+1} = \underset{\theta\in\Theta}{\texttt{argmax}} \log g(\mathbf{y}_{t+1}|\theta). \end{aligned}$$

On construit ainsi une suite $(\theta_t, t \geq 0)$ qui est une chaîne de Markov, dont il faut montrer l'ergodicité et le bon comportement de la loi stationnaire pour le problème d'estimation (Diebolt et Celeux, 1993). Comme pour les méthodes MCMC, l'estimation est donnée par une moyenne empirique si la LFGN s'applique.

SAEM Dans les situations où l'information disponible est faible par rapport à la fraction manquante, la variance des perturbations de la suite générée par SEM devient grande. Ceci a conduit à la détermination de l'algorithme SAEM (*Simulated Annealing EM*, Celeux et Diebolt, 1992). Il s'agit d'un hybride de EM et SEM, défini par

$$\theta_{t+1} = \gamma_{t+1} SEM(\theta_t) + (1 - \gamma_{t+1}) EM(\theta_t),$$

où la suite (γ_t) décroît de 1 à 0. SAEM évolue ainsi d'un "pur SEM" jusqu'à un "pur EM", générant une suite de moins en moins bruitée, et mimant les algorithmes de recuit simulé d'où il tire son nom. En choisissant bien la vitesse de décroissance de (γ_t), il est possible dans certains cas de montrer la convergence p.s. de (θ_t) vers un maximum local de L.

MCEM L'algorithme MCEM (Monte-Carlo EM, Wei and Tanner, 1990) est une autre manière de résoudre le problème du calcul d'espérance dans l'algorithme EM. Il consiste à remplacer Q par son approximation de Monte-Carlo. L'itération $\theta_{t+1} = MCEM(\theta_t)$ s'écrit :

$$\begin{aligned}
\texttt{Restauration} \quad &: \quad \texttt{simuler } \mathbf{y}^m \sim k(\cdot\,|\mathbf{x}, \theta_t), \quad m = 1, \ldots, B \\
\texttt{Monte-Carlo} \quad &: \quad \widehat{Q}(\theta|\theta_t) = \frac{1}{m} \sum_{j=1}^{m} \log g(\mathbf{y}^j|\theta) \\
\texttt{Maximisation} \quad &: \quad \theta_{t+1} = \underset{\theta \in \Theta}{\texttt{argmax}}\, \widehat{Q}(\theta|\theta_t).
\end{aligned}$$

Cet algorithme revient à SEM lorsque $m = 1$, et "tend vers" EM lorsque $m \to \infty$. Si, comme le proposent les auteurs, on fait croître $m = m(t)$ avec t, on obtient une autre version de type SAEM.

Mélange de lois

L'estimation des paramètres d'un mélange de lois est typiquement un problème d'information incomplète, où les observations sont le vecteur $\mathbf{x}$, et les composantes de provenance des observations sont les *variables latentes*, $\mathbf{z}$, non observées. L'estimateur $\hat{\theta}_{\mathbf{y}}$ est facile à déterminer pour $\mathbf{y} = (\mathbf{x}, \mathbf{z})$, mais $\hat{\theta}_{\mathbf{x}}$ est inaccessible. Ce problème a servi de base pour l'étude du comportement de EM : voir, e.g., Redner et Walker (1984), Titterington, Smith et Makov (1985), Celeux et Diebolt (1985), McLachlan et Basford (1989). Lorsque la famille de lois utilisée est exponentielle, EM et toutes les versions stochastiques ci-dessus sont explicites. Nous avons rappelé leurs définitions, ainsi que les résultats de convergence adaptés à ce contexte dans [5].

5.2 Comparaisons par simulation

Nous avons proposé une expérimentation de ces algorithmes par des simulations que nous avons voulues intensives et complètes, dans le cadre

des mélanges de lois gaussiennes unidimensionnelles. Trois mélanges de lois ont été testés, reflétant des situations pathologiques classiques telles que (i) composantes de même moyenne et variances différentes (loi symétrique à queues lourdes) ; (ii) composantes très imbriquées donnant une loi unimodale biaisée ; (iii) 4 composantes dont 2 séparées donnant des modes distants.

