Université Montpellier II
— Sciences et Techniques du Languedoc —

THÈSE

pour obtenir le grade de

Docteur de l'Université Montpellier II

Discipline : Mathématiques Appliquées

École Doctorale : Information Structures et Systèmes

Modèles à Facteurs Conditionnellement Hétéroscédastiques et à Structure Markovienne Cachée pour les Séries Financières

présentée et soutenue publiquement le 05 juillet 2006

par

Mohamed SAIDANE

Composition du jury

Président :	Jean-Noël Bacro	Université des Sciences, Montpellier II
Rapporteurs :	Jean-Pierre Florens	Université des Sciences Sociales, Toulouse I
	Christian Francq	Université Charles-de-Gaulle, Lille III
Examinateur :	Ali Gannoun	CNAM Paris
Directeur de Thèse :	Christian Lavergne	Université Paul Valéry, Montpellier III

Institut de Mathématiques et de Modélisation de Montpellier — UMR-CNRS 5149

Je dédie cette thèse à tous ceux que j'aime
à mes chers parents,
à mon frère et mes soeurs,
et à la grande famille SAIDANE.

Remerciements

Cette thèse est le fruit de travaux menés au sein des projets IS2 et MISTIS de l'INRIA Rhône-Alpes et de l'Institut de Mathématiques et de Modélisation de Montpellier. J'ai eu la chance d'y bénéficier d'un encadrement enrichissant et dynamique que j'ai longtemps cherché, et qui m'a permis de réaliser ce travail. À cette occasion, j'exprime ma profonde gratitude à Gilles Celeux et Florence Forbes pour m'avoir accueilli à l'INRIA et au Professeur Gilles Ducharme pour m'avoir accueilli à l'équipe de Probabilités & Statistique de l'I3M.

Le bon déroulement de cette thèse, jusqu'à son heureux dénouement, sont en grande partie imputables à mon directeur de thèse Christian Lavergne. Dans les périodes difficiles, il a su prendre du temps pour m'aider à avancer. Il m'a laissé aussi une grande liberté pour aborder ce travail. Ses conseils et son soutien ont été particulièrement précieux pour son accomplissement. Je le remercie donc très chaleureusement, aussi bien pour avoir dirigé mes travaux avec talent que pour m'avoir accompagné amicalement dans ce cheminement et même, à l'occasion, en dehors de mes activités professionnelles. Merci beaucoup Christian et j'espère que notre collaboration ne s'arrêtera pas avec cette thèse.

Je tiens à remercier également :

Le professeur Jean-Noël Bacro de l'Université Montpellier II, pour l'intérêt qu'il a porté à mes travaux et pour m'avoir fait l'honneur de bien vouloir présider le jury de cette thèse.

Les professeurs Jean-Pierre Florens de l'Université des Sciences Sociales - Toulouse I et Christian Francq de l'Université Charles-de-Gaulle - Lille III pour avoir accepté de rapporter cette thèse et pour avoir relu mes travaux avec une grande attention, leurs précieuses remarques m'ont permis de corriger et compléter mon manuscrit.

Le professeur Ali Gannoun du CNAM-Paris pour avoir accepté d'examiner cette thèse et pour m'avoir fait l'honneur de venir le jour de la soutenance et de faire partie du jury.

La période passée à l'INRIA m'a beaucoup apporté. J'ai rencontré des personnes remarquables, d'un point de vue personnel et professionnel. Merci à tous les membres de IS2 et MISTIS. Merci à Matthieu Vignes, Benjamin Esterni, Edwige Allain et Juliette Blanchet avec qui j'ai partagé le bureau D113 pendant mes deux premières années de thèse, et avec qui j'ai eu tant de discussions fructueuses. J'ai eu le plaisir de côtoyer, aussi, Henri Bertholon, Jean-Baptiste Durand, Grégory Noulin, Stéphane Girard, Paulo Gonçalvès, Myriam Garrido, Emilie Lebarbier, Franck Corset, Ollivier Taramasco, Guillaume Bouchard, Julien Jacques et Charles Bouveyron. Ces remerciements s'adressent aussi à mes voisins de l'INRIA, dont Claude Lemaréchal, Jérôme Malick, Aris Daniilidis, Hamoudi Kalla, Navneet Dalal, Bendehiba Bouksara, Chantal Baudin, Elodie Toihein et Françoise de Coninck.

Je remercie également tous les chercheurs, enseignants et membres du personnel de l'Institut de Mathématiques et de Modélisation de Montpellier, aussi bien que les membres du groupe de travail "Modèles Statistiques à Structures Cachées" pour leur amitié et leur aide pendant cette dernière année de thèse. Merci à mes collègues de bureau et désormais amis, Rémi Landri et Faiza Bessaoud, qui m'ont longuement soutenu et encouragé lors des moments difficiles et avec qui j'ai partagé les pires et les meilleurs moments de la thèse. Mes remerciements vont également à Catherine Trottier, Marie-José Martinez, Xavier Bry, Mohamed Mellouk, Ludovic Menneteau, Ahmad Younso, Florence Chaubert, Yann Guédon, Frederic Mortier, Gérard Biau, Alain Berlinet, Michel Nguiffo Boyom, Paolo Oliveira, Patrick Redont, Baptiste Chapuisat, Nicole Grachet, Mireille Piquet, Bernadette Lacan et Eric Hugounenq.

Tout au long de cette thèse, les moments de détente ont été aussi nombreux. Tous mes remerciements à mes chers amis Gérard Boudjema et Ikram Ben Amor avec qui j'ai eu l'occasion de découvrir les jolis coins de Grenoble, les merveilleux massifs de la Chartreuse et du Vercors et la station de Chamrousse. Merci beaucoup à Ikram pour les inoubliables soirées grenobloises au café de l'Olympia à la place de Notre Dame.

Enfin, merci profondément à tout ceux que j'aime, tous mes amis en Tunisie, sans qui certains moments m'auraient semblé bien plus difficiles : mes anciens camarades du Lycée Borj El-Baccouch de l'Ariana, mes amis de l'IHEC de Carthage et de l'ISG, et en particulier mon cher ami et collègue Mhamed-Ali Elaroui qui m'a beaucoup encouragé tout au long de ma thèse.

Table des matières

CHAPITRE 1

Introduction : La Théorie Factorielle en Finance

Ce premier chapitre a comme but de poser le cadre de ce travail et d'esquisser les directions générales dans lesquelles on a voulu orienter la recherche. A l'aide de références historiques et des exemples simples on montre quelles sont les limites des modèles actuels et quelles sont les développements possibles pour une meilleure approche statistique des donnés financières.

1.1 Notes Historiques

Le mathématicien français Louis Bachelier [1870,1946] est aujourd'hui considéré comme un précurseur de la théorie moderne des probabilités et comme le fondateur de la théorie économique des marchés financiers efficients. Dans sa thèse intitulée Théorie de la spéculation soutenue le 29 mars 1900, il a introduit la continuité dans les problèmes de probabilité en prenant le temps comme une variable. En particulier, il a élaboré une théorie mathématique du mouvement brownien cinq ans avant le grand physicien Albert Einstein et qui est aujourd'hui à la base de la plupart des modèles de prix en finance, notamment la formule de Black-Scholes [1973]. C'était donc la première fois qu'on consacre un travail académique en Mathématiques pour expliquer le comportement des marchés boursiers. Bachelier était un scientifique à part entière et il regardait l'évolution des actions de la même façon que le comportement des particules dans l'espace après des chocs aléatoires. Ce dernier mot est important car il désigne une notion centrale en probabilités et en statistique et représente la matière première de ces sciences. L'intuition de Bachelier était qu'il est impossible de prédire le prix futur des actifs financiers. "L'espérance mathématique d'un spéculateur est nulle car il a autant de chances de gagner que de perdre car le marché est un jeu juste" écrivait-t-il dans sa thèse.

Si Bachelier a le mérite d'avoir introduit la finance comme sujet de recherche pour les mathématiques, il faut attendre un demi siècle pour assister à une nouvelle rencontre marquante des deux sciences. C'est en 1952 qu'un jeune étudiant doctorant, Harry

Markowitz, publiait un petit article de quatorze pages qui allait révolutionner la finance. Sous le nom "Sélection de portefeuille" il a été publié dans le seul journal de spécialité à l'époque, le désormais fameux Journal of Finance. Cette fois-ci c'est un économiste qui utilise les outils statistiques simples comme la moyenne et la variance pour formaliser les notions de rendement espéré et le risque des actions. La notion de rendement d'une action est centrale en finance. Markowitz a pu montrer que les investisseurs ont intérêt à investir dans plusieurs actions au lieu d'une seule car ainsi ils réduisent le risque de leur portefeuille. C'est le principe de la diversification. Il a montré aussi qu'avec les mêmes actions on peut construire des portefeuilles avec le même risque mais avec des rendements espérés différents. Il a introduit la notion d'efficience. Un portefeuille est dit efficient si parmi tous les portefeuilles avec le même rendement espéré il est celui avec le risque minimum. On peut figurer tous ces portefeuilles sur un graphique moyenne-variance et la courbe qui contient tous les portefeuilles efficients porte le nom de frontière efficiente.

Parmi tous les portefeuilles construits à partir des actions d'un marché quelconque, il y en a un qui reçoit beaucoup d'attention de la part des investisseurs : c'est l'indice boursier. Ce portefeuille est théorique et contient toutes les actions émises par les entreprises qui entrent dans la composition de l'indice. En dépit du fait qu'il est virtuel, on peut calculer sa valeur et donc suivre son évolution au cours du temps. C'est le cas de beaucoup de gestionnaires qui doivent comparer les performances de leur portefeuille avec celle de l'indice. En général il est très difficile d'obtenir une performance supérieure à ce portefeuille très diversifié sans prendre des risques importants. Le choix des titres qui font partie de l'indice se base sur des critères comme la liquidité des titres ou la taille des entreprises et de temps en temps, la composition de l'indice change. Son rendement est calculé comme étant une moyenne pondérée des rendements des titres qui le composent avec des poids proportionnels à la capitalisation boursière de chaque entreprise. Étant donné qu'il contient en général tous les titres cotés sur le marché qu'il représente, on l'associe au marché lui même. Cette dernière notion est beaucoup utilisée dans le langage boursier. On dit souvent "aujourd'hui le marché monte" ou "le marché est déprimé en ce moment" mais il n'est pas facile de définir ce qu'est le marché et encore plus difficile de l'observer car il n'est pas quantifiable. Par définition il contient tous les biens échangeables et non échangeables comme par exemple les maisons ou le capital humain.

1.2 Les Modèles d'Évaluation des Actifs Financiers

Le marché est une notion centrale en finance et donc beaucoup de modèles l'intègrent mais à sa place ils prennent l'indice comme substitut. Pour illustrer ces dernières affirmations, considérons le CAPM (Capital Asset Pricing Model) un fameux modèle développé par Sharpe [1963, 1964] et Treynor [1961] dans les années soixante en utilisant des notions statistiques simples comme la moyenne et la variance introduites par Markowitz dans l'étude de la finance.

1.2.1 Le CAPM

Sans entrer dans les détails concernant les arguments économiques qui se trouvent derrière ce modèle on peut donner la relation simple qui relie le rendement espéré d'une action individuelle et le rendement espéré du marché :

$$\mathbb{E}(R_i - R_f) = \beta_i \mathbb{E}(R_M - R_f)$$

où R_i est le rendement de l'action i, R_M est le rendement du marché et R_f est le taux sans risque qu'on peut obtenir en plaçant l'argent sur un compte d'épargne. Si l'économie est en équilibre on peut montrer que le coefficient β_i est donné par l'expression suivante :

$$\beta_i = \frac{Cov(R_i, R_M)}{Var(R_M)}$$

On associe β_i au risque que l'action prend par rapport au marché. Un portefeuille qui se comporte de manière identique au marché aura un β_i égal à l'unité alors qu'un portefeuille avec un rendement constant aura un β_i nul.

L'estimation du β_i a aussi une interprétation statistique car elle est égale à l'estimation de la pente de la droite de régression linéaire de R_i sur R_M par la méthode des moindres carrés. On peut remarquer que R_M est commun à toutes les actions. Ce qui est spécifique à chacune est le β_i qui représente leur sensibilité par rapport à l'évolution du marché qui peut être vu comme un facteur commun influençant toutes les actions. Prédire son évolution n'est pas chose facile comme l'a montré Bachelier il y a plus de cent ans mais si on a une bonne connaissance des β_i on peut avoir une idée du comportement relatif espéré de deux actions en particulier. Autrement dit, si deux actions ont des betas de signe contraire et si le marché enregistre une forte hausse il est probable que l'action avec le beta positif enregistrera une hausse de son prix alors que celle avec un beta négatif enregistrera une baisse de son prix. Le modèle peut être généralisé pour prendre en compte plusieurs facteurs et arriver à une formule similaire avec celle du CAPM. Ross [1976] a formalisé cette idée en introduisant en 1976 le modèle APT (Arbitrage Pricing Theory).

En statistique il existe un modèle similaire appelé l'analyse factorielle. Il est utilisé pour modéliser la dépendance linéaire d'un grand nombre de variables par rapport à un petit nombre de facteurs. La différence avec l'APT est que les facteurs ne sont pas observés et doivent être estimés par le modèle. Cela peut avoir un intérêt dans le cas du CAPM car on sait qu'à la place du marché on utilise un indice qui est une moyenne pondérée. Dans la littérature on utilise souvent les indices boursiers Isakov [1999] mais également des indices équi-pondérés comme par exemple Fama et MacBeth [1973] dans leur article qui fait référence pour tous les travaux qui visent a tester le CAPM.

1.2.2 Critique de Roll et CAPM conditionnel

Il faut remarquer que la formule du CAPM contient des moments qui doivent être estimés à partir des réalisations passées des rendements et pour cela il n'y a pas une méthode unique d'estimation. Cette formulation du modèle est dite ex-ante. Il y a une autre formulation du CAPM qui utilise les valeurs observées des rendements à la place de leur moments et est dite la formulation ex-post. Cette transformation est possible si on fait l'hypothèse qu'on est en présence d'un jeu juste ("fair game" en anglais) à savoir la réalisation du rendement de n'importe quelle action est en moyenne égale à la valeur espérée par l'investisseur. On peut écrire cela sous la forme de l'égalité suivante :

$$\begin{aligned} R_{i,t} &= \mathbb{E}(R_{i,t}) + \beta_i \left[R_{M,t} - \mathbb{E}(R_{M,t})\right] + \varepsilon_{i,t} \\ &= R_{f,t} + \beta_i(R_{M,t} - R_{f,t}) + \varepsilon_{i,t} \end{aligned}$$

En soustrayant $R_{f,t}$ des deux cotées on obtient la forme ex-post du CAPM :

$$(R_{i,t} - R_{f,t}) = \beta_i(R_{M,t} - R_{f,t}) + \varepsilon_{i,t}$$

Le CAPM a eu beaucoup de succès à cause de sa forme simple et facile à interpréter. Il est enseigné dans tous les cours de finance mais les critiques qui ont été émises à son égard sont nombreuses. On a plusieurs fois annoncé la mort du CAPM après avoir fait des testes statistiques qui l'ont rejeté. La validité de ses critiques a été mise en doute par Roll en 1977. Il a critiqué les calculs effectués pour tester la validité du modèle car ils prenaient en compte un indice pour représenter le marché. Roll montre que l'indice n'est pas un portefeuille efficient et donc on ne peut pas le substituer au marché. Pour cette raison les calculs qui mettent en cause le CAPM sont en fait la preuve que le choix de l'indice n'est pas valide. Par la suite on proposera un modèle statistique qui prendra en compte cette critique et ne supposera plus que le marché est observable mais l'estimera en même temps que les β_i.

La version du CAPM présentée jusqu'ici porte le nom de CAPM inconditionnel. Elle suppose que les β_i sont constants au cours du temps et sont calculés à l'aide des moyennes, variances et covariances obtenues sur la base des données historiques. Cela suppose que les distributions des rendements sont stables dans le temps. Il y a une version du CAPM qui porte le nom de CAPM conditionnel qui suppose que les agents prennent leur décisions en tenant compte de l'information disponible au debut de chaque période d'investissement. Dans ce cas les β_i changent à chaque période t en fonction de $\mathcal{D}_{t-1}$, l'information disponible. En voici deux exemples.

Le Modèle de de Bollerslev, Engle, et Wooldrigde [1988]

Soit $\mathbf{y}_t$, le vecteur des rendements excédentaires réels de tous les actifs du marché, mesuré comme étant le rendement nominal de la période t moins le taux de rendement nominal sur un actif sans risque. Soit μ_t et $\mathbf{\Sigma}_t$, le vecteur de moyenne conditionnelle et la matrice des covariances conditionnelles de ces rendements, étant donnée l'information disponible à la période $t-1$. Soit ω_{t-1}, le vecteur de poids (market weights) à la fin de la période précédente tel que le rendement excédentaire sur le marché est défini de la façon suivante :

$$\mathbf{y}_M = \mathbf{y}_t'\omega_{t-1}$$

II s'ensuit que le vecteur des covariances avec le marché est $\boldsymbol{\Sigma}_t\omega_{t-1}$. En utilisant la formulation de Jensen [1972], Bollerslev, Engle, et Wooldrigde [1988] obtiennent un CAPM de la forme :

$$\mu_t = \delta\boldsymbol{\Sigma}_t\omega_{t-1}$$

où ω_{t-1} est le vecteur des poids à la fin de la période $t-1$. δ est un scalaire qui correspond au coefficient d'aversion relative au risque. Bollerslev, Engle, et Wooldrigde supposent δ constant sur toute la période. Ils soutiennent également l'hypothèse d'un même coefficient pour tous les actifs. La variance conditionnelle du rendement excédentaire du marché est égale à :

$$\sigma^2_{M,t} = \omega'_{t-1}\boldsymbol{\Sigma}_t\omega_{t-1}$$

et la moyenne conditionnelle est égale à :

$$\mu_{M,t} = \omega'_{t-1}\mu_t$$

On peut réécrire cette expression comme étant égale à :

$$\mu_{M,t} = \delta\sigma^2_{M,t}$$

de telle sorte que δ est considéré comme étant la pente du "trade-off" du marché entre la moyenne et la variance.

En utilisant la définition usuelle du bêta d'un actif, la covariance entre cet actif et le marché divisé par la variance du portefeuille du marché,

$$\beta_t = \frac{\boldsymbol{\Sigma}_t\omega_{t-1}}{\sigma^2_{M,t}}$$

et en substituant dans l'équation du CAPM et de la relation entre la variance et la moyenne conditionnelle du marché on obtient l'expression familière :

$$\mu_t = \beta_t\mu_{M,t}$$

Cela implique que puisque la matrice des covariances des rendements varie dans le temps, les rendements moyens et les bêtas seront également variables dans le temps. Le système d'équations estimé par Bollerslev et al., est le suivant :

$$\begin{cases} y_{it} = b_i + \delta\sum\limits_j \omega_{jt}h_{ijt} + \varepsilon_{it} \\ h_{ijt} = \gamma_{ij} + \alpha_{ijt}\varepsilon_{it-1}\varepsilon_{jt-1} + \beta_{ij}h_{ijt-1} \\ \varepsilon_t|\mathcal{D}_{t-1} \sim \mathcal{N}(0, \Omega_t) \end{cases}$$

Le modèle de Ng [1991]

Examinons maintenant un modèle d'évaluation d'actifs pour lequel le CAPM de Sharpe-Lintner et le zéro-beta CAPM sont des cas spéciaux. Le modèle de Ng [l991] constitue une alternative beaucoup plus riche que les modèles développés par Bollerslev et al. [1988]. Nous reprenons ici la dérivation de son modèle.

Soit $\{R_{it},\ i = 1, ..., q\}$, le taux de rendement sur les actifs risqués de la période $t-1$ à t ; $R_{M,t}$ est le rendement sur le portefeuille de marché. Soit $R_{z,t}$, le taux de rendement sur un portefeuille à beta-zéro à variance minimum, lequel est non corrélé avec R_{it} et $R_{M,t}$ $\forall i$. Nous pouvons exprimer la relation d'équilibre entre les taux de rendement anticipés d'actif risqué et du portefeuille à beta-zéro de la façon suivante :

$$\mathbb{E}\left[(R_{it} - R_{z,t})\,|\mathcal{D}_{t-1}\right] = \lambda_{ot} Cov\left(R_{M,t}, R_{it}|\mathcal{D}_{t-1}\right) \tag{1.1}$$

où λ_{ot} est un scalaire relié à l'aversion relative au risque de l'économie. L'opérateur d'espérance, la covariance des rendements avec le marché et λ_{ot} sont conditionnels à l'ensemble d'information $\mathcal{D}_{t-1}$, disponible dans le marché à la période $t-1$. À partir de la relation (1.l), il s'ensuit que

$$\mathbb{E}\left[(R_{M,t} - R_{z,t})\,|\mathcal{D}_{t-1}\right] = \lambda_{ot} Var\left(R_{M,t}|\mathcal{D}_{t-1}\right) \tag{1.2}$$

le rendement anticipé en excédent du taux beta-zéro est proportionel a λ_{ot}. Les modèles (1.1) et (1.2) supposent qu'il n'existe pas de taux sans risque. En faisant l'hypothèse qu'un taux sans risque est présent dans l'économie. en remplaçant $R_{z,t}$ par R_{ft} on obtient le résultat bien connu du CAPM de Sharpe et Lintner et cela va dans le sens de l'analyse des primes de risque de Merton [1980] suivant l'hypothèse d'aversion au risque constant. Suivant l'hypothèse que la variance du changement de la richesse de l'investisseur est beaucoup plus grande que la variance du changement dans les variables d'état, Merton dérive l'équation (1.2), où $\lambda_{ot} = \lambda_o$, est une constante. Il interprète λ_o comme la mesure d'aversion au risque d'Arrow-Pratt. Par conséquent, la prime de risque est strictement positive lorsque la fonction d'utilité des investisseurs est croissante et strictement concave.

En combinant (1.1) et (1.2) et en reformulant le tout, on obtient :

$$\mathbb{E}\left[(R_{it} - R_{z,t})\,|\mathcal{D}_{t-1}\right] = \frac{\left[\delta_o + \lambda_o Var\left(R_{M,t}|\mathcal{D}_{t-1}\right)\right] Cov\left(R_{M,t}, R_{it}|\mathcal{D}_{t-1}\right)}{Var\left(R_{M,t}|\mathcal{D}_{t-1}\right)} \tag{1.3}$$

Suivant l'hypothèse que λ_{ot} est stable dans le temps et que le rendement anticipé en excédent du portefeuille beta-zéro de la relation (1.2) est linéaire dans sa variance, la relation (1.3) peut être interprétée comme une variante du CAPM à beta-zéro où le paramètre constant δ_o représente les coûts de transactions relevant du différentiel entre les taux prêteurs et les taux emprunteurs. On peut obtenir un δ_o différent de zéro en raison des dividendes ou des coûts de transactions qui ne sont pas explicitement inclus dans le modèle.

On fait l'hypothèse que le rendement espéré sur le portefeuille beta-zéro a variance minimale est constant dans le temps et que le taux sans risque est observable au temps $t-1$. En reformulant (1.3) en notation matricielle, on obtient :

$$\mathbb{E}\left[r_t|\mathcal{D}_{t-1}\right] = \alpha_o\mathbf{I} + \left(\delta_o + \lambda_o\omega'_{t-1}\Omega_t\omega_{t-1}\right)\left(\omega'_{t-1}\Omega_t\omega_{t-1}\right)^{-1}\Omega_t\omega_{t-1} \tag{1.4}$$

où α_o est un scalaire représentant la prime de risque anticipée d'un portefeuille beta-zéro.

En faisant l'hypothèse que les rendements excédentaires réalisés sont les prévisions non-biaisées des investisseurs, on peut reformuler la relation (1.4) :

$$r_t = \alpha_o\mathbf{I} + \left(\delta_o + \lambda_o\omega'_{t-1}\Omega_t\omega_{t-1}\right)\left(\omega'_{t-1}\Omega_t\omega_{t-1}\right)^{-1}\Omega_t\omega_{t-1} + \varepsilon_t \tag{1.5}$$

$$(\varepsilon_t|\mathcal{D}_{t-1}) \sim \mathcal{N}(0, \Omega_t)$$

où ε_t est le vecteur des différences entre les rendements excédentaires réalisés et les rendements excédentaires espérés.

On peut reformuler le système en (1.5) :

$$r_t = \alpha + \left(\delta + \lambda\omega'_{t-1}\Omega_t\omega_{t-1}\right)\left(\omega'_{t-1}\Omega_t\omega_{t-1}\right)^{-1}\Omega_t\omega_{t-1} + \varepsilon_t \tag{1.6}$$

$$(\varepsilon_t|\mathcal{D}_{t-1}) \sim \mathcal{N}(0, \Omega_t)$$

La nouvelle équation permet aux paramètres α, δ et λ de varier selon les actifs, mais d'être constants sur la période d'estimation (Brown et Weinstein [1983]). Cette approche complémente celle de Bollerslev et al. [1988] qui font I'hypothèse que le coefficient de pente du risque de covariance est constant et que $\delta = 0$. De plus, ils contraignent λ à être le même pour tous les actifs. Il est utile d'utiliser cette paramétrisation alternative, puisqu'à l'équilibre, "le trade-off" entre le rendement excédentaire conditionnel et la variance conditionnelle du marché est probablement différent entre les différents marchés boursiers. La constance des paramètres α, δ et λ de chaque marché sur la période d'estimation correspond à l'hypothèse selon laquelle les goûts et préférences des consommateurs demeurent les mêmes durant la période d'estimation. Étant donnée la courte période considérée, cette hypothèse ne semble pas farfelue.

Le système d'équations estimé par Ng [1991] est le suivant :

$$\left\{\begin{array}{c} r_t = \alpha + \left(\delta + \lambda\omega'_{t-1}\Omega_t\omega_{t-1}\right)\left(\omega'_{t-1}\Omega_t\omega_{t-1}\right)^{-1}\Omega_t\omega_{t-1} + \varepsilon_t \\ h_{ijt} = \gamma_{ij} + \alpha_{ij}\varepsilon_{it-1}\varepsilon_{jt-1} + \beta_{ij}h_{ijt-1}, \quad i,j = 1,...,q \\ (\varepsilon_t|\mathcal{D}_{t-1}) \sim \mathcal{N}(0, \Omega_t) \end{array}\right.$$

Cette version dynamique du modèle semble plus réaliste. Le modèle statistique qu'on proposera plus loin, prendra en compte ce point de vue mais en proposant une modélisation différente. Les rendements seront vus comme des combinaisons linéaires de facteurs conditionnellement hétéroscédastiques plus un terme idéosyncratique.

1.2.3 Les Modèles à Facteurs

La notion de "modèle à facteur" (où modèle à index ou à coefficients bêtas,...) est ancienne en Finance. Ces modèles sont issus à la fois des théories d'évaluation des actifs financiers et de l'analyse des séries temporelles. Ces deux courants de la littérature font appel à deux notions différentes, qui sont toutes deux utiles pour réduire la dimension du modèle statistique. Dans ce type de modèles on montre que c'est la covariance avec certaines variables directrices qui explique la différence entre les rendements espérés et qui s'interprète comme la quantité de risque rémunéré. Ces variables directrices sont souvent appelées facteurs en tant que variables explicatives des rendements. La réduction de dimension s'opère donc en coupe transversale, grâce à une hypothèse d'indépendance conditionnelle entre les rendements d'un grand nombre d'actifs financiers étant donné un petit nombre de facteurs comme dans l'analyse factorielle standard.

En général, ces modèles supposent que le rendement d'un actif financier y_{it} ($i \in \{1, ..., q\}$ et $t \in \{1, ..., n\}$) peut être exprimé comme une somme d'une partie **anticipée** et une partie **non anticipée**. La partie non anticipée du rendement peut être aussi exprimée comme une somme de deux composantes : une composante **systématique** qui ne peut pas être diversifiée et une composante **non systématique** spécifique à l'actif en question. La partie systématique et non anticipée du rendement est supposée suivre une structure à facteurs. Le modèle général avec k facteurs et q actifs peut être écrit dans sa version standard sous la forme suivante :

Structure du Modèle Standard

Pour $\quad i = 1, ..., q \quad$ et $\quad t = 1, ..., n$

$$y_{it} = \overbrace{\underbrace{\mathbb{E}(y_{it})}_{\theta_i}}^{\text{partie anticipée}} + \overbrace{\underbrace{\sum_{j=1}^{k} x_{ij} f_{jt}}_{\text{partie systématique}} + \underbrace{\varepsilon_{it}}_{\text{partie non systématique}}}^{\text{partie non anticipée}}$$

Les f_{jt} sont des variables aléatoires non observables et indépendantes appelées **facteurs communs**, les x_{ij} qui leurs sont associés sont les **pondérations** et les ε_{it} sont aussi des variables aléatoires non observables et indépendantes appelées facteurs **spécifiques**. Afin de réduire la dimension du modèle statistique et de simplifier le calcul de la matrice de covariance des rendements, dans une structure moyenne-variance de sélection de portefeuilles, le nombre de facteurs k doit être beaucoup plus petit que le nombre d'actifs q. Cette méthode tente donc de représenter les variables étudiées dans un cadre linéaire, en fonction d'un certain nombre assez réduit de variables aléatoires non observables appelées facteurs communs. Ces facteurs détiennent une part importante de l'information sur les caractéristiques communes des variables initiales aussi bien que sur les relations complexes qui existent entre elles. Le modèle suppose alors que toutes les corrélations sont expliquées par les facteurs communs et que la variation résiduelle provient d'une source de variables spécifiques non corrélées appelées facteurs spécifiques, uniques ou idiosyncratiques.

Dans la littérature financière, différentes méthodes ont été considérées pour l'identification des facteurs. Certaines approches ont utilisé des facteurs spécifiés par avance en se basant sur des données macro-économiques telles que le taux d'inflation, le taux d'intérêt,... (King, Sentana et Wadhwani [1994]) D'autres ont utilisé des combinaisons linéaires des séries observées (technique d'analyse en composantes principales, voir par exemple Ng, Engle et Rothschild [1992] et Kaiser [1997]).

1.3 Incertitude, Risque et Volatilité

Conformément à la logique des modèles d'évaluation des actifs financiers, comme l'APT ou le CAPM, la volatilité joue un rôle essentiel dans la détermination du rendement. En particulier, un actif plus risqué étant supposé offrir un rendement supérieur à celui de l'actif sans risque. Aujourd'hui on admet que ces rendements sont des séries qui présentent des comportements de type hétéroscédastique avec très souvent de la persistance. Cet effet a été mesuré notamment à travers des modèles dans lesquels la volatilité est directement introduite dans l'équation de l'espérance conditionnelle comme variable explicative du rendement (Engle et al. [1987] ou French et al. [1987]). Certains auteurs (par exemple, Schwert [1990]) ont montré que, inversement, le rendement peut intervenir dans l'explication de la volatilité. Il s'agit alors d'effet d'asymétrie (ou d'effets de levier), car la réaction de la volatilité à un choc sur le rendement est différente selon le signe du choc : on observe généralement qu'un choc à la baisse sur le rendement accroît beaucoup plus la volatilité, toutes choses égales par ailleurs, qu'un choc à la hausse. Ces différentes interactions semblent relativement robustes pour rendre compte de la dynamique de la plupart des prix des actifs financiers.

1.3.1 Des Perceptions du Risque Différentes

La diversité des acteurs financiers (des théoriciens aux praticiens) préoccupés par le concept de volatilité explique la diversité des approches pour traiter ce concept et les débats qui peuvent en résulter. Sur les marchés, chacun a sa propre perception du risque (fonction d'aversion envers le risque). Toute la difficulté réside dans la réconciliation entre les concepts théoriques du risque et son estimation par les investisseurs qui ont adopté la notion de volatilité. La typologie proposée par Granger [2002] permet de distinguer plusieurs acteurs.

- Les mathématiciens qui s'intéressent à la théorie d'évaluation des options, avec une approche en temps continu. La nécessité d'intégrer une prévision de volatilité des cours pour obtenir le prix d'une option a elle-même conduit à une modélisation approfondie de cette prévision avec une mise en évidence de caractéristiques, telles que, par exemple, celle d'une structure par terme décroissante de la volatilité. La volatilité étant la seule variable non observable dans le prix d'une option, il est équivalent de raisonner sur celle-ci directement ou sur les prix. C'est ainsi que, à partir du prix des options cotées, est calculée une volatilité dite implicite qui correspond à la volatilité moyenne anticipée par les intervenants de marchés.

- Les économètres et les statisticiens empiriques. Les modèles (ARCH, GARCH, etc.) ont permis de souligner les phénomènes d'hétérocédasticité et de persistance de la volatilité. Ces approches ont également mis en évidence les limites du postulat d'une distribution normale des rendements – et donc celles de la volatilité historique – pour l'évaluation des risques de marché. Leur démarche a également permis, notamment, de mettre en évidence les phénomènes de retour à la moyenne de la volatilité.

- Les économistes de la théorie de l'incertain. Ils travaillent sur la théorie du portefeuille, les effets bénéfiques de la diversification, via la distinction entre risque spécifique et risque systématique, et le CAPM, modèle dans lequel la volatilité joue un rôle essentiel dans la détermination du rendement.

- Les gérants d'OPCVM et les traders (les "professionnels"). Leur objectif est de maximiser le rendement de leurs transactions (certes avec un horizon différent). Pour eux, la prévisibilité des cours dépend de la volatilité, voire de la volatilité de la volatilité... Leur comportement est lui-même parfois accusé d'être un facteur explicatif de la volatilité. Quant aux fonds spéculatifs (hedge funds), la diversité de leurs stratégies ne permet pas de conclure sur l'incidence éventuelle de leur comportement sur la volatilité : il n'en demeure pas moins généralement admis que, par leurs opérations d'arbitrage, ils concourent à l'efficience des marchés, et que, par leurs transactions, ils contribuent à la liquidité des marchés.

- Les investisseurs individuels. Typiquement, ils sont préoccupés par la chute des cours, a fortiori lorsque leur retraite repose sur un système de capitalisation. Ils le sont également plus par la volatilité individuelle des titres que par celle des indices boursiers, laquelle va être d'autant plus faible que la corrélation entre les titres est réduite.

À cette classification, il convient d'ajouter les autorités prudentielles et les banques centrales préoccupées par les conséquences potentielles d'une hausse de la volatilité sur le risque systémique et la stabilité financière.

1.3.2 Les Modèles d'Hétéroscédasticité Dynamique

Il a été montré depuis longtemps que la volatilité conditionnelle des rendements est, au moins partiellement, prévisible. On observe en particulier que des variations importantes des prix (positives ou négatives) sont généralement suivies de variations importantes des rendements, indiquant une hétéroscédasticité dans la volatilité de ces rendements. Au cours de ces dernières années une littérature abondante a été consacrée au mode de formation de la volatilité financière et plusieurs approches ont été proposées pour décrire sa dynamique à travers le temps. Toutefois ce sont des spécifications de type "Autoregressive Conditional Heteroscedasticity" qui sont généralement utilisées pour décrire cette évolution. Le Prix Nobel 2003 a donc plus particulièrement récompensé le Professeur Engle pour ses méthodes d'analyse des séries temporelles à volatilité non constante. Il a en effet révolutionné l'économétrie en proposant une classe de modèles permettant de prévoir correctement le comportement de ce type de séries : les modèles ARCH. Ces modèles permettent d'effectuer la prévision de variables économiques dont la volatilité varie au cours du temps. Ils sont donc particulièrement utiles en finance car les cours boursiers se caractérisent par une variabilité pouvant être très instable. Le succès du modèle ARCH fut consacré en 1982 par la publication dans la revue Econometrica d'un article où Robert Engle étudiait l'inflation du Royaume-Uni à l'aide de ce nouvel outil. Cet article eut un succès retentissant et Bollerslev, l'un des étudiants en thèse de Robert Engle, proposa ensuite en 1986 une version généralisée de ce modèle, qui fut baptisée modèle GARCH (autorégressif conditionnellement hétéroscédastique généralisé). Ces modèles connurent ensuite de très nombreuses extensions dans les années quatre-vingt-dix, sous les acronymes les plus divers. On peut citer, sans être exhaustif, les modèles EGARCH (Exponential GARCH), TGARCH (Threshold GARCH), GQARCH (Quadratic GARCH), ARCH-M (ARCH in Mean), FIGARCH (Fractionally Integrated GARCH). Ces travaux sont, encore aujourd'hui, à la base de très nombreuses recherches en économie, en économétrie et en finance. Robert Engle continue lui-même à explorer de nouvelles voies extrêmement prometteuses, notamment pour l'analyse du risque et l'étude de la microstructure des marchés financiers.

Ces modèles sont définis par deux équations : une équation de moyenne qui décrit l'évolution de la variable dépendante en fonction d'une ou de plusieurs variables indépendantes, et une équation qui décrit la nature de la variabilité temporelle de la variance conditionnelle ou de l'hétéroscédasticité. Ces équations sont données par :

$$y_t = \gamma_0 + \sum_i \gamma_i x_i + u_t$$

le terme d'erreur u_t a une moyenne nulle et une variance h_t^2 variable à travers le temps :

$$h_t^2 = w + \sum_j \alpha_j z_j$$

Dans les équations ci-dessus, y_t représente la série qu'on cherche à modéliser, x_i les variables explicatives de l'équation de moyenne (qui peuvent être des variables exogènes ou bien des valeurs retardées de y ou bien aussi des valeurs actuelles et/ou retardées de la spécification hétéroscédastique), et les z_j sont les variables explicatives de la spécification hétéroscédastique. Étant donné que la variance du terme d'erreur u_t évolue au cours du temps, ce dernier peut donc être exprimé sous la forme suivante :

$$u_t = h_t v_t \quad \text{ou bien} \quad v_t = \frac{u_t}{h_t}$$

où v_t est l'erreur standardisée vérifiant la propriété : $v_t \sim iid(0,1)$.

ARCH

Nous considérons ici la forme particulière de volatilité conditionnelle proposée par Engle [1982]. Dans sa version la plus simple le modèle ARCH(q) suppose que la variance des résidus u_t de l'équation de moyenne évolue selon le processus :

$$Var(u_t/\mathcal{D}_{1:t-1}) = h_t^2 = w + \sum_{i=1}^{q} \alpha_i u_{t-i}^2$$

où $\mathcal{D}_{1:t} = \{u_{t-s},\, s \geq 0\}$. Dans cette spécification la variance conditionnelle, h_t^2, est exprimée comme étant une fonction des q valeurs retardées des carrés des résidus de l'équation de moyenne.

L'application de ces modèles sur des données réelles nécessite souvent la prise en compte d'un grand nombre de retards q. Généralement pour éviter le problème de négativité de la variance conditionnelle, nous utilisons souvent une structure de retards artificielle fixe avec des pondérations décroissantes dans le temps.

GARCH

Le modèle GARCH proposé par Bollerslev [1986] permet de résoudre ce problème en introduisant directement des valeurs retardées de la variance conditionnelle dans la spécification de la volatilité conditionnelle. Cette nouvelle spécification conduit à une représentation GARCH(p,q) pour la variance conditionnelle de u_t :

$$Var(u_t/\mathcal{D}_{1:t-1}) = h_t^2 = w + \sum_{i=1}^{q} \alpha_i u_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}^2$$

Dans ce cas, h_t^2, est une fonction des q valeurs retardées des carrés des résidus et des p valeurs retardées de la variance conditionnelle. Cette spécification nécessite souvent moins de paramètres et permet d'un meilleur ajustement. Elle ne nécessite donc pas la structure artificielle proposé par Engle. La forme la plus utilisée est celle d'une spécification GARCH(1,1) :

$$h_t^2 = w + \alpha\, u_{t-1}^2 + \beta\, h_{t-1}^2$$

Bien que ces modèles fournissent des prévisions de futures périodes, il est à remarquer que la formulation ARCH repose sur des hypothèses qui peuvent s'écarter plus ou moins des situations réelles. En effet, les modèles ARCH et GARCH sont tout à fait symétriques, c'est-à-dire que les effets des chocs ne sont pas différenciés selon leurs signes, et pourtant l'asymétrie représente une hypothèse très réaliste pour des séries monétaires ou financières. Ce problème a donc préoccupé les économistes et cela a conduit à une fructueuse littérature où la famille des modèles GARCH a été alimentée par de nombreux modèles asymétriques dérivés des GARCH dans une tentative de résoudre le problème d'asymétrie. Les auteurs précurseurs dans cette littérature sont Nelson, Donaldson et Kamstra, Lundbergh et Terasvirta, Glosten, Jagannathan, Runkle et Hagerud et Sentana.

EGARCH

La seconde grande approche couvre les modèles ARCH non linéaires et plus particulièrement la prise en compte des phénomènes asymétries. L'idée est toute simple : l'effet hétéroscédastique n'est sans doute pas le même suivant que l'erreur précédente est positive ou négative. Nelson [1991] a proposé le processus GARCH exponentiel ou EGARCH(p,q) qui donne à la variance conditionnelle la définition suivante : Un processus u_t satisfait une représentation EGARCH(p,q) si et seulement si :

$$\log(h_t^2) = w + \sum_{i=1}^{q} \alpha_i g(v_{t-i}) + \sum_{j=1}^{p} \beta_j \log(h_{t-j}^2)$$

où le résidu normalisé v_t est un bruit faible et où la fonction $g(.)$ vérifie :

$$g(v_{t-i}) = \theta\, v_{t-i} + \gamma(|v_{t-i}| - \mathbb{E}|v_{t-i}|)$$

Si l'on pose $a_i = \theta\alpha_i$ et $b_i = \alpha_i\gamma$, la variance conditionnelle de u_t peut se réécrire sous la forme :

$$\log(h_t^2) = w + \sum_{i=1}^{q} a_i v_{t-i} + \sum_{i=1}^{q} b_i \left(|v_{t-i}| - \mathbb{E}|v_{t-i}|\right) + \sum_{j=1}^{p} \beta_j \log(h_{t-j}^2)$$

Dans le cas d'un processus EGARCH(1,1), nous avons donc :

$$\log(h_t^2) = w + a v_{t-1} + b\left(|v_{t-1}| - \mathbb{E}|v_{t-1}|\right) + \beta \log(h_{t-1}^2)$$

Deux remarques doivent être faites à ce niveau :

1. L'écriture porte sur le logarithme de la variance conditionnelle h_t^2 de u_t, en conséquence aucune restriction n'a besoin d'être imposée sur les différents paramètres de l'équation pour assurer la positivité de h_t^2.
2. La variance conditionnelle h_t^2 fait apparaître un effet de signe, correspondant à $a v_{t-1}$, et un effet d'amplitude mesuré par $b\left(|v_{t-1}| - \mathbb{E}|v_{t-1}|\right)$.

GQARCH

Le processus GQARCH (Q pour Quadratic) suppose également des asymétries dans la réponse de la volatilité conditionnelle aux innovations. Il a été introduit par Engle et Ng [1993] et Sentana [1995].

Un processus u_t satisfait une représentation GQARCH(1,1) si et seulement si :

$$u_t = v_t h_t$$
$$h_t^2 = \omega + \gamma\, u_{t-1} + \alpha\, u_{t-1}^2 + \beta\, h_{t-1}^2$$

La variance conditionnelle est donc définie comme une forme quadratique en u_{t-1}, et elle sera positive lorsque ω, α, $\beta > 0$ et $\gamma^2 \leq 4\alpha\omega$. Il faut remarquer aussi que la forme quadratique $f(u_{t-1}) = \gamma\, u_{t-1} + \alpha\, u_{t-1}^2$ étant minimale en

$$-\frac{\gamma}{2\alpha}$$

la symétrie de la réponse n'est donc pas obtenue en zéro mais en ce point : à amplitude donnée de l'innovation passée, on a bien un impact sur h_t^2 différent selon le signe de u_{t-1}. Si $u_{t-1} > 0$, son impact sur h_t sera beaucoup plus grand que dans le cas où $u_{t-1} < 0$. Par ailleurs, Sentana [1995] et He et Teräsvirta [1999] ont montré que les conditions pour la stationnarité de la covariance sont identiques à celles dérivées dans le cadre du modèle GARCH(1,1), à savoir :

$$\alpha + \beta < 1$$

Notons ici que la stationnarité au nineau de la covariance ne dépend pas du paramètres d'asymétrie γ. La somme $p = \alpha + \beta$ peut aussi être considérée comme une mesure de la persistance des chocs de volatilité. Sentana [1995] a montré aussi que les conditions d'existence des moments non conditionnels d'ordre quatre sont exactement les mêmes que celles d'un modèle GARCH(1,1). De plus, comme u est un processus centré, les expressions de son espérance et de sa variance non conditionnelles sont également identiques à celles obtenues avec un GARCH(1,1) — la moyenne non conditionnelle est nulle, alors que la variance est donnée par $h_u^2 = \frac{\omega}{1-p}$. Nous pouvons démontrer aussi que les moments impairs sont toujours nuls, et que la série u_t est non corrélée. Les corrélations croisées entre u_t^2 et u_{t-k} sont aussi nulles pour $k \neq 1$. Dans le cas où $k = 1$, $Cov(u_t^2, u_{t-1}) = \gamma\, h_u^2$ pour un modèle GQARCH(1,1) et zéro pour un modèle GARCH(1,1). En se basant sur les résultats de He et Teräsvirta [1999], on peut démontrer que la kurtosis de u_t est donnée par

$$\mathbb{k}_u = \mathbb{k}_v \left[1 - \frac{\alpha^2\,(\mathbb{k}_v - 1)}{1 - p^2}\right]^{-1} + \mathbb{k}_v \frac{A^*}{1 - \alpha^2\,(\mathbb{k}_v - 1) - p^2} \tag{1.7}$$

où $A^* = (\gamma/h_u)^2$ et $\mathbb{k}_v$ la kurtosis de v_t. Nous remarquons donc que la kurtosis est croissante avec la valeur absolue de γ, et naturellement égale à celle afférente au GARCH lorsque les deux processus sont confondus, soit pour $\gamma = 0$. Ce gain explique que le GQARCH domine souvent empiriquement le GARCH, ce dernier ayant tendance à sous-estimer l'épaisseur des queues de distribution.

La fonction d'autocorrélation de u_t^2 est donnée par :

$$\varphi_2(\tau) = \begin{cases} \frac{2\alpha(1-p^2+\alpha\,p)+A^*(\mathbb{k}_v\alpha+\beta)}{2(1-p^2+\alpha^2)+\mathbb{k}_v A^*}, & \tau = 1 \\ p^{\tau-1}\varphi_2(1), & \tau > 1 \end{cases} \tag{1.8}$$

Cette fonction décroît de la même façon que celle d'un modèle GARCH(1,1). Pour une faible valeur de A^*, l'autocorrélation d'ordre un devient presque la même dans les deux modèles. Après quelques transformations algébriques des équations (1.7) et (1.8), nous pouvons exprimer l'autocorrélation d'ordre un en fonction des kurtosis et de la persistence, soit

$$\varphi_2(1) = \frac{2\sqrt{\frac{(\mathbb{k}_u-\mathbb{k}_v)(1-p^2)-\mathbb{k}_v A^*}{(\mathbb{k}_v-1)\mathbb{k}_u}}\left[1 - p^2 + p\sqrt{\frac{(\mathbb{k}_u-\mathbb{k}_v)(1-p^2)-\mathbb{k}_v A^*}{(\mathbb{k}_v-1)\mathbb{k}_u}}\right]}{2\left[1 - p^2 + \frac{(\mathbb{k}_u-\mathbb{k}_v)(1-p^2)-\mathbb{k}_v A^*}{(\mathbb{k}_v-1)\mathbb{k}_u}\right] + \mathbb{k}_v A^*} + \frac{A^*\left[p + (\mathbb{k}_v - 1)\sqrt{\frac{(\mathbb{k}_u-\mathbb{k}_v)(1-p^2)-\mathbb{k}_v A^*}{(\mathbb{k}_v-1)\mathbb{k}_u}}\right]}{2\left[1 - p^2 + \frac{(\mathbb{k}_u-\mathbb{k}_v)(1-p^2)-\mathbb{k}_v A^*}{(\mathbb{k}_v-1)\mathbb{k}_u}\right] + \mathbb{k}_v A^*}$$

La figure 1.1 représente cette relation pour un modèle GQARCH(1,1) gaussien, dans le cas où $\alpha+\beta = 0.9$, $A^* = 0$ et $A^* = 0.99$, aussi bien que dans le cas où $\alpha+\beta = 0.99$ pour les mêmes valeurs de A^*. Ce graphique montre donc que dans le cas usuel où la kurtosis prend des valeurs entre 5 et 10, l'introduction d'un terme d'asymétrie au niveau de la spécification GARCH(1,1), qui nécessite à son tour l'introduction d'autres contraintes permettant de garantir une kurtosis finie, n'aura aucun effet significatif sur la relation entre les trois quantités d'intérêt. Par exemple, si $\alpha+\beta = 0.99$ et $A^* = 0.1$, la kurtosis d'un GQARCH(1,1) sera toujours supérieure à 17.

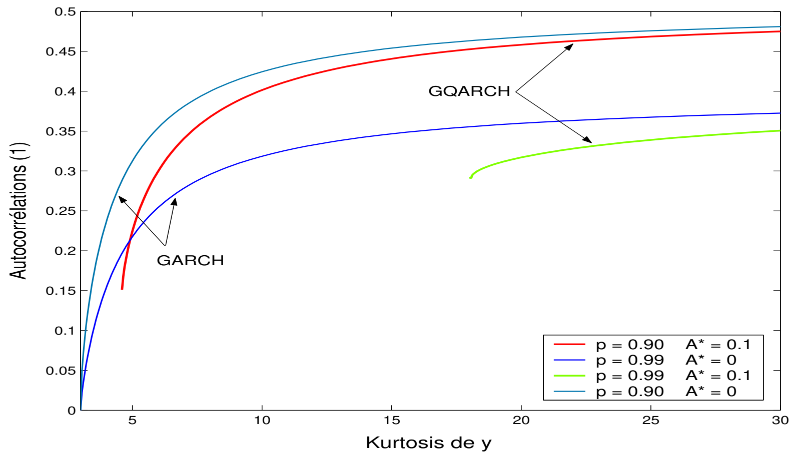

FIG. 1.1 – Relation entre l'autocorrélation d'ordre un, la kurtosis et la persistence d'un modèle GQARCH(1,1) asymétrique.

1.3.3 Les Modèles à Variance Stochastique

La classe des modèles de volatilité stochastique est apparue comme une approche alternative pour les modèles de type ARCH. Cette approche consiste à formuler un modèle contenant une composante de variance non observable, son logarithme est modélisé directement comme un processus stochastique d'autorégression linéaire. Ces modèles exploitent donc la prévisibilité de la volatilité à partir des variances conditionnelles passées pour déterminer les rendements d'actifs. Une approche plus ancienne explique la dynamique des rendements par le flux d'information (voir Clark [1973]). Cette idée est justifiée par plusieurs études empiriques où l'on observe l'effet de publication de données économiques importantes sur la volatilité (voir, par exemple, Baillie et Bollerslev [1991]). Nous pouvons introduire aussi les valeurs absolues des rendements passés pour modéliser une asymétrie dans le comportement de la volatilité suite à un accroissement ou une baisse des prix. Les liens entre la dynamique des rendements financiers et celles du volume d'échanges peuvent être analysés d'avantage dans le cadre d'un modèle bi-varié de volatilité stochastique.

1.3.4 L'Approche Factorielle des Modèles à Variance Dynamique

Dans la littérature financière et jusqu'au début des années quatre-vingt-dix, les modèles d'évaluation des actifs ont été considérés dans un cadre statique. Cependant et avec le développement de cette nouvelle famille de modèles d'hétéroscédasticité dynamique des variances conditionnelles, la recherche en finance de marché a porté beaucoup plus ces dernières années sur la modélisation de l'inter-dépendance entre les processus de volatilité à travers des modèles inter-temporelles. Ces modèles sont basés sur l'hypothèse que les réactions des agents financiers sont essentiellement fondées sur la distribution des rendements conditionnellement à leur ensemble informationnel supposé variable à travers le temps.

Ces modèles sont basés sur les mêmes principes des modèles à facteurs standards : on suppose toujours que chacune des variables observables y_{it} est une combinaison linéaire de k $(k < q)$ facteurs communs non observables f_{it} plus un terme idiosyncratique ε_{it}, mais la seul différence avec l'approche classique c'est que cette fois-ci les facteurs communs sont supposés suivre des processus conditionnellement hétéroscédastiques de type ARCH. Dans ce cas nous pouvons aussi obtenir une représentation parcimonieuse pour les moments conditionnels de second ordre en terme d'un nombre de facteurs beaucoup plus petit que la dimension du vecteur des observations. Une telle spécification nous permettra, aussi, d'éviter les problèmes de calcul liés au grand nombre de paramètres à estimer engendrés par les modèles de volatilité multi-variés. Parmi les travaux qui ont été menés dans ce sens nous pouvons citer, sans être exhaustif, le modèle à facteurs GARCH de Engle [1987] ; le modèle GARCH à facteurs latents de Diebold et Nerlove [1989] et les modèles de Kroner [1987] ; Harvey, Ruiz et Sentana [1992] ; Lin [1992] ; Ng, Engle et Rothschild [1992] ; Bollerslev et Engle [1994] qui ont étudié les conditions de stationnarité de la covariance des modèles GARCH à k-facteurs ; King, Sentana et Wadhwani [1994] ; Sentana et Shah [1994] ; Demos et Parissi [1998] ; Demos et Sentana [1998] ; Sentana [1998] ; Sentana [2000] et enfin le modèle ARCH généralisé à structure latente de Fiorentini, Sentana, et Shephard [2004] et qui ont proposé une approche purement bayésienne pour l'estimation de ses paramètres.

Structure Générale du Modèle

$$y_{it} = \mu_{it} + \eta_{it}$$

$$\eta_{it} = \underbrace{x_{i1}f_{1t} + x_{i2}f_{2t}}_{\text{risque systématique}} + \underbrace{\varepsilon_{it}}_{\text{risque spécifique}}$$

$$\mu_{it} = x_{i1}h_{1t}\tau_1 + x_{i2}h_{2t}\tau_2$$

$$Var_{t-1}(r_{it}) = x_{i1}^2 h_{1t} + x_{i2}^2 h_{2t} + \psi_{it}$$

$$Cov_{t-1}(r_{it}, r_{jt}) = x_{i1}x_{j1}h_{1t} + x_{i2}x_{j2}h_{2t}$$

Certains auteurs ont proposé des modèles à facteurs qui tiennent en compte à la fois l'effet de certaines variables économiques observables et d'un certain nombre de facteurs non observables. Les facteurs communs ont été supposés suivre des processus GQARCH(1,1) univariés. Dans ce cas, la prime du risque associée à chacun des actifs peut être modélisée comme une combinaison linéaire des volatilités associées aux différents facteurs. La structure de ce modèle (dans le cas d'un seul facteur observable et un seul facteur non observable) est donnée par l'encadré ci-dessus, où y_{it} est l'excès de rendement de l'actif i durant la période t, μ_{it} la prime du risque de l'actif i, η_{it} le rendement non anticipé de l'actif i, f_{1t} le facteur commun "observable" lié aux innovations des variables exogènes, f_{2t} le facteur commun "non observable", x_{i1} la sensibilité du rendement de l'actif i à f_{1t}, x_{i2} la sensibilité du rendement de l'actif i à f_{2t}, ε_{it} le risque spécifique à l'actif i, h_{1t} la variance conditionnelle du facteur "observable", h_{2t} la variance conditionnelle du facteur "non observable", τ_1 est le prix de risque du facteur "observable" alors que τ_2 est le prix de risque du facteur "non observable", et ψ_{it} la variance conditionnelle idiosyncratique de l'actif i. Ce modèle peut, donc, être considéré comme une version dynamique de l'APT.

1.4 Généralisation Espace-État Dynamique

Dans la section 1.2.2 on a vu que les paramètres β_i peuvent varier au cours du temps. Dans le cadre du CAPM, ils ont une interprétation en termes de moments car ils sont le rapport d'une covariance et d'une variance. Si on accepte que les distributions des rendements ne sont pas les mêmes d'une période à l'autre on obtient des paramètres variables. D'une manière plus générale on peut se dire que le modèle qu'on a estimé hier a changé aujourd'hui car même si sa structure est toujours la même, ses paramètres ont changé. En régression ce problème est bien connu. On peut régresser les mesures d'une variable par rapport au temps sur une période de deux années consécutives et trouver des coefficients de régression totalement différents d'une année à l'autre alors qu'on estime à chaque fois le même modèle. On peut tester l'éventuel changement à l'aide d'une statistique F introduite par Chow [1960].

Dans un article publié par Econometrica en 1989, Hamilton a introduit le concept de changements de régimes dans la recherche empirique en économétrie afin de prendre en compte un certain type de non stationnarité présente dans de nombreuses séries à caractère économique et financier. Ayant observé que ce type de séries présente souvent des ruptures dans leur moyenne, l'idée originale d'Hamilton fut de modéliser cette non stationnarité à l'aide d'un processus linéaire par morceaux. En particulier, on suppose que la série observée peut être approchée à l'aide d'un modèle dont les paramètres évoluent au cours du temps. De plus, Hamilton émet l'hypothèse que l'évolution de ces paramètres est gouvernée par une variable inobservable que l'on peut modéliser à l'aide d'une chaîne de Markov à m régimes. Ainsi, la série change dans son comportement en fonction de l'état prévalant.

Les changements au niveau de la structure interne des rendements financiers peuvent être le résultat de plusieurs aléas, que ceux soient de nature quantitative ou qualitative. Cecchetti, Lam et Mark [1990], par exemple, ont proposé un modèle d'évaluation d'actifs tenant compte des situations de forte et faible croissance économique. Ils ont montré que les transitions entre ces deux états de l'économie affectent certains comportements fondamentaux des rendements financiers, tels que l'aspect leptokurtique et le phénomène de retournement à la moyenne de la volatilité. D'autre part, Blanchard et Watson [1982] ont proposé un modèle pour étudier si la présence de bulles stochastiques (surévaluation de la valeur d'un titre, ou bien de l'ensemble des valeurs d'un secteur) peut provoquer un changement de régime de la courbe de rendement des valeurs boursières. Par la suite, Schaller et van Norden [1997] ont développé une nouvelle approche empirique, fondée sur l'emploi de méthodes de régression avec changement de régime, en vue de différencier deux modèles d'évaluation des actifs, soit le modèle des bulles et le modèle des engouements. Ils ont démontré par ailleurs que, si on part de l'hypothèse que l'hétéroscédasticité varie selon l'état, le modèle de Cutler, Poterba et Summers [1991] relatif aux engouements peut également déboucher sur un changement de régime.

En poursuivant la démarche de Hamilton [1990], le modèle que nous nous proposons d'étudier sera une combinaison entre les modèles à facteurs conditionnellement hétéroscédastiques et les modèles de Chaînes de Markov Cachées. Cette nouvelle spécification peut être considérée comme une généralisation espace-état dynamique des modèles standards. Les facteurs communs seront donc générés par une chaîne de Markov cachée à états gaussiens et les vecteurs d'observations par des modèles d'analyse factorielle conditionnellement hétéroscédastiques.

1.5 Conclusion

Contrairement aux modèles déjà existants, notre spécification admet des coefficients qui varient dans le temps et permet ainsi de mieux modéliser les divers aspects d'hétérogénéité dans la dynamique des séries financières. Le fait que les paramètres du modèle peuvent changer au cours du temps est un aspect non négligeable du traitement statistique des séries temporelles et il sera au coeur des développements proposés par ce travail sous la dénomination de paramétrisation dynamique. Bien que la technique de la fenêtre glissante apporte une première solution au problème elle implique un choix subjectif de la largeur de la fenêtre qui ne semble pas satisfaisante. On verra plus tard comment on peut intégrer dans les modèles le fait que les paramètres varient au cours du temps. Un deuxième problème traité par ce travail est l'estimation d'un facteur commun qui influence un ensemble de séries temporelles observables comme par exemple les rendements des actions. On a vu dans cette introduction que cela peut être utile étant donné qu'on voudrait avoir une estimation de l'évolution du marché. D'une manière plus générale, on va montrer comment on peut exprimer l'inter-dépendance d'un grand nombre de variables à l'aide d'une combinaison linéaire d'un petit nombre de facteurs et on parlera alors d'une structure factorielle des données.

CHAPITRE 2

Les Modèles à Facteurs Standards

Après une présentation générale de la théorie factorielle en finance, ce chapitre a pour objectif de décrire les modèles à facteurs standards et de faire un tour d'horizon des différentes méthodes d'estimation existantes dans la littérature. Les deux premières sections décrivent la structure générale du modèle, les conditions nécessaires d'identification et l'approche de maximum de vraisemblance proposée par Jöreskog. Les différents éléments pouvant contribuer à une comparaison entre les techniques de l'analyse factorielle et de l'analyse en composantes principales, seront aussi présentés. Dans la troisième section un accent particulier sera mis sur l'approche itérative et les algorithmes de type EM qui constituent une des bases de notre travail. Le problème particulier d'une structure factorielle oblique sera discuté dans la dernière section.

2.1 Introduction

Conçue à l'origine pour l'analyse de tests psychométriques, l'analyse factorielle a été introduite par Spearman [1904], Kelley [1928] et Thurstone [1931]. Il ont introduit la représentation de l'espace factoriel, les rotations de facteurs à l'intérieur de cet espace, l'usage du calcul matriciel, la notion de structure simple, les facteurs obliques, et les facteurs de second ordre. Depuis lors ces méthodes n'ont cessé de se développer et de se diversifier, notamment sous l'impulsion de Hotelling [1933] en économétrie. Au début c'était le besoin de condenser les données statistique de grande dimension sur plusieurs variables en un nombre beaucoup plus petit d'indices où de facteurs communs qui était à l'origine des modèles à facteurs dans les sciences sociales. Ces indices sont dans la plupart du temps considérés comme des variables latentes. L'analyse factorielle exprime, donc, la corrélation entre un grand nombre de variables observées à l'aide d'un petit nombre de facteurs. Les variables sont décrites comme des combinaisons linéaires des facteurs auxquels on ajoute du bruit. Les facteurs sont des variables non observables ou tout simplement l'information les concernant manque et doivent être estimés en même temps que les paramètres du modèle.

2.2 Les Modèles à Facteurs Orthogonaux

Dans les applications financières, les modèles à facteurs ont été utilisés pour la première fois au début des années 1960 comme une approche alternative au CAPM. Les différentes techniques qui ont été mises en place pour l'estimation de ces modèles sont essentiellement basées sur la méthode de maximum de vraisemblance avec des contraintes d'identification sur les paramètres.

2.2.1 Modèle de Base et Structure des Facteurs

Ce modèle a été initialement crée pour l'étude de données individuelles. Il repose sur la modélisation suivante : on note q le nombre de variables étudiées, n le nombre d'observations dont on dispose pour chaque variable, et y_{it} la valeur de la t-ème observation de la variable y_i ; le modèle décrivant les variables $y_1, ..., y_q$ en fonction de k facteurs communs $f_1, ..., f_k$, $k < q$, s'écrit :

$$y_{it} = \theta_i + x_{i1} f_{1t} + ... + x_{ik} f_{kt} + \varepsilon_{it} \quad \text{pour} \quad t = 1, ..., n$$

On suppose que les $\{\varepsilon_{it}\}$ sont indépendants entre eux et indépendants des facteurs. En outre, comme le modèle est destiné à l'étude de données individuelles, on suppose que les différentes observations d'une même variable (indicées ici par t) sont non corrélées entre elles. Dans le cadre d'une modélisation multivariée, si on désigne par $\mathbf{y}_t$ le vecteur des variables observables de dimension $q \times 1$; $\theta = [\theta_1, \theta_2, ..., \theta_q]'$ le vecteur des moyennes de dimension $q \times 1$. La forme matricielle de ce modèle sera donnée par :

$$\mathbf{y}_t = \theta + \mathbf{X}\mathbf{f}_t + \varepsilon_t \tag{2.1}$$

où $\mathbf{X}$ est une matrice déterministe de dimension $q \times k$, à coefficients inconnus appelée matrice des pondérations (loadings). Les éléments du vecteur aléatoire $\mathbf{f}_t$ de dimension $k \times 1$ sont les facteurs communs ou les facteurs scores. ε_t est un vecteur aléatoire de dimension $q \times 1$ dont les éléments sont les facteurs spécifiques, appelés aussi facteurs uniques ou "idiosyncratiques". La variance de ce vecteur représente la variabilité des observations non expliquée par les facteurs communs. Enfin les ε_t sont supposés mutuellement indépendants $\forall\, t$. Les hypothèses classiques de ce modèle sont :

- $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$ où $\mathbf{\Psi} = diag(\psi_1, ..., \psi_q)$, appelée matrice des variances "idiosyncratiques",
- des facteurs non corrélés et standardisés $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, et
- ε_t et $\mathbf{f}_s$ sont mutuellement indépendants pour tout t, s.

Ce modèle est donc construit de telle façon que par conditionnement sur les facteurs communs, les variables observables sont indépendantes ce qui implique que toute la corrélation entre les variables de départ passe par les k facteurs. En effet, la loi de la séquence complète des variables de départ $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ conditionnellement à la séquence des facteurs communs $\mathcal{F} = \{\mathbf{f}_1, ..., \mathbf{f}_n\}$ se factorise sous la forme :

$$p(\mathcal{Y}/\mathcal{F};\Theta) = \prod_{t=1}^{n} p(\mathbf{y}_t/\mathbf{f}_t;\Theta)$$

Ce modèle consiste, donc, à chercher un nombre k minimal de facteurs tel que cette propriété soit vérifiée. Dans le cas Gaussien, une telle propriété caractérise de façon claire les modèles à facteurs. En effet, si l'on suppose que

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{f}_t \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \theta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}\mathbf{X}' + \mathbf{\Psi} & \mathbf{X} \\ \mathbf{X}' & \mathbf{I}_k \end{pmatrix}\right]$$

la loi de $\mathbf{y}$ sachant $\mathbf{f}$ est de la forme $\mathcal{N}(\theta + \mathbf{X}\mathbf{f}_t, \mathbf{\Psi})$ et on a :

$$p(\mathbf{y}_t/\mathbf{f}_t) = \prod_{i=1}^{q} p(y_{it}/\mathbf{f}_t)$$

si et seulement si $\mathbf{\Psi}$ est diagonale. Dans ce cas $\mathbf{X}$ et $\mathbf{\Psi}$ seront définis par : $\mathbf{X} = \mathbb{E}(\mathbf{y}_t\mathbf{f}_t')$ et $\mathbf{\Psi} = Var(\mathbf{y}_t - \theta - \mathbf{X}\mathbf{f}_t)$. Mais la définition de Bartholomew [1987] s'étend évidemment à un cadre non Gaussien sous la forme suivante :

Définition On dit que $\mathbf{y} = [\mathbf{y}_1', \mathbf{y}_2', ..., \mathbf{y}_q']'$ est un vecteur aléatoire vérifiant un modèle à k facteurs, si et seulement s'il existe un vecteur aléatoire $\mathbf{f}$ à valeurs dans $\mathbb{R}^k$ tel que, conditionnellement à $\mathbf{f}$, les variables aléatoires $\mathbf{y}_1, ..., \mathbf{y}_q$ soient indépendantes.

Pour deux éléments quelconques y_{it} et y_{jt} de $\mathbf{y}_t$, les moments sont caractérisés par :

$$\begin{aligned} Var(y_{it}/\theta, \mathbf{X}, \mathbf{f}, \mathbf{\Psi}) &= \psi_i \qquad \text{et} \\ Var(y_{it}/\theta, \mathbf{X}, \mathbf{\Psi}) &= \sum_{l=1}^{k} x_{il}^2 + \psi_i \qquad \forall\, i \end{aligned}$$

d'autre part on a :

$$\begin{aligned} Cov(y_{it}, y_{jt}/\theta, \mathbf{X}, \mathbf{f}, \mathbf{\Psi}) &= 0 \qquad \text{et} \\ Cov(y_{it}, y_{jt}/\theta, \mathbf{X}, \mathbf{\Psi}) &= \sum_{l=1}^{k} x_{il}x_{jl} \qquad \forall\, i, j \quad i \neq j \end{aligned}$$

En se basant sur ces propriétés, nous pouvons exprimer autrement le modèle à k-facteurs par une simple condition sur la matrice de variance-covariance $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \mathbf{X}\mathbf{X}' + \mathbf{\Psi} \tag{2.2}$$

où les éléments diagonaux de la matrice de variance-covariance associés aux facteurs $\mathbf{XX}'$ sont appelés les **communalités**, $\mathbf{x}_i^2 = \sum_{j=1}^{k} x_{ij}^2$, pour $i = 1, ..., q$ alors que les éléments de $\mathbf{\Psi}$ sont appelés les **spécificités** ou les **unicités**. Si nous prenons l'exemple d'évaluation d'actifs, cette écriture signifie que le rendement de tout titre primaire est expliqué par des titres communs d'une part et par un titre spécifique d'autre part. On peut considérer que les titres communs sont formés d'un panier de titres émis par les entreprises des principaux secteurs économiques. On peut, par exemple, utiliser les 40 valeurs du CAC, avec des coefficients de pondération variables, pour expliquer le rendement de tous les titres du marché boursier parisien. Cette écriture suppose donc que même s'il existe un nombre infini de titres de base, il est possible pour l'évaluation de la prime de risque de ne retenir qu'un petit nombre de titres communs qui représenteront les facteurs macro-économiques des aléas économiques constatés sur les marchés financiers, tandis qu'un titre unique représentera le risque spécifique.

2.2.2 La Méthode d'Analyse en Composantes Principales

Si la parenté entre les techniques de l'analyse factorielle et de l'analyse en composantes principales (ACP) est reconnue par l'ensemble des auteurs, il semble qu'il y ait toujours eu une certaine dimension polémique entre les tenants de l'une et l'autre démarche. En particulier il est habituel de voir opposer l'ACP, technique purement descriptive, à l'analyse factorielle reposant sur un modèle probabiliste. Cependant, il existe des liens entre les deux approches. En particulier, dans certains cas précis, l'analyse factorielle peut être interprétée comme une ACP.

Approche Théorique

Étant donné un vecteur aléatoire $\mathbf{y}$ de taille q, le but de l'ACP est de construire des variables aléatoires $f_1, ..., f_k$ linéaires en $\mathbf{y}$, deux à deux non corrélées et de variance maximale[1]. Ces variables sont construites de façon itérative : on cherche d'abord $f_1 = \beta_1'\mathbf{y}$ de variance maximale sous la contrainte $\beta_1'\beta_1 = 1$, puis $f_2 = \beta_2'\mathbf{y}$ de variance maximale sous les contraintes $\beta_2'\beta_2 = 1$ et $Cov(f_1, f_2) = 0$, et, de façon générale, $f_k = \beta_k'\mathbf{y}$ de variance maximale sous les contraintes $\beta_k'\beta_k = 1$ et $Cov(f_i, f_k) = 0 \; \forall \, i < k$.

Les résultats standards sur la diagonalisation des matrices symétriques permettent de montrer aisément que $\beta_1, ..., \beta_k$ sont des vecteurs propres de norme 1 de la matrice $\mathbf{\Sigma}$, associés respectivement aux k plus grandes valeurs propres de cette matrice.

Si $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k ... \geq \lambda_q$ sont les valeurs propres de $\mathbf{\Sigma}$, on obtient en outre :

$$\forall \, j = 1, ..., k \quad Var(f_j) = Var(\beta_j'\mathbf{y}) = \lambda_j$$

En pratique, on choisit k de façon à ce que la "part de la variance expliquée" par $f_1, ..., f_k$ soit suffisamment grande, c'est-à-dire de façon que $\sum_{i=1}^{q} Var(y_i) - \sum_{i=1}^{k} Var(f_i)$ soit inférieure à un nombre ε fixé d'avance et proche de zéro. Ceci peut s'écrire encore :

[1] On se limite ici à la présentation de l'ACP pour la métrique identité, l'objectif essentiel est d'introduire la comparaison avec l'analyse factorielle.

$$tr\,\mathbf{\Sigma} - \sum_{i=1}^{k} \lambda_i < \varepsilon \quad \text{où} \quad \sum_{i=k+1}^{q} \lambda_i < \varepsilon$$

Pour faciliter la comparaison ultérieure avec le modèle de l'analyse factorielle, il est utile de détailler un peu plus la démarche qui est faite ici, et de noter en particulier que cette démarche consiste à approximer la matrice $\mathbf{\Sigma}$ par une matrice de rang plus petit. En effet, si l'on note $\beta^* = [\beta_1, ..., \beta_q]$ la matrice orthogonale dont les colonnes sont constituées par une base orthonormée de vecteurs propres de $\mathbf{\Sigma}$, on peut écrire $\mathbf{\Sigma} = \beta^* \mathbf{\Delta} \beta^{*\prime}$ avec

$$\mathbf{\Delta} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_q \end{bmatrix} \quad \lambda_1 \geq ... \geq \lambda_q > 0$$

Si l'on note : $\beta_1^* = [\beta_1, ..., \beta_k]$, $\beta_2^* = [\beta_{k+1}, ..., \beta_q]$, et

$$\mathbf{\Delta}_1 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{bmatrix}, \quad \mathbf{\Delta}_2 = \begin{bmatrix} \lambda_{k+1} & & 0 \\ & \ddots & \\ 0 & & \lambda_q \end{bmatrix}$$

on a donc : $\mathbf{\Sigma} = \beta_1^* \mathbf{\Delta}_1 \beta_1^{*\prime} + \beta_2^* \mathbf{\Delta}_2 \beta_2^{*\prime}$.

Cette démarche revient donc à approximer $\mathbf{\Sigma}$ par $\beta_1^* \mathbf{\Delta}_1 \beta_1^{*\prime}$ et à considérer que cette approximation est satisfaisante dès lors que $tr\mathbf{\Delta}_2 = tr\,[\beta_2^* \mathbf{\Delta}_2 \beta_2^{*\prime}] < \varepsilon$. En posant $\Lambda = \beta_1^* \mathbf{\Delta}_1^{\frac{1}{2}}$, nous pouvons décomposer $\mathbf{\Sigma}$ sous la forme :

$$\mathbf{\Sigma} = \Lambda\Lambda' + \mathbf{D}_2 \quad \text{avec} \quad \mathbf{D}_2 = \beta_2^* \mathbf{\Delta}_2 \beta_2^{*\prime} \quad \text{et} \quad tr\mathbf{D}_2 < \varepsilon$$

Si maintenant on note $\mathbf{f} = [f_1, ..., f_k]' = \beta_1^{*\prime}\mathbf{y}$ le vecteur aléatoire constitué des k premières composantes principales, et si l'on note $\varphi = \mathbf{\Delta}_1^{-\frac{1}{2}}\mathbf{f}$, on peut aussi effectuer une décomposition de $\mathbf{y}$ sous la forme suivante :

$$\mathbf{y} = \beta_1^* \mathbf{f} + \mathbf{u} = \Lambda\varphi + \mathbf{u}$$

avec $\mathbf{u} = \mathbf{y} - \beta_1^* \beta_1^{*\prime}\mathbf{y} = \beta_2^* \beta_2^{*\prime}\mathbf{y}$ et $\Lambda = \beta_1^* \mathbf{\Delta}_1^{\frac{1}{2}}$. Dans cette décomposition, $\Lambda\varphi = \beta_1^* \beta_1^{*\prime}\mathbf{y}$ est la projection orthogonale de $\mathbf{y}$ sur le sous espace vectoriel engendré par les vecteurs colonnes de β_1^* (puisque $\beta_1^{*\prime}\beta_1^* = \mathbf{I}_k$), qui coincide avec le sous espace vectoriel engendré par les vecteurs colonnes de Λ. On vérifie d'ailleurs que $\varphi = (\Lambda'\Lambda)^{-1}\Lambda'\mathbf{y}$.

Cette décomposition vérifie les propriétés suivantes :

- $Var(\varphi) = \mathbf{\Delta}_1^{-\frac{1}{2}} Var(\mathbf{f}) \mathbf{\Delta}_1^{-\frac{1}{2}} = \mathbf{\Delta}_1^{-\frac{1}{2}} \mathbf{\Delta}_1 \mathbf{\Delta}_1^{-\frac{1}{2}} = \mathbf{I}_k$
- $Var(\mathbf{u}) = \beta_2^* \beta_2^{*\prime} \mathbf{\Sigma} \beta_2^* \beta_2^{*\prime} = \beta_2^* \mathbf{\Delta}_2 \beta_2^{*\prime} = \mathbf{D}_2$

- $\mathbb{E}\left[\mathbf{u}\varphi'\right] = \mathbb{E}\left[\beta_2^*\beta_2^{*\prime}\mathbf{y}\mathbf{y}'\beta_1^*\mathbf{D}_1^{-\frac{1}{2}}\right] = \beta_2^*\beta_2^{*\prime}\mathbf{\Sigma}\beta_1^*\mathbf{\Delta}_1^{-\frac{1}{2}} = \mathbf{0}$

On a supposé ici $\mathbb{E}(\mathbf{y}) = 0$ pour simplifier les notations. La différence entre cette écriture et le modèle de l'analyse factorielle tient donc dans les propriétés de $\mathbf{u}$. L'hypothèse faite ici est seulement que $Var(\mathbf{u})$ est une matrice, de rang $(q-k)$, dont la trace est inférieure à un nombre ε fixé à l'avance.

Approche Empirique

Lorsqu'on s'intéresse à un échantillon $\mathbf{y}_1, ..., \mathbf{y}_n$ de la variable $\mathbf{y}$, on définit les composantes principales $f_{it} = \widehat{\beta}_i'\mathbf{y}_t$ de façon analogue. Les $\widehat{\beta}_i$ pour $i = 1$ à k sont obtenus comme vecteurs propres associés aux k plus grandes valeurs propres de la matrice de variance-covariance empirique.

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t - \overline{\mathbf{y}})(\mathbf{y}_t - \overline{\mathbf{y}})'$$

Il est intéressant de noter que, lorsque les $\mathbf{y}_t$ sont supposés suivre une loi $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, les $\widehat{\beta}_i$ sont les estimateurs du maximum de vraisemblance des β_i.

Il est possible (voir par exemple Anderson [2003]) de calculer les lois limites des valeurs propres et des vecteurs propres de $\mathbf{\Sigma}$. Ceci peut permettre de mener des tests sur le nombre k de composantes principales à retenir, afin de ne pas s'en tenir à des critères purement descriptifs pour effectuer ce choix. On peut en effet effectuer des tests d'hypothèses de la forme :

$$\left\{\begin{array}{l} H_0 : \sum\limits_{i=k+1}^{q} \lambda_i \geq \varepsilon \quad \text{contre} \quad H_1 : \sum\limits_{i=k+1}^{q} \lambda_i < \varepsilon \quad \text{ou} \\ H_0 : \dfrac{\sum_{i=k+1}^{q} \lambda_i}{\sum_{i=1}^{q} \lambda_i} \geq \delta \quad \text{contre} \quad H_1 : \dfrac{\sum_{i=k+1}^{q} \lambda_i}{\sum_{i=1}^{q} \lambda_i} < \delta \end{array}\right.$$

L'ACP Comme un cas Particulier du Modèle à Facteurs

Supposons que les variables observées vérifient un modèle à k facteurs, c'est-à-dire, de façon équivalente que leur matrice de variance-covariance s'écrit $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}' + \mathbf{\Psi}$ avec $\mathbf{X}$ matrice $(q \times k)$ de rang k et $\mathbf{\Psi}$ diagonale définie positive.

Si la matrice $\mathbf{\Psi}$ était connue, on pourrait écrire $\mathbf{\Sigma} - \mathbf{\Psi} = \mathbf{X}\mathbf{X}'$. La matrice $\mathbf{\Sigma} - \mathbf{\Psi}$ étant alors de rang k exactement, elle admettrait $q-k$ valeurs propres nulles et une ACP menée sur la matrice $\mathbf{\Sigma} - \mathbf{\Psi}$ fournirait la matrice $\mathbf{X}$ de façon exacte. En effet, en reprenant les notations précédentes, on aurait :

$$\mathbf{\Sigma} - \mathbf{\Psi} = \beta^*\mathbf{\Delta}\beta^{*\prime} \quad \text{avec} \quad \mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}_1 & 0 \\ 0 & 0 \end{bmatrix}$$

donc $\mathbf{\Sigma} - \mathbf{\Psi} = \beta_1^* \mathbf{\Delta}_1 \beta_1^{*\prime} = (\beta_1^* \mathbf{\Delta}_1^{\frac{1}{2}})(\mathbf{\Delta}_\mathbf{1}^{\frac{1}{2}} \beta_1^{*\prime}) = \Lambda\Lambda'$

Lorsque $\mathbf{\Sigma}$ est la matrice de corrélation, on dit que $\mathbf{\Sigma} - \mathbf{\Psi}$ est la matrice de corrélation réduite, et on peut donc énoncer que l'analyse factorielle est équivalente à une ACP sur la matrice de corrélation réduite.

Bien sûr, en pratique, $\mathbf{\Psi}$ n'est pas connue par avance. Cependant, ce qui vient d'être dit est intéressant d'un point de vue pratique. Pour l'estimation d'un modèle à facteurs, on commence souvent la procédure par l'application d'une ACP sur une approximation de la matrice de corrélation réduite, qui permet d'obtenir une évaluation quantitative du nombre de facteurs communs à retenir.

2.3 Les Contraintes d'Identification

La structure du modèle à k-facteurs déjà présentée par l'équation (2.1) est caractérisée par un nombre assez important de paramètres. Une telle caractéristique va conduire à des problèmes d'identification (multiplicité de solutions), d'où la nécessité d'imposer certaines restrictions sur la structure des corrélations (2.2).

2.3.1 Rang de la Matrice des Pondérations

Le problème lié au rang de la matrice des pondérations $\mathbf{X}$ n'a pas été suffisamment évoqué dans les travaux de recherche antérieurs (portant sur des applications financières). En effet, la plupart de ces travaux ont supposé d'une manière implicite que le nombre de facteurs k est connu par avance. Si ce n'est pas le cas et $\mathbf{X}$ n'est pas de plein rang, donc le modèle ne sera pas complètement identifié.

Supposons, par exemple, que $rang(\mathbf{X}) = r$ avec $r < k$, alors il existe une matrice $\mathbf{Q}$ de dimension $k \times (k-r)$ tel que $\mathbf{XQ} = \mathbf{0}$ et $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_{k-r}$. Si $\mathbf{M}$ est une matrice quelconque de dimension $q \times (k-r)$ choisie de telle façon que $\mathbf{MM}'$ soit diagonale, donc la matrice de variance-covariance $\mathbf{\Sigma}$ peut être exprimée sous la forme :

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{XX}' + \mathbf{\Psi} \\
&= (\mathbf{XX}' + \mathbf{MM}') + \mathbf{\Psi} - \mathbf{MM}' \\
&= (\mathbf{X} + \mathbf{MQ}')(\mathbf{X} + \mathbf{MQ}')' + \mathbf{\Psi} - \mathbf{MM}'
\end{aligned}$$

Ceci implique que $\mathbf{\Sigma} = \widehat{\mathbf{X}}\widehat{\mathbf{X}}' + \widehat{\mathbf{\Psi}}$ où $\widehat{\mathbf{X}} = \mathbf{X} + \mathbf{MQ}'$ et $\widehat{\mathbf{\Psi}} = \mathbf{\Psi} - \mathbf{MM}'$ et, par conséquent, le modèle ne sera pas identifié d'une façon unique.

Solution : *L'existence et l'unicité du modèle à facteurs ne seront garanties que si* $rang(\mathbf{X}) = k$.

