Méthodes de Monte Carlo EM et approximations particulaires : Application à la calibration d'un modèle de volatilité stochastique.

Mouhamad M. Allaya

UNIVERSITÉ PARIS I PANTHÉON-SORBONNE
&
LABORATOIRE STATISTIQUE, ANALYSE, MODÉLISATION MULTIDISCIPLINAIRE.

T H È S E

présentée pour obtenir

LE GRADE DE DOCTEUR EN MATHÉMATIQUES APPLIQUÉES DE L'UNIVERSITÉ PARIS I PANTHÉON-SORBONNE

Spécialité : Probabilités et Statistique

par

Mouhamad M. Allaya

Méthodes de Monte-Carlo EM et approximations particulaires : Application à la calibration d'un modèle de volatilité stochastique.

soutenue publiquement le 09 décembre 2013, devant le jury composé de :

Jean-Marc Bardet	Professeur, université Paris 1 Panthéon-Sorbonne	**Directeur**
Aliou Diop	Professeur, université Gaston Berger de St-Louis	**Rapporteur**
Michel Verleysen	Professeur, université de Louvain-la-Neuve	**Rapporteur**
Marie Cottrell	Professeur, université Paris 1 Panthéon-Sorbonne	**Présidente**
Jean-Bernard Baillon	Professeur, université Paris 1 Panthéon-Sorbonne	**Examinateur**
Pierre Bertrand	Professeur, université de Clermont-Ferrand	**Examinateur**

Remerciements

Je tiens tout d'abord à dire merci à mon encadreur Jean-Marc Bardet pour avoir accepté de diriger cette thèse malgré toutes les péripéties par lesquelles je suis passé. Je tiens aussi à remercier la directrice du laboratoire SAMM, Marie Cottrell, pour avoir été constamment un soutien matériel et moral dans ces diverses épreuves avec l'aide de tout le laboratoire. Qu'elle trouve en ces quelques mots toute ma gratitude à son égard.

Je témoigne de toute mon estime à l'endroit de Michel Verleysen qui a accepté d'être rapporteur de cette thèse. Et je remercie Aliou Diop, pour avoir également accepté d'être rapporteur, mais aussi pour m'avoir formé dans les classes antérieures et initié à la recherche au sein du laboratoire LERSTAD de Saint-Louis. J'exprime aussi mes remerciements à l'endroit de Pierre Bertrand pour avoir accepté d'examiner cette thèse.

Je remercie également Jean-Bernard Baillon, Jean-Marc Bardet et Marie Cottrell pour avoir entre autres facilité mes démarches administratives pendant tout ce temps. Je remercie les membres du SAMM pour les moments de bonheur partagé autour de mets excellents d'origines diverses, à l'image du laboratoire. Je tiens aussi à dire merci aux camarades communistes, idéalistes, philosophes, capitalistes et utopistes (Fania, Béchir, Omar, Abdelkader, Souhelia, Anthikos, Ibrahima, Dhaou, Djihed, Tzirizo, Alexis, Bienvenue, William) pour les débats et quêtes de sens ô combien passionnants et passionnés.

Je suis également redevable à Madalina, le train bondé de monde, quelque part en direction de St-Rémy lès Chevreuse, à Annie, Sandie, Julien, Sonia, Charles, Pierre, Patrice, Patrick, Cécile, Fabrice, Nicolas, collègues au quotidien, pour leurs preuves d'amitiés.

En outre, je tiens à remercier vivement tous les enseignants que j'ai eus du primaire à l'université et qui m'ont transmis leurs savoirs et leur pédagogie entre autres. A l'endroit de mes promotionnaires, de mes amis (Thiam, Konté, Boubacar, Katy, Alioune, Abou, Khaly, etc.) ainsi que de ma famille adoptive des Ulis et Morangis (Ismael, Coumba, Maimouna), je leur témoigne toute ma gratitude pour leur accueil et sollicitude.

Je suis aussi redevable aux amis et promotionnaires de Montréal chez qui je termine ces quelques lignes, pour leur accueil chaleureux (Kao, Souleymane, Nenema, Racine, Khady, Albano, Aïcha et tous les autres). Qu'ils soient témoins de toute ma reconnaissance.

Enfin, mes remerciements vont à l'endroit de ma famille, soutien indéfectible : ma mère et mes frères (Ibrahim, Samsou et Noréini) qu'ils soient remerciés de toute l'affection ininterrompue dont je suis encore le récipiendaire. Les mots ne sauraient suffire pour décrire le soutien de tout ordre qu'ils m'ont apporté. Je leur suis infiniment redevable.

A mon père.

Résumé

Ce travail de thèse poursuit une perspective double dans l'usage conjoint des méthodes de Monte Carlo séquentielles (MMS) et de l'algorithme Espérance-Maximisation (EM) dans le cadre des modèles de Markov cachés présentant une structure de dépendance markovienne d'ordre supérieur à 1 au niveau de la composante inobservée. Tout d'abord, nous commençons par un exposé succinct de l'assise théorique des deux concepts statistiques à travers les chapitres 1 et 2 qui leurs sont consacrés. Dans un second temps, nous nous intéressons à la mise en pratique simultanée des deux concepts au chapitre 3 et ce dans le cadre usuel où la structure de dépendance est d'ordre 1.

L'apport des méthodes MMS dans ce travail réside dans leur capacité à approximer efficacement des fonctionnelles conditionnelles bornées, notamment des quantités de filtrage et de lissage dans un cadre non linéaire et non gaussien. Quant à l'algorithme EM, il est motivé par la présence à la fois de variables observables et inobservables (ou partiellement observées) dans les modèles de Markov Cachés et singulièrement les modèles de volatilité stochastique étudié.

Après avoir présenté aussi bien l'algorithme EM que les méthodes MCs ainsi que quelques unes de leurs propriétés dans les chapitres 1 et 2 respectivement, nous illustrons ces deux outils statistiques au travers de la calibration d'un modèle de volatilité stochastique. Cette application est effectuée pour des taux change ainsi que pour quelques indices boursiers au chapitre 3. Nous concluons ce chapitre sur un léger écart du modèle de volatilité stochastique canonique utilisé ainsi que des simulations de Monte Carlo portant sur le modèle résultant.

Enfin, nous nous efforçons dans les chapitres 4 et 5 à fournir les assises théoriques et pratiques de l'extension des méthodes Monte Carlo séquentielles notamment le filtrage

et le lissage particulaire lorsque la structure markovienne est plus prononcée. En guise d'illustration, nous donnons l'exemple d'un modèle de volatilité stochastique dégénéré dont une approximation présente une telle propriété de dépendance.

Abstract

This thesis pursues a double perspective in the joint use of sequential Monte Carlo methods (SMC) and the Expectation-Maximization algorithm (EM) under hidden Markov models having a Markov dependence structure of order grater than one in the unobserved component signal. Firstly, we begin with a brief description of the theoretical basis of both statistical concepts through Chapters 1 and 2 that are devoted. In a second hand, we focus on the simultaneous implementation of both concepts in Chapter 3 in the usual setting where the dependence structure is of order 1.

The contribution of SMC methods in this work lies in their ability to effectively approximate any bounded conditional functional in particular, those of filtering and smoothing quantities in a non-linear and non-Gaussian settings. The EM algorithm is itself motivated by the presence of both observable and unobservable (or partially observed) variables in Hidden Markov Models and particularly the stochastic volatility models in study.

Having presented the EM algorithm as well as the SMC methods and some of their properties in Chapters 1 and 2 respectively, we illustrate these two statistical tools through the calibration of a stochastic volatility model. This application is done for exchange rates and for some stock indexes in Chapter 3. We conclude this chapter on a slight departure from canonical stochastic volatility model as well Monte Carlo simulations on the resulting model.

Finally, we strive in Chapters 4 and 5 to provide the theoretical and practical foundation of sequential Monte Carlo methods extension including particle filtering and smoothing when the Markov structure is more pronounced. As an illustration, we give the example of a degenerate stochastic volatility model whose approximation has

such a dependence property.

Table des matières

Introduction **1**

Notations **5**

1 Algorithme Espérance Maximisation **7**
1.1 Introduction 8
1.2 Formulation et propriété 9
1.2.1 Notations, définitions 9
1.2.2 Formalisme 9
1.3 Applications à la famille exponentielle 15
1.3.1 Distribution multinomiale 16
1.3.2 Distribution exponentielle 18
1.4 Éléments de convergence 19
1.4.1 Convergence de la suite $\left\{L(\theta^{(k)}), k \geq 0\right\}$ 19
1.4.2 Convergence de la suite $\left\{\theta^{(k)}, k \geq 0\right\}$ 22
1.4.3 Vitesse de convergence 23
1.4.4 Critères d'arrêt 24
1.5 Quelques extensions 24
1.5.1 Monte Carlo EM 24
1.5.2 α-EM 25
1.6 Conclusion 27

2 Méthodes de Monte Carlo Séquentielles **29**
2.1 Introduction 30
2.2 Échantillonnage d'importance 31

2.2.1 Échantillonnage préférentiel 31
2.2.2 Échantillonnage préférentiel auto-normalisé 34
2.2.3 Échantillonnage préférentiel avec rééchantillonnage 38
2.3 MMC & Approximations particulaires 40
2.3.1 Filtrage particulaire 43
2.3.2 Lissage particulaire 48
2.4 Quelques résultats théoriques 53
2.4.1 Le Cadre 53
2.4.2 Analyse du filtrage 54
2.4.3 Analyse du lissage 61
2.5 Conclusion 65

3 Calibration d'un modèle de volatilité canonique 67
3.1 Genèse du modèle de volatilité 68
3.2 Données synthétiques 70
3.2.1 Jeu de données 1 70
3.2.2 jeu de données 2 71
3.3 Données réelles 73
3.3.1 Taux de change GBP/USD 73
3.3.2 Taux de change USD/YEN 74
3.3.3 Taux de change EURO/USD 74
3.3.4 Indice S&P 500 75
3.3.5 Indice Dow Jones 76
3.3.6 Indice FTSE 100 77
3.3.7 Indice Nikkei 225 78
3.4 Extension du modèle canonique 83
3.5 Conclusion 91

4 On some extensions of the SMC methods in higher-order HMM 93
4.1 Introduction 94
4.2 SMC methods in one-order HMM 95
4.2.1 Filtering recursions 96
4.2.2 Smoothing recursions 97
4.3 SMC methods in higher-order HMM 101
4.3.1 $\ell-$order Filtering recursions 102
4.3.2 ℓ-order smoothing recursions 105
4.4 Parameter estimation 108
4.5 Convergence issues 111
4.5.1 L_q mean error 111
4.6 Proof of Prop.4.3.2 113

5 Stochastic Volatility Model through GEM algorithm and SMC methods. 117
5.1 Framework 118
5.2 EM algorithm under latent data Model 119
5.2.1 Model specification 120
5.2.2 Generalized EM algorithm for parameter estimation 121
5.3 Sequential Monte Carlo approximations 125
5.3.1 Forward sampling 126
5.3.2 Forward-Backward algorithm 128
5.3.3 Application 139
5.3.4 GEM convergence 139

Conclusion 143

A Annexe A 145
A.1 Variables aléatoires et chaînes de Markov 145
A.1.1 Sommes de variables aléatoires 145
A.1.2 Chaînes de Markov 147
A.2 Mesure de dégénérescence 148
A.2.1 Coefficient de variation 149
A.2.2 Nombre de particules efficaces 149
A.2.3 Entropie de Shannon 149
A.3 Schémas de rééchantillonnage 150
A.3.1 Redistribution Multinomiale 150
A.3.2 Redistribution Résiduelle 151
A.3.3 Redistribution Systématique 152

Bibliographie 153

Table des figures

1.1 Mélange gaussien et itérations de l'EM. 15

2.1 Trajectoire du modèle de croissance sur un horizon de $T = 500$ réalisations. 42

3.1 Trajectoire de (3.7) sur un horizon de $T = 500$ réalisations & Itérations du MCEM pour $N = 200$ particules. 71
3.2 Trajectoire de (3.7) sur un horizon de $T = 4000$ réalisations & Itérations du MCEM pour $N = 200$ particules. 72
3.3 Analyse du taux de change GBP/USD. 73
3.4 Analyse du taux de change USD/YEN. 74
3.5 Analyse du taux de change EURO/USD. 75
3.6 Analyse de l'indice S&P 500. 76
3.7 Analyse de l'indice Dow Jones. 77
3.8 Analyse de l'indice FTSE 100. 78
3.9 Analyse de l'indice Nikkei 225. 79
3.10 Trajectoire du modèle (3.11) sur un horizon de $T = 500$ réalisations. . . 81
3.11 Histogrammes des paramètres sur 150 expériences de Monte Carlo. . . 82
3.12 150 expériences de Monte Carlo : Évolution du MCEM en fonction du nombre de particules pour $T = 500$ fixé. 83
3.13 *de haut en bas* 150 expériences de Monte Carlo : Évolution des 50 dernières itérations du MCEM en fonction du nombre de particules pour $T = 500$ fixé avec les histogrammes correspondants. 84
3.14 150 expériences de Monte Carlo : Évolution du MCEM en fonction de la longueur de la trajectoire pour $N = 250$ fixé. 88

3.15 *de haut en bas* 150 expériences de Monte Carlo : Évolution des 50 dernières itérations du MCEM en fonction de la longueur de la trajectoire pour un nombre de particules fixé à $N = 250$ ainsi que les histogrammes correspondants. 91

4.1 MCEM iterations for (4.58) generated with $\theta^* = (0.7, -0.15, 0.2, 0.3)$. 110
4.2 MCEM iterations for 4.60 with $\theta^* = (0.8, 0.1, \sqrt{0.3}, -0.8612)$ 111

5.1 Sample path and MCEM iterations 140

Introduction

Ce travail de thèse a pour objet d'étendre les méthodes de Monte Carlo séquentielles au cadre de modèles à espace d'états généraux ayant un certain degré de mémoire. Typiquement, nous nous intéressons aux modèles de Markov cachés (MMC) d'ordre strictement plus grand que un. Le prétexte de cette extension est dicté par le fait que certaines situations physiques modélisées par des MMC peuvent présenter une forme de dépendance basée sur une mesure de l'éloignement du signal caché avec ses retards. Ce qui peut se traduire par une structure markovienne plus prononcée du signal inobservé et pouvant également se retrouver dans les observations faites du signal recueilli ou observable.

Problématique

Il s'agit donc d'estimer un modèle à espace d'états discret non linéaire se prêtant à une structure de dépendance markovienne d'ordre $\ell > 1$ au niveau du signal caché. Notons que l'on ne s'intéresse pas à l'estimation de cet ordre bien que celui-ci puisse être fait par pénalisation de vraisemblance (voir à ce propos Finesso [1990], Van Handel [2011]). Mais, on suppose que celui-ci est connu et l'on s'intéresse aux problématiques de filtrage et de lissage dans de tels modèles. L'approche classique de Kalman [1960] pour solutionner de telles problématiques dans le cadre de modèles à espace d'états se trouve éprouvée lorsque l'on se départit des hypothèses de linéarité des modèles ou de gaussianité de leurs bruits d'états ou d'observations. Malgré les extensions du filtre Kalman basique telles que le filtre de Kalman étendu ou le filtre de Kalman linéarisé pour contourner certaines de ces limitations, il convient de souligner que celles-ci sont prises en défaut lorsque par exemple la non-linéarité ou la non-guassianité est très prononcée. D'où la nécessité d'user de méthodes d'approximation telles que les

méthodes de Monte Carlo séquentielles (MCs) qui sont non seulement indépendantes de la nature du bruit d'état ou d'observation mais aussi de la linéarité ou non du modèle étudié. En guise d'exemple, nous illustrerons ces problématiques au travers d'un modèle à volatilité stochastique dégénéré nécessitant de les résoudre.

Outils utilisés

Ce travail de thèse s'appuie essentiellement sur deux piliers : l'algorithme Espérance-Maximisation et les méthodes de Monte Carlo séquentielles. L'algorithme EM, issu des travaux de Dempster et al. [1977] est un procédé d'optimisation itératif de vraisemblance de modèles à données manquantes ou partiellement observées. Sa simplicité, sa modularité ont fini de le populariser à travers diverses problématiques d'inférence statistique où la vraisemblance du modèle en étude est difficile à mettre en œuvre par les outils habituels de maximisation. L'idée sous-jacente de cet algorithme est de "compléter" les données manquantes afin d'en déduire une vraisemblance plus malléable se prêtant plus facilement à une maximisation. Les méthodes de Monte Carlo Séquentielles quant à elles constituent une extension des méthodes de Monte Carlo par Chaînes de Markov dans un cadre itératif où la dimension des fonctionnelles approximées croit avec le temps. Elles constituent par conséquent la version en ligne des MCMC qui elles sont adaptées aux fonctionnelles à dimension statique. Depuis, l'article séminal de Gordon et al. [1993], les méthodes de MC séquentielles n'ont cessé de susciter de l'intérêt tant leur champs d'applications est vaste (ingénierie industrielle, militaire, etc.). Elles permettent entre autres d'estimer, de prévoir l'état d'un système non directement observable à un instant donné.

Démarche entreprise

Il s'agit dans un premier temps d'étendre les méthodes de Monte Carlo séquentielles aux problématiques de filtrage et de lissage lorsque la chaîne cachée sous-jacente est markovienne d'ordre supérieur à un. Dans un deuxième temps d'utiliser une extension de l'algorithme EM comme outil d'inférence sur les paramètres des modèles de Markov cachés, singulièrement un modèle de volatilité stochastique dégénéré. De prime abord, la problématique des modèles de Markov d'ordre $\ell > 1$ est relativement simple. Puisqu'une chaîne de Markov d'ordre $\ell > 1$ peut naturellement être vue comme une chaîne de Markov d'ordre 1. Pour se faire, il suffit de prendre suffisamment de recul dans le passé de la chaîne pour former une nouvelle chaîne de Markov vectorielle dont chaque réalisation est formée de l'empilement des retards de la chaîne originelle. D'une part, cette résultante a l'avantage de s'inscrire dans le schéma habituel pour les problématiques d'inférence. Cependant, cette réécriture introduit au moins deux contraintes

additionnelles : d'une part, la résultante des bruits de mesure issue d'une telle transformation est très souvent dégénérée. De plus, outre la dépendance inter-composante de la chaîne on a également une dépendance intra-composante issue d'une telle transformation. Dans notre exemple, le modèle de volatilité étudié est tout simplement dégénéré. Il faut par conséquent recourir à des transformations du modèle originel afin d'obtenir une réécriture approchée pouvant se prêter à une inférence. Il faut donc trouver un bon compromis entre complexité et estimabilité du modèle approché eu égards à la particularité de chacune des réécritures possibles pour ce type de modèle.

Plan

Nous commençons tout d'abord par un exposé succinct des deux fondements de cette étude à savoir l'EM et les méthodes de MCs. Ainsi, les chapitres 1 et 2 sont consacrés à l'assise théorique de ces deux concepts statistiques sans être exhaustif. Nous ponctuons cet exposé par un chapitre applicatif (chapitre 3) où nous calibrons le modèle de volatilité canonique aux données synthétiques et réelles. Nous en profitons pour faire une petite extension de ce modèle de base dans le but d'éprouver la méthode d'estimation. Quant aux chapitres 4 et 5 ils se proposent de donner une extension des méthodes de MCs pour le cas des modèles de Markov cachés d'ordre supérieur. Le résumé des points développés est donné ci-dessous.

Chapitre 1

Ce chapitre est dédié à l'étude de l'algorithme EM à travers quelques unes de ses applications ainsi que de ses extensions. Nous abordons les fondements théoriques de cet algorithme par le biais de résultats de convergence sur les estimés qui en découlent. Nous donnons aussi un petit aperçu d'extensions possibles de l'EM de base et notamment celle qui nous est utile dans la suite.

Chapitre 2

Dans ce chapitre nous mettons en relief les méthodes de Monte Carlo séquentielles dans le cadre d'un MMC d'ordre 1. Singulièrement, nous traitons du filtrage et du lissage particulaires comme recours lorsque les approches classiques sont limitées. Nous explorons quelques résultats d'analyse théorique qui permettent de les sous-tendre.

Chapitre 3

Nous nous attelons de prime abord en la mise en oeuvre conjointe de l'algorithme EM et des méthodes MCs à travers le modèle de volatilité stochastique canonique. Dans un second temps nous adjoignons une contribution qui consiste à se départir du modèle

de base en introduisant un terme perturbateur en $X^2 + \cos(X^2)$ dans la dynamique du signal. Tout comme le modèle canonique, ce modèle est également estimé par MCEM. Quelques simulations de Monte Carlo permettent de le valider.

Chapitre 4

Dans ce chapitre nous donnons un aperçu théorique de l'extension des méthodes MCs au cadre de chaînes de Markov d'ordre supérieur à 1. Nous commençons par un rappel des rudiments basiques utilisés lors de l'établissement des équations de filtrage et de lissage particulaires. Nous donnons une adaptation permettant de tenir compte de plus de mémoire dans la structure du modèle caché.

Chapitre 5

Nous approchons un modèle de volatilité stochastique multi-échelles dégénéré par un modèle de Markov caché avec une structure de dépendance plus accrue et ce, au moyen de l'extension faite précédemment au chapitre 4. Ce chapitre est donc une forme d'applications du chapitre 4.

Notations

Abréviations :
a.k.a : *also known as* ;
EM : Espérance Maximisation ;
EP : Échantillonnage Préférentiel ;
EPR : Échantillonnage Préférentiel avec Rééchantillonnage ;
EQM : erreur quadratique moyenne ;
LFGN : Loi Forte des Grands Nombres ;
MAP : Maximum *a posteriori* ;
MCEM : Monte Carlo EM ;
MCs : Monte Carlo séquentielles ;
MMC : Modèles de Markov Cachés ;
REQM : racine de l'erreur quadratique moyenne ;
RHS : *right hand side* ;
SMC : *Sequential Monte Carlo* ;
TLC : Théorème Limite Central ;
w.r.t : *with respect to* ;

Ensembles :
$X_{r:m} := (X_r, X_{r+1}, X_{r+2}, \ldots, X_m)$ vecteur de variables aléatoires ;
$\mathcal{X}$: espace d'états ;
$\mathcal{Y}$: espace des observations ;
$\mathbb{F}_b(\mathcal{X})$: ensemble des fonctions mesurables et bornées sur $\mathcal{X}$;
$\mathcal{C}_b(\mathcal{X})$: Ensemble des fonctions continues et bornées sur $\mathcal{X}$;

Algorithme EM :
θ : vecteur de paramètres ;

Θ : espace des paramètres ;
$Q(\theta, \theta')$: quantité intermédiaire de l'EM ;
$L(\theta)$: Vraisemblance des données incomplètes sous le paramètre θ ;
$g_c(x, y, \theta)$: vraisemblance des données complètes ;
∇_θ : opérateur gradient au point θ ;

Particules :
$\pi_{m:n|n}$: distribution conditionnelle de $X_{m:n}$ sachant $Y_{0:n}$;
π_k : distribution de filtrage à l'instant k ;
ρ_k : distribution prédictive à l'instant k ;
$\hat{\pi}_k$: approximation particulaire de π_k ;
K : noyau de transition ;
Q : noyau instrumental dans l'échantillonnage d'importance ;
$q(\cdot|\cdot)$: densité instrumental dans l'échantillonnage d'importance ;

Autres notations :
$a \propto b \Leftrightarrow a = b$ à une constante de normalisation près ;
$\| \cdot \|$: norme euclidienne ;
$u.v$: produit scalaire de u par v ;
$\| \cdot \|_q$: norme L_q ;
$\xrightarrow{p.s}$: Convergence presque sure ;
$\xrightarrow{\mathcal{D}}$: Convergence en loi ;

CHAPITRE 1

Algorithme Espérance Maximisation

Sommaire

1.1 Introduction 8

1.2 Formulation et propriété 9

1.2.1 Notations, définitions 9

1.2.2 Formalisme 9

1.3 Applications à la famille exponentielle 15

1.3.1 Distribution multinomiale 16

1.3.2 Distribution exponentielle 18

1.4 Éléments de convergence 19

1.4.1 Convergence de la suite $\left\{L(\theta^{(k)}), k \geq 0\right\}$ 19

1.4.1.1 Notations et Définitions 20

1.4.2 Convergence de la suite $\left\{\theta^{(k)}, k \geq 0\right\}$ 22

1.4.3 Vitesse de convergence 23

1.4.4 Critères d'arrêt 24

1.5 Quelques extensions 24

1.5.1 Monte Carlo EM 24

1.5.2 α-EM 25

1.6 Conclusion 27

1.1 Introduction

Issu des travaux de Dempster et al. [1977], l'algorithme Espérance-Maximisation (EM) est une méthode itérative d'optimisation permettant de trouver le Maximum de vraisemblance ou le Maximum *a posteriori* dans les modèles probabilistes et statistiques faisant intervenir des données manquantes, latentes ou partiellement observées. La simplicité et la modularité de l'algorithme EM ont fini de le populariser à travers diverses problématiques d'inférence statistique où la vraisemblance du modèle en étude est difficile voire impossible à mettre en œuvre par les outils habituels de maximisation. L'algorithme (EM) est d'usage dans de nombreux domaines divers et variés tels que l'imagerie médicale, le traitement du signal, l'apprentissage statistique, etc.

L'idée sous-jacente étant de faciliter la recherche d'optimum de la fonction objectif en 'complétant' celle-ci des variables inobservables. En d'autre terme, on augmente l'espace des observations afin que celui-ci incorpore ces dites données manquantes tout en sachant que l'inférence ne peut se faire que sur les données effectivement observées. En pratique, deux phases s'enchaînent. La première phase, dite E pour Espérance, dans laquelle on calcule l'espérance d'une certaine fonction appelée *quantité intermédiaire* de l'EM. La deuxième, dite M pour Maximisation, consiste en la maximisation de cette espérance pour obtenir un nouveau jeu de paramètres à injecter lors d'un nouvel appel de la phase E. On montre qu'à l'issu d'une itération donnée de l'algorithme EM, la fonction de vraisemblance est améliorée dans le sens où, celle-ci, calculée sous le jeu de paramètres courant, est supérieure ou égale à celle calculée à l'itération immédiatement avant. Ces deux étapes E et M sont réitérées jusqu'à ce que l'on ait jugé de la proximité du jeu de paramètres rendant optimal la fonction objectif.

Dans ce chapitre, nous ne pourrons guère être exhaustif dans le traitement de l'algorithme EM. Cependant, nous allons mettre en relief quelques propriétés remarquables qui permettent d'appréhender cet algorithme ainsi que quelques unes de ses nombreuses variantes. Notamment, nous mettons l'accent sur son usage dans le cas des chaînes de Markov cachées (CMC) utile dans la suite. Ainsi, nous avons adopté le plan suivant pour ce chapitre : la première section est dédiée au formalisme de l'algorithme EM, la deuxième section met l'accent sur les propriétés de convergence, la troisième porte sur le cas de la famille exponentielle courbe. Nous ponctuons cet exposé par quelques exemples d'application et concluons par des extensions de l'algorithme EM.

1.2 Formulation et propriété

1.2.1 Notations, définitions

Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, μ une mesure $\sigma-$finie sur $(\mathcal{Y}^n, \mathbf{B}(\mathcal{Y}^n))$ espace mesurable. On se donne un modèle paramétrique $(\mathcal{Y}^n, \mathbf{B}(\mathcal{Y}^n), \mathbb{P}_\theta, \theta \in \Theta)$ où $\Theta \subset \mathbb{R}^d$ est l'espace des paramètres. On suppose qu'il existe une fonction positive $f(\cdot, \theta)$ vérifiant

$$\frac{d\mathbb{P}_\theta}{d\mu}(y) := f(y, \theta) \tag{1.1}$$

appelée dérivée de Randon-Nykodim de la mesure $\mathbb{P}_\theta$ par rapport à la mesure $\sigma-$finie μ. Nous noterons par $(Y_1, Y_2, \ldots, Y_n)$ un $n-$échantillon c'est-à-dire un n-uplet i.i.d de densité $Y_i \sim f(y_i, \theta), i = 1, 2, \ldots, n$ notée simplement $L(y_i; \theta)$. La densité jointe d'un $n-$échantillon $(Y_1, Y_2, \ldots, Y_n)$ appelée vraisemblance de l'échantillon est notée par :

$$L(y_1, y_2, \ldots, y_n; \theta) = \prod_{i=1}^{n} L(y_i; \theta). \tag{1.2}$$

Définition 1.2.1. *On appelle estimateur du maximum de vraisemblance (EMV) noté $\hat{\theta}_{MV}$ la valeur du paramètre θ rendant maximale la vraisemblance, c'est-à-dire :*

$$\hat{\theta}_{MV} := \arg\max_{\theta \in \Theta} L(y_1, y_2, \ldots, y_n; \theta). \tag{1.3}$$

Pour des raisons de simplicité, on préfère maximiser la log-vraisemblance notée simplement $l(\theta) := \log L(y_1, y_2, \ldots, y_n; \theta)$ au lieu de la vraisemblance, notée quant à elle $L(\theta)$. Il arrive très souvent que pour une raison donnée, le modèle initial dépende de variables partiellement observées ou inobservables rendant difficile voire impossible la maximisation de $l(\theta)$. Nous nous plaçons dans ce cadre et notons par $\mathcal{X}$ le sous-espace des variables inobservables. Dans toute la suite, nous supposons que $L(\theta) > 0$ et que toutes les variables qui interviennent sont continues. Nous travaillons avec la mesure de Lebesgue sur $\mathbb{R}^p, p \in \mathbb{N}^*$ comme mesure de référence. Le cas des variables discrètes étant obtenu en prenant une mesure de comptage comme mesure de référence.

1.2.2 Formalisme

On se donne un espace d'échantillonnage $\mathcal{Z}$ dit *espace des données complètes* sur lequel évolue des observations ayant une composante observable et une composante inobservable. On note par $\mathcal{Y}$ le sous-espace de variables observables dit *espace des données manquantes*. On suppose qu'il existe une surjection $s : \mathcal{Z} \to \mathcal{Y}$ et l'on note par

$$\mathcal{X}(y) := \{z \in \mathcal{Z} : s(z) = y, y \in \mathcal{Y}\}$$

la restriction de $\mathcal{Z}$ interagissant avec $\mathcal{Y}$. Par abus de notation, on utilise $\mathcal{X}$ à la fois pour référer cette restriction et le sous-espace des variables inobservables. On désigne par $Y = (Y_1, Y_2, \ldots, Y_n)$ un $n-$échantillon généré par un modèle statistique de fonction de vraisemblance $g(Y, \theta)$ et dépendant de variables inobservables notées aussi $X = (X_1, X_2, \ldots, X_q)$. Considérons la vraisemblance des données complètes (X, Y) notée $g_c(X, Y; \theta)$. L'EM se propose de maximiser la fonction de vraisemblance

$$L(\theta) = \int_{\mathcal{X}} g_c(X, Y; \theta) dX \tag{1.4}$$

qui, typiquement est une intégrale multidimensionnelle difficile voire impossible à évaluer analytiquement. Comme le vecteur de variables X est inconnu, il convient de trouver un estimateur à la nouvelle fonction objectif $\log g_c(X, Y; \cdot)$ sur les différentes réalisations de X à la lumière des observations Y. Le plus naturel étant l'espérance mathématique de cette fonction et dont il convient de maximiser dans un deuxième temps. Ainsi, définit-on la famille de fonctions auxiliaires $\{Q(\cdot, \theta'), \theta' \in \Theta\}$, point focal de l'algorithme EM.

Définition 1.2.2. *On appelle quantité intermédiaire ou Q-fonction de l'algorithme EM la fonction définie par :*

$$Q(\theta, \theta') := \mathbb{E}\left[\log g_c(X, Y; \theta) \middle| Y; \theta'\right] \tag{1.5}$$

où l'espérance est évaluée sous la densité conditionnelle de X sachant Y, à θ' fixé.

Fondamentalement, l'EM se décompose en deux étapes :

1. **Expectation** calculer $Q(\theta, \theta^{(k)}) = \mathbb{E}\left[\log g_c(X, Y; \theta) \middle| Y; \theta^{(k)}\right]$
2. **Maximization** $\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k)})$

Ainsi, on construit la suite $\{\theta^{(k)}\}_{k \geq 1}$ avec la donnée de $\theta^{(0)}$ telle que la vraisemblance incomplète est améliorée à chaque itération.

Proposition 1.2.3. *Monotonie de l'EM (Dempster et al. [1977])*
Pour tout $(\theta^{(k+1)}, \theta^{(k)}) \in \Theta \times \Theta$,

$$Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)}) \Rightarrow L(\theta^{(k+1)}) \geq L(\theta^{(k)}). \tag{1.6}$$

Cette propriété de monotonie résume à elle seule un premier attrait que peut susciter l'EM. Remplacer une maximisation quasi-impossible par une suite d'estimateurs de paramètres du modèle faisant croître la vraisemblance de ce dernier jusqu'à une précision donnée *a priori*.

Démonstration. Notons par $p(x|y,\theta)$ la densité conditionnelle de $X|Y,\theta$. La règle de de Bayes fournit l'égalité suivante :

$$p(x|y,\theta) = \frac{g_c(x,y,\theta)}{g(y,\theta)}. \tag{1.7}$$

D'où :

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \log g(y,\theta) \\ &= \log g_c(x,y,\theta) - \log p(x|y,\theta). \end{aligned} \tag{1.8}$$

En prenant l'espérance conditionnelle de part et d'autre de l' égalité (1.8) par rapport à la distribution conditionnelle de $X|Y,\theta^{(k)}$ on obtient :

$$\begin{aligned} l(\theta) &= \mathbb{E}\left[\log g_c(X,Y,\theta)\Big|Y,\theta^{(k)}\right] - \mathbb{E}\left[\log p(X|Y,\theta)\Big|Y,\theta^{(k)}\right] \\ &= Q(\theta,\theta^{(k)}) - \mathcal{H}(\theta,\theta^{(k)}), \end{aligned} \tag{1.9}$$

où $\mathcal{H}(\theta,\theta^{(k)}) := \mathbb{E}\left[\log p(X|Y,\theta)\Big|Y,\theta^{(k)}\right]$. De plus, à partir de l'égalité (1.9) on obtient :

$$\begin{aligned} l(\theta^{(k+1)}) - l(\theta^{(k)}) = Q(\theta^{(k+1)},\theta^{(k)}) - Q(\theta^{(k)},\theta^{(k)}) \\ - \{\mathcal{H}(\theta^{(k+1)},\theta^{(k)}) - \mathcal{H}(\theta^{(k)},\theta^{(k)})\}. \end{aligned} \tag{1.10}$$

Il suffit alors de montrer que $\mathcal{H}(\theta^{(k+1)},\theta^{(k)}) - \mathcal{H}(\theta^{(k)},\theta^{(k)}) \leq 0$ pour aboutir à la propriété voulue. Ainsi, pour tout $\theta \in \Theta$,

$$\begin{aligned} \mathcal{H}(\theta,\theta^{(k)}) - \mathcal{H}(\theta^{(k)},\theta^{(k)}) &= \mathbb{E}\left[\log \frac{p(X|Y,\theta)}{p(X|Y,\theta^{(k)})}\Big|Y,\theta^{(k)}\right] \\ &\leq \log \mathbb{E}\left[\frac{p(X|Y,\theta)}{p(X|Y,\theta^{(k)})}\Big|Y,\theta^{(k)}\right] \quad \text{(inégalité de Jensen)} \\ &= \log \int p(x|y,\theta)dx \\ &= 0. \end{aligned} \tag{1.11}$$

Ce qui donne le résultat. □

Un premier aperçu de l'EM est donné au tableau suivant.

Algorithme 1 Algorithme Espérance-Maximisation

1: Choisir un paramètre initial $\theta^{(0)}$
2: **Pour** $k = 0, 1, 2, \ldots$ **faire**
— **Étape E :** Évaluer $Q(\theta, \theta^{(k)}) = \mathbb{E}\left[\log g_c(X, Y; \theta) \middle| Y; \theta^{(k)}\right]$
— **Étape M :** Trouver $\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k)})$
3: **FinPour.**

Par souci de parcimonie, nous faisons quelques hypothèses pour asseoir les fondements de ce qui précède ainsi que pour étayer la suite de l'exposé.

Hypothèses 1.2.4.

1. *L'espace des paramètres* Θ *est un ouvert de* $\mathbb{R}^d$ *avec* $d \in \mathbb{N}^*$ *;*
2. *Pour tout* $\theta \in \Theta$, $0 < L(\theta) < +\infty$ *;*
3. $\theta \mapsto L(\theta)$ *est continûment différentiable sur* Θ *;*
4. *Pour tout* $(\theta, \theta') \in \Theta \times \Theta$,
$$\int_{\mathcal{X}} \left|\nabla_\theta \log p(x|y, \theta)\right| p(x|y, \theta') dx < +\infty \tag{1.12}$$
avec ∇_θ *est l'opérateur gradient appliqué au point* $\theta = \theta'$
5. *Pour tout* $\theta' \in \Theta$, $\theta \mapsto \mathcal{H}(\theta, \theta')$ *est continûment différentiable sur* Θ.

Un deuxième attrait suscité par l'EM est que ce dernier permet de calculer indirectement le gradient de la log-vraisemblance en tout point $\theta \in \Theta$.

Proposition 1.2.5. *Pour tout* $\theta \in \Theta$, $\theta \mapsto Q(\theta, \theta^{(k)})$ *est continûment différentiable et on a :*
$$\nabla_\theta Q(\theta, \theta^{(k)})\big|_{\theta=\theta^{(k)}} = \nabla_\theta l(\theta)|_{\theta=\theta^{(k)}}\,, \tag{1.13}$$
avec $\theta^{(k)}$ *fixé.*

Démonstration. De l'égalité (1.9) et en combinant les points 3. et 5. de Hypothèses1.2.4 la fonction $\theta \mapsto Q(\theta, \theta^{(k)})$ est différentiable comme différence de 2 fonctions différentiables et on a :
$$\nabla_\theta l(\theta)|_{\theta=\theta^{(k)}} = \nabla_\theta Q(\theta, \theta^{(k)})\big|_{\theta=\theta^{(k)}} - \nabla_\theta \mathcal{H}(\theta, \theta^{(k)})\big|_{\theta=\theta^{(k)}}\,. \tag{1.14}$$
Il suffit alors que le second membre de droite de l'égalité (1.14) soit nul pour obtenir l'égalité voulue. En utilisant (1.11), la fonction $\theta \mapsto \mathcal{H}(\theta, \theta^{(k)})$ est maximale en $\theta = \theta^{(k)}$. D'où
$$\nabla_\theta \mathcal{H}(\theta, \theta^{(k)})\big|_{\theta=\theta^{(k)}} = 0.$$
□

Cette deuxième propriété caractérise les points de stabilité de l'algorithme EM comme étant des points stationnaires, c'est-à-dire des points $\theta^{(*)}$ tels que

$$\nabla_\theta l(\theta)|_{\theta=\theta^{(*)}} = 0.$$

Notons que certaines difficultés supplémentaires peuvent subsister dans la phase E et/ou dans la phase M. Les difficultés spécifiques à la phase E dans les chaînes de Markov cachées seront traitées dans le chapitre suivant. Au niveau de la phase M, il peut s'avérer difficile de trouver le paramètre $\theta^{(k+1)}$ qui réalise le maximum de la fonction objectif $Q(\theta, \theta^{(k)})$. C'est la raison pour laquelle Wu [1983] a proposé de choisir $\theta^{(k+1)}$ de sorte que la relation suivante soit satisfaite :

$$Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)}). \tag{1.15}$$

Ce qui donne lieu à l'algorithme EM généralisé (GEM en anglais). En effet, (1.15) est une condition suffisante pour satisfaire à la monotonie de l'EM (proposition 1.2.3). En d'autre terme, la vraisemblance ne décroit pas après une itération de l'EM généralisé. Ainsi, en lieu et place d'un maximum local ou global, on utilise uniquement ce critère améliorant la fonction objectif. Le résumé de cet algorithme est donné au tableau suivant. En guise d'illustration, nous donnons un exemple traitant de mélange

Algorithme 2 Algorithme Espérance-Maximisation Généralisé

1: Choisir un paramètre initial $\theta^{(0)}$
2: **Pour** $k = 0, 1, 2, \ldots$ **faire**
   — **Étape E :** Évaluer $Q(\theta, \theta^{(k)}) = \mathbb{E}\left[\log g_c(X, Y; \theta) \middle| Y; \theta^{(k)}\right]$
   — **Étape M :** Trouver $\theta^{(k+1)}$ tel que $Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)})$
3: **FinPour.**

de distributions qui est fréquent en classification.

Exemple 1.2.6. *Mélange gaussien*
On se donne un $N-$échantillon $Y = (Y_1, Y_2, \ldots, Y_N)$ tel que la densité de l'observation Y_i est donnée par la combinaison convexe

$$f_{Y_i}(y_i, \theta) = \sum_{r=1}^{K} \lambda_r f_r(y_i, \theta_r) \tag{1.16}$$

avec $\sum_{r=1}^{K} \lambda_r = 1$ et $\lambda_r \geq 0$ pour tout $(i, r) \in \{1, 2, \ldots, N\} \times \{1, 2, \ldots, K\}$. Prenant $f_r(\cdot, \theta_r) = \mathcal{N}(\cdot|\mu_r, \Sigma_r)$, on obtient un mélange de K lois gaussiennes. Du point de vue de la classification, cette densité peut être comprise en supposant l'existence de K classes ou groupes distincts numérotés $r = 1, 2, \ldots, K$ et dont chacun est caractérisé par la densité marginale $f_r(\cdot, \theta_r)$. On note par X_i la variable aléatoire égale à r si Y_i est issue

du $r^{ième}$ groupe pour tout $(i, r) \in \{1, 2, \ldots, N\} \times \{1, 2, \ldots, K\}$. L'observation Y_i peut alors s'exprimer à la lumière de cette segmentation par la densité conditionnelle

$$f_{Y_i|X_i}(y_i, \theta) = \sum_{r=1}^{K} f_r(y_i, \theta_r) 1_{\{X_i=r\}}. \tag{1.17}$$

En d'autre terme, la réalisation de la variable Y_i est conditionnée par celle de X_i. C'est à dire que l'on commence par générer une réalisation $r \in \{1, 2, \ldots, K\}$ de X_i, avec $\mathbb{P}(X_i = r) = \lambda_r$, puis Y_i est tirée suivant $\mathcal{N}(\cdot|\mu_r, \Sigma_r)$ conditionnellement à X_i. En application de la règle de Bayes, la densité jointe du couple (X_i, Y_i) est alors donnée par :

$$\begin{aligned} f_{(X_i,Y_i)}(x_i, y_i; \theta) &= f_{Y_i|X_i,\theta}(y_i, \theta) f_{X_i,\theta}(x_i, \theta) \\ &= \sum_{r=1}^{K} f_r(y_i, \theta_r) 1_{\{X_i=r\}} \sum_{s=1}^{K} \lambda_s 1_{\{X_i=s\}} \\ &= \sum_{r=1}^{K} \lambda_r f_r(y_i, \theta_r) 1_{\{X_i=r\}} \end{aligned} \tag{1.18}$$

Il est alors facile d'écrire la vraisemblance des données complètes en tenant compte du vecteur des variables inobservables $X = (X_1, X_2, \ldots, X_N)$:

$$\begin{aligned} &g_c(X, Y; \theta) = \\ &\prod_{i=1}^{N} \left[\sum_{r=1}^{K} 1_{\{X_i=r\}} \mathbb{P}(X_i = r) \frac{1}{\sqrt{2\pi}\det(\Sigma_r)^{1/2}} \exp -\frac{1}{2}(y_r - \mu_r)^T \Sigma_r^{-1} (y_r - \mu_r) \right] \end{aligned} \tag{1.19}$$

et en tirer la $log-$vraisemblance complète :

$$\begin{aligned} &\log g_c(X, Y; \theta) = \\ &\sum_{i=1}^{N} \log \left[\sum_{r=1}^{K} 1_{\{X_i=r\}} \mathbb{P}(X_i = r) \frac{1}{\sqrt{2\pi}\det(\Sigma_r)^{1/2}} \exp -\frac{1}{2}(y_r - \mu_r)^T \Sigma_r^{-1} (y_r - \mu_r) \right] \end{aligned} \tag{1.20}$$

A titre illustratif, prenons $N = 1000$ $K = 2$, $\lambda \sim \mathcal{B}er(2/5)$, $\sigma_1 = 8$, $\sigma_2 = 7$, $\mu_1 = 124$ et $\mu_2 = 157$. Les valeurs d'initialisation sont $\mu_1^{(0)} = 110$, $\sigma_1^{(0)} = 5$, $\mu_2^{(0)} = 170$, $\sigma_2^{(0)} = 5$, $\lambda_1^{(0)} = 0.2$ et $\lambda_2^{(0)} = 0.8$. A l'issu de 40 itérations de l'EM, les paramètres estimés sont consignés dans les graphiques ci-après. L'on constate la convergence effective de l'EM vers les valeurs escomptées de manière satisfaisante.

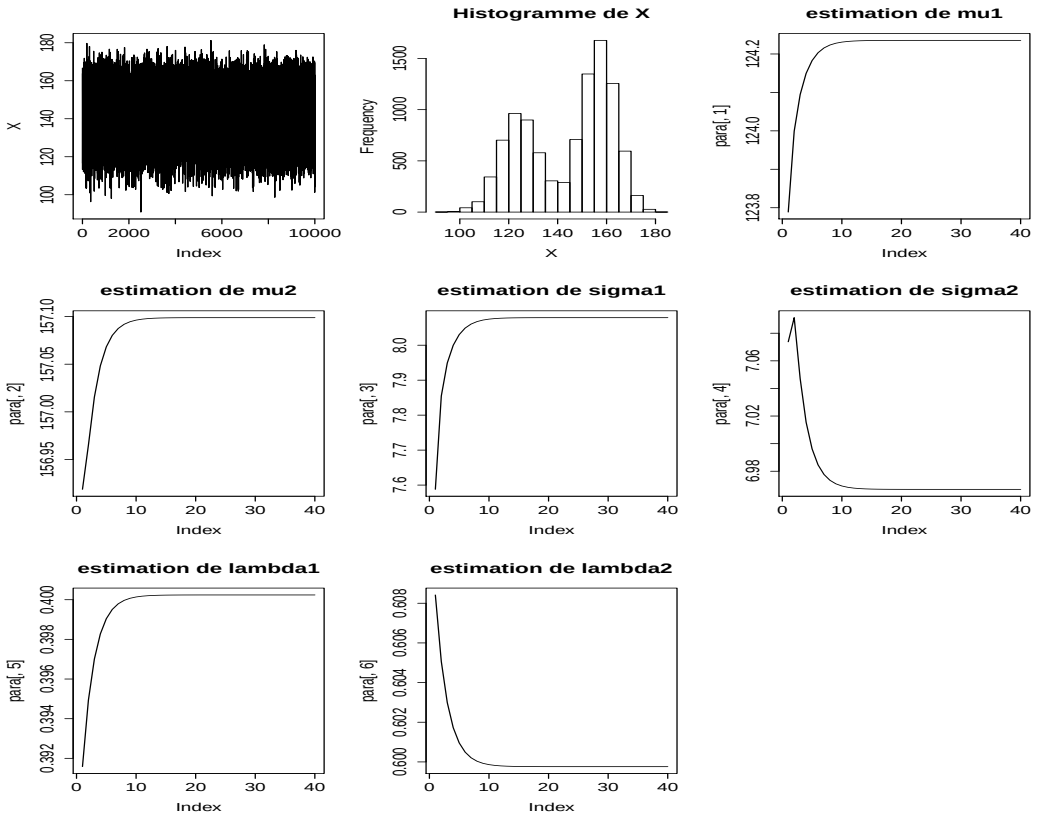


FIGURE 1.1 – Mélange gaussien et itérations de l'EM.

1.3 Applications à la famille exponentielle

Il existe beaucoup d'application de l'EM en médecine, notamment en analyse de survie. On cherche par exemple le temps de survie de patients après avoir reçu un certain traitement. Dans ce cas de figure une famille de distribution est particulièrement utilisée. Il s'agit de la famille exponentielle.

Définition 1.3.1. *Soit Θ l'espace des paramètres. On se donne μ une mesure $\sigma-$finie sur $\mathcal{Y}$, a et b des fonctions définies respectivement sur $\mathcal{Y}$ et Θ à valeurs dans $\mathbb{R}^+$, c et d des fonctions définies respectivement sur $\mathcal{Y}$ et Θ à valeurs dans $\mathbb{R}^n$. On dit qu'une distribution appartient à la famille exponentielle de dimension n si sa fonction de densité par rapport à μ s'écrit :*

$$g(y,\theta) = c(y)d(\theta)\exp(b(\theta)\cdot a(y)). \tag{1.21}$$

Lorsqu'on a les inclusions $\mathcal{Y}, \Theta \subset \mathbb{R}^n$ cette densité s'écrit :

$$g(y,\theta) = c(y)d(\theta)\exp(\theta\cdot y). \tag{1.22}$$

On convient alors de la qualifier de famille "naturelle".

Notons que la famille exponentielle compte entre autres les distributions gaussienne, multinomiale, de Poisson, etc. Dans ce cas de figure, en supposant que les données complètes notées $z = (x, y)$ sont issues de cette famille, la quantité intermédiaire s'écrit simplement :

$$\begin{aligned} Q(\theta, \theta^k) &= \mathbb{E}\left[\log g_c(Z;\theta)|Y, \theta^{(k)}\right] \\ &= \mathbb{E}\left[\log c(Y)|Y, \theta^{(k)}\right] + b(\theta)\cdot a^{(k)} + \log d(\theta) \end{aligned}$$

avec $a^{(k)} := \mathbb{E}\left[a(Z)|Y, \theta^{(k)}\right]$ un estimateur de la statistique exhaustive. Ainsi s'achève la partie E de l'EM. Étant donné que les termes $a^{(k)}$ et $\mathbb{E}\left[\log c(Y)|Y, \theta^{(k)}\right]$ sont indépendants de θ, la maximisation se réduit à trouver $\theta^{(k+1)} \in \Theta$ tel que :

$$\theta^{(k+1)} = \max_{\theta\in\Theta}\left[b(\theta)\cdot a^{(k)} + \log d(\theta)\right].$$

Ce qui se résume à itérer sur k les deux étapes suivantes jusqu'à un critère d'arrêt.
Étape E : Calculer $a^{(k)} := \mathbb{E}\left[a(Z)|Y, \theta^{(k)}\right]$
Étape M : Trouver $\theta^{(k+1)}$ tel que

$$\theta^{(k+1)} = \max_{\theta\in\Theta}\left[b(\theta)\cdot a^{(k)} + \log d(\theta)\right].$$

1.3.1 Distribution multinomiale

Cette première illustration de la famille exponentielle a une portée historique. Elle traite de l'identification de gènes.

Exemple 1.3.2. *(Fisher and Balmukand [1928])*
Cet exemple repris dans plusieurs travaux a trait aux modèles génétiques précisément à la fréquence de certains gènes. Un descriptif complet du problème pourra être trouvé entre autres dans Fisher and Balmukand [1928], Dempster et al. [1977]. Le problème se pose en ces termes. Étant donné un vecteur d'observations $Y = (Y_1, Y_2, Y_3, Y_4)^T$ issu de la distribution multinomiale

$$\mathcal{M}\text{ulti}\left(n; \frac{1}{2} + \frac{1}{4}\psi, \frac{1}{4}(1-\psi), \frac{1}{4}(1-\psi), \frac{1}{4}\psi\right) \tag{1.23}$$

avec $0 \leq \psi \leq 1$ inconnu et modélisant la répartition de $n = 197$ animaux dans 4 catégories. L'objectif est alors d'estimer ψ à partir de Y. Une première solution consiste à utiliser l'estimateur du maximum de vraisemblance $\hat{\psi}_{MV}$ en formant la vraisemblance :

$$L(\psi) = \frac{n!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2} + \frac{1}{4}\psi\right)^{y_1}\left(\frac{1}{4}(1-\psi)\right)^{y_2}\left(\frac{1}{4}(1-\psi)\right)^{y_3}\left(\frac{1}{4}\psi\right)^{y_4} \tag{1.24}$$

d'où l'on tire la log-vraisemblance :

$$l(\psi) = C + y_1 \log(2+\psi) + (y_2+y_3)\log(1-\psi) + y_4 \log(\psi) \tag{1.25}$$

avec C *constante indépendante de* ψ*. En dérivant* l *par rapport à* ψ *on a :*

$$\frac{\partial l(\psi)}{\partial \psi} = 0 \Leftrightarrow \frac{y_1}{2+\psi} - \frac{y_2+y_3}{1-\psi} + \frac{y_4}{\psi} = 0. \tag{1.26}$$

Avec la donnée de $(y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$ *on a une estimation de* ψ *:*

$$\hat{\psi}_{MV} = \frac{1+\sqrt{53809}}{34} \approx 0.626821.$$

Supposons maintenant que la première catégorie soit composée de 2 sous-catégories notées y_{11} *et* y_{12} *avec les probabilités respectives* $\frac{1}{2}$ *et* $\frac{1}{4}\psi$*. Dans cette nouvelle configuration seule la somme* $y_1 = y_{11} + y_{12}$ *est accessible,* y_{11}*(resp.* y_{12}*) demeurant inconnue. Afin d'intégrer cette nouvelle donne, on considère les données complètes symbolisées par le vecteur* $x := (y_{11}, y_{12}, y_2, y_3, y_4)$ *avec* (y_{11}, y_{12}) *la partie inobservée de* x*. Partant de la vraisemblance complète donnée par :*

$$g_c(x,\psi) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!}\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{1}{4}\psi\right)^{y_{12}}\left(\frac{1}{4}(1-\psi)\right)^{y_2+y_3}\left(\frac{1}{4}\psi\right)^{y_4} \tag{1.27}$$

on déduit la log-vraisemblance complète :

$$\log g_c(x,\psi) = C' + (y_{12}+y_4)\log(\psi) + (y_2+y_3)\log(1-\psi) \tag{1.28}$$

d'où l'on tire la quantité intermédiaire :

$$\begin{aligned} Q(\psi,\psi') &= \mathbb{E}\left[\log g_c(X,\psi)|Y,\psi'\right] \\ &= C' + (y_4 + \mathbb{E}\left[Y_{12}|Y,\psi'\right])\log(\psi) + (y_2+y_3)\log(1-\psi) \end{aligned} \tag{1.29}$$

avec C' *constante indépendante de* ψ*. Comme* $Y_{12}|Y_1,\psi'$ *suit une loi binomiale* $\mathcal{B}\text{in}(y_1, \frac{1/4\psi'}{1/2+1/4\psi'})$*, il s'en suit que la phase E de l'EM est calculable explicitement. A l'itération* $(k+1)$*, la quantité intermédiaire s'écrit alors :*

$$\begin{aligned} Q(\psi,\psi^{(k)}) &= C^{(k)} + \left(y_4 + \mathbb{E}\left[Y_{12}|Y_{11},\psi^{(k)}\right]\right)\log(\psi) + (y_2+y_3)\log(1-\psi) \\ &= C^{(k)} + \left(y_4 + y_1 \times \frac{\psi^{(k)}}{2+\psi^{(k)}}\right)\log(\psi) + (y_2+y_3)\log(1-\psi) \end{aligned} \tag{1.30}$$

où $C^{(k)}$ *est une constante indépendante de* ψ*.*

La phase M s'écrit en dérivant (1.30) par rapport à ψ. D'où l'équation de mise à jour du paramètre donnée par :

$$\psi^{(k+1)} = \frac{y_4 + y_1 \times \frac{\psi^{(k)}}{2+\psi^{(k)}}}{y_2 + y_3 + y_4 + y_1 \times \frac{\psi^{(k)}}{2+\psi^{(k)}}}. \tag{1.31}$$

1.3.2 Distribution exponentielle

L'exemple suivant est d'ordre pédagogique mais ne nécessite pas un recours systématique à l'EM. Il a trait aux données de survie censurées avec comme fonction de survie $\Xi(X) := \mathbb{P}(X > x)$. Ce qui revient à s'intéresser à la queue de distribution de X. Dans ce qui suit, cette dernière modélise le fait qu'un individu survive au delà de l'instant x.