Deux versions de SEM ont été utilisées, se distinguant par la manière de construire l'estimateur (à l'aide de la moyenne empirique, ou en appliquant EM après une "exploration" utilisant SEM, baptisée SEM-EM). Ceci a donné cinq algorithmes à comparer (EM compris). Chaque stratégie a été appliquée sur 50 replications de Monte Carlo, pour trois tailles d'échantillons, plusieurs durées de simulation, et plusieurs méthodes d'initialisation. Les comparaisons ont portées notamment sur le *pourcentage de biens classés*, indicateur couramment utilisé en classification (e.g., Celeux et Govaert, 1993).

Un problème crucial est apparu clairement lors de nos essais : le besoin de détecter les permutations d'étiquetage des composantes du mélange qui peuvent se produire lorsque les algorithmes visitent des modes de L différents, mais équivalents au changement de numérotation des composantes du mélange près. L'existence de $k!$ modes pour un mélange à k composantes avait déjà été noté par, e.g., Redner et Walker (1984), mais n'avait pas été vraiment pris en compte lors de l'utilisation de ces algorithmes. Une séquence simulée par SEM peut ainsi explorer alternativement plusieurs modes équivalents. Si l'on calcule un estimateur moyenne empirique sur une telle suite, l'estimation peut être très mauvaise (par exemple à mi-chemin de deux modes équivalents). Dans notre cas, où l'estimateur est calculé sur replications, ce problème est encore plus gênant. Nous avons proposé quelques critères permettant la détection de ces permutations et le retour à l'étiquetage initial.

Pour le mélange avec même moyennes, les versions de SEM se sont montrées nettement plus performantes que les autres algorithmes. Le mélange très imbriqué, plus difficile, a également été mieux estimé (et surtout mieux "reclassé") par les versions SEM, mais de manière moins nette, notamment en ce qui concerne la séparation des moyennes très proches. Le mélange à 4 composantes a aussi été mieux estimé par les stratégies de type SEM. Nous avons en complément représenté l'exploration par EM et SEM de la surface de L, illustrant ainsi notamment le problème d'étiquetage.

Enfin, nous avons testé SEM sur un jeu de données réelles tiré d'une étude de cas, où deux choix de modélisation font que le paramètre θ est en dimensions 8 ou 11 (Basford et McLachlan, 1985). La stratégie SEM-EM s'est là aussi révélée plus performante que l'algorithme EM utilisé par les auteurs, en découvrant un maximum supérieur à ceux trouvés par les auteurs dans un cas, et en découvrant un autre point fixe de EM dans l'autre cas.

Chapitre 6

Problèmes mal posés en statistique

L'outil technique de résolution des problèmes mal posés est la *régularisation* d'inverse d'opérateurs. Cette régularisation est nécessaire lorsque l'opérateur inverse n'est pas continu, et qu'il est appliqué à une perturbation de la transformation initiale. L'inverse du "signal" perturbé peut alors être très différent de l'inverse du signal non perturbé, qui est l'objet que l'on cherche à reconstruire. C'est le cas notamment en statistique, où l'on peut donner pour préciser les idées le cadre formel ci-dessous.

Supposons que l'on observe $\hat{g}$, une perturbation de g, ou plus précisément en statistique un estimateur $\hat{g}_n$ de g fondé sur n observations. On sait que g est une transformation $g = Kf$ d'un "paramètre" inconnu f qui est l'objet à estimer. Le principe est de construire une suite $(K_m^{-1}, m \geq 0)$ vérifiant $||K_m^{-1}g - f|| \to 0$ lorsque $m \to \infty$ pour une certaine norme. On l'appelle suite *d'inverses régularisés* de K. Carrol *et al.* (1991) donnent des méthodes de constructions de telles suites dans différent contextes, notamment pour les opérateurs de convolution dans des espaces de Hilbert.

Lorsqu'on sait construire une suite régularisante, on utilise $\hat{f}_n = K_m^{-1}\hat{g}_n$ pour estimer f. En contrôlant la vitesse de la suite $m(n)$, et à partir d'hypothèses sur la vitesse avec laquelle $\mathbb{E}(||\hat{g}_n - g||) \to 0$, Carrol *et al.* (1991) donnent également des résultats de consistance de la suite $(\hat{f}_n = K_{m(n)}^{-1}\hat{g}_n)$ d'estimateurs de f.