2.3.2 Rotations Orthogonales

Maintenant si on suppose que $rang(\mathbf{X}) = k$, la validité du modèle à k-facteurs reste toujours vérifiée même dans le cas où ces derniers (c-à-d les facteurs) obéiront à une rotation. En effet, si on désigne par $\mathbf{P}$ une matrice orthogonale de dimension $k \times k$, le vecteur des observations $\mathbf{y}_t$ peut s'écrire sous la forme :

$$\mathbf{y}_t = \theta + \mathbf{X}^*\mathbf{f}_t^* + \varepsilon_t \tag{2.3}$$

où, d'une part, les facteurs qui ont obéit à une rotation $\mathbf{f}_t^* = \mathbf{P}'\mathbf{f}_t$ et la matrice des pondération qui leurs correspond $\mathbf{X}^* = \mathbf{XP}$ vérifient toujours un modèle à k-facteurs sans affecter la distribution de $\mathbf{y}_t$. D'autre part, les deux premiers moments $\mathbb{E}(\mathbf{f}_t^*) = \mathbf{0}$ et $Var(\mathbf{f}_t^*) = \mathbf{P}'\mathbf{P} = \mathbf{I}_k$, permettent aussi de vérifier la relation $\mathbf{\Sigma} = \mathbf{X}^*\mathbf{X}^{*\prime} + \mathbf{\Psi}$. On a ainsi une infinité de solutions possibles basées sur des transformations orthogonales près des facteurs. Ce problème est, essentiellement, lié à l'invariance de la fonction de vraisemblance sous des transformations linéaires inversibles des facteurs. Plusieurs solutions ont été proposées dans la littérature, dans ce chapitre nous allons envisager deux. Cependant, il faut noter que chacune d'entre elles a ses propres lacunes surtout en ce qui concerne l'interprétation des facteurs (Press et Shigemasu, [1989]).

Solutions Possibles

1. La solution la plus simple proposée par Geweke et Zhou [1996] consiste à imposer des contraintes "hiérarchiques" sur la matrice des pondérations. Dans ce cas si on suppose sans aucune perte de généralité que les k premières lignes de $\mathbf{X}$ sont indépendantes, la matrice des pondérations peut, donc, être exprimée sous la forme $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ où $\mathbf{X}_1$ est une matrice de dimension $k \times k$ formée par les k premières lignes de $\mathbf{X}$ et $\mathbf{X}_2$ la matrice de dimension $(q-k) \times k$ des lignes restantes. Étant donné que $\mathbf{X}_1$ est non singulière, donc il existe une seule et unique matrice $\mathbf{P}$ orthogonale permettant d'obtenir une matrice triangulaire inférieure $\mathbf{X}_1\mathbf{P}'$ dont les éléments diagonaux sont positifs. En effet, si on désigne par $\mathbf{A} = \mathbf{X}_1\mathbf{X}_1'$ une matrice symétrique et définie positive, nous pouvons effectuer la décomposition suivante : $\mathbf{A} = \mathbf{LDU}$ où $\mathbf{L}$ est une matrice triangulaire inférieure avec des éléments diagonaux unitaires, $\mathbf{D}$ une matrice diagonale dont les éléments sont positifs et $\mathbf{U} = \mathbf{L}'$. Si $\mathbf{L}_1 = \mathbf{LD}^{1/2}$, donc $\mathbf{L}_1$ est l'unique matrice triangulaire inférieure dont les éléments diagonaux sont positifs qui satisfait la décomposition de Cholesky $\mathbf{A} = \mathbf{L}_1\mathbf{L}_1'$. Par conséquent, $\mathbf{P} = \mathbf{L}_1^{-1}\mathbf{X}_1$ est une matrice orthogonale unique. Ainsi, pour garantir l'identification du modèle nous allons supposer que $\mathbf{X}$ est de la forme suivante :

$$\mathbf{X} = \begin{pmatrix} x_{11} & 0 & 0 & \dots & 0 \\ x_{21} & x_{22} & 0 & \dots & 0 \\ x_{31} & x_{32} & x_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kk} \\ x_{k+1,1} & x_{k+1,2} & x_{k+1,3} & \dots & x_{k+1,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{q1} & x_{q2} & x_{q3} & \dots & x_{qk} \end{pmatrix} \tag{2.4}$$

où $x_{i,i} > 0$ pour $i = 1, ..., k$ et $x_{i,j} = 0$ pour $i < j$, $i, j = 1, ..., k$. Cette condition nécessite $\frac{1}{2}k(k-1)$ contraintes et permet d'identifier les pondérations et les facteurs qui leurs sont associés. Nous remarquons ici que l'ordre choisi des séries observées dans le vecteur $\mathbf{y}_t$ peut conduire à des problèmes d'interprétation.

2. La deuxième solution est basée sur une transformation des pondérations permettant de satisfaire une contrainte arbitraire telle que

$$\mathbf{X}'\mathbf{D}^{-1}\mathbf{X} \qquad \text{soit diagonale} \tag{2.5}$$

où $\mathbf{D}$ est une matrice diagonale, elle pourrait être l'identité ou même $\mathbf{\Psi}$ et dont les éléments diagonaux doivent être classés par ordre décroissant. Cette solution suppose que les colonnes de $\mathbf{X}$ sont orthogonales relativement à la fonction de poids. Nous remarquons, aussi, que cette contrainte est invariante par échelle et, à l'exception d'un changement des signes des colonnes, $\mathbf{X}$ sera définie d'une manière unique. Dans ce cas, le nombre des restrictions est égale aussi à $\frac{1}{2}k(k-1)$. Cependant, cette solution est beaucoup plus restrictive que la précédente surtout de point de vue interprétation du fait que les colonnes de la matrice des pondérations doivent forcement satisfaire les contraintes d'orthogonalité.[2]

2.3.3 Parcimonie

La structure des corrélations (l'équation (2.2)) entraîne un autre problème d'identification lié au nombre des facteurs communs. En effet, le nombre des éléments distincts de la matrice de variance-covariance des observations $\mathbf{\Sigma}$ est égale à $\frac{1}{2}q(q+1)$, alors que le nombre des paramètres libres dans le modèle est égale à $qk+q$ appartenant à $\mathbf{X}$ et $\mathbf{\Psi}$ respectivement, moins $\frac{1}{2}k(k-1)$ éléments qu'on a déjà fixé par les contraintes (2.4) ou (2.5). Afin d'obtenir une solution unique, la différence d entre le nombre d'équations et le nombre d'inconnus doit être positive.

- Si $d<0$: Il y a beaucoup plus de paramètres que d'équations et il y aura ainsi une infinité de solutions possibles pour $\mathbf{X}$ et $\mathbf{\Psi}$.
- Si $d=0$: Nous pouvons généralement trouver une solution. Toutefois, le modèle aura autant de paramètres que d'équations et de ce fait aucun gain de parcimonie ne sera obtenu.
- Si $d>0$: Il y a plus d'équations que de paramètres. Dans ce cas le modèle à facteurs nous permettra d'une explication plus simple que celle de la matrice de variance-covariance complète concernant le comportement de $\mathbf{y}_t$.

Solution : *L'utilisation de l'une des solutions (2.4) ou (2.5) présentées ci-dessus, revient à imposer des contraintes d'identification sur le nombre des facteurs à retenir vérifiant l'inégalité $d \geq 0$, où*

$$d = \frac{1}{2}q(q+1) - \left[qk + q - \frac{1}{2}k(k-1)\right]$$

[2] La solution des composantes principales pour le modèle à facteurs orthogonaux suppose que $\mathbf{\Psi} \to \mathbf{0}$. Dans ce cas on aura $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}'$ avec $\mathbf{X} = \mathbf{\Gamma}\mathbf{D}^{1/2}$ où $\mathbf{\Gamma}$ est une matrice d'ordre $q \times k$ dont les colonnes sont les vecteurs propres normalisés correspondants aux k plus grandes valeurs propres de $\mathbf{D}$.

TAB. 2.1 – Nombre maximal de facteurs k pour q séries.

q	k max						
01 à 07	0	0	1	1	2	3	3
08 à 14	4	5	6	6	7	8	9
15 à 21	10	10	11	12	13	14	15
22 à 27	15	16	17	18	19	20	21

2.4 L'Approche d'Estimation de Jöreskog

Nous présentons ci-dessous le calcul des estimateurs de maximum de vraisemblance des paramètres du modèle, lorsque le nombre k des facteurs est fixé. Cette approche est inspirée des des travaux de Jöreskog [1967, 1969]. Depuis les années 1940, plusieurs méthodes d'estimation basées essentiellement sur l'analyse de corrélation canonique ont été développées. On cite principalement les travaux de Lawley [1940, 1942, 1943, 1967], Rao [1955], Howe [1955] et Bargmann [1957].

2.4.1 La Fonction de Vraisemblance

On se place ici dans le cadre du modèle standard dans lequel on suppose en outre que les facteurs communs et spécifiques suivent indépendamment des lois normales. Les $\mathbf{y}_t$ suivent alors indépendamment une loi $\mathcal{N}(\theta, \mathbf{\Sigma})$ avec $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}' + \mathbf{\Psi}$. La vraisemblance d'une séquence d'observations $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ sera donc donnée par :

$$\begin{aligned}\mathcal{L}(\Theta/\mathcal{Y}) &= p\left(\mathbf{y}_1, ..., \mathbf{y}_n/\theta, \mathbf{X}, \mathbf{\Psi}\right) \\ &= (2\pi)^{-nq/2}\left|\mathbf{\Sigma}\right|^{-n/2}\exp\left[-\frac{1}{2}\sum_{t=1}^{n}\left(\mathbf{y}_t - \theta\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{y}_t - \theta\right)\right] \\ &= (2\pi)^{-nq/2}\left|\mathbf{\Sigma}\right|^{-n/2}\exp\left[-\frac{n}{2}tr\left(\mathbf{S}\mathbf{\Sigma}^{-1}\right)\right]\end{aligned}$$

où $\mathbf{S} = \frac{1}{n}\sum\limits_{t=1}^{n}\left(\mathbf{y}_t - \theta\right)'\left(\mathbf{y}_t - \theta\right)$ et $\Theta = \{\theta, \mathbf{X}, \mathbf{\Psi}\}$. La maximisation de cette fonction par rapport à θ donne $\widehat{\theta} = \overline{\mathbf{y}}$. Comme nous l'avons déjà mentionné précédemment, et afin de grantir une solution unique pour les paramètres du modèle, on va imposer la contrainte de diagonalité sur la matrice $\mathbf{\Gamma} = \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}$. La fonction à maximiser est, donc, équivalente à

$$\mathcal{L}(\Theta/\mathcal{Y}) = -\frac{n}{2}\left[\log\left|\mathbf{\Sigma}\right| + tr(\mathbf{S}\mathbf{\Sigma}^{-1})\right] \tag{2.6}$$

où bien à la minimisation de la fonction f_k, avec

$$f_k(\mathbf{X}, \mathbf{\Psi}) = \log\left|\mathbf{\Sigma}\right| + tr(\mathbf{S}\mathbf{\Sigma}^{-1}) - \log\left|\mathbf{S}\right| - q \tag{2.7}$$

La minimisation de cette fonction consiste, tout d'abord, à chercher le minimum conditionnel pour une matrice $\mathbf{\Psi}$ connue et, par la suite, le minimum global. La dérivée partielle de f_k par rapport à $\mathbf{X}$ est donnée par :

$$\frac{\partial f_k}{\partial \mathbf{X}} = 2\mathbf{\Sigma}^{-1}(\mathbf{\Sigma} - \mathbf{S})\mathbf{\Sigma}^{-1}\mathbf{X} \tag{2.8}$$

ceci implique que pour une valeur fixée de $\mathbf{\Psi}$, la valeur de $\mathbf{X}$ qui minimise cette fonction doit satisfaire l'égalité

$$\mathbf{\Sigma}^{-1}(\mathbf{\Sigma} - \mathbf{S})\mathbf{\Sigma}^{-1}\mathbf{X} = \mathbf{0} \tag{2.9}$$

En utilisant l'identité (Lawley et Maxwell [1971]),

$$\mathbf{\Sigma}^{-1} = \mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{X}(\mathbf{I} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Psi}^{-1} \tag{2.10}$$

on démontre que

$$(\mathbf{\Sigma} - \mathbf{S})\,\mathbf{\Psi}^{-1}\mathbf{X}\left(\mathbf{I} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}\right)^{-1} = \mathbf{0} \tag{2.11}$$

La pré-multiplication de cette équation par $(\mathbf{I} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X})$ donne

$$(\mathbf{\Sigma} - \mathbf{S})\mathbf{\Psi}^{-1}\mathbf{X} = \mathbf{0} \tag{2.12}$$

ou bien

$$\mathbf{S}\mathbf{\Psi}^{-1}\mathbf{X} = \mathbf{X}(\mathbf{I} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}) \tag{2.13}$$

Finalement, la multiplication à gauche de l'équation (2.13) par $\mathbf{\Psi}^{-\frac{1}{2}}$ donne

$$(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}})(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{X}) = (\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{X})(\mathbf{I} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}) \tag{2.14}$$

Cette dernière équation nous montre que les colonnes de $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{X})$ sont les vecteurs propres de la matrice $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}})$, et les éléments diagonaux de $(\mathbf{I}+\mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X})$ sont les valeurs propres correspondantes. Étant donné que les éléments diagonaux de $\mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}$ sont donnés par les sommes des carrés des éléments des différentes colonnes de la matrice $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{X})$, chaque élément diagonal sera égale à la valeur propre correspondante moins 1. À ce niveau, il faut remarquer que $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}})$ est une matrice de dimension $q \times q$, elle a donc q valeurs et vecteurs propres, cependant k vecteurs seulement sont nécessaires pour déterminer les colonnes de $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{X})$. Dans ce cas, et à condition de considérer seulement des valeurs réelles pour les éléments de $\mathbf{X}$, nous pouvons démontrer que le minimum de f_k pour une matrice $\mathbf{\Psi}$ donnée, sera obtenu lorsque les vecteurs sont choisis de telle façon qu'ils correspondent aux plus grandes valeurs propres.

Dans toute la suite, nous allons désigner par $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq .... \geq \widetilde{\lambda}_q$ les valeurs propres ordonnées de la matrice $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}})$ et par $\widetilde{v}_1, \widetilde{v}_2, ..., \widetilde{v}_k$ les vecteurs propres correspondant aux k plus grandes valeurs propres. Dans ce cas, étant donné que la matrice $(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}})$ est symétrique, les vecteurs $\widetilde{v}_1, \widetilde{v}_2, ..., \widetilde{v}_k$ sont mutuellement orthogonaux. Si on désigne par $\widetilde{\mathbf{\Theta}}$ la matrice diagonale formée par les valeurs $\widetilde{\lambda}_1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_k$ et $\widetilde{\mathbf{\Omega}}$ la matrice formée par les vecteurs $\widetilde{v}_1, \widetilde{v}_2, ..., \widetilde{v}_k$, on aura

$$\widetilde{\mathbf{\Omega}}'\widetilde{\mathbf{\Omega}} = \mathbf{I} \tag{2.15}$$

ce qui implique

$$\mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{X}} = \widetilde{\mathbf{\Omega}}(\widetilde{\mathbf{\Theta}} - \mathbf{I})^{\frac{1}{2}} \tag{2.16}$$

L'estimation de maximum de vraisemblance conditionnelle de $\mathbf{X}$ sera obtenue en multipliant à gauche l'équation (2.16) par $\mathbf{\Psi}^{\frac{1}{2}}$, soit

$$\widetilde{\mathbf{X}} = \mathbf{\Psi}^{\frac{1}{2}}\widetilde{\mathbf{\Omega}}(\widetilde{\mathbf{\Theta}} - \mathbf{I})^{\frac{1}{2}} \tag{2.17}$$

il faut noter ici que, lorsque une ou plusieurs valeurs propres (parmi les k les plus grandes) sont inférieures à un, cette méthode ne donne pas une solution réelle pour $\widetilde{\mathbf{X}}$. Les applications empiriques ont montré que ce problème survient seulement lorsque le nombre de facteurs k est très grand.

Maintenant, nous allons exprimer le minimum conditionnel de f_k en fonction des $(q-k)$ valeurs propres les plus petites. Plus précisément, nous allons démontrer que

$$f_k^*(\mathbf{\Psi}) = -\log(\widetilde{\lambda}_{k+1}\widetilde{\lambda}_{k+2}...\widetilde{\lambda}_q) + (\widetilde{\lambda}_{k+1} + \widetilde{\lambda}_{k+2} + ... + \widetilde{\lambda}_q) - (q-k) \tag{2.18}$$

Pour ce faire, nous calculons tout d'abord le déterminant de $(\mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{\Sigma}}\mathbf{\Psi}^{-\frac{1}{2}})$, soit

$$\begin{aligned}
|\mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{\Sigma}}\mathbf{\Psi}^{-\frac{1}{2}}| &= |\mathbf{\Psi}^{-\frac{1}{2}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' + \mathbf{\Psi})\mathbf{\Psi}^{-\frac{1}{2}}| \\
&= |\mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'\mathbf{\Psi}^{-\frac{1}{2}} + \mathbf{I}| \\
&= |\widetilde{\mathbf{X}}'\mathbf{\Psi}^{-1}\widetilde{\mathbf{X}} + \mathbf{I}| \\
&= \widetilde{\lambda}_1\widetilde{\lambda}_2...\widetilde{\lambda}_k
\end{aligned} \tag{2.19}$$

l'utilisation de la formule (2.10) donne

$$\mathbf{\Psi}^{\frac{1}{2}}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{\Psi}^{\frac{1}{2}} = \mathbf{I} - \mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{X}}\widetilde{\mathbf{\Theta}}^{-1}\widetilde{\mathbf{X}}'\mathbf{\Psi}^{-\frac{1}{2}} \tag{2.20}$$

on a aussi

$$|\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}| = \widetilde{\lambda}_1\widetilde{\lambda}_2 \ldots \widetilde{\lambda}_q \tag{2.21}$$

donc le rapport des déterminants $|\widetilde{\mathbf{\Sigma}}|$ par $|\mathbf{S}|$ donne

$$\begin{aligned}
\frac{|\widetilde{\mathbf{\Sigma}}|}{|\mathbf{S}|} &= \frac{|\mathbf{\Psi}^{-\frac{1}{2}}|}{|\mathbf{\Psi}^{-\frac{1}{2}}|}\frac{|\widetilde{\mathbf{\Sigma}}|}{|\mathbf{S}|}\frac{|\mathbf{\Psi}^{-\frac{1}{2}}|}{|\mathbf{\Psi}^{-\frac{1}{2}}|} \\
&= \frac{|\mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{\Sigma}}\mathbf{\Psi}^{-\frac{1}{2}}|}{|\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}|} \\
&= \frac{\widetilde{\lambda}_1\widetilde{\lambda}_2...\widetilde{\lambda}_k}{\widetilde{\lambda}_1\widetilde{\lambda}_2...\widetilde{\lambda}_q} = \frac{1}{\widetilde{\lambda}_{k+1}\widetilde{\lambda}_{k+2}...\widetilde{\lambda}_q}
\end{aligned} \tag{2.22}$$

et par conséquent

$$\log|\widetilde{\mathbf{\Sigma}}| - \log|\mathbf{S}| = -\log[\widetilde{\lambda}_{k+1}\widetilde{\lambda}_{k+2}\ldots\widetilde{\lambda}_q] \tag{2.23}$$

par la suite, nous calculons $tr[\mathbf{S}\widetilde{\mathbf{\Sigma}}^{-1}]$, soit

$$\begin{aligned} tr[\mathbf{S}\widetilde{\mathbf{\Sigma}}^{-1}] &= tr\left[\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Psi}^{\frac{1}{2}}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{\Psi}^{\frac{1}{2}}\mathbf{\Psi}^{-\frac{1}{2}}\right] \\ &= tr\left[\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Psi}^{\frac{1}{2}}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{\Psi}^{\frac{1}{2}}\right] \\ &= tr\left[\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}\left(\mathbf{I} - \mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{X}}\widetilde{\mathbf{\Theta}}^{-1}\widetilde{\mathbf{X}}'\mathbf{\Psi}^{-\frac{1}{2}}\right)\right] \\ &= tr\left[\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}\right] - tr\left[\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Psi}^{-\frac{1}{2}}\widetilde{\mathbf{X}}\widetilde{\mathbf{\Theta}}^{-1}\widetilde{\mathbf{X}}'\mathbf{\Psi}^{-\frac{1}{2}}\right] \end{aligned} \tag{2.24}$$

en se basant sur l'équation (2.13), nous pouvons démontrer que

$$\begin{aligned} tr(\mathbf{S}\widetilde{\mathbf{\Sigma}}^{-1}) &= tr(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}) - tr(\widetilde{\mathbf{X}}'\mathbf{\Psi}^{-1}\widetilde{\mathbf{X}}) \\ &= \sum_{i=1}^{q}\widetilde{\lambda}_i - \sum_{i=1}^{k}(\widetilde{\lambda}_i - 1) \\ &= \sum_{i=k+1}^{q}\widetilde{\lambda}_i + k \end{aligned} \tag{2.25}$$

Enfin, la substitution des équations (2.23) et (2.25) dans l'équation (2.7) donne (2.18).

2.4.2 Choix des vecteurs propres

Soient M un ensemble formé par k valeurs quelconques parmi $\widetilde{\lambda}_1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_q$, et $\overline{M}$ l'ensemble complémentaire contenant les $(q-k)$ valeurs restantes. Soit $\widetilde{\mathbf{\Omega}}$ la matrice formées par les vecteurs qui correspondent aux valeurs de M. On démontre que

$$f_k^*(\mathbf{\Psi}) = -\log\left(\prod\widetilde{\lambda}_j\right) + \sum\widetilde{\lambda}_j - (q-k) \tag{2.26}$$

où le produit et la somme sont appliqués sur toutes les $\widetilde{\lambda}_j$ de $\overline{M}$.

Si on remplace maintenant $\widetilde{\lambda}_{k+b}$ de l'équation (2.18) par $\widetilde{\lambda}_\alpha$, où $\alpha \leq k$ et $b \geq 1$ et si on suppose en plus que $\widetilde{\lambda}_\alpha$ et $\widetilde{\lambda}_{k+b}$ sont les deux supérieures ou égales à 1, la valeur de $f_k^*(\mathbf{\Psi})$ va nécessairement changer

$$f^\alpha(\mathbf{\Psi}) - f^{k+b}(\mathbf{\Psi}) = (\widetilde{\lambda}_\alpha - \log\widetilde{\lambda}_\alpha) - (\widetilde{\lambda}_{k+b} - \log\widetilde{\lambda}_{k+b}) \tag{2.27}$$

du fait que $\widetilde{\lambda}_\alpha > \widetilde{\lambda}_{k+b}$ et $(x - \log x)$ une fonction croissante et monotone sur l'intervalle $]1, \infty[$, la quantité précédente est toujours positive. Ainsi, l'utilisation de n'importe quel ensemble de $(q-k)$ valeurs propres dans la formule (2.18) autre que les $(q-k)$ valeurs les plus petites, va nécessairement augmenter la valeur de $f_k^*(\mathbf{\Psi})$.

Si les valeurs $\widetilde{\lambda}_{k+1}, \widetilde{\lambda}_{k+2}, ..., \widetilde{\lambda}_q$ sont très proches de 1, (2.18) sera équivalente à

$$\begin{aligned} f_k^*(\mathbf{\Psi}) &= -\sum_{i=k+1}^{q} \log \widetilde{\lambda}_i + \sum_{i=k+1}^{q} \widetilde{\lambda}_i - (q-k) \\ &= -\sum \log \left[1 + (\widetilde{\lambda}_i - 1)\right] + \sum \widetilde{\lambda}_i - (q-k) \\ &= -\sum \left[(\widetilde{\lambda}_i - 1) - \frac{1}{2}(\widetilde{\lambda}_i - 1)^2 + ...\right] + \sum \widetilde{\lambda}_i - (q-k) \\ &\approx \frac{1}{2} \sum \left[\widetilde{\lambda}_i - 1\right]^2 \end{aligned} \tag{2.28}$$

Ainsi la fonction $f_k^*(\mathbf{\Psi})$ paraît comme une mesure de la variation des valeurs propres par rapport à la valeur 1. Les estimations de maximum de vraisemblance seront, donc, obtenues lorsque ces racines seront les plus proches que possible de 1.

Notons enfin que la minimisation de $f_k^*(\mathbf{\Psi})$ nécessite le calcul de ses dérivées partielles. En se basant sur l'équation (2.2), la forme de $\widehat{\mathbf{\Sigma}}^{-1}$ donnée par l'équation (2.10) et l'égalité de l'équation (2.12), nous pouvons démontrer que

$$\frac{\partial f_k^*}{\partial \mathbf{\Psi}} = diag\left[\mathbf{\Psi}^{-1}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' + \mathbf{\Psi} - \mathbf{S})\mathbf{\Psi}^{-1}\right] \tag{2.29}$$

or cette quantité n'est égale à zéro que lorsque les éléments diagonaux de la matrice $(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' + \mathbf{\Psi} - \mathbf{S})$ sont tous nuls étant donné que $\mathbf{\Psi}^{-1}$ est diagonale non nulle, d'où

$$\mathbf{\Psi} = diag\left(\mathbf{S} - \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'\right) \tag{2.30}$$

Cette équation ne donne pas le minimum de f_k par rapport à $\mathbf{\Psi}$ pour une valeur donnée de $\mathbf{X}$, mais tout simplement, une relation qui se vérifie au minimum global absolu de f_k.

Les estimations de maximum de vraisemblance de $\mathbf{X}$ et $\mathbf{\Psi}$ doivent, donc, satisfaire les équations (2.17) et (2.30) ou des équations qui leurs sont équivalentes. La résolution de ces équations peut se faire d'une manière itérative.

2.4.3 Méthode numérique pour le calcul des estimations

Les algorithmes itératifs qui ont été présentés dans la littérature pour la minimisation de f consistent à trouver des estimations $\Theta^{(1)}, \Theta^{(2)}, ...,$ vérifiant

$$f\left(\mathbf{S}, \mathbf{\Sigma}(\Theta^{(r+1)})\right) < f\left(\mathbf{S}, \mathbf{\Sigma}(\Theta^{(r)})\right) \tag{2.31}$$

Pour l'implémentation de cet algorithme il faut tout d'abord choisir une valeur initiale pour Θ, soit $\Theta^{(1)}$. Les itérations seront arrêtées lorsque, par exemple, les valeurs absolues des dérivées du premier ordre de f par rapport aux paramètres seront toutes inférieures à une valeur positive proche de zéro.

L'algorithme itératif est donc de la forme suivante :

$$\Theta^{(r+1)} = \Theta^{(r)} + \alpha_r \left[H_r(\Theta^{(r)}) \right]^{-1} g_r \left(\Theta^{(r)} \right) \tag{2.32}$$

où $\Theta^{(r)}$ est l'ensemble des paramètres de la r *ième* itération ; α_r est un paramètre spécifique à l'itération (avec $0 < \alpha_r \leq 1$) ; $H_r\left(\Theta^{(r)}\right)$ et $g_r\left(\Theta^{(r)}\right)$ sont, respectivement, la matrice Hessienne et le gradient négatif de f évalués en $\Theta^{(r)}$.

Les dérivées du premier et du second ordre de f sont :

$$\frac{\partial f}{\partial \Theta_i} = tr \left[\mathbf{A}(\mathbf{\Sigma} - \mathbf{S})\mathbf{A} \frac{\partial \mathbf{\Sigma}}{\partial \Theta_i} \right] \quad \text{et} \quad \frac{\partial^2 f}{\partial \Theta_i \partial \Theta_j} = tr \left[\frac{\mathbf{A}\partial \mathbf{\Sigma}}{\partial \Theta_i} \frac{\mathbf{A}\partial \mathbf{\Sigma}}{\partial \Theta_j} \right] \tag{2.33}$$

où $\mathbf{A} = \mathbf{\Sigma}^{-1}$ pour le maximum de vraisemblance.

Enfin il faut noter que cette méthode donne parfois des résultats contradictoires tels que, par exemple, des pondérations complexes ou bien des variances spécifiques négatives. Ces problèmes sont essentiellement liés au choix des valeurs d'initialisation de l'algorithme itératif. Pour une révue de littérature, voir Lawley [1942], Rao [1955], Howe [1955], Bargmann [1957], Emmett [1949], Lawley et Maxwell [1971], Lord [1956], Maxwell [1961] et Jöreskog [1967].

2.5 Estimation par les Algorithmes de type EM

L'algorithme EM de Dempster, Laird et Rubin [1977] est une procédure générale pour maximiser la vraisemblance. Elle est adaptée à de nombreuses situations décrites sous forme de problèmes avec données incomplètes. Depuis cet article de référence, de nombreux auteurs ont décrit cet algorithme, ses propriétés et parfois ses variantes, par exemple McLachlan et Krishnan [1997] ou encore, dans le cas des modèles d'analyse factorielle, Rubin et Thayer [1982, 1983]. Dans le cadre des modèles à facteurs que nous avons présentés dans la section 2, les données manquantes correspondent aux facteurs communs supposés non observables. Partant d'un paramètre initial Θ_0, cet algorithme procède en deux étapes successives, l'étape E (pour Expectation) qui consiste à calculer l'espérance de la log-vraisemblance des données complétées conditionnellement aux variables observables et l'étape M (pour Maximization) dont l'objectif est de maximiser cette espérance afin de mettre à jour les paramètres du modèle.

2.5.1 Structure Générale de l'Algorithme

Soit Y, le vecteur aléatoire correspondant aux données observées y, ayant une fonction de densité dénotée $p(y|\theta)$, où $\theta = [\theta_1, ..., \theta_d]'$ est un vecteur de paramètres inconnus dans l'espace Θ. Le vecteur des valeurs observées y est incomplet ; c'est-à-dire que certaines de ses données sont manquantes. Si la situation était idéale, toutes les données seraient présentes. Dans ce cas, ce serait le vecteur x qui serait observé. Mais dans les cas qui nous intéressent, c'est y qui est observé et ce dernier a des valeurs manquantes qui sont contenues dans le vecteur z. Donc si on ajoutait le vecteur z au vecteur y, toutes les données seraient présentes et ainsi, le vecteur x serait formé. La méthode de maximum de vraisemblance consiste à maximiser la quantité

$$\mathcal{L}_c(\theta|y,z) = \log p(y,z|\Theta) \tag{2.34}$$

appelée log-vraisemblance complétée dans le contexte des algorithmes de type EM.

Mais dans ce cas, seulement le vecteur y est observé et donc la log-vraisemblance complétée est une quantité aléatoire qui ne peut pas être maximisée directement. Cependant, si on utilise une "distribution de moyenne" de la forme $q(z/y)$ afin de calculer la moyenne par rapport à z, nous pouvons éliminer la partie aléatoire. Dans ce cas, l'espérance de la log-vraisemblance complétée sera donnée par :

$$\mathbb{E}\left[\mathcal{L}_c(\theta|y,z)\right] = \int_z q(z|y,\theta)\log p(y,z|\theta)dz \tag{2.35}$$

c'est, donc, une quantité déterministe qui dépend de θ.

Comme nous l'avons déjà mentionné ci-dessus, les étapes de l'algorithme EM permettent d'augmenter la vraisemblance des observations. Afin de prouver un tel argument, nous allons démontrer dans un premier temps que pour une "distribution de moyenne" $q(z|y,\theta)$ arbitraire, la log-vraisemblance est minorée, soit

$$\begin{aligned}
\mathcal{L}(\theta|y) &= \log p(y|\theta) \\
&= \log \int_z p(y,z|\theta)dz \\
&= \log \int_z q(z|y,\theta)\frac{p(y,z|\theta)}{q(z|y,\theta)}dz \\
&\geq \int_z q(z|y,\theta)\log\left\{\frac{p(y,z|\theta)}{q(z|y,\theta)}\right\}dz \\
&= \ell(q,\theta)
\end{aligned} \tag{2.36}$$

Dans cette équation nous avons appliqué l'inégalité de Jensen en se basant sur la concavité de la fonction logarithme. Nous remarquons ici que pour une distribution arbitraire $q(z|y,\theta)$, la fonction auxiliaire $\ell(q,\theta)$ est un minorant pour la log-vraisemblance. L'algorithme EM consiste donc à maximiser dans un premier temps, à l'étape *(i+1)*, la fonction $\ell(q,\theta^{(i)})$ par rapport à q, afin de trouver $q^{(i+1)}$, et par la suite à maximiser $\ell(q^{(i+1)},\theta)$ par rapport à θ afin de mettre à jour la valeur de $\theta^{(i)}$. Les itérations de cet algorithme se résument par les 2 étapes suivantes :

Étape E : $q^{(i+1)} = arg\max_q \ell(q,\theta^{(i)})$

Étape M : $\theta^{(i+1)} = arg\max_\Theta \ell(q^{(i+1)},\theta)$

Dans ce cas, l'étape M est équivalente à la maximisation de l'espérance conditionnelle de la log-vraisemblance complétée. En effet, la fonction auxiliaire $\ell(q,\theta)$ peut être écrite sous la forme suivante :

$$\begin{aligned}
\ell(q,\theta) &= \int_z q(z|y,\theta)\log\left\{\frac{p(y,z|\theta)}{q(z|y,\theta)}\right\}dz \\
&= \int_z q(z|y,\theta)\log p(y,z|\theta)dz - \int_z q(z|y,\theta)\log q(z|y,\theta)dz \\
&= \mathbb{E}\left[\mathcal{L}_c(\theta|y,z)\right] - \int_z q(z|y,\theta)\log q(z|y,\theta)dz \qquad (2.37)
\end{aligned}$$

le deuxième terme de cette équation ne dépend pas de θ, donc la maximisation de $\ell(q,\theta)$ revient à maximiser $\mathbb{E}(\mathcal{L}_c(\theta|y,z))$ par rapport à θ.

Nous remarquons aussi qu'au niveau de l'étape E, la maximisation de $\ell(q,\theta^{(i)})$ par rapport à q peut toujours être menée en prenant $q^{(i+1)}(z|y) = p(z|y,\theta^{(i)})$. En effet,

$$\begin{aligned}
\ell(p(z|y,q,\theta^{(i)}),\theta^{(i)}) &= \int_z p(z|y,\theta^{(i)})\log\left\{\frac{p(y,z/\theta)}{p(z|y,\theta^{(i)})}\right\}dz \\
&= \int_z p(z|y,\theta^{(i)})\log p(y|\theta^{(i)})dz \\
&= \log p(y|\theta^{(i)}) \\
&= \mathcal{L}(\theta^{(i)}|y) \qquad (2.38)
\end{aligned}$$

et étant donné que $\mathcal{L}(\theta|y)$ est un majorant pour $\ell(q,\theta^{(i)})$, cette fonction auxiliaire sera maximisée lorsqu'on prend $q(z|y) = p(z|y,\theta^{(i)})$. Au niveau de l'étape E on utilise, donc, cette distribution afin de calculer l'espérance conditionnelle de la log-vraisemblance complétée. Par la suite, au niveau de l'étape M on maximise cette espérance conditionnelle par rapport aux paramètres afin de trouver une nouvelle valeur $\theta^{(i+1)}$. Cette nouvelle valeur $\theta^{(i+1)}$ va nous permettre de mettre à jour la distribution $p(z|y,\theta^{(i+1)})$ que l'on va utiliser dans les prochaines itérations.

Algorithme 1 : Espérance-Maximisation

Répéter

Étape E : $\widehat{q}(z) = \mathbb{E}\left[p(x|\theta)|y,\theta^{(i)}\right]$

Étape M : $\widehat{\theta} = \arg\max_{\Theta} \mathcal{Q}\left(\theta,\theta^{(i)}\right)$

$\widehat{\theta} \longrightarrow \theta^{(i+1)}, \quad i = i+1$

Jusqu'à $\mathcal{L}(\theta^{(i+1)}|y) - \mathcal{L}(\theta^{(i)}|y) < \varepsilon$

Finalement, il faut noter que l'étape M donne des paramètres qui augmentent seulement le minorant de la vraisemblance. Cependant, l'augmentation d'un minorant d'une fonction ne conduit pas nécessairement à une augmentation de la fonction elle même, s'il y a un gap entre les deux. Au niveau de l'étape E ce gap a été rempli par un choix approprié de la distribution q. En effet, pour $q(z|y,\theta) = p(z|y,\theta^{(i+1)})$ on aura :

$$\mathcal{L}(\theta^{(i)}|y) = \ell(q^{(i+1)}, \theta^{(i)}) \tag{2.39}$$

et par conséquent, l'étape M qui augmente $\ell(q^{(i+1)}, \theta)$ va conduire nécessairement à une augmentation de la vraisemblance non complétée $\mathcal{L}(\theta|y)$.

Au delà de ses propriétés théoriques, l'algorithme EM est largement apprécié pour sa simplicité d'implantation, ses itérations généralement peu gourmandes en temps de calcul, le peu de mémoire nécessaire pour le faire fonctionner (il nécessite peu de stockage) et enfin son principe assez naturel heuristiquement. Ces divers points apparaissent lorsqu'on passe en revue chacune de ses deux étapes.

2.5.2 L'Algorithme EM et les Modèles à Facteurs

Dans le cas des modèles à facteurs standards, l'espérance conditionnelle de la log-vraisemblance complétée d'une séquence de n vecteurs d'observations indépendants $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ est donnée par :

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \mathbb{E}\left\{\log p(\mathcal{Y}/\mathbf{f}, \Theta^{(i)})/\mathcal{Y}, \Theta\right\} = \sum_{t=1}^{n} \int p(\mathbf{f}/\mathbf{y}_t, \Theta) \log p(\mathbf{y}_t/\mathbf{f}, \Theta^{(i)})\, d\mathbf{f}$$

où $\Theta^{(i)} = \left\{\mathbf{X}^{(i)}, \theta^{(i)}, \mathbf{\Psi}^{(i)}\right\}$ est l'ensemble des nouveaux paramètres du modèle. Le calcul de cette espérance conditionnelle nécessite la détermination de

- La densité jointe $p(\mathbf{y}_t, \mathbf{f}_t/\Theta)$ des données complétées,
- La densité marginale $p(\mathbf{y}_t/\Theta)$ des données observées et
- La densité conditionnelle $p(\mathbf{f}_t/\mathbf{y}_t; \Theta)$.

Étape E

Cette étape nécessite le calcul des moyennes et des matrices de variance-covariance conditionnelles des facteurs communs. Dans ce qui précède, nous avons déjà démontré que la distribution d'un vecteur d'observations quelconque $\mathbf{y}_t$ est Gaussienne de la forme $\mathcal{N}(\theta\,,\,\mathbf{XX}' + \mathbf{\Psi})$. Nous avons indiqué aussi que la distribution jointe des observations et des facteurs communs est Gaussienne, soit

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{f}_t \end{pmatrix} /\Theta \sim \mathcal{N}\left[\begin{pmatrix} \theta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{XX}' + \mathbf{\Psi} & \mathbf{X} \\ \mathbf{X}' & \mathbf{I}_k \end{pmatrix}\right] \tag{2.40}$$

En se basant sur les propriétés de la loi normale multivariée, on démontre que

$$\mathbf{f}_t/\mathbf{y}_t, \Theta \sim \mathcal{N}\left[\gamma(\mathbf{y}_t - \theta)\,,\, \mathbf{I}_k - \gamma\mathbf{X}\right] \tag{2.41}$$

où $\gamma = \mathbf{X}'(\mathbf{XX}' + \mathbf{\Psi})^{-1}$. Les statistiques exhaustives du premier et second ordre seront donc données par :

$$\begin{aligned} \widetilde{\mathbf{f}}_t &= \mathbb{E}\left(\mathbf{f}_t/\mathbf{y}_t, \Theta\right) = \gamma(\mathbf{y}_t - \theta) & (2.42)\\ \widetilde{\mathbf{R}}_t &= \mathbb{E}\left(\mathbf{f}_t\mathbf{f}_t'/\mathbf{y}_t, \Theta\right) = \mathbf{I}_k - \gamma\mathbf{X} + \widetilde{\mathbf{f}}_t\widetilde{\mathbf{f}}_t' & (2.43) \end{aligned}$$

Étape M

La fonction auxiliaire qu'on cherche à maximiser est donnée par :

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = -\frac{1}{2}\sum_{t=1}^{n}\left[\log|\boldsymbol{\Psi}| + \mathbb{E}\left\{(\mathbf{y}_t - \mathbf{X}\mathbf{f}_t - \theta)'\boldsymbol{\Psi}^{-1}(\mathbf{y}_t - \mathbf{X}\mathbf{f}_t - \theta)/\mathcal{Y}, \Theta\right\}\right] \quad (2.44)$$

la dérivée de l'équation (2.44) par rapport à θ et la résolution des conditions du premier ordre permettent de trouver

$$\begin{aligned} \frac{\partial}{\partial\theta}\mathcal{Q}(\Theta, \Theta^{(i)}) &= \boldsymbol{\Psi}^{-1}\sum_{t=1}^{n}(\mathbf{y}_t - \mathbf{X}\widetilde{\mathbf{f}}_t - \theta) = \mathbf{0}\\ \theta^{(i+1)} &= \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t - \mathbf{X}^{(i)}\widetilde{\mathbf{f}}_t) & (2.45) \end{aligned}$$

La maximisation de (2.44) par rapport à $\mathbf{X}$, après avoir remplacé θ par $\theta^{(i+1)}$, nous permettra de trouver :

$$\frac{\partial}{\partial\mathbf{X}}\mathcal{Q}(\Theta, \Theta^{(i)}) = \boldsymbol{\Psi}^{-1}\sum_{t=1}^{n}\left[\mathbf{y}_t\widetilde{\mathbf{f}}_t' - \theta^{(i+1)}\widetilde{\mathbf{f}}_t' - \mathbf{X}\widetilde{\mathbf{R}}_t\right] = \mathbf{0}$$

$$\mathbf{X}^{(i+1)} = \left[\sum_{t=1}^{n}\mathbf{y}_t\widetilde{\mathbf{f}}_t' - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t\sum_{t=1}^{n}\widetilde{\mathbf{f}}_t'\right]\left[\sum_{t=1}^{n}\widetilde{\mathbf{R}}_t - \frac{1}{n}\sum_{t=1}^{n}\widetilde{\mathbf{f}}_t\sum_{t=1}^{n}\widetilde{\mathbf{f}}_t'\right]^{-1} \quad (2.46)$$

La matrice des pondérations $\mathbf{X}$ et le vecteur des moyennes θ peuvent être estimés simultanément. En effet, si on pose

$$\boldsymbol{\Gamma}_1 = \sum_{t=1}^{n}\mathbf{y}_t\widetilde{\mathbf{f}}_t' \quad , \quad \boldsymbol{\Gamma}_2 = \sum_{t=1}^{n}\widetilde{\mathbf{R}}_t \quad , \quad \zeta_1 = \sum_{t=1}^{n}\mathbf{y}_t \quad \text{et} \quad \zeta_2 = \sum_{t=1}^{n}\widetilde{\mathbf{f}}_t$$

nous pouvons démontrer que :

$$\begin{bmatrix} \mathbf{X}^{(i+1)} & \theta^{(i+1)} \end{bmatrix} = \left[\sum_{t=1}^{n}\begin{bmatrix} \mathbf{y}_t\widetilde{\mathbf{f}}_t' & \mathbf{y}_t \end{bmatrix}\right]\left[\sum_{t=1}^{n}\begin{bmatrix} \widetilde{\mathbf{R}}_t & \widetilde{\mathbf{f}}_t \\ \widetilde{\mathbf{f}}_t' & 1 \end{bmatrix}\right]^{-1} = \begin{bmatrix} \boldsymbol{\Gamma}_1 & \zeta_1 \end{bmatrix}\begin{bmatrix} \boldsymbol{\Gamma}_2 & \zeta_2 \\ \zeta_2' & n \end{bmatrix}^{-1}$$

Le complément Schur $(\boldsymbol{\Gamma}|n)$ de la matrice inversée est

$$(\mathbf{\Gamma}|n) = \mathbf{\Gamma}_2 - \frac{1}{n}\zeta_2\zeta_2'$$

et le premier élément du produit matriciel de l'avant dernière équation est donné par

$$\mathbf{\Gamma}_1(\mathbf{\Gamma}|n)^{-1} - \frac{1}{n}\zeta_1\zeta_2'(\mathbf{\Gamma}|n)^{-1} = \left[\mathbf{\Gamma}_1 - \frac{1}{n}\zeta_1\zeta_2'\right]\left[\mathbf{\Gamma}_2 - \frac{1}{n}\zeta_2\zeta_2'\right]^{-1}$$

ce qui donne exactement l'estimation de maximum de vraisemblance de $\mathbf{X}$. Le deuxième élément du produit matriciel de cette même équation est :

$$-\frac{1}{n}\mathbf{\Gamma}_1(\mathbf{\Gamma}|n)^{-1}\zeta_2 + \frac{1}{n}\zeta_1 + \frac{1}{n^2}\zeta_1\zeta_2'(\mathbf{\Gamma}|n)^{-1}\zeta_2 = \frac{1}{n}\left[\zeta_1 - (\mathbf{\Gamma}_1(\Gamma|n)^{-1} - \frac{1}{n}\zeta_1\zeta_2'(\mathbf{\Gamma}|n)^{-1})\zeta_2\right]$$

soit, donc, l'estimation de maximum de vraisemblance de la moyenne θ.

Finalement, pour la mise à jour de la matrice des variances idiosyncratiques, on maximise la fonction (2.44) par rapport à l'inverse de $\mathbf{\Psi}$, soit

$$\begin{aligned}
\frac{\partial}{\partial\mathbf{\Psi}^{-1}}\,\mathcal{Q}(\Theta,\Theta^{(i)}) &= \frac{1}{2}\sum_{t=1}^{n}\left[\mathbf{\Psi} - \mathbb{E}\left\{(\mathbf{y}_t - \mathbf{X}\mathbf{f}_t - \theta)(\mathbf{y}_t - \mathbf{X}\mathbf{f}_t - \theta)'/\mathcal{Y},\Theta\right\}\right] = \mathbf{0} \\
\mathbf{\Psi}^{(i+1)} &= \frac{1}{n}\sum_{t=1}^{n} diag\left[\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X}^{(i+1)} & \theta^{(i+1)} \end{bmatrix}\begin{bmatrix} \mathbf{y}_t\widetilde{\mathbf{f}}_t' & \mathbf{y}_t \end{bmatrix}'\right] \qquad (2.47)
\end{aligned}$$

Nous remarquons, aussi, que cette nouvelle valeur dépend des valeurs estimées de la moyenne et des pondérations. Cette dernière formule est obtenue après avoir remplacé l'espérance conditionnelle de l'équation (2.47) par :

$$\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X} & \theta \end{bmatrix}\begin{bmatrix} \widetilde{\mathbf{f}}_t\mathbf{y}_t' \\ \mathbf{y}_t' \end{bmatrix} - \begin{bmatrix} \mathbf{y}_t\widetilde{\mathbf{f}}_t' & \mathbf{y}_t \end{bmatrix}\begin{bmatrix} \mathbf{X}' \\ \theta' \end{bmatrix} + \begin{bmatrix} \mathbf{X} & \theta \end{bmatrix}\begin{bmatrix} \widetilde{\mathbf{R}}_t & \widetilde{\mathbf{f}}_t \\ \widetilde{\mathbf{f}}_t' & 1 \end{bmatrix}\begin{bmatrix} \mathbf{X}' \\ \theta' \end{bmatrix}$$

Notons enfin que si on pose $C_{\mathbf{yy}} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t - \overline{\mathbf{y}})(\mathbf{y}_t - \overline{\mathbf{y}})'$ et $\mathbf{\Delta}^{(i)} = \gamma^{(i)}\mathbf{X}^{(i)}$, nous pouvons démontrer, aussi, que :

$$\theta^{(i+1)} = \overline{\mathbf{y}}$$

$$\mathbf{X}^{(i+1)} = C_{\mathbf{yy}}\gamma^{(i)\prime}\left[\gamma^{(i)}C_{\mathbf{yy}}\gamma^{(i)\prime} + \mathbf{\Delta}^{(i)}\right]^{-1}$$

$$\mathbf{\Psi}^{(i+1)} = diag\left[C_{\mathbf{yy}} - C_{\mathbf{yy}}\gamma^{(i)\prime}\left(\gamma^{(i)}C_{\mathbf{yy}}\gamma^{(i)\prime} + \mathbf{\Delta}^{(i)}\right)^{-1}\gamma^{(i)}C_{\mathbf{yy}}\right]$$

2.5.3 Estimation Sous Contraintes

Si l'on souhaite travailler avec une matrice $\mathbf{X}$ définie de façon unique alors il est nécessaire d'imposer certaines restrictions supplémentaires sur la structure de cette dernière. En général, on impose que $\mathbf{X}'\mathbf{X}$ soit une matrice diagonale, ou que $\mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}$ soit une matrice diagonale. Ceci revient à imposer la contrainte que les vecteurs colonnes de $\mathbf{X}$ soient orthogonaux pour la métrique usuelle dans le premier cas, et pour la métrique $\mathbf{\Psi}^{-1}$ dans le deuxième cas.

Une seconde approche consiste à se donner des contraintes a priori sur la matrice $\mathbf{X}$, en nombre suffisant pour qu'elle soit définie de façon unique, mais en choisissant ces contraintes de sorte qu'elles aient une interprétation. Le plus souvent, il s'agit d'imposer que la matrice $\mathbf{X}$ contienne un certain nombre d'éléments nuls, en des positions déterminées par avance, c'est à dire à supposer a priori que certaines variables sont non corrélées avec certains facteurs. Il s'agit de la démarche dite de l'analyse factorielle confirmatoire par opposition à la démarche de l'analyse factorielle dite exploratoire qui n'impose aucune contrainte a priori sur les relations entre facteurs et variables.

Comme nous l'avons déjà cité précédemment, la structure générale qui a été proposée par Geweek et Zhou [1996] et Aguilar et West [2000] en imposant des contraintes sur les éléments de la matrice $\mathbf{X}$ de type $x_{ii} > 0$ pour $i = 1, ..., k$ et $x_{ij} = 0$ pour $i < j$; $i, j = 1, ..., k$ permet de garantir l'existence d'une solution unique. Donc afin d'estimer les paramètres du modèle en tenant compte de ces contraintes, nous pouvons adapter la solution générale proposée par Rubin et Thayer [1982] à notre cas. Le principe de cette solution est basé sur le fait que par conditionnement sur les facteurs latents $\mathbf{f}$, les variables observables $\mathbf{y}_i$, $i = 1, ..., q$ sont indépendantes. Dans ce cas nous pouvons traiter chaque variable séparément, mais en pratique toutes les variables $\mathbf{y}$ ayant une même structure (en termes de zéros a priori au niveau de la matrice $\mathbf{X}$) seront traitées simultanément.

Considérons la *i-ème* variable $\mathbf{y}_i$ avec les coefficients de régression $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{0i})$ où $\mathbf{x}_{0i}$ représente les coefficient nuls qu'on a déjà fixé a priori et $\mathbf{x}_{1i}$ les coefficients qu'on cherche à estimer. Dans ce cas, nous pouvons aussi décomposer les matrices $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})$ et $(C_{\mathbf{yy}}\gamma)$ d'une manière similaire ; soit $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})_{1i}$ et $(C_{\mathbf{yy}}\gamma)_{1i}$ qui correspondent aux facteurs dont les coefficients pour la *i-ème* variable observable sont non nuls. L'estimation de maximum de vraisemblance de $\mathbf{x}_i$ basée sur les statistiques exhaustives que nous avons déjà calculé sera donnée par :

$$\begin{aligned} \mathbf{x}_i^* &= (\mathbf{x}_{1i}^*, \mathbf{x}_{0i}^*) \quad \text{où} \quad \mathbf{x}_{0i}^* = (0, ..., 0) \\ \text{et} \quad \mathbf{x}_{1i}^* &= (C_{\mathbf{yy}}\gamma)_{1i} \left[(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})_{1i} \right]^{-1} \end{aligned} \tag{2.48}$$

et l'estimation de maximum de vraisemblance de $\mathbf{\Psi}$ par

$$\psi_i^* = C_{\mathbf{yy}i} - (C_{\mathbf{yy}}\gamma)_{1i} \left[(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})_{1i} \right]^{-1} (\gamma' C_{\mathbf{yy}})_{1i} \tag{2.49}$$

où $C_{\mathbf{yy}i}$ est le *i-ème* élément diagonal de $C_{\mathbf{yy}}$. Ainsi, $\mathbf{X}^* = \left[\mathbf{x}_1^{*\prime}, \mathbf{x}_2^{*\prime}, ..., \mathbf{x}_q^{*\prime} \right]'$ et $\mathbf{\Psi}^* = diag \left[\psi_1^*, \psi_2^*, ..., \psi_q^* \right]$.

2.5.4 L'Algorithme ECME

L'algorithme EM que nous avons présenté est souvent considéré comme un algorithme convergeant assez lentement (voir, par exemple, Louis [1982], Laird, Lange et Stram [1987], Lange [1995] et McLachlan et Krishnan [1997]). Ce taux de convergence est linéaire au voisinage d'un point stationnaire θ^* de la vraisemblance (voir McLachlan et Krishnan [1997] Chap. 3.9), contrairement à des méthodes de type Newton qui bénéficient d'une convergence quadratique localement. Chaque itération EM correspond à une application g de Θ dans Θ tel que $\theta^{(i+1)} = g(\theta^{(i)})$. Si $\theta^{(i)}$ converge vers un point θ^* et que g est une application continue, alors $\theta^* = g(\theta^*)$. Un développement de Taylor de $g(\theta^*)$ au voisinage de θ^* permet d'écrire

$$\theta^{(i+1)} - \theta^* \approx \mathbf{H}(\theta^*)\left[\theta^{(i)} - \theta^*\right]$$

avec $\mathbf{H}(\theta^*)$ la matrice jacobienne $d \times d$ de $g(\theta)$. Ainsi, une itération de EM est quasiment linéaire au voisinage de la convergence avec matrice de convergence $\mathbf{H}(\theta^*)$. Comme le taux global de convergence est généralement donné par

$$\delta = \lim_{i \longrightarrow \infty} \frac{\|\theta^{(i+1)} - \theta^*\|}{\|\theta^{(i)} - \theta^*\|}$$

pour n'importe quelle norme $\|.\|$ de $\mathbb{R}^d$, il correspond ainsi à la plus grande valeur propre de $\mathbf{H}(\theta^*)$. La vitesse de convergence de EM sera donc dépendante de la valeur de δ, une grande valeur imposant une convergence lente.

Comme cet algorithme peut s'avérer assez lent dans certaines situations, de nombreux auteurs ont récemment proposé des versions modifiées de celui-ci pour accélérer sa convergence tout en préservant la simplicité de ses itérations. Dans ce contexte Liu et Rubin [1998] considèrent, dans le cadre des modèles à facteurs standards, l'algorithme ECME (Expectation Conditional Maximization of Either) qu'ils ont développé en 1994. Dans ECME, l'étape E de EM est inchangée mais l'étape M de EM est remplacée par l'étape CM (Conditional Maximization) qui maximise, au choix en fonction des paramètres, soit l'espérance conditionnelle de la log-vraisemblance complétée comme c'est déjà le cas dans EM, soit directement la log-vraisemblance.

La Structure Générale de l'Algorithme ECME

Soient $x \in \mathcal{X}$ la variable qui désigne les données complétées avec une densité $f(x/\theta)$, $y \in \mathcal{Y}$ une variable désignant les données observées non complétées, où $\theta \in \Theta$ et $y = y(x)$ une surjection de $\mathcal{X}$ vers $\mathcal{Y}$. Si on note, aussi, par $g(y/\theta)$ la densité de y et $K(x/y, \theta)$ la densité conditionnelle de x étant donnée y, on aura

$$g(y/\theta) = \int_{\mathcal{X}(y)} f(x/\theta)dx$$

où $\mathcal{X}(y) = \{x : x \in \mathcal{X},\, y(x) = y\}$ et $f(x/\theta) = g(y/\theta)K(x/y, \theta)$.

L'objectif est de trouver l'estimation de maximum de vraisemblance de θ, $\widehat{\theta}$ qui maximise la log-vraisemblance actuelle donnée par

$$\mathcal{L}(\theta) \equiv \log g(y/\theta) = \mathcal{Q}(\theta/\theta') - \mathcal{H}(\theta/\theta')$$

où $\mathcal{Q}(\theta/\theta') = \mathbb{E}\Big[\log f(x/\theta)/y, \theta'\Big]$ est l'espérance de la log-vraisemblance des données complétées, et $\mathcal{H}(\theta/\theta') = \mathbb{E}\Big[\log K(x/y, \theta)/y, \theta'\Big]$ l'espérance de la log-vraisemblance des données manquantes.

Algorithme 2 : ECME

répéter

Étape E : $\widehat{q}(\mathbf{x}) = \mathbb{E}\Big[p(x)/z, \theta^{(i)}\Big]$

Étape CM 1 : $\widehat{\theta}_1 = \arg\max_{\theta_1} \mathcal{L}\left(\theta, \theta^{(i)}\right)$

Étape CM S : $\widehat{\theta}_s = \arg\max_{\theta_s} \mathcal{L}\left(\theta, \theta^{(i)}\right)$

$\theta^{(i+(s-1)/S)} \longrightarrow \theta^{(i+s/S)}, \quad i = i+1$

Jusqu'à $\mathcal{L}(\Theta^{(i+1)}) - \mathcal{L}(\Theta^{(i)}) < \varepsilon$

L'algorithme EM maximise $\mathcal{L}(\theta)$ à travers une maximisation itérative de $\mathcal{Q}(\theta/\theta')$ par rapport à θ. La *i ème* itération $\theta^{(i)} \rightarrow \theta^{(i+1)}$ de cet algorithme est définie par une espérance, ou une étape E permettant de calculer $\mathcal{Q}(\theta/\theta^{(i)})$ comme fonction de θ, suivie par une maximisation, ou une étape M permettant de trouver $\theta = \theta^{(i+1)}$ en maximisant $\mathcal{Q}(\theta/\theta^{(i)})$. Chacune de ces itérations permet d'augmenter $\mathcal{L}(\theta)$, et d'une manière plus générale si l'algorithme EM converge vers une valeur θ^*, cette valeur sera un maximum local de $\mathcal{L}(\theta)$. L'algorithme ECM (Expectation Conditional Maximization) remplace l'étape M de chaque itération EM par $S > 1$ étapes de maximisations conditionnelles où l'on contraint un certain nombre de paramètres à chaque fois $(h_s(\theta), s = 1, ..., S)$, de façons que l'ensemble de θ ait été estimé à l'issue des S sous-étapes. Meng et Rubin [1993] ont montré que l'ECM est un algorithme EM généralisé. Lorsque la maximisation globale lors de l'étape M n'est pas directement réalisable, on la remplace alors par une maximisation itérative. Dans ce cas il suffit de trouver un $\theta^{(i)}$ tel que

$$\mathcal{Q}(\theta^{(i)}/\theta^{(i-1)}) \geq \mathcal{Q}(\theta^{(i-1)}/\theta^{(i-1)})$$

L'algorithme ECME vient pour remplacer certaines étapes CM de l'algorithme ECM par des étapes qui maximisent la fonction de vraisemblance actuelle $\mathcal{L}(\theta)$ avec des contraintes sur θ. Soit $s \in \mathbb{L}_{\mathcal{Q}} \cup \mathbb{L}_{\mathcal{L}} = \{1, ..., S\}$. L'algorithme ECME est, donc, une approche itérative, $\theta^{(i)} \rightarrow \theta^{(i+1)}$, qui consiste en une étape E permettant de calculer $\mathcal{Q}(\theta/\theta^{(i)})$ et en S étapes de maximisations conditionnelles indexées par s ayant comme input $\theta^{(i+(s-1)/S)}$ et comme output $\theta^{(i+s/S)}$. Pour $s \in \mathbb{L}_{\mathcal{Q}}$, $\mathcal{Q}(\theta^{(i+s/S)}/\theta^{(i)}) \geq \mathcal{Q}(\theta/\theta^{(i)})$ pour tout θ satisfaisant $h_s(\theta) = h_s(\theta^{(i+(s-1)/S)})$ et pour $s \in \mathbb{L}_{\mathcal{L}}$, $\mathcal{L}(\theta^{(i+s/S)}) \geq \mathcal{L}(\theta)$ pour tout θ satisfaisant $h_s(\theta) = h_s(\theta^{(i+(s-1)/S)})$.

L'Algorithme ECME et les Modèles à Facteurs

Dans ce cas cet algorithme consiste à décomposer l'ensemble des paramètres Θ en deux parties, soit $\Theta_1 = \{\theta, \mathbf{X}\}$ et $\Theta_2 = \{\mathbf{\Psi}\}$. Il paraît donc plus facile de maximiser $\mathcal{L}(\Theta)$ par rapport aux éléments de la matrice $\mathbf{\Psi}$ (de dimension q) plutôt que de la maximiser par rapport aux éléments de la matrice $\mathbf{X}$ (de dimension $q \times k$), ou bien par rapport aux éléments de $\mathbf{X}$ et $\mathbf{\Psi}$ simultanément. Ainsi, chaque itération de cet algorithme se décompose en trois étapes, une étape E et deux étapes CM.

Étape E

Cette étape est la même que celle de l'algorithme EM. Il s'agit de calculer l'espérance conditionnelle de la log-vraisemblance complétée par rapport aux observations et à l'estimation actuelle des paramètres (équation (2.44)).

Étape CM1

Cette étape est la même que l'étape M de l'algorithme EM en ce qui concerne l'estimation de la moyenne θ et de la matrice des pondérations $\mathbf{X}$.

Étape CM2

Cette étape consiste à estimer les variances spécifiques $\widehat{\psi}_i$, $i = 1, ..., q$ en maximisant la vraisemblance actuelle $\mathcal{L}(\Theta/\mathcal{Y})$ étant données les valeurs déjà trouvées pour θ et $\mathbf{X}$ au niveau de l'étape CM1. Une telle maximisation pourra se faire en utilisant un algorithme de type Newton-Raphson. La fonction à maximiser est donnée par

$$f(\psi_i) = -\log|\mathbf{X}\mathbf{X}' + \mathbf{\Psi}| - tr\left[C_{\mathbf{yy}}(\mathbf{X}\mathbf{X}' + \mathbf{\Psi})^{-1}\right]$$

Dans le cas où la log-vraisemblance est une fonction quadratique de ψ, la convergence sera obtenue après une itération. Dans le cas où elle est concave et uni-modale, la séquence, $\psi^{(1)}, \psi^{(2)}, ...$ converge vers $\widehat{\mathbf{\Psi}}$. Il est reconnu que cette approche converge en général au voisinage de la solution mais, si la solution initiale en est trop éloignée, une divergence peut advenir. En pratique on utilise cette méthode pour optimiser une solution approchée, suffisamment proche de la solution optimale.

L'algorithme itératif qu'on va utiliser pour la maximisation de la fonction $f(\psi)$ est, donc, donné par la formule suivante :

$$\psi^{(i+1)} = \psi^{(i)} + \left[H_{(i)}(\psi^{(i)})\right]^{-1} g_{(i)}(\psi^{(i)}) \tag{2.50}$$

où $\psi^{(i)}$ est le vecteur des paramètres de la *i-ème* itération ; $H_{(i)}(\psi^{(i)})$ est la matrice Hessienne (matrice des dérivées secondes de f par rapport aux paramètres, évaluée en $\psi^{(i)}$) et $g_{(i)}(\psi^{(i)})$ le gradient négatif de f évalué en $\psi^{(i)}$.

Les dérivées du premier et du second ordre de f sont données par

$$\frac{\partial f(\mathbf{\Psi})}{\partial \psi_i} = -\left[\sigma_{ii} - B_{ii}\right] \quad \text{et} \quad \frac{\partial^2 f(\mathbf{\Psi})}{\partial \psi_i \partial \psi_j} = \sigma_{ij}\left[\sigma_{ij} - 2B_{ij}\right]$$

TAB. 2.2 – Les paramètres de simulation

θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$
1.0000	1.0000	2.0000	1.0000
2.0000	2.0000	3.0000	2.0000
3.0000	3.0000	4.0000	3.0000
4.0000	4.0000	5.0000	4.0000
5.0000	5.0000	6.0000	5.0000
6.0000	6.0000	7.0000	6.0000

où $\mathbf{\Sigma}^{-1} = (\sigma_{ij})$ et $\mathbf{B} = (B_{ij})$ avec $\mathbf{B} = \mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1}$. La stabilité d'un tel processus ne peut être garantie théoriquement surtout dans le cas où q est grand. Plus le nombre de paramètres est important, plus cette stabilité sera difficile à obtenir. En pratique, seulement une ou deux étapes de Newton-Raphson seront largement suffisantes lorsque l'algorithme ECME est très proche de la solution optimale.

2.6 Exemples d'Application

L'analyse empirique que nous allons effectuer au cours de cette section sur les méthodes d'estimation et les critères de choix de modèles sera basée sur deux jeux de données. Dans un premier temps nous allons étudier certaines propriétés des algorithmes que nous avons déjà présentés en se basant sur des simulations. Par la suite, ces algorithmes seront appliqués sur des données financières et plus précisément sur les rendements en excès de certaines devises.

2.6.1 Simulation I

Nous avons appliqué les deux algorithmes, EM et ECME sur des données simulées (un échantillon de 600 observations). Dans ce cas, nous avons adopté une spécification avec $q = 6$ variables observables et $k = 2$ facteurs communs. Pour l'initialisation de la matrice des variances idiosyncratiques $\mathbf{\Psi}$, nous avons suivi la démarche de Jöreskog et Sörbom [1988] en prenant $\mathbf{\Psi}^{(0)} = diag(\psi_1^{(0)}, ..., \psi_6^{(0)})$, où $\psi_i^{(0)} = (1 - (1/2)k/q)\,(1/s_{ii})$ pour $i = 1, ..., 6$ et les s_{ii} sont les éléments diagonaux de la matrice $\mathbf{S} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t - \overline{\mathbf{y}})(\mathbf{y}_t - \overline{\mathbf{y}})'$. Pour la génération des observations $\mathbf{y}_t$, nous avons utilisé les valeurs données dans le tableau 2.2. Les trajectoires simulées de ces séries et leurs distributions empiriques sont représentées dans la figure 2.1.

Les résultats obtenus pour 3 critères de convergence différents et pour une seule replication sont donnés dans le tableau 2.3. Nous remarquons que l'algorithme ECME permet d'accélérer la convergence en réduisant le temps de calcul (en termes de nombre d'iterations pour les 3 critères que nous avons utilisé) tout en gardant la monotonie de convergence. Nous avons calculé aussi la vraisemblance à chaque itération en utilisant les algorithmes ECME et EM et nous avons constaté qu'au niveau des premières itérations, la vraisemblance calculée par l'ECME est toujours supérieur à celle calculée par l'algorithme EM, ce qui justifie notre argument à propos de l'ECME comme un algorithme permettant d'accélérer la convergence (voir la figure 2.2).

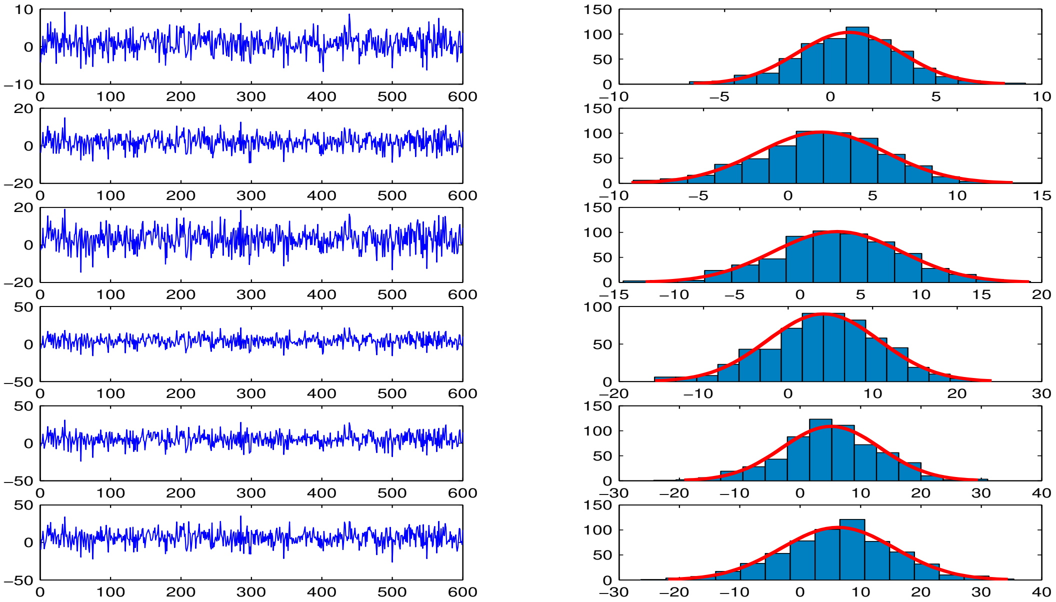

FIG. 2.1 – Les séries d'observations et leurs distributions empiriques

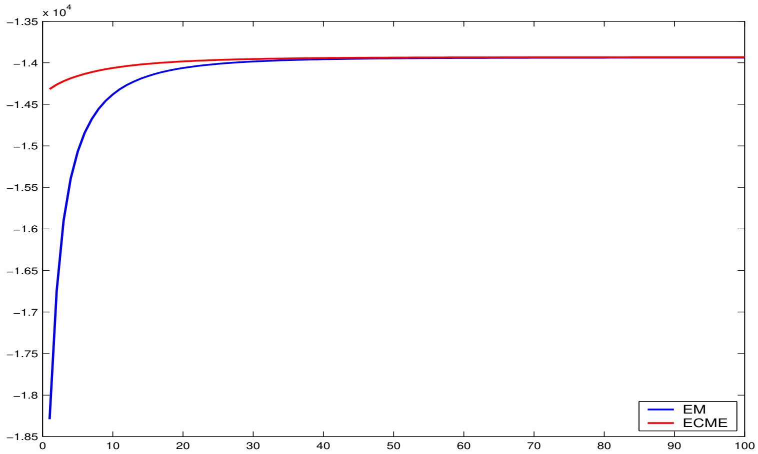

FIG. 2.2 – Les deux fonctions de vraisemblance

Tab. 2.3 – Les résultats des simulations

Méthode	EM	EM	EM	ECME	ECME	ECME
Critère de convergence	100 ité	10^{-5}	10^{-10}	100 ité	10^{-5}	10^{-10}
Nombre d'itérations	100	476	1599	100	183	413
log-vraisemblance	$-1.39.10^4$	$-1.39.10^4$	$-1.38.10^4$	$-1.39.10^4$	$-1.39.10^4$	$-1.38.10^4$
Valeurs initiales pour les x_{ij}	Solutions	Solutions	Solutions	Solutions	Solutions	Solutions
0.5	1.0095	0.9712	0.9660	0.9973	0.9802	0.9784
1	2.0874	2.0520	2.0552	2.0012	2.0534	2.0758
1	3.0837	3.1092	3.1119	3.0735	3.0906	3.1026
1.5	4.0917	4.0983	4.1001	4.0986	4.1097	4.0910
2	5.1363	5.1451	5.1259	5.1407	5.1264	5.1270
3	6.1026	6.1065	6.0961	6.1011	6.0976	6.1174
1	2.0101	2.0012	2.0658	2.0176	2.0122	2.0257
0.8	2.9633	2.9769	2.9942	2.9719	2.9858	2.9837
1.5	3.9246	3.9699	3.9677	3.9267	3.9787	3.9771
2	4.9205	4.9664	4.9549	4.9152	4.9652	4.9641
2.5	5.9444	5.9622	5.9715	5.9511	5.9710	5.9705
3	6.9590	6.9505	6.9436	6.9600	6.9493	7.0894
Valeurs estimées de	0.9316	0.9367	0.9272	0.9387	0.9341	0.9272
la matrice $\mathbf{\Psi}$	2.0047	2.0140	2.0363	2.0063	2.0246	2.0363
	3.0814	3.0908	3.0617	3.0750	3.0911	3.0617
	4.0325	4.0284	4.0081	4.0312	4.0183	4.0081
	4.9107	4.9283	4.9284	4.9291	4.9172	4.9284
	5.9267	5.9252	5.9272	5.9314	5.9449	5.9272
Estimation de						
la moyenne θ	1.0666					
	2.0514					
	3.1391					
	4.0904					
	5.1083					
	5.9689					

2.6.2 Simulation II : Sélection de Modèles

Le problème de sélection de modèles consiste à choisir une structure adéquate, contenant un nombre suffisant de paramètres, permettant d'assurer un ajustement réaliste à l'ensemble de données d'apprentissage. Lorsque le modèle est fixé, la théorie de l'information fournit un cadre rigoureux pour l'élaboration d'estimateurs performants. Mais dans plusieurs situations, les connaissances a priori sur les données ne permettent pas de déterminer un unique modèle dans lequel se placer pour réaliser l'inférence. C'est pourquoi depuis la fin des années 70 les méthodes pour la sélection de modèles à partir des données ont été développées.

Dans la littérature existante, les critères de sélection de modèles traditionnels basés sur la vraisemblance rassemblent une variante de critères tels que le critère de Akaike AIC [1974], le critère de Schwarz [1978] ou critère Bayésien, ou BIC, et les critères d'information qui leurs sont reliés telles que les méthodes ICOMP de Bozdogan et Ramirez [1987] et Bozdogan et Shigemasu [1998]. Ces différents critères respectent les principes fondamentaux du choix d'un modèle : bon ajustement, parcimonie et objectivité. Le critère d'Akaike implique que plus les données sont en grande quantité, plus le modèle retenu sera compliqué. En terme mathématique, la dimension du modèle retenu tend vers l'infini quand le nombre de données fait de même. D'un point de vue pratique, ce critère consiste à minimiser la distance entre les densités de probabilité vraie et estimée des données et il se calcule en soustrayant au χ^2 de vraisemblance deux fois le nombre de degrés de liberté du modèle étudié.

Diverses adaptations de l'AIC sont disponibles. Schwarz [1978] a suggérer le BIC qui augmente l'information sur le nombre de paramètres avec le nombre d'observations. Les deux critères sont équivalents lorsque le nombre de variables à sélectionner, au niveau du modèle, est fixé. Le choix du critère est à l'inverse déterminant lorsqu'il s'agit de comparer des modèles de niveaux différents. Le critère BIC, par exemple, nous pouvons le considérer dans une structure bayésienne comme une approximation de la log-vraisemblance intégrée $\mathcal{L}(\mathcal{Y}) = \int \mathcal{L}(\Theta/\mathcal{Y})\pi(\Theta)d\Theta$, où $\pi(\Theta)$ est une distribution a priori non informative sur le paramètre Θ (voir Kass et Raftery, [1995]). Pour un modèle quelconque $\mathcal{M}$, ce critère est donné par :

$$\mathrm{BIC}(\mathcal{M}) = -2\,\mathcal{L}(\widehat{\Theta}/\mathcal{Y}) + v_{\mathcal{M}} \log n$$

où $v_{\mathcal{M}}$ désigne le nombre de paramètres libres du modèle $\mathcal{M}$. Le critère AIC est basé sur l'utilisation du terme de pénalité le moins rigoureux $2\,v_{\mathcal{M}}$. Le critère le plus performant est le plus bas.

Pour tester l'aptitude de ces deux critères à choisir la spécification convenable, nous avons mis en competition différents modèles à facteurs qui diffèrent par leurs structures cachées. Nous avons donc estimé sur des données simulées quatre modèles à facteurs avec un, deux, trois et quatre facteurs communs respectivement. Dans la première simulation nous avons utilisé $q = 6$ variables observables. Dans ce cas et étant donnée la contrainte de parcimonie trois facteurs au maximum peuvent être retenus. Dans la deuxième simulation, nous avons considéré le cas de $q = 9$ variables observables et où le nombre de facteurs communs ne doit pas dépasser 5.

Première Étude : Dans cette première étude, nous avons considéré un modèle à un seul facteur de dimension six pour la génération de 600 observations (avec 1000 réplications). Ainsi, $q = 6$, $k = 1$ et $n = 600$. Dans chacune des réplications, n observations ont été générées en utilisant les paramètres suivants :

$$\mathbf{X}' = [\ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6\ ] \quad \text{et} \quad diag(\mathbf{\Psi}) = [\ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6\ ]$$

par la suite, nous avons appliqué sur chaque réplication l'algorithme EM standard et sa version conditionnelle ECME afin d'estimer les paramètres de trois spécifications différentes (avec $k = 1$, 2 et 3 facteurs communs) et de calculer les critères de sélection qui leurs correspondent. Le tableau 2.4 donne le nombre de fois qu'un modèle à k-facteurs est choisi par chacun de ces critères. Par exemple en utilisant l'algorithme EM (avec des contraintes sur les pondérations), les critères AIC et BIC sélectionnent toujours le vrai modèle (modèle à un seul facteur). L'utilisation de EM ou de ECME donne exactement le même résultat.

TAB. 2.4 – Résultats de la première Simulation

* Algorithme EM sans contraintes

Critère	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC	1000	0	0	0
BIC	1000	0	0	0

* Algorithme EM avec contraintes

Critère	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC	1000	0	0	0
BIC	1000	0	0	0

Pour présenter la procédure générale d'estimation des modèles à facteurs avec des contraintes sur les pondérations (équations (2.48) et (2.49)), nous allons considérer le cas d'un modèle à 3 facteurs et 6 variables observables. Nous supposons, aussi, que la matrice des pondérations $\mathbf{X}$ a une structure identique à celle donnée par (2.4) ; la première série a des coefficients a priori nuls sur les facteurs 2, et 3 ; la deuxième variable a un seul coefficient a priori nul sur le facteur 3 ; et tous les autres coefficients sont non nuls. En poursuivant la démarche générale de Rubin et Thayer [1982], on va traiter les deux premières variables séparément et les variables restantes simultanément. Pour la première variable $\mathbf{y}_1$, $\mathbf{x}^*_{11}$ est le coefficient correspondant au premier facteur et $\mathbf{x}^*_{01}$ contenant les coefficients nuls associés aux facteurs 2 et 3, $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})_{11}$ est la sous matrice de dimension 1×1 de la matrice $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})$ c'est, donc, l'élément de la première ligne et la première colonne de cette matrice, et $(C_{\mathbf{yy}}\gamma)_{11}$ l'élément de la première ligne et la première colonne de $(C_{\mathbf{yy}}\gamma)$. En ce qui concerne la variable $\mathbf{y}_2$, $\mathbf{x}^*_{12}$ est le vecteur contenant les coefficients associés aux facteurs 1 et 2, alors que $\mathbf{x}^*_{02}$ contient le coefficient nul associé au troisième facteur ; $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})_{12}$ est la sous matrice de dimension 2×2 de la matrice $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})$ et qui consiste en ses deux premières lignes et ses deux premières colonnes, d'autre part $(C_{\mathbf{yy}}\gamma)_{12}$ est la sous matrice de dimension 1×2 de $(C_{\mathbf{yy}}\gamma)$ et qui consiste en sa deuxième ligne et ses deux premières colonnes. Finalement, pour $i = 3, ..., 6$ il n'y a aucune restriction a priori sur les coefficients de pondération. Dans ce cas $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})_{1i}$ est la matrice $(\gamma' C_{\mathbf{yy}}\gamma + \mathbf{\Delta})$ elle même et $(C_{\mathbf{yy}}\gamma)_{1i}$ la sous matrice de dimension 3×3 de $(C_{\mathbf{yy}}\gamma)$ et qui consiste en ses 4 dernières lignes et ses 3 colonnes. En poursuivant la même logique, nous pouvons calculer aussi les éléments de la matrice des variances spécifiques $\mathbf{\Psi}$.

Deuxième Étude : Dans cette deuxième étude, nous avons simulé des données en se basant sur un modèle avec $q = 9$ variables observables, $k = 3$ facteurs communs et un nombre d'observations $n = 800$. Les paramètres de cette simulations sont donnés par : $\theta = diag(\mathbf{\Psi}) = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix}'$ et

$$\mathbf{X}' = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 0 & 3 & 4 & 1 & 6 & 7 & 3 & 2 & 2 \\ 0 & 0 & 1 & 2 & 3 & 8 & 1 & 7 & 3 \end{bmatrix}$$

Les résultats pour 1000 réplications sont donnés dans le tableau 2.5. Nous remarquons que les deux critères sont en faveur du vrai modèle. Les résultats de l'ECME

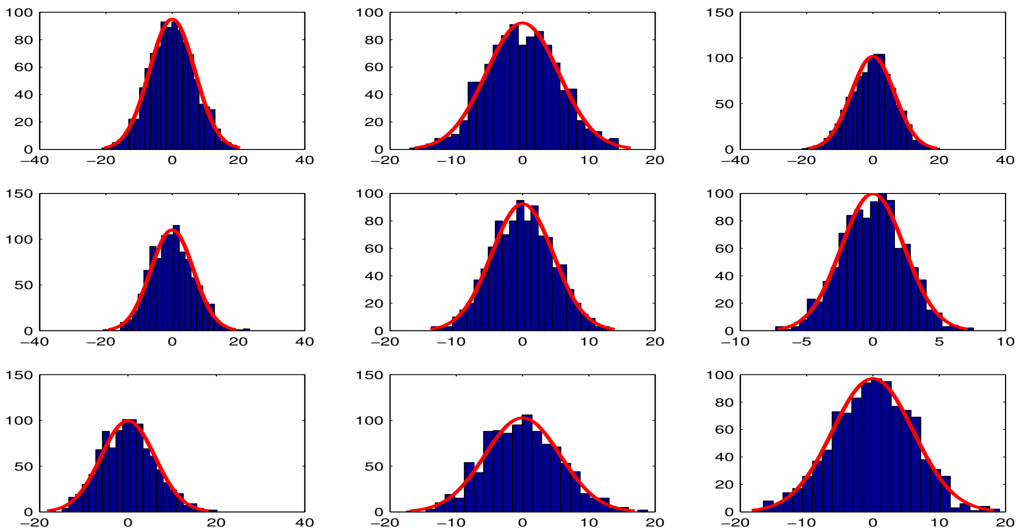

FIG. 2.3 – Les distributions empiriques des erreurs spécifiques de l'estimation d'un modèle à 4 facteurs (avec des contraintes sur les pondérations) sur des données générées par un modèle à un seul facteur commun.

sont, aussi, identiques à ceux de l'EM (avec et sans contraintes).

TAB. 2.5 – Résultats de la deuxième Simulation

* Algorithme EM sans contraintes

Critère	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC	0	0	1000	0
BIC	0	0	1000	0

* Algorithme EM avec contraintes

Critère	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC	0	0	1000	0
BIC	0	0	1000	0

Dans la pratique lorsque le nombre de facteurs k n'est pas conforme avec la structure réelle des données (k est plus ou moins grand), nous pouvons retrouver les problèmes de multi-modalité discutés par Lopes et West [2004]. À titre d'exemple, nous avons estimé un modèle avec un nombre de facteurs $k = 4$ et des contraintes sur les pondérations en utilisant des données générées par un modèle à un seul facteur commun. Dans ce cas, la figure 2.3 montre que les erreurs spécifiques ont des distributions multi-modales.

2.6.3 Application sur les rendements des taux de change

Dans cette section nous allons étudier la structure factorielle de 6 séries de rendements de taux de change. Il s'agit des rendements mensuels des cours en volume (évalués par rapport à la livre sterling) du Dollar Américain (USD), le Dollar Canadien (CAD), le Yen Japonais (JPY), le Franc Français (FRF), la Lire Italienne (ITL) et le Deutsche

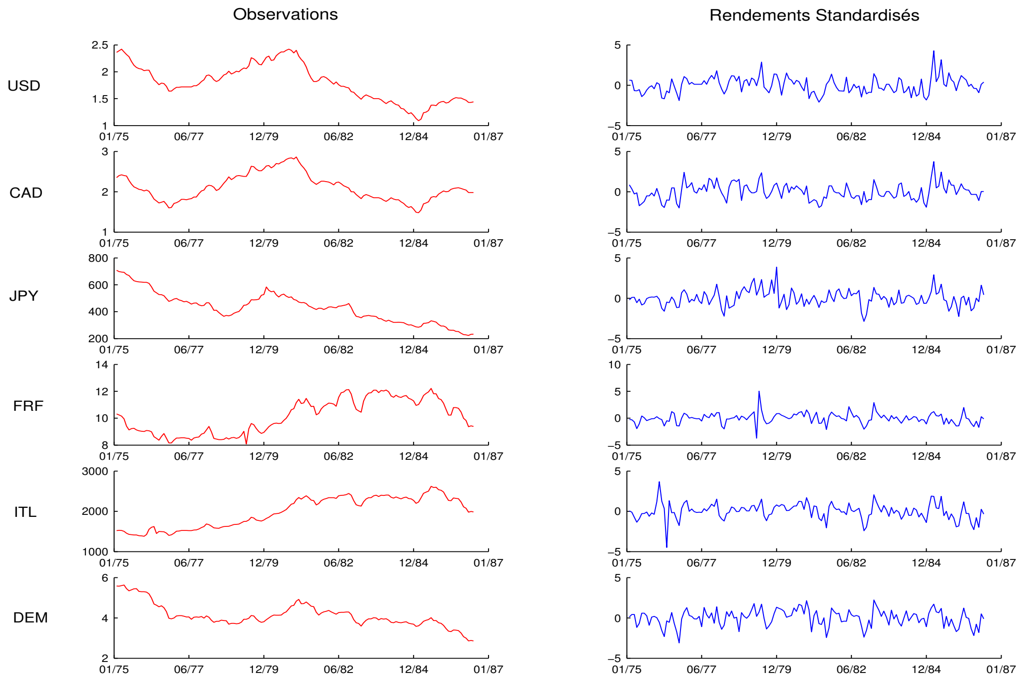

FIG. 2.4 – Les séries réelles et leurs rendements standardisés.

