# Méthodes de Monte Carlo pour l'inférence dans les modèles à données latentes
## — Mémoire d'habilitation à diriger des recherches —

Olivier Cappé

Soutenue le 5 décembre 2006

Université de Marne la Vallée
Spécialité : Traitement du signal et des images

devant le jury composé de
Gilles Celeux, Directeur de Recherche INRIA, univ. Paris XI *(rapporteur)*
Guy Demoment, Professeur univ. Paris XI, LSS *(président)*
Hans Rudolf Künsch, Professeur ETH Zürich *(rapporteur)*
Jean-Pierre Le Cadre, Directeur de Recherche CNRS, IRISA *(rapporteur)*
Philippe Loubaton, Professeur univ. Marne la Vallée, IGM

# Avant propos

Sortant d'un travail de rédaction et de synthèse de grande ampleur en anglais, la monographie [Cappé et al., 2005] rédigée avec Eric Moulines et Tobias Rydén, qui a mobilisé la majeur partie de mon énergie de l'été 2002 au printemps 2005, j'ai souhaité conserver ce mémoire d'habilitation aussi concis que possible. J'ai par ailleurs adopté le parti pris d'éviter de dupliquer à l'identique les aspects traités dans [Cappé et al., 2005] pour privilégier une présentation parfois assez différente, que j'ai en général voulue plus aisée d'accès, ainsi que des aspects plus prospectifs concernant mon activité actuelle de recherche.

Sur ce dernier point, je dois dire que paradoxalement un des aspects de la vie de chercheur CNRS (comparée à celle des enseignant-chercheurs que je côtoie le plus souvent) que j'ai appris à apprécier, ou tout au moins à ne pas considérer comme une simple contrainte dénuée de sens, est celui de devoir écrire régulièrement un rapport d'activité. En particulier, le fait de devoir présenter de façon cohérente des travaux qui, souvent, tiennent plus d'un enchaînement, parfois assez incontrôlable, d'occasions, d'engagements ou de rencontres m'a toujours été profitable. L'autre aspect appréciable, pour moi en tout cas, de ce rapport d'activité consiste à tenter de formuler des questions concernant l'avenir en terme de directions de recherche, de projets, voire de stratégie. C'est bien évidemment également une des limites de l'exercice puisque ce type de rapport étant essentiellement destiné à exister plus qu'autre chose, les questions posées restent essentiellement, et le plus souvent uniquement, des questions que l'on se pose à soi-même.

A l'encontre des meilleures traditions, ce document d'habilitation ne fait pas le point sur les travaux de recherche effectués depuis ma thèse (soutenue le 27 septembre 1993) mais plutôt sur des thématiques de recherche que j'avais mises en avant dans le projet de recherche sur lequel j'ai été recruté au CNRS (à l'automne 1996). Inertie de l'existant aidant, il s'agit de thèmes sur lesquels je n'ai réellement commencé à travailler que depuis 1998–1999. Ce document d'habilitation fait donc le point sur une période de mon activité de recherche relativement courte s'étendant de 1999 à 2006 (la liste exacte des publications référencées dans le document est donnée en page 9). Il fait donc l'impasse sur les travaux effectués au cours de ma thèse, concernant la réduction de bruit de fond dans les enregistrements audio [Cappé, 1994][1], [Cappé & Laroche, 1995; Godsill et al., 1998] ainsi que, plus généralement, en traitement de la parole [Cappé & Moulines, 1996; Cappé et al., 1998c; Stylianou et al., 1998; Campedel-Oudot et al., 2001]. C'est également la raison pour laquelle on ne trouvera, de façon peut être plus gênante, que fort peu d'écho des thèses que j'ai encadrées (Stéphanie Dubost, 2001 ; Guillaume Picard, 2005) ou co-encadré avec Eric Moulines (Ioannis Stylianou, 1996 ; Marine Campedel-Oudot, 1998 ; Vincent Buchoux) dans ce document, à l'exception de celle de Loïs Rigouste (en cours, co-encadré avec François Yvon) plus directement liée à l'utilisation de modèles

[1] IEEE Signal Processing Society 1995 Young Author Best Paper Award

à données latentes. De façon plus anecdotique, j'ai également soustrait du champ du document quelques travaux moins personnels et dont la thématique se raccrochait de surcroît beaucoup plus difficilement aux thèmes développés ici comme [Buchoux et al., 2000b] et [Cappé et al., 2002][2].

Etant donné que les travaux présentés ici s'inscrivent dans le cadre de mon projet de recherche CNRS de 1996, il m'a semblé utile et intéressant, ne serait-ce que pour observer un exemple de l'écart entre les prédictions et la réalité en matière de stratégie de recherche, de citer le résumé que j'en faisait alors :

> *Processus de Markov cachés pour le traitement du signal*
> *Programme de recherche présenté par Olivier Cappé (janvier 1996)*
>
> *Les processus de Markov cachés se sont révélés être des outils extrêmement puissants pour modéliser divers phénomènes présentant des changements de dynamique au cours du temps. L'estimation et l'utilisation de ce type de modèles soulève un ensemble de questions ardues qui stimulent un domaine de recherche très actif. De plus, l'essor récent de techniques nouvelles de calcul statistique laisse présager d'importants développements scientifiques dans ce domaine.*
>
> *Le programme de recherche présenté s'attache à l'étude des processus de Markov cachés dans le cadre de trois applications centrales dans le domaine du traitement du signal (restauration de signaux dégradés, traitement de la parole et modélisation de trafic téléinformatique).* [...]
>
> *Etant donné l'importance des problèmes de nature méthodologique qui se posent, il me semble également nécessaire de considérer des objectifs à plus long terme. Le paragraphe 3* [du projet] *propose trois axes de réflexion (prise en compte des connaissances a priori, initialisation et algorithmes d'optimisation robustes, adéquation du modèle) destinés à compléter et à enrichir les travaux portant sur chacune des applications considérées.*

A la (re-)lecture de ces paragraphes, il me semble que trois constatations s'imposent. Tout d'abord le plan général a été tenu et [Cappé et al., 2005] constitue en quelque sorte l'aboutissement de la logique annoncée ci-dessus consistant à participer, à mon niveau, au regain d'intérêt autour du thème des modèles de Markov cachés qui s'est manifesté de façon assez claire à partir de la seconde moitié des années 1990. La seconde constatation est que les applications citées étaient surtout guidées par une certaine continuité avec mes travaux antérieurs sur l'audio et la parole, continuité que j'ai plutôt cherché à tarir qu'à entretenir. J'ai en fait donc fort peu travaillé sur les applications en question. Le cas du télétrafic qui était en fait la seule piste réellement nouvelle mentionnée parmi ces applications est un peu à part puisque j'ai travaillé autour de ce type données [Cappé et al., 1998b; Cappé, 2002; Cappé et al., 2002; Cappé & Roueff, 2003] mais d'une façon que je juge toutefois comme faiblement aboutie et, dans certains cas, assez éloignée du thème des modèles de Markov cachés stricto sensu. Enfin, parmi les pistes méthodologiques mentionnées dans le troisième paragraphe, un aspect que je n'avais manifestement pas anticipé était l'importance que prendrait dans ma recherche le thème des méthodes de simulation. Ce thème de l'utilisation de méthodes de simulation, désignées sous le nom de *méthodes de Monte Carlo* (au sens large), pour l'inférence dans les modèles à données latentes constitue le point commun de tous les travaux présentés dans ce mémoire.

[2] IEEE Signal Processing Society 2005 Signal Processing Magazine Best Paper Award.

Pour conclure, je dois avouer que j'ai toujours eu une certaine réticence vis à vis des remerciements dans les documents scientifiques dont j'ai parfois l'impression qu'ils relèvent plus de la vie privée de l'auteur que de ce que les lecteurs potentiels viennent y chercher (mais je peux me tromper). Pour une juste appréciation de ce qui suit, cependant, je ne peux passer sous silence le nom de deux personnes qui ont eu une influence particulièrement décisive sur mes travaux de recherche. Tout d'abord, Eric Moulines qui m'a permis le virage thématique opéré à l'issue de ma thèse (en 1995) notamment en m'enseignant, par l'exemple, les premiers rudiments de probabilité et statistiques. Je dois également à Eric et à son enthousiasme le fait d'avoir su — et dans certains cas, d'avoir dû — vaincre mon hésitation naturelle vis à vis des incertitudes de l'avenir. En second lieu, Christian Robert, que j'ai rencontré en 1997, est bien évidemment à l'origine de ma conversion à l'utilisation des techniques de simulation. Travailler avec Christian c'est également être confronté à une efficacité et une réactivité parfois stupéfiante qui m'a beaucoup donné à réfléchir. Je les remercie affectueusement tous deux.

# Table des matières

# Liste des travaux annexés au mémoire

| | |
|---|---|
| [Cappé et al., 1999] O. Cappé, A. Doucet, M. Lavielle & E. Moulines. Simulation-based methods for blind maximum-likelihood filter identification. *Signal Processing*, 73(1–2):3–25, 1999. URL `http://www.tsi.enst.fr/~cappe/papers/spdcv.ps.gz`. | Section 3.3 |
| [Cappé, 2001a] O. Cappé. Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. *Monte Carlo Methods and Applications*, 7(1–2):81–92, 2001a. URL `http://www.tsi.enst.fr/~cappe/papers/ma_rmlpa.ps.gz`. | Section 2.3 |
| [Cappé, 2002] O. Cappé. A bayesian approach for simultaneous segmentation and classification of count data. *IEEE Trans. Signal Processing*, 50(2):400–410, February 2002. URL `http://www.tsi.enst.fr/~cappe/papers/cntdat.ps.gz`. | Section 3.3 |
| [Cappé et al., 2003] O. Cappé, C. Robert & T. Rydén. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. Royal Statist. Soc. Ser. B*, 65(3):679–700, 2003. URL `http://www.tsi.enst.fr/~cappe/papers/crr01ct.ps.gz`. | Section 3.4 |
| [Cappé et al., 2004] O. Cappé, A. Guillin, J.-M. Marin & C. P. Robert. Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4):907–929, 2004. URL `http://www.tsi.enst.fr/~cappe/papers/ion02.ps.gz`. | Section 3.2 |
| [Cappé et al., 2005] O. Cappé, E. Moulines & T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005. URL `http://www.tsi.enst.fr/~cappe/ihmm/`. | |
| [Rigouste et al., 2006a] L. Rigouste, O. Cappé & F. Yvon. Inference and evaluation of the multinomial mixture model for text clustering. Rapport technique 2006D004, Télécom Paris, 2006a. URL `http://arxiv.org/cs.IR/0606069`. | Section 3.3 |
| [Olsson et al., 2006b] J. Olsson, O. Cappé, R. Douc & E. Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models. Rapport technique, Lund University, 2006b. URL `http://arxiv.org/math.ST/0609514`. | Section 2.3 |

Table 1: Liste des principaux travaux personnels auxquels il est fait référence dans le mémoire et sections correspondantes.

# Chapitre 1

# Les modèles à données latentes

Ce chapitre introduit de façon aussi concise que possible le contexte général des modèles à données latentes. Les lecteurs au fait du domaine, ou ceux qui ont déjà lu le chapitre 10 de [Cappé et al., 2005], peuvent se dispenser de lire ce chapitre. La section 1.5, qui est plus détaillée, concerne une question importante dans certains modèles mais non traitée dans [Cappé et al., 2005], celle de l'approximation de la vraisemblance elle-même.

## 1.1 Les données latentes

Un modèle à données latentes, cachées, manquantes ou non observées — *latent, hidden, missing* ou *incomplete data model* en anglais[1] — correspond à un modèle de représentation dans lequel le comportement d'ensemble des données observées, notées $Y$, est décrit par la loi de probabilité jointe de $Y$ et de variables non observables, notées $X$. Dans ce document, toutes les lois seront supposées à densité et l'on notera $f(x, y)$ la densité de probabilité jointe de $X$ et de $Y$ par rapport à une mesure $\lambda \otimes \mu$. Le plus souvent seule $\lambda$ sera explicitement figurée, ne serait-ce que pour rendre plus clair les différences existant entre le cas où $X$ est élément d'un ensemble X dénombrable et ceux où X est continu. La mesure de domination $\mu$ joue peu de rôle dans cet exposé et l'on supposera tacitement que l'on a affaire à la mesure de Lebesgue (sur Y ensemble dans lequel $Y$ prend ses valeurs) que l'on notera simplement $dy$ plutôt que $\mu(dy)$.

L'énoncé du principe général appelle quelques remarques. Tout d'abord, un modèle à données latentes est manifestement un modèle probabiliste des données dans lequel on représente la variabilité des observations par une loi de probabilité. La densité de probabilité correspondante s'obtient par marginalisation des variables latentes :

$$\ell(y) \stackrel{\text{def}}{=} \int f(x, y)\lambda(dx) \; . \tag{1.1}$$

Du point de vue de la modélisation, seule la loi $\ell$ correspond à une réalité observable et il existe clairement une infinité de manière de définir, à la fois, les données latentes $Y$ et la loi jointe $f$ pour une loi $\ell$ donnée. Cette remarque est mise à profit notamment dans la

[1] Il y a en fait une nuance subtile entre certains de ces termes et le cadre des *incomplete data models*, tel que défini par [Dempster et al., 1977], inclut des exemples dans lesquels ont est obligé d'écrire (1.1) de façon plus générale (voir [Cappé et al., 2006] pour un exemple). Stricto sensu, on considère ici le cas des modèles à données manquantes (ou cachées). Je préfère toutefois utiliser en français le terme de données ou variables *latentes* qui me semble moins sujet à confusion.

méthode d'accélération de l'algorithme EM proposée par [Fessler & Hero, 1995]. Toutefois, la représentation à partir de données latentes n'est réellement pertinente que si la loi jointe $f$ est simple, c'est à dire, à la fois une loi dans laquelle les relations de dépendance entre les différentes variables sont modélisées de façon élémentaire (indépendance, chaîne de Markov, etc.) ainsi qu'une loi paramétrée de façon très structurée ; cas notamment des lois appartenant à une famille paramétrique de type exponentielle. Même dans ce contexte relativement contraint, la loi des observations possède une structure plus complexe du fait de l'opération de marginalisation. Par rapport à d'autres modèles utilisés pour le traitement statistique des données, les modèles à données latentes sont donc en premier lieu confrontés à des questions de faisabilité puisque même les préoccupations de base en statistique comme le calcul de moments de $Y$ ou l'évaluation de la vraisemblance $\ell(Y)$ sont en général problématiques.

Il n'est donc pas surprenant que les contributions qui aient fortement marqué le domaine soient avant tout des contributions de nature plus algorithmique que réellement statistique, avec, pour citer les étapes les plus importantes : [Kalman & Bucy, 1961] pour le modèle d'état linéaire dans le cas Gaussien ; [Baum et al., 1970] pour le modèle de Markov caché à état fini ; [Dempster et al., 1977] pour l'algorithme EM (*Expectation-Maximization*) ; [Gelfand & Smith, 1990] (entre autres) pour l'utilisation des techniques de Monte Carlo par chaîne de Markov (MCMC) combinée avec le principe de l'augmentation de données ; [Green, 1995] pour le traitement algorithmique des modèles dont la dimension est inconnue. Cette prévalence des contributions de nature algorithmique est logique dans la mesure où la popularité et les succès réellement remarquables des modèles à données latentes sont précisément liés à l'existence de procédures efficaces permettant de mener à bien les deux grandes tâches qui feront l'objet des chapitres 2 et 3 : l'inférence sur les données latentes et sur les paramètres. L'importance des questions liées à l'existence de procédures de calcul efficaces se mesure également au fait que des modèles comme les modèles de Markov cachés (pour lesquels ces aspects sont maîtrisés) sont fréquemment utilisés en lieu et place de modèles supposés plus exacts du comportement des variables observés [Hodgson, 1998; Chib, 1998; Bolot & Grossglauser, 1996]. Dans le cas cas où le modèle naturel consisterait à considérer que l'état est un processus de renouvellement général, on peut par exemple obtenir des approximations raisonnables en considérant des regroupement d'états dans les modèles de Markov cachés [Hodgson, 1998; Bolot & Grossglauser, 1996; Andersen & Nielsen, 1998; Yoshihara et al., 2001; Robert & LeBoudec, 1997].

Cette réputation, parfois un peu excessive, de modèles nécessitant l'utilisation de moyens algorithmiques importants et se prêtant peu à l'étude théorique suscite des résistances dans certains domaines d'application. En économétrie ou dans le domaine des réseaux informatiques, par exemple, les modèles à données latentes sont globalement peu utilisés en dehors du modèle à volatilité stochastique [Hull & White, 1987] et du MMPP (*Markov Modulated Poisson Process*) [Fischer & Meier-Hellstern, 1993]. Il est vrai que dans ces deux domaines, comme probablement dans d'autres, la possibilité de représenter les données à temps continu pour modéliser des phénomènes d'échelle et des flux de données très irréguliers est primordiale. Or, les aspects algorithmiques liés au traitement (calcul des lois de filtrage, de la vraisemblance...) des modèles à données latentes à temps continu et à observations discrètes restent mal connus, sorti du cas linéaire gaussien (voir, notamment, [Elerian et al., 2001; Roberts et al., 2004] et [Beskos et al., 2005]).

## 1.2 Les paramètres

Dans la plupart des cas, le modèle à données latentes comporte des paramètres que l'on notera génériquement par $\theta$ en indiquant par $f(x, y\,;\theta)$, $\ell(y\,;\theta)$, etc. la dépendance des différentes quantités vis à vis du paramètre. Il est clair que cette notation est mieux appropriée au cas de l'approche classique ou « fréquentiste » de l'estimation. Dans le cas de l'approche bayésienne où les paramètres inconnus $\theta$ sont munis d'une densité de probabilité a priori $\pi_0(\theta)$, la notation la plus naturelle est d'utiliser la barre de conditionnement $f(x, y \,|\, \theta)$ de façon à pouvoir écrire la loi des observations sous la forme

$$\ell(y) \stackrel{\text{def}}{=} \iint f(x, y \,|\, \theta)\pi_0(\theta)\lambda(dx)d\theta \; . \tag{1.2}$$

Cette écriture montre bien le lien intrinsèque existant entre les modèles à données latentes et l'approche bayésienne puisque, même en l'absence de données latentes $X$, le simple traitement bayésien des paramètres implique que la loi des observations s'écrive sous la forme d'une loi marginale similaire à (1.1).

Dans ces conditions, et surtout dans le cadre bayésien, il n'est pas aisé de donner une définition très claire de ce qui fait la différence entre les paramètres $\theta$ et les données manquantes. Une fausse piste serait de considérer qu'il s'agit des quantités qui, in fine, sont modélisées de façon déterministe (typiquement les paramètres des lois a priori des paramètres) puisque dans le jargon bayésien ces paramètres de haut niveau sont désignés par le terme d'hyperparamètres. Mon impression est, qu'en règle générale, on ne désigne sous le nom de paramètres que les quantités de dimension constante, tandis que les données latentes ont, en général, une dimension qui croît avec le nombre d'observations. Les « paramètres » sont donc en général, mais pas nécessairement, des quantités que l'on a espoir de pouvoir estimer de façon consistante. Inversement, les données latentes correspondent à des quantités qui ne peuvent être déterminées qu'avec une incertitude probabiliste, quelle que soit la quantité d'information effectivement observée.

Pour prendre un exemple concret, considérons le cas du modèle de mélange fini pour lequel $f(x, y) = w_x g_x(y)$ où $x$ est élément d'un ensemble discret $\mathsf{X}$, $\sum_{x\in\mathsf{X}} w_x = 1$, et $\{g_x\}_{x\in\mathsf{X}}$ sont des densités de probabilité. On suppose de plus que les observations $(Y_0, \ldots, Y_n)$ et les états latents correspondants (souvent appelées indicatrices du mélange dans ce contexte) forment des séquences conjointement IID telle que la loi marginale du couple $(X_k, Y_k)$ soit donnée par $f$. Typiquement, la séquence $(X_k)_{k\in\mathbb{N}}$ constitue les données latentes tandis que les paramètres sont, par exemple, les poids $\{w_x\}_{x\in\mathsf{X}}$ du mélange ainsi que, souvent, des quantités définissant les densités $\{g_x\}_{x\in\mathsf{X}}$. Dans cet exemple, les paramètres sont donc estimable de façon consistante (en le nombre d'observations) sous des conditions peu contraignantes, tout au moins dans le cas où les densités $g_x$ sont représentées de façon paramétrique [Titterington et al., 1985]. Inversement, l'information la plus pertinente que l'on puisse espérer obtenir sur un état particulier est sa distribution de probabilité conditionnelle aux observations $\mathrm{P}\left[X_k = x \,|\, \{Y_k\}_{k\in\mathbb{N}}\right]$[2] qui, dans la plupart des modèles, correspond à une loi non dégénérée (non déterministe) avec $\mathrm{P}_{\{Y_k\}_{k\in\mathbb{N}}}$ probabilité 1.