6.1 Inversion de transformée de Laplace bruitée

Dans [2], nous construisons une séquence d'inverses régularisés pour la transformée de Laplace. L'un des exemples d'utilisation en statistique est l'estimation de la densité d'un mélange continu de lois exponentielles. Supposons que l'on dispose d'un n-échantillon issu d'un tel mélange, dont la

fonction de répartition est

$$G(t) = \int_0^\infty (1 - e^{-tx}) f(x)\, dx, \quad t \in (0, \infty), \tag{6.1}$$

où f est la densité inconnue du mélange, sur $(0, \infty)$. L'approche par inversion consiste à utiliser le fait que G s'exprime par une transformée de Laplace $G(t) = 1 - (\mathcal{L}f)(t)$. On estime alors G par la fonction de répartition empirique $\hat{G}_n$, puis on inverse (6.1) en utilisant $\hat{G}_n$ à la place de G.

Pour reprendre les notations du cadre général, on cherche à reconstruire f à partir d'une observation bruitée $\hat{g}$ de $g = \mathcal{L}f$. Supposons $f \in L^2(\lambda_+)$, où λ_+ est la mesure de Lebesgue sur $(0, \infty)$. L'opérateur $\mathcal{L}$ est borné mais seulement injectif, de sorte que son inverse n'est pas défini sur tout $L^2(\lambda_+)$. Il n'est pas non plus continu, ce qui fait que la détermination de f à partir de $\hat{g}$ est un problème mal posé. Il s'agit donc de déterminer une suite d'opérateurs linéaires bornés $(\mathcal{L}_m^{-1})$ sur $L^2(\lambda_+)$, telle que

$$\left\| \mathcal{L}_m^{-1}(\mathcal{L}f) - f \right\| \to 0, \quad \text{lorsque } m \to \infty, \quad f \in L^2(\lambda_+). \tag{6.2}$$

Nous construisons $(\mathcal{L}_m^{-1})$ en exprimant la transformée de Laplace comme un opérateur de convolution des fonctions sur le groupe multiplicatif $(0, \infty)$ muni de la mesure (de Haar) $d\mu_+(x) = x^{-1} d\lambda_+(x)$. Ceci permet d'utiliser un résultat de Carrol *et al.* (1991), théorème 3.1, sur la régularisation de la déconvolution.

Nous précisons ensuite les vitesses de convergence et le choix de la suite $m(n)$. Sous une hypothèses de régularité de f et l'hypothèse suivante sur l'estimateur de g,

$$\mathbb{E}(||\hat{g}_n - g||) = \mathcal{O}(n^{-p}), \quad \text{lorsque } n \to \infty, \text{ pour } p > 0,$$

le choix $m(n) = \left[n^p (\log n)^{-1/2} \right]$ avec d'autres conditions techniques assure que

$$\mathbb{E}(||\hat{f}_n - f||) = \mathcal{O}\left((\log n)^{-1/2} \right), \quad \text{lorsque } n \to \infty.$$

Nous illustrons cette technique par un exemple dans le cadre statistique à partir de données simulées issues du mélange continu (6.1), en choisissant pour la densité de mélange f elle-même une loi exponentielle. La mise en œuvre demande la résolution numérique d'une intégrale multiple assez délicate (oscillante), et n'est donc pas très directe (elle a demandé l'utilisation de techniques d'intégration numérique *ad hoc*, voir [2]). La reconstruction de f est tout de même satisfaisante pour des tailles d'échantillon modérées ($n = 200$). Nous étudions enfin le comportement de l'erreur d'estimation en fonction de m. Les essais montrent que le choix de petites valeurs de m semble préférable pour les tailles d'échantillon raisonnables testées.