Mark (DEM) [3]. Les données s'étalent sur la période 1/1975 à 12/1986 inclue (voir figure 2.4). Chacune des séries a été standardisée par rapport à sa moyenne et à son écart-type à travers la période d'étude afin de neutraliser l'éventuel effet d'hétéroscédasticité dynamique qui caractérise d'une manière générale les séries à caractère économique ou financier. Les études antérieures menées par West et Harrison [1997] sur cette même base de données (en utilisant la technique d'analyse en composantes principales) ont montré que l'utilisation d'au plus 3 composantes principales est largement suffisante pour expliquer une grande part de la variance totale. En partant de ce résultat et afin de satisfaire aussi la contrainte de parcimonie, les modèles que nous allons estimer (dans toute la suite) ne doivent retenir qu'un nombre de facteurs $k \leq 3$.

Pour le choix du nombre de facteurs adéquat, nous avons appliqué les critères de sélection de modèles AIC et BIC sur des modèles avec $k = 1$, 2 et 3 facteurs communs. L'application de ces critères nécessite tout d'abord l'estimation des paramètres pour différentes valeurs de k. Pour ce faire, nous avons utilisé les algorithmes EM et ECME que nous avons déjà présentés en imposant à chaque fois une structure triangulaire inférieure sur la matrice des pondérations $\mathbf{X}$. Les différentes devises ont été aussi étudiées dans l'ordre donné par la figure 2.4. Enfin, le tableau 2.6 donne les résultats de cette première expérimentation (en utilisant des spécifications avec et sans contraintes) et montre qu'un nombre de facteurs $k = 2$ est largement suffisant pour expliquer toute la corrélation entre les rendements des différentes devises [4]. Les résultats de l'estimation utilisant un algorithme EM pour un modèle à 2 facteurs sont, aussi, donnés dans

[3] PACIFIC EXCHANGE RATE SERVICE, Sauder School of Business, http ://fx.sauder.ubc.ca/.

[4] L'utilisation d'un algorithme ECME avec et sans contraintes donne exactement les mêmes résultats de EM pour les valeurs de AIC et BIC.

TAB. 2.6 – Les valeurs des critères d'information

* Algorithme EM sans contraintes

Critère	$k = 1$	$k = 2$	$k = 3$
AIC	2081.9	**1884.7**	2588.5
BIC	2135.2	**1952.8**	2668.5

* Algorithme EM avec contraintes

Critère	$k = 1$	$k = 2$	$k = 3$
AIC	2081.9	1902.6	1897.5
BIC	2135.2	**1970.8**	1977.5

TAB. 2.7 – Modèle à 2 facteurs avec contraintes sur les pondérations.

θ (10^{-15})	$\mathbf{X}$		$diag(\mathbf{\Psi})$
0.0734	0.9971	0.0000	0.0680
0.0207	0.9771	0.0001	0.1047
−0.1265	0.5247	0.3894	0.6362
0.0699	0.5186	0.6524	0.4089
−0.0194	0.5292	0.5730	0.4841
0.0323	0.5528	0.9249	0.0005

le tableau 2.7.

La représentation graphique des distributions empiriques des erreurs d'estimation dans le cas d'un modèle à 2 facteurs (figure 2.5) montre que ces dernières peuvent être approximées par des distributions Gaussiennes. Les trajectoires des moyennes conditionnelles des deux facteurs sont données dans la figure 2.6. Le premier facteur est représenté avec les séries des rendements USD et CAD et le deuxième facteur avec les rendements de la monnaie Japonaise JPY et les autres monnaies européennes. Pour chacune des séries de rendements $i = 1, ..., 6$, nous avons calculé aussi le pourcentage de la variance conditionnelle expliquée par chaque facteur $j = 1, 2$, soit $100 \left[1 + \frac{x_{ij}^2}{\psi_i}\right]$. Le tableau 2.8 nous donne les valeurs estimées de ces quantités aussi bien que celles du facteur spécifique en utilisant l'algorithme EM contraint.

TAB. 2.8 – Pourcentage de la variance de chacune des séries expliquée par f_1, f_2 et ε.

Devise	Facteur 1	Facteur 2	ε
USD	93.5986	00.0000	06.4014
CAD	90.1155	00.0000	09.8845
JPY	25.8945	14.2636	59.8419
FRF	24.3702	38.5710	37.0588
ITL	25.6386	30.0532	44.3082
DEM	26.3111	73.6477	00.0412

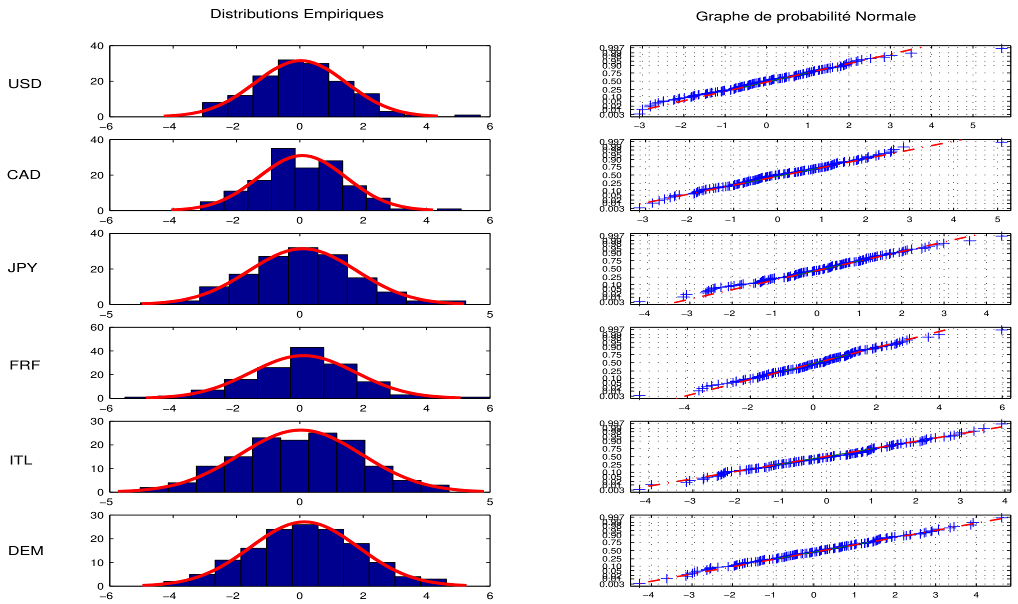

FIG. 2.5 – Les distributions des erreurs d'estimation.

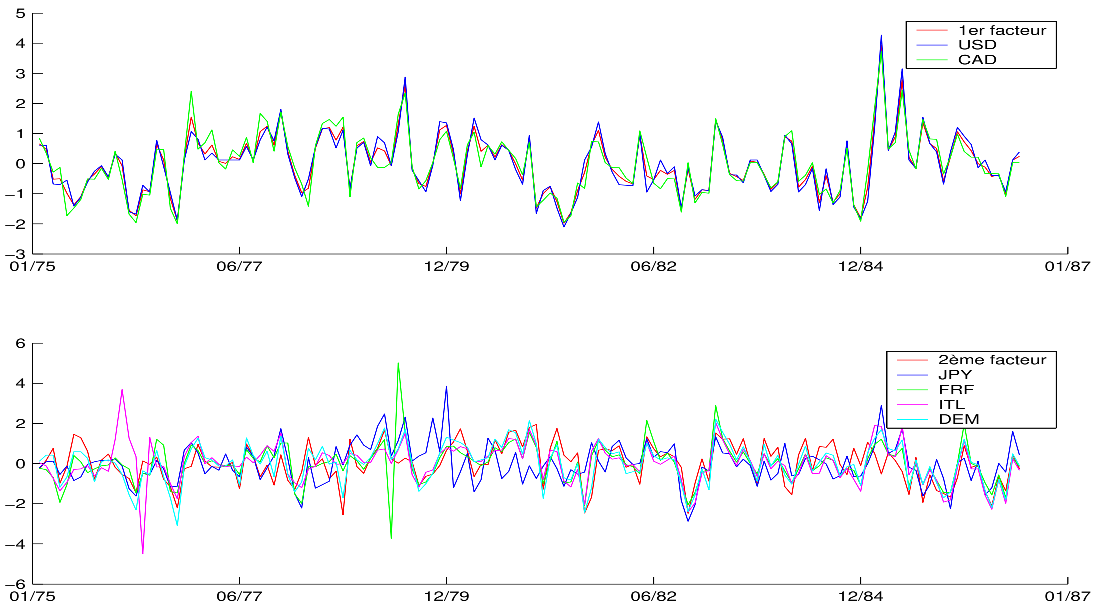

FIG. 2.6 – Moyennes conditionnelles des facteurs et rendements des taux de change. Premier facteur plus USD et CAD (premier graphique) et le deuxième facteur plus JPY, FRF, ITL et DEM (deuxième graphique).

TAB. 2.9 – Modèle à 3 facteurs avec contraintes sur les pondérations.

θ (10^{-15})		$\mathbf{X}$		$diag(\mathbf{\Psi})$
0.0734	1.0043	0.0000	0.0000	0.0620
0.0207	0.9805	0.0434	0.0000	0.1093
−0.1265	0.5320	0.3706	0.2723	0.6024
0.0699	0.5254	0.6990	0.4111	0.2440
−0.0194	0.5374	0.5625	0.3893	0.3965
0.0323	0.5755	0.4798	0.8571	0.0001

Ces résultats montrent que le premier facteur représente la valeur de la Livre Sterling relativement à un panier de devises dans lequel les monnaies de l'Amérique du Nord sont dominantes. Nous remarquons aussi que le Dollar Américain et le Dollar Canadien ont approximativement le même poids : c'est le résultat de l'intégration économique nord-américaine qui a entraîné une certaine harmonisation de l'inflation et des cycles commerciaux dans les deux pays. Ce premier facteur peut, donc, être considéré comme un facteur purement Nord Américain. Le deuxième facteur, par contre, pourra être considéré comme un facteur spécifique aux pays de la communauté économique Européenne. Enfin, nous remarquons que la variabilité des rendements de la monnaie Japonaise est fortement expliquée par des facteurs spécifiques (soit le 2/3 de la variabilité totale). Dans ce cas, un modèle avec $k = 3$ facteurs communs peut éventuellement déplacer une certaine partie de cette variabilité spécifique dans le troisième facteur (que l'on peut appeler, par exemple, facteur Japon). L'estimation d'un modèle à 3 facteurs communs avec des contraintes sur les pondérations donne les résultats du tableau 2.9.

La figure 2.5 nous montre que les distributions des erreurs d'estimation dans le cas d'un modèle à 2 facteurs sont Gaussiennes (des distributions uni-modales). Cependant, la figure 2.7 montre une certaine multi-modalité dans les distributions des erreurs d'estimation d'un modèle à 3 facteurs communs. Ce problème de multi-modalité est le résultat d'une mauvaise spécification du modèle (généralement obtenu lorsque le nombre de facteurs k n'est pas conforme avec la structure réelle des données).

Finalement il faut noter que la structure triangulaire inférieure que nous avons imposé sur la matrice des pondérations (afin de garantir l'identification du modèle) peut conduire à des problèmes d'interprétation des facteurs. L'ordre que nous avons choisi pour les différentes devises dans le vecteur $\mathbf{y}_t$ at-il donc un effet sur l'estimation des paramètres ? Ou bien en d'autres termes : La forte dépendance entre le CAD et le USD est-elle due à ce choix bien particulier ? Pour répondre à cette question, nous avons inter-changé l'ordre du CAD et JPY. Par la suite nous avons estimé un modèle à deux facteurs communs en imposant la même structure de contraintes sur la matrice $\mathbf{X}$. Les résultats de cette estimation sont donnés dans le tableau 2.10. Une comparaison avec les résultats de l'analyse originale (tableau 2.7) montre que l'ordre des variables dans le vecteur $\mathbf{y}_t$ n'a aucun effet sur l'estimation. Ainsi, la matrice de variance-covariance des observations $\mathbf{\Sigma}$ ne sera pas affectée. Dans ce cas le modèle nous donnera les mêmes prévisions pour les $\mathbf{y}_t$ quelque soit l'ordre des y_{it} $(i = 1, ..., q)$.

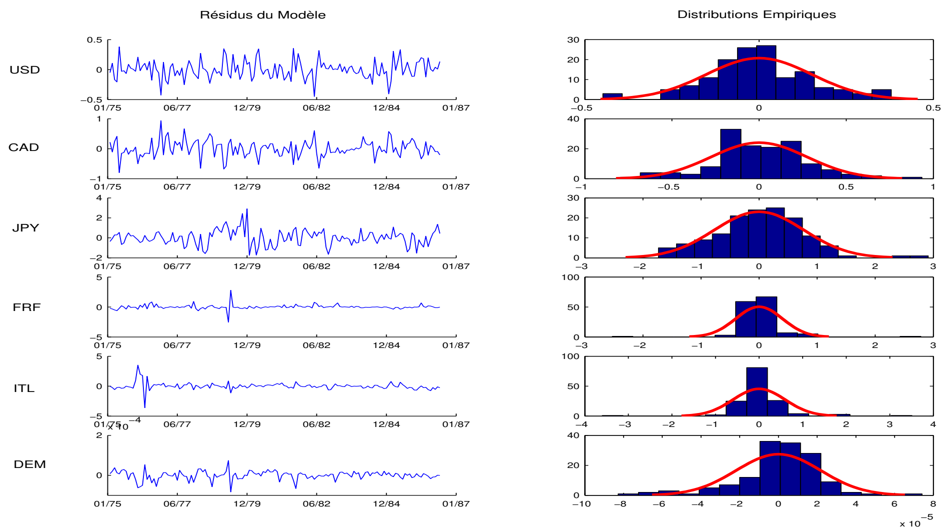

FIG. 2.7 – Modèle à 3 facteurs : Les distributions empiriques des erreurs d'estimation.

TAB. 2.10 – Modèles à 2 facteurs avec contraintes sur les pondérations. Dans ce cas nous avons inter-changé l'ordre de CAD et JPY.

θ (10^{-15})	$\mathbf{X}$		$diag(\mathbf{\Psi})$
0.0734	0.9971	0.0000	0.0680
0.0207	0.5247	0.3894	0.6362
−0.1265	0.9771	0.0001	0.1047
0.0699	0.5186	0.6524	0.4089
−0.0194	0.5292	0.5730	0.4841
0.0323	0.5528	0.9249	0.0005

2.7 Les Modèles à Facteurs Obliques

Dans le but de faciliter l'interprétation des facteurs latents extraits par l'analyse factorielle, il est fortement suggéré de procéder à une rotation de ces facteurs. Rappelons que la décision concernant le choix d'une rotation orthogonale ou oblique donne lieu à des débats assez virulents. Les tenants de la rotation orthogonale soulignent sa simplicité mathématique, alors que les défenseurs de la rotation oblique affirment que seule une rotation oblique est en mesure de bien refléter la réalité des phénomènes étudiés. En parlant de la rotation orthogonale, ces auteurs affirment que de telles solutions sont, la plupart du temps, des représentations naïves et irréalistes des phénomènes étudiés et que tout se ramène à la question suivante : Les aspects que nous postulons à propos d'un construit multidimensionnel sont-ils intercorrélés ? La réponse à cette question est reléguée à un simple statut de supposition lorsque nous employons une rotation orthogonale. Pour cette raison, certains auteurs recommandent vivement de procéder aux deux types de rotation ; si la rotation oblique démontre une corrélation importante entre les dimensions et que cet état de fait correspond à la position théorique entretenue à l'égard du construit étudié, il faut alors privilégier cette solution plus représentative de la réalité. Si par ailleurs, la solution oblique démontre l'absence de corrélation (ou une corrélation négligeable) entre les facteurs, il est alors approprié de se rabattre sur la solution orthogonale plus simple.

Dans la section 2.3.2 nous avons déjà démontré que la matrice des pondérations $\mathbf{X}$ et ainsi les facteurs $\mathbf{f}_t$ peuvent être transformés par une rotation orthogonale sans affecter la distribution des observations (équation (2.3)). Nous pouvons démontrer aussi que la rotation des facteurs peut être effectuée en transformant le modèle à l'aide d'une certaine matrice non singulière. Dans ce cas, on suppose que le modèle à k-facteurs est toujours vérifié pour les n observations $\mathbf{y}_t$ comme dans (2.1) avec la matrice des pondérations $\mathbf{X}^*$ donnée par (2.4) et les facteurs $\mathbf{f}_t^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. Par la suite si on désigne par $\mathbf{H} \neq \mathbf{I}_k$, une matrice définie positive de dimension $(k \times k)$ et que l'on peut décomposer par la formule $\mathbf{H} = \mathbf{LL}'$[5], les nouveaux facteurs seront donc définis par $\mathbf{f}_t = \mathbf{Lf}_t^*$ et la matrice des pondérations correspondante par $\mathbf{X} = \mathbf{X}^*\mathbf{L}^{-1}$. Enfin, la nouvelle spécification pour $\mathbf{y}_t$, $t = 1, ..., n$ sera définie par

$$\mathbf{y}_t = \theta + \mathbf{X}\mathbf{f}_t + \varepsilon_t \quad \text{où} \tag{2.51}$$

$$\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{H}) \quad \text{et} \tag{2.52}$$

$$\mathbf{\Sigma} = \mathbf{XHX}' + \mathbf{\Psi} = \mathbf{X}^*\mathbf{X}^{*'} + \mathbf{\Psi} \tag{2.53}$$

Dans la littérature sur les modèles d'analyse factorielle, cette dernière spécification est appelée **Modèle à Facteurs Obliques**. La décomposition de la matrice de variance-covariance $\mathbf{\Sigma}$ donnée par l'équation (2.2) dans le cas d'une structure orthogonale, montre que les communalités dépendent seulement des éléments diagonaux de la matrice des pondérations $\mathbf{X}$. Cependant, dans les applications réelles (lorsque les facteurs communs peuvent être corrélés) la structure oblique s'avère beaucoup plus intéressante de point de vue prévision et interprétation des résultats. D'après l'équation (2.53), nous remarquons aussi que la décomposition de $\mathbf{\Sigma}$ ne sera pas affectée par cette transformation. Dans ce cas, la source commune de variabilité $\mathbf{XHX}'$ est construite par les contributions des variances des facteurs communs (coefficients de la matrice $\mathbf{H}$) et celles de la nouvelle matrice des pondérations.

Plusieurs méthodes ont été proposées pour l'ajustement de ces modèles. La plus simple est celle qui consiste dans un premier temps à estimer un modèle à facteurs orthogonaux, pour lui appliquer par la suite une transformation convenable. Cette transformation est basée sur un choix bien particulier d'une matrice $\mathbf{H}$ permettant de fournir des facteurs aussi intuitivement significatifs que possible.

Notons enfin qu'une estimation simultanée des paramètres de ce modèle nécessite l'introduction d'autres contraintes sur la matrice des pondérations. Dans ce cas il sera beaucoup plus facile de travailler avec une matrice $\mathbf{H}$ diagonale, qui pourra être obtenue en transformant les pondérations de la spécification orthogonale définie par l'équation

[5] $\mathbf{L}$ pourrait être la décomposition de Cholesky de $\mathbf{H}$ ou la décomposition en valeur singulière $\mathbf{H} = \mathbf{EDE}'$, dans ce cas nous pouvons prendre $\mathbf{L} = \mathbf{ED}^{1/2}$ où $\mathbf{D}$ représente la matrice diagonale des valeurs propres et $\mathbf{E}$ la matrice des vecteurs propres qui lui correspond.

(2.1). En effet, si on désigne par $\mathbf{X}^o = \left[x_{ij}^o\right]$ (pour $i = 1, ..., q$ et $j = 1, ..., k$) la matrice des pondérations de la forme (2.4) ; $\mathbf{f}_t^o$ les facteurs correspondants, où $\mathbf{f}_t^o \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ et par $\mathbf{H}$ la matrice diagonale d'ordre $(k \times k)$ de la forme :

$$\mathbf{H}^{1/2} = diag\left[x_{1,1}^o, x_{2,2}^o, ..., x_{k,k}^o\right]$$

nous pouvons définir des nouvelles pondérations, soit $\mathbf{X} = \mathbf{X}^o\mathbf{H}^{-1/2}$ et des nouveaux facteurs donnés par $\mathbf{f}_t = \mathbf{H}^{1/2}\mathbf{f}_t^o$. Les nouveaux facteurs sont toujours non corrélés mais cette fois-ci, ils ont des variances qui ne sont soumises à aucune contrainte, $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$. Dans ce cas, la nouvelle matrice des pondérations $\mathbf{X}$ aura une structure légèrement différente, donnée par :

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ x_{21} & 1 & 0 & \dots & 0 \\ x_{31} & x_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \dots & 1 \\ x_{k+1,1} & x_{k+1,2} & x_{k+1,3} & \cdots & x_{k+1,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{q1} & x_{q2} & x_{q3} & \dots & x_{qk} \end{pmatrix} \tag{2.54}$$

où $x_{i,i} = 1$ pour $i = 1, ..., k$ et $x_{i,j} = 0$ pour $i < j$, $i, j = 1, ..., k$. En utilisant cette structure, la fonction de vraisemblance ne sera pas affectée et le modèle sera complètement identifié en se basant sur le même nombre de contraintes comme dans le cas orthogonal. Pour calculer les éléments de $\mathbf{H}$, nous pouvons par exemple considérer les éléments positifs de la matrice des pondérations $\mathbf{X}^o$ du modèle orthogonal comme étant les écart-types des facteurs dans un modèle équivalent avec une matrice de covariance diagonale et dont les éléments ne sont soumis à aucune contrainte.

Cette transformation permet aussi d'exprimer le modèle sous une forme tenant compte d'une certaine dynamique au niveau de la variance des facteurs communs. Il s'agit donc d'une généralisation du modèle APT proposé par Ross [1976], en considérant des facteurs à variances variables dans le temps, ou d'une manière équivalente un modèle avec une matrice de pondérations dynamique dans le temps (les covariances ou "betas" des différents actifs avec un facteur particulier changent à travers le temps). Dans ce cas la prime du risque de chacun des actifs varie dans le temps chaque fois où le risque d'un facteur particulier change.

2.8 Conclusion

L'analyse factorielle est la base de la modélisation de ce travail. Elle contient l'idée principale qui a motivé cette recherche, à savoir estimer les facteurs communs qui influencent les données et déterminer une structure linéaire qui reflète cette dépendance. Dans certaines situations, le fait que l'analyse factorielle suppose que les données sont les réalisations d'un certain nombre de variables peut ne pas être approprié car elle ne prend pas en compte leur éventuelle structure temporelle. Dans ce travail on traitera les données comme étant des séries temporelles et donc on devra tenir compte du temps. Cela est valable aussi pour la structure linéaire qui est considérée comme fixe dans l'Analyse factorielle. Au cours de ce travail, on introduira une structure dynamique qui permettra de tenir compte des deux caractéristiques recherchées par ce travail : la structure factorielle conditionnellement hétéroscédastique qui caractérise les séries temporelles à caractère économique ou financier et la paramétrization dynamique. Cette nouvelle spécification pose donc une structure markovienne sur les paramètres du modèle ce qui permet de tenir compte des modifications structurelles des données qui risquent d'arriver au cours du temps.

CHAPITRE 3

Les Modèles à Facteurs Conditionnellement Hétéroscédastiques

Dans ce chapitre, nous étudions une classe de modèles à facteurs conditionnellement hétéroscédastiques. Nous introduisons tout d'abord la structure générale du modèle de base. Par la suite, nous discutons ses propriétés et ses conditions d'identification. Dans une structure espace-état en séries temporelles, on obtient des estimations pour les facteurs communs non observables et leurs variances en utilisant une version modifiée du filtre de Kalman. Un algorithme EM conditionnel sera aussi proposé pour l'estimation de l'ensemble des paramètres du modèle. Finalement, nous présentons trois algorithmes différents permettant de calculer la fonction de vraisemblance, son gradient, et les estimateurs des facteurs qui sont numériquement efficaces et fiables, et statistiquement justifiés.

3.1 Introduction

Les modèles à facteurs dynamiques sont actuellement utilisés dans de nombreux domaines de l'économie. On peut notamment mentionner des exemples en macroéconomie (voir Geweke [1977], Stock et Watson [1989, 1993], Quah et Sargent [1993], Forni, Hallin, Lippi et Reichlin [2004] pour citer des travaux récents), mais aussi en économétrie financière (par exemple Diebold et Nerlove [1989], Engle, Ng et Rothschild [1990], King, Sentana et Wadhwani [1994], Demos et Sentana [1998], Aguilar et West [2000] et Fiorentini, Sentana, et Sephard [2004]). Dans de tels modèles, le terme facteurs viens de l'analyse factorielle. Dans ce cas les variables observées sont supposées dépendre linéairement d'un petit nombre de variables sous-jacentes inobservables, appelées facteurs. La confusion vient plutôt du terme "dynamique" qui a plusieurs interprétations dans la littérature. Il peut caractériser l'évolution des paramètres du modèle qui ne sont plus considérés constants ou bien il peut se référer au fait que les facteurs suivent soit des processus auto régressifs, soit aussi des processus de volatilité purement stochastique ou bien des processus conditionnellement hétéroscédastiques de type ARCH, qui ont été introduits par Engle en 1982 puis généralisés par Bollersev en 1986.

Dans les applications financières et jusqu'à la fin des années 80, ces modèles ont été considérés dans un cadre statique. Ces dernières années plusieurs travaux de recherche, portant essentiellement sur le marché américain, ont montré l'existence de risques idiosyncratiques élevés pour la plupart des actions quelque soit le modèle d'évaluation utilisé (CAPM ou APT). La présence de ces risques idiosyncratiques élevés peut empêcher une évaluation correcte des facteurs générant les rendements, lorsqu'une méthode d'analyse factorielle classique est utilisée. De plus, il est aujourd'hui bien établi que les corrélations entre les rendements ne sont pas stables dans le temps. Pour parvenir à une évaluation correcte des facteurs, différentes spécifications dynamiques ont été proposées. L'idée était, donc, de prendre en compte la majeure partie de l'information contenue dans les distributions des rendements des actions, en utilisant dans la plupart des cas des spécifications conditionnellement hétéroscédastique pour la modélisation de la dynamique des facteurs communs. Il s'agit de modèles introduisant une modélisation explicite de la variance des facteurs, variance qui suit un processus temporel particulier. Ainsi, étant donnée l'information passée, la distribution conditionnelle des facteurs est normale, de moyenne nulle et de variance $\mathbf{H}_t$ elle-même fonction de la variance passée, ce qui permet d'introduire une corrélation non constante entre les rendements et donc de formaliser les phénomènes de persistance et de co-mouvements.

Dans le cadre des modèles dynamiques, où les facteurs communs sont supposés suivre des processus autorégressifs, deux méthodes ont été principalement utilisées pour leur estimation. La première se situe dans le domaine des fréquences et revient à effectuer une décomposition particulière de la densité spectrale du processus vectoriel constitué par l'ensemble des variables étudiées. La seconde se situe dans le domaine des temps et suppose une modélisation de la dynamique des facteurs, puis une estimation par filtre de Kalman.

En ce qui concerne les modèles à facteurs conditionnellement hétéroscédastique, la première approche d'estimation proposée dans la littérature est constituée principalement de trois étapes (voir Kroner [1987] ; Engle, Ng et Rothschild [1990] ; Lin, Engle et Ito [1991] ; Sentana, Shah et Wadhwani [1992] ; King, Sentana et Wadhwani [1994] ; et Kaiser [1997]). La première étape consiste à identifier les facteurs communs moyennant la technique d'analyse en composantes principales appliquée à une approximation de la matrice de corrélation réduite. Dans une deuxième étape, un algorithme de type Newton est utilisé pour estimer les paramètres de la composante conditionnellement hétéroscédastique. Ces nouveaux paramètres seront utilisés par la suite dans une troisième étape afin d'estimer les paramètres du modèle (les moyennes, les pondérations et les variances idiosyncratiques) par le maximum de vraisemblance. Dans ce cas et étant donnée la complexité de calcul engendrée par cette première méthode, une approche itérative basée sur le principe généralisé de l'algorithme EM proposé par Dempster et al. [1977], semble beaucoup plus appropriée. Pour obtenir une estimation des facteurs et par la suite la fonction de vraisemblance, il est utile d'introduire ici une version un peu modifiée du filtre de Kalman appliqué à ce modèle en particulier. Une description plus détaillée se trouve dans la section 3.3 et le chapitre 4 de ce travail.

3.2 Modèle de base et Structure des Facteurs

Ce modèle est inspiré par l'analyse factorielle qui exprime un grand nombre de variables observées comme combinaisons linéaires d'un petit nombre de variables latentes, donc non observés, appelées facteurs. Le modèle qu'on se propose d'étudier prend en compte des séries chronologiques et k facteurs supposés aléatoires et à variances dynamiques qui sont partagés par toutes les variables observées, raison pour laquelle on a intégré le mot conditionnellement hétéroscédastique dans le titre du modèle. Les auteurs precurseurs dans cette littérature sont Engle, Ng et Rothschild [1990] qui ont utilisé cette structure pour la modélisation des bons de trésor. Un modèle similaire a été utilisé, par la suite, par Engle et Ng [1993] pour étudier le comportement dynamique de la structure à terme des taux d'intérêt. Diebold et Nerlove [1989] ont utilisé aussi un modèle conditionnellement hétéroscédastique pour étudier la dynamique des marchés de change. La dynamique et l'intégration des marchés financiers ont été aussi étudiées dans un cadre factorielle conditionnellement hétéroscédastique par Engle et Susmel [1993] et King, Sentana et Wadhwani [1994].

3.2.1 Le Modèle

Considérons le modèle multivarié suivant :

$$\mathbf{y}_t = \mathbf{B}\mathbf{z}_t + \mathbf{X}\mathbf{f}_t + \varepsilon_t \quad \text{où} \quad \mathbf{f}_t = \mathbf{H}_t^{1/2}\mathbf{f}_t^*, \quad \text{et} \tag{3.1}$$

$$\begin{pmatrix} \mathbf{f}_t^* \\ \varepsilon_t \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi} \end{pmatrix}\right] \tag{3.2}$$

où $\mathbf{y}_t$ est un vecteur aléatoire de variables observables de dimension $(q \times 1)$, $\mathbf{z}_t$ est un vecteur de variables exogènes ou de variables explicatives dépendantes retardées, de dimension $(m \times 1)$, $\mathbf{B}$ est la matrice des coefficients de régression associés aux éléments de $\mathbf{z}$ et de dimension $(q \times m)$, $\mathbf{f}_t$ est le vecteur des facteurs communs non observales de dimension $(k \times 1)$, ε_t est le vecteur des erreurs idiosyncratiques de dimension $(q \times 1)$, $\mathbf{X}$ est la matrice des pondérations de dimension $(q \times k)$, avec $k \leq q$ et $rang\,[\mathbf{B}, \mathbf{X}] = m + k$, $\mathbf{\Psi}$ est une matrice semi-définie positive des variances idiosyncratiques supposée constante et de dimension $(q \times q)$, et $\mathbf{H}_t$ une matrice diagonale définie positive de dimension $(k \times k)$ dont les éléments sont les variances des facteurs communs supposées variables dans le temps. En particulier, nous supposons que les variances des facteurs communs suivent des processus GQARCH(1,1). Le *i-ème* élément de la diagonale de cette matrice $\mathbf{H}_t$ est donné par

$$h_{it} = 1 + \gamma_i f_{it-1} + \alpha_i f_{it-1}^2 + \delta_i h_{it-1} \tag{3.3}$$

lorsque $\gamma_i = 0$ on retrouve la spécification GARCH(1,1), si en plus $\delta_i = 0$ on retrouve la spécification ARCH(1) et si tous les coefficients sont nuls ($\gamma_i = \delta_i = \alpha_i = 0$), on retrouve le cas homoscédastique. Dans ce cas la généralisation des processus GQARCH pour un ordre plus élevé ne pose aucun problème de point de vue estimation par l'algorithme EM. Cependant, Harvey, Ruiz et Sentana [1992] ont démontré que généralement l'utilisation du filtre de Kalman pour l'estimation des facteurs communs et leurs variances, en adoptant une spécification de type ARCH non quadratique, ne fournit pas des estimateurs asymptotiquement efficaces. Ce problème peut être levé en considérant un grand nombre de variables observables q (voir Sentana [2004]).

Pour garantir la positivité de la variance des facteurs lors de l'estimation, h_{it} pourra aussi être exprimée sous la forme suivante :

$$h_{it} = 1 + \beta_i \left[f_{it-1} - \mu_i \right]^2 + \delta_i h_{it-1} \tag{3.4}$$

où $\beta_i\,,\,\delta_i > 0\ \forall\, i = 1, ..., k$. Étant donnée que cette spécification est définie à un paramètres d'échelle près, donc pour ramener tous les facteurs communs à la même échelle, nous pouvons considérer soit des facteurs à variances marginales unitaires (voir Sentana [1995]), ou bien aussi comme dans ce cas, en supposant que le premier terme constant de la spécification GQARCH est égale à 1.

Une spécification beaucoup plus générale que celle donnée par [3.1 - 3.2], considère que les erreurs idiosyncratiques ε_t sont des variables qui suivent aussi des processus d'hétéroscédasticité dynamique. Soit $\mathcal{D}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ...\}$, l'ensemble d'informations disponible jusqu'à l'instant $t-1$, la distribution conditionnelle des facteurs communs et spécifiques est gaussienne de la forme suivante :

$$\begin{pmatrix} \mathbf{f}_t \\ \varepsilon_t \end{pmatrix} / \mathcal{D}_{t-1} \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{H}_{t/t-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_{t/t-1} \end{pmatrix} \right]$$

où $\mathbf{H}_{t/t-1}$ est la matrice diagonale définie positive de dimension $(k \times k)$ des variances conditionnelles des facteurs communs, et $\mathbf{\Psi}_{t/t-1}$ la matrice semi-définie positive de dimension $(q \times q)$ des variances conditionnelles des facteurs spécifiques. La forme diagonale de $\mathbf{H}_{t/t-1}$ implique ici que les facteurs sont conditionnellement orthogonaux. Cette hypothèse ajouté à la constance de la matrice des pondérations $\mathbf{X}$, a des implications d'identifiabilité très importantes. Afin d'étudier l'identification des ces modèles aucune restriction supplémentaire sur la forme fonctionnelle de $\mathbf{H}_{t/t-1}$ et $\mathbf{\Psi}_{t/t-1}$ (autre que la restriction d'être mesurable par rapport à $\mathcal{D}_{t-1}$) ne sera ajoutée.

Les hypothèses de ce modèle impliquent que la distribution de $\mathbf{y}_t$ conditionnellement à $\mathcal{D}_{t-1}$ a une moyenne $\mathbf{Bz}_t$ et une matrice de variance-covariance $\mathbf{\Sigma}_{t/t-1}$, soit

$$\mathbf{\Sigma}_{t/t-1} = \mathbf{X}\mathbf{H}_{t/t-1}\mathbf{X}' + \mathbf{\Psi}_{t/t-1}$$

Cette spécification est le cas général de plusieurs modèles étudiés dans la littérature économétrique. Tous ces travaux supposent que les facteurs communs suivent des processus de type ARCH, mais diffèrent par la modélisation des éléments idiosyncratiques. Par exemple, Diebold et Nerlove [1989] ont supposé que la variance de ε_t est constante et diagonale, alors que King, Sentana et Wadhwani [1994] ont retenu la forme diagonale mais avec des éléments dynamiques. D'une manière alternative, le modèle ARCH à facteurs proposé par Engle [1987] suppose que la matrice $\mathbf{\Psi}_t$ est constante, non nécessairement diagonale, mais singulière (voir Nijman et Sentana [1996]). Finalement, il faut noter que si $\mathbf{f}_t$ et ε_t sont conditionnellement homoscédastiques et orthogonaux, le modèle ci-dessus se réduira au modèle à facteurs standard que nous avons déjà présenté dans le chapitre 2. Au contraire, lorsque $\mathbf{f}_t$ et ε_t sont conditionnellement hétéroscédastiques, mais stationnaires au niveau de la covariance, le modèle ci-dessus impliquera une structure à k-facteurs non conditionnelle pour $\mathbf{y}_t$. La matrice de covariance non conditionnelle sera donnée par

$$\mathbf{\Sigma} = \mathbf{XHX}' + \mathbf{\Psi}$$

où $Var(\mathbf{f}_t) = \mathbb{E}(\mathbf{H}_{t/t-1}) = \mathbf{H}$ et $Var(\varepsilon_t) = \mathbb{E}(\mathbf{\Psi}_{t/t-1}) = \mathbf{\Psi}$. La spécification ci-dessus peut être aussi considérée comme l'un des cas particuliers du modèle étudié par Harvey, Ruiz et Sentana [1992], et qui tient compte d'une certaine dynamique au niveau de la moyenne. Ce modèle suppose que la prime du risque associée à chacun des facteurs est aussi variable à travers le temps. Les modèles à facteurs dynamiques ou à tendances communes, aussi bien que les processus ARMA vectoriels et les modèles à facteurs étudiés par Engle, Ng et Rothschild [1990] et King, Sentana et Wadhwani [1994] sont des cas particuliers de cette spécification.

3.2.2 Conditions suffisantes d'identification

Les propriétés statistiques des modèles à facteurs ont été étudiées notamment par Engle [1987] ; Kroner [1987] ; Engle, Ng et Rothschild [1990] ; Harvey, Ruiz et Sentana [1992] ; Lin [1992] ; Bollerslev et Engle [1994] ; Gourieroux, Monfort et Renault [1995] ; et Nijman et Sentana [1996]. Cependant, le problème d'identification du modèle dans le cas où les facteurs communs suivent des processus conditionnellement hétéroscédastiques n'a pas été suffisamment évoqué. La plupart des travaux ont supposé, soit que les facteurs sont connus par avance, soit aussi l'existence d'un seul facteur.

Comme on l'a dit au début du chapitre 2, le but du modèle à facteurs est de donner une description simplifiée des covariances entre les variables, et seulement des covariances. En effet l'écriture $\mathbf{\Sigma} = \mathbf{XX}' + \mathbf{\Psi}$ avec $\mathbf{\Psi}$ diagonale revient à imposer des contraintes seulement sur les termes non diagonaux de $\mathbf{\Sigma}$. Le fait que le modèle à facteurs soit orienté sur une approximation optimale des covariances entre les variables étudiées, lui confère des propriétés d'invariance par changement d'échelle. En particulier, les résultats obtenus en décomposant la matrice de variance-covariance par le modèle à facteurs sont identiques, à un changement d'échelle près, à ceux que l'on obtient en décomposant la matrice de corrélation.

Supposons en effet que $\mathbf{\Sigma} = \mathbf{XX}' + \mathbf{\Psi}$ avec $\mathbf{X}$ matrice de dimension $(q \times k)$ de rang k et $\mathbf{\Psi}$ diagonale, définie positive. Notons $\widetilde{\mathbf{\Sigma}} = diag\,(\mathbf{\Sigma})$ et $\mathbf{R} = \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{\Sigma}\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}$ la matrice de corrélation. On peut alors écrire :

$$\mathbf{R} = \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\left[\mathbf{XX}' + \mathbf{\Psi}\right]\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}} = \mathbf{X}^*\mathbf{X}^{*\prime} + \mathbf{\Psi}^*$$

avec $\mathbf{X}^* = \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{X}$ et $\mathbf{\Psi}^* = \widetilde{\mathbf{\Sigma}}^{-1}\mathbf{\Psi}$, matrice diagonale définie positive. Les matrices $\mathbf{X}$ et $\mathbf{\Psi}$ ne sont donc modifiées que par un changement d'échelle.

En outre, on peut écrire :

$$\mathbf{y}_t = \mathbf{X}\mathbf{f}_t + \varepsilon_t \quad \text{avec} \quad \mathbf{f}_t = \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{y}_t + \nu_t$$

où $\nu_t \sim \mathcal{N}\left[\mathbf{0}, \mathbf{I} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right]$. Il en résulte que les variables $\widetilde{\mathbf{y}}_t = \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{y}_t$ vérifient :

$$\widetilde{\mathbf{y}}_t = \mathbf{X}^*\mathbf{f}_t + \widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\varepsilon_t$$

avec

$$\begin{aligned}
\mathbf{f}_t &= \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{y}_t = \mathbf{X}'\left[\widetilde{\mathbf{\Sigma}}^{\frac{1}{2}}\mathbf{R}\widetilde{\mathbf{\Sigma}}^{\frac{1}{2}}\right]^{-1}\mathbf{y}_t + \nu_t \\
&= \mathbf{X}'\widetilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{R}^{-1}\mathbf{y}_t + \nu_t \\
&= \mathbf{X}^{*\prime}\mathbf{R}^{-1}\widetilde{\mathbf{y}}_t + \nu_t
\end{aligned}$$

et

$$\nu_t \sim \mathcal{N}\left[\mathbf{0}, \mathbf{I} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right] = \mathcal{N}\left[\mathbf{0}, \mathbf{X}^{*\prime}\mathbf{R}^{-1}\mathbf{X}^*\right]$$

Les facteurs communs sont donc inchangés, leur construction en fonction des variables réduites est identique à celle qui a été faite à partir des variables initiales, et les facteurs spécifiques ont été réduits par le même changement d'unité que les variables initiales. En ce sens, on peut parler de l'invariance du modèle à facteurs par changement d'échelle. Ainsi, l'indétermination des facteurs ne peut pas être levée par un simple changement d'échelle. Dans le chapitre 2 nous avons démontré aussi la possibilité de générer un modèle équivalent à [3.1 - 3.2] en se basant sur une transformation orthogonale près des facteurs, soit :

$$\mathbf{y}_t = \mathbf{B}\mathbf{z}_t + \mathbf{X}^*\mathbf{f}_t^* + \varepsilon_t \tag{3.5}$$

où $\mathbf{X}^* = \mathbf{X}\mathbf{Q}'$, $\mathbf{f}_t^* = \mathbf{Q}\mathbf{f}_t$, et $\mathbf{Q}$ une matrice orthogonale arbitraire de dimension $(k \times k)$, et la matrice de covariance non conditionnelle, $\mathbf{X}^*\mathbf{X}^{*\prime} + \mathbf{\Psi}$ reste inchangée. Dans ce cas certaines restrictions de type zéro (matrice triangulaire inférieure) ont été imposées sur $\mathbf{X}$ afin que la seule matrice orthogonale admissible $\mathbf{Q}$ sera l'identité.

Dans le cas où les éléments de $\mathbf{H}_t$ sont dynamiques dans le temps, l'ensemble des matrices orthogonales $\mathbf{Q}$ admissibles sera beaucoup plus petit étant donné que la matrice de variance-covariance conditionnelle $\mathbf{H}_t^*$ des facteurs transformés $\mathbf{f}_t^* = \mathbf{Q}\mathbf{f}_t$ doit rester diagonale $\forall t$. Sans aucune perte de généralité, on va diviser les facteurs en deux groupes, le deuxième groupe, s'il existe, il sera caractérisé pour tout t par une matrice de variance-covariance scalaire (de dimension au moins égale à 2), soit

$$\mathbf{H}_{t/t-1} = \begin{bmatrix} \mathbf{H}_{1t/t-1} & \mathbf{0} \\ \mathbf{0} & h_{2t/t-1}\mathbf{I}_{k_2} \end{bmatrix} \tag{3.6}$$

Si on décompose la matrice $\mathbf{X}$ d'une manière équivalente, soit

$$\mathbf{X} = \left[\mathbf{X}_1 \,|\, \mathbf{X}_2\right] \tag{3.7}$$

nous pouvons établir le résultat suivant :

Proposition : Si $\mathbf{H}_{t/t-1}$ et $\mathbf{X}$ prennent les formes (3.6) et (3.7) et si $Var(\mathbf{f}_{2t}) = h_{2t/t-1}\mathbf{I}_{k_2}$ avec $1 < k_2 < k$, donc $\mathbf{X}_1$ est unique sous n'importe quelle transformation orthogonale (exception faite pour les signes des colonnes).

Preuve : Soit $\mathbf{H}^*_{t/t-1}$ la matrice de variance-covariance des facteurs transformés $\mathbf{f}^*_t = \mathbf{Q}\mathbf{f}_t$, où $\mathbf{Q}$ est une matrice orthogonale arbitraire. Dans ce cas nous pouvons décomposer $\mathbf{Q}$ (par analogie avec (3.6) et (3.7)) sous la forme :

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}$$

où les matrices $\mathbf{Q}_{11}$, $\mathbf{Q}_{12}$, $\mathbf{Q}_{21}$ et $\mathbf{Q}_{22}$ sont, respectivement, de dimension $(k-k_2 \times k-k_2)$, $(k-k_2 \times k_2)$, $(k_2 \times k-k_2)$ et $(k_2 \times k_2)$. Afin de démontrer cette proposition, il suffit de montrer que la seule transformation admissible est donnée par :

$$\widetilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{I}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} \end{bmatrix}$$

où les matrices $\mathbf{I}^{1/2}$, $\mathbf{0}$ et $\mathbf{Q}_{22}$ sont, respectivement, de dimension $(k-k_2 \times k-k_2)$, $(k-k_2 \times k_2)$ et $(k_2 \times k_2)$ avec $\mathbf{I}^{1/2}\mathbf{I}^{1/2} = \mathbf{I}$. Pour ce faire, nous allons décomposer la matrice $\mathbf{H}^*_{t/t-1} = \mathbf{Q}\mathbf{H}_{t/t-1}\mathbf{Q}'$ sous la forme :

$$\mathbf{H}^*_{t/t-1} = \begin{bmatrix} \mathbf{Q}_{11}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{11} + h_{2t/t-1}\mathbf{Q}_{12}\mathbf{Q}'_{12} & \mathbf{Q}_{11}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{21} + h_{2t/t-1}\mathbf{Q}_{12}\mathbf{Q}'_{22} \\ \mathbf{Q}_{11}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{21} + h_{2t/t-1}\mathbf{Q}_{12}\mathbf{Q}'_{22} & \mathbf{Q}_{21}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{21} + h_{2t/t-1}\mathbf{Q}_{22}\mathbf{Q}'_{22} \end{bmatrix}$$

Conditions d'Identification

1. $\mathbf{Q}_{11}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{11}$ diagonale,
2. $\mathbf{Q}_{12}\mathbf{Q}'_{12}$ diagonale,
3. $\mathbf{Q}_{11}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{21}$ nulle,
4. $\mathbf{Q}_{12}\mathbf{Q}'_{22}$ nulle,
5. $\mathbf{Q}_{21}\mathbf{H}_{1t/t-1}\mathbf{Q}'_{21}$ scalaire,
6. $\mathbf{Q}_{22}\mathbf{Q}'_{22}$ scalaire.

Étant donnée la dynamique à travers le temps des éléments de $\mathbf{H}_{1t/t-1}$, et afin que $\mathbf{H}^*_{t/t-1}$ garde sa forme diagonale donnée par l'équation (3.6) pour tout t, les conditions ci-dessus doivent être vérifiées. Dans ce cas si on désigne par q_{21i} la *i-ème* colonne de $\mathbf{Q}_{21}$ et $h_{1it/t-1}$ le *i-ème* élément de la diagonale de $\mathbf{H}_{1t/t-1}$ $(i = 1, ..., k-k_2)$, la condition (5.) pourra donc être réécrite sous la forme :

$$\sum_{i=1}^{k-k_2} h_{1it/t-1} q_{21i} q'_{21i}$$

ici $h_{1it/t-1}$ varie avec i et t en même temps, l'expression donnée par (5.) sera donc scalaire si et seulement si $q_{21i}q'_{21i}$ est scalaire pour tout i. Ceci est équivalent à $q_{21i} = 0$, et ainsi $\mathbf{Q}_{21} = \mathbf{0}$. Dans ce cas la condition (3.) sera aussi vérifiée.

Nous pouvons aussi réécrire la condition (6.) sous la forme : $\mathbf{Q}_{22}\mathbf{Q}'_{22} = \mathbf{I}$, ce qui implique une matrice $\mathbf{Q}_{22}$ orthogonale. Dans ce cas la condition (4.) ne sera vérifiée que si et seulement si $\mathbf{Q}_{12} = \mathbf{0}$, ce qui rend aussi la condition (2.) vérifiée.

Finalement, si on désigne par q_{11i} la *i-ème* colonne de $\mathbf{Q}_{11}$ $(i = 1, k - k_2)$, la condition (1.) pourra aussi être réécrite sous la forme :

$$\sum_{i=1}^{k-k_2} h_{1it/t-1} q_{11i} q'_{11i} \quad \text{diagonale}$$

cette condition ne sera vérifiée que si et seulement si chaque q_{11i} a un seul élément non nul. La propriété de positivité de la variance et l'exclusion des permutations des facteurs impliquent que $\mathbf{Q}_{11}$ doit être (la racine carré de) la matrice unitaire.

Toutefois, il faut noter la généralité de cette proposition étant donné qu'elle est obtenue sans supposer aucune paramétrisation particulière pour l'hétéroscédasticité dynamique des facteurs. Cependant, elle suppose l'orthogonalité conditionnelle des facteurs, la dynamique de leurs variances et la constance de la matrice $\mathbf{X}$. Ainsi dans le cas où il n'y a aucun ou bien un seul facteur conditionnellement homoscédastique, la matrice $\mathbf{X}$ sera identifiée d'une façon unique et sans aucune restriction supplémentaire. En effet, si on suppose que tous les éléments de $\mathbf{H}_{t/t-1}$ sont dynamiques dans le temps (c-à-d $k_2 = 0$), dans ce cas les facteurs transformés $\mathbf{f}_t^* = \mathbf{Q}\mathbf{H}_{t/t-1}^{-1/2}\mathbf{f}_t$ et les matrices de pondérations qui leur sont associées $\mathbf{X}_t^* = \mathbf{X}\mathbf{H}_{t/t-1}^{1/2}\mathbf{Q}'$, permettront de générer la même matrice de covariance conditionnelle pour $\mathbf{y}_t$, mais contrairement au cas homoscédastique, différentes rotations orthogonales seront nécessaires pour chaque instant t. Cependant, les transformations orthogonales (3.5) sont invariantes dans le temps ce qui implique que la matrice $\mathbf{X}$ est identifiable d'une façon unique.

Ce résultat peut aussi être démontré en exprimant le modèle comme un modèle à facteurs conditionnellement homoscédastiques à pondérations variables (voir Engle, Ng et Rothschild [1990]), soit

$$\mathbf{y}_t = \mathbf{B}\mathbf{z}_t + \mathbf{X}_{t/t-1}\widetilde{\mathbf{f}}_t + \varepsilon_t$$

où $Var_{t-1}(\widetilde{\mathbf{f}}_t) = \mathbf{I}_k$ et $\mathbf{X}_{t/t-1} = \mathbf{X}\mathbf{H}_{t/t-1}^{1/2}$. Dans une telle structure, la proposition que nous avons déjà avancé affirme, tout simplement, que les colonnes de $\mathbf{X}$ dont les coefficients de proportionnalité, $h_{jt/t-1}^{1/2}$, sont actuellement variables à travers le temps seront directement identifiables. Sous sa forme actuelle, cette proposition indique alors que la sous identification vient des facteurs ayant une même variance, plutôt que des variances constantes. Lorsque le nombre des facteurs conditionnellement homoscédastiques est au moins égale à 2, certaines restrictions doivent donc être imposées sur la structure des pondérations pour que le modèle soit complètement identifiable.

Afin de tester cette proposition, Sentana [2002] a estimé un modèle à deux facteurs pour étudier l'intégration de 11 marchés financiers européens avec et sans contraintes d'identification ($x_{12} = 0$). Au début il a estimé un modèle à facteurs standards et il a constaté qu'une estimation sans contraintes n'améliore pas les résultats. Au contraire, lorsque la variance du premier facteur était considérée variable dans le temps, l'estimation sans contraintes ($x_{12} \neq 0$) a conduit à une certaine amélioration au niveau de la fonction de vraisemblance.

3.2.3 Représentation Espace-État et Estimation des Facteurs

Pour estimer un modèle structurel à composantes non observables, on a recours à sa représentation espace-état. Ce type de représentation permettra d'extraire les différentes composantes du modèle en utilisant le filtre de Kalman. Dans le cadre d'une modélisation espace-état, une série temporelle est donc générée par un système qui transforme l'information, contenue dans des signaux exogènes présents et passés, en observations futurs. Les "états" du modèle sont autant de résumés de l'information dans le signal exogène, transmise par la dynamique interne qui gouverne la série. Malgré leur attrait, l'utilisation de ces modèles était limitée jusqu'à tout récemment par la contrainte voulant que la distribution des innovations obéisse à une loi normale conditionnelle. Il n'était donc pas possible de modéliser des séries conditionnellement hétéroscédastiques dans un cadre espace-état. Harvey, Ruiz et Sentana [1992] ont levé cette contrainte. Ils ont montré comment le cadre espace-état permet de tenir compte des effets ARCH, que ceux-ci touchent les équations de mesure ou de transition. Pour arriver à leurs fin, les auteurs ont modifié le filtre habituel de Kalman et mis au point un filtre approché (ou quasi-optimal) permettant d'estimer ces modèles.

Le modèle à facteurs conditionnellement hétéroscédastiques [3.1 - 3.2] peut être considéré comme un processus stochastique bidimensionnel (ou un champ aléatoire) avec les indices $i = 1, ..., q$ et $t = 1, ..., n$. Ainsi, nous pouvons l'exprimer par deux représentations différentes : une représentation espace-état en séries temporelles et une représentation espace-état en coupe transversale.

I. Représentation espace-état en séries temporelles

Dans cette représentation, nous considérons les facteurs communs comme une variable d'état. Les équations de mesure et de transition sont, donc, données par :

$$\text{[équation de mesure]} \qquad \mathbf{y}_t = \mathbf{B}\mathbf{z}_t + \mathbf{X}\mathbf{f}_t + \varepsilon_t$$

$$\text{[équation de transition]} \qquad \mathbf{f}_t = \mathbf{0}.\mathbf{f}_{t-1} + \mathbf{f}_t$$

où $\varepsilon_t/\mathcal{D}_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$ et $\mathbf{f}_t/\mathcal{D}_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{H}_t)$ avec $\mathcal{D}_{t-1} = \{\mathcal{Y}_{t-1}, \mathcal{Z}_t, \mathcal{F}_{t-1}\}$; $\mathcal{Y}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ...\}$; $\mathcal{Z}_t = \{\mathbf{z}_t, \mathbf{z}_{t-1}, ...\}$ et $\mathcal{F}_{t-1} = \{\mathbf{f}_{t-1}, \mathbf{f}_{t-2}, ...\}$. Afin de simplifier l'analyse, les paramètres [1] définissant le modèle espace-état sont supposés connus. La question consiste alors à estimer à chaque instant t les variables cachées (le vecteur d'état) conditionnellement aux variables observées jusqu'à la date t (le vecteur de mesure). Dans une première étape nous calculons les 3 prévisions suivantes :

$$\mathbb{E}\left[\mathbf{f}_t/\mathcal{D}_{t-1}\right] = \mathbf{f}_{t/t-1} = \mathbf{0}$$

$$\mathbb{E}\left[\mathbf{y}_t/\mathcal{D}_{t-1}\right] = \mathbf{y}_{t/t-1} = \mathbf{B}\mathbf{z}_t$$

$$Var\left[f_{it}/\mathcal{D}_{t-1}\right] = h_{it/t-1} = 1 + \gamma_i f_{it-1/t-1} + \alpha_i \left[f^2_{it-1/t-1} + h_{it-1/t-1}\right] + \delta_i h_{it-1/t-2}$$

où $h_{it-1/t-1}$ le *i-ème* élément de la diagonale de $\mathbf{H}_{t-1/t-1}$. La prévision consiste donc à rechercher la meilleure approximation de l'état $\mathbf{f}_t$ sachant les observations passées.

Au temps t, on dispose d'une nouvelle observation de $\mathbf{y}$, soit $\mathbf{y}_t$. On peut alors mettre à jour $\mathbf{f}_t$ et sa variance $\mathbf{H}_t$:

$$\begin{aligned}
\mathbf{f}_{t/t} &= \mathbf{f}_{t/t-1} + \mathbf{H}_{t/t-1}\mathbf{X}' \left[\mathbf{X}\mathbf{H}_{t/t-1}\mathbf{X}' + \mathbf{\Psi}\right]^{-1} \left(\mathbf{y}_t - \mathbf{X}\mathbf{f}_{t/t-1} - \mathbf{B}\mathbf{z}_t\right) \\
&= \mathbf{H}_{t/t-1}\mathbf{X}'\mathbf{\Sigma}^{-1}_{t/t-1}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)
\end{aligned}$$

et

$$\begin{aligned}
\mathbf{H}_{t/t} &= \mathbf{H}_{t/t-1} - \mathbf{H}_{t/t-1}\mathbf{X}' \left[\mathbf{X}\mathbf{H}_{t/t-1}\mathbf{X}' + \mathbf{\Psi}\right]^{-1} \mathbf{X}\mathbf{H}_{t/t-1} \\
&= \mathbf{H}_{t/t-1} - \mathbf{H}_{t/t-1}\mathbf{X}'\mathbf{\Sigma}^{-1}_{t/t-1}\mathbf{X}\mathbf{H}_{t/t-1}
\end{aligned}$$

Notons ici que la matrice de variance-covariance $\mathbf{\Sigma}_{t/t-1} = \mathbf{X}\mathbf{H}_{t/t-1}\mathbf{X}' + \mathbf{\Psi}$ a une forme très particulière qui peut être mieux exploitée moyennant la formule de Woodbury. Ainsi, pour inverser la matrice $\mathbf{\Sigma}_{t/t-1}$, de dimension $(q \times q)$, il suffit d'inverser $\mathbf{\Psi}$, et $[\mathbf{H}^{-1}_{t/t-1} + \mathbf{X}'\mathbf{\Psi}\mathbf{X}]$ de dimension $(k \times k)$ seulement. Dans le cas d'un modèle espace-état exacte, ces deux derniers estimateurs sont les estimateurs conditionnellement non biaisés qui minimisent la variance de ceux-ci. Le filtre de Kalman est donc optimal en ce sens qu'il est le meilleur estimateur dans la classe des estimateur linéaires.

La dernière étape est celle du lissage qui consiste à rechercher la meilleure approximation de l'état $\mathbf{f}_t$ sachant les observations passées, présentes et futures $\mathcal{Y}_{1:n}$. Dans ce cas bien particulier et étant donnée la nature dégénérée de l'équation de transition, les équations de lissage seront données par :

$$\mathbf{f}_{t/n} = \mathbf{f}_{t/t} \quad \text{et} \quad \mathbf{H}_{t/n} = \mathbf{H}_{t/t}$$

Une description plus détaillée des algorithmes de filtrage et de lissage (Rauch-Tung-Striebel [1965]) aussi bien que du filtre d'information pour les modèles espace-état linéaires et gaussiens se trouve dans le chapitre 4 de ce travail.

[1] Il s'agit ici principalement des matrices $\mathbf{B}$, $\mathbf{X}$, $\mathbf{\Psi}$, γ_i, α_i et δ_i pour $i = 1, ..., k$.

II. Représentation espace-état en coupe transversale

Dans ce cas et pour t fixé, l'équation de mesure sera donnée par :

$$y_{it} = \mathbf{b}_i'\mathbf{z}_t + \mathbf{x}_i'\mathbf{f}_{it} + \varepsilon_{it} \tag{3.8}$$

où les ε_{it} suivent des lois $\mathcal{N}(0, \psi_i)$ pour tout $i = 1, ..., q$, $\mathbf{x}_i' = [x_{i1}, ..., x_{ik}]$ est la *i-ème* ligne de $\mathbf{X}$, avec $\mathbf{x} = \left[\mathbf{x}_1', ..., \mathbf{x}_q'\right]' = vec\left(\mathbf{X}'\right)$, $\mathbf{b}_i'$ la *i-ème* ligne de la matrice $\mathbf{B}$ et ψ_i le *i-ème* élément de la diagonale de $\mathbf{\Psi}$, telle que $\psi = [\psi_1, ..., \psi_q]' = vecd\left(\mathbf{\Psi}\right)$. Puisque les facteurs sont les mêmes pour toutes les y_{it}, l'équation de transition sera tout simplement donnée par $\mathbf{f}_{it} = \mathbf{f}_{i-1t}$, avec la condition initiale $\mathbf{f}_{0t} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{H}_t\right)$. Ces équations correspondent au modèle à tendances communes en coupe transversale sans innovations dans l'équation de transition (voir Harvey [1989]), et elles permettent de mieux caractériser la dépendance en coupe transversale dans $\mathbf{y}_t$[2].

En se basant sur cette dernière représentation, nous pouvons appliquer le filtre de Kalman transversalement afin d'obtenir à chaque période t "les scores de régression", qui sont les meilleures (dans le sens de l'erreur conditionnelle quadratique moyenne) estimations pour les facteurs, $\mathbf{f}_{t/t} = \mathbb{E}\left(\mathbf{f}_t/\mathcal{Y}_t\right)$, aussi bien que les erreurs quadratiques moyennes associées, $\mathbf{H}_{t/t} = Var\left(\mathbf{f}_t/\mathcal{Y}_t\right)$. En commençant les itérations avec $\mathbf{f}_{0t/0t} = \mathbf{0}$ et $\mathbf{H}_{0t/0t} = \mathbf{H}_t$, les équations de mise à jour seront données par :

$$\begin{aligned} \mathbf{f}_{it/it} &= \mathbf{f}_{i-1t/i-1t} + \delta_{it}^{-1}\mathbf{H}_{i-1t/i-1t}\mathbf{x}_i\eta_{it} \\ \mathbf{H}_{it/it} &= \mathbf{H}_{i-1t/i-1t} - \delta_{it}^{-1}\mathbf{H}_{i-1t/i-1t}\mathbf{x}_i\mathbf{x}_i'\mathbf{H}_{i-1t/i-1t} \end{aligned} \tag{3.9}$$

où

$$\begin{aligned} \eta_{it} &= y_{it} - \mathbf{b}_i'\mathbf{z}_t - \mathbf{x}_i'\mathbf{f}_{i-1t/i-1t} \\ \delta_{it} &= \mathbf{x}_i'\mathbf{H}_{i-1t/i-1t}\mathbf{x}_i + \psi_i \end{aligned} \tag{3.10}$$

sont les erreurs de prévision et leurs variances. Dans cette représentation en coupe transversale, étant donné que $\mathbf{f}_{it/qt} = \mathbf{f}_{qt/qt} = \mathbf{f}_{t/t}$ et $\mathbf{H}_{it/qt} = \mathbf{H}_{qt/qt} = \mathbf{H}_{t/t}$, l'étape de lissage n'est pas nécessaire.

3.3 Estimation de Maximum de Vraisemblance

La méthode du maximum de vraisemblance est à la fois l'une des plus utilisées et des plus controversées en statistique. Elle a en effet un attrait à la fois intuitif, parce que la vraisemblance semble bien contenir toute l'information fournie par les observations, et théorique, à cause des bonnes propriétés asymptotiques des estimateurs correspondants sous certaines conditions de régularité. Dans le cas des modèles à structure cachée et en particulier les modèles à facteurs où les variances communes sont supposées dynamiques dans le temps, une approche de maximum de vraisemblance itérative basée sur le principe de l'algorithme EM généralisé semble beaucoup plus appropriée.

[2] D'une manière alternative, si on considère les vecteurs $\mathbf{f}_{it}$ comme des paramètres et les $\mathbf{x}_i'$ comme des regrésseurs, nous pouvons aussi les interpréter comme une représentation espace-état d'un modèle de régression linéaire pondérée, qui utilise $\mathbf{f}_{0t} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{H}_t\right)$ comme a priori informative.

3.3.1 Les Méthodes d'Optimisation basées sur les Dérivés

Dans ce modèle, les paramètres d'intérêt $\Theta' = \{\mathbf{b}', \mathbf{x}', \psi', \phi'\}$ peuvent, toujours, être estimés en maximisant la log-vraisemblance des variables observables, $\mathbf{y}_t$. Dans ce cas, $rang\left(\mathbf{\Sigma}_t\right) = q$, donc la log-vraisemblance pour n observations (en ignorant les conditions initiales) sera donnée par $\sum_{t=1}^{n} \mathcal{L}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right)$, où

$$\mathcal{L}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) = -\frac{q}{2}\log 2\pi - \frac{1}{2}\log\left|\mathbf{\Sigma}_t\right| - \frac{1}{2}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)'\mathbf{\Sigma}_t^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)$$

avec $\mathbf{\Sigma}_t = \mathbf{X}\mathbf{H}_t\mathbf{X}' + \mathbf{\Psi}$ et $\mathbf{H}_t = diag\left[\mathbf{h}_t\left(\phi\right)\right]$. Étant donnée la non linéarité du modèle, une approche d'optimisation numérique est nécessaire pour le calcul des dérivées du premier ordre et l'estimation du maximum de vraisemblance des paramètres. Mais dans ce cas-ci nous pouvons également obtenir une expression analytique pour le score. En effet, la fonction score (voir Bollerslev et Wooldridge [1992]) $\ell\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) = \partial\mathcal{L}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right)/\partial\Theta$ pour tous modèles conditionnellement gaussiens de moyenne μ_t et de matrice de variance-covariance $\mathbf{\Sigma}_t$, est donnée par :

$$\begin{gathered}\ell\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) = \frac{\partial\mu_t'}{\partial\Theta}\mathbf{\Sigma}_t^{-1}\left(\mathbf{y}_t - \mu_t\right) + \\ \frac{1}{2}\frac{\partial\, vec'\left[\mathbf{\Sigma}_t\right]}{\partial\Theta}\left[\mathbf{\Sigma}_t^{-1}\otimes\mathbf{\Sigma}_t^{-1}\right] vec\left[\left(\mathbf{y}_t - \mu_t\right)\left(\mathbf{y}_t - \mu_t\right)' - \mathbf{\Sigma}_t\right]\end{gathered}$$

Dans ce cas $\mu_t = \mathbf{B}\mathbf{z}_t$ et le différentiel de $\mathbf{\Sigma}_t$ est donné par :

$$d\left[\mathbf{X}\mathbf{H}_t\mathbf{X}' + \mathbf{\Psi}\right] = \left[d\mathbf{X}\right]\mathbf{H}_t\mathbf{X}' + \mathbf{X}\left[d\mathbf{H}_t\right]\mathbf{X}' + \mathbf{X}\mathbf{H}_t\left[d\mathbf{X}'\right] + d\mathbf{\Psi}$$

Les trois termes du Jacobien qui correspondent à $\mathbf{x}$, ψ et ϕ seront donnés par :

$$\begin{aligned}\frac{\partial\, vec\left[\mathbf{\Sigma}_t\right]'}{\partial\,\mathbf{x}} &= \left[\mathbf{I} + \mathbf{K}_{qq}\right]\left[\mathbf{I}\otimes\mathbf{X}\mathbf{H}_t\right] \\ \frac{\partial\, vec\left[\mathbf{\Sigma}_t\right]'}{\partial\,\phi} &= \left[\mathbf{X}\otimes\mathbf{X}\right]\mathbf{E}_k\frac{\partial\,\mathbf{h}_t\left(\Theta\right)'}{\partial\phi} \\ \frac{\partial\, vec\left[\mathbf{\Sigma}_t\right]'}{\partial\,\psi} &= \mathbf{E}_q\end{aligned}$$

où $\mathbf{E}_n$ est l'unique matrice de "diagonalisation" de dimension $n^2 \times n$ qui transforme $vec(\mathbf{A})$ en $vecd(\mathbf{A})$, soit $vecd(\mathbf{A}) = \mathbf{E}_n' vec(\mathbf{A})$, et $\mathbf{K}_{mn}$ la matrice de commutation d'ordres m et n (voir Magnus et Neudecker [1988]).

Après quelques transformations algébriques, on obtient :

$$\begin{aligned}\ell_{\mathbf{b}}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) &= vec\left[\mathbf{\Sigma}_t^{-1}\mathbf{y}_t\mathbf{z}_t' - \mathbf{\Sigma}_t^{-1}\mathbf{B}\,\mathbf{z}_t\mathbf{z}_t'\right] \\ \ell_{\mathbf{x}}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) &= vec\left[\mathbf{H}_t\mathbf{X}'\mathbf{\Sigma}_t^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)'\mathbf{\Sigma}_t^{-1} - \mathbf{H}_t\mathbf{X}'\mathbf{\Sigma}_t^{-1}\right] \\ \ell_{\psi}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) &= \frac{1}{2}\, vecd\left[\mathbf{\Sigma}_t^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)'\mathbf{\Sigma}_t^{-1} - \mathbf{\Sigma}_t^{-1}\right] \\ \ell_{\phi}\left(\mathbf{y}_t/\mathcal{D}_{t-1};\Theta\right) &= \frac{1}{2}\frac{\partial\,\mathbf{h}_t'(\phi)}{\partial\phi}\, vecd\left[\mathbf{X}'\mathbf{\Sigma}_t^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)'\mathbf{\Sigma}_t^{-1}\mathbf{X} - \mathbf{X}'\mathbf{\Sigma}_t^{-1}\mathbf{X}\right]\end{aligned}$$

3.3.2 Les Cas Heywood

Pour effectuer une estimation des paramètres $\Theta = \{\mathbf{B}, \mathbf{X}, \mathbf{\Psi}, \alpha_i, \gamma_i, \delta_i, \, i = 1, ...k\}$ et afin d'obtenir une solution valable, en utilisant la méthode du maximum de vraisemblance, il faut tout d'abord imposer certaines restrictions sur ces paramètres avant la résolution des conditions du premier ordre : $\sum_{t=1}^{n} \ell \left(\mathbf{y}_t / \mathcal{D}_{t-1}, \Theta\right) = \mathbf{0}$. Il s'agit de contraintes permettant de garantir une valeur positive pour les variances idiosyncratiques ψ_i, et ainsi une matrice de covariance $\mathbf{\Sigma}_t$ définie positive. Cela correspond à des contraintes de positivité traduites par les conditions de Kuhn-Tucker suivantes :

$$\begin{aligned}
\widetilde{\psi} &\geq \mathbf{0} \\
\sum_{t=1}^{n} vecd \left[\widetilde{\mathbf{\Sigma}}_t^{-1} (\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)' \widetilde{\mathbf{\Sigma}}_t^{-1} - \widetilde{\mathbf{\Sigma}}_t^{-1} \right] &\leq \mathbf{0} \\
\sum_{t=1}^{n} vecd \left[\widetilde{\mathbf{\Sigma}}_t^{-1} (\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)' \widetilde{\mathbf{\Sigma}}_t^{-1} - \widetilde{\mathbf{\Sigma}}_t^{-1} \right] \odot \widetilde{\psi} &= \mathbf{0}
\end{aligned} \tag{3.11}$$

où $\widetilde{}$ désigne les estimations du maximum de vraisemblance, et $\odot$ le produit matriciel de Hadamard. La deuxième ligne de (3.11) fournie les multiplicateurs de Kuhn-Tucker (moins) associés aux q restrictions de l'inégalité $\psi \geq \mathbf{0}$.

Ceci signifie que des variances idiosyncratiques nulles, qui sont sur la frontière de l'espace d'admissibilité, peuvent satisfaire les conditions du premier ordre de la maximisation (3.11) même si $\sum_{t=1}^{n} \ell_\psi \left(\mathbf{y}_t / \mathcal{D}_{t-1}, \Theta\right) \neq \mathbf{0}$. Cette solution, connue dans la littérature sur les modèles à facteurs statiques comme "cas de Heywood" (Heywood [1931]), est fréquemment rencontrée dans les cas pratiques. Étant donné que le nombre maximal des cas Heywood lorsque $rang(\mathbf{\Sigma}_t) = q$ est égal à k, ceci implique que sur les solutions intérieures, il y a $\sum_{j=1}^{k} \mathcal{C}_q^j$ solutions "corner" potentielles. L'évaluation du score et de la fonction de vraisemblance avec ces solutions permet de vérifier, respectivement, si elles constituent des maxima locaux et des maxima globaux.[3]

En principe, le modèle conditionnellement hétéroscédastique [3.1 - 3.2] reste toujours bien défini même si la matrice $\mathbf{\Psi}$ n'est pas de plein rang. Dans ce cas, on dit que certaines variables observables y_{it} sont parfaitement expliquées par les facteurs communs. On dit aussi que la distribution conditionnelle des facteurs cachés sachant les observations est dégénérée. Par exemple, dans le cas limite où $rang\left(\mathbf{\Psi}\right) = q - k$ et $\mathbf{H}_{t/t} = \mathbf{0}$, tous les facteurs communs seront effectivement observables.

Notons enfin que dans le cadre des modèles à facteurs standards, Bartholomew [1987] a montré que les variances idiosyncratiques ψ_i seront nulles lorsque la corrélation

[3] Par exemple, lorsque $q = 2$, $k = 1$ et $h_{1t} = 1 \; \forall \, t$, les conditions (3.11) seront toujours vérifiées par les deux solutions possibles avec cas Heywood. En fait, toutes les solutions corner doivent être des maxima globaux, puisqu'un modèle à un seul facteur statique est sous identifié avec deux séries seulement, mais une variance idiosyncratique singulière peut le rendre complètement identifié.

linéaire entre une variable observée et les variables restantes est assez élevée. Cet argument est basé sur le fait que les éléments diagonaux de $\mathbf{\Sigma}_t^{-1}$ sont les réciproques des variances résiduelles dans les régressions (conditionnelles) de chaque y_{it} sur les $q-1$ séries restantes.

3.3.3 L'Algorithme EM

D'après Dempster et al. [1977], l'algorithme EM est une approche générale qui fait un calcul itératif pour trouver des estimateurs du maximum de vraisemblance lorsque les données sont incomplètes. Cet algorithme a connu un grand essor dans plusieurs domaines de l'économétrie appliquée (voir Engle et Watson [1981], Watson et Engle [1983], Hamilton [1990]). Ce succès est expliqué par le fait que

- L'algorithme EM est stable numériquement et la vraisemblance croît à chaque itération (sauf à un point fixe de l'algorithme).
- L'algorithme EM converge globalement sous certaines conditions. En effet, en partant d'un point arbitraire Θ_0 dans l'espace du paramètre, la convergence se fait presque toujours à un maximum local. Il peut arriver que ce ne soit pas le cas, mais cela arrive très rarement ; soit que le choix de Θ_0 ait été très malchanceux ou encore qu'il y ait une pathologie locale dans la fonction de log-vraisemblance.
- L'algorithme EM est facilement mis en application parce qu'il s'appuie sur le calcul des données complètes. En effet, l'étape E ne prend que l'espérance sur la distribution conditionnelle des données complètes à chaque itération, tandis que l'étape M n'exige, pour sa part, que l'estimation du maximum de vraisemblance des données complètes à chaque itération, qui est souvent sous une forme simple.
- L'algorithme EM est souvent facile à programmer, puisque ni l'évaluation de sa vraisemblance des données observées ni celle de ses dérivées ne sont nécessaires.
- L'algorithme EM demande peu d'espace de stockage et peut généralement être utilisé sur un petit ordinateur. Par exemple, il n'a pas besoin d'emmagasiner la matrice d'information ni son inverse.
- Le coût par itération étant généralement bas, un plus grand nombre d'itérations que les autres méthodes peut donc être exécuté par l'algorithme EM pour un coût donné.
- En observant la croissance monotone de la vraisemblance à chaque itération, il est facile de contrôler sa convergence et les erreurs de programmation.
- L'algorithme EM peut être utilisé pour fournir des valeurs estimées des données manquantes.

Certaines critiques peuvent aussi être adressées à cet algorithme, notamment dans le cas des modèles à facteurs avec hétéroscédasticité dynamique, à savoir :

- L'algorithme EM n'a pas de procédure incluse qui pourrait produire la matrice de variance-covariance des paramètres estimés.
- L'algorithme EM peut converger lentement même pour les problèmes qui semblent inoffensifs. Il peut converger lentement aussi lorsqu'il y a beaucoup d'information manquante.

- Il n'est pas certain que l'algorithme EM convergera à un maximum global ou local lorsqu'il y a plusieurs maxima.
- Le travail analytique nécessaire est souvent plus simple que celui des autres méthodes puisque seulement l'espérance conditionnelle de la log-vraisemblance pour les données complètes a besoin d'être maximisée. Cependant il y a une certaine quantité de travail analytique à faire pour exécuter l'étape E, et dans certains cas cette étape peut être analytiquement impossible à trouver. C'est le cas d'ailleurs du modèle à facteurs [3.1 - 3.2] où l'étape E nécessite le calcule des espérances conditionnelles et des matrices de variance-covariance conditionnelles de certaines fonctions non linéaires de $\mathbf{f}_t$ et qui ne peuvent pas être obtenues directement par le filtre de Kalman.

L'application directe de cet algorithme est donc assez compliquée du fait que l'estimation des paramètres liés aux processus GQARCH, nécessite le calcul des moments conditionnels de certaines fonctions non linéaires des facteurs communs. L'approche itérative qu'on va présenter, par la suite, suppose que l'algorithme EM standard pour les modèles à facteurs peut toujours être appliqué, même en présence d'effets de type ARCH, et ce à condition que les paramètres de la variance conditionnelle seront connus. Il s'agit d'une approche en deux étapes qui, dans un premier temps, utilise l'algorithme EM pour estimer les coefficients des variables explicatives, aussi bien que les éléments de la matrice des pondérations et les variances idiosyncratiques et ce conditionnellement aux valeurs des paramètres ARCH. Dans une seconde étape, nous appliquons une méthode basée sur les dérivées du premier ordre afin d'estimer les paramètres de la variance conditionnelle. Nous allons, donc, combiner l'approche EM avec un algorithme de type Newton pour estimer l'ensemble des paramètres de ce modèle.

I. Pseudo-Maximum de Vraisemblance

Dans la littérature financière plusieurs paramétrisations pour la structure générale [3.1 - 3.2] ont été adoptées. Diebold et Nerlove [1989] ont proposé le modèle à facteurs ARCH avec une matrice de variances idiosyncratiques $\mathbf{\Psi}$ diagonale. Engle [1987] a proposé aussi un modèle ARCH multivarié à structure factorielle dont la matrice des variances idiosyncratiques n'est pas diagonale, mais singulière. La différence entre les deux modèles réside par conséquent dans le rang de $\mathbf{\Psi}$, et donc le degré d'observabilité des facteurs. Si $rang(\mathbf{\Psi}) \leq q - k$ les facteurs seront complètement dévoilés par les variables observées $\mathcal{Y}$. Autrement, ces derniers ne seront que partiellement dévoilés (voir King, Sentana et Wadhwani [1994]).

En supposant que les facteurs communs $\mathbf{f}_t$ sont observables, on obtient :

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{f}_t \end{pmatrix} / \mathcal{Y}_{t-1}, \mathcal{F}_{t-1}, \mathcal{Z}_t \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{B}\mathbf{z}_t \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}\mathbf{H}_t\mathbf{X}' + \mathbf{\Psi} & \mathbf{X}\mathbf{H}_t \\ \mathbf{H}_t\mathbf{X}' & \mathbf{H}_t \end{pmatrix} \right] \tag{3.12}$$

où $\mathcal{Y}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ...\}$, $\mathcal{F}_{t-1} = \{\mathbf{f}_{t-1}, \mathbf{f}_{t-2}, ...\}$, et $\mathcal{Z}_t = \{\mathbf{z}_t, \mathbf{z}_{t-1}, ...\}$: c'est l'ensemble informationnel disponible à la date $t-1$.

Dans ce travail le système [3.1 - 3.2] suppose implicitement que ces facteurs sont non observables. Cependant, les paramètres $\mathbf{B}$, $\mathbf{X}$, $\mathbf{\Psi}$, $\gamma = \{\gamma_1, \gamma_2, ..., \gamma_k\}$, $\alpha = \{\alpha_1, ..., \alpha_k\}$

et $\delta = \{\delta_1, ..., \delta_k\}$ pourront toujours être estimés en se basant sur les données observées $\mathcal{Y}_n = \{\mathbf{y}_n, \mathbf{y}_{n-1}, ..., \mathbf{y}_1\}$ et $\mathcal{Z}_n = \{\mathbf{z}_n, \mathbf{z}_{n-1}, ..., \mathbf{z}_1\}$, mais pour un nombre fini de séries observées q, la distribution des $\mathbf{y}_t$ conditionnellement à l'information disponible jusqu'à la date $t-1$ est inconnue. Afin de résoudre ce problème, Harvey, Ruiz et Sentana [1992] et Demos et Sentana [1998] ont proposé l'approximation suivante :

$$\mathbf{y}_t/\mathcal{Y}_{t-1}, \mathcal{Z}_t \approx \mathcal{N}\left(\mathbf{B}\mathbf{z}_t, \mathbf{\Sigma}_{t/t-1}\right) \tag{3.13}$$

où $\mathbf{\Sigma}_{t/t-1} = \mathbf{X}\mathbf{H}_{t/t-1}\mathbf{X}' + \mathbf{\Psi}$; "$\approx$" signifie "approximativement distribuée", et $\mathbf{H}_{t/t-1}$ l'espérance de $\mathbf{H}_t$, conditionnellement à $\mathcal{Y}_{t-1}$ et $\mathcal{Z}_t$ obtenue par le filtre de Kalman[4].

En ignorant les conditions initiales, la pseudo log-vraisemblance sera donnée par :

$$\mathcal{L}(\Theta/\mathcal{Y}) = c - \frac{1}{2}\sum_{t=1}^{n} \log|\mathbf{\Sigma}_{t/t-1}| - \frac{1}{2}\sum_{t=1}^{n}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)'\mathbf{\Sigma}_{t/t-1}^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t) \tag{3.14}$$

Cette fonction peut être maximisée par rapport au vecteur des paramètres $\Theta' = [vec(\mathbf{B})', vec(\mathbf{X})', vech(\mathbf{\Psi})', \gamma', \alpha', \delta']$, en résolvant les conditions du premier ordre qui leurs sont associées. Mais étant donné que ces dernières sont très compliquées dans ce cas, une approche numérique est nécessaire. En utilisant une méthode de maximum de vraisemblance basée sur les dérivées premières, le filtre de Kalman doit être utilisé pour estimer les facteurs non observables $\mathbf{f}_t$, aussi bien que leurs variances supposées dynamiques dans le temps une fois pour chaque paramètres et à chaque itération. Cette procédure nécessite alors un temps de calcul assez important, qui peut augmenter d'une manière disproportionnée lorsque le nombre des séries considérées augmente aussi. Ce n'est donc pas surprenant que les applications empiriques portant sur ce type de modèles ont été limitées au cas où q est relativement petit. En effet, étant données les dimensions des matrices $\mathbf{B}$, $\mathbf{X}$ et $\mathbf{\Psi}$, une méthode basée sur les dérivées premières utilise le filtre de Kalman $q\,[m + k + (q+1)/2]$ fois à chaque itération alors que l'algorithme EM n'utilise le filtre qu'une seule fois. Par exemple, dans le cas d'un modèle où $q = 200$, sans variables exogènes, avec deux facteurs communs et une matrice $\mathbf{\Psi}$ diagonale, à chaque itération la méthode basée sur les dérivées premières utilise le filtre de Kalman 600 fois beaucoup plus que l'algorithme EM.