La situation n'est pas toujours aussi simple notamment dans le cas des modèles de Markov cachés où les états sous-jacents forment une chaîne de Markov de densité initiale $\nu$ et de densité de transition (densité du noyau de transition par rapport à $\lambda$) $q$. Dans

[2] Dans l'approche bayésienne, cette probabilité dépend réellement de toutes les observations puisque l'on marginalise le paramètre inconnu $\theta$. Dans l'approche classique, on aura plutôt tendance à estimer $\theta$ par $\hat{\theta}$ puis à considérer $\mathrm{P}\left[X_k = x \,|\, Y_k\,;, \hat{\theta}\right]$.

ce cas, il est relativement clair que $\nu$ qui correspond à la densité de probabilité de $X_0$ ne saurait être estimée de façon consistante à partir d'une unique séquence d'observations $\{X_k\}_{k\in\mathbb{N}}$, sauf si l'on suppose que $\nu$ est liée à $q$, par exemple, lorsque le modèle est supposé stationnaire et que $\nu$ correspond à la loi stationnaire associé à $q$. Pour ce modèle, et même dans le cadre bayésien où $\nu$ est vue comme une quantité aléatoire, celle-ci est en général désignée sous le nom de paramètre bien qu'il s'agisse d'un paramètre, le plus souvent, non identifiable à partir des données. Dans cet exemple particulier, la distribution initiale $\nu$ ne joue, dans le cas stationnaire, qu'un rôle marginal de toute façon puisqu'il est avantageux du point de vue algorithmique et asymptotiquement équivalent du point de vue statistique de fixer ce paramètre à une valeur arbitraire, typiquement uniforme en $x$ lorsque c'est possible (cf. chapitres 10 et 11 de [Cappé et al., 2005]).

## 1.3 Les modèles à données latentes en traitement du signal

Dans le domaine du traitement du signal, les modèles à données latente se sont longtemps réduits à leur variante linéaire Gaussienne — lorsque la distribution jointe de $X$ et de $Y$ est gaussienne multivariée — qui est à l'origine de deux des avancées emblématiques du domaine : le filtrage de Wiener dans le cas stationnaire [Wiener, 1949], et le filtrage de Kalman dans le cas général des modèles d'état linéaires gaussiens [Kalman & Bucy, 1961]. Tout au long des années 1960, 1970 et jusqu'à la fin des années 1980, un des champs d'intérêt fondamentaux du domaine des signaux et systèmes a consisté à explorer les possibilités de ce modèle de base (extension à des tâches autre que le filtrage stricto sensu), à améliorer le comportement numérique des algorithmes ou à proposer des stratégies approximatives pour des variantes du modèle (notamment l'EKF, *Extended Kalman Filter*). On trouvera un panorama très complet de l'état des connaissances dans ce domaine dans [Kailath et al., 2000]. Bien évidemment, à coté de ces travaux plus méthodologiques, le modèle linéaire gaussien a donné lieu a quantité de travaux plus appliqués dont les techniques de débruitage étudiées au cours de ma thèse [Cappé, 1993] constituaient l'un des avatars. La caractéristique fondamentale du modèle linéaire gaussien est que, tout au moins lorsque ses paramètres sont entièrement connus, la technique d'inférence sur les données latentes (en l'occurrence le calcul de leur moyenne et covariance conditionnelles) est totalement explicite, ne requérant que des opérations linéaires en les données et implémentable de façon récursive. Un des apports méthodologiques du domaine est d'ailleurs d'avoir su reformuler cette technique de façon élégante et plus propice à l'intuition en termes de projection dans l'espace $L^2(\mathrm{P})$. Pour le filtrage proprement dit, l'exposé le plus pédagogique de cette réinterprétation reste [Brockwell & Davis, 1991], le lissage est traité, notamment, dans [Kailath et al., 2000].

Curieusement, l'autre classe de modèles à données latentes pour laquelle l'inférence sur les données latentes peut avoir une solution numérique calculable a longtemps gardé un caractère plus marginal en traitement du signal. Il s'agit bien sûr du cas où les données latentes prennent leur valeur dans un ensemble fini. Le modèle emblématique du domaine, le modèle de Markov caché à état fini, est longtemps resté un outil spécifique de la reconnaissance de parole [Rabiner, 1989]. A dire vrai, ce modèle faisait également partie des centres d'intérêt majeurs du domaine des communications numériques [Bahl et al., 1974], pas forcément sous la même dénomination. Même le modèle plus simple de mélange est resté peu utilisé jusqu'à la fin des années 1980, à l'exception de travaux portant sur la problématique de la déconvolution, notamment pour des applications géophysiques [Kormylo & Mendel, 1982]. A la fin des années 1980, le champ le plus

actif d'utilisation de modèles à données latentes pour le traitement du signal était, si l'on excepte le domaine un peu particulier de la reconnaissance de la parole, celui du traitement d'image autour de l'utilisation des modèles a priori d'image dit « champs de Gibbs » (modèles de Potts et d'Ising), notamment dans les applications de débruitage.

A partir des années 1990, l'utilisation des modèles à données latentes se développe de façon considérable, d'abord essentiellement à partir du cas où les données latentes sont discrètes dans des domaines aussi variés que l'économétrie [Hamilton, 1989], la bio-informatique [Churchill, 1989; Krogh et al., 1994], l'analyse de signaux biomédicaux [Fredkin & Rice, 1992; Ball & Rice, 1992], la reconnaissance d'écriture manuscrite [Kundu et al., 1989], les communications numériques [Kaleh & Vallet, 1994], la détection et le suivi de fréquences [Streit & Barrett, 1990], la robotique [Zhu, 1991], les tâches de reconnaissance [Reynolds, 1995] et de transformation [Stylianou et al., 1998] de locuteurs en traitement de la parole, la séparation de source [Moulines et al., 1997], ... On trouvera une liste raisonnablement exhaustive des travaux des années 1990 pour ce qui concerne plus spécifiquement le cas des modèles de Markov cachés dans [Cappé, 2001b]. A partir de la fin des années 1990, la plus large diffusion des techniques d'inférence par simulation, qu'elles soient en bloc de type MCMC [Gilks et al., 1996; Robert & Casella, 2004] ou en ligne [Liu & Chen, 1998; Doucet et al., 2000, 2001a], a contribué à une véritable explosion du type de modèles considérés avec une ouverture réelle vers des modèles à espace d'état continus. Pour ne citer que quelques exemples mentionnons le modèle à volatilité stochastique en économétrie et finance quantitative [Jacquier et al., 1994; Shephard & Pitt, 1997; Kim et al., 1998; Sandmann & Koopman, 1998], les modèles de ruptures [Green, 1995; Chib, 1998; Lavielle & Lebarbier, 2001; Cappé, 2002; Fearnhead & Clifford, 2003], les modèles d'état linéaires à saut (en anglais, *jump markov models* ou *conditionally Gaussian linear state-space models*) [Carter & Kohn, 1994, 1996; Cappé et al., 1999; Doucet & Andrieu, 2001; Doucet et al., 2001b], les modèles d'état non-linéaires de poursuite [Hue et al., 2002; Ristic et al., 2004; Vermaak et al., 2005].

## 1.4 Les outils de base : l'augmentation de données, l'algorithme EM

Dans le domaine des modèles à données latentes, la contribution qui a eu l'influence la plus durable est très certainement [Dempster et al., 1977], qui si l'on en croit [Meng & Van Dyk, 1997] est entré haut la main dans le hit parade des travaux publiés au cours des trente dernières années. Avec le recul, les apports les plus remarquables de [Dempster et al., 1977] sont d'une part d'avoir démontré la très grande généricité d'une approche qui n'avait été auparavant utilisée que de façon nettement moins transparente, en particulier par [Baum & Eagon, 1967; Baum et al., 1970]. Le second apport particulièrement remarquable, bien que relativement intuitif du point de vue bayésien, et d'avoir montré que l'estimation de paramètres dans les modèles à données latentes est intimement liée à la tâche d'identification de la loi des données latentes conditionnellement aux observations. Le troisième apport est bien sûr l'algorithme EM proprement dit et notamment l'idée d'approximation locale de la vraisemblance par une fonction auxiliaire qui a des prolongements jusque dans les approches d'estimation approximative dites *variationnelles* qui s'inspirent notamment de l'utilisation de l'inégalité de Jensen faite dans l'EM [Jordan et al., 1999].

### 1.4.1 La quantité intermédiaire de l'algorithme EM

On supposera que $\ell(y\,;\theta)$ est strictement positive sur $\mathsf{Y} \times \Theta$ de façon à pouvoir définir la log-vraisemblance

$$\mathrm{L}(y\,;\theta) \stackrel{\text{def}}{=} \log \ell(y\,;\theta)\ . \tag{1.3}$$

La densité conditionnelle des données latentes sachant les observations est définie par

$$p(x \,|\, y\,;\theta) \stackrel{\text{def}}{=} f(x, y\,;\theta)/\ell(y\,;\theta)\ . \tag{1.4}$$

On désigne sous le nom de quantité intermédiaire de l'algorithme EM, la fonction $\Theta \to \mathbb{R}$ indexée par $\theta' \in \Theta$ définie par

$$\theta \mapsto \mathcal{Q}(\theta\,;\theta') \stackrel{\text{def}}{=} \int \log f(x, y\,;\theta) p(x \,|\, y\,;\theta')\, \lambda(dx)\ . \tag{1.5}$$

L'algorithme EM est une technique de maximisation itérative de la vraisemblance où les valeurs successives du paramètre s'obtiennent par la relation de récurrence

$$\theta^{i+1} = \arg\max_{\theta \in \Theta} \mathcal{Q}(\theta\,;\theta^i)\ , \tag{1.6}$$

$\theta^0$ étant une initialisation arbitraire. De nombreuses variantes existent et en particulier le fait de réaliser exactement la maximisation dans (1.6) n'est pas indispensable [Meng & Rubin, 1993; Fessler & Hero, 1995; Lange, 1995]. Le calcul de la quantité intermédiaire de l'algorithme EM (1.5) est dite « étape E » de l'algorithme et la maximisation de (1.6) « étape M ».

A partir de (1.3) et de (1.4), on peut réécrire la quantité intermédiaire de l'EM sous la forme

$$\mathcal{Q}(\theta\,;\theta') = \mathrm{L}(y\,;\theta) - \mathcal{H}(\theta\,;\theta')\ , \tag{1.7}$$

où

$$\mathcal{H}(\theta\,;\theta') \stackrel{\text{def}}{=} -\int \log p(x \,|\, y\,;\theta) p(x \,|\, y\,;\theta')\, \lambda(dx)\ . \tag{1.8}$$

La quantité intermédiaire de l'algorithme EM diffère donc de la log-vraisemblance par un terme que l'on reconnaît aisément comme l'entropie de la loi conditionnelle $p(\cdot \,|\, y\,;\theta')$ [Cover & Thomas, 1991] (il ne s'agit pas de l'entropie conditionnelle à proprement parler puisqu'on n'intègre pas sous la loi de $Y$). Le point le plus remarquable est que les incréments de $\mathcal{H}(\theta\,;\theta')$,

$$\mathcal{H}(\theta\,;\theta') - \mathcal{H}(\theta'\,;\theta') = -\int \log \frac{p(x \,|\, y\,;\theta)}{p(x \,|\, y\,;\theta')} p(x \,|\, y\,;\theta')\, \lambda(dx)\ , \tag{1.9}$$

correspondent à la divergence de Kullback-Leibler (aussi appelé entropie relative en théorie de l'information) entre les lois conditionnelles $p$ indexées respectivement par $\theta$ et $\theta'$. La positivité de cette quantité implique la propriété bien connue d'accroissement monotone de la log-vraisemblance par l'algorithme EM.

En dérivant (1.8) par rapport à $\theta$ (en supposant les conditions de régularité suffisantes pour pouvoir dériver sous l'intégrale), on vérifie aisément que

$$\nabla_\theta \mathcal{H}(\theta\,;\theta')\big|_{\theta=\theta'} = 0$$

où $\nabla_\theta$ désigne le gradient par rapport au paramètre $\theta$. Cette égalité montre que l'accroissement de la vraisemblance à chaque itération de l'algorithme EM est strictement positif, sauf si $\theta'$ est un point stable (tel que $\nabla_\theta \mathrm{L}(y\,;\theta) = 0$) de la log-vraisemblance.

### 1.4.2 Les relations de Fisher et Louis

L'égalité précédente implique que le gradient de la log-vraisemblance peut s'écrire sous la forme

$$\nabla_\theta \mathrm{L}(y\,;\theta') = \int \nabla_\theta \log f(x,y\,;\theta)|_{\theta=\theta'}\, p(x\,|\,y\,;\theta')\,\lambda(dx)\;, \tag{1.10}$$

dite *relation de Fisher*, probablement du fait des commentaires de B. Effron dans la discussion de [Dempster et al., 1977]. Sous conditions de différentiabilité au second ordre, on obtient une seconde relation, dite *de Louis*, ou également *missing information principle* [Louis, 1982] :

$$\begin{aligned}\nabla^2_\theta \mathrm{L}(y\,;\theta') = &\int \nabla^2_\theta \log f(x,y\,;\theta)\big|_{\theta=\theta'}\, p(x\,|\,y\,;\theta')\,\lambda(dx)\\ &- \int \nabla^2_\theta \log p(x\,|\,y\,;\theta)\big|_{\theta=\theta'}\, p(x\,|\,y\,;\theta')\,\lambda(dx)\;,\end{aligned} \tag{1.11}$$

que l'on peut réécrire un peu différemment de façon à en faire ressortir la symétrie :

$$\begin{aligned}\nabla^2_\theta \mathrm{L}(y\,;\theta') + \left[\nabla_\theta \mathrm{L}(y\,;\theta')\right]\left[\nabla_\theta \mathrm{L}(y\,;\theta')\right]^t = \int \Big\{ &\nabla^2_\theta \log f(x,y\,;\theta)\big|_{\theta=\theta'}\\ &+ \left[\nabla_\theta \log f(x,y\,;\theta)|_{\theta=\theta'}\right]\left[\nabla_\theta \log f(x,y\,;\theta)|_{\theta=\theta'}\right]^t\Big\} p(x\,|\,y\,;\theta')\,\lambda(dx)\;.\end{aligned} \tag{1.12}$$

### 1.4.3 Cas des familles exponentielles

Dans la plupart des contextes où l'algorithme EM est utilisé, et en particulier pour que la maximisation associée à (1.6) soit réalisable de façon exacte, la loi jointe associée au modèle complet $\{f(x,y\,;\theta)\}_{\theta\in\Theta}$ appartient à une famille exponentielle telle que

$$f(x,y\,;\theta) = \exp\left[\psi(\theta)^t S(x,y) - c(\theta)\right] h(x,y)\;, \tag{1.13}$$

où $S$ est $\psi$ sont des fonctions à valeur vectorielle dite respectivement *statistiques suffisantes* et *paramétrisation naturelle* du modèle (et l'exposant $t$ désigne la transposition).

Dans ce contexte, la quantité intermédiaire de l'algorithme EM s'écrit

$$\mathcal{Q}(\theta\,;\theta') = \psi(\theta)^t \left[\int S(x,y) p(x\,|\,y\,;\theta')\,\lambda(dx)\right] - c(\theta) + \mathcal{C}(y,\theta')\;, \tag{1.14}$$

où le dernier terme ne joue aucun rôle dans la mesure ou il ne dépend pas de $\theta$.

De même, les relations de Fisher et Louis prennent une forme plus simple, donnée ici dans le cas où $\psi(\theta) = \theta$ (paramétrisation naturelle) pour simplifier les expressions,

$$\nabla_\theta \mathrm{L}(y\,;\theta) = \left[\int S(x,y) p(x\,|\,y\,;\theta)\,\lambda(dx)\right] - \nabla_\theta c(\theta)\;, \tag{1.15}$$

$$\begin{aligned}\nabla^2_\theta \mathrm{L}(y\,;\theta) + \left[\nabla_\theta \mathrm{L}(y\,;\theta)\right]\left[\nabla_\theta \mathrm{L}(y\,;\theta)\right]^t = &\\ - \nabla^2_\theta c(\theta) + \nabla_\theta c(\theta)\nabla_\theta c(\theta)^t + &\int S(x,y)S(x,y)^t p(x\,|\,y\,;\theta)\lambda(dx)\\ - \nabla_\theta c(\theta)\left[\int S(x,y) p(x\,|\,y\,;\theta)\right]^t &- \left[\int S(x,y) p(x\,|\,y\,;\theta)\right]\nabla_\theta c(\theta)^t\;.\end{aligned} \tag{1.16}$$

La remarque importante est que pour évaluer ces quantités, il suffit de calculer les moments conditionnels $\mathrm{E}(S(X,Y) \,|\, Y\,;\theta]$ et $\mathrm{E}\left[S(X,Y)S(X,Y)^t \,|\, Y\,;\theta\right]$. Dans le cas des familles exponentielles, la praticabilité de l'algorithme EM ou de l'évaluation des dérivées de la log-vraisemblance se réduit donc à celle du calcul de l'espérance conditionnelle des statistiques suffisantes du modèles complet. On peut en particulier dans ce cas décrire de façon totalement équivalente les trajectoires de l'algorithme EM soit par la donnée de la séquence de paramètres $\{\theta_i\}_{i\in\mathbb{N}}$ soit par celle des espérances conditionnelles de la statistique suffisante $\{S^i\}_{i\in\mathbb{N}}$, où $S^i = \mathrm{E}\left[S(X,Y) \,|\, Y\,;\theta^i\right]$, principe qui est utilisée notamment dans l'algorithme SAEM (*Stochastic Approximation EM*) de [Delyon et al., 1999].

Même s'il s'agit d'une remarque moins directement liée à notre propos, il est intéressant de noter que pour une famille en paramétrisation naturelle ($\psi(\theta) \equiv \theta$) on peut réécrire (1.15) et (1.16) sous une forme très révélatrice de la nature de l'algorithme EM :

$$\nabla_\theta \mathrm{L}(Y\,;\theta) = \mathrm{E}\left[S(X,Y) \,|\, Y\,;\theta\right] - \mathrm{E}\left[S(X,Y)\,;\theta\right] \ . \tag{1.17}$$

Où l'on retrouve en particulier le fait que l'algorithme EM prend la forme très simple

$$\mathrm{E}\left[S(X,Y)\,;\theta^{i+1}\right] = \mathrm{E}\left[S(X,Y) \,|\, Y\,;\theta^i\right] \ ,$$

c'est à dire que $\theta^{i+1}$ s'obtient comme l'estimateur du maximum de vraisemblance dans le modèle complet (c'est à dire celui où l'on observerait à la fois $X$ et $Y$) pour une pseudo réalisation de la statistique suffisante égale à $\mathrm{E}\left[S(X,Y) \,|\, Y\,;\theta^i\right]$.

### 1.4.4 Difficultés et limitations

Une critique fréquemment faite à l'algorithme EM tel que présenté par [Dempster et al., 1977] est qu'il converge lentement. Cette vitesse de convergence parfois lente est la contrepartie de son extrême simplicité d'implémentation : c'est en effet un des seuls algorithmes d'optimisation numérique implémentable intégralement par un novice, notamment dans le cas où les paramètres $\theta$ du modèle sont contraints ; ce problème étant en quelque sorte caché dans l'EM puisque la maximisation de (1.6) est en général explicite, même en tenant compte des contraintes. C'est également un algorithme qui conduit à des implémentations particulièrement simples à valider du fait de la propriété d'augmentation monotone de la vraisemblance (« implémentation » étant compris ici à la fois comme incluant les formules utilisées ainsi que le code censé implémenter ces formules)[3].

La question de la vitesse de convergence, qui a fait l'objet d'un certain nombre de travaux tout au long des années 19990 [Lange, 1995; Jamshidian & Jennrich, 1997], me semble moins pertinente aujourd'hui. D'une part, l'augmentation de la vitesse des processeurs et l'utilisation plus courante d'algorithmes intensifs de type Monte Carlo on rendu la « lenteur » de l'EM toute relative. Par ailleurs, au moins dans les contextes où l'étape E est réalisé exactement, la relation de Fisher (1.10) montre que le calcul du gradient de la log-vraisemblance est d'un coût de calcul équivalent à l'étape E (voir aussi le cas des modèles exponentiels dans la section 1.4.3 ci-dessus). Il est donc possible d'utiliser, à la place de l'étape M, une routine d'optimisation de type quasi-Newton ou gradient

[3] Même s'il peut sembler plus anecdotique, ce dernier point m'apparaît avec le recul comme particulièrement important, notamment comparé aux techniques de type MCMC (*Monte Carlo par Chaîne de Markov*) pour lesquelles, outre la question du réglage des paramètres de l'algorithme, il existe une difficulté inhérente de validation de l'algorithme (liée à l'absence de relation déterministe satisfaite par les résultats d'exécution) qui s'avère parfois insurmontable sans un certain niveau d'expertise.

conjugué qui permet effectivement d'atteindre une convergence d'ordre supérieur (quadratique), au prix parfois d'une certaine difficulté à gérer les contraintes portant sur les paramètres, cf. exemples dans [Cappé et al., 1998a, 2005].

Les cas où l'étape E est réalisable exactement mais où l'étape M ne l'est pas correspondent en général à des modèles complets n'appartenant pas à une famille exponentielle ou bien à l'utilisation d'une paramétrisation spécifique se traduisant par l'existence de contraintes sur $\theta$. Une méthodologie souvent applicable dans ce cas est l'approche proposée par [Meng & Rubin, 1993; Fessler & Hero, 1995] dans laquelle l'étape M est réalisée partiellement par relaxation sur les paramètres. La limitation la plus préoccupante de l'algorithme EM concerne plutôt les cas où l'étape E n'est pas réalisable exactement. Concrètement, c'est la situation la plus fréquente dès lors que les données latentes ne sont pas à valeurs dans un ensemble fini (et de cardinal faible), en excluant quelques situations particulières comme le modèle linéaire gaussien. A mon sens, cette question reste complètement d'actualité même si un certain nombre d'approches ont d'ores et déjà été étudiées ; approches dont la caractéristique commune consiste à remplacer l'étape E par diverses formes d'approximation de type Monte Carlo [Wei & Tanner, 1991; Booth & Hobert, 1999; Fort & Moulines, 2003] ou Robbins-Monro (où la moyennisation se fait de façon récursive) [Celeux & Diebolt, 1985; Delyon et al., 1999].