Exemple 1.3.3. *On se donne un échantillon $Y = (Y_1, Y_2, \ldots, Y_{n_1})$ d'observations (non censurées) et $X = (X_1, X_2, \ldots, X_{n_2})$ un échantillon d'observations censurées à la date T. L'instant de survie est donc minoré par T. On suppose que la densité des observations effectives est donnée par*

$$g(y_i, \theta) = \theta^{-1} \exp(-y_i/\theta), \quad i = 1, 2, \ldots, n_1 \tag{1.32}$$

avec θ la moyenne du temps de survie. Afin de calculer la quantité intermédiaire on suppose également que les données complétées (censurées ou non) sont i.i.d de même densité. On en déduit la densité des données complétées :

$$g_c(x, y; \theta) = \frac{1}{\theta^{n_2}} \exp\left(-\sum_{i=1}^{n_2} x_i/\theta\right) \frac{1}{\theta^{n_1}} \exp\left(-\sum_{i=1}^{n_1} y_i/\theta\right).$$

Ainsi, à l'étape $(k+1)$ la quantité intermédiaire est donnée par :

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= \mathbb{E}\left[\log g_c(X, Y; \theta) | Y, \theta^{(k)}\right] \\ &= \mathbb{E}\left[-n_2 \log\theta - \sum_{i=1}^{n_2} X_i - n_1 \log\theta - \sum_{i=1}^{n_1} Y_i \Big| Y, \theta^{(k)}\right] \\ &= -(n_1 + n_2)\log\theta - \frac{1}{\theta}\sum_{i=1}^{n_2} \mathbb{E}\left[X_i | \theta^{(k)}\right] - \frac{1}{\theta}\sum_{i=1}^{n_1} Y_i \\ &= -(n_1 + n_2)\log\theta - \frac{n_2(T + \theta^{(k)})}{\theta} - \frac{1}{\theta}\sum_{i=1}^{n_1} Y_i. \end{aligned} \tag{1.33}$$

La dernière égalité vient du fait que $\mathbb{E}\left[X_i | \theta^{(k)}\right] = T + \theta^{(k)}$. En effet, il est facile de voir que la densité d'une donnée censurée est donnée par :

$$g(x_i, \theta) = \theta^{-1} \exp((T - x_i)/\theta) \quad , x_i \geq T, \tag{1.34}$$

D'où l'on tire

$$\begin{aligned} \mathbb{E}\left[X_i | \theta^{(k)}\right] &= \int_T^\infty x_i \frac{1}{\theta^{(k)}} \exp((T - x_i)/\theta^{(k)}) dx_i \\ &= \frac{1}{\theta^{(k)}} \exp(T/\theta^{(k)}) \int_T^\infty x_i \exp(-x_i/\theta^{(k)}) dx_i \\ &= T + \theta^{(k)}. \end{aligned}$$

En dérivant $Q(\theta, \theta^{(k)})$ *par rapport à* θ*, on obtient l'équation suivante :*

$$\frac{\partial Q(\theta, \theta^{(k)})}{\partial \theta^{(k)}} = \frac{-(n_1 + n_2)}{\theta} + \frac{n_2(T + \theta^{(k)})}{\theta^2} + \frac{1}{\theta^2}\sum_{i=1}^{n_1} Y_i = 0, \tag{1.35}$$

dont la résolution donne alors

$$\theta^{(k+1)} = \frac{n_1}{n_1 + n_2}\bar{y} + \frac{n_2 T}{n_1 + n_2} + \frac{n_2}{n_1 + n_2}\theta^{(k)}.$$

Ce qui achève l'étape M. Notons au passage qu'un calcul simple permet d'établir le maximum de vraisemblance

$$\hat{\theta}_{ML} = \bar{y} + \frac{n_2}{n_1} T$$

qui est également la limite de la suite précédente.

1.4 Éléments de convergence

Dans cette section, nous abordons quelques critères permettant de jauger la convergence ou la non convergence de l'algorithme EM et sa généralisation. Notons que l'on peut à la fois s'intéresser à la convergence effective de la suite des paramètres $\{\theta^{(k)}, k \geq 0\}$ ainsi que de $\{L(\theta^{(k)}), k \geq 0\}$ vers les points stationnaires θ^* et $L(\theta^*)$ respectivement. Il faut noter que la convergence de la suite de vraisemblance $\{L(\theta^{(k)}), k \geq 0\}$ revêt plus d'importance comparée à celle des $\{\theta^{(k)}, k \geq 0\}$, itérés de l'EM. De plus, la convergence de l'une des deux suites n'entraîne pas forcement celle l'autre. Il faut également souligner que l'on ne peut garantir la convergence effective vers un maximum global de la fonction objectif $L(\cdot)$. Les deux sous-sections suivantes inspirées par Wu [1983] répondent essentiellement à deux questions :

1. A supposer que $\{L(\theta^{(k)}), k \geq 0\}$ converge vers un point L^*, est-ce que ce dernier est un maximum global, local, un point de scelle ou tout autre point et dans quelles conditions ?
2. Dans quelles mesures les itérés $\{\theta^{(k)}, k \geq 0\}$ de l'algorithme convergent-ils vers un point θ^* ?

1.4.1 Convergence de la suite $\{L(\theta^{(k)}), k \geq 0\}$

Comme noté précédemment, la convergence de la suite des vraisemblances $\{L(\theta^{(k)})\}$ revêt une importance capitale puisque c'est d'elle dont découle la convergence ou non de l'EM. Notons d'abord, que lorsque la vraisemblance est bornée supérieurement on

peut montrer que la suite $\left\{L(\theta^{(k)}), k \geq 0\right\}$ converge vers un point stationnaire L^*, c'est-à-dire qu'il existe un point $\theta^* \in \Theta$ vérifiant :

$$L^* := L(\theta^*) = \lim_{k\to\infty} L(\theta^{(k)}) \tag{1.36}$$

et

$$\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = \left.\frac{\partial \log L(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0. \tag{1.37}$$

En général, L^* est soit un maximum local voire global lorsque L est unimodal, soit un minimum, soit un point de scelle. La nature du point L^* est donc tributaire de la forme de la fonction de vraisemblance $L(\cdot)$, du choix de l'initialisation de l'EM. C'est par exemple le cas lorsque L^* est un point de scelle, il suffit alors de changer l'initialisation de $L(\theta^{(0)})$ en $L(\theta'^{(0)})$ pour voir l'EM diverger de L^*. En d'autre terme, une petite perturbation aléatoire de $L(\theta^{(0)})$ suffit pour se départir de l'attraction exercée par L^*. Un autre constat peut être fait lorsque $L(\cdot)$ possède plusieurs points stationnaires. Dans ce cas de figure, la convergence vers l'un ou l'autre des points L^* dépend fortement de l'initialisation $L(\theta^{(0)})$. Le seul cas où la convergence est indépendante de celle-ci est obtenue lorsque $L(\cdot)$ est unimodale dans Ω moyennant une condition de différentiabilité de $L(\cdot)$.

1.4.1.1 Notations et Définitions

Commençons par définir l'ensemble des valeurs maximisant la quantité intermédiaire donné par :

$$\mathcal{M}\left(\theta^{(k)}\right) := \left\{\theta \in \Theta : Q(\theta, \theta^{(k)}) \geq Q(\theta', \theta^{(k)}), \forall \theta' \in \Theta\right\} \tag{1.38}$$

pour le GEM. Il faut noter qu'en général l'ensemble $\mathcal{M}$ se réduit à un singleton. Cependant, cette définition autorise l'éventualité de disposer de plusieurs maxima pour la fonction objectif.

Définition 1.4.1. *Soit T une application de Θ à valeurs dans les sous-ensembles de Θ. On dit que T est fermée sur $S \subseteq \Theta$ si pour toutes suites convergentes $\left\{\theta^{(i)}\right\}_{i\geq 0}$ et $\left\{\theta^{(i)'}\right\}_{i\geq 0}$ telles que :*

1. $\theta^{(i)} \to \theta \in S$,
2. $\theta^{(i)'} \to \theta'$, *avec* $\theta^{(i)'} \in T(\theta^{(i)})$ *pour tout* i,

on a : $\theta' \in T(\theta)$.

Comme on peut s'en douter, les résultats de convergence de l'EM utilisent essentiellement des résultats d'Analyse. Nous commençons par donner un premier résultat de Zangwill [1969] sur les convergences d'algorithmes itératifs avant d'en donner l'adaptation de Wu [1983] à l'algorithme EM. Notons par ailleurs que les algorithmes itératifs considérés ici sont ceux qualifiés d'algorithmes "de montée" [1] construits par la donnée

1. *ascent algorithms* en anglais

de $\theta^{(0)} \in \Theta$ et $\theta^{(k+1)} \in T(\theta^{(k)})$ où $T : \Theta \to 2^{\Theta}$ est une application donnée.

Théorème 1.4.2. *(Zangwill [1969])*
Considérons l'algorithme itératif suivant :

$$\begin{cases} \theta^{(0)} \in \Theta, \\ \theta^{(k+1)} \in T(\theta^{(k)}), k \in \mathbb{N} \end{cases} \tag{1.39}$$

sous les hypothèses :

1. *pour tout $k \in \mathbb{N}, \theta^{(k)} \in \Delta$ compact ;*
2. *T est fermée sur $\Theta \setminus \Delta$, complémentaire de Δ dans Θ ;*
3. *il existe une fonction réelle α définie sur Θ et continue telle que*
   — $\theta \notin \Delta \Rightarrow \alpha(\theta) < \alpha(\eta)$, $\forall \eta \in \mathcal{M}(\theta)$
   — $\theta \in \Delta \Rightarrow \alpha(\theta) \leq \alpha(\eta)$, $\forall \eta \in \mathcal{M}(\theta)$, *avec Δ sous ensemble de Θ;*

Alors, toutes les limites de $\{\theta^{(k)}\}$ sont dans Δ et $\alpha(\theta^{(k)})$ converge de manière monotone vers $\alpha(\theta)$, $\theta \in \Delta$.

Démonstration. Soit θ une limite de $\{\theta^{(k)}\}$ i.e

$$\theta^{(k)} \to \theta. \tag{1.40}$$

Alors il existe une sous-suite notée $\{\eta^{(k)}\}$ de $\{\theta^{(k)}\}$ définie par :

$$\eta^{(k)} := \theta^{(a(k)+1)} \in T(\theta^{(a(k))}) \tag{1.41}$$

telle que :

$$\eta^{(k)} \to \theta,$$

avec $a : \mathbb{N} \to \mathbb{N}^*$ strictement croissante. De cette sous-suite, on peut extraire une sous-suite $\{\beta^{(k)}\}$ définie par :

$$\beta^{(k)} := \eta^{(b(k)+1)} \in T(\eta^{(b(k))}) \subset T(\eta^{(k)}) \tag{1.42}$$

$b : \mathbb{N} \to \mathbb{N}^*$ strictement croissante telle que :

$$\beta^{(k)} \to \beta. \tag{1.43}$$

Ainsi,

$$\begin{cases} \eta^{(k)} \to \theta, \\ \beta^{(k)} \to \beta, \quad \beta^{(k)} \in T(\eta^{(k)}). \end{cases} \tag{1.44}$$

Supposons que $\theta \notin \Delta$. Étant donnée que T est fermée en θ alors $\beta \in T(\theta)$. D'où

$$\alpha(\beta) > \alpha(\theta).$$

De plus,

$$\alpha(\beta^{(k)}) \leq \alpha(\eta^{(k)}) \leq \alpha(\beta^{(k+1)}). \tag{1.45}$$

Par passage à la limite :

$$\alpha(\eta^{(k)}) \to \alpha(\theta). \tag{1.46}$$

Ainsi,

$$\alpha(\theta) = \alpha(\beta) \ \text{(Absurde)}. \tag{1.47}$$

Par suite, $\theta \in \Delta$. □

En application du théorème de Théorème 1.4.2, Wu fournit l'adaptation suivante à l'algorithme EM généralisé.

Théorème 1.4.3. *(Wu [1983])*
Considérons le GEM généré par la donnée de $\left\{\theta^{(k)}\right\}_{k\geq 0}$ *avec* $\theta^{(k)} \in \mathcal{M}(\theta^{(k+1)})$. *Supposons que :*

1. $\mathcal{M}(\theta^{(k+1)})$ *est fermée sur le complémentaire de* S *ensemble des points stationnaires de* $L(\cdot)$ *dans l'intérieur de* Θ *;*
2. $L(\theta^{(k+1)}) > L(\theta^{(k)})$ *pour tout* $L(\theta^{(k)}) \notin S$ *;*

Alors toutes les limites de $\left\{\theta^{(k)}\right\}_{k\geq 0}$ *sont des points points stationnaires et* $\left\{L(\theta^{(k)})\right\}_{k\geq 0}$ *converge de manière monotone vers* $L^* = L(\theta^*)$, $\theta^* \in S$ *quelconque.*

Notons qu'un affaiblissement de la condition de fermeture du Théorème 1.4.2 peut être obtenue pour l'algorithme EM. En effet, une condition suffisante pour satisfaire la fermeture de l'ensemble $\mathcal{M}(\cdot)$ est la continuité aux points θ et θ' de $Q(\theta, \theta')$. D'où le second résultat de Wu vérifiable pour une large famille de distributions dont la famille exponentielle.

Théorème 1.4.4. *(Wu [1983])*
Supposons que $Q(\theta, \theta')$ *soit continu en* (θ, θ'). *Alors toutes les limites de la suite* $\left\{\theta^{(k)}\right\}_{k\geq 0}$ *généré par l'EM sont des points stationnaires de* $L(\cdot)$ *et* $\left\{L(\theta^{(k)})\right\}_{k\geq 0}$ *converge de manière monotone vers* $L^* = L(\theta^{(*)})$, *avec* $\theta^{(*)}$ *point stationnaire quelconque.*

Ce résultat découle d'une application directe du précédent résultat.

1.4.2 Convergence de la suite $\left\{\theta^{(k)}, k \geq 0\right\}$

Dans la même lancée, il est possible d'établir la convergence des itérés $\left\{\theta^{(k)}, k \geq 0\right\}$ de l'EM sous des conditions un peu plus drastique. Définissons $S(\theta^*) := \{\theta \in \Theta, L(\theta) = \theta^*\}$ comme l'ensemble des points stationnaires de $L(\cdot)$ à l'intérieur de Θ tels que $L(\theta) = \theta^*$.

Théorème 1.4.5.
S'il existe un seul point stationnaire pour $L(\cdot)$ *ie,* $S(\theta^*) = \{\theta^*\}$*, alors*

$$\theta^{(k)} \to \theta^*.$$

Démonstration. Il suffit d'appliquer Théorème 1.4.3 pour $S(\theta^*) = \{\theta^*\}$. □

Étant donné le caractère contraignant de l'hypothèse précédente, Wu [1983] remarque qu'une condition nécessaire de cette convergence est donnée par

$$\| \theta^{(k+1)} - \theta^{(k)} \| \to 0, \quad \text{à mesure que} \quad k \to \infty.$$

Ce qui donne l'allègement suivant.

Théorème 1.4.6. *Soit* $\left\{\theta^{(k)}\right\}$ *une instance du GEM telle que :*

$$\| \theta^{(k+1)} - \theta^{(k)} \| \to 0, k \to \infty$$

Alors, toutes les limites de cette suite sont dans un sous-ensemble connexe et compact de $S(L^*)$*. En particulier, si* $S(L^*)$ *est discret alors* $\theta^{(k)} \to \theta^* \in S(L^*)$.

1.4.3 Vitesse de convergence

Afin de discuter de la vitesse de convergence de l'EM il convient d'introduire la notion de taux de convergence de l'EM. Théoriquement, le taux global de convergence de l'EM se définit comme la variation relative limite notée r suivante :

$$r := \lim_{k \to +\infty} \frac{\| \theta^{(k+1)} - \theta^* \|}{\| \theta^{(k)} - \theta^* \|}. \tag{1.48}$$

Étant donné que θ^* est inconnu pour des données réelles, il convient en pratique de remplacer cette mesure par la variation relative

$$r' := \lim_{k \to +\infty} \frac{\| \theta^{(k+1)} - \theta^{(k)} \|}{\| \theta^{(k)} - \theta^{(k-1)} \|} \tag{1.49}$$

qui elle est disponible à partir de la seconde itération de l'EM. Cette mesure est parfois appelée taux de convergence linéaire. On pourra consulter McLachlan and Krishnan [2008] pour d'avantage de développement.

1.4.4 Critères d'arrêt

Il n'existe pas de règles spécifiques d'arrêt pour le GEM. Néanmoins, il existe un certain nombre de critères heuristiques utilisés afin de mettre un terme aux itérations du GEM. Un premier critère peut être choisi en considérant une variation relative des paramètres. Etant donnée la $k^{(\text{ième})}$ itération du GEM, si

$$\frac{\| \theta^{(k)} - \theta^{(k-1)} \|}{\| \theta^{(k)} \| + \delta} < \epsilon \tag{1.50}$$

avec δ et ϵ convenablement choisis alors on peut considérer avoir atteint avec $\theta^{(k+k_0)}$ le M.V avec k_0 itérations supplémentaires [2]. Un autre critère utilisé est la différence de vraisemblances (voir dans Chan and Ledolter [1995])

$$l(\theta^{(k)}) - l(\theta^{(k-1)}) \tag{1.51}$$

où cette différence est prise stochastiquement très petite. Enfin, il existe aussi dans la littérature la différence

$$Q(\theta^{(k)}, \theta^{(k-1)}) - Q(\theta^{(k-1)}, \theta^{(k-1)}) \tag{1.52}$$

à comparer à un seuil fixé. Par ailleurs, il convient de souligner aussi que l'on peut se fixer un nombre *a priori* d'itérations avec l'avantage de ne faire aucun test mais nécessitant un compromis entre le temps de calcul et la précision des estimés.

1.5 Quelques extensions

Dans cette section nous évoquons sans exhaustivité quelques variantes et extensions possibles de l'algorithme EM. Nous avons déjà introduit le GEM qui permet entre autres de contourner les difficultés inhérentes au problème d'optimisation. Certaines de ces extensions interviennent dans la phase E par exemple en aidant à approximer les espérances conditionnelles intervenant dans cette phase. D'autres peuvent aider à minimiser le risque de tomber sur des optimum locaux dans la phase M.

1.5.1 Monte Carlo EM

Il arrive très souvent que la phase E soulève beaucoup de difficultés eu égard à la forme de la vraisemblance complète. Une façon simple de contourner ce problème consiste à remplacer l'espérance théorique par sa contre-partie empirique. A l'instant k, supposons disposer d'un $N-$échantillon $(X_1, X_2, \ldots, X_N)$ tiré de la densité des données

2. Les k_0 itérations en plus assurent entre autres que le critère est réellement atteint sans que cela ne soit dû aux erreurs de Monte Carlo

manquantes $p(\cdot|Y, \theta^{(k-1)})$. Wei and Tanner [1990] préconise de remplacer l'évaluation de $Q(\theta, \theta^{(k-1)})$ par son approximation de Monte Carlo :

$$\hat{Q}_N(\theta, \theta^{(k-1)}) = \frac{1}{N} \sum_{i=1}^{N} \log g_c(X_i, Y, \theta) \tag{1.53}$$

puis de la maximiser dans la partie M. Cette approximation admet de propriétés de convergence notamment sous certaines hypothèses dont celle qui fait que $\{X_i\}$ forme une chaîne de Markov de loi invariante la distribution conditionnelle de X sachant $Y = y$

$$\hat{Q}_N(\theta, \theta^{(k-1)}) \to Q(\theta, \theta^{(k-1)}) \tag{1.54}$$

presque sûrement, à mesure que $N \to +\infty$. Signalons le cas particulier de $N = 1$ appelé SEM (*Stochastic EM*). Ce dernier aborde dans le sens de minimiser le risque de tomber sur des optimum locaux avec une étape supplémentaire S (*Stochastic*) pour l'EM basique (voir Celeux and Diebolt [1992] pour une présentation détaillée). Par ailleurs, notons qu'il existe plusieurs façons de simuler l'échantillon $(X_1, X_2, \ldots, X_N)$, notamment l'échantillonneur de Gibbs (voir Chan and Ledolter [1995]) ou une version séquentielle des méthodes de Monte Carlo auxquelles on va d'avantage s'intéresser.

1.5.2 α-EM

L'$\alpha-$ EM est une généralisation de l'EM de base s'appuyant sur une réinterprétation du logarithme de la vraisemblance complète dont on évalue l'espérance conditionnelle. En effet, il est facile de voir que la $Q-$fonction de l'EM peut être vue comme une divergence de Kulback-Leibler, cas particulier de la $\alpha-$divergence ou divergence d'ordre α qui elle à son tour est aussi un cas particulier de la $f-$ divergence entre deux mesures de probabilité dans notre cas. On pourra se référer à entre autres à Réyni [1960], Havrda and Chavat [1967], Csiszár [1967] ou à Csiszár [1972] pour une classe plus large de mesure divergence. Néanmoins nous donnons quelques exemples pour introduire l'$\alpha-$ EM. [3]

3. Rappelons qu'une divergence D est une fonction définie sur un espace E vérifiant :
   — $\forall s, t \in E, D(s||t) \geq 0$;
   — $D(s||t) = 0$ si et seulement si $s = t$;
   — pour une petite variation ds, le développement de Taylor de D vérifie :

$$D(s + ds||s) \approx \frac{1}{2} \sum_{i,j \in E} g_{i,j} ds_i ds_j$$

   $(g_{i,j}(s))_{i,j \in E}$ une matrice définie positive.

Exemple 1.5.1. *On considère 2 mesures discrètes* $p = \{p_j\}_{j\in J}$ *et* $q = \{p_j\}_{j\in J}$ *sur* E *espace des mesures discrètes avec* $J \subset \mathbb{N}$. *On définit la* $f-$*divergence de Csiszár entre* p *et* q *par :*

$$D_f(p||q) := \sum_{j\in J} p_j f\left(\frac{q_j}{p_j}\right) \tag{1.55}$$

où f *est une fonction convexe 2 fois différentiables telle que* $f(1) = f'(1) = 0$ *et* $f''(1) = 1$ *En prenant une paramétrisation particulière de* f *donnée par :*

$$f_\alpha(x) = \begin{cases} x\log x - (x-1) & si\ \alpha = +1 \\ -\log x + (x-1) & si\ \alpha = -1 \\ \frac{4}{1-\alpha^2}(1 - x^{(1+\alpha)/2}) - \frac{2}{1-\alpha}(x-1) & si\ \alpha \neq \pm 1 \end{cases} \tag{1.56}$$

on obtient la $\alpha-$*divergence* $D^{(\alpha)}$ *associée :*

$$D^{(\alpha)}(p||q) = \begin{cases} K(p||q) & si\ \alpha = +1 \\ K(q||p) & si\ \alpha = -1 \\ \frac{4}{1-\alpha^2}(1 - x^{(1+\alpha)/2}) - \frac{2}{1-\alpha}(x-1) & si\ \alpha \neq \pm 1 \end{cases} \tag{1.57}$$

où $K(\cdot||\cdot)$ *est la divergence de Kulback-leibler définie par :*

$$K(p||q) := \sum_{j\in J} p_j \log\frac{p_j}{q_j}.$$

Notons que pour des mesures continues, on obtient une définition analogue en remplaçant les sommes discrètes par des intégrales. On pourra consulter avec intérêt la note bibliographique de Basseville [2013].

La quantité intermédiaire s'écrit :

$$Q^{(\alpha)}(\theta,\theta) = \frac{2}{1+\alpha}\left\{S^{(\alpha)}(\theta,\theta) - 1\right\} \tag{1.58}$$

avec

$$\begin{aligned} S^{(\alpha)}(\theta,\theta) &= \int_{\mathcal{X}(\mathcal{Y})} p(x|y,\theta')\left(\frac{p(y|x,\theta)}{p(y|x,\theta')}\right)^{\frac{1+\alpha}{2}} dx \\ &= \mathbb{E}\left[\left(\frac{p(Y|X,\theta)}{p(Y|X,\theta')}\right)^{\frac{1+\alpha}{2}} \Big| Y,\theta'\right]. \end{aligned} \tag{1.59}$$

L'$\alpha-$EM consiste donc à réitérer le calculer de $S^{(\alpha)}$ et ainsi que de son optimisation. Il suffit alors de voir qu'en prenant $\alpha = -1$ on retombe sur l'EM classique (voir Matsuyama [2003]).

1.6 Conclusion

Dans ce chapitre, nous avons survolé quelques grandes lignes de l'algorithme EM. Nous avons décrit sa propriété de monotonie comme pouvant être son plus grand atout. Nous l'avons illustré au travers d'exemples-jouets. Nous avons également décrit quelques unes des extensions de l'EM, singulièrement la version de Monte Carlo dont nous aurons à utiliser dans la suite. Nous avons aussi ponctué cette description par quelques résultats de convergence de l'EM adaptés de l'Analyse. Il faut cependant noter que cette présentation est loin d'être exhaustive. Nous n'avons pas parlé entre autres de l'EM bayésien, l'EM variationnel, le *Supplemented* EM (Meng and Rubin [1991]) ou de l'EM Conditionnel qui sont tous des variantes de l'EM basique. Celles-ci se plaçant dans une perspective de lever certaines limitations inhérentes aux problèmes spécifiques rencontrés dans son utilisant et en vue de son amélioration. De plus, nous avons aussi mentionné et non illustré les limitations de l'EM par le biais de cas pathologiques (voir Boyles [1983]). Par conséquent, nous renvoyons à McLachlan and Krishnan [2008] pour un aperçu de l'étendue de l'EM, de quelques unes de ses extensions possibles ainsi que les points non abordés dans cet exposé.

CHAPITRE 2

Méthodes de Monte Carlo Séquentielles

Sommaire

2.1 Introduction 30
2.2 Échantillonnage d'importance 31
2.2.1 Échantillonnage préférentiel 31
2.2.2 Échantillonnage préférentiel auto-normalisé 34
2.2.3 Échantillonnage préférentiel avec rééchantillonnage 38
2.3 MMC & Approximations particulaires 40
2.3.1 Filtrage particulaire 43
2.3.1.1 Échantillonnage d'importance séquentiel 44
2.3.1.2 Filtre de Bootstrap 47
2.3.2 Lissage particulaire 48
2.3.2.1 Lissage Forward-Backward 49
2.3.2.2 Lissage 2-filtres 50
2.3.2.3 Lissage joint 52
2.4 Quelques résultats théoriques 53
2.4.1 Le Cadre 53
2.4.2 Analyse du filtrage 54
2.4.2.1 Erreur quadratique moyenne 57
2.4.2.2 Convergence presque-sure 58
2.4.2.3 TCL 59
2.4.3 Analyse du lissage 61
2.4.3.1 Formulation 61

2.4.3.2 Inégalités de déviation 62
2.4.3.3 Erreur L_q 64
2.4.3.4 TCL 65
2.5 Conclusion 65

2.1 Introduction

Les méthodes de Monte Carlo séquentielles (MCs) constituent un ensemble d'outils très élaborés destinés à la résolution de problématiques d'inférence inhérentes aux modèles probabilistes et statistiques, et ce au moyen de simulations. Ces méthodes peuvent être considérées comme la version séquentielle des méthodes de Monte Calo par Chaînes de Markov qui elles, datent de la deuxième guerre mondiale avec les travaux de von Neumann et Ulam, puis formalisés par Metropolis et Ulam à partir de 1949 dans Metropolis and Ulam [1949]. Les méthodes MCs quant à elles, sont plus récentes et datent de la fin des années 60 avec les travaux de Handschin and Mayne [1969] et Handschin [1970]. Les méthodes MCs furent redécouvertes durant les années 80-90, notamment grâce à l'essor des outils informatiques qui jusque là faisait défaut à l'avancée de celles-ci avec comme article précurseur celui de Gordon et al. [1993].

De manière synthétique, supposons disposer d'un système en évolution dans le temps à travers divers états possibles. Le système est supposé non directement observable. A chaque état, le système émet un signal qui génère une observation bruitée quantifiable. Celle-ci peut prendre la forme d'un rendement instantané lorsque l'on s'intéresse à la volatilité d'une action ou d'un taux de change en bourse, une cordonnée (position, altitude, ...) lorsqu'il s'agit de la localisation ou la poursuite d'un mobile [1] etc. Dès lors, on comprend que les domaines d'application demeurent nombreux et variés : industrie (détection précoce de panne), traitement du signal (reconnaissance vocale, débruitage de signaux sonores), navigation (positionnement GPS), militaire (poursuite de cible en mouvement) etc. Il s'agit alors d'associer un modèle *a priori* de déplacement ou modèle d'état du système aux données recueillies via un modèle de mesure. Fort de cette association, les méthodes MCs peuvent sous certaines hypothèses aider à estimer l'état du système à chaque instant au vu des observations disponibles on parle alors de *filtrage*. De plus, on peut à la lumière des observations disponibles essayer de lisser l'état du système (resp. prévoir l'état du système dans un avenir proche), on parle alors de *lissage* (resp. de *prédiction*). Par ailleurs, il est d'usage courant que les méthodes MCs servent aussi bien dans l'estimation en ligne comme hors ligne de modèles probabilistes ou statistiques là où les approches habituelles sont difficiles voire impossibles à mettre en œuvre.

1. On parle de modèle numérique de terrain, voir Dahia [2005]

Étant donnée l'étendue du domaine, nous nous contenterons d'exposer les rudiments nécessaires à la compréhension des méthodes MCs, de leurs applications et quelques unes de leurs propriétés. Ainsi, nous avons opté pour le plan suivant : tout d'abord nous nous plaçons dans le cadre d'un modèle de Markov caché dominé même si on peut avoir une vue plus générale, puis nous y exposons quelques techniques de filtrage et de lissage non linéaire. Dans une dernière phase nous donnons quelques éléments d'analyse théorique qui fondent ces méthodes.

2.2 Échantillonnage d'importance

Par souci de simplicité, nous ne considérons que des densités de probabilités. Nous commençons par exposer quelques techniques d'approximation d'intégrale via des simulations de variables aléatoires. Nous mettons l'accent sur l'échantillonnage d'importance et l'échantillonnage d'importance séquentiel utilisé en approximation particulaire. Pour un exposé des techniques de génération de variables aléatoires on pourra se reporter avec intérêt à entre autres Devroye [1986], Ripley [1987], Robert and Cassella [2004] ou Millet [2006].

2.2.1 Échantillonnage préférentiel

L'échantillonnage préférentiel (EP) ou échantillonnage d'importance (EI) est une méthode de simulation basée sur des fonctions dites d'*importance* ou poids d'*importance*. Son origine est difficile à dater cependant elle semble avoir émergé en physique nucléaire avec les travaux de Goertzel [1949], Kahn [1949] et Kahn and Harris [1949]. Pour le formaliser, considérons l'évaluation de l'intégrale suivante :

$$I(f) = \int_{\mathcal{X}} f(x)p(x)dx, \tag{2.1}$$

que l'on suppose impossible à calculer analytiquement et numériquement avec f fonction test quelconque. Supposons de plus que $p(\cdot)$ soit une densité sur $\mathcal{X}$. Une interprétation probabiliste de (2.1) permet alors de voir que :

$$\begin{aligned} I(f) &= \int_{\mathcal{X}} f(x)p(x)dx \\ &= \mathbb{E}_p[f(X)] \\ &= \int_{\mathcal{X}} f(x)\frac{p(x)}{\tilde{p}(x)}\tilde{p}(x)dx \\ &= \int_{\mathcal{X}} f(x)\omega(x)\tilde{p}(x)dx \\ &= \mathbb{E}_{\tilde{p}}[f(X)\omega(X)], \end{aligned} \tag{2.2}$$

où $\omega(x) = \frac{p(x)}{\tilde{p}(x)}$ est la fonction d'importance et $\mathbb{E}_p$ (resp. $\mathbb{E}_{\tilde{p}}$) est l'espérance mathématique évaluée sous la densité $p(\cdot)$ (resp. $\tilde{p}(\cdot)$). Ainsi, en lieu et place d'une espérance sous la densité d'intérêt ou densité cible $p(\cdot)$, l'EP permet d'évaluer (2.1) sous une densité *instrumentale* $\tilde{p}(\cdot)$.

Définition 2.2.1. *On appelle méthode d'EP toute technique basée sur la génération d'un $N-$échantillon $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(N)})$ tiré d'une densité instrumentale $\tilde{p}(\cdot)$ et permettant d'approximer* (2.1) *par :*

$$\hat{I}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi^{(i)}) \omega(\xi^{(i)}). \tag{2.3}$$

Le procédé est résumé dans le tableau suivant. De là, il est naturel de convenir de

Algorithme 3 Échantillonnage d'importance

1: **Choisir** $\tilde{p}$ telle que $supp(\tilde{p}) \supset supp(f.p)$
2: **Pour** $i = 1, 2, \ldots, N$ **faire**
   — **Générer** $\xi^{(i)} \sim \tilde{p}(\cdot)$
   — **Poser** $\omega(\xi^{(i)}) = \frac{p(\xi^{(i)})}{\tilde{p}(\xi^{(i)})}$
3: **FinPour**
4: **Retourner** $\hat{I}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi^{(i)}) \omega(\xi^{(i)})$

la condition d'existence de la fonction d'importance servant dans cette estimation. Ainsi, l'approximation (2.3) reste valide tant que le support de la densité instrumentale contient celui de la densité cible i.e $supp(\tilde{p}) \supset supp(f.p)$. Ce qui est une condition naturelle.

Propriétés 2.2.2. *Le biais et la variance de* (2.3) *sont donnés par :*

$$b(\hat{I}(f)) = \mathbb{E}_{\tilde{p}}[\hat{I}(f) - I(f)] = 0 \tag{2.4}$$

et

$$Var_{\tilde{p}}[\hat{I}(f)] = \frac{Var_{\tilde{p}}[f(X)\omega(X)]}{N}. \tag{2.5}$$

Sous cette hypothèse de support et d'intégrabilité i.e $\mathbb{E}_{\tilde{p}} |f(X)\omega(X)| < +\infty$, on a la consistance et la normalité asymptotique.

Théorème 2.2.3. *Pour toute fonction test f,*

$$\hat{I}(f) \xrightarrow{p.s} I(f), \quad \text{à mesure que } N \text{ tend vers } \quad +\infty. \tag{2.6}$$

De plus, si

$$\mathbb{E}_{\tilde{p}}\left[f^2(X)\omega^2(X)\right] = \int_{\mathcal{X}} f^2(x)\frac{p^2(x)}{\tilde{p}(x)}dx < +\infty, \tag{2.7}$$

alors :

$$\sqrt{N}\left(\hat{I}(f) - I(f)\right)\sigma_f^{-1} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \tag{2.8}$$

où $\sigma_f = \sqrt{Var_{\tilde{p}}[f(X)\omega(X)]}$.

Démonstration. C'est une application directe de la LFGN et du TCL. □

Notons que la condition (2.7) garantit la finitude de la variance de l'estimateur (2.3). De plus, on peut exercer un contrôle sur cet estimateur sur un horizon fini comme par exemple disposer d'une borne pour les moments d'ordre k ou sur les queues de distributions. Le résultat suivant utilise entre autres les inégalités de déviations type Hoeffding et exploite le fait que (2.3) est une somme de variables i.i.d. Nous donnons en annexe quelques un des ces résultats (pour un exposé détaillé sur des résultats de convergence de sommes de variables aléatoires i.i.d on peut se référer à Petrov [1995].)

Théorème 2.2.4. *Quels que soient f fonction test, $k \geq 2$ et $N \geq 1$ (2.3) vérifie :*

$$\mathbb{E}_{\tilde{p}}\left|\hat{I}(f) - I(f)\right|^k \leq C(k)N^{-k/2}\,\mathbb{E}_{\tilde{p}}\left|f(\xi)\omega(\xi) - I(f)\right|^k \tag{2.9}$$

de plus, pour tout $t \geq 0$,

$$\mathbb{P}\left[\left|\hat{I}(f) - I(f)\right| \geq t\right] \leq 2\exp\left[-2Nt^2/osc(f\omega)\right] \tag{2.10}$$

où $C(k)$ est une constante dépendant uniquement de k et

$$osc(f) := \sup_{(x,x')\in\mathcal{X}\times\mathcal{X}} |f(x) - f(x')| = 2\inf_{c\in\mathbb{R}}\|f - c\|_\infty. \tag{2.11}$$

Démonstration. C'est une application directe des Théorème A.1.3 et Théorème A.1.4 de l'Annexe A. □

Remarque 2.2.5. *Si la fonction d'importance ω n'est pas bornée, les poids $\omega(\xi^i), i = 1, 2, \cdots, N$ seront très irrégulièrement répartis si bien que un nombre très petit des variables ξ^i a un poids non nul et élevé. Les autres, majoritaires et avec des poids nuls ne vont par conséquent guère contribuer aux estimations subséquentes.*

Par ailleurs, on peut également convenir du choix optimal de la densité instrumentale en terme de variance minimale pour cet estimateur.

Théorème 2.2.6. *La densité instrumentale qui minimise la variance de l'estimateur* (2.3) *est donnée par :*

$$\tilde{p}_{optim}(x) = \frac{|f(x)|p(x)}{\int_X |f(z)|p(z)dz}. \tag{2.12}$$

Démonstration. Il suffit de trouver une borne inférieure à la variance de $f(X)\omega(X)$ et de montrer que pour le choix de la densité instrumentale énoncée, cette borne inférieure est atteinte. En effet, pour toute densité instrumentale $\tilde{p}(\cdot)$ on a :

$$Var_{\tilde{p}}[f(X)\omega(X)] = \mathbb{E}_{\tilde{p}}[f^2(X)\omega(X)^2] - \mathbb{E}_{\tilde{p}}[f(X)\omega(X)]^2 \tag{2.13}$$

et d'après l'inégalité de Jensen il vient que :

$$\begin{aligned} \mathbb{E}_{\tilde{p}}[f^2(X)\omega(X)^2] &\geq \mathbb{E}_{\tilde{p}}[f(X)\omega(X)]^2 \\ &= \left(\int_X |f(x)|p(x)dx\right)^2. \end{aligned} \tag{2.14}$$

Il est alors immédiat de voir que cette borne inférieure est atteinte par le choix de $\tilde{p} = \tilde{p}_{\text{optim}}$. □

En pratique, ce choix optimal n'est guère d'un grand secours puisque ce dernier dépend de la constante de normalisation $\int_X |f(z)|p(z)dz$ qui se trouve être à un signe près l'intégrale dont on cherche à estimer. Néanmoins, il sert de guide parmi la classe des densités d'importance sous-optimales.

2.2.2 Échantillonnage préférentiel auto-normalisé

Le problème de la constante de normalisation dont souffre l'EP peut facilement être levé. Moyennant une réécriture, on peut voir que :

$$\begin{aligned} I(f) &= \frac{\int_X f(z)\omega(z)\tilde{p}(z)dz}{\int_X \omega(z)\tilde{p}(z)dz} \\ &= \frac{\mathbb{E}_{\tilde{p}}[f(X)\omega(X)]}{\mathbb{E}_{\tilde{p}}[\omega(X)]}. \end{aligned} \tag{2.15}$$

En prenant la contre partie empirique, on obtient un nouvel estimateur de l'EP dit auto-normalisé donné par :

$$\begin{aligned} \hat{\tilde{I}}(f) &= \frac{\frac{1}{N}\sum_{i=1}^N f(\xi^{(i)})\omega(\xi^{(i)})}{\frac{1}{N}\sum_{i=1}^N \omega(\xi^{(i)})} \\ &= \sum_{i=1}^N f(\xi^{(i)})\tilde{\omega}(\xi^{(i)}), \end{aligned} \tag{2.16}$$

où

$$\tilde{\omega}(\xi^{(i)}) = \frac{\omega(\xi^{(i)})}{\sum_{j=1}^{N} \omega(\xi^{(j)})}, \quad i = 1, 2, \cdots, N$$

sont les poids d'importance auto-normalisés, p et $\tilde{p}$ pouvant être connues à des constantes multiplicatives près. Le résumé de cette procédure est donné au tableau ci-dessous. Étant donné que (2.16) est un rapport d'estimateurs, on en déduit qu'il a un biais.

Algorithme 4 Échantillonnage d'importance auto-normalisé

1: **Choisir** $\tilde{p}$ telle que $supp(\tilde{p}) \supset supp(f \cdot p)$
2: **Pour** $i = 1, 2, \ldots, N$ **faire**
   — **Générer** $\xi^{(i)} \sim \tilde{p}(\cdot)$
   — **Poser** $\omega(\xi^{(i)}) = \frac{p(\xi^{(i)})}{\tilde{p}(\xi^{(i)})}$
3: **FinPour**
4: **Retourner** $\hat{\hat{I}}(f) = \frac{\sum_{i=1}^{N} f(\xi^{(i)})\omega(\xi^{(i)})}{\sum_{i=1}^{N} \omega(\xi^{(i)})}$

Mais, ce dernier disparaît asymptotiquement comme le montre le résultat suivant.

Propriétés 2.2.7. *Le biais et la variance de* (2.16) *sont donnés respectivement par :*

$$\begin{aligned} b(\hat{\hat{I}}(f)) &= \mathbb{E}_{\tilde{p}}\left[\hat{\hat{I}}(f) - I(f)\right] \\ &= \frac{I(f)Var_{\tilde{p}}[\omega(X)] - Cov_{\tilde{p}}[\omega(X), f(X)\omega(X)]}{N} + O(N^{-2}) \end{aligned} \tag{2.17}$$

et

$$\begin{aligned} Var_{\tilde{p}}[\hat{\hat{I}}(f)] = \frac{Var_{\tilde{p}}[f(X)\omega(X)] - 2I(f)Cov_{\tilde{p}}[\omega(X), f(X)\omega(X)]}{N} \\ + \frac{I(f)^2 Var_{\tilde{p}}[\omega(X)]}{N} + O(N^{-2}). \end{aligned} \tag{2.18}$$

Démonstration. La preuve est basée sur une application directe de la Delta-méthode (Voir Liu [2001], page 36). □

Notons que l'optimalité donnée par le Théorème 2.2.6 n'est plus atteinte. Néanmoins, dans certaines configurations (voir Van Dijk and Kloeck [1985], Casella and Robert [1998]) cet estimateur se compare favorablement sous l'erreur quadratique moyenne. Par ailleurs, on a également des résultats de consistance et de normalité asymptotique analogues à ceux de l'estimateur de l'EP.

Théorème 2.2.8. *Pour toute fonction test* f,

$$\hat{\hat{I}}(f) \xrightarrow{p.s} I(f), \quad \text{à mesure que } N \text{ tend vers } +\infty. \tag{2.19}$$

De plus, si

$$\mathbb{E}_{\tilde{p}}\left[(1+f^2(X))\omega^2(X)\right] = \int_{\mathcal{X}}(1+f^2(x))\frac{p^2(x)}{\tilde{p}(x)}dx < +\infty, \tag{2.20}$$

alors :

$$\sqrt{N}\left(\hat{\hat{I}}(f) - I(f)\right)\sigma_f'^{-1} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{2.21}$$

avec $\sigma_f'^2 = \mathbb{E}_{\tilde{p}}\left[(f(X)-I(f))^2\omega^2(X)\right]$.

Démonstration. Il suffit de voir que :

$$\sqrt{N}\left(\hat{\hat{I}}(f) - I(f)\right) = \frac{N^{-1/2}\sum_{i=1}^{N} G_f(\xi^{(i)})}{N^{-1}\sum_{i=1}^{N}\omega(\xi^{(i)})} \tag{2.22}$$

où $G_f(\xi^{(i)}) := \left[f(\xi^{(i)}) - I(f)\right]\omega(\xi^{(i)})$ une suite de variables aléatoires centrées i.i.d et de variance $\sigma_f'^2$. Étant donnés que le numérateur converge en distribution vers $\mathcal{N}(0,\sigma_f'^2)$ par le TCL et le dénominateur converge presque sûrement vers 1 par la LFGN, il s'en suit d'après le lemme de Slutsky que le rapport converge en distribution vers $\mathcal{N}(0,\sigma_f'^2)$. □

Comme pour (2.3), on peut obtenir une borne pour le moment d'ordre k avec $k \geq 2$ de l'estimateur de l'EP auto-normalisé (2.16) et avoir une inégalité de déviation de type Hoeffding.

Théorème 2.2.9. *Supposons que* $\mathbb{E}_{\tilde{p}}\left[\omega^k(X)\right] < +\infty$. *Il existe une constante* $C < +\infty$ *telle pour toute fonction test* f *et* $N \geq 1$ (2.16) *vérifie :*

$$\mathbb{E}_{\tilde{p}}\left|\hat{\hat{I}}(f) - I(f)\right|^k \leq CN^{-k/2}osc^k(f) \tag{2.23}$$

de plus, pour tout $t \geq 0$,

$$\mathbb{P}\left(\left|\hat{\hat{I}}(f) - I(f)\right| \geq t\right) \leq 4\exp\{-8Nt^2/9\|\omega\|_\infty^2 osc^2(f)\}. \tag{2.24}$$

Démonstration. Voir Cappé et al. [2005] page 294-295. □

Cependant, l'EP aussi bien dans sa forme non normalisée que auto-normalisée n'est adéquat lorsque l'on procède à une estimation de type récursif ou en ligne. En effet, on est obligé de recalculer l'estimé à partir non seulement de la nouvelle observation immédiatement disponible, mais aussi sur toute l'historique du processus d'observation. [2]

2. Ce dernier problème sera levé à la section Filtrage particulaire

Exemple 2.2.10. *(Quantiles et queues de distribution)*
L'EP peut facilement être appliqué à l'approximation des queues de distributions ainsi qu'aux quantiles. Supposons que l'on s'intéresse à l'estimation de la queue de distribution de X*,*

$$\theta := \mathbb{P}(X > s) = 1 - F(s) \quad \textit{avec} \quad X \sim F(\cdot) \ \textit{et} \ \ s \ \textit{un seuil donné.}$$

En se plaçant dans les conditions habituelles de l'EP avec $F(\cdot)$ *fonction de distribution cible et* $\tilde{F}(\cdot)$ *fonction de distribution instrumentale supposée absolument continue par rapport à* $F(\cdot)$*. Alors l'estimateur de* θ *par EP est donné par :*

$$\hat{\theta}^{EP} = \frac{1}{N}\sum_{i=1}^{N} 1_{\{Z_i > s\}}\omega(Z_i), \quad où \quad \omega(Z_i) := \frac{dF}{d\tilde{F}}(Z_i)$$

avec Z_i *i.i.d de fonction de distribution* $\tilde{F}(\cdot)$*. Dans la même lancée, on peut s'intéresser à l'estimation d'un* $\alpha-$*quantile défini par :*

$$y_\alpha = \inf\{y : F(y) > \alpha\}.$$

L'approche classique consiste à utiliser l'estimateur empirique de la fonction de distribution noté $\hat{F}_n$ *pour en déduire celle de* Y_α*,*

$$\hat{y}_\alpha = \inf\left\{y : \hat{F}_n(y) > \alpha\right\}.$$

L'approche par EP procède de la même idée sauf qu'en lieu et place de l'estimateur empirique de la fonction de distribution F*, l'estimateur par EP lui est substitué. En effet, étant donnée une fonction de distribution instrumentale* $\tilde{F}$ *absolument continue par rapport à* F*, en notant* $\omega(y) := \frac{dF}{d\tilde{F}}(y)$ *la dérivée de Radon-Nykodim on obtient une estimation par EP de la fonction de distribution* F *donnée par :*

$$\hat{\tilde{F}}_n(y) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{Z_i \leq y\}}\omega(Z_i).$$

Un estimateur du quantile d'ordre α *est alors donné par :*

$$\hat{y}_\alpha = \inf\left\{y : \hat{\tilde{F}}_n(y) > \alpha\right\}$$

Enfin, soulignons qu'en pratique il convient d'utiliser la forme auto-normalisée de ces estimateurs pour éviter des cas de divergences. On pourra consulter Cannamela et al. [2008] pour quelques résultats de convergence de cet estimateur ainsi que des variantes de ce dernier.

2.2.3 Échantillonnage préférentiel avec rééchantillonnage

Une des limitations de l'EP est que certaines variables aléatoires ont un poids d'importance quasi-nul voire nul si bien qu'elles n'apportent aucune contribution dans les estimations subséquentes. Une façon assez simple d'y remédier est de sélectionner les variables les plus contributrices par un procédé de tirage de celles-ci. L'échantillonnage préférentiel avec rééchantillonnage (EPR) ou échantillonnage d'importance avec rééchantillonnage (EIS) procède de cette idée. Introduit par Rubin [1987, 1988], l'EPR est une méthode en deux étapes qui permet de simuler un échantillon $(\xi_1, \xi_2, \ldots, \xi_N)$ asymptotiquement i.i.d d'une densité cible $p(\cdot)$. Tout comme l'EP, l'EPR utilise une densité instrumentale $\tilde{p}$ dans laquelle un premier échantillon i.i.d $(\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)})$ de taille $M \geq N$ avec les poids d'importance auto-normalisés définis par :

$$\tilde{\omega}^{(1)} := \frac{\omega(\tilde{\xi}^{(1)})}{\sum_{r=1}^{M} \omega(\tilde{\xi}^{(r)})}, \quad \tilde{\omega}^{(2)} := \frac{\omega(\tilde{\xi}^{(2)})}{\sum_{r=1}^{M} \omega(\tilde{\xi}^{(r)})}, \quad \ldots, \quad \tilde{\omega}^{(M)} := \frac{\omega(\tilde{\xi}^{(M)})}{\sum_{r=1}^{M} \omega(\tilde{\xi}^{(r)})} \tag{2.25}$$

tout en tenant compte de la contrainte habituelle sur les supports. Cette première phase d'échantillonnage est donc identique en tout point à l'EP. La deuxième phase quant à elle, consiste en la prise en compte effective de la contribution individuelle de chaque variable à travers le poids d'importance qui lui est associée. Il existe plusieurs façons de prendre en compte cette contribution. Une manière simple de le faire est d'utiliser des tirages avec remise des variables $\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)}$ proportionnellement aux poids d'importance $\tilde{\omega}^{(1)}, \tilde{\omega}^{(2)}, \ldots, \tilde{\omega^{(M)}}$ afin d'obtenir un nouvel échantillon $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(N)})$. Concrètement, chaque variable $\tilde{\xi}^{(j)}$ est tirée N^j fois selon la loi binomiale $\mathcal{B}(N, \tilde{\omega}^{(j)})$. Si bien que le vecteur $(N^1, N^2, \ldots, N^M)$ représentant le nombre de fois où chacune des variables $\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)}$ est tirée suit une loi multinomiale $\mathcal{M}(N, \tilde{\omega}^{(1)}, \tilde{\omega}^{(2)}, \ldots, \tilde{\omega}^{(M)})$ de probabilités de succès $(\tilde{\omega}^{(1)}, \tilde{\omega}^{(2)}, \ldots, \tilde{\omega}^{(M)})$. Ainsi, une fois les futurs indices $I^1, I^2, \ldots, I^N$ sont simulés indépendamment avec les probabilités :

$$\mathbb{P}(I^i = j | \tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)}) = \tilde{\omega}^{(j)}, \quad j = 1, 2, \ldots, M \tag{2.26}$$

on pose alors $\xi^{(i)} = \tilde{\xi}^{(I^i)}$, pour tout $i = 1, 2, \ldots, N$. D'où l'on déduit un nouvel estimateur de l'EP de (2.1) avec rééchantillonnage donné par :

$$\hat{\hat{I}}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi^{(i)}) = \frac{1}{N} \sum_{i=1}^{M} N^i f(\tilde{\xi}^{(i)}). \tag{2.27}$$

Ainsi, cette procédure privilégie les variables les plus contributrices en terme de poids d'importance. Il arrive aussi que la taille de l'échantillon originel M soit très grande devant celle de l'échantillon résultant N afin d'avoir une large palette de choix lors de la sélection. En d'autre terme, les variables au poids élevé sont dupliquées au détriment des variables de poids faible qui elles sont systématiquement éliminées. Le résumé de

la procédure est complètement décrite au tableau ci-dessous.