6.2 Mélange signé de lois exponentielles

Dans [4], nous utilisons une technique d'inversion assez proche de manière à estimer les paramètres d'un mélange de lois exponentielles, lorsque la mesure de mélange est discrète, mais avec des poids non nécessairement positifs (mesure discrète signée). Ce type de loi est aussi appelée *distribution hyper-exponentielle généralisée* (voir Botta *et al.*, 1987), et a des applications en théorie du contrôle (voir aussi Martin et Miller, 1992). Ces problèmes étaient l'une des motivations de ce travail, l'autre étant l'application statistique au mélange discret (avec poids positifs) de lois exponentielles. Plus précisément, l'objet d'intérêt est la densité

$$f(x) = \sum_{k\geq 1} \alpha_k e^{-x/\theta_k}, \quad x \in (0,\infty), \tag{6.3}$$

avec $\sum_{k\geq 1} |\alpha_k| < \infty$, et $0 < \theta_1 < \theta_2 < \cdots$, et l'objectif est d'estimer les poids $(\alpha_k, k \geq 1)$ et les paramètres $(\theta_k, k \geq 1)$.

Pour un noyau arbitraire $\varphi \in L^1(\mu_+)$, on considère le "mélange"

$$f(x) = \sum_{k\geq 1} \alpha_k \varphi(\frac{x}{\theta_k}), \quad x > 0. \tag{6.4}$$

Comme précédemment, le principe est d'interpréter (6.4) comme une convolution avec le noyau φ, de fonctions sur le groupe $(0,\infty)$ muni de la multiplication $x \oplus y = x \cdot y$. La transformée de Fourier sur $L^1(\mu_+)$ est définie par $(\mathcal{F}^+\varphi)(t) \triangleq (\mathcal{F}(\varphi \circ \exp))(t)$, et on note $\varphi^+ = \mathcal{F}^+\varphi$. En l'appliquant à f, on vérifie que

$$f^+(t) = \Big(\sum_{k\geq 1} \alpha_k e^{it\log\theta_k}\Big)\varphi^+(t),$$

de sorte que f^+/φ^+ est une combinaison linéaire d'exponentielles complexes. Dans le cas (théorique) où f^+ est parfaitement connue, et avec une "condition de séparation" des (θ_k), les paramètres peuvent être restitués de façon exacte par une transformée de Fourier inverse. En réalité, on dispose seulement d'une approximation $\hat{f}_n^+$ (un estimateur construit à partir de n observations dans l'exemple statistique). On montre (théorème 2.3) que l'on peut également estimer les paramètres si l'estimateur satisfait la condition

$$\sup_{-M\leq t\leq M} |\hat{f}_n^+(t) - f^+(t)| \xrightarrow{\mathbb{P}} 0 \quad \text{lorsque } n \to \infty, \text{ pour tout } M > 0. \tag{6.5}$$

Dans le cas d'intérêt $\varphi(x) = e^{-x}$, un problème se pose car $\varphi \notin L^1(\mu_+)$. Il est possible de modifier le noyau, de manière à retomber sur le problème précédent avec une relation équivalente $h(x) = \sum_{k\geq 1} \beta_k \rho(x/\theta_k)$, où $\beta_k = \sqrt{\theta_k}\alpha_k$, et $\rho \in L^1(\mu_+)$ avec ρ^+ explicite. Si l'on peut construire un estimateur $\hat{h}_n^+$ de h^+ vérifiant la condition (6.5), le problème est résolu. On montre finalement que c'est le cas pour les deux exemples traités, dont celui de nature statistique.

Chapitre 7

Liste de travaux

7.1 Articles et ouvrages

Travaux publiés ou à paraître

[1] Chauveau, D. (1992), Algorithmes EM et SEM pour un mélange censuré de distributions de défaillances, application à la fiabilité. *Rev. Statistique Appliquée*, **40**, 67–76.

[2] Chauveau, D., van Rooij, A., Ruymgaart, F. (1994), Regularized inversion of noisy Laplace transforms. *Advances in Applied Math.*, **15**, 186–201.

[3] Chauveau, D. (1995), A stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning Inference*, **46**, 1–25.

[4] Chauveau, D., Martin, C. F., van Rooij, A. C. and Ruymgaart, F. H. (1996), Discrete signed mixtures of exponentials. *Commun. Statist. – Stochastic Models*, **12**, n° 2, 245–263.

[5] Celeux, G., Chauveau, D. and Diebolt, J. (1996), Stochastic versions of the EM algorithm : An Experimental Study in the Mixture Case. *J. Statist. Comput. Simul.* **55**, 287–314.