II. Structure Générale de l'Algorithme

L'algorithme EM proposé par Rubin et Thayer [1982, 1983] reste toujours valable dans le cas où les paramètres de la variance conditionnelle sont non nuls mais connus. En effet, si on suppose que les facteurs $\mathbf{f}_t$ sont observables et sous l'hypothèse de normalité, les équations [3.1 - 3.2] impliquent :

[4] Sentana [1994] a indiqué que cette approximation s'améliore avec l'augmentation du nombre des séries observées q, étant donné que les facteurs non observables peuvent être estimés d'une manière consistante par des combinaisons linéaires des $\mathbf{y}_t$. Si on procède par une transformation du processus génératif des données [3.1 - 3.2] de sorte que $\mathbf{f}_t = \mathbf{H}_{t/t-1}^{1/2}\mathbf{f}_t^*$, alors la distribution de $\mathbf{y}_t/\mathcal{Y}_{t-1}, \mathcal{Z}_t$ sera exactement gaussienne (voir Harvey, Ruiz et le Sentana [1992]). Dans ce cas les deux modèles ne peuvent donc se distinguer l'un de l'autre sur la base de la distribution des $\mathbf{y}_t$.

$$\mathbf{y}_t/\mathbf{f}_t, \mathcal{F}_{t-1}, \mathcal{Z}_t \sim \mathcal{N}\left[\mathbf{A}\widetilde{\mathbf{y}}_t, \mathbf{\Psi}\right] \tag{3.15}$$

où $\mathbf{A} = [\mathbf{B}, \mathbf{X}]$, la matrice des coefficients de "régression" de dimension $q \times (m+k)$ et $\widetilde{\mathbf{y}}'_t = [\mathbf{z}'_t, \mathbf{f}'_t]$. Ainsi, la vraisemblance de la *t-ème* observation, conditionnellement à l'information "disponible" à la date t, peut être exprimée sous la forme :

$$\begin{aligned} p\left(\mathbf{y}_t, \mathbf{f}_t/\mathcal{Y}_{t-1}, \mathcal{F}_{t-1}, \mathcal{Z}_t\right) &= p\left(\mathbf{y}_t/\mathbf{f}_t, \mathcal{Y}_{t-1}, \mathcal{F}_{t-1}, \mathcal{Z}_t\right) p\left(\mathbf{f}_t/\mathcal{Y}_{t-1}, \mathcal{F}_{t-1}, \mathcal{Z}_t\right) \\ &= p\left(\mathbf{y}_t/\mathcal{Y}_{t-1}, \mathcal{F}_{t-1}, \mathcal{Z}_t\right) p\left(\mathbf{f}_t/\mathcal{Y}_{t}, \mathcal{F}_{t-1}, \mathcal{Z}_t\right) \end{aligned} \tag{3.16}$$

en ignorant les conditions initiales, et en supposant que $\mathbf{\Psi}$ est de plein rang, la fonction de log-vraisemblance jointe sera donnée par :

$$\begin{aligned} \mathcal{L}(\Theta/\mathcal{Y}, \mathcal{F}) = &- \frac{nq}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{\Psi}| - \frac{1}{2}\sum_{t=1}^{n}\left(\mathbf{y}_t - \mathbf{A}\widetilde{\mathbf{y}}_t\right)' \mathbf{\Psi}^{-1}\left(\mathbf{y}_t - \mathbf{A}\widetilde{\mathbf{y}}_t\right) \\ &- \frac{1}{2}\sum_{i=1}^{k}\left(\sum_{t=1}^{n}\log(h_{it}) + \sum_{t=1}^{n}\frac{f_{it}^2}{h_{it}}\right) \end{aligned} \tag{3.17}$$

Pour l'estimation des paramètres, et étant donné que les $\mathbf{f}_t$ sont non observables, nous pouvons appliquer l'algorithme EM en calculant dans une première étape (étape E) l'espérance de la log-vraisemblance complétée donnée par (3.17) en conditionnant par rapport à $\mathcal{D}_{ni} = \left\{\mathcal{Y}_n, \mathcal{Z}_n, \Theta^{(i)}\right\}$, où $\Theta^{(i)}$ est l'estimation actuelle des paramètres. Dans une deuxième étape (étape M), cette espérance conditionnelle sera maximisée par rapport aux paramètres du modèle $\mathbf{B}$, $\mathbf{X}$ et $\mathbf{\Psi}$.

Étape E :

L'espérance conditionnelle de la log-vraisemblance complétée est donnée par :

$$\begin{aligned} \mathcal{Q}(\Theta/\Theta^{(i)}) \simeq\ & c - \frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{\Psi}| - \frac{1}{2}\sum_{t=1}^{n} tr\left[\mathbf{\Psi}^{-1}\mathbb{E}\left((\mathbf{y}_t - \mathbf{A}\widetilde{\mathbf{y}}_t)(\mathbf{y}_t - \mathbf{A}\widetilde{\mathbf{y}}_t)'\right)/\mathcal{D}_{ni}\right] \\ &- \frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}\mathbb{E}\left(\log(h_{jt}) + \frac{f_{jt}^2}{h_{jt}}/\mathcal{D}_{ni}\right) \end{aligned} \tag{3.18}$$

Étape M :

Dans cette étape, la maximisation de la fonction $\mathcal{Q}(\Theta/\Theta^{(i)})$ par rapport à $\mathbf{A}$ et $\mathbf{\Psi}$ peut être menée en ignorant le dernier terme de (3.18).[5] Les conditions du premier ordre sont données par :

[5] Si on suppose que $\mathbf{f}_t = \mathbf{H}_{t/t-1}^{1/2}\mathbf{f}_t^*$, ceci ne serait plus vrai parce que $h_{it/t-1}$ dépend indirectement de $\mathbf{A}$ et $\mathbf{\Psi}$. Dans ce cas il est conceptuellement possible que les valeurs des paramètres qui maximisent la première partie de (3.18) pourraient réellement diminuer la deuxième partie. Néanmoins, à condition que ces paramètres augmentent l'expression en général, le principe de l'algorithme EM généralisé reste toujours vérifié.

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\Theta/\Theta^{(i)})}{\partial \mathbf{A}} &= \sum_{t=1}^{n}\Big[-2\mathbf{y}_t\mathbb{E}\left(\widetilde{\mathbf{y}}_t'/\mathcal{D}_{ni}\right)+2\mathbf{A}\mathbb{E}\left(\widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t'/\mathcal{D}_{ni}\right)\Big]\\
\frac{\partial \mathcal{Q}(\Theta/\Theta^{(i)})}{\partial \mathbf{\Psi}^{-1}} &= -n\,tr(\mathbf{\Psi})+tr\left[\sum_{t=1}^{n}\mathbb{E}\Big((\mathbf{y}_t-\mathbf{A}\widetilde{\mathbf{y}}_t)(\mathbf{y}_t-\mathbf{A}\widetilde{\mathbf{y}}_t)'/\mathcal{D}_{ni}\Big)\right]
\end{aligned}$$

La résolution de ces conditions donne :

$$\mathbf{A}^{(i+1)}=\left[\sum_{t=1}^{n}\mathbf{y}_t\mathbb{E}\left(\widetilde{\mathbf{y}}_t'/\mathcal{D}_{ni}\right)\right]\left[\sum_{t=1}^{n}\mathbb{E}\left(\widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t'/\mathcal{D}_{ni}\right)\right]^{-1} \tag{3.19}$$

$$\mathbf{\Psi}^{(i+1)}=\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big[(\mathbf{y}_t-\mathbf{A}\widetilde{\mathbf{y}}_t)(\mathbf{y}_t-\mathbf{A}\widetilde{\mathbf{y}}_t)'/\mathcal{D}_{ni}\Big] \tag{3.20}$$

pour calculer ces valeurs, il faut tout d'abord calculer les espérances conditionnelles qui apparaissent dans les équations (3.19) et (3.20). Ces espérances conditionnelles peuvent être fournies par le filtre de Kalman[6].

Dans ce cas si on désigne par $\mathbb{E}\left[\widetilde{\mathbf{y}}_t'/\mathcal{D}_t\right]=\widetilde{\mathbf{y}}_{t/t}^{(i)\prime}=\left[\mathbf{z}_t',\mathbf{f}_{t/t}^{(i)\prime}\right]$ et $\mathbb{E}\left[\widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t'/\mathcal{D}_t\right]=\mathbf{\Omega}_{t/t}^{(i)}$, on peut démontrer que :

$$\mathbf{\Omega}_{t/t}^{(i)}=\mathbb{E}\left[\begin{pmatrix}\mathbf{z}_t\mathbf{z}_t' & \mathbf{z}_t\mathbf{f}_t'\\ \mathbf{f}_t\mathbf{z}_t' & \mathbf{f}_t\mathbf{f}_t'\end{pmatrix}/\mathcal{D}_t\right]=\begin{bmatrix}\mathbf{z}_t\mathbf{z}_t' & \mathbf{z}_t\mathbf{f}_{t/t}^{(i)\prime}\\ \mathbf{f}_{t/t}^{(i)}\mathbf{z}_t' & \mathbf{H}_{t/t}^{(i)}+\mathbf{f}_{t/t}^{(i)}\mathbf{f}_{t/t}^{(i)\prime}\end{bmatrix} \tag{3.21}$$

donc

$$\mathbf{A}^{(i+1)}=\left[\sum_{t=1}^{n}\mathbf{y}_t\widetilde{\mathbf{y}}_{t/t}^{(i)\prime}\right]\left[\sum_{t=1}^{n}\mathbf{\Omega}_{t/t}^{(i)}\right]^{-1} \tag{3.22}$$

et en se basant sur cette équation, nous pouvons déterminer $\mathbf{\Psi}^{(i+1)}$, soit

$$\mathbf{\Psi}^{(i+1)}=\frac{1}{n}\sum_{t=1}^{n}\left[\mathbf{y}_t\mathbf{y}_t'-\mathbf{A}^{(i+1)}\widetilde{\mathbf{y}}_{t/t}^{(i)}\mathbf{y}_t'\right] \tag{3.23}$$

Si on suppose que les paramètres de la variance conditionnelle sont nuls (des facteurs homoscédastiques et orthogonaux), ces équations seront exactement les mêmes que celles déjà trouvées dans le chapitre 2. Pour l'estimation des paramètres de la composante conditionnellement hétéroscédastique γ, α et δ, en utilisant un algorithme EM exacte, l'équation (3.18) nécessite le calcul des espérances et variances conditionnelles de certaines fonctions non linéaires des facteurs communs $\mathbf{f}_t$. L'implémentation de cet algorithme a été donc entravée par l'impossibilité de calculer ces moments conditionnels d'une manière analytique exacte (voir Fiorentini, Sentana et Shephard [2004] pour

[6] Si $\mathbf{f}_t=\mathbf{H}_{t/t-1}^{1/2}\mathbf{f}_t^*$, le modèle sera conditionnellement gaussien et le filtre de Kalman fournira les espérances conditionnelles exactes. Cependant, si $\mathbf{H}_t$ est une fonction des variables non observables, comme dans [3.1 - 3.2], le filtre produira seulement des valeurs approchées.

une approximation par simulations). La maximisation directe de la log-vraisemblance des variables observées par rapport aux paramètres GQARCH conduit, aussi, à des équations simultanées qui n'ont pas une solution analytique exacte.

Pour surmonter ces problèmes de calcul, diverses solutions ont été proposées comme par exemple, la méthode en "zig-zag" de Demos et Sentana [1998]. Cette méthode consiste à maximiser, dans une première étape, l'espérance conditionnelle de la log-vraisemblance complétée (3.18) par rapport aux paramètres, $\mathbf{B}$, $\mathbf{X}$ et $\mathbf{\Psi}$ moyennant l'algorithme EM, en utilisant les paramètres de la variance conditionnelle qu'on a déjà trouvé dans l'itération précédente. Par la suite et dans une seconde étape, on utilise les nouvelles valeurs de $\mathbf{B}$, $\mathbf{X}$ et $\mathbf{\Psi}$ pour maximiser la log-vraisemblance des variables observées (3.14) par rapport aux paramètres de la composante conditionnellement hétéroscédastique γ_i, α_i et δ_i, pour $i = 1, 2, ..., k$.

Une approche alternative, particulièrement intéressante quand un seul paramètre de la variance conditionnelle est inconnu (soit, par exemple, α_1), consiste à estimer les paramètres inconnus (c-à-d, $\mathbf{B}$, $\mathbf{X}$ et $\mathbf{\Psi}$) en maximisant pour différentes valeurs de α_1 l'espérance conditionnelle de la log-vraisemblance complétée moyennant un algorithme EM. La valeur α_1^* pour la quelle la fonction de vraisemblance est la plus élevée sera, donc, considérée comme une estimation de maximum de vraisemblance de ce paramètre. L'intérêt pratique de cette méthode reste très limité surtout pour un nombre de paramètres inconnus supérieur à un.

Une troisième approche consiste à approximer les espérances conditionnelles de l'équation (3.18). Dans ce cas, nous supposons que les matrices $\mathbf{A}$ et $\mathbf{\Psi}$ sont maintenues constantes à leurs valeurs de la dernière itération. La première partie de la fonction de log-vraisemblance sera donc considérée comme une constante :

$$\mathcal{Q}(\Theta/\Theta^{(i)}) = c^* - \frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}\mathbb{E}\left[\log h_{jt} + \frac{f_{jt}^2}{h_{jt}}/\mathcal{Y}_n, \mathcal{Z}_n, \Theta^{(i)}\right] \tag{3.24}$$

cette équation nécessite le calcul des espérances conditionnelles de certaines fonctions non linéaires de $\mathbf{f}_t$. Étant donné que ces dernières n'ont pas une forme analytique exacte, nous pouvons les approximer en ignorant les implications de l'inégalité de Jensen (voir Demos et Sentana [1998]), soit

$$\begin{aligned}\mathcal{Q}(\Theta/\Theta^{(i)}) &= c^* - \frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}\left[\log\left(\mathbb{E}\left(h_{jt}/\mathcal{D}_{ni}\right)\right) + \mathbb{E}(f_{jt}^2/\mathcal{D}_{ni})/\mathbb{E}(h_{jt}/\mathcal{D}_{ni})\right] \\ &= c^* - \frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}\left(\log h_{jt/t-1}^{(i)} + \frac{f_{jt/t}^{(i)\,2} + h_{jt/t}^{(i)}}{h_{jt/t-1}^{(i)}}\right) \end{aligned} \tag{3.25}$$

où $h_{jt/t-1}^{(i)} = 1 + \gamma_j f_{jt-1/t-1}^{(i)} + \alpha_j\left(f_{jt-1/t-1}^{(i)\,2} + h_{jt-1/t-1}^{(i)}\right) + \delta_j h_{jt-1/t-2}^{(i)}$, $h_{jt/t}^{(i)}$ est le *j-ème* élément de la diagonale de $\mathbf{H}_{t/t}$ et $f_{jt/t}^{(i)}$ le *j-ème* élément de $\mathbf{f}_{t/t}$ les deux évalués en utilisant les paramètres de la *i-ème* itération. La mise à jour des paramètres $\gamma_j^{(i+1)}$,

$\alpha_j^{(i+1)}$, et $\delta_j^{(i+1)}$ peut être menée en maximisant d'une manière itérative l'espérance conditionnelle (3.25). Pour chacun des facteurs, cette maximisation est équivalente à l'estimation d'un modèle GQARCH(1,1) univarié. Cependant, et étant donné qu'on n'a pas utilisé l'expression analytique exacte de l'espérance conditionnelle, cette approche ne conduit pas nécessairement à un maximum de $\mathcal{L}(\Theta/\mathcal{Y})$.

Dans le cas où [3.1 - 3.2] est caractérisé par une structure factorielle de la forme $\mathbf{f}_t = \mathbf{H}_{t/t-1}^{1/2}\mathbf{f}_t^*$, la log-vraisemblance complétée sera donnée par :

$$\begin{aligned}\mathcal{L}(\Theta/\mathcal{Y},\mathcal{F}) = & -\frac{nq}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{\Psi}| - \frac{1}{2}\sum_{t=1}^{n}(\mathbf{y}_t - \mathbf{A}\widetilde{\mathbf{y}}_t)'\,\mathbf{\Psi}^{-1}\,(\mathbf{y}_t - \mathbf{A}\widetilde{\mathbf{y}}_t) \\ & - \frac{1}{2}\sum_{i=1}^{k}\left(\sum_{t=1}^{n}\log h_{it/t-1} + \sum_{t=1}^{n} f_{it}^2/h_{it/t-1}\right)\end{aligned}$$

Dans une première étape, les paramètres $\mathbf{A}$ et $\mathbf{\Psi}$ seront estimés en utilisant l'algorithme EM. Par la suite, étant donné que $h_{jt/t-1} = 1 + \gamma_j f_{jt-1/t-1} + \alpha_j(f_{jt-1/t-1}^2 + h_{jt-1/t-1}) + \delta_j h_{jt-1/t-2}$ est une fonction mesurable de $\mathcal{Y}_{t-1}$, nous reprenons l'expression précédente et nous calculons les espérances conditionnelles des facteurs :

$$\begin{aligned}\mathcal{Q}(\Theta/\Theta^{(i)}) &= c^* - \frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}\Big[\log h_{jt/t-1} + \mathbb{E}(f_{jt}^2/\mathcal{D}_{ni})/h_{jt/t-1}\Big] \\ &= c^* - \frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}\left(\log h_{jt/t-1} + \frac{f_{jt/t}^{(i)\,2} + h_{jt/t}^{(i)}}{h_{jt/t-1}}\right) \qquad (3.26)\end{aligned}$$

Les paramètres, $\gamma_j^{(i+1)}$, $\alpha_j^{(i+1)}$, et $\delta_j^{(i+1)}$ seront obtenus par maximisation numérique de l'espérance conditionnelle de la log-vraisemblance (3.26). Dans chacune des itérations et pour chaque paramètre on doit mettre à jour les valeurs de $f_{jt-1/t-1}$ et $h_{jt-1/t-1}$. Par conséquent, le filtre de Kalman sera utilisé assez souvent comme si la maximisation était menée directement sur la log vraisemblance non complétée (3.14). Cette approximation s'améliore avec l'augmentation du nombre des variables observées. En effet, l'estimateur optimal des facteurs latents fourni par le filtre de Kalman (basé sur des combinaisons linéaires des $\mathbf{y}_t$) est asymptotiquement plus efficace pour q grand. Dans ce cas le modèle devient un modèle de régression multiple classique avec k modèles GQARCH univariés.

3.4 Calcul de la Fonction de Vraisemblance et des Scores

Le principe du maximum de vraisemblance constitue un cadre théorique bien connu, et son application aux modèles à facteurs dynamiques est une avancée importante dans le domaine. Malheureusement, cette approche est coûteuse en temps de calculs et la plupart des applications qui ont porté sur la modélisation de la dynamique des moments conditionnels de second ordre en adoptant une structure factorielle, ont été limiées à l'analyse de jeux de données de tailles réduites (voir Kroner [1987] ; Lin, Engle et Ito [1991] ; Sentana, Shah et Wadhwani [1992] et King, Sentana et Wadhwani [1994]). Nous présentons ici trois méthodes alternatives pour calculer la fonction de vraisemblance, son gradient, et les estimateurs des facteurs qui sont numériquement efficaces et fiables, et statistiquement justifiées. Pour la simplicité de l'exposé nous considérons le cas des modèles sans variables explicatives.

3.4.1 L'algorithme Récursif

Nous utilisons la structure espace-état (3.8) pour calculer la log-vraisemblance moyennant une décomposition en coupe transversale de l'erreur de prévision :

$$\mathcal{L}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) = -\frac{q}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{q}\log\left|\delta_{it}(\Theta)\right| - \frac{1}{2}\sum_{i=1}^{q}\frac{\eta_{it}^2(\Theta)}{\delta_{it}(\Theta)} \tag{3.27}$$

où η_{it} et δ_{it} sont les paramètres déjà calculés par (3.10).

Cette décomposition vérifie implicitement la factorisation de type Cholesky de la matrice $\mathbf{\Sigma}_t$. En effet, nous pouvons trouver une matrice triangulaire inférieure unitaire Θ_t, telle que $\Theta_t\eta_t = \mathbf{y}_t$ avec $\Theta_t\Delta_t\Theta_t'$ la factorisation symétrique de $\mathbf{\Sigma}_t$ et où $\Delta_t = diag\left[\delta_{1t}, ..., \delta_{qt}\right]$. Cette dernière démarche est cependant plus efficace puisqu'elle n'entraîne aucune opération sur les matrices, mais seulement une étape de filtrage de Kalman pour estimer les facteurs et leurs variances. Elle constitue donc une solution plus simple et moins coûteuse en termes d'erreurs numériques qui peuvent être engendrées par l'inversion des $\mathbf{\Sigma}_t$ (voir, Bauer et Reinsch [1971]). Cet algorithme ne sera pas aussi affecté par les valeurs de certaines ψ_i qui peuvent être nulles (dont le nombre ne dépasse pas k), et ainsi, reste valide même dans le cas où les valeurs de ces paramètres atteignent la limite de l'espace d'admissibilité durant le processus d'optimisation. Toutefois, cette procédure ne garantit pas une factorisation symétrique lorsque la matrice $\mathbf{\Sigma}_t$ n'est pas semi-définie positive, mais un tel résultat apparaîtra seulement dans le cas où les paramètres sont inadmissibles.

En ce qui concerne la fonction score, l'application de l'expression générale développée par Bollerslev et Wooldridge [1992] à la formule (3.27) donne :

$$\ell\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) = -\sum_{i=1}^{q}\frac{\partial\eta_{it}(\Theta)}{\partial\Theta}\frac{\eta_{it}(\Theta)}{\delta_{it}(\Theta)} + \frac{1}{2}\sum_{i=1}^{q}\left[\frac{1}{\delta_{it}(\Theta)}\frac{\partial\delta_{it}(\Theta)}{\partial\Theta}\left(\frac{\eta_{it}^2(\Theta)}{\delta_{it}(\Theta)} - 1\right)\right] \tag{3.28}$$

où

$$\begin{aligned}
\frac{\partial\eta_{it}(\Theta)}{\partial\Theta_j} &= -\left(\mathbf{x}_i'\frac{\partial\mathbf{f}_{i-1t/i-1t}(\Theta)}{\partial\Theta_j} + \frac{\partial\mathbf{x}_i'}{\partial\Theta_j}\mathbf{f}_{i-1t/i-1t}(\Theta)\right) \quad \text{et} \\
\frac{\partial\delta_{it}\left(\Theta\right)}{\partial\Theta_j} &= 2\frac{\partial\mathbf{x}_i'}{\partial\Theta_j}\mathbf{H}_{i-1t/i-1t}(\Theta)\mathbf{x}_i + \mathbf{x}_i'\frac{\partial\mathbf{H}_{i-1t/i-1t}(\Theta)}{\partial\Theta_j}\mathbf{x}_i + \frac{\partial\psi_i}{\partial\Theta_j}
\end{aligned}$$

Nous pouvons calculer $\partial\mathbf{f}_{i-1t/i-1t}(\Theta)/\partial\Theta_j$ et $\partial\mathbf{H}_{i-1t/i-1t}(\Theta)/\partial\Theta_j$ en utilisant les équations de mise à jour données par (3.9). Dans ce cas, $\partial\mathbf{f}_{0t/0t}(\Theta)/\partial\Theta_j = \mathbf{0}$ et $\partial\mathbf{H}_{0t/0t}(\Theta)/\partial\Theta_j = \partial\mathbf{H}_t\left(\phi\right)/\partial\Theta_j$ seront considérées comme valeurs initiales (voir, Harvey [1989]). Notons enfin que, étant données les hypothèses imposées sur $\mathbf{h}_t$, les paramètres $\mathbf{x}_i'$ et ψ_i n'apparaissent pas dans la décomposition en coupe transversale de l'erreur de prévision avant que y_{it} ne soit traitée.

3.4.2 La Méthode non Récursive

Malgré l'intérêt porté par la méthode récursive, nous pouvons toujours développer un algorithme beaucoup plus efficace à partir de l'identité (3.16). Après avoir regroupé les termes et à condition que les inverses nécessaires existent, on aura :

$$\begin{aligned}
\mathcal{L}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) &= -\frac{q}{2}\log 2\pi - \frac{1}{2}\log\left(|\boldsymbol{\Psi}|\cdot|\mathbf{H}_t|\cdot\left|\mathbf{H}_{t/t}^{-1}\right|\right)\\
&- \frac{1}{2}\left(\mathbf{y}_t'\boldsymbol{\Psi}^{-1}\mathbf{y}_t - \mathbf{f}_{t/t}'\mathbf{H}_{t/t}^{-1}\mathbf{f}_{t/t}\right) - \frac{1}{2}\mathbf{f}_t'\left(\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X} - \mathbf{H}_{t/t}^{-1}\right)\mathbf{f}_t\\
&- \mathbf{f}_t'\left(\mathbf{H}_{t/t}^{-1}\mathbf{f}_{t/t} - \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}_t\right) \qquad (3.29)
\end{aligned}$$

Étant donné que les deux derniers termes doivent être identiquement nuls pour toutes les valeurs de $\mathbf{f}_t$, l'équation (3.29) donne :

$$\begin{aligned}
\mathbf{H}_{t/t} &= \left(\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\\
\mathbf{f}_{t/t} &= \left(\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}_t\\
|\boldsymbol{\Sigma}_t| &= |\mathbf{H}_t|\cdot|\boldsymbol{\Psi}|\cdot\left|\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right|\\
\mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{y}_t &= \mathbf{y}_t'\boldsymbol{\Psi}^{-1}\mathbf{y}_t - \mathbf{y}_t'\boldsymbol{\Psi}^{-1}\mathbf{X}\left(\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}_t \qquad (3.30)
\end{aligned}$$

Dans ce cas la factorisation de la matrice $\boldsymbol{\Sigma}_t$ (une matrice de dimension $(q\times q)$) sera remplacée par une factorisation de la matrice $\left[\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right]$ de dimension $(k\times k)$ et de la matrice $\boldsymbol{\Psi}$ de dimension $(q\times q)$. L'efficacité de cette approche non récursive en termes de temps de calcul se traduit donc à travers la structure bien particulière de la matrice $\boldsymbol{\Psi}$. D'une part $\boldsymbol{\Psi}$ est une matrice diagonale, son inverse et son déterminant seront alors faciles à calculer. D'autre part les ψ_i ne varient pas à travers le temps ce qui implique une matrice $\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}$ invariante aussi[7].

L'expression (3.16) peut aussi être utilisée afin de simplifier le calcul du score. D'après l'inégalité de Kullback on a :

$$\mathbb{E}\left[\sum_t \ell\left(\mathbf{f}_t/\mathbf{y}_t,\mathcal{Y}_{t-1};\Theta\right)/\mathcal{Y},\Theta\right] = \mathbf{0}$$

donc $\ell(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta)$ peut être obtenue en appliquant l'espérance conditionnelle (sachant $\mathcal{Y}_n$ et Θ) sur la somme des scores non observables qui correspondent à $\mathcal{L}(\mathbf{f}_t/\mathcal{Y}_{t-1};\Theta)$ et $\mathcal{L}(\mathbf{y}_t/\mathbf{f}_t,\mathcal{Y}_{t-1};\Theta)$. Si on suppose que $\psi > \mathbf{0}$, ceci implique

$$\begin{aligned}
\ell_{\mathbf{x}}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) &= vec\left\{\left[\mathbf{f}_{t/t}\mathbf{y}_t' - \left(\mathbf{f}_{t/t}\mathbf{f}_{t/t}' + \mathbf{H}_{t/t}\right)\mathbf{X}'\right]\boldsymbol{\Psi}^{-1}\right\}\\
\ell_{\psi}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) &= \frac{1}{2}vecd\left\{\boldsymbol{\Psi}^{-1}\left[\left(\mathbf{y}_t - \mathbf{X}\mathbf{f}_{t/t}\right)\left(\mathbf{y}_t - \mathbf{X}\mathbf{f}_{t/t}\right)' + \mathbf{X}\mathbf{H}_{t/t}\mathbf{X}' - \boldsymbol{\Psi}\right]\boldsymbol{\Psi}^{-1}\right\}\\
\ell_{\phi}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) &= \frac{1}{2}\partial\mathbf{h}_t'(\phi)/\partial\phi\cdot vecd\left\{\mathbf{H}_t^{-1}\left[\mathbf{f}_{t/t}\mathbf{f}_{t/t}' + \mathbf{H}_{t/t} - \mathbf{H}_t\right]\mathbf{H}_t^{-1}\right\} \qquad (3.31)
\end{aligned}$$

[7] Afin d'éviter les erreurs d'ordre numériques liées à l'inversion des matrices, il faut mener tout d'abord la factorisation symétrique de $\left[\mathbf{H}_t^{-1} + \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right] = \mathbf{F}_{Lt}\mathbf{F}_{Dt}\mathbf{F}_{Lt}'$ par la suite, nous pouvons calculer les expressions nécessaire telles que $|\boldsymbol{\Sigma}_t| = |\mathbf{H}_t|\cdot|\boldsymbol{\Psi}|\cdot|\mathbf{F}_{Dt}|$, $\mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{y}_t = \mathbf{y}_t'\boldsymbol{\Psi}^{-1}\mathbf{y}_t - \mathbf{y}_t'\mathbf{F}_{Dt}^{-1}\mathbf{y}_t$, $\mathbf{H}_{t/t} = \mathbf{F}_{Lt}^{-1'}\mathbf{F}_{Dt}^{-1}\mathbf{F}_{Lt}^{-1}$ et $\mathbf{f}_{t/t} = \mathbf{r}_t$, avec $\mathbf{y}_t$, $\mathbf{r}_t$ et $\mathbf{F}_{Lt}^{-1}$ obtenues comme solutions des systèmes triangulaires unitaires des équations linéaires : $\mathbf{F}_{Lt}\mathbf{y}_t = \mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}_t$, $\mathbf{r}_t = \mathbf{F}_{Dt}^{-1}\mathbf{y}_t$ et $\mathbf{F}_{Lt}\mathbf{F}_{Lt}^{-1} = \mathbf{I}_k$.

Comme dans le cas de l'algorithme récursif, ces expressions peuvent être obtenues en appliquant la formule de Woodbury généralisée à $\mathbf{\Sigma}_t$.

Malheureusement, les avantages offerts par cet algorithme pour calculer (avec (3.29) et (3.31)) les valeurs de $\mathcal{L}(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta)$ et $\ell(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta)$ seront perdus même s'il y a un seul élément nul de ψ. Pour cette raison, Sentana [2000] a proposé un autre algorithme permettant de combiner les avantages des deux précédents.

3.4.3 L'algorithme Récursif en Bloc

On se limite à deux blocs de tailles N_a et N_b pour la simplicité de l'exposé, mais on peut généraliser sur un nombre quelconque de blocs [8]. Tous les vecteurs et les matrices seront décomposés de telle façon que l'on puisse écrire (3.1) sous la forme :

$$\begin{aligned} \mathbf{y}_{at} &= \mathbf{X}_a\mathbf{f}_t + \varepsilon_{at} \\ \mathbf{y}_{bt} &= \mathbf{X}_b\mathbf{f}_t + \varepsilon_{bt} \end{aligned} \tag{3.32}$$

et on défini $\mathbf{x}_a$, $\mathbf{x}_b$, ψ_a et ψ_b les éléments correspondants à $\mathbf{x}$ et ψ.

Nous pouvons décomposer la log-vraisemblance jointe de $\mathbf{y}_{at}$ et $\mathbf{y}_{bt}$, soit :

$$\mathcal{L}(\mathbf{y}_{at},\mathbf{y}_{bt}/\mathcal{Y}_{t-1};\Theta) = \mathcal{L}(\mathbf{y}_{at}/\mathcal{Y}_{t-1};\Theta) + \mathcal{L}(\mathbf{y}_{bt}/\mathbf{y}_{at},\mathcal{Y}_{t-1};\Theta) \tag{3.33}$$

Si $\psi_a > \mathbf{0}$, $\mathcal{L}(\mathbf{y}_{at}/\mathcal{Y}_{t-1};\Theta)$ peut être calculée en utilisant l'algorithme de la section 3.4.2. Sinon, nous pouvons ordonner de nouveau les variables de telle façon que les h éléments nuls de ψ apparaissent dans les N_b dernières positions. Une telle méthode est équivalente à la pré-multiplication de l'équation (3.1) par une matrice de permutation symétrique U'. Cette matrice peut être déterminée en inter-changeant le premier élément nul (dans un sens descendant) de ψ avec le premier élément positif (dans un sens ascendant), et on répète la même procédure pour les éléments restants.

D'un autre côté on a :

$$\mathcal{L}(\mathbf{y}_{bt}/\mathbf{y}_{at},\mathcal{Y}_{t-1};\Theta) = -\frac{N_b}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_{b.at}| - \frac{1}{2}\eta'_{bt}\mathbf{\Sigma}^{-1}_{b.at}\eta_{bt}$$

où

$$\begin{aligned} \eta_{bt} &= \mathbf{y}_{bt} - \mathbf{X}_b\mathbf{H}_t\mathbf{X}'_a\mathbf{\Sigma}^{-1}_{at}\mathbf{y}_{at} = \mathbf{y}_{bt} - \mathbf{X}_b\mathbf{f}_{at/at} \quad \text{et} \\ \mathbf{\Sigma}_{b.at} &= \mathbf{X}_b\mathbf{H}_t\mathbf{X}'_b + \mathbf{\Psi}_b - \mathbf{X}_b\mathbf{H}_t\mathbf{X}'_a\mathbf{\Sigma}^{-1}_{at}\mathbf{X}_a\mathbf{H}_t\mathbf{X}'_b = \mathbf{X}_b\mathbf{H}_{at/at}\mathbf{X}'_b + \mathbf{\Psi}_b \end{aligned}$$

A ce niveau, étant donné que $\eta_{bt} = \mathbf{X}_b\left(\mathbf{f}_t - \mathbf{f}_{at/at}\right) + \varepsilon_{bt}$, la matrice de covariance conditionnelle de η_{bt} aura la même structure factorielle que celle de la matrice de covariance conditionnelle de $\mathbf{y}_{bt}$, mais en remplaçant tout simplement $\mathbf{H}_t$ par $\mathbf{H}_{at/at}$. Il faut noter aussi que la définie positivité de $\mathbf{\Sigma}_t$ implique l'existence de l'inverse de $\mathbf{\Sigma}_{b.at}$ même dans le cas limite où $\psi_b = \mathbf{0}$, et ce à condition que $N_b \leq k$.

Finalement et en se basant sur la représentation espace-état en coupe transversale (3.8), nous pouvons développer les équations de mise à jour en bloc suivantes :

[8] Avec un seul bloc, cet algorithme est équivalent à l'algorithme non récursif de la section 3.4.2, mais avec q blocs, nous obtenons l'algorithme récursif de la section 3.4.1.

$$\begin{aligned}
\mathbf{f}_{t/t} &= \mathbf{f}_{at/at} + \mathbf{H}_{at/at}\mathbf{X}_b'\mathbf{\Sigma}_{b.at}^{-1}\eta_{bt} \quad \text{et} \\
\mathbf{H}_{t/t} &= \mathbf{H}_{at/at} - \mathbf{H}_{at/at}\mathbf{X}_b'\mathbf{\Sigma}_{b.at}^{-1}\mathbf{X}_b\mathbf{H}_{at/at}
\end{aligned} \tag{3.34}$$

Ces expressions peuvent être obtenues, aussi, en appliquant la formule de Woodbury (voir Annexe). Malgré leur complexité apparente, des simulations numériques effectuées par Sentana ont confirmé les avantages de calcul de cet algorithme récursif en bloc par rapport à celui basé sur la décomposition de Cholesky de $\mathbf{\Sigma}_t$, surtout lorsque N_b est relativement petit par rapport à q. Par exemple, si $N_b = 1$, $\mathbf{\Sigma}_{b.at}$ est un scalaire. Notons bien que ces expressions restent toujours valident même lorsque les valeurs des paramètres tendent vers la limite de l'espace d'admissibilité.

En ce qui concerne le gradient, l'équation (3.33) implique

$$\ell\left(\mathbf{y}_{at}, \mathbf{y}_{bt}/\mathcal{Y}_{t-1};\Theta\right) = \ell\left(\mathbf{y}_{at}/\mathcal{Y}_{t-1};\Theta\right) + \ell\left(\mathbf{y}_{bt}/\mathbf{y}_{at}, \mathcal{Y}_{t-1};\Theta\right)$$

Étant données nos hypothèses concernant $\mathbf{h}_t$, $\mathbf{x}_b$ et ψ_b affectent seulement la seconde composante. Les dérivées de la fonction de log-vraisemblance par rapport à ces paramètres seront données par :

$$\begin{aligned}
\ell_{\mathbf{x}_b}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) &= vec\left[\mathbf{f}_{at/at}\eta_{bt}'\mathbf{\Sigma}_{b.at}^{-1} + \mathbf{H}_{at/at}\mathbf{X}_b'\mathbf{\Sigma}_{b.at}^{-1}\left(\eta_{bt}\eta_{bt}'\mathbf{\Sigma}_{b.at}^{-1} - \mathbf{I}\right)\right] \\
\ell_{\psi_b}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right) &= \frac{1}{2}\,vecd\left[\mathbf{\Sigma}_{b.at}^{-1}\eta_{bt}\eta_{bt}'\mathbf{\Sigma}_{b.at}^{-1} - \mathbf{\Sigma}_{b.at}^{-1}\right]
\end{aligned}$$

Notons ici que $\sum_t \ell_{\psi_b}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right)$ sont les dérivées dont le signe va être vérifié pour décider si une solution limite satisfait les condition du premier ordre de Kuhn-Tucker (3.11). D'un autre côté, puisque $\mathbf{x}_a$ et ψ_a affectent la première composante directement et la seconde à travers $\mathbf{f}_{at/at}$ et $\mathbf{H}_{at/at}$, nous pouvons combiner les expressions de la section 3.4.2 avec les équations de mise à jour en bloc (3.34) afin de trouver $\ell_{\mathbf{x}_a}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right)$ et $\ell_{\psi_a}\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right)$. Notons enfin que l'expression de $\ell\left(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta\right)$ donnée par (3.31) ne sera pas affectée par les éléments nuls de ψ_b.

Si le nombre des cas Heywood h est égal au nombre des facteurs, les séries correspondantes seront classées dans les k premières positions, avec $N_a = k$. En effet, pour garantir une matrice de covariance $\mathbf{\Sigma}_t$ définie positive, $\mathbf{X}_a$ doit être de plein rang ce qui implique des facteurs complètement observés donnés par : $\mathbf{f}_t = \mathbf{X}_a^{-1}\mathbf{y}_{at}$. La distribution conditionnelle de $\mathbf{y}_{bt}$ sachant $\mathbf{y}_{at}$ et $\mathcal{Y}_{t-1}$ est donc gaussienne de moyenne $\mathbf{X}_b^*\mathbf{y}_{at}$, avec $\mathbf{X}_b^* = \mathbf{X}_b\mathbf{X}_a^{-1}$, et de matrice de covariance $\mathbf{\Psi}_b$ diagonale. Étant donné que cette re-paramétrisation est bijective, moyennant la propriété d'invariance du maximum de vraisemblance, nous pouvons combiner les estimations de $\mathbf{x}_a$ et ϕ obtenues à partir du modèle marginal de $\mathbf{y}_{at}$ avec les estimations MCO de $\mathbf{x}_b^*$ et ψ_b obtenues par la régression de chaque élément de $\mathbf{y}_{bt}$ sur $\mathbf{y}_{at}$ (voir, Sentana [1997]). À moins que les conditions de Kuhn-Tucker (3.11) ne seront satisfaites, les paramètres résultants peuvent être considérés comme des estimations de maximum de vraisemblance d'un modèle sous la contrainte d'égalité de k ψ_j à 0. Dans ce cas, l'estimation d'un modèle sans contraintes de positivité sur certaines de ces variances idiosyncratiques, augmentera la log-vraisemblance jointe. Cette méthode peut aussi être utilisée si le nombre des cas Heywood est strictement inférieur à k, à condition que les variables dont les variances idiosyncratiques sont nulles ne dépendent que de h facteurs (voir, Lawley et Maxwell [1971]). Bien que cette condition peut toujours être satisfaite dans les modèles à facteurs statiques à travers les rotations orthogonales dues à l'indétermination de la matrice des pondérations, elle ne sera pas vérifiée dans le cas général où les variances des facteurs communs sont supposées dynamiques dans le temps.

3.5 Simulations de Monte Carlo

Nous avons testé la qualité des estimations des algorithmes que nous avons présenté dans ce chapitre. Pour ce faire nous avons simulé des modèles à facteurs conditionnellement hétéroscédastiques qui diffèrent par leurs structures de volatilité, en supposant dans un premier temps une moyenne nulle, par la suite une moyenne qui dépend d'un certain nombre de variables exogènes. Pour choisir la structure de volatilité convenable, deux critères de sélection de modèles ont été utilisés. En se basant sur des données simulées aussi bien que sur des données financières, nous avons testé enfin le pouvoir prévisionnel de ces modèles. Plusieurs modèles compétitifs ont été donc utilisés comme benchmarks pour la mise en évidence des performances du modèle à facteurs en se basant sur plusieurs critères d'évaluation. Dans tous les cas étudiés, nous avons supposé une structure homoscédastique pour les variances idiosyncratiques.

3.5.1 Stabilité et exactitude des Estimations

- Modèles à un seul facteur : Nous avons appliqué l'algorithme, EM sur des données simulées (un échantillon de 800 observations avec 50 réplications). Dans ce cas, nous avons adopté une spécification avec 6 variables observables et un seul facteur latent sans variables exogènes. Pour la génération des données et l'initialisation de l'algorithme EM, nous avons utilisé les valeurs du tableau 3.1. Pour l'estimation proprement dite, nous avons implémenté tout d'abord l'algorithme EM pour estimer les éléments de la matrice des pondérations $\mathbf{X}$ aussi bien que les variances idiosyncratiques ψ_i, et par la suite un algorithme de Newton pour maximiser dans un premier temps la log-vraisemblance non complétée (3.14), et dans une seconde étape l'espérance conditionnelle de la deuxième composante de la log-vraisemblance complétée (3.18), et ce afin d'estimer les coefficients de la variance conditionnelle. Tous les résultats sont donnés aussi dans le tableau 3.1.

À cette spécification nous avons rajouté, par la suite, trois variables explicatives exogènes. Les paramètres de cette simulation, les valeurs initiales, les moyennes des différents paramètres et leurs coefficients de variation obtenus pour 800 observations et 50 replications sont donnés dans le tableau 3.2.

- Modèles à deux facteurs : Dans cette deuxième simulation nous avons considéré un modèle à deux facteurs communs sans variables exogènes. Le premier facteur est

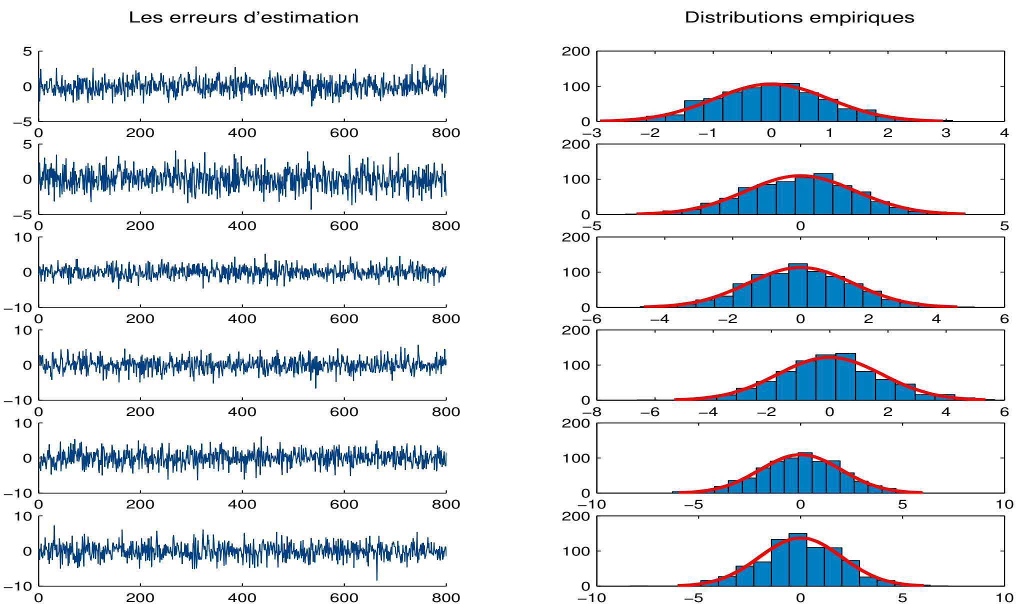

FIG. 3.1 – Modèle à un seul facteur : Les erreurs d'estimation et leurs distributions empiriques pour une seule replication.

TAB. 3.1 – Modèle à un seul facteur commun

	$\mathbf{X}$	$diag(\mathbf{\Psi})$	γ
Paramètres de simulation	1.0000 $\langle$1.0000$\rangle$	1.0000 $\langle$1.0000$\rangle$	0.2000 $\langle$1.0000$\rangle$
	2.0000 $\langle$1.0000$\rangle$	2.0000 $\langle$1.0000$\rangle$	0.2000 $\langle$1.0000$\rangle$
	3.0000 $\langle$1.0000$\rangle$	3.0000 $\langle$1.0000$\rangle$	0.5000 $\langle$1.0000$\rangle$
	4.0000 $\langle$1.0000$\rangle$	4.0000 $\langle$1.0000$\rangle$	
	5.0000 $\langle$1.0000$\rangle$	5.0000 $\langle$1.0000$\rangle$	
	6.0000 $\langle$1.0000$\rangle$	6.0000 $\langle$1.0000$\rangle$	
Vraisemblance non complétée	0.9970 (0.0308)	0.9945 (0.0260)	0.2165 (0.0604)
	1.9875 (0.0653)	1.9974 (0.0531)	0.1913 (0.0294)
	2.9809 (0.0917)	2.9794 (0.0838)	0.5022 (0.0472)
	3.9792 (0.1239)	4.0182 (0.0897)	
	4.9783 (0.1571)	4.9576 (0.1734)	
	5.9539 (0.2051)	6.0654 (0.2758)	
Vraisemblance complétée	0.8973 (0.0315)	0.9412 (0.0289)	0.1833 (0.0651)
	1.8861 (0.0667)	1.9642 (0.0596)	0.1944 (0.0321)
	2.9547 (0.0821)	2.9544 (0.0794)	0.4853 (0.0451)
	3.9613 (0.1311)	3.8952 (0.0911)	
	4.9534 (0.1598)	4.8871 (0.1693)	
	5.8752 (0.1824)	5.9577 (0.2816)	

(.) écart-types des estimations, $\langle . \rangle$ Paramètres d'initialisation

conditionnellement hétéroscédastique, suivant un processus GQARCH(1,1), alors que le deuxième est conditionnellement homoscédastique. Pour la génération des données et l'initialisation de l'algorithme EM, nous avons utilisé les paramètres donnés dans le

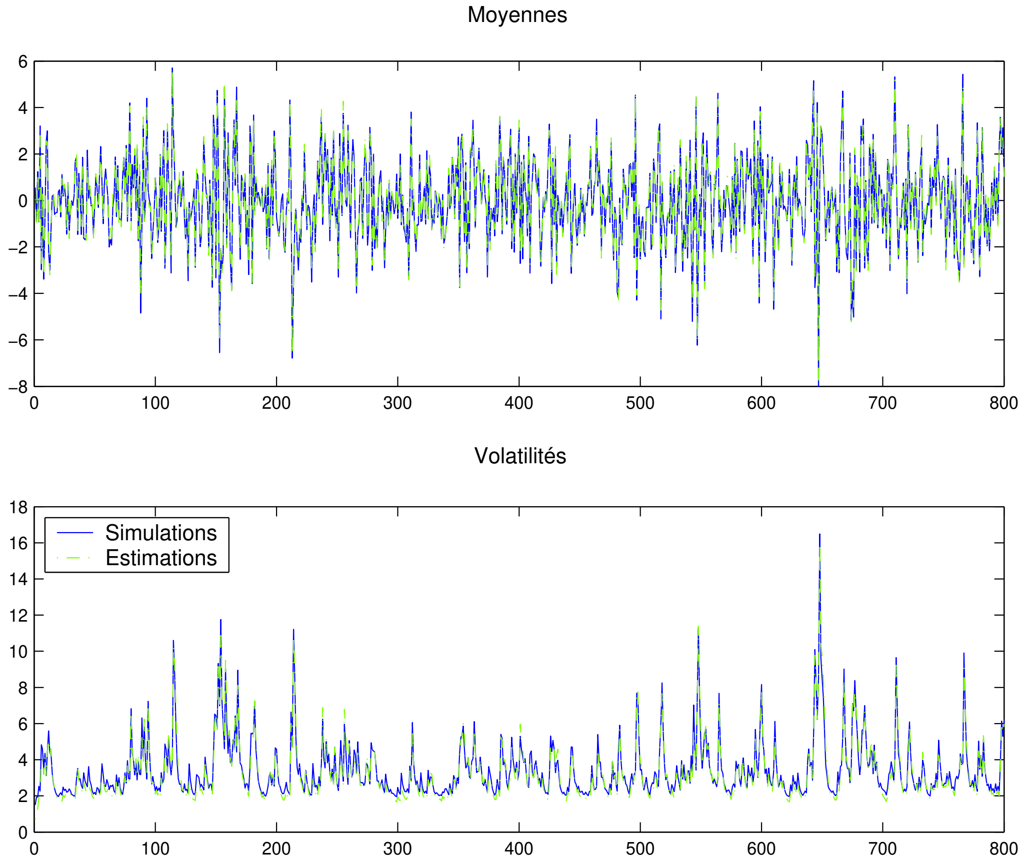

FIG. 3.2 – Modèle à un seul facteur : Moyenne du facteur et sa volatilité.

tableau 3.3. Les résultats pour 1000 observations et 100 replications sont donnés, aussi, dans le tableau 3.3.

- Modèles à trois facteurs : Nous avons étudié enfin un modèle à trois facteurs communs sans variables exogènes. Les deux premiers facteurs sont supposés conditionnellement hétéroscédastiques ayant des spécifications GQARCH(1,1) différentes, alors que le troisième à une structure conditionnellement homoscédastique. Pour la génération des données et l'initialisation de l'algorithme EM, nous avons utilisé les paramètres donnés dans le tableau 3.4. Les résultats pour 1500 observations et 100 réplications sont donnés, aussi, dans le tableau 3.4.

- Sélection de Modèles Pour le choix de la structure de volatilité convenable, nous avons utilisé les critères, AIC et BIC. La même démarche présentée dans le chapitre 2 a été suivie, à savoir trouver le critère minimum dans un certain nombre de modèles. Pour chacun, il suffit de calculer la vraisemblance, puis le critère et de choisir le critère minimum pour discriminer entre les modèles. On comptabilise alors pour chaque critère le nombre de fois qu'un modèle a été choisi et on considère que le modèle choisi est celui que les minimisations de critère ont sélectionné le plus souvent. Nous avons donc

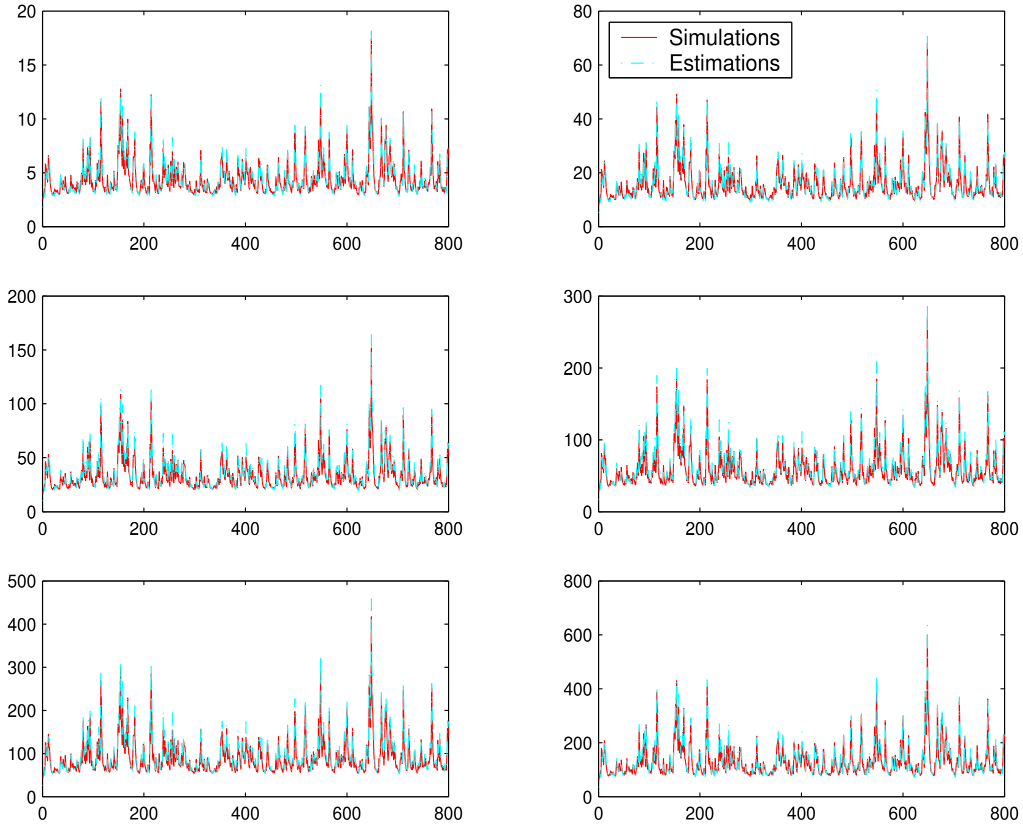

FIG. 3.3 – Modèle à un seul facteur : Volatilités des 6 séries.

mené trois expériences :

1. La première est basée sur la simulation d'un modèle à un seul facteur conditionnellement hétéroscédastique (CHF1) avec 100 réplications, en utilisant les paramètres du tableau 3.1.
2. La deuxième sur un modèle à deux facteurs dont le premier est conditionnellement hétéroscédastique alors que le deuxième est homoscédastique (CHF2) avec 100 réplications aussi, en utilisant les paramètres du tableau 3.3.
3. La troisième sur un modèle à deux facteurs homoscédastiques (FA2) avec 100 réplications, en utilisant les paramètres du tableau 2.2 (chapitre 2).

Les résultats de ces trois expériences sont donnés dans le tableau 3.5.

3.5.2 Prévision

Tous les modèles que nous avons déjà présenté peuvent être utilisés pour prévoir les variances des actifs individuels. L'application des modèles à facteurs dans une perspective de prévision a été déjà considérée dans les travaux de Kaiser [1997], où il a proposé des modèles GARCH à un seul facteur pour la modélisation de la dynamique des prix sur le marché boursier allemand. Déjà pour appliquer ces modèles aux problèmes de

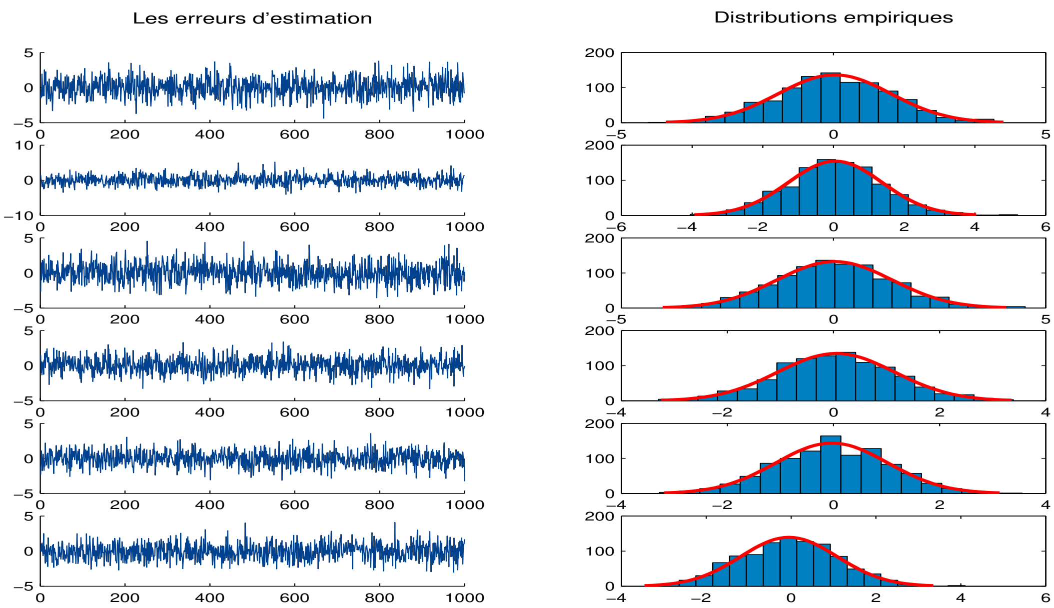

FIG. 3.4 – Modèle à deux facteurs : Les erreurs d'estimation et leurs distributions empiriques pour une seule replication.

TAB. 3.2 – Modèle à un seul facteur commun avec des variables explicatives.

	B			**X**	$diag(\mathbf{\Psi})$	γ
Paramètres de simulation	3.0000	4.0000	1.0000	1.0000	1.0000	0.3000
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨0.1000⟩
	3.0000	4.0000	1.0000	2.0000	2.0000	0.2000
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨0.1000⟩
	3.0000	4.0000	1.0000	3.0000	3.0000	0.6000
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨0.1000⟩
	2.0000	4.0000	5.0000	4.0000	4.0000	
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	
	2.0000	4.0000	5.0000	5.0000	5.0000	
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	
	2.0000	4.0000	5.0000	6.0000	6.0000	
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	
Estimation des paramètres	3.0007	3.9766	0.9882	1.0431	1.0200	0.3149
	(0.0011)	(0.0005)	(0.0004)	(0.0399)	(0.0406)	(0.3864)
	2.9951	3.9582	0.9949	2.0376	1.9681	0.1951
	(0.0019)	(0.0008)	(0.0006)	(0.0355)	(0.0354)	(0.1841)
	3.0079	3.9399	0.9939	3.0504	3.0688	0.5865
	(0.0032)	(0.0013)	(0.0011)	(0.0381)	(0.0357)	(0.0628)
	1.9974	3.9063	4.9982	4.0793	4.0879	
	(0.0062)	(0.0018)	(0.0010)	(0.0382)	(0.0381)	
	2.0042	3.8793	4.9966	5.0854	4.8256	
	(0.0073)	(0.0021)	(0.0018)	(0.0368)	(0.0466)	
	2.0026	3.8643	5.0088	6.0651	5.9588	
	(0.0092)	(0.0025)	(0.0014)	(0.0367)	(0.0388)	

(.) Coefficients de variation, ⟨.⟩ Paramètres d'initialisation

prévision, il faut tout d'abord estimer leurs paramètres, par la suite on calcule des valeurs prévues pour les variances conditionnelles des facteurs communs et enfin on déduit

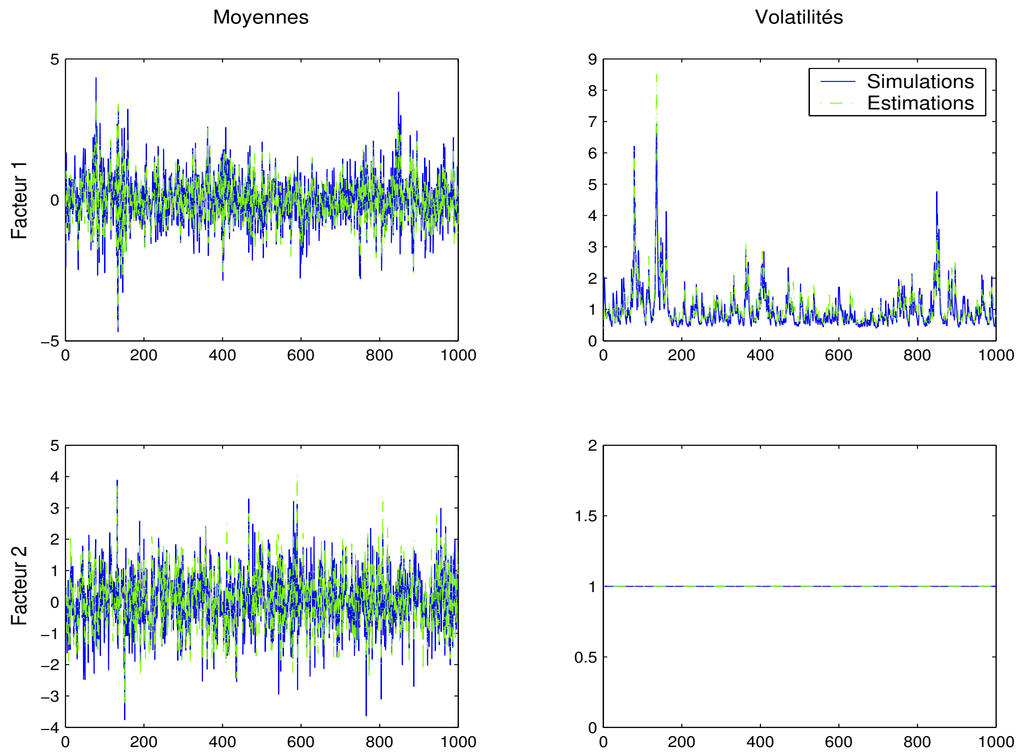

FIG. 3.5 – Modèle à deux Facteurs : Estimation des deux Facteurs et leurs volatilités.

TAB. 3.3 – Modèle à deux facteurs communs

	$\mathbf{X}$		$diag(\mathbf{\Psi})$	γ
Paramètres de simulation	1.0000 ⟨1.0000⟩	2.0000 ⟨1.0000⟩	1.0000 ⟨1.0000⟩	0.2000 ⟨0.1000⟩
	2.0000 ⟨1.0000⟩	3.0000 ⟨1.0000⟩	3.0000 ⟨1.0000⟩	0.2000 ⟨0.1000⟩
	3.0000 ⟨1.0000⟩	1.0000 ⟨1.0000⟩	2.0000 ⟨1.0000⟩	0.2000 ⟨0.4000⟩
	2.0000 ⟨1.0000⟩	4.0000 ⟨1.0000⟩	4.0000 ⟨1.0000⟩	0.6000 ⟨0.2000⟩
	4.0000 ⟨1.0000⟩	2.0000 ⟨1.0000⟩	3.0000 ⟨1.0000⟩	
	4.0000 ⟨1.0000⟩	2.0000 ⟨1.0000⟩	2.0000 ⟨1.0000⟩	
Estimation des paramètres	0.8315 (0.1001)	2.1051 (0.0630)	1.0054 (0.0666)	0.1871 (0.0307)
	1.9512 (0.1415)	3.1172 (0.1323)	3.0023 (0.1003)	0.1923 (0.0423)
	3.0815 (0.0836)	0.9523 (0.1119)	1.9779 (0.0867)	0.2124 (0.0455)
	1.9643 (0.1811)	3.9647 (0.1024)	3.9859 (0.1231)	0.5908 (0.0650)
	4.0857 (0.1120)	2.0956 (0.1765)	2.9816 (0.1072)	
	3.9477 (0.1021)	1.9378 (0.1471)	1.9977 (0.0937)	

(.) Écart-types, ⟨.⟩ Paramètres d'initialisation

directement la valeur prévue de la matrice de variance-covariance des observations. Les moments conditionnels de la distribution prédictive, $\mathbf{y}_{t+1}/\mathcal{D}_t$ pour différents horizons de prévision : $s = 2, ..., n$, sont donnés par :

$$\mathbb{E}(\mathbf{y}_{t+s}/\mathcal{D}_t) = \theta \qquad \text{et} \qquad Var(\mathbf{y}_{t+s}/\mathcal{D}_t) = \mathbf{X}\mathbf{H}_{t+s/t}\mathbf{X}' + \mathbf{\Psi}$$

Dans le cas des modèles à facteurs standards, les prévisions de la variance conditionnelle des facteurs : $\widetilde{h}_{t+1/t} = \mathbb{E}(h_{t+1}/\mathcal{Y}_{1:t})$, sont données par :

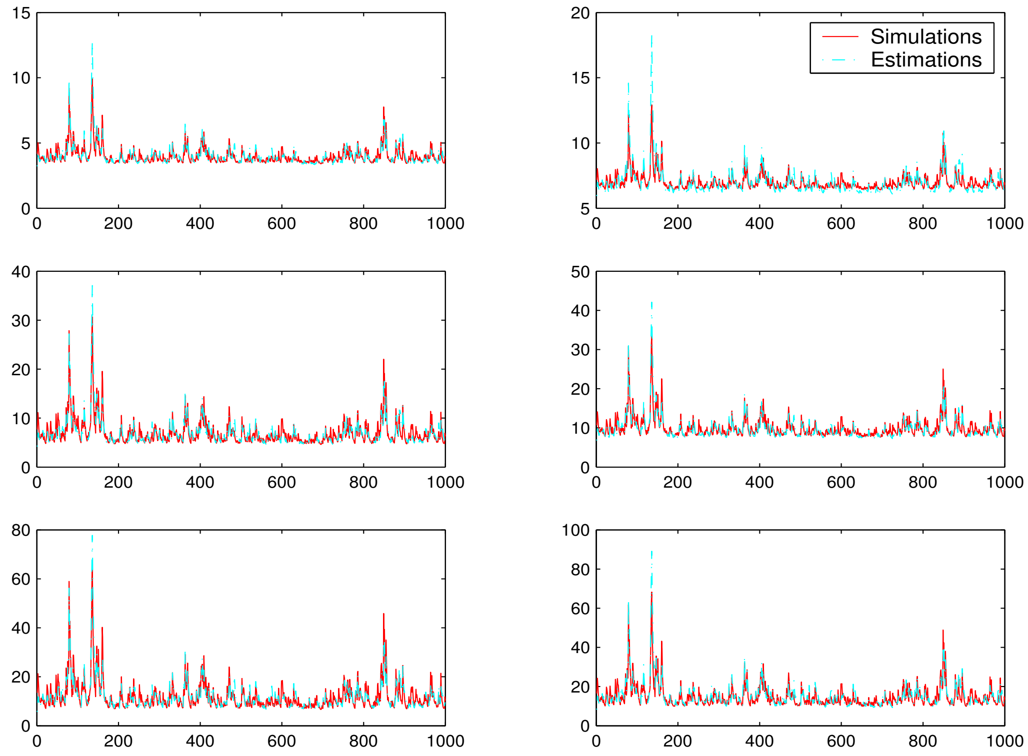

FIG. 3.6 – Modèle à deux facteurs : Volatilités des 6 séries

$$\widetilde{h}_{t+s/t} = 1 \qquad \forall\, s \geq 1 \tag{3.35}$$

quelque soit la période et quelque soit l'horizon s, la volatilité anticipée des facteurs est toujours la même, ce qui implique aussi un rendement anticipé constant pour tous les actifs. Cependant, dans le cas des modèles admettant une variance dynamique pour les facteurs communs (une spécification GQARCH(1,1)), cette volatilité anticipée pour un horizon de prévision s quelconque sera donnée par :

$$\begin{aligned}
\widetilde{h}_{t+1/t} &= w + \gamma\, \mathbb{E}(\mathbf{f}_t/\mathcal{D}_t) + \alpha\, \mathbb{E}(\mathbf{f}_t^2/\mathcal{D}_t) + \delta\, \mathbb{E}(h_t/\mathcal{D}_t) \\
&= w + \gamma\, f_{t/t} + \alpha\, f_{t/t}^2 + \delta\, h_{t/t} \\
\\
\widetilde{h}_{t+2/t} &= w + \gamma\, \mathbb{E}(\mathbf{f}_{t+1}/\mathcal{D}_t) + \alpha\, \mathbb{E}(\mathbf{f}_{t+1}^2/\mathcal{D}_t) + \delta\, \mathbb{E}(h_{t+1}/\mathcal{D}_t) \\
&= w + (\alpha + \delta)\widetilde{h}_{t+1/t} \\
&\;\;\vdots
\end{aligned}$$

et pour $s > 2$, on a :

$$\widetilde{h}_{t+s/t} = w \left[\sum_{i=0}^{s-2} (\alpha + \delta)^i \right] + (\alpha + \delta)^{s-1} \widetilde{h}_{t+1/t} \tag{3.36}$$

Tab. 3.4 – Modèle à trois facteurs communs.

	$\mathbf{X}$			$diag(\mathbf{\Psi})$	γ_1	γ_2
	1.0000	4.0000	0.0000	1.0000	0.2000	0.1000
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨0.1000⟩	⟨0.1000⟩
	2.0000	4.0000	0.0000	2.0000	0.2000	0.2000
	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨1.0000⟩	⟨0.1000⟩	⟨0.1000⟩
	3.0000	0.0000	0.0000	3.0000	0.2000	0.3000
Paramètres de simulation	⟨1.0000⟩	⟨0.0000⟩	⟨0.0000⟩	⟨2.0000⟩	⟨0.1000⟩	⟨0.1000⟩
	0.0000	0.0000	1.0000	4.0000	0.6000	0.5000
	⟨0.0000⟩	⟨0.0000⟩	⟨1.0000⟩	⟨2.0000⟩	⟨0.2000⟩	⟨0.2000⟩
	0.0000	5.0000	2.0000	5.0000		
	⟨0.0000⟩	⟨2.0000⟩	⟨1.0000⟩	⟨1.0000⟩		
	0.0000	5.0000	3.0000	6.0000		
	⟨0.0000⟩	⟨2.0000⟩	⟨1.0000⟩	⟨2.0000⟩		
	1.0084	3.8655	0.0812	1.1023	0.2301	0.0906
	(0.0399)	(0.0011)	(0.1140)	(0.0406)	(0.0307)	(0.0573)
	1.9658	3.9186	0.0098	1.9851	0.2061	0.1889
	(0.0855)	(0.0019)	(0.1413)	(0.0354)	(0.0423)	(0.0423)
	3.0252	0.0179	0.1024	2.9779	0.2109	0.2879
Estimation des paramètres	(0.0381)	(0.0032)	(0.0563)	(0.0357)	(0.0455)	(0.0047)
	0.0640	0.0951	1.1241	3.9479	0.5789	0.4877
	(0.0382)	(0.0062)	(0.0288)	(0.0381)	(0.0650)	(0.1108)
	0.0982	4.8842	1.9541	4.9116		
	(0.0368)	(0.0073)	(0.1866)	(0.0466)		
	0.1424	4.9369	3.1478	6.1232		
	(0.0367)	(0.0092)	(0.0011)	(0.0388)		

(.) Coefficients de variation, ⟨.⟩ Paramètres d'initialisation

Tab. 3.5 – Sélection de Modèles.

		FA1	FA2	FA3	CHF1	CHF2
	Expérience 1	0	0	2	86	12
AIC	Expérience 2	0	0	0	7	93
	Expérience 3	0	100	0	0	0
		0	0	2	91	7
BIC		0	0	0	7	93
		0	100	0	0	0

sous la contrainte $\alpha + \delta < 1$ (condition de stationnarité), on a aussi

$$\lim_{s\to\infty} \widetilde{h}_{t+s/t} \sim \frac{w}{1-\alpha-\delta}$$

Étant données les prévisions de la variance conditionnelle des facteurs communs, nous pouvons calculer des prévisions pour les variances conditionnelles des actifs individuels $\widehat{\sigma}^2_{i,t+s/t}$, les éléments de la diagonale de $\mathbf{\Sigma}_{t+s/t}$.

Méthodes de Prévision Alternatives

Afin de tester le pouvoir prédictif du modèle à facteurs conditionnellement hétéroscédastique nous l'avons mis en competition avec une méthode de prévision naive, un modèle à facteurs standard et des modèles GQARCH univariés. La méthode de prévision

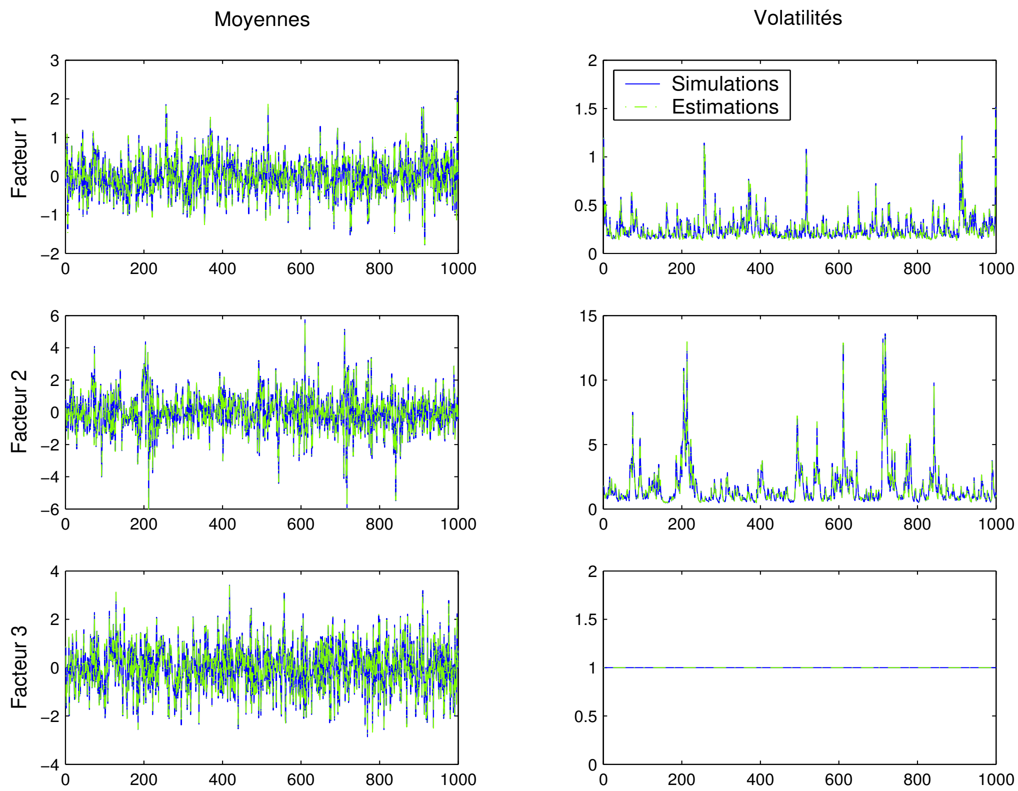

FIG. 3.7 – Modèle à trois facteurs : Les facteurs et leurs volatilités.

naive est basée, tout simplement, sur la moyenne historique de toutes les observations donnée par :

$$\widehat{v}_{it}^2 = \frac{1}{t}\sum_{j=1}^{t}(y_{ij} - \overline{y}_{it})^2 \qquad \text{avec} \qquad \overline{y}_{it} = \frac{1}{t}\sum_{j=1}^{t} y_{ij}$$

Le fait stylisé à l'origine du modèle GQARCH est que la volatilité du rendement des actifs financiers évolue de manière prévisible. Dans le modèle retenu, la volatilité conditionnelle d'un jour dépend de la volatilité de la veille, d'un terme représentant l'asymétrie entre volatilité et rendement et du carré du rendement observé la veille :

$$y_{it} = \theta_i + \sqrt{h_{it}}\,\varepsilon_{it} \quad \text{avec}$$

$$h_{it} = w_i + \alpha_i y_{it-1} + \gamma_i y_{it-1}^2 + \delta_i h_{it-1}$$

$$\text{pour } i = 1, ..., q$$

Pour chaque instant t et $\forall\ s = 1, ...,$ la volatilité anticipée est donnée par :

$$h_{it+s/t} = w_i + \theta_i(\alpha_i + \gamma_i\theta_i) + (\alpha_i + \gamma_i)h_{it+s-1/t}$$

Méthodes pour la Comparaison des Modèles de Prévision

Dans la littérature financière plusieurs critères ont été utilisés afin de comparer l'exactitude hors échantillon des modèles de prévision. Le critère le plus utilisé est celui de l'erreur carré moyenne ou sa racine carrée. Ce dernier prend en compte le carré de l'écart de la variance prévue par rapport à la variance observée. Ce critère désigné par (RMSE) utilisant N valeurs prévues est de la forme suivante :

$$\mathrm{RMSE}(\widehat{v}_i^2) = \sqrt{\sum_{t=1}^{N} \left(\widehat{v}_{it}^2 - v_{it}^2\right)^2}$$

Une autre méthode pour la mesure de la performance d'un modèle de prévision est basée sur l'écart absolu relatif par rapport à la vraie valeur. L'erreur absolue moyenne en pourcentage (MAPE) est calculée selon la formule suivante :

$$\mathrm{MAPE}(\widehat{v}_i^2) = \sum_{t=1}^{N} \frac{|\widehat{v}_{it}^2 - v_{it}^2|}{v_{it}^2}$$

Une mesure robuste contre la divergence vis à vis des hypothèses de normalité est donnée par la médiane des carrés des erreurs (MedSE), soit

$$\mathrm{MedSE}(\widehat{v}_i^2) = Mediane\ \left[\widehat{v}_{it}^2 - v_{it}^2\right]^2$$

Ces trois mesures seront comparées en utilisant un indice de performance inspiré de la théorie de la décision, i.e. le critère de Savage-Niehans :

$$\mathrm{Perf}_i = \sum_{j=1}^{q} \frac{\mathrm{EC}_i - \min_i \mathrm{EC}_i}{\min_i \mathrm{EC}_i}$$

où EC est l'un des critères d'erreur que nous avons décrit ci-dessus et q le nombre d'actifs que l'on veut prévoir le rendement. Cet indice peut être interprété comme une perte relative au niveau de l'exactitude des prévisions engendrée par l'un des modèles spécifiques en comparaison avec le meilleur modèle pour l'actif j.

Finalement, il faut noter l'intérêt de la méthode de régression des variances observées sur les variances prévues, soit la régression :

$$v_{it} = v + w\widehat{v}_{it} + \phi_{it}$$

permettant de tester si la constante v est égale à zéro et le coefficient de la ponte w est égale à un (ce qui résulte en une prévision non biaisée).

Applications

les estimations des différents modèles sont effectuées ici à l'aide d'une fenêtre mobile de 1000 observations. Après chacune, les prévisions de volatilité sont calculées pour les 5 jours suivants. L'échantillon est ensuite décalé de 5 observations et l'opération est recommencée pour l'estimation et l'algorithme de prévision.

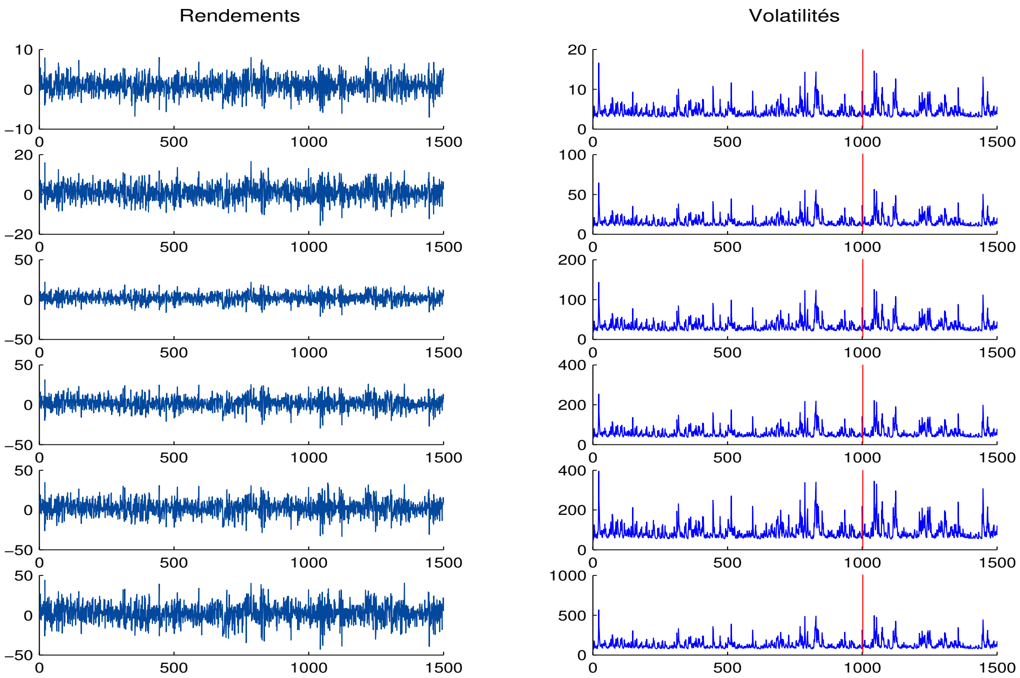

FIG. 3.8 – Rendements et volatilités simulés. La ligne verticale représente la date de commencement des prévisions.

Jeux de Données Simulées : Dans cette première application nous avons simulé un modèle à un seul facteur conditionnellement hétéroscédastique (CHFA), avec $q = 6$ variables observées, $k = 1$ et $n = 1500$. Sur cette même base de données, nous avons estimé un modèle à un seul facteur standard (FA M.), des modèles GQARCH(1,1) univariés pour chacune des séries. Ces modèles ont été utilisés par la suite pour calculer la volatilité anticipée pour chaque série. Sur la même base de données, nous avons appliqué aussi la méthode de prévision naive. Les paramètres de cette simulation sont donnés dans le tableau 3.1, en ajoutant à cette spécification une moyenne $\theta = [1\ 1\ 2\ 2\ 3\ 3]'$. Les séries de rendements simulés et leurs volatilités sont données dans la figure 3.8. Les résultats pour les différents critères et les différents modèles sont donnés dans la figure 3.9 et le tableau 3.5. Ce tableau montre que les différents critères de comparaison sont en faveur du vrai modèle (le modèle CHFA qui est à la base des simulations).

Rendements des taux de change : Dans cette application nous avons considéré les rendements journaliers des cours en valeurs (évalués par rapport à la livre sterling) du Dollar Américain (USD), le Dollar Canadien (CAD), le Franc Français (FRF), la Lire Italienne (ITL), le Deutsche Mark (DEM) et le Yen Japonais (JPY)[9]. Les données s'étalent sur la période 03/01/1983 à 22/12/1988 inclue. Pour le calcul des rendements, nous avons utilisé la formule des rendements composés continus :

$$r_t = \log p_t - \log p_{t-1} \approx \frac{p_t - p_{t-1}}{p_{t-1}}$$

[9] PACIFIC EXCHANGE RATE SERVICE, Sauder School of Business, http ://fx.sauder.ubc.ca/.