Algorithme 5 Échantillonnage préférentiel avec rééchantillonnage

Échantillonnage

1. **Choisir** $\tilde{p}$ telle que $supp(\tilde{p}) \supset supp(f \cdot p)$
2. **Pour** $j = 1, 2, \ldots, M$ **faire**
   — **Générer** $\tilde{\xi}^{(j)} \overset{\text{i.i.d}}{\sim} \tilde{p}(\cdot)$
   — **Poser** $\tilde{\omega}^{(j)} = \frac{p(\tilde{\xi}^{(j)})}{\tilde{p}(\tilde{\xi}^{(j)})} / \sum_{r=1}^{M} \frac{p(\tilde{\xi}^{(r)})}{\tilde{p}(\tilde{\xi}^{(r)})}$
3. **FinPour**

Rééchantillonage

1. **Pour** $i = 1, 2, \ldots, N$ **faire**
   — **Tirer** les indices I^i indépendamment avec les probabilités

$$\mathbb{P}(I^i = j | \tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)}) = \tilde{\omega}^{(j)}, \quad j = 1, 2, \ldots, M \tag{2.28}$$

   — **Poser** $\xi^{(i)} = \tilde{\xi}^{(I^i)}$
2. **FinPour**

Notons que l'estimateur (2.27) ainsi construit est sans biais. En effet,

$$\begin{aligned}
\mathbb{E}_{\tilde{p}}\left[\hat{\hat{I}}(f)\right] &= \mathbb{E}_{\tilde{p}}\left[\mathbb{E}_{\tilde{p}}\left[\hat{\hat{I}}(f)\middle|\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)}\right]\right] \\
&= \mathbb{E}_{\tilde{p}}\left[\frac{1}{N}\sum_{i=1}^{M} f(\tilde{\xi}^{(i)})\,\mathbb{E}_{\tilde{p}}\left[N^i\middle|\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \ldots, \tilde{\xi}^{(M)}\right]\right] \\
&= \mathbb{E}_{\tilde{p}}\left[\frac{1}{N}\sum_{i=1}^{M} f(\tilde{\xi}^{(i)}) N\tilde{\omega}^{(i)}\right] = \mathbb{E}_{\tilde{p}}\left[\sum_{i=1}^{M} f(\tilde{\xi}^{(i)})\tilde{\omega}^{(i)}\right] \\
&= \mathbb{E}_{\tilde{p}}\left[\hat{I}(f)\right].
\end{aligned} \tag{2.29}$$

Cependant, l'erreur quadratique moyenne de (2.27) est toujours supérieure à celle de (2.16) eu égard à la décomposition de l'erreur d'estimation

$$\hat{\hat{I}}(f) - I(f) = \left(\hat{\hat{I}}(f) - \hat{I}(f)\right) + \left(\hat{I}(f) - I(f)\right) \tag{2.30}$$

donnant celle de la variance :

$$\mathbb{E}_{\tilde{p}}\left[\hat{\hat{I}}(f) - I(f)\right]^2 = \mathbb{E}_{\tilde{p}}\left[\hat{\hat{I}}(f) - \hat{I}(f)\right]^2 + \mathbb{E}_{\tilde{p}}\left[\hat{I}(f) - I(f)\right]^2 \tag{2.31}$$

où l'on identifie aisément l'erreur associée à l'EP comme le second terme du membre de droite de cette égalité et le premier terme est assimilé au coût résultant de la transformation de l'EP en EPR. Par ailleurs, le rééchantillonnage rend caduque l'hypothèse i.i.d utilisée jusqu'ici afin d'étudier le comportement asymptotique des estimateurs. Une approche plus élaborée est requise pour étudier (2.27). On trouvera des résultats sur la consistance et la normalité asymptotique dans Douc and Moulines [2008].

Soulignons par ailleurs, qu'il existe d'autres formes de rééchantillonnage. Nous en traitons quelques unes en Annexe A.3. Comme mentionné en fin de sous-section 2.2.3 , l'EPR n'est pas adéquat pour les estimations de type récursif. C'est donc une version séquentielle qui est utilisée dans ce cas de figure. Notamment, dans les modèles de Markov cachés (MMC) où certaines problématiques soulevées font appel à des solutions récursives. Nous allons introduire les MMC à la section suivante et revenir plus amplement sur la formulation de la version séquentielle de l'EPR en sous-section Filtrage particulaire.

2.3 MMC & Approximations particulaires

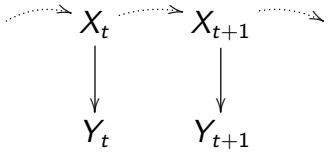
De façon générale, un modèle de Markov caché (MMC en abrégé) est un modèle statistique dans lequel un automate ou un système donné évolue selon une dynamique markovienne à travers des états dits cachés du fait que ces derniers ne sont pas directement observables. C'est donc au travers d'un autre mécanisme de mesure que l'on parvient à quantifier les réalisations d'un tel automate. De manière formelle, un MMC est la donnée d'un processus bivarié $(X_t, Y_t)_{t\geq 0}$ discret à valeurs dans un espace produit $\mathcal{X} \times \mathcal{Y}$ pour lequel $\mathcal{X}$ (resp. $\mathcal{Y}$) peut être très général[3] et tel que :

— $(X_t)_{t\geq 0}$ est une chaîne de Markov de distribution initiale ν et de noyau de transition K_t. Ce processus modélise l'automate à ces différents états cachés ;

— $(Y_t)_{t\geq 0}$ est un processus de mesure ou d'observation servant à quantifier les réalisations du processus $(X_t)_{t\geq 0}$ et vérifiant la Proposition 2.3.1 dite de *Canal d'observations sans mémoire.*

Proposition 2.3.1. *Les observations* $\{Y_t, t \geq 0\}$ *sont mutuellement indépendantes conditionnellement aux états* $\{X_t, t \geq 0\}$ *et la distribution* Y_t *conditionnellement aux variables* $\{X_t, t \geq 0\}$ *ne dépend que* X_t.

Schématiquement, on a le graphe de dépendance donné ci-dessous.



3. dénombrable, continu, voire hybride

Il faut cependant souligner que l'on peut facilement complexifier cette structure de dépendance en fonction de la complexité du système étudié. Notons que la formulation la plus courante d'un MMC est celle donnée par sa représentation à espace d'états :

$$\begin{cases} X_{t+1} = f_{t+1}(X_t, W_{t+1}) \\ Y_t = h_t(X_t, V_t) \end{cases} \tag{2.32}$$

où la suite $\{V_t, t \geq 0\}$ (resp. $\{W_t, t \geq 0\}$) est le bruit d'état (resp. d'observation) i.i.d et indépendante de X_0 et $\{f_t, t \geq 0\}$ (resp. $\{h_t, t \geq 0\}$) est une suite de fonctions mesurables. Dans la formulation (2.32), la première équation représente l'équation de transition ou d'états et la deuxième est l'équation d'observations ou de mesure. Une autre formulation similaire à celle par (2.32) obtenue par la donnée de la loi initiale de la chaîne $\nu(\cdot)$, le noyau de transition $K_t(\cdot,\cdot)$ ainsi que la loi d'émission $\psi(\cdot,\cdot)$. Le résumé de cette deuxième formulation est donné par :

$$\begin{cases} X_0 \sim \nu(\cdot) \\ K_t(X_{t-1}, X_t) := \mathbb{P}(X_t | X_{t-1}), \quad t \geq 1 \\ \psi(X_t, Y_t) := \mathbb{P}(Y_t | X_t) \end{cases} \tag{2.33}$$

Ainsi, un MMC est complètement caractérisé par la donnée du triplé (ν, K_t, ψ). De plus, lorsque l'espace d'état $\mathcal{X}$ (resp. d'observation $\mathcal{Y}$) est de dimension finie alors K_t (resp. ψ) est une matrice stochastique. Par ailleurs, selon la nature du noyau on peut avoir des simplifications possibles :

- MMC homogène lorsque le noyau K_t est indépendant du temps ;
- MMC partiellement dominé lorsque K_t admet une densité de probabilité ;
- MMC totalement dominé lorsque K_t et ψ admettent chacune une densité de probabilité ;

La mesure de référence étant la mesure de Lebesgue dans le cas continu. Dans les applications que nous effectuons nous nous plaçons dans le dernier cas de figure où les densités de transition et de mesure sont complètement spécifiées.

Exemple 2.3.2. *(Modèle de croissance)*
Considérons la représentation à espace d'états du MMC suivant :

$$\begin{cases} X_t = \frac{X_{t-1}}{2} + 25\frac{X_{t-1}}{1-X_{t-1}^2} + 8\cos(1.2t) + V_t \\ Y_t = \frac{X_t^2}{20} + W_t \end{cases} \tag{2.34}$$

où $X_0 \sim \mathcal{N}(0, 15)$, $\begin{pmatrix} V_t \\ W_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{V_t}^2 & 0 \\ 0 & \sigma_{W_t}^2 \end{pmatrix}\right)$ *avec la donnée des variances* $\sigma_V^2 = 10$ *et* $\sigma_W^2 = 1$. *Le signal caché est modélisé par* X *tandis que le processus des observations est modélisé par* Y. *On peut facilement voir que les noyaux de transition et d'émission admettent des densités gaussiennes. Une trajectoire du couple (signal,observation) est donnée à la Figure 2.1. On peut déjà noter la forte non-linéarité de ce couple. Cet exemple dû à Kitagawa sert souvent d'exemple test aux différentes méthodes de MCs utilisées.*

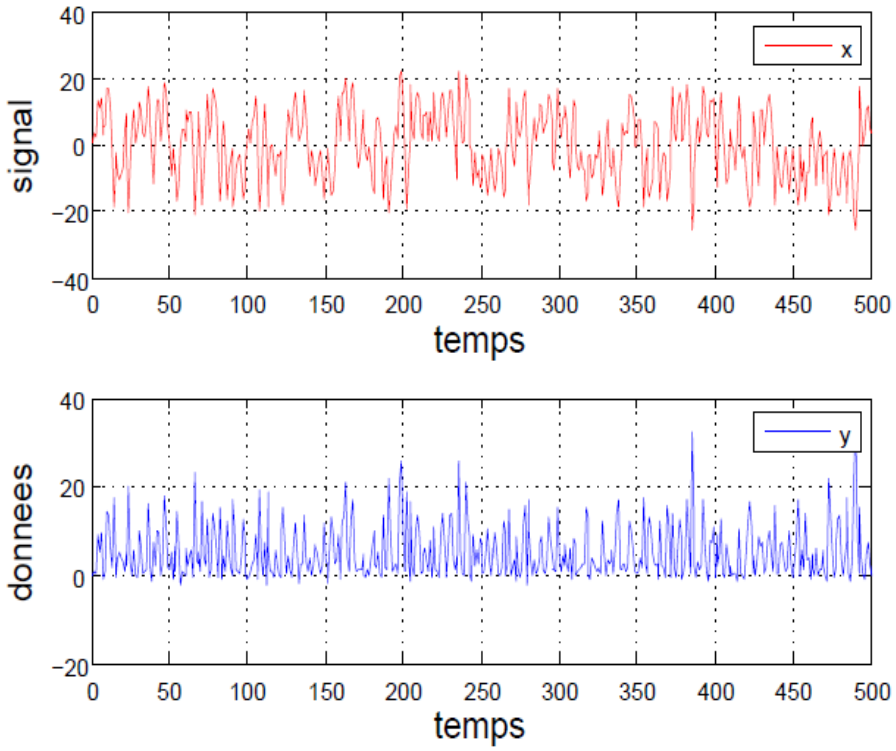


FIGURE 2.1 – Trajectoire du modèle de croissance sur un horizon de $T = 500$ réalisations.

Problématique des MMC

Les problèmes de base que soulèvent les MMC lorsque $\mathcal{X}$ est au plus dénombrable peuvent être scindés en trois catégories :

1. pour un MMC de paramètre (ν, K_t, ψ) donné, évaluer la vraisemblance d'une séquence d'observations $Y_{0:k} := (Y_0, Y_1, \ldots, Y_k)$?
2. considérons un MMC de paramètre (ν, K_t, ψ) et $Y_{0:k}$ une suite d'observations. Quelle la séquence d'états $X_{0:k} := (X_0, X_1, \ldots, X_k)$ la plus probable permettant de générer les observations $Y_{0:k}$?
3. pour une séquence d'observations $Y_{0:k}$, trouver le MMC de paramètre (ν, K_t, ψ) le plus vraisemblablement à l'origine de celles-ci c'est à dire qui ajuste le mieux $Y_{0:k}$?

Solutions

Les solutions à ces problèmes canoniques sont connues comme étant les algorithmes *Forward*, *Forward-Backward* de Baum, et les algorithmes de Viterbi (maximisation du

MAP). On peut entre autres consulter le tutoriel de Rabiner [1989] et le livre Rabiner and Juang [1993] pour un exposé détaillé ou le cours de Campillo [2006] pour un résumé concis de ces solutions. Notons simplement, que ces dernières partagent en commun l'usage de la règle de Bayes eu égard aux fonctionnelles conditionnelles intervenant dans leurs formulations. Dans la section suivante nous focalisons notre attention sur la description des méthodes MCs comme solutions adaptées aux problèmes de filtrage, lissage et d'estimation dans les MMC non linéaires et/ou non gaussiens en ayant pris soin de les décrire au préalable.

2.3.1 Filtrage particulaire

Le filtrage particulaire est une méthode de simulation séquentielle mettant en compétition des variables aléatoires appelées *particules* qui explorent de manière indépendantes l'espace d'états tout en interagissant entre elles par le biais d'un mécanisme de sélection. Le filtrage particulaire fait partie de la classe des MCs, analogue récursive des méthodes de Monte Carlo par chaînes de Markov. Concrètement, étant donné un MMC

$$\begin{cases} X_t = F_t(X_{t-1}, \Omega_t) \\ Y_t = H_t(X_t, V_t) \end{cases} \tag{2.35}$$

où Ω_t, V_t sont des bruits blancs indépendants mutuellement et indépendants de X_0 de loi initiale $\nu(\cdot)$. L'objet du filtrage est alors de calculer les distributions *a posteriori* $\mathbb{P}(X_{0:t}|Y_{1:t})$ et en particulier les distributions marginales $\mathbb{P}(X_t|Y_{1:t-1})$ et $\mathbb{P}(X_t|Y_{1:t})$ en tout $t \geq 0$. Dans toute la suite, on admet que les distributions étudiées admettent des densités par rapport à une mesure de référence. Typiquement, le filtrage particulaire permet de solutionner des équations faisant intervenir des fonctionnelles de la densité *a posteriori* $p(x_{0:t}|y_{1:t})$ et en particulier la densité marginale $p(x_t|y_{1:t})$ en temps réel. Pour ce faire, le théorème de Bayes fournit la formule récursive suivante :

$$\begin{aligned} p\left(x_{0:t}|y_{1:t}\right) &= \frac{p\left(y_{1:t}|x_{0:t}\right) p(x_{0:t})}{\int_{\mathcal{X}^{t+1}} p\left(y_{1:t}|x_{0:t}\right) p(x_{0:t}) dx_{0:t}} && (2.36) \\ &= p\left(x_{0:t-1}|y_{1:t-1}\right) \frac{p\left(y_t|x_t\right) p\left(x_t|x_{t-1}\right)}{p\left(y_t|y_{1:t-1}\right)} && (2.37) \\ &\propto p\left(x_{0:t-1}|y_{1:t-1}\right) p\left(y_t|x_t\right) p\left(x_t|x_{t-1}\right) && (2.38) \end{aligned}$$

et les densités marginales satisfont :

$$p\left(x_t|y_{1:t-1}\right) = \int_{\mathcal{X}} p\left(x_t|x_{t-1}\right) p(x_{t-1}|y_{1:t-1}) dx_{t-1} \tag{2.39}$$

$$p\left(x_t|y_{1:t}\right) = \frac{p\left(y_t|x_t\right) p\left(x_t|y_{1:t-1}\right)}{\int_{\mathcal{X}} p\left(y_t|x_t\right) p\left(x_t|y_{1:t-1}\right) dx_t} \tag{2.40}$$

sous l'hypothèse de *canal sans mémoire* 2.3.1. L'équation (2.39) est appelée équation de prédiction tandis que (2.40) est l'équation de correction. Les équations (2.38), (2.39) et (2.40) admettent rarement des solutions analytiques sauf pour de rares cas incluant les MMC linéaires et gaussiens. Dans ce cas, les solutions sont données par le filtre de Kalman [1960], Kalman and Bucy [1961]. Autrement, ces formules font intervenir des intégrales complexes de grande dimension [4] impossibles à évaluer analytiquement. C'est la raison pour laquelle il convient de les approcher. Dans cette optique, le filtrage particulaire permet d'apporter une solution séquentielle indépendamment de la dimension des intégrales.

2.3.1.1 Échantillonnage d'importance séquentiel

Commençons par reformuler les problématiques d'intégration rencontrées dans les équations (2.38), (2.39) et (2.40) avant de donner une adaptation séquentielle de l'EP pour les approximer. Considérons une fonction $f_t : \mathcal{X}^{t+1} \to \mathbb{R}^{n_{f_t}}$ intégrable par rapport à $p(x_{0:t}|y_{1:t})$. Soit alors à évaluer l'intégrale :

$$I(f_t) = \int_{\mathcal{X}^{t+1}} f_t(x_{0:t}) p(x_{0:t}|y_{1:t}) dx_{0:t}. \tag{2.41}$$

La démarche de la section précédente impose que l'on se donne une densité instrumentale notée $q(x_{0:t}|y_{1:t})$ dont le support contient celui de la densité cible $p(x_{0:t}|y_{1:t})$ puis de former le poids d'importance non normalisé

$$\omega_t(x_{0:t}) := \frac{p(x_{0:t}|y_{1:t})}{q(x_{0:t}|y_{1:t})}. \tag{2.42}$$

On en déduit une réécriture de l'intégrale (2.41) donnée par :

$$I(f_t) = \frac{\int_{\mathcal{X}^{t+1}} f_t(x_{0:t}) \omega_t(x_{0:t}) q(x_{0:t}|y_{1:t}) dx_{0:t}}{\int_{\mathcal{X}^{t+1}} \omega_t(x_{0:t}) q(x_{0:t}|y_{1:t}) dx_{0:t}}. \tag{2.43}$$

En se donnant une séquence de N particules $\{\xi_{0:t}^{(i)}\}_{i=1}^N$ i.i.d et de distribution $q(\cdot|y_{1:t})$, on tire une estimation de (2.41) par EP :

$$\hat{I}_N(f_t) = \frac{\frac{1}{N}\sum_{i=1}^N f_t(\xi_{0:t}^{(i)}) \omega_t(\xi_{0:t}^{(i)})}{\frac{1}{N}\sum_{i=1}^N \omega_t(\xi_{0:t}^{(i)})} = \sum_{i=1}^N f_t(\xi_{0:t}^{(i)}) \tilde{\omega}_t^{(i)} \tag{2.44}$$

avec

$$\tilde{\omega}_t^{(i)} := \frac{\omega_t(\xi_{0:t}^{(i)})}{\sum_{j=1}^N \omega_t(\xi_{0:t}^{(j)})}, i = 1, 2, \ldots, N$$

4. par exemple la constante de normalisation $p(y_{1:t}) = \int p(y_{1:t}|x_{0:t}) p(x_{0:t}) dx_{0:t}$, la distribution marginale $p(x_t|y_t)$, toute intégrale de la forme $\int f_t(x_{0:t}) p(x_{0:t}|y_{1:t}) dx_{0:t}$

les poids d'importance auto-normalisés. Cet estimateur, malgré les propriétés qu'on lui connaît, n'est pas adéquat lorsqu'il s'agit de l'évaluer pour tout t. Dès lors, il convient de rendre cette évaluation séquentielle en t. Pour cela, l'hypothèse suivante est capitale.

Hypothèses 2.3.3. *La densité instrumentale admet la factorisation suivante :*

$$q(x_{0:t}|y_{1:t}) = q(x_{0:t-1}|y_{1:t-1})q(x_t|x_{0:t-1}, y_{1:t}). \tag{2.45}$$

Cette hypothèse stipule tout simplement que la densité instrumentale admette une densité marginale à l'instant $t-1$. En itérant cette récurrence sur t, on obtient l'égalité suivante :

$$q(x_{0:t}|y_{1:t}) = q(x_0)\prod_{k=1}^{t} q(x_k|x_{0:k-1}, y_{1:k}). \tag{2.46}$$

Notons que parmi les densités instrumentales qui satisfont cette hypothèse on compte l' *a priori* de transition de la chaîne de Markov. En combinant (2.38) et (2.45) le poids d'importance non normalisé admet la relation de récurrence suivante :

$$\begin{aligned}
\omega_t(x_{0:t}) &= \frac{p(x_{0:t}|y_{1:t})}{q(x_{0:t}|y_{1:t})} \\
&\propto \frac{p\left(x_{0:t-1}|y_{1:t-1}\right) p\left(y_t|x_t\right) p\left(x_t|x_{t-1}\right)}{q(x_{0:t-1}|y_{1:t-1})q(x_t|x_{0:t-1}, y_{1:t})} \\
&\propto \omega_{t-1}(x_{0:t-1})\frac{p\left(y_t|x_t\right) p\left(x_t|x_{t-1}\right)}{q(x_t|x_{0:t-1}, y_{1:t})}
\end{aligned} \tag{2.47}$$

D'où une version trajectorielle particulaire de cette récurrence donnée par :

$$\begin{aligned}
&\forall (i,t) \in \{1,2,\ldots,N\} \times \{1,2,\ldots,T\} \\
&\omega_t^{(i)} \propto \omega_{t-1}^{(i)} \frac{p(y_t|\xi_t^{(i)})p(\xi_t^{(i)}|\xi_{t-1}^{(i)})}{q(\xi_t^{(i)}|\xi_{0:t-1}^{(i)}; y_{1:t})}
\end{aligned} \tag{2.48}$$

On obtient ainsi une approximation séquentielle en t de la densité cible $p(x_{0:t}|y_{1:t})$ en terme de nuage pondéré $\{\tilde{\omega}_t^{(i)}, \xi_{0:t}^{(i)}\}_{i=1}^N$ si bien qu'une estimation particulaire de (2.41) par EPS est donnée par :

$$\hat{\hat{I}}_N(f_t) = \sum_{i=1}^{N} \tilde{\omega}_t^{(i)} f_t(\xi_{0:t}^{(i)}) \tag{2.49}$$

où

$$\tilde{\omega}_t^{(i)} := \frac{\omega_t^{(i)}}{\sum_{j=1}^N \omega_t^{(j)}}, \quad i = 1, 2, \cdots, N$$

sont les poids auto-normalisés.

Remarque 2.3.4. *La différence entre* (2.44) *et* (2.49) *réside d'une part dans la séquentialité de l'évaluation des poids d'importance ainsi que dans la formation des particules.Pour* (2.44) *l'échantillon servant d'estimation est tiré d'un seul coup contrairement à ce qui est fait dans* (2.49).

Le résumé de cette approche est donné ci-dessous.

Algorithme 6 Échantillonnage préférentiel séquentiel

Initialisation

$t = 0$

Pour $i = 1, 2, \ldots, N$ **faire**

— Tirer $\xi_0^{(i)} \sim \nu(\cdot)$

— Poser $\tilde{\omega}_0^{(i)} = \frac{1}{N}$

FinPour

$t = t + 1$

Échantillonnage d'importance

1. **Pour** $i = 1, 2, \ldots, N$ **faire**
   — **Tirer** $\xi_t^{(i)} \sim q(\cdot|\xi_{0:t-1}^{(i)}, y_{1:t})$
   — Évaluer les poids d'importance

$$\omega_t^{(i)} \propto \omega_{t-1}^{(i)} \frac{p(y_t|\xi_t^{(i)}) p(\xi_t^{(i)}|\xi_{t-1}^{(i)})}{q(\xi_t^{(i)}|\xi_{0:t-1}^{(i)}, y_{1:t})}$$

2. **FinPour**

Normaliser les poids

1. **Pour** $i = 1, 2, \ldots, N$ **faire**

$$\tilde{\omega}_t^{(i)} = \frac{\omega_t^{(i)}}{\sum_{j=1}^{N} \omega_t^{(j)}}$$

2. **FinPour**
3. **Poser** $t = t + 1$ et aller à l'étape Échantillonnage d'importance.

Notons que parmi la classe des lois instrumentales admissibles, un choix particulier confert une variance minimale aux poids d'importance auto-normalisés.

Proposition 2.3.5. *Conditionnellement à* $\xi_{0:t-1}^{(i)}$ *et* $y_{1:t}$*, la densité instrumentale qui minimise la variance du poids d'importance auto-normalisé* $\tilde{\omega}_t^{(i)}$ *est donnée par :*

$$q(x_t|\xi_{0:t-1}^{(i)}, y_{1:t}) = p(x_t|\xi_{t-1}^{(i)}, y_t). \tag{2.50}$$

Démonstration. Il suffit de voir que la variance de $\tilde{\omega}_t^{(i)}$ est donnée par

$$V_{q(x_t|\xi_{0:t-1}^{(i)},y_{1:t})}(\tilde{\omega}_t^{(i)}) = (\tilde{\omega}_{t-1}^{(i)})^2 \left[\int_{\mathcal{X}} \frac{\left(p(y_t|x_t)p(x_t|\xi_{t-1}^{(i)})\right)^2}{q(x_t|\xi_{0:t-1}^{(i)}, y_{1:t})} dx_t - \left(p(y_t|\xi_{t-1}^{(i)})\right)^2 \right] \quad (2.51)$$

qui s'annule pour le choix de $q(x_t|\xi_{0:t-1}^{(i)}, y_{1:t}) = p(x_t|\xi_{t-1}^{(i)}, y_t)$, voir Doucet et al. [2000] pour le détail du calcul. □

Soulignons qu'en pratique cette densité optimale est difficile voire impossible à obtenir. De plus, l'on peut montrer que la variance des poids augmente en fonction du temps.D'où le problème de dégénérescence suivant.

Remarque 2.3.6. *Il est connu que l'EP n'est pas adéquat pour les espaces de grandes dimensions (voir Bengtsson et al. [2008]). En effet, la distribution des poids devient de plus en plus asymétrique approchant celle d'une Dirac. Si bien qu'au bout de quelques itérations toutes les particules ont des poids d'importance nuls sauf une avec un poids valant 1. Une autre façon d'appréhender ce phénomène de dégénérescence réside dans le fait que la variance des poids d'importance augmente avec le temps. Afin de corriger cette imperfection, on introduit une étape de sélection appelé aussi rééchantillonnage. Elle consiste à sélectionner parmi les particules proposées par la partie EP de l'algorithme celles qui vont survivre à la prochaine génération et celles qui seront simplement éliminées. Les particules au poids d'importance élevé sont dupliquées au détriment de celles au poids faibles selon un mécanisme de redistribution des poids. Nous en traitons quelques uns en section A.3.*

2.3.1.2 Filtre de Bootstrap

Dans le choix de la densité instrumentale, l'usage courant est de recourir à l'*a priori* de transition de la chaîne de Markov comme densité instrumentale

$$q(x_t|\xi_{0:t-1}^{(i)}, y_{1:t}) = p(x_t|\xi_{t-1}^{(i)}). \quad (2.52)$$

Ce choix donne lieu au filtre de *Bootstrap* connu aussi sous le nom de *méthode de condensation*. Il est de loin le plus simple. L'avantage manifeste de ce dernier étant la disponibilité immédiate de cette densité, puisqu'elle est une donnée du modèle. Si bien qu'aucune approximation n'est nécessaire afin de l'obtenir. Nous en donnons un résumé ci-après. Notons que diverses généralisations existent. Cependant, dans la suite nous ne les traiterons pas. On pourra consulter avec intérêt Pitt and Shephard [1999] pour une généralisation appelé *filtre auxiliaire*.

Algorithme 7 Filtre de Bootstrap

Initialisation

$t = 0$

Pour $i = 1, 2, \ldots, N$ **faire**

— Tirer $\xi_0^{(i)} \sim \nu(\cdot)$

— Poser $\tilde{\omega}_0^{(i)} = \frac{1}{N}$

FinPour

$t = t + 1$

Échantillonnage d'importance

1. **Pour** $i = 1, 2, \ldots, N$ **faire**
   — **Tirer** $\xi_t^{(i)} \sim q(\cdot|\xi_{0:t-1}^{(i)}, y_{1:t})$ **et Poser** $\tilde{\xi}_{0:t}^{(i)} = (\tilde{\xi}_{0:t-1}^{(i)}, \xi_t^{(i)})$
   — Évaluer les poids d'importance

$$\omega_t^{(i)} \propto p(y_t|\xi_t^{(i)}) \tag{2.53}$$

2. **FinPour**

Normaliser les poids

1. **Pour** $i = 1, 2, \ldots, N$ **faire**

$$\tilde{\omega}_t^{(i)} = \frac{\omega_t^{(i)}}{\sum_{j=1}^{N} \omega_t^{(j)}}$$

2. **FinPour**

Rééchantillonage

1. Appliquer un algorithme de rééchantillonnage section A.3 de l'appendice pour convertir $\left\{\xi_{0:t}^{(i)}, \tilde{\omega}_t^{(i)}\right\}_{i=1}^{N}$ en $\left\{\tilde{\xi}_{0:t}^{(i)}, 1/N\right\}_{i=1}^{N}$
2. **Poser** $t = t + 1$ et aller à l'étape Échantillonnage d'importance.

On peut remarquer que les particules $\tilde{\xi}_{0:t}^{(i)}$ obtenues après rééchantillonnage sont équi-pondérés. Ce qui simplifie d'avantage l'écriture du poids d'importance constatée en (2.53).

2.3.2 Lissage particulaire

De façon générale, le lissage consiste à approcher un état ou une séquence d'états d'un système donné à la lumière d'observations. Celles-ci peuvent être constituées de toute la trajectoire du processus d'observation ou une partie de celle-ci. Tout comme le filtrage particulaire, le lissage particulaire est motivé par l'absence de solutions analytiques de fonctionnelles conditionnelles qui interviennent dans son élaboration. Le

cadre restant celui des MMC non linéaires et/ou non gaussiens. L'exposé suivant s'appuie sur le fait la chaîne X reste markovienne lorsqu'on inverse l'axe des temps. En effet :

Proposition 2.3.7. *Conditionnellement à $Y_{1:n}$, $\{X_{n-k}, k \geq 0\}$ est une chaîne de Markov de densités de transition données par :*

$$p(x_k|x_{k+1:n}, y_{1:n}) = p(x_k|x_{k+1}, y_{1:k}). \tag{2.54}$$

Démonstration. Voir, Cappé et al. [2005], p.70. □

2.3.2.1 Lissage Forward-Backward

Le lissage marginal a pour objet d'approcher la distribution conditionnelle de X_k connaissant toutes les observations jusqu'à l'instant n, i.e $Y_{1:n}$ pour tout $k < n$ et ce à rebours. En terme de densité, il faut donc approximer la densité de lissage $p(x_k|y_{:n})$ en utilisant une approximation de la densité marginale antérieure $p(x_{k+1}|y_{:n})$. Pour ce faire, la relation suivante permet d'établir une telle récurrence.

Lemme 2.3.8. *Pour tout $k < n$,*

$$p(x_k|y_{1:n}) = \int_{\mathcal{X}} \frac{p(x_{k+1}|y_{1:n})p(x_{k+1}|x_k)}{\int_{\mathcal{X}} p(x_{k+1}|x_k)p(x_k|y_{1:k})dx_k} dx_{k+1}. \tag{2.55}$$

Supposons d'une part disposer des approximations particulaires de la phase *Forward* des densités de filtrage $p(x_k|y_{1:k})$:

$$\hat{p}(x_k|y_{1:k}) = \sum_{i=1}^{N} \tilde{\omega}_k^{(i)} \delta_{\xi_k^{(i)}}(x_k), \quad k = 1, 2, \ldots, n \tag{2.56}$$

avec $\tilde{\omega}_k^{(i)}$ le poids d'importance auto-normalisé associé à la particule $\xi_k^{(i)}$. D'autre part, supposons également disposer de l'approximation *Backward* de la densité de lissage $p(x_{k+1}|y_{1:n})$ donnée par

$$\hat{p}(x_{k+1}|y_{1:n} = \sum_{j=1}^{N} \tilde{\omega}_{k+1}^{(j)} \delta_{\xi_{k+1}^{(j)}}(x_{k+1}). \tag{2.57}$$

En combinant (2.56) et (2.57), une approximation particulaire de la densité d'intérêt est donnée par :

$$\hat{p}(x_k|y_{1:n}) = \sum_{i=1}^{N} \tilde{\omega}_{k|n}^{(i)} \delta_{\xi_k^{(i)}}(x_k) \tag{2.58}$$

où

$$\tilde{\omega}_{k|n}^{(i)} = \tilde{\omega}_k^{(i)} \left(\sum_{j=1}^{N} \tilde{\omega}_{k+1|n}^{(j)} \frac{p\left(\xi_{k+1}^{(j)} | \xi_k^{(i)}\right)}{\sum_{r=1}^{N} \tilde{\omega}_k^{(r)} p\left(\xi_{k+1}^{(j)} | \xi_k^{(r)}\right)} \right)$$

avec la convention $\tilde{\omega}_{n|n}^{(i)} = \tilde{\omega}_n^{(i)}$. Un résumé de cette approche est donné au tableau suivant.

Algorithme 8 Lissage marginal Forward-Backward

1: Phase *Forward* : **Pour** $k = 1, \ldots, n$
   — **Exécuter** un algorithme de filtrage particulaire pour obtenir des particules pondérées $\left\{\xi_k^{(i)}, \omega_k^{(i)}\right\}_{i=1}^{N}$.
2: Phase *Backward* :
   — **Pour** $i = 1, \ldots, N$ **poser** $\tilde{\omega}_{n|n}^{(i)} = \tilde{\omega}_n^{(i)}$
   — **Pour** $k = n-1, n-2, \ldots, 0$ **et** $i = 1, \ldots, N$ **Poser**

$$\tilde{\omega}_{k|n}^{(i)} = \tilde{\omega}_k^{(i)} \left(\sum_{j=1}^{N} \tilde{\omega}_{k+1|n}^{(j)} \frac{p\left(\xi_{k+1}^{(j)} | \xi_k^{(i)}\right)}{\sum_{r=1}^{N} \tilde{\omega}_k^{(r)} p\left(\xi_{k+1}^{(j)} | \xi_k^{(r)}\right)} \right)$$

Remarque 2.3.9. *La complexité de cet algorithme est de* $O(N^2)$ *pour chaque pas de temps. Ce qui rend son usage délicat lorsque le nombre de particules* N *augmente. De plus, son efficacité est assujettie à celle de la phase Forward de production des particules. En effet, les positions de celles-ci demeurent inchangées après l'exécution de la phase Backward. Puisque seuls, les poids d'importance sont mis à jour dans cette phase. Par conséquent, si les particules de la phase Forward explorent une zone de faible intérêt, les estimations utilisant des fonctionnelles de la densité de lissage ne seront guère très utiles. Ce qui se traduit par des estimations éloignées des grandeurs estimées.*

2.3.2.2 Lissage 2-filtres

Dans le cadre du lissage marginal, une autre approche est envisageable afin de limiter la dépendance de performance de ce dernier qu'à la seule phase *Forward*. En

effet, partant de la factorisation de la densité marginale $p(x_k|y_{1:n})$ suivante :

$$\begin{aligned} p(x_k|y_{1:n}) &= p(x_k|y_{1:k-1}, y_{k:n}) \\ &= \frac{p(x_k|y_{1:k-1})p(y_{k:n}|y_{1:k-1}, x_k)}{p(y_{k:n}|y_{1:k-1})} \\ &\propto p(x_k|y_{1:k-1})p(y_{k:n}|x_k) \\ &\propto \underbrace{p(x_k|y_{1:k})}_{filtre1}\overbrace{p(y_{k:n}|x_k)}^{filtre2} \end{aligned} \tag{2.59}$$

appelé *2-filtre* à la suite de Fraser and Protter [1969] qui l'ont appliqué aux modèles linéaires gaussiens. C'est donc une combinaison du filtre *Forward* classique (filtre 1) et d'un second filtre appelé le filtre d'information arrière (*Backward Information filter* en anglais) dans Mayne [1966]. Afin d'évaluer récursivement le filtre 2, la relation suivante est utilisée :

$$p(y_{k:n}|x_k) = \int_{\mathcal{X}} p(y_{k+1:n}|x_{k+1})p(x_{k+1}|x_k)p(y_k|x_k)dx_{k+1} \tag{2.60}$$

Cependant, $p(y_{k:n}|x_k)$ n'est pas une densité en x_k. Si bien que son intégrale sur $\mathcal{X}$, n'est pas forcement bornée [5]. Ainsi, une renormalisation proposée dans Briers et al. [2010] permet de gérer cette éventualité. L'idée sous-jacente est d'introduire une densité artificielle $\gamma_k(x_k)$ telle que :

$$p(y_{k:n}|x_k) \propto \frac{\tilde{p}(x_k|y_{k:n})}{\gamma_k(x_k)} \tag{2.61}$$

$$\tilde{p}(x_k|y_{k:n}) = \int_{\mathcal{X}^{n-k}} \tilde{p}(x_{k:n}|y_{k:n})dx_{k+1:n} \tag{2.62}$$

et satisfaisant

$$\tilde{p}(x_{k:n}|y_{k:n}) \propto \gamma_k(x_k)\prod_{r=k}^{n} p(x_{r+1}|x_r)\prod_{r=k+1}^{n} p(y_r|x_r) \tag{2.63}$$

rendant la quantité d'intérêt $p(y_{k:n}|x_k)$ normalisable. La relation (2.60) devient alors :

$$\begin{aligned} p(y_{k:n}|x_k) &= \int_{\mathcal{X}} p(y_{k+1:n}|x_{k+1})p(x_{k+1}|x_k)p(y_k|x_k)dx_{k+1} \\ &\propto \int_{\mathcal{X}} \frac{\tilde{p}(x_{k+1}|y_{k+1:n})}{\gamma_{k+1}(x_{k+1})}p(x_{k+1}|x_k)p(y_k|x_k)dx_{k+1} \\ &= \frac{p(y_k|x_k)}{\gamma_k(x_k)}\int_{\mathcal{X}} \tilde{p}(x_{k+1}|y_{k+1:n})\frac{p(x_{k+1}|x_k)\gamma_k(x_k)}{\gamma_{k+1}(x_{k+1})}dx_{k+1} \end{aligned} \tag{2.64}$$

5. Certaines références dont Kitagawa [1996] font explicitement l'hypothèse de la bornitude de $p(y_{k:n}|x_k)$.

Remarque 2.3.10. *Le choix de γ_k est certes arbitraire mais il est tel que le rapport* (2.61) *ait un sens. Un choix naturel est l' a priori $\gamma_k(x_k) = p(x_k)$ ou de son approximation. Cependant, le lissage 2-filtre a la même complexité de $O(nN^2)$ que le lissage marginal Forward-Backward.*

2.3.2.3 Lissage joint

Toujours, en faisant l'observation que la densité marginale $p(x_k|y_{1:n})$ peut être obtenue par marginalisation de la densité jointe $p(x_{k:n}|y_{1:n})$, on peut obtenir une généralisation du lissage marginal au cas du lissage joint. En effet, comme précédemment, on peut établir une relation de récurrence liant les densités jointes $p(x_{k:n}|y_{1:n})$ et $p(x_{k+1:n}|y_{1:n})$ de la manière suivante.

Lemme 2.3.11. *Pour tout $k < n$, la densité de lissage joint $p(x_{k:n}|y_{1:n})$ se factorise comme suit :*

$$\begin{aligned} p(x_{k:n}|y_{1:n}) &= p(x_k|x_{k+1:n}, y_{1:n})p(x_{k+1:n}|y_{1:n}) \\ &= p(x_k|x_{k+1}, y_{1:k})p(x_{k+1:n}|y_{1:n}). \end{aligned} \tag{2.65}$$

De plus, en itérant cette relation on obtient :

$$p(x_{k:n}|y_{1:n}) = p(x_n|y_{1:n}) \prod_{r=k}^{n-1} p(x_r|x_{r+1}, y_{1:r}). \tag{2.66}$$

Démonstration. Il suffit d'appliquer la règle de Bayes avec la Prop. 2.3.7. □

De cette relation, on peut voir qu'une approximation de la densité d'intérêt $p(x_{k:n}|y_{1:n})$ requiert celle de $p(x_r|x_{r+1}, y_{1:r})$ pour tout $k \leq r \leq n-1$. Pour se faire, la règle de Bayes fournit la relation :

$$p(x_r|x_{r+1}, y_{1:r}) \propto p(x_r|y_{1:r})p(x_{r+1}|x_r) \tag{2.67}$$

d'où l'on tire une approximation particulaire de cette densité :

$$\hat{p}(x_r|x_{r+1}, y_{1:r}) = \sum_{i_r=1}^{N} \kappa_r^{(i_r)} \delta_{\xi_r^{(i_r)}}(x_r) \tag{2.68}$$

avec

$$\kappa_r^{(i_r)} = \frac{\tilde{\omega}_r^{(i_r)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(i_r)})}{\sum_{m=1}^{N} \tilde{\omega}_r^{(m)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(m)})}, \quad r = k, k+1, \ldots, n-1. \tag{2.69}$$

Partant de (2.68), on peut obtenir une approximation particulaire de $p(x_{k:n}|y_{1:n})$ donnée par :

$$\begin{aligned} \hat{p}(x_{k:n}|y_{1:n}) = \sum_{i_k=1}^{N} \sum_{i_{k+1}=1}^{N} \cdots \sum_{i_n=1}^{N} \tilde{\omega}_n^{i_n} \prod_{r=k}^{n-1} \frac{\tilde{\omega}_r^{(i_r)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(i_r)})}{\sum_{m=1}^{N} \tilde{\omega}_r^{(m)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(m)})} \\ \times \delta_{\xi_k^{(i_k)}, \xi_{k+1}^{(i_{k+1})}, \ldots, \xi_n^{(i_n)}}(x_{k:n}), \end{aligned} \tag{2.70}$$

avec $\left\{\xi_r^{(i_r)}, \tilde{\omega}_r^{(i_r)}\right\}_{i_r=1}^{N}$ les nuages de particules approximant les densités de filtrage $p(x_r|y_{1:r})$ aux instants $r = k, \ldots, n-1$.

Remarque 2.3.12. *Un premier avantage avéré du lissage joint est qu'il englobe plusieurs cas de figure de lissage. Par exemple, les densités de lissage du type $p(x_r, x_m|y_{1:n})$ avec $r \leq r, m \leq n$ sont aussi des marginales de la densité jointe de lissage. Un deuxième avantage est plutôt d'ordre théorique. Puisque les lissages qui dérivent du lissage joint héritent de ses propriétés. Cependant, tout comme le lissage marginal, le point faible du lissage joint réside dans sa complexité algorithmique qui devient très vite explosive au vu des N^{n-k+1} trajectoires de particules.*

2.4 Quelques résultats théoriques

Dans cette section, nous mettons en exergue quelques résultats de convergence pour les estimés du filtrage et du lissage joint traités plus haut. Pour cela, nous commençons par reformuler les problématiques du filtrage ainsi que du lissage dans le cadre général de distributions conditionnelles. Puis, nous exposons quelques éléments de théorie asymptotique inhérents aux approximations particulaires. Cette présentation se base entre autres sur les articles Crisan and Doucet [2000], Crisan and Doucet [2002] Chopin [2004].

2.4.1 Le Cadre

Afin de reformuler la problématique du filtrage en terme de distributions *a posteriori*, on introduit quelques notations. Étant donnés μ une mesure, f une fonction bornée K un noyau de transition et $A \in \mathcal{B}(\mathcal{X})$ on définit les opérations suivantes :

$$(\mu, f) = \int f d\mu$$

$$\mu K(A) = \int \mu(dx) K(x, A)$$

$$\mu f(x) = \int K(x, dz) f(z)$$

On se donne un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ sur lequel on considère un modèle à espace d'états général défini par la donnée d'un processus bivarié discret $\{(X, Y)\}$ à valeurs dans $\mathcal{X} \times \mathcal{Y}$. $\{X_k, k \in \mathbb{N}\}$ modélise l'état caché d'un système étudié et $\{Y_k, k \in \mathbb{N}^*\}$ les observations effectives. On munit les espaces d'états $\mathcal{X}$ et d'observations $\mathcal{Y}$ de leurs tribus boréliennes respectives $\mathcal{B}(\mathcal{X})$ et $\mathcal{B}(\mathcal{Y})$. On note par $M_F(\mathcal{X})$(resp. $\mathcal{P}(\mathcal{X})$) l'ensemble des mesures finies (resp. de probabilité) sur $\mathcal{B}(\mathcal{X})$. On munit $\mathcal{M}_F(\mathcal{X})$ (resp.

$\mathcal{P}(\mathcal{X})$) de la topologie faible, i.e la plus petite topologie rendant les applications $\mu \in \mathcal{M}_F(\mathcal{X}) \mapsto \mu(f)$ continues pour toute fonction $f \in \mathcal{C}_b(\mathcal{X})$.

2.4.2 Analyse du filtrage

On suppose que le système caché est une chaîne de Markov régie par une distribution $\pi_0(dx_0)$ sur l'état initial X_0 et une transition entre états évoluant selon la transition :

$$\mathbb{P}(X_k \in A_k | X_{0:k-1} = x_{0:k-1}, Y_{0:k-1} = y_{0:k-1}) = \int_{A_k} M_k(y_{0:k-1}, x_{0:k-1}, dx_k), \quad \forall A_k \in \mathcal{B}(\mathcal{X})$$

avec $M_k(y_{0:k-1}, \cdot, \cdot)$ un noyau de transition tel que :

$$(M_k \mu)(dx_{0:k}) = M_k(y_{0:k-1}, x_{0:k-1}, dx_k)\mu(dx_{0:k-1})$$

pour tout $\mu \in \mathcal{P}(\mathcal{X})$. Conditionnellement à $\{X_k, k \in \mathbb{N}\}$, les observations sont indépendantes et évoluent selon la dynamique

$$\mathbb{P}(Y_k \in B_k | Y_{0:k-1} = y_{0:k-1}, X_{0:k} = x_{0:k}) = \int_{B_k} g_k(y_{0:k}, x_{0:k}) dy_k, \qquad B_k \in \mathcal{B}(\mathcal{Y})$$

avec $g_k(\cdot, \cdot)$ fonction bornée. On définit les mesures de probabilités d'intérêt $(\pi_k)_{k\in\mathbb{N}}$ et $(\rho_k)_{k\in\mathbb{N}}$ par :

$$\rho_k(A_{0:k}) = \begin{cases} \mathbb{P}(X_{0:k} \in A_{0:k}) & \text{si } k = 0 \\ \mathbb{P}(X_{0:k} \in A_{0:k} | Y_{1:k-1} = y_{1:k-1}) & \text{si } k > 0 \end{cases}$$

et

$$\pi_k(A_{0:k}) = \begin{cases} \mathbb{P}(X_{0:k} \in A_{0:k}) & \text{si } k = 0 \\ \mathbb{P}(X_{0:k} \in A_{0:k} | Y_{1:k} = y_{1:k}) & \text{si } k > 0 \end{cases}$$

avec $A_i \in \mathcal{B}(\mathcal{X}), i = 1, 2, \cdots, k$.

Proposition 2.4.1. *Pour tout $k \geq 1$, les distributions de prédiction et de correction satisfont respectivement :*

$$\rho_k(A_{0:k}) = \int_{A_{0:k-1}} M(y_{1:k}, x_{k-1}, dx_k)\pi_{k-1}(dx_{0:k-1}), \tag{2.71}$$

et

$$\pi_k(A_{0:k}) = \frac{g_k(y_{1:k}, x_{0:k})\rho_k(dx_{0:k})}{C_k} \tag{2.72}$$

avec $C_k = \int_{\mathcal{X}^{k+1}} g_k(y_{1:k}, x_{0:k})\rho_k(dx_{0:k})$.

(2.71) et (2.72) définissent les étapes de prédiction et de mise à jour de la distribution jointe. Avec les notations introduites plus haut on déduit une réécriture en terme de flots suivante :

Corollaire 2.4.2. *Pour tout* $k \geq 1$, $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$ *on a :*

$$(\rho_k, \varphi_k) = (\pi_{k-1}, M_k \varphi_k) \tag{2.73}$$

et

$$(\pi_k, \varphi_k) = \frac{(\rho_k, \varphi_k g_k)}{(\rho_k, g_k)}. \tag{2.74}$$

Afin de mettre en relief l'idée d'EI, on introduit la distribution d'importance

$$\tilde{\rho}_k = \pi_{k-1} \Xi_k$$

où Ξ_k est un noyau de transition tel que ρ_k est absolument continue par rapport à $\tilde{\rho}_k$ et l'on pose

$$\tau_k := \frac{d\rho_k}{d\tilde{\rho}_k}$$

la dérivée de Radon-Nykodim avec $\tau_k(\cdot) = \tau(y_{1:k}, \pi_{k-1}, \cdot)$. De plus, à partir de Corollaire 2.4.2 on a également

$$\frac{d\pi_k}{d\tilde{\rho}_k} \propto g_k \tau_k.$$

Avec ces considérations, on peut donner l'algorithme suivant.

Algorithme 9 Filtrage particulaire

Initialisation

$k = 0$

Pour $i = 1, 2, \ldots, N$

— Tirer $\xi_0^{(i)} \sim \pi_0(dx_0)$

— Poser $k = k + 1$

FinPour

Échantillonnage préférentiel

1. Pour $i = 1, 2, \ldots, N$
   — Tirer $\tilde{\xi}_{0:k}^{(i)} \sim \Xi_k(y_{1:k}, \xi_{0:k-1}^{(i)}, \hat{\pi}_{k-1}, d\tilde{x}_{0:k})$
   — Évaluer et Normaliser les poids d'importance

$$\tilde{\omega}_k^{(i)} \propto g_k(y_{1:k}, \tilde{\xi}_{0:k}^{(i)}) \tau(y_{1:k}, \hat{\pi}_{k-1}, \tilde{\xi}_{0:k}^{(i)})$$

   — Considérer les mesures empiriques

$$\hat{\hat{\rho}}_k(dx_{0:k}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\xi}_{0:k}^{(i)}}(dx_{0:k})$$

$$\hat{\hat{\pi}}_k(dx_{0:k}) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\omega}_k^{(i)} \delta_{\tilde{\xi}_{0:k}^{(i)}}(dx_{0:k})$$

2. FinPour

Rééchantillonage

1. Pour $i = 1, 2, \ldots, N$ utiliser un mécanisme de sélection pour transformer $\left\{\tilde{\xi}_{0:k}^{(i)}, \tilde{\omega}_k^{(i)}\right\}_{i=1}^{N}$ en $\left\{\tilde{\xi}_{0:k}^{(i)}, 1/N\right\}_{i=1}^{N}$ de sorte que

$$\hat{\pi}_k(dx_{0:k}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\xi}_{0:k}^{(i)}}(dx_{0:k})$$

2. FinPour

Poser $k = k + 1$ et aller à l'étape Échantillonnage d'importance.

Remarque 2.4.3. *On peut noter que le noyau Ξ_k est assez général pour englober un bon nombre de filtres particulaires utilisés en pratique; Par exemple, en prenant*

$$\Xi_k(y_{1:k}, \tilde{\xi}_{0:k-1}^{(i)}, \hat{\pi}_{k-1}, d\tilde{x}_{0:k}) = \delta_{\xi_{0:k-1}^{(i)}}(d\tilde{x}_{0:k-1}) M_k(y_{1:k-1}, \xi_{0:k-1}^{(i)}, d\tilde{x}_k)$$

on obtient le filtre de Bootstrap, i.e $\tilde{\omega}_k^{(i)} \propto g_k(y_{1:k}, \xi_{0:k-1}^{(i)})$.

2.4.2.1 Erreur quadratique moyenne

Dans cette section, on s'intéresse à borner l'EQM. Notons que les hypothèses nécessaires pour aborder cette quantité ne sont pas très contraignantes. On trouvera le descriptif complet dans entre autres Crisan and Doucet [2000]. Nous donnons une série de 3 lemmes qui permettent d'énoncer le résultat. On note par c_k une constante dépendant du temps.

Lemme 2.4.4. *Soit* $\varphi_{k-1} \in \mathbb{F}_b(\mathcal{X}^k)$ *telle que*

$$\mathbb{E}\left[\left((\hat{\pi}_{k-1}, \varphi_{k-1}) - (\hat{\pi}_{k-1}, \varphi_{k-1})\right)^2\right] \leq c_{k-1}\frac{||\varphi_{k-1}||^2}{N}. \tag{2.75}$$

Alors il existe une constante $\tilde{c}_k$ *telle que pour tout* $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$,

$$\mathbb{E}\left[\left((\hat{\tilde{\rho}}_k, \varphi_k) - (\tilde{\rho}_k, \varphi_k)\right)^2\right] \leq \tilde{c}_k\frac{||\varphi_{k-1}||^2}{N}. \tag{2.76}$$

Lemme 2.4.5. *Soient* $\varphi_{k-1} \in \mathbb{F}_b(\mathcal{X}^k)$ *et* $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$ *telles que*

$$\mathbb{E}\left[\left((\hat{\pi}_{k-1}, \varphi_{k-1}) - (\hat{\pi}_{k-1}, \varphi_{k-1})\right)^2\right] \leq c_{k-1}\frac{||\varphi_{k-1}||^2}{N}. \tag{2.77}$$

et

$$\mathbb{E}\left[\left((\hat{\tilde{\rho}}_k, \varphi_k) - (\tilde{\rho}_k, \varphi_k)\right)^2\right] \leq \tilde{c}_k\frac{||\varphi_k||^2}{N}. \tag{2.78}$$

Alors il existe une constante $\bar{c}_k$ *telle que pour tout* $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$,

$$\mathbb{E}\left[\left((\hat{\tilde{\pi}}_k, \varphi_k) - (\tilde{\pi}_k, \varphi_k)\right)^2\right] \leq \bar{c}_k\frac{||\varphi_k||^2}{N}. \tag{2.79}$$

Lemme 2.4.6. *Soit* $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$ *telle que*

$$\mathbb{E}\left[\left((\hat{\tilde{\pi}}_k, \varphi_k) - (\tilde{\pi}_k, \varphi_k)\right)^2\right] \leq \bar{c}_k\frac{||\varphi_k||^2}{N}. \tag{2.80}$$

Alors il existe une constante c_k *telle que pour tout* $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$,

$$\mathbb{E}\left[\left((\hat{\pi}_k, \varphi_k) - (\pi_k, \varphi_k)\right)^2\right] \leq c_k\frac{||\varphi_k||^2}{N}. \tag{2.81}$$

On peut alors énoncer le résultat.