[6] Chauveau, D., Diebolt, J. and Robert, C.P. (1998), Control by the Central Limit Theorem. In *Discretization and MCMC convergence assessment* (C.P. Robert Ed.), Lecture Notes in Statistics n° 135, Springer-Verlag, New York, **Chap. 5**, 99–126.

[7] Muri, F., Chauveau, D., Cellier, D. (1998), Convergence assessment in latent variable models : DNA Applications. In *Discretization and MCMC convergence assessment* (C.P. Robert Ed.), Lecture Notes in Statistics n° 135, Springer-Verlag, New York **Chap. 6**, 127–146.

[8] Chauveau, D. et Vandekerkhove, P. (1999), Un Algorithme de Hastings-Metropolis avec apprentissage séquentiel. *C. R. Acad. Sci. Paris*, **t. 329**, Série I, p. 173–176.

[9] Chauveau, D. and Diebolt, J. (1999), An automated stopping rule for MCMC convergence assessment. *Computational Statistics*, **14**, 3, 419–442.

[10] Chauveau, D. and Vandekerkhove, P. (2001), Improving convergence of the Hastings-Metropolis algorithm with an adaptive proposal. *Scandinavian Journal of Statistics* (to appear).

[11] Altaleb, A. and Chauveau, D. (2001), Bayesian analysis of the Logit model and comparison of two Metropolis-Hastings strategies. *Computational Statistics & Data Analysis* (to appear).

[12] Chauveau, D. et Vandekerkhove, P. (2001), Algorithmes de Hastings-Metropolis en interaction. *C. R. Acad. Sci. Paris* (à paraître).

Discussions d'articles

[13] Chauveau, D. (1997), in discussion of : The EM algorithm – An old folk song sung to a fast new tune, by Meng, X.L. and Van Dyk, D., *J. Royal Statistical Society*, B, **59 :3**, 511–567.

Soumis pour publication

[14] Chauveau, D. et Vandekerkhove, P. (09/2000), An entropy estimator to control stability of Markovian dynamical systems.

[15] Chauveau, D. and Diebolt, J. (12/2000), Estimation of the limiting variance for Markov chains.

[16] Chauveau, D. et Vandekerkhove, P. (07/2001), Interacting Hastings-Metropolis algorithms.

7.2 Thèse et rapports techniques

[RT1] Chauveau, D., Raoult, J.P. (1990), *Comportement asymptotique de deux approximations poissoniennes du taux de défaillance d'un matériel électronique*. Université Paris-Sud Orsay et Alcatel CIT, Dpt ATC, Ormes.

[RT2] Chauveau, D. (1991), *Extension des algorithmes EM et SEM à la reconnaissance de mélanges censurés de distributions de défaillances*. Thèse de Doctorat, Université Paris-Sud, Orsay, France.

[RT3] Bon, J.L., Bretagnolle, J., Chauveau, D., Jakubowicz, P., Pamphile, P. et Raoult, J.P. (1993), *Calcul séquentiel de fiabilité à partir d'approximations exponentielles*. Rapport technique Université Paris-Sud et Electricité de France, groupe ESF.

[RT4] J.P. Raoult, D. Chauveau, C. Cocozza, M. Roussignol (1995), *Modèles de durée de survie applicables à la mécanique sous contraintes d'environnement*. Rapport technique Université de Marne-la-Vallée et Electricité de France.

[RT5] Chauveau, D. and Diebolt, J. (1998) *An automated stopping rule for MCMC convergence assessment*. Rapport de Recherche RR-3566, INRIA Rhône-Alpes.

[RT6] Chauveau, D. et Vandekerkhove, P. (1999) *Improving convergence of the Hastings-Metropolis algorithm with a learning proposal*. Prépublication no 14/99, Université Marne-la-Vallée.

[RT7] Chauveau, D. et Vandekerkhove, P. (2000) *An entropy estimator to control stability of Markovian dynamical systems*. Prépublication no 05/2000, Université Marne-la-Vallée.

[RT8] Chauveau, D. and Diebolt, J. (2000), *Stability properties for a product Markov chain*. Prépublication no 06/2000, Université Marne-la-Vallée.