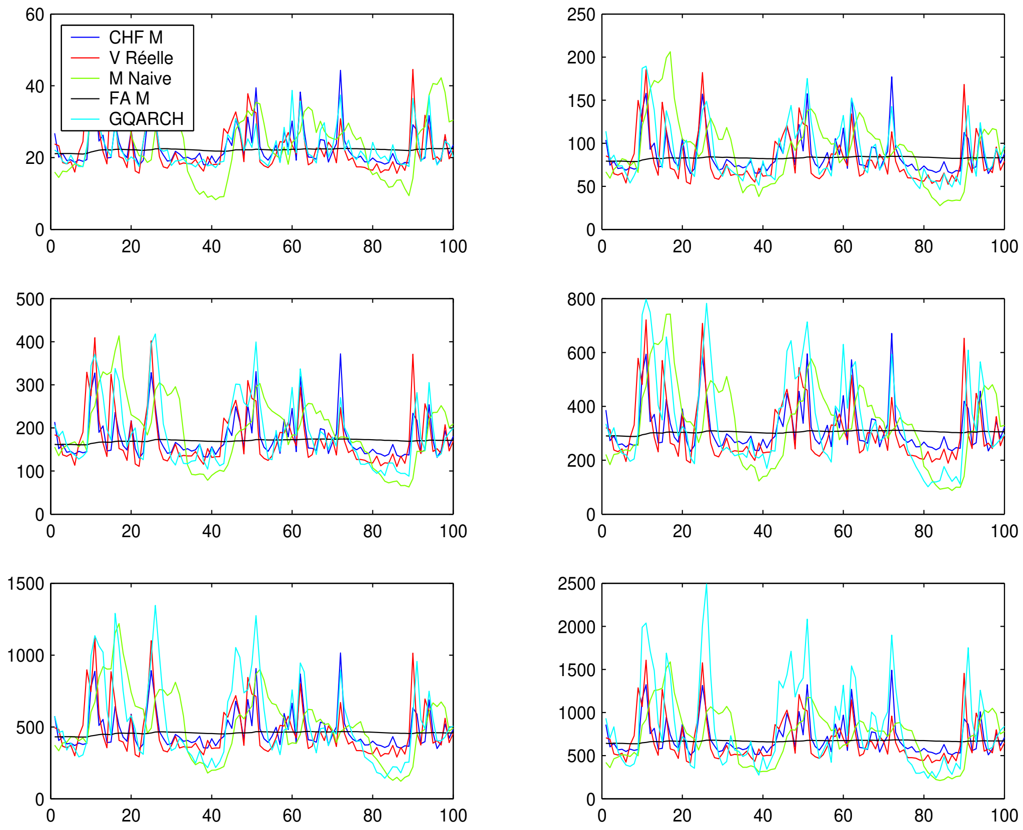

FIG. 3.9 – Simulations : Prévision de la volatilité par les différents modèles.

où p_t est le cours de cloture du taux de change journalier à la date t. La figure 3.10 représente l'évolution des cours et leurs rendements (soit 1500 observations).

Les résultats pour les différents critères sont donnés dans le tableau 3.6 et la figure 3.11. Quant à la volatilité réelle des rendements et étant donné qu'elle est effectivement non observée, pour son calcul nous avons utilisé l'approximation suivante :

$$v_{it}^2 = \sum_{j=t}^{t+4} (y_{ij} - \overline{y}_{it})^2 \qquad \text{avec} \qquad \overline{y}_{it} = \frac{1}{5} \sum_{j=t}^{t+4} y_{ij}$$

Tous les résultats sont aussi en faveur du modèle à facteurs avec hétéroscédasticité dynamique. La régression $v_{it} = v + w\widehat{v}_{it} + \phi_{it}$, où $\widehat{v}_{it}$ sont les volatilités anticipées calculées en utilisant le modèle CHFA, montre que les t de Student sont inférieurs à 1.96 pour le coefficient w : dans ce cas on ne rejette pas $H_0 : w = 1$ pour un risque $\alpha = 5\%$. Cependant pour le coefficient v, l'hypothèse $H_0 : v = 0$ est rejetée pour les séries FRF et ITL. Les valeurs de t pour $\alpha = 5\%$ sont données dans le tableau 3.7.

TAB. 3.6 – Résultats pour les différents critères

	Simulations				Données réelles			
	CHF	FAM	ARCH	Naive	CHF	FAM	ARCH	Naive
RMSE	49.524	71.856	56.516	117.04	2.3799	3.4338	3.8784	3.3223
	205.38	286.76	289.82	430.19	2.7570	3.7125	4.3456	3.6779
	448.39	648.34	646.29	925.85	0.8694	1.8379	1.7691	1.7397
	804.96	1146.2	1457.3	1736.2	0.9045	1.9847	1.7921	1.8752
	1247.8	1801.2	2424.8	2731.8	0.9364	1.7806	1.8533	1.6989
	1804.0	2575.5	4331.6	3675.6	1.2495	2.5172	2.5195	2.3647
MAPE	13.981	21.425	15.694	41.432	92.626	218.93	231.83	205.03
	18.887	27.245	25.685	43.812	99.205	186.13	253.12	171.02
	16.820	25.712	26.658	44.093	178.90	770.16	585.46	726.34
	18.822	27.920	35.637	47.737	142.38	680.28	463.00	640.85
	17.596	26.830	37.682	48.271	526.86	1493.6	1474.7	1412.6
	19.057	28.413	45.224	46.762	221.10	591.55	597.99	556.71
MedSE	5.7182	15.316	6.2975	56.142	0.0119	0.0843	0.0909	0.0745
	176.56	376.06	236.32	762.87	0.0196	0.0584	0.1002	0.0480
	579.67	1335.7	1131.6	2844.0	0.0011	0.0334	0.0171	0.0288
	2175.0	5157.1	5642.3	10432	0.0010	0.0397	0.0193	0.0362
	4363.4	10358	17383	24331	0.0008	0.0269	0.0186	0.0233
	10408	25142	58246	53416	0.0032	0.0664	0.0540	0.0549
Perf	0.0000	2.5881	4.1484	6.9062	0.0000	5.0138	5.2176	4.5110
	0.0000	3.0026	4.4752	9.6376	0.0000	12.803	11.082	11.697
	0.0000	8.2730	10.566	28.550	0.0000	128.13	82.130	112.23

TAB. 3.7 – Tests par la régression.

	Devises					
	USD	CAD	FRF	ITL	DEM	JPY
v	1.0983	1.8997	-2.1191	-2.5736	-1.5184	-0.3288
w	-1.2083	-1.9105	1.3884	1.7341	0.9615	-0.3872

3.6 Conclusion

Dans ce chapitre, nous avons discuté l'estimation d'une classe de modèles à facteurs conditionnellement hétéroscédastiques par le maximum de vraisemblance à information complète. Nous avons déterminé la fonction de vraisemblance et le score, aussi bien que les conditions du premier ordre de Kuhn-Tucker en utilisant des contraintes d'inégalité sur les paramètres du modèle, permettant de garantir la positivité des variances idiosyncratiques par l'algorithme d'optimisation. Nous avons expliqué, par la suite, l'application d'un algorithme EM conditionnel pour l'estimation de l'ensemble des paramètres du modèle. Cet algorithme est basé sur une version modifiée du filtre de Kalman permettant d'obtenir les meilleurs (dans le sens de l'erreur quadratique moyenne) estimations pour les facteurs non observables et leurs variances. Les simulations que nous avons effectué ont permis de souligner que la convergence se fait presque toujours à un maximum local. Toutefois, cette convergence paraît un peut lente vue la

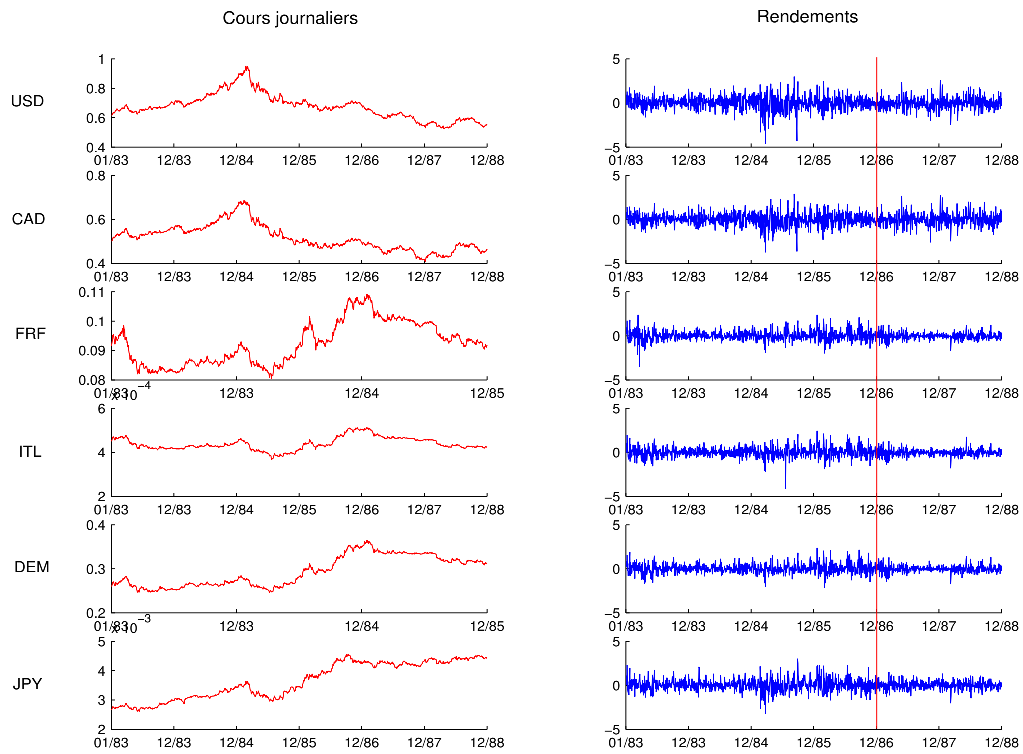

Fig. 3.10 – Les cours journaliers et leurs rendements. La ligne verticale représente la date de commencement des prévisions.

quantité plus où moins importante d'information manquante. Finalement, nous avons présenté trois algorithmes numériquement efficaces permettant de calculer la fonction de vraisemblance, son gradient, et les meilleures estimations filtrées des facteurs.

Le modèle qui sera présenté dans le chapitre cinq est une prolongation de plusieurs idées déjà présentées ici. Il tentera de construire une structure factorielle conditionnellement hétéroscédastique avec des paramètres variables. En particulier, nous allons considérer le cas où la dynamique de ces paramètres est gouvernée par une variable non observable que l'on peut modéliser à l'aide d'une chaîne de Markov cachée à m régimes. Dans ce cas le filtre de Kalman doit être modifiée encore afin de tenir compte du caractère aléatoire de la nouvelle variable d'état markovien. Cette nouvelle spécification va donc nous permettre de modéliser simultanément la dynamique de la volatilité conditionnelle des facteurs communs, et la dynamique de l'ensemble des paramètres du modèle afin de tenir compte des éventuels changements de régime qui peuvent affecter les séries à caractère économique et financier.

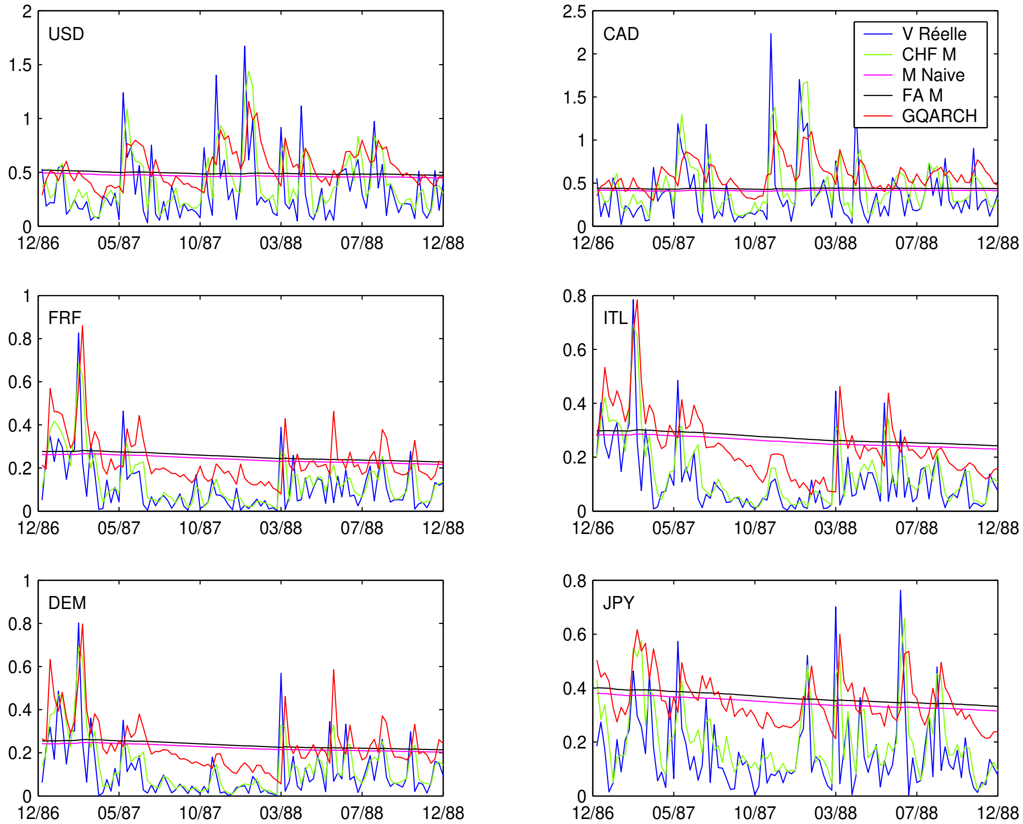

FIG. 3.11 – Taux de change : Prévision de la volatilité par les différents modèles.

3.7 Annexe : La Formule de Woodbury Généralisée

Soient $\mathbf{A}_{q\times q}$, $\mathbf{B}_{q\times k}$, $\mathbf{L}_{k\times k}$ et $\mathbf{D}_{q\times k}$ des matrices complexes, et soit

$$\mathbf{E}_{q\times q} = \mathbf{A} + \mathbf{BLD}^H \tag{3.37}$$

où $\mathbf{D}^H$ est le transposé conjugué de $\mathbf{D}$. À condition que les inverses nécessaires existent, la formule de Woodbury implique :

$$\mathbf{E}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BF}^{-1}\mathbf{D}^H\mathbf{A}^{-1} \tag{3.38}$$

$$|\mathbf{E}| = |\mathbf{A}| \cdot |\mathbf{L}| \cdot |\mathbf{F}| \tag{3.39}$$

où

$$\mathbf{F}_{k\times k} = \mathbf{L}^{-1} + \mathbf{D}^H\mathbf{A}^{-1}\mathbf{B}$$

(voir Householder [1964]. Lorsque $k = 1$, la formule de Woodbury est équivalente à celle de Sherman-Morrison. Ces formules sont particulièrement utiles lorsque la valeur de $\mathbf{A}^{-1}$ est déjà calculée et $k < q$, $\mathbf{F}$ sera donc facile à inverser.

Étant donné que (3.38) implique $\mathbf{F}^{-1} = \mathbf{L} - \mathbf{LD}^H\mathbf{E}^{-1}\mathbf{BL}$, donc l'inverse de $\mathbf{F}$ existe si et seulement si l'inverse de $\mathbf{E}$ existe aussi. Une considération particulière sera donnée au cas où $\mathbf{A}$ et $\mathbf{L}$ sont singulières. En effet, la décomposition en valeurs singulières de ces matrices est donnée par :

$$\mathbf{A} = \mathbf{U}\Delta\mathbf{V}^H = [\mathbf{U}_1 \;\; \mathbf{U}_2] \begin{bmatrix} \Delta_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^H \\ \mathbf{V}_2^H \end{bmatrix}$$

$$\mathbf{L} = \mathbf{P}\Lambda\mathbf{Q}^H = [\mathbf{P}_1 \;\; \mathbf{P}_2] \begin{bmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1^H \\ \mathbf{Q}_2^H \end{bmatrix}$$

avec $\mathbf{U}$, $\mathbf{V}$, $\mathbf{P}$, $\mathbf{Q}$ des matrices unitaires, $\Delta_1\,, \Lambda_1 > 0$, $rang(\Delta_1) = q_1$ et $rang(\Lambda_1) = k_1$. Il est claire pour que $\mathbf{A} + \mathbf{BLD}^H$ soit de plein rang q, il faut satisfaire la condition nécessaire : $q_1 + k_1 \geq q$. Maintenant si on désigne par

$$\mathbf{R} = \mathbf{U}^H\mathbf{BP} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \quad \text{et} \quad \mathbf{S} = \mathbf{V}^H\mathbf{DQ} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

et si on utilise par la suite la formule de l'inverse des matrices par bloc (à condition que les inverses nécessaires existent), on aura :

$$
\begin{aligned}
\mathbf{E}^{-1} &= \left[\mathbf{U}\left(\Delta + \mathbf{R}\Lambda_1\mathbf{S}^H\right)\mathbf{V}^H\right]^{-1} \\
&= \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \begin{bmatrix} \Delta_1 + \mathbf{R}_{11}\Lambda_1\mathbf{S}_{11}^H & \mathbf{R}_{11}\Lambda_1\mathbf{S}_{21}^H \\ \mathbf{R}_{21}\Lambda_1\mathbf{S}_{11}^H & \mathbf{R}_{21}\Lambda_1\mathbf{S}_{21}^H \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^H \\ \mathbf{U}_2^H \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} \\
&\quad \begin{bmatrix} \mathbf{E}_{11}^{-1} + \mathbf{E}_{11}^{-1}\mathbf{R}_{11}\Lambda_1\mathbf{S}_{21}^H\mathbf{G}^{-1}\mathbf{R}_{21}\Lambda_1\mathbf{S}_{11}^H\mathbf{E}_{11}^{-1} & -\mathbf{E}_{11}^{-1}\mathbf{R}_{11}\Lambda_1\mathbf{S}_{21}^H\mathbf{G}^{-1} \\ -\mathbf{G}^{-1}\mathbf{R}_{21}\Lambda_1\mathbf{S}_{11}^H\mathbf{E}_{11}^{-1} & \mathbf{G}^{-1} \end{bmatrix} \\
&\quad \begin{bmatrix} \mathbf{U}_1^H \\ \mathbf{U}_2^H \end{bmatrix} = \mathbf{V}_1\mathbf{E}_{11}^{-1}\mathbf{U}_1^H \\
&\qquad + \left(\mathbf{V}_2 - \mathbf{V}_1\mathbf{E}_{11}^{-1}\mathbf{R}_{11}\Lambda_1\mathbf{S}_{21}^H\right)\mathbf{G}^{-1}\left(\mathbf{U}_2^H - \mathbf{R}_{21}\Lambda_1\mathbf{S}_{11}^H\mathbf{E}_{11}^{-1}\mathbf{U}_1^H\right)
\end{aligned}
$$

où

$$
\begin{aligned}
\mathbf{G} &= \mathbf{R}_{21}\Lambda_1\mathbf{S}_{21}^H - \mathbf{R}_{21}\Lambda_1\mathbf{S}_{11}^H\mathbf{E}_{11}^{-1}\mathbf{R}_{11}\Lambda_1\mathbf{S}_{21}^H \\
\mathbf{E}_{11} &= \Delta_1 + \mathbf{R}_{11}\Lambda_1\mathbf{S}_{11}^H
\end{aligned}
$$

Une application répétée de la formule de Woodbury standard

$$
\begin{aligned}
\mathbf{E}_{11}^{-1} &= \Delta_1^{-1} - \Delta_1^{-1}\mathbf{R}_{11}\mathbf{F}_{11}^{-1}\mathbf{S}_{11}^H\Delta_1^{-1} \\
\mathbf{F}_{11} &= \Lambda_1^{-1} + \mathbf{S}_{11}^H\Delta_1^{-1}\mathbf{R}_{11}
\end{aligned}
$$

donne après quelques simplifications

$$
\begin{gathered}
\mathbf{E}^{-1} = \mathbf{V}_1\left[\Delta_1^{-1} - \Delta_1^{-1}\mathbf{R}_{11}\mathbf{F}_{11}^{-1}\mathbf{S}_{11}^H\Delta_1^{-1}\right]\mathbf{U}_1^H \\
+ \left(\mathbf{V}_2 - \mathbf{V}_1\Delta_1^{-1}\mathbf{R}_{11}\mathbf{F}_{11}^{-1}\mathbf{S}_{21}^H\right)\left[\mathbf{R}_{21}\mathbf{F}_{11}^{-1}\mathbf{S}_{21}^H\right]^{-1}\left(\mathbf{U}_2^H - \mathbf{R}_{21}\mathbf{F}_{11}^{-1}\mathbf{S}_{11}^H\Delta_1^{-1}\mathbf{U}_1^H\right)
\end{gathered} \tag{3.40}
$$

Notons ici que si $m_1 = m$ et $k_1 = k$, le second terme de (3.40) disparaît, et le premier sera équivalent à (3.38). D'une manière similaire, la formule de calcul des déterminants des matrices par bloc et la formule (3.39) donnent :

$$
|\mathbf{E}| = |\mathbf{E}_{11}| \cdot |\mathbf{G}| = |\Delta_1| \cdot |\Lambda_1| \cdot |\mathbf{F}_{11}| \cdot |\mathbf{G}| \tag{3.41}
$$

Cette généralisation peut se faire à travers une factorisation alternative de $\mathbf{A}$ et $\mathbf{L}$, qui peut être plus convaincante si ces matrices ont une forme bien particulière. Dans le cas des modèles à facteurs conditionnellement hétéroscédastiques, la matrice de covariance $\mathbf{\Sigma}_t$ est de la forme (3.37), avec $\mathbf{A} = \mathbf{\Psi}$, $\mathbf{B} = \mathbf{D} = \mathbf{X}$ et $\mathbf{L} = \mathbf{H}_t$. La formule de Woodbury peut donc être appliquée directement pour calculer la log-vraisemblance et le score. Toutefois, il est difficile de démontrer directement que $\mathbf{\Psi}$ est de plein rang, les expressions (3.29) et (3.31) pour $\mathcal{L}(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta)$ et $\ell(\mathbf{y}_t/\mathcal{Y}_{t-1};\Theta)$ peuvent être obtenues algébriquement en utilisant (3.38) et (3.39) avec $\mathbf{F} = \mathbf{H}_t^{-1} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}$. Enfin, nous pouvons démontrer aussi que les expressions dérivées dans la section (3.4.3) coincident avec celles obtenues en appliquant la formule de Woodbury modifiée (3.40) et (3.41), où $\mathbf{P} = \mathbf{Q} = \mathbf{I}_k$, $\mathbf{U} = \mathbf{V}$ est une matrice de permutation qui déplace les séries avec des ψ_i nulles aux dernières positions, et $\mathbf{G} = \mathbf{\Sigma}_{b,at}$.

CHAPITRE 4

Systèmes Dynamiques à Structure Markovienne Cachée

Ce chapitre présente les notions nécessaires pour l'élaboration et l'estimation du modèle factoriel dynamique à états-mixtes que nous présentons dans le chapitre 5. Dans une première section, nous définissons les modèles de Markov cachés. Nous présentons ensuite les algorithmes d'inférence dans le cas des chaînes homogènes. Une méthode itérative basée sur l'algorithme EM pour l'estimation de maximum de vraisemblance de ces modèles sera aussi discutée. Dans une deuxième section, nous étudions les modèles espace-état linéaires. Les algorithmes de filtrage et lissage seront développés en deux versions différentes, et enfin un algorithme EM pour l'estimation des paramètres sera proposé. Une structure dynamique hybride plus générale tenant compte de la possibilité de changement de régime sera présentée dans la dernière section. Cette nouvelle spécification est construite par une combinaison des modèles espace-état linéaires avec les modèles de Markov cachés. Pour l'inférence des structures cachées et l'estimation des paramètres, un algorithme EM basé sur une version quasi-optimale du filtre de Kalman combiné avec une méthode pseudo-bayésienne généralisée sera présenté.

4.1 Les Chaînes de Markov Cachées

Les modèles à données latentes (ou manquantes ou cachées) constituent des outils puissants pour modéliser des systèmes dont la dynamique effectue des transitions entre différents états impossible à observer directement. L'étude de ces modèles a réellement débuté dans les années soixante, par l'analyse des modèles d'états linéaires gaussiens, qui a suscité un engouement fort dans la communauté automatique et traitement du signal. En parallèle se sont développées dès la fin des années soixante des études sur les Modèles de Markov cachés à états discrets (le processus latent étant une chaîne de Markov prenant ses valeurs dans un ensemble fini d'états).

Ces modèles ont connu un vif succès tant en traitement de parole (les HMM[1] forment l'élément de base des systèmes de reconnaissance de parole) qu'en bioinformatique (où les HMM sont utilisés pour la segmentation et le séquençage de génomes). Dans la littérature économétrique ces modèles ont été introduits par Hamilton [1989] afin de prendre en compte un certain type de non stationnarité présente dans de nombreuses séries à caractère économique et financier.

4.1.1 Définition

Dans une chaîne de Markov cachée, les différents états d'un système peuvent être caractérisés par un nombre fini de valeurs. Les transitions entre les états se produisent entre deux instants discrets consécutifs, selon une certaine loi de probabilité. La probabilité de chaque état ne dépend que de l'état qui le précède immédiatement. Un modèle HMM représente de la même façon qu'une chaîne de Markov un ensemble de séquences d'observations dont l'état de chacune n'est pas observé, mais associé à une fonction de densité de probabilité. Il s'agit donc d'un processus doublement stochastique, dans lequel les observations sont une fonction aléatoire de l'état et dont l'état change à chaque instant en fonction des probabilités de transition issues de l'état antérieur. Ce modèle partage donc avec le modèle de mélange la caractéristique essentielle de faire intervenir une structure sous-jacente (non observable) sous la forme d'une variable indicatrice (ou étiquette), associée à chaque observation, et prenant un nombre fini de valeurs. Le modèle HMM est toutefois plus riche que le modèle de mélange dans le sens où il permet de rendre compte des interactions temporelles en substituant à l'hypothèse d'indicatrices *iid* celle d'une évolution markovienne.

Plus précisément, on dira qu'un processus aléatoire $\{\mathcal{Y}_t\}_{t\geq 1}$ (éventuellement vectoriel) a une structure de modèle HMM, si il existe un processus aléatoire $\{\mathcal{S}_t\}_{t\geq 1}$ (défini sur le même espace de probabilité), prenant un nombre fini m de valeurs, tel que :

1. Les indicatrices S_t ont une évolution markovienne "homogène" (c-à-d indépendante de l'indice temporel)
$$p(S_2/S_1) = p(S_t/\mathcal{S}_{1:t-1}) = p(S_t/S_{t-1})$$
où la notation $\mathcal{S}_{1:t-1}$ désigne la séquence $\{S_1, S_2, ..., S_{t-1}\}$.
2. Les observations $\mathbf{y}_t$ sont indépendantes conditionnellement aux indicatrices S_t.
$$p(\mathcal{Y}_{1:n}/\mathcal{S}_{1:n}) = \prod_{t=1}^{n} p(\mathbf{y}_t/S_t)$$

La manière usuelle de paramétrer un tel modèle consiste

1. pour la partie markovienne, à spécifier la distribution initiale $\pi = [\pi_1, \pi_2, ..., \pi_m]'$ où $\pi_i = p(S_1 = i)$ et la matrice de transition $\mathbf{P}$

[1] Dans la suite nous utiliserons plutôt l'appellation anglaise de HMM pour Hidden Markov Model qui est la plus largement employée.

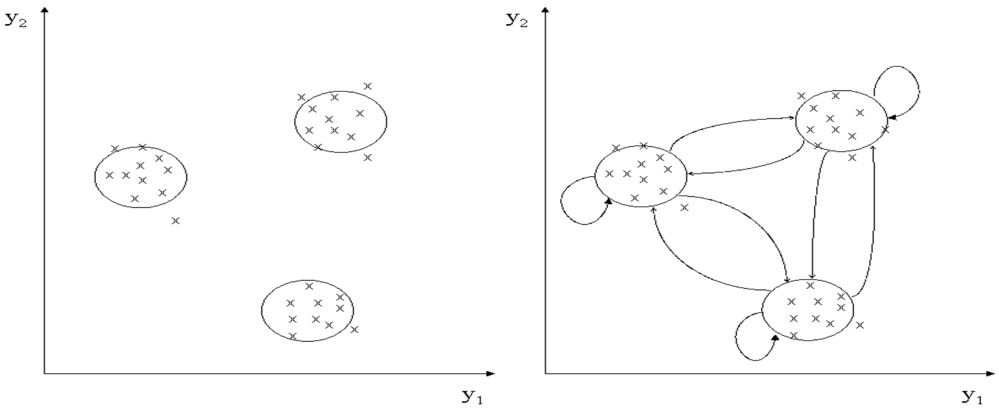

FIG. 4.1 – Du Modèle de Mélange aux Modèles HMM

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m-1} & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m-1} & p_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{m-11} & p_{m-12} & \cdots & p_{m-1m-1} & p_{m-1m} \\ p_{m1} & p_{m2} & \cdots & p_{mm-1} & p_{mm} \end{bmatrix}$$

où $p_{ij} = p(S_{t+1} = j|S_t = i)$. On note que la matrice $\mathbf{P}$ possède une structure particulière, dite de matrice stochastique, pour laquelle $\sum_{j=1}^{m} p_{ij} = 1$ (pour toutes les lignes). Ainsi, on suppose que la probabilité d'un état à une période t quelconque dépend seulement de l'état choisi à l'instant $t-1$.

2. pour la partie observation, la loi $b_j(\mathbf{y}_t) = p(\mathbf{y}_t|S_t = j)$ appartient en général à une même famille paramétrique de paramètre Θ_j. Dans le cas particulier qui va nous retenir, celui des HMM conditionnellement gaussiens, la loi $b_j(\mathbf{y}_t)$ est une loi normale multivariée paramétrée par son vecteur moyen μ_j et sa matrice de variance-covariance $\mathbf{\Sigma}_j$.

La flexibilité de ces modèles les rend très intéressants en pratique, mais la présence de variables cachées complique l'inférence statistique. Le calcul, pour un modèle entièrement spécifié, de la loi jointe d'un ensemble de variables observées (typiquement le calcul de la vraisemblance), nécessite des algorithmes de complexité au mieux polynômial, au pire exponentielle en fonction du nombre de variables du modèle. L'estimation des paramètres basée sur la vraisemblance est rendue difficile par l'absence de formule explicite pour le maximum de vraisemblance, et requiert en général des algorithmes itératifs, comme l'algorithme EM. De plus, dans ce cas, chaque itération met elle-même en jeu les algorithmes de calcul de probabilités évoqués ci-dessus.

4.1.2 Le Modèle Graphique

Les modèles graphiques sont le mariage entre la théorie des probabilités et celles des graphes. Ils fournissent des outils intuitifs et naturels pour traiter des problèmes dans lesquelles l'incertitude et la complexité des données jouent un rôle important. L'idée fondamentale des modèles graphiques est la modularité : un système complexe est construit en combinant des parties plus simples. La théorie des probabilités combine alors ces parties assurant une cohérence à l'ensemble du système.

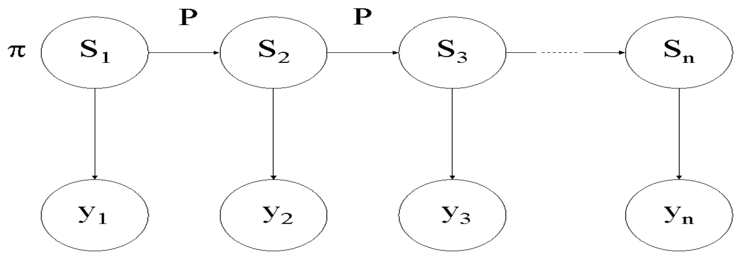

FIG. 4.2 – Représentation graphique d'un HMM. Chaque morceau vertical représente une période bien déterminée. Le noeud en haut de chaque morceau représente la variable multinomiale S_t et le noeud en bas représente les variables observées $\mathbf{y}_t$.

La structure d'une chaîne de Markov cachée d'ordre 1 est définie par le graphe d'indépendance conditionnelle de la figure 4.2. Les sommets de ce graphe sont les variables aléatoires S_t (prenant leurs valeurs dans un ensemble discret) et $\mathbf{y}_t$. Le graphe est toujours orienté et sans cycle. Les arcs orientés représentent un lien de dépendance directe (lien de causalité). Ainsi un arc allant de A à B exprimera le fait que B dépend directement de A. L'absence d'arc ne renseigne alors que sur la non-existence d'une dépendance directe. Les paramètres exprimant le poids donné à ces relations sont les probabilités conditionnelles des variables sachant leurs parents (exemple : $p(B|A)$) ou les probabilités a priori si la variable n'a pas de parents.

En se basant sur cette structure graphique, certaines propriétés fort utiles des HMM pourront être établies. Notons tout d'abord que

$$p(\mathbf{y}_t, S_t|\mathbf{y}_{t-1}, S_{t-1}) = \frac{p(\mathbf{y}_{t-1:t}|S_{t-1:t})p(S_{t-1:t})}{p(\mathbf{y}_{t-1}|S_{t-1})p(S_{t-1})} = p(\mathbf{y}_t|S_t)p(S_t|S_{t-1})$$

et que de même

$$p(\mathbf{y}_t, S_t|\mathbf{y}_t, S_t) = \frac{p(\mathbf{y}_{1:t}|S_{1:t})p(S_{1:t})}{p(\mathbf{y}_{1:t-1}|S_{1:t-1})p(S_{1:t-1})} = p(\mathbf{y}_t|S_t)p(S_t|S_{t-1})$$

c'est à dire que le processus joint $\{\mathbf{y}_t, S_t\}$ est markovien homogène, tout comme $\{S_t\}$, à la seule différence que son espace d'états (l'espace dans lequel il prend ses valeurs) n'est pas fini. Par contre, il est important de garder à l'esprit le fait que le processus observé $\{\mathbf{y}_t\}$ seul n'est pas markovien puisque

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \sum_{i=1}^{m} p(\mathbf{y}_t, S_t = i|\mathbf{y}_{1:t-1}) = \sum_{i=1}^{m} p(\mathbf{y}_t|S_t = i)p(S_t = i|\mathbf{y}_{1:t-1})$$

Cette dernière équation montre bien que la loi de $\mathbf{y}_t$ conditionnellement à son passé est un modèle de mélange dont les poids $p(S_t = i|\mathbf{y}_{1:t-1})$ dépendent du passé complet du signal (et pas seulement de $\mathbf{y}_{t-1}$).

Nous rappelons par ailleurs un résultats classique des processus markoviens qui est que

$$p(f(S_{t_1:t_2}), h(S_{t_4:t_5})|g(S_{t_3})) = p(f(S_{t_1:t_2})|g(S_{t_3}))p(h(S_{t_4:t_5})|g(S_{t_3}))$$

dès que $t_1 \leq t_2 \leq t_3 < t_4 \leq t_5$ ou $t_1 \leq t_2 < t_3 \leq t_4 \leq t_5$ (où f, g et h sont des fonctions mesurables). Ce qu'on résume souvent en disant que le passé et le future d'une chaîne de Markov sont conditionnellement indépendant lorsque l'on conditionne par rapport au point courant.

4.1.3 Le Problème d'Inférence

Étant donnée une suite d'observations $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ et un modèle $\Theta = (\pi, \mathbf{P}, \mu, \mathbf{\Sigma})$, comment peut-on calculer efficacement la probabilité que la suite d'observations $\mathcal{Y}$ soit produite par Θ, c'est-à-dire $p(\mathcal{Y}/\Theta)$. Autrement dit, comment évaluer le modèle afin de choisir parmi plusieurs celui qui génère le mieux cette suite d'observations. Plusieurs techniques permettent de résoudre ce problème : méthode d'évaluation directe, procédure "Avant-Arrière" et algorithme de Viterbi.

Évaluation Directe

La probabilité $p(\mathcal{Y}/\Theta)$ d'une suite d'observations $\mathcal{Y}$, sachant qu'un modèle Θ est donné, est la somme sur tous les chemins d'états, $\mathcal{S}$ possibles des probabilités conjointes de $\mathcal{Y}$ et de $\mathcal{S}$ par rapport à ce modèle :

$$\begin{aligned}
p(\mathcal{Y}|\Theta) &= \sum_{s_1=1}^{m}\sum_{s_2=1}^{m}...\sum_{s_n=1}^{m} p(\mathcal{Y}, \mathcal{S}_{1:n} = s_{1:n}|\Theta) \\
&= \sum_{s_1=1}^{m}\sum_{s_2=1}^{m}...\sum_{s_n=1}^{m} p(\mathcal{Y}|\mathcal{S}_{1:n} = s_{1:n}, \Theta)p(\mathcal{S}_{1:n} = s_{1:n}|\Theta) \\
&= \sum_{S_1}\sum_{S_2}...\sum_{S_n} p(S_1)\prod_{t=1}^{n-1} p(S_{t+1}|S_t)\prod_{t=1}^{n} p(\mathbf{y}_t|S_t, \Theta)
\end{aligned}$$

1. Initialement à $t = 1$ l'état initial est S_1 avec une probabilité $p(S_1)$ et une observation $\mathbf{y}_1$ est générée avec une probabilité $p(\mathbf{y}_1|S_1, \Theta)$;
2. à $t = t + 1$, $(t = 2)$, une transition est effectuée à l'état S_2 à partir de l'état S_1 avec une probabilité de transition $p(S_2|S_1)$ et une observation $\mathbf{y}_2$ est générée avec une probabilité $p(\mathbf{y}_2|S_2, \Theta)$;

3. Ce processus continue de la même manière jusqu'à la dernière transition $t = n$ de l'état S_{n-1} à S_n avec une probabilité de transition $p(S_n|S_{n-1})$ et une observation $\mathbf{y}_n$ est générée avec une probabilité $p(\mathbf{y}_n|S_n, \Theta)$.

Pour calculer la probabilité $p(\mathcal{Y}/\mathcal{S}, \Theta)$ par cette méthode, il faut $(2n-1)m^n$ multiplications et $m^n - 1$ additions soit environ $2nm^n$ opérations. Cet ordre de calcul est non faisable même pour des petites valeurs de n et m. Par exemple, pour $m = 5$ et $n = 100$ on obtient environ 10^{72} opérations.

La procédure "Avant-Arrière"

Dans cette approche, on considère que l'observation peut se faire en deux étapes : d'abord, l'émission de la suite d'observations $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t$ et la réalisation de l'état i à la date t, puis l'émission de la suite d'observations $\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, ..., \mathbf{y}_n$ en partant de l'état $S_t = i$. Dans ce cas, l'évaluation de l'observation est

$$p(\mathcal{Y}|\Theta) = \sum_{i=1}^{m} \alpha_t(i)\beta_t(i)$$

où $\alpha_t(i)$ est la probabilité d'émettre la suite $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t$ et d'aboutir à l'état i à l'instant t sachant le modèle et $\beta_t(i)$ la probabilité d'émettre la suite $\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, ..., \mathbf{y}_n$ en partant de l'état i à l'instant t sachant le modèle. Le calcul de $\alpha_t(i)$ se fait avec t croissant tandis que celui de $\beta_t(i)$ se fait avec t décroissant, d'où l'expression Avant-Arrière.

Calcul de α Soit la variable Avant $\alpha_t(j)$

$$\alpha_t(j) = p(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t, S_t = j|\Theta), \quad 1 \leq j \leq m, \ 1 \leq t \leq n$$

1. Initialisation, $t = 1$

$$\alpha_1(i) = \pi_i p(\mathbf{y}_1|S_1 = i, \Theta), \quad i = 1, 2, ..., m$$

2. Induction

$$\begin{aligned}
\alpha_t(j) &= p(\mathbf{y}_1, ..., \mathbf{y}_t, S_t = j) \\
&= p(\mathbf{y}_1, ..., \mathbf{y}_t|S_t = j)p(S_t = j) \\
&= p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}|S_t = j)p(\mathbf{y}_t|S_t = j)p(S_t = j) \\
&= p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}, S_t = j)p(\mathbf{y}_t|S_t = j) \\
&= \sum_{i=1}^{m} p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}, S_{t-1} = i, S_t = j)p(\mathbf{y}_t|S_t = j) \\
&= \sum_{i=1}^{m} p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}, S_t = j|S_{t-1} = i)p(S_{t-1} = i)p(\mathbf{y}_t|S_t = j) \\
&= \sum_{i=1}^{m} p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}|S_{t-1} = i)p(S_t = j|S_{t-1} = i)p(S_{t-1} = i)p(\mathbf{y}_t|S_t = j) \\
&= \sum_{i=1}^{m} p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}, S_{t-1} = i)p(S_t = j|S_{t-1} = i)p(\mathbf{y}_t|S_t = j)
\end{aligned}$$

Ceci implique

$$\alpha_t(j) = \left[\sum_{i=1}^{m} \alpha_{t-1}(i) p_{ij}\right] p(\mathbf{y}_t|S_t = j, \Theta), \quad j = 1, 2, ..., m, \; t = 2, 3, ..., n$$

Cette étape montre comment l'état j peut être visité à la date $t+1$ à partir de m états possibles i, $1 \leq i \leq m$ à la date t.

3. Terminaison

$$p(\mathcal{Y}|\Theta) = \sum_{j=1}^{m} \alpha_n(j)$$

Pour calculer la probabilité de l'observation par cette méthode $m(m+1)(n-1)+m$ multiplications et $m(m-1)(n-1)$ additions, soit environ $m^2 n$ opérations sont effectuées. Par exemple, pour $m = 5$ et $n = 100$ on obtient environ 3000 opérations au lieu de 10^{72} opérations demandées par la méthode directe.

Calcul de β Soit la variable Arrière $\beta_t(i)$ définie par

$$\beta_t(i) = p(\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, ..., \mathbf{y}_n|S_t = i, \Theta), \quad 1 \leq i \leq m, \; 1 \leq t \leq n$$

1. Initialisation, $t = n$

$$\beta_n(i) = 1, \quad \forall i = 1, 2, ..., m$$

Cette étape définie arbitrairement $\beta_n(i) = 1$ pour tous les états i.

2. Induction

$$\begin{aligned}
\beta_t(i) &= p(\mathbf{y}_{t+1}, ..., \mathbf{y}_n|S_t = i) \\
&= \sum_{j=1}^{m} p(\mathbf{y}_{t+1}, ..., \mathbf{y}_n, S_{t+1} = j|S_t = i) \\
&= \sum_{j=1}^{m} p(\mathbf{y}_{t+1}, ..., \mathbf{y}_n|S_{t+1} = j, S_t = i) p(S_{t+1} = j|S_t = i) \\
&= \sum_{j=1}^{m} p(\mathbf{y}_{t+2}, ..., \mathbf{y}_n|S_{t+1} = j) p(\mathbf{y}_{t+1}|S_{t+1} = j) p(S_{t+1} = j|S_t = i)
\end{aligned}$$

Ceci implique

$$\beta_t(i) = \sum_{j=1}^{m} \beta_{t+1}(j) p_{ij} p(\mathbf{y}_{t+1}|S_{t+1} = j), \quad i = 1, 2, ..., m, \;\; t = n-1, n-2, ..., 1$$

Pour être dans l'état i à l'instant t, et pour tenir compte de la suite d'observations de $t+1$ à n, nous devons considérer tous les états possibles j (toutes les transitions p_{ij}) aussi bien que l'observation $\mathbf{y}_{t+1}$ dans l'état j (les $p(\mathbf{y}_{t+1}|S_{t+1}=j,\Theta)$), puis de tenir compte de la suite d'observations partielle restante à partir de l'état j ($\beta_{t+1}(j)$). Pour calculer la probabilité $p(\mathcal{Y}|\Theta)$ par cette méthode $m(m+1)(n-1)+m$ multiplications et $m(m-1)(n-1)$ additions soit environ m^2n opérations sont effectuées.

Les deux variables $\alpha_t(i)$ et $\beta_t(j)$ peuvent être utilisées pour calculer $p(\mathcal{Y}|\Theta)$ à chaque instant t, avec $1 \leq t \leq n$:

$$p(\mathcal{Y}|\Theta) = \sum_{i=1}^{m} \alpha_t(i)\beta_t(i)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_t(i) p_{ij} p(\mathbf{y}_{t+1}|S_{t+1}=j,\Theta)\beta_{t+1}(j)$$

Cette formule sera utilisée par la suite pour résoudre le problème d'estimation des paramètres.

4.1.4 Estimation de la Suite Cachée

Étant donnée une suite d'observations $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$, et un modèle Θ, comment peut-on choisir une suite d'états $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ qui soit optimale selon un critère convenable. La difficulté réside dans la définition de la suite optimale d'états, c'est-à-dire qu'il existe plusieurs critères d'optimalité possibles. Selon le choix du critère nous proposons trois solutions :

Estimation de l'état par les Probabilités de Lissage

Cette méthode consiste à choisir l'état S_t qui est le plus probable et ceci indépendamment des autres états, ce qui revient à choisir à l'instant t l'état i^* qui maximise $p(S_t = i|\mathcal{Y}, \Theta)$ pour $i = 1, ..., m$. En utilisant ce critère, il serait donc nécessaire de déterminer les probabilités a posteriori $\gamma_t(i)$, soit

$$\begin{aligned}
\gamma_t(i) &= p(S_t = i|\mathbf{y}_1, ..., \mathbf{y}_n), \quad i = 1, ..., m, \; t = 1, ..., n \\
&= \sum_{j=1}^{m} p(S_t = i, S_{t+1} = j|\mathbf{y}_1, ..., \mathbf{y}_n) \\
&= \sum_{j=1}^{m} p(S_t = i|S_{t+1} = j, \mathbf{y}_1, ..., \mathbf{y}_n) p(S_{t+1} = j|\mathbf{y}_1, ..., \mathbf{y}_n) \\
&= \sum_{j=1}^{m} p(S_t = i|S_{t+1} = j, \mathbf{y}_1, ..., \mathbf{y}_t) p(S_{t+1} = j|\mathbf{y}_1, ..., \mathbf{y}_n) \\
&= \sum_{j=1}^{m} \frac{p(S_t = i, S_{t+1} = j, \mathbf{y}_1, ..., \mathbf{y}_t)}{\sum_{i=1}^{m} p(S_t = i, \mathbf{y}_1, ..., \mathbf{y}_t) p(S_{t+1} = j|S_t = i)} p(S_{t+1} = j|\mathbf{y}_1, ..., \mathbf{y}_n)
\end{aligned}$$

Ceci implique

$$\gamma_t(i) = \sum_{j=1}^{m} \frac{\alpha_t(i)p_{ij}}{\sum_{i=1}^{m} \alpha_t(i)p_{ij}} \gamma_{t+1}(j) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{m} \alpha_t(i)\beta_t(i)}$$

Pour l'initialisation de ce calcul, on prend $\gamma_n(i) = \alpha_n(i)$. En utilisant ainsi $\gamma_t(i)$ nous pouvons estimer l'état individuel S_t le plus probable au temps t :

$$S_t = \arg\max_{1 \leq i \leq m} [\gamma_t(i)]$$

Bien que cette équation maximise le nombre espéré des états individuels en sélectionnant l'état le plus vraisemblable à chaque instant t, elle peut conduire à une séquence incorrecte dans le cas où le HMM possède des transitions d'états nulles pour certains états i et j ($p_{ij} = 0$).

Prise en compte des transitions 2 à 2

Dans certaines applications, nous choisissons des états qui ont le plus de chance deux à deux. Ceci revient à maximiser les probabilités a posteriori $p(S_{t-1} = i, S_t = j/\mathcal{Y})$ pour $i, j = 1, ..., m$. Le calcul de ces probabilités est basé sur les récurrence que nous avons déjà développé pour les variables α et β, soit

$$\begin{aligned}
\xi_t(i,j) &= p(S_{t-1} = i, S_t = j|\mathcal{Y}) \\
&= \frac{p(\mathcal{Y}|S_{t-1} = i, S_t = j)p(S_t = j|S_{t-1} = i)p(S_{t-1} = i)}{p(\mathcal{Y})} \\
&= p(\mathbf{y}_1, ..., \mathbf{y}_{t-1}|S_{t-1} = i)p(\mathbf{y}_t|S_t = j)p(\mathbf{y}_t, ..., \mathbf{y}_n|S_t = j) \\
&\qquad \times \frac{p(S_t = j|S_{t-1} = i)p(S_{t-1} = i)}{p(\mathcal{Y})} \\
&= \frac{\alpha_{t-1}(i)p(\mathbf{y}_t|S_t = j)\beta_t(j)p_{ij}}{p(\mathcal{Y})}
\end{aligned}$$

Nous pouvons l'exprimer, aussi, en fonction des variables α et γ :

$$\xi_t(i,j) = \frac{\alpha_{t-1}(i)p(\mathbf{y}_t|S_t = j)\gamma_t(j)p_{ij}}{\alpha_t(j)}$$

Algorithme de Viterbi

Le critère le plus utilisé est celui de trouver l'unique trajectoire optimale de la suite d'états, c'est-à-dire de maximiser $p(\mathcal{S}|\mathcal{Y}, \Theta)$ ou maximiser $p(\mathcal{S}, \mathcal{Y}|\Theta)$. Une technique formelle pour trouver le chemin optimal est basée sur les méthodes de programmation dynamique, c'est l'algorithme de Viterbi (1967).

C'est un algorithme récursif qui permet de trouver à partir d'une suite d'observations provenant d'un canal sans mémoire, une solution optimale au problème d'estimation de la suite d'états d'un processus de Markov à temps discret qui produit cette suite d'observations.

Pour trouver une trajectoire unique et optimale de la suite d'états, $\mathcal{S} = \{S_1, ..., S_n\}$ produisant la suite d'observations $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ nous définissons la quantité

$$\delta_t(i) = \max_{S_1, S_2, ..., S_{t-1}} \log p(S_1, S_2, ..., S_t = i, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t | \Theta), \quad t \geq 2$$

qui représente le meilleur score (la probabilité maximale) correspondant à une trajectoire unique jusqu'au temps t et qui prend en compte les premières "t-observations" et s'arrête à l'état i. Par itération

$$\delta_t(j) = \max_{1 \leq i \leq m} [\delta_{t-1}(i) + \log p_{ij}] + \log b_j(\mathbf{y}_t), \quad 1 \leq j \leq m$$

Pour retrouver la suite optimale d'états, nous devons garder une trace des arguments qui maximise l'équation ci-dessus pour chaque t et j.

Le principe de cet algorithme consiste donc à maximiser la probabilité conjointe $p(\mathcal{Y}, \mathcal{S})$ donnée par :

$$\begin{aligned}
p(\mathcal{Y}, \mathcal{S}) &= p(\mathcal{S}) p(\mathcal{Y}|\mathcal{S}) \\
&= p(S_1 = l) p(\mathbf{y}_1 | S_1 = l) \prod_{t=2}^{n} p(S_t = j | S_{t-1} = i) \prod_{t=2}^{n} p(\mathbf{y}_t | S_t = j) \\
&= \pi_l b_l(\mathbf{y}_1) \prod_{t=2}^{n} p_{ij} b_j(\mathbf{y}_t)
\end{aligned}$$

On a alors

$$\log p(\mathcal{Y}, \mathcal{S}) = \log [\pi_l b_l(\mathbf{y}_1)] + \sum_{t=2}^{n} \delta(S_t = j)$$

qui représente le coût total pour le chemin $\mathcal{S}$, où δ est le coût d'un segment (une transition d'un état à un autre) de chemin $\mathcal{S}$:

$$\delta(S_t = j) = \log p_{ij} + \log b_j(\mathbf{y}_t)$$

Nous définissons $\psi_t(j)$ comme étant le chemin le plus court correspondant au noeud $S_t = j$ (surviveur). À chaque instant t, il existe m surviveurs (un pour chaque noeud). L'algorithme nécessite, à chaque instant t, la mémorisation de ces m surviveurs ainsi que leurs coûts.

1. Initialisation, $t = 1$

Si S_1 est connu a priori, alors

$$\begin{cases} \delta_1(i) = 0,\ \forall\, i & \text{(coût du surviveur } i\text{)} \\ \psi_i = i & \text{(cette variable stocke l'état optimal à l'instant } t\text{)} \end{cases}$$

Autrement, si S_1 est inconnu a priori, alors

$$\begin{cases} \delta_1(i) = \log\left[\pi_i b_i(\mathbf{y}_1)\right] & i = 1, 2, ..., m \\ \psi_i = 0 \end{cases}$$

2. Induction

$$\begin{cases} \delta_t(j) = \max\limits_{1 \le i \le m} \left[\delta_{t-1}(i)\right] b_j(\mathbf{y}_t), \quad 1 \le j \le m,\ 2 \le t \le n \\ \psi_t(j) = \arg\max\limits_{1 \le i \le m} \left[\delta_t(i) + \log p_{ij}\right] \end{cases}$$

3. Terminaison

$$\begin{cases} \log p^* = \max\limits_{1 \le i \le m} \left[\delta_n(i)\right] \\ S_n^* = \arg\max\limits_{1 \le i \le m} \left[\delta_n(i)\right] \end{cases}$$

Chemin obtenu "Retrograde"

$$S_t^* = \psi_{t+1}(S_{t+1}^*), \quad t = n-1, n-2, ..., 1$$

4.1.5 Optimisation des Paramètres du Modèle

Comment peut-on ajuster les paramètres du modèle $\Theta = (\pi, \mathbf{P}, \mu, \mathbf{\Sigma})$ pour maximiser $p(\mathcal{Y}|\Theta)$? Le fait que la longueur de la suite d'observations (données d'apprentissage) est finie, il n'existe pas de solutions analytiques directes (d'optimisation globale) pour construire le modèle. Cependant, nous pouvons choisir $\Theta = (\pi, \mathbf{P}, \mu, \mathbf{\Sigma})$ tel que $p(\mathcal{Y}|\Theta)$ est un maximum local en utilisant une procédure itérative telle que celle de Baum-Welch (voir Baum et Eagon [1967], et Baum [1972]) ou d'une façon équivalente l'algorithme d'identification de mélange de type EM (Dempster et al., 1977) ou en utilisant aussi les techniques de gradient telle que la méthode de Liporace [1982].

Pour l'implémentation d'un algorithme EM, il faut tout d'abord calculer la vraisemblance des données complétées. Dans le cas des HMM, la probabilité jointe d'une séquence complète d'observations $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ et une séquence complète d'états cachés $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ peut être obtenue en calculant le produit des probabilités conditionnelles locales, soit

$$p(\mathcal{S}, \mathcal{Y}|\Theta) = p(S_1) \prod_{t=1}^{n-1} p(S_{t+1}|S_t, \Theta) \prod_{t=1}^{n} p(\mathbf{y}_t|S_t, \Theta)$$

Les probabilités de transition sont définies par :

$$p(S_{t+1} = j/S_t = i) = \prod_{i,j=1}^{m} [p_{ij}]^{S_t^i S_{t+1}^j}$$

où $S_\tau^i = 1$ si à la date τ le système est à l'état i et 0 autrement. De même la probabilité de l'état initial sera définie par :

$$\pi = p(S_1) = \prod_{i=1}^{m} [\pi_i]^{S_1^i}$$

L'espérance conditionnelle de la log-vraisemblance complétée est donnée par :

$$\begin{aligned}
\mathcal{Q}(\Theta, \Theta^{(i)}) &= \mathbb{E}\left[\log p(\mathcal{Y}, \mathcal{S}|\Theta^{(i)})|\mathcal{Y}, \Theta\right] \\
&= \sum_{\{S_n\}} p(\mathcal{S}|\mathcal{Y}, \Theta) \log p(\mathcal{Y}, \mathcal{S}|\Theta^{(i)}) \\
&= \sum_{\{S_n\}} p(\mathcal{S}|\mathcal{Y}, \Theta) \log \left[p(S_1) \prod_{t=2}^{n} p(S_t|S_{t-1}, \Theta) \prod_{t=1}^{n} p(\mathbf{y}_t|S_t, \Theta)\right]
\end{aligned}$$

L'étape E consiste simplement à calculer les probabilités conditionnelles $\gamma_t(j)$ et $\xi_t(i,j)$ en utilisant la procédure "Avant-Arrière" décrite précédemment. Ce calcul se fait en considérant la valeur courante $\Theta^{(i)}$ des paramètres du modèle. En utilisant ces probabilités, l'équation ci-dessus peut être exprimée sous la forme suivante :

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \sum_{t=1}^{n} \sum_{i=1}^{m} \log p(\mathbf{y}_t|S_t = i, \mu_i, \mathbf{\Sigma}_i) \mathbb{E}\left[S_t = i|\mathcal{Y}; \Theta^{(i)}\right]$$

$$+ \sum_{t=1}^{n-1} \sum_{i=1}^{m} \sum_{j=1}^{m} \log p_{ij} \mathbb{E}\left[S_t = i, S_{t+1} = j|\mathcal{Y}; \Theta^{(i)}\right] + \sum_{i=1}^{m} \log \pi_i \mathbb{E}\left[S_t = i|\mathcal{Y}; \Theta^{(i)}\right]$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{m} \gamma_t(i) \log p(\mathbf{y}_t|S_t = i, \mu_i, \mathbf{\Sigma}_i) + \sum_{t=1}^{n-1} \sum_{i=1}^{m} \sum_{j=1}^{m} \xi_t(i,j) \log p_{ij} + \sum_{i=1}^{m} \gamma_t(i) \log \pi_i$$

La maximisation de cette fonction par rapport aux paramètres de l'ensemble Θ en tenant compte de la contrainte $\sum_{i=1}^{m} \pi_i = 1$ et des contraintes de normalisation de chacune des lignes de la matrice de transition $\mathbf{P}$, $\sum_{j=1}^{m} p_{ij} = 1$ pour $i = 1, ..., m$ (en introduisant autant de multiplicateurs de Lagrange), nous permet de trouver :

$$\widehat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{m} \gamma_1(i)}$$

$$\widehat{p}_{ij} = \frac{\sum_{t=2}^{n} \xi_t(i,j)}{\sum_{t=2}^{n} \gamma_{t-1}(i)}$$

$$\widehat{\mu}_j = \frac{\sum_{t=1}^{n} \gamma_t(j)\mathbf{y}_t}{\sum_{t=1}^{n} \gamma_t(j)}$$

$$\widehat{\mathbf{\Sigma}}_j = \frac{\sum_{t=1}^{n} \gamma_t(j)\mathbf{y}_t\mathbf{y}_t'}{\sum_{t=1}^{n} \gamma_t(j)} - \widehat{\mu}_j\widehat{\mu}_j'$$

4.2 Introduction aux Modèles espace-état

L'étude de systèmes physiques émettant au cours du temps des signaux déterminés par des états internes non observés, a conduit à développer en traitement du signal les modèles dits espace-état. L'émergence de ces modèles, appelés aussi modèles dynamiques à facteurs, est relativement récente dans la recherche empirique en finance. De nombreuses procédures statistiques fréquemment utilisées dans la branche empirique de la recherche économique peuvent aujourd'hui se reformuler dans le cadre des modèles espace-état, notamment les modèles à composantes inobservables, les modèles à tendance stochastique et les modèles à coefficients aléatoires.

4.2.1 Présentation générale des modèles espace-état

Soit un processus multidimensionnel $\mathbf{y}_t$, on appelle modèle espace-état de ce processus, le système décrit par les équations suivantes :

Représentation espace-état

$$\left\{ \begin{array}{ll} \mathbf{y}_t = \theta + \mathbf{X}\mathbf{f}_t + \varepsilon_t & \text{Équation de mesure} \\ \mathbf{f}_{t+1} = \mathbf{A}\mathbf{f}_t + \mathbf{G}\omega_{t+1} & \text{Équation d'état} \end{array} \right.$$

$$\text{où} \quad \begin{pmatrix} \varepsilon_t \\ \omega_t \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix} \right]$$

Ces modèles sont constitués : *(i)* d'une ou plusieurs équation(s) de mesure décrivant la manière dont les variables observées sont générées par les variables cachées et les résidus *(ii)* d'une ou plusieurs équation(s) d'état décrivant la manière dont les variables cachées sont générées à partir de leur retard et d'innovations. La variable $\mathbf{y}_t$ est appelé observation ou variable de mesure, $\mathbf{f}_t$ est la variable d'état à la date t, ε_t est le vecteur des innovations à la date t, ω_t est le vecteur des erreurs de mesures à la date t, $\mathbf{A}$ est la matrice de transition, $\mathbf{X}$ est la matrice de mesure et $\mathbf{Xf}_t$ le signal à la date t. Les matrices $\mathbf{A}$, $\mathbf{X}$ sont de taille $(k \times k)$ et $(q \times k)$, $\mathbf{G}$ est une matrice déterministe de taille $(k \times k)$ et $\mathbf{f}_0$ est un vecteur aléatoire de loi $\mathcal{N}(\mathbf{0}, \mathbf{H}_0)$ indépendant du bruit blanc normal. Étant donné que la somme de variables gaussienne est toujours une variable gaussienne, la distribution de $\mathbf{f}_{t+1}$ sera, par conséquent, gaussienne. En conditionnant par rapport à $\mathbf{f}_t$, sa moyenne sera donnée par $\mathbf{Af}_t$ et sa matrice de covariance par $\mathbf{GQG'}$. Conditionnellement à $\mathbf{f}_t$, la distribution de $\mathbf{y}_t$ est aussi gaussienne de moyenne $\mathbf{Xf}_t$ et de matrice de covariance $\mathbf{\Psi}$.

Dans leur version élémentaire, ces modèles reposent sur un certain nombre d'hypothèses principales : les équations de mesure et d'état sont linéaires ; les bruits d'observation et d'innovation sont des bruits blancs [2] ; les variables cachées suivent à un instant initial donné une loi gaussienne. À ces dernières, se sont ajoutées des hypothèses secondaires permettant de déterminer la forme canonique : l'indépendance entre les bruits d'observation et d'innovation (condition d'inversibilité) et l'indépendance entre la variable cachée initiale et ces bruits (condition de causalité).Toutes ces hypothèses sont destinées à simplifier les procédures d'estimation.

Ce système est dit sous forme canonique si et seulement si :

$$\mathbb{E}\left[\varepsilon_t \omega_s\right] = \mathbb{E}\left[\varepsilon_t \mathbf{f}_0\right] = \mathbb{E}\left[\omega_t \mathbf{f}_0\right] = 0 \quad \forall\, t, s = 1, ..., n$$

Le modèle espace-état est alors dit causal et inversible.

Pour autant, on peut associer à un processus donné $\mathbf{y}_t$ plusieurs représentations espace-état. En effet, s'il existe une représentation de vecteur d'état $\mathbf{f}_t$, on peut formuler facilement une autre représentation $\mathbf{f}_t^* = \mathbf{M}_t \mathbf{f}_t$, $\mathbf{M}_t$ étant une matrice inversible quelconque. De même, au lieu de modéliser $\mathbf{f}_{t+1}$ dans l'équation d'état, on pourrait sans difficulté adapter l'estimation à un modèle d'état de $\mathbf{f}_t$. Enfin, diverses dimensions du vecteur d'état sont possibles et il convient de rechercher un modèle de dimension minimale, de manière à ne pas alourdir la procédure d'estimation.

Le modèle graphique visant à exprimer la structure et la dynamique de ce système est donné dans la figure 4.3. Cette représentation nous montre que le modèle espace-état a une structure identique à celle d'un HMM ; seulement le type des noeuds (vecteurs continus) et le modèle probabiliste (modèle linéaire et gaussien) changent. Les relations d'indépendance conditionnelles qui caractérisent ce modèle sont aussi identiques à celles caractérisant les HMM. Étant donné l'état à un instant t quelconque, les états futurs seront conditionnellement indépendants des états passés.

4.2.2 Filtrage de Kalman

Pour calculer des estimations filtrées du vecteur d'état $\mathbf{f}_t$ en se basant sur une séquence d'observations $\mathcal{Y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t\}$, l'algorithme optimal[3], appelé filtre de

[2] Un bruit blanc (au sens faible) est un processus aléatoire d'espérance et d'auto-covariances nulles, dont la distribution n'est pas toujours supposée gaussienne.

[3] Sous le terme "meilleure approximation" ou "optimal", on pense ici à deux critères d'optimalité qui s'avèrent être équivalents dans le cas gaussien : la maximisation de la vraisemblance du vecteur

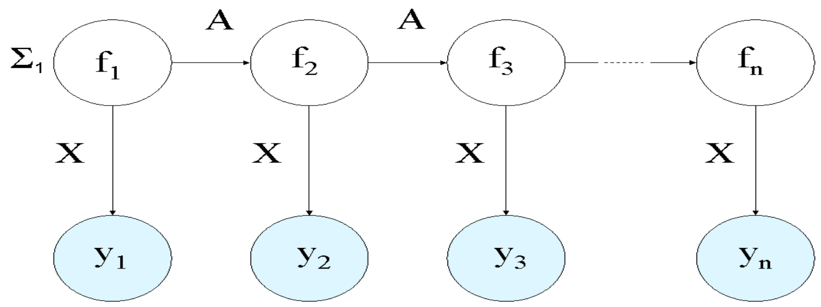

FIG. 4.3 – Modèle graphique d'une structure espace-état linéaire. Chaque morceau verticale représente un instant t quelconque.

Kalman, est utilisé. L'algorithme est structuré en deux étapes reprises d'itération en itération. Les deux premières équations sont des équations de "mises à jour des mesures" (actualisation) et les deux suivantes de "mise à jour du temps" (prévision). La première étape concerne les lois de probabilité a posteriori qui tiennent compte de l'information à la date t, $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$[4]. La seconde étape, à la différence de la première, ne dépend pas des observations à la date t : le calcul peut être fait "hors-ligne", c'est-à-dire sans utiliser les signaux $\mathbf{y}_t$. Enfin, la dernière équation actualise la matrice de gain[5] K_t qui intervient dans les équations précédentes.

Pour pouvoir introduire le filtre de Kalman appliqué aux données normales et expliquer comment on effectue l'initialisation de l'algorithme, on doit présenter quelques notions préliminaires. On commence par un résultat, énoncé sous la forme d'un lemme, utile pour le calcul des espérances et des variances conditionnelles en fonction des moments non conditionnels dans le cas de la loi normale. Il s'énonce comme suit :

Lemme 4.1. Si $\mathbf{x}$ et $\mathbf{y}$ sont deux vecteurs aléatoires normalement distribués avec moyennes $\mu_{\mathbf{x}}$ et $\mu_{\mathbf{y}}$, variances $\mathbf{\Sigma_{xx}}$ et $\mathbf{\Sigma_{yy}}$ et covariance $\mathbf{\Sigma_{xy}}$, alors on peut écrire :

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] &= \mu_{\mathbf{x}} + \mathbf{\Sigma_{xy}}\mathbf{\Sigma_{yy}^{-1}}\left[\mathbf{y} - \mu_{\mathbf{y}}\right] \\ Var\left[\mathbf{x}|\mathbf{y}\right] &= \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xy}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{\Sigma_{xy}'}\end{aligned}$$

Une démonstration de ce résultat peut être trouvée dans Anderson [2003]. Le filtre de Kalman se base sur les résultats énoncés par ce lemme. A chaque fois qu'on obtient une information supplémentaire liée à une variable aléatoire normalement distribuée, on ajuste ses moments pour tenir compte de toute l'information disponible. Dans toute la suite, nous allons désigner par $\mathbf{f}_{t/t}$ la moyenne conditionnelle de $\mathbf{f}_t$ par rapport à la

d'état conditionnellement au vecteur de mesure ou la minimisation des carrés des erreurs réalisées sur le vecteur d'état. Dans le cas non-gaussien, le filtre de Kalman reste uniquement optimal parmi les estimateurs linéaires.

[4]Notons que cette quantité est l'analogue de la variable α normalisée des HMM.

[5]La matrice K_t est dénommée matrice de gain car, comme cela sera expliqué plus loin, sa prise en compte engendre un gain en précision de l'estimation $\mathbf{f}_{t/t}$ de la variable cachée, relativement à $\mathbf{f}_{t/t-1}$.

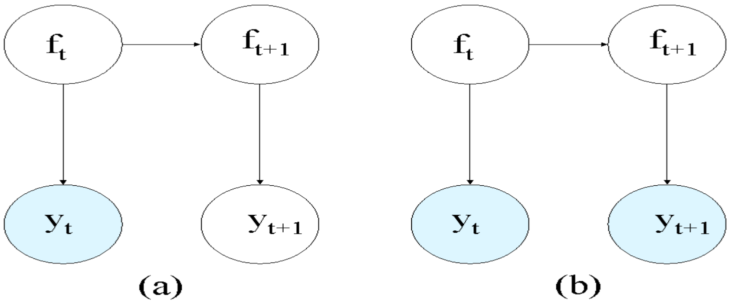

FIG. 4.4 – (a) fragment d'un modèle espace-état avant la mise à jour des observations et (b) après mise à jour.

séquence d'observations $\mathcal{Y}_{1:t}$ et par $\mathbf{H}_{t/t}$ sa matrice de variance-covariance conditionnellement à $\mathcal{Y}_{1:t}$, soient

$$\begin{aligned}\mathbf{f}_{t/t} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:t}\right]\\ \mathbf{H}_{t/t} &= \mathbb{E}\left[(\mathbf{f}_t-\mathbf{f}_{t/t})(\mathbf{f}_t-\mathbf{f}_{t/t})'/\mathcal{Y}_{1:t}\right]\end{aligned}$$

L'implémentation de cet algorithme nécessite aussi le calcul de la distribution de probabilité de $\mathbf{f}_t$ conditionnellement à $\mathcal{Y}_{1:t-1}$. En utilisant cette nouvelle notation, cette distribution aura une moyenne $\mathbf{f}_{t/t-1}$ et une matrice de covariance $\mathbf{H}_{t/t-1}$.

Afin d'illustrer les relations de récurrence nécessaire pour l'implémentation du filtre de Kalman, nous allons utiliser les fragments du modèle graphique de la figure 4.4. Dans le fragment gauche, où on conditionne par rapport à $\mathcal{Y}_{1:t}$, on suppose qu'on a déjà calculé $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$; et ainsi, on a calculé $\mathbf{f}_{t/t}$ et $\mathbf{H}_{t/t}$. On veut déplacer cette distribution vers le fragment à droite, où on conditionne par rapport à $\mathcal{Y}_{1:t-1}$. Pour ce faire, on va décomposer cette transition en deux étapes :

$$\begin{array}{ll}\text{Mise à jour du temps :} & p(\mathbf{f}_t/\mathcal{Y}_t) \rightarrow p(\mathbf{f}_{t+1}/\mathcal{Y}_t)\\ \text{Mise à jour des mesures :} & p(\mathbf{f}_{t+1}/\mathcal{Y}_t) \rightarrow p(\mathbf{f}_{t+1}/\mathcal{Y}_{t+1})\end{array}$$

Au niveau de l'étape, *mise à jour du temps*, la distribution sera tout simplement propagée dans le temps jusqu'à l'observation suivante. Par la suite, la nouvelle moyenne et la nouvelle matrice de covariance seront calculées en se basant sur leurs anciennes valeurs, sans utiliser les nouvelles mesures (i.e., les observations). Au niveau de l'étape, *mise à jour des mesures*, les nouvelles observations $\mathbf{y}_{t+1}$ seront utilisées pour mettre à jour la distribution de probabilité de $\mathbf{f}_{t+1}$. Ces deux étapes aboutissent à la conception d'un schéma de filtrage adaptatif permettant de trouver la meilleure approximation de l'état et de sa matrice de covariance à l'instant $t+1$ ($\mathbf{f}_{t+1/t+1}$ et $\mathbf{H}_{t+1/t+1}$) sachant les observations présentes et passées, en se basant sur la meilleure approximation obtenue à l'instant t ($\mathbf{f}_{t/t}$ et $\mathbf{H}_{t/t}$).

En utilisant les propriétés de la loi normale multivariée et le lemme 4.1, on peut obtenir à chaque période t les relations suivantes :

$$\mathbf{f}_{t+1/t} = \mathbf{A}\mathbf{f}_{t/t} \tag{4.1}$$

c'est la prévision de $\mathbf{f}_{t+1}$ au temps t, soit l'espérance conditionnelle de $\mathbf{f}_{t+1}$ étant donnée l'information disponible au temps t. D'une manière équivalente, les variances conditionnelles seront données par :

$$\begin{aligned}
\mathbf{H}_{t+1/t} &= \mathbb{E}\left[(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})'/\mathcal{Y}_{1:t}\right] \\
&= \mathbb{E}\left[(\mathbf{A}\mathbf{f}_t + \mathbf{G}w_t - \mathbf{A}\mathbf{f}_{t/t})(\mathbf{A}\mathbf{f}_t + \mathbf{G}w_t - \mathbf{A}\mathbf{f}_{t/t})'/\mathcal{Y}_{1:t}\right] \\
&= \mathbf{A}\mathbf{H}_{t/t}\mathbf{A}' + \mathbf{G}\mathbf{Q}\mathbf{G}'
\end{aligned} \tag{4.2}$$

où on a utilisé le fait que $\mathbf{f}_{t+1/t}$ est une constante dans la distribution conditionnelle, que ω_t a une moyenne nulle, et que ω_t et $\mathbf{f}_t$ sont indépendants.

La forme de l'équation de mesure et les statistiques de prédiction ci-dessus, nous permettent de trouver :

$$\begin{aligned}
\mathbb{E}\left[\mathbf{y}_{t+1}/\mathcal{Y}_{1:t}\right] &= \mathbb{E}\left[\mathbf{X}\mathbf{f}_{t+1} + \varepsilon_{t+1}/\mathcal{Y}_{1:t}\right] \\
&= \mathbf{X}\mathbf{f}_{t+1/t}
\end{aligned} \tag{4.3}$$

et

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{y}_{t+1} - \mathbf{y}_{t+1/t})(\mathbf{y}_{t+1} - \mathbf{y}_{t+1/t})'/\mathcal{Y}_{1:t}\right] \\
= \ &\mathbb{E}\left[(\mathbf{X}\mathbf{f}_{t+1} + \varepsilon_{t+1} - \mathbf{X}\mathbf{f}_{t+1/t})(\mathbf{X}\mathbf{f}_{t+1} + \varepsilon_{t+1} - \mathbf{X}\mathbf{f}_{t+1/t})'/\mathcal{Y}_{1:t}\right] \\
= \ &\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}
\end{aligned} \tag{4.4}$$

et enfin,

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{y}_{t+1} - \mathbf{y}_{t+1/t})(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})'/\mathcal{Y}_{1:t}\right] \\
&\qquad = \mathbb{E}\left[(\mathbf{X}\mathbf{f}_{t+1} + \varepsilon_{t+1} - \mathbf{y}_{t+1/t})(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})'/\mathcal{Y}_{1:t}\right] \\
&\qquad = \mathbf{X}\mathbf{H}_{t+1/t}
\end{aligned} \tag{4.5}$$

La distribution conditionnelle conjointe de $\mathbf{f}_{t+1}$ et $\mathbf{y}_{t+1}$ sachant l'information disponible jusqu'à la date t, $\mathcal{Y}_{1:t}$, est gaussienne

$$\begin{bmatrix} \mathbf{f}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} /\mathcal{Y}_{1:t} \sim \mathcal{N}\left[\begin{bmatrix} \mathbf{f}_{t+1/t} \\ \theta + \mathbf{X}\mathbf{f}_{t+1/t} \end{bmatrix}, \begin{bmatrix} \mathbf{H}_{t+1/t} & \mathbf{H}_{t+1/t}\mathbf{X}' \\ \mathbf{X}\mathbf{H}_{t+1/t} & \mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi} \end{bmatrix}\right] \tag{4.6}$$

en utilisant le résultat du lemme 4.1, on démontre que

$$\begin{aligned}
\mathbf{f}_{t+1/t+1} &= \mathbf{f}_{t+1/t} + \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1}\left(\mathbf{y}_{t+1} - \mathbf{X}\mathbf{f}_{t+1/t} - \theta\right) &(4.7) \\
\mathbf{H}_{t+1/t+1} &= \mathbf{H}_{t+1/t} - \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1}\mathbf{X}\mathbf{H}_{t+1/t} &(4.8)
\end{aligned}$$

Supposons qu'à la date t on a déjà les estimations de la moyenne $\mathbf{f}_{t/t}$ et de la matrice de covariance $\mathbf{H}_{t/t}$, chaque récursion de l'algorithme de filtrage se résume alors par les quatres équations suivantes :

$$\begin{aligned}
\mathbf{f}_{t+1/t} &= \mathbf{A}\mathbf{f}_{t/t} && (4.9)\\
\mathbf{H}_{t+1/t} &= \mathbf{A}\mathbf{H}_{t/t}\mathbf{A}' + \mathbf{G}\mathbf{Q}\mathbf{G}' && (4.10)\\
\mathbf{f}_{t+1/t+1} &= \mathbf{f}_{t+1/t} + \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1}\left(\mathbf{y}_{t+1} - \mathbf{X}\mathbf{f}_{t+1/t} - \theta\right) && (4.11)\\
\mathbf{H}_{t+1/t+1} &= \mathbf{H}_{t+1/t} - \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1}\mathbf{X}\mathbf{H}_{t+1/t} && (4.12)
\end{aligned}$$

Cet algorithme sera initialisé en prenant $\mathbf{f}_{0/-1} = \mathbf{0}$ et $\mathbf{H}_{0/-1} = \mathbf{H}_0$. Les équations de mise à jour pourront aussi être écrites sous une forme plus compacte en utilisant la matrice de gain de Kalman, définie par :

$$K_{t+1} = \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1} \tag{4.13}$$

En utilisant cette notation on obtient :

$$\begin{aligned}
\mathbf{f}_{t+1/t+1} &= \mathbf{f}_{t+1/t} + K_{t+1}\left[\mathbf{y}_{t+1} - \mathbf{X}\mathbf{f}_{t+1/t} - \theta\right] && (4.14)\\
\mathbf{H}_{t+1/t+1} &= \mathbf{H}_{t+1/t} - K_{t+1}\mathbf{X}\mathbf{H}_{t+1/t} && (4.15)
\end{aligned}$$

Nous pouvons, aussi, utiliser la formule d'inversion des matrices de Woodbury afin d'exprimer autrement la matrice de gain, soit

$$\begin{aligned}
K_{t+1} &= \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t} + \mathbf{\Psi}\right]^{-1}\\
&= \left[\mathbf{H}_{t+1/t}^{-1} + \mathbf{X}'\mathbf{\Psi}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{\Psi}^{-1}\\
&= \left[\mathbf{H}_{t+1/t} + \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1}\mathbf{X}\mathbf{H}_{t+1/t}\right]\mathbf{X}'\mathbf{\Psi}^{-1}\\
&= \mathbf{H}_{t+1/t+1}\mathbf{X}'\mathbf{\Psi}^{-1} && (4.16)
\end{aligned}$$

qui exprime la matrice de gain en fonction de la valeur actualisée $\mathbf{H}_{t+1/t+1}$.

L'équation (4.14) calcule l'estimation courante du vecteur d'état $\mathbf{f}_{t+1/t+1}$ comme la somme pondérée de la prévision à la date t du vecteur d'état $\mathbf{f}_{t+1}$ et de l'erreur de prévision calculée à partir de la dernière valeur observée $\mathbf{y}_{t+1}$. La pondération K_{t+1} est actualisée à chaque itération par l'équation (4.13). L'équation (4.9) permet de calculer la prévision de $\mathbf{f}_{t+1}$ à la date t, $\mathbf{f}_{t+1/t}$, comme la projection de $\mathbf{f}_{t+1}$ sur son passé (passé synthétisé par $\mathbf{f}_{t/t}$).

Les équations (4.10) et (4.12) sur les matrices de covariance sont appelées "équations de Riccati". Ces équations permettent de calculer la suite des gains de Kalman K_t et ce calcul peut être fait "hors-ligne". La matrice de covariance a posteriori $\mathbf{H}_{t+1/t+1}$ connaît généralement un gain en précision par rapport à la matrice de covariance a priori $\mathbf{H}_{t+1/t}$ grâce au terme $K_{t+1}\mathbf{X}\mathbf{H}_{t+1/t}$. La matrice de covariance a priori en $t+1$, $\mathbf{H}_{t+1/t}$, prend en compte les erreurs liées aux innovations de l'état avec la matrice $\mathbf{G}\mathbf{Q}\mathbf{G}'$, mais est aussi augmentée d'un terme $\mathbf{A}\mathbf{H}_{t/t}\mathbf{A}'$ associé aux erreurs sur l'état à la date t (équation (4.10)). Lorsque les variables d'état sont stationnaires, la covariance prévue $\mathbf{H}_{t+1/t}$ qui part d'une incertitude a priori P, tend vers une constante $\mathbf{H}_\infty$ (voir Harvey, [1989]). Après une période transitoire, les intervalles de confiance entourant des variables cachées stationnaires ont donc une largeur à peu près constante.

4.2.3 Le Filtre d'Information

L'algorithme de filtrage qu'on vient de présenter est basé sur les moments de la distribution normale. Un algorithme équivalent pourra aussi être implémenté en utilisant les paramètres canoniques de la distribution normale. Cet algorithme est connu sous sous l'appelation filtre d'information.

Les paramètres canoniques d'une distribution gaussienne sont définis par la transformation inverse : $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ et $\xi = \mathbf{\Sigma}^{-1}\mu$. Dans ce cas, si on désigne par $\widetilde{\mathbf{f}}_{t/t-1}$ et $\widetilde{\mathbf{H}}_{t/t-1}$ les paramètres canoniques de la distribution conditionnelle de $\mathbf{f}_t$ par rapport à $\mathcal{Y}_{1:t-1}$ et par $\widetilde{\mathbf{f}}_{t/t}$ et $\widetilde{\mathbf{H}}_{t/t}$ les paramètres canoniques conditionnellement à $\mathcal{Y}_{1:t}$, nous pouvons obtenir un algorithme adaptatif similaire à celui donné par les équations [4.9 - 4.12]. Pour ce faire, nous allons commencer tout d'abord par l'inversion des matrices de covariance. Afin de simplifier les calculs, on pose $\mathbf{D} = \mathbf{GQG}'$.

$$
\begin{aligned}
\widetilde{\mathbf{H}}_{t+1/t} &= \mathbf{H}_{t+1/t}^{-1} \\
&= \left[\mathbf{A}\mathbf{H}_{t/t}\mathbf{A}' + \mathbf{D}\right]^{-1} \\
&= \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{A}\left[\mathbf{H}_{t/t}^{-1} + \mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1} \\
&= \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{A}\left[\widetilde{\mathbf{H}}_{t/t} + \mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}
\end{aligned}
$$

Nous pouvons démontrer, aussi, que :

$$
\begin{aligned}
\widetilde{\mathbf{H}}_{t+1/t+1} &= \mathbf{H}_{t+1/t+1}^{-1} \\
&= \left[\mathbf{H}_{t+1/t} - \mathbf{H}_{t+1/t}\mathbf{X}'\left[\mathbf{X}\mathbf{H}_{t+1/t}\mathbf{X}' + \mathbf{\Psi}\right]^{-1}\mathbf{X}\mathbf{H}_{t+1/t}\right]^{-1} \\
&= \mathbf{H}_{t+1/t}^{-1} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X} \\
&= \widetilde{\mathbf{H}}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}
\end{aligned}
$$

Par la suite et en ce qui concerne les paramètres $\widetilde{\mathbf{f}}$, on a :

$$
\begin{aligned}
\widetilde{\mathbf{f}}_{t+1/t} &= \mathbf{H}_{t+1/t}^{-1}\mathbf{f}_{t+1/t} \\
&= \mathbf{H}_{t+1/t}^{-1}\mathbf{A}\mathbf{f}_{t/t} \\
&= \mathbf{H}_{t+1/t}^{-1}\mathbf{A}\mathbf{H}_{t/t}\widetilde{\mathbf{f}}_{t/t} \\
&= \left[\mathbf{A}\mathbf{H}_{t/t}\mathbf{A}' + \mathbf{D}\right]^{-1}\mathbf{A}\mathbf{H}_{t/t}\widetilde{\mathbf{f}}_{t/t} \\
&= \mathbf{D}^{-1}\mathbf{A}\left[\mathbf{H}_{t/t}^{-1} + \mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\widetilde{\mathbf{f}}_{t/t} \\
&= \mathbf{D}^{-1}\mathbf{A}\left[\widetilde{\mathbf{H}}_{t/t} + \mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\widetilde{\mathbf{f}}_{t/t}
\end{aligned}
$$

et

$$\begin{aligned}
\widetilde{\mathbf{f}}_{t+1/t+1} &= \mathbf{H}^{-1}_{t+1/t+1}\mathbf{f}_{t+1/t+1} \\
&= \mathbf{H}^{-1}_{t+1/t+1}\left[\mathbf{f}_{t+1/t} + \mathbf{H}_{t+1/t+1}\mathbf{X}'\mathbf{\Psi}^{-1}\left(\mathbf{y}_{t+1} - \mathbf{X}\mathbf{f}_{t+1/t} - \theta\right)\right] \\
&= \left[\mathbf{H}^{-1}_{t+1/t+1} - \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}\right]\mathbf{H}_{t+1/t}\widetilde{\mathbf{f}}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}(\mathbf{y}_{t+1} - \theta) \\
&= \left[\mathbf{H}^{-1}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X}\right]\mathbf{H}_{t+1/t}\widetilde{\mathbf{f}}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}(\mathbf{y}_{t+1} - \theta) \\
&= \widetilde{\mathbf{f}}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}(\mathbf{y}_{t+1} - \theta)
\end{aligned}$$

donc étant données les valeurs estimées $\widetilde{\mathbf{f}}_{t/t}$ et $\widetilde{\mathbf{H}}_{t/t}$, nous pouvons calculer d'une manière récursive les estimations $\widetilde{\mathbf{f}}_{t+1/t+1}$ et $\widetilde{\mathbf{H}}_{t+1/t+1}$ à travers les équations suivantes :

$$\widetilde{\mathbf{f}}_{t+1/t} = \mathbf{D}^{-1}\mathbf{A}\left[\widetilde{\mathbf{H}}_{t/t} + \mathbf{A}'\mathbf{D}\mathbf{A}\right]^{-1}\widetilde{\mathbf{f}}_{t/t} \tag{4.17}$$

$$\widetilde{\mathbf{f}}_{t+1/t+1} = \widetilde{\mathbf{f}}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}(\mathbf{y}_{t+1} - \theta) \tag{4.18}$$

$$\widetilde{\mathbf{H}}_{t+1/t} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{A}\left[\widetilde{\mathbf{H}}_{t/t} + \mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1} \tag{4.19}$$

$$\widetilde{\mathbf{H}}_{t+1/t+1} = \widetilde{\mathbf{H}}_{t+1/t} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X} \tag{4.20}$$

l'algorithme sera initialisé par $\widetilde{\mathbf{f}}_{1/0} = \overline{\mathbf{f}}_1$ et $\widetilde{\mathbf{H}}_{1/0} = \mathbf{H}_0$.