Théorème 2.4.7. *Pour tout* $k \geq 0$, *il existe une constante* c_k *telle que pour tout* $\varphi_k \in \mathbb{F}_b(\mathcal{X}^{k+1})$,

$$\mathbb{E}\left[\left((\hat{\pi}_k, \varphi_k) - (\pi_k, \varphi_k)\right)^2\right] \leq c_k\frac{||\varphi_k||^2}{N}. \tag{2.82}$$

Démonstration. pour $k = 0$, $\hat{\pi}_0$ est constituée d'un échantillon iid tiré de π_0. D'où

$$\mathbb{E}\left[\left((\hat{\pi}_0, \varphi_0) - (\pi_0, \varphi_0)\right)^2\right] \leq c_0 \frac{||\varphi_0||^2}{N}.$$

Pour $k \geq 0$, on utilise les lemmes 2.4.4, 2.4.5 et 2.4.6. □

Remarque 2.4.8. *Dans le cas du filtrage marginal, on peut obtenir une borne (constante c) indépendante du temps k. On peut consulter Del Moral and Guionnet [1998] pour plus de détails.*

2.4.2.2 Convergence presque-sure

Afin d'établir l'utilité ou non de cet algorithme, il convient de préciser la nature de la convergence à laquelle on peut s'attendre.

Définition 2.4.9. *Soit $(\mu_n)_{n\geq 1}$ une suite de mesures aléatoires de probabilité dans $\mathcal{M}_F(\mathcal{X})$ et $\mu \in \mathcal{M}_F(\mathcal{X})$ une mesure déterministe. On dit que $(\mu_n)_{n\geq 1}$ converge faiblement vers μ et on note*

$$\lim_{n\to\infty} \mu_n = \mu \tag{2.83}$$

si

$$\lim_{n\to\infty} (\mu_n, f) = (\mu, f) \tag{2.84}$$

pour toute fonction $f \in \mathcal{C}_b(\mathcal{X})$.

Ceci permet alors de définir la manière dont une mesure de probabilité est approchée au sens faible. On s'assurera que l'égalité (2.83) ait lieu $\mathbb{P} - p.s.$ pour s'assurer de la convergence presque sûre. Et pour cela, quelques hypothèses contraignantes sont nécessaires. Comme précédemment, nous ne détaillons pas ces conditions, cependant on peut retenir qu'elles sont relatives d'une part aux poids d'importance afin de s'assurer qu'ils sont bornés supérieurement. Et d'autre part que la redistribution choisie n'engendre pas trop divergence entre distribution cible et instrumentale. Sous les hypothèses 1-B, 2-B et 3-B de Crisan and Doucet [2000] on a le résultat suivant.

Théorème 2.4.10. *Pour tout $k \geq 0$,*

$$\lim_{N\to+\infty} \hat{\pi}_k = \pi_k \quad \mathbb{P}\, p.s. \tag{2.85}$$

Pour prouver cette convergence, on peut s'aider de deux résultats intermédiaires.

Lemme 2.4.11. *Pour tout $k \geq 0$, supposons que*

$$\lim_{N\to+\infty} \hat{\pi}_{k-1} = \pi_{k-1} \quad \mathbb{P}\, p.s. \tag{2.86}$$

Alors,

$$\lim_{N\to+\infty} \hat{\rho}_k = \rho_k \quad \mathbb{P}\, p.s. \tag{2.87}$$

Lemme 2.4.12. *Pour tout $k \geq 0$, supposons que*

$$\lim_{N\to+\infty} \hat{\pi}_{k-1} = \pi_{k-1} \quad \mathbb{P}\,p.s. \tag{2.88}$$

et

$$\lim_{N\to+\infty} \hat{\rho}_k = \rho_k \quad \mathbb{P}\,p.s. \tag{2.89}$$

Alors,

$$\lim_{N\to+\infty} \hat{\pi}_k = \pi_k \quad \mathbb{P}\,p.s. \tag{2.90}$$

La preuve du Théorème 2.4.10 est alors simple à établir moyennant ces deux résultats.

Démonstration. On procède par induction sur k. □

2.4.2.3 TCL

Commençons par constater les convergences presque sûre suivantes. Pour tout $k \geq 0$ et $\varphi_k \in \mathcal{C}_b(\mathcal{X})$,

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \varphi_k(\tilde{\xi}_{0:k}^{(i)}) &\xrightarrow{ps} (\tilde{\rho}_k, \varphi_k) \\ \frac{1}{N}\sum_{i=1}^{N} \tilde{\omega}_k^{(i)} \varphi_k(\tilde{\xi}_{0:k}^{(i)}) &\xrightarrow{ps} (\pi_k, \varphi_k) \\ \frac{1}{N}\sum_{i=1}^{N} \varphi_k(\xi_{0:k}^{(i)}) &\xrightarrow{ps} (\pi_k, \varphi_k) \end{aligned} \tag{2.91}$$

à mesure que $N \to \infty$. Définissons les variances asymptotiques au travers de relations récursives suivantes :

$$\begin{aligned} \tilde{V}_k(\varphi_k) &= \hat{V}_k(\Xi_{k-1}\varphi_k) + (\phi_{k-1}, Var_{\Xi_k}(\varphi_k)) \\ V_k(\varphi_k) &= \tilde{V}_k(C_k^{-1} g_k \tau_k [\varphi_k - (\pi_k, \varphi_k)]) \\ \hat{V}_k(\varphi_k) &= V_k(\varphi_k) + Var_{\pi_k}(\varphi_k). \end{aligned}$$

Tout comme les démarches antérieures, on commence par énoncer quelques lemmes qui permettent de donner le résultat.

Lemme 2.4.13. *Soit $\varphi_{k-1} \in \mathcal{C}_b(\mathcal{X}^k)$ telle que :*

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N} \varphi_{k-1}(\xi_{0:k-1}^{(i)}) - (\pi_{k-1}, \varphi_{k-1})\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \hat{V}_{k-1}(\varphi_{k-1})). \tag{2.92}$$

Alors pour tout $\varphi_k \in \mathcal{C}_b(\mathcal{X}^{k+1})$,

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_k(\tilde{\xi}_{0:k}^{(i)}) - (\tilde{\rho}_k, \varphi_k)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{V}_k(\varphi_k)). \tag{2.93}$$

Lemme 2.4.14. *Soit* $\varphi_k \in \mathcal{C}_b(\mathcal{X}^{k+1})$ *telle que :*

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_k(\tilde{\xi}_{0:k}^{(i)}) - (\tilde{\rho}_k, \varphi_k)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{V}_k(\varphi_k)). \tag{2.94}$$

Alors pour tout $\varphi_k \in \mathcal{C}_b(\mathcal{X}^{k+1})$,

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\omega}_k^{(i)}\varphi_k(\tilde{\xi}_{0:k}^{(i)}) - (\pi_k, \varphi_k)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_k(\varphi_k)). \tag{2.95}$$

Lemme 2.4.15. *Supposons que* $\hat{V}_k(\varphi_k) = V_k(\varphi_k) + Var_{\pi_k}(\varphi_k)$ *et que pour tout* $\varphi_k \in \mathcal{C}_b(\mathcal{X}^{k+1})$

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\omega}_k^{(i)}\varphi_k(\tilde{\xi}_{0:k}^{(i)}) - (\pi_k, \varphi_k)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_k(\varphi_k)). \tag{2.96}$$

Alors pour une redistribution multinomiale on a :

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_k(\xi_{0:k}^{(i)}) - (\pi_k, \varphi_k)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \hat{V}_k(\varphi_k)). \tag{2.97}$$

On peut alors donner le résultat.

Théorème 2.4.16. *Supposons* $g_k\tau_k$ *continue et bornée et que la redistribution multinomiale est utilisée. Alors pour tout* $\varphi_k \in \mathcal{C}_b(\mathcal{X}^{k+1})$, (π_k, φ_k) *et* $\hat{V}_k(\varphi_k)$ *sont finis de plus*

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_k(\xi_{0:k}^{(i)}) - (\pi_k, \varphi_k)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \hat{V}_k(\varphi_k)). \tag{2.98}$$

Démonstration. $k = 0$, $\left\{\xi_0^{(i)}\right\}_{i=1}^{N}$ est un nuage iid de loi π_0. Avec la donnée de $\varphi_0 \in \mathcal{C}_b(\mathcal{X})$ on a

$$N^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi_k(\xi_0^{(i)}) - (\pi_0, \varphi_0)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \hat{V}_k(\varphi_0)).$$

Pour $k \geq 0$, on procède par induction moyennant les lemmes 2.4.13, 2.4.14 et 2.4.15. □

2.4.3 Analyse du lissage

Cette section est basée entre autres sur Douc et al. [2011], Nous commençons par reformuler le lissage en terme de distribution conditionnelle pour en déduire quelques éléments de convergence des contre-parties empiriques.

2.4.3.1 Formulation

Soit μ une mesure de probabilité sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. On note par B_μ le noyau de lissage *Backward* défini sur $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ par

$$B_\mu(x, h) := \frac{\int_{\mathcal{X}} \mu(dx')M(x', x)h(x)}{\int_{\mathcal{X}} \mu(dx')M(x', x)}$$

pour tous $x, x' \in \mathcal{X}$, h une fonction mesurable bornée et M un noyau de transition de transition de $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ vers $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. La motivation du noyau B_μ tient au fait que lorsqu'on renverse l'axe du temps, $\{X_k, k \in \mathbb{N}\}$ reste markovien.

Proposition 2.4.17. *$\{X_{n-k}, k \in \mathbb{N}\}$ est une chaîne de Markov non homogène de noyau de transition B_μ de $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ dans $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ satisfaisant pour tout $f \in \mathbb{F}_b(\mathcal{X})$,*

$$B_\mu(X_{k+1}, f) := \mathbb{E}_\mu\left[f(X_k)|X_{k+1:n}, Y_{0:n}\right] = \mathbb{E}_\mu\left[f(X_k)|X_{k+1:n}, Y_{0:n}\right] \tag{2.99}$$

avec $\mathbb{E}_\mu$ mettant en exergue μ comme loi initiale.

Démonstration. voir, Cappé et al. [2005]. □

On introduit la distribution de lissage joint

$$\pi_{k:n|n}(dx_{k:n}) := \mathbb{P}(X_{k:n} \in dx_{k:n}|Y_{1:n} = y_{1:n}), k < n.$$

De cette définition et par la règle de Bayes on déduit la relation de récurrence suivante.

Proposition 2.4.18. *Pour toute fonction $h \in \mathbb{F}_b(\mathcal{X}^{n-k+1})$,*

$$\pi_{k:n|n}(h) = \int_{\mathcal{X}^{n-k+1}} B_{\pi_k}(x_{k+1}, dx_k)\pi_{k+1:n|n}(dx_{k+1:n})h(x_{k:n}) \tag{2.100}$$

et par itération sur k, on obtient la factorisation

$$\pi_{k:n|n}(h) = \int_{\mathcal{X}^{n-k+1}} h(x_{k:n})\pi_n(dx_n)\prod_{r=k}^{n-1} B_{\pi_r}(x_{r+1}, dx_r). \tag{2.101}$$

On déduit aisément une relation analogue pour la distribution marginale de lissage.

Corollaire 2.4.19. *Pour toute fonction* $h \in \mathbb{F}_b(\mathcal{X})$,

$$\pi_{k|n}(h) = \int_{\mathcal{X}^2} B_{\pi_k}(x_{k+1}, dx_k)\pi_{k+1|n}(dx_{k+1})h(x_k). \tag{2.102}$$

Afin d'approximer les distributions de lissage on se donne d'une part les nuages de points $\left\{\xi_r^{(i_r)}\tilde{\omega}_r^{(i_r)}\right\}_{i_r=1}^N$ approximant π_r dans la phase *Forward* par :

$$\tilde{\pi}_r(dx_r) = \sum_{i_r=1}^N \tilde{\omega}_r^{(i_r)}\delta_{\xi_r^{(i_r)}}(dx_r)$$

pour $r = 1, 2, \ldots, n$.

2.4.3.2 Inégalités de déviation

On peut obtenir des inégalités de déviations exponentielles sur un horizon temporel n fini ou non. On se place dans le cas $n < \infty$ et on suppose les fonctions d'importance $\omega_k(\cdot)$ ainsi et les vraisemblances marginales $g_k(\cdot)$ sont uniformément bornées avec $g_k > 0$ [6] alors :

Théorème 2.4.20. *Pour tout* $h \in \mathbb{F}_b(\mathcal{X}^{n+1})$, *il existe des constantes* $0 < b_n, c_n < \infty$ *telles que pour tous* $\epsilon, N > 0$,

$$\mathbb{P}\left(|\hat{\pi}_{0:n|n}(h) - \pi_{0:n|n}(h)| \geq \epsilon\right) \leq b_n \exp\left(-\frac{c_n N\epsilon^2}{osc^2(h)}\right). \tag{2.103}$$

Avant de donner les grandes lignes de cette preuve [7] nous commençons par introduire quelques noyaux permettant une réécriture en somme de termes télescopiques bornés de l'erreur d'estimation. Pour $k \geq 0$, définissons le noyau $\mathbb{L}_{k,n} : \mathcal{X}^{k+1} \times \mathcal{B}(\mathcal{X})^{\otimes(n+1)} \to [0, 1]$ tel que :

$$\mathbb{L}_{k,n}(x_{0:k}, h) := \int_{\mathcal{X}^{n-k}} \prod_{u=k+1}^n M(x_{u-1}, dx_u)g_u(x_u)h(x_{0:n}). \tag{2.104}$$

6. Hypothèses $A1 - A2$ dans Douc et al. [2011]
7. On pourra se référer à Douc et al. [2011] pour exposé détaillé de ce résultat

avec $\mathbb{L}_{n,n}(x_{0:n}, h) := h(x_{0:n})$, où $h \in \mathbb{F}_b(\mathcal{X}^{n+1})$. En utilisant (2.104), pour tout $0 \leq k \leq n$

$$\begin{aligned}\hat{\pi}_{0:k|k}\left[\mathbb{L}_{k,n}(x_{0:k}, h)\right] &= \int_{\mathcal{X}^{k+1}} \hat{\pi}_{0:k|k}(dx_{0:k})\mathbb{L}_{k,n}(x_{0:k}, h) \\ &= \int_{\mathcal{X}^{k+1}} \hat{\pi}_k(dx_k) \\ &\times \prod_{r=0}^{k-1} \hat{B}_r(x_{r+1}, dx_r)\mathbb{L}_{k,n}(x_{0:k}, h) \\ &= \int_{\mathcal{X}^l} \hat{\pi}_k(dx_k)\hat{\mathfrak{L}}_{k,n}(x_k, h)\end{aligned} \tag{2.105}$$

où l'on définit les noyaux $\hat{\mathfrak{L}}_{k,n}$ et $\mathfrak{L}_{k,n}$ sur $\mathcal{X} \times \mathcal{B}(\mathcal{X})^{\otimes(n+1)}$ respectivement par :

$$\hat{\mathfrak{L}}_{k,n}(x_k, h) := \int_{\mathcal{X}^{k+1}} \prod_{r=0}^{k-1} \hat{B}_r(x_{r+1}, dx_r)\mathbb{L}_{k,n}(x_{0:k}, h) \tag{2.106}$$

et

$$\mathfrak{L}_{k,n}(x_k, h) := \int_{\mathcal{X}^{k+1}} \prod_{r=0}^{k-1} B_r(x_{r+1}, dx_r)\mathbb{L}_{k,n}(x_{0:k}, h) \tag{2.107}$$

pour tout $x_k \in \mathcal{X}$. Par application de la règle de Bayes on a :

$$\pi_{0:n|n}(h) = \frac{\int_{\mathcal{X}^{n+1}} \nu(dx_0) \prod_{u=0}^{n} M(x_{u-1}, dx_u) g_u(x_u) h(x_{0:n}) dx_{0:n}}{\int_{\mathcal{X}^{n+1}} \nu(dx_0) \prod_{u=0}^{n} M(x_{u-1}, dx_u) g_u(x_u) dx_{0:n}} \tag{2.108}$$

si bien que

$$\pi_{0:n|n}(h) = \frac{\pi_{0:k|k}\left[\mathbb{L}_{k,n}(., h)\right]}{\pi_{0:k|k}\left[\mathbb{L}_{k,n}(., 1)\right]}, \quad \forall k \geq 0.. \tag{2.109}$$

En fin, nous introduisons $\hat{G}_{k,n}$ un noyau sur $\mathcal{X} \times \mathcal{B}(\mathcal{X})^{\otimes(n+1)}$ défini par

$$\hat{G}_{k,n}(x, h) := \hat{\mathfrak{L}}_{k,n}(x, h) - \frac{\hat{\phi}_{k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot, h)\right]}{\hat{\pi}_{k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot, 1)\right]} \hat{\mathfrak{L}}_{k,n}(x, 1) \tag{2.110}$$

pou tout $x \in \mathcal{X}$, $h \in \mathbb{F}_b(\mathcal{X}^{n+1})$.

Démonstration. La preuve est alors faite par induction sur $n \geq 0$ en utilisant la dé-

composition successive de l'erreur de lissage suivante :

$$\begin{aligned}
\Delta_n^N(h) &= \hat{\pi}_{0:n|n}(h) - \pi_{0:n|n}(h) \\
&= \sum_{k=0}^{n} \left\{ \frac{\hat{\pi}_{0:k|k}\left[\mathbb{L}_{k,n}(\cdot,h)\right]}{\hat{\pi}_{0:k|k}\left[\mathbb{L}_{k,n}(\cdot,1)\right]} - \frac{\hat{\pi}_{0:k-1|k-1}\left[\mathbb{L}_{k-1,n}(\cdot,h)\right]}{\hat{\pi}_{0:k-1|k-1}\left[\mathbb{L}_{k-1,n}(\cdot,1)\right]} \right\} \\
&= \sum_{k=0}^{n} \left\{ \frac{\hat{\pi}_{0:k|k}\left[\hat{\mathfrak{L}}_{k,n}(\cdot,h)\right]}{\hat{\pi}_{0:k|k}\left[\hat{\mathfrak{L}}_{k,n}(\cdot,1)\right]} - \frac{\hat{\pi}_{0:k-1|k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot,h)\right]}{\hat{\pi}_{0:k-1|k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot,1)\right]} \right\} \\
&= \sum_{k=0}^{n} \frac{N^{-1}\sum_{i=1}^{N} \omega_k^{(i)} \hat{G}_{k,n}(\xi_k^{(i)},h)}{N^{-1}\sum_{i=1}^{N} \omega_k^{(i)} \mathfrak{L}_{k,n}(\xi_k^{(i)},1)}
\end{aligned} \tag{2.111}$$

avec les conventions

$$\frac{\hat{\pi}_{0:-1|-1}\left[\mathbb{L}_{-1,n}(.,h)\right]}{\hat{\pi}_{0:-1|-1}\left[\mathbb{L}_{-1,n}(.,1)\right]} = \frac{\pi_0\left[\mathbb{L}_{0,n}(.,h)\right]}{\pi_0\left[\mathbb{L}_{0,n}(.,1)\right]} = \pi_{0:n|n}(h)$$

et $\{\omega_k^{(i)}, \xi_k^{(i)}\}_{i=1}^N$ un nuage de particules pondéré approximant π_k. Le reste de la preuve consiste alors à vérifier les 3 points du LemmeA.1.5 de Hoeffding généralisé à cette décomposition et ce à chaque étape de la récurrence pour conclure. □

Lorsque l'horizon temporel n n'est pas borné, on peut obtenir une inégalité exponentielle analogue Théorème 2.4.20 avec une borne indépendante de n.

2.4.3.3 Erreur L_q

La borne de déviation établie précédemment permettent d'établir une convergence dans l'espace $L_q, q \geq 2$. Pour une variable aléatoire X, on définit par $\| X \|_q$ la norme L_q de X.

Théorème 2.4.21. *Il existe une constante $C(n,N)$ telle que l'erreur L_q de l'algorithme de lissage satisfait*

$$\| \hat{\pi}_{0:n|n}(h) - \pi_{0:n|n}(h) \|_q \leq C(n,N) \times osc(h) \tag{2.112}$$

pour tout $h \in \mathbb{F}_b(\mathcal{X}^{n+1})$.

Notons que ce résultat est une reformulation du Th.1 dans Dubarry and Le Corff [2010] pour une fonction $h \in \mathbb{F}_b(\mathcal{X}^{n+1})$ quelconque au lieu d'une forme additive pour h. Leur preuve peut également s'appliquer à quelques détails près sur la constante $C(n,N)$.

2.4.3.4 TCL

On peut également obtenir une convergence en loi de l'estimateur de la distribution jointe de lissage afin de disposer d'intervalle de confiance pour ce dernier. En effet,

Théorème 2.4.22. *Pour tout* $h \in \mathbb{F}_b(\mathcal{X}^{n+1})$,

$$\sqrt{N}\left(\hat{\pi}_{0:n|n}(h) - \pi_{0:n|n}(h)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_{0:n|n}(h)) \tag{2.113}$$

à mesure que N *tend vers l'infini, avec* $\Gamma_{0:n|n}$ *la variance asymptotique définie par récurrence sur* n.

La preuve de ce résultat utilise essentiellement le Lemme de Slutsky en se basant sur une décomposition de l'erreur de lissage. On pourra consulter Douc et al. [2011], Th. 8 pour les détails de ce calcul.

2.5 Conclusion

Dans ce chapitre, nous avons parcouru quelques unes des méthodes de Monte Carlo séquentielles appliquées à des processus discrets et notamment aux Chaînes de Markov cachées. Nous nous somme intéressé au filtrage et lissage particulaires lorsque les hypothèses de linéarité et de gaussianité sont prises au dépourvues dans la représentation à espace d'états de ces modèles. Nous avons ponctué cette présentation de quelques résultats théoriques qui sous-tendent ces méthodes. Cependant, il clair que cette présentation est loin d'être complète. Par exemple, nous n'avons pas abordé dans le cadre du filtrage particulaire le filtre *auxiliaire* de Pitt and Shephard [1999] qui peut être vu comme une généralisation du filtre de *Bootstrap*. Dans le cadre du lissage, nous n'avons abordé qu'un type de lissage : le lissage à point fixe. Par ailleurs, nous n'avons parlé de filtre prédictive. Il faut cependant noté que sur ce dernier point les récursions sont analogues à celles du filtrage ou du lissage. En fin, concernant l'aspect théorique, remarquons que seules des fonctionnelles bornées ont été considérées. Des résultats de convergence dans le cadre du filtrage auraient pu être abordées pour une classe de fonctionnelles non bornées (voir Hu and Schön [2011]).

CHAPITRE 3

Calibration d'un modèle de volatilité canonique

Sommaire

3.1 Genèse du modèle de volatilité 68
3.2 Données synthétiques 70
3.2.1 Jeu de données 1 70
3.2.2 jeu de données 2 71
3.3 Données réelles 73
3.3.1 Taux de change GBP/USD 73
3.3.2 Taux de change USD/YEN 74
3.3.3 Taux de change EURO/USD 74
3.3.4 Indice S&P 500 75
3.3.5 Indice Dow Jones 76
3.3.6 Indice FTSE 100 77
3.3.7 Indice Nikkei 225 78
3.4 Extension du modèle canonique 83
3.5 Conclusion 91

Nous donnons une application à la calibration d'un modèle de volatilité stochastique discret aux données de taux de change ainsi que des indices boursiers. Nous nous intéressons aux taux GBP/USD, YEN/USD et EURO/USD. Puis, nous étendons cette étude aux indices S&P 500, Dow Jones et NIKKEI 225. Pour se faire, nous avons adopté le plan suivant. Dans un premier temps nous commençons par utiliser des données simulées afin de jauger la méthode d'estimation. L'avantage dans ce cas de figure est de pouvoir contrôler les paramètres restitués à l'issu de la procédure d'estimation.

Dans un deuxième temps, nous considérons la calibration de ce modèle sur des jeux de données réelles et éventuellement faire une comparaison avec des estimations obtenues de certains auteurs. En dernier lieu, nous nous départissons du modèle basique de volatilité afin de mettre la méthode d'estimation à l'épreuve.

3.1 Genèse du modèle de volatilité

On considère un espace probabilisé filtré. Sur ce dernier on se donne un actif financier risqué dont la dynamique de prix suit un brownien géométrique standard :

$$\frac{dS_t}{S_t} = \varpi dt + \sigma dW_t \tag{3.1}$$

ϖ étant le *drift* ou la moyenne, σ est la volatilité qui est constante $(W_t)_{t\geq 0}$ est un mouvement brownien standard. Une application du théorème d'Itô fournit la solution de cette équation différentielle stochastique donnée par :

$$S_t = S_0 e^{(\varpi - \frac{1}{2}\sigma^2)t + \sigma W_t}. \tag{3.2}$$

Ce modèle a montré ses limites entre autres dans la capture des faits dits *stylisés*. Notamment, la non constance de la volatilité est un point crucial dans la modélisation financière. Parmi les diverses extensions on peut citer celles qui confèrent une dynamique propre à la volatilité (Modèle ARCH, GARCH, Heston etc.). Singulièrement, le point de vue exploré est celui qui confert à la volatilité une dynamique de processus latent donc inobservable ou partiellement observable. Au quel cas, il faut lui associer un processus d'observation servant à quantifier ses réalisations. En adoptant un processus exponentiel Orstein-Uhlenbeck pour la volatilité, le modèle précédent se réécrit :

$$\begin{cases} \frac{dS_t}{S_t} = \varpi dt + \sigma_t dW_t \\ \log \sigma_t^2 = U_t \\ dU_t = \gamma(\delta - U_t)dt + \zeta dW_t^* \end{cases} \tag{3.3}$$

avec $(W_t^*)_{t\geq 0}$ un mouvement brownien pouvant être corrélé avec $(W_t)_{t\geq 0}$, δ la moyenne à long terme, γ la vitesse de retour à la moyenne assurant également la stationnarité de U lorsque $|\gamma| < 1$ et $\zeta > 0$ un terme de variance appelé aussi volatilité de la volatilité. Par souci de simplicité, on impose une corrélation nulle entre les deux browniens. Afin de pouvoir calibrer le modèle continu (3.3) aux données réelles, on utilise la discrétisation d'Euler suivante. Posons $\Delta := t_{k+1} - t_k > 0$ le pas de discrétisation pour deux instants consécutifs discrets indexés par $k \in \mathbb{N}$. (3.3) s'écrit alors :

$$\begin{cases} \frac{S_{t_{k+1}} - S_{t_k}}{S_{t_k}} = \varpi\Delta + \sigma_{t_k}\sqrt{\Delta}V_{t_k} \\ \log \sigma_{t_k}^2 = U_{t_k} \\ U_{t_{k+1}} - U_{t_k} = \gamma(\delta - U_{t_k})\Delta + \sqrt{\Delta}\zeta W_{t_k} \end{cases} \tag{3.4}$$

ce qui est équivalent à :

$$\begin{cases} \frac{1}{\sqrt{\Delta}}\left(\frac{S_{t_{k+1}}-S_{t_k}}{S_{t_k}} - \varpi\Delta\right) = \sigma_{t_k} V_{t_k} \\ \mid \sigma_{t_k} \mid = e^{U_{t_k}/2} \\ U_{t_{k+1}} - \delta = (1-\gamma\Delta)(U_{t_k}-\delta) + \sqrt{\Delta}\zeta W_{t_k}. \end{cases} \tag{3.5}$$

Définissons les processus discrets $\left\{ Y_{t_k} := \frac{1}{\sqrt{\Delta}}\left(\frac{S_{t_{k+1}}-S_{t_k}}{S_{t_k}} - \varpi\Delta\right), k \in \mathbb{N} \right\}$ et $\{X_{t_k} := (U_{t_k} - \delta), k \in \mathbb{N}\}$. En posant $\beta := e^{\delta/2}$, $\sigma := \sqrt{\Delta}\zeta$ et $\alpha := (1 - \gamma\sqrt{\Delta})$ avec la convention $\sigma_{t_k} \geq 0$ pour tout k, on obtient

$$\begin{cases} X_{t_k} = \alpha X_{t_{k-1}} + \sigma W_{t_k} \\ Y_{t_k} = \beta e^{X_{t_k}/2} V_{t_k} \end{cases} \tag{3.6}$$

En prenant le pas $\Delta = 1$, on obtient le modèle discret suivant

$$\begin{cases} X_k = \alpha X_{k-1} + \sigma W_k \\ Y_k = \beta e^{X_k/2} V_k, \quad k \geq 1 \end{cases} \tag{3.7}$$

avec (V_k) et (W_k) des bruits indépendants gaussiens et indépendants de $X_0 \sim \mathcal{N}(0, \sigma_0^2)$, $|\alpha| < 1$. Le vecteur de paramètres étant $\theta = (\alpha, \beta, \sigma)$. Notons que dans cette version discrétisée de (3.3), $\{Y_k\}$ est le processus des rendements issus des variations du prix du sous-jacent donc observable et $\{X_k\}$ est le processus de log-volatilité centré et qui demeure inobservé. Remarquons enfin, qu'il est d'usage de linéariser (3.7) afin d'obtenir un modèle à espace d'état linéaire et quasi-gaussien suivant :

$$\begin{cases} X_k = \alpha X_{k-1} + \sigma W_k \\ \log Y_k^2 = \log \beta^2 + m + X_k + \log V_k^2 - m \end{cases} \tag{3.8}$$

où $m := \mathbb{E}(\log V_k^2) = -1.27049$ et $\log V_k^2$ suit une loi $log\chi^2$. [1] Sous ce modèle linéarisé, la quantité intermédiaire donnée à une constante près indépendante de k s'écrit :

$$\begin{aligned} Q(\theta^{(k)}, \theta^{(k-1)}) &:= \mathbb{E}_{\theta^{(k)}}\left[\log p_{\theta^{(k-1)}}(X_{0:n}, Y_{1:n}) \mid Y_{1:n}\right] \approx -\frac{n}{2}\log[\sigma^{(k)}]^2 \\ &- \frac{1}{2}\left\{ \frac{\sum_{r=1}^{n} \mathbb{E}_{\theta^{(k)}}\left[(X_r - \alpha^{(k)} X_{r-1})^2 | Y_{1:n}\right]}{[\sigma^{(k)}]^2} \right\} \\ &- \frac{1}{2}\left\{ \sum_{r=1}^{n} \mathbb{E}_{\theta^{(k)}}\left[\exp(X_r - Y_r - \log[\beta^{(k)}]^2 + m) - (X_r - Y_r - \log[\beta^{(k)}]^2 + m) | Y_{1:n}\right] \right\} \end{aligned}$$

1. Cette loi est bien approximée par celle d'un mélange de 7 lois gaussiennes (voir Kim et al. [1998]).

Par dérivation de la quantité intermédiaire par rapport aux différents paramètres, le mécanisme de mise à jour des paramètres obéit alors au schéma de récurrence suivant :

$$\begin{cases} \alpha^{(k+1)} = \frac{\sum_{r=1}^{n} \mathbb{E}_{\theta^{(k)}}[X_{r-1}X_r|Y_{1:n}]}{\sum_{r=1}^{n} \mathbb{E}_{\theta^{(k)}}\left[X_{r-1}^2|Y_{1:n}\right]} \\ \log[\beta^{(k+1)}]^2 + m = \log\left[\frac{1}{n}\sum_{r=1}^{n} \mathbb{E}_{\theta^{(k)}}[\exp(y_r - X_r + m)|Y_{1:n}]\right] \\ \sigma^{(k+1)} = \sqrt{\frac{1}{n}\sum_{r=1}^{n}\left(\mathbb{E}_{\theta^{(k)}}\left[X_r|Y_{1:n}\right] - \alpha^{(k+1)}\,\mathbb{E}_{\theta^{(k)}}\left[X_{r-1}|Y_{1:n}\right]\right)^2} \end{cases} \tag{3.9}$$

Nous donnons à présent quelques illustrations numériques au travers de séries de données simulées dans un premier temps, puis réelles dans un second temps.

3.2 Données synthétiques

3.2.1 Jeu de données 1

On produit une première trajectoire issue du modèle (3.8) de $T = 500$ observations. Ces données ont été générées sous le vecteur de paramètres $\theta^* = (0.9, \sqrt{0.1}, -0.8612)$ avec $\alpha^* = 0.9$, $\sigma^* = \sqrt{0.1}$ et $\log(\beta^*)^2 = -0.8612$. La procédure du MCEM est lancée avec les paramètres d'initialisation $(\alpha^{(0)}, \log(\beta^{(0)})^2, \sigma^{(0)}) = (0.6, -0.3, \sqrt{0.3})$. A l'issue de la procédure du MCEM, nous disposons de l'évolution des différents paramètres au travers des 500 itérations. Le tracé joint de la trajectoire des séries ainsi que les différentes itérations du MCEM est donnée à la Figure 3.1. On peut aussi s'intéresser au risque quadratique moyen pour les trois paramètres estimés à l'issue des itérations du MCEM. Le tableau suivant est un résumé de cette grandeur.

Estimateurs	$\hat{\alpha}$	$\log[\hat{\beta}]^2$	$[\hat{\sigma}]^2$
REQM	0.0251	0.0655	0.0489

TABLE 3.1 – Racine carrée de l'erreur quadratique moyenne sur les 500 itérations

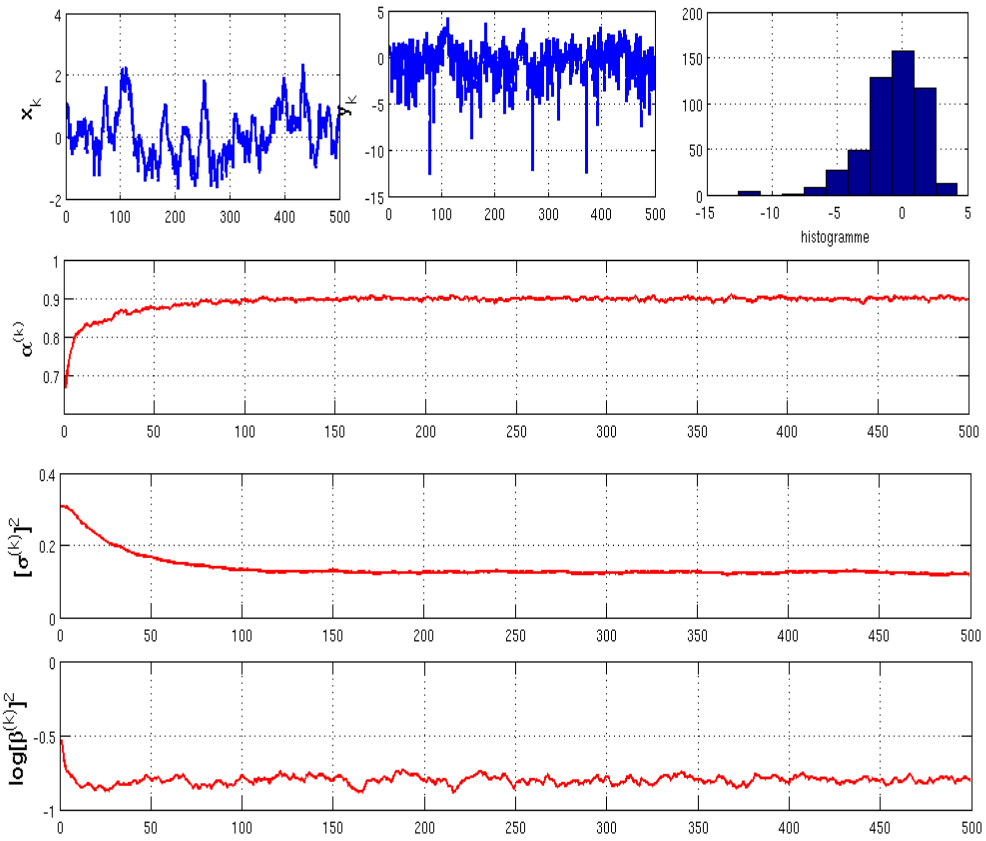


FIGURE 3.1 – Trajectoire de (3.7) sur un horizon de $T = 500$ réalisations & Itérations du MCEM pour $N = 200$ particules.

3.2.2 jeu de données 2

Nous donnons un deuxième jeux de données avec cette fois un horizon temporel plus long $T = 4000$ avec comme paramètres $\theta^* = (0.92, \sqrt{0.4}, -0.7)$ où $\alpha^* = 0.92$, $\sigma^* = \sqrt{0.4}$ et $\log(\beta^*)^2 = -0.7$. Le nombre de particules reste inchangé ($N = 200$). Une conclusion analogue à celle obtenue plus haut peut être déduite. A la Figure 3.2, nous avons un tracé conjoint de la trajectoire des séries ainsi que les itérations du MCEM.

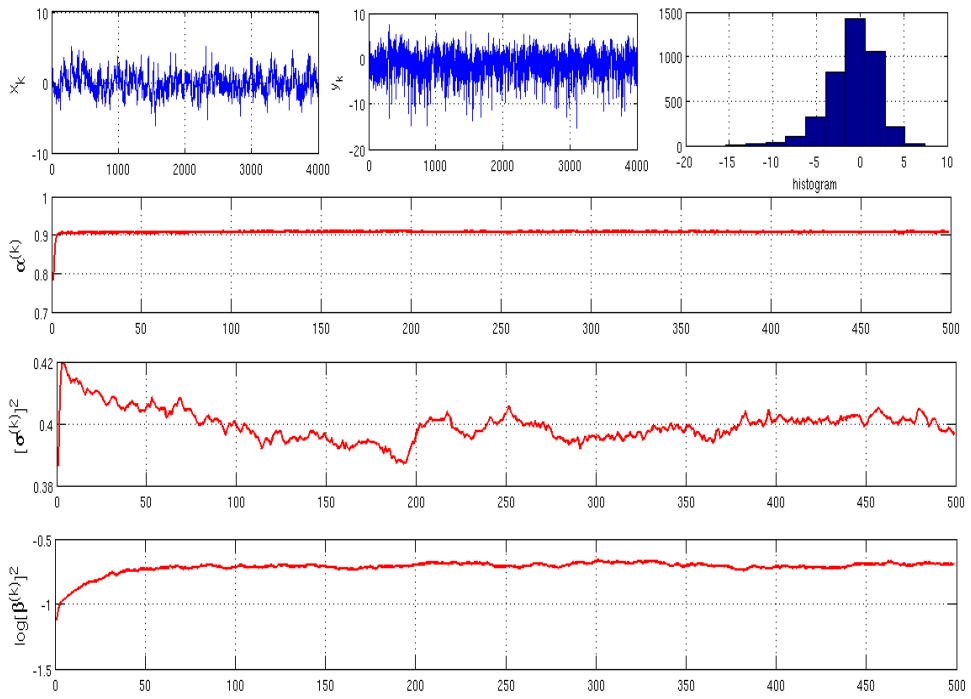


FIGURE 3.2 – Trajectoire de (3.7) sur un horizon de $T = 4000$ réalisations & Itérations du MCEM pour $N = 200$ particules.

On peut également établir le tableau du REQM relativement aux estimés.

Estimateurs	$\hat{\alpha}$	$\log[\hat{\beta}]^2$	$[\hat{\sigma}]^2$
REQM	0.0139	0.0050	0.0519

TABLE 3.2 – Racine carrée de l'erreur quadratique moyenne sur les 500 itérations

Remarque 3.2.1. *Comme on peut le constater aux Figure 3.1 et Figure 3.2, les itérations peuvent être arrêtées plus tôt étant donné le caractère relativement coûteux du temps de calculs*[2]*. En utilisant un critère heuristique d'arrêt comme une variation relative des paramètres seuillée ou tout simplement en ayant recours au rapport de vraisemblance comme dans Chan and Ledolter [1995] ou Kim and Stoffer [2006]. Par ailleurs, dans un souci d'harmonie et de concision le nombre d'itérations du MCEM a été arbitrairement fixé à 500 aussi bien pour les données réelles que synthétiques.*

2. Tous les calculs ont été effectués avec des PC Intel Core 2 Duo CPU 2.20 GHz

3.3 Données réelles

Dans la suite et sauf mention contraire, nous maintenons $N = 300$ particules et $\theta^{(0)} = (0.2, 0.1, -0.01)$ comme paramètres d'initialisation du MCEM avec $\alpha^{(0)} = 0.2$, $\sigma^{(0)} = \sqrt{0.1}$ et $2 \times \log(\beta^{(0)}) = -0.01$.

3.3.1 Taux de change GBP/USD

Nous disposons de l'historique du taux de change journalier GBP/USD disponible gratuitement sur le site de la *Federal Reserve System*. Nous avons retenu la période du 1 octobre 1981 au 28 juin 1985. En effet, cette série a entre autres été étudiée dans Harvey et al. [1994], Durban and Koopman [2000], Doucet and Tadic [2003]. A la Figure 3.3, nous avons représenté le taux de change GBP/USD ainsi que les paramètres estimés par calibration du modèle (3.8). De (a) à (c) on a respectivement le taux de change journalier noté p_k, le logarithme du carré de son rendement corrigé[3] de sa moyenne noté y_k et l'histogramme de ce dernier.

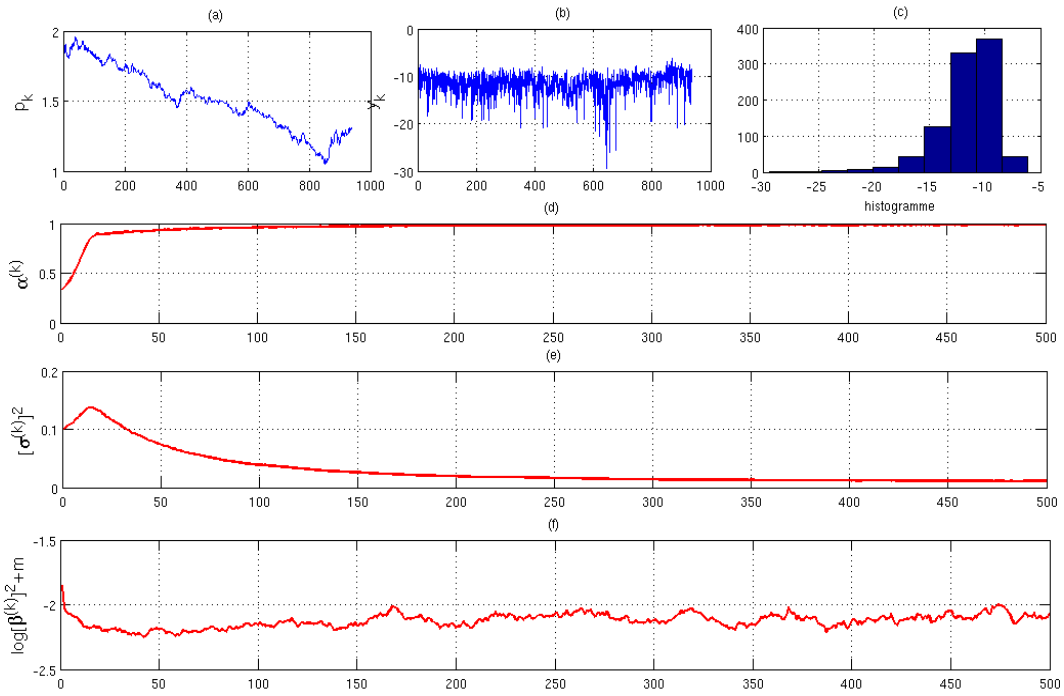


FIGURE 3.3 – Analyse du taux de change GBP/USD.

3. comme suggéré dans Harvey et al. [1994]

3.3.2 Taux de change USD/YEN

Nous avons également recueilli le taux de change journalier USD/YEN pour la période allant du 31 mai 2005 au 1 juin 2012. A la Figure 3.4, nous avons résumé quelque caractéristique de ce dernier. Au tracé (a), nous avons noté par p_t le taux journalier. (b) est celui du logarithme du carré des rendements corrigé de leur moyenne et noté y_t et en (c) l'histogramme de ce dernier. Les tracés (d), (e) et (f) représentent les trajectoires des paramètres estimés à l'issu de 500 itérations du MCEM.

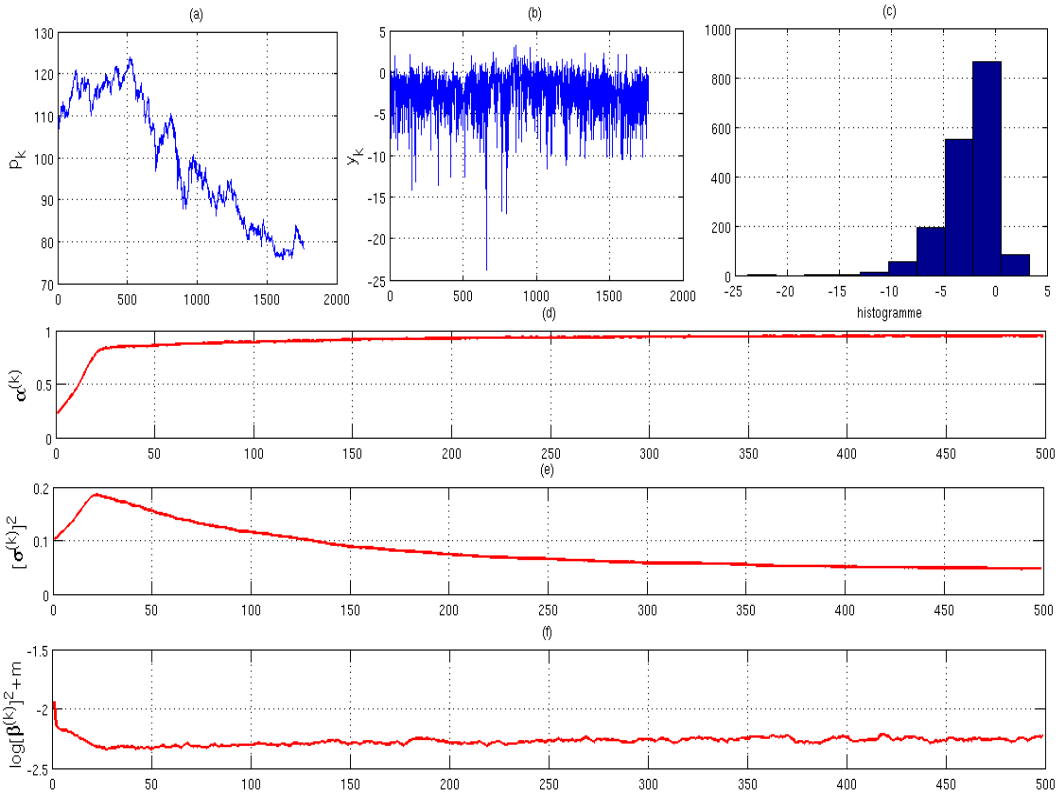


FIGURE 3.4 – Analyse du taux de change USD/YEN.

3.3.3 Taux de change EURO/USD

La dernière série de taux de change concerne EURO/USD. Celle-ci consiste en 3375 observations couvrant la période allant du 04 janvier 1999 au 01 juin 2012. Nous avons effectué une analyse analogue au taux USD/YEN dont le résumé est consigné à la Figure 3.5.

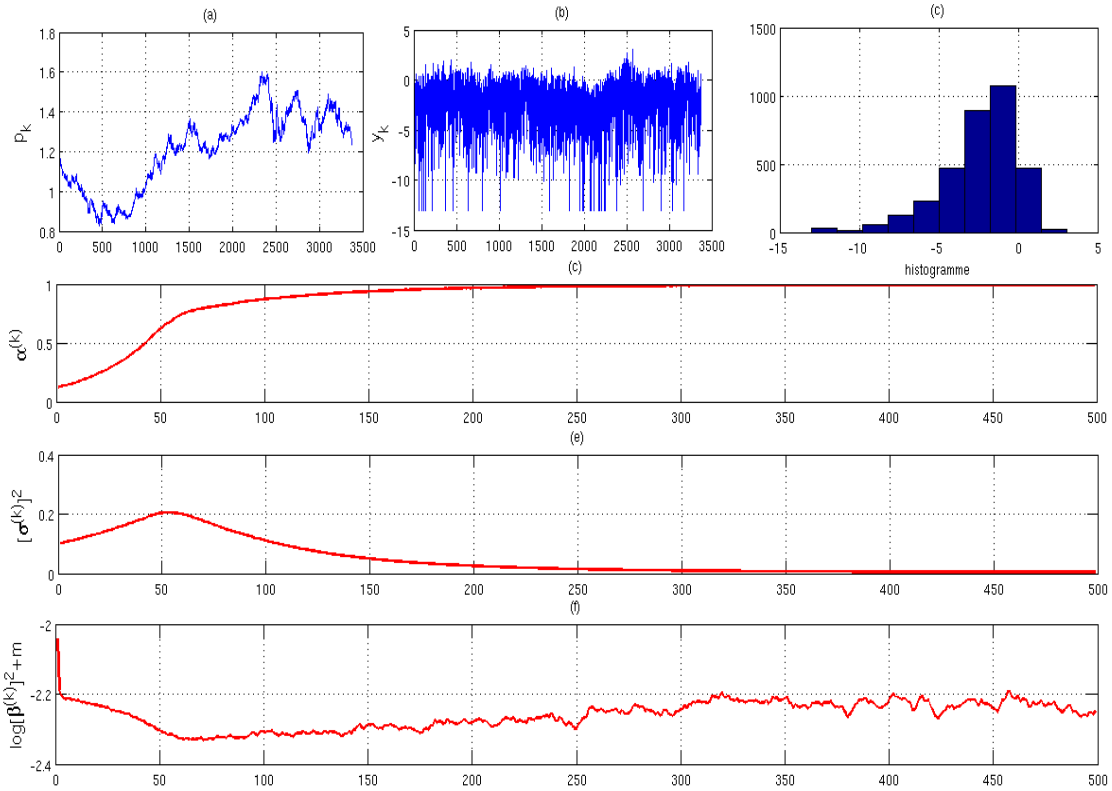


FIGURE 3.5 – Analyse du taux de change EURO/USD.

3.3.4 Indice S&P 500

Nous analysons les données journalières de l'indice S&P 500 pour la période allant de 20 mai 2008 au 08 mai 2012. Le rendement journalier est formé sur les cotations à l'ouverture et à la fermeture. Une calibration du modèle de volatilité basique est aussi effectuée. Le tracé de cette dernière est donné à la Figure 3.6.

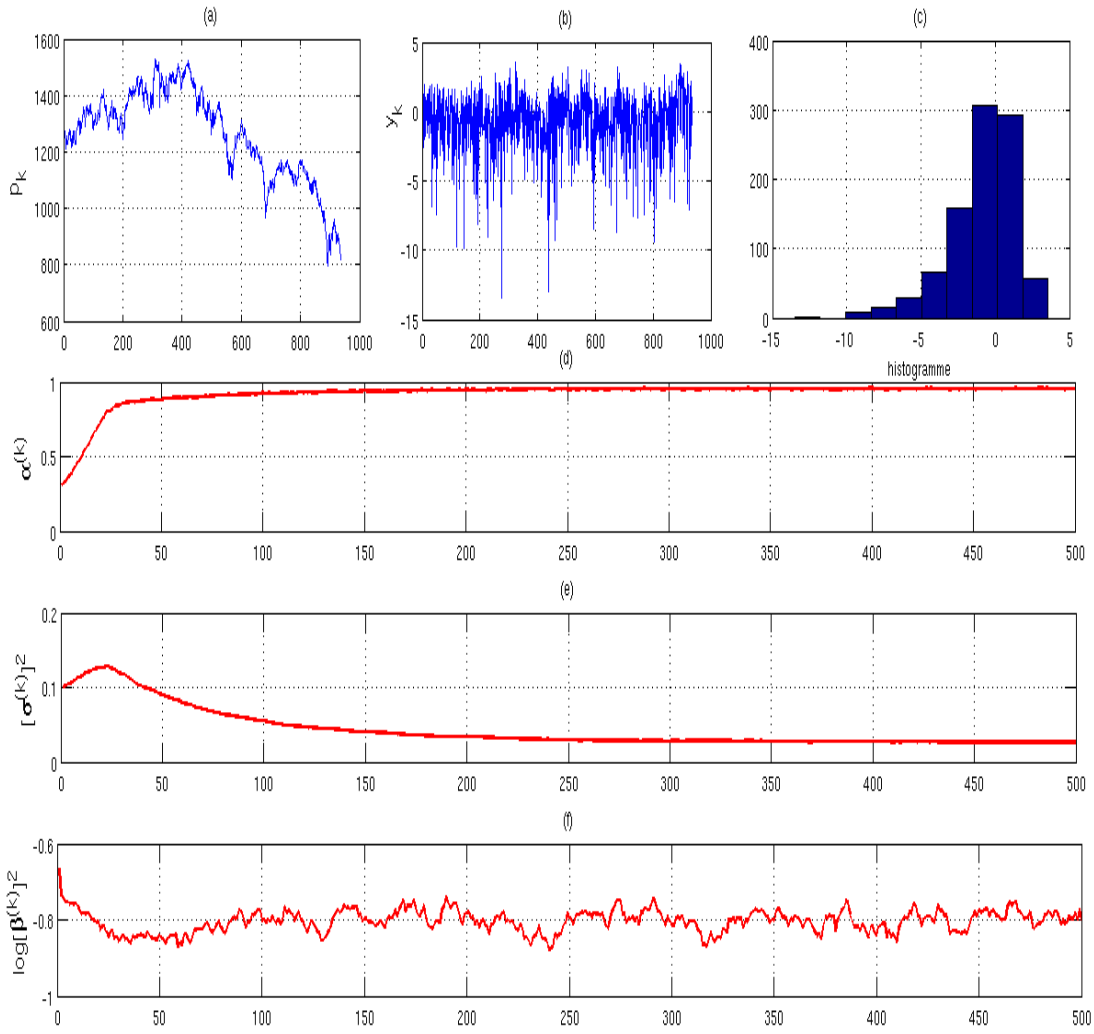


FIGURE 3.6 – Analyse de l'indice S&P 500.

3.3.5 Indice Dow Jones

Nous disposons des cotations journalières (à la fermeture) de l'indice Dow Jones pour la période allant du 04 janvier 1999 au 24 septembre 2002. Une analyse analogue au précédent indice est réalisé.