Une difficulté supplémentaire dans ce contexte est lié au fait que la simulation exacte de variables indépendantes sous les lois conditionnelles $p(x \mid y\,;\theta)$ ou même la simple évaluation des valeurs de la densité $p(x \mid y\,;\theta)$ est le plus souvent impossible dans les cas où l'étape E de l'EM n'est pas directement réalisable. En effet si $f(x, y\,;\theta)$ est bien connu analytiquement, la densité conditionnelle $p(x \mid y\,;\theta)$ en diffère par le facteur $\ell(y\,;\theta)$ dont le calcul implique précisément d'être capable d'intégrer $f(x, y\,;\theta)$ par rapport à $x$. La densité conditionnelle $p(x \mid y\,;\theta)$ n'est donc, dans ces situations, connue qu'à une constante de proportionnalité près. A part la méthode d'acceptation-rejet, dont l'applicabilité est tout de même très limitée, les techniques envisageables dans cette situation, qu'il s'agisse de simulations MCMC [Robert & Casella, 2004], d'échantillonnage préférentiel (*importance sampling* en anglais) autonormalisé ou « bayésien » [Geweke, 1989], ou de techniques de Monte Carlo séquentielles [Cappé et al., 2004], conduisent à des simulations non indépendantes, sous une loi qui ne fait qu'approximer la loi cible $p(x \mid y\,;\theta)$.

## 1.5 Evaluation de la vraisemblance

Dans les cas mentionnés ci-dessus, où l'étape E de l'algorithme EM n'est pas réalisable exactement, le calcul de la vraisemblance $\ell(Y\,;\theta)$ (ou de son logarithme) est également impossible pour la même raison : la difficulté d'intégrer par rapport à $x$ la densité $f(x, y\,;\theta)$. Dans ce contexte l'estimation au sens du maximum de vraisemblance implique donc de maximiser une fonction qui n'est pas calculable explicitement. Nous avons toutefois vu ci-dessus que l'approche EM (ou d'autres méthodes imaginables comme des méthodes de gradient — cf. chapitres 10 et 11 de [Cappé et al., 2005] — permettent d'aborder le problème de maximisation sans chercher à résoudre celui de l'évaluation de la vraisemblance. Dans certains contextes cependant, l'évaluation de la vraisemblance est un objectif d'intérêt en lui-même, notamment lorsqu'il s'agit de tester l'adéquation d'un modèle avec des données. C'est en particulier le cas lorsque le modèle à donnée latente est utilisé à des fins de classification de données.

Le pendant bayésien de cette question, c'est à dire, le calcul du *facteur de Bayes*

$$\frac{\int \ell_1(Y\,;\theta)d\theta}{\int \ell_2(Y\,;\theta)d\theta}\,,$$

où $\ell_1$ et $\ell_2$ correspondent à deux modèles à comparer, est également connu pour être un problème délicat. De même que l'approche EM ne résout par la question de l'évaluation de la vraisemblance, les méthodes MCMC qui permettent de simuler le paramètres $\theta$ sous sa loi a posteriori sachant les observations $Y$ dans le $i$-ème modèle ne permettent pas d'évaluer $\int \ell_i(Y\,;\theta)d\theta$. La réponse la plus communément admise à cette question consiste faire coexister les deux modèles à comparer au sein d'un formalisme bayésien intégré de façon à utiliser un algorithme MCMC susceptible de visiter simultanément les deux modèles [Green, 1995; Robert & Casella, 2004]. On obtient alors une estimation du facteur de Bayes grâce aux fréquences empiriques de visite, par l'échantillonneur, des modèles concurrents.

Cette section fait le point sur la principale approche qui a été utilisée pour répondre à cette difficulté (autre que l'approche MCMC à laquelle il est fait référence ci-dessus) qui consiste à utiliser l'échantillonnage préférentiel. Il s'agit d'un sujet qui n'a pas été abordé dans [Cappé et al., 2005] dans la mesure où, dans le cas du traitement classique (non bayésien) des modèles de Markov cachés, il est possible de décomposer la log-vraisemblance sous une forme qui permet son calcul exact, dans les cas les plus simples, ou tout au moins son approximation par des techniques de Monte Carlo séquentielles (cf. section 2.3.3). Pour cette raison, cette section est un peu plus développée que les précédentes.

L'échantillonnage préférentiel consiste à utiliser des simulations $\xi^1,\ldots,\xi^m$ sous une loi, dite d'*instrumentale*, $qd\lambda$ de façon à approximer $\int f(x)p(x)\lambda(dx)$, où $p$ est une densité de probabilité, par

$$\frac{1}{m}\sum_{i=1}^{m} f(\xi^i)\frac{p(\xi^i)}{q(\xi^i)}\,.$$

On peut substituer à cet estimateur sa version *auto-normalisée*

$$\frac{\sum_{i=1}^m f(\xi^i)\frac{p(\xi^i)}{q(\xi^i)}}{\sum_{i=1}^m \frac{p(\xi^i)}{q(\xi^i)}}$$

qui a l'avantage de pouvoir être utilisée également dans les cas où $p$ et/ou $q$ ne sont connus qu'à un facteur d'échelle près. L'étude des propriétés de base (consistance, normalité asymptotique) est très classique, et je me réfère à [Geweke, 1989] et à la section 9.1 de [Cappé et al., 2005] pour les formules qui seront utilisées ci-dessous.

Dans le cas qui nous intéresse, l'échantillonnage préférentiel consiste à simuler des variables $\xi^1,\ldots,\xi^m$ sous une loi $qd\lambda$ de façon à approximer la vraisemblance par

$$\hat{\ell}^m(Y\,;\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{f(\xi^i,Y\,;\theta)}{q(\xi^i)}\,. \tag{1.18}$$

Cette idée a été popularisée dans la littérature statistique par [Geyer & Thompson, 1992; Geyer, 1996] (avec une façon particulière de choisir $q$) mais elle est connue depuis le début des années 1980 en recherche opérationnelle sous le nom de *méthode de la fonction de score* [Glynn, 1990; Rubinstein & Shapiro, 1993]. Curieusement d'ailleurs, l'aspect qui vaut à cette approche le nom de méthode de la fonction de score n'est pas exploité

par [Geyer & Thompson, 1992] : il s'agit de la remarque que la maximisation numérique de l'approximation $\hat{\ell}^m(Y\,;\theta)$ est facilitée par le fait que le gradient de cette fonction est calculable via une approximation du même type que (1.21) dans la mesure où

$$\nabla_\theta \hat{\ell}^m(Y\,;\theta) = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log f(\xi^i, Y\,;\theta) \frac{f(\xi^i, Y\,;\theta)}{q(\xi^i)} \;. \tag{1.19}$$

Les équations (1.19) et (1.18) suggèrent d'ailleurs naturellement une approximation du *gradient de la log-vraisemblance* sous la forme suivante

$$\nabla_\theta \hat{\mathrm{L}}^m(Y\,;\theta) = \frac{\sum_{i=1}^m \nabla_\theta \log f(\xi^i, Y\,;\theta) \frac{f(\xi^i,Y\,;\theta)}{q(\xi^i)}}{\sum_{i=1}^m \frac{f(\xi^i,Y\,;\theta)}{q(\xi^i)}} \;, \tag{1.20}$$

que l'on reconnaît également comme une approximation de la formule de Fisher (1.10) par l'approche d'échantillonnage préférentiel auto-normalisée. L'intérêt des approximations (1.18)–(1.20) et d'être compatibles entre elles : (1.20) est bien le gradient du logarithme de (1.18), ce qui permet, par exemple, d'utiliser des procédures standards d'optimisation numérique pour maximiser l'approximation (1.18) vis à vis du paramètre $\theta$.

La question principale que pose cette approche est bien sûr celui du choix de la densité instrumentale $q$. Pour progresser sur cette question, dans le cas où les tirages $\xi^i$ sont des variables IID de densité $q$ le théorème central limite usuel donne directement l'expression suivante de la variance asymptotique de (1.18)[4] :

$$v(\theta) = \ell^2(Y\,;\theta) \left( \int \frac{p^2(x \,|\, Y\,;\theta)}{q^2(x)}\, q(x)\lambda(dx) - 1 \right) \;,$$

sous réserve que le terme intégral soit effectivement bien défini. Ce terme possède une interprétation probabiliste puisque

$$\log \int \frac{p^2(x \,|\, Y\,;\theta)}{q^2(x)}\, q(x)\lambda(dx) \stackrel{\text{def}}{=} D_2\left[p(\cdot \,|\, Y\,;\theta) \parallel q\right]$$

est la divergence de Rényi, parfois également appelée $\alpha$-divergence, d'ordre $\alpha = 2$, entre $p(\cdot \,|\, Y\,;\theta)$ et $q$ [Basseville, 1989]. Par application de l'inégalité de l'inégalité de Jensen, on montre facilement que

$$\begin{aligned} D_2\left[p(\cdot \,|\, Y\,;\theta) \parallel q\right] &= \log \int \frac{p(x \,|\, Y\,;\theta)}{q(x)}\, p(x \,|\, Y\,;\theta)\lambda(dx) \\ &\geq \int \log \frac{p(x \,|\, Y\,;\theta)}{q(x)}\, p(x \,|\, Y\,;\theta)\lambda(dx) \stackrel{\text{def}}{=} K\left[p(\cdot \,|\, Y\,;\theta) \parallel q\right] \;, \end{aligned}$$

où $K$ désigne la divergence de Kullback-Leibler, ce qui implique notamment que la variance normalisée $v(\theta)/\ell^2(Y\,;\theta)$ est minoré par $\exp K\left[p(\cdot \,|\, Y\,;\theta) \parallel q\right] - 1$. Ainsi sous la condition essentielle, et non triviale, que la divergence de Rényi $D_2\left[p(\cdot \,|\, Y\,;\theta) \parallel q\right]$ soit finie, la variance asymptotique de l'estimateur de la vraisemblance $\hat{\ell}^m(Y\,;\theta)$ est contrôlé par des quantités qui s'interprètent fondamentalement comme des mesures de proximité entre la loi conditionnelle $p(\cdot \,|\, Y\,;\theta)$ et la loi instrumentale $q$.

Cette constatation suggère manifestement d'utiliser comme loi instrumentale la loi conditionnelle $p(\cdot \,|\, Y\,;\theta)$, solution adoptée, par exemple, dans [Buntine & Jakulin, 2004].

[4] On parle bien évidemment ici d'un résultat conditionnel aux observations $Y$.

Cette proposition rencontre malheureusement de sévères difficultés. Tout d'abord, l'utilisation directe de $q = p(\cdot \mid Y ; \theta)$ dans (1.18) conduit à un résultat paradoxal qui reflète simplement le fait que si l'on connaît $p(\cdot \mid Y ; \theta)$ exactement, on connaît également la vraisemblance $\ell(Y ; \theta)$ sans avoir besoin de recourir à des simulations. Conformément à ce qui a été dit à la fin du paragraphe 1.4.4, la situation la plus courante (en dehors de celle où le calcul de la vraisemblance est explicite) est celle où l'on sait générer approximativement des variables $\xi^i$ distribuées selon la densité $p(\cdot \mid Y ; \theta')$, par exemple par des approches MCMC, mais où cette dernière n'est connue qu'à un facteur de normalisation près, qui est précisément égal à la vraisemblance $\ell(Y ; \theta)$ à évaluer. Pour remédier à cette difficulté, on peut songer à utiliser la variante auto-normalisée de (1.18) sous la forme

$$\hat{\ell}^m(Y ; \theta) = \frac{\sum_{i=1}^m \frac{f(\xi^i, Y ; \theta)}{f(\xi^i, Y ; \theta)}}{\sum_{i=1}^m \frac{f_X(\xi^i ; \theta)}{f(\xi^i, Y ; \theta)}} = \left( \frac{1}{m} \frac{f_x(\xi^i ; \theta)}{f(\xi^i, Y ; \theta)} \right)^{-1} ,$$

où $f_X$ désigne la densité marginale de $f$, c'est à dire la loi *a priori* des données latentes $X$. Malheureusement, l'expression même de l'estimateur obtenu montre qu'il repose en fait sur un calcul d'échantillonnage préférentiel avec comme loi cible la loi a priori $f_X$ des données latentes et comme loi instrumentale la loi $p(\cdot \mid Y ; \theta)$ conditionnelle aux observations. Compte tenu de ce qui a été dit précédemment, un tel estimateur ne peut être asymptotiquement normal que si $D_2 \left[ f_X \parallel p(\cdot \mid Y ; \theta) \right]$ est finie, ce qui ne sera quasiment jamais le cas dès que X n'est pas fini (ou compact) dans la mesure où la loi a posteriori $p(\cdot \mid Y ; \theta)$ est plus concentrée que la loi à priori $f_X$.

En pratique, l'estimateur précédent n'est donc quasiment jamais utilisable, même si on peut envisager des stratégie visant à le rendre plus robuste [Hesterberg, 1995; Raftery et al., 2006]. On peut d'ailleurs noter que le choix inverse, celui d'utiliser (1.18) avec comme loi instrumentale $q = f_X$, la loi a priori des données latentes, conduit à la condition $D_2 \left[ p(\cdot \mid Y ; \theta) \parallel f_X \right] < \infty$ qui elle est plus raisonnable pour la raison mentionnée ci-dessus. Ce choix de l'a priori pour simuler les données latentes $\xi^i$ — c'est à dire, de l'utilisation de la méthode de Monte Carlo classique — est notamment préconisé par [Griffiths & Steyvers, 2002] pour calculer la vraisemblance dans le modèle LDA (*Latent Dirichlet Association*) que nous étudions dans le cadre de la thèse de Loïs Rigouste[5]. Ce choix de l'a priori comme loi instrumentale est toutefois peu efficace en grande dimension.

La solution proposée par [Geyer & Thompson, 1992; Geyer, 1996] consiste à utiliser comme loi instrumentale la loi conditionnelle $p(\cdot \mid Y ; \theta')$ paramétrée par une autre valeur $\theta'$, si possible proche de $\theta$. La vraisemblance est alors approximée par

$$\hat{\ell}^m(Y ; \theta) = \frac{1}{m} \sum_{i=1}^m \frac{f(\xi^i, Y ; \theta)}{p(\xi^i \mid Y ; \theta')} = \ell(Y ; \theta') \frac{1}{m} \sum_{i=1}^m \frac{f(\xi^i, Y ; \theta)}{f(\xi^i, Y ; \theta')} . \tag{1.21}$$

Pour les raisons déjà évoqué ci-dessus, seule la seconde forme de cet estimateur est exploitable en pratique dans le cas où la vraisemblance n'est pas calculable directement. On obtient donc un estimateur du rapport du rapport $\ell(Y ; \theta) / \ell(Y ; \theta')$, d'où le nom de *simulated likelihood ratio* parfois utilisé pour décrire cette façon de procéder. Malgré

[5] Il est d'ailleurs à noter que dans la littérature d'apprentissage statistique, l'approche la plus populaire pour calculer la vraisemblance dans ce type de modèles consiste à utiliser des approximations déterministes de type « variationnel » [Blei et al., 2002; Minka & Lafferty, 2002]. L'intégration de ce types d'approches dans un discours statistique pose toutefois question dans la mesure où l'on ne dispose d'aucun moyen, fusse-t-il principalement théorique comme l'est, dans un certain sens, l'asymptotique en $m$ dans les méthodes de Monte Carlo, pour contrôler l'erreur d'approximation.

la présence de la constante $\ell(Y\,;\theta')$ inconnue la seconde forme de (1.21) permet manifestement d'utiliser l'approximation $\hat{\ell}^m(Y\,;\theta)$ pour des tests d'hypothèse de type rapport de vraisemblance, voire pour l'estimation de paramètre en maximisant l'approximation $\hat{\ell}^m(Y\,;\theta)$ par rapport à $\theta$ ($\theta'$ étant fixé). On vérifie d'ailleurs que la présence d'une constante inconnue n'invalide pas les remarques faites précédemment à propos du calcul du gradient et que l'on peut directement utiliser (1.20) comme approximation du score.

Dans un contexte statistique tout au moins, le problème posé par cette approche est clairement celui du choix du point de référence $\theta'$. Malheureusement, si les observations $Y$ sont informatives sur le paramètre $\theta$ et les données latentes $X$, les lois $p(x\,|\,Y\,;\theta)$ et $p(x\,|\,Y\,;\theta')$ ne sont comparables que pour des valeurs du paramètre $\theta$ très proches de $\theta'$, et ce d'autant plus que le nombre d'observations augmente. C'est cet effet que nous avons mis en évidence, dans le cadre d'hypothèses très simples, dans [Douc et al., 2002]. En pratique, la solution qui consiste à explorer tout l'espace des paramètres à partir d'un unique point de référence $\theta'$ est le plus souvent illusoire et la question du (ou des) choix de $\theta'$ garantissant une bonne approximation de la vraisemblance reste un problème ouvert. En s'inspirant des idées exposées dans [Rubinstein & Kroese, 2004] on peut d'ailleurs imaginer des aproches adaptatives dans lesquels $\theta'$ est ajusté, à partir des simulations, de façon à améliorer l'approximation de la vraisemblance dans certaines régions de l'espace des paramètres.

## 1.6 Les modèles conditionnels en apprentissage

Dans de nombreux domaines d'application des modèles de Markov cachés (type de modèle à données latentes avec lequel j'ai le plus d'expérience) ceux-ci sont utilisés, au moins pendant une phase dite d'*apprentissage*, de façon *semi-supervisée* : typiquement l'état latent est lié à une quantité observable qui est une étiquette (en anglais *label*) associée à l'observation dans une application de type classification, ou bien, à la valeur d'une variable de réponse dans un application de type régression. Au cours de la phase d'apprentissage, on considère donc que l'état $X$ est au moins partiellement observé, en sus des observations $Y$ proprement dites. Dans la phase de *test*, seules les observations $Y$ sont supposés disponibles et le modèle, équipé des paramètres ajustés lors de la phase d'apprentissage, est sollicité pour évaluer les probabilités à posteriori $p(x\,|\,Y\,;_h\,at\theta)$ ou déterminer l'état le plus probable $\arg\max_{x\in\mathsf{X}} p(x\,|\,Y\,;\hat{\theta})$.

En reconnaissance de la parole, par exemple, l'apprentissage se fait à partir de données segmentées et étiquetées, au minimum en mots, et souvent en phonèmes. Cette segmentation n'est pas suffisante pour déterminer entièrement les états du modèle (qui d'ailleurs dans ce cas n'ont pas une réalité physique totalement incontestable) mais elle permet de réduire fortement l'éventail des états possibles lors de l'apprentissage. Typiquement, les données correspondant à la prononciation d'un mot particulier ne servent à estimer que les paramètres du modèle de ce mot. Toutefois chaque modèle de mot étant un modèle de Markov caché, il reste une certaine incertitude probabiliste sur l'état *à l'intérieur* du modèle. On trouvera dans [Rigouste et al., 2005b] un autre exemple de ce type d'apprentissage semi-supervisé pour une application en traitement du langage naturel.

Ce contexte a donné lieu à l'utilisation de modèles qui bien qu'étant très liés aux modèles évoqués précédemment ne comportent pas à proprement parler de données latentes. Il s'agit des modèles dits « conditionnels », « discriminants » [Ng & Jordan, 2001] ou *maximum entropy Markov models* [McCallum et al., 2000], *conditional random fields (CRF)* [Lafferty et al., 2001] pour les modèles liés aux modèles de Markov cachés. Dans

cette approche, on spécifie uniquement la forme de la loi conditionnelle $p(x \mid y\,; \theta)$ sans qu'il soit nécessaire de spécifier la loi complète du modèle $f(x, y\,; \theta)$. Si on ne spécifie pas la loi complète $f(x, y\,; \theta)$, il ne s'agit plus à proprement parler d'un modèle probabiliste des observations — dit aussi modèle « génératif » dans le domaine de l'apprentissage statistique — mais d'un modèle statistique conditionnel (l'exemple le plus simple étant le cas de la régression logistique considéré dans [Ng & Jordan, 2001]). Typiquement, dans un contexte complètement supervisé, les paramètres $\theta$ de ce modèle conditionnel sont estimés durant la phase d'apprentissage en maximisant, par rapport aux paramètres $\theta$, $p(X \mid Y\,; \theta)$ c'est à dire la probabilité conditionnelle des variables d'états (supposées ici observées) sachant les observations. Le calcul de la log-vraisemblance et de son gradient dans ces modèles conditionnels présente souvent une certaine ressemblance formelle avec le cas des modèles à données latentes (voir notamment [Lafferty et al., 2001; Wallach, 2002] pour le cas des *conditional random fields* à structure markovienne).