Le filtre de Kalman et le filtre d'information sont mathématiquement équivalents ; la différence d'ordre pratique entre les deux est essentiellement numérique. Étant donné que la condition de nombre d'une matrice est la réciproque de la condition de nombre de son inverse, donc si l'état initial est connu avec certitude on doit prendre $\mathbf{H}_1 = 0$, et dans ce cas $\widetilde{\mathbf{H}}_1$ sera indéfinie ce qui nous oblige à utiliser le filtre de Kalman. En revanche, lorsque l'état initial est absolument inconnu on prend $\widetilde{\mathbf{H}}_1 = 0$, ce qui rend $\mathbf{H}_1$ indéfinie et dans ce cas on sera obligé à utiliser le filtre d'information.

4.2.4 L'Algorithme de Lissage

Dans ce qui précède, nous avons présenté une technique de filtrage en deux versions différentes permettant d'obtenir la meilleure approximation de l'état $\mathbf{f}_t$ du système à la date t, conditionnellement à l'information disponible jusqu'en t. Maintenant, nous allons présenter l'algorithme de lissage qui donne l'approximation optimale du vecteur d'état à l'instant t, conditionnellement à toute l'information disponible sur l'ensemble de la période, $\mathcal{Y}_{1:n}$. Pour $t = 1, ..., n-1$, cet algorithme consiste en une paire de relations récursives utilisant comme conditions initiales (pour $t = n-1$) les quantités finales $\mathbf{f}_{n/n}$ et $\mathbf{H}_{n/n}$ données par l'étape de filtrage (filtre de Kalman appliqué jusqu'à la date finale n). Les relations de lissage fournissent alors, récursivement en remontant le temps, les quantités : $\mathbf{f}_{t/n}$ et $\mathbf{H}_{t/n}$, $t = n-1, ..., 1$.

Algorithme de Rauch-Tung-Striebel (RTS 1965)

Le développement de cet algorithme repose sur le fragment du modèle graphique de la figure 4.5. Nous commençons tout d'abord par le calcul de la distribution conjointe

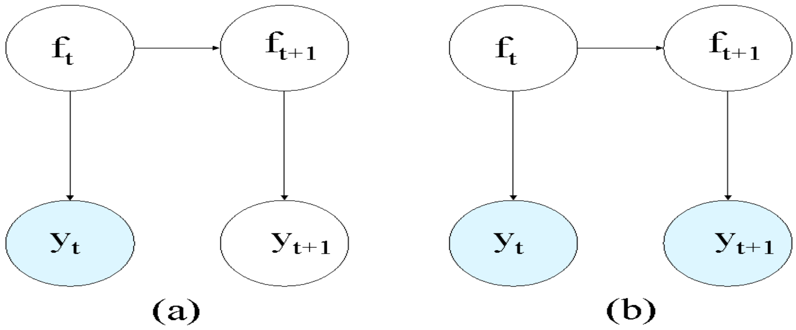

FIG. 4.5 – (a) fragment d'un modèle espace-état où les observations $\mathcal{Y}_{1:t}$ sont disponibles, et (b) le même fragment mais avec les observations $\mathcal{Y}_{t+1:n}$.

de $\mathbf{f}_t$ et $\mathbf{f}_{t+1}$, conditionnellement à $\mathcal{Y}_{1:t}$. En utilisant l'identité $\mathbf{f}_{t+1/t} = \mathbf{A}\mathbf{f}_{t/t}$ et les estimations de l'état et de sa matrice de covariance basées sur l'algorithme de filtrage, on aura :

$$\mathbb{E}\left[(\mathbf{f}_t - \mathbf{f}_{t/t})(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})'/\mathcal{Y}_{1:t}\right] = \mathbf{H}_{t/t}\mathbf{A}' \tag{4.21}$$

ceci implique

$$\begin{bmatrix} \mathbf{f}_t \\ \mathbf{f}_{t+1} \end{bmatrix} /\mathcal{Y}_{1:t} \sim \mathcal{N}\left[\begin{bmatrix} \mathbf{f}_{t/t} \\ \mathbf{f}_{t+1/t} \end{bmatrix}, \begin{bmatrix} \mathbf{H}_{t/t} & \mathbf{H}_{t/t}\mathbf{A}' \\ \mathbf{A}\mathbf{H}_{t/t} & \mathbf{H}_{t+1/t} \end{bmatrix}\right] \tag{4.22}$$

Nous calculons ensuite la probabilité de $\mathbf{f}_t$ en conditionnant par rapport à $\mathbf{f}_{t+1}$ et $\mathcal{Y}_{1:t}$. Dans ce cas, les propriétés de la loi normale multivariée impliquent :

$$\begin{aligned} \mathbb{E}\left[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:t}\right] &= \mathbf{f}_{t/t} + \mathbf{H}_{t/t}\mathbf{A}'\mathbf{H}^{-1}_{t+1/t}\left(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t}\right) \\ &= \mathbf{f}_{t/t} + J_t\left[\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t}\right] \end{aligned} \tag{4.23}$$

où $J_t = \mathbf{H}_{t/t}\mathbf{A}'\mathbf{H}^{-1}_{t+1/t}$, et

$$\begin{aligned} Var\left[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:t}\right] &= \mathbf{H}_{t/t} - \mathbf{H}_{t/t}\mathbf{A}'\mathbf{H}^{-1}_{t+1/t}\mathbf{A}\mathbf{H}_{t/t} \\ &= \mathbf{H}_{t/t} - J_t\mathbf{H}_{t+1/t}J'_t \end{aligned} \tag{4.24}$$

Le conditionnement par rapport à $\mathbf{f}_{t+1}$ rend la variable d'état $\mathbf{f}_t$ indépendante des observations futures $\mathcal{Y}_{t+1:n}$. L'utilisation de cette propriété nous permettra d'écrire :

$$\begin{aligned} \mathbb{E}[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:n}] &= \mathbb{E}[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:t}] \\ &= \mathbf{f}_{t/t} + J_t(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t}) \end{aligned} \tag{4.25}$$

et

$$\begin{aligned} Var[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:n}] &= Var[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:t}] \\ &= \mathbf{H}_{t/t} - J_t\mathbf{H}_{t+1/t}J_t' \end{aligned} \tag{4.26}$$

En utilisant la formule de l'espérance totale, on aura :

$$\begin{aligned} \mathbf{f}_{t/n} &= \mathbb{E}[\mathbf{f}_t/\mathbf{y}_1, ..., \mathbf{y}_n] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:n}\right]/\mathcal{Y}_{1:n}\right] \\ &= \mathbb{E}\left[\mathbf{f}_{t/t} + J_t(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})/\mathcal{Y}_{1:n}\right] \\ &= \mathbf{f}_{t/t} + J_t(\mathbf{f}_{t+1/n} - \mathbf{f}_{t+1/t}) \end{aligned} \tag{4.27}$$

dans cette dernière équation nous avons considéré que toutes les quantités autres que $\mathbf{f}_{t+1}$ sont des constantes lorsqu'on conditionne par rapport à $\mathcal{Y}_{1:n}$. Cette équation de mise à jour montre qu'une estimation de $\mathbf{f}_t$ basée sur la séquence complète d'observations peut être obtenue en corrigeant les estimations de filtrage $\mathbf{f}_{t/t}$ par un terme d'erreur tenant compte de la différence entre l'estimation de lissage de $\mathbf{f}_{t+1}$ et son estimation de filtrage $\mathbf{f}_{t+1/t}$. La matrice de gain J_t dépend seulement des matrices calculées au niveau des récurrences "avant"[6].

L'utilisation de la formule de variance totale, nous permettra aussi de trouver :

$$\begin{aligned} \mathbf{H}_{t/n} &= Var[\mathbf{f}_t/\mathbf{y}_1, ..., \mathbf{y}_n] \\ &= Var\left[\mathbb{E}\left[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:n}\right]/\mathcal{Y}_{1:n}\right] + \mathbb{E}\left[Var\left[\mathbf{f}_t/\mathbf{f}_{t+1}, \mathcal{Y}_{1:n}\right]/\mathcal{Y}_{1:n}\right] \\ &= Var\left[\mathbf{f}_{t/t} + J_t(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})/\mathcal{Y}_{1:n}\right] + \mathbb{E}\left[\mathbf{H}_{t/t} - J_t\mathbf{H}_{t+1/t}J_t'/\mathcal{Y}_{1:n}\right] \\ &= J_tVar\left[(\mathbf{f}_{t+1} - \mathbf{f}_{t+1/t})/\mathcal{Y}_{1:n}\right]J_t' + \mathbf{H}_{t/t} - J_t\mathbf{H}_{t+1/t}J_t' \\ &= J_tVar\left[\mathbf{f}_{t+1}/\mathcal{Y}_{1:n}\right]J_t' + \mathbf{H}_{t/t} - J_t\mathbf{H}_{t+1/t}J_t' \\ &= J_t\mathbf{H}_{t+1/n}J_t' + \mathbf{H}_{t/t} - J_t\mathbf{H}_{t+1/t}J_t' \\ &= \mathbf{H}_{t/t} + J_t\left(\mathbf{H}_{t+1/n} - \mathbf{H}_{t+1/t}\right)J_t' \end{aligned} \tag{4.28}$$

dans tous ces calculs intermédiaires nous avons considéré les espérances conditionnelles par rapport à $\mathcal{Y}_{1:t}$ comme des constantes lorsqu'on conditionne par rapport à la séquence complète d'observations $\mathcal{Y}_{1:n}$.

Finalement, étant données les statistiques de prédiction $\mathbf{f}_{t+1/t}$ et $\mathbf{H}_{t+1/t}$ et les statistiques de filtrage $\mathbf{f}_{t/t}$ et $\mathbf{H}_{t/t}$, l'algorithme de lissage RTS pourra être exprimé par les deux équations suivantes :

$$\mathbf{f}_{t/n} = \mathbf{f}_{t/t} + J_t\left[\mathbf{f}_{t+1/n} - \mathbf{f}_{t+1/t}\right] \tag{4.29}$$

$$\mathbf{H}_{t/n} = \mathbf{H}_{t/t} + J_t\left[\mathbf{H}_{t+1/n} - \mathbf{H}_{t+1/t}\right]J_t' \tag{4.30}$$

[6] Par analogie avec la procédure "Avant-Arrière" des modèles HMM.

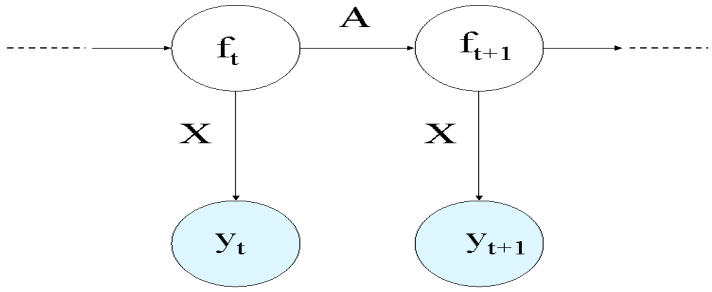

FIG. 4.6 – Fragment d'un modèle espace-état sans observations.

Algorithme de Lissage à Filtrage Double

Dans cette section, nous introduisons une approche alternative pour le lissage dans les modèles espace-état. Cette approche, appelée "algorithme à deux filtres", peut être considérée comme l'analogue de l'algorithme "Avant-Arrière" dans le cas des HMM (voir Jordan [1998]). Son principe est basé sur la combinaison d'une variable "Avant" (la probabilité $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$) avec une variable "Arrière" (la probabilité $p(\mathbf{f}_t/\mathcal{Y}_{t+1:n})$). Pour obtenir des estimations de filtrage avec la variable "Arrière", il suffit "d'inverser la dynamique" et d'appliquer, par la suite, un algorithme de filtrage "Avant". En termes de modèle graphique, il s'agit d'inverser le sens des flèches dans le graphe.

Dans toute la suite, la matrice d'autorégression $\mathbf{A}$ est supposée inversible. La forme inverse de l'équation de transition sera, donc, donnée par :

$$\mathbf{f}_t = \mathbf{A}^{-1}\mathbf{f}_{t+1} - \mathbf{A}^{-1}\mathbf{G}\omega_{t+1} \tag{4.31}$$

où t varie dans le sens inverse du temps. Ramarquons ici que la variable ω_{t+1} dépend de tous les états "passés" ; c-à-d., $\mathbf{f}_{t+1}, ..., \mathbf{f}_n$. Dans ce cas l'hypothèse fondamentale qu'on a utilisé pour l'implémentation du filtre de Kalman ne sera plus valable ce qui rend impossible l'application d'un tel algorithme pour l'équation (4.31).

En se basant sur le fragment du modèle graphique de la figure 4.6, nous pouvons inverser la dynamique du système. La matrice de covariance non conditionnelle du couple $(\mathbf{f}_t, \mathbf{f}_{t+1})$ est donnée par :

$$\begin{bmatrix} \mathbf{H}_t & \mathbf{H}_t\mathbf{A}' \\ \mathbf{A}\mathbf{H}_t & \mathbf{A}\mathbf{H}_t\mathbf{A}' + \mathbf{G}\mathbf{Q}\mathbf{G}' \end{bmatrix} \tag{4.32}$$

L'inversion de la relation entre $\mathbf{f}_t$ et $\mathbf{f}_{t+1}$ nous permettra d'exprimer la matrice de covariance non conditionnelle de $\mathbf{f}_t$ en fonction de celle de $\mathbf{f}_{t+1}$, soit

$$\mathbf{H}_t = \mathbf{A}^{-1}\mathbf{H}_{t+1}\mathbf{A}^{-1'} - \mathbf{A}^{-1}\mathbf{G}\mathbf{Q}\mathbf{G}'\mathbf{A}^{-1'} \tag{4.33}$$

Cette équation implique, aussi,

$$\mathbf{A}\mathbf{H}_t = \mathbf{H}_{t+1}\mathbf{A}^{-1\prime} - \mathbf{G}\mathbf{Q}\mathbf{G}'\mathbf{A}^{-1\prime} \tag{4.34}$$

La matrice de covariance (4.32) pourra donc être exprimée sous la forme suivante :

$$\begin{bmatrix} \mathbf{A}^{-1}\mathbf{H}_{t+1}\mathbf{A}^{-1\prime} - \mathbf{A}^{-1}\mathbf{G}\mathbf{Q}\mathbf{G}'\mathbf{A}^{-1\prime} & \mathbf{A}^{-1}\mathbf{H}_{t+1} - \mathbf{A}^{-1}\mathbf{G}\mathbf{Q}\mathbf{G}' \\ \mathbf{H}_{t+1}\mathbf{A}^{-1\prime} - \mathbf{G}\mathbf{Q}\mathbf{G}'\mathbf{A}^{-1\prime} & \mathbf{H}_{t+1} \end{bmatrix} \tag{4.35}$$

Notons que le premier élément de la deuxième colonne de cette matrice peut être écrit sous la forme $\mathbf{A}^{-1}\left[\mathbf{I} - \mathbf{G}\mathbf{Q}\mathbf{G}'\mathbf{H}_{t+1}^{-1}\right]\mathbf{H}_{t+1}$. Si on pose

$$\widetilde{\mathbf{A}} = \mathbf{A}^{-1}\left[\mathbf{I} - \mathbf{G}\mathbf{Q}\mathbf{G}'\mathbf{H}_{t+1}^{-1}\right] \tag{4.36}$$

les covariances seront données par $\widetilde{\mathbf{A}}\mathbf{H}_{t+1}$ et $\mathbf{H}_{t+1}\widetilde{\mathbf{A}}'$, et la relation inversée par :

$$\mathbf{f}_t = \widetilde{\mathbf{A}}\mathbf{f}_{t+1} + \widetilde{\mathbf{G}}\widetilde{\omega}_{t+1} \tag{4.37}$$

avec

$$\begin{aligned} \widetilde{\mathbf{G}} &= -\mathbf{A}^{-1}\mathbf{G} && (4.38)\\ \widetilde{\omega}_{t+1} &= \omega_{t+1} - \mathbf{Q}\mathbf{G}'\mathbf{H}_{t+1}^{-1}\mathbf{f}_{t+1} && (4.39) \end{aligned}$$

Nous obtenons, aussi :

$$\widetilde{\mathbf{Q}} = \mathbb{E}\left[\widetilde{\omega}_{t+1}\widetilde{\omega}'_{t+1}\right] = \mathbf{Q} - \mathbf{Q}\mathbf{G}'\mathbf{H}_{t+1}^{-1}\mathbf{G}\mathbf{Q} \tag{4.40}$$

et ainsi

$$\mathbf{H}_t = \widetilde{\mathbf{A}}\mathbf{H}_{t+1}\widetilde{\mathbf{A}}' + \widetilde{\mathbf{G}}\widetilde{\mathbf{Q}}\widetilde{\mathbf{G}}' \tag{4.41}$$

Finalement, nous pouvons démontrer que le terme $\widetilde{\omega}_{t+1}$ est indépendant des états "passés" $\mathbf{f}_{t+1}, ..., \mathbf{f}_n$, c-à-d., $Cov(\widetilde{\omega}_{t+1}, \mathbf{f}_{t+k}) = \mathbf{0}\ \forall\, k \geq 1$.

Par analogie avec les équations de filtrage [4.17 - 4.20], l'utilisation des paramètres canoniques et la relation des dynamiques inversées nous permettra de trouver :

$$\begin{aligned} \widetilde{\mathbf{H}}_{t/t+1} &= \mathbf{A}'\mathbf{D}\mathbf{A} + \mathbf{H}_t^{-1} - \mathbf{A}'\mathbf{D}^{-1}\left[\widetilde{\mathbf{H}}_{t+1/t+1} + \mathbf{D}^{-1} - \mathbf{H}_{t+1}^{-1}\right]^{-1}\mathbf{D}^{-1}\mathbf{A} && (4.42)\\ \widetilde{\mathbf{H}}_{t/t} &= \widetilde{\mathbf{H}}_{t/t+1} + \mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X} && (4.43)\\ \widetilde{\mathbf{f}}_{t/t+1} &= \mathbf{A}'\mathbf{D}^{-1}\left[\widetilde{\mathbf{H}}_{t+1/t+1} + \mathbf{D}^{-1} - \mathbf{H}_{t+1}^{-1}\right]^{-1}\widetilde{\mathbf{f}}_{t+1/t+1} && (4.44)\\ \widetilde{\mathbf{f}}_{t/t} &= \widetilde{\mathbf{f}}_{t/t+1} + \mathbf{X}'\mathbf{\Psi}^{-1}(\mathbf{y}_t - \theta) && (4.45) \end{aligned}$$

nous pouvons aussi transformer ces équations afin d'obtenir une version de filtrage "Arrière" basée sur les moments. Dans ce cas, les moments de la distribution prédictive seront donnés par : $\mathbf{f}_{t/t+1} = \widetilde{\mathbf{H}}_{t/t+1}^{-1}\widetilde{\mathbf{f}}_{t/t+1}$ et $\mathbf{H}_{t/t+1} = \widetilde{\mathbf{H}}_{t/t+1}^{-1}$.

Jusqu'ici, nous avons présenté des algorithmes adaptatifs permettant de mettre à jour les moments conditionnels de $\mathbf{f}_t$ sachant les séquences d'observations $\mathcal{Y}_{1:t}$ et $\mathcal{Y}_{t+1:n}$. Afin d'estimer le vecteur d'état et sa matrice de covariance en se basant sur la séquence complète des observations $\mathcal{Y}_{1:n}$, on doit calculer $p(\mathbf{f}_t/\mathcal{Y}_{1:n})$ qui nécessite à son tour la fusion des distributions conditionnelles $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$ et $p(\mathbf{f}_t/\mathcal{Y}_{t+1:n})$. En adoptant la méthodologie de Jordan [1998], ces estimations seront données par :

$$\mathbf{f}_{t/n} = \mathbf{H}_{t/n}\left[\mathbf{H}_{t/t}^{-1}\mathbf{f}_{t/t} + \mathbf{H}_{t/t+1}^{-1}\mathbf{f}_{t/t+1}\right] \tag{4.46}$$

$$\mathbf{H}_{t/n} = \left[\mathbf{H}_{t/t}^{-1} + \mathbf{H}_{t/t+1}^{-1} - \mathbf{H}_{t}^{-1}\right]^{-1} \tag{4.47}$$

4.2.5 Optimisation des paramètres et Algorithme EM

Dans la partie précédente, les matrices $\mathbf{A}$, $\mathbf{X}$, $\mathbf{Q}$, $\boldsymbol{\Psi}$, $\mathbf{H}_0$ ainsi que le vecteur θ étaient supposés connus. En pratique, ces matrices sont inconnues et doivent être estimées. L'algorithme EM est couramment utilisé pour déterminer les Estimateurs du Maximum de Vraisemblance des paramètres d'un modèle espace-état. Cet algorithme itératif a le mérite d'être simple, même s'il est relativement lent à converger par rapport à des algorithmes plus sophistiqués.

L'Algorithme EM

Pour procéder à une estimation par maximum de vraisemblance des paramètres d'un modèle espace-état, il est nécessaire d'avoir l'expression de la fonction de vraisemblance. Pour chaque jeu de paramètres Θ, la log-vraisemblance complétée associée à un échantillon $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ et à une séquence complète d'états cachés $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ d'un modèle espace-état est donnée par :

$$\mathcal{L}(\mathcal{Y}, \mathcal{F}|\Theta) = \log\left[p(\mathbf{f}_1|\Theta)\prod_{t=2}^{n} p(\mathbf{f}_t|\mathbf{f}_{t-1}, \Theta)\prod_{t=1}^{n} p(\mathbf{y}_t|\mathbf{f}_t, \Theta)\right] \tag{4.48}$$

L'algorithme EM est alors un algorithme itératif qui génère une séquence d'estimations $(\Theta^{(i)})_{i=1,2,...}$ à partir d'une condition initiale Θ_0. Chaque itération se décompose en deux étapes qui s'écrivent :

1. Étape E : l'espérance conditionnelle de la log-vraisemblance complétée $\mathcal{Q}(\Theta, \Theta^{(i)})$ se déduit de $\mathbf{f}_{t/n}$ et de $\mathbf{H}_{t/n}$, calculés par l'algorithme de lissage.
2. Étape M : la maximisation de $\mathcal{Q}(\Theta, \Theta^{(i)})$ par rapport à Θ conduit à $\Theta^{(i+1)}$.

La première étape E calcule une espérance conditionnelle de la log-vraisemblance complétée à partir de la formule précédente. Ces formules mobilisent en particulier l'application des algorithmes de filtrage et de lissage de Kalman pour connaître l'espérance conditionnelle de l'état $\mathbf{f}_{t/n}$ et de sa covariance $\mathbf{H}_{t/n}$ à paramètres $\Theta^{(i)}$ et observations $\mathcal{Y}_{1:n}$ fixés. Cette espérance conditionnelle $\mathcal{Q}(\Theta, \Theta^{(i)})$ est donnée par :

$$\begin{aligned}
\mathcal{Q}(\Theta,\Theta^{(i)}) &= \mathbb{E}\left[\log p(\mathcal{Y},\mathcal{F}|\Theta^{(i)})|\mathcal{Y},\Theta\right] = \int p(\mathcal{F}|\mathcal{Y},\Theta)\log p(\mathcal{Y},\mathcal{F}|\Theta^{(i)})d\mathcal{F} \\
&\simeq -\frac{1}{2}\sum_{t=1}^{n}\Big[\log|\boldsymbol{\Psi}| + \mathbb{E}\left\{(\mathbf{y}_t-\mathbf{X}\mathbf{f}_t-\theta)'\boldsymbol{\Psi}^{-1}(\mathbf{y}_t-\mathbf{X}\mathbf{f}_t-\theta)|\mathcal{Y},\Theta\right\}\Big] \\
&\quad -\frac{1}{2}\sum_{t=2}^{n}\Big[\log|\mathbf{Q}| + \mathbb{E}\left\{(\mathbf{f}_t-\mathbf{A}\mathbf{f}_{t-1})'\mathbf{Q}^{-1}(\mathbf{f}_t-\mathbf{A}\mathbf{f}_{t-1})|\mathcal{Y},\Theta\right\}\Big] \\
&\quad -\frac{1}{2}\Big[\log|\mathbf{H}_0| + \mathbb{E}\left\{\mathbf{f}_1'\mathbf{H}_0^{-1}\mathbf{f}_1|\mathcal{Y},\Theta\right\}\Big] \qquad (4.49)
\end{aligned}$$

pour la simplification nous avons supposé que $\mathbf{G} = \mathbf{I}_k$. La seconde étape M, consiste à rechercher un jeu de paramètres maximisant la vraisemblance estimée dans l'étape E. Cette maximisation peut-être analytique ou numérique selon la complexité du problème. Après un cycle "Étape E/Étape M", on obtient $\Theta^{(i+1)}$ et on peut montrer que la vraisemblance a augmenté $(\mathcal{L}(\mathcal{Y}_{1:n}|\Theta^{(i+1)}) > \mathcal{L}(\mathcal{Y}_{1:n}|\Theta^{(i)}))$. En itérant ces étapes E et M, les paramètres estimés par l'algorithme convergent généralement vers le maximum de vraisemblance. Les formules de mise à jour sont données par :

$$\widehat{\mathbf{X}} = \left[\sum_{t=1}^{n}\mathbf{y}_t\mathbf{f}_{t/n}' - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t\sum_{t=1}^{n}\mathbf{f}_{t/n}'\right]\left[\sum_{t=1}^{n}\widetilde{\mathbf{R}}_{t/n} - \frac{1}{n}\sum_{t=1}^{n}\mathbf{f}_{t/n}\sum_{t=1}^{n}\mathbf{f}_{t/n}'\right]^{-1}$$

$$\widehat{\boldsymbol{\Psi}} = \frac{1}{n}\sum_{t=1}^{n}\left\{\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix}\widehat{\mathbf{X}} & \widehat{\theta}\end{bmatrix}\begin{bmatrix}\mathbf{y}_t\mathbf{f}_{t/n}' & \mathbf{y}_t\end{bmatrix}'\right\}$$

$$\widehat{\theta} = \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t - \widehat{\mathbf{X}}\mathbf{f}_{t/n}\right)$$

et les paramètres de l'équation de transition seront mis à jour par :

$$\widehat{\mathbf{A}} = \left[\sum_{t=2}^{n}\widetilde{\mathbf{R}}_{t-1,t/n} - \frac{1}{n-1}\sum_{t=2}^{n}\mathbf{f}_{t/n}\sum_{t=2}^{n}\mathbf{f}_{t-1/n}'\right]\left[\sum_{t=2}^{n}\widetilde{\mathbf{R}}_{t-1/n} - \frac{1}{n-1}\sum_{t=2}^{n}\mathbf{f}_{t-1/n}\sum_{t=2}^{n}\mathbf{f}_{t-1/n}'\right]^{-1}$$

$$\widehat{\mathbf{Q}} = \frac{1}{n-1}\sum_{t=2}^{n}\left[\widetilde{\mathbf{R}}_{t/n} - \begin{bmatrix}\widehat{\mathbf{A}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{R}}_{t-1,t/n} & \mathbf{f}_{t/n}\end{bmatrix}'\right]$$

avec $\widehat{\mathbf{H}}_0 = \widetilde{\mathbf{R}}_{1/n}$; $\widetilde{\mathbf{R}}_{t/n} = \mathbb{E}\left[\mathbf{f}_t\mathbf{f}_t'|\mathcal{Y}\right]$ et $\widetilde{\mathbf{R}}_{t-1,t/n} = \mathbb{E}\left[\mathbf{f}_{t-1}\mathbf{f}_t'|\mathcal{Y}\right]$. Ces différentes valeurs nécessitent la connaissance de la matrice de covariance de la distribution a posteriori jointe de deux vecteurs d'états successifs. Cette distribution, $p(\mathbf{f}_t, \mathbf{f}_{t-1}/\mathbf{y}_{1:n})$, est aussi gaussienne et sa matrice de covariance peut être écrite sous la forme :

$$\mathbf{H}_{t-1,t/n} = \mathbf{H}_{t/n}\mathbf{H}_{t/t-1}^{-1}\mathbf{A}\mathbf{H}_{t-1/t-1}$$

Quelques Limites Pratiques

Les propriétés statistiques de l'estimateur du maximum de vraisemblance ne sont pas abordées ici mais certaines difficultés de la phase d'estimation sont présentées. Trois problèmes sont brièvement étudiés : le choix des conditions initiales, l'importance du ratio signal/bruit et les propriétés de convergence de l'algorithme EM.

La mise en oeuvre du filtre de Kalman nécessite généralement de spécifier les conditions initiales du vecteur d'état. En effet, si tous les éléments du vecteur d'état initial $\mathbf{f}_0$ sont exactement connus a priori, alors $\mathbf{f}_0$ a une distribution a priori correcte, c'est-à-dire dont tous les moments sont finis, avec une moyenne connue et une matrice de covariance bornée. Le filtre de Kalman fournit alors la fonction de vraisemblance exacte des observations par la décomposition de l'erreur de prévision. Une telle information a priori est cependant rarement disponible. Dans cette perspective, nous pouvons fixer arbitrairement les valeurs initiales du vecteur d'état $\mathbf{f}_0$. Le problème est que les estimations vont dépendre de ces valeurs. Il s'agit alors de tester la sensibilité aux conditions initiales et cela d'autant que l'algorithme EM fournit des maxima locaux.

Un deuxième problème concerne le traitement des matrices $\mathbf{Q}$ et $\mathbf{\Psi}$, qui représentent respectivement les matrices de variance-covariance du vecteur des innovations et du vecteur des erreurs de mesure. En effet, un élément fondamental dans l'estimation des modèles espace-état est le degré de lissage des variables non observées, qui dépend des deux matrices précédentes. Par exemple, dans le cas univarié, un ratio $\mathbf{\Psi}/\mathbf{Q}$ élevé (appelé ratio signal/bruit) contribue à accroître le pouvoir explicatif de la variable latente et l'équation de mesure sera donc mieux estimée. À la limite, pour de grandes valeurs de $\mathbf{Q}$, la variable non observée absorbe toute la variation des résidus dans l'équation de mesure. Alternativement, si $\mathbf{Q}$ est une matrice nulle et si $\mathbf{A}$ est la matrice identité, les estimations filtrées (respectivement lissées) correspondront à la méthode des moindres carrés récursifs (respectivement des moindres carrés). Il est donc particulièrement important de déterminer ce ratio. Dans la pratique, la plupart des études fixent ce ratio de telle sorte que l'estimation de la variable latente soit suffisamment lisse, avec des fluctuations jugées raisonnable d'une période à l'autre. Des tests de sensibilité sont alors utilisés en spécifiant différentes valeurs pour ce ratio.

4.3 Modèles Espace-État et Changement de Régime

Les modèles espace-état à changement de régime ont été introduits il y a 15 ans par Shumway et Stoffer [1991] en économétrie, puis ont été ensuite largement utilisés en économétrie (Kim [1994]) et en traitement automatique de la parole (Lee, Attias, et Deng [2003] et Rosti et Gales [2003, 2004] ainsi que les références citées dans ces articles). Le modèle proposé par Kim [1994] est une extension du modèle de changement de régime markovien étudié par Hamilton [1988, 1989] pour les modèles espace-état linéaires. Cette nouvelle spécification est basée sur la combinaison des modèles espace-état avec les modèles de chaînes de Markov cachées, en supposant que les différents états de l'économie aussi bien que la transition d'un état à un autre ne sont pas observables. Ces modèles sont définis au paragraphe 4.3.1. Dans ce paragraphe, nous introduisons diverses notations qui seront utilisées dans la suite de ce chapitre.

4.3.1 Définition et Notations

Shumway et Stoffer [2000] passent en revue de la littérature traitant la modélisation du changement de régime dans les séries temporelles dans leur livre Time Series Analysis and its Applications. Ils présentent notamment le modèle suivant :

Représentation espace-état multi-régime

$$S_t \sim P(S_t = j / S_{t-1} = i)$$

$$\begin{cases} \mathbf{y}_t = \theta_{s_t} + \mathbf{X}_{s_t}\mathbf{f}_t + \varepsilon_t & \text{Équation de mesure} \\ & \text{Équation de transition} \\ \mathbf{f}_{t+1} = \mathbf{A}_{s_t}\mathbf{f}_t + \omega_{t+1} & \text{Si } S_{t+1} = S_t \\ \mathbf{f}_{t+1} \sim \mathcal{N}\left[\mathbf{0}, \mathbf{H}_{s_{t+1}}\right] & \text{Si } S_{t+1} \neq S_t \end{cases}$$

$$\text{où} \quad \begin{pmatrix} \varepsilon_t \\ \omega_t \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Psi}_{s_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{s_t} \end{pmatrix}\right]$$

L'équation de mesure décrit l'évolution du vecteur $\mathbf{y}_t$ $(q \times 1)$ des variables observées en fonction du vecteur $\mathbf{f}_t$ $(k \times 1)$ des variables inobservées. L'équation d'état ou de transition décrit la dynamique des variables inobservées. Les matrices θ_{s_t} $(q \times 1)$, $\mathbf{X}_{s_t}$ $(q \times k)$, $\mathbf{A}_{s_t}$ $(k \times k)$, $\mathbf{Q}_{s_t}$ $(k \times k)$, $\mathbf{\Psi}_{s_t}$ $(q \times q)$ sont les matrices de paramètres qui dépendent du régime S_t inobservable à valeurs discrètes ($S_t = j$, $j = 1, 2, ..., m$) suivant un processus markovien d'ordre 1 admettant une matrice de transition $\mathbf{P} = [p_{ij}]$ (chaîne de Markov homogène) et un vecteur de probabilités de l'état initial π[7]. $\omega_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{s_t})$ et $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}_{s_t})$ sont deux termes de bruit gaussiens indépendamment distribués. Les paramètres du modèle $\Theta_j = \{\theta_j, \mathbf{A}_j, \mathbf{X}_j, \mathbf{Q}_j, \mathbf{\Psi}_j\}$ pour $j = 1, 2, ..., m$ sont supposés constants à l'intérieur de chaque régime. On définit $M_{t/t}(j)$ par la relation $M_{t/t}(j) = p(S_t = j | \mathcal{Y}_{1:t})$ où $\mathcal{Y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y_t}\}$ est l'information disponible à la période t. Chaque état j implique un modèle état-mesure différent et une estimation de $M_{t/t}(j)$ doit être effectuée. D'une manière générale, le filtre de Kalman doit être modifié pour tenir compte du caractère aléatoire de S_t. On suppose qu'à chaque période t la probabilité que S_t se trouve dans l'état j est égale à $p_t(j)$. Si on n'as pas de raison de préférer un état plutôt qu'un autre au temps t alors on pose $p_t(j) = m^{-1}$.

4.3.2 Les Méthodes d'Inférence Approximatives

La difficulté d'estimation des modèles espace-état avec changement de régime provient du fait que le nombre des séquences d'états possibles augmente d'une manière exponentielle avec le temps. À l'instant $t = 1$, $p(\mathbf{f}_1/\mathbf{y}_1)$ est un mélange de m gaussiennes (une composante pour chaque valeur possible de S_1). Chacune de ces composantes sera propagée à l'instant $t = 2$ en m composantes. D'une manière générale, à un instant t quelconque, la probabilité de l'état $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$ est un mélange de m^t gaussiennes, une pour chaque séquence d'états possible $S_1, ..., S_t$. Dans la littérature récente

[7] Ainsi $1/p_{ij}$ est le temps espéré de rester à l'état/régime i avant de passer à un autre état j. Nous pouvons changer la distribution sur chaque segment en modélisant d'une manière explicite la persistence de chaque régime (Rabiner, [1989] et Kulp, Reese, Haussler, et Eckman [1996]).

plusieurs méthodes d'approximation, aussi bien déterministes que stochastiques, ont été proposées afin de résoudre ce problème (voir Murphy [2002]).

La méthode pseudo Bayésienne Généralisée : La méthode pseudo-bayésienne généralisée d'ordre r (GPB(r)) consiste à approximer à un instant t quelconque m^t composantes de mélange par un mélange de r gaussiennes en utilisant la technique dite "moment matching" (voir par exemple, Bar-Shalom et Li [1993], Kim [1994] et Murphy [2002)). Compte tenu des applications qui ont utilisé cette méthode, la plupart ont considéré un ordre $r = 1$ ou bien $r = 2$. Dans ce dernier cas l'algorithme combine des gaussiennes qui diffèrent par leurs structures de retard de deux périodes. En général plus le retard est grand, plus la méthode donne une meilleure approximation (Smith et Markov [1980]). Notons enfin que l'optimalité de cette méthode dans le sens de Kullback-Leibler, et la convergence de l'erreur d'approximation ont été aussi déjà preuvé (Lauritzen [1996] et Boyen et Koller [1998]).

L'Algorithme des modèles multiples interagissant (IMM) Dans cet algorithme, plusieurs filtres opèrent en parallèle, chaque filtre est adapté à un modèle pour la dynamique de la variable étudiée. Par la suite, les états estimés à partir de ces filtres sont combinés sur une base probabiliste pour former l'état estimé global. L'algorithme IMM est constitué de six étapes principales. Dans une première étape, les probabilités de mélange des modèles seront calculées. La probabilité que le modèle $\mathcal{M}_i$ était effectif à l'instant $t-1$, étant donné que le modèle $\mathcal{M}_j$ est effectif à l'instant t, conditionnellement aux mesures $\mathcal{Y}_{1:t-1}$ reçues jusqu'à l'instant $t-1$ est calculée à partir de :

$$\begin{aligned}
\mu_{i/j}(t-1|t-1) &= p(S_{t-1} = i|S_t = j, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c_j} p(S_t = j|S_{t-1} = i, \mathcal{Y}_{1:t-1}) p(S_{t-1} = i|\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c_j} p_{ij} \mu_i(t-1), \; i, j = 1, ..., m
\end{aligned}$$

où p_{ij} est la probabilité a priori de transition de l'état i à l'état j, $\mu_i(t-1)$ est la probabilité que le modèle i soit effectif à l'instant $t-1$ et c_j sont des constantes de normalisation calculées à partir de :

$$c_j = \sum_{i=1}^{m} p_{ij} \mu_i(t-1), \; j = 1, ..., m$$

où m représente le nombre de modèles en interaction. Dans une deuxième étape, à partir de $\mathbf{f}^j_{t-1/t-1}$, l'état estimé par le filtre adapté au modèle $\mathcal{M}_j(t)$, sa covariance $\mathbf{H}^j_{t-1/t-1}$ et la probabilité $\mu_{i/j}(t-1/t-1)$, l'estimée initiale $\mathbf{f}^{0j}_{t-1/t-1}$ et sa covariance $\mathbf{H}^{0j}_{t-1/t-1}$ pour le filtre adapté au modèle $\mathcal{M}_j(t)$ seront calculées selon :

$$\begin{aligned}
\mathbf{f}^{0j}_{t-1/t-1} &= \sum_{i=1}^{m} \mathbf{f}^{j}_{t-1/t-1}\mu_{i/j}(t-1/t-1) \\
\mathbf{H}^{0j}_{t-1/t-1} &= \sum_{i=1}^{m} \mu_{i/j}(t-1/t-1)\left[\mathbf{H}^{j}_{t-1/t-1} + \left(\mathbf{f}^{j}_{t-1/t-1} - \mathbf{f}^{0j}_{t-1/t-1}\right)\right. \\
&\qquad \left.\left(\mathbf{f}^{j}_{t-1/t-1} - \mathbf{f}^{0j}_{t-1/t-1}\right)'\right] \quad \text{pour } j = 1, ..., m
\end{aligned}$$

La troisième étape, c'est une étape de filtrage conditionnel aux modèles. En utilisant l'état estimé initial, sa matrice de covariance et les mesures reçues à l'instant t, on calcule les estimations $\left\{\mathbf{f}^{j}_{t/t},\ j = 1, ..., m\right\}$ conditionnelles aux modèles et leurs matrices de covariance $\left\{\mathbf{H}^{j}_{t/t},\ j = 1, ..., m\right\}$. Dans une quatrième étape les fonctions de vraisemblance seront calculées. Au niveau de l'étape cinq, les probabilités du modèle $\mu_j(t)$ seront mises à jour en utilisant la formule :

$$\mu_j(t) = p(S_t = j|\mathcal{Y}_{1:t}) = \frac{1}{c}\Lambda_j(t)c_j$$

où $\Lambda_j(t)$ est la vraisemblance du modèle j à l'instant t et $c = \sum_{j=1}^{m} \Lambda_j(t)c_j$. Dans une dernière étape, l'état estimé global $\mathbf{f}_{t/t}$ et sa matrice de covariance seront calculés par :

$$\begin{aligned}
\mathbf{f}_{t/t} &= \sum_{j=1}^{m} \mu_j(t)\mathbf{f}^{j}_{t/t} \\
\mathbf{H}_{t/t} &= \sum_{j=1}^{m} \mu_j(t)\left[\mathbf{H}^{j}_{t/t} + \left(\mathbf{f}^{j}_{t/t} - \mathbf{f}_{t/t}\right)\left(\mathbf{f}^{j}_{t/t} - \mathbf{f}_{t/t}\right)'\right]
\end{aligned}$$

L'Algorithme de Viterbi Approximé : Étant donnée la dépendance de la vraisemblance des observations de tout l'historique des états cachés, l'utilisation d'un algorithme de Viterbi exacte (Viterbi [1967]) n'est plus envisageable dans ce cas. Il n'est donc pas possible de reconstruire le chemin le plus probable qu'en introduisant certaines approximations pour déterminer d'une manière récursive les états $\widehat{S}_t$ permettant de maximiser la probabilité $p(\mathcal{F}_{1:n}, \mathcal{S}_{1:n}|\mathcal{Y}_{1:n})$. Pour une application de cette version approximée aux modèles espace-état avec changement de régime, voir Pavlovic, Rehg, Cham, et Murphy [1999] et Murphy [2002]. Dans le chapitre 5 de ce travail, nous présentons aussi une approche basée sur l'approximation de viterbi pour l'inférence des structures cachées et l'estimation des paramètres d'un modèle à facteurs conditionnellement hétéroscédastiques avec changement de régime.

Les Méthodes Itératives : Ces méthodes sont basées sur les techniques de Monte Carlo par chaînes de Markov (MCMC) pour simuler les lois de probabilité a posteriori des états cachés. L'idée consiste à proposer une structure "universelle" de simulation permettant d'obtenir un échantillon d'une loi quelconque sans jamais simuler directement suivant celle-ci, en faisant appel à une chaîne de Markov ergodique de loi stationnaire la loi d'intérêt (Smith et Markov [1980], Carter et Kohn [1994, 1996], Billio, Monfort et Robert [1998], Doucet et Andrieu [2001] et Rosti et Gales [2004]). Dans le cas des modèles espace-état avec changement de régime, Rosti et Gales [2004] ont proposé un algorithme de Rao-Blackwell à la Gibbs permettant dans une première étape de reconstruire efficacement la séquence des états markoviens en utilisant la distribution $p(S_t|\mathcal{Y}_{1:n}, \mathcal{S}_{\bar{t}})$. Dans une seconde étape, la segmentation optimale sera utilisée pour l'inférence des états continus en se basant sur la distribution $p(\mathbf{f}_t|\mathcal{Y}_{1:n}, \{S_i\})$. Rosti et Gales ont démontré aussi que cet algorithme converge presque sûrement vers les vraies statistiques a posteriori $p(S_t = j|\mathcal{Y}_{1:n})$, $\widehat{\mathbf{f}}_t = \mathbb{E}(\mathbf{f}_t|\mathcal{Y}_{1:n})$ et $\widehat{\mathbf{H}}_t = \mathbb{E}(\mathbf{f}_t\mathbf{f}_t'|\mathcal{Y}_{1:n})$.

Les Méthodes d'Approximation Variationnelles : La technique générale d'inférence variationnelle structurée pour les réseaux bayésiens dynamiques a été déjà présentée par Saul et Jordan [1996], Jordan, Ghahramani, Jaakkola, et Saul [1999] et par la suite par Pavlovic, Rehg, Cham et Murphy [1999], Pavlovic, Rehg et MacCormick [2000] et Ghahramani et Hinton [2000] pour le cas particulier des modèles espace-état linéaires avec changement de régime markovien. L'idée derrière cette méthode consiste à trouver une distribution de probabilité $Q(\mathcal{F}, \mathcal{S}/\mathcal{Y})$, permettant d'approcher la distribution a posteriori $p(\mathcal{F}, \mathcal{S}/\mathcal{Y})$ des états cachés, et ainsi de faciliter le calcul, en éliminant certaines relations de dépendance conditionnelles. Pour ce faire, Ghahramani et Hinton [2000] ont proposé un algorithme en deux étapes qui alterne des phases "Avant-Arrière" (pour le HMM), avec des phases d'inférence (pour chacun des modèles espace-état $\mathcal{M}_j$, $j = 1, ..., m$). Les paramètres de la chaîne de Markov seront estimés en utilisant les statistiques exhaustives obtenues au niveau de la deuxième étape, et les paramètres des différents modèles espace-états seront estimés en utilisant les probabilités a posteriori des états discrets obtenues au niveau de la première étape.

4.3.3 Inférence des Structures Cachées : Méthode GPB(1)

L'inférence des structures cachées dans les modèles espace-état avec changement de régime est beaucoup plus compliquée que celle des modèles linéaires. Cette complexité provient essentiellement de la diversité des composants, de la diversité des structures et de la diversité des interactions mises en jeu. Dans ce cas, l'état du système à un instant t quelconque dépend de la séquence complète des états cachés jusqu'à la date t. Le calcul exacte de la vraisemblance des observations implique, donc, une somme sur un nombre exponentiel de séquences d'états possibles. Lorsque la séquence complète des états cachés est connue, on peut alors appliquer les techniques habituelles de filtrage et de lissage de Kalman permettant, dans une première étape, d'estimer les états continus et leurs variance. Dans une seconde étape, ces nouvelles valeurs seront utilisées pour l'estimation de l'ensemble des paramètres du modèle en utilisant, par exemple, un algorithme de type EM.

Dans toute la suite, nous allons utiliser la méthode GPB(1) pour l'implémentation des algorithmes de filtrage et de lissage. Afin de pouvoir décrire ces algorithmes, nous allons introduire tout d'abord les notations suivantes :

$$\begin{aligned}
\mathbf{f}_{t/\tau}^{i(j)} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_{t-1}=i, S_t=j\right] \\
\mathbf{f}_{t/\tau}^{(j)k} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t=j, S_{t+1}=k\right] \\
\mathbf{f}_{t/\tau}^{j} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t=j\right]
\end{aligned}$$

Si $\tau = t$, ces dernières seront appelées statistiques de filtrage ; si $\tau > t$, on les appelle statistiques de lissage ; et si $\tau < t$ on les appelle statistiques de prédiction. Notons aussi que l'indexe entre parenthèses représente la valeur du noeud de changement à la date t ; le terme à gauche représente la valeur de S_{t-1}, et celui qui se trouve à droite, représente la valeur de S_{t+1}. Ces distinctions sont nécessaires pour mieux caractériser les termes de covariance. Nous définissons aussi les statistiques suivantes :

$$\begin{aligned}
V_{t/\tau}^{j} &= Cov\left(\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t=j\right) \\
V_{t,t-1/\tau}^{j} &= Cov\left(\mathbf{f}_t, \mathbf{f}_{t-1}/\mathcal{Y}_{1:\tau}, S_t=j\right) \\
V_{t,t-1/\tau}^{i(j)} &= Cov\left(\mathbf{f}_t, \mathbf{f}_{t-1}/\mathcal{Y}_{1:\tau}, S_{t-1}=i, S_t=j\right) \\
M_{t-1,t/\tau}(i,j) &= p\left(S_{t-1}=i, S_t=j/\mathcal{Y}_{1:\tau}\right) \\
M_{t/\tau}(j) &= p(S_t=j/\mathcal{Y}_{1:\tau}) \\
L_t(i,j) &= p(\mathbf{y}_t/\mathcal{Y}_{1:t-1}, S_{t-1}=i, S_t=j)
\end{aligned}$$

où $L_t(i,j)$ est la vraisemblance de l'innovation à l'instant t, lorsque le système est dans le régime j.

L'Algorithme de Filtrage

Nous effectuons les opérations suivantes successivement :

$$\begin{aligned}
\mathbf{f}_{t/t-1}^{i(j)} &= \mathbf{A}_j \mathbf{f}_{t-1/t-1}^{(i)} \\
V_{t/t-1}^{i(j)} &= \mathbf{A}_j V_{t-1/t-1}^{(i)} \mathbf{A}_j' + \mathbf{Q}_j
\end{aligned}$$

Nous calculons par la suite l'erreur de prédiction (l'innovation), la variance de l'erreur, la matrice de gain de Kalman, et la vraisemblance de cette observation :

$$e_t(i,j) = \mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{t/t-1}^{i(j)} - \theta_j$$

$$\mathbf{\Sigma}_t^{i(j)} = \mathbf{X}_j V_{t/t-1}^{i(j)} \mathbf{X}_j' + \mathbf{\Psi}_j$$

$$K_t(i,j) = V_{t/t-1}^{i(j)} \mathbf{X}_j' \mathbf{\Sigma}_t^{i(j)\,-1}$$

$$L_t(i,j) = \mathcal{N}\left[\mathbf{0}, \mathbf{\Sigma}_t^{i(j)}\right]$$

Ensuite, nous mettons à jour nos estimations des moyennes, des variances, et des covariances, soient

$$\begin{aligned}
\mathbf{f}_{t/t}^{i(j)} &= \mathbf{f}_{t/t-1}^{i(j)} + K_t(i,j)e_t(i,j) \\
V_{t/t}^{i(j)} &= \left[\mathbf{I}_k - K_t(i,j)\mathbf{X}_j\right] V_{t/t-1}^{i(j)} = V_{t/t-1}^{i(j)} - K_t(i,j)\mathbf{\Sigma}_t^{i(j)} K_t(i,j)' \\
V_{t,t-1/t}^{i(j)} &= \left[\mathbf{I}_k - K_t(i,j)\mathbf{X}_j\right] \mathbf{A}_j V_{t-1/t-1}^{(i)}
\end{aligned}$$

La méthode GPB(1) basée sur la fusion des moments conditionnel nécessite, aussi, le calcul des probabilités suivantes :

$$M_{t-1,t/t}(i,j) = p(S_{t-1} = i, S_t = j/\mathcal{Y}_{1:t}) = \frac{L_t(i,j)p_{ij}M_{t-1/t-1}(i)}{\sum_{i=1}^{m}\sum_{j=1}^{m} L_t(i,j)p_{i,j}M_{t-1/t-1}(i)}$$

étant donné que :

$$\begin{aligned}
M_{t-1,t/t}(i,j) &= p(S_{t-1} = i, S_t = j/\mathcal{Y}_{1:t}) \\
&= p(S_{t-1} = i, S_t = j/\mathbf{y}_t, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(S_{t-1} = i, S_t = j, \mathbf{y}_t/\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(\mathbf{y}_t/S_{t-1} = i, S_t = j, \mathcal{Y}_{1:t-1}) p(S_{t-1} = i, S_t = j/\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(\mathbf{y}_t/S_{t-1} = i, S_t = j, \mathcal{Y}_{1:t-1}) p(S_{t-1} = i/\mathcal{Y}_{1:t-1}) \times \\
&\qquad p(S_t = j/S_{t-1} = i, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} L_t(i,j)p_{ij}M_{t-1/t-1}(i)
\end{aligned}$$

où c est la constante de normalisation donnée par :

$$c = \sum_{i=1}^{m}\sum_{j=1}^{m} L_t(i,j)p_{ij}M_{t-1/t-1}(i)$$

Nous calculons aussi les probabilités

$$M_{t/t}(j) = \sum_{i=1}^{m} M_{t-1,t/t}(i,j)$$

$$Z_{i/j}(t) = p(S_{t-1} = i/S_t = j, \mathcal{Y}_{1:t}) = M_{t-1,t/t}(i,j)/M_{t/t}(j)$$

En dernière étape, les moyennes et les variances seront mises à jour à travers les équations suivantes :

$$\begin{aligned}
\mathbf{f}_{t/t}^{j} &= \sum_{i=1}^{m} Z_{i/j}(t)\mathbf{f}_{t/t}^{i(j)} \\
V_{t/t}^{j} &= \sum_{i=1}^{m} Z_{i/j}(t)V_{t/t}^{i(j)} + \sum_{i=1}^{m} Z_{i/j}(t)\left[\mathbf{f}_{t/t}^{i(j)} - \mathbf{f}_{t/t}^{j}\right]\left[\mathbf{f}_{t/t}^{i(j)} - \mathbf{f}_{t/t}^{j}\right]'
\end{aligned}$$

Pour l'initialisation de cet algorithme on prend $\mathbf{f}_{1/0}^{j} = \mathbb{E}(\mathbf{f}_1/S_1 = j) = \theta^j$ et $V_{1/0}^{j} = Cov(\mathbf{f}_1/S_1 = j) = \mathbf{\Sigma}^j$, et on pose $M_{0/0} = \pi$.

L'Algorithme de Lissage

Les statistiques de prédiction seront, tout d'abord, récupérées à partir de l'algorithme de filtrage, soient

$$\begin{aligned}\mathbf{f}_{t+1/t}^{i(j)} &= \mathbf{A}_j\mathbf{f}_{t/t}^{i}\\ V_{t+1/t}^{i(j)} &= \mathbf{A}_j V_{t/t}^{i}\mathbf{A}_j' + \mathbf{Q}_j\end{aligned}$$

par la suite, nous calculons les matrices de gain de lissage

$$J_t^{(j)k} = V_{t/t}^{j}\mathbf{A}_k' V_{t+1/t}^{(j)k\,-1}$$

et nous mettons à jour les estimations des moyennes, des variances et des covariances, soient

$$\mathbf{f}_{t/n}^{(j)k} = \mathbf{f}_{t/t}^{j} + J_t^{(j)k}\left[\mathbf{f}_{t+1/n}^{k} - \mathbf{f}_{t+1/t}^{j(k)}\right]$$

$$V_{t/n}^{(j)k} = V_{t/t}^{j} + J_t^{(j)k}\left[V_{t+1/n}^{k} - V_{t+1/t}^{j(k)}\right]J_t^{(j)k'}$$

$$V_{t+1,t/n}^{j(k)} = V_{t+1,t/t+1}^{j(k)} + \left[V_{t+1/n}^{k} - V_{t+1/t+1}^{k}\right]V_{t+1/t+1}^{k\,-1}V_{t+1,t/t+1}^{j(k)}$$

Les termes de covariances calculés au niveau de cette étape de lissage, $V_{t,t-1/n}$, pourront être obtenus sans recourir aux termes de filtrage qui leurs correspondent (voir par exemple, Shumway et Stoffer [1991] et Ghahramani et Hinton [1996]). Dans ce cas la fonction de lissage sera donnée par :

$$\left(\mathbf{f}_{t/n}^{(j)k}, V_{t/n}^{(j)k}, V_{t,t-1/n}^{i(j)}\right) = \mathcal{L}iss\left[\mathbf{f}_{t+1/n}^{k}, V_{t+1/n}^{k}, V_{t+1,t/n}^{j(k)}, \mathbf{f}_{t/t}^{j}, V_{t/t}^{j}, V_{t-1/t-1}^{i}, \mathbf{A}_{t+1}, \mathbf{Q}_{t+1}, \mathbf{Q}_t\right]$$

avec

$$V_{t,t-1/n}^{i(j)} = V_{t/t}^{j}J_{t-1}^{(i)j\prime} + J_{t-1}^{(i)j}\left[V_{t+1,t/n}^{j(k)} - \mathbf{A}_{t+1}V_{t/t}^{j}\right]J_{t-1}^{(i)j\prime}$$

et où la condition au borne est donnée par :

$$V_{n,n-1/n} = \left[\mathbf{I} - K_n\mathbf{X}_n\right]\mathbf{A}_n V_{n-1/n-1}$$

Nous calculons par la suite les probabilités,

$$U_{t/t+1}^{j/k} = p(S_t = j/S_{t+1} = k, \mathcal{Y}_{1:n}) \simeq \frac{M_{t/t}(j)p_{jk}}{\sum\limits_{j'=1}^{m} M_{t/t}(j')p_{j'k}}$$

où l'approximation provient du fait que S_t n'est pas conditionnellement indépendante du futur $\mathbf{y}_{t+1}, ..., \mathbf{y}_n$ étant donné l'état S_{t+1} (voir Pearl [1988]).

$$\begin{aligned} p(S_t = j/S_{t+1} = k, \mathcal{Y}_{1:n}) &\simeq p(S_t = j/S_{t+1} = k, \mathcal{Y}_{1:t}) \\ &= \frac{p(S_t = j/\mathcal{Y}_{1:t})p(S_{t+1} = k/S_t = j)}{p(S_{t+1} = k/\mathcal{Y}_{1:t})} \end{aligned}$$

Pour l'implémentation de l'algorithme de lissage et la fusion des moments conditionnels, nous calculons aussi les probabilités suivantes :

$$M_{t,t+1/n}(j,k) = U^{j/k}_{t/t+1} M_{t+1/n}(k)$$

$$M_{t/n}(j) = \sum_{k=1}^{m} M_{t,t+1/n}(j,k)$$

$$Z_{k/j}(t+1) = p(S_{t+1} = k/S_t = j, \mathcal{Y}_{1:n}) = M_{t,t+1/n}(j,k)/M_{t/n}(j)$$

et les statistiques a posteriori

$$\begin{aligned} \mathbf{f}^j_{t/n} &= \sum_{k=1}^{m} Z_{k/j}(t+1)\mathbf{f}^{(j)k}_{t/n} \\ \mathbf{f}_{t/n} &= \sum_{j=1}^{m} M_{t/n}(j)\mathbf{f}^j_{t/n} \\ V^j_{t/n} &= \sum_{k=1}^{m} Z_{k/j}(t+1)V^{(j)k}_{t/n} + Z_{k/j}(t+1)\left[\mathbf{f}^{(j)k}_{t/n} - \mathbf{f}^j_{t/n}\right]\left[\mathbf{f}^{(j)k}_{t/n} - \mathbf{f}^j_{t/n}\right]' \\ V_{t/n} &= \sum_{j=1}^{m} M_{t/n}(j)V^j_{t/n} + \sum_{j=1}^{m} M_{t/n}(j)\left[\mathbf{f}^j_{t/n} - \mathbf{f}_{t/n}\right]\left[\mathbf{f}^j_{t/n} - \mathbf{f}_{t/n}\right]' \end{aligned}$$

par la suite, on pose

$$\mathbf{f}^{j(k)}_{t+1/n} = \mathbb{E}\left[\mathbf{f}_{t+1}/\mathcal{Y}_{1:n}, S_{t+1} = k, S_t = j\right] \simeq \mathbf{f}^k_{t+1/n}$$

Enfin, on calcule

$$\begin{aligned} V^k_{t+1,t/n} &= \sum_{j=1}^{m} U^{j/k}_{t/t+1}V^{j(k)}_{t+1,t/n} + \sum_{j=1}^{m} U^{j/k}_{t/t+1}\left[\mathbf{f}^{(j)k}_{t/n} - \mathbf{f}^{j(k)}_{t+1/n}\right]\left[\mathbf{f}^{(j)k}_{t/n} - \mathbf{f}^{j(k)}_{t+1/n}\right]' \\ \mathbf{f}^{()k}_{t/n} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:n}, S_{t+1} = k\right] = \sum_{j=1}^{m} \mathbf{f}^{(j)k}_{t/n} U^{j/k}_{t/t+1} \\ V_{t+1,t/n} &= \sum_{k=1}^{m} M_{t+1/n}(k)V^k_{t+1,t/n} + \sum_{k=1}^{m} M_{t+1/n}(k)\left[\mathbf{f}^{()k}_{t/n} - \mathbf{f}^k_{t+1/n}\right]\left[\mathbf{f}^{()k}_{t/n} - \mathbf{f}^k_{t+1/n}\right]' \end{aligned}$$

4.3.4 Optimisation des Paramètres et Algorithme EM

La log-vraisemblance complétée d'un modèle espace-état avec changement de régime, $\mathcal{L}(\mathcal{Y},\mathcal{F},\mathcal{S}) = p(\mathcal{F}_{1:n},\mathcal{S}_{1:n},\mathcal{Y}_{1:n}|\Theta)$, est donnée par :

$$\begin{aligned}
\mathcal{L}(\mathcal{Y},\mathcal{F},\mathcal{S}) = & - \frac{1}{2}\sum_{t=1}^{n}(\mathbf{y}_t - \mathbf{X}_{s_t}\mathbf{f}_t - \theta_{s_t})'\mathbf{\Psi}_{s_t}^{-1}(\mathbf{y}_t - \mathbf{X}_{s_t}\mathbf{f}_t - \theta_{s_t}) - \frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{\Psi}_{s_t}| \\
& - \frac{1}{2}\sum_{t=2}^{n}(\mathbf{f}_t - \mathbf{A}_{s_t}\mathbf{f}_{t-1})'\mathbf{H}_t^{-1}(\mathbf{f}_t - \mathbf{A}_{s_t}\mathbf{f}_{t-1}) - \frac{1}{2}\sum_{t=2}^{n}\log|\mathbf{H}_t| \\
& - \frac{1}{2}(\mathbf{f}_1 - \mu_1)'\mathbf{H}_1^{-1}(\mathbf{f}_1 - \mu_1) - \frac{1}{2}\log|\mathbf{H}_1| - \frac{n(q+k)}{2}\log 2\pi \\
& + \log\pi_1 + \sum_{t=2}^{n}\log p(S_t|S_{t-1}) \qquad (4.50)
\end{aligned}$$

où π_1 est la probabilité de l'état initial.

Étape E : La quantité qu'on cherche à maximiser (l'espérance conditionnelle de la log-vraisemblance complétée) est donnée par :

$$\begin{aligned}
\mathcal{Q}(\Theta,\Theta^{(i)}) &= \mathbb{E}_{p(\mathcal{S}_{1:n},\mathcal{F}_{1:n}/\mathcal{Y}_{1:n})}\Big[\log p(\mathcal{F}_{1:n},\mathcal{S}_{1:n},\mathcal{Y}_{1:n})\Big] \\
&= \mathbb{E}_{p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})}\Big[\mathbb{E}_{p(\mathcal{F}_{1:n}/\mathcal{S}_{1:n},\mathcal{Y}_{1:n})}\Big[\log p(\mathcal{F}_{1:n},\mathcal{S}_{1:n},\mathcal{Y}_{1:n})\Big]\Big] \\
&\simeq \mathbb{E}_{p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})}\Big[\mathbb{E}_{p(\mathcal{F}_{1:n}/\mathcal{Y}_{1:n})}\Big[\log p(\mathcal{F}_{1:n},\mathcal{S}_{1:n},\mathcal{Y}_{1:n})\Big]\Big] \\
&= p(\mathcal{Y}_{1:n})\sum_{t=2}^{n}\sum_{S_t}\left[\sum_{\{S_\tau,\,\tau\neq t\}} p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})\right]\widetilde{\mathbb{E}}\Big[\log p(\mathbf{f}_t/\mathbf{f}_{t-1},S_t)\Big] + \ldots. \\
&= p(\mathcal{Y}_{1:n})\sum_{t=2}^{n}\sum_{S_t=j} M_{t/n}(j)\widetilde{\mathbb{E}}\Big[\log p(\mathbf{f}_t/\mathbf{f}_{t-1},S_t)\Big] + \ldots. \qquad (4.51)
\end{aligned}$$

où $\widetilde{\mathbb{E}}[.] = \mathbb{E}[.|\mathcal{Y}_{1:n}]$. L'approximation est justifiée par le fait que nous avons utilisé $\mathbb{E}[\mathbf{f}_t/\mathcal{Y}_{1:n}]$ au lieu de $\mathbb{E}[\mathbf{f}_t/\mathcal{Y}_{1:n},\mathcal{S}_{1:n}]$, étant donné que le dernier terme est un nombre exponentiel de vecteurs (un pour chaque segmentation).

Étape M : Pour la mise à jour des paramètres du modèle, nous pouvons utiliser directement les valeurs $\mathbf{f}_{t/n}^j$, $V_{t/n}^j$, et $V_{t,t-1/n}^{i(j)}$ déjà calculées au niveau de l'étape de filtrage. Ceci nous permet d'éviter le calcul des termes de covariances $V_{t,t-1/n}$ et ainsi de ne pas alourdir les calculs avec les deux dernières opérations de fusion (les quatre dernières équations de la section précédente).

Pour l'implémentation des formules de mise à jour, nous introduisons tout d'abord les notations suivantes :

$$\begin{aligned}
\widetilde{\mathbf{f}}_t &= \widetilde{\mathbb{E}}[\mathbf{f}_t] \\
\widetilde{\mathbf{R}}_t &= \widetilde{\mathbb{E}}\left[\mathbf{f}_t\mathbf{f}_t'\right] = V_{t/n} + \mathbf{f}_{t/n}\mathbf{f}_{t/n}' \\
\widetilde{\mathbf{R}}_{t,t-1} &= \widetilde{\mathbb{E}}\left[\mathbf{f}_t\mathbf{f}_{t-1}'\right] = V_{t,t-1/n} + \mathbf{f}_{t/n}\mathbf{f}_{t-1/n}'
\end{aligned}$$

1- Mise à jour de la matrice de transition des états continus

Les dérivées de la fonction auxiliaire (4.51) par rapport aux matrices $\mathbf{A}_j$, pour $j = 1, ..., m$, sont données par :

$$\begin{aligned}\frac{\partial \mathcal{Q}(\Theta, \Theta^{(i)})}{\partial \mathbf{A}_j} &= -\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbb{E}}\left[\mathbf{Q}_j^{-1}(\mathbf{f}_t - \mathbf{A}_j\mathbf{f}_{t-1})\mathbf{f}'_{t-1}\right] \\ &= -\sum_{t=2}^{n} M_{t/n}(j)\mathbf{Q}_j^{-1}\widetilde{\mathbf{R}}_{t,t-1} + \sum_{t=2}^{n} M_{t/n}(j)\mathbf{Q}_j^{-1}\mathbf{A}_j\widetilde{\mathbf{R}}_{t-1}\end{aligned}$$

La résolution des conditions du premier ordre nous permettra de trouver :

$$\widehat{\mathbf{A}}_j = \left[\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_{t,t-1}\right]\left[\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_{t-1}\right]^{-1}$$

2- Mise à jour de la matrice de covariance de l'équation de transition

Les dérivées premières de (4.51) par rapport à $\mathbf{Q}_j^{-1}$ sont données par :

$$\begin{aligned}\frac{\partial \mathcal{Q}(\Theta, \Theta^{(i)})}{\partial \mathbf{Q}_j^{-1}} &= -\frac{1}{2}\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbb{E}}\left[(\mathbf{f}_t - \mathbf{A}_j\mathbf{f}_{t-1})(\mathbf{f}_t - \mathbf{A}_j\mathbf{f}_{t-1})'\right] + \frac{1}{2}\sum_{t=2}^{n} M_{t/n}(j)\mathbf{Q}_j \\ = -\frac{1}{2}\sum_{t=2}^{n} M_{t/n}(j) &\left[\widetilde{\mathbf{R}}_t - \mathbf{A}_j\widetilde{\mathbf{R}}'_{t,t-1} - \widetilde{\mathbf{R}}_{t/t-1}\mathbf{A}'_j + \mathbf{A}_j\widetilde{\mathbf{R}}_{t-1}\mathbf{A}'_j\right] + \frac{1}{2}\sum_{t=2}^{n} M_{t/n}(j)\mathbf{Q}_j\end{aligned}$$

En utilisant la nouvelle valeur de $\mathbf{A}_j$ et le fait que $\widetilde{\mathbf{R}}_t$ est symétrique, la résolution des conditions du premier ordre nous permettra de trouver :

$$\begin{aligned}&\mathbf{A}_j\left(\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_{t-1}\right)\mathbf{A}'_j = \\ &\quad\left(\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_{t,t-1}\right)\left(\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_{t-1}\right)^{-1}\left(\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}'_{t,t-1}\right) \\ &\quad= \mathbf{A}_j\left(\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}'_{t,t-1}\right) = \left(\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_{t,t-1}\right)\mathbf{A}'_j\end{aligned}$$

ce qui implique,

$$\widehat{\mathbf{Q}}_j = \frac{1}{\sum\limits_{t=2}^{n} M_{t/n}(j)}\left\{\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}_t - \widehat{\mathbf{A}}_j\sum_{t=2}^{n} M_{t/n}(j)\widetilde{\mathbf{R}}'_{t,t-1}\right\}$$

3- Mise à jour des moyennes θ_j

Les dérivées de la fonction auxiliaire (4.51) par rapport à θ_j donnent :

$$\frac{\partial \mathcal{Q}(\Theta,\Theta^{(i)})}{\partial \theta_j} = \mathbf{\Psi}_j^{-1} \sum_{t=1}^{n} M_{t/n}(j) \left(\mathbf{y}_t - \mathbf{X}_j \widetilde{\mathbf{f}}_t - \theta_j \right)$$

et la résolution des conditions du premier ordre donne :

$$\widehat{\theta}_j = \frac{\sum_{t=1}^{n} M_{t/n}(j) \left(\mathbf{y}_t - \mathbf{X}_j \widetilde{\mathbf{f}}_t \right)}{\sum_{t=1}^{n} M_{t/n}(j)}$$

4- Mise à jour de la matrice de mesure

Les dérivées premières de (4.51) par rapport à $\mathbf{X}_j$ sont données par :

$$\frac{\partial \mathcal{Q}(\Theta,\Theta^{(i)})}{\partial \mathbf{X}_j} = -\frac{1}{2} \sum_{t=1}^{n} M_{t/n}(j) \widetilde{\mathbb{E}} \left[2\mathbf{\Psi}_j^{-1} (\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t - \theta_j) \mathbf{f}_t' \right]$$

et la résolution des conditions du premier ordre permet de trouver :

$$\widehat{\mathbf{X}}_j = \left[\sum_{t=1}^{n} M_{t/n}(j) \left(\mathbf{y}_t - \widehat{\theta}_j \right) \widetilde{\mathbf{f}}_t' \right] \left[\sum_{t=1}^{n} M_{t/n}(j) \widetilde{\mathbf{R}}_t \right]^{-1}$$

5- Mise à jour de la matrice de covariance de l'équation de mesure

Les dérivées de (4.51) par rapport à $\mathbf{\Psi}_j^{-1}$ donnent :

$$\begin{aligned} &\frac{\partial \mathcal{Q}(\Theta,\Theta^{(i)})}{\partial \mathbf{\Psi}_j^{-1}} = \\ &\frac{1}{2} \sum_{t=1}^{n} \widetilde{\mathbb{E}} \left[M_{t/n}(j) \left[(\mathbf{y}_t - \theta_j)(\mathbf{y}_t - \theta_j)' - 2\mathbf{X}_j \mathbf{f}_t (\mathbf{y}_t - \theta_j)' + \mathbf{X}_j \mathbf{f}_t \mathbf{f}_t' \mathbf{X}_j' \right] \right] \\ &+ \frac{1}{2} \mathbf{\Psi}_j \sum_{t=1}^{n} M_{t/n}(j) \end{aligned}$$

et en utilisant les nouvelles valeurs de θ_j et $\mathbf{X}_j$, on obtient :

$$\left(\sum_{t=1}^{n} M_{t/n}(j) \widetilde{\mathbf{R}}_t \right) \mathbf{X}_j' = \sum_{t=1}^{n} M_{t/n}(j) \widetilde{\mathbf{f}}_t (\mathbf{y}_t - \theta_j)' \stackrel{def}{=} \mathcal{Z}$$

ce qui implique

$$\begin{aligned}\frac{\partial \mathcal{Q}(\Theta,\Theta^{(i)})}{\partial \mathbf{\Psi}_j^{-1}} &= \frac{1}{2}\sum_{t=1}^{n}\left[\sum_{t=1}^{n} M_{t/n}(j)(\mathbf{y}_t-\theta_j)(\mathbf{y}_t-\theta_j)' - 2\mathbf{X}_j\mathcal{Z} + \mathbf{X}_j\mathcal{Z}\right] \\ &+ \frac{1}{2}\mathbf{\Psi}_j\sum_{t=1}^{n} M_{t/n}(j)\end{aligned}$$

La résolution des conditions du premier ordre donne :

$$\widehat{\mathbf{\Psi}}_j = \frac{1}{\sum\limits_{t=1}^{n} M_{t/n}(j)}\left\{\sum_{t=1}^{n} M_{t/n}(j)\left[\left(\mathbf{y}_t-\widehat{\theta}_j\right)\left(\mathbf{y}_t-\widehat{\theta}_j\right)' - \widehat{\mathbf{X}}_j\widetilde{\mathbf{f}}_t\left(\mathbf{y}_t-\widehat{\theta}_j\right)'\right]\right\}$$

6- Estimation des paramètres de la chaîne de Markov

Si on suppose que la chaîne a commencé à l'état j, l'utilisation du multiplicateur de Lagrange[8], sous la contrainte $\sum\limits_{j=1}^{m}\pi_j = 1$, où $\pi_j = p(S_1 = j)$ et $p(S_1) = \pi = [\pi_1, \pi_2, ..., \pi_m]'$, nous permet de trouver :

$$\widehat{\pi}_j = \frac{M_{1/n}(j)}{\sum\limits_{i=1}^{m} M_{1/n}(i)}$$

La maximisation de la fonction auxiliaire (4.51) par rapport aux probabilités de transition p_{ij}, en utilisant aussi le multiplicateur de Lagrange, sous la contrainte $\sum\limits_{j=1}^{m} p_{ij} = 1$ nous permettra de trouver :

$$\widehat{p}_{ij} = \frac{\sum\limits_{t=2}^{n} M_{t-1,t/n}(i,j)}{\sum\limits_{t=2}^{n} M_{t-1/n}(i)}$$

[8] Pour plus de détails, voir par exemple Rabiner [1989], Hamilton [1990], Bishop [1995] et Xu et Jordan [1996]. Dans l'annexe du chapitre 5, on dérive exactement les mêmes formules pour le cas des modèles à facteurs conditionnellement hétéroscédastiques.

CHAPITRE 5

Modèles à Facteurs Dynamiques et Changement de Régime

Dans ce chapitre, nous étudions une classe de modèles à facteurs dynamiques et à structure markovienne cachée pour les séries financières conditionnellement hétéroscédastiques avec changement de régime. Pour la modélisation de ces changements, nous avons proposé une nouvelle approche basée sur la combinaison des modèles à facteurs, déjà présentés dans le chapitre 3, avec les modèles de chaînes de Markov cachés. L'idée originale de ce travail est la modélisation de cette non stationnarité à l'aide d'un processus multivarié et linéaire par morceaux que l'on peut considérer, aussi, comme un système linéaire et dynamique à états mixtes. En particulier, nous avons supposé que les séries observées peuvent être approchées à l'aide d'un modèle dont les paramètres évoluent au cours du temps. Nous avons émis, aussi, l'hypothèse que l'évolution de ces paramètres est gouvernée par une variable inobservable que l'on peut modéliser à l'aide d'une chaîne de Markov à m régimes. Pour l'inférence des structures cachées et l'estimation des paramètres nous, avons proposé deux approches différentes fondées sur le principe de l'algorithme EM généralisé. Les différents régimes, les facteurs communs et leurs volatilités sont supposés non observables et l'inférence doit être menée à partir du processus observable.

5.1 Introduction

Le phénomène de variance non constante dans les séries chronologiques de rendements d'actifs financiers est connu depuis longtemps. Ces dernières ont tendance à exhiber des successions de phases de relative tranquillité et de phases de forte volatilité. Dans le cas général, tout ce qui est variation dans le régime de taux de change, déréglementation, ouverture financière, les débâcles des marchés financiers, chocs de politique tels que les réformes fiscales et les réformes du commerce, peuvent être modéliser en tant que des changements de la variance d'un des chocs. Sur la base de ces observations, la question-clé aujourd'hui est de savoir si cette non constance correspond à un changement structurel de tendance ou à une phase de turbulences conjoncturelles.

Dans un cadre univarié plusieurs approches ont été proposées pour la modélisation de cette non constance (voir Lamoureux et Lastrapes [1990]). Une première approche consiste à travailler au niveau de la volatilité conditionnelle en utilisant, pour modéliser celle-ci, un modèle de la famille des ARCH. Une autre approche consiste à travailler au niveau la volatilité non conditionnelle en utilisant des modèles à changement ou à saut de régime (Hamilton [1988] et [1989]). L'on peut également envisager une approche qui fait la synthèse des deux approches précédentes, dans le sens où l'on introduit des changements de régime dans un modèle du type ARCH pour tenir compte de la non constance de la volatilité non conditionnelle (Lastrapes [1989], Lamoureux et Lastrapes [1990], Gray [1996], Aggarval, Inclan et Leal [1999]).

Le modèle qu'on se propose d'étudier dans ce chapitre est une généralisation du modèle à facteurs conditionnellement hétéroscédastiques du chapitre 3. Dans ce cas et au lieu de tenir compte de la dynamique des variances des facteurs communs seulement, nous allons considérer le cas où tous les paramètres du modèle sont dynamiques à travers le temps. Le reste de ce chapitre est organisé comme suit : Dans une deuxième section, nous introduisons la forme générale du modèle dans sa version la plus simple. Il s'agit en fait d'un modèle d'analyse factorielle standard combiné avec un processus markovien non observable d'ordre un. Nous étudions, par la suite, sa fonction de vraisemblance pour enfin estimer ses paramètres en utilisant un algorithme EM exacte inspiré de l'algorithme de Baum et Welch pour les HMM. Dans la troisième section nous allons étendre le modèle standard pour l'étude des co-mouvements des séries financières caractérisées par une hétéroscédasticité dynamique au niveau de la variance. Nous allons l'étudier par la suite dans une structure espace-état multi-régime, afin d'aboutir finalement à des estimations pour les facteurs en utilisant une version quasi-optimale du filtre de Kalman basée sur la technique de "moment matching" (appelée aussi méthode pseudo-bayésienne généralisée dans la littérature sur les modèles espace-état avec changement de régime). Dans la section 4, nous allons présenter une autre approche alternative pour l'inférence des structures cachées et l'estimation des paramètres de ces modèles fondée sur l'approximation de viterbi. La fonction de vraisemblance et l'algorithme EM seront présentés dans la cinquième section, où nous allons discuter l'estimation des paramètres de la composante conditionnellement hétéroscédastique basée sur la restauration des états cachées de la chaîne de Markov en utilisant, soit les probabilités a posteriori déjà fournies par l'algorithme de lissage, soit la séquence optimale obtenue par l'approximation de Viterbi. Finalement, dans les sections 6 et 7 plusieurs expérimentations basées aussi bien sur des simulations que sur l'analyse d'une base de données financières (rendements journaliers de taux de change), seront menées en utilisant différentes spécifications, afin d'étudier certaines propriétés des algorithmes d'estimation et d'inférence des structures latentes qu'on a proposé.

5.2 Structure Markovienne à Facteurs Statiques

Dans le chapitre 2, le modèle d'analyse factorielle a été présenté dans un cadre gaussien statique comme étant une méthode multivariée qui vise à expliquer des rapports parmi plusieurs variables corrélées et difficiles à interpréter avec des facteurs relativement indépendants mais conceptuellement peu significatifs. Ce modèle s'approche beaucoup de la méthode d'analyse en composantes principales mais la différence fondamentale est que les composantes principales n'impliquent aucun modèle mathématique, tandis que l'analyse factorielle est une méthode plus avancée qui en emploie un en le choisissant par des procédures probabilistes telles que le maximum de vraisemblance (Everitt et Dunn [1991]). La généralisation de ces modèles peut se faire à travers les modèles de mélange gaussiens pour les facteurs spécifiques. Un tel modèle est connu sous l'appellation modèle d'analyse factorielle partagé. Dans les modèles d'analyse factorielle à structure markovienne cachée (FAHMM), les facteurs communs et spécifiques sont générés par un modèle HMM à états gaussiens. Les paramètres du modèle changent donc en fonction de l'état prévalant du HMM.

5.2.1 La Structure Générale du FAHMM

Le modèle d'analyse factorielle à structure markovienne cachée est une généralisation espace-état dynamique des modèles standards. Les vecteurs d'état continu de dimension k (facteurs communs) seront générés par des modèles HMM à états gaussiens. Les vecteurs d'observations de dimension q seront, ainsi, générés par des modèles d'analyse factorielle à composantes de bruit multiples. La structure générale des modèles FAHMM est donnée par :

$$\boxed{\begin{array}{c} S_t \sim P(S_t = j / S_{t-1} = i) \\[1em] \text{pour} \quad t = 1, ..., n \quad \text{et} \quad i, j = 1, ..., m \\[1em] \mathbf{y}_t = \mathbf{X}_{s_t} \mathbf{f}_{s_t} + \varepsilon_{s_t} \quad \text{avec} \quad \left\{ \begin{array}{l} \varepsilon_j \sim \mathcal{N}(\theta_j \, , \mathbf{\Psi}_j) \\[1em] \mathbf{f}_j \sim \mathcal{N}(\mathbf{0} \, , \mathbf{H}_j) \end{array} \right. \end{array}}$$

où S_t est une chaîne de Markov homogène à états cachés,[1] les p_{ij} sont les probabilités de transition d'un état i à un état j. Les paramètres $\mathbf{0}$ et θ_j sont, respectivement, les moyennes du vecteur des facteurs communs et du vecteur des facteurs spécifiques à un état quelconque $S_t = j$. Les matrices de pondérations et les matrices de variance-covariance diagonales des facteurs spécifiques ε_t et des facteurs communs $\mathbf{f}_t$ pour chaque état j sont, respectivement, désignées par $\mathbf{X}_j$, $\mathbf{\Psi}_j$ et $\mathbf{H}_j$. Un réseau bayésien dynamique décrivant les modèles FAHMM est donné dans la figure 5.1. [2]

Dans ce cas on suppose l'indépendance conditionnelle entre les variables qui ne sont pas connectées par des arcs directes. Ainsi, comme pour les HMM, on suppose l'indépendance conditionnelle entre les outputs $\mathbf{y}_t$ sachant les états cachés.

Une fois que notre modèle est établi et que nous avons une séquence d'observations, il reste à pouvoir passer de l'observation au modèle. Pour cela, il faut en fait résoudre trois problèmes : le premier problème consiste à calculer $p(\mathcal{Y}/\Theta)$, la probabilité de la séquence d'observations, étant donné le modèle et les observations $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$.