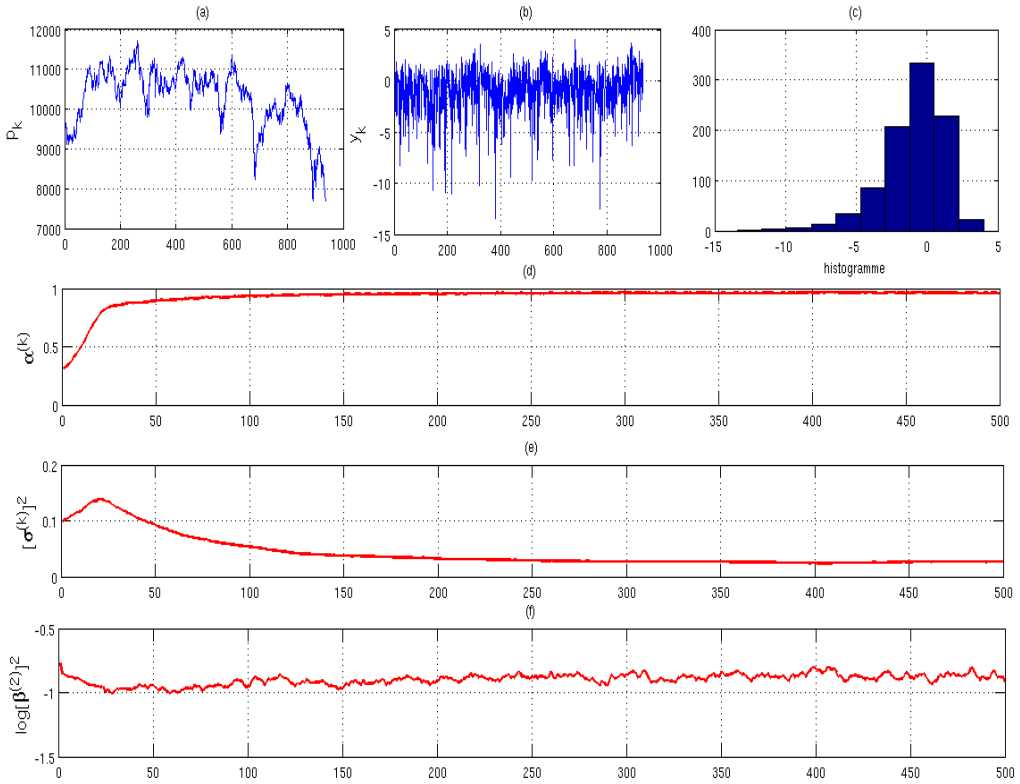


FIGURE 3.7 – Analyse de l'indice Dow Jones.

3.3.6 Indice FTSE 100

Nous examinons également l'indice FTSE 100 pour la période du 04 janvier 1999 au 24 septembre 2002. Les résultats sont consignés à la Figure 3.8.

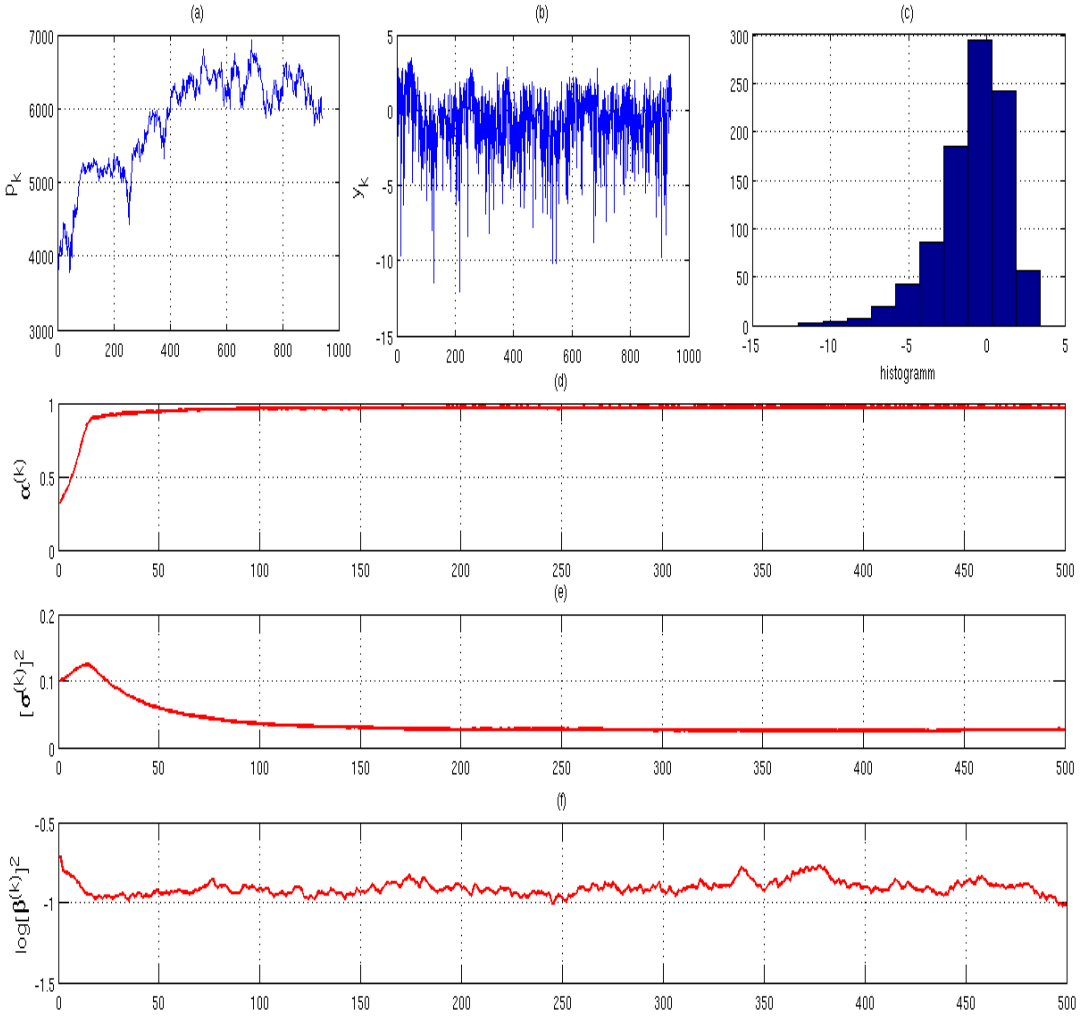


FIGURE 3.8 – Analyse de l'indice FTSE 100.

3.3.7 Indice Nikkei 225

Une dernière application est effectuée sur l'indice Nikkei 225. La même période retenue pour l'indice FTSE 100 est également considérée. Une analyse similaire est aussi menée.

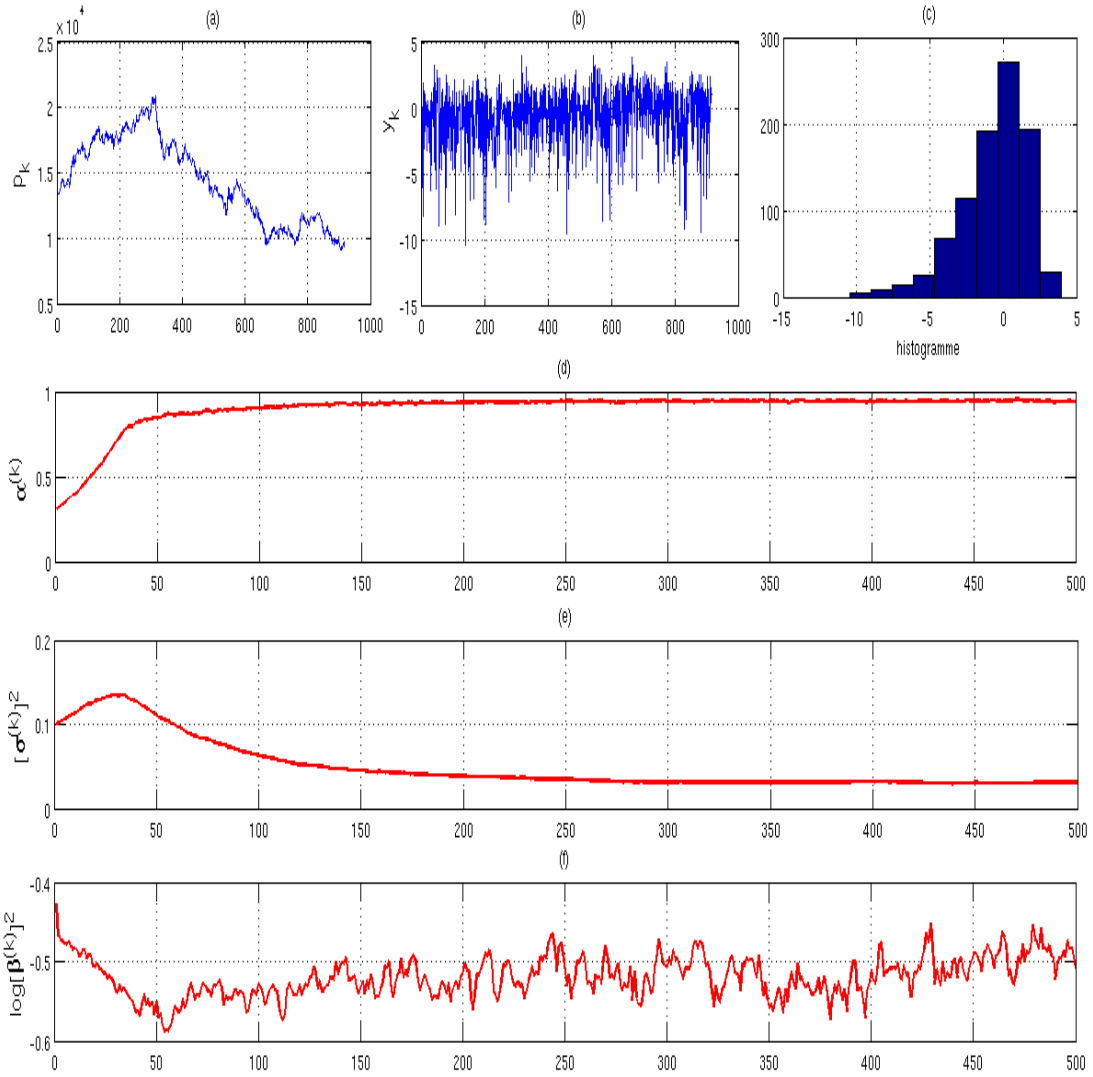


FIGURE 3.9 – Analyse de l'indice Nikkei 225.

Comparaison des indices boursiers

Nous reprenons les résultats d'estimation des trois paramètres sur les indices Dow Jones, FTSE et Nikkei 225 dans Krichene [2003] par les méthodes MCMC. Les moyennes *a posteriori* de ces paramètres sont résumées au tableau suivant :

Estimations \ Indices	Nikkei 225	Dow Jones	FTSE 100
$\hat{\alpha}$	0.944	0.961	0.974
$\hat{\sigma}$	0.178	0.165	0.157
$\exp(\hat{\varpi})$	1.387	1.154	1.182

TABLE 3.3 – Estimations par MCMC

[4] On peut directement comparer la dernière itération du MCEM avec les moyennes estimées par MCMC. Cependant, nous avons moyenné les 40 dernières itérations du MCEM afin de les comparer aux moyennes *a posteriori* de ces paramètres estimés par MCMC dans Krichene [2003]. Le résumé est consigné dans le tableau suivant :

Estimations \ Indices	Nikkei 225	Dow Jones	FTSE 100
$\hat{\alpha}$	0.9447	0.9609	0.9692
$\hat{\sigma}$	0.1744	0.1628	0.1631
$\exp(\hat{\varpi})$	1.4770	1.2196	1.1884

TABLE 3.4 – Moyennes des 40 dernières itérations du MCEM

On constate que les résultats des deux méthodes sont sensiblement proches.

Remarque 3.3.1. *Dans ces différentes applications aux données de taux de change et d'indices boursiers il ressort que le paramètre α est très proche de 1. Ce qui tend à confirmer une hypothèse de persistance de la volatilité ainsi que de son retour à la moyenne de long terme. Les valeurs estimées de σ sont raisonnablement petites (inférieures à 20%). Ce qui permet de garantir une certaine stabilité de la calibration de la volatilité. Enfin, le paramètre $\exp(\varpi) = \beta^{-2}$ dont les valeurs estimées sont relativement grandes traduit la quantifiabilité de l'apport de nouvelles informations sur la volatilité.*

Afin d'éprouver le modèle de volatilité basique ainsi que la linéarisation effectuée dans l'équation des observations, nous avons repris le modèle non linéarisé auquel nous avons adjoint un terme additif en $X^2 + \cos(X^2)$ hautement non-linéaire, augmentant de fait la complexité du modèle. Si bien que l'évolution des observations est régie par :

$$Y_k = \beta \exp\left(\frac{X_k^2 + X_k + \cos(X_k^2)}{2}\right) V_k \tag{3.10}$$

4. Le paramètre noté $\exp(\varpi)$ dans Krichene [2003] est lié à notre paramètre β par la relation $\beta = \exp(-\varpi/2)$.

avec comme dynamique complète du modèle s'écrivant :

$$\begin{cases} X_k = \alpha X_{k-1} + \sigma W_k \\ Y_k = \beta \exp\left(\frac{X_k^2 + X_k + \cos(X_k^2)}{2}\right) V_k. \end{cases} \tag{3.11}$$

Dans cette nouvelle configuration, nous avons répliqué 150 expériences de Monte Carlo consistant à estimer le modèle (3.11) par MCEM avec des paramètres d'initialisation aléatoires. Pour chaque expérience, nous avons lancé 500 fois la procédure MCEM avec dans un premier temps $N = 200$ particules, $\theta^{(0)} = (\alpha^{(0)}, \sigma^{(0)}, \beta^{(0)})$ avec $\alpha^{(0)}, \beta^{(0)} \sim \mathcal{U}([0,1])$ et $\sigma^{(0)} \sim \mathcal{U}([0.01, 1.01])$. Rappelons que la procédure du MCEM obéit au schéma de récurrence habituel. Étant donnée la $(k+1)^{\text{ème}}$ itération le mécanisme de mise à jour séquentiel des paramètres suit alors la récurrence suivante :

$$\begin{cases} \alpha^{(k+1)} = \frac{\sum_{r=2}^{n} \mathbb{E}_{\theta^{(k)}}[X_{r-1} X_r | Y_{1:n}]}{\sum_{r=2}^{n} \mathbb{E}_{\theta^{(k)}}\left[X_{r-1}^2 | Y_{1:n}\right]} \\ \sigma^{(k+1)} = \sqrt{\frac{1}{n} \sum_{r=2}^{n} \mathbb{E}_{\theta^{(k)}}\left[(X_r - \alpha^{(k+1)} X_{r-1})^2 | Y_{1:n}\right]} \\ \beta^{(k+1)} = \sqrt{\sum_{r=1}^{n} y_i^2 \, \mathbb{E}_{\theta^{(k)}}\left[e^{-(X_r^2 + \cos(X_r^2) + X_r)} | Y_{1:n}\right]}. \end{cases} \tag{3.12}$$

Les paramètres du modèle synthétique utilisé étant $\theta^* = (0.7, 0.2, 5)$ avec $\alpha^* = 0.7$, $\sigma^* = 0.2$ et $\beta^* = 5$ dont une trajectoire est donnée à la Figure 3.10.

A la fin de chaque expérience de Monte Carlo (500 $^{\text{ème}}$ itération du MCEM), nous

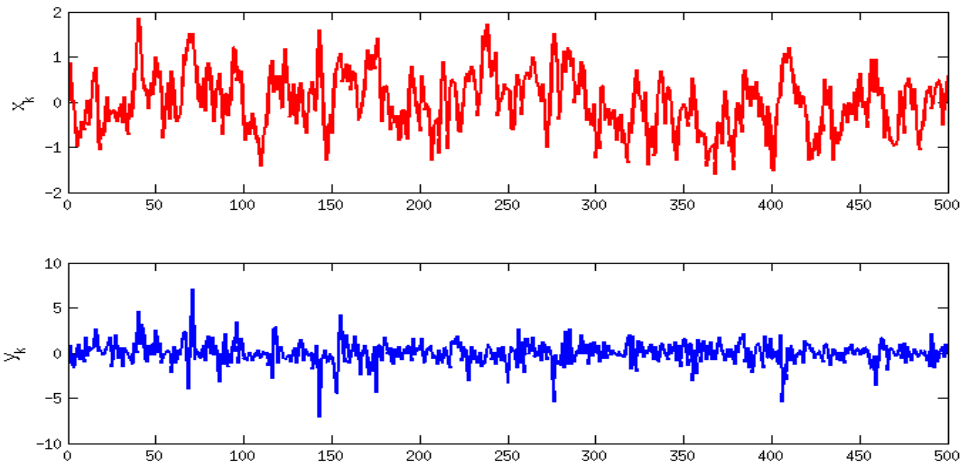


FIGURE 3.10 – Trajectoire du modèle (3.11) sur un horizon de $T = 500$ réalisations.

moyennons les 50 dernières itérations du MCEM pour chacun des 3 paramètres estimés. Un premier résumé des expériences de Monte Carlo est donné au tableau suivant

Dans le Tableau 3.5, nous avons mis respectivement les moyenne, biais, écart-type et

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.7344	0.2351	0.5215
Biais	0.0344	0.0351	0.0215
écart-type	0.0754	0.0674	0.1167
REQM	0.0830	0.0761	0.1187

TABLE 3.5 – Résumé des 150 réplications de Monte Carlo

la racine de l'erreur quadratique moyenne obtenus à l'issu des expériences de Monte Carlo pour chacun des 3 paramètres. Un diagnostic graphique des erreurs (Figure 3.11) est donné par les histogrammes des 3 paramètres sur les 150 expériences de Monte Carlo.

Nous avons constaté un taux de divergence commun aux 3 paramètres de 6% pouvant

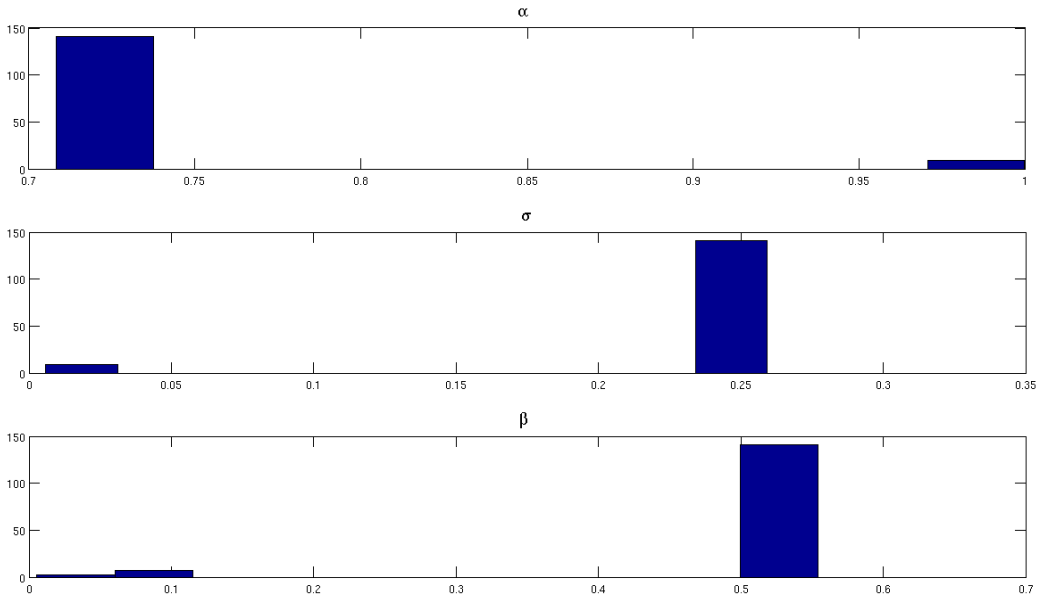


FIGURE 3.11 – Histogrammes des paramètres sur 150 expériences de Monte Carlo.

résulter entre autres des problèmes d'initialisation, de modalités de la vraisemblance, etc. Ce qui se traduit par l'existence de classes supplémentaires à très faibles effectifs dans les histogrammes. La classe dominante de chaque histogramme étant celle traduisant la convergence effective du paramètre concerné. On constate une restitution assez fidèle des paramètres injectés en entrée du modèle et ce de manière satisfaisante.

3.4 Extension du modèle canonique

Afin d'éprouver encore plus le modèle précédent, nous avons réalisé 2 autres séries d'expériences de Monte Carlo. Dans la première série de 150 réplications, nous faisons varier le nombre de particules utilisé de 250 à 2000. La longueur de trajectoire des données reste fixée à $T = 500$. Cette fois, le jeu de paramètres utilisé pour la génération des données est $\theta^* = (0.9, 0.2, 0.6)$ et l'initialisation du MCEM est $\theta^{(0)} = (0, 0.1, 0)$ Un premier résumé est consigné à la Figure 3.12.

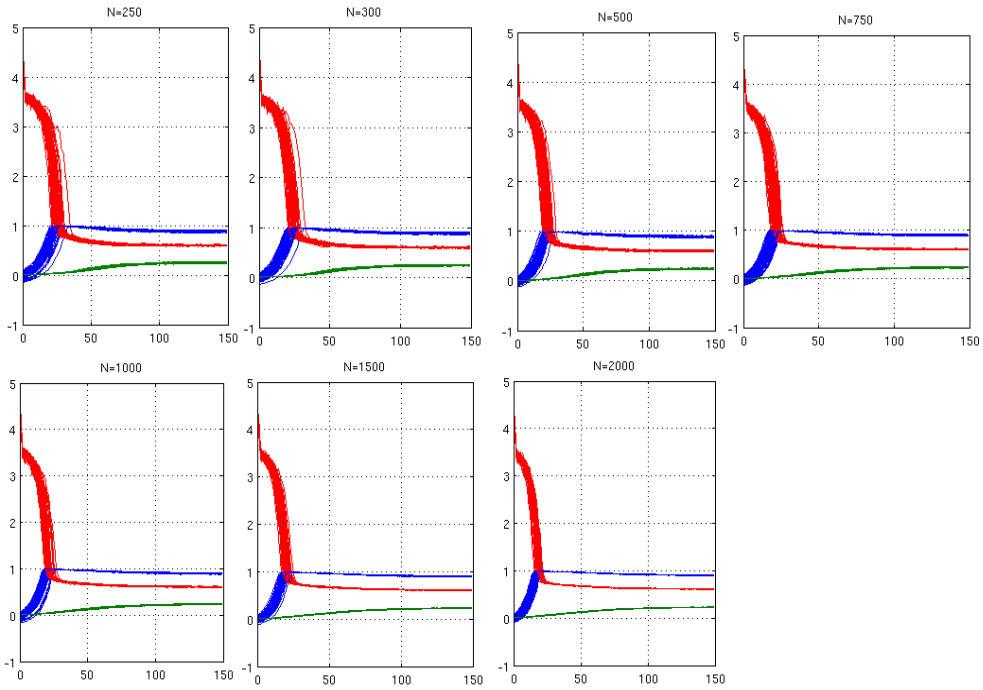


FIGURE 3.12 – 150 expériences de Monte Carlo : Évolution du MCEM en fonction du nombre de particules pour $T = 500$ fixé.

Remarque 3.4.1. *Un constat que l'on peut faire est qu'à mesure que le nombre de particules croit, on observe une plus grande stabilité au niveau des paramètres estimés à T fixé. Ce qui conforte l'idée de la convergence effective des statistiques exhaustives, fonction des distributions conditionnelles approximées par les systèmes de particules. Cependant, cela n'augure en rien de la convergence effective des paramètres estimés vers le vecteur de paramètres θ^* au vu des points soulevés au chapitre 1 sur la convergence de l'EM en général.*

Par soucis de concision, nous avons aussi agrégé les 50 dernières itérations du MCEM en les moyennant. Une évolution comparative des ces paramètres estimés est illustrée à la Figure 3.13.

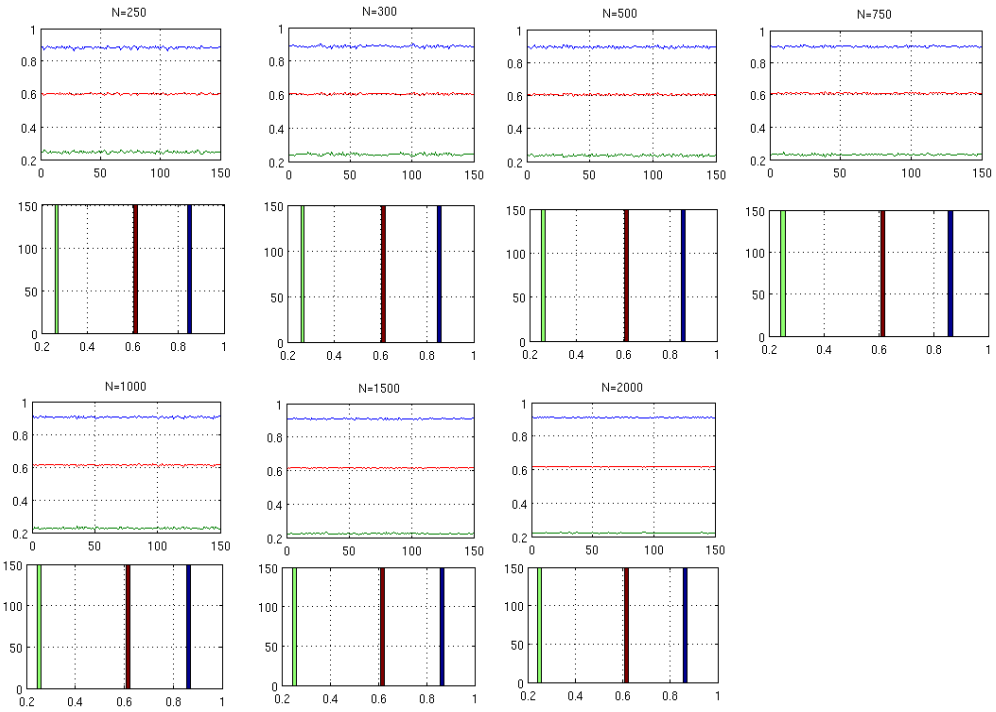


FIGURE 3.13 – *de haut en bas* 150 expériences de Monte Carlo : Évolution des 50 dernières itérations du MCEM en fonction du nombre de particules pour $T = 500$ fixé avec les histogrammes correspondants.

Remarque 3.4.2. *Le biais apparent constaté dans le premier jeux de données demeure. Cependant, l'on peut constater que pour une composante θ_i du vecteur de paramètres il existe une itération k_i permettant de réduire voire d'annuler ce biais. Le problème se pose alors de trouver, s'il existe la $k^{ème}$ itération annulant le biais de toutes les composantes du vecteur de paramètres.*

Parallèlement aux illustrations graphiques, nous avons calculé quelques grandeurs numériques pour une meilleure appréhension des simulations. Les tableaux suivants illustrent un résumé agrégé des 150 réplications de Monte Carlo à travers les moyennes, biais, erreurs quadratiques moyennes et écarts-types constitués à partir des 50 dernières itérations du MCEM.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.8835	0.2452	0.6008
Biais	−0.0164	0.0452	0.0088
écart-type	0.0754	0.0674	0.1167
REQM	0.0198	0.0457	0.0082

TABLE 3.6 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.8876	0.2421	0.6028
Biais	−0.0123	0.0421	0.0028
écart-type	0.0095	0.0057	0.0067
REQM	0.0159	0.0426	0.0082

TABLE 3.7 – Agrégation des 150 réplications de Monte Carlo avec $N = 300$ et $T = 500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.8937	0.2367	0.6056
Biais	−0.0062	0.0367	0.0056
écart-type	0.0084	0.0059	0.0060
REQM	0.0112	0.0372	0.0088

TABLE 3.8 – Agrégation des 150 réplications de Monte Carlo avec $N = 500$ et $T = 500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9008	0.2299	0.6103
Biais	0.0008	0.0299	0.0103
écart-type	0.0071	0.0060	0.0052
REQM	0.0082	0.0306	0.0117

TABLE 3.9 – Agrégation des 150 réplications de Monte Carlo avec $N = 750$ et $T = 500$.

La seconde série de 150 réplications de Monte Carlo consiste à observer le comportement des paramètres estimés sur une longueur de trajectoire variable tout en gardant le nombre de particules constant. Le résumé des différentes simulations est donné à la figure Figure 3.14.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9046	0.2265	0.6131
Biais	0.0046	0.0265	0.0131
écart-type	0.0063	0.0057	0.0046
REQM	0.0087	0.0272	0.0140

TABLE 3.10 – Agrégation des 150 réplications de Monte Carlo avec $N = 1000$ et $T = 500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9086	0.2228	0.6159
Biais	0.0086	0.0228	0.0159
écart-type	0.0052	0.0054	0.0040
REQM	0.0103	0.0235	0.0164

TABLE 3.11 – Agrégation des 150 réplications de Monte Carlo avec $N = 1500$ et $T = 500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9103	0.2211	0.6172
Biais	0.0103	0.0211	0.0172
écart-type	0.0046	0.0054	0.0035
REQM	0.0113	0.0218	0.0176

TABLE 3.12 – Agrégation des 150 réplications de Monte Carlo avec $N = 2000$ et $T = 500$.

Comme précédemment, pour chaque réplication nous nous sommes intéressé aux 50 dernières itérations du MCEM. Une vue de celles-ci est donnée à la Figure 3.15.

Remarque 3.4.3. *Un premier constat que l'on peut faire est que l'on est en mesure d'estimer correctement les paramètres à mesure que l'on augmente la longueur de la trajectoire malgré que l'on ait gelé le nombre de particules à* $N = 250$. *Le seul inconvénient apparent est l'effort supplémentaire fourni pour traiter les observations en plus.*

Nombre de particules N / Longueur de la trajectoire T	250	300	500	750	1000	1500	2000
500	0.0198	0.0159	0.0112	0.0082	0.0087	0.0103	0.0113

TABLE 3.13 – Récapitulatif des REQM de $\hat{\alpha}$ en fonction de N pour $T = 500$ réalisations

Nombre de particules N / Longueur de la trajectoire T	500	300	500	750	1000	1500	2000
500	0.0457	0.0426	0.0372	0.0306	0.0272	0.0235	0.0218

TABLE 3.14 – Récapitulatif des REQM de $\hat{\sigma}$ en fonction de N pour $T = 500$ réalisations

Nombre de particules N / Longueur de la trajectoire T	250	300	500	750	1000	1500	2000
250	0.0082	0.0082	0.0088	0.0117	0.0140	0.0164	0.0176

TABLE 3.15 – Récapitulatif des REQM de $\hat{\beta}$ en fonction de N pour $T = 500$ réalisations

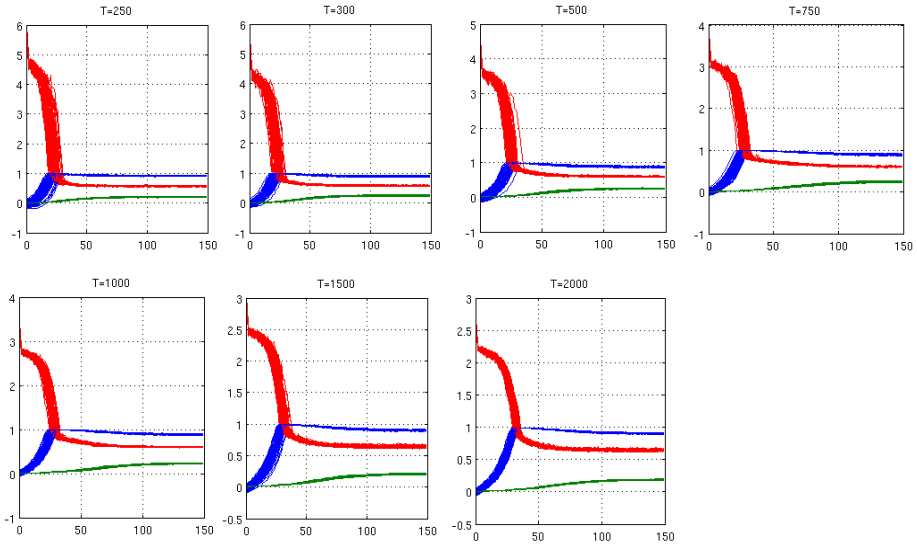


FIGURE 3.14 – 150 expériences de Monte Carlo : Évolution du MCEM en fonction de la longueur de la trajectoire pour $N = 250$ fixé.

De manière analogue aux réplications antérieures, un point de vue numérique des 50 dernières itérations du MCEM peut être considéré. Les tableaux suivant en constituent le résumé agrégé.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9205	0.2017	0.5705
Biais	0.0205	0.0017	−0.0294
écart-type	0.0111	0.0048	0.0088
REQM	0.0235	0.0067	0.0308

TABLE 3.16 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 250$.

Remarque 3.4.4. *On peut constater dans les précédentes simulations l'utilité d'arrêter les itérations du MCEM plus tôt. En prenant par exemple le rapport de vraisemblances incomplètes entre paramètres consécutifs comme critère d'arrêt. On se fixe alors un seuil de précision qui, dès qu'il est atteint impose la fin des itérations du MCEM. Cependant, ce choix du seuil certes arbitraire se doit d'être judicieux. En effet, un seuil trop faible (une grande précision) entraînera plus d'effort computationnel pour*

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.8947	0.2497	0.5825
Biais	−0.0052	0.0497	−0.0174
écart-type	0.0118	0.00659	0.0083
REQM	0.0146	0.0501	0.0195

TABLE 3.17 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 300$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.8835	0.2452	0.6008
Biais	−0.0164	0.0452	0.0088
écart-type	0.0754	0.0674	0.1167
REQM	0.0198	0.0457	0.0082

TABLE 3.18 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9013	0.2316	0.6115
Biais	0.0013	0.0316	0.0115
écart-type	0.0104	0.01092	0.0081
REQM	0.0122	0.03365	0.0146

TABLE 3.19 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 750$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.8896	0.2212	0.6038
Biais	−0.0103	0.0212	0.0038
écart-type	0.0091	0.0081	0.0058
REQM	0.0142	0.0229	0.0074

TABLE 3.20 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 1000$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9042	0.2012	0.6385
Biais	0.0042	0.0012	0.0385
écart-type	0.0069	0.0066	0.0115
REQM	0.0085	0.0075	0.0402

TABLE 3.21 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 1500$.

Paramètres estimés	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\beta}$
Moyenne	0.9080	0.1826	0.6408
Biais	0.0080	−0.0173	0.0408
écart-type	0.0067	0.0066	0.0094
REQM	0.0104	0.0187	0.0419

TABLE 3.22 – Agrégation des 150 réplications de Monte Carlo avec $N = 250$ et $T = 2000$.

Nombre de particules N \ Longueur de trajectoire T	250	300	500	750	1000	1500	2000
250	0.0235	0.0146	0.0198	0.0122	0.0142	0.0085	0.0104

TABLE 3.23 – Récapitulatif des REQM de $\hat{\alpha}$ en fonction de T pour $N = 250$ particules

Nombre de particules N \ Longueur de trajectoire T	250	300	500	750	1000	1500	2000
250	0.0067	0.0501	0.0457	0.03365	0.0229	0.0075	0.0187

TABLE 3.24 – Récapitulatif des REQM de $\hat{\sigma}$ en fonction de T pour $N = 250$ particules

Nombre de particules N \ Longueur de trajectoire T	250	300	500	750	1000	1500	2000
250	0.0308	0.0195	0.0082	0.0146	0.0074	0.0402	0.0419

TABLE 3.25 – Récapitulatif des REQM de $\hat{\beta}$ en fonction de T pour $N = 250$ particules

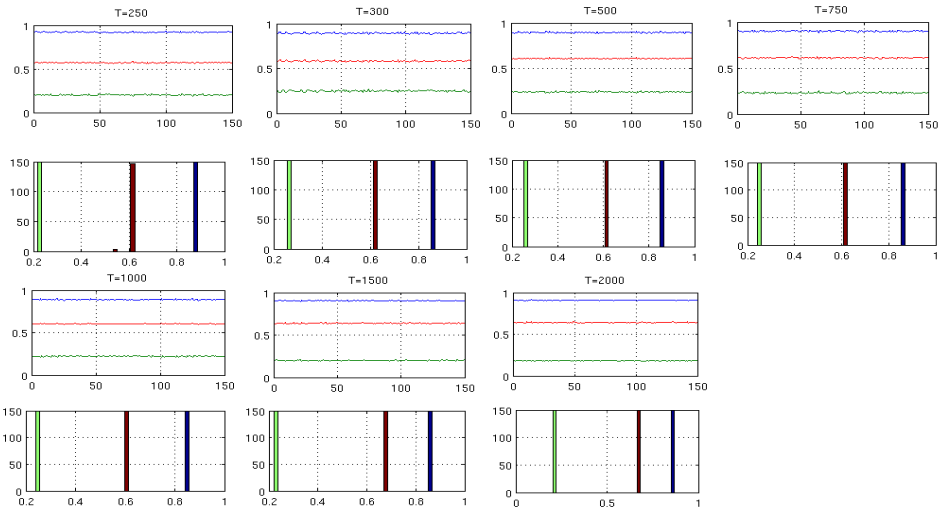


FIGURE 3.15 – *de haut en bas* 150 expériences de Monte Carlo : Évolution des 50 dernières itérations du MCEM en fonction de la longueur de la trajectoire pour un nombre de particules fixé à $N = 250$ ainsi que les histogrammes correspondants.

être atteint. Tandis que le contraire aura l'effet inverse sans pour autant garantir la proximité du vrai vecteur de paramètres cible θ^*.

3.5 Conclusion

Dans ce chapitre, à travers le modèle de volatilité stochastique canonique nous avons mis en relief l'usage commun de l'algorithme EM et les méthodes de MCs. Nous nous sommes intéressés à sa calibration aux données financières notamment aux indices boursiers et aux taux de change. Nous nous sommes également comparé aux MCMC traitant des mêmes données. Enfin, nous avons également éprouvé le modèle de base en adjoignant une composante fortement non-linéaire. D'une part, afin de se départir du modèle de base et d'autre part de jauger de la robustesse du MCEM dans les paramètres estimés. Ce travail est loin d'être complet. Il demeure encore plusieurs points qui nécessitent plus d'attention. Un point qui mérite que l'on s'y arrête d'avantage est la réduction du coût de calcul. Par exemple, une idée sur la réduction du coût de calcul du MCEM serait de rendre adaptatif le nombre de particules utilisées combiné à une fonction de forçage pour obliger l'arrêt d'itérations non requises. Étant données que certaines composantes du vecteur de paramètres convergent plus vite que d'autres. Ce qui peut se traduire par un gain substantiel en temps de calcul.

CHAPITRE 4

On some extensions of the SMC methods in higher-order HMM

Sommaire

4.1 Introduction 94
4.2 SMC methods in one-order HMM 95
4.2.1 Filtering recursions 96
4.2.2 Smoothing recursions 97
4.2.2.1 Marginal smoothing 98
4.2.2.2 Joint smoothing 100
4.2.2.3 Particle smoother 101
4.3 SMC methods in higher-order HMM 101
4.3.1 $\ell-$order Filtering recursions 102
4.3.2 ℓ-order smoothing recursions 105
4.3.2.1 Joint smoothing 106
4.3.2.2 Particle smoother 107
4.4 Parameter estimation 108
4.5 Convergence issues 111
4.5.1 L_q mean error 111
4.6 Proof of Prop.4.3.2 113

We analyze some extensions of the Sequential Monte Carlo (SMC) methods in the context of nonlinear state space models. Namely, we adapt the SMC methods to handle higher-order HMM through the usual recursions of posterior distributions. It proceeds on mimicking the two-step procedure, that is the prediction step and the update step, in the derivation of the filter distribution. Once stated, we extend some

smoothing recursions as the Forward-Backward algorithm and the Backward smoother to deal with the effective smoothing distributions in higher-order HMM. We give a toy example as an application of these recursions. Finally, we also consider studying the convergence issue of smoothing estimates, particularly those obtained in the forward and backward passes when dealing with higher-order HMM.

4.1 Introduction

The literature of SMC methods is recent and can be dated from the paper by Gordon et al. [1993]. Although, several attempts had preceded including the work by Handschin [1970], Handschin and Mayne [1969] among others. The main obstacle to the SMC's growth was the limitation of computing power. Since then, several efforts have been made both in theory and in practice to lay down the foundations of SMC methods. One may consult review articles such as Cappé et al. [2007], Doucet et al. [2000] or books by Cappé et al. [2005], Del Moral [2004] or Doucet et al. [2001] which include several theoretical results and a range of rich and varied applications in many areas.

So far, the SMC methods apply to hidden Markov models of order 1, commonly called one-order HMM. Specifically, a $\mathcal{X} \times \mathcal{Y}-$valued bivariate process $\{(X_k, Y_k)\}$, where $\{X_k\}$ is an unobservable dynamic Markov model of order 1. $\{Y_k\}$ represents the observation process used indirectly to quantify the realizations of the process $\{X_k\}$ and satisfying the channel without memory property's. However, it may happen that the signal process $\{X_k\}$ depends on more than one of its lags that is, the memory process of the signal is more persistent. Thus, a direct application of SMC methods still a little tricky. To overcome this difficulty, one may think at first glance that a trivial rewriting of the process $\{X_k\}$ according to its lags is enough and may help to fall in the usual case of Markov chain of order 1. However, this formulation is not without causing additional difficulty. In fact, one may face among other the degeneracy problem of the state noise resulting from this state transformation. A new approach is needed. In this perspective, we derive a new approach that helps handling higher-order HMM without any modification of the former kind . To achieve it, we just mimic the different stages in the establishment of the filtering and smoothing equations in non-linear and non-Gaussian state space models. Singularly, we mimic the prediction and the correction steps of the filter distribution in one-order HMM and adapt it to HMM of order strictly greater than one. Once done, we derive analogous recursions to those of the Forward-Backward smoother and the particle smoother by Godsill et al. [2004]. In the sequel, we show a use of the SMC methods extension in an example a stochastic volatility with an l memory depth. As a final point, we end up with some convergence issues of the resulting approximations.

4.2 SMC methods in one-order HMM

Particle filter and smoother belong to SMC methods that aim at generating samples realizations from actual and historical state sequences given the whole data set or a part of it. The main idea being that any given measure on a measurable space can be approximated by a sum of empirical measures. Particle filter aims at computing recursively in time, the conditional distribution of the current state given the whole data up to current time k, that is the filtering distribution. Smoothing is more branched, however, most cases can be plugged into the joint smoothing distribution. When classical approaches fail because of lack of analytical solutions or for a non-linear or non Gaussian purpose, the SMC methods can help in a certain way to get rid off most of these limitations. As long as some minimal requirements are met, the SMC methods are set of powerful tools that approximate any function of the state sequences even for a class of unbounded functions [1] given the data up to a given time. To state the general idea of particle filter and smoother, consider the following one-order HMM :

$$\begin{cases} X_k &= a_k(X_{k-1}, V_k) \\ Y_k &= b_k(X_k, W_k) \end{cases} \tag{4.1}$$

where $a_k(\cdot,\cdot)$ and $b_k(\cdot,\cdot)$ are possibly non-linear functions, $\{X_k\}$ is a $1-$order Markov chain with initial state X_0 distributed according to a diffuse prior distribution $\nu(\cdot)$ and transition kernel M from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We assume that M admits a density function m w.r.t a dominating measure λ. $(V_k)_{k\geq 1}$ and $(W_k)_{k\geq 1}$ are i.i.d disturbance noises independent of X_0, respectively the state noise and the measurement noise. We also assume that the observation process $\{Y_k\}$, constructed on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is conditionally independent given $\{X_k\}$ with a marginal distribution admitting a density function g such that

$$\forall A \in \mathcal{B}(\mathcal{Y}), \quad \mathbb{P}(Y_k \in A | X_k) = \int_A g(X_k, y)\mu(dy),$$

where μ is a $\sigma-$finite measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. To sum up, the model is given by

$$\begin{cases} X_0 \sim \nu(\cdot) \\ X_k | X_{k-1} = x_{k-1} \sim m(x_{k-1}, \cdot) \quad k \geq 1 \\ Y_k | X_k = x_k \sim g(\cdot, x_k) \end{cases} \tag{4.2}$$

For the sake of simplicity the data are fixed that is, $Y_k = y_k$ for all time indexes. The Lebesgue measure is used as a reference measure in order to lighten the notations. We also omit the dependence of the so called marginal likelihood function $g(\cdot)$ to the data by using the shortened notation $g_k(x_k) := g(x_k, y_k)$ and $p(\cdot)$ denotes a generic

1. see Hu and Schön [2011] for details of such unboundedness.

symbol for densities. The following notations will be used to introduce the quantities of interest. Let $\mathbf{F}_b(\mathcal{X}^{k+1})$ be the set of bounded and measurable functions on $\mathcal{X}^{k+1}$. Given an HMM with initial state X_0 distributed according to $\nu(\cdot)$, define

$$\phi_{\nu,0:k|k}(f) := \mathbb{E}_\nu \left[f(X_{0:k}) | Y_{1:k} \right], \quad k \geq 0, f \in \mathbf{F}_b(\mathcal{X}^{k+1}) \tag{4.3}$$

as the conditional distribution of $f(X_{0:k})$ given $Y_{1:k}$ with $X_0 \sim \nu(\cdot)$. Whenever f depends only on X_k, it is usual to simplify the notation to :

$$\phi_{\nu,k}(f) := \mathbb{E}_\nu \left[f(X_k) | Y_{1:k} \right], \quad k \geq 0, f \in \mathbf{F}_b(\mathcal{X}) \tag{4.4}$$

and we refer to this as the filter distribution that is, the conditional distribution of $f(X_k)$ given $Y_{1:k}$. We also introduce the $1-$step predictive distribution

$$\phi_{\nu,k|k-1}(f) := \mathbb{E}_\nu \left[f(X_k) | Y_{1:k-1} \right], \quad k \geq 0, f \in \mathbf{F}_b(\mathcal{X}) \tag{4.5}$$

with the convention $\phi_{\nu,0|-1} := \nu$, where $\mathbb{E}_\nu$ is the expectation taken with the underlying law and emphasizing ν as the initial distribution of X_0. We also denote similarly the corresponding conditional densities of the later distributions as a slight abuse of notation. Their arguments help discriminate between these functions. For example, $\phi_{\nu,k}(x_k)$ is used to denote the filtering density while $\phi_{\nu,k|k-1}(x_k)$ is the $1-$step predictive density.

4.2.1 Filtering recursions

Particle filtering goal is to compute recursively in time the joint posterior distribution (4.3) or some of its features such as (4.4) a.k.a the filtering distribution. In terms of operator, (4.3) admits the compact recursive formula

$$\phi_{\nu,0:k|k}(f) = \frac{\phi_{\nu,0:k-1|k-1}(f g_k M)}{\phi_{\nu,0:k-1|k-1}(g_k M)}, \quad \forall f \in \mathbf{F}_b(\mathcal{X}^{k+1}) \tag{4.6}$$

and (4.4) satisfies the recursive formulas :

$$\phi_{\nu,k|k-1} = \phi_{\nu,k-1} M \tag{4.7}$$

and

$$\phi_{\nu,k}(f) = \frac{\phi_{\nu,k|k-1}(f g_k)}{\phi_{\nu,k|k-1}(g_k)}, \quad \forall f \in \mathbf{F}_b(\mathcal{X}). \tag{4.8}$$

Note that (4.6) and (4.8) are obtained via Bayes rule and (4.7) is a direct application of Kolmogorov equation. A more intuitive interpretation of these relations can be given in terms of the corresponding conditional densities given by :

$$\phi_{\nu,0:k|k}(x_{0:k}) = \frac{\phi_{\nu,0:k-1|k-1}(x_{0:k-1}) m(x_{k-1}, x_k) g_k(x_k)}{\int_{\mathcal{X}^{k+1}} \phi_{\nu,0:k-1|k-1}(x_{0:k-1}) m(x_{k-1}, x_k) g_k(x_k) dx_{0:k}} \tag{4.9}$$

for the joint posterior density and

$$\phi_{\nu,k|k-1}(x_k) = \int_{\mathcal{X}} m(x_{k-1}, x_k)\phi_{\nu,k-1|k-1}(x_{k-1})dx_{k-1} \tag{4.10}$$

$$\phi_{\nu,k}(x_k) = \frac{g_k(x_k)\phi_{\nu,k|k-1}(x_k)}{\int_{\mathcal{X}} g_k(x_k)\phi_{\nu,k|k-1}(x_k)dx_k} \tag{4.11}$$

for the predictive and the filtering density respectively. So, particle filter is a two-step procedure that uses (4.7) as a prediction step for the next state and (4.11) as an update step according to the new observation. Within a Sequential Importance sampling procedure, one can get a particle filter estimate of (4.9) :

$$\hat{\phi}_{\nu,0:k|k}(dx_{0:k}) = \Omega_k^{-1} \sum_{i=1}^{N} \omega_k^{(i)} \delta_{\xi_{0:k}^{(i)}}(dx_{0:k}) \tag{4.12}$$

and deduce an estimate of (4.11) as marginal distribution of the latter :

$$\hat{\phi}_{\nu,k}(dx_k) = \Omega_k^{-1} \sum_{i=1}^{N} \omega_k^{(i)} \delta_{\xi_k^{(i)}}(dx_k) \tag{4.13}$$

where $\Omega_k := \sum_{i=1}^{N} \omega_k^{(i)}$, $\delta_x(\cdot)$ is the delta-Dirac mass located at x and $\omega_k^{(i)}$ is the importance weight associated to the particles position $\xi_{0:k}^{(i)}$. The detail derivation of these weights may be found in Doucet et al. [2001], Doucet and Johansen [2011]. A summary of particle filter is given below. $q(\cdot)$ is a generic notation for instrumental densities in the Importance Sampling procedure. Note that the resampling step is done only if the degeneracy problem appears, for example when using the effective sample size approximation as a quantifier of this phenomena. Before moving towards, note that one can have an approximation of the joint posterior distribution $p(dx_{0:n}|y_{1:n})$ just on storing the outputs at each time step of the generic particle filter.

4.2.2 Smoothing recursions

The general idea shared by most of smoothing recursions is the nature of the reversed time of the dynamic model $\{X_k\}$. In fact, $\{X_k\}$ still a Markov chain, backward in time. The following result makes clear that assertion.

Proposition 4.2.1. *Given the data, $\{X_k\}$ is a Markov chain backward in time with transition backward kernels from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ defined by :*

$$\begin{aligned} B_{k,\nu}(X_{k+1}, f) :&= \mathbb{E}\left[f(X_k)|X_{k+1:n}, Y_{0:n}\right] \\ &= \mathbb{E}\left[f(X_k)|X_{k+1}, Y_{0:k}\right] \end{aligned} \tag{4.14}$$

for any $f \in \mathbf{F}_b(\mathcal{X})$.

Algorithme 10 Generic particle filter

1: Initialization : For $i = 1, 2, \ldots, N$ draw $\xi_0^{(i)} \sim q(\cdot)$ and set $\omega_0^{(i)} = 1/N$;
2: Set $k \leftarrow 1$
3: Importance Sampling step : For $i = 1, 2, \ldots, N$
— draw $\bar{\xi}_k^{(i)} \sim q\left(\cdot|\xi_{0:k-1}^{(i)}, y_{1:k}\right)$
— Evaluate and Normalize the importance weights :

$$\omega_k^{(i)} \propto \omega_{k-1}^{(i)} \frac{m\left(\xi_{k-1}^{(i)}, \xi_k^{(i)}\right) g_k\left(\xi_k^{(i)}\right)}{q\left(\xi_k^{(i)}|\xi_{0:k-1}^{(i)}, y_{1:k}\right)}$$

4: Resampling step : (if necessary)
— Multiply/Discard $\bar{\xi}_k^{(i)}$ w.r.t $\omega_k^{(i)}$ to get $\xi_k^{(i)}$ approximately distributed according to $\phi_{\nu,k}$
— For $i = 1, 2, \ldots, N$ Set $\omega_k^{(i)} = 1/N$
5: Set $k \leftarrow k + 1$ and go to the importance sampling step

Proof. See Cappé et al. [2005], p.70. □

Under this backward dynamic, one can make use of smoothing recursions.

4.2.2.1 Marginal smoothing

The problem in concern is to compute backward and recursively in time the smoothed distribution

$$\phi_{\nu,k|n}(f) := \mathbb{E}\left[f(X_k)|Y_{1:n}\right], \quad k < n \tag{4.15}$$

for any $f \in \mathbf{F}_b(\mathcal{X})$.

Lemma 4.2.2. *For any $1 \leq k < n$, the smoothed distribution factorizes as :*

$$\begin{aligned}\phi_{\nu,k|n}(f) &= \int_{\mathcal{X}} f(x_k) \left[\int_{\mathcal{X}} \frac{\phi_{\nu,k}(x_k) m(x_k, x_{k+1})}{\int_{\mathcal{X}} \phi_{\nu,k}(x_k) m(x_k, x_{k+1}) dx_k} \phi_{\nu,k+1|n}(dx_{k+1})\right] dx_k \\ &= \int_{\mathcal{X}^2} f(x_k) B_{\nu,k}(x_{k+1}, dx_k) \phi_{\nu,k+1|n}(dx_{k+1})\end{aligned} \tag{4.16}$$

where $B_{k,\nu}(X_{k+1}, .)$ is the Backward kernel for any function $f \in \mathbf{F}_b(\mathcal{X})$.

Consider the generic particle filter gathering the weighted sample $\left\{\xi_k^{(i)}, \omega_k^{(i)}\right\}_{i=1}^N$ that target the filtering distribution $p(dx_k|y_{1:k})$ in the sense of (4.13), at time k with $k =$

$1, 2, \ldots, n$. In addition, assume at time $k+1$ one has weighted sample $\left\{\xi_{k+1}^{(i)}, \omega_{k+1|n}^{(i)}\right\}_{i=1}^{N}$ targeting the distribution $\phi_{k+1|n}$ in the sense :

$$\hat{\phi}_{\nu,k+1|n}(dx_{k+1}) = \sum_{j=1}^{N} \omega_{k+1|n}^{(j)} \delta_{\xi_{k+1}^{(j)}}(dx_{k+1}). \tag{4.17}$$

Combining the former and latter outputs, one can achieve a particle estimate of the smoothed distribution given by :

$$\hat{\phi}_{\nu,k|n}(f) = \sum_{i=1}^{N} \omega_{k|n}^{(i)} f(\xi_k^{(i)}), \quad \text{for } k < n \tag{4.18}$$

where the smoothed importance weights are given by :

$$\omega_{k|n}^{(i)} = \omega_k^{(i)} \left(\sum_{j=1}^{N} \omega_{k+1|n}^{(j)} \frac{m\left(\xi_k^{(i)}, \xi_{k+1}^{(j)}\right)}{\sum_{r=1}^{N} \omega_k^{(r)} m\left(\xi_k^{(r)}, \xi_{k+1}^{(j)}\right)} \right) \tag{4.19}$$

The resume of the procedure is given below :

Algorithme 11 Forward-Backward algorithm

1: Forward filtering step : For $k = 0, \ldots, n$

— run the particles filtering algorithm to get the weighted particles $\left\{\xi_k^{(i)}, \omega_k^{(i)}\right\}_{i=1}^{N}$.