# Chapitre 2

# Inférence sur les variables latentes dans les modèles de Markov cachés

La première partie de ce chapitre (sections 2.1 et 2.2) présente le contenu des chapitres 3 et 4 de [Cappé et al., 2005] à propos des équations de lissage à horizon fixe dans les modèles de Markov cachés. Il ne s'agit pas à proprement parler d'un thème nouveau et la contribution de [Cappé et al., 2005] consiste surtout en un travail d'unification permettant de généraliser le cadre introduit par [Baum et al., 1970] afin de retrouver les trois variantes du lissage utilisées en pratique. Un des intérêts de ce travail est de montrer, qu'en dehors du cas où l'espace d'état est fini où toutes les façons de procéder sont relativement équivalentes, ces trois variantes ont des intérêts ou défauts propres qui les rendent plus ou moins appropriées à certaines utilisations, tant algorithmiques que théoriques. La section 2.2 présente une vision originale, due à [Zeitouni & Dembo, 1988], qui permet d'implémenter certains calculs de lissage de façon récursive.

La section 2.3 est consacrée à l'utilisation de méthodes de Monte Carlo séquentielles et présente la contribution de [Cappé, 2001a; Olsson et al., 2006a] sur le thème de l'approximation de fonctionnelles de l'état appliquée à l'estimation de paramètres fixes.

## 2.1 Lissage dans les modèles de Markov cachés [Cappé et al., 2005, chapitre 3]

Pour donner les arguments principaux du chapitre 3 de [Cappé et al., 2005], nous considérons ici le cas très simple d'une chaîne de Markov $\{X_n\}_{n\in\mathbb{N}}$ sur un espace X discret, de loi initiale $\nu$ et de matrice de transition $Q$. Une tâche, apparemment très simple, qui met en évidence l'ensemble des arguments concernant le lissage dans les modèles de Markov cachés consiste à évaluer les probabilités conditionnelles $\mathrm{P}_\nu(X_k = x | X_0 \in G_0, \ldots, X_n \in G_n)$ où $G_0, \ldots, G_n$ sont des ensembles arbitraires de $X$. L'exemple les plus simple consistant à déterminer la loi de la chaîne accrochée aux deux extrémités $\mathrm{P}_\nu(X_k = x | X_0 = x_0, X_n = x_n)$. De façon plus générale, on cherche à déterminer, à tous instants, la loi de la chaîne sachant qu'elle passe par (ou qu'elle évite) certaines régions de l'espace d'état.

Bien que très élémentaire ce cadre constitue bien un exemple de lissage dans les modèles de Markov cachés puisque qu'en définissant les « observations » $Y_n$ comme les variables aléatoires $Y_n = \mathbb{1}_{G_n}(X_n)$, le couple $(X_n, Y_n)$ forme bien une chaîne de Markov *contrôlée par* $X_n$, où seul l'état $X_n$ influence la transition vers $(X_{n+1}, Y_{n+1})$. Bien évidemment, dans le cas présent $Y_n$ est même $X_n$ mesurable mais il est aisé de vérifier

que l'on peut toujours se ramener à ce cas en élargissant l'état du système. Cependant, le but recherché ici n'est pas de répéter les équations générale de lissage que l'on trouvera dans la chapitre 3 de [Cappé et al., 2005] mais bien d'expliciter les principes qui les sous-tendent dans un cas élémentaire.

Par ailleurs, cet exemple très simple met également bien en évidence l'une des difficultés posés par le lissage : pour que $\mathrm{P}_\nu(X_k|Y_{0:n})$[1] soit bien définie, il est nécessaire que $\mathrm{P}_\nu(Y_{0:n})$ soit strictement positif, quel que soit le choix de la loi initiale $\nu$. Dans notre cas, cela ne sera possible — sauf à spécifier plus précisément les ensembles $G_0, \ldots, G_n$ — que si $Q(x, x') > 0$, c'est à dire que la chaîne est irréductible à un pas, ce qui constitue une condition assez restrictive[2].

### 2.1.1 La décomposition forward-backward

Cette décomposition, telle que proposée par [Baum et al., 1970], consiste à écrire

$$\begin{aligned}\phi_{\nu,k|n}(x) &\stackrel{\mathrm{def}}{=} \mathrm{P}_\nu(X_k = x|Y_{0:n}) \\ &\propto \mathrm{P}_\nu(X_k = x, Y_{0:n}) = \underbrace{\mathrm{P}(Y_{k+1:n}|X_k = x)}_{\beta_{k|n}(x)} \underbrace{\mathrm{P}_\nu(X_k = x, Y_{0:k})}_{\alpha_{\nu,k}(x)} , \quad (2.1)\end{aligned}$$

en utilisant la propriété de Markov. Les quantités $\alpha_{\nu,k}(x)$ et $\beta_{k|n}(x)$ constituent respectivement les « variables » *forward* et *backward* telles qu'introduites par [Baum et al., 1970]. Par rapport aux notations de [Baum et al., 1970], nous indiquons toutefois explicitement le fait que la variable avant $\alpha_{\nu,k}(x)$ dépend de la loi initiale $\nu$ et de $Y_{0:k}$, tandis que la variable arrière $\beta_{k|n}(x)$ dépend de l'ensemble des observations futures $Y_{k+1:n}$. Par ailleurs, le terme de « variable » n'est pas très approprié puisqu'il est clair que ces deux quantités sont de natures différentes. La quantité avant $\alpha_{\nu,k}(x)$ est une mesure, normalisable en une mesure de probabilité puisque

$$\phi_{\nu,k|k}(x) \stackrel{\mathrm{def}}{=} \mathrm{P}_\nu(X_k = x|Y_{0:k}) = \frac{\alpha_{\nu,k}(x)}{\sum_{x' \in \mathsf{X}} \alpha_{\nu,k}(x')} ,$$

est la loi dite de *filtrage*. De fait, la relation (2.1) aurait très bien pu être écrite en utilisant la loi de filtrage $\phi_{\nu,k|k}$ à la place de la mesure avant (non-normalisée) $\alpha_{\nu,k}(x)$. C'est d'ailleurs la façon dont est calculée cette quantité en pratique pour éviter les problèmes de dépassement des possibilités de représentation numérique (procédure dite de « scaling » bien connue en reconnaissance de la parole, bien que son interprétation probabiliste soit souvent ignorée [Rabiner, 1989]). Par contre, $\beta_{k|n}(x)$ correspond fondamentalement à une loi conditionnelle, difficilement normalisable — $\beta_{k|n}(x)$ n'a aucune raison particulière d'être sommable en $x$ — et n'ayant pas d'interprétation probabiliste directe, même si on peut évidemment l'interpréter, dans les cas où elle est normalisable, comme une loi a posteriori dans un pseudo modèle où $X_k$ possède une loi a priori, éventuellement impropre, uniforme sur X. Pour souligner cette distinction, qui devient très importante lorsque X est un espace d'état général, nous avons proposé d'utiliser les termes de *mesure avant* pour $\alpha_{\nu,k}(x)$ et de *fonction arrière* pour $\beta_{k|n}(x)$.

[1] On notera dans la suite la collection de variables $Y_l, \ldots, Y_k$ par $Y_{l:k}$.

[2] Ceci-dit, c'est également souvent une hypothèse nécessaire pour garantir les propriétés d'oubli des conditions initiales des relations de filtrage et de lissage, qui elles mêmes sont nécessaire pour garantir la consistance de l'estimateur du maximum de vraisemblance (voir notamment la fin de la section ci-dessous 2.3). L'exemple 4.3.28 du chapitre 4 de [Cappé et al., 2005] montre d'ailleurs qu'en elle même l'irréductibilité à un pas ne suffit pas toujours à garantir l'oubli des relations de filtrage.

On montre facilement que $\alpha_{\nu,k}(x)$ et $\beta_{k|n}(x)$ sont calculables par des récursions (respectivement, pour $k = 0, \ldots, n$ et $k = n, n-1, \ldots, 0$) qui font que le calcul de l'ensemble des lois de lissage, pour $k \in \{0, \ldots, n\}$, ne croît que linéairement avec $n$ (voir, notamment, le chapitre 3 de [Cappé et al., 2005]).

Bien que constituant la méthode classique de calcul des lois de lissage, la décomposition (2.1) a le défaut de ne pas être stable numériquement du fait de l'absence de façon simple de normaliser $\beta_{k|n}(x)$ [Ephraim & Merhav, 2002]. Dans le cas où X est fini, on trouve bien sûr toujours des façons heuristiques d'éviter les problèmes numériques au détriment de l'interprétabilité des récursions avant et arrière [Rabiner, 1989] (voir aussi la section 3.4 de [Cappé et al., 2005]). Dans le cas des modèles d'états linéaires gaussiens, la récursion arrière est également calculable grâce à la forme dite « information » des relations de filtrage de Kalman qui permet de manipuler des densités gaussiennes non normalisées et, éventuellement, impropres (cf. chapitre 5.2.5 de [Cappé et al., 2005])[3].

La décomposition forward-backward a paradoxalement également le défaut de ne pas explicitement reposer sur un aspect fondamental du problème qui est que, conditionnellement à $Y_{0:n}$, $(X_k)_{k\in\mathbb{N}}$ possède une structure de chaîne de Markov inhomogène. En effet,

$$\begin{aligned}
&\mathrm{P}_\nu(X_{0:n} = x_{0:n}|Y_{0:n}) \\
&\qquad \propto \mathrm{P}_\nu(X_{0:n} = x_{0:n}, Y_{0:n}) = \nu(x_0)\mathbb{1}_{G_0}(x_0) \prod_{k=0}^{n-1} Q(x_k, x_{k+1})\mathbb{1}_{G_{k+1}}(x_{k+1}) \ , \quad (2.2)
\end{aligned}$$

où l'on reconnaît, au terme de normalisation manquant près, une structure markovienne (produit de termes ne faisant interagir que deux variables d'état successives).

### 2.1.2 La décomposition markovienne avant

Pour réécrire (2.2) de façon plus standard, il est nécessaire de faire apparaître les noyaux de transitions qui relient les lois de lissage $\phi_{\nu,k|n}$ et $\phi_{\nu,k+1|n}$ pour $k = 0, \ldots, n-1$ :

$$\begin{aligned}
\phi_{\nu,k+1|n}(x) &\stackrel{\mathrm{def}}{=} \mathrm{P}_\nu(X_{k+1} = x|Y_{0:n}) \\
&\propto \sum_{x'\in\mathsf{X}} \mathrm{P}_\nu(X_{k+1} = x, X_k = x', Y_{0:n}) \\
&= \sum_{x'\in\mathsf{X}} \mathrm{P}_\nu(X_{k+1} = x|X_k = x', Y_{0:n}) \underbrace{\mathrm{P}_\nu(X_k = x', Y_{0:n})}_{\propto \phi_{\nu,k|n}(x')} \\
&\propto \sum_{x'\in\mathsf{X}} \underbrace{\mathrm{P}_\nu(Y_{k+2:n}|X_{k+1} = x)}_{\beta_{k+1|n}(x)} \underbrace{\mathrm{P}_\nu(Y_{k+1}|X_{k+1} = x)}_{\mathbb{1}_{G_{k+1}}(x)} \underbrace{\mathrm{P}_\nu(X_{k+1} = x|X_k = x')}_{Q(x',x)} \phi_{\nu,k|n}(x') \ .
\end{aligned} \quad (2.3)$$

D'où, en posant

$$\mathrm{F}_{k|n}(x', x) \stackrel{\mathrm{def}}{=} \frac{Q(x', x)\mathbb{1}_{G_{k+1}}(x)\beta_{k+1|n}(x)}{\sum_{x''\in\mathsf{X}} Q(x', x'')\mathbb{1}_{G_{k+1}}(x'')\beta_{k+1|n}(x'')} \ , \quad (2.4)$$

[3] De façon générale, la décomposition forward-backward est connue, dans le cas des modèles d'états linéaires gaussiens, sous le nom de *two-filter decomposition* [Kailath et al., 2000] ; dénomination quelque peu malheureuse dans la mesure où, sauf dans les cas particuliers où $\beta_{k|n}(x)$ peut effectivement s'interpréter comme un prédicteur correspondant à un système dont les indices temporels ont été inversés, la fonction arrière $\beta_{k|n}(x)$ ne correspond pas, en général, à une opération de filtrage.

$$\phi_{\nu,k+1|n}(x) = \sum_{x' \in \mathsf{X}} \phi_{\nu,k|n}(x') \mathrm{F}_{k|n}(x', x) \ .$$

Les *noyaux avant* définis en (2.4) permettent donc d'écrire la loi de la chaîne cachée conditionnée aux observations $Y_{0:n}$ sous une forme markovienne usuelle, quoi que non homogène. Les propriétés des noyaux avant sont capitales pour garantir les propriétés d'oublis des relations de lissage qui jouent un rôle central pour assurer la consistance de l'estimateur du maximum de vraisemblance ou la stabilité des approximations particulaires (cf. chapitres 4, 9 et 11 de [Cappé et al., 2005]).

Du point de vue algorithmique par contre, la décomposition markovienne avant n'est pas plus intéressante que la décomposition forward-backward dans la mesure où elle repose entièrement sur la détermination des fonctions arrière $\beta_{k|n}$ et ne permet donc pas d'éviter la question de la normalisation.

### 2.1.3 La décomposition markovienne arrière

Il est toutefois possible de factoriser (2.2) différemment de façon à faire apparaître un produit de *noyaux de transition arrière* opérant sur les variables indicées dans le sens temporel inversé. En effet,

$$\begin{aligned} \phi_{\nu,k|n}(x) \stackrel{\mathrm{def}}{=} \mathrm{P}_\nu(X_k = x | Y_{0:n}) &= \sum_{x' \in \mathsf{X}} \mathrm{P}_\nu(X_k = x, X_{k+1} = x' | Y_{0:n}) \\ &= \sum_{x' \in \mathsf{X}} \mathrm{P}_\nu(X_k = x | X_{k+1} = x', Y_{0:k}) \underbrace{\mathrm{P}_\nu(X_{k+1} = x' | Y_{0:n})}_{\phi_{\nu,k+1|n}(x')} \ . \end{aligned} \quad (2.5)$$

L'évaluation du premier terme du membre de droite (2.5) de correspond à un problème classique de retournement temporel conduisant à définir le noyau arrière de la façon suivante

$$\mathrm{B}_{\nu,k}(x', x) \stackrel{\mathrm{def}}{=} \mathrm{P}_\nu(X_k = x | X_{k+1} = x', Y_{0:k}) = \frac{\phi_{\nu,k|k}(x) Q(x, x')}{\sum_{x'' \in \mathsf{X}} \phi_{\nu,k|k}(x'') Q(x'', x')} \ . \quad (2.6)$$

Les équations (2.5) et (2.6) montrent qu'une fois les lois de filtrage $\phi_{\nu,k|k}$ calculées pour $k = 0, \ldots, n$, les lois de lissage s'en déduisent par une récursion arrière n'impliquant plus que les lois de filtrage $\phi_{\nu,k|k}$ elles-mêmes et le noyau $Q$ (en particulier, les observations $Y_k$ n'interviennent plus). Cette décomposition markovienne arrière permet donc d'éviter totalement les problèmes de normalisation et constitue l'approche de choix pour l'évaluation numériques des probabilités de lissage. De façon curieuse toutefois, alors que cette approche est bien connue dans le cas des modèles d'états linéaires gaussiens (c'est les principe commun des approches de Rauch-Tung-Striebel, dite également, *forward filtering, backward smoothing* et de Bryson-Frazier ou *disturbance smoothing*), elle est totalement méconnue dans le cas des modèles de Markov cachés à état fini [Ephraim & Merhav, 2002].

Bien que préférable d'un point de vue algorithmique, cette seconde décomposition implique le retournement temporel — équation (2.6) — qui n'admet pas nécessairement une formulation explicite dans le cas d'un espace d'état général. Notons que pour démontrer des propriétés d'oubli « bi-directionnelles » (à la fois vis à vis des observations futures et passées) des lois de lissage, il est nécessaire d'avoir recours simultanément aux deux types de décompositions [Douc et al., 2004].

## 2.2 Estimation de fonctionnelles lissées [Cappé et al., 2005, sections 4.1 et 10.2-10.3]

Une limitation fondamentale des trois procédures de lissage décrites ci-dessus est le fait qu'elles impliquent toutes deux passages successifs à travers les observations (dans le sens des indices croissant puis décroissants, ou inversement). Cette particularité rend le lissage coûteux lorsqu'il s'agit de traiter de très grandes masses de données — il faut en particulier stocker de l'ordre de $|\mathsf{X}| \times n$ variables intermédiaires —, voire totalement impraticable lorsqu'il s'agit d'estimer les paramètres *en ligne*, sans stockage des données.

On peux en fait nuancer ce propos grâce à une remarque peu connue, apparemment due à [Zeitouni & Dembo, 1988] et reprise essentiellement par les auteurs de [Elliott et al., 1995] et leur co-auteurs : si le calcul complet des lois de lissage nécessite bien un double passage à travers les données, il est par contre possible de calculer de façon récursive en $n$ l'espérance conditionnelle de fonctions *fixées* de l'état, dès lors qu'elles ont une certaine structure, que nous avons baptisé *fonctionnelles de lissage* dans [Cappé et al., 2005]. Ma contribution personnelle sur ce sujet a consisté à monter

- d'une part, que la classe de fonctionnelles de l'état qu'il est possible de considérer est plus générale que les seules fonctionnelles additives et qu'il est en particulier possible de calculer ainsi la matrice d'information observée (hessien de la log-vraisemblance) en utilisant la formule de Louis (1.12) (voir sections 4.1 et 10.3 de [Cappé et al., 2005] ainsi que [Cappé & Moulines, 2005b]) ;
- d'autre part, que l'application de cette approche de calcul récursif à la fonctionnelle additive définie par l'équation de Fisher (1.10) (pour le calcul du gradient de la log-vraisemblance) permet de retrouver exactement l'approche dite des *équations de sensibilité* utilisée par [Campillo & Le Gland, 1989; Le Gland & Mevel, 1997; Cappé et al., 1998a; Collings & Rydén, 1998] (voir section 10.2 de [Cappé et al., 2005] ainsi que [Cappé & Moulines, 2005b]).

Pour donner simplement l'argument principal, considérons une famille de fonctions $t_n$ ayant la structure suivante :

$$t_{n+1}(x_{0:n+1}) = m_n(x_n, x_{n+1}) t_n(x_{0:n}) + s_n(x_n, x_{n+1}) \ , \tag{2.7}$$

où $m_n$ et $s_n$ sont des fonctions sur $\mathsf{X}^2$. Cette structure inclut en particulier les fonctionnelles additives $t_n(x_{0:n}) = \sum_{k=0}^{n-1} s_k(x_k, x_{k+1})$ ainsi que leur carré ($[\sum_{k=0}^{n-1} s_k(x_k, x_{k+1})]^2$), elle permet donc notamment de traiter tous les termes qui apparaissent dans les équations de Fisher (1.10) et Louis (1.12). L'idée de base est que s'il est impossible de mettre directement à jour $\gamma_n \stackrel{\text{def}}{=} \mathrm{E}\left[t_n(X_{0:n}) | Y_{0:n}\right]$ récursivement en $n$, il est par contre possible de le faire pour la quantité intermédiaire suivante

$$\tau_{\nu,n}(x) \stackrel{\text{def}}{=} \mathrm{E}\left[ \mathbb{1}_{\{x\}}(X_n) t_n(X_{0:n}) \middle| Y_{0:n} \right] \ . \tag{2.8}$$

En effet, on vérifie sans difficulté que[4]

$$\tau_{\nu,0}(x) = \phi_{\nu,0|0}(x) t_0(x)$$

[4] On donne ici l'expression applicable dans le cas d'un modèle de Markov caché à observations et états discrets avec les notations définies en début de la section 2.1, voir la section 4.1 de [Cappé et al., 2005] pour le cas général.

et

$$\tau_{\nu,n+1}(x) = c_{\nu,n+1}^{-1} \sum_{x' \in \mathsf{X}} \Big[ \tau_n(x')\, Q(x',x)\, g_{n+1}(x) m_n(x',x) + \phi_{\nu,n|n}(x')\, Q(x',x)\, g_{n+1}(x) s_n(x',x) \Big] \quad (2.9)$$

où $c_{\nu,n+1}$ est le facteur de normalisation qui apparaît dans la récursion usuelle de filtrage :

$$c_{\nu,n+1} \stackrel{\text{def}}{=} \sum_{(x',x) \in \mathsf{X}^2} \phi_{\nu,n|n}(x') Q(x',x) g_{n+1}(x) \ .$$

Ainsi $\tau_{\nu,n}$ peut être mis à jour récursivement, en même temps que le filtre usuel $\phi_{\nu,n|n}$, à l'aide de récursions très similaires dans les deux cas. La quantité d'intérêt $\gamma_n$ s'obtient simplement par intégration de $\tau_{\nu,n}$

$$\gamma_n = \sum_{x \in \mathsf{X}} \tau_{\nu,n}(x) \ .$$

Dans le cas général, la quantité intermédiaire $\tau_{\nu,n}$ correspond à une mesure signée sur X qui, du fait sa définition en (2.8), est absolument continue par rapport à la mesure de filtrage $\phi_{\nu,n|n}$.

L'élégance de cette réécriture récursive ne doit pas masquer un problème fondamental lié au fait qu'elle n'est possible que parce que la fonctionnelle $t_n$ a été entièrement spécifiée : s'il y a plusieurs fonctionnelles, il faudra mettre en œuvre plusieurs récursions parallèles. Pour cette raison, et sauf si l'on l'a affaire à de très grands volume de données ou si l'on cherche à faire de l'estimation en ligne, la réécriture récursive est un général plus coûteuse que les décompositions présentées dans la section 2.1 dès lors que l'on a affaire à une fonctionnelle additive, c'est à dire en particulier dans le cas des statistiques de l'EM ou du gradient de la log-vraisemblance (voir discussion en fin des sections 4.1 et 10.3 de [Cappé et al., 2005]). Par contre, dans le cas du hessien de la log-vraisemblance, la réécriture récursive permet effectivement de calculer une quantité qui n'est pas directement disponible à partir des décompositions de la section 2.1 [Cappé & Moulines, 2005b].