[RT9] Chauveau, D. and Diebolt, J. (2001), *Estimation of the limiting variance for Markov chains*. Prépublication no 01/2001, Université Marne-la-Vallée.

[RT10] Chauveau, D. et Vandekerkhove, P. (2001), Interacting Hastings-Metropolis algorithms. Prépublication no 08/2001, Université Marne-la-Vallée.

[RT11] Chauveau, D. (2001), *User's guide to the CLTC software*, Université Marne-la-Vallée. (Manuel d'utilisation de [L1] distribué en ligne sur `http://math.univ-mlv.fr/~chauveau/pgm/cltc/cltc.html`).

7.3 Réalisations informatiques

[L1] Réalisation et publication en ligne du logiciel "`CLTC`" pour le diagnostic automatique de la convergence des algorithmes MCMC (1998). Logiciel en `C` et `Mathematica`, distribué et documenté sur le site `http://math.univ-mlv.fr/~chauveau/pgm/cltc/cltc.html`.

[L2] Réalisation du logiciel "`IHM`" de type boîte noire en `C` pour l'implémentation d'algorithmes de Hastings-Metropolis en interaction (2001). Travail en cours sur `http://math.univ-mlv.fr/~chauveau/pgm/ihm`.

Bibliographie

Altaleb, A. (1999), *Méthodes d'échantillonnage par mélanges et algorithmes MCMC*. Thèse de doctorat de l'Université de Rouen (direction : C.P. Robert).

Basford, K. E. and McLachlan, G. J. (1985), Likelikood estimation with normal mixture models. *Applied Statistics*, **34**, 282–289.

Best, N.G., Cowles, M.K. et Vines, K. (1995), *CODA : Convergence diagnosis and output analysis software for Gibbs sampling output*. Version 0.30. Tech. Report, MRC Biostatistics Unit, Univ. of Cambridge.

Billingsley, P. (1968), *Convergence of probability measures*. John Wiley & Sons, New York.

Biscarat, J.C. (1994), Almost sure convergence of a class of stochastic algorithms. *Stochastic Processes and their Applications*, **50**, 83–99.

Bosq, D. (1996), *Nonparametric statistics for stochastic processes*. Lecture Notes in Statistics no 110, Springer-Verlag, New-York.

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

Botta, R.F., Harris, C.M. and Marchal, W.G. (1987), Characterizations of generalized hyperexponential distribution functions. Commun. Statist. – Stochastic Models, **3**, 115–148.

Brooks, S.P., and Roberts, G. (1998), Assessing convergence of Markov Chain Monte Carlo algorithms. *Statistics and Computing*, **8**(4), 319–335.

Broniatowski, M., Celeux, G. and Diebolt, J. (1983), Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data Analysis and Informatics*, (Diday E. *et al;* eds.) **3**, 359–374, Amsterdam, North Holland

Caroll, R.J., van Rooij, A.C.M. and Ruymgaart, F.H. (1991), Theoretical aspects of ill-posed problems in statistics. *Acta Applicandae Mathematicae*, **24**, 113–140.

Celeux, G. and Diebolt, J. (1985), The SEM Algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly*, **2**, 73–82

Celeux, G. and Diebolt, J. (1992), A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports*, **41**, 119–134.

Celeux, G. and Govaert, G. (1993), Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statis. Comput. Simul.*, **47**, 127–146.

Dacunha-Castelle, D. and Duflo, M. (1986), *Probability and Statistics*, vol. 2. Springer-Verlag, New York.

Del Moral P., and Guionnet A. (1999), Central Limit Theorem for nonlinear filtering and Interacting Particle Systems. *Annals of Applied Probability*, **9**, no 2, 275–297.

Del Moral P., and Miclo L. (2000), *Branching and Interacting Particle Systems approximations of Feynman-Kac formulae with applications to nonlinear filtering*. Séminaire de Probabilités XXXIV, Ed. J. Azéma and M. Emery and M. Ledoux and M. Yor, Lecture Notes in Mathematics, Springer-Verlag Berlin, **Vol. 1729**, 1–145 .

Delong, D. M. (1981), Crossing probabilities for a square root boundary by a Bessel process. *Communication in Stat. Theory and Methods*, **A10**, 2197-2213.