[1] Le symbole $\sim$ dans $S_t \sim P(S_t/S_{t-1})$ est utilisé pour représenter une chaîne de Markov discrète. Normalement ce symbole indique que la variable du membre gauche est distribuée selon la fonction de densité de probabilité du membre droite.

[2] Les réseaux bayésiens dynamiques peuvent être présentés en conjonction avec les modèles génératifs afin d'illustrer l'hypothèse d'indépendance conditionnelle dans un modèle statistique.

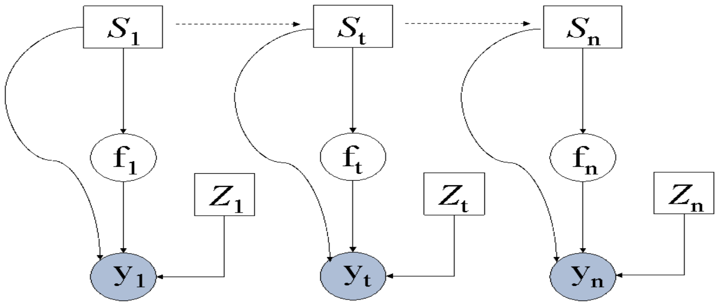

FIG. 5.1 – Modèle Graphique d'un FAHMM. Les noeuds rectangulaires représentent les variables aléatoires discrètes, c-à-d les états HMM $\{S_t\}$. Les variables aléatoires continues, c-à-d les facteurs communs $\mathbf{f}_t$, sont représentées par des noeuds arrondies. Les noeuds hachurés désignent les variables observables, $\mathbf{y}_t$. Les $\mathbf{z}_t$ sont des variables exogènes (observables) que l'on peut, éventuellement, introduire dans le modèle comme étant des variables explicatives.

Le deuxième problème est celui de l'ajustement des paramètres du modèle permettant de maximiser $p(\mathcal{Y}/\Theta)$ et enfin le troisième consiste à chercher la séquence d'états optimale qui correspond le mieux aux observations.

5.2.2 Calcul de la Fonction de Vraisemblance

L'aspect important de n'importe que modèle génératif est la complexité de calcul de sa fonction de vraisemblance. Le modèle génératif ci-dessus peut être exprimé par les deux distributions gaussiennes suivantes :

$$p(\mathbf{f}_t/S_t = j) = \mathcal{N}(\mathbf{0}, \mathbf{H}_j) \quad (5.1)$$

$$p(\mathbf{y}_t/\mathbf{f}_t, S_t = j) = \mathcal{N}(\theta_j + \mathbf{X}_j\mathbf{f}_t, \mathbf{\Psi}_j) \quad (5.2)$$

La vraisemblance d'une observation $\mathbf{y}_t$ sachant l'état actuel $S_t = j$ peut être obtenue en intégrant par rapport au vecteur d'état $\mathbf{f}_t$ le produit des deux gaussiennes [5.1 - 5.2]. La vraisemblance résultante est aussi gaussienne et peut être écrite sous la forme :

$$b_j(\mathbf{y}_t) = p(\mathbf{y}_t/S_t = j) = \mathcal{N}(\mathbf{y}_t \,/\, \theta_j, \mathbf{\Sigma}_j) \quad (5.3)$$

où

$$\mathbf{\Sigma}_j = \mathbf{X}_j\mathbf{H}_j\mathbf{X}'_j + \mathbf{\Psi}_j \quad (5.4)$$

Le calcul de la fonction de vraisemblance nécessite l'inversion de matrices de dimension $(q \times q)$ données par (5.4). Si le problème de capacité de mémoire ne se pose pas, nous pouvons calculer tous les déterminants et les inverses qui leurs correspondent pour tous les états du système avant d'entamer les étapes de mise à jour. Cependant, pour les systèmes de grande dimension un tel calcul devient très lourd. Pour éviter ce problème, il faut calculer tous les déterminants et les inverses à chaque instant t. L'utilisation de l'égalité de Woodbury peut aussi simplifier ce calcul.

$$\left[\mathbf{X}_j\mathbf{H}_j\mathbf{X}_j' + \mathbf{\Psi}_j\right]^{-1} = \mathbf{\Psi}_j^{-1} - \mathbf{\Psi}_j^{-1}\mathbf{X}_j\left[\mathbf{X}_j'\mathbf{\Psi}_j^{-1}\mathbf{X}_j + \mathbf{H}_j^{-1}\right]^{-1}\mathbf{X}_j'\mathbf{\Psi}_j^{-1} \tag{5.5}$$

où les inverses des matrices $\mathbf{\Psi}_j$ et $\mathbf{H}_j$ sont faciles à calculer étant donné qu'elles sont diagonales. La matrice complète $\left[\mathbf{X}_j'\mathbf{\Psi}_j^{-1}\mathbf{X}_j + \mathbf{H}_j^{-1}\right]$ nécessite l'inversion d'une matrice de dimension $(k \times k)$ seulement. Celle-ci est, donc, plus facile et plus rapide à inverser qu'une matrice de dimension $(q \times q)$ si $k \lll q$. Les déterminants nécessaires pour les calculs de la vraisemblance sont donnés par la formule :

$$|\mathbf{X}_j\mathbf{H}_j\mathbf{X}_j' + \mathbf{\Psi}_j| = |\mathbf{\Psi}_j||\mathbf{H}_j||\mathbf{X}_j'\mathbf{\Psi}_j^{-1}\mathbf{X}_j + \mathbf{H}_j^{-1}| \tag{5.6}$$

où les déterminants des matrices de covariance diagonales sont faciles à calculer. Le déterminant d'une matrice de dimension $(k \times k)$ est souvent obtenu comme sous-produit de son inverse en utilisant, par exemple, une décomposition de type Cholesky.

De la même façon, comme dans le cas des HMM, l'algorithme de Viterbi peut aussi être utilisé pour l'identification de la séquence d'états optimale. On utilisera cette technique plus tard dans les applications décrites dans la section 5.2.4. Toute implémentation de l'algorithme de Viterbi, tel que l'algorithme "token passing",[3] peut aussi être adaptée et appliquée dans le cas des FAHMM.

5.2.3 Optimisation des Paramètres d'un FAHMM

L'estimation des paramètres de ce modèle peut être menée en utilisant un algorithme d'apprentissage discriminant telle que l'erreur de classification minimum (voir L. Saul et M. Rahim [1999] pour une application en reconnaissance automatique de parole), mais ici on se contentera d'une approche de maximum de vraisemblance. Comme dans le cas des modèles HMM (chapitre 4) et les modèles à facteurs (chapitres 2 et 3), nous allons développer une approche itérative de maximum de vraisemblance basée sur le principe de l'algorithme EM. La vraisemblance complétée d'une séquence d'observations, $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$, d'une séquence de vecteurs d'états continus, $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$, et d'une séquence d'états HMM, $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ est donnée par :

$$p(\mathcal{Y}, \mathcal{F}, \mathcal{S}/\Theta) = p(S_1)\prod_{t=2}^{n} p(S_t/S_{t-1})\prod_{t=1}^{n} p(\mathbf{f}_t/S_t;\Theta)p(\mathbf{y}_t/\mathbf{f}_t, S_t;\Theta) \tag{5.7}$$

où $p(S_1) = \pi_{s_1}$ est la probabilité de l'état initial, $p(S_t/S_{t-1}) = p_{s_{t-1}s_t}$ sont les probabilités de transition et $\Theta = \{\pi, p_{ij}, \theta_j, \mathbf{X}_j, \mathbf{H}_j, \mathbf{\Psi}_j\}$.

[3] Voir Viterbi [1967] et Young, Russell, et Thornton [1989] pour une application dans le domaine de reconnaissance de parole.

L'Algorithme EM

Cet algorithme consiste en deux étapes, une étape E (Espérance) et une étape M (Maximisation). À la différence des algorithmes développés dans les chapitres précédents, les données manquantes dans ce cas sont de deux types : les états continus (les facteurs communs) et les états discrets (les états HMM).

Étape E :

Dans cette première étape, l'espérance conditionnelle de la log-vraisemblance des données complétées sera calculée, soit

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \mathbb{E}\left[\log p\left(\mathcal{Y}, \mathcal{F}, \mathcal{S}/\Theta^{(i)}\right)/\mathcal{Y}, \Theta\right] \tag{5.8}$$

L'ensemble actuel de tous les paramètres du modèle est désigné par $\Theta^{(i)}$. L'équation (5.8) montre que cette étape nécessite le calcul de certaines statistiques exhaustives que ce soit pour les vecteurs d'états continus ou bien pour les états de la chaîne de Markov. À chaque itération ces statistiques seront évaluées en utilisant les paramètres de l'itération précédente.

Étape M :

Au niveau de l'étape maximisation, un ensemble de paramètres, $\widehat{\Theta}$, maximisant la fonction auxiliaire $\mathcal{Q}$ sera calculé, soit

$$\widehat{\Theta} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{(i)})$$

Ces paramètres seront par la suite utilisés comme l'ensemble des anciens paramètres au niveau de l'itération $(i+1)$, $\widehat{\Theta} \longrightarrow \Theta^{(i+1)}$. Ces deux étapes sont répétées jusqu'à ce que la différence entre la fonction de vraisemblance de l'itération $(i+1)$ et celle de l'itération (i) ne change pratiquement plus.

Dans toute la suite, on va calculer les statistiques a posteriori nécessaire pour l'implémentation de l'étape E. Les équations de mise à jour des paramètres résultants de l'étape M seront présentées dans l'annexe.

Les Statistiques a Posteriori

Étant données les distributions conditionnelles (5.1) et (5.2), la vraisemblance marginale de $\mathbf{y}_t$, sachant l'état actuel $S_t = j$, sera donnée par :

$$\begin{aligned} b_j(\mathbf{y}_t) = p(\mathbf{y}_t/S_t = j) &= \int \mathcal{N}(\theta_j + \mathbf{X}_j\mathbf{f}_t, \mathbf{\Psi}_j)\mathcal{N}(\mathbf{0}, \mathbf{H}_j)\, d\mathbf{f}_t \\ &= \mathcal{N}\left[\theta_j, \mathbf{X}_j\mathbf{H}_j\mathbf{X}_j' + \mathbf{\Psi}_j\right] \end{aligned}$$

L'algorithme Avant-Arrière : La probabilité jointe d'une séquence d'observations $\mathcal{Y}_{1:t}$ et de l'état actuel $S_t = j$ est représentée par la variable "Avant" $\alpha_j(t) = p(S_t = j, \mathcal{Y}_{1:t})$. Si on suppose en plus que la première observation est générée par le premier état discret, la variable "Avant" sera donc initialisée par :

$$\begin{cases} b_1(\mathbf{y}_1) & , \quad j = 1 \\ 0 & , \quad j \neq 1 \end{cases}$$

En se basant sur les propriétés de l'indépendance conditionnelle dans les modèles de chaînes de Markov cachées, nous pouvons développer la formule de récurrence suivante pour la variable "Avant" à l'instant t :

$$\begin{aligned} \alpha_j(t) &= p(S_t = j, \mathcal{Y}_{1:t}) = p(\mathbf{y}_t/S_t = j)p(S_t = j, \mathcal{Y}_{1:t-1}) \\ &= p(\mathbf{y}_t/S_t = j)\sum_{i=1}^{m} p(S_t = j, S_{t-1} = i, \mathcal{Y}_{1:t-1}) \\ &= p(\mathbf{y}_t/S_t = j)\sum_{i=1}^{m} p(S_t = j/S_{t-1} = i)p(S_{t-1} = i, \mathcal{Y}_{1:t-1}) \\ &= b_j(\mathbf{y}_t)\sum_{i=1}^{m} p_{ij}\alpha_i(t-1) \end{aligned} \tag{5.9}$$

La probabilité d'une séquence d'observations allant de $t+1$ jusqu'à n conditionnellement à l'état actuel $S_t = j$ est représentée par la variable "Arrière", $\beta_i(t) = p(\mathcal{Y}_{t+1:n}/S_t = i)$. Cette variable sera initialisée par $\beta_i(n) = 1, \ \forall \ i \in [1, m]$. Les propriétés de l'indépendance conditionnelle dans les modèles de chaînes de Markov cachées impliquent dans ce cas, aussi, une formule de récurrence qui exprime la variable "Arrière" à la date $t-1$ en fonction de toutes ses valeurs futures, soit

$$\begin{aligned} \beta_i(t-1) &= p(\mathcal{Y}_{t:n}/S_{t-1} = i) = \sum_{j=1}^{m} p(S_t = j, \mathcal{Y}_{t:n}/S_{t-1} = i) \\ &= \sum_{j=1}^{m} p(S_t = j/S_{t-1} = i)p(\mathbf{y}_t/S_t = j)p(\mathcal{Y}_{t+1:n}/S_t = j) \\ &= \sum_{j=1}^{m} p_{ij}b_j(\mathbf{y}_t)\beta_j(t) \end{aligned} \tag{5.10}$$

Maintenant, nous pouvons exprimer la vraisemblance de la séquence complète d'observations, $\mathcal{Y}$, en fonction des variables "Arrière" et "Avant", soit

$$p(\mathcal{Y}) = \sum_{i=1}^{m} p(S_t = i, \mathcal{Y}_{1:t})p(\mathcal{Y}_{t+1:n}/S_t = i) = \sum_{i=1}^{m} \alpha_i(t)\beta_i(t) \tag{5.11}$$

Probabilités a Posteriori des États Discrets : L'étape E de l'algorithme EM nécessite le calcul des probabilités a posteriori des états markoviens $S_t = j$ pour $j = 1, ..., m$. Ces probabilités peuvent être exprimées en fonction des variables "Arrière-Avant" de la manière suivante :

$$\begin{aligned}
\gamma_j(t) = p(S_t = j/\mathcal{Y}) &= \frac{p(S_t = j, \mathcal{Y})}{p(\mathcal{Y})} \\
&= \frac{p(S_t = j, \mathcal{Y}_{1:t})p(\mathcal{Y}_{t+1:n}/S_t = j)}{p(\mathcal{Y})} \\
&= \frac{\alpha_j(t)\beta_j(t)}{\sum_{i=1}^{m} \alpha_i(t)\beta_i(t)}
\end{aligned}$$

Les probabilités jointes de l'état actuel $S_t = j$ et l'état $S_{t-1} = i$ sachant la séquence complète des observations est nécessaire, aussi, pour implémenter les formules de mise à jour des probabilités de transition. Ces probabilités peuvent être, aussi, exprimées en fonction des variables "Arrière-Avant", soit

$$\begin{aligned}
\xi_{ij}(t) &= p(S_{t-1} = i, S_t = j/\mathcal{Y}) \\
&= \frac{p(S_{t-1} = i, \mathcal{Y}_{1:t-1})p(S_t = j/S_{t-1} = i)p(\mathbf{y}_t/S_t = j)p(\mathcal{Y}_{t+1:n}/S_t = j)}{p(\mathcal{Y})} \\
&= \frac{\alpha_i(t-1)p_{ij}b_j(\mathbf{y}_t)\beta_j(t)}{\sum_{i=1}^{m} \alpha_i(t)\beta_i(t)}
\end{aligned} \tag{5.12}$$

Statistiques a Posteriori des États Continus : Étant donné l'état actuel $S_t = j$, la distribution jointe du vecteur des observations et du vecteur des états continus à l'instant t est gaussienne. En se basant sur les propriétés de la loi normale multivariée et les résultats précédents, nous pouvons écrire :

$$p(\mathbf{y}_t, \mathbf{f}_t/S_t = j) = \mathcal{N}\left[\begin{pmatrix} \theta_j \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}_j\mathbf{H}_j\mathbf{X}'_j + \mathbf{\Psi}_j & \mathbf{X}_j\mathbf{H}_j \\ \mathbf{H}_j\mathbf{X}'_j & \mathbf{H}_j \end{pmatrix}\right] \tag{5.13}$$

Dans ce cas on démontre que :

$$p(\mathbf{f}_t/\mathbf{y}_t, S_t = j) = \mathcal{N}\left[\mathbf{K}_j(\mathbf{y}_t - \theta_j), \mathbf{H}_j - \mathbf{K}_j\mathbf{X}_j\mathbf{H}_j\right] \tag{5.14}$$

où $\mathbf{K}_j = \mathbf{H}_j\mathbf{X}'_j\left[\mathbf{X}_j\mathbf{H}_j\mathbf{X}'_j + \mathbf{\Psi}_j\right]^{-1}$. Les statistiques nécessaires pour l'implémentation de l'étape E et la mise à jour des paramètres sont les suivantes :

$$\begin{aligned}
\widetilde{\mathbf{f}}_{jt} &= \mathbf{K}_j\left[\mathbf{y}_t - \theta_j\right] \\
\widetilde{\mathbf{R}}_{jt} &= \mathbf{H}_j - \mathbf{K}_j\mathbf{X}_j\mathbf{H}_j + \widetilde{\mathbf{f}}_{jt}\widetilde{\mathbf{f}}'_{jt}
\end{aligned}$$

5.2.4 Identification des États Cachés

Dans les applications des modèles HMM, la variable d'état cachée a toujours une signification relative au phénomène étudié (par exemple en reconnaissance de parole, les états cachés sont reliés aux différentes parties du mot prononcé, voir Jelinek et al., [1975]). Dans notre cas, l'état caché indique la transition d'un régime à un autre qui peut être due à un événement bien particulier. Pour des raisons d'interprétation du modèle, étant donnée une séquence d'observations $\mathcal{Y}$, il est donc utile d'identifier la séquence optimale d'états $\mathcal{S}$ qui lui correspond. Pour l'identification de cette séquence, plusieurs critères d'optimalité existent. L'algorithme de probabilité a posteriori maximale MAP, par exemple, permet de résoudre ce problème de restauration. Cette méthode estime la séquence $\mathcal{S}$ par les états qui maximisent la probabilité a posteriori $p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n};\widehat{\Theta})$, où $\widehat{\Theta}$ est l'estimation de maximum de vraisemblance de Θ.

Estimation des États par les Probabilités de Lissage

Une méthode de restauration du processus caché alternative au MAP consiste également à restaurer les états cachés à partir de leur valeur la plus probable, mais sur la base d'un critère local, c'est-à-dire en déterminant individuellement chaque état le plus probable, soit

$$\widehat{S}_{t/n} = \arg\max_j p(S_t = j/\mathcal{Y};\Theta)\,, \quad 1 \leq t \leq n$$

Ces probabilités (dites probabilités de lissage) sont obtenues par l'algorithme Avant-Arrière de la section 5.2.3.

$$p(S_t = j/\mathcal{Y};\Theta) = \gamma_j(t)$$

Une discussion sur ces méthodes dans le cadre des chaînes de Markov cachées est disponible dans Ephraim et Mehrav [2002] et dans Fredkin et Rice [1992].

Estimation des États par les Probabilités de Filtrage

Dans le cas où on considère seulement l'information passée et présente de la séquence observée $\mathcal{Y}_{1:t}$, on doit maximiser les probabilités de filtrage pour estimer la séquence optimale d'états cachés, soit

$$\widehat{S}_{t/t} = \arg\max_j p(S_t = j/\mathcal{Y}_{1:t};\Theta)\,, \quad 1 \leq t \leq n$$

Dans ce cas on maximise les probabilités a posteriori

$$\begin{aligned} p(S_t = j/\mathcal{Y}_{1:t}) &= \frac{p(\mathbf{y}_t/S_t = j, \mathcal{Y}_{1:t-1})p(S_t = j/\mathcal{Y}_{1:t-1})}{p(\mathcal{Y}_{1:t}/\mathcal{Y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t/S_t = j)p(S_t = j, \mathcal{Y}_{1:t-1})}{p(\mathcal{Y}_{1:t})} \\ &= \frac{\alpha_j(t)}{\sum\limits_{i=1}^{m} \alpha_i(t)} \end{aligned}$$

Estimation des États par les Probabilités de Prédiction

Dans les applications financières, du point de vue des investisseurs, la méthode la plus intéressante pour estimer les états est celle qui permet de prévoir l'état de la période suivante S_{t+1} en se basant sur l'information disponible à la date t, soit $\mathcal{Y}_{1:t}$. Une telle méthode est basée sur la maximisation des probabilités de prédiction :

$$\widehat{S}_{t+1/t} = \arg\max_j p(S_{t+1} = j/\mathcal{Y}_{1:t};\Theta)\,, \quad 1 \leq t \leq n-1$$

Les probabilités a posteriori maximisées par $\widehat{S}_{t+1/t}$ sont données par :

$$\begin{aligned}
p(S_{t+1} = j/\mathcal{Y}_{1:t}) &= \sum_{i=1}^{m} P(S_{t+1} = j, S_t = i/\mathcal{Y}_{1:t}) \\
&= \frac{\sum\limits_{i=1}^{m} p(\mathcal{Y}_{1:t}/S_{t+1} = j, S_t = i)p(S_{t+1} = j, S_t = i)}{p(\mathcal{Y}_{1:t})} \\
&= \frac{\sum\limits_{i=1}^{m} p(\mathcal{Y}_{1:t}, S_t = i)p_{ij}}{p(\mathcal{Y}_{1:t})} \\
&= \frac{\sum\limits_{i=1}^{m} \alpha_i(t)p_{ij}}{\sum\limits_{i=1}^{m} \alpha_i(t)}
\end{aligned}$$

Nous pouvons, aussi, calculer les probabilités de prédiction pour un horizon de h périodes, par exemple la prédiction de l'état pour deux périodes est donnée par :

$$\widehat{S}_{t+2/t} = \arg\max_j p(S_{t+2} = j/\mathcal{Y}_{1:t};\Theta)\,, \quad 1 \leq t \leq n-2$$

et qui sera obtenue en maximisant à travers $1 \leq j \leq m$ les probabilités suivantes :

$$p(S_{t+2} = j/\mathcal{Y}_{1:t};\Theta) = \frac{\sum\limits_{r=1}^{m}\sum\limits_{i=1}^{m} \alpha_r(t)p_{ri}p_{ij}}{\sum\limits_{h=1}^{m} \alpha_h(t)}$$

Nous remarquons ici que les différents critères que nous avons défini maximisent seulement le nombre d'états individuels corrects. Ces méthodes peuvent donc aboutir à des erreurs dans certains cas particuliers. Par exemple lorsque le modèle de Markov caché possède des probabilités de transition égales à zéro, la séquence optimale obtenue pourrait en fait ne pas être une séquence d'états possible puisque le critère considéré ne tient pas compte des probabilités des changements d'états. Une solution possible est de modifier le critère d'optimalité. On pourrait par exemple chercher la séquence d'états qui maximise les paires d'états (S_t, S_{t+1}) ou même les triplets d'états (S_t, S_{t+1}, S_{t+2}). Le critère le plus utilisé est celui qui cherche la meilleure séquence d'états globale (le meilleur chemin), c'est-à-dire qui maximise $p(\mathcal{S}, \mathcal{Y}/\Theta)$.

L'Algorithme de Viterbi

Si ces critères sont tous adaptés à certaines applications, le critère le plus utilisé est donc celui qui cherche la meilleure séquence d'états globale, ce qui revient à maximiser $p(\mathcal{S}, \mathcal{Y}/\Theta)$ ou bien $p(\mathcal{S}/\mathcal{Y}, \Theta)$. L'algorithme de Viterbi est une technique qui permet de calculer ce chemin optimal. Dans cette sous-section nous rappelons le principe de cet algorithme, qui est l'algorithme du MAP pour les chaînes de Markov cachées. L'algorithme de Viterbi est un algorithme de programmation dynamique, c'est-à-dire une méthode de résolution de problèmes d'optimisation qui repose sur une propriété de décomposabilité de la fonction à optimiser.

Notre objectif est donc de trouver :

$$\begin{aligned}
\widehat{\mathcal{S}}_{1:n} &= \arg\max_{\mathcal{S}_{1:n}} p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n}) \\
&= \arg\max_{\mathcal{S}_{1:n}} \frac{p(\mathcal{S}_{1:n}, \mathcal{Y}_{1:n})}{p(\mathcal{Y}_{1:n})} \\
&= \arg\max_{\mathcal{S}_{1:n}} p(\mathcal{S}_{1:n}, \mathcal{Y}_{1:n}) \qquad (5.15)
\end{aligned}$$

Pour ce faire, on va définir la variable suivante qui peut être calculée récursivement :

$$\begin{aligned}
\delta_t(S_t) &= \arg\max_{S_{1:t-1}} p(\mathcal{S}_{1:t}, \mathcal{Y}_{1:t}) \qquad (5.16) \\
\delta_j(t) &= \arg\max_{\mathcal{S}_{1:t-1}} p(\mathcal{S}_{1:t-1}, S_t = j, \mathcal{Y}_{1:t}) \qquad (5.17)
\end{aligned}$$

Notons aussi qu'on est en train de maximiser par rapport à la séquence allant jusqu'à la date $t-1$, $\mathcal{S}_{1:t-1}$, et que

$$\begin{aligned}
\delta_1 &= p(S_1, \mathbf{y}_1) = p(\mathbf{y}_1/S_1)p(S_1) \\
\delta_j(1) &= p(\mathbf{y}_1/S_1 = j)p(S_1 = j) \qquad \text{et} \qquad (5.18) \\
\max_{\mathcal{S}_{1:n}} p(\mathcal{S}_{1:n}, \mathcal{Y}_{1:n}) &= \max_{\mathcal{S}_{1:n}} \delta_{1:n} = \max_j \delta_j(n) \qquad (5.19)
\end{aligned}$$

Dans ce cas, la variable δ_t peut être exprimée sous la forme suivante :

$$\begin{aligned}
\delta_{t+1} &= \max_{\mathcal{S}_{1:t}} p(\mathcal{S}_{1:t+1}, \mathcal{Y}_{1:t+1}) \\
&= \max_{\mathcal{S}_{1:t}} \left[p(\mathbf{y}_{t+1}/S_{t+1})p(S_{t+1}/S_t)p(\mathcal{S}_{1:t}, \mathcal{Y}_{1:t})\right] \\
&= p(\mathbf{y}_{t+1}/S_{t+1})\max_{S_t} \left[p(S_{t+1}/S_t)\max_{\mathcal{S}_{1:t-1}} [p(\mathcal{S}_{1:t}, \mathcal{Y}_{1:t})]\right] \\
&= p(\mathbf{y}_{t+1}/S_{t+1})\max_{S_t} \left[p(S_{t+1}/S_t)\delta_t\right]
\end{aligned}$$

et donc

$$\delta_j(t+1) = p(\mathbf{y}_{t+1}/S_{t+1} = j)\max_i \left[p_{ij}\ \delta_i(t)\right] \qquad (5.20)$$

Ainsi, pour trouver le maximum de $p(\mathcal{S}_{1:n}, \mathcal{Y}_{1:n})$ nous initialisons l'algorithme avec (5.18). Par la suite on calcule $\delta_2, ..., \delta_n$ en utilisant (5.20). Finalement on calcule le maximum global par (5.19). À ce niveau, il faut noter que la valeur de δ_t diminue lorsque t augmente (on multiplie des probabilités). Afin d'éviter les problèmes d'ordre numérique, on doit normaliser δ_t à chaque itération, par exemple à la longueur unitaire.[4] Pour obtenir la séquence optimale, nous allons définir une variable permettant de stocker les valeurs de S_t qui maximisent la fonction récurrente $p(S_{t+1}/S_t)\delta_t(S_t)$ de l'équation (5.20) pour toutes les valeurs de S_{t+1}, soit

$$\begin{aligned} \digamma_{t+1}(S_{t+1}) &= \arg\max_{S_t} \Big[p(S_{t+1}/S_t)\delta_t(S_t)\Big] \\ \digamma_j(t+1) &= \arg\max_{i} \Big[p_{ij}\delta_i(t)\Big] \quad \text{pour} \quad t = 1, ..., n-1 \end{aligned} \tag{5.21}$$

La séquence d'états optimale est alors extraite par la procédure de recherche rétrograde (backtracking en anglais) suivante :

$$\widehat{S}_n = \arg\max_j \delta_j(n) \tag{5.22}$$

$$\widehat{S}_t = \digamma_{t+1}(\widehat{S}_{t+1}) \quad \text{pour} \quad t = 1, ..., n-1 \tag{5.23}$$

Ainsi la procédure de trouver la séquence d'états la plus probable commence par le calcul utilisant la récurrence (5.20) tandis qu'on garde toujours un pointeur sur "l'état gagnant" dans l'opération de recherche du maximum. Finalement l'état j_n^* sera trouvé par (5.19) et commençant par cet état, la séquence des états est poursuivie comme un pointeur dans chaque état indiqué. Cela donne l'ensemble des états recherchés.

L'algorithme global peut s'interprété comme une recherche dans un graphe dans les noeuds sont formés par les états du HMM à chaque instant t, $1 \leq t \leq n$.

Application des Différents Critères

Nous avons appliqué les algorithmes d'identification de la séquence optimale pour étudier leur aptitude à détecter les points de changement de régime en considérant des données simulées et une base de données réelles.

Simulations : Pour les simulations nous avons généré des données à partir d'un modèle à facteurs standards avec $k = 2$ facteurs communs, $q = 6$ séries d'observations, $n = 700$ observations et deux régimes markoviens. Les paramètres de cette simulation sont donnés dans le tableau 5.1. La date du changement de régime est $t^* = n/2 + 1$. Nous avons généré une centaine de réplications et sur chacune nous avons estimé le modèle pour appliquer par la suite les différents critères en comptabilisant à chaque fois la date du changement. Le tableau 5.2 donne tous les résultats. Ce tableau nous montre que seulement l'algorithme de Viterbi et l'algorithme de lissage sont capables de détecter les points de changement. Les autres algorithmes donnent dans la plupart

[4] Ce qui nous intéresse ici c'est la séquence qui maximise la probabilité globale et non pas la maximisation de la probabilité elle même. Dans ce cas, la normalisation de δ_t n'affecte que la dernière.

TAB. 5.1 – Paramètres et résultats de la simulation

Les paramètres de simulation
. Les vrais paramètres,
(.) Les valeurs d'initialisation de l'algorithme EM

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$
État 1	2.0000 (0.0000)	4.0000 (1.0000)	6.0000 (2.0000)	5.0000 (2.0000)
	2.0000 (1.0000)	3.0000 (1.0000)	5.0000 (2.0000)	4.0000 (2.0000)
	1.0000 (0.0000)	5.0000 (2.5000)	3.0000 (1.0000)	5.0000 (2.5000)
	1.0000 (0.0000)	5.0000 (2.0000)	3.0000 (1.0000)	6.0000 (3.0000)
	2.0000 (0.0000)	3.0000 (1.0000)	2.0000 (1.0000)	7.0000 (3.0000)
	2.0000 (1.0000)	4.0000 (1.0000)	4.0000 (2.0000)	9.0000 (3.5000)
État 2	1.0000 (0.0000)	2.0000 (0.5000)	1.0000 (0.0000)	1.0000 (0.5000)
	1.0000 (0.0000)	3.0000 (1.0000)	1.0000 (0.0000)	2.0000 (0.5000)
	2.0000 (1.0000)	2.0000 (0.5000)	2.0000 (0.5000)	1.0000 (0.5000)
	2.0000 (1.0000)	3.0000 (1.0000)	2.0000 (0.5000)	2.0000 (0.5000)
	1.0000 (0.0000)	2.0000 (0.5000)	3.0000 (1.0000)	2.0000 (0.5000)
	2.0000 (0.0000)	3.0000 (0.5000)	3.0000 (1.0000)	1.0000 (0.5000)

Résultats de l'estimation
. Moyenne, (.) écart-types

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$
État 1	2.0545 (0.3711)	3.9214 (0.2321)	6.1124 (0.2757)	4.9534 (1.2795)
	2.0541 (0.3125)	2.9478 (0.2044)	5.1057 (0.2206)	3.9145 (0.9521)
	1.0573 (0.3084)	4.9456 (0.2677)	3.0994 (0.2043)	5.0539 (0.8413)
	1.0378 (0.3017)	4.9841 (0.2096)	3.1620 (0.2484)	5.9720 (0.8274)
	2.0470 (0.2342)	2.9875 (0.1812)	2.1304 (0.1582)	6.9239 (0.5578)
	2.0386 (0.2776)	3.9354 (0.2182)	4.1473 (0.2311)	9.0187 (0.7569)
État 2	0.9913 (0.1198)	2.0014 (0.0891)	0.9734 (0.0819)	0.9874 (0.0872)
	0.9844 (0.1621)	2.9828 (0.1311)	0.9763 (0.1288)	1.9984 (0.2817)
	1.9997 (0.1375)	2.0166 (0.0877)	1.9884 (0.0796)	0.9981 (0.0897)
	2.0047 (0.1613)	3.0331 (0.1064)	1.9816 (0.1104)	1.9964 (0.1766)
	1.0079 (0.1791)	2.0545 (0.1283)	2.9694 (0.1168)	1.9898 (0.2724)
	1.9998 (0.1907)	3.0403 (0.1205)	2.9891 (0.1115)	1.0169 (0.1431)

des cas des résultats décalés d'une ou de deux périodes. Nous remarquons, aussi, que les algorithmes de filtrage et de prédiction pour un horizon d'une et de deux périodes donnent exactement les mêmes résultats, et c'est pour cette raison que nous avons donné les résultats de l'algorithme de filtrage seulement dans le tableau 5.2.

Données Financières : Dans cette deuxième application nous avons considéré les rendements journaliers des cours en valeurs (évalués par rapport à la livre sterling) du Dollar Américain (USD), le Dollar Canadien (CAD), le Franc Français (FRF), le Franc Suisse, la Lire Italienne (ITL), le Deutsche Mark (DEM), le Yen Japonais (JPY) et le Dollar de Hong Kong (HKD)[5]. Les données s'étalent sur la période 10/10/1990 à 26/11/1993 inclue (soit 800 observations couvrant la période de la crise financière qui a frappé les marchés de change dans les pays membres du système monétaire européen

[5] PACIFIC EXCHANGE RATE SERVICE, Sauder School of Business, http ://fx.sauder.ubc.ca/.

Tab. 5.2 – Identification de la séquence optimale

Méthode	$t-4$	$t-3$	$t-2$	$t-1$	t^*	$t+1$	$t+2$
Al. de Viterbi	00	02	01	07	89	01	00
Al. de Lissage	00	01	03	07	88	01	00
Al. de Filtrage	00	00	00	00	02	50	38

SME et qui s'est déclenché vers la fin du mois de septembre 1992 lorsque la Livre Sterling et la Lire Italienne ont quitté le SME). Pour le calcul des rendements, nous avons utilisé la formule des rendements standardisés par rapport à la moyenne et l'écart-type de chaque séries afin de neutraliser l'effet d'hétéroscédasticité dynamique éventuelle qui caractérise d'une manière générale les séries financières.

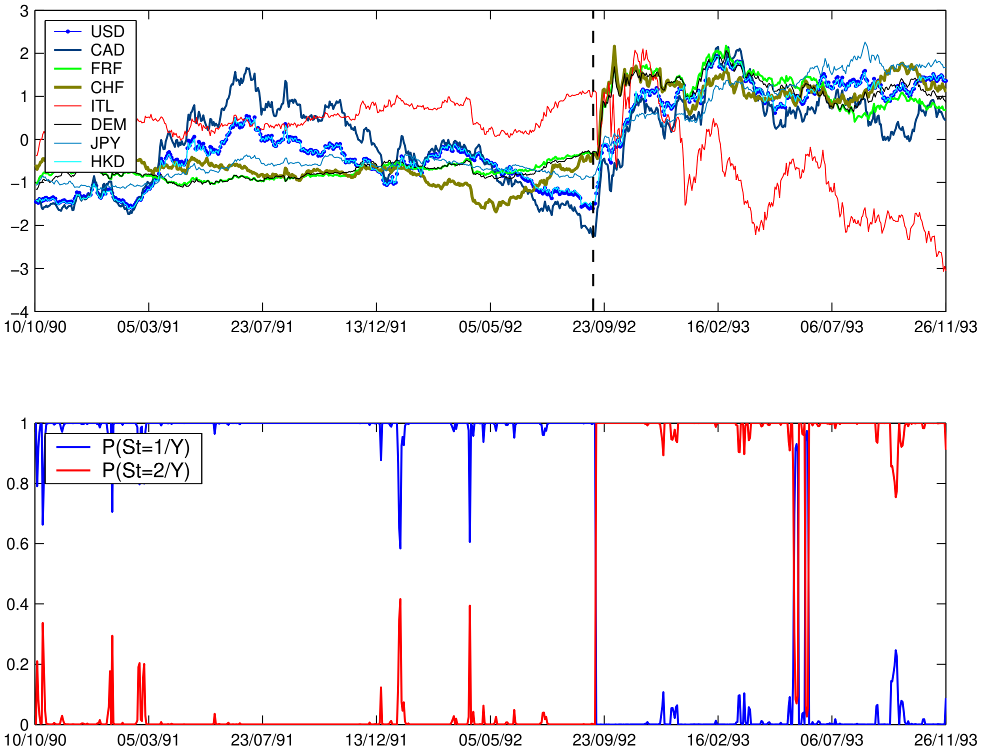

Fig. 5.2 – *Graphique 1 :* Prix spot en valeurs des différentes devises par rapport à la Livre Sterling. La ligne verticale représente la date du changement de régime (déclenchement de la crise financière dans les marchés de change des pays membres du SME). *Graphique 2 :* Les probabilités a posteriori $\gamma_j(t)$ des états cachés estimées par un modèle FAHMM à deux états markoviens et deux facteurs communs.

Sur cette base de données nous avons estimé des modèles FAHMM à deux états markoviens avec 1, 2 et 3 facteurs communs. Dans ce cas, les critères de sélection AIC et BIC ont favorisé la deuxième spécification. Les résultats d'estimation de ce modèle sont donnés dans le tableau 5.3. La représentation graphique des probabilités de lissage, $\gamma_j(t)$, montre que le modèle est capable de détecter le point de changement et que ces probabilités donnent la même séquence optimale identifiée par un algorithme de Viterbi. Nous remarquons aussi que les algorithmes de filtrage et de prédiction donnent le même résultat (figure 5.3).

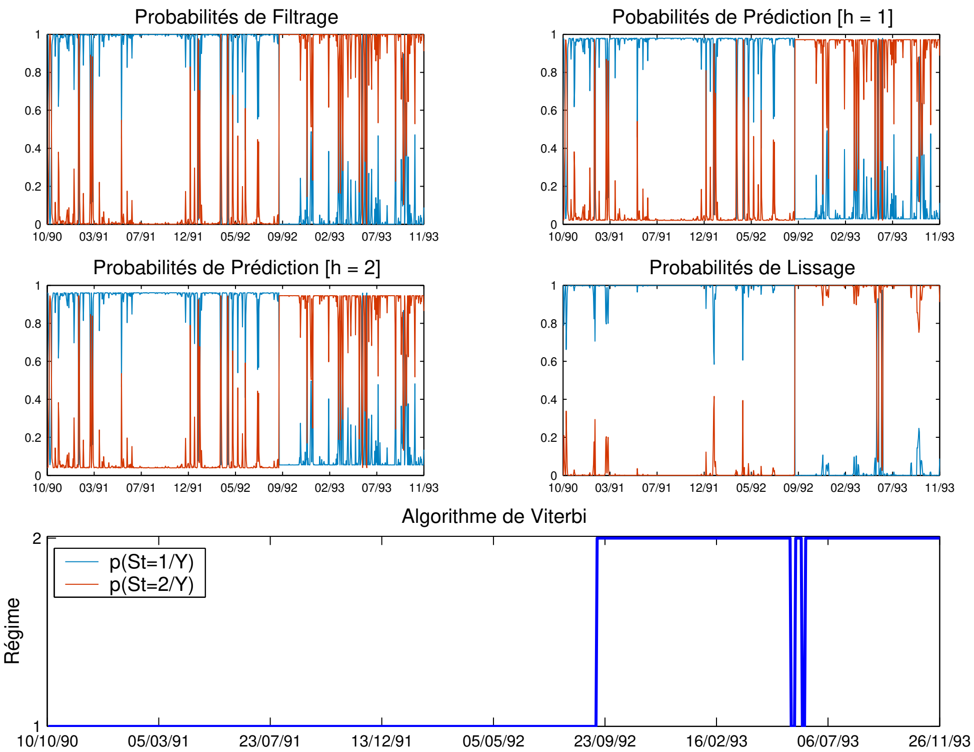

FIG. 5.3 – Identification de la séquence d'états optimale.

5.3 Modèles Conditionnellement Hétéroscédastiques

Le modèle qu'on se propose d'étudier maintenant est construit par :

- une structure Markovienne cachée pour les paramètres du modèle permettant de tenir compte des différents états de la nature qui peuvent affecter la dynamique des séries étudiées. Dans ce cas, les propriétés des différentes séries à un instant quelconque t, dépendent du régime prévalant. Un régime bien particulier est la réalisation d'une chaîne de Markov homogène à état fini.
- un modèle à facteurs linéaire pour les rendements en excès. Les paramètres de cette spécification sont supposés constants à l'intérieur de chaque régime.
- des processus GQARCH univariés pour la modélisation de la volatilité des facteurs communs, fondée sur l'idée que celle-ci est hautement persistante (le phénomène dit de "volatility clustering").

5.3.1 Le Modèle de base

Soit $\mathbf{y}_t$ le vecteur des rendements en excès des différents actifs (de dimension $q \times 1$) et $\mathbf{f}_t$ le vecteur des facteurs communs non observables de dimension $k \times 1$. Notre modèle

TAB. 5.3 – Modèle FAHMM à deux facteurs

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$
État 1	-0.0088	1.9529	1.6736	0.0010
	0.0030	1.7949	1.5554	0.0681
	-0.0163	-0.3482	0.3236	0.0148
	-0.0208	-0.2977	0.1713	0.3398
	0.0390	-0.2140	0.2772	0.0400
	-0.0189	-0.5149	0.2078	0.0238
	-0.0505	0.9898	0.8442	0.3715
	-0.0084	1.9332	1.6637	0.0058
État 2	0.0892	0.5019	0.8806	0.0028
	0.0663	0.4536	0.8709	0.1736
	0.0478	1.3327	0.0719	0.3078
	0.0482	1.1989	-0.0595	0.3630
	-0.0431	0.8588	0.1367	1.4879
	0.0471	1.4229	-0.0327	0.0157
	0.1268	0.6281	0.5831	0.7676
	0.0884	0.5012	0.8859	0.0078

à facteurs avec changement de régime suppose que le rendement en excès d'un actif quelconque pourra être exprimé comme étant la somme de son rendement anticipé, de k chocks systématiques et d'un chock idiosyncratique. La forme matricielle de cette nouvelle spécification Markovienne dynamique est donnée par :

$$\boxed{\begin{array}{c} S_t \sim P(S_t = j/S_{t-1} = i) \\[1ex] t = 1, ..., n \quad \text{et} \quad i, j = 1, ..., m \\[1ex] \mathbf{f}_{s_t} = \mathbf{H}_{s_t}^{1/2}\mathbf{f}_t^* \qquad \text{où} \qquad \mathbf{f}_t^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k) \\[1ex] \mathbf{y}_t = \mathbf{X}_{s_t}\mathbf{f}_{s_t} + \varepsilon_{s_t} \quad \text{avec} \quad \varepsilon_{s_t} \sim \mathcal{N}(\theta_{s_t}, \mathbf{\Psi}_{s_t}) \end{array}}$$

Les mêmes notations sont utilisées que dans la section 5.2.1. $\mathbf{y}_t$ est toujours un vecteur aléatoire de dimension $(q \times 1)$, c'est le vecteur des variables observables. Contrairement au cas standard, les variances des facteurs communs (les éléments diagonaux de $\mathbf{H}_{jt}$) sont maintenant supposées variables à travers le temps et leurs paramètres changent avec le régime. Nous supposons, en particulier, que les facteurs communs sont des processus GQARCH(1,1) avec changement de régime. Le *l-ème* élément de la diagonale de la matrice $\mathbf{H}_{jt}$ sous un régime bien particulier $S_t = j$ étant donné que $S_{t-1} = i$, sera donné par :

$$h_{lt}^{(j)} = w_j^l + \gamma_j^l f_{lt-1}^{(i)} + \alpha_j^l f_{lt-1}^{(i)\,2} + \delta_j^l h_{lt-1}^{(i)} \quad \text{pour} \quad l = 1, ..., k$$

La variance conditionnelle $h_{lt}^{(j)}$ sera positive lorsque w_j^l, α_j^l, $\delta_j^l > 0$ et ${\gamma_j^l}^2 \leq 4\alpha_j^l w_j^l$ pour tout $j = 1, ..., m$ et $l = 1, ..., k$. Ce processus sera stationnaire au niveau de la covariance lorsque $\alpha_j^l + \delta_j^l < 1$, $\forall\, j, l$. Pour garantir l'identification du modèle, nous supposons toujours que $q \geq k$ et $rang(\mathbf{X}_j) = k$, $\forall\, j$. Nous supposons aussi que les facteurs communs et les facteurs spécifiques ne sont pas corrélés, et que les $\mathbf{f}_t$ et $\varepsilon_{t'}$ sont mutuellement indépendants pour tout t, t'.

Notre modèle est assez général dans le sens où il permet de tenir compte de tous les changements structurels, c'est-à-dire les changements dans les relations entre les variables étudiées, sans imposer aucune restriction supplémentaire sur la nature de ces changements ou sur leur date d'occurrence. Il nous permet de tenir compte, simultanément, du comportement dynamique usuel de la volatilité commune due à certaines forces économiques communes, aussi bien que de la variation discrète brusque au niveau de la volatilité commune et spécifique qui peut être due à certains événements anormaux liés, par exemple, aux changements de la conjoncture ou bien aux cycles économiques. Dans une perspective d'analyse et de prévision des rendements financiers, cette nouvelle spécification nous permettra de mieux caractériser la dynamique des prix et de résoudre les problèmes liés aux changements de la structure interne des données financières. Il s'agit de problèmes de type :

- Peut-on distinguer différents régimes caractérisant les rendements financiers ?
- Comment les régimes se diffèrent-ils ?
- Quelle est la fréquence de ces changements de régime et quelles sont leurs dates d'occurrence ?
- Est-ce que le degré des co-mouvements a augmenté ou bien diminué ?
- Les fluctuations communes et spécifiques sont beaucoup ou moins volatiles ?
- Les changements de régime sont-ils prédictibles ?

5.3.2 Représentation Espace-état Multi-Régime

Le modèle à facteurs conditionnellement hétéroscédastiques que nous venons de définir ci-dessus peut être considéré comme un processus stochastique mltidimensionnel (ou comme un champ aléatoire) avec les indices $i = 1, ..., q$, $t = 1, ..., n$ et $j = 1, ...m$. Ainsi, nous pouvons l'exprimer par une représentation espace-état en séries temporelles à plusieurs états. Dans cette représentation, nous considérons les facteurs communs comme une variable d'état continue. Les équations de mesure et de transition sont, donc, données par :

$$\begin{array}{ll} \text{[Équation de Mesure]} & \mathbf{y}_t = \theta_{s_t} + \mathbf{X}_{s_t}\mathbf{f}_{s_t} + \varepsilon_{s_t} \\ \text{[Équation de Transition]} & \mathbf{f}_{s_t} = \mathbf{0}.\mathbf{f}_{s_{t-1}} + \mathbf{f}_{s_t} \end{array}$$

Pour la dérivation des équations de filtrage et de lissage nous allons utiliser la méthode pseudo bayésienne généralisée d'ordre un (GPB(1)), basée sur la technique de "Moment Matching". Ces statistiques seront, par la suite introduites dans un algorithme EM conditionnel afin d'estimer tous les paramètres du modèle. Pour l'implémentation de ces algorithmes, nous allons utiliser les mêmes notations introduites dans le chapitre 4, à savoir :

$$\begin{aligned}
\mathbf{f}_{t/\tau}^{i(j)} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_{t-1}=i, S_t=j\right] \\
\mathbf{f}_{t/\tau}^{(j)k} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t=j, S_{t+1}=k\right] \\
\mathbf{f}_{t/\tau}^{j} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t=j\right]
\end{aligned}$$

et

$$\begin{aligned}
h_{lt/\tau}^{j} &= Var\left(f_{lt}/\mathcal{Y}_{1:\tau}, S_t=j\right) \\
h_{lt/t-1}^{i(j)} &= Var\left(f_{lt}/\mathcal{Y}_{1:t-1}, S_{t-1}=i, S_t=j\right) \\
M_{t-1,t/\tau}(i,j) &= p\left(S_{t-1}=i, S_t=j/\mathcal{Y}_{1:\tau}\right) \\
M_{t/\tau}(j) &= p(S_t=j/\mathcal{Y}_{1:\tau}) \\
L_t(i,j) &= p(\mathbf{y}_t/\mathcal{Y}_{1:t-1}, S_{t-1}=i, S_t=j)
\end{aligned}$$

où $L_t(i,j)$ est la vraisemblance de l'innovation à l'instant t, lorsque le système est dans le régime j.

L'Algorithme de Filtrage

Nous effectuons les opérations suivantes successivement :

$$\begin{aligned}
\mathbf{f}_{t/t-1}^{i(j)} &= \mathbf{0}.\mathbf{f}_{t-1/t-1}^{i} = \mathbf{0} \qquad \forall \quad i,j=1,...,m \quad \text{et} && (5.24)\\
h_{lt/t-1}^{i(j)} &= w_{lj} + \gamma_{lj} f_{lt-1/t-1}^{i} + \alpha_{lj}\left[f_{lt-1/t-1}^{i\,2} + h_{lt-1/t-1}^{i}\right] + \delta_{lj} h_{lt-1/t-2}^{i} && (5.25)\\
\mathbf{H}_{t/t-1}^{i(j)} &= diag\left[h_{lt/t-1}^{i(j)}\right] \quad \text{avec} \quad l=1,2,...,k
\end{aligned}$$

Nous calculons par la suite l'erreur de prédiction, la variance de l'erreur, la matrice de gain de Kalman, et la vraisemblance de cette observation

$$\mathbf{e}_t(i,j) = \mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}_{t/t-1}^{i(j)}$$

$$\mathbf{\Sigma}_{t/t-1}^{i(j)} = \mathbf{X}_j \mathbf{H}_{t/t-1}^{i(j)} \mathbf{X}_j' + \mathbf{\Psi}_j$$

$$K_t(i,j) = \mathbf{H}_{t/t-1}^{i(j)} \mathbf{X}_j' \mathbf{\Sigma}_{t/t-1}^{i(j)\,-1}$$

$$L_t(i,j) = \mathcal{N}\left[\,\mathbf{0}\,,\mathbf{\Sigma}_{t/t-1}^{i(j)}\right]$$

Ensuite, nous mettons à jour nos estimations de la moyenne et de la variance, soient

$$\begin{aligned}
\mathbf{f}_{t/t}^{i(j)} &= \mathbf{f}_{t/t-1}^{i(j)} + K_t(i,j)\mathbf{e}_t(i,j) && (5.26)\\
\mathbf{H}_{t/t}^{i(j)} &= \left[\mathbf{I}_k - K_t(i,j)\mathbf{X}_j\right]\mathbf{H}_{t/t-1}^{i(j)} = \mathbf{H}_{t/t-1}^{i(j)} - K_t(i,j)\mathbf{\Sigma}_{t/t-1}^{i(j)} K_t(i,j)' && (5.27)
\end{aligned}$$

Le problème fondamental inhérent au filtre de Kalman multi-régime, c'est que le nombre de séquences d'états possibles à chaque instant t augmente d'une manière exponentielle avec le temps. Supposons que la distribution initiale $p(\mathbf{f}_1)$ est un mélange de m gaussiennes, une pour chaque valeur de S_1. Chaque composante sera propagée par la suite à travers m équations différentes (une pour chaque valeur de S_2), de telle façon que $p(\mathbf{f}_2)$ devient un mélange de m^2 gaussiennes. En général, à un instant t quelconque, la probabilité de l'état $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$ devient un mélange de m^t gaussiennes, une pour chaque séquence d'états possible $S_1, ..., S_t$. Afin de résoudre ce problème de croissance exponentielle nous avons utilisé la technique de fusion (collapsing technique en Anglais). Cette technique consiste à approcher le mélange de m^t gaussiennes par un mélange de r gaussiennes. Une telle méthode est appelée méthode pseudo bayésienne généralisée d'ordre r (GPB(r)). Lorsque $r = 1$, on approxime le mélange par une seule gaussienne en utilisant la technique dite "moment matching".

Pour l'implémentation de cet algorithme, nous calculons les probabilités suivantes :

$$M_{t-1,t/t}(i,j) = \frac{L_t(i,j)p_{ij}M_{t-1/t-1}(i)}{\sum_{i=1}^{m}\sum_{j=1}^{m} L_t(i,j)p_{ij}M_{t-1/t-1}(i)}$$

étant donné que

$$\begin{aligned}
M_{t-1,t/t}(i,j) &= p(S_{t-1} = i, S_t = j/\mathcal{Y}_{1:t}) \\
&= p(S_{t-1} = i, S_t = j/\mathbf{y}_t, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(S_{t-1} = i, S_t = j, \mathbf{y}_t/\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(\mathbf{y}_t/S_{t-1} = i, S_t = j, \mathcal{Y}_{1:t-1}) p(S_{t-1} = i, S_t = j/\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(\mathbf{y}_t/S_{t-1} = i, S_t = j, \mathcal{Y}_{1:t-1}) p(S_{t-1} = i/\mathcal{Y}_{1:t-1}) \times \\
&\quad p(S_t = j/S_{t-1} = i, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} L_t(i,j)p_{ij}M_{t-1/t-1}(i)
\end{aligned}$$

où c est la constante de normalisation donnée par :

$$c = \sum_{i=1}^{m}\sum_{j=1}^{m} L_t(i,j)p_{ij}M_{t-1/t-1}(i)$$

Nous calculons, aussi, les probabilités

$$M_{t/t}(j) = \sum_{i=1}^{m} M_{t-1,t/t}(i,j)$$

$$Z_{i/j}(t) = p(S_{t-1} = i/S_t = j, \mathcal{Y}_{1:t}) = M_{t-1,t/t}(i,j)/M_{t/t}(j)$$

En dernière étape, les moyennes, les volatilités et les volatilités prédites seront mises à jour à travers les équations suivantes :

$$
\begin{aligned}
\mathbf{f}^j_{t/t} &= \sum_{i=1}^m Z_{i/j}(t)\mathbf{f}^{i(j)}_{t/t} \\
h^j_{lt/t} &= \sum_{i=1}^m Z_{i/j}(t)h^{i(j)}_{lt/t} + \sum_{i=1}^m Z_{i/j}(t)\left[f^{i(j)}_{lt/t} - f^j_{lt/t}\right]\left[f^{i(j)}_{lt/t} - f^j_{lt/t}\right]' \\
h^j_{lt/t-1} &= \sum_{i=1}^m Z_{i/j}(t)h^{i(j)}_{lt/t-1} + \sum_{i=1}^m Z_{i/j}(t)\left[f^{i(j)}_{lt/t-1} - f^j_{lt/t-1}\right]\left[f^{i(j)}_{lt/t-1} - f^j_{lt/t-1}\right]' \\
\mathbf{H}^j_{t/t} &= diag\left[h^j_{lt/t}\right] \quad \text{et} \quad \mathbf{H}^j_{t/t-1} = diag\left[h^j_{lt/t-1}\right] \quad \text{pour} \quad l = 1, 2, ..., k
\end{aligned}
$$

L'Algorithme de Lissage

Étant donnée la nature dégénérée de l'équation de transition, la matrice de gain de lissage $J^{(j)k}_t$ est toujours nulle, soit

$$J^{(j)k}_t = \mathbf{H}^j_{t/t}\mathbf{0}'_k\mathbf{H}^{(j)k\,-1}_{t+1/t} = \mathbf{0}$$

ce qui implique :

$$
\begin{aligned}
\mathbf{f}^{(j)k}_{t/n} &= \mathbf{f}^j_{t/t} + J^{(j)k}_t\left[\mathbf{f}^k_{t+1/n} - \mathbf{f}^{j(k)}_{t+1/t}\right] = \mathbf{f}^j_{t/t} \\
\mathbf{H}^{(j)k}_{t/n} &= \mathbf{H}^j_{t/t} + J^{(j)k}_t\left[\mathbf{H}^k_{t+1/n} - \mathbf{H}^{j(k)}_{t+1/t}\right]J^{(j)k\prime}_t = \mathbf{H}^j_{t/t}
\end{aligned}
$$

Nous calculons par la suite les probabilités,

$$U^{j/k}_{t/t+1} = p(S_t = j/S_{t+1} = k, \mathcal{Y}_{1:n}) \simeq \frac{M_{t/t}(j)p_{jk}}{\sum\limits_{j'=1}^m M_{t/t}(j')p_{j'k}}$$

où l'approximation provient du fait que S_t n'est pas conditionnellement indépendante du futur $\mathbf{y}_{t+1}, ..., \mathbf{y}_n$ étant donné l'état S_{t+1}. Une telle approximation n'est pas, aussi, mauvaise à condition que le futur ne contient pas plus d'informations sur S_t autres que celles continues dans S_{t+1} (voir Kim [1994]).

Pour la mise à jour des paramètres, nous avons besoin aussi des probabilités

$$
\begin{aligned}
M_{t,t+1/n}(j,k) &= U^{j/k}_{t/t+1}M_{t+1/n}(k) \\
M_{t/n}(j) &= \sum_{k=1}^m M_{t,t+1/n}(j,k)
\end{aligned}
$$

5.4 Inférence basée sur l'Approximation de Viterbi

L'application de la méthode de Viterbi est connue depuis longtemps dans le cas des modèles de Markov cachés à états discrets (voir, par exemple, Rabiner et Juang [1993] et le chapitre 4) aussi bien que dans le cas des modèles de Gauss-Markov à états continus (Kalman [1960] et Kalman et Bucy [1961]). Dans le cas de notre modèle à facteurs conditionnellement hétéroscédastiques, cet algorithm consiste à identifier la meilleure séquence d'états cachés $\{S_t,\ t = 1, ..., n\}$, et de facteurs communs $\{\mathbf{f}_t\}$ permettant de minimiser le coût Hamiltonien donné par :

$$\begin{aligned}
&\mathcal{H}(\mathcal{F}_{1:n}, \mathcal{S}_{1:n}, \mathcal{Y}_{1:n}) \simeq Constante + \sum_{t=2}^{n} S_t'(-\log \mathbf{P})S_{t-1} + S_1'(-\log \pi) \\
&+\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m}\left[(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}_{jt} - \theta_j)'\mathbf{\Psi}_j^{-1}(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}_{jt} - \theta_j) + \log|\mathbf{\Psi}_j|\right] S_t(j) \\
&+\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m}\left[\mathbf{f}_{jt}'\mathbf{H}_{jt}^{-1}\mathbf{f}_{jt} + \log|\mathbf{H}_{jt}|\right] S_t(j) \qquad (5.28)
\end{aligned}$$

où $\mathcal{Y}_{1:n}$ est séquence complète d'observations, π le vecteur des probabilités de l'état initial et $\mathbf{P}$ la matrice de transition des états HMM. La *i-ème* ligne de cette matrice est donnée par $[p_{i1} \dots p_{im}]$ et $S_t = [S_t(1), ..., S_t(m)]'$, avec $S_t(j) = 1$ si $S_t = j$ et 0 sinon.

Si maintenant on désigne par $\mathcal{S}_{1:n}^*$ la meilleure séquence d'états Markoviens, nous pouvons approcher la distribution a posteriori $p(\mathcal{F}_{1:n}, \mathcal{S}_{1:n}/\mathcal{Y}_{1:n})$ par [6] :

$$\begin{aligned}
p(\mathcal{F}_{1:n}, \mathcal{S}_{1:n}/\mathcal{Y}_{1:n}) &= p(\mathcal{F}_{1:n}/\mathcal{S}_{1:n}, \mathcal{Y}_{1:n})p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n}) \\
&\simeq p(\mathcal{F}_{1:n}/\mathcal{S}_{1:n}, \mathcal{Y}_{1:n})\mu(\mathcal{S}_{1:n} - \mathcal{S}_{1:n}^*)
\end{aligned}$$

où la probabilité a posteriori $p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})$ a été approchée par son mode. D'une manière plus formelle, la séquence optimale d'états Markoviens $\mathcal{S}_{1:n}^*$ vérifie la propriété :

$$\mathcal{S}_{1:n}^* = \arg\max_{\mathcal{S}_{1:n}} p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})$$

Nous pouvons démontrer, aussi, qu'une solution "sous-optimale" à ce problème peut être obtenue par une optimisation récursive de la probabilité de la meilleure séquence à la date t.

$$\begin{aligned}
J_{t,j} &= \max_{\mathcal{S}_{1:t-1}} p(\mathcal{S}_{1:t-1}, S_t = j, \mathcal{Y}_{1:t}) \\
&\simeq \max_i \Big\{ p(\mathbf{y}_t/S_t = j, S_{t-1} = i, \mathcal{S}_{1:t-2}^*(i), \mathcal{Y}_{1:t-1})p(S_t = j/S_{t-1} = i) \\
&\qquad \times \max_{\mathcal{S}_{1:t-2}} p(\mathcal{S}_{1:t-2}, S_{t-1} = i, \mathcal{Y}_{1:t-1}) \Big\}
\end{aligned}$$

[6] $\mu(x) = 1$ pour $x = \emptyset$ et zéro autrement.

où $\mathcal{S}^*_{1:t-2}(i) = \arg\max_{\mathcal{S}_{1:t-2}} J_{t-1,i}$ est la "meilleure" séquence d'états Markoviens jusqu'à la date $t-1$ lorsque le système est à l'état i à la date $t-1$.

On définit tout d'abord le "meilleur" coût partiel jusqu'à la date t de la séquence $\mathcal{Y}_{1:t}$ lorsque le système est à l'état j à l'instant t :

$$J_{t,j} = \min_{\mathcal{S}_{1:t-1},\mathcal{F}_{1:t}} \mathcal{H}\left[\mathcal{F}_{1:t}, \{\mathcal{S}_{1:t-1}, S_t = j\}, \mathcal{Y}_{1:t}\right] \tag{5.29}$$

Notons que ce coût, est le coût minimal pour toutes les séquences possibles d'états Markoviens $\mathcal{S}_{1:t-1}$ et d'états continus du modèle à facteurs $\mathcal{F}_{1:t}$. Ce coût partiel est indispensable pour l'implémentation d'une inférence de Viterbi qui minimise un coût total. Pour une transition $i \to j$ quelconque, nous pouvons maintenant facilement établir une relation entre les estimations de filtrage et de prédiction (équations [5.24-5.25]). D'après la théorie de l'estimation de Kalman (Anderson et Moore [1979]), lorsque une nouvelle observation $\mathbf{y}_t$ devient disponible à la date t, chacune de ces estimations de prédiction seront filtrées en utilisant l'algorithme de mise à jour de Kalman (équations [5.26-5.27]). Ainsi, chacune de ces transitions $i \to j$ a un certain coût d'innovation $J_{t,t-1,i,j}$ qui lui est associé, et qui est donné par :

$$\begin{aligned} J_{t,t-1,i,j} &= \frac{1}{2}\left[\mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}^{i(j)}_{t/t-1}\right]' \mathbf{\Sigma}^{i(j)\,-1}_{t/t-1} \left[\mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}^{i(j)}_{t/t-1}\right] \\ &+ \frac{1}{2}\log\left|\mathbf{\Sigma}^{i(j)}_{t/t-1}\right| - \log p_{ij} \end{aligned} \tag{5.30}$$

Une partie de ce coût d'innovation reflète la transition de l'état continu (les facteurs), soient les termes d'innovation de l'équation (5.26). Le coût restant, $-\log p_{ij}$, est dû à la transition HMM de l'état i à l'état j.

Par ailleurs, pour chaque état actuel j, il y a m états possibles qui peuvent être à son origine. Afin minimiser le coût total à chaque instant t et pour chacun des états j, on doit sélectionner le "meilleur" état précédent i, soit

$$\begin{aligned} J_{t,j} &= \min_i \{J_{t,t-1,i,j} + J_{t-1,i}\} \\ \delta_{t-1,j} &= \arg\min_i \{J_{t,t-1,i,j} + J_{t-1,i}\} \end{aligned}$$

L'indexe de cet état sera récupéré dans la variable $\delta_{t-1,j}$. Ainsi, les m meilleures estimations de filtrage des états du modèle à facteurs et leurs variances à la date t seront données par : $\mathbf{f}^j_{t/t} = \mathbf{f}^{\delta_{t-1,j}\,(j)}_{t/t}$ et $\mathbf{H}^j_{t/t} = \mathbf{H}^{\delta_{t-1,j}\,(j)}_{t/t}$ avec $h^j_{lt/t-1} = h^{\delta_{t-1,j}\,(j)}_{lt/t-1}$ pour $l = 1, ..., k$. Une fois toutes les n observations $\mathcal{Y}_{1:n}$ seront traitées, le meilleur coût global sera obtenu par :

$$J^*_n = \min_j J_{n,j}$$

Pour décoder la meilleure séquence d'états, on utilise l'indexe du meilleur état final, $j_n^* = \arg\min_j J_{n,j}$, et par la suite on procède d'une manière retrograde à travers la variable qui contient tout l'historique des meilleurs états de transition $\delta_{t-1,j}$, afin d'obtenir l'état optimal à chaque instant t :

$$j_t^* = \delta_{t,j_{t+1}^*}$$

Les statistiques exhaustives nécessaires pour l'implémentation de l'algorithme EM seront, tout simplement, données par $\mathbb{E}(S_t/.) = S_t(j^*)$ et $\mathbb{E}(S_t S_{t-1}'/.) = S_t(j^*)S_{t-1}(j^*)'$.[7] Étant donnée la meilleure séquence d'états, les statistiques exhaustives du modèle à facteurs peuvent être obtenues directement moyennant l'algorithme de lissage de Rauch-Tung-Striber [1965] (voir aussi le chapitre 4 et Rosti et Gales [2001] pour une revue de la littérature plus récente). Par exemple,

$$\mathbb{E}(\mathbf{f}_t, S_t(j)/.) = \begin{cases} \mathbf{f}_{t/n}^{j_t^*} & j = j_t^* \\ \mathbf{0} & \text{autrement} \end{cases}$$

L'approche de Viterbi peut donc être résumée à travers les itérations suivantes :

Algorithme de VITERBI

Initialisation des statistiques de prédiction $\mathbf{f}_{0/-1}^j$ et $\mathbf{H}_{0/-1}^j$; et du coût $J_{0,j}$.

pour $t = 1 : n$
- **pour** $j = 1 : m$
  - **pour** $i = 1 : m$
    - Algorithme de filtrage de kalman
    - Calcul de $\mathbf{f}_{t/t}^{i(j)}$ et $\mathbf{H}_{t/t}^{i(j)}$
    - Calculer le coût d'innovation $J_{t/t-1,i,j}$
  - **fin**
  - Calculer le "meilleur" coût partiel $J_{t,j}$, l'état de transition $\delta_{t-1,j}$, et les estimations des statistiques $\mathbf{f}_{t/t}^j$ et $\mathbf{H}_{t/t}^j$.
- **fin**

fin

Identification du "meilleur" état de transition j_n^*.

Récurrences arrières → la "meilleure" séquence d'états, j_t^*.

Calculer les statistiques exhaustives du modèle.

---

[7] L'opérateur $\mathbb{E}(/.)$ désigne l'espérance conditionnelle par rapport à la distribution a posteriori, par exemple, $\mathbb{E}(\mathbf{f}_t/.) = \sum_{\mathcal{S}} \int_{\mathcal{F}} \mathbf{f}_t p(\mathcal{F}, \mathcal{S}/\mathcal{Y})$.

5.5 Algorithme EM

Comme pour les modèles à facteurs standards, il paraît naturel d'envisager l'estimation des paramètres du modèle à facteurs conditionellement hétéroscédastiques avec changement de régime à l'aide de la méthode du maximum de vraisemblance et d'utiliser pour ceci une version généralisée de l'algorithme EM. Rappelons que pour maximiser la log-vraisemblance, cet algorithme fait appel à la notion de données complétées (le vecteur $[\mathbf{y}, \mathbf{f}, \mathcal{S}]$ dans notre cas) et s'appuie sur la log-vraisemblance $\mathcal{L}(\Theta/\mathcal{Y}, \mathcal{F}, \mathcal{S})$ de ces données complétées qui s'écrit :

$$\mathcal{L}(\Theta/\mathcal{Y}, \mathcal{F}, \mathcal{S}) = \log \left[p(S_1) \prod_{t=2}^{n} p(S_t/S_{t-1}) \prod_{t=1}^{n} p(\mathbf{f}_t/S_t, \mathcal{D}_{1:t-1}) p(\mathbf{y}_t/\mathbf{f}_t, S_t, \mathcal{D}_{1:t-1}) \right]$$

où $\mathcal{D}_{1:t-1} = \{\mathcal{Y}_{1:t-1}, \mathcal{F}_{1:t-1}, \mathcal{S}_{1:t-1}\}$, est l'ensemble informationnel disponible à la date $t-1$. Le principe de l'algorithme est de maximiser de manière itérative l'espérance de cette log-vraisemblance complétée conditionnellement aux données $\mathcal{Y}$ et à la valeur du paramètre courant $\Theta^{(i)}$:

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \mathbb{E}\left[\log p(\mathcal{Y}, \mathcal{F}, \mathcal{S}/\Theta^{(i)})/\mathcal{Y}, \Theta \right]$$
$$= \sum_{\forall \mathcal{S}} \int p(\mathcal{F}/\mathcal{Y}, \mathcal{S}, \Theta) p(\mathcal{S}/\mathcal{Y}, \Theta) \log p(\mathcal{Y}, \mathcal{F}, \mathcal{S}/\Theta^{(i)})\, d\mathcal{F}$$

Les étapes de maximisation permettent de trover $\Theta^{(i+1)}$, valeur de Θ qui maximise $\mathcal{Q}(\Theta, \Theta^{(i)})$ à travers toutes les valeurs possibles de Θ. $\Theta^{(i+1)}$ remplace par la suite $\Theta^{(i)}$ au niveau de l'étape E et $\Theta^{(i+2)}$ sera choisi comme maximum de $\mathcal{Q}(\Theta, \Theta^{(i+1)})$. La procédure sera répétée jusqu'à ce que la séquence $\Theta^{(0)}$, $\Theta^{(1)}$, $\Theta^{(2)}$, ... converge. L'algorithme EM est construit de telle façon que la séquence des $\Theta^{(i)}$ convergera vers l'estimateur de maximum de vraisemblance de Θ.

Malheureusement, la maximisation de cette fonction $\mathcal{Q}(\Theta, \Theta^{(i)})$ n'est pas directe comme pour le modèle standard ; les difficultés résultent de la structure de la dépendance du modèle et la détermination de l'espérance conditionnelle de certaines fonctions non linéaires de $\mathbf{f}_t$ pose des problèmes. Cette situation est voisine de celle des modèles conditionellement hétéroscédastiques étudiés dans le chapitre 3 qui nécessitent la subdivision de l'étape de maximisation en deux sous-étapes de maximisation conditionnelle. Ainsi, pour résoudre ce problème, nous proposons ici une démarche en trois étapes :

Étape E :

Soit $\mathcal{D}_n^{(i)} = \left\{\mathcal{Y}_{1:n}, \Theta^{(i)}\right\}$ et $\widetilde{\mathbf{y}}_{jt} = \mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j$. L'espérance conditionnelle de la log-vraisemblance complétée est donnée par :

$$
\begin{aligned}
\mathcal{Q}(\Theta/\Theta^{(i)}) \simeq & \sum_{j=1}^{m} M_{1/n}(j)\log p(S_1) - \sum_{t=2}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m} M_{t-1,t/n}(i,j)\log p_{ij} \\
- & \frac{1}{2}\sum_{j=1}^{m}\sum_{t=1}^{n} M_{t/n}(j)\left[\log|\mathbf{\Psi}_j| + \mathbb{E}\left\{(\widetilde{\mathbf{y}}_{jt}-\theta_j)'\mathbf{\Psi}_j^{-1}(\widetilde{\mathbf{y}}_{jt}-\theta_j)/\mathcal{D}_n^{(i)}\right\}\right] \\
- & \frac{1}{2}\sum_{j=1}^{m}\sum_{l=1}^{k}\sum_{t=1}^{n} M_{t/n}(j)\mathbb{E}\left[\log(h_{lt}^{j}) + \frac{f_{lt}^2}{h_{lt}^{j}}/\mathcal{D}_n^{(i)}\right] \qquad (5.31)
\end{aligned}
$$

Étape CM1 :

Maintenant, étant données les statistiques exhaustives qu'on a déjà calculé moyennant les algorithmes GPB(1) ou de Viterbi, l'optimisation des paramètres du modèle peut être menée en maximisant la log-vraisemblance complétée (5.31) par rapport aux probabilités de l'état initial π_j, les probabilités de transition p_{ij}, les moyennes des facteurs spécifiques θ_j, les pondérations $\mathbf{X}_j$ et les variances idiosyncratiques $\mathbf{\Psi}_j$. Une description plus détaillée de ces calculs sera présentée dans l'annexe.

$$\widehat{\pi}_j = \frac{M_{1/n}(j)}{\sum_{i=1}^{m} M_{1/n}(i)}$$

$$\widehat{p}_{ij} = \frac{\sum_{t=2}^{n} M_{t-1,t/n}(i,j)}{\sum_{t=2}^{n} M_{t-1/n}(i)}$$

$$\widehat{\theta}_j = \frac{1}{\sum_{t=1}^{n} M_{t/n}(j)} \sum_{t=1}^{n} M_{t/n}(j)\left(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}_{t/n}^{j}\right)$$

$$\widehat{\mathbf{x}}_{jl} = \left[\sum_{t=1}^{n} M_{t/n}(j)(y_{tl}-\theta_{jl})\mathbf{f}_{t/n}^{j}\right]'\left[\sum_{t=1}^{n} M_{t/n}(j)\left[\mathbf{H}_{t/n}^{j} + \mathbf{f}_{t/n}^{j}\mathbf{f}_{t/n}^{j\prime}\right]\right]^{-1}$$

$$
\begin{aligned}
\widehat{\mathbf{\Psi}}_j = \frac{1}{\sum_{t=1}^{n} M_{t/n}(j)} \sum_{t=1}^{n} M_{t/n}(j) diag & \left\{\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{f}_{t/n}^{j}\mathbf{y}_t' \\ \mathbf{y}_t' \end{bmatrix} - \begin{bmatrix} \mathbf{y}_t\mathbf{f}_{t/n}^{j\prime} & \mathbf{y}_t \end{bmatrix}\right. \\
& \left. \times \begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix} + \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{H}_{t/n}^{j} + \mathbf{f}_{t/n}^{j}\mathbf{f}_{t/n}^{j\prime} & \mathbf{f}_{t/n}^{j} \\ \mathbf{f}_{t/n}^{j\prime} & 1 \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix}\right\}
\end{aligned}
$$

où $\mathbf{x}_{jl}$ est le *l-ème* vecteur ligne de $\mathbf{X}_j$, y_{tl} et θ_{jl} sont, respectivement, les *l-ème* composantes du vecteur des observations actuelles et du vecteur des moyennes spécifiques sous le régime j.

Étape CM2 :

Étant données les nouvelles valeurs de π_j, p_{ij}, θ_j, $\mathbf{X}_j$ et $\mathbf{\Psi}_j$, si les facteurs et les différents états de la chaîne de Markov seront observés on aura :

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{f}_t \end{pmatrix} /\mathcal{D}_{1:t-1}, S_t = j \sim \mathcal{N}\left[\begin{pmatrix} \theta_j \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}_j\mathbf{H}_{jt}\mathbf{X}_j' + \mathbf{\Psi}_j & \mathbf{X}_j\mathbf{H}_{jt} \\ \mathbf{H}_{jt}\mathbf{X}_j' & \mathbf{H}_{jt} \end{pmatrix}\right]$$

Cependant, les états $\mathbf{f}_t$ et S_t sont cachés, mais dans ce cas et afin d'estimer les paramètres du modèle, nous pouvons comme dans le chapitre 3 approximer la distribution des $\mathbf{y}_t$, conditionnellement à l'ensemble informationnel disponible à la date $t-1$ en utilisant la distribution suivante :

$$\mathbf{y}_t/\mathcal{Y}_{1:t-1}, S_t = j, \mathcal{S}_{1:t-1} \approx \mathcal{N}\left[\theta_j, \mathbf{\Sigma}_{t/t-1}^{(j)}\right]$$

où $\mathbf{\Sigma}_{t/t-1}^{(j)} = \mathbf{X}_j\mathbf{H}_{t/t-1}^{(j)}\mathbf{X}_j' + \mathbf{\Psi}_j$ avec $\mathbf{H}_{t/t-1}^{(j)}$ l'espérance de $\mathbf{H}_t$, conditionnellement à $\mathcal{Y}_{1:t-1}$ et $\mathcal{S}_{1:t}$, obtenue via une version quasi-optimale du filtre de Kalman. Le *l-ème* élément de la diagonale de $\mathbf{H}_{t/t-1}^{(j)}$ sera donné par $h_{lt/t-1}^{j} = h_{lt/t-1}^{\delta_{t-1,j}(j)}$. Ainsi, en ignorant les conditions initiales, la pseudo log-vraisemblance peut s'écrire sous la forme :

$$\mathcal{L}^* = c - \frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m} S_t(j)\left[\log|\mathbf{\Sigma}_{t/t-1}^{(j)}| + (\mathbf{y}_t - \theta_j)'\mathbf{\Sigma}_{t/t-1}^{(j)-1}(\mathbf{y}_t - \theta_j)\right] \qquad (5.32)$$

Dans la deuxième étape de maximisation conditionnelle, en utilisant les nouvelles valeurs des paramètres θ_j, $\mathbf{X}_j$ et $\mathbf{\Psi}_j$ déjà trouvées au niveau de l'étape *CM1*, on maximise(5.32) par rapport aux paramètres de la composante conditionnellement hétéroscédastique, w_j, γ_j, α_j et δ_j. Il faut ensuite recommencer les étapes E et *CM1* avec ces nouvelles solutions. Le procédé sera donc répété jusqu'à la convergence souhaitée. Cependant, l'implémentation de cet algorithme d'optimisation nécessite l'identification de la séquence optimale des états cachés. Ce problème peut être résolu en utilisant soit les probabilités a posteriori $M_{t/n}(j)$ déjà fournies par l'algorithme de lissage, ou bien la séquence optimale obtenue par l'approximation de Viterbi. Une fois que cette séquence sera connue, sur chaque segment de données on maximise la pseudo log-vraisemblance $\mathcal{L}^*$ en utilisant un algorithme de type Newton.

Les dérivés premières de la pseudo log-vraisemblance $\mathcal{L}^*$ par rapport aux paramètres $\phi_{lj} = \{w_{lj}, \gamma_{lj}, \alpha_{lj}, \delta_{lj}\}$, $j = 1, ...m$ et $l = 1, 2, ..., k$, sont données par :

$$\frac{\partial\mathcal{L}^*(\Theta/\mathcal{Y})}{\partial\phi_j} = -\frac{1}{2}\sum_{t=1}^{n}\frac{\partial\,\mathbf{h}_{t/t-1}^{(s_t)\prime}}{\partial\phi_j}\, vecd\left[\mathbf{X}_{s_t}'\mathbf{\Sigma}_{t/t-1}^{(s_t)-1}\left[\mathbf{\Sigma}_{t/t-1}^{(s_t)} - (\mathbf{y}_t - \theta_{s_t})(\mathbf{y}_t - \theta_{s_t})'\right]\mathbf{\Sigma}_{t/t-1}^{(s_t)-1}\mathbf{X}_{s_t}\right]$$

avec

$$\begin{aligned}
\frac{\partial\, h^{(s_t)}_{lt/t-1}}{\partial w_{lj}} &= 1 \qquad \text{si } \ S_t = j \ \text{ et } \ 0 \ \text{ sinon} \\
\frac{\partial\, h^{(s_t)}_{lt/t-1}}{\partial \alpha_{lj}} &= \left[f^{(s_{t-1})\,2}_{lt-1/t-1} + h^{(s_{t-1})}_{lt-1/t-1}\right] + \delta_{lj}\frac{\partial\, h^{(s_{t-1})}_{lt-1/t-2}}{\partial \alpha_{lj}} \\
\frac{\partial\, h^{(s_t)}_{lt/t-1}}{\partial \gamma_{lj}} &= f^{(s_{t-1})}_{lt-1/t-1} + \delta_{lj}\frac{\partial\, h^{(s_{t-1})}_{lt-1/t-2}}{\partial \gamma_{lj}} \\
\frac{\partial\, h^{(s_t)}_{lt/t-1}}{\partial \delta_{lj}} &= h^{(s_{t-1})}_{lt-1/t-2} + \delta_{lj}\frac{\partial\, h^{(s_{t-1})}_{lt-1/t-2}}{\partial \delta_{lj}}
\end{aligned}$$

L'algorithme itératif utilisé pour la maximisation de la fonction $\mathcal{L}^*$ sans contraintes sur les paramètres ϕ_j est donné par la formule suivante :

$$\phi_j^{(i+1)} = \phi_j^{(i)} + \left[H_{(i)}\left(\phi_j^{(i)}\right)\right]^{-1} g_{(i)}\left(\phi_j^{(i)}\right)$$

où $\phi_j^{(i)}$ est le vecteur contenant les paramètres de la *i-ème* itération ; $H_{(i)}\left(\phi_j^{(i)}\right)$ est une approximation de la matrice Hessienne de $\mathcal{L}^*$ par rapport aux paramètres, évaluée à $\phi_j^{(i)}$; et $g_{(i)}\left(\phi_j^{(i)}\right)$ est le gradient négatif de $\mathcal{L}^*$ évalué à $\phi_j^{(i)}$. Cependant, afin de tenir compte des contraintes de positivité de la variance conditionnelle et de stationnarité au niveau de la covariance du processus GQARCH, nous pouvons utilisé directement la fonction *fmincon* de matlab.

Le modèle à facteurs standards est un cas particulier du système dynamique présenté dans la section 5.3. Dans ce cas, les formules de mise à jour des paramètres sont exactement les mêmes que celles du modèle conditionellement hétéroscédastique, à l'exception de la matrice de covariance des facteurs communs $\mathbf{H}_j$ qui est donnée par :

$$\widehat{\mathbf{H}}_j = \frac{1}{\sum\limits_{t=1}^{n}\gamma_j(t)} diag\left\{\sum_{t=1}^{n}\gamma_j(t)\left[\widetilde{\mathbf{R}}_j + \widetilde{\mathbf{f}}_{jt}\widetilde{\mathbf{f}}'_{jt}\right]\right\}$$

Les autres paramètres c-à-d, les matrices des pondérations $\mathbf{X}_j$, les vecteurs des moyennes θ_j et les variances idiosyncratiques $\mathbf{\Psi}_j$ seront obtenus en remplaçant, tout simplement, $\mathbf{f}^j_{t/n}$ et $\mathbf{H}^j_{t/n}$ par $\widetilde{\mathbf{f}}_{jt}$ et $\widetilde{\mathbf{R}}_j$.

5.6 Simulations de Monte Carlo

Dans cette section nous proposons de mener une série d'expérimentations afin d'étudier certaines propriétés des algorithmes que nous avons déjà présenté et des estimations résultantes du modèle proposé. Nous allons donc mener trois expérimentations qui vont nous permettre de répondre aux questions suivantes :

1. La question la plus importante qu'on se pose à propos des estimations est de savoir si ces dernières sont des estimations consistantes de Θ et quelle est aussi la taille raisonnable de la séquence d'observations permettant d'obtenir des estimations stables et exactes ?
2. La question la plus classique est de savoir si les estimations sont asymptotiquement distribuées selon la loi normale, et quelle est la taille de la séquence à partir de la quelle une telle approximation sera vérifiée.
3. Comment peut-on choisir un modèle fiable contenant un nombre suffisant de paramètres permettant d'assurer un ajustement réaliste à l'ensemble des données d'apprentissage. Pour répondre à cette question, nous allons utiliser deux critères de sélection : le AIC et le BIC.