2: Backward smoothing step

— For $i = 1, \ldots, N$ set $\omega_{n|n}^{(i)} = \omega_n^{(i)}$

— For $k = n-1$ down to 0 and $i = 1, \ldots, N$ set

$$\omega_{k|n}^{(i)} = \omega_k^{(i)} \left(\sum_{j=1}^{N} \omega_{k+1|n}^{(j)} \frac{m\left(\xi_k^{(i)}, \xi_{k+1}^{(j)}\right)}{\sum_{r=1}^{N} \omega_k^{(r)} m\left(\xi_k^{(r)}, \xi_{k+1}^{(j)}\right)} \right)$$

Remarque 4.2.3. *As one can notice, the F-B algorithm is nothing but a weight update since particle positions generated in the forward pass are kept in the backward pass. Moreover, it is an $O(N^2)$ expensive algorithm at each time step.*

4.2.2.2 Joint smoothing

An extension of the F-B algorithm is reachable for the joint posterior density $p(x_{k:n}|y_{1:n})$. Using similar argument as in the marginal smoothing one can obtain the following recursions :

Lemma 4.2.4. *Under* (4.14)*, for any* $k < n$ *the joint smoothed density* $p(x_{k:n}|y_{1:n})$ *factorizes backward in time as :*

$$p(x_{k:n}|y_{1:n}) = p(x_k|x_{k+1}, y_{1:k})p(x_{k+1:n}|y_{1:n}) \tag{4.20}$$

which iterates to :

$$p(x_{k:n}|y_{1:n}) = p(x_n|y_{1:n}) \prod_{r=k}^{n-1} p(x_r|x_{r+1}, y_{1:r}). \tag{4.21}$$

From this result, one is able to compute the conditional expectation

$$\begin{aligned} \phi_{k:n|n}(f) &:= \mathbb{E}\left[f(X_{k:n})|Y_{1:n}\right] \\ &= \int_{\mathcal{X}^{n-k+1}} f(x_{k:n}) p(x_n|y_{1:n}) \prod_{r=k}^{n-1} p(x_r|x_{r+1}, y_{1:r}) dx_{k:n} \end{aligned} \tag{4.22}$$

for any $f \in \mathbf{F}_b(\mathcal{X}^{n-k+1})$. Note at first that :

$$p(x_r|x_{r+1}, y_{1:r}) \propto p(x_r|y_{1:r})p(x_{r+1}|x_r) \tag{4.23}$$

From a particle estimate of the density $p(x_r|x_{r+1}, y_{1:r})$:

$$\hat{p}(x_r|x_{r+1}, y_{1:r}) = \sum_{i_r=1}^{N} \kappa_r^{(i_r)} \delta_{\xi_r^{(i_r)}}(x_r) \tag{4.24}$$

where

$$\kappa_r^{(i_r)} = \frac{\omega_r^{(i_r)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(i_r)})}{\sum_{l=1}^{N} \omega_r^{(l)} p(\xi_{r+1}^{i_{r+1}}|\xi_r^{(l)})}, \quad r = k, k+1, \ldots, n-1 \tag{4.25}$$

one can achieve a particle estimate of (4.22) :

$$\begin{aligned} \hat{\phi}_{\nu,k:n|n}(f) = \sum_{i_k=1}^{N} \sum_{i_{k+1}=1}^{N} \cdots \sum_{i_n=1}^{N} \omega_n^{i_n} \prod_{r=k}^{n-1} \frac{\omega_r^{(i_r)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(i_r)})}{\sum_{l=1}^{N} \omega_r^{(l)} p(\xi_{r+1}^{i_{r+1}}|\xi_r^{(l)})} \\ \times f(\xi_k^{(i_k)}, \xi_{k+1}^{(i_{k+1})}, \ldots, \xi_n^{(i_n)}), \end{aligned} \tag{4.26}$$

where $\left\{\xi_r^{(i_r)}, \omega_r^{(i_r)}\right\}_{i_r=1}^{N}, r = k, \ldots, n-1$ are sets of weighted particles targeting the filtering distribution $\phi_{\nu,r}$. Note that (4.26) has not a practical interest since its complexity is exponential. Nevertheless, it is of great interest in a theoretical perspective. In fact, the deriving marginal smoother estimates inherit the convergence properties of the latter.

4.2.2.3 Particle smoother

One of the limitations of the F-B algorithm is its computational cost. In fact, it requires $O(N^2)$ operations at each time step to compute the smoothed weights. Following Godsill et al. [2004] it is easy to get smoothed distribution estimate with a linear computational effort at each time step under (4.14). Extending lemma 4.2.4 to whole time indexes one get :

Lemma 4.2.5. *The joint posterior density factorizes as :*

$$p(x_{0:n}|y_{1:n}) = p(x_n|y_{1:n}) \prod_{k=0}^{n-1} p(x_k|x_{k+1}, y_{1:k}). \tag{4.27}$$

Consider a particle estimate of the distribution $p(dx_k|x_{k+1}, y_{1:k})$:

$$\hat{p}(dx_k|x_{k+1}, y_{1:n}) = \sum_{i=1}^{N} \kappa_k^{(i)} \delta_{\xi_k^{(i)}}(dx_k) \tag{4.28}$$

where

$$\kappa_k^{(i)} = \frac{\omega_k^{(i)} p(x_{k+1}|\xi_k^{(i)})}{\sum_{j=1}^{N} \omega_k^{(j)} p(x_{k+1}|\xi_k^{(j)})}. \tag{4.29}$$

Using the particle revision, one can draw consecutive states backward in time as follows. Assume $\tilde{\xi}_{k+1:n}$ to be a random sample drawn from $p(x_{k+1:n}|y_{1:n})$. Step back in time and draw $\tilde{\xi}_k$ from $p(x_k|\tilde{\xi}_{k+1:n}, y_{1:n})$. The sample $(\tilde{\xi}_k, \tilde{\xi}_{k+1:n})$ is an approximate random realization from $p(x_{k:n}|y_{1:n})$. Iterating the mechanism down to $k = 0$, one get a random sample from the joint smoothing density. The overall algorithm is given below. This

Algorithme 12 Smoothing algorithm

1: For $i = 1, 2, \ldots, N$ choose $\tilde{\xi}_n = \xi_n^{(i)}$ with probability $\omega_n^{(i)}$
2: For $k = n-1$ down to 0 and $i = 1, 2, \ldots, N$
   — Evaluate $\kappa_k^{(i)} \propto \omega_k^{(i)} p(\tilde{\xi}_{k+1}|\xi_k^{(i)})$;
   — Choose $\tilde{\xi}_k = \xi_k^{(i)}$ with probability $\kappa_k^{(i)}$;
3: $\tilde{\xi}_{0:n}$ is an approximate random realization from $p(x_{0:n}|y_{0:n})$.

algorithm is an $O(N)$ expensive at each time step.

4.3 SMC methods in higher-order HMM

Consider the following state space model

$$\begin{cases} X_k &= a_k(X_{k-\ell:k-1}, V_k) \\ Y_k &= b_k(X_{k-\ell:k}, W_k) \end{cases} \tag{4.30}$$

where $a_k(\cdot,\cdot)$ and $b_k(\cdot,\cdot)$ are possibly non-linear functions, $\{X_k\}$ is an $\ell-$order Markov chain with initial state sequences $X_{-\ell:-1}$ distributed according to a diffuse prior distribution $\nu(\cdot)$ and transition kernel M from $(\mathcal{X}^\ell, \mathcal{B}(\mathcal{X})^{\otimes\ell})$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We assume that M admits a density function m w.r.t a dominating measure λ. $(V_k)_{k\geq 0}$ and $(W_k)_{k\geq 0}$ are i.i.d disturbance noises possibly correlated with $corr(V_i, W_j) = \rho 1_{i=j}$ and independent of $X_{-\ell:-1}$. We also assume that the observation process $\{Y_k\}$, constructed on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is conditionally independent given $\{X_k\}$ with a marginal distribution admitting a density function g such that

$$\forall A \in \mathcal{B}(\mathcal{Y}), \quad \mathbb{P}(Y_k \in A | X_{k-\ell:k-1}, X_k) = \int_A g(X_{k-\ell:k-1}, X_k, y)\mu(dy),$$

where μ is a $\sigma-$finite measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. For the sake of simplicity, the data are fixed that is, $Y_k = y_k$ for all time indexes. We also omit the dependence of likelihood function g to the data by using the short hand notation $g_k(\underline{x}_{k-1}, x_k) := g(x_{k-\ell:k-1}, x_k, \cdot)$, with $\underline{x}_{k-1} := x_{k-\ell:k-1}$. To sum up

$$\begin{cases} X_{-\ell:-1} \sim \nu(\cdot) \\ X_k | \underline{X}_{k-1} \sim m(\underline{x}_{k-1}, \cdot) \qquad k \geq 0 \\ Y_k | \underline{X}_{k-1}, X_k \sim g_k(\underline{x}_{k-1}, x_k) \end{cases} \tag{4.31}$$

4.3.1 $\ell-$order Filtering recursions

Consider the problem of computing recursively in time the following quantity :

$$\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) := \mathbb{E}_\nu \left[f(X_{k:k+\ell-1}) | Y_{0:k+\ell-1} \right] \tag{4.32}$$

where $-\ell + 1 \leq k \leq n - \ell$, for any $f \in \mathbf{F}_b(\mathcal{X}^\ell)$. Notice that on taking $\ell = 1$ and $\rho = 0$, we fall in the classical nonlinear filtering problem in $1-$order HMM. Since we deal with state sequences, we shall call it in the sequel an ℓ-filtering problem and the resulting particle solution as an $\ell-$ particle filter to emphasize the overlapping $\ell-$ size vectors in concern. A common way to approximate such a distribution is to use a cloud of weighted particles $\left\{ \xi^{(i)}_{k:k+\ell-1}, \omega^{(i)}_{k+\ell-1} \right\}_{i=1}^N$ through the estimate :

$$\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(dz_{1:\ell}) = \Omega^{-1}_{k+\ell-1} \sum_{i=1}^N \omega^{(i)}_{k+\ell-1} \delta_{\xi^{(i)}_{k:k+\ell-1}}(dz_{1:\ell}) \tag{4.33}$$

where $\Omega_{k+\ell-1} = \sum_{i=1}^N \omega^{(i)}_{k+\ell-1}$ and $\omega^{(i)}_{k+\ell-1}$ is obtained within an importance sampling procedure. The following result give a way to solve the ℓ-filtering problem recursively in time.

Proposition 4.3.1. *For any index $-\ell \leq k \leq n-\ell$ and $f \in \mathbf{F}_b(\mathcal{X}^\ell)$, the distribution $\phi_{\nu,k:k+\ell-1|k+\ell-1}$ satisfies the recursive relation :*

$$\begin{aligned}\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) &\propto \int_{\mathcal{X}^{\ell+1}} f(x_{k:k+\ell-1})\phi_{\nu,k-1:k+\ell-2|k+\ell-2}(\underline{x}_{k+\ell-2}) \\ &\times m(\underline{x}_{k+\ell-2}, x_{k+\ell-1})g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1})dx_{k-1:k+\ell-1}\end{aligned} \tag{4.34}$$

with the convention $\phi_{\nu,-\ell:-1|-1} := \nu$.

Proof. It suffices to see that :

$$\begin{aligned}&p(x_{k:k+\ell-1}|y_{0:k+\ell-1}) \\ &= \int_{\mathcal{X}} p(x_{k-1:k+\ell-1}|y_{0:k+\ell-1})dx_{k-1} \propto \int_{\mathcal{X}} p(x_{k-1:k+\ell-1}, y_{0:k+\ell-1})dx_{k-1} \\ &\propto \int_{\mathcal{X}} p(x_{k-1:k+\ell-2}|y_{0:k+\ell-2})m(\underline{x}_{k+\ell-2}; x_{k+\ell-1})g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1})dx_{k-1}\end{aligned}$$

so that,

$$\begin{aligned}&\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) \\ &\propto \int_{\mathcal{X}^\ell} f(x_{k:k+\ell-1}) \left(\int_{\mathcal{X}} p(x_{k-1:k+\ell-1}, y_{0:k+\ell-1})dx_{k-1} \right) dx_{k:k+\ell-1} \\ &\propto \int_{\mathcal{X}^{\ell+1}} f(x_{k:k+\ell-1})\phi_{\nu,k-1:k+\ell-1|k+\ell-2}(x_{k-1:k+\ell-2}) \\ &\times m(\underline{x}_{k+\ell-2}; x_{k+\ell-1})g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1})dx_{k-1:k+\ell-1}.\end{aligned} \tag{4.35}$$

which leads to the result. □

In order to highlight the two-step procedure mentioned above, the following operator formulation is given :

$$\phi_{\nu,k:k+\ell-1|k+\ell-2} = \phi_{\nu,k-1:k+\ell-2|k+\ell-2}M \tag{4.36}$$

as the prediction step and

$$\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) = \frac{\phi_{\nu,k:k+\ell-1|k+\ell-2}(fg_{k+\ell-1})}{\phi_{\nu,k:k+\ell-1|k+\ell-2}(g_{k+\ell-1})} \tag{4.37}$$

as the correction step, for any $f \in \mathbf{F}_b(\mathcal{X}^\ell)$ and $-\ell+1 \leq k \leq n-\ell$. At time $(k+\ell-2)$, assume one has a cloud of weighted sample $\left\{\xi^{(i)}_{k-1:k+\ell-2}, \omega^{(i)}_{k+\ell-2}\right\}_{i=1}^N$ approximating the ℓ-filter distribution $\phi_{\nu,k-1:k+\ell-2|k+\ell-2}$ in the sense

$$\hat{\phi}_{\nu,k-1:k+\ell-2|k+\ell-2}(dx_{k-1:k+\ell-2}) = \Omega^{-1}_{k+\ell-2}\sum_{i=1}^N \omega^{(i)}_{k+\ell-2}\delta_{\xi^{(i)}_{k-1:k+\ell-2}}(dx_{k-1:k+\ell-2}) \tag{4.38}$$

where $\Omega_{k+\ell-2} = \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)}$. A particle estimate at the next time step $k+\ell-1$ of the ℓ-filter distribution is achieved by :

$$\begin{aligned}
&\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(f) \\
&\propto \int_{\mathcal{X}} \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} f(\xi_{k:k+\ell-2}^{(i)}, x_{k+\ell-1}) \omega_{k+\ell-2}^{(i)} m(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) \\
&\times g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) dx_{k+\ell-1} \qquad (4.39) \\
&= \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)} \int_{\mathcal{X}} f(\xi_{k:k+\ell-2}^{(i)}, x_{k+\ell-1}) m(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) \\
&\times g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) dx_{k+\ell-1}.
\end{aligned}$$

where the last integral of (4.39) can be thought as expectation under either the transition density function or the likelihood density function. Note also that a mixture argument can be considered to evaluate it. Notice that these recursive weights are obtained within a classical bootstrap filter Gordon et al. [1993] or the general framework of the auxiliary particle filter Pitt and Shephard [1999]. Following Douc et al. [2011], the auxiliary filter proceeds as follows. Let $\left\{\xi_{-\ell:-1}^{(i)}\right\}_{i=1}^{N}$ be i.i.d random variables distributed according to an instrumental distribution χ. Define the associated unnormalized importance weights $\omega_{-1}^{(i)} := \frac{d\nu}{d\chi}(\xi_{-\ell:-1}^{(i)})$. With the weighted sample $\left\{\xi_{-\ell:-1}^{(i)}, \omega_{-1}^{(i)}\right\}_{i=1}^{N}$ one can approximate $\phi_{\nu,-\ell:-1|-1}$ by :

$$\hat{\phi}_{\nu,-\ell:-1|-1}(f) = \frac{\sum_{i=1}^{N} \omega_{-1}^{(i)} f(\xi_{-\ell:-1}^{(i)})}{\sum_{i=1}^{N} \omega_{-1}^{(i)}} \qquad (4.40)$$

for any $f \in \mathbf{F}_b(\mathcal{X}^\ell)$. Assume one has a weighted sample $\left\{\xi_{k-1:k+\ell-2}^{(i)}, \omega_{k+\ell-2}^{(i)}\right\}_{i=1}^{N}$ targeting $\phi_{\nu,k-1:k+\ell-2|k+\ell-2}$ in the same way and consider the auxiliary target distribution

$$\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}^{a}(i,f) := \frac{\omega_{k+k-2}^{(i)} M(\xi_{k-1:k+\ell-2}^{(i)} g_{k+\ell-1} f)}{\sum_{i=1}^{N} \omega_{k+k-2}^{(i)} M(\xi_{k-1:k+\ell-2}^{(i)} g_{k+\ell-1} f)} \qquad (4.41)$$

on the space $\{1, 2, \cdots, N\} \times \mathcal{X}^\ell$, for any $f \in \mathbf{F}_b(\mathcal{X}^\ell)$. One can notice that the target distribution (4.37) is a marginal of (4.41) on integrating out the indexes. Thus, to produce samples from the target distribution one has in a first stage to simulate random samples according to (4.41) and then cancel the indexes. Clearly, consider the random sample $\left\{I_{k+\ell-1}^{(i)}, \xi_{k:k+\ell-1}^{(i)}\right\}_{i=1}^{N}$ simulated from the instrumental distribution

$$\pi_{k:k+\ell-1|k+\ell-1}(i,f) \propto \omega_{k+\ell-2}^{(i)} \vartheta_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}) Q_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, f) \qquad (4.42)$$

on the space $\{1,2,\cdots,N\}\times\mathcal{X}^{\ell}$ where $\left\{\vartheta_{k+\ell-1}(\xi^{(i)}_{k-1:k+\ell-2})\right\}_{i=1}^{N}$ is the *adjustment multiplier* weights and $Q_{k+\ell-1}$ is an $\ell-$order Markov transition kernel taken as *proposal* with transition kernel density w.r.t to λ denoted by $q_{k+\ell-1}(\underline{x},\cdot)$. At each particle position $\xi^{(i)}_{k:k+\ell-1}$ we define the corresponding importance weight by

$$\omega^{(i)}_{k+\ell-1} := \frac{m(\xi^{(I^{(i)}_{k+\ell-1})}_{k-1:k+\ell-2},\xi^{(i)}_{k+\ell-1})g_{k+\ell-1}(\xi^{(i)}_{k+\ell-1})}{\vartheta_{k+\ell-1}(\xi^{(I^{(i)}_{k+\ell-1})}_{k-1:k+\ell-2})q_{k+\ell-1}(\xi^{(I^{(i)}_{k+\ell-1})}_{k-1:k+\ell-2},\xi^{(i)}_{k+\ell-1})} \tag{4.43}$$

such that

$$\omega^{(i)}_{k+\ell-1} \propto \frac{d\hat{\phi}^{a}_{\nu,k:k+\ell-1|k+\ell-1}}{d\pi_{k:k+\ell-1|k+\ell-1}}(I^{(i)}_{k+\ell-1},\xi^{(i)}_{k:k+\ell-1}) \tag{4.44}$$

Finally, the index are discarded and $\left\{\xi^{(i)}_{k:k+\ell-1},\omega^{(i)}_{k+\ell-1}\right\}_{i=1}^{N}$ is a weighted sample that targets $\phi_{\nu,k:k+\ell-1|k+\ell-1}$, the distribution of interest. Moreover on setting $\vartheta_{k+\ell-1}(\underline{x})\equiv 1$ and $q_{k+\ell-1}(\underline{x},\cdot)\equiv m(\underline{x},\cdot)$ for all $\underline{x}\in\mathcal{X}^{\ell}$ we get the classical *Bootstrap filter* in $\ell-$order HMM. We also define $\mathcal{F}^{N}_{k}$ as the $\sigma-$field conjointly generated by the data and the particles at time $k\geq 0$ by

$$\mathcal{F}^{N}_{k} := \sigma\left\{Y_{0:n},\left(\xi^{(i)}_{s-\ell+1:s},\omega^{(i)}_{s}\right)_{i=1}^{N},0\leq s\leq k\right\}\vee\sigma\left\{\left(\xi^{(i)}_{-\ell:-1},\omega^{(i)}_{-1}\right)_{i=1}^{N}\right\} \tag{4.45}$$

4.3.2 ℓ-order smoothing recursions

Before stating ℓ-order smoothing, we precise some smoothing quantities that can be easily handled :

$$\phi_{\nu,k|n}(f) := \mathbb{E}_{\nu}\left[f(X_k)\middle|Y_{0:n}\right], k<n, \tag{4.46}$$

$$\phi_{\nu,m,p|n}(g) := \mathbb{E}_{\nu}\left[g(X_p,X_m)\middle|Y_{0:n}\right], \mid p-m\mid\leq \ell+1, \tag{4.47}$$

$$\phi_{\nu,-\ell:n|n}(h) := \mathbb{E}_{\nu}\left[h(X_{-\ell:n})\middle|Y_{0:n}\right], \tag{4.48}$$

for any $f\in\mathbf{F}_b(\mathcal{X})$, $g\in\mathbf{F}_b(\mathcal{X}^2)$ and $h\in\mathbf{F}_b(\mathcal{X}^{n+\ell+1})$. Since (4.46) and (4.47) are particular cases of (4.48) we do not mention them here. In order to derive similar recursions as in $1-$order HMM, one needs to give the reversed time dynamic of the Markov chain through backward transition kernels. The following result shows that the hidden process still Markovian backward in time given the data.

Proposition 4.3.2. *Let ν be an initial distribution on $X_{-\ell:-1}$, $f\in\mathbf{F}_b(\mathcal{X})$, $n>0$ and $-\ell+1\leq p\leq n-\ell$. Then $\{X_{n-k}\}_{k\geq 0}$ is a Markov chain with backward transition kernels*

defined by :

$$\begin{aligned} B_{\nu,p+\ell-1}(X_{p+1:p+\ell}, f) &:= \mathbb{E}_\nu \left[f(X_p) \middle| X_{p+1:n}, Y_{0:n} \right] \\ &= \mathbb{E}_\nu \left[f(X_p) \middle| X_{p+1:p+\ell}, Y_{0:p+\ell-1} \right] \end{aligned} \tag{4.49}$$

Proof. see section 4.6 □

4.3.2.1 Joint smoothing

To deal with (4.48) one needs the following factorization.

Lemma 4.3.3. *For any function $f \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$, the joint smoothing distribution satisfies the backward kernels decomposition :*

$$\begin{aligned} \phi_{\nu,-\ell:n|n}(f) &= \int_{\mathcal{X}^{n+\ell+1}} f(x_{-\ell:n}) B_{\nu,-1}(x_{-\ell+1:0}, dx_{-\ell}) \phi_{\nu,-\ell+1:n|n}(dx_{-\ell+1:n}) \\ &= \int_{\mathcal{X}^{n+\ell+1}} f(x_{-\ell:n}) \phi_{\nu,n-\ell+1:n|n}(dx_{n-\ell+1:n}) \\ &\times \prod_{p=-\ell}^{n-\ell} B_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p). \end{aligned} \tag{4.50}$$

To get a particle estimate of (4.48), one needs to run the following two steps. In the first step, the $\ell-$filter distributions are approximated by

$$\hat{\phi}_{\nu,p:p+\ell-1|p+\ell-1}(dx_{p:p+\ell-1}) = \Omega_{p+\ell-1}^{-1} \sum_{i_p=1}^{N} \omega_{p+\ell-1}^{(i_p)} \delta_{\xi_{p:p+\ell-1}^{(i_p)}}(dx_{p:p+\ell-1}), \tag{4.51}$$

with $\left\{ \omega_{p+\ell-1}^{(i_p)}, \xi_{p:p+\ell-1}^{(i_p)} \right\}_{i_p=1}^{N}$ being the targeting weighted samples of the $\ell-$filter distributions $\phi_{\nu,p:p+\ell-1|p+\ell-1}(dz_{1:\ell})$, $p = -\ell, -\ell+1, \ldots, n-\ell$. The second step consists in approximating the backward kernels $B_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p)$ by :

$$\hat{B}_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p) = \sum_{i_p=1}^{N} \frac{\omega_{p+\ell-1}^{(i_p)} m(\xi_{p:p+\ell-1}^{(i_p)}, x_{p+\ell})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, x_{p+\ell})} \delta_{\xi_p^{(i_p)}}(dx_p) \tag{4.52}$$

$p = -\ell, -\ell+1, \ldots, n-\ell$. Plugging (4.51) and (4.52) into (4.50), a particle estimate of (4.48) is given by :

$$\begin{aligned} \hat{\phi}_{\nu,-\ell:n|n}(f) = \Omega_n^{-1} \sum_{i_n=1} \Bigg[& \sum_{i_{-\ell}=1}^{N} \cdots \sum_{i_{n-\ell}=1}^{N} f(\xi_{-\ell}^{(i_{-\ell})}, \ldots, \xi_{n-\ell}^{(i_{n-\ell})}, \xi_{n-\ell+1:n}^{(i_n)}) \\ & \times \prod_{p=-\ell}^{n-\ell} \frac{\omega_{p+\ell-1}^{(i_p)} m(\xi_{p:p+\ell-1}^{(i_p)}, \xi_{p+\ell}^{(i_{p+\ell})})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, \xi_{p+\ell}^{(i_{p+\ell})})} \Bigg] \omega_n^{(i_n)}, \end{aligned} \tag{4.53}$$

for $f \in \mathbf{F}_b\left(\mathcal{X}^{n+\ell+1}\right)$. Before moving towards the theoretical properties of this estimator, we give a summary description of the former procedure. Once the two passes performed,

Algorithme 13 Smoothing in $\ell-$order HMM

1: Forward pass : For $p = -\ell, -\ell+1, \ldots, n-\ell+1$ approximate $\phi_{\nu,p:p+\ell-1|p+\ell-1}$ by

$$\hat{\phi}_{\nu,p:p+\ell-1|p+\ell-1}(dx_{p:p+\ell-1}) = \Omega_{p+\ell-1}^{-1} \sum_{i=1}^{N} \omega_{p+\ell-1}^{(i)} \delta_{\xi_{p:p+\ell-1}^{(i)}}(dx_{p:p+\ell-1})$$

2: Backward pass : For $p = n-\ell$ down to $-\ell$ approximate $B_{\nu,p+\ell-1}$ by :

$$\hat{B}_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p) = \sum_{i_p=1}^{N} \frac{\omega_{p+\ell-1}^{(i_p)} m(\xi_{p:p+\ell-1}^{(i_p)}, x_{p+\ell})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, x_{p+\ell})} \delta_{\xi_p^{(i_p)}}(dx_p)$$

one can approximate (4.48) using (4.53) and deduce approximation for (4.46) and (4.47) as marginal of the latter.

4.3.2.2 Particle smoother

One may also achieve similar particle smoother to those of Godsill et al. [2004] using the following identity.

Lemma 4.3.4. *Under Prop.4.3.2, the joint smoothing density factorizes as*

$$p(x_{-\ell:n}|y_{0:n}) = p(x_{n-\ell+1:n}|y_{0:n}) \prod_{k=-\ell}^{n-\ell} p(x_k|x_{k+1:n}; y_{0:n}) \tag{4.54}$$

where

$$\begin{aligned} p(x_k|x_{k+1:n}, y_{0:n}) &= p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \\ &\propto p(x_{k+\ell}|x_{k:k+\ell-1}) p(x_{k:k+\ell-1}|y_{0:k+\ell-1}). \end{aligned}$$

Assume one has run the $\ell-$order filter mentioned previously to get the weighted particles

$$\left\{\omega_{k+\ell-1}^{(i)}, \xi_{k:k+\ell-1}^{(i)}\right\}_{i=1}^{N}, \quad -\ell+1 \leq k \leq n-\ell$$

approximating the $\ell-$filter densities $p(x_{k:k+\ell-1}|y_{0:k+\ell-1})$. Using the previous weighted sample one could get a particle estimate of $p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1})$:

$$p(dx_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \approx \sum_{i=1}^{N} \kappa_k^{(i)} \delta_{\xi_k^{(i)}}(dx_k) \tag{4.55}$$

where the modified weights are given by

$$\kappa_k^{(i)} = \frac{\omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, x_{k+\ell})}{\sum_{j=1}^{N} \omega_{k+\ell-1}^{(j)} m(\xi_{k:k+\ell-1}^{(j)}, x_{k+\ell})} \tag{4.56}$$

With these modified weights, one can simulate consecutive states in the reverse-time as follows. Let $\tilde{x}_{k+1:n}$ be a random sample drawn from $p(x_{k+1:n}|y_{0:n})$, step back in time and

Algorithme 14 Particle smoother in $\ell-$order HMM

1: Choose $\tilde{\xi}_{n-\ell+1:n} = \xi_{n-\ell+1:n}^{(i)}$ with probability $\omega_n^{(i)}$
2: For $k = n-\ell$ down to $-\ell$ do
   — Evaluate $\kappa_k^{(i)} \propto \omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, \tilde{\xi}_{k+\ell})$, for $i = 1, \ldots, N$;
   — Choose $\tilde{\xi}_k = \xi_k^{(i)}$ with probability $\kappa_k^{(i)}$
3: EndFor
4: $\tilde{\xi}_{-\ell:n} = (\tilde{\xi}_{-\ell}, \tilde{\xi}_{-\ell+1}, \ldots, \tilde{\xi}_n)$ is an approximate random realization from $p(x_{-\ell:n}|y_{0:n})$.

draw $\tilde{x}_k$ from $p(x_k|\tilde{x}_{k+1:n}, y_{0:n})$. The pair $(\tilde{x}_k, \tilde{x}_{k+1:n})$ is an approximate random realization of $p(x_{k:n}|y_{0:n})$. Iterating this mechanism backward in time one gets the smoothing algorithm 4.3.2.2. The computational complexity is $O(N)$ at each time step which compares favorably to the $O(N^2)$ of the marginal smoothing.

4.4 Parameter estimation

MCEM as a combination of the GEM with SMC is a tool that can be used to estimate HMM when dealing with latent process. We do not fully detail the GEM algorithm since it is well documented (see Dempster et al. [1977] or McLachlan and Krishnan [2008] for a review). However, the main idea is depicted below :

Algorithme 15 Generalized EM algorithm

1: Choose an initial guess $\theta^{(0)}$
2: **For** $m = 1, 2, \ldots$ **do**
   1. **E-Step :** Compute $Q(\theta, \theta^{(m-1)})$
   2. **M-Step :** Find $\theta^{(m)}$ s.t $Q(\theta^{(m)}, \theta^{(m-1)}) \geq Q(\theta^{(m-1)}, \theta^{(m-1)})$
3: **EndFor.**

The E-step consists in computing the *intermediate quantity*, that is the conditional expectation of the logarithm of the complete data likelihood given the data and the current value of the parameter vector $\theta^{(m-1)}$:

$$Q(\theta^{(m)}, \theta^{(m-1)}) = \mathbb{E}_{\theta^{(m-1)}} \left[\log p_{\theta^{(m)}} (X_{-2:n}, Y_{0:n}) \,|\, Y_{0:n} \right] \tag{4.57}$$

where $(n+1)$ is the sample size of the data indexed from 0 to n, $p_{\theta^{(m)}}$ a generic notation for densities depending on parameter $\theta^{(m)}$ and X is a hidden signal initialized to a diffuse prior distribution ν on $X_{-2:-1}$ and Y the observation process. As a first illustration, consider the following toy example

$$\begin{cases} X_k = \pi_1 X_{k-1} + \pi_2 X_{k-2} + \sigma_W W_k \\ \\ Y_k = X_k + \sigma_V V_k, \end{cases} \tag{4.58}$$

We assume that $(V_k, W_k)_{k\geq 0}$ are i.i.d and independent of $X_{-2:-1}$ with $(V_k, W_k) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$, where $|\pi_1 \pm \pi_2| < 1$ and $|\pi_2| < 1$ ensuring the stationarity of X. At iteration m of the GEM, the parameters are updated through the recursive scheme

$$\begin{cases} \pi_1^{(m)} & = \dfrac{\sum_{k=2}^n \mathbb{E}_{\theta^{(m-1)}}\left[\left(X_{k-1}X_k - \pi_2^{(m)} X_{k-1}X_{k-2}\right)\middle| Y_{0:n}\right]}{\sum_{k=1}^n \mathbb{E}_{\theta^{(m-1)}}\left[X_{k-1}^2 | Y_{0:n}\right]} \\ \\ \pi_2^{(m)} & = \dfrac{\sum_{k=2}^n \mathbb{E}_{\theta^{(m-1)}}\left[\left(X_{k-2}X_k - \pi_1^{(m)} X_{k-1}X_{k-2}\right)\middle| Y_{0:n}\right]}{\sum_{k=2}^n \mathbb{E}_{\theta^{(m-1)}}\left[X_{k-2}^2 | Y_{0:n}\right]} \\ \\ \left[\sigma_V^{(m)}\right]^2 & = \frac{1}{n}\sum_{k=2}^n \mathbb{E}_{\theta^{(m-1)}}\left[\left(Y_k - X_k\right)^2 \middle| Y_{0:n}\right] \\ \\ \left[\sigma_W^{(m)}\right]^2 & = \frac{1}{n}\sum_{k=2}^n \mathbb{E}_{\theta^{(m-1)}}\left[\left(X_k - \pi_1^{(m)} X_{k-1} - \pi_2^{(m)} X_{k-2}\right)^2 \middle| Y_{0:n}\right] \end{cases}$$

As a second illustration, consider the following discrete stochastic volatility model

$$\begin{cases} X_k = \pi_1 X_{k-1} + \pi_2 X_{k-2} + \sigma W_k \\ \\ Y_k = \beta \exp\left(X_k/2\right) V_k, \end{cases} \tag{4.59}$$

under the same assumptions as in the former model. Since (V_k) and (W_k) are independent and Gaussian it's common to use a linearized version of (4.59) given by :

$$\begin{cases} X_{k+1} = \pi_1 X_k + \pi_2 X_{k-1} + \sigma W_{k+1} \\ \\ Y'_{k+1} = \alpha + X_{k+1} + \eta_{k+1} - \zeta \end{cases} \tag{4.60}$$

where $Y'_{k+1} := \log Y_{k+1}^2$, $\eta_k := \log V_k^2$ are i.i.d noises independent of (W_k) with a $\log\chi^2(1)$ distribution, $\zeta := \mathbb{E}(\log V_k^2) = -1.27049$, $\alpha := \log\beta^2 + \zeta$ and $\theta := (\pi_1, \pi_2, \sigma, \alpha)$

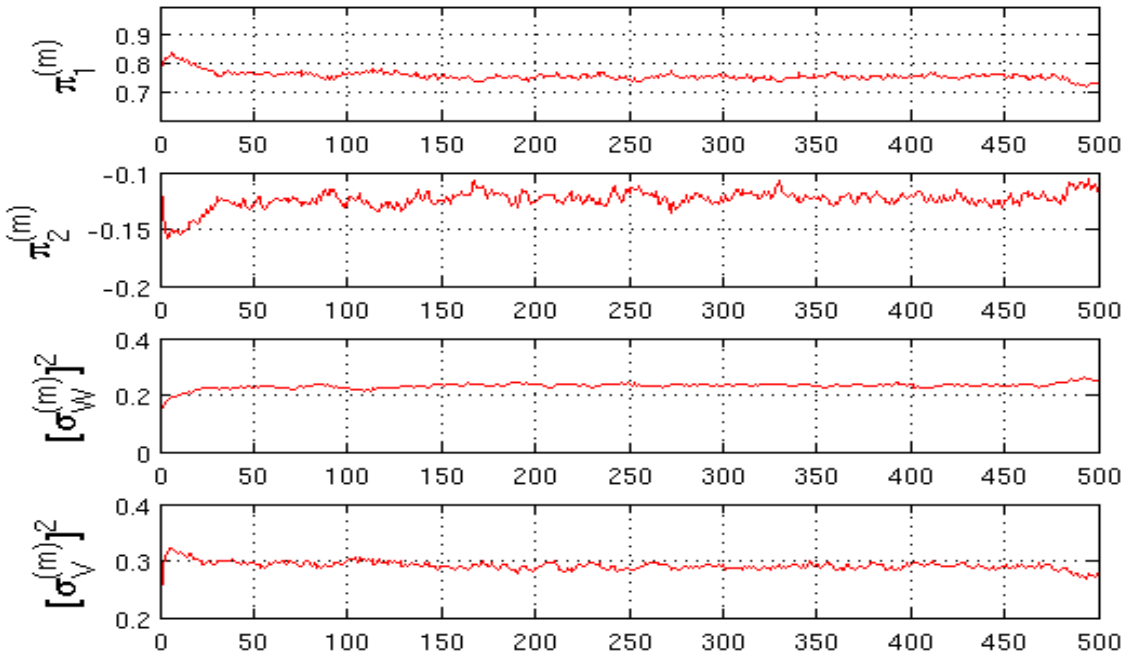


FIGURE 4.1 – MCEM iterations for (4.58) generated with $\theta^* = (0.7, -0.15, 0.2, 0.3)$

is the parameter vector. On taking the derivatives of $Q(\theta^{(m)}, \theta^{(m-1)})$ with respect to each parameter on gets the recursive following parameter update :

$$\begin{cases} \alpha^{(m)} &= \log\left[\frac{1}{n}\sum_{k=0}^{n}\mathbb{E}_{\theta^{(m-1)}}[\exp(Y_k - X_k + \zeta)|Y'_{0:n}]\right] \\ \pi_1^{(m)} &= \dfrac{\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\left(X_{k-1}X_k - \pi_2^{(m)}X_{k-1}X_{k-2}\right)\middle|Y'_{0:n}\right]}{\sum_{k=1}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[X_{k-1}^2|Y'_{0:n}\right]} \\ \pi_2^{(m)} &= \dfrac{\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\left(X_{k-2}X_k - \pi_1^{(m)}X_{k-1}X_{k-2}\right)\middle|Y'_{0:n}\right]}{\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[X_{k-2}^2|Y'_{0:n}\right]} \\ \sigma^{(m)} &= \sqrt{\frac{1}{n}\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\left(X_k - \pi_1^{(m)}X_{k-1} - \pi_2^{(m)}X_{k-2}\right)^2\middle|Y'_{0:n}\right]} \end{cases} \tag{4.61}$$

As a synthetic example, figure (4.2) is generated using the true parameter vector $(\pi_1^* = 0.8, \pi_2^* = 0.1, \sigma^* = \sqrt{0.3}, \log[\beta^*]^2 = -0.8612)$.

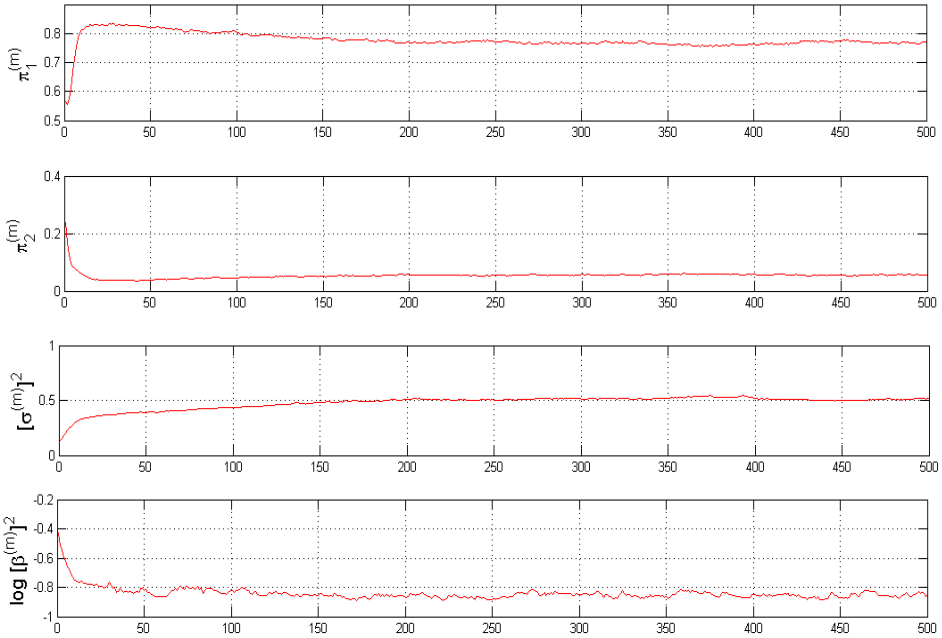


FIGURE 4.2 – MCEM iterations for 4.60 with $\theta^* = (0.8, 0.1, \sqrt{0.3}, -0.8612)$

As point perspective, the adjustment of the smoothing weights is required. Indeed, one can notice that the estimates are not entirely satisfactory for some parameters in this example. More attention is needed to correct this shortcoming.

4.5 Convergence issues

In this section, we are concerned with the effective convergence of (4.39) and (4.53). In other to deal with these estimates we mimic some results by Douc et al. [2011] and Dubarry and Le Corff [2010] to the context of higher-order HMM.

4.5.1 L_q mean error

Let's consider the following three kernels and a particular smoothing error decomposition that will be made clear latter. For $k \geq 0$, define the kernel $\mathbb{L}_{k,n} : \mathcal{X}^{k+\ell+1} \times \mathcal{B}(\mathcal{X})^{\otimes(n+\ell+1)} \to [0,1]$ such that :

$$\mathbb{L}_{k,n}(x_{-\ell:k}, h) := \int_{\mathcal{X}^{n-k}} \prod_{u=k+1}^{n} M(\underline{x}_{u-1}, dx_u) g_u(\underline{x}_{u-1}, x_u) h(x_{-\ell:n}). \tag{4.62}$$

with $\mathbb{L}_{n,n}(x_{-\ell:n}, h) := h(x_{-\ell:n})$, where $h \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$. Using, (4.62), for any $0 \leq k \leq n$

$$\begin{aligned}\hat{\phi}_{\nu,-\ell:k|k}\left[\mathbb{L}_{k,n}(x_{-\ell:k}, h)\right] &= \int_{\mathcal{X}^{k+\ell+1}} \hat{\phi}_{\nu,-\ell:k|k}(dx_{-\ell:k})\mathbb{L}_{k,n}(x_{-\ell:k}, h) \\ &= \int_{\mathcal{X}^{k+\ell+1}} \hat{\phi}_{\nu,k-\ell+1:k|k}(dx_{k-\ell+1:k}) \\ &\times \prod_{r=-\ell}^{k-\ell} \hat{B}_{r+\ell-1}(x_{r+1:r+\ell}, dx_r)\mathbb{L}_{k,n}(x_{-\ell:k}, h) \\ &= \int_{\mathcal{X}^{\ell}} \hat{\phi}_{\nu,k-\ell+1:k|k}(dx_{k-\ell+1:k})\hat{\mathfrak{L}}_{k,n}(x_{k-\ell+1:k}, h)\end{aligned} \tag{4.63}$$

where we define the kernels $\hat{\mathfrak{L}}_{k,n}$ and $\mathfrak{L}_{k,n}$ on $\mathcal{X}^{\ell} \times \mathcal{B}(\mathcal{X})^{\otimes(n+\ell+1)}$ respectively by :

$$\hat{\mathfrak{L}}_{k,n}(x_{k-\ell+1:k}, h) := \int_{\mathcal{X}^{k+1}} \prod_{r=-\ell}^{k-\ell} \hat{B}_{r+\ell-1}(x_{r+1:r+\ell}, dx_r)\mathbb{L}_{k,n}(x_{-\ell:k}, h) \tag{4.64}$$

and

$$\mathfrak{L}_{k,n}(x_{k-\ell+1:k}, h) := \int_{\mathcal{X}^{k+1}} \prod_{r=-\ell}^{k-\ell} B_{r+\ell-1}(x_{r+1:r+\ell}, dx_r)\mathbb{L}_{k,n}(x_{-\ell:k}, h) \tag{4.65}$$

for all $x_{k-\ell+1:k} \in \mathcal{X}^{\ell}$. Applying Bayes rule one could notice that :

$$\phi_{\nu,-\ell:n|n}(h) = \frac{\int_{\mathcal{X}^{n+\ell+1}} \nu(dx_{-\ell:-1}) \prod_{u=0}^{n} M(\underline{x}_{u-1}, dx_u) g_u(\underline{x}_{u-1}, x_u) h(x_{-\ell:n}) dx_{-\ell:n}}{\int_{\mathcal{X}^{n+\ell+1}} \nu(dx_{-\ell:-1}) \prod_{u=0}^{n} M(\underline{x}_{u-1}, dx_u) g_u(\underline{x}_{u-1}, x_u) dx_{-\ell:n}} \tag{4.66}$$

so that,

$$\phi_{\nu,-\ell:n|n}(h) = \frac{\phi_{\nu,-\ell:k|k}\left[\mathbb{L}_{k,n}(., h)\right]}{\phi_{\nu,-\ell:k|k}\left[\mathbb{L}_{k,n}(., 1)\right]}, \quad \forall k \geq 0. \tag{4.67}$$

Consider the smoothing error

$$\Delta_n^N(h) := \hat{\phi}_{\nu,-\ell:n|n}(h) - \phi_{\nu,-\ell:n|n}(h), \quad h \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1}),$$

with the convention

$$\frac{\hat{\phi}_{\nu,-\ell:-1|-1}\left[\mathbb{L}_{-1,n}(., h)\right]}{\hat{\phi}_{\nu,-\ell:-1|-1}\left[\mathbb{L}_{-1,n}(., 1)\right]} = \frac{\phi_{\nu,-\ell:0|0}\left[\mathbb{L}_{0,n}(., h)\right]}{\phi_{\nu,-\ell:0|0}\left[\mathbb{L}_{0,n}(., 1)\right]}.$$

The following decomposition plays a key role in the sequel.

Lemma 4.5.1. *For any $h \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$, the smoothing error expands as*

$$\begin{aligned}\Delta_n^N(h) &= \sum_{k=0}^{n} \left\{ \frac{\hat{\phi}_{\nu,-\ell:k|k}\left[\mathbb{L}_{k,n}(\cdot,h)\right]}{\hat{\phi}_{\nu,-\ell:k|k}\left[\mathbb{L}_{k,n}(\cdot,1)\right]} - \frac{\hat{\phi}_{\nu,-\ell:k-1|k-1}\left[\mathbb{L}_{k-1,n}(\cdot,h)\right]}{\hat{\phi}_{\nu,-\ell:k-1|k-1}\left[\mathbb{L}_{k-1,n}(\cdot,1)\right]} \right\} \\ &= \sum_{k=0}^{n} \left\{ \frac{\hat{\phi}_{\nu,k-\ell+1:k|k}\left[\hat{\mathfrak{L}}_{k,n}(\cdot,h)\right]}{\hat{\phi}_{\nu,k-\ell+1:k|k}\left[\hat{\mathfrak{L}}_{k,n}(\cdot,1)\right]} - \frac{\hat{\phi}_{\nu,k-\ell:k-1|k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot,h)\right]}{\hat{\phi}_{\nu,k-\ell:k-1|k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot,1)\right]} \right\} \\ &= \sum_{k=0}^{n} \frac{N^{-1}\sum_{i=1}^{N} \omega_k^{(i)} \hat{G}_{k,n}(\xi_{k-\ell+1:k}^{(i)}, h)}{N^{-1}\sum_{i=1}^{N} \omega_k^{(i)} \mathfrak{L}_{k,n}(\xi_{k-\ell+1:k}^{(i)}, 1)}\end{aligned} \tag{4.68}$$

where $\hat{G}_{k,n}$ is a kernel on $\mathcal{X}^{\ell} \times \mathcal{B}(\mathcal{X})^{\otimes(n+\ell+1)}$ defined by

$$\hat{G}_{k,n}(\underline{x}, h) := \hat{\mathfrak{L}}_{k,n}(\underline{x}, h) - \frac{\hat{\phi}_{\nu,k-\ell:k-1|k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot,h)\right]}{\hat{\phi}_{\nu,k-\ell:k-1|k-1}\left[\hat{\mathfrak{L}}_{k-1,n}(\cdot,1)\right]} \hat{\mathfrak{L}}_{k,n}(\underline{x}, 1) \tag{4.69}$$

for all $\underline{x} \in \mathcal{X}^{\ell}$, $h \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$ and $\{\omega_k^{(i)}, \xi_{k-\ell+1:k}^{(i)}\}_{i=1}^{N}$ a weighted sample targeting $\phi_{\nu,k-\ell+1:k|k}$.

Proof. The RHS of the first equality in (4.68) is a telescoping sum which reduces to

$$\frac{\hat{\phi}_{\nu,-\ell:n|n}\left[\mathbb{L}_{n,n}(.,h)\right]}{\hat{\phi}_{\nu,-\ell:n|n}\left[\mathbb{L}_{n,n}(.,1)\right]} - \frac{\hat{\phi}_{\nu,-\ell:-1|-1}\left[\mathbb{L}_{-1,n}(.,h)\right]}{\hat{\phi}_{\nu,-\ell:-1|-1}\left[\mathbb{L}_{-1,n}(.,1)\right]}. \tag{4.70}$$

The second equality is obtained via Lemma 4.50, since :

$$\begin{aligned}\phi_{\nu,-\ell:k|k}(f) = &\int_{\mathcal{X}^{k+\ell+1}} f(x_{-\ell:k}) \phi_{\nu,k-\ell+1:k|k}(dx_{k-\ell+1:k}) \\ &\times \prod_{r=-\ell}^{k-\ell} B_{\nu,r+\ell-1}(x_{r+1:r+\ell}, dx_r)\end{aligned} \tag{4.71}$$

for all $k \geq 0$ and $f \in \mathbf{F}_b(\mathcal{X}^{k+\ell+1})$. □

The last step is to compute upper bound for the L_q mean error decomposition.

4.6 Proof of Prop.4.3.2

One may use the following intermediate result.

Lemma 4.6.1. *For any function $f \in \mathbf{F}_b(\mathcal{X}^\ell)$ and index $k \geq 1-\ell$,*

$$\begin{aligned}&\phi_{\nu,k:k+\ell-1|k+\ell-1}(f)L_{k+\ell-1} = \\ &\int_{\mathcal{X}^{k+2\ell}} f(x_{k:k+\ell-1})\nu(x_{-\ell:-1}) \prod_{i=0}^{k+\ell-1} m(\underline{x}_{i-1}; x_i) g_i(\underline{x}_{i-1}, x_i) dx_{-\ell:k+\ell-1}\end{aligned} \tag{4.72}$$

where $L_{k+\ell-1}$ denotes the likelihood density of $y_{0:k+\ell-1}$.

Proof. It suffices to see that :

$$\begin{aligned}p(x_{k:k+\ell-1}|y_{0:k+\ell-1}) &= \int_{\mathcal{X}^{k+\ell}} p(x_{-\ell:k+\ell-1}|y_{0:k+\ell-1}) dx_{-\ell:k-1} \\ &= \int_{\mathcal{X}^{k+\ell}} \frac{p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1})}{p(y_{0:k+\ell-1})} dx_{-\ell:k-1} \\ &= L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+\ell}} p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1}) dx_{-\ell:k-1}\end{aligned} \tag{4.73}$$

Using (4.73), the expectation of $f(X_{k:k+\ell-1})$ conditional on $Y_{0:k+\ell-1}$ is given by :

$$\begin{aligned}&\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) \\ &= \int_{\mathcal{X}^\ell} f(x_{k:k+\ell-1}) \left(L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+\ell}} p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1}) dx_{-\ell:k-1} \right) dx_{k:k+\ell-1} \\ &= L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+2\ell}} f(x_{k:k+\ell-1}) p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1}) dx_{-\ell:k+\ell-1} \\ &= L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+2\ell}} f(x_{k:k+\ell-1}) \nu(x_{-\ell:-1}) \prod_{i=0}^{k+\ell-1} m(\underline{x}_{i-1}, x_i) g_i(\underline{x}_{i-1}, x_i) dx_{-\ell:k+\ell-1}\end{aligned} \tag{4.74}$$

which leads to the identity. □

Note that this identity is extensible up to the final time index n :

$$\phi_{\nu,k:n|n}(f)L_n = \int_{\mathcal{X}^{n+\ell+1}} f(x_{k:n}) \nu(x_{-\ell:-1}) \prod_{i=0}^{n} m(\underline{x}_{i-1}; x_i) g_i(\underline{x}_{i-1}, x_i) dx_{-\ell:n} \tag{4.75}$$

for any function $f \in \mathbf{F}_b(\mathcal{X}^{n-k+1})$.

Proof. From previous lemma, for any functions $e \in \mathbf{F}_b(\mathcal{X}^{\ell-1})$, $f \in \mathbf{F}_b(\mathcal{X})$ and $h \in$

$\mathbf{F}_b(\mathcal{X}^{n-k-\ell+1})$,

$$
\begin{aligned}
&\mathbb{E}\left[f(X_k)e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\Big|Y_{0:n}\right]\\
&=\int_{\mathcal{X}^{n-k+1}} f(x_k)e(x_{k+1:k+\ell-1})h(x_{k+\ell:n})\Phi_{\nu,k:n|n}(dx_{k:n})\\
&=L_n^{-1}\int_{\mathcal{X}^{k+2\ell+1}} f(x_k)e(x_{k+1:k+\ell-1})\nu(x_{-\ell:-1})\prod_{i=0}^{k+\ell-1} m(\underline{x}_{i-1},x_i)g_i(\underline{x}_{i-1},x_i)\\
&\times m(\underline{x}_{k+\ell-1},x_{k+\ell})g_{k+\ell}(\underline{x}_{k+\ell-1},x_{k+\ell})\\
&\times\left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n})\prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1},x_i)g_i(\underline{x}_{i-1},x_i)dx_{k+\ell+1:n}\right]dx_{-\ell:k+\ell}\\
&=\frac{L_{k-\ell+1}}{L_n}\int_{\mathcal{X}^{\ell+1}} f(x_k)e(x_{k+1:k+\ell-1})\Phi_{\nu,k:k+\ell-1|k+\ell-1}(dx_{k:k+\ell-1})m(\underline{x}_{k+\ell-1};x_{k+\ell})\\
&\times g_{k+\ell}(\underline{x}_{k+\ell-1},x_{k+\ell})\left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n})\prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1};x_i)g_i(\underline{x}_{i-1},x_i)dx_{k+\ell+1:n}\right]dx_{k+\ell}
\end{aligned}
$$

using the implicit definition of the backward kernel (5.51) applied to the function

$$
\begin{aligned}
r(x_{k:k+\ell-1},x_{k+\ell}) &= f(x_k)e(x_{k+1:k+\ell-1})g_{k+\ell}(\underline{x}_{k+\ell-1},x_{k+\ell})\\
&\times\left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n})\prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1};x_i)g_i(\underline{x}_{i-1},x_i)dx_{k+\ell+1:n}\right]dx_{k+\ell}
\end{aligned}
\tag{4.76}
$$

one could get

$$
\begin{aligned}
&\mathbb{E}\left[f(X_k)e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\Big|Y_{0:n}\right]\\
&=\frac{L_{k-\ell+1}}{L_n}\int_{\mathcal{X}^{\ell+1}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell},dx_k)f(x_k)e(x_{k+1:k+\ell-1})\\
&\times\Phi_{\nu,k+1:k+\ell|k+\ell-1}(dx_{k+1:k+\ell})g_{k+\ell}(\underline{x}_{k+\ell-1},x_{k+\ell})\\
&\times\left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n})\prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1};x_i)g_i(\underline{x}_{i-1},x_i)dx_{k+\ell+1:n}\right]
\end{aligned}
\tag{4.77}
$$

taking $f\equiv 1$, for any functions $h'\in\mathbf{F}_b(\mathcal{X}^{n-k-\ell+1})$ and $e'\in\mathbf{F}_b(\mathcal{X}^{\ell-1})$

$$
\begin{aligned}
&\mathbb{E}\left[e'(X_{k+1:k+\ell-1})h'(X_{k+\ell:n})\Big|Y_{0:n}\right]\\
&=\frac{L_{k-\ell+1}}{L_n}\int_{\mathcal{X}^{\ell}} e'(x_{k+1:k+\ell-1})\Phi_{\nu,k+1:k+\ell|k+\ell-1}(dx_{k+1:k+\ell})g_{k+\ell}(\underline{x}_{k+\ell-1},x_{k+\ell})\\
&\times\left[\int_{\mathcal{X}^{n-k-\ell}} h'(x_{k+\ell:n})\prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1},x_i)g_i(\underline{x}_{i-1},x_i)dx_{k+\ell+1:n}\right]
\end{aligned}
\tag{4.78}
$$

Identifying $e'h'$ with $e(x_{k+1:k+\ell-1})h(x_{k+\ell:n})\int_{\mathcal{X}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, x)f(x)dx$, (4.77) may be rewritten as

$$\begin{aligned}&\mathbb{E}\left[f(X_k)e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\middle| Y_{0:n}\right]\\&= \mathbb{E}\left[e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\int_{\mathcal{X}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, x)f(x)dx\middle| Y_{0:n}\right]\end{aligned} \tag{4.79}$$

which leads to the result. □

CHAPITRE 5

Stochastic Volatility Model through GEM algorithm and SMC methods.