## 2.3 Approximations particulaires [Cappé, 2001a; Olsson et al., 2006a; Cappé et al., 2005, sections 8.3 et 11.1]

Conformément à ce qui a été dit dans le chapitre précédent, dès que l'espace d'état X n'est plus fini (et sauf dans quelques rares cas particuliers comme le modèle linéaire gaussien), le calcul de l'espérance conditionnelle des variables latentes devient problématique. La structure des décompositions de lissage, décrites dans les sections 2.1 et 2.2, reste valable mais elles impliquent des opérations d'intégration qui ne peuvent plus être réalisées numériquement de façon exacte. Depuis une dizaine d'années, ce problème a été abordé de façon novatrice par l'utilisation de techniques de simulation dite *séquentielles* [Liu & Chen, 1998; Doucet et al., 2000, 2001a; Arulampalam et al., 2002; Ristic et al., 2004; Künsch, 2005].

Dans ce cadre, ma contribution personnelle a porté essentiellement — outre un travail plus ponctuel sur la comparaison des méthodes de ré-échantillonnage [Douc et al., 2005a] — sur l'utilisation des méthodes de Monte Carlo séquentielles pour l'approximation de quantités lissées de la forme de celle discutées dans la section 2.2, et partant de

là, de leur utilisation pour l'estimation de paramètres fixes ([Cappé, 2001a], chapitres 8 et 11 de [Cappé et al., 2005] ainsi que [Olsson et al., 2006a]). Dans la suite de cette section, j'expose de façon très succincte le principe des méthodes de Monte Carlo séquentielles (ou de filtrage particulaire) ainsi que leur utilisation proposée pour l'estimation de paramètres fixes.

### 2.3.1 Le modèle de Markov caché paramétrique

On considère ici un modèle de Markov caché ou modèle d'état général qui comporte deux composantes :

**Les états** $\{X_k\}_{k\in\mathbb{N}}$ forme une chaîne de Markov homogène de densité de probabilité initiale $X_0 \sim \nu_\theta$ et de densité de transition $X_{k+1}|X_{0:k} \sim q_\theta(X_k, \cdot)$

**Les observations** $\{Y_k\}_{k\in\mathbb{N}}$ sont supposées conditionnellement indépendantes sachant $\{X_k\}_{k\in\mathbb{N}}$ et telles que $Y_k|X_{0:n} \sim g_\theta(X_k, \cdot)$.

Le but est d'estimer le vecteur de paramètres $\theta$ à partir d'un ensemble observations fixé $Y_{0:n}$, contexte dit (en anglais) d'estimation *batch* ou *off-line*. Cette distinction est importante puisque bien que l'on considère des approches utilisant les simulations particulaires et ne compromettant pas le caractère séquentiel des simulations[5], nous n'avons pas explicitement abordé la question de l'estimation en ligne.

### 2.3.2 Les méthodes de Monte Carlo séquentielles

L'algorithme de base consiste en une réécriture récursive de l'échantillonnage préférentiel à loi instrumentale markovienne qui remonte à [Handschin & Mayne, 1969; Handschin, 1970]. Il s'agit mettre à jour un système de « particules » pondérées $(\xi_k^{1:m}, \omega_k^{1:m})$ de la façon suivante

- Les nouvelles positions des particules $\xi_{k+1}^{1:m}$ sont tirées conditionnellement sachant le passé $\xi_{0:k}^{1:m}, \omega_{0:k}^{1:m}, Y_{0:k+1}$ et tels que

$$\xi_{k+1}^i|\xi_{0:k}^{1:m}, \omega_{0:k}^{1:m}, Y_{0:k+1} \sim r_{k+1}(\xi_k^i, \cdot) \ .$$

- Les nouveaux poids d'importance (normalisés) sont définis par

$$\omega_{k+1}^i = \frac{\omega_k^i w_{k+1}(\xi_k^i, \xi_{k+1}^i)}{\sum_{j=1}^m \omega_k^j w_{k+1}(\xi_k^j, \xi_{k+1}^j)} \ ,$$

où $w_k(x, x') \stackrel{\text{def}}{=} q_\theta(x, x') g_\theta(x, Y_k)/r_k(x, x')$.

Notons que les notations ci-dessus masquent le fait que la fonction d'importance $w_k$ et donc les positions des particules $\xi_k^i$ et les poids $\omega_k^i$ dépendent de $\theta$. Dans cette approche, le système de particules dépend du paramètre $\theta$ et l'on compte l'exploiter pour évaluer des informations « locales » (typiquement, la valeur de la log-vraisemblance ou de son gradient ou bien encore l'espérance des statistiques de l'EM) en $\theta$. En particulier, on ne considère pas ici les méthodes dans lesquelles le paramètre $\theta$ est inclus dans l'espace d'état de la chaîne cachée [Liu & West, 2001; Fearnhead, 2002; Storvik, 2002].

Il est possible d'avoir une interprétation trajectorielle de l'algorithme précédent en définissant la trajectoire de chaque particule par $\xi_{0:k+1}^i = (\xi_{0:k}^i, \xi_{k+1}^i)$. Malheureusement

[5] On n'a en particulier pas considéré l'utilisation de méthodes non séquentielles réminiscences des décompositions évoquées dans la section 2.1 telles que proposées par [Kitagawa, 1996; Godsill et al., 2004; Briers et al., 2004].

cet algorithme de base dégénère rapidement, lorsque $n$ augmente, dans des configurations où une majorité des poids normalisés $\omega_n^i$ sont égaux à zéros (situation analysée dans le cas, plus simple, d'observations IID dans la section 7.3 de [Cappé et al., 2005] et dans [Douc et al., 2002]). La solution proposée dans [Gordon et al., 1993] consiste à effectuer régulièrement un rééchantillonnage que l'on présente ici dans sa version la plus simple, dite « rééchantillonnage multinomial » :

- $\xi_k^i$ est remplacé $\xi_k^{J_k^i}$ où $J_k^{1:m}|\xi_{0:k}^{1:m}, \omega_{0:k}^{1:m}, Y_{0:k+1} \sim \mathrm{Mult}(m, \omega_k^{1:m})$.
- La trajectoire $\xi_{0:k}^i$ est remplacée par $\xi_{0:k}^{J_k^i}$.
- Les poids $\omega_k^i$ sont tous réinitialisés à $1/m$.

Du fait du rééchantillonnage et pour éviter les ambiguïtés, la position d'indice $l$ dans la $k$ème trajectoire est désignée par $\xi_{0:k}^i(l)$ et on continue à noter, pour simplifier, $\xi_k^i = \xi_{0:k}^i(k)$.

Sous des conditions assez générales, il est possible de montrer que l'algorithme précédent fournit de fait une approximation consistante, en le nombre de particules, des probabilités de lissage dans le sens où [Del Moral, 2004]

$$\left\| \sum_{i=1}^m \omega_n^i f(\xi_{0:n}^i) - \mathrm{E}_\theta \left[ f(X_{0:n}) | Y_{0:n} \right] \right\|_{p|Y_{0:n}} \leq \frac{\|f\|_\infty \mathcal{C}(n)}{\sqrt{m}} \ ,$$

pour $f$ une fonction quelconque (bornée) de la trajectoire et où les normes $L^p$ sont définies conditionnellement aux observations $Y_{0:n}$ (en particulier la borne $\mathcal{C}(n)$ dépend en général des observations $Y_{0:n}$). Bien que rassurant, eu égard à sa généralité, ce type de résultats n'est malheureusement pas nécessairement suffisant dans un contexte statistique où il est nécessaire de contrôler la croissance de l'erreur avec le nombre d'observations $n$.

### 2.3.3 Estimation directe de la vraisemblance

Pour ce qui concerne l'estimation du paramètre $\theta$, une première technique fiable consiste à approximer la log-vraisemblance. En effet, on montre facilement en utilisant les propriétés markoviennes du modèle que si $(\tilde{\xi}_{0:k}^{1:m}, \tilde{\omega}_k^{1:m})$ est un système de particules pondérées qui vise la distribution prédictive jointe $\mathrm{P}_\theta(dx_{0:k}|Y_{0:k-1})$ (plutôt que la distribution de lissage comme précédemment), la log-vraisemblance peut être approximée par

$$\hat{\mathrm{L}}_n^m(\theta) = \sum_{k=0}^n \log \left( \sum_{i=1}^m \tilde{\omega}_k^i g_\theta(\tilde{\xi}_k^i, Y_k) \right) \ .$$

Cette approximation est intéressante dans la mesure où elle repose sur les propriétés markoviennes du système et se présente sous la forme d'une somme de $n$ termes n'impliquant que les approximations successives de la loi *marginale de prédiction* $\mathrm{P}_\theta(dx_k|Y_{0:k-1})$. Sous des conditions additionnelles de *mélangeance uniforme* (définies ci dessous) il est notamment possible de montrer que [Olsson & Rydén, 2005]

$$n^{-1} \left\| \hat{\mathrm{L}}_n^m(\theta) - \mathrm{L}_n(\theta) \right\|_{p|Y_{0:n}} \leq \frac{\mathcal{C}}{\sqrt{m}} \ ,$$

où $\mathcal{C}$ est une constante (qui dépend des observations), ce qui permet d'envisager une approximation stable, en le nombre $n$ d'observations, de la log-vraisemblance normalisée $\mathrm{L}_n(\theta)/n$ qui constitue la quantité pertinente du point de vue statistique (qui converge en particulier vers un contraste limite déterministe). L'utilisation directe de l'approximation

séquentielle de la log-vraisemblance constitue donc une première approche efficace pour l'estimation de $\theta$ [Shephard & Pitt, 1997; Hürzeler & Künsch, 1998; Olsson & Rydén, 2005]. Malheureusement, la maximisation numérique à partir uniquement de l'évaluation (approchée) de la log-vraisemblance devient difficilement praticable lorsque la dimension de $\theta$ est plus importante (typiquement supérieure à 3).

### 2.3.4 Estimation de fonctionnelles additives lissées

Pour aller plus loin, qu'il s'agisse de calculer le gradient de la log-vraisemblance ou l'espérance conditionnelle des statistiques suffisantes de l'algorithme EM, il nous faudra conformément à la discussion de la section 2.2, estimer l'espérance d'une fonctionnelle additive de l'état de la forme

$$\gamma_n = \sum_{k=0}^{n-1} \mathrm{E}_\theta \left[ s_k(X_k, X_{k+1}) | \, Y_{0:n} \right] \ ,$$

où $s_k$ sont des fonctions dépendant des observations et possiblement de la valeur de $\theta$. A priori, il est difficile d'imaginer approcher cette quantité autrement que par l'estimateur trajectoriel naturel [Cappé, 2001a]

$$\hat{\gamma}_n^m = \sum_{i=1}^{m} \omega_n^i \sum_{k=0}^{n-1} s_k \left( \xi_{0:n}^i(k), \xi_{0:n}^i(k+1) \right) \ .$$

La forme même de cette approximation montre qu'à la différence du cas de la log-vraisemblance, nous avons maintenant affaire à l'approximation d'une fonction impliquant *l'ensemble de la trajectoire de chaque particule*. Les travaux récents dans le domaine montrent d'ailleurs que cette constatation est fondamentalement due à la nature du problème et pas à l'utilisation de la formule de Fisher puisque toutes les façons d'approximer le gradient de la log-vraisemblance sont confrontées à la même question [Andrieu et al., 2005]. On constate facilement que le calcul de $\hat{\gamma}_n^m$ ne requiert en aucun cas le stockage des trajectoires complètes des particules et qu'il suffit de mémoriser pour chaque trajectoire, en plus de la position courante de la particule $\xi_n^i$ et du poids trajectoriel $\omega_n^i$, la valeur cumulée de la fonctionnelle le long de la trajectoire

$$\sum_{k=0}^{n-1} s_k \left( \xi_{0:n}^i(k), \xi_{0:n}^i(k+1) \right) \ ,$$

qui se calcule elle aussi manifestement de façon récursive. On retrouve d'ailleurs les notions introduites dans la section 2.2 dans la mesure où le système de particules $\xi_n^i$, pour $i = 1, \ldots, m$, pondéré par les poids signés

$$\omega_n^i \left[ \sum_{k=0}^{n-1} s_k \left( \xi_{0:n}^i(k), \xi_{0:n}^i(k+1) \right) \right]$$

constitue une approximation de la mesure signée $\tau_{\nu,n}$ définie en (2.8) [Cappé, 2001a; Cérou et al., 2001; Poyiadjis et al., 2005].

L'examen des trajectoires de particules sur différents modèles (cf. figure 2.1 dans le cas du modèle de volatilité stochastique) suggère l'idée que l'approximation de fonctionnelles lissées pâtit du fait que le système de particules est très dégénéré pour les valeurs situées loin dans l'historique : pour $k \ll n$, la population $\xi_{0:n}^{1:m}(k)$ est en fait réduite à très

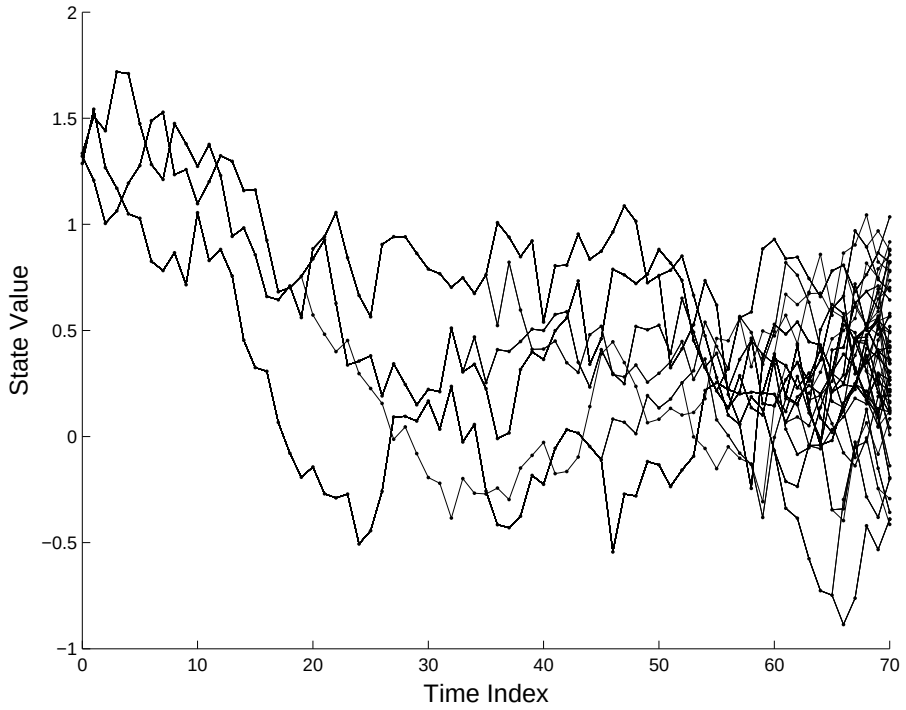


FIG. 2.1 – Trajectoires des particules pour $n = 70$ et $N = 50$ (modèle de volatilité stochastique) [Cappé et al., 2005].

peu de points distincts du fait des rééchantillonnages successifs. Ainsi la qualité de l'approximation de $\mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:n}\right]$ n'est pas uniforme le long de la trajectoire : elle est bonne lorsque $k$ diffère peu de $n$ et se dégrade pour les faibles valeurs de $n$.

La solution que nous avons proposé dans la section 8.3 de [Cappé et al., 2005] (voir aussi [Cappé & Moulines, 2005a]) s'inspire des idées d'oubli qui sont la clé des preuves tant de la consistance du maximum de vraisemblance dans les modèles de Markov cachés [Douc et al., 2004] que de la stabilité, en le nombre d'observations, des méthodes de Monte Carlo séquentielles [Del Moral & Guionnet, 2001]. Le travail récent [Andrieu et al., 2005] propose, partant du même constat, une approche dans laquelle on utilise un principe similaire pour modifier, non pas l'approximation séquentielle, mais la nature même de la quantité que l'on cherche à approximer (selon une construction due à [Rydén, 1997]). Le principe de l'estimateur à délai fixe est qu'en choisissant $\Delta$ suffisamment grand, $\mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:n}\right]$ et en fait proche de $\mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:k+\Delta}\right]$. Plus précisément, sous les hypothèses de mélangeance uniforme (voir ci dessous), on montre que

$$\left|\mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:n}\right] - \mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:k+\Delta}\right]\right| \leq \|s_k\|_\infty \rho^\Delta \,,$$

pour $n \geq k + \Delta$. Par contre, au vu de la figure 2.1, il y a tout lieu de penser que pour des valeurs raisonnables de $\Delta$, l'estimation particulaire de $\mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:k+\Delta}\right]$ sera bien meilleure que celle de $\mathrm{E}_\theta\left[s_k(X_k, X_{k+1})|\, Y_{0:n}\right]$. Par conséquent, nous avons proposé pour estimer $\gamma_n$ l'estimateur *à délai fixe* (*fixed-lag* en anglais) suivant :

$$\hat{\gamma}_n^{m,\Delta} = \sum_{k=0}^{n-1} \sum_{i=1}^{m} \omega^i_{(k+\Delta)\wedge n}\; s_k\left(\xi^i_{0:(k+\Delta)\wedge n}(k), \xi^i_{0:(k+\Delta)\wedge n}(k+1)\right) \,. \tag{2.10}$$

Pas plus que l'estimateur précédent, cet estimateur à délai fixe ne requiert de stocker les trajectoires complètes des particules et il a exactement la même complexité de calcul. Il devient simplement nécessaire de stocker l'histoire récente du système de particules, c'est à dire $\xi^i_{0:k}(k-\Delta:k)$, pour $i=1,\ldots,m$, ce qui est raisonnable pour des valeurs de $\Delta$ de l'ordre de quelques dizaines.

Empiriquement, nous avons constaté que le choix de $\Delta$ relève d'une problématique assez courante en statistique de compromis biais/variance : si $\Delta$ est trop faible (typiquement inférieur à 10) l'estimateur à délai fixe présente un biais sensible et, à l'inverse, la variance de l'estimateur croît avec $\Delta$. Dans tous les cas, en utilisant une valeur de $\Delta$ de l'ordre de quelques dizaines, la variance de $\hat{\gamma}_n^{m,\Delta}$ est significativement plus faible que celle de l'estimateur de référence $\hat{\gamma}_n^m$ sans que le biais ne soit significatif. A titre d'exemple, la comparaison des figures 2.2 et 2.3 montre que l'estimateur proposé conduit à une réduction très significative de la variance des estimations de paramètres obtenues avec les versions Monte Carlo de l'algorithme EM de type *Monte Carlo EM* [Wei & Tanner, 1991] ou *Stochastic Approximation EM* [Delyon et al., 1999] (voir détails concernant ces simulations dans [Olsson et al., 2006a]).

Avec Jimmy Olsson, doctorant de Tobias Rydén à Lund, nous avons récemment essayé de donner une base théorique à ces constatations empirique sous les conditions de *mélangeance uniforme* utilisées, avec quelques variantes, par [Del Moral & Guionnet, 2001; Le Gland & Oudjane, 2004; Douc et al., 2004; Douc & Moulines, 2005; Olsson & Rydén, 2005; Künsch, 2005].

**Hypothèse 1** (Mélangeance uniforme)**.**

1. $0<\sigma_-<q_\theta(x,x')<\sigma_+<\infty$, pour tout $\theta$.
2. Pour tout $y$, $\sup_\theta \|g_\theta(\cdot,y)\|_\infty<\infty$, $\inf_\theta \int g_\theta(x,y)dx>0$.
3. Pour tout $y_{0:\infty}$, $\theta$, et $k$, $\|w_k(\cdot,\cdot)\|_\infty<\infty$.