Delyon, B., Lavielle, M. and Moulines, E. (1999), On a stochastic approximation version of the EM algorithm. *Annals of Statistics*, **27** (1), 94–128.

Dempster, A., Laird, N. and Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc.*, B, **39**, 1–38.

Diebolt, J. and Celeux, G. (1993), Asymptotic properties of a stochastic EM algorithm for estimating mixture proportions. *Stochastic Models*, **9**, 599–613.

Diebolt, J. and Robert, C.P. (1994), Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical Society*, B, **56**, 363–375.

Duflo, M. (1996), *Algorithmes stochastiques*. Mathématiques et applications no 23, Springer-Verlag.

Eggermont, P. P. B. and LaRiccia, V. N. (1999), Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE trans. Inform. Theory*, **45**, no. 4, 1321–1326.

Fuk, D. Kh., and Nagaev, S. V. (1971, 1976), Probability inequalities for sums of independent random variables. *Th. Probab. Appl.* **16**, 643–660, **21**, 875.

Gelfand, A.E. and Sahu, S.K. (1994), On Markov chain Monte Carlo acceleration. *Journal of Computational and Graphical Statistics* **3**, 261–276.

Gelfand, A.E. and Smith, A.F.M. (1990), Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. and Rubin, D. B. (1992), Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, no. 4, 457–511.

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

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996), *Markov Chain Monte Carlo in practice*. Chapman & Hall, London.

Green, P. J. (1995), Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82** (4), 711–732.

Guihenneuc-Jouyaux, C. and Robert, C.P. (1998), Valid discretization via renewal theory, In *Discretization and MCMC convergence assessment* (C.P. Robert Ed.). Lecture Notes in Statistics n° 135, Springer-Verlag, New York.

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

Holden, L. (1998), Geometric convergence of the Metropolis-Hastings simulation algorithm. *Statistics and Probability Letters*, **39**, 4, 371–377.

Kemeny, J.G. and Snell, J.L. (1960), *Finite Markov Chains*. Springer-Verlag, New York.

Lehmann, E. L. (1975), *Nonparametrics : Statistical methods based on rank*. Holden-Day series in Probability and Statistics. Mc Graw-Hill.

Martin, C.F. and Miller, J. (1992), Observer based design for robust stabilization of nonlinear systems. *The mathematics of control theory* (N.K. Nichols and D.H. Owens Eds.), Clarendon, Oxford.

McLachlan, G.J. and Basford, K.E. (1989), *Mixture models - inference and applications to clustering*. New York, Marcel Dekker.

Meng, X.L. et Van Dyk, D. (1997), The EM algorithm – An old folk song sung to a fast new tune. *J. Royal Statistical Society*, B, **59 :3**, 511–567.

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

Meyn, S.P. and Tweedie, R.L. (1993), *Markov chains and stochastic stability*. Springer-Verlag, London.

Raftery, A.E., and Lewis, S. (1992), How many iterations in the Gibbs sampler ?, in *Bayesian Statistics*, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (eds.), 4, 763–773. Oxford University Press, Oxford.

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

Robert, C.P. (1996), *Méthodes de Monte Carlo par chaînes de Markov*. Economica, Paris.

Roberts, G.O. and Tweedie, R.L. (1996), Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

Roberts, G.O. and Tweedie, R.L. (1999, 2001), Bounds on regeneration times and convergence rates for Markov chains. *Stoch. Proc. Applic*, **80**, 211–229, with correction **91** 337–338.

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

Shapiro, S.S., and Wilk, M.B. (1965) An analysis of variance test for normality. *Biometrika* **52**, 591–611.

Tanner, M. A. and Wong, W. H. (1987), The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550.

Titterington, D. M., Smith, A. F. M. and Makov U. E. (1985), *Statistical analysis of finite mixture distribution*. New York, Wiley.

Tsybakov, A. B. and Van Der Meulen, E. C. (1994), Root t consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.*, **23**, 75–83.

Wei, G. C. G. and Tanner, M. A. (1990), A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.

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

Ycart, B. (1999), Cutoff for samples of Markov chains. *ESAIM-PS*, **3**, 89–

107.