Sommaire

5.1 Framework 118
5.2 EM algorithm under latent data Model 119
5.2.1 Model specification 120
5.2.2 Generalized EM algorithm for parameter estimation 121
5.2.2.1 The Expectation Step 122
5.2.2.2 The Maximization Step 123
5.3 Sequential Monte Carlo approximations 125
5.3.1 Forward sampling 126
5.3.2 Forward-Backward algorithm 128
5.3.2.1 One order HMM 129
5.3.2.2 High order HMM 132
5.3.3 Application 139
5.3.4 GEM convergence 139

We deal with application of the Sequential Monte Carlo (SMC) methods in the context of nonlinear state space models. The scope of this analysis is an approximate non-degenerate and stationary state space representation of a degenerated multi-scale stochastic volatility model. Due to the latent structure of the specified model, we use the Generalized Expectation-Maximization algorithm (GEM) to estimate model's parameters. Since we end up with sufficient statistics in this estimation that can not

be computed analytically, we resort to SMC methods that help approximate them. We conclude with example of application to simulated data and discuss convergence issues.

5.1 Framework

We consider the dynamic of an asset price $(S_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathbb{F}, \mathcal{F}, \mathbb{P})$ usually modeled by the stochastic differential equation

$$\frac{dS_t}{S_t} = \mu dt + \sigma_t dW_t^{(0)} \tag{5.1}$$

where $(W_t^{(0)})_{t\geq 0}$ is the standard Brownian motion built on the same probability space, $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ the natural filtration generated by the Brownian motion, μ a given constant rate of return and σ_t the volatility of the price. In our context, this volatility is driven by p mean reverting OU (Ornstein-Ulhenbeck) processes under a single Brownian motion $(W_t^{(*)})_{t\geq 0}$:

$$\begin{cases} \ln(\sigma_t^2) &= \sum_{i=1}^{p} U_{t,i} \\ dU_{t,i} &= \alpha_i(\mu_i - U_{t,i})dt + \beta_i dW_t^{(*)}, i = 1, \ldots, p \end{cases} \tag{5.2}$$

where $W_t^{(0)}, W_t^{(*)}$ are possibly correlated Brownian motions and for each OU process $U_{t,i}$, α_i is the rate of it's mean reversion, μ_i the long-run mean and β_i the volatility of the volatility. In order to achieve model's calibration to real data and parameters estimation, one needs to use a discretized version of the continuous time model. Consider the following Euler-Maruyama discretization of the stochastic volatility model given by

$$\begin{cases} X_{k+1} &= \alpha X_k + \sigma W_{k+1} \\ Y_k &= \beta e^{1_p^T X_k/2} V_k \end{cases} \tag{5.3}$$

where $\alpha := \mathbf{diag}(\alpha_1, \ldots, \alpha_p)$, $\sigma := (\sigma_1, \ldots, \sigma_p)^T$, $|\alpha_i| < 1$ for $1 \leq i \leq p$, W_{k+1} is independent of (X_i, Y_i) for $0 \leq i \leq k$, V_k is independent of (X_i, Y_i) for $0 \leq i \leq k-1$, $\mathbf{Corr}(W_i, V_j) = \rho 1_{\{i=j\}}$. In order to make manageable the state degeneracy, model (5.3) is rewritten in an approximate stationary regime that will be specified later. Since each component process $X(m)$ follows an AR(1) model given by

$$(1 - \alpha_m B) X_{k+1}(m) = \sigma_m W_{k+1}, \quad m = 1, 2, \ldots, p$$

where B is the backshift operator, the signal X with $X_k := 1_p^T X_k$ can be seen as the output when passing the signal W through a linear filter with the overall transfer function given by

$$T(z) := \sum_{m=1}^{p} \frac{\sigma_m}{1 - \alpha_m z} = \frac{\psi_0 + \psi_1 z + \psi_1 z^2 + \ldots + \psi_q z^q}{1 - \varphi_1 z - \varphi_2 z^2 - \ldots - \varphi_p z^p}, \quad z \in \mathbb{C} \tag{5.4}$$

where $p = q + 1$, the coefficients $\varphi_i := \varphi_i(\theta)$, $1 \leq i \leq p$ and $\psi_j := \psi_j(\theta), 0 \leq j \leq q$ are given respectively by

$$\varphi_i(\theta) = (-1)^{i+1} \sum_{J \subseteq \{1,\dots,p\}} \prod_{j \in J} \alpha_j \tag{5.5}$$

$$\psi_j(\theta) = (-1)^j \sum_{m=1}^{p} \sigma_m \sum_{\substack{K \subseteq \{1,\dots,p\} \setminus \{m\} \\ \#K = j}} \prod_{j \in K} \alpha_k. \tag{5.6}$$

One can easily deduce the ARMA(p, q) process associated to $T(\cdot)$ by

$$\varphi(B)X_{k+1} = \Psi(B)W_{k+1}$$

where $\varphi(\cdot)$ (resp. $\Psi(\cdot)$) is the corresponding p^{th} (resp. q^{th}) order polynomials of the AR (resp. MA) part. For sake of simplicity, we also use the AR(∞) representation of X.

Lemma 5.1.1. *Assume that $\varphi(\cdot)$ and $\Psi(\cdot)$ have no common zeros and $\Psi(z) \neq 0$ for $|z| \leq 1$. Then the ARMA(p, q) process X satisfies an AR(∞) representation given by*

$$X_{k+1} = \sum_{j=1}^{\infty} \tilde{\pi}_j X_{k+1-j} + \sigma W_{k+1} \tag{5.7}$$

with $\sum_{j=0}^{\infty} |\tilde{\pi}_j| < \infty$, where the sequence $\{\tilde{\pi}_j\}$, $\tilde{\pi}_j := \frac{\pi_j(\theta)}{\pi_0(\theta)}$ are determined by the equations

$$\pi_j(\theta) = \frac{\varphi_j(\theta)}{\psi_0(\theta)} - \sum_{k=1}^{\min(q, j-1)} \frac{\psi_k(\theta)}{\psi_0(\theta)} \pi_{j-k}(\theta) + \pi_0(\theta) \frac{\psi_j(\theta)}{\psi_0(\theta)}, \quad j = 1, 2, \dots \tag{5.8}$$

where $\pi_0(\theta) = \psi_0(\theta)^{-1}$, $\sigma := \psi_0(\theta) = \sum_{m=1}^{p} \sigma_m$, $\varphi_j(\theta) = 0$ for $j > p$, and with the convention that $\sum_m^r = 0$ if $m > r$.

Proof. See Brockwell and Davis [1991]. □

The paper is organized as follows. In section 2 we estimate an approximate version the later model within EM algorithm. Section 3 is dedicated to SMC methods that handle sufficient statistics approximation in the estimation step. We end by illustrating with a simulation study and discuss some convergence issues in section 4.

5.2 EM algorithm under latent data Model

In the sequel, as one can not take into account all the infinite lags of the AR(∞) process, we choose a suitable truncation order, say ℓ to represent X.

5.2.1 Model specification

Assume one has initialized the process X for negative time indexes from $k = -\ell$ up to $k = -1$ according to a diffuse prior measure ν_θ on $\mathcal{X}^\ell$, the $\ell-$ Cartesian product of the state space $\mathcal{X}$. Consider the following approximate model :

$$\begin{cases} X_k = \pi^T \underline{X}_{k-1} + \sigma W_k \\ \\ Y_k = \beta \exp(X_k/2)\, V_k, \;\; k \geq 0 \end{cases} \tag{5.9}$$

where $(V_k, W_k)_{k\geq 0}$ are i.i.d and independent of $X_{-\ell:-1}$ with $(V_k, W_k) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$, $\theta := (\pi^T, \sigma, \beta, \rho)$ is the parameter vector such that the roots of the associated lag polynomial lie outside the unit circle with the convention $\sigma > 0, \beta > 0$, $|\rho| < 1$, $\pi^T := (\pi_1, \pi_2, \ldots \pi_\ell)$ and $\underline{X}_k := (X_k, X_{k-1}, \ldots, X_{k-\ell+1})^T$. We shall consider the filtration $\left(\mathcal{F}_k^{X,Y}\right)_{k\geq 0}$ generated by the joint sample path of the process $(X_k, Y_k)_{k\geq 0}$ that is $\mathcal{F}_k^{X,Y} = \sigma\left((X_s, Y_s) : 0 \leq s \leq k\right) \vee \sigma\left(X_{-\ell:-1}\right)$. Define also $\mathcal{Y}$ as the observations' space and $\mathbf{F}_b(\mathcal{X})$ the set of bounded measurable functions on $\mathcal{X}$.

Lemma 5.2.1. *$(X_k, k \geq 0)$ is an $\ell-$order Markov chain w.r.t. the filtration $\left(\mathcal{F}_k^{X,Y}\right)_{k\geq 0}$, with initial distribution ν_θ on the sequence $X_{-\ell:-1}$ and transition kernel density*

$$m_\theta(\underline{x}, z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(z - \pi^T \underline{x})^2}{2\sigma^2}\right\} \tag{5.10}$$

w.r.t. the Lebesgue measure on $\mathbb{R}$, where $\underline{x} \in \mathbb{R}^\ell$ and $z \in \mathbb{R}$.

Proof. $\underline{X}_{k-1}$ is $\mathcal{F}_{k-1}^{X,Y}-$measurable and W_k is independent of $\mathcal{F}_{k-1}^{X,Y}$ thus of $\underline{X}_{k-1}$. From (5.9), $X_k = h(\underline{X}_{k-1}, W_k)$; so, for any $f \in \mathbf{F}_b(\mathcal{X})$:

$$\mathbb{E}[f(X_k)|\mathcal{F}_{k-1}^{X,Y}] = \mathbb{E}[f \circ h(\underline{X}_{k-1}, W_k)|\mathcal{F}_{k-1}^{X,Y}] = \varphi(\underline{X}_{k-1})$$

where $\varphi(\underline{x}_{k-1}) = \mathbb{E}[f \circ h(\underline{x}_{k-1}, W_k)]$. □

Lemma 5.2.2. *Under model (5.9), for any $f \in \mathbf{F}_b(\mathcal{Y})$,*

$$\mathbb{E}\left[f(Y_k)\middle|\mathcal{F}_{k-1}^{X,Y}, X_k\right] = \int_{\mathcal{Y}} f(y) g^\theta(y, \underline{X}_{k-1}, X_k) dy \tag{5.11}$$

where

$$g^\theta(y, \underline{x}_{k-1}, x_k) = \frac{1}{\sqrt{2\pi\beta^2(1-\rho^2)e^{x_k}}} \exp\left\{-\frac{\left(y - \sigma^{-1}\beta\rho e^{x_k/2}(x_k - \pi^T \underline{x}_{k-1})\right)^2}{2\beta^2(1-\rho^2)e^{x_k}}\right\}. \tag{5.12}$$

Proof. From (5.9), for any function $f \in \mathbf{F}_b(\mathcal{Y})$,

$$\mathbb{E}\left[f(Y_k)\middle|\mathcal{F}_{k-1}^{X,Y}, X_k\right] = \mathbb{E}\left[f\left(\beta \exp(X_k/2)V_k\right)\middle|\mathcal{F}_{k-1}^{X,Y}, X_k\right]. \tag{5.13}$$

Since V_k, W_k are correlated, they verify the following identity :

$$V_k = \rho W_k + \sqrt{1-\rho^2}\zeta_k \tag{5.14}$$

where $(\zeta_j)_{j\geq 0}$ are standard Gaussian i.i.d random variables independent of $(W_j)_{j\geq 0}$ and $X_{-\ell:-1}$. Moreover, $X_{-\ell:k}$ and $Y_{0:k-1}$ are measurable w.r.t $\sigma(X_{-\ell:-1}) \vee \sigma(W_j, 0 \leq j \leq k) \vee \sigma(V_j, 0 \leq j \leq k-1)$, implying that ζ_k is independent of $X_k \vee \mathcal{F}_{k-1}^{X,Y}$. So

$$\begin{aligned}
&\mathbb{E}\left[f\left(\beta \exp(X_k/2)V_k\right)\middle|\mathcal{F}_{k-1}^{X,Y}, X_k\right] \\
=&\mathbb{E}\left[f\left(\beta \exp\left(X_k/2\right)\left(\rho\sigma^{-1}(X_k - \pi^T \underline{X}_{k-1}) + \sqrt{1-\rho^2}\zeta_k\right)\right)\middle|\mathcal{F}_{k-1}^{X,Y}, X_k\right] \\
=&\mathbb{E}\left[f\left(\beta \exp\left(X_k/2\right)\left(\rho\sigma^{-1}(X_k - \pi^T \underline{X}_{k-1}) + \sqrt{1-\rho^2}\zeta_k\right)\right)\middle|_{X_k = x_k, \underline{X}_{k-1} = \underline{x}_{k-1}}\right]
\end{aligned}$$

Using the last equality and the Gaussianity of ζ_k, the kernel density expression (5.12) is then easily derived. □

Since the data are fixed $\{Y_k = y_k, k \geq 0\}$, the latter kernel density may be simply rewritten as $g_k^\theta(\underline{x}_{k-1}, x_k)$ instead of $g^\theta(y_k, \underline{x}_{k-1}, x_k)$ omitting the dependence to the data. With previous specifications, model (5.9) can be seen as an HMM with

$$\begin{cases} \nu_\theta(x_{-\ell:-1}) & \text{the prior density on the initial state's sequences,} \\ m_\theta(\underline{x}_{k-1}; x_k) & \text{the transition kernel density,} \\ g_k^\theta(\underline{x}_{k-1}, x_k) & \text{the measurement density, for } k \geq 0. \end{cases}$$

We can now face the parameter estimation of model (5.9).

5.2.2 Generalized EM algorithm for parameter estimation

Since we are dealing with partially observable data due to the latent process X, a natural procedure for estimating parameters is the well known Expectation-Maximization (EM) algorithm formalized in Dempster et al. [1977]. Recall that the EM algorithm is an iterative procedure suitable for computing the MLE in the presence of missing data or more generally, in incomplete data case where the classical maximization of the likelihood fails because of the likelihood structure of the model embedded with incompleteness. The EM algorithm consists in two steps : the Expectation step and the Maximization step. Both the steps are repeated till a given convergence criterion occurs. The generic pseudo code of the generalized EM algorithm is given below.

Algorithme 16 Generalized EM algorithm

1: Choose an initial guess $\theta^{(0)}$
2: **For** $k = 0, 1, 2, \ldots$ **do**
   1. **E-Step :** Compute $Q(\theta^{(k+1)}, \theta^{(k)})$
   2. **M-Step :** Find $\theta^{(k+1)}$ such that $Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)})$
3: **EndFor.**

5.2.2.1 The Expectation Step

The E-step consists in computing the conditional expectation of the logarithm of the complete data likelihood in the light of the data named the *intermediate quantity* for the current value of the parameter vector θ'.The intermediate quantity is then given by :

$$Q(\theta, \theta') = \mathbb{E}_{\theta'} \left[\log p_\theta \left(X_{-\ell:n}, Y_{0:n} \right) | Y_{0:n} \right] \tag{5.15}$$

where $(n+1)$ is the sample size of the data indexed from 0 to n and p_θ is a generic notation for densities depending on parameter θ.

Lemma 5.2.3. *Under model's assumptions* (5.9)*, the logarithm of the complete data likelihood is given by*

$$\log p_\theta \left(x_{-\ell:n}, y_{0:n} \right) = \log \nu_\theta \left(x_{-\ell:-1} \right) - (n+1) \log \left[2\pi\beta(1-\rho^2)^{1/2} \right] - (n+1) \log \sigma$$
$$- \frac{1}{2} \sum_{k=0}^{n} \left\{ x_k + \frac{1}{(1-\rho^2)} \left\{ \frac{e^{-x_k} y_k^2}{\beta^2} + \frac{\left[x_k - \pi^T \underline{x}_{k-1} \right]^2}{\sigma^2} - \frac{2\rho}{\sigma\beta} e^{-x_k/2} y_k \left[x_k - \pi^T \underline{x}_{k-1} \right] \right\} \right\}.$$

Proof. Since the complete data likelihood iterates as follows :

$$\begin{aligned} p_\theta \left(x_{-\ell:n}, y_{0:n} \right) &= p_\theta \left(y_n | y_{0:n-1}, x_{0:n} \right) p_\theta \left(x_n | x_{0:n-1}, y_{0:n-1} \right) p_\theta \left(x_{-\ell:n-1}, y_{0:n-1} \right) \\ &= p_\theta \left(y_n | x_{n-\ell:n} \right) p_\theta \left(x_n | x_{n-\ell:n-1} \right) p_\theta \left(x_{-\ell:n-1}, y_{0:n-1} \right) \\ &= g_n^\theta \left(\underline{x}_{n-1}, x_n \right) m_\theta \left(\underline{x}_{n-1}; x_n \right) p_\theta \left(x_{-\ell:n-1}, y_{0:n-1} \right), \end{aligned} \tag{5.16}$$

one could obtain the factorization below :

$$p_\theta \left(x_{-\ell:n}, y_{0:n} \right) = \nu_\theta \left(x_{-\ell:-1} \right) \prod_{k=0}^{n} m_\theta \left(\underline{x}_{k-1}; x_k \right) g_k^\theta \left(\underline{x}_{k-1}, x_k \right), \tag{5.17}$$

where m_θ (resp. g_k^θ) are given by (5.10) (resp. (5.12)). Taking the logarithm leads to

$$\begin{aligned} \log p_\theta \left(x_{-\ell:n}, y_{0:n} \right) = \log \nu_\theta \left(x_{-\ell:-1} \right) &+ \sum_{k=0}^{n} \log m_\theta \left(\underline{x}_{k-1}; x_k \right) \\ &+ \sum_{k=0}^{n} \log g_k^\theta \left(\underline{x}_{k-1}, x_k \right). \end{aligned} \tag{5.18}$$

□

Let's introduce some additional notations for convenience purposes :

$$\tilde{X}_k := \left[X_k, \underline{X}_{k-1}\right]^T \in \mathbb{R}^{\ell+1};$$

define also the sufficient statistics by :

$$S_1 := \frac{1}{n+1}\sum_{k=0}^{n} Y_k^2 \mathbb{E}_{\theta'}\left[\exp\left(-X_k\right)\Big| Y_{0:n}\right], S_2 := \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{E}_{\theta'}\left[\tilde{X}_k\tilde{X}_k^T | Y_{0:n}\right],$$
$$S_3 := \frac{1}{n+1}\sum_{k=0}^{n} Y_k \mathbb{E}_{\theta'}\left[\exp\left(-X_k/2\right)\odot \tilde{X}_k | Y_{0:n}\right],$$

where $\odot$ denotes the multiplication of the left hand side by each component of the right hand side. With these notations and up to additive constant independent of θ, the intermediate quantity is given by :

$$Q(\theta,\theta') = -(n+1)\left\{\log\left[\beta(1-\rho^2)^{1/2}\right] + \log\sigma\right\}$$
$$-\frac{(n+1)}{2(1-\rho^2)}\left\{\frac{1}{\beta^2}S_1 + \frac{1}{\sigma^2}\left[1; -\pi^T\right] S_2 \begin{bmatrix}1\\ -\pi\end{bmatrix} - \frac{2\rho}{\sigma\beta}\left[1; -\pi^T\right] S_3\right\}.$$

5.2.2.2 The Maximization Step

We now face the evaluation of the M step of the GEM algorithm. In many situations, the close solution to the M step, that is the value that globally maximizes the intermediate quantity $Q(\theta,\theta')$ is quite harder to find. An easy way to tackle the problem is, instead of searching for θ_{k+1} that realizes the global maximum, to choose θ_{k+1} such that

$$Q(\theta_{k+1},\theta_k) \geq Q(\theta_k,\theta_k). \tag{5.19}$$

(5.19) is a sufficient condition ensuring an increase of the likelihood function (see McLachlan and Krishnan [2008], section 3.3) after one iteration of the M step. To get such value of θ_{k+1}, one can resort to the Alternating Optimization (AO) algorithm [1] (see Bezdek and Hathaway [2003a] and Bezdek and Hathaway [2003b] for more details). In our context, the parameter vector is $\theta := (\rho, \beta, \sigma, \pi_1, \pi_2, \ldots, \pi_l)$. θ is partitioned into four blocks. The overall parameters are carried out by combining the outputs of the overall runs.

One can now make use of the AO by differentiating the intermediate quantity w.r.t. the parameters to get at best the optimal parameters values or at worst, improvement in the parameter values at each iteration.

1. One could also run one step of Newton-Raphson procedure if possible

Maximum with respect to ρ

Computing the derivative of $Q(\theta, \theta')$ w.r.t. ρ leads to

$$\frac{\partial Q(\theta, \theta')}{\partial \rho} = 0 \Leftrightarrow -\rho^3 + a_2\rho^2 + a_1\rho + a_0 = 0 \tag{5.20}$$

with $a_2 = \frac{1}{\sigma\beta}\begin{bmatrix}1; & -\pi^T\end{bmatrix} S_3$, $a_1 = 1 - \frac{S_1}{\beta^2} - \frac{1}{\sigma^2}\begin{bmatrix}1; & -\pi^T\end{bmatrix} S_2 \begin{bmatrix}1\\ -\pi\end{bmatrix}$ and $a_0 = \frac{1}{\sigma\beta}\begin{bmatrix}1; & -\pi^T\end{bmatrix} S_3$; which is a third order polynomial. One has to chose the root ρ^* which is in absolute value smaller than one.

Maximum with respect to β

Computing the derivative w.r.t. β leads to

$$\frac{\partial Q(\theta, \theta')}{\partial \beta} = 0 \Leftrightarrow b_2\beta^2 + b_1\beta + b_0 = 0 \tag{5.21}$$

where $b_2 = -(1-\rho^2)$, $b_1 = -\frac{\rho}{\sigma}\begin{bmatrix}1; & -\pi^T\end{bmatrix} S_3$, $b_0 = S_1$. Since the discriminant of this quadratic polynomial is non-negative, there are two real roots with opposite signs that is $\beta_1 \times \beta_2 = \frac{b_0}{b_2} < 0$.

Maximum with respect to σ

Computing the derivative w.r.t σ leads to

$$\frac{\partial Q(\theta, \theta')}{\partial \sigma} = 0 \Leftrightarrow c_2\sigma^2 + c_1\sigma + c_0 = 0 \tag{5.22}$$

with $c_2 = -(1-\rho^2), c_1 = -\frac{\rho}{\beta}\begin{bmatrix}1; & -\pi^T\end{bmatrix} S_3(\theta)$, $c_0 = \begin{bmatrix}1; & -\pi^T\end{bmatrix} S_2 \begin{bmatrix}1\\ -\pi\end{bmatrix}$. Once again, the discriminant is positive and there are two real roots with opposite signs given by $\sigma_1 \times \sigma_2 = \frac{c_0}{c_2} < 0$. So, the positive root is chosen in successive runs of the M-step.

Maximum with respect to π

Computing the derivative with respect to π_i leads to

$$\begin{aligned}\frac{\partial Q(\theta, \theta')}{\partial \pi_i} = 0 \quad &\Leftrightarrow \quad S_7^{(i)}\pi_i - S_4^{(i)}(\theta) + \sum_{j\neq i}\pi_j S_5^{(i,j)} - \frac{\rho\sigma}{\beta}S_6^{(i)} = 0 \\ &\Rightarrow \quad \pi_i = \frac{S_4^{(i)} - \sum_{j\neq i}\pi_j S_5^{(i,j)} - \frac{\rho\sigma}{\beta}S_6^{(i)}}{S_7^{(i)}},\end{aligned} \tag{5.23}$$

where for all $i = 1, 2, \cdots, \ell$

$$S_4^{(i)} := \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{E}_{\theta'}\left[X_k X_{k-i} | Y_{0:n}\right], \quad S_5^{(i,j)} := \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{E}_{\theta'}\left[X_{k-i} X_{k-j} | Y_{0:n}\right]$$

$$S_6^{(i)} := \frac{1}{n+1}\sum_{k=0}^{n} Y_k \mathbb{E}_{\theta'}\left[\exp\left(-X_k/2\right) X_{k-i} | Y_{0:n}\right], \quad S_7^{(i)} := \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{E}_{\theta'}\left[X_{k-i}^2 | Y_{0:n}\right].$$

Note that these new statistics do not require to be computed since they can be obtained from the formers. The generic pseudo code of the generalized EM algorithm in this context is summarized below.

Algorithme 17 Generalized EM algorithm

1: Choose an initial guess $\theta^{(0)} = (\rho^{(0)}, \beta^{(0)}, \sigma^{(0)}, \pi^{\tau(0)})$
2: **For** $m = 0, 1, 2, \ldots$ **do**
   1. **E-Step :** Given $\theta^{(m)}$
      — Compute S_1, S_2 and S_3
      — Deduce $S_4^{(i)}, S_5^{(i,j)}, S_6$ and $S_7^{(i)}$ for $i, j = 1, 2, \cdots, \ell$
   2. **M-Step :** Find $\theta^{(m+1)}$ such that $Q(\theta^{(m+1)}, \theta^{(m)}) \geq Q(\theta^{(m)}, \theta^{(m)})$
3: **EndFor.**

5.3 Sequential Monte Carlo approximations

The parameter updates in the GEM algorithm requires approximation of the sufficient statistics appearing in (5.20), (5.21), (5.22) and (5.23) that is, the evaluation of smoothing distributions of the form

$$\phi_{\nu,s,t|n}(f) := \mathbb{E}[f(X_s, X_t)|Y_{0:n}], \tag{5.24}$$

where $0 \leq |s-t| \leq \ell+1$ with $-\ell \leq s, t \leq n$ and $f \in \mathbf{F}_b(\mathcal{X}^2)$. A simple rewrite of these distributions shows that both there admit densities w.r.t. to the Lebesgue measure and one can easily derive them from the joint smoothing density $p_\theta(x_{-\ell:n}|y_{0:n})$ as marginal densities in the following way

$$\begin{aligned}\mathbb{E}[f(X_s, X_t)|Y_{0:n}] &= \int_{\mathcal{X}^2} f(x_s, x_t) p_\theta(x_s, x_t|y_{0:n}) dx_s dx_t \\ &= \int_{\mathcal{X}^{n+\ell+1}} f(x_s, x_t) p_\theta(x_{-\ell:n}|y_{0:n}) dx_{-\ell:n}. \end{aligned} \tag{5.25}$$

Other derivations may be achieved but require evaluation of joint smoothing densities which evolve complex multidimensional densities preventing a direct computation. The

reset of this section is devoted to the SMC methods that help to approximate them. In order to simplify the presentation, the explicit dependence on parameters is omitted for the densities. We also use the notations L_k for the incomplete likelihood density of $y_{0:k}$, $\phi_{\nu,k} := \phi_{\nu,k|k}$ as the filtering distribution and the filtering density at time k, as an abuse of notation.

5.3.1 Forward sampling

A first smoothing approximation is achieved when thinking these distributions as marginals of the joint smoothing distribution. So, one just needs to approximate the latter in a forward pass, with means of Sequential Importance Sampling with Resampling (SIRS) algorithm. To do so, consider a suitable instrumental distribution density $q_n(x_{-\ell:n})$ which may depend on the observations $y_{0:n}$, targeting the joint smoothing density $p(x_{-\ell:n}|y_{0:n})$. Define the unnormalized importance weights

$$\omega_n(x_{-\ell:n}) := \frac{p(x_{-\ell:n}, y_{0:n})}{q_n(x_{-\ell:n})}. \tag{5.26}$$

For any function $h \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$, the following identity holds :

$$\mathbb{E}(h(X_{-\ell:n})|Y_{0:n}) = \frac{\mathbb{E}_{q_n}[h(X_{-\ell:n})\omega_n(X_{-\ell:n})]}{\mathbb{E}_{q_n}[\omega_n(X_{-\ell:n})]} \tag{5.27}$$

where the expectation $\mathbb{E}_{q_n}$ is taken with respect to the instrumental density $q_n(x_{-\ell:n})$. Replacing with empirical counterparts, one can get an estimate of (5.27).

Approximation of (5.27)

Let $\left(\xi_{-\ell:n}^{(i)}, 1 \leq i \leq N\right)$ be i.i.d. sample of size N drawn from the instrumental density $q_n(\cdot)$, then :

$$\begin{aligned}\mathbb{E}(h(X_{-\ell:n})|Y_{0:n}) &\approx \frac{\frac{1}{N}\sum_{i=1}^{N}\omega_n\left(\xi_{-\ell:n}^{(i)}\right)h(\xi_{-\ell:n}^{(i)})}{\frac{1}{N}\sum_{j=1}^{N}\omega_n\left(\xi_{-\ell:n}^{(j)}\right)} \\ &= \sum_{i=1}^{N}\overline{\omega}_n\left(\xi_{-\ell:n}^{(i)}\right)h(\xi_{-\ell:n}^{(i)})\end{aligned} \tag{5.28}$$

where

$$\overline{\omega}_n\left(\xi_{-\ell:n}^{(i)}\right) = \frac{\omega_n\left(\xi_{-\ell:n}^{(i)}\right)}{\sum_{j=1}^{N}\omega_n\left(\xi_{-\ell:n}^{(j)}\right)}, \quad i = 1, 2, \ldots, N$$

are the normalized importance weights. But this estimate still in a batch mode in the sense that, one needs to recompute these weights on the whole data set as soon as a new observation becomes available. Which is not suitable for recursive estimate. Since the computational effort is increasing with time.

Recursive weights

In order to avoid loosing previous simulations, one may render the weights' evaluation sequential in time. To achieve this, if there exists $\tilde{q}_n$ such that the proposal density expands as follows :

$$q_n\left(x_{-\ell:n}\right) = q_{n-1}\left(x_{-\ell:n-1}\right)\tilde{q}_n\left(x_n|x_{-\ell:n-1}\right) \tag{5.29}$$

one could get a recursive scheme for the unnormalized importance weights :

$$\begin{aligned}
\omega_n\left(x_{-\ell:n}\right) &= \frac{p\left(x_{-\ell:n}; y_{0:n}\right)}{q_n\left(x_{-\ell:n}\right)} \\
&= \frac{p\left(x_{-\ell:n-1}; y_{0:n-1}\right) p\left(y_n|x_{-\ell:n}\right) p\left(x_n|x_{-\ell:n-1}\right)}{q_{n-1}\left(x_{-\ell:n-1}\right)\tilde{q}_n\left(x_n|x_{-\ell:n-1}\right)} \\
&= \omega_{n-1}\left(x_{-\ell:n-1}\right)\frac{p\left(y_n|x_{n-\ell:n}\right) p\left(x_n|x_{n-\ell:n-1}\right)}{\tilde{q}_n\left(x_n|x_{-\ell:n-1}\right)} \\
&= \omega_{n-1}\left(x_{-\ell:n-1}\right)\underbrace{\frac{g_n\left(\underline{x}_{n-1}, x_n\right) m\left(\underline{x}_{n-1}; x_n\right)}{\tilde{q}_n\left(x_n|x_{-\ell:n-1}\right)}}_{\textit{the incremental ratio}}
\end{aligned} \tag{5.30}$$

where we use (5.16) in the second equality. The common choice of a proposal density that meets the requirement (5.29) is the Markov transition density of the model that is on taking $\tilde{q}_n$ to be the transition density of the hidden Markov chain or equivalently $q_n(x_{-\ell:n}) = p(x_{-\ell:n})$ where $p(x_{-\ell:n})$ is the joint density of the states $x_{-\ell:n}$. The corresponding algorithm is the Sequential Importance Sampling (SIS) algorithm. However, the SIS algorithm is not suitable in high dimensional spaces since the distribution of the importance weights becomes more and more skewed. To prevent such degeneracy problem, a correction step is introduced which consists in discarding particles with low weights and duplicating the ones with high weights. This resampling step is done only if the Effective Sample Size (ESS) falls below a given threshold. Recall that ESS is given by

$$\mathrm{ESS}_k = \left[\sum_{i=1}^{N}\left(\overline{\omega}_k^{(i)}\right)^2\right]^{-1}.$$

The resulting algorithm is known as the Sequential Importance Sampling Resampling (SISR) which differs from the SIS algorithm only for the resampling step. We give a sketch of the algorithm below. The correction step consists in line 6 and 7. Other lines are the SIS part of the SISR algorithm.

Algorithme 18 Sequential Importance Sampling Resampling algorithm

1: **For** $i = 1, 2, \ldots, N$
   — Ensure $\xi_{-\ell:-1}^{(i)} \sim \nu(\cdot)$, and Draw $\xi_0^{(i)} \sim m\left(\xi_{-\ell:-1}^{(i)}; \cdot\right)$
2: **EndFor**
3: Set $k \leftarrow 1$
4: **For** $i = 1, 2, \ldots, N$
   — Draw $\bar{\xi}_k^{(i)} \sim \tilde{q}_k\left(\cdot|\xi_{-\ell:k-1}^{(i)}\right)$ and Set $\bar{\xi}_{0:k}^{(i)} \leftarrow \left(\xi_{-\ell:k-1}^{(i)}; \bar{\xi}_k^{(i)}\right)$
   — Evaluate and Normalize the importance weights :

$$\omega_k^{(i)} \propto \omega_{k-1}^{(i)} \underbrace{\frac{g_k\left(\bar{\xi}_{k-\ell:k-1}^{(i)}, \bar{\xi}_k^{(i)}\right) m\left(\bar{\xi}_{k-\ell:k-1}^{(i)}; \bar{\xi}_k^{(i)}\right)}{\tilde{q}_k\left(\bar{\xi}_k^{(i)}|\bar{\xi}_{-\ell:k-1}^{(i)}\right)}}_{\textit{the incremental ratio}} \tag{5.31}$$

5: **EndFor**
6: **If** $N_{\text{eff}} <$ Threshold
   — Multiply/Discard particles $\left\{\bar{\xi}_{-\ell:k}^{(i)}\right\}_{i=1}^{N}$ with respect to their normalized weights to get N particles $\left\{\xi_{-\ell:k}^{(i)}\right\}_{i=1}^{N}$ approximately distributed according to $p_\theta\left(\xi_{-\ell:k}^{(i)}|y_{0:k}\right)$
   — Set $\omega_k^{(i)} = N^{-1}$ For $i = 1, \ldots, N$
7: **EndIf**
8: Set $k \leftarrow k + 1$

5.3.2 Forward-Backward algorithm

Another way of improving sufficient statistics computation is to consider the approximation of the joint smoothing distribution in a Forward and Backward passes where in the former, the weighted particles are produced using a Bootstrap filter algorithm. In the latter, particles' weights are updated according to a recursive backward dynamic. In order to state backward recursions, one needs a representation of the Markov chain in the reverse time direction. More precisely, one should design reverse moves for consecutive states X_{k+1}, X_k that is the conditional law of X_k given $X_{k+1:n}$ and $Y_{0:n}$. This conditional law may be reached through the conditional law of $X_k, X_{k+1}|Y_{0:n}$ in a $one-$order HMM and $X_{k:k+\ell}|Y_{0:n}$ for the $\ell-$order HMM, $\ell > 1$. In a first stage, we state the F-B algorithm for one-order HMM. In a second stage, we see how to adapt it for higher order HMM.

5.3.2.1 One order HMM

We consider a fully dominated one-order HMM with an initial distribution ν on X_0, a transition kernel M admitting a density with respect to the Lebesgue measure $m(x_{k-1}, x_k) := p(x_k|x_{k-1})$ and emission density $g_k(x_k) := p(y_k|x_k)$ for $k \geq 0$. We also assume that $\{Y_k, k \geq 0\}$ are conditionally independent given $\{X_k, k \geq 0\}$ and Y_k depends only on X_k. Under these assumptions, we have analogous lemmas to those of 2.1 and 2.2.

Lemma 5.3.1. *$(X_k, k \geq 0)$ is a one−order Markov chain with respect to the filtration $\left(\mathcal{F}_k^{X,Y}\right)_{k\geq 0}$ with initial distribution ν on the state X_0 and transition kernel density*

$$m(x_{k-1}, x_k) \tag{5.32}$$

with respect to the Lebesgue measure on $\mathbb{R}$.

Lemma 5.3.2. *Under previous model assumptions, for any $f \in \mathbf{F}_b(\mathcal{Y})$,*

$$\mathbb{E}\left[f(Y_k)\middle|\mathcal{F}_k^{X,Y}\right] = \int_{\mathcal{Y}} f(y)g_k(X_k)dy. \tag{5.33}$$

The problem is to compute the joint smoothing distribution or some of it's features as the marginal smoothing distribution backward in time. To do so, one may need to have implicit definition of the backward kernel of the reverse Markov chain for consecutive states X_{k+1}, X_k. For any function $f \in \mathbf{F}_b(\mathcal{X}^2)$, the distribution of (X_k, X_{k+1}) given the observation $Y_{0:k}$, under the initial distribution ν on X_0 is given by

$$\mathbb{E}\left[f(X_k, X_{k+1})|Y_{0:k}\right] = \int_{\mathcal{X}^2} f(x_k, x_{k+1})\phi_{\nu,k}(dx_k)m(x_k, x_{k+1})dx_{k+1} \tag{5.34}$$

where $\phi_{\nu,k}$ is the filtering distribution at time k. If one can design a reverse kernel $B_{\nu,k}$ from $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$ to $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$ such that :

$$\mathbb{E}\left[f(X_k, X_{k+1})|Y_{0:k}\right] = \int_{\mathcal{X}^2} f(x_k, x_{k+1})\phi_{\nu,k+1|k}(dx_{k+1})B_{\nu,k}(x_{k+1}, dx_k) \tag{5.35}$$

where $\phi_{\nu,k+1|k} := \phi_{\nu,k}M$ is the one step predictive distribution, then the following result holds.

Lemma 5.3.3. *Given an initial distribution ν, a strictly positive index n and index $k \in \{0, 1, \ldots, n-1\}$*

$$\mathbb{E}\left[f(X_k)|X_{k+1:n}, Y_{0:n}\right] = \mathbb{E}\left[f(X_k)|X_{k+1}, Y_{0:k}\right] = B_{\nu,k}(X_{k+1}, f) \tag{5.36}$$

where $B_{k,\nu}(X_{k+1}, .)$ is the Backward kernel for any function $f \in \mathbf{F}_b(\mathcal{X})$.

Proof. see (Cappé et al. [2005], page 70). □

Lemma 5.3.4. *Given an initial distribution* ν, $n > 0$ *and* $k \in \{0, 1, \ldots, n-1\}$

$$\begin{aligned} \phi_{\nu,k|n}(f) &:= \mathbb{E}\left[f(X_k)\middle| Y_{0:n}\right] \\ &= \int_{\mathcal{X}^2} f(x_k) B_{\nu,k}(x_{k+1}, dx_k)\phi_{\nu,k+1|n}(dx_{k+1}), \end{aligned} \tag{5.37}$$

for any function $f \in \mathbf{F}(\mathcal{X}^2)$.

Proof. It suffices to establish a recurrent formula between $\phi_{\nu,k|n}$ and $\phi_{\nu,k+1|n}$:

$$\begin{aligned} p(x_k|y_{0:n}) &= \int_{\mathcal{X}} p(x_k, x_{k+1}|y_{0:n})dx_{k+1} \\ &= \int_{\mathcal{X}} p(x_k|x_{k+1}, y_{0:n})p(x_{k+1}|y_{0:n})dx_{k+1} \\ &= \int_{\mathcal{X}} B_{\nu,k}(x_{k+1}, x_k)p(x_{k+1}|y_{0:n})dx_{k+1} \end{aligned} \tag{5.38}$$

Using (5.38), the conditional expectation of $f(X_k)$ given $Y_{0:n}$ is given by :

$$\phi_{\nu,k|n}(f) = \int_{\mathcal{X}^2} f(x_k) B_{\nu,k}(x_{k+1}, dx_k)\phi_{\nu,k+1|n}(dx_{k+1}) \tag{5.39}$$

which leads to the result. □

Combining a weighted particle filter approximation $\{\xi_k^{(i)}, \omega_k^{(i)}\}_{i=1}^N$ at current time index k and particles smoother $\{\xi_{k+1}^{(j)}, \omega_{k+1|n}^{(j)}\}_{j=1}^N$ at time $k+1$, one could get an estimate of the marginal smoothing distribution by :

$$\hat{\phi}_{\nu,k|n}(f) = \sum_{i=1}^{N} f(\xi_k^{(i)})\omega_{k|n}^{(i)}, \text{ for } \quad k < n \tag{5.40}$$

where

$$\omega_{k|n}^{(i)} := \sum_{j=1}^{N} \frac{m(\xi_k^{(i)}, \xi_{k+1}^{(j)})\omega_k^{(i)}}{\sum_{l=1}^{N} m(\xi_k^{(l)}, \xi_{k+1}^{(j)})\omega_k^{(l)}} \omega_{k+1|n}^{(j)}. \tag{5.41}$$

Note that the derivation of (5.41) is obtained from the following identity :

$$\phi_{\nu,k|n}(f) = \int_{\mathcal{X}} f(x_k) \left[\int_{\mathcal{X}} \frac{\phi_{\nu,k}(x_k)m(x_k, x_{k+1})}{\int_{\mathcal{X}} \phi_{\nu,k}(x_k)m(x_k, x_{k+1})dx_k} \phi_{\nu,k+1|n}(dx_{k+1})\right] dx_k \tag{5.42}$$

The overall algorithm is summarized below. Note that the evaluation of (5.41) is ex-

Algorithme 19 FB algorithm for marginal smoothing distribution

1: Forward filtering
2: **For** $k = 0, \ldots, n$
— run the particles filtering algorithm to get the weighted particles $\left\{\xi_k^{(i)}, \omega_k^{(i)}\right\}_{i=1}^{N}$.
3: **EndFor**
4: Backward smoothing
— Set $\omega_{n|n}^{(i)} = \omega_n^{(i)}$ For $i = 1, \ldots, N$
— **For** $k = n-1$ down to 0 and $i = 1$ up to N

$$\omega_{k|n}^{(i)} = \omega_k^{(i)} \left(\sum_{j=1}^{N} \omega_{k+1|n}^{(j)} \frac{m\left(\xi_k^{(i)}, \xi_{k+1}^{(j)}\right)}{\sum_{l=1}^{N} \omega_k^{(l)} m\left(\xi_k^{(l)}, \xi_{k+1}^{(j)}\right)} \right) \tag{5.43}$$

— **EndFor**

pensive since it requires $O(N^2)$ operations at each time step by observing that the denominator in (5.43) can be performed independently. So, the resulting estimates can be computed in a reasonable delay only when the number of particles are moderate. It should be noted that the marginal smoothing densities are not the only elements involved in the EM algorithm. The joint smoothing distributions $\phi_{\nu,k:n|n}$ or $\phi_{\nu,k,k+1|n}$ are of great interest specifically the latter.

Joint smoothing approximation

The joint smoothing distribution can be computed backward in time since

$$\begin{aligned} \phi_{\nu,k:n|n}(f) &:= \mathbb{E}\left[f(X_{k:n}) \middle| Y_{0:n} \right] \\ &= \int_{\mathcal{X}^{n-k+1}} f(x_{k:n}) B_{\nu,k}(x_{k+1}, dx_k) \phi_{\nu,k+1:n|n}(dx_{k+1:n}), \quad f \in \mathbf{F}_b\left(\mathcal{X}^{n-k+1}\right) \end{aligned} \tag{5.44}$$

which iterates to :

$$\begin{aligned} \phi_{\nu,k:n|n}(f) &= \int_{\mathcal{X}^{n-k+1}} f(x_{k:n}) \phi_{\nu,n}(dx_n) B_{\nu,n-1}(x_n, dx_{n-1}) \times \ldots \times B_{\nu,k}(x_{k+1}, dx_k) \\ &= \int_{\mathcal{X}^{n-k+1}} f(x_{k:n}) \phi_{\nu,n}(dx_n) \prod_{r=k}^{n-1} B_{\nu,r}(x_{r+1}, dx_r) \end{aligned}$$

In a similar way, the joint smoothing distribution $\phi_{\nu,k,k+1|n}$ iterates as :

$$\begin{aligned} \phi_{\nu,k,k+1|n}(f) &:= \mathbb{E}\left[f(X_k, X_{k+1}) \middle| Y_{0:n} \right] \\ &= \int_{\mathcal{X}^2} f(x_k, x_{k+1}) B_{\nu,k}(x_{k+1}, dx_k) \phi_{\nu,k+1|n}(dx_{k+1}), \quad f \in \mathbf{F}_b\left(\mathcal{X}^2\right) \end{aligned} \tag{5.45}$$

Consider the sets of weighted particles $\left\{\xi_r^{(i_r)}, \omega_r^{(i_r)}\right\}_{i_r=1}^N, r = k, \ldots, n$ each of them approximating the filtering distribution $\phi_{\nu,k}$. Using the backward kernel estimate $\hat{B}_{\nu,r}$ of $B_{\nu,r}$ given by

$$\hat{B}_{\nu,r}(x_{r+1}, dx_r) = \sum_{i_r=1}^N \frac{\omega_r^{(i_r)} m(\xi_r^{(i_r)}, x_{r+1})}{\sum_{l=1}^N \omega_r^{(l)} m(\xi_r^{(l)}, x_{r+1})} \delta_{\xi_r^{(i_r)}}(dx_r). \tag{5.46}$$

One can easily get a particle estimate of (5.44) :

$$\begin{aligned} \hat{\phi}_{\nu,k:n|n}(f) = &\sum_{i_k=1}^N \cdots \sum_{i_n=1}^N f(\xi_k^{(i_k)}, \ldots, \xi_n^{(i_n)}) \\ &\times \frac{\omega_n^{i_n}}{\Omega_n} \prod_{r=k}^{n-1} \frac{\omega_r^{(i_r)} m(\xi_r^{(i_r)}, \xi_{r+1}^{(i_{r+1})})}{\sum_{l=1}^N \omega_r^{(l)} m(\xi_r^{(l)}, \xi_{r+1}^{(i_{r+1})})}, \end{aligned} \tag{5.47}$$

where $\Omega_r = \sum_{r=1}^N \omega_r^{(r)}$, $f \in \mathbf{F}_b\left(\mathcal{X}^{n-k+1}\right)$. Assume one has a set of weighted particle $\left\{\xi_{k+1}^{(j)}, \omega_{k+1|n}^{(j)}\right\}_{j=1}^N$ targeting the smoothing distribution $\phi_{\nu,k+1|n}$. A particle estimate of (5.45) is then given by :

$$\hat{\phi}_{\nu,k,k+1|n}(f) = \sum_{i=1}^N \sum_{j=1}^N f(\xi_k^{(i)}, \xi_{k+1}^{(j)}) \frac{\omega_k^{(i)} m(\xi_k^{(i)}, \xi_{k+1}^{(j)})}{\sum_{l=1}^N \omega_k^{(l)} m(\xi_k^{(l)}, \xi_{k+1}^{(j)})} \omega_{k+1|n}^{(j)} \tag{5.48}$$

where $f \in \mathbf{F}_b\left(\mathcal{X}^2\right)$. As one can see it, (5.47) remains highly impractical since it's computational cost is exponential in the number of operations (see Douc et al. [2011] for more details and extensions). Alternative Monte Carlo smoothing can also be achieved with less computational effort see for instance Godsill et al. [2004], Cappé et al. [2007] among others. (5.45) also inherits the $O(N^2)$ computational cost at each time step.

5.3.2.2 High order HMM

We now turn back to the adaptation of the FB algorithm to high-order-HMM. In the following we give some elementary keys required before stating the alternative FB algorithm for higher order HMM.

Lemma 5.3.5. *The joint smoothing density $p(x_{-\ell:n}|y_{0:n})$ can be factorized as*

$$p(x_{-\ell:n}|y_{0:n}) = p(x_{n-\ell+1:n}|y_{0:n}) \prod_{k=-\ell}^{n-\ell} p(x_k|x_{k+1:n}; y_{0:n}) \tag{5.49}$$

Proof. This factorization is easily obtained via Bayes rule and model's properties. □

To state the smoothing problem, note that for any function $f \in \mathbf{F}_b(\mathcal{X}^{\ell+1})$

$$\begin{aligned}&\mathbb{E}\left[f(X_{k:k+\ell-1}; X_{k+\ell})|\, Y_{0:k+\ell-1}\right] = \\ &\int_{\mathcal{X}^{\ell+1}} f(x_{k:k+\ell-1}; x_{k+\ell}) m(\underline{x}_{k+\ell-1}; x_{k+\ell}) \phi_{\nu,k:k+\ell-1|k+\ell-1}(dx_{k:k+\ell-1}) dx_{\ell+1}\end{aligned} \tag{5.50}$$

where $\phi_{\nu,k:k+\ell-1|k+\ell-1}$ is the conditional distribution of $X_{k:k+\ell-1}$ given $Y_{0:k+\ell-1}$. With (5.49), if one can design a reverse kernel $B_{\nu,k+\ell-1}$ from $(\mathcal{X}^{\ell}, \mathbb{B}(\mathcal{X}^{\ell}))$ to $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$ such that :

$$\begin{aligned}\mathbb{E}\left[f(X_{k:k+\ell-1}; X_{k+\ell})\middle| Y_{0:k+\ell-1}\right] &= \int_{\mathcal{X}^{\ell+1}} f(x_{k:k+\ell-1}; x_{k+\ell}) \\ &\times \phi_{\nu,k+1:k+\ell|k+\ell-1}(dx_{k+1:k+\ell}) B_{\nu,k+\ell-1}(x_{k+1:k+\ell}; dx_k)\end{aligned} \tag{5.51}$$

where $\phi_{\nu,k+1:k+\ell|k+\ell-1} = \phi_{\nu,k:k+\ell-1|k+\ell-1} M$ is the one-step predictive distribution of $\phi_{\nu,k:k+\ell-1|k+\ell-1}$, then the following result holds.

Proposition 5.3.6. *Given an initial distribution ν on the state sequences $X_{-\ell:-1}$, a strictly positive index n and index $k \in \{-\ell+1, -\ell+2, \ldots, n-\ell\}$,*

$$\mathbb{E}\left[f(X_k)\middle| X_{k+1:n}, Y_{0:n}\right] = B_{\nu,k+\ell-1}(X_{k+1:k+\ell}, f) \tag{5.52}$$

where $B_{\nu,k+\ell-1}$ is the backward kernel defined in (5.51) *for any $f \in \mathbf{F}_b(\mathcal{X})$.*

Proof. see chapitre 4, Prop. 4.3.2. □

Prop. 5.3.6 shows that conditionally on the whole observation $Y_{0:n}$ the process $(X_{n-k})_{k\geq 0}$ is a Markov chain backward in time with transition densities given by

$$\begin{aligned}B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, x_k) &= p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \\ &\propto p(x_{k+\ell}|x_{k:k+\ell-1}) p(x_{k:k+\ell-1}|y_{0:k+\ell-1}).\end{aligned} \tag{5.53}$$

As a consequence of Prop. 5.3.6, the joint smoothing distribution $\phi_{-\ell:n|n}$ can be stated in terms of backward kernels.

Corollary 5.3.7. *For any initial distribution ν on the state sequences $X_{-\ell:-1}$, $n > 0$ and $f \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$,*

$$\begin{aligned}\mathbb{E}\left[f(X_{-\ell:n})\middle| Y_{0:n}\right] = \int_{\mathcal{X}^{\ell}} &\left[\int_{\mathcal{X}^{n+1}} f(x_{-\ell:n}) \prod_{k=-\ell}^{n-\ell} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, dx_k)\right] \\ &\times \phi_{\nu,n-\ell+1:n|n}(dx_{n-\ell+1:n}).\end{aligned} \tag{5.54}$$

Remark 5.3.8. *From previous results, it is clear that approximating smoothing quantities backward in time in the context of higher-order HMM requires the evaluation forward in time of conditional distributions of size ℓ that is $\phi_{\nu,k:k+\ell-1|k+\ell-1}$. To achieve this end, we have at least in theory two options :*

1. *a natural solution can be designed directly from the SISR algorithm as an output ;*
2. *another solution is achieved when mimicking the* two-stages like *as in the derivation of the filtering distribution approximations (prediction and update steps) to be able to approximate the distributions*

$$\phi_{\nu,k:k+\ell-1|k+\ell-1}, \quad \forall -\ell+1 \leq k \leq n-\ell+1;$$

In the following, we investigate the feasibility of these two options.