La troisième condition est souvent impliquée par les deux précédentes, notamment dans le cas du *bootstrap filter* où le noyau de proposition $r_k$ est égal à $q_\theta$. Avec ces hypothèses, et en supposant que le rééchantillonnage est effectué systématiquement, nous avons obtenu des bornes sur la norme $L^p$ (conditionnelle aux observations) de l'erreur $\|\hat{\gamma}_n^{m,\Delta}-\gamma_n\|_{p|Y_{0:n}}$ ainsi que sur le biais $|\operatorname{E}[\hat{\gamma}_n^{m,\Delta}|Y_{0:n}]-\gamma_n|$ (théorème 4.5 de [Olsson et al., 2006a]). Dans le cas général, l'expression des bornes n'est pas très simple car elle dépend des ratios d'importance $w_k$ et donc des observations $Y_{0:n}$. Néanmoins, on vérifie aisément que les bornes sont bien meilleures pour l'estimateur à délai fixe $\hat{\gamma}_n^{m,\Delta}$ que pour l'estimateur classique $\hat{\gamma}_n^m$. Sous des hypothèses additionnelles de bornitude uniforme de certains moments, on peut intégrer les observations de façon à obtenir des bornes de la forme (proposition 4.7 de [Olsson et al., 2006a]) :

$$n^{-1}\|\hat{\gamma}_n^{m,\Delta}-\gamma_n\|_p \leq \mathcal{C}_1\rho^\Delta + \frac{\mathcal{C}_2\Delta+\mathcal{C}_3}{\sqrt{m}}\ ,$$
$$n^{-1}\left|\operatorname{E}[\hat{\gamma}_n^{m,\Delta}-\gamma_n]\right| \leq \mathcal{C}_1\rho^\Delta + \frac{\mathcal{C}_4\Delta+\mathcal{C}_5}{m}\ ,$$

où $\mathcal{C}_i$ sont des constantes (qui, cette fois-ci, ne dépendent plus des observations mais peuvent dépendre de $\theta$) et $\rho=1-\sigma_-/\sigma_+$. En particulier, on constate qu'il est suffisant de faire croître le délai $\Delta$ très légèrement avec le nombre d'observations $n$, par exemple $\Delta_n\propto\log(n)$, pour éliminer les termes dûs à l'approximation à délai fixe et obtenir pour la norme $L^p$ une borne de l'erreur de l'ordre de $O(m^{-1/2}\log(n))$. Au facteur $\log(n)$ près et au fait qu'il faut choisir $\Delta_n\propto\log(n)$ s'accroissant légèrement avec $n$, on retrouve

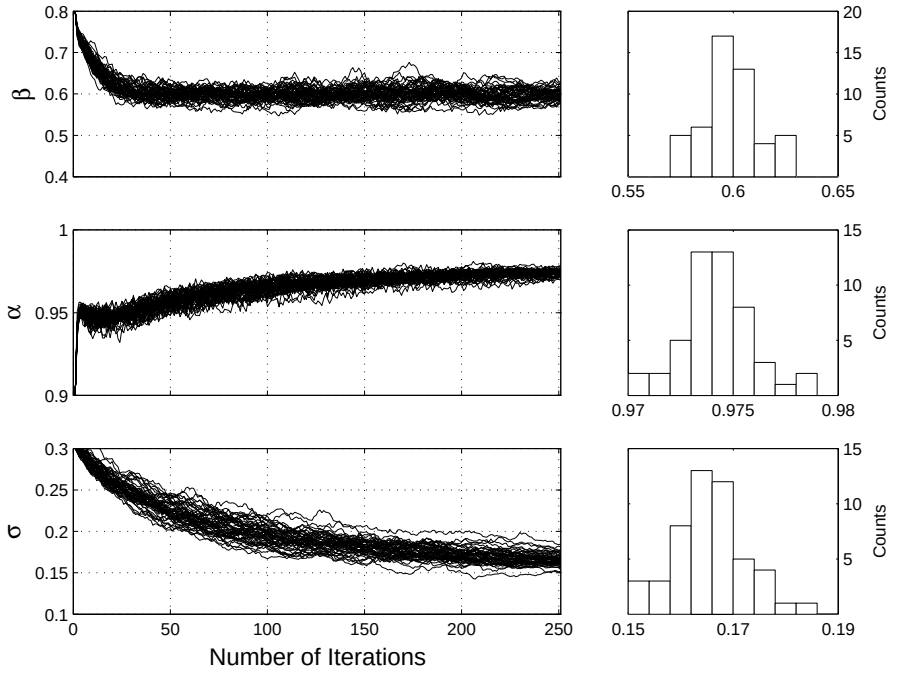


FIG. 2.2 – Trajectoires des paramètres estimés par Monte Carlo EM en utilisant l'estimateur $\hat{\gamma}_n^m$ (modèle de volatilité stochastique, 5000 observations) [Olsson et al., 2006b].

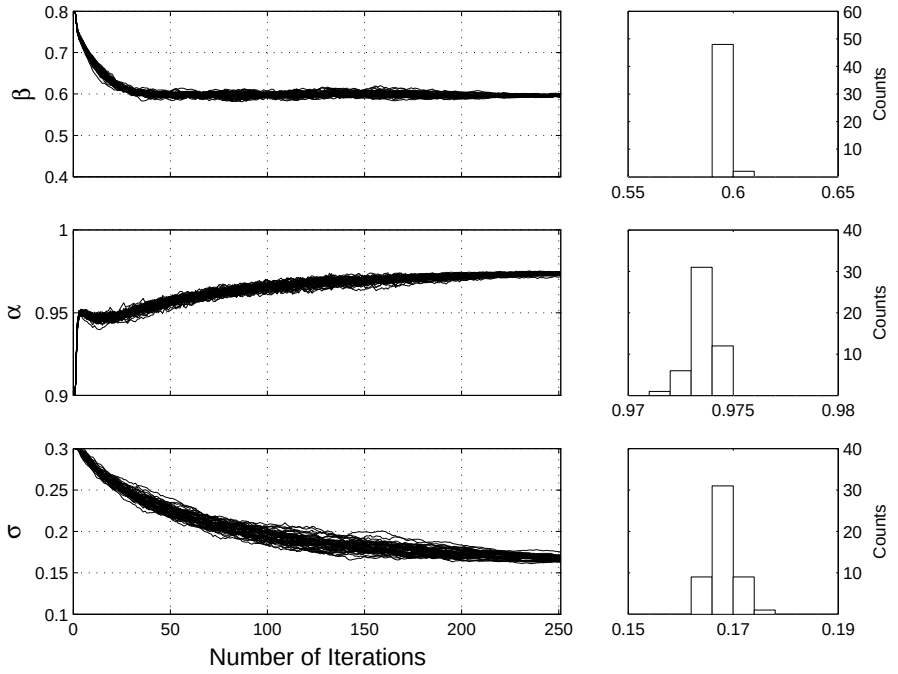


FIG. 2.3 – Trajectoires des paramètres estimés par Monte Carlo EM en utilisant l'estimateur à délai fixe $\hat{\gamma}_n^{m,\Delta}$ pour $\Delta = 40$ (même modèle, données et nombre de particules que pour la figure 2.2) [Olsson et al., 2006b].

quasiment un résultat de stabilité semblable à celui établi dans le cas de l'approximation de la log-vraisemblance. Par comparaison, la meilleur borne que nous ayons pu obtenir, sous les mêmes hypothèses, pour l'estimateur trajectoriel classique est de la forme

$$n^{-1}\|\hat{\gamma}_n^{m,\Delta} - \gamma_n\|_p \leq \frac{n\mathcal{C}_6}{\sqrt{m}} ,$$

même si les simulations que nous avons pu effectuer ne confirment pas nécessairement une dégradation aussi sensible des performances d'estimation lorsque $n$ croît.

Une prolongation naturelle de ces travaux consiste à se demander comment utiliser l'approximation à délai fixe (2.10) pour faire de l'estimation en ligne ? Il existe une façon de procéder assez directe qui s'inspire du principe de l'algorithme SAEM (Stochastic Approximation EM) de [Delyon et al., 1999] plutôt que de celui de l'algorithme MCEM utilisé dans le cas des figures 2.2 et 2.3. Avant d'aborder le cas, plus complexe, des modèles à dépendance markovienne nous avons entrepris d'étudier l'algorithme correspondant dans le cas d'observations IID (travail en cours avec Eric Moulines et Christophe Andrieu, voir aussi [Cappé et al., 2006]).

# Chapitre 3

# Inférence bayésienne sur les paramètres

On considère dans ce chapitre le sujet déjà abordé dans la fin de la section 2.3 du chapitre précédent, à savoir, l'estimation de paramètres dans les modèles à données latentes. L'angle d'approche est un peu différent cependant, dans la mesure où l'on se place ici exclusivement du point de vue bayésien en utilisant des techniques de Monte Carlo pour simuler directement sous la loi à posteriori des paramètres sachant les observations. Ce faisant, il est souvent, mais pas toujours, nécessaire de simuler également les variables latentes.

La section 3.2 présente un travail très directement lié à la section 2.3 du chapitre précédent dans lequel on propose d'appliquer des idées issues des techniques de Monte Carlo séquentielles au cas statique, c'est à dire, pour une simuler une loi a posteriori fixée. On discute ensuite brièvement de plusieurs travaux appliquant la méthodologie MCMC (Monte Carlo par chaîne de Markov) à des problèmes issus du traitement de signal (section 3.3). Enfin, la section 3.4 décrit le travail réalisé à propos de l'algorithme de simulation « à temps continu », proposé par [Stephens, 2000], qui se veut une alternative à l'approche *à saut réversible* de [Green, 1995].

Avant toute chose, la section 3.1 ci-dessous introduit, de façon schématique, les principales notions dont nous aurons besoin concernant l'approche MCMC, en renvoyant aux chapitres 6 et 13 de [Cappé et al., 2005] ou à [Robert & Casella, 2004] pour un traitement plus complet.

## 3.1 Introduction aux méthodes de Monte Carlo par chaîne de Markov

Les algorithmes MCMC ont pour principe, à partir de la donnée d'une loi dite *cible* $\pi$ *connue à un facteur de proportionnalité près*, de construire une chaîne de Markov $\{X_i\}_{i\in\mathbb{N}}$ ayant pour loi stationnaire $\pi$. Sous des conditions additionnelles d'ergodicité, on pourra

1. voir $X_m$ pour $m$ « suffisamment grand » comme un tirage approximatif sous la loi $\pi$ (convergence en loi) ;
2. utiliser $1/m \sum_{i=1}^{m} f(X_i)$ comme approximation Monte Carlo de l'espérance $\mathrm{E}_\pi[f(X)]$ de fonctions $f$ sous $\pi$ (théorème ergodique).

En général, la loi cible $\pi$ est la loi a posteriori du paramètre définie, à une constante près, par la règle de Bayes

$$\pi(\theta) \stackrel{\text{def}}{=} p(\theta|Y) \propto \ell(Y|\theta)\pi_0(\theta) \ .$$

Mais dans les modèles à données latentes, $\pi$ correspond souvent plutôt à la loi a posteriori jointe des paramètres $\theta$ et des données latentes $X$, ou d'une partie d'entre elle. On utilise ici la notation générique $\mathsf{X}$ pour désigner l'espace sur lequel est défini $\pi$. Pour simplifier les notations, on supposera dans la que la loi $\pi$ est définie sur un espace discret $\mathsf{X}$.

Pour garantir que $\pi$ est bien loi stationnaire de $\{X_i\}_{i\in\mathbb{N}}$, les algorithmes MCMC se fondent, en général, une condition plus forte dite de *balance détaillée* ou $\pi$*-réversibilité* :

$$\pi(x)K(x,x') = \pi(x')K(x',x) \qquad \forall x, x' \in \mathsf{X} \ , \tag{3.1}$$

où $K$ désigne le noyau de transition correspondant à l'algorithme utilisé pour simuler la chaîne. L'algorithme de *Metropolis-Hastings*, qui constitue l'approche MCMC la plus générique (applicable sans aucune autre connaissance sur $\pi$) réalise ce programme à partir d'un noyau de proposition $Q$ quelconque grâce à la règle d'acceptation-rejet classique

$$\begin{aligned} X'|X_{0:i} &\sim Q(X_i,\cdot) \ , \\ U|X_{0:i}, X' &\sim \text{uniforme}([0,1]) \ , \\ X_{i+1} &= \begin{cases} X' & \text{if } U < \frac{\pi(X')Q(X',X_i)}{\pi(X_i)Q(X_i,X')} \ , \\ X_i & \text{sinon .} \end{cases} \end{aligned} \tag{3.2}$$

L'autre grande famille d'algorithmes, moins générique car elle nécessite la détermination explicite de lois conditionnelles, est désignée sous le nom d'*échantillonneur de Gibbs*. On parle de *rao-blackwellisation*[1] lorsqu'il est possible d'utiliser la connaissance de ces lois conditionnelles pour intégrer une partie de l'état $X$. Typiquement, si $X = (U, V)$ est partitionné en deux parties et la loi conditionnelle $\pi(u|v)$ est connue exactement — on ne peut ici tolérer une constante de normalisation inconnue, qui en général, dépendrait de la valeur de $v$ —, on a

$$\pi(v) = \frac{\pi(u,v)}{\pi(u|v)} \ ,$$

ce qui montre que $\pi(v)$ est connue (à une constante près) et peut donc être adoptée comme loi cible de l'algorithme MCMC. En pratique la faisabilité de la rao-blackwellisation dépend essentiellement du coût de calcul associé à la détermination de la loi conditionnelle $\pi(u|v)$ pour une valeur de $v$ donnée. Un point important est que dans la plupart des modèles d'intérêt l'algorithme complet de simulation compose plusieurs type de propositions utilisées de façon concomitante (souvent appliquées successivement, dans un ordre prescrit, ou bien dans un ordre aléatoire).

Une variante de l'algorithme de Metropolis-Hastings qui a rencontré un très grand succès, car elle permet d'aborder des questions de choix de modèles, est l'algorithme *à saut réversible* proposé par [Green, 1995]. Dans ce cas, l'espace $\mathsf{X}$ s'écrit comme une union disjointe d'ensembles $\mathsf{X}_d$ de « dimensions différentes ». Typiquement, lorsque $x \in \mathsf{X}_d$, $x$ est un vecteur composée de $d$ coordonnées et $d$ décrit plusieurs valeurs possibles

[1] Je m'abrite ici derrière l'autorité morale de Christian Robert pour utiliser ce néologisme à la fois comme un nom (rao-blackwellisation) et comme un adjectif (rao-blackwellisé).

(voir, notamment, la section 13.2 de [Cappé et al., 2005]). Devant la difficulté d'imaginer des lois de proposition efficaces dans ce contexte, la solution proposée par [Green, 1995] consiste à n'autoriser que des mouvements susceptibles de mettre en correspondance deux ensembles $\mathsf{X}_d$ et $\mathsf{X}_{d'}$ particuliers en utilisant, de surcroît, des mouvements très simples (voir la section 3.4 qui discute de cet aspect plus en détail). L'algorithme complet est obtenu par composition de ces différents mouvements élémentaires.

## 3.2 Algorithmes de simulation de type population [Cappé et al., 2004]

En lien direct avec la fin du chapitre précédent, on considère ici l'utilisation de techniques de Monte Carlo séquentielles pour simuler sous la loi a posteriori d'un modèle bayésien non-dynamique. Cette idée a pour origine à la fois les travaux de [Chopin, 2002] qui suggéraient que l'utilisation de techniques particulaires, avec incorporation progressive des données, constituaient une solution efficace aux problèmes rencontrés par les techniques MCMC lorsque le volume de données est important, mais également les premiers travaux sur les méthodes MCMC adaptatives [Andrieu & Robert, 2001; Haario et al., 1999] dans lesquels apparaissait clairement la difficulté de garantir l'invariance de la loi stationnaire lorsque les lois de proposition sont continuellement adaptées. L'idée est ici d'approcher la loi cible par une large population de particules à partir de laquelle il serait possible de calibrer les paramètres des lois de proposition et dont l'évolution obéit à des règles moins contraintes que les méthodes de simulation multiples basées sur le principe des approches MCMC (en particulier, la condition de balance détaillée) comme [Mengersen & Robert, 2003].

En notant $\pi(x)$ la loi cible, que l'on suppose connue à une constante près, l'algorithme dit *Population Monte Carlo* proposé dans [Cappé et al., 2004] fonctionne comme suit :

**A l'itération** $k = 0$ On simule $m$ particules $\xi_0^1, \ldots, \xi_0^m$ indépendamment sous une loi instrumentale $\nu$ et l'on définit les poids d'importance par

$$\omega_0^i = \frac{\pi(\xi_0^i)}{\nu(\xi_0^i)} \; .$$

**A l'itération** $k \geq 1$ On simule $m$ particules $\xi_k^1, \ldots, \xi_k^m$, conditionnellement indépendantes sachant les simulations précédentes, sous des densités instrumentales $q_k^i$ qui peuvent éventuellement dépendre de l'ensemble des simulations précédentes. On définit alors les poids d'importance par

$$\omega_k^i = \frac{\pi(\xi_k^i)}{q_k^i(\xi_k^i)} \; .$$

L'intérêt de cet algorithme est que quelle que soit la façon de choisir $q_k^i$, on a toujours

C1. $\xi_k^1, \ldots, \xi_k^m$ sont conditionnellement indépendantes sachant $\mathcal{F}_{k-1}$,

C2. pour toute fonction bornée $f$ et tout $i$, $\mathrm{E}[\omega_k^i f(\xi_k^i)|\mathcal{F}_{k-1}] \propto \mathrm{E}_\pi[f(X)]$,

où $\mathcal{F}_{k-1}$ désigne la tribu engendrée par l'ensemble des simulations jusqu'à l'itération $k-1$. En utilisant les résultats disponibles sur la convergence des méthodes particulaires [Del Moral, 2004; Künsch, 2005; Douc & Moulines, 2005], on peut donc s'attendre à ce que les estimateurs du type

$$\frac{\sum_{i=1}^m \omega_n^i f(\xi_n^i)}{\sum_{i=1}^m \omega_n^i} \; ,$$

voire

$$\sum_{k=1}^{n} \alpha_k \frac{\sum_{i=1}^{m} \omega_k^i f(\xi_k^i)}{\sum_{i=1}^{m} \omega_k^i} ,$$

où les poids $\alpha_k$ sont introduits pour tenir compte des différences de variances des estimateurs obtenus aux itérations successives, soient des approximations consistantes et asymptotiquement normales de $\mathrm{E}_\pi[f(X)]$ (lorsque $m$ tends vers l'infini, le nombre d'itérations $n$ étant fixé).

Restait à imaginer des façons efficaces de construire les loi propositions $q_k^i$ afin d'obtenir des estimateurs de variance aussi faible que possible. Dans [Cappé et al., 2004], nous nous sommes limités au *proof of concept* en montrant, sur des exemples, qu'à partir de choix usuels dans le domaine des méthodes de Monte Carlo séquentielles (par exemple en construisant $q_k^i$ à partir du résultat du rééchantillonnage dans la population $\xi_{k-1}^{1:m}$, basé sur les poids d'importance $\omega_k^{1:m}$, suivi d'une perturbation aléatoire locale markovienne), il était possible d'obtenir des résultats satisfaisants dans des contextes où les méthodes MCMC usuelles rencontraient des difficultés.

Ce travail, qui pour ma part c'est arrêté là, à été poursuivi par [Celeux et al., 2003] et [Douc et al., 2005b,c] qui ont étendu le cadre proposé et défini des règles de mise à jour des lois $q_k^i$ permettant effectivement une adaptation vis à vis d'un objectif, notamment, de minimisation de la variance d'estimation de $\mathrm{E}_\pi[f(X)]$ dans les cas où la fonction $f$ est connue. Par ailleurs, [Del Moral et al., 2006] ont étendu cette idée à des constructions plus générales où la condition C2 ci-dessus n'est plus vérifiée, ce qui permet notamment d'utiliser des loi de propositions pour lesquelles le ratio d'importance $\pi/q_k^i$ ne serait pas directement défini (par exemple, des propositions correspondant à une application de l'algorithme de Metropolis-Hastings).

## 3.3 Applications des méthodes de Monte Carlo par chaîne de Markov [Cappé et al., 1999; Cappé, 2002; Rigouste et al., 2006a]

Depuis la fin des années 1990, les techniques de simulation de type Monte Carlo par chaîne de Markov (puis de Monte Carlo séquentiel) ont fait l'objet d'un intérêt grandissant dans le domaine du traitement du signal, et plus généralement, des sciences de l'information et de la communication. A l'origine portées essentiellement par quelques groupes pionniers, en particulier celui du Department of Engineering de l'université de Cambridge, ces approches font maintenant l'objet d'une large diffusion dans le domaine. Mes travaux en la matière on porté principalement sur :

**la déconvolution de signaux non-gaussiens** en utilisant des techniques caractéristiques des modèles conditionnellement linéaires gaussiens qui mélangent simulation et calcul explicites de filtrage et de lissage de Kalman [Cappé et al., 1999] (voir aussi [Buchoux et al., 2000a]) ;

**la segmentation et la classification de données de comptage** en utilisant l'approche à saut réversible de [Green, 1995] pour simuler des modèles à nombre de segments et nombre de types de segments inconnus [Cappé, 2002] (voir aussi [Cappé et al., 1998b]) ;

**l'analyse exploratoire non supervisée de documents textuels** où l'utilisation de diverses variantes de l'échantillonneur de Gibbs a été comparée, sur une application de grande échelle, à des techniques plus simples (utilisation de l'algorithme EM, voire

algorithmes heuristiques de type « k-moyennes ») [Rigouste et al., 2006a] (voir aussi [Rigouste et al., 2006b, 2005a]).

Sans rentrer dans le détail de ces travaux, je souhaite ici exposer les principales réflexions que j'en tire en ce qui concerne mes directions de recherche actuelles et futures.

Tout d'abord, l'utilisation des techniques MCMC, et plus généralement des techniques à base de simulations, permet effectivement d'aborder des questions difficiles en traitement de signal — on trouvera de nombreux exemples dans les numéros spéciaux consacrées à ces sujets par les revues *IEEE Transactions on Signal Processing* (publié en janvier et février 2002, eds. S. Godsill et P. Djuric) et *Signal Processing* (publié en janvier 2001, eds. O. Cappé et J-Y. Tourneret). Les questions essentielles dans ce domaine restent, d'une part, la comparaisons avec des approches sous-optimales, souvent plus simple en terme de coût d'implémentation, et d'autre part, l'intégration dans une chaîne de traitement totalement automatisé dans laquelle le destinataire ultime de l'inférence n'est pas un expert statisticien mais une étape ultérieure de traitement.