Forward Sampling 1

Assume one has run the SISR algorithm on the whole data set forward in time, to get a cloud of weighted particles $\left\{\omega_n^{(i)}, \xi_{-\ell:n}^{(i)}\right\}_{i=1}^N$ targeting the joint smoothing $\phi_{\nu,-\ell:n|n}$. Recycling draws from the former run, one could get a particle estimate of the smoothing distributions of interest $\phi_{\nu,k:k+\ell-1|k+\ell-1}$:

$$\forall -\ell+1 \leq k \leq n-\ell+1, \quad \hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(dz_{1:\ell}) = \Omega_n^{-1} \sum_{i=1}^N \omega_n^{(i)} \delta_{\xi_{k:k+\ell-1}^{(i)}}(dz_{1:\ell}), \quad (5.55)$$

where $\Omega_n = \sum_{i=1}^N \omega_n^{(i)}$, that is when extracting the required sub-sequences from the previous path particle, $\left\{\xi_{k:k+\ell-1}^{(i)}\right\}_{i=1}^N$ with the same weights $\left\{\omega_n^{(i)}\right\}_{i=1}^N$. This approximation inherits among others advantages (cheap computational cost) and drawbacks (sample depletion induced by the resampling step) of the SISR algorithm.

Forward Sampling 2

In order to derive similar recursions as in the filtering distribution approximations, one may need to establish a recursive relationship between $\phi_{\nu,k:k+\ell-1|k+\ell-1}$ and $\phi_{\nu,k-1:k+\ell-2|k+\ell-2}$ provided they are well defined for suitable indexes. The following recursion holds :

Lemma 5.3.9. *For any function $f \in \mathbf{F}_b(\mathcal{X}^\ell)$,*

$$\begin{aligned}\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) \propto \int_{\mathcal{X}^{\ell+1}} & f(x_{k:k+\ell-1}) \phi_{\nu,k-1:k+\ell-2|k+\ell-2}(x_{k-1:k+\ell-2}) \\ & \times m(\underline{x}_{k+\ell-2}; x_{k+\ell-1}) g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1}) dx_{k-1:k+\ell-1}.\end{aligned} \quad (5.56)$$

Proof. see chapitre 4, Prop.4.3.1. □

Assume one has a cloud of weighted particles $\left\{\omega_{k+\ell-2}^{(i)}, \xi_{k-1:k+\ell-2}^{(i)}\right\}_{i=1}^{N}$ targeting $\phi_{\nu,k-1:k+\ell-2|k+\ell-2}$ in the sense that :

$$\hat{\phi}_{\nu,k:k+\ell-2|k+\ell-2}(dz_{1:\ell}) = \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)} \delta_{\xi_{k-1:k+\ell-2}^{(i)}}(dz_{1:\ell}), \tag{5.57}$$

where $\Omega_{k+\ell-2} = \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)}$. Substituting the latter approximation in (5.56) one could get a first particle estimate of the distribution of interest :

$$\begin{aligned}
&\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(f) \\
&\propto \int_{\mathcal{X}} \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} f(\xi_{k:k+\ell-2}^{(i)}, x_{k+\ell-1}) \omega_{k+\ell-2}^{(i)} m(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) \\
&\times g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) dx_{k+\ell-1} \\
&= \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)} \int_{\mathcal{X}} f(\xi_{k:k+\ell-2}^{(i)}, x_{k+\ell-1}) m(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) \\
&\times g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) dx_{k+\ell-1}.
\end{aligned} \tag{5.58}$$

Note that the appearing integral in (5.58) may be easily evaluated by considering the latter either as expectation under the density $g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, \cdot)$ or the density $m(\xi_{k-1:k+\ell-2}^{(i)}, \cdot)$ which are both Gaussian.

The required steps for the two approaches are summarized in the following result.

Proposition 5.3.10. *(Alternative Forward/Backward sampling)*
FORWARD SMOOTHING : *Compute, forward in time the joint smoothing distributions* $\phi_{\nu,-\ell+1:0|0}, \phi_{\nu,1-\ell+1:1|1}, \phi_{\nu,2-\ell+1:2|2}, \ldots, \phi_{\nu,n-\ell+1:n|n}$ *using* (5.55) *or* (5.58).
BACKWARD SMOOTHING : *For* $k = n - \ell$ *down to* $-\ell + 1$, *using* (5.51), *compute*

$$\phi_{\nu,k:k+\ell|k+\ell-1} = \phi_{\nu,k+1:k+\ell|k+\ell-1} B_{\nu,k+\ell-1}$$

where $\phi_{\nu,k+1:k+\ell|k+\ell-1} = \phi_{\nu,k:k+\ell-1|k+\ell-1} M$.

Once the backward kernel stated, one may face the evaluation of the smoothing distributions in higher order Markov chain specifically, those needed in the EM algorithm.

Marginal smoothing

In order to have particle estimate of the smoothing distribution, one may use the following link between $\phi_{\nu,k|n}$ and $\phi_{\nu,k+1:k+\ell|n}$.

Lemma 5.3.11. *Let n, ℓ, k be indexes such that $n > 0$, $\ell \geq 1$ and $-\ell \leq k \leq n-\ell$. For any function $f \in \mathbf{F}_b(\mathcal{X})$,*

$$\begin{aligned}\phi_{\nu,k|n}(f) &= \int_{\mathcal{X}^{\ell+1}} f(x_k) B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, dx_k) \\ &\times \phi_{\nu,k+1:k+\ell|n}(dx_{k+1:k+\ell}) \\ &= \int_{\mathcal{X}^{\ell}} \left[\int_{\mathcal{X}} f(x_k) B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, dx_k) \right] \phi_{\nu,k+1:k+\ell|n}(dx_{k+1:k+\ell}).\end{aligned} \tag{5.59}$$

Proof. This recursion is straightforward observing that :

$$\begin{aligned}p(x_k|y_{0:n}) &= \int_{\mathcal{X}^{\ell}} p(x_{k:k+\ell}|y_{0:n}) dx_{k+1:k+\ell} \\ &= \int_{\mathcal{X}^{\ell}} p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) p(x_{k+1:k+\ell}|y_{0:n}) dx_{k+1:k+\ell} \\ &= \int_{\mathcal{X}^{\ell}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, x_k) \phi_{\nu,k+1:k+\ell|n}(dx_{k+1:k+\ell})\end{aligned} \tag{5.60}$$

Using (5.60), the conditional expectation of $f(X_k)$ given $Y_{0:n}$ follows. □

Assume one has an approximation $\hat{B}_{\nu,k+\ell-1}$ of the backward kernel $B_{\nu,k+\ell-1}$ given by

$$\hat{B}_{\nu,k+\ell-1}(x_{k+1:k+\ell}, dx_k) = \sum_{i=1}^{N} \frac{\omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, x_{k+\ell})}{\sum_{r=1}^{N} \omega_{k+\ell-1}^{(r)} m(\xi_{k:k+\ell-1}^{(r)}, x_{k+\ell})} \delta_{\xi_k^{(i)}}(dx_k) \tag{5.61}$$

where $\left\{ \omega_{k+\ell-1}^{(i)}, \xi_{k:k+\ell-1}^{(i)} \right\}_{i=1}^{N}$ is a cloud of weighted particles targeting the distribution $\phi_{\nu,k:k+\ell-1|k+\ell-1}$ in the sense that :

$$\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(dz_{1:\ell}) = \Omega_{k+\ell-1}^{-1} \sum_{i=1}^{N} \omega_{k+\ell-1}^{(i)} \delta_{\xi_{k:k+\ell-1}^{(i)}}(dz_{1:\ell}), \tag{5.62}$$

where $\Omega_{k+\ell-1} = \sum_{i=1}^{N} \omega_{k+\ell-1}^{(i)}$. Combining the later estimate with a weighted particles $\left\{ \omega_{k+1:k+\ell|n}^{(j)}, \xi_{k+1:k+\ell}^{(j)} \right\}_{j=1}^{N}$ targeting the distribution $\phi_{\nu,k+1:k+\ell|n}$, one can easily derive a particle estimate of the marginal smoothing distribution given by :

$$\hat{\phi}_{\nu,k|n}(f) = \sum_{j=1}^{N} \sum_{i=1}^{N} f(\xi_k^{(i)}) \frac{\omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, \xi_{k+\ell}^{(j)})}{\sum_{r=1}^{N} \omega_{k+\ell-1}^{(r)} m(\xi_{k:k+\ell-1}^{(r)}, \xi_{k+\ell}^{(j)})} \omega_{k+1:k+\ell|n}^{(j)} \tag{5.63}$$

As it appears, the computational cost of this estimate is of the same magnitude as in a $1-order-$ HMM since it requires $O(N^2)$ operations at each time step.

Joint smoothing

The next joint smoothing distribution is the one effectively required in the EM algorithm. It iterates backward in time as follows.

Lemma 5.3.12. *Let $-\ell \leq p, m \leq n$ with $|p-m| \leq \ell+1$, $p \leq m$ and $n > 0$. For any function $f \in \mathbf{F}_b(\mathcal{X}^2)$,*

$$\phi_{\nu,p,m|n}(f) = \int_{\mathcal{X}^{\ell+1}} f(x_p, x_m) B_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p) \phi_{\nu,p+1:p+\ell|n}(dx_{p+1:p+\ell}) \tag{5.64}$$

Proof. It is exactly analogous in all respects to the previous lemma. □

As in the marginal case, given both an approximation $\hat{B}_{\nu,p+\ell-1}$ of the backward kernel $B_{\nu,p+\ell-1}$:

$$\hat{B}_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p) = \sum_{i=1}^{N} \frac{\omega_{p+\ell-1}^{(i)} m(\xi_{p:p+\ell-1}^{(i)}, x_{p+\ell})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, x_{p+\ell})} \delta_{\xi_p^{(i)}}(dx_p) \tag{5.65}$$

where $\left\{\omega_{p+\ell-1}^{(i)}, \xi_{p:p+\ell-1}^{(i)}\right\}_{i=1}^{N}$ and $\left\{\omega_{p+1:p+\ell|n}^{(j)}, \xi_{p:p+\ell}^{(j)}\right\}_{j=1}^{N}$ are sets of weighted particles targeting respectively the distribution $\phi_{\nu,p:p+\ell-1|p+\ell-1}$ and $\phi_{\nu,p+1:p+\ell|n}$. Then, a particle estimate of the joint smoothing distribution is easily derived by :

$$\hat{\phi}_{\nu,p,m|n}(f) = \sum_{j=1}^{N} \sum_{i=1}^{N} f(\xi_p^{(i)}, \xi_m^{(j)}) \frac{\omega_{p+\ell-1}^{(i)} m(\xi_{p:p+\ell-1}^{(i)}, \xi_{p+\ell}^{(j)})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, \xi_{p+\ell}^{(j)})} \omega_{p+1:p+\ell|n}^{(j)}. \tag{5.66}$$

From Cor. 5.3.7, a particle estimate of the joint smoothing distribution (5.54) can be achieved. In fact, consider at first particle approximations of the backward kernels $B_{\nu,k+\ell-1}$, $k = -\ell, -\ell+1, \ldots, n-\ell$ given by

$$\hat{B}_{\nu,k+\ell-1}(x_{k+1:k+\ell}, dx_k) = \sum_{i_k=1}^{N} \frac{\omega_{k+\ell-1}^{(i_k)} m(\xi_{k:k+\ell-1}^{(i_k)}, x_{k+\ell})}{\sum_{r=1}^{N} \omega_{k+\ell-1}^{(r)} m(\xi_{k:k+\ell-1}^{(r)}, x_{k+\ell})} \delta_{\xi_k^{(i_k)}}(dx_k). \tag{5.67}$$

where $\left\{\xi_{k:k+\ell-1}^{(i_k)}, \omega_{k+\ell-1}^{(i_k)}\right\}_{i_k=1}^{N}$ are sets of weighted particles targeting the smoothing distribution $\phi_{\nu,k:k+\ell-1|k+\ell-1}$. Consider also a particle approximation of the smoothing distribution $\phi_{\nu,n-\ell+1:n|n}$ given by

$$\hat{\phi}_{\nu,n-\ell+1:n|n}(dz_{1:\ell}) = \Omega_n^{-1} \sum_{i_n=1}^{N} \omega_n^{(i_n)} \delta_{\xi_{n-\ell+1:n}^{(i_n)}}(dz_{1:\ell}). \tag{5.68}$$

Combining the former estimates (5.67) and (5.68), a particle estimate of (5.54) is then easily derived by

$$\hat{\phi}_{\nu,-\ell:n|n}(f) = \Omega_n^{-1} \sum_{i_n=1} \left[\sum_{i_{-\ell}=1}^{N} \cdots \sum_{i_{n-\ell}=1}^{N} f(\xi_{-\ell}^{(i_{-\ell})}, \ldots, \xi_{n-\ell}^{(i_{n-\ell})}, \xi_{n-\ell+1:n}^{(i_n)}) \times \prod_{k=-\ell}^{n-\ell} \frac{\omega_{k+\ell-1}^{(i_k)} m(\xi_{k:k+\ell-1}^{(i_k)}, \xi_{k+\ell}^{(i_{k+\ell})})}{\sum_{r=1}^{N} \omega_{k+\ell-1}^{(r)} m(\xi_{k:k+\ell-1}^{(r)}, \xi_{k+\ell}^{(i_{k+\ell})})} \right] \omega_n^{(i_n)} \tag{5.69}$$

where $\Omega_r = \sum_{r=1}^{N} \omega_r^{(r)}$, $f \in \mathbf{F}_b\left(\mathcal{X}^{n+\ell+1}\right)$. As on can noted, the evaluation of the joint smoothing distribution backward in time is expensive. To have moderate computational cost one may use similar derivation as in Godsill et al. [2004].

Alternative joint smoothing

Assume one has run one of the SMC methods mentioned above to get the weighted particles

$$\left\{ \omega_{k+\ell-1}^{(i)}, \xi_{k:k+\ell-1}^{(i)} \right\}_{i=1}^{N}, \quad -\ell+1 \leq k \leq n-\ell$$

approximating the smoothing densities $p(x_{k:k+\ell-1}|y_{0:k+\ell-1})$. From Lemma 5.3.5 and (5.53) the joint smoothing density is given by

$$p(x_{-\ell:n}|y_{0:n}) = p(x_{n-\ell+1:n}|y_{0:n}) \prod_{k=-\ell}^{n-\ell} p(x_k|x_{k+1:n}; y_{0:n}) \tag{5.70}$$

where

$$p(x_k|x_{k+1:n}, y_{0:n}) = p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \propto p(x_{k+\ell}|x_{k:k+\ell-1}) p(x_{k:k+\ell-1}|y_{0:k+\ell-1}).$$

Using the previous weighted sample one could get a particle estimate of $p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1})$:

$$p(dx_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \approx \sum_{i=1}^{N} \kappa_k^{(i)} \delta_{\xi_k^{(i)}}(dx_k) \tag{5.71}$$

where the modified weights are given by

$$\kappa_k^{(i)} = \frac{\omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, x_{k+\ell})}{\sum_{j=1}^{N} \omega_{k+\ell-1}^{(j)} m(\xi_{k:k+\ell-1}^{(j)}, x_{k+\ell})} \tag{5.72}$$

With these modified weights, one can simulate consecutive states in the reverse-time as follows : Let $\tilde{x}_{k+1:n}$ be a random sample drawn from $p(x_{k+1:n}|y_{0:n})$,

step back in time and draw $\tilde{x}_k$ from $p(x_k|\tilde{x}_{k+1:n}, y_{0:n})$. The pair $(\tilde{x}_k, \tilde{x}_{k+1:n})$ is an approximate random realization of $p(x_{k:n}|y_{0:n})$. Iterating this mechanism backward in time one gets the following smoothing algorithm :

Algorithme 20 Alternative Smoothing algorithm

1: Choose $\tilde{\xi}_{n-\ell+1:n} = \xi^{(i)}_{n-\ell+1:n}$ with probability $\omega^{(i)}_n$
2: For $k = n-\ell$ down to $-\ell$ do
— Evaluate $\kappa^{(i)}_k \propto \omega^{(i)}_{k+\ell-1} m(\xi^{(i)}_{k:k+\ell-1}, \tilde{\xi}_{k+\ell})$, for $i = 1, \ldots, N$;
— Choose $\tilde{\xi}_k = \xi^{(i)}_k$ with probability $\kappa^{(i)}_k$
3: EndFor
4: $\tilde{\xi}_{-\ell:n} = (\tilde{\xi}_{-\ell}, \tilde{\xi}_{-\ell+1}, \ldots, \tilde{\xi}_n)$ is an approximate random realization from $p(x_{-\ell:n}|y_{0:n})$.

The computational complexity is $O(N)$ at each time step which compares favorably to the $O(N^2)$ of the marginal smoothing.

5.3.3 Application

In this section we are look at the actual calibration of this model to simulated data. We choose $\ell = 3$. So, the parameter vector is given by $\theta^{(*)} = (\pi_1^{(*)}, \pi_2^{(*)}, \pi_3^{(*)}, \sigma^{(*)}, \beta^{(*)}, \rho^{(*)})$ with $\pi_1^{(*)} = 0.5$, $\pi_2^{(*)} = 0.29$, $\pi_3^{(*)} = -0.1$, $\sigma^{(*)} = 0.2$, $\beta^{(*)} = 0.5$ and $\rho^{(*)} = -0.4$. A sample path of the generated model is given at Figure 5.1. The MCEM was initialized with $\theta^{(0)} = (0.1, 0.1, 0.1, 0.01, 0.2, -0.7)$.

One could notice that the last 3 parameters are well estimated in contrast to the AR coefficients. It seems that the smoothing weights have to be readjusted. This point will be investigated further.

5.3.4 GEM convergence

In order to assess at somehow $Q(\theta_{k+1}, \theta_k)$ evolves as k goes to infinity, that is to know how close θ_{k+1} is to the optimum (stationary) parameter value, θ^* say, one may need some kind of convergence of the sequence $(L(\theta_k))_{k\geq 0}$ towards L^* that realizes $L^* = L(\theta^*)$. In a practical view, since the incomplete likelihood data is not available one may use heuristic measures of closeness. A simple way is to fix arbitrary the number of iterations of the MCEM. Another view is to use the relative likelihood as given in Kim and Stoffer [2006] originally from Chan and Ledolter [1995]. Even if one could compute it explicitly, because of the approximations done in the E-step, the monotonicity

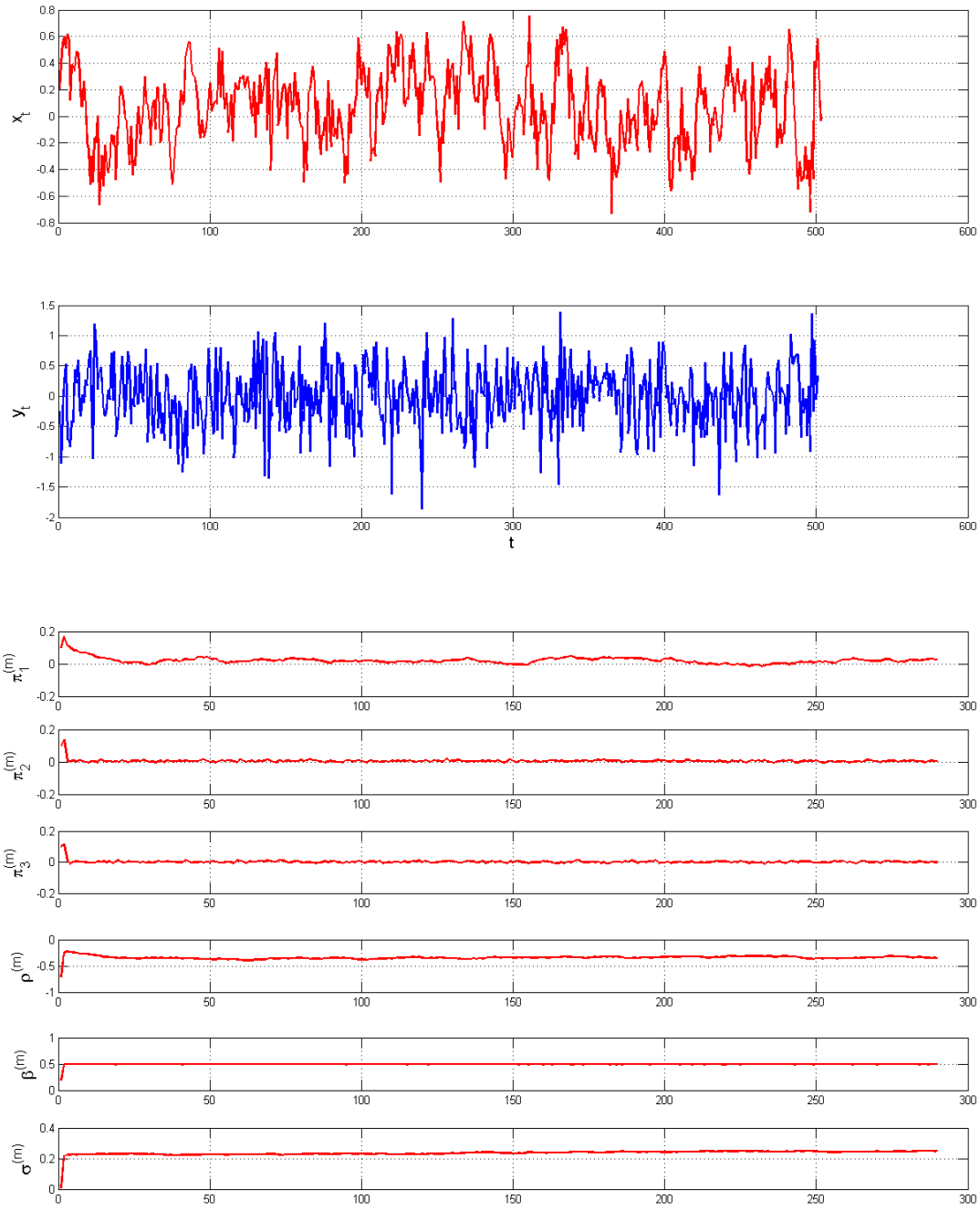


FIGURE 5.1 – Sample path and MCEM iterations

property of the incomplete likelihood is not ensured. Let $L_n(\theta_k)$ (resp. $L_n(\theta_{k+1})$) be the incomplete likelihood data at iteration k (resp. $k+1$) of the data $y_{0:n}$ emphasizing its dependence to the model's parameter. The relative likelihood $R(\theta_k, \theta_{k+1})$ is defined by

$$R(\theta_k, \theta_{k+1}) := \frac{L_n(\theta_{k+1})}{L_n(\theta_k)} \tag{5.73}$$

which could be used for monitoring the GEM convergence observing that :

$$\begin{aligned} R(\theta_k, \theta_{k+1}) &= \frac{p_{\theta_{k+1}}(x_{-\ell:n}, y_{0:n})}{p_{\theta_{k+1}}(x_{-\ell:n}|y_{0:n})} \times \frac{p_{\theta_k}(x_{-\ell:n}|y_{0:n})}{p_{\theta_k}(x_{-\ell:n}, y_{0:n})} \\ &= \int_{\mathcal{X}^{n+\ell+1}} \left[\frac{p_{\theta_{k+1}}(x_{-\ell:n}, y_{0:n})}{p_{\theta_{k+1}}(x_{-\ell:n}|y_{0:n})} \times \frac{p_{\theta_k}(x_{-\ell:n}|y_{0:n})}{p_{\theta_k}(x_{-\ell:n}, y_{0:n})} \right] p_{\theta_{k+1}}(x_{-\ell:n}|y_{0:n}) dx_{-\ell:n} \\ &= \int_{\mathcal{X}^{n+\ell+1}} \left[\frac{p_{\theta_{k+1}}(x_{-\ell:n}, y_{0:n})}{p_{\theta_k}(x_{-\ell:n}, y_{0:n})} \right] p_{\theta_k}(x_{-\ell:n}|y_{0:n}) dx_{-\ell:n} \\ &= \mathbb{E}\left[\frac{p_{\theta_{k+1}}(X_{-\ell:n}, Y_{0:n})}{p_{\theta_k}(X_{-\ell:n}, Y_{0:n})} \Big| Y_{0:n}; \theta_k \right]. \end{aligned} \tag{5.74}$$

Taking the log-likelihood of the ratio leads to

$$\begin{aligned} &\Delta l(\theta_k, \theta_{k+1}) := \log R(\theta_k, \theta_{k+1}) \\ &= \log L_n(\theta_{k+1}) - \log L_n(\theta_k) = \log \mathbb{E}\left[\frac{p_{\theta_{k+1}}(X_{-\ell:n}, Y_{0:n})}{p_{\theta_k}(X_{-\ell:n}, Y_{0:n})} \Big| Y_{0:n}; \theta_k \right]. \end{aligned} \tag{5.75}$$

Assume one has a weighted particle $\left\{ \omega_n^{(i)}, \xi_{-\ell:n}^{(i)} \right\}_{i=1}^N$ targeting the smoothing density $p_{\theta_k}(X_{-\ell:n}|Y_{0:n})$, one could get an estimate $\hat{\Delta} l(\theta_k, \theta_{k+1})$ of $\Delta l(\theta_k, \theta_{k+1})$ given by :

$$\hat{\Delta} l(\theta_k, \theta_{k+1}) = \log \left[\sum_{i=1}^N \omega_n^{(i)} \frac{p_{\theta_{k+1}}(\xi_{-\ell:n}^{(i)}, y_{0:n})}{p_{\theta_k}(\xi_{-\ell:n}^{(i)}, y_{0:n})} \right]. \tag{5.76}$$

Conclusion

Dans cette thèse nous avons parcouru les grandes lignes de l'algorithme EM ainsi que celles des méthodes MCs au travers de quelques unes de leurs propriétés. Nous avons également mis en relief le MCEM comme méthode d'estimation dans les MMC et qui consiste en la mise en commun de ces deux concepts.

Une première illustration a été faite sur le modèle de volatilité stochastique canonique. La deuxième illustration s'est portée sur une extension de ce modèle qui met à rude épreuve le MCEM. Cette extension a été ponctuée par des simulations de Monte Carlo à des fins de validation. Nous avons aussi exploré l'extension des méthodes de MCs au cas des MMC d'ordre supérieur. Notamment, lorsqu'on est en présence d'une forme de dépendance plus accrue dans le signal caché ou observé et ce par l'intermédiaire de problématique de filtrage et de lissage particulaires. L'exemple illustratif a été celui d'un modèle de volatilité stochastique dégénéré qui a été approché puis estimé par MCEM.

Cependant, il faut dire que ce travail est loin d'être achevé. En effet, l'étude de la stabilité des distributions conditionnelles de filtrage et de lissage dans les MMC d'ordre supérieur est plus délicate. En effet, dans le cas habituel on peut facilement mettre en exergue des différences de martingales en évoquant des propriétés de mélanges, d'oubli ou d'ergodicité du modèle. Il semble qu'un autre point de vue soit requis pour aborder cette stabilité.

ANNEXE A

Annexe A

A.1 Variables aléatoires et chaînes de Markov

Nous donnons quelques résultats utilisés dans le chapitre 2 et basés entre autres sur Petrov [1995], Cappé et al. [2005].

A.1.1 Sommes de variables aléatoires

On considère un espace probabilisé (Ω, A, P). Soient $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires arbitraires définies sur ce dernier. On note par $S_N := \sum_{i=1}^{N} X_i$.

Théorème A.1.1.

$$\mathbb{E}\,|S_N|^k \leq \sum_{i=1}^{N} \mathbb{E}\,|X_i|^k \quad si \quad 0 < k \leq 1 \tag{A.1}$$

et

$$\mathbb{E}\,|S_N|^k \leq N^{k-1} \sum_{i=1}^{N} \mathbb{E}\,|X_i|^k \quad si \quad k > 1 \tag{A.2}$$

Démonstration. Il suffit d'appliquer les inégalités élémentaires suivantes :

$$\left|\sum_{i=1}^{N} a_i\right|^k \leq \sum_{i=1}^{N} |a_i|^k \quad \text{pour} \quad 0 < k \leq 1 \tag{A.3}$$

et,

$$\left|\sum_{i=1}^{N} a_i\right|^k \leq N^{k-1} \sum_{i=1}^{N} |a_i|^k \quad \text{pour} \quad k > 1 \tag{A.4}$$

où $a_i, i = 1, 2, \ldots, N$ sont des réels quelconques. □

Dans la suite les variables $X_1, X_2, \ldots, X_N$ sont i.i.d.

Théorème A.1.2. *Supposons que* $\mathbb{E}\, X_i = 0$ *pour tout* i. *Soit* $k \geq 2$. *On pose* $M_{k,N} := \sum_{i=1}^{N} |X_i|^k$ *et* $B_N := \sum_{i=1}^{N} \mathbb{E}\, X_i^2$. *Alors*

$$\mathbb{E}\,|S_N|^k \leq C(k) \left(M_{k,N} + B_N^{k/2}\right) \quad \textit{Inégalité de Rosenthal} \tag{A.5}$$

où $C(k)$ *est une constante dépendant uniquement de* k.

Démonstration. Voir Petrov [1995], page 59. □

Les deux résultats suivants sont ceux utilisés au chapitre 2 pour établir le moment d'ordre k ainsi l'inégalité de Hoeffding pour les estimateurs de l'EP et l'EP autonormalisé.

Théorème A.1.3. *Soient* $X_1, X_2, \ldots, X_N$ *une suite de variables aléatoires centrées et indépendantes et* $k \geq 2$. *Alors,*

$$\mathbb{E}\,|S_N|^k \leq C(k) N^{k/2-1} \sum_{i=1}^{N} \mathbb{E}\,|X_i|^k \tag{A.6}$$

où $C(k)$ *est une constante dépendant uniquement de* k.

Démonstration. D'après l'inégalité de Lyaponov, pour toute variable aléatoire Y

$$[\mathbb{E}\,|Y|^r]^{1/r} \leq [\mathbb{E}\,|Y|^s]^{1/s} \quad \text{avec} \quad 0 < r < s. \tag{A.7}$$

Posons $Y := \frac{1}{N} \sum_{i=1}^{N} X_i$ et notons par $V(x_i)$ la densité de la variable $X_i, i = 1, 2, \ldots, N$. Alors, soit

$$(B_N/N)^{1/2} \leq (M_{k,N}/N)^{1/k} \quad \text{pour tout} \quad k \geq 2 \tag{A.8}$$

soit

$$(B_N)^{k/2} \leq N^{k/2-1} (M_{k,N}) \tag{A.9}$$

D'où en combinant avec l'inégalité de Rosenthal (A.5) on aboutit au résultat. □

Théorème A.1.4. *Soient $X_1, X_2, \ldots, X_N$ une suite de variables aléatoires indépendantes telles qu'il existe $a_1, b_1, a_2, b_2, \ldots, a_N, b_N \in \mathbb{R}$ vérifiant $P(a_i \leq X_i \leq b_i) = 1$ pour $i = 1, 2, \ldots, N$. Alors pour tout $t \geq 0$,*

$$P(S_N - \mathbb{E}\, S_N \geq t) \leq \exp\left(-2t^2 / \sum_{i=1}^{N} (b_i - a_i)^2\right) \tag{A.10}$$

et

$$P(S_N - \mathbb{E}\, S_N \leq -t) \leq \exp\left(-2t^2 / \sum_{i=1}^{N} (b_i - a_i)^2\right). \tag{A.11}$$

Démonstration. On peut se reporter à Hoeffding [1963]. □

Nous donnons un dernier lemme de Hoeffding généralisé utilisé pour établir une inégalité de déviation pour l'erreur de lissage

Lemme A.1.5. *Soient a_N, b_N et b des variables aléatoires définies sur le même espace probabilisé telles qu'il existe des constantes positives β, B, C et M satisfaisant :*

$$(I) \quad |a_N / b_N| \leq M \ \ \mathbb{P} \ \ p.s \ \ \text{et } b \geq \beta \ \ \mathbb{P} \ \ p.s$$

$$(II) \quad \text{Pour tout } \ \epsilon > 0, N \geq 1, \ \ \mathbb{P}\left(|a_N - b_N| > \epsilon\right) \leq B e^{-CN\epsilon^2}$$

$$(III) \quad \text{Pour tout } \ \epsilon > 0, N \geq 1, \ \ \mathbb{P}\left(|a_N| > \epsilon\right) \leq B e^{-CN(\epsilon/M)^2}$$

alors,

$$\mathbb{P}\left(|a_N / b_N| > \epsilon\right) \leq B e^{-CN(\epsilon/2M)^2}.$$

Démonstration. Voir Douc et al. [2011], Lemma 4. □

A.1.2 Chaînes de Markov

Soient $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ et $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ deux espaces mesurables.

Définition A.1.6. *On appelle noyau de transition non normalisé de $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ vers $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ une fonction $K : \mathcal{X} \times \mathcal{B}(\mathcal{Y}) \to [0, \infty]$ vérifiant :*

- *$\forall x \in \mathcal{X}$, $K(x, \cdot)$ est une mesure positive sur $\mathcal{Y}$;*
- *$\forall A \in \mathcal{B}(\mathcal{Y})$, la fonction $x \mapsto K(x, A)$ est mesurable.*

Définition A.1.7. *Un noyau de transition non normalisé est dit admettre une densité par rapport à une mesure positive μ sur $\mathcal{Y}$, s'il existe une fonction positive $k : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ mesurable par rapport à la tribu produit $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$ vérifiant :*

$$K(x, A) = \int_A k(x, y) \mu(dy), \quad A \in \mathcal{B}(\mathcal{Y}). \tag{A.12}$$

Remarque A.1.8. *Lorsque $K(x, \mathcal{Y}) = 1$, K est appelé noyau de transition. Si en plus on a $\mathcal{X} = \mathcal{Y}$, on convient d'appeler K noyau de transition de Markov.*

Remarque A.1.9. *Lorsque $\mathcal{X}$ et $\mathcal{Y}$ sont dénombrables et de cardinaux finis, K est appelé matrice de transition (de probabilité ou stochastique).*

Nous allons donner quelques propriétés relatives aux opérations sur les noyaux.

Propriétés A.1.10. *Soit K (resp. Q) un noyau de transition non-normalisé de $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ vers $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ (resp. $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ vers $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$).*

1. *le produit KQ définie par :*

$$KQ(x, A) := \int K(x, dy)Q(x, A), \quad x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{Y}) \tag{A.13}$$

est un noyau de transition non-normalisé de $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ vers $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

2. *Si K est un noyau de transition ses itérés définies par la relation*

$$\begin{cases} K^n = KK^{n-1}, & pour \quad n \geq 1 \\ K^0(x, .) = \delta_x(.), & x \in \mathcal{X} \end{cases} \tag{A.14}$$

satisfont l'équation de Chapman-Kolmogorov :

$$K^{n+m} = K^n K^m, \quad m, n \geq 0 \tag{A.15}$$

et pour tous $x \in \mathcal{X}$ et $A \in \mathcal{B}(\mathcal{X})$,

$$K^{n+m}(x, A) = \int K^n(x, dy)K^m(y, A) \tag{A.16}$$

avec δ_x une mesure de Dirac localisé en x ;

Définition A.1.11. *On dit que le noyau K_t satisfait à la propriété de Feller si pour tout $t > 0$ la fonction $K_t f$ définie par :*

$$K_t f(x) := \int f(y)K_t(x, dy) \tag{A.17}$$

est continue pour toute fonction continue bornée f.

A.2 Mesure de dégénérescence

Il existe plusieurs mesures heuristiques permettant de jauger le phénomène de dégénérescence des particules. On peut citer entre autres le coefficient de variation (voir Kong et al. [1994]), le nombre de particules efficaces appelé *the Effective Sample Size* en anglais (voir Liu and Chen [1995] et Liu [1996]) ou l'entropie de Shannon (voir Shannon [1948]).

A.2.1 Coefficient de variation

Il est donné par la formule suivante

$$CV_N(t) := \frac{1}{N}\left[\sum_{i=1}^{N}\left(N\frac{\omega_t^{(i)}}{\sum_{j=1}^{N}\omega_t^{(j)}}-1\right)^2\right]^{1/2} \tag{A.18}$$

Cette mesure est maximale lorsque :

$$\exists i_0 \in \{1,2,\ldots,N\} \text{ tel que } \omega_t^{(i_0)} = 1 \text{ et } \omega_t^{(i)} = 0 \forall i \neq i_0 \tag{A.19}$$

au quel cas $CV_N(t) = \sqrt{N-1}$. Cette situation correspond à l'état de dégénérescence complète où une seule particule à un poids valant 1 et toutes les autres particules ont des poids nuls. Inversement, CV_N est minimale lorsque

$$\forall i \in\in \{1,2,\ldots,N\}, \omega_t^{(i)} = \frac{1}{N}. \tag{A.20}$$

Par conséquent, $CV_N(t) = 0$. Dans ce cas de figure les poids sont équi-répartis. Ainsi, plus CV_N est proche de 0, mieux la répartition des poids est meilleure.

A.2.2 Nombre de particules efficaces

Il est défini par

$$\begin{aligned}\text{Neff}(t) :&= \left[\sum_{i=1}^{N}\left(\frac{\omega_t^{(i)}}{\sum_{j=1}^{N}\omega_t^{(j)}}\right)^2\right]^{-1}\\ &= \frac{N}{1+CV_N^2(t)}\end{aligned} \tag{A.21}$$

Cette mesure prend les valeurs entre 1 et N. Contrairement à CV_N, plus Neff est grand (proche de N) mieux les poids sont diversifiés.

A.2.3 Entropie de Shannon

Elle est définie par :

$$\text{Ent}(t) := -\sum_{i=1}^{N}\frac{\omega_t^{(i)}}{\sum_{j=1}^{N}\omega_t^{(j)}}\log_2\left(\frac{\omega_t^{(i)}}{\sum_{j=1}^{N}\omega_t^{(j)}}\right) \tag{A.22}$$

où $\log_2$ est le logarithme à base 2. Ces valeurs varient entre 0 (dégénérescence totale) et $\log_2(N)$ (équi-répartition des poids).

Remarque A.2.1. *En pratique, Neff est préféré car simple et moins coûteux à évaluer. De là, ou bien l'on rééchantillonne à tout instant. Ce qui a pour avantage d'éliminer toute dégénérescence de poids. Cependant cette approche induit un nouveau problème : absence de diversité des particules (une particule est dupliquée plusieurs fois). Ou bien l'on se fixe un seuil arbitraire (un pourcentage du nombre de particules, ex : 75% de N) en deçà duquel on suppose qu'il y'a dégénérescence au quel cas, on procède au rééchantillonnage. Ce seuillage a pour effet de favoriser la diversité des particules en limitant le nombre d'appels à la procédure de rééchantillonnage.*

A.3 Schémas de rééchantillonnage

Dans ces quelques notes, nous exposons quelques schémas de rééchantillonnage utilisés en MMC. principalement ceux qui conservent le nombre de particules rééchantillonnées fixe. On note par $N^{(i)}$ le nombre de descendants de la particule $\xi_{0:k}^{(i)}$ de sorte que $\sum_{i=1}^{N} N^{(i)} = N$. Rappelons que l'objectif du rééchantillonnage est d'éliminer le problème dû à la dégénérescence des particules. Le choix étant fait parmi la classe des solutions non biaisées dans le sens où $\mathbb{E}(N^{(i)}) = N\tilde{\omega}_k^{(i)}$, celles qui confèrent d'une part une complexité algorithmique linéaire en le nombre de particules. Et d'autre part, celles qui pourraient réduire la variance induite $\mathbb{V}(N^{(i)})$. Rappelons également que le rééchantillonnage systématique peut induire un appauvrissement de l'échantillon et qu'il convient de fournir un critère permettant de jauger du nombre particules efficaces

$$N_{eff}(t) := \frac{N}{1 + CV_N(t)^2}$$

et approchée par :

$$\hat{N}_{eff}(t) := \frac{1}{\sum_{i=1}^{N}(\omega_t^{(i)})^2}.$$

A.3.1 Redistribution Multinomiale

L'idée du rééchantillonnage multinomial est de tirer à l'instant k donné, N particules parmi $\left\{\xi_{0:k}^{(1)}, \xi_{0:k}^{(2)}, \ldots, \xi_{0:k}^{(N)}\right\}$ avec les probabilités $\left\{\tilde{\omega}_k^{(1)}, \tilde{\omega}_k^{(2)}, \ldots, \tilde{\omega}_k^{(N)}\right\}$. Le problème se résume alors au choix des indices des particules. Ce qui peut être fait par la méthode d'inversion. Les particules résultantes sont alors équi-pondérées de poids $\frac{1}{N}$. Le résume de cette procédure est donné ci-dessous. Du fait de la lenteur de cette procédure, une première amélioration peut être obtenue en triant au préalable les poids d'importance. Cependant la complexité qui était de $O(N\log(N))$ demeure inchangée. Le gain est observé au niveau du nombre de tests effectué qui diminue. Une deuxième amélioration est obtenue en ayant recours aux statistiques d'ordre en utilisant le résultat suivant.

Algorithme 21 Redistribution Multinomiale

1: **Entrée :** $\tilde{\omega}_k^{(1:N)}, \xi_k^{(1:N)}$
2: Pour $i = 1, 2, 3, \ldots N$ faire
— Générer $u \sim \mathcal{U}([0,1])$
— $j \leftarrow 1$
— Tant que $\tilde{\omega}_k^{(1)} + \tilde{\omega}_k^{(2)} + \cdots + \tilde{\omega}_k^{(j)} < u$ Faire
— $j \leftarrow j + 1$
— FinTant que
— $\tilde{\xi}_{0:k}^{(i)} = \xi_{0:k}^{(j)}$
3: FinPour
4: **Sortie :** $\tilde{\xi}_{0:k}^{(1:N)}$

Proposition A.3.1. *(Malmsquit [1950])*
Soient $U_1, U_2, \ldots, U_N$ *une suite de variables aléatoires i.i.d de distribution* $\mathcal{U}([0,1])$ *et* $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(N)}$ *les statistiques d'ordre associées. Alors*

$$U_N^{1/N}, U_N^{1/N} U_{N-1}^{1/(N-1)}, \ldots, U_N^{1/N} U_{N-1}^{1/(N-1)} \ldots U_1^{1/1}$$

a même distribution que $U_{(N)}, U_{(N-1)}, \ldots, U_{(1)}$.

Avec ce résultat, il est alors aisé de générer ces statistiques d'ordre. Par ailleurs, il clair que cette procédure est non biaisée eu égard à la définition précédente et sa variance est donnée par $\mathbb{V}(N^{(i)}) = N\tilde{\omega}_k^{(i)}(1-\tilde{\omega}_k^{(i)})$. De plus, sa complexité algorithmique est de $\mathrm{O}(N)$.

A.3.2 Redistribution Résiduelle

La procédure précédente fait beaucoup d'appels au générateur pseudo-aléatoire. Une façon de limiter ces appels est d'utiliser la redistribution résiduelle. L'idée est dans une première étape choisir de façon déterministe une partie des particules, puis dans la deuxième étape utiliser la redistribution multinomiale pour sélectionner les particules résiduelles. Elle procède donc dans un premier temps en affectant un nombre $\tilde{N}^{(i)}$ de descendants à la particule $\xi_{0:k}^{(i)}$ conformément à son poids d'importance $\tilde{\omega}_k^{(i)}$ avec $\tilde{N}^{(i)} := [N\tilde{\omega}_k^{(i)}]$. Puis, une redistribution multinomiale est faite sur les poids $\overline{\omega}_k^{(i)} = \frac{(N\tilde{\omega}_k^{(i)} - \tilde{N}^{(i)})}{\overline{N}_k}$ afin de choisir les $\overline{N}_k := N - \sum_{i=1}^N \tilde{N}^{(i)}$. La variance de cette procédure est donnée par $\mathbb{V}(N^{(i)}) = \overline{N}_k \overline{\omega}_k^{(i)}(1 - \overline{\omega}_k^{(i)})$ donc inférieure à celle de la multinomiale. Enfin, notons que cette approche trouve son importance lorsque le nombre de particules résiduel est faible. Au quel cas, il sera moins coûteux en terme de calcul.

A.3.3 Redistribution Systématique

la redistribution *systématique* connue aussi sous le nom de redistribution de variance minimale est la plus attrayante parmi la classe des procédures de rééchantillonnage non biaisées. On génère N points ordonnés suivants :

$$\frac{\tilde{u}}{N}, \frac{\tilde{u}+1}{N}, \frac{\tilde{u}+2}{N}, \ldots, \frac{\tilde{u}+N-1}{N}$$

avec $\tilde{u} \sim \mathcal{U}([0,1])$. Les particules sont alors rééchantillonnées au moyen de la redistribution multinomiale précédente avec les N points ordonnés comme poids. L'avantage manifeste est l'usage d'une seule uniforme. Cependant, il faut noter que les résultats de convergence du filtrage ne concernent que les redistributions multinomiale et résiduelle. Par ailleurs, on montre aussi que la variance de cette procédure est donnée par $\mathbb{V}(N^{(i)}) = \overline{N}_k \overline{\omega}_k^{(i)}(1 - \overline{N}_k \overline{\omega}_k^{(i)})$ et que sa complexité est de $O(N)$. Enfin, soulignons qu'il existe d'autres schémas de rééchantillonnage tels que la redistribution stratifiée, la redistribution de Baker, etc.

Bibliographie

M. Basseville. Divergence measures for statistical data processing—An annotated bibliography. *Signal Processing*, 93(4) :621–633, 2013.

T. Bengtsson, P. Bickel, and Bo Li. Curse-of-dimensionality revisited : Collapse of the particle filter in very large scale systems. *Probability and Statistics : Essays in Honor of David A. Freedman (Beachwood, Ohio, USA : Institute of Mathematical Statistics)*, 2 :316–334, 2008.

J. C. Bezdek and R. J. Hathaway. Some notes on Alternating Optimization. *Lecture notes in computer science*, 2003a. ISSN 0302-974.

J. C. Bezdek and R. J. Hathaway. Convergence of Alternating Optimization. *Neural, Parallel & Scientific Computations*, 11(4), 2003b.

R. A. Boyles. On the Convergence of EM algoritm. *Journal of the Royal Statistical Society, Series B*, 45 :47–50, 1983.

M. Briers, A. Doucet, and S. Maskell. Smoothing Algorithms for State-Space Models. *Annals of Institute of Statistical Mathematics*, 62(1) :61–89, 2010.

P. J. Brockwell and R. A. Davis. *Time Series : Theory and Methods.* Springer Series in Statistics, 1991.

F. Campillo. *Filtrage Particulaire et Modèles de Markov Cachés*, 2006. URL `ftp://ftp.irisa.fr/local/sigma2/campillo/cours/2006-master2-toulon.pdf`.

C. Cannamela, J. Garnier, and B. Iooss. Controlled stratification for quantile estimation. *Ann. Appl. Stat*, 2 :1554–1580, 2008.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models.* Springer, 2005.

O. Cappé, S. J. Godsill, and E. Moulines. An overwiew of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5) :899–924, 2007.

G. Casella and C. P. Robert. Post-processing accept-reject samples : recycling and rescaling. *J. Compt. Graph. Statist.*, 7(2) :139–157, 1998.

G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports*, 41(1-2) :119–134, 1992.

K. S. Chan and J. Ledolter. Monte Carlo EM Estimation for Times Series Models Involving Counts. *Journal of American Statistical Association*, 90(429) :242–252, 1995.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6) :2385–2411, 2004.

D. Crisan and A. Doucet. Convergence of sequential Monte Carlo methods. Technical report, 2000.

D. Crisan and A. Doucet. A Survey of Convergence Results on Particle Filtering Methods for Practitioners. *IEEE Transactions on Signal Processing*, 50(3), March 2002.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2 :299–318, 1967.

I. Csiszár. A class of measure of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4) :191–213, 1972.

K. Dahia. *Nouvelles méthodes en filtrage particulaire-Application au recalage de navigation inertielle par mesures altimétriques.* PhD thesis, LMC/IMAG, Université Joseph Fourier, Grenoble I., 2005.

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

P. Del Moral and A. Guionnet. On the stability of measure valued processes. Applications to nonlinear filtering and interacting particle systems. *Publication du laboratoire de Statistique et Probabilités 3-98, Université Paul Sabatier, Toulouse*, 1998.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 1977.

L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New-York, 1986.

R. Douc and E. Moulines. Limit theorems for weighted samples with applications to Sequential Monte Carlo Methods. *Ann. Statist.*, 36(5) :2344–2376, 2008.

R. Douc, A. Garivier, E. Moulines, and J. Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab*, 21(6) :2109–2145, 2011.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing : Fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

A. Doucet and B. B. Tadic. Parameter estimation in general state-space models using particle methods. *Annals of Institute of Statistical Mathematics*, 55(2) :409–422, 2003.

A. Doucet, S. J. Godsill, and C. Andrieu. On Sequential Monte Carlo Sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3) :197–208, 2000.

A. Doucet, N. de Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice : Statistics for Engineering and Information Science*. Springer-Verlag, New York, 2001.

C. Dubarry and S. Le Corff. Non-asymptotic deviation inequalities for smoothed additive functionals in non-linear state-space models with applications to parameter estimation. *Submitted*, 2010.

J. Durban and S. J. Koopman. Time series analysis of non-Gaussian observations based on state space models form both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Series B*, 62(1) :3–56, 2000.

L. Finesso. *Consistent estimation of the order for Markov and hidden Markov chains*. PhD thesis, University of Maryland, 1990.

R. A. Fisher and B Balmukand. The estimation linkage from the offspring of selfed heterozygotes. *Journal of Genetics*, 20 :79–92, 1928.

D. Fraser and J. Protter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 4 :387–390, 1969.

S. J. Godsill, A. Doucet, and M. West. Monte Carlo Smoothing for Nonlinear Time Series. *Journal of the American Statistical Association*, 99(465) :156–168, 2004.

G. Goertzel. Quota Sampling and Importance Function in Stochastic Solution of Particle Problem. Technical Report 434, Oak Ridge National Laboratory, jun 1949.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2) :107–113, 1993.

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

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

A. Harvey, E. Ruiz, and N. Shephard. Multivariate Stochastic Variance Models . *The Review of Economic Studies*, 61(2) :247–264, 1994.

J.H. Havrda and F. Chavat. Qualification methods of classification processes : Concept of structural α entropy. *Kybernetika*, 3(1) :30–35, 1967.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301) :13–30, 1963.

X. Hu and T. B. Schön. A General Convergence Result for Particle Filtering. *IEEE Transactions on Signal Processing*, 2011.

H. Kahn. Modifications of the Monte Carlo method. Technical report, Rand Corporation, Nov 1949.

H. Kahn and T. E. Harris. Estimation of Particle Transmission by Random Sampling. In *Monte Carlo Method*, Applied Mathematics Series, 1949.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering*, 82 :35–45, 1960.

R. E. Kalman and R. S. Bucy. New Results in Linear Filtering and Prediction Theory. *Transactions of the ASME-Journal of Basic Engineering*, 83 :95–107, 1961.

J. Kim and D. S. Stoffer. Fitting Stochastic Volatility Models in the Presence of Irregular Sampling Via Particle Methods and the EM Algorithm. *Journal of Time Series Analysis*, 29(5) :811–833, 2006.

S. Kim, N. Shephard, and S. Chib. Stochastic Volatility : Likelihood Inference and Comparison with ARCH Models . *The Review of Economic Studies*, 65(3) :361–393, jul 1998.

G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1) :1–25, 1996.

A. Kong, J. S. Liu, and W. H. Wong. Sequential Imputation and Bayesian missing Data Problems. *Journal of the American Statistical Association*, 89(425) :278–288, 1994.

N. Krichene. Modeling Stochastic Volatility with Application to Stok Returns. Technical Report No. 03/125, International Monetary Fund, 2003.

J. S. Liu. Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling. *Statistics and Computing*, 6(2) :113–119, 1996.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer : New York, 2001.

J. S. Liu and R. Chen. Blind Deconvolution via Sequential Imputations. *Journal of the American Statistical Association*, 90(430) :567–576, 1995.

S. Malmsquit. On a property of order statistics from a rectangular distribution . *Skand Akt.*, 33 :214–222, 1950.

Y. Matsuyama. The α-EM algorithm : Surrogate likelihood maximization using α-logarithmic information measures. *IEEE Transactions on Information Theory*, 49 (3) :692–706, 2003.

D. Q. Mayne. A solution of the smoothing problem for linear dynamic systems. *Automatica*, 4 :73–92, 1966.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics, 2008.

X.-L. Meng and D. B. Rubin. Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm. *Journal of American Statistical Association*, 86(416) : 899–909, Dec. 1991.

N. Metropolis and S. Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247) :335–341, 1949.

A. Millet. *Méthodes de Monte Carlo*. Université Paris 7, Paris 1, 2006. URL `http://www.proba.jussieu.fr/pageperso/millet/montecarlo.pdf`.

V. V. Petrov. *Limit Theorems of Probability Theory : Sequence of Independent Random Variables*. Number 4. Oxford Studies in Probability Series, 1995.

M. K. Pitt and N. Shephard. Filtering via Simulation : Auxiliary Particle Filters. *Journal of the American Statistical Association*, 94(446) :590–599, 1999.

L R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

L. R. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, englewood cliffs. edition, 1993.

A. Réyni. On measures of entropy and information . *Proc.* 4^{th} *Berkley Symposium on Math. Stat. and Prob.*, 10 :547–561, 1960.

B. D. Ripley. *Stochastic Simulation*. Wiley Series in Probability and Statistics, 1987.

C. P. Robert and G. Cassella. *Monte Carlo Statiscal Methods*. Springer, 2004.

D. B. Rubin. The Calculation of Posterior Distributions by Data Augmentation : Comment : A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fraction of missing information is modest : the SIR algorithm (discussion of Tanner and Wong). *J. Am. Statist. Assoc.*, 82(398) :543–546, Jun. 1987.

D. B. Rubin. Using the SIR algorithm to simulate posterior distribution. In *Bayesian Statistics 3*. Bernardo, J.M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds), 1988.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27 :379–423 and 623–656, 1948.

H.K. Van Dijk and T. Kloeck. Experiments with some alternatives for simple importance. In *Bayesian Statistics 2*. Bernardo, J.M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds), 1985.

R. Van Handel. On the minimal penalty for Markov order estimation. *Probability Theory and Related Fields*, 150(3-4) :709–738, August 2011.

G. C. G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411) :699–704, sep. 1990.

C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1) :95–103, 1983.

W. I. Zangwill. *Nonlinear Programming : a Unified Approach.* Prentice-Hall International Series in Management, Englewood Cliffs : N.J. Prentice-Hall, 1969.