Dans ce contexte, on met parfois en évidence des comportements problématiques des algorithmes MCMC — absence de communication entre certaines zones de l'espace des paramètres ou convergence très lente — qui sont difficiles à surmonter (voir [Celeux et al., 2000] pour un exemple prototypique de ces difficultés). En jouant sur la structure des modèles, il existe cependant souvent des possibilités très importantes d'amélioration par rao-blackwellisation (voir, par exemple, la section 6.3 de [Cappé et al., 2005] ou [Rigouste et al., 2006a]). Cet état de fait conduit d'ailleurs à des questions souvent difficiles à trancher sur le fait de savoir jusqu'à quel point il est légitime de modifier la structure d'un modèle, surtout quand cela concerne uniquement la modélisation a priori : dans l'exemple du modèle considéré dans [Cappé, 2002], selon le choix du modèle a priori des ruptures, il est ou non possible d'utiliser des algorithmes MCMC plus robustes utilisant des calculs exacts comme ceux proposés par [Chib, 1998] ou [Fearnhead, 2006]. En l'absence d'information plus précise sur les processus physiques qui génèrent les données, on est obligé de s'en remettre à l'examen des conséquences entraînées sur les résultats d'analyse des données, qui sont souvent empiriquement faibles pour ce type de paramètres (a priori sur les durées) par rapport à l'influence d'autres paramètres a priori (notamment sur l'amplitude des sauts lors des ruptures).

Le développement et l'implémentation des techniques MCMC est en général un travail d'ingénierie délicat, souvent très long par rapport à celui nécessaire pour d'autres techniques, et qui nécessite des compétences multiples (allant de la compréhension des principes et des modèles à des capacités de programmation). C'est en particulier vrai pour les techniques à saut réversible où les possibilités d'erreurs sont nombreuses — du calcul des ratios d'acceptation jusqu'à la mise en œuvre – et difficiles à détecter du fait, *(1)* de l'aspect (pseudo) aléatoire des résultats de l'algorithme, *(2)* du caractère réellement inconnu du comportement attendu, même en présence de données simulées (en général le seul aspect analysable reste le comportement attendu en l'absence de données). Une direction de recherche intéressantes qui répond, en partie, à cette remarque est celle des méthodes de simulations « adaptatives » qui se veulent d'une applicabilité plus générique et dans lesquelles l'algorithme calibre lui-même, au vu des résultats, ses paramètres de simulation de façon à maintenir des performances de convergence satisfaisantes (thème du workshop Adap'ski organisé par Christian Robert en janvier 2005 et du projet ADAP'MC financé par l'Agence Nationale de la Recherche et animé par Eric Moulines depuis début 2006).

Enfin, dernier point qui me semble important : le cas des données de très grande dimension. Je fais ici référence principalement au cas où les observations sont multi-

variées, avec des dimensions de l'ordre de plusieurs milliers voire plus, même si le cas où le nombre de données est grand — plusieurs centaines de milliers ou plus — est également important. Nos premiers travaux dans ce domaine [Rigouste et al., 2006a] (en utilisant un modèle de mélange de lois multinomiales avec quelques dizaines de milliers de dimensions) suggèrent que tant les modèles probabilistes que leur traitement bayésien via les techniques MCMC restent pertinents dans ce cadre, même si on observe une multiplication des comportements problématiques sous la forme de zones de l'espace des paramètres séparées par des creux de vraisemblances que les algorithmes MCMC n'arrivent pas à mettre en communication. Même le formalisme bayésien ne semble plus si efficace dans un contexte où la très grande dimensionalité des observations conduit à des facteurs de vraisemblance qui ont tendance à systématiquement prendre le pas sur les termes liés à l'a priori sur les paramètres (voir notamment dans [Rigouste et al., 2006a], l'évaluation, plutôt décevante de la version « complètement bayésienne » du classificateur *naive-bayes* utilisé traditionnellement en classification de textes). Dans ce cadre, il me semble surtout essentiel de commencer par mieux comprendre comment des modèles probabilistes très simples (multinomial pour les données discrètes, gaussiens pour les données continues) ainsi que leur version bayésienne se comportent dans des conditions où le nombre d'observations est du même ordre, voire inférieur, à la dimensionalité du modèle.

## 3.4 Algorithmes de simulation à temps continu [Cappé et al., 2003]

Pour terminer j'évoque une travail qui n'a pas eu de suite dans la mesure où il se solde par un constat plutôt négatif quoique peut être pas totalement définitif.

Le point de départ est une question assez naturelle et générale consistant à se demander dans quelle mesure il est possible de remplacer le mécanisme d'acceptation/rejet des algorithmes MCMC par la modulation des durées de séjour opérée dans les chaînes de Markov à temps continu. Avec les notations de la section 3.1, les algorithmes MCMC génèrent une chaîne de Markov à temps discret $\{X_k\}_{k\in\mathbb{N}}$ dont le noyau de transition $K$ vérifie la condition de balance détaillée (3.1) qui implique que $\pi$ est distribution stationnaire de la chaîne. L'algorithme de Metropolis-Hastings réalise ce programme dans un cadre très général, malheureusement la règle d'acceptation-rejet (3.2) implique que l'algorithme peut rester bloqué, parfois assez longtemps, en un point et qu'il a une très grande réticence à se déplacer vers des valeurs $x'$ pour lesquelles $\pi(x') \ll \pi(x)$ conduisant l'algorithme à rester éventuellement bloqué dans des régions de l'espace.

Par comparaison, dans un algorithme de simulation « à temps continu », on simule un processus de Markov à temps continu $\{X(t)\}_{t\in\mathbb{R}_+}$ où la chaîne incluse $\{X_i\}_{i\in\mathbb{N}}$ (voir notations sur le schéma de la figure 3.1) a pour noyau de transition $Q$ tandis que, conditionnellement à $X_i = x$, la durée de séjour $D_i$ est distribué selon une loi exponentielle de paramètre $\lambda(x)$. Pour le processus de Markov à temps continu, la condition de balance détaillée s'écrit

$$\pi(x)\lambda(x)Q(x,x') = \pi(x')\lambda(x')Q(x',x) \ , \quad \text{pour } x, x' \in \mathsf{X} \text{ et } x \neq x' \ , \tag{3.3}$$

où l'on a décomposé (pour les valeurs de $x \neq x'$) le générateur infinitésimal $\Lambda(x,x')$ du processus sous la forme $\lambda(x)Q(x,x')$ afin de faire apparaître le noyau de la chaîne incluse et le paramètre de la durée de séjour. La relation (3.3) laisse envisager des cas où $X(t)$ se déplace librement dans l'espace des paramètres selon le noyau $Q$, tandis

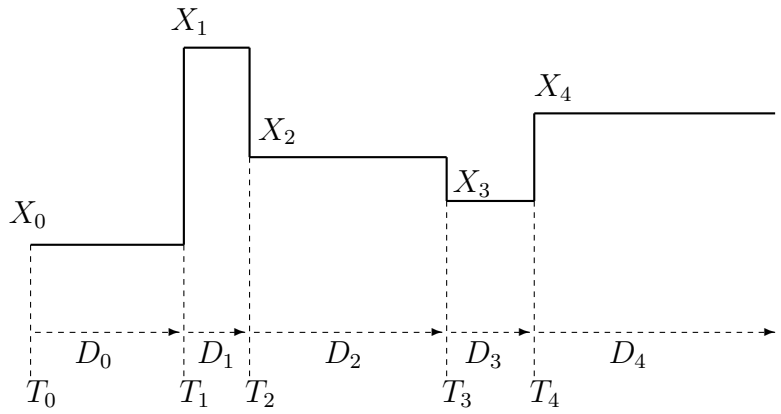


FIG. 3.1 – Principe de l'algorithme de simulation à temps continu [Cappé et al., 2003].

que la loi stationnaire $\pi$ serait rétablie à travers un choix judicieux des paramètres $\lambda(x)$. Une remarque importante (dont nous n'avions pas conscience au préalable et qui nous a été faite par Gareth Roberts) est que le caractère « à temps continu » est ici assez factice dans la mesure où l'on simule le processus $X(t)$, ce qui implique de connaître les paramètres de durée de vie $\lambda(x)$. Or dans ce cas, l'estimateur trajectoriel

$$\frac{1}{t}\int f(X(t))dt$$

de $\mathrm{E}_\pi[h(X)]$ a une variance asymptotique plus élevée que l'estimateur rao-blackwellisé[2]

$$\frac{\sum_{i=0}^{m(t)} \lambda^{-1}(X_i) f(X_i)}{\sum_{i=0}^{m(t)} \lambda^{-1}(X_i)} \,, \tag{3.4}$$

où $m(t)/t \to \mathrm{E}_\pi[\lambda^{-1}(X)]$ lorsque $t \to \infty$, dans lequel on remplacé les durées $D_i$ par leur espérance $\lambda^{-1}(X_i)$. Ainsi, une façon alternative de décrire l'algorithme proposé consiste à y voir une version markovienne de l'échantillonnage préférentiel dans laquelle on propose des valeurs successives de $X_i$, en utilisant le noyau $Q$, qui sont corrigées à l'aide des poids d'importance $\omega(X_i) = \lambda^{-1}(X_i)$. Il est clair qu'exposé de cette façon, le principe des algorithmes à temps continu est lié à celui des méthodes considérées dans la section 3.2, à la différence près que l'on ne dispose plus de populations successives de points en interaction mais d'une unique chaîne.

Malheureusement, il est difficile de trouver un cadre très général dans lequel l'équation de balance détaillée à temps continu (3.3) admet une solution satisfaisante. La première solution simple correspond au cas où les propositions $\{X_i\}_{i\in\mathbb{N}}$ sont IID ($Q(x,x') = q(x')$), auquel cas il suffit de prendre $\lambda(x) = q(x)/\pi(x)$ et (3.4) correspond exactement à l'estimateur d'échantillonnage préférentiel (auto-normalisé) habituel. Un autre cas trivial se présente lorsque le noyau $Q$ correspond à une marche aléatoire symétrique ($Q(x,x') = q(x-x')$) auquel cas on garantit (3.3) en prenant $\lambda(x) = 1/\pi(x)$. Malheureusement, si X n'est pas un espace d'état fini, la chaîne $\{X_i\}$ générée par $Q$ sera, au mieux, nulle-récurrente voir transiente, et donc (3.4) n'a aucune chance de converger rapidement vers $\mathrm{E}_\pi[h(X)]$. Cette limitation n'est toutefois pas forcément rédhibitoire en soi dans la mesure où cet algorithme pourrait être composé avec d'autres types de propositions, à même de rétablir le caractère récurrent de la chaîne.

[2]Dans le cas où les propositions $\{X_i\}$ sont IID, on vérifie aisément que la variance asymptotique de l'estimateur rao-blackwellisé est deux fois plus faible.

L'algorithme proposé par [Stephens, 2000] s'applique dans le contexte particulier de la simulation de modèles de dimension variable évoquée en fin de la section 3.1. Dans ce cadre, l'approche de [Green, 1995] conduit, dans sa forme la plus simple, à l'utilisation de mouvements quasi-déterministes où

$$Q(x, x') = \frac{1}{d(x)} \sum_{x'' \in \mathsf{X}_-(x)} \delta_{\{x''\}}(x') \ ,$$

où $d(x)$ est une dimension (entière) qui dépend de $x$ est $\mathsf{X}_-(x)$ est un sous-ensemble de X, de cardinal $d(x)$, qui lui aussi dépend de $x$, mais ne contient pas le point $x$. Pour fixer les idées, on rencontre un tel mécanisme dans un mouvement de « mort » (dans les propositions de type « naissance ou mort »), lorsque $x$ correspond à un modèle de dimension $d(x)$ et que l'on décide de supprimer au hasard une de ses composantes afin d'obtenir un modèle de dimension $d(x) - 1$, les éléments de $\mathsf{X}_-(x)$ correspondent alors aux $d(x)$ configurations obtenues à partir de $x$ en supprimant une coordonnée au hasard. Les mouvements de type « fusion » dans les propositions de type « fusion ou scission » (*split or merge* en anglais) fonctionnent sur le même principe.

En conservant cette idée d'un mouvement qui, partant de $x$, est restreint à un nombre de configurations limitées de $x' \in \mathsf{X}_-(x)$, la condition de balance détaillée (3.3) s'écrit

$$\pi(x)\Lambda(x, x') = \pi(x')\Lambda(x', x) \ , \quad \text{pour } x \in \mathsf{X} \text{ et } x' \in \mathsf{X}_-(x) \ .$$

Pour obtenir une expression totalement explicite du taux $\Lambda(x, x')$, il est nécessaire de spécifier la forme de $\Lambda(x', x)$. La solution la plus simple consiste à choisir le taux de mouvements hors de $x' \in \mathsf{X}_-(x)$ constant (on note $\lambda_+$ la valeur commune de $\lambda(x')$ pour $x' \in \mathsf{X}_-(x)$) et de spécifier le noyau de transition $Q(x', \cdot)$ utilisé lors de ces mouvements. On obtient alors

$$\Lambda(x, x') = \frac{\pi(x')}{\pi(x)} \lambda_+ Q(x', x) \ , \tag{3.5}$$

pour un choix arbitraire du noyau instrumental $Q$ utilisé lors du mouvement inverse de $x'$ vers $x$ (dit mouvement de « naissance » dans le cas des mouvements de naissance ou mort mentionnés ci-dessus). L'équation (3.5) est exactement la condition d'équilibre utilisée par [Stephens, 2000] dans le cas des mouvement de naissance ou mort mais elle s'applique manifestement de façon plus générale.

Le paramètre de la durée de séjour en $x$ s'obtient, à partir de (3.5), par $\lambda(x) = \sum_{x' \in \mathsf{X}_-(x)} \Lambda(x, x')$ tandis que la probabilité de transition correspondante est donnée par $\mathrm{P}(X_{i+1} = x' | X_i = x) = \Lambda(x, x')/\lambda(x)$. En intégrant la remarque de rao-blackwellisation des durées mentionnées ci-dessus (remarque qui n'apparaissait pas dans l'article de [Stephens, 2000]) l'algorithme correspondant est le suivant :

**Naissance** Affecter à $X_i$ le poids d'importance $\omega(X_i) = \lambda_+^{-1}$ et simuler $X_{i+1}$ selon $Q(X_i, \cdot)$.

**Mort** Affecter à $X_i$ le poids d'importance

$$\omega(X_i) = \left( \sum_{x' \in \mathsf{X}_-(X_i)} \frac{\pi(x')}{\pi(X_i)} \lambda_+ Q(x', X_i) \right)^{-1}$$

et simuler $X_{i+1}$ avec probabilité

$$\mathrm{P}(X_{i+1} = x' | X_i) = \omega(X_i) \times \frac{\pi(x')}{\pi(X_i)} \lambda_+ Q(x', X_i) \ ,$$

pour $x' \in \mathsf{X}_-(X_i)$.

Cet algorithme est correct lorsque, pour chaque configuration de $x$ donnée, un seul type de mouvement est possible. C'est notamment le cas quand $\pi(x)$ est la loi a posteriori dans une situation où le nombre de modèles en compétition est égal à deux. Dans des cas plus généraux, on vérifie qu'il est possible de mettre en compétition l'ensemble des mouvements possibles en un point $x$ donné en sommant les intensités — c'est à dire les valeurs de $\Lambda(x, x')$ — correspondant à ces différents types de mouvements [Cappé et al., 2003].

Comparé à sa variante MCMC, cet algorithme présente le défaut — signalé dans l'article non publié de [Clifford & Nicholls, 1994] — de nécessiter l'évaluation de la loi cible $\pi$ pour l'ensemble des configurations $x'$ atteignables depuis $X_i$ lors d'un mouvement de mort. Par comparaison, l'algorithme MCMC à saut réversible propose (avec une probabilité uniforme) l'*une de ces configurations* et seul le calcul de la loi cible pour cette configuration sera nécessaire pour évaluer la probabilité d'acceptation. La variante à temps continu est donc, en général, plus coûteuse à implémenter. Malgré ce défaut, les résultats présentés dans [Stephens, 2000] pour le modèle de mélange de lois gaussiennes (univariées) à nombre de composantes inconnu étaient plutôt encourageants. La comparaison présentée par [Stephens, 2000] est toutefois assez indirecte puisque l'algorithme MCMC de référence est celui de [Richardson & Green, 1997] et qu'entre les deux algorithmes de nombreux détails d'implémentation diffèrent : le choix des a priori (échangeables dans [Stephens, 2000], ordonnés dans [Richardson & Green, 1997]), l'absence de mouvements de type fusion ou scission dans [Stephens, 2000] et surtout l'utilisation de l'augmentation de données (simulation des indicatrices d'appartenance aux composantes de mélange) dans [Richardson & Green, 1997] alors que [Stephens, 2000] a recours au calcul direct de la vraisemblance (sans introduire les variables indicatrices).

Dans ce contexte, les contributions de [Cappé et al., 2003] sont essentiellement

1. de replacer la méthode de [Stephens, 2000] dans un cadre plus général en montrant que l'équation d'équilibre (3.5) peut être appliquée à la plupart des mouvements utilisés dans les algorithmes de type MCMC à saut réversible, avec de fort liens entre les algorithmes correspondant[3] ;
2. de comparer les deux types d'algorithmes (MCMC et à temps continu) en utilisant des mouvements strictement équivalents dans les deux cas, avec comme conclusion le fait que les deux types d'algorithmes produisent des résultats qui ne sont pas statistiquement distinguables (voir un exemple typique sur la figure 3.2) ;
3. de montrer, pour justifier l'observation précédente, que l'algorithme à temps continu peut être vu (en utilisant une construction très standard) comme la limite d'une séquence d'algorithmes MCMC ralentis (dans lesquels le nombre moyens de mouvements acceptés tend vers zéro) rééchantillonnés à un rythme de plus en plus lent (théorème 1 de [Cappé et al., 2003]).

Dans ces conditions, et compte tenu des questions de coût d'implémentation de l'algorithme à temps continu mentionnées ci-dessus, la conclusion de [Cappé et al., 2003] reste nettement favorable à l'algorithme MCMC à saut réversible classique.

[3] En particulier l'apparition d'un terme de Jacobien souvent abordée de façon assez obscure, notamment dans [Stephens, 2000], se pose de façon identique pour les deux types de méthodes.

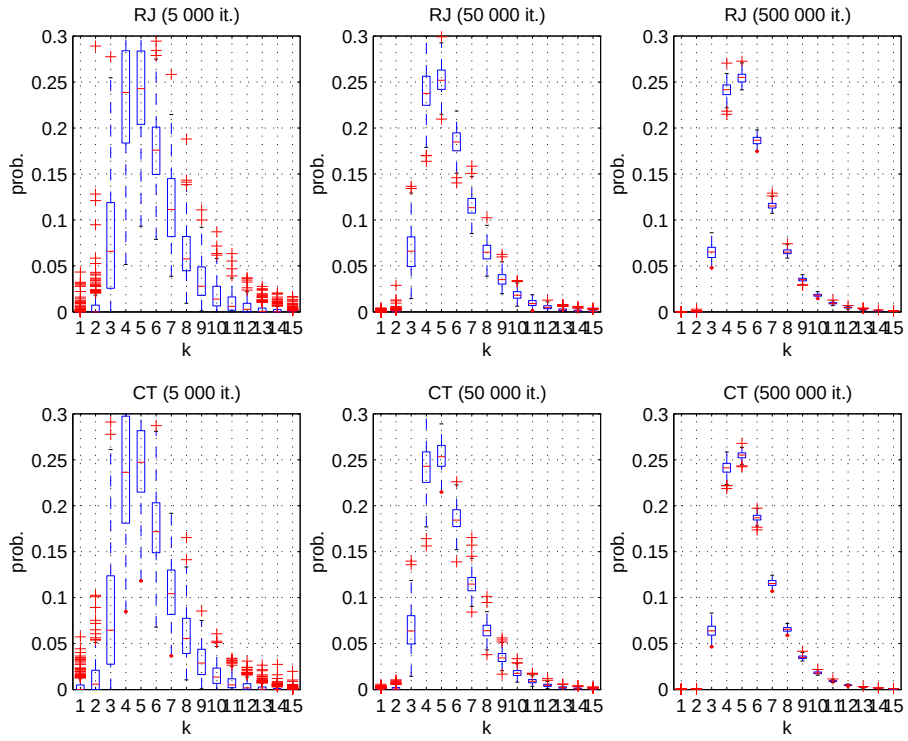


FIG. 3.2 – Estimation de la loi a posteriori du nombre de composantes dans un modèle de mélange de gaussiennes à nombre de composantes inconnues en utilisant l'algorithme MCMC à saut réversible (figures du haut — label « RJ ») ou l'algorithme à temps continu (en bas – label « CT »). Le nombre de simulations varie de, de gauche à droite, 5000 à 500000 itérations. Chaque figure résume 200 réalisations indépendantes sous forme de boxplots [Cappé et al., 2003].

# Bibliographie


A. T. Andersen & B. F. Nielsen. A Markovian approach for modeling packet traffic with long-range dependence. *IEEE J. Selected Areas Commun.*, 16(5) :719–732, 1998.

C. Andrieu, A. Doucet & V. B. Tadic. Online simulation-based methods for parameter estimation in non linear non gaussian state-space models. In *Proc. IEEE Conf. Decis. Control*, 2005.

C. Andrieu & C. Robert. Controlled markov chain monte carlo methods for optimal sampling. Rapport technique 125, Cahiers du Ceremade, 2001.

M. Arulampalam, S. Maskell, N. Gordon & T. Clapp. A tutorial on particle filters for on line non-linear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.*, 50 :241–254, 2002.

L. Bahl, J. Cocke, F. Jelinek & J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Inform. Theory*, 20(2) :284–287, Mar. 1974.

F. G. Ball & J. H. Rice. Stochastic models for ion channels : Introduction and bibliography. *Math. Biosci.*, 112 :189–206, 1992.

M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Process.*, 18(4) :349–369, 1989.

L. E. Baum & J. A. Eagon. An inequality with applications to statistical estimation for probalistic functions of Markov processes and to a model for ecology. *Bull. Am. Math. Soc.*, 73 :360–363, 1967.

L. E. Baum, T. P. Petrie, G. Soules & N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41(1) :164–171, 1970.

A. Beskos, O. Papaspiliopoulos, G. O. Roberts & P. Fearnhead. Exact and efficient likelihood-based estimation for discretely observed diffusions processes. To appear in J. R. Stat. Soc. Ser. B Stat. Methodol., 2005.

D. M. Blei, A. Y. Ng & M. I. Jordan. Latent Dirichlet allocation. In T. G. Dietterich, S. Becker & Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA, 2002. MIT Press.

J.-C. Bolot & M. Grossglauser. Parsimonious Markov modeling of processes with long range dependence. Rapport technique 2835, INRIA, 1996.

J. Booth & J. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 61 :265–285, 1999.

M. Briers, A. Doucet & S. Maskell. Smoothing algorithms for state-space models. Rapport technique TR-CUED-F-INFENG 498, University of Cambridge, Department of Engineering, 2004.

P. J. Brockwell & R. A. Davis. *Time Series : Theory and Methods*. Springer, 2nd edition, 1991.

V. Buchoux, O. Cappé & E. Moulines. Turbo multiuser detection for coded ds-cdma systems : A gibbs sampling approach. In *34th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, October 29 - November 1 2000a. URL http://www.tsi.enst.fr/~cappe/papers/bchx_mud2000.ps.gz.

V. Buchoux, O. Cappé, E. Moulines & A. Gorokhov. On the performance of semi-blind subspace-based channel estimation. *IEEE Trans. Signal Processing*, 48(6) :1750–1759, 2000b. URL http://dx.doi.org/10.1109/78.845932.

W. Buntine & A. Jakulin. Applying discrete PCA in data analysis. In M. Chickering & J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04)*, pages 59–66. AUAI Press, 2004.

M. Campedel-Oudot, O. Cappé & E. Moulines. Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach. *IEEE Trans. Speech and Audio Processing*, 9(5) :469–481, July 2001. URL http://www.tsi.enst.fr/~cappe/papers/env99.ps.gz.

F. Campillo & F. Le Gland. MLE for patially observed diffusions : Direct maximization vs. the EM algorithm. *Stoch. Proc. App.*, 33 :245–274, 1989.

O. Cappé. *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux*. Thése, Ecole Nationale Supérieure des Télécommunications, Paris, septembre 1993. URL http://www.tsi.enst.fr/~cappe/papers/these.ps.gz.

O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Trans. Speech and Audio Processing*, 2(2) :345–349, April 1994. URL http://dx.doi.org/10.1109/89.279283.

O. Cappé. Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. *Monte Carlo Methods and Applications*, 7(1–2) : 81–92, 2001a. URL http://www.tsi.enst.fr/~cappe/papers/ma_rmlpa.ps.gz.

O. Cappé. Ten years of HMMs (online bibliography 1989–2000), Mar. 2001b. URL http://www.tsi.enst.fr/~cappe/docs/hmmbib.html.

O. Cappé. A bayesian approach for simultaneous segmentation and classification of count data. *IEEE Trans. Signal Processing*, 50(2) :400–410, February 2002. URL http://www.tsi.enst.fr/~cappe/papers/cntdat.ps.gz.

O. Cappé, V. Buchoux & E. Moulines. Quasi-Newton method for maximum likelihood estimation of hidden Markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, volume 4, pages 2265–2268, 1998a.

O. Cappé, M. Charbit & E. Moulines. Recursive EM algorithm with applications to DOA estimation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Toulouse, France, may 2006. URL http://www.tsi.enst.fr/~cappe/papers/06icassp_ccm.pdf.

O. Cappé, R. Douc, E. Moulines & C. Robert. Bayesian analysis of overdispersed count data with application to teletraffic monitoring. In R. Payne & P. Green, editors, *COMPSTAT 1998 Proceedings in Computational Statistics*, pages 215–220. Physica-Verlag, 1998b.

O. Cappé, A. Doucet, M. Lavielle & E. Moulines. Simulation-based methods for blind maximum-likelihood filter identification. *Signal Processing*, 73(1–2) :3–25, 1999. URL `http://www.tsi.enst.fr/~cappe/papers/spdcv.ps.gz`.

O. Cappé, A. Guillin, J.-M. Marin & C. P. Robert. Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4) :907–929, 2004. URL `http://www.tsi.enst.fr/~cappe/papers/ion02.ps.gz`.

O. Cappé & J. Laroche. Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings. *IEEE Trans. Speech and Audio Processing*, 3(1) : 84–93, January 1995. URL `http://dx.doi.org/10.1109/89.365378`.

O. Cappé, C. Mokbel, D. Jouvet & E. Moulines. An algorithm for maximum likelihood estimation of hidden Markov models with unknown state-tying. *IEEE Trans. Speech and Audio Processing*, 6(1) :61–70, January 1998c. URL `http://dx.doi.org/10.1109/89.650312`.

O. Cappé & E. Moulines. Regularization techniques for discrete cepstrum estimation. *IEEE Signal Process. Lett.*, 3(4) :100–102, April 1996. URL `http://dx.doi.org/10.1109/97.489060`.

O. Cappé & E. Moulines. On the use of particle filtering for maximum likelihood parameter estimation. In *European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, September 2005a. URL `http://www.tsi.enst.fr/~cappe/papers/05eusipco_cm.pdf`.

O. Cappé & E. Moulines. Recursive computation of the score and observed information matrix in hidden markov models. In *IEEE Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, France, July 2005b. URL `http://www.tsi.enst.fr/~cappe/05ssp_cm.pdf`.

O. Cappé, E. Moulines, J.-C. Pesquet, A. Petropulu & Z. Yang. Long-range dependence and heavy-tail modeling for teletraffic data. *IEEE Signal Processing Magazine*, 19(3) : 14–27, May 2002. URL `http://www.tsi.enst.fr/~cappe/papers/lrd.pdf`.

O. Cappé, E. Moulines & T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005. URL `http://www.tsi.enst.fr/~cappe/ihmm/`.

O. Cappé, C. Robert & T. Rydén. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. Royal Statist. Soc. Ser. B*, 65 (3) :679–700, 2003. URL `http://www.tsi.enst.fr/~cappe/papers/crr01ct.ps.gz`.

O. Cappé & F. Roueff. Evaluation numérique de l'information de Fisher pour des observations irrégulières de l'état d'une file d'attente. In *Actes du colloque du GRETSI*, Paris, France, septembre 2003. URL `http://www.tsi.enst.fr/~cappe/papers/metro02-gretsi03.ps.gz`.

C. K. Carter & R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3) : 541–553, 1994.

C. K. Carter & R. Kohn. Markov chain Monte Carlo in conditionnaly Gaussian state space models. *Biometrika*, 83(3) :589–601, 1996.

G. Celeux & J. Diebolt. The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist.*, 2 :73–82, 1985.

G. Celeux, M. Hurn & C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Assoc.*, 95(3) :957–979, 2000.

G. Celeux, J. M. Marin & C. P. Robert. Iterated importance sampling in missing data problems. Rapport technique 326, Univ. Paris Dauphine, 2003.

F. Cérou, F. Le Gland & N. Newton. Stochastic particle methods for linear tangent filtering equations. In J.-L. Menaldi, E. Rofman & A. Sulem, editors, *Optimal Control and PDE's - Innovations and Applications, in Honor of Alain Bensoussan's 60th Anniversary*, pages 231–240. IOS Press, Amsterdam, 2001.

S. Chib. Estimation and comparison of multiple change point models. *J. Econometrics*, 86 :221–241, 1998.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89 :539–552, 2002.

G. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin Mathematical Biology*, 51 :79–94, 1989.

P. Clifford & G. Nicholls. Comparison of birth-and-death and Metropolis-Hastings Markov chain Monte Carlo for the Strauss process. Rapport technique, Department of Statistics, Oxford University, 1994.

I. B. Collings & T. Rydén. A new maximum likelihood gradient algorithm for on-line hidden Markov model identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, volume 4, pages 2261–2264, 1998.

T. M. Cover & J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

P. Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

P. Del Moral, A. Doucet & A. Jasra. Sequential monte carlo samplers. *J. Roy. Statist. Soc. Ser. B*, 68(3) :411, 2006.

P. Del Moral & A. Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. *Annales de l'Institut Henri Poincaré*, 37 :155–194, 2001.

B. Delyon, M. Lavielle & E. Moulines. On a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1), 1999.

A. P. Dempster, N. M. Laird & D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1) :1–38 (with discussion), 1977.

R. Douc, O. Cappé & E. Moulines. Comparison of resampling schemes for particle filtering. In *4th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Zagreb, Croatia, September 2005a. URL `http://arxiv.org/cs.CE/0507025`.

R. Douc, O. Cappé, E. Moulines & C. P. Robert. On the convergence of the Monte Carlo maximum likelihood method for latent variable models. *Scandinavian Journal of Statistics*, 29(4), 2002. URL `http://www.tsi.enst.fr/~cappe/papers/mcml.ps.gz`.

R. Douc, A. Guillin, J.-M. Marin & C. P. Robert. Convergence of adaptive sampling schemes. Rapport technique 2005-6, CEREMADE, 2005b.

R. Douc, A. Guillin, J.-M. Marin & C. P. Robert. Minimum variance adaptive sampling. Rapport technique, CEREMADE, 2005c.

R. Douc & E. Moulines. Limit theorems for weighted samples with applications to sequential Monte Carlo. Preprint, July 2005.

R. Douc, E. Moulines & T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32(5) :2254–2304, 2004.

A. Doucet & C. Andrieu. Iterative algorithms for state estimation of jump Markov linear systems. *IEEE Trans. Signal Process.*, 49(6) :1216–1227, 2001.

A. Doucet, N. De Freitas & N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001a.

A. Doucet, S. Godsill & C. Andrieu. On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, 10 :197–208, 2000.

A. Doucet, N. Gordon & V. Krishnamurthy. Particle filters for state estimation of jump Markov linear systems. *IEEE Trans. Signal Process.*, 49 :613–624, 2001b.

O. Elerian, S. Chib & N. Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4) :959–993, 2001.

R. J. Elliott, L. Aggoun & J. B. Moore. *Hidden Markov Models : Estimation and Control*. Springer, New York, 1995.

Y. Ephraim & N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48 : 1518–1569, June 2002.

P. Fearnhead. Markov chain Monte Carlo, sufficient statistics and particle filter. *J. Comput. Graph. Statist.*, 11(4) :848–862, 2002.

P. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Stat. Comput.*, 16 :203–213, 2006.

P. Fearnhead & P. Clifford. On-line inference for hidden Markov models via particle filters. *J. Roy. Statist. Soc. Ser. B*, 65 :887–899, 2003.

J. A. Fessler & A. O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized em algorithms. *IEEE Trans. Image Process.*, 4(10) : 1417–1429, 1995.

W. Fischer & K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance evaluation*, 18(2), 1993.

G. Fort & E. Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.*, 31(4) :1220–1259, 2003.

D. R. Fredkin & J. A. Rice. Maximum-likelihood-estimation and identification directly from single-channel recordings. *Proc. Roy. Soc. London Ser. B*, 249 :125–132, 1992.

A. E. Gelfand & A. F. M. Smith. Sampling based approaches to calculating marginal densities. *J. Am. Statist. Assoc.*, 85 :398–409, 1990.

J. Geweke. Bayesian inference in econometric models using Monte-Carlo integration. *Econometrica*, 57(6) :1317–1339, 1989.

C. J. Geyer. Estimation and optimization of functions. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman, 1996.

C. J. Geyer & E. A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B*, 54(3) :657–699, 1992.

W. R. Gilks, S. Richardson & D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics Series. Chapman & Hall, 1996.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Commun. ACM*, 33(10) :75–84, 1990.

S. Godsill, P. Rayner & O. Cappé. Digital audio restoration. In M. Kahrs & K. Brandenburg, editors, *Applications of Signal Processing to Audio and Acoustics*. Kluwer Academic Publishers, 1998.

S. J. Godsill, A. Doucet & M. West. Monte carlo smoothing for non-linear time series. *J. Am. Statist. Assoc.*, 50 :438–449, 2004.

N. Gordon, D. Salmond & A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, 140 :107–113, 1993.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82 :711–732, 1995.

T. L. Griffiths & M. Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.

H. Haario, E. Saksman & J. Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14 :375–395, 1999.

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

J. Handschin. Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6 :555–563, 1970.

J. Handschin & D. Mayne. Monte Carlo techniques to estimate the conditionnal expectation in multi-stage non-linear filtering. In *Int. J. Control*, volume 9, pages 547–559, 1969.

T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2) :185–194, 1995.

M. E. A. Hodgson. *Reversible jump Markov chain Monte Carlo and inference for ion channel data*. Thése, University of Bristol, 1998.

C. Hue, J.-P. Le Cadre & P. Pérez. Sequential Monte Carlo methods for multiple target tracking and data fusion. *IEEE Trans. Signal Process.*, 50(2) :309–325, 2002.

J. Hull & A. White. The pricing of options on assets with stochastic volatilities. *J. Finance*, 42 :281–300, 1987.

M. Hürzeler & H. R. Künsch. Monte Carlo approximations for general state-space models. *J. Comput. Graph. Statist.*, 7 :175–193, 1998.

E. Jacquier, N. G. Polson & P. E. Rossi. Bayesian analysis of stochastic volatility models (with discussion). *J. Bus. Econom. Statist.*, 12 :371–417, 1994.

M. Jamshidian & R. J. Jennrich. Acceleration of the EM algorithm using quasi-Newton methods. *J. Roy. Statist. Soc. Ser. B*, 59(3) :569–587, 1997.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola & L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

T. Kailath, A. Sayed & B. Hassibi. *Linear Estimation*. Prentice-Hall, 2000.

G. K. Kaleh & R. Vallet. Joint parameter estimation and symbol detection for linear or nonlinear unknown channels. *IEEE Trans. Commun.*, 42(7) :2406–2413, 1994.

R. E. Kalman & R. Bucy. New results in linear filtering and prediction theory. *J. Basic Eng., Trans. ASME, Series D*, 83(3) :95–108, 1961.

S. Kim, N. Shephard & S. Chib. Stochastic volatility : Likelihood inference and comparison with ARCH models. *Rev. Econom. Stud.*, 65 :361–394, 1998.

G. Kitagawa. Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, 1 :1–25, 1996.

J. Kormylo & J. M. Mendel. Maximum-likelihood detection and estimation of Bernoulli-Gaussian processes. *IEEE Trans. Inform. Theory*, 28 :482–488, 1982.

A. Krogh, I. S. Mian & D. Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res.*, 22 :4768–4778, 1994.

A. Kundu, Y. He & P. Bahl. Recognition of handwritten word : first and second order hidden Markov model based approach. *Pattern recognition*, 22(3), 1989.

H. R. Künsch. Recursive Monte-Carlo filters : algorithms and theoretical analysis. *Ann. Statist.*, 33(5) :1983–2021, 2005.

J. Lafferty, A. McCallum & F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 57(2) :425–437, 1995.

M. Lavielle & E. Lebarbier. An application of MCMC methods to the multiple change-points problem. *Signal Process.*, 81 :39–53, 2001.

F. Le Gland & L. Mevel. Recursive estimation in HMMs. In *Proc. IEEE Conf. Decis. Control*, pages 3468–3473, 1997.

F. Le Gland & N. Oudjane. Stability and uniform approximation of nonlinear filters using the hilbert metric and application to particle filters. *Ann. Appl. Probab.*, 14 :144–187, 2004.

J. Liu & R. Chen. Sequential Monte-Carlo methods for dynamic systems. *J. Roy. Statist. Soc. Ser. B*, 93 :1032–1044, 1998.

J. Liu & M. West. Combined parameter and state estimation in simulation-based filtering. In N. D. F. A. Doucet & N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 44 :226–233, 1982.

A. McCallum, D. Freitag & F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning (ICML)*, 2000.

X.-L. Meng & D. B. Rubin. Maximum likelihood estimation via the ECM algorithm : A general framework. *Biometrika*, 80(2) :267–278, 1993.

X.-L. Meng & D. Van Dyk. The EM algorithm — an old folk song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B*, 59(3) :511–567, 1997.

K. L. Mengersen & C. P. Robert. The pinball sampler. In J. M. Bernardo, A. P. Dawid, J. O. Berger & M. West, editors, *Bayesian Statistics 7*. Oxford University Press, 2003.

T. Minka & J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

E. Moulines, J.-F. Cardoso & E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997.

A. Y. Ng & M. I. Jordan. On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. In T. Dietterich, S. Becker & Z. Ghahramani, editors, *Adv. Neural Inf. Process. Syst.*, volume 14. MIT Press, 2001.

J. Olsson, O. Cappé, R. Douc & E. Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models. Rapport technique, Lund University, 2006a. URL `http://arxiv.org/math.ST/0609514`.

J. Olsson, O. Cappé, R. Douc & E. Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models. Rapport technique, Lund University, 2006b.

J. Olsson & T. Rydén. Asymptotic properties of the bootstrap particle filter maximum likelihood estimator for state space models. Rapport technique LUTFMS-5052-2005, Lund University, 2005.

G. Poyiadjis, A. Doucet & S. S. Singh. Particle methods for optimal filter derivative : application to parameter estimation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages v/925–v/928, 18-23 March 2005.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2) :257–285, Feb. 1989.

A. E. Raftery, M. A. Newton, J. Satagopan & P. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Rapport technique 499, University of Washington, Department of Statistics, 2006.

D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Commun.*, 17(2) :91–108, Aug. 1995.

S. Richardson & P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B*, 59 :731–792, 1997.

L. Rigouste, O. Cappé & F. Yvon. Inference for Probabilistic Unsupervised Text Clustering. In *IEEE Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, France, July 2005a. URL `http://www.tsi.enst.fr/publications/enst/inproceedings-2005-5737.pdf`.

L. Rigouste, O. Cappé & F. Yvon. Modèle de mélange multi-thématique pour la Fouille de Textes. In *Traitement Automatique des Langues Naturelles — Atelier DÉfi Fouille de Textes*, Dourdan, France, juin 2005b. URL `http://www.tsi.enst.fr/publications/enst/inproceedings-2005-5613.pdf`.

L. Rigouste, O. Cappé & F. Yvon. Inference and evaluation of the multinomial mixture model for text clustering. Rapport technique 2006D004, Télécom Paris, 2006a. URL `http://arxiv.org/cs.IR/0606069`.

L. Rigouste, O. Cappé & F. Yvon. Quelques observations sur le modèle LDA. In *Journées Internationales d'Analyse statistique des données textuelles*, pages 819–830, Besançon, France, avril 2006b. URL `http://www.tsi.enst.fr/publications/enst/inproceedings-2006-6308.pdf`.

B. Ristic, M. Arulampalam & A. Gordon. *Beyond Kalman Filters : Particle Filters for Target Tracking*. Artech House, 2004.

C. P. Robert & G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

S. Robert & J. LeBoudec. New models for self-similar traffic. *Performance Evaluation*, 30 (1) :57–68, 1997.

G. O. Roberts, O. Papaspiliopoulos & P. Dellaportas. Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *J. Roy. Statist. Soc. Ser. B*, 66(2) :369–393, 2004.

R. Y. Rubinstein & D. P. Kroese. *The Cross-Entropy Method*. Springer, 2004.

R. Y. Rubinstein & A. Shapiro. *Discrete Event Systems : Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley and Sons, 1993.

T. Rydén. On recursive estimation for hidden Markov models. *Stochastic Process. Appl.*, 66(1) :79–96, 1997. ISSN 0304-4149.

G. Sandmann & S. J. Koopman. Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics*, 87(2) :271–301, 1998.

N. Shephard & M. Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3) :653–667, 1997. Erratum in volume 91, 249–250, 2004.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Ann. Statist.*, 28 :40–74, 2000.

G. Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.*, pages 281–289, 2002.

R. Streit & R. Barrett. Frequency line tracking using hidden Markov models. *IEEE Trans. Acoust., Speech, Signal Process.*, 38(4), 1990.

Y. Stylianou, O. Cappé & E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, 6(2) :131–142, March 1998. URL `http://dx.doi.org/10.1109/89.661472`.

D. M. Titterington, A. F. M. Smith & U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester, 1985.

J. Vermaak, N. Ikoma & S. J. Godsill. Sequential Monte Carlo framework for extended object tracking. *IEE Proc. Radar Sonar Navig.*, 152(5) :353–363, 2005.

H. M. Wallach. Conditional random fields : An introduction. Rapport technique MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania, 2002.

G. C. G. Wei & M. A. Tanner. A Monte-Carlo implementation of the EM algorithm and the poor man's Data Augmentation algorithms. *J. Am. Statist. Assoc.*, 85 :699–704, 1991.

N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, New York, 1949.

T. Yoshihara, S. Kasahara & Y. Takah. Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telco. Systems*, 17(1-2) :185–211, 2001.

O. Zeitouni & A. Dembo. Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes. *IEEE Trans. Inform. Theory*, 34(4), July 1988.

Q. Zhu. Hidden Markov model for dynamic obstacle avoidance of mobile robot navigation. *IEEE Trans. Robot. Autom.*, 7(3), 1991.