5.6.1 Exactitude et Stabilité des Estimations

Les simulations que nous allons présenter maintenant sont basées sur des modèles avec $q = 6$ séries d'observations, trois états Markoviens cachés et un seul facteur commun suivant un processus GQARCH(1,1). Les dates de changement de régime sont $t_1^* = n/3+1$ et $t_2^* = 2n/3+1$. Les itérations de l'algorithme EM s'arrêteront lorsque le changement relatif de la fonction de vraisemblance entre deux itérations successives devient inférieur à une valeur seuil choisie, par exemple, égale à $= 10^{-4}$. Notre objectif est d'étudier le comportement des estimations lorsque la taille de la séquence n augmente de 600 à 1500. Pour ce faire, nous avons généré des séquences d'observations de tailles $n = 600$, 900, 1200 et 1500 (avec une centaine de replications pour chaque simulation). Ici nous avons considéré le cas où la constante de la spécification GQARCH est connue ($w_j = 1 \ \forall \ j = 1, 2, 3$). Les paramètres de cette simulation, aussi bien que les valeurs d'initialisation $\Theta^{(0)}$ de l'algorithme EM, sont donnés dans le tableau 5.4.

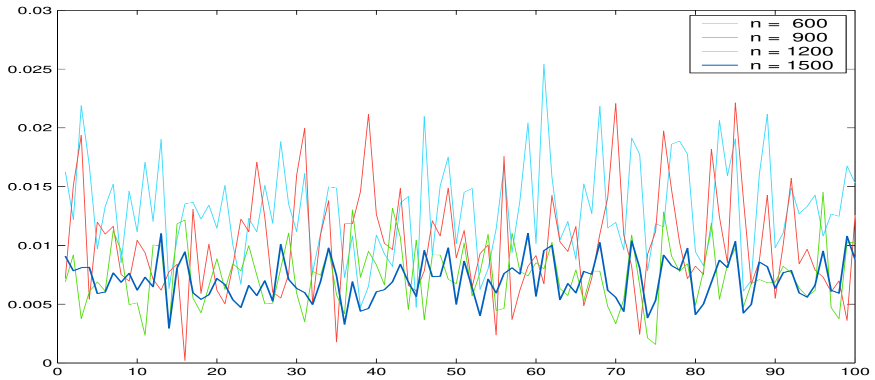

FIG. 5.4 – Les Divergences de Kullback-Leibler $\widetilde{K}(\Theta_0, \widetilde{\Theta}_n)$.

Pour mesurer la distance entre les estimations $\widetilde{\Theta}$ et les vrais paramètres Θ_0, nous avons utilisé la divergence de Kullback-Leibler (voir Juang et Rabiner [1985] pour une application sur les HMM) donnée par :

$$K(\Theta_0, \Theta) \stackrel{def}{=} \lim_{n} \frac{1}{n} \left\{ \log \mathcal{L}(\mathbf{y}_1, ..., \mathbf{y}_n; \Theta_0) - \log \mathcal{L}(\mathbf{y}_1, ..., \mathbf{y}_n; \Theta) \right\}$$

Tab. 5.4 – Paramètres de Simulation.

	θ	$\mathbf{X}$	$diag(\mathbf{\Psi})$	ϕ
	1.0000 (0.0000)	1.0000 (0.5000)	1.0000 (0.5000)	0.5000 (0.1200)
	1.0000 (1.0000)	2.0000 (1.0000)	1.0000 (0.5000)	0.1000 (0.1800)
État 1	1.0000 (0.5000)	3.0000 (1.0000)	1.0000 (0.5000)	0.8000 (0.3800)
	2.0000 (1.0000)	4.0000 (1.5000)	1.0000 (0.5000)	
	2.0000 (0.0000)	5.0000 (1.5000)	1.0000 (0.5000)	
	2.0000 (0.5000)	6.0000 (2.5000)	1.0000 (0.5000)	
	1.0000 (1.0000)	2.0000 (1.0000)	2.0000 (0.5000)	0.1000 (0.2900)
	2.0000 (1.0000)	2.0000 (0.5000)	2.0000 (0.5000)	0.3000 (0.1200)
État 2	1.0000 (1.0000)	2.0000 (0.5000)	2.0000 (0.5000)	0.4000 (0.7800)
	2.0000 (1.0000)	3.0000 (1.0000)	2.0000 (0.5000)	
	1.0000 (1.0000)	3.0000 (0.5000)	2.0000 (0.5000)	
	2.0000 (1.0000)	3.0000 (0.5000)	2.0000 (0.5000)	
	2.0000 (1.0000)	1.0000 (1.0000)	3.0000 (0.5000)	0.2000 (0.6000)
	3.0000 (1.0000)	3.0000 (0.5000)	3.0000 (0.5000)	0.2000 (0.5400)
État 3	2.0000 (1.0000)	1.0000 (0.5000)	3.0000 (0.5000)	0.6000 (0.2000)
	3.0000 (1.0000)	2.0000 (1.0000)	3.0000 (0.5000)	
	2.0000 (1.0000)	4.0000 (0.5000)	3.0000 (0.5000)	
	3.0000 (1.0000)	4.0000 (0.5000)	3.0000 (0.5000)	

. Les vrais paramètres du modèle, (.) Les valeurs d'initialisation de l'algorithme EM.

où Θ_0 est l'ensemble des vrais paramètres. Pour une séquence finie de longueur n, on définit la divergence de Kullback-Leibler empirique entre deux ensembles de paramètres par la formule :

$$K_n(\Theta_0, \Theta) \stackrel{def}{=} \frac{1}{n} \Big\{ \log\ \mathcal{L}(\mathbf{y}_1, ..., \mathbf{y}_n; \Theta_0) - \log\ \mathcal{L}(\mathbf{y}_1, ..., \mathbf{y}_n; \Theta) \Big\}$$

Pour chacune des valeurs de n, nous avons appliqué notre procédure d'estimation une centaine de fois, et les distances $\widetilde{K}_n(\Theta_0, \widetilde{\Theta}_n)$ entre chacun des cent estimateurs et le vrai paramètre Θ_0 ont été évaluées sur une nouvelle séquence, indépendante des cent premières qui ont été utilisées pour obtenir les estimateurs. Une telle procédure nous permet d'éviter la sous estimation potentielle de la distance qui peut résulter de l'estimation des paramètres et l'évaluation de leurs performances sur la même séquence. Dans le tableau 5.5 on donne les moyennes et les écart-types des estimations pour $n = 1500$. Les résultats montrent que notre méthode d'estimation fonctionne bien étant donné que les moyennes sont très proches des vrais paramètres et les écart-types sont faibles. La représentation graphique de $\widetilde{K}_n(\Theta_0, \widetilde{\Theta}_n)$ pour les différentes valeurs de n est donnée sous une même échelle dans la figure 5.4. Ce graphique montre une décroissance en moyenne et en écart de ces distances lorsque le nombre d'observations n augmente. Étant donné que des petites valeurs de $\widetilde{K}_n$ impliquent une très forte similarité entre Θ_0 et $\widetilde{\Theta}_n$, les résultats de cette expérimentation montrent une amélioration au niveau de l'exactitude et la stabilité de la séquence des estimations lorsque n augmente.

5.6.2 Distribution Asymptotique des Estimations

Afin d'étudier la distribution asymptotique des estimations $\widetilde{\Theta}_n$, nous avons utilisé la statistique de Shapiro-Francia [1972] permettant de tester la normalité de chacune

TAB. 5.5 – Moyennes et écart-types (.) des estimations avec $n = 1500$.

	θ	$\mathbf{X}$	$diag(\boldsymbol{\Psi})$	ϕ
	0.9833 (0.0983)	1.9880 (0.0682)	0.9789 (0.0472)	0.4988 (0.0736)
	1.0284 (0.0974)	1.9973 (0.0667)	1.0216 (0.0457)	0.1073 (0.0496)
État 1	1.0197 (0.0857)	2.9752 (0.0589)	0.9878 (0.0593)	0.7824 (0.0371)
	1.9875 (0.0861)	3.9940 (0.0571)	0.9958 (0.0607)	
	1.9914 (0.0973)	4.9945 (0.0577)	0.9980 (0.0572)	
	2.0841 (0.0866)	5.9652 (0.0604)	1.0106 (0.0486)	
	0.9932 (0.0718)	1.9961 (0.0583)	2.0162 (0.0615)	0.1017 (0.0765)
	1.9917 (0.0745)	2.0214 (0.0618)	2.0256 (0.0592)	0.3022 (0.0483)
État 2	1.0754 (0.0773)	2.0108 (0.0579)	1.9914 (0.0622)	0.3971 (0.0366)
	1.9886 (0.0852)	2.9972 (0.0564)	1.9947 (0.0638)	
	1.0381 (0.0836)	3.0127 (0.0591)	2.0082 (0.0676)	
	1.9914 (0.0794)	3.0394 (0.0538)	1.9928 (0.0606)	
	1.9726 (0.0833)	1.0134 (0.0475)	2.9988 (0.0584)	0.2046 (0.0776)
	2.9759 (0.0872)	3.0297 (0.0481)	3.0047 (0.0561)	0.1992 (0.0377)
État 3	1.9681 (0.0867)	1.0099 (0.0463)	2.9797 (0.0692)	0.5876 (0.0281)
	2.9726 (0.0954)	2.0192 (0.0454)	2.9783 (0.0667)	
	1.9718 (0.0988)	4.0154 (0.0508)	3.0146 (0.0689)	
	2.9690 (0.0826)	4.0205 (0.0511)	2.9792 (0.0712)	

des composantes de $\widetilde{\Theta}_n$ dans un cadre univarié. Ce test est généralement considéré comme étant relativement plus puissent par rapport à d'autres tests, et meilleur que le test de Shapiro-Wilk [1965] pour les échantillons Leptokurtiques. Le test de Shapiro-Francia est basé sur une idée proposée (sans demonstration) par Gupta [1952] (voir aussi Stephens [1975]) selon laquelle on obtient la statistique

$$\mathcal{W} = \frac{\left(m'\widetilde{\Theta}^{(v)}\right)^2}{(m'm)\sum_{i=1}^{v}(\widetilde{\Theta}_{(i)} - \overline{\Theta})^2} \quad \text{où} \quad m_i = \left(\frac{i-3/8}{n+1/4}\right)^{-1}, \quad i = 1, ..., v$$

Dans ce cas $m' = [m_1, m_2, ..., m_v]$, $\widetilde{\Theta}^{(v)} = \left(\widetilde{\Theta}_{(1)}, ..., \widetilde{\Theta}_{(v)}\right)$ la statistique ordonnée correspondante à $\widetilde{\Theta} = \left(\widetilde{\Theta}_1, ..., \widetilde{\Theta}_v\right)$ et v le nombre des réplications. Tous les résultats présentés dans le tableau 5.6 montrent que le test de Shapiro-Francia ne rejette pas l'hypothèse nulle (les Θ_i forment un échantillon aléatoire de la loi $\mathcal{N}(\mu, \sigma)$, avec μ et σ inconnus) pour un niveau de signification $\alpha = 5\%$.

5.6.3 Sélection de Modèles

Pour le choix de la structure de volatilité convenable, nous allons utiliser les critères de sélection AIC et BIC. Sur le plan empirique, les performances du critère BIC ont été mises en évidence dans plusieurs études portant sur les modèles de mélange (voir par exemple Roeder et Wasserman, [1997]) et, sur le plan théorique, il a été déjà démontré que ce critère fournira une estimation consistante de la dimension du modèle de Markov caché sous certaines conditions de rigueur (Gassiat, [2002]). Cependant, le critère AIC

TAB. 5.6 – Test de Shapiro-Francia (simulation avec $n = 900$).

Statistique						
						Vecteurs des moyennes
pval	*0.4964	0.4103	0.3976	0.4184	0.1838	0.4413
	**0.3114	0.4601	0.4819	0.3187	0.3489	0.2653
	***0.2500	0.1668	0.3310	0.1834	0.3513	0.2108
$\mathcal{W}$ statistic	0.0089	0.2267	-0.2594	0.2061	-0.9010	0.1478
	0.4920	0.1001	0.0454	0.4713	0.3882	0.6271
	0.6745	0.9667	0.4372	0.9027	0.3819	0.8035
						Les Pondérations
	0.3450	0.2838	0.1668	0.3760	0.3877	0.3819
	0.1997	0.4091	0.2831	0.3023	0.3190	0.2546
	0.4767	0.3227	0.2908	0.3868	0.1926	0.1367
	-0.3988	-0.5717	-0.9669	-0.3159	0.2853	0.3006
	0.8427	0.2299	0.5737	-0.5179	0.4705	-0.6601
	0.0586	-0.4602	-0.5509	0.2878	0.8683	-1.0954
						Les Variances Idiosyncratiques
	0.4870	0.2921	0.2474	0.2510	0.2269	0.4519
	0.4104	0.4838	0.4740	0.3962	0.3703	0.2766
	0.3707	0.3742	0.2888	0.2860	0.2778	0.3131
	0.0326	0.5473	-0.6828	0.6714	0.7491	-0.1208
	-0.2265	0.0406	0.0652	0.2632	-0.3311	0.5930
	-0.3299	0.3208	-0.5570	0.5652	-0.5895	-0.4871
						Les paramètres GQARCH
	0.3224	0.1812	0.2117			
	0.3138	0.4761	0.3436			
	0.4356	0.4967	0.4386			
	0.4611	0.9109	0.8006			
	-0.4852	-0.0599	-0.4027			
	-0.1622	0.0084	-0.1545			

* Régime 1, ** Régime 2, *** Régime 3.

permet dans la plupart des cas de sélectionner des modèles, aussi, complexes (voir Burnham et Anderson [1998]).

Dans cette expérience nous considérons deux situations différentes avec des modèles à facteurs qui diffèrent par leurs structures cachées. Dans le premier cas, le vrai modèle est celui que nous avons déjà utilisé dans la simulation 5.5.1 avec un nombre d'observations égale à 900. Dans le second cas, le vrai modèle est construit par deux états Markoviens cachés et deux facteurs conditionnellement hétéroscédastiques suivant des processus GQARCH(1,1). Le nombre d'observations dans ce deuxième exemple est égal à 800, et la date du changement de régime est $t^* = n/2 + 1$ (les paramètres de cette deuxième simulation sont donnés dans le tableau 5.7). Les étapes de la procédure de sélection de modèles sont comme suit. Pour chaque critère de sélection, nous utilisons l'algorithme EM pour estimer sur la même base de données plusieurs configurations (obtenues en changeant le nombre des états aussi bien que celui des facteurs). Dans le deuxième exemple, nous avons utilisé des initialisations aléatoires pour l'implémentation de l'algorithme EM[8]. La minimisation du critère de sélection – calculé après les itérations EM – nous permettra de trouver le meilleur modèle parmi tous les candidats. Les résultats

[8]Les valeurs initiales de l'algorithme EM, ont été obtenues moyennant une perturbation aléatoire allant jusqu'à 20% des vrais paramètres du modèle

pour les deux exemples sont donnés dans le tableau 5.8. Dans le premier exemple, le critère BIC choisi 3 états et un seul facteur. C'est en fait la meilleure classification, étant donné que l'utilisation de un ou deux états ne représente pas la structure réelle des données, et le choix de deux facteurs conduit à un sur-ajustement. Dans le deuxième exemple, le BIC choisi aussi la spécification adéquate avec deux états et deux facteurs conditionnellement hétéroscédastiques.

Le critère de l'erreur carrée moyenne donné par

$$\widehat{e} = \frac{1}{n} \sum_{i=1}^{q} \sum_{t=1}^{n} \|y_{it} - \widehat{y}_{it}\|^2$$

où $\widehat{\mathbf{y}}_t = \sum_{j=1}^{m} S_t(j) \left[\widehat{\theta}_j + \widehat{\mathbf{X}}_j \mathbf{f}^j_{t/n} \right]$ montre aussi que $k = 1$ et $m = 3$ est fortement favorisé dans le premier exemple (figure 5.5).

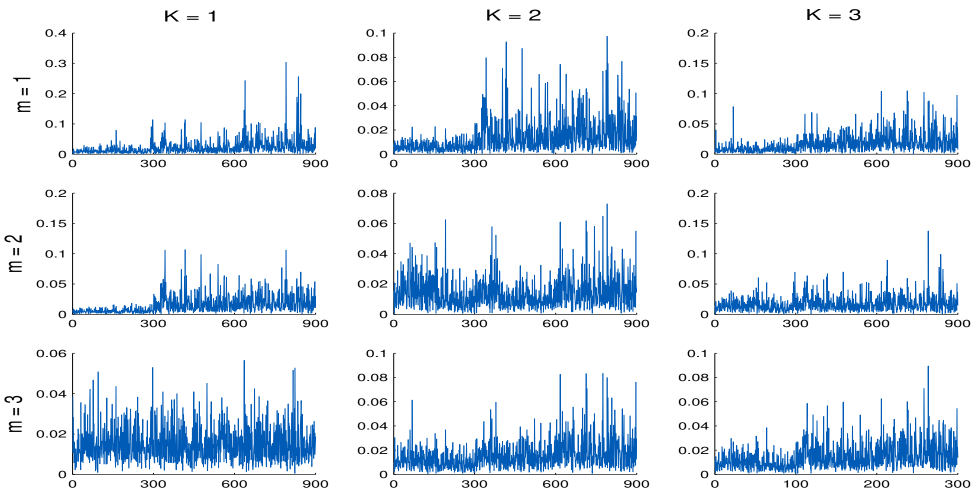

FIG. 5.5 – Calcul de l'erreur d'estimation (premier exemple) pour 9 configurations différentes avec hétéroscédasticité dynamique.

Pour mettre en valeur l'évolution des estimations du modèle obtenues par la méthode EM, la figure 5.6 montre les estimations des états HMM aux itérations 2, 5, 10 et 15. Chaque figure représente la trajectoire de la variable d'état discrète (le régime) obtenue en utilisant le vrai modèle. Il est claire donc qu'après 15 itérations, l'algorithme nous donnera la meilleure segmentation. En utilisant les valeurs initiales du tableau 5.4, la figure 5.7 montre que l'algorithme EM convergera vers les estimations des processus GQARCH après environ 50 itérations. Les figures 5.8 et 5.9 montrent qu'à l'exception du vrai modèle, tous les autres modèles conduisent soit à une sur-estimation soit à une sous-estimation. Les figures 5.10 et 5.11 montrent, respectivement, que la log-vraisemblance la plus élevée est celle qui correspond au vrai modèle et que les erreurs d'estimation ne sont pas corrélées. Ainsi, toute la corrélation entre les variables observées est complètement expliquée par les facteurs communs et spécifiques.

TAB. 5.7 – Paramètres de simulation (Exemple 2).

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$	ϕ	
	0.5000	1.0000	1.0000	0.1000	0.1000	0.5000
	0.5000	2.0000	1.0000	0.1000	0.3000	0.1000
État 1	0.7000	3.0000	2.0000	0.1000	0.4000	0.8000
	1.0000	4.0000	2.0000	0.1000		
	0.5000	5.0000	3.0000	0.1000		
	0.7000	6.0000	3.0000	0.1000		
	1.0000	1.0000	1.0000	0.4000	0.3000	0.2000
	0.9000	1.0000	2.0000	0.4000	0.2000	0.1000
État 2	1.0000	4.0000	3.0000	0.4000	0.7000	0.6000
	0.7000	4.0000	3.0000	0.4000		
	1.1000	2.0000	2.0000	0.4000		
	1.5000	2.0000	1.0000	0.4000		

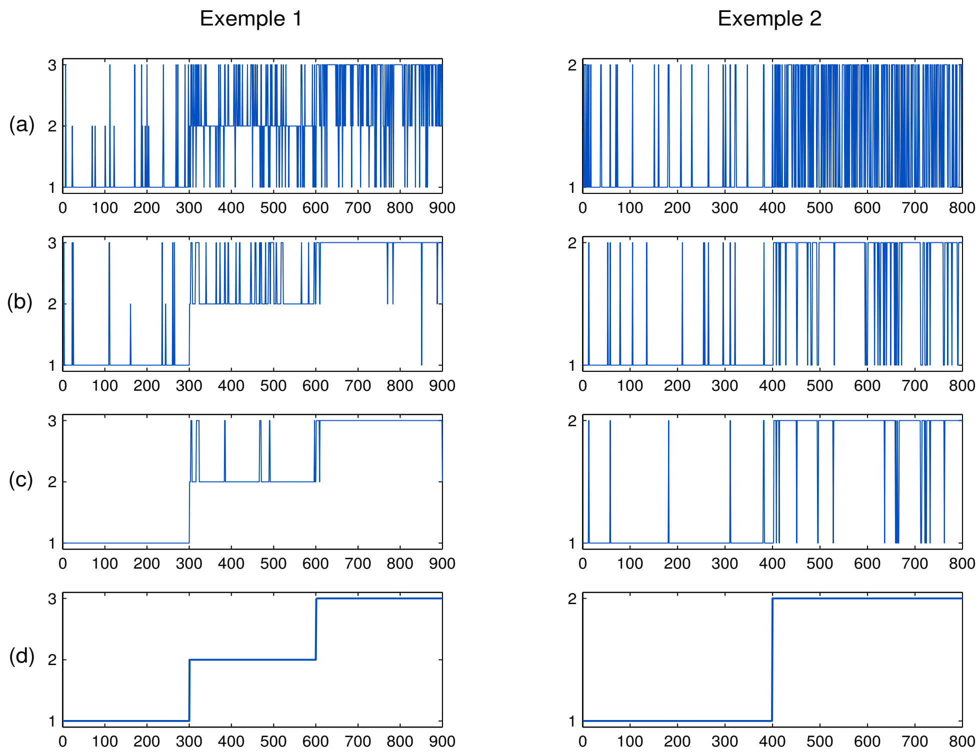

FIG. 5.6 – Évolution de l'estimation des états HMM en utilisant le vrai modèle : (a) itération 2, (b) itération 5, (c) itération 10, (d) itération (15).

TAB. 5.8 – Valeurs des critères AIC et BIC pour différents modèles à facteurs estimés sur la même base de données. Les valeurs entre parenthèses sont les critères de sélection de l'exemple 2.

Critère	$k=1$	$k=2$	$k=3$
		$m=1$	
AIC	24310 (22610)	24082 (22494)	24016 (22414)
BIC	24411 (22708)	24226 (22635)	24203 (22597)
		$m=2$	
	23398 (22332)	23248 (22240)	23160 (22312)
	23629 (22557)	23565 (**22549**)	23563 (22706)
		$m=3$	
	23190 (22324)	23412 (22248)	23544 (22380)
	23550 (22675)	23902 (22726)	24164 (22984)

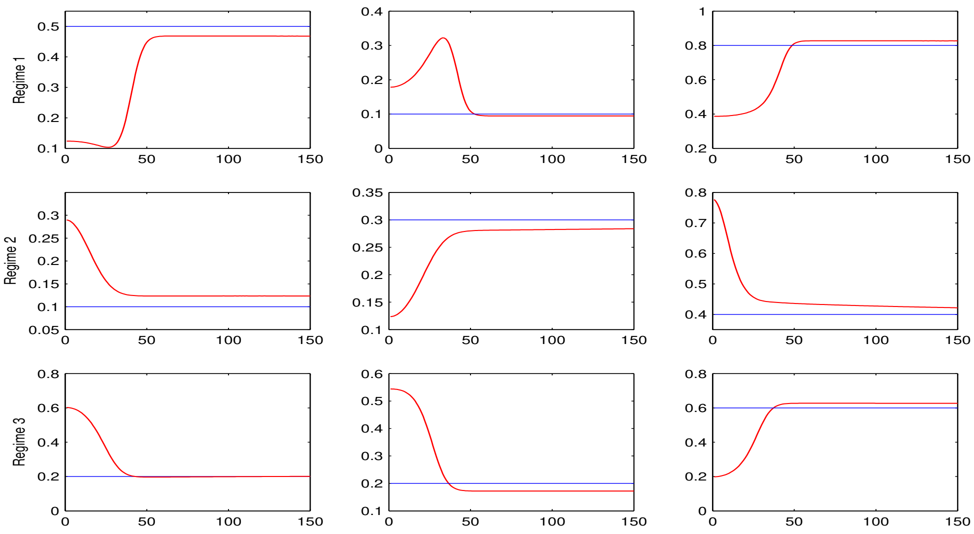

FIG. 5.7 – Évolution de l'estimation des paramètres de la composante conditionnellement hétéroscédastique durant les itérations EM dans le premier exemple : γ_j (colonne 1), α_j (colonne 2) and δ_j (colonne 3).

Pour confirmer les résultats précédents, nous avons mené des expérimentations de Monte Carlo. Pour ce faire, nous avons généré 100 réplications à partir du vrai modèle dans chacun des deux exemples. Par la suite, nous avons appliqué le critère BIC afin de choisir le meilleur nombre de facteurs communs et d'états Markoviens. La figure 5.12 donne les fréquences de choix pour chacune des spécifications. Dans les deux exemples, nous remarquons que le BIC préfère dans la plupart du temps le vrai modèle.

5.7 Application Empirique

Notre modèle conditionellement hétéroscédastique sera maintenant appliqué pour la modélisation des comouvements de huit devises durant la période de crise financière

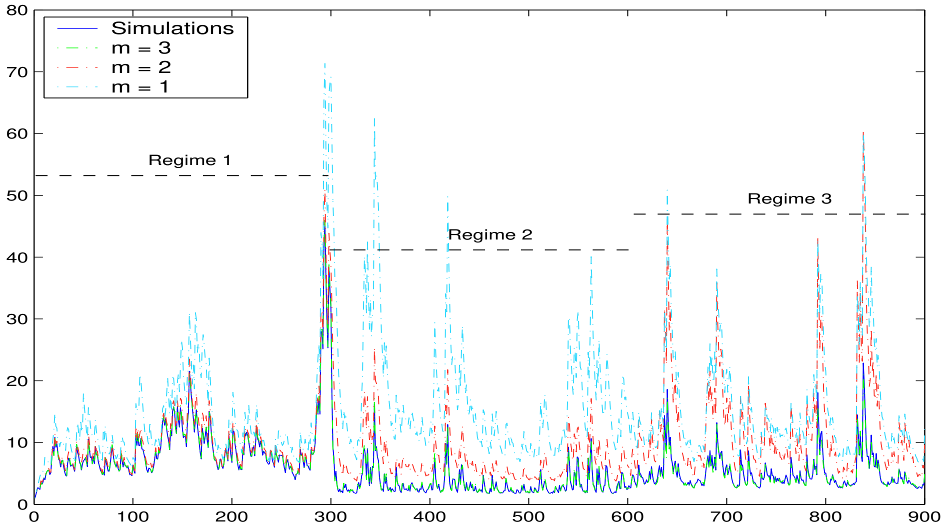

FIG. 5.8 – Exemple 1 : Volatilité des facteurs communs pour différentes spécifications.

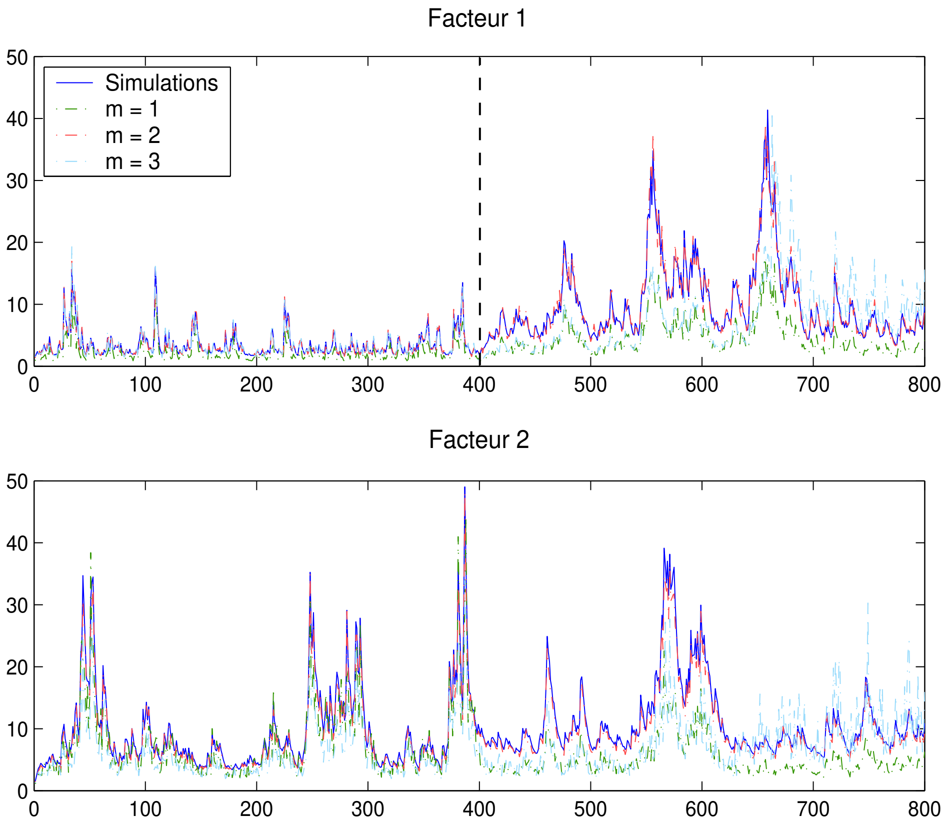

FIG. 5.9 – Exemple 2 : Volatilité des facteurs communs pour différentes spécifications. La ligne verticale représente la date de changement de régime.

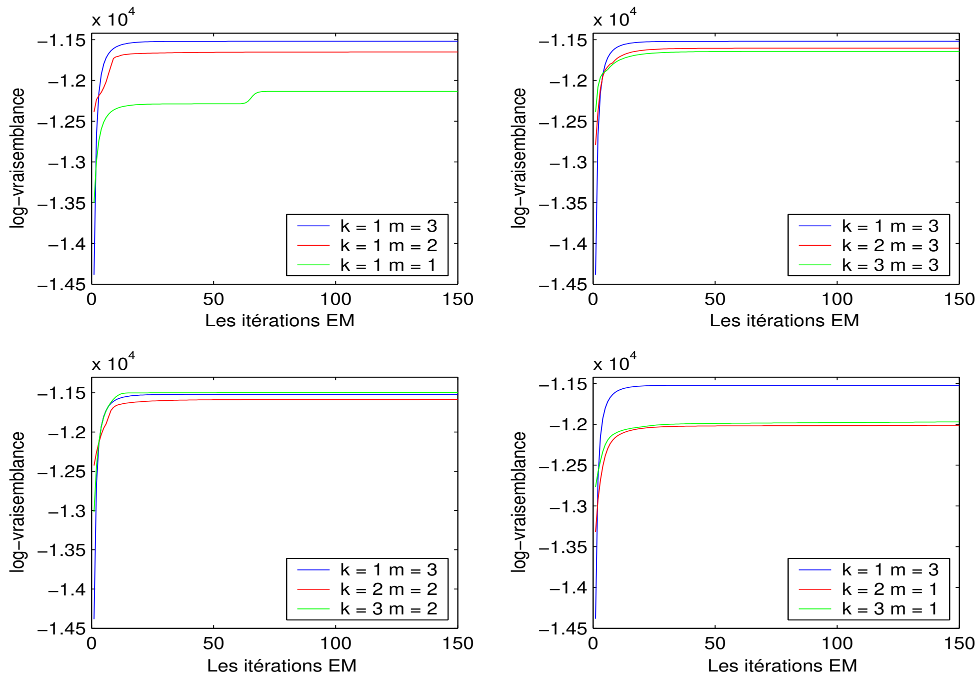

FIG. 5.10 – Exemple 1 : Log-vraisemblances des différentes spécifications.

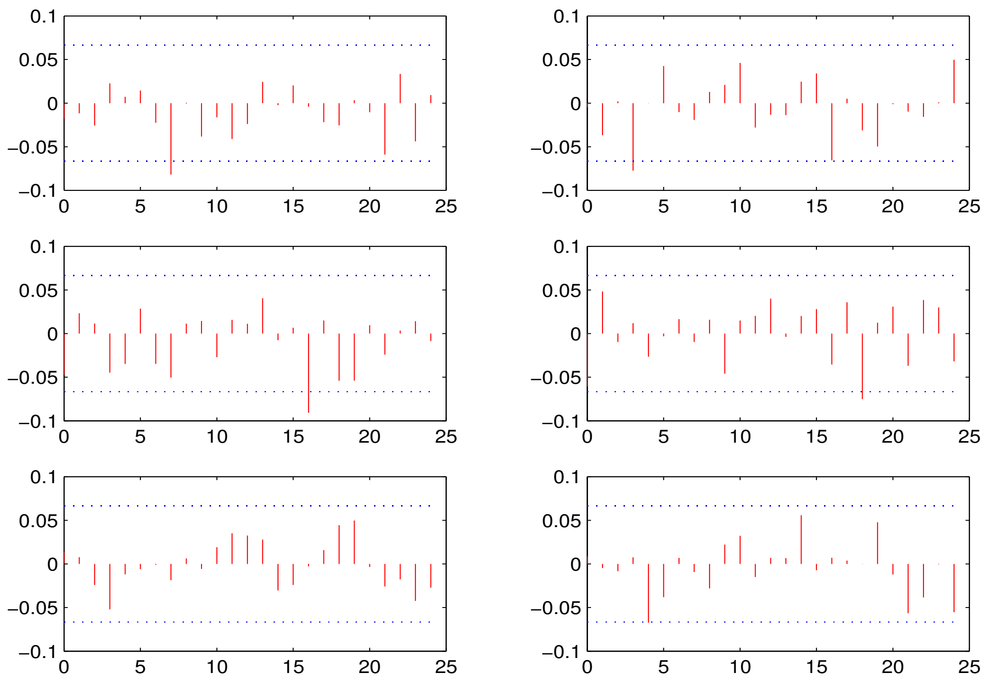

FIG. 5.11 – Exemple 1 : Les fonctions d'autocorrélation empiriques des erreurs d'estimation basées sur le vrai modèle.

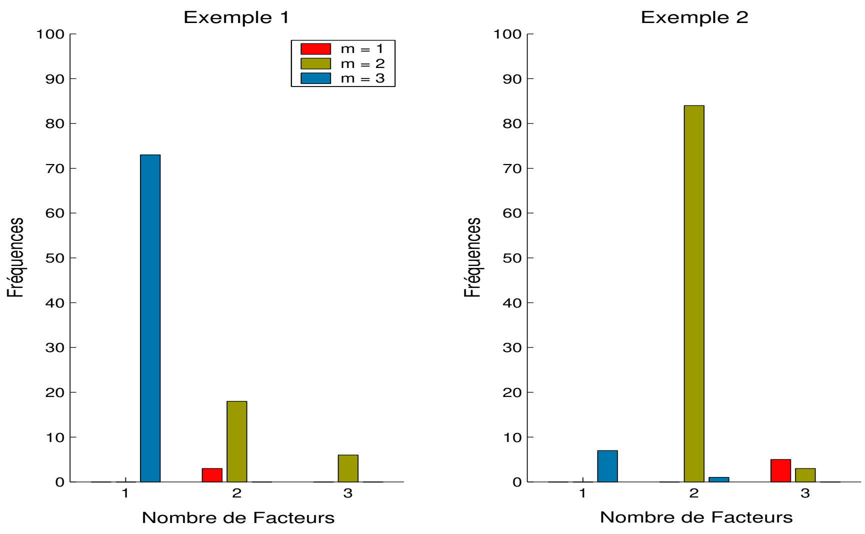

FIG. 5.12 – Les fréquences de choix pour chaque modèle selon le critère BIC.

qui a frappé le système monétaire Européen entre 1992 et 1993. Durant cette période le système monétaire Européen a été fortement perturbé par la violente tourmente qui s'est abattue sur les marchés des changes Européens en septembre et octobre 1992, issue des difficultés de ratification du traité de Maastricht au Danemark et en France. La livre sterling et la lire ont dû quitter le mécanisme de change en septembre 1992 et en novembre de la même année, la peseta et l'escudo ont été dévalués de 6% par rapport aux autres monnaies. En janvier 1993, la livre irlandaise a été dévaluée de 10% ; en mai, la peseta et l'escudo ont subi une nouvelle dévaluation. Enfin, en août 1993, les ministres des Finances ont tiré les conclusions de la crise en portant les marges de fluctuation à 15%.

Quel est l'impact de ces changements sur la nature de la volatilité ? Le degré des comouvements a augmenté ou bien diminué ? Les fluctuations communes sont devenus beaucoup ou moins volatiles ? L'impact de ces crises sur les pays individuels a-t-il évolué au cours du temps ? La réponse à ces questions paraît cruciale pour les dirigeants des politiques économiques et notamment pour les responsables des politiques de change afin de trouver des solutions pour déceler les crises financières avant qu'elles ne se produisent et de protéger, ainsi, l'économie nationale contre l'effet contagion. Le fait de savoir si la volatilité commune a augmenté ou bien diminué, et si les différents pays sont devenus beaucoup ou moins symétriques, permet d'agir au niveau de la politique monétaire par des moyens fiscaux et réglementaires adaptés. La réponse à ces question paraît aussi cruciale pour les chercheurs et les académiques intéressés par les questions de développement économique et l'impact de l'intégration monétaire et financière sur la synchronisation entre taux de change.

5.7.1 Les Données

Les données que nous allons analyser sont les rendements journaliers des cours spot de huit devises cotées en Livres Sterling.[9] Notre base de données contient 601 observations allant de 05/03/1991 jusqu'à 05/07/1993. Les 601 observations ont été transformées afin de calculer des rendements journaliers, ce qui a résulté en la perte de la première observation :

$$r_t = \log p_t - \log p_{t-1} \approx \frac{p_t - p_{t-1}}{p_{t-1}}$$

où p_t est le cours de change journalier à la date t (cours de cloture). Cette quantité peut être considérée comme le logarithme du taux de croissance géométrique, connu en finance sous l'appellation rendement composé continu. La représentation graphique des différentes séries et leurs rendements considérées dans l'ordre : Dollar Américain (USD), Dollar Canadien (CAD), Franc Français (FRF), Franc Suisse (CHF), Lire Italienne (ITL), Deutsche Mark (DEM), Yen Japonais (JPY), et le Dollar de Hong Kong (HKD) est donnée dans la figure 5.13.

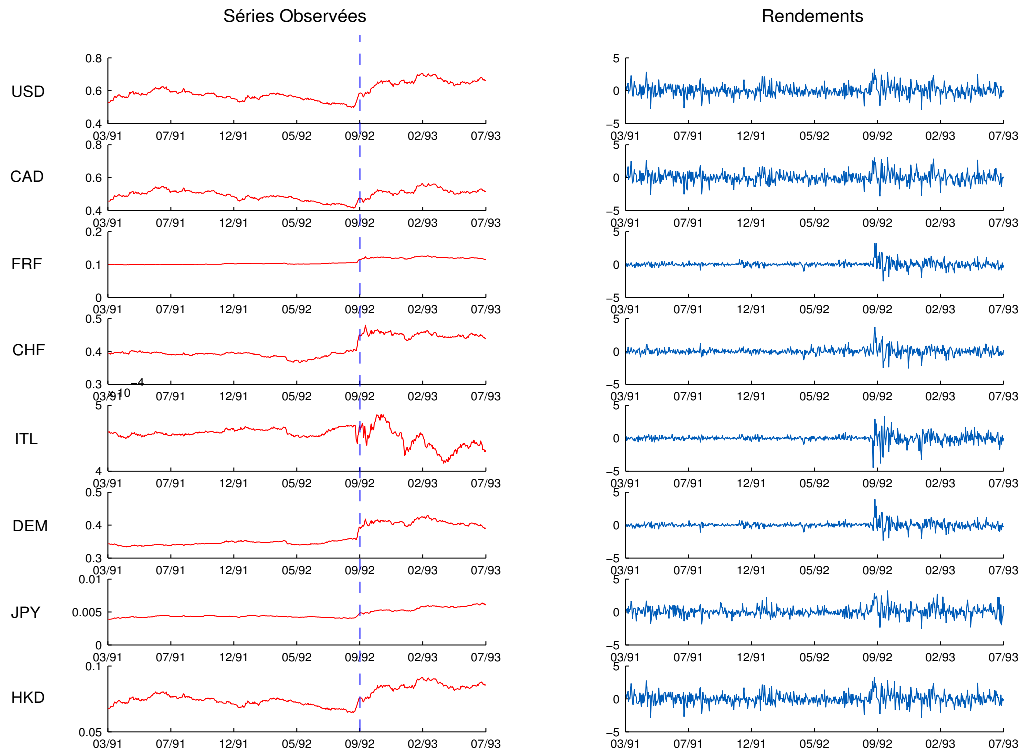

FIG. 5.13 – Cours journaliers des taux de change et leurs rendements allant du 05/03/1991 jusqu'à 05/07/1993 (600 observations). La ligne verticale représente la date $t^* = 31/08/1992$.

[9] PACIFIC EXCHANGE RATE SERVICE, Sauder School of Business, http ://fx.sauder.ubc.ca/. Pour les cours nous avons utilisé la notation en valeurs.

TAB. 5.9 – Caractéristiques statistiques des séries de rendement entre 05/03/1991 et 05/07/1993. Q_1 et Q_3 désignent, respectivement, le premier et le troisième quartile. BJ est le test de normalité basé sur la skewness et la kurtosis suivant une distribution de Chi-deux avec deux degrés de liberté. LB(12) est de test de Ljung et Box estimé pour une corrélation sérielle d'ordre 12 calculé sur les carrées des rendements.

Statistique	USD	CAD	FRF	CHF	ITL	DEM	JPY	HKD
Moyennes	0.0385	0.0205	0.0224	0.0179	-0.0110	0.0213	0.0758	0.0393
écarts-types	0.8236	0.8290	0.4651	0.5431	0.6038	0.4743	0.7315	0.8262
Skewness	0.2991	0.3298	0.9279	0.5182	-1.0302	1.0804	0.4624	0.2946
Kurtosis	4.4333	4.6104	13.4776	10.2500	15.8776	15.4744	4.9780	4.5005
Test BJ 10^3	0.0601	0.0754	2.8212	1.3365	4.2378	3.9936	0.1188	0.0647
Maximum	3.2860	3.0359	3.2270	3.6562	3.3113	3.9079	3.2273	3.2676
Q_3	0.5021	0.4692	0.1946	0.2507	0.1893	0.1824	0.4159	0.4956
Médiane	0	-0.0140	0.0005	-0.0087	-0.0098	0.0029	0.0147	0.0060
Q_1	-0.4648	-0.4691	-0.1534	-0.2465	-0.1824	-0.1693	-0.3103	-0.4460
Minimum	-2.8506	-2.8345	-2.5251	-2.5592	-4.4431	-2.3295	-2.5374	-2.8564
LB(12)	33.916	35.982	54.356	37.223	37.125	58.206	48.727	30.562

5.7.2 Analyse Exploratoire

Dans la table 5.9 sont représentées les différentes caractéristiques statistiques des séries étudiées sur la période couvrant les années 1991 à 1993. Afin de tester l'hypothèse de normalité de la distribution de ces séries, la skewness et la kurtosis ont été ajoutées. Les résultats obtenus doivent en principe se rapprocher des hypothèses couramment émises dans la théorie financière, à savoir que les cours doivent être des variables aléatoires indépendantes et identiquement distribuées. D'autre part, leur distribution n'est pas normale mais plutôt leptokurtique et asymétrique. Dans notre cas, la skewness de chacune des séries est proche de zéro alors que la kurtosis est très grande. L'hypothèse de normalité a été aussi rejetée par le test de Bera et Jarque [1982] (test BJ). L'examen de chacune de ces séries dans la figure 5.13 ne montre pas la présence d'une corrélation sérielle significative, mais il paraît qu'il y a une persistence au niveau de la variance conditionnelle.

Pour mieux caractériser la structure des données et fixer un cadre d'analyse pour une étude plus approfondie basée sur des modèles avec changement de régime et à facteurs conditionnellement hétéroscédastiques, nous avons développé tout d'abord une analyse exploratoire traditionnelle. Des modèles à facteurs standards ont été estimés sur cette base de données en considérant $k = 1$, 2 et 3 facteurs communs. Tous les résultats sont donnés dans le tableau 5.10. Ces estimations ne tiennent pas en compte la structure de dépendance dynamique qui caractérise généralement les séries de nature économique ou financière. Cependant, elles dégagent certains résultats intéressants que nous présentons ci-dessous.

1. Les pondérations associées au premier facteur (première colonne de la matrice des pondérations) ont essentiellement la même structure lorsque un modèle avec un, deux ou trois facteurs est estimé sur cette base de données.
2. Le troisième facteur paraît moins important que les autres.

TAB. 5.10 – Modèles à facteurs standards avec différents nombres de facteurs.

Nombre de facteurs	θ	$diag(\Psi)$	$\mathbf{X}$		
	0.0385	0.0015	0.9229		
	0.0205	0.0634	0.8861		
	0.0224	0.1917	0.1748		
$k=1$	0.0179	0.2767	0.1497		
	-0.0110	0.3453	0.1535		
	0.0213	0.2076	0.1467		
	0.0758	0.2633	0.5844		
	0.0393	0.0043	0.9239		
	0.0385	0.0003	0.9876	0.0000	
	0.0205	0.0648	0.9070	0.0074	
	0.0224	0.0129	0.7809	1.4996	
$k=2$	0.0179	0.1037	0.7037	1.4434	
	-0.0110	0.2313	0.4536	1.1920	
	0.0213	0.0011	0.8289	1.7703	
	0.0758	0.2449	0.7096	0.1256	
	0.0393	0.0058	0.9938	0.0212	
	0.0385	0.0001	0.9967	0.0000	0.0000
	0.0205	0.0643	0.9612	0.0312	0.0000
	0.0224	0.0124	0.8923	1.6482	0.1207
$k=3$	0.0179	0.0996	0.8538	1.6833	0.2337
	-0.0110	0.1669	0.7279	1.1453	0.0364
	0.0213	0.0002	0.9163	1.7929	0.1072
	0.0758	0.2445	0.8571	0.5501	0.0841
	0.0393	0.0057	0.9998	0.0192	0.0015

3. L'analyse de la skewness et de la kurtosis conduit aux conclusions usuelles dans les études des cours boursiers. Elles sont différentes de 0 et 3, ce qui signifie que la distribution n'est pas normale mais plutôt asymétrique avec des queues épaisses caractérisant une distribution Leptokurtique.
4. L'hypothèse de variables indépendantes est aussi rejetée, car la statistique LB de Ljung-box [1978] calculée avec 12 retards indique des autocorrélations au niveau des carrées des rendements. Ce résultat est interprété par Bollerslev [1987] comme un signe de présence d'hétéroscédasticité dynamique et du phénomène de regroupement (clustering) des volatilités. La figure 5.14 nous montre aussi une dépendance des carrés des rendements entre eux, traduites par des autocorrélations significatives durables pour toutes les séries. Ceci nous conduit au rejet de l'hypothèse d'absence d'autocorrélation des cours et met en évidence la présence d'une hétéroscédasticité, confirmée également par le test ARCH.

5.7.3 Analyse à Facteurs Dynamiques

Dans cette section, une série d'ajustements a été réalisée par des modèles à facteurs standards et conditionnellement hétéroscédastiques (avec et sans changement de régime). Pour ce faire nous avons utilisé les 600 observations des rendements des taux de change allant du 05/03/1991 jusqu'à 05/07/1993. La meilleure stratégie d'initialisation consiste à commencer avec une classification des données à l'aide de la segmentation des observations dans les différents états obtenus grâce au critère de Viterbi,

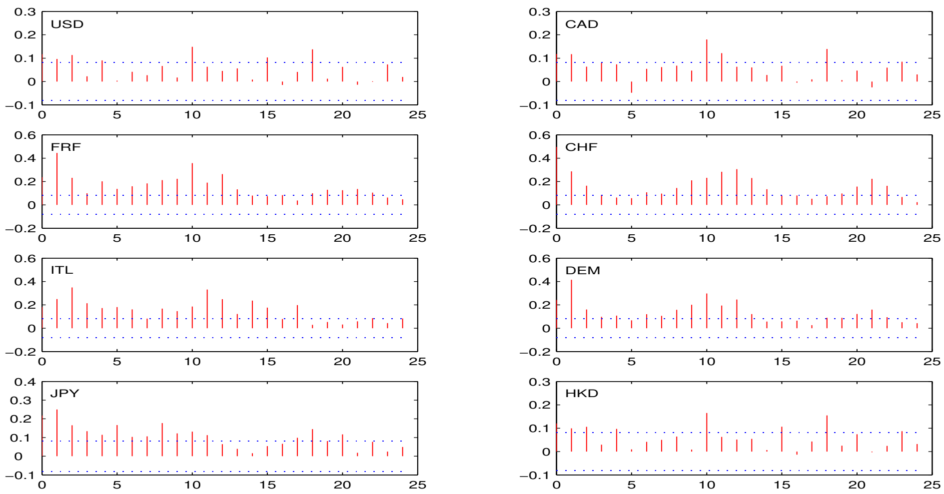

FIG. 5.14 – Autocorrélations empiriques des carrés des rendements basées sur les données allant du 05/03/1991 jusqu'à 05/07/1993.

qui permet de mesurer une distance donc une similitude entre 2 trajectoires. Dans le cas de notre modèle, une classification satisfaisante peut être menée en estimant dans un premier temps un modèle à facteurs standards (on suppose que les facteurs sont homoscédastiques)[10]. Par la suite, nous pouvons soit implémenter un algorithme de Viterbi, soit utiliser directement les probabilités a posteriori $p(S_t = j/\mathcal{Y})$, $j = 1, ..., m$ fournies par l'output EM afin d'obtenir la séquence d'états optimale. Dans une deuxième étape, des modèles à facteurs conditionnellement hétéroscédastiques simples seront initialisés pour chaque segment de données. Pour ce faire, on utilise la matrice de covariance empirique comme estimation de la matrice des variances idiosyncratiques $\mathbf{\Psi}_j$ et la moyenne empirique comme estimation de la moyenne θ_j. Les paramètres de la variance conditionnellement hétéroscédastiques seront initialisés en appliquant un modèle GQARCH(1,1) sur chaque segment de données. Finalement, l'initialisation de la matrice des probabilités de transition $\mathbf{P}$ sera menée en divisant le nombre de transitions de l'état i à l'état j, $(i, j = 1, ..., m)$, par le nombre de transitions de l'état i à n'importe quel autre état.

Afin d'identifier le nombre de facteurs communs et d'états Markoviens, permettant de mieux décrire la structure cachée des données, nous avons estimé différents modèles, en supposant que chaque variable d'état (continue ou discrète) prend une valeur de 1 à 3. Les critères AIC et BIC présentés dans le chapitre 2 ont été utilisés, par la suite, pour choisir la structure convenable. Pour la simplicité, nous avons supposé que les coefficients des spécifications GQARCH ne varient pas avec le régime.

Afin d'identifier la première série avec le premier facteur et pour assurer l'identification du modèle et l'existence d'une solution unique, certaines restrictions supplémentaires doivent être imposées sur les pondérations. Ces contraintes ont été déjà étudiées dans le chapitre 2. En effet, la matrice $\mathbf{X}_t$ doit être de plein rang k, $\forall\ t$, ce qui nous per-

[10] En pratique, une vingtaine d'itérations de l'algorithme EM est largement suffisante.

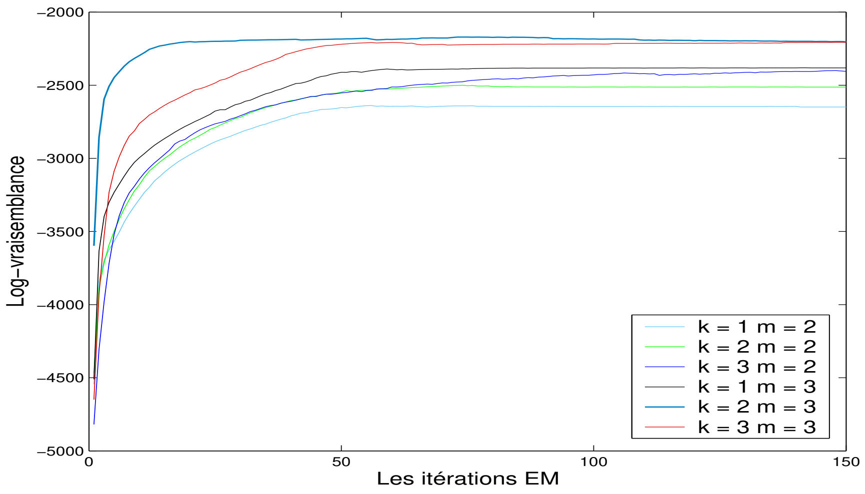

FIG. 5.15 – log-vraisemblances des différentes spécifications avec hétéroscédasticité conditionnelle.

met d'éviter les problèmes d'identification liées à l'invariance du modèle suite à une transformation orthogonale de la matrice des pondérations (voir, par exemple, Geweke et Singleton [1980]). Une contrainte de parcimonie doit, aussi, être imposée sur les pondération afin d'éviter les problèmes de sur-paramétrisation – le nombre des paramètres libres à une date t quelconque ne doit pas dépasser $q(q+1)/2$ (paramètres libres de $\mathbf{\Sigma}_t$). Finalement, l'invariance des vecteurs de facteurs communs sous des transformations linéaires inversibles, doit aussi être assurée (Press [1985], chapitre 10). Dans ce sens, notre travail suivra celui de Geweke et Zhou [1996], en adoptant des contraintes "hiérarchiques" sur la structure des pondérations ayant la forme suivante :

$$\mathbf{X}_j = \begin{pmatrix} x_{11j} & 0 & 0 & \dots & 0 \\ x_{21j} & x_{22j} & 0 & \dots & 0 \\ x_{31j} & x_{32j} & x_{33j} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1j} & x_{k2j} & x_{k3j} & \dots & x_{kkj} \\ x_{k+1,1j} & x_{k+1,2j} & x_{k+1,3j} & \dots & x_{k+1,kj} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{q1j} & x_{q2j} & x_{q3j} & \dots & x_{qkj} \end{pmatrix}$$

où $x_{i,ij} > 0$ pour $i = 1, ..., k$; $j = 1..., m$ et $x_{i,lj} = 0$ pour $i < l$, $i, l = 1, ..., k$. Cette forme permet de garantir directement une matrice $\mathbf{X}_j$ de plein rang k et permet aussi d'identifier le premier facteur avec la première série.

Le tableau 5.11 donne les résultats de tous les critères de sélection, aussi bien que la log-vraisemblance de chaque spécification. La figure 5.15 représente l'évolution de la fonction log-vraisemblance de chacune des spécifications considérées durant les itérations EM. Tous ces résultats indiquent qu'une spécification avec 2 facteurs et 3 états Markoviens permet de mieux représenter la structure des données. Les résultats

TAB. 5.11 – Valeurs de la log-vraisemblance et des critères AIC et BIC pour les différentes spécifications estimées sur la période 05/03/91 – 05/07/93.

Critère	$k = 1$	$k = 2$	$k = 3$
		$m = 1$	
log-vraisemblance (-)	3904.5 (3921.2)	4321.6 (3755.9)	4279.4 (3759.6)
AIC	7865.0 (7890.3)	8699.2 (7575.9)	8638.8 (7599.2)
BIC	7988.1 (7995.8)	8822.3 (7716.6)	8814.7 (7775.1)
		$m = 2$	
	2648.7 (2660.4)	2512.6 (2531.6)	2404.7 (2506.3)
	5413.5 (5428.8)	5145.1 (5203.2)	4961.3 (5184.6)
	5668.5 (5666.2)	5408.9 (5511.0)	5295.5 (5562.8)
		$m = 3$	
	2381.2 (2400.7)	**2202.7** (2225.7)	2207.5 (2253.4)
	4938.4 (4969.3)	**4631.4** (4667.4)	4685.0 (4722.8)
	5325.3 (5338.7)	**5128.3** (5142.2)	5278.6 (5197.7)

. Modèles conditionnellement hétéroscédastiques, (.) Modèles standards

de l'analyse empirique avec $k = 2$ et 3 facteurs conditionnellement hétéroscédastiques dans une structure Markovienne à 3 états cachés sont, respectivement, donnés dans les tableaux 5.12 et 5.13 et les figures 5.16 jusqu'à 5.23. Dans le premier cas (où $k = 2$), la matrice des probabilités de transition et le vecteur des probabilités de l'état initial sont donnés par :

$$\mathbf{P} = \begin{bmatrix} 0.9773 & 0.0227 & 0.0000 \\ 0.0000 & 0.9698 & 0.0302 \\ 0.0834 & 0.2478 & 0.6688 \end{bmatrix} \quad \text{et} \quad \pi = \begin{bmatrix} 1.0000 \\ 0.0000 \\ 0.0000 \end{bmatrix}$$

En utilisant cette spécification à deux facteurs, la figure 5.16 montre comment le modèle permet une bonne reconstruction des changements brusques qui ont touché la série des rendements DEM, et en particulier la violente tourmente qui s'est abattue sur les marchés des changes européens en septembre et octobre 1992. La figure 5.17 montre clairement que la troisième variable d'état correspond aux périodes de forte volatilité, la deuxième variable correspond à la période avant août 1992, et la première à la période caractérisée par une volatilité plus ou moins faible qui vient juste après octobre 1992. la figure 5.16 nous montre aussi que le temps de séjour moyen écoulé dans le premier régime est d'environ 37.8 semaines contre 76 dans le deuxième et 6.2 dans le troisième. D'autres résultats intéressants que nous résumons ci-dessous ont pu aussi être obtenus par cette analyse.

1. La figure 5.18 montre que la variabilité et la dynamique des devises Européennes, FRF, CHF, ITL et DEM sont, essentiellement, expliquées par le deuxième facteur. La figure 5.19 montre que la contribution du troisième facteur à l'ajustement du modèle est très faible, voir même négligeable.
2. La dynamique des variances communes (figures 5.20 et 5.21) montre que les deux premiers facteurs ont un pouvoir explicatif plus important que celui du troisième. La figure 5.17 montre des changements remarquables et une volatilité de plus

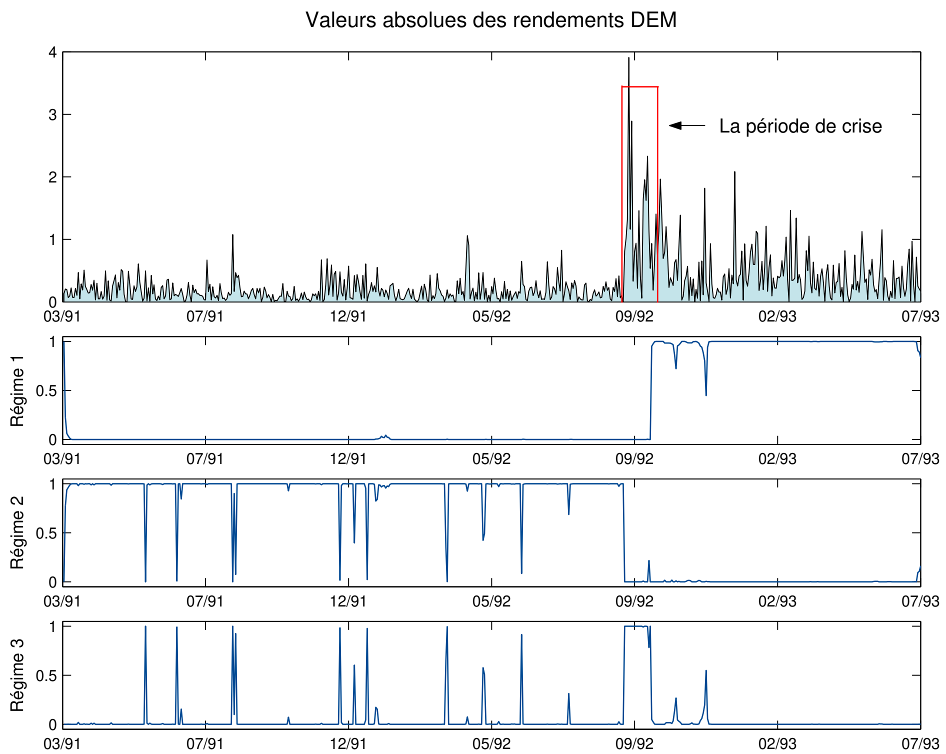

FIG. 5.16 – *Graphiques 2,3,4 :* Probabilités a posteriori des états cachés $M_{t/n}(j)$ données par l'algorithme de lissage.

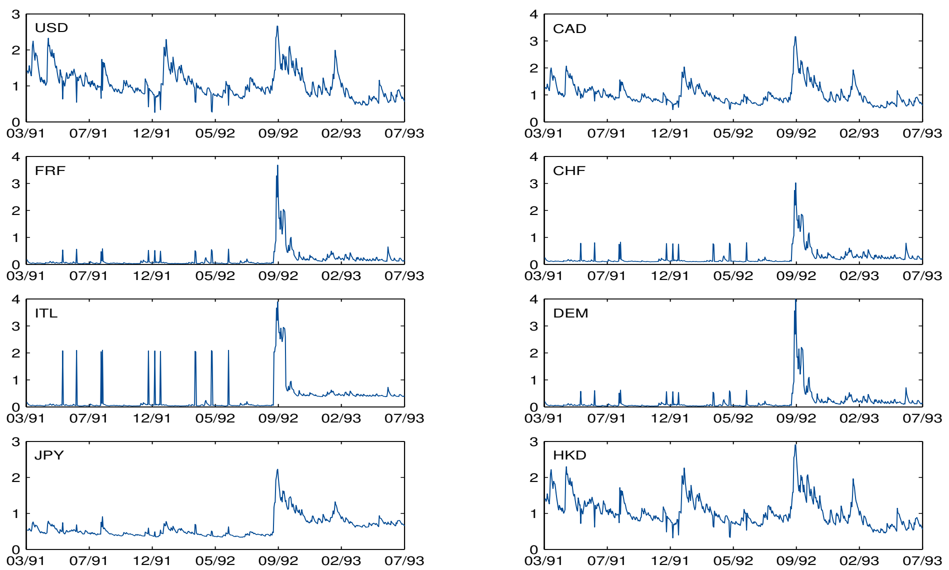

FIG. 5.17 – Volatilités des différentes séries en utilisant une spécification à 3 états et 2 facteurs conditionnellement hétéroscédastiques.

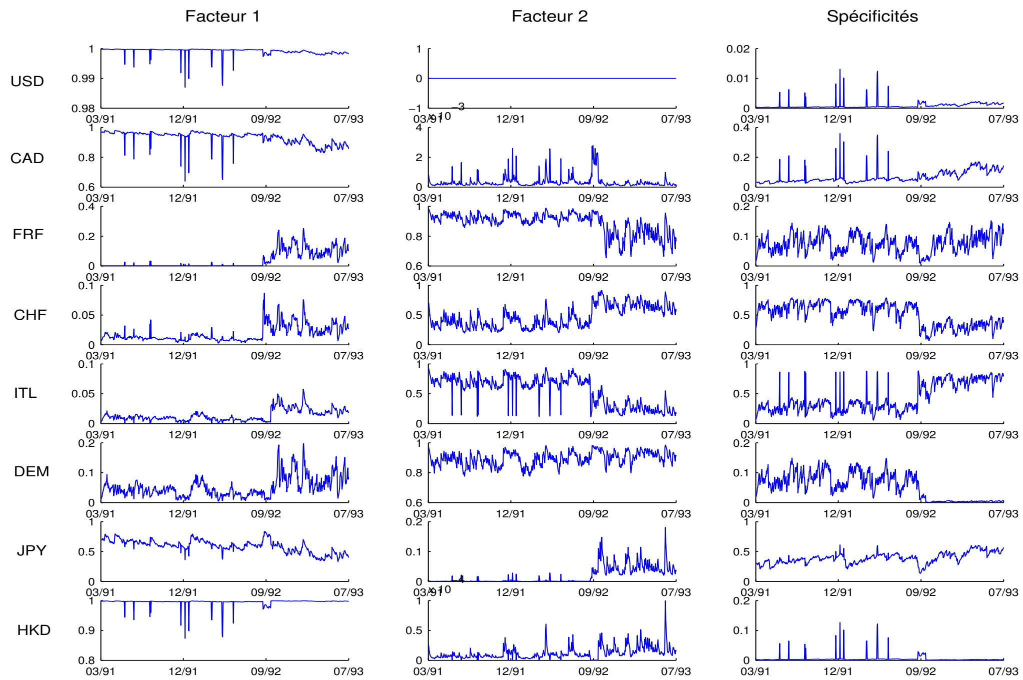

FIG. 5.18 – Modèle à deux facteurs : Proportion de la variance de chacune des séries expliquée par les trois facteurs (communs et spécifique), sur la période allant du 05/03/1991 jusqu'à 05/07/1993.

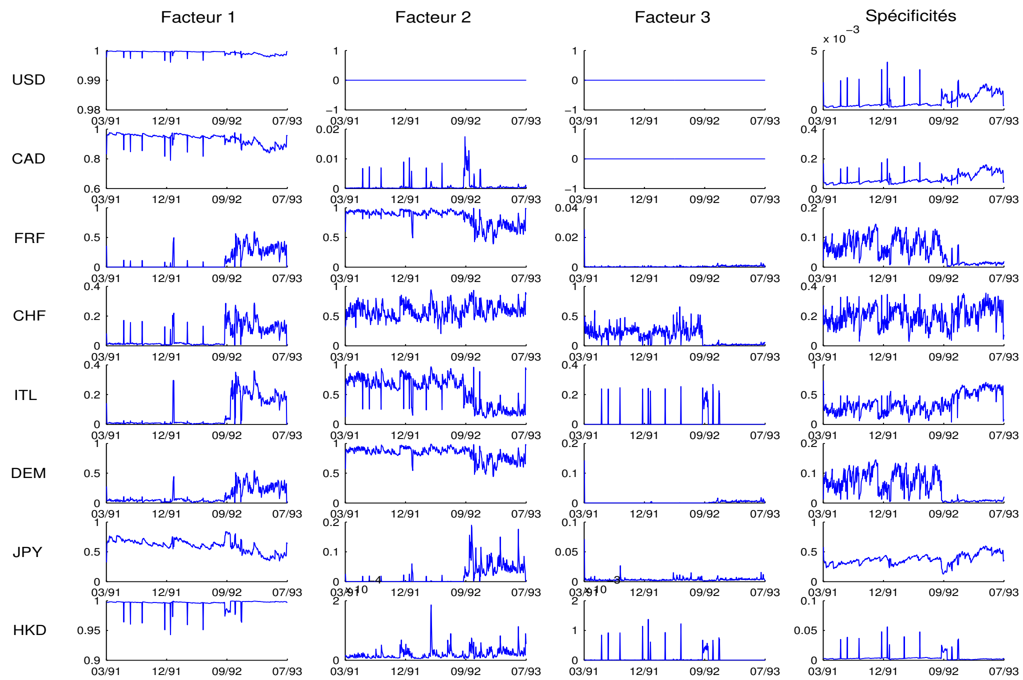

FIG. 5.19 – Modèle à trois facteurs : Proportion de la variance de chacune des séries expliquée par les facteurs (communs et spécifique), sur la période allant du 05/03/1991 jusqu'à 05/07/1993.

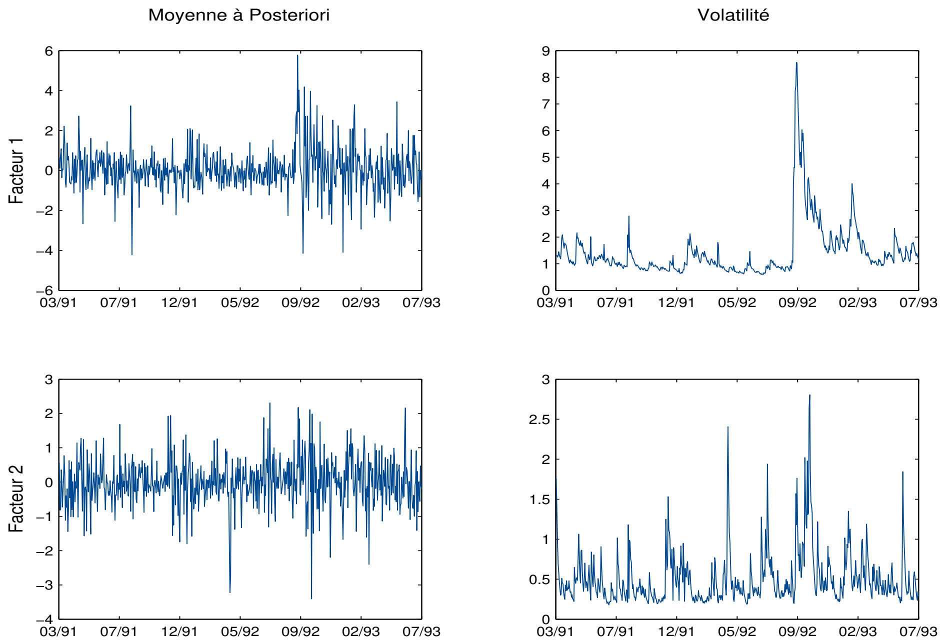

FIG. 5.20 – Modèle à deux facteurs : Moyenne des facteurs communs et leurs volatilités, les éléments de la diagonale de $\mathbf{H}_t$ (du 05/03/1991 jusqu'à 05/07/1993).

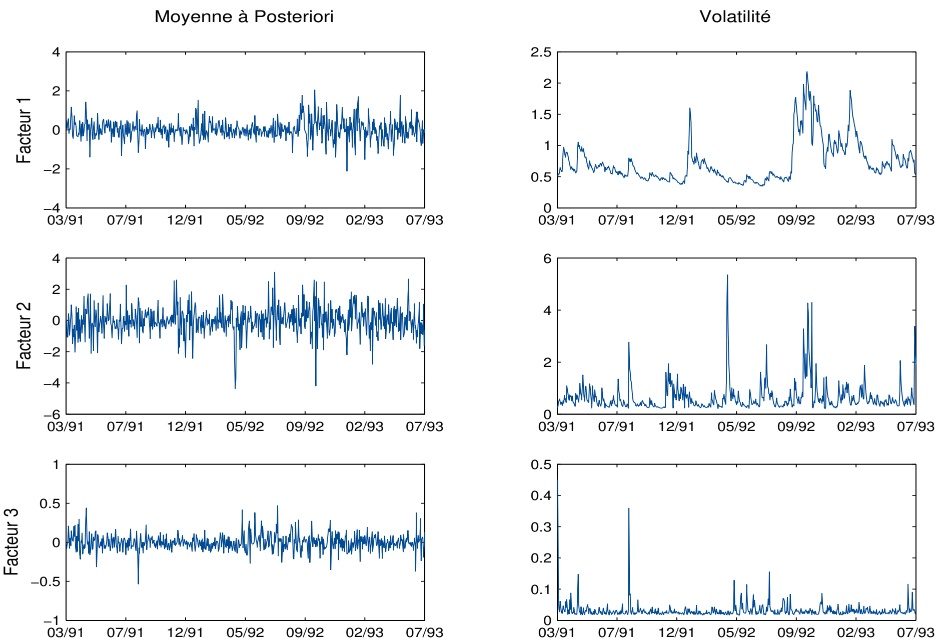

FIG. 5.21 – Modèle à trois facteurs : Moyenne des facteurs communs et leurs volatilités, les éléments de la diagonale de $\mathbf{H}_t$ (du 05/03/1991 jusqu'à 05/07/1993).

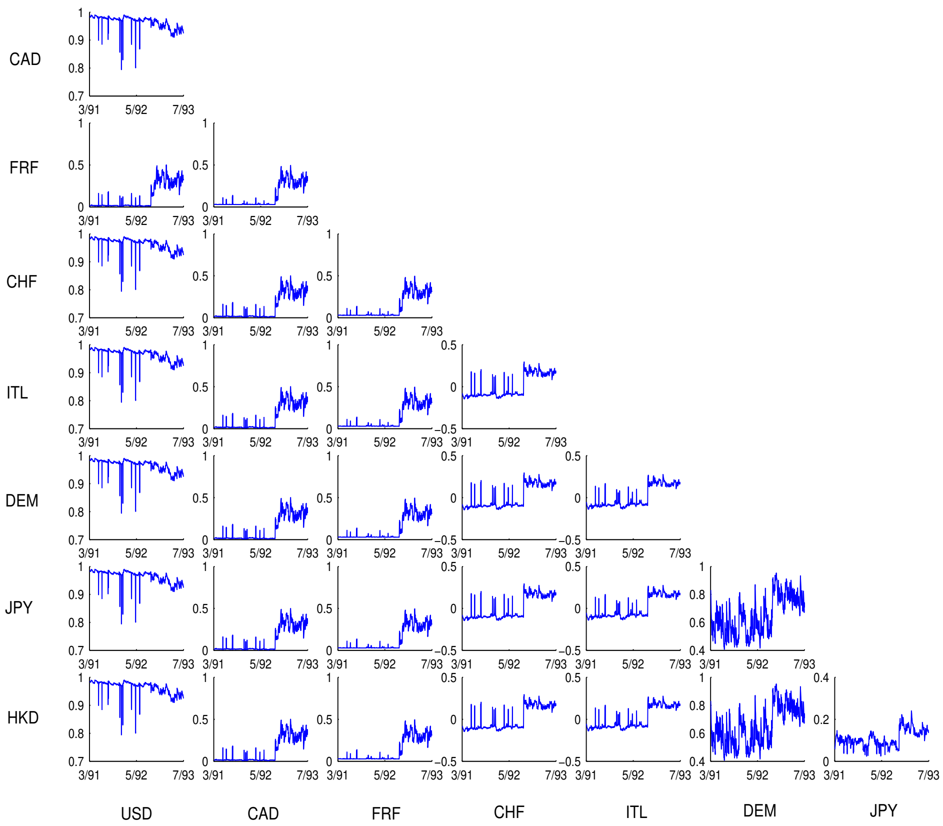

FIG. 5.22 – Modèle à deux facteurs : Structure de co-dépendance de chacune des séries pour la période 05/03/1991 jusqu'à 05/07/1993.

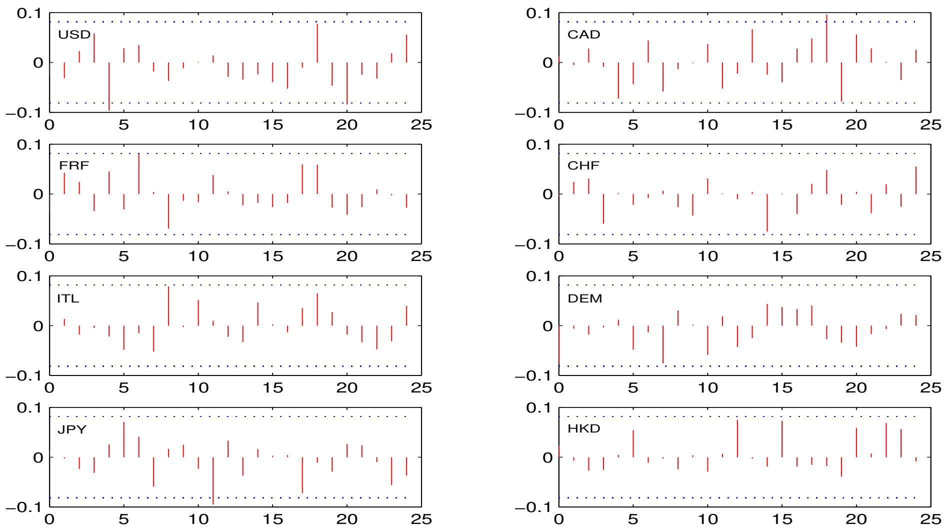

FIG. 5.23 – Modèle à deux facteurs : Fonctions d'autocorrélations des résidus.

TAB. 5.12 – Modèle à deux facteurs conditionnellement hétéroscédastiques

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$	ϕ_1	ϕ_2
	0.0127	0.7044	0.0000	0.0010	0.0860	0.0826
	-0.0082	0.6778	0.0178	0.0869	0.1071	0.1504
	-0.0414	0.1107	0.5790	0.0157	0.0826	0.1919
État 1	-0.0467	0.0693	0.6167	0.0817	0.8742	0.6332
	-0.0548	0.0790	0.4661	0.3194		
	-0.0457	0.0953	0.6159	0.0007		
	0.0915	0.4798	0.2731	0.3692		
	0.0120	0.6997	-0.0054	0.0015		
	-0.0035	1.0375	0.0000	0.0004	0.0860	0.0826
	-0.0159	0.9688	0.0237	0.0402	0.1071	0.1504
	0.0188	0.0030	0.3176	0.0031	0.0826	0.1919
État 2	0.0116	-0.0347	0.3163	0.0725	0.8742	0.6332
	0.0131	0.0211	0.3039	0.0134		
	0.0212	-0.0419	0.3289	0.0037		
	0.0129	0.5367	0.0277	0.1599		
	-0.0009	1.0302	0.0044	0.0032		
	0.0492	0.5576	0.0000	0.0034	0.0860	0.0826
	0.0042	0.5914	-0.0706	0.1615	0.1071	0.1504
	0.0290	0.0833	1.4265	0.0318	0.0826	0.1919
État 3	0.0562	0.1109	1.1904	0.4202	0.8742	0.6332
	-0.3031	0.0384	1.0897	1.7904		
	0.0067	0.0654	1.4910	0.0348		
	0.1477	0.4625	0.2418	0.2974		
	0.0315	0.5788	0.0009	0.0400		

en plus forte vers la fin de 1992 suite à la spéculation enclenchée par le résultat négatif du premier référendum danois (juin 1992) et les incertitudes qui ont entouré le référendum français (septembre 1992) et qui ont engendré des turbulences monétaires spéculatives et ont obligé en fin de compte les autorités italiennes et britanniques à retirer leurs monnaies du mécanisme de change Européen. L'impact de cet événement paraît évident à travers les trajectoires estimées de la volatilité des différentes séries aussi bien que celle des facteurs communs.

3. Notons aussi que l'impact des changements de la volatilité vers la fin de 1992 sur les facteurs communs ont renforcé le besoin d'utiliser une spécification conditionnellement hétéroscédastique pour la modélisation des facteurs et une structure Markovienne pour les paramètres du modèle. Pour ces deux facteurs la somme de α_i et δ_i estimés est proche de un. Ceci indique la présence d'un effet GARCH fort et une persistence au niveau de la volatilité des taux de change.
4. La figure 5.18 montre que le premier facteur explique au moins 95% de la variance des devises USD, CAD et HKD pour toute la période (et au moins 99% avant la crise de 1992 pour le USD et le HKD). Ce facteur explique aussi 70% de la variance de la monnaie Japonaise avant août 1992 et 50% après cette date. La contribution du deuxième facteur dans l'explication de la variance de ces devises est pratiquement négligeable, à l'exception du JPY où la contribution est au tour de 10% après août 1992. La dynamique de la variance de cette devise est en fait expliquée à raison de 50% (au plus) par la composante idiosyncratique qui lui est associée après la crise de 1992.

TAB. 5.13 – Modèle à trois facteurs conditionnellement hétéroscédastiques

	θ	$\mathbf{X}$			$diag(\mathbf{\Psi})$	ϕ_1	ϕ_2	ϕ_2
	0.0226	0.6805	0.0000	0.0000	0.0011	0.0692	0.5184	0.0788
	0.0001	0.6541	0.0090	0.0000	0.0875	0.0825	0.1122	0.1312
	-0.0379	0.2243	0.2386	-0.0894	0.0032	0.0605	0.2956	0.1586
État 1	-0.0381	0.1490	0.2396	0.4421	0.0732	0.9090	0.5884	0.1866
	-0.0514	0.2467	0.1897	0.0148	0.3175			
	-0.0396	0.2054	0.2434	0.2207	0.0018			
	0.1041	0.4463	0.1039	0.3130	0.3588			
	0.0219	0.6764	-0.0024	0.0008	0.0013			
	-0.0033	0.9921	0.0000	0.0000	0.0004	0.0692	0.5184	0.0788
	-0.0176	0.9312	0.0076	0.0000	0.0415	0.0825	0.1122	0.1312
	0.0202	0.0039	0.1181	0.0173	0.0032	0.0605	0.2956	0.1586
État 2	0.0106	-0.0260	0.1130	0.7365	0.0125	0.9090	0.5884	0.1866
	0.0151	0.0185	0.1139	-0.0002	0.0135			
	0.0229	-0.0400	0.1224	0.0121	0.0037			
	0.0110	0.5178	0.0059	0.2120	0.1573			
	-0.0006	0.9847	0.0022	-0.0155	0.0032			
	0.1452	0.8805	0.0000	0.0000	0.0024	0.0692	0.5184	0.0788
	0.1197	0.9175	-0.0775	0.0000	0.1673	0.0825	0.1122	0.1312
	0.0855	0.2714	0.6790	0.1739	0.0578	0.0605	0.2956	0.1586
État 3	0.1267	0.3729	0.6019	-0.6542	0.3393	0.9090	0.5884	0.1866
	-0.2990	0.1667	0.5007	3.4690	0.6658			
	0.0571	0.2660	0.7232	0.0800	0.0101			
	0.2352	0.7487	0.1151	0.2919	0.2972			
	0.1249	0.9215	-0.0026	-0.1860	0.0394			

5. Cette figure nous montre aussi que le deuxième facteur explique à peu près 90% de la variance des devises FRF, ITL et DEM avant la crise financière de 1992. La contribution de ce facteur dans l'explication de la variance du CHF est au tour de 80% après août 1992 et 40% avant cette date. En revanche, la contribution du premier facteur dans l'explication de la variance de ces devises Européennes est en particulier negligeable avant août 1992.

D'une manière générale, les résultats montrent que toutes les corrélations entre les devises Européennes ont augmenté juste après août 1992 (figure 5.22). Une telle augmentation est due à ce qu'on appelle l'effet contagion et qui se traduit par une augmentation significative des co-mouvements des prix (tels que les taux de change, taux d'intérêt, prix des actifs,...) et des quantités à travers des marchés, suite à une crise se produisant dans un marché ou un groupe de marchés. Ce phénomène peut être expliqué par les liens financiers, économiques et politiques des différents pays. Sur le plan économique, ces liens sont généralement représentés par le commerce international. Lorsque deux pays sont en concurrence sur un marché étranger, une dévaluation du taux de change dans l'un des deux pays va détériorer l'avantage comparatif de l'autre. Par conséquent, ces pays finissent par dévaluer afin de relancer leurs exportations. le premier facteur représente la valeur de la Livre Sterling relativement à un panier de devises dans lequel le HKD, USD et CAD sont dominants. Le tableau 5.12 montre que le USD, CAD et HKD ont approximativement le même poids, ceci est dû au fait que la détermination du cours du Dollar Canadien et du Dollar de Hong Kong sur les marchés internationaux est fortement liée à celle du Dollar Américain. Ce premier facteur peut, donc, être considéré comme un facteur purement Nord Américain. Le deuxième facteur pourra, aussi, être considéré comme un facteur spécifique aux pays de la communauté économique Européenne. Il représente un panier restreint de devises dominées par les monnaies de l'Union Européenne, avec un poids relativement faible du Yen Japonais. La part de la variabilité totale expliquée par ce facteur est pratiquement négligeable pour le Dollar Américain, le Dollar Canadien et le Dollar de Hong Kong. Pour ces trois devises les pondérations $x_{1,2j}$, $x_{2,2j}$ et $x_{8,2j}$ sont plus ou moins faibles $\forall j = 1, 2, 3$. Le calcul de la part de la variance totale expliquée par les facteurs spécifiques donne des valeurs très faibles pour les États-Unis et l'Allemagne. Un tel résultat montre, donc, le rôle fondamental joué par les monnaies de ces deux pays dans la détermination de leurs secteurs de facteurs. Nous remarquons aussi que le Franc Français et la Lire Italienne ont les parts de variances spécifiques les plus grandes (durant la période de crise), ce qui indique leur éloignement de leurs secteurs de facteurs. Finalement, la représentation graphique des fonctions d'autocorrélations empiriques des erreurs d'estimation basées sur le modèle à deux facteurs conditionnellement hétéroscédastiques et trois états Markoviens (figure 5.23) montre l'absence de corrélation. Le test de Ljung-Box ne rejette pas, aussi, l'hypothèse nulle d'absence de corrélation sérielle au niveau des résidus. Enfin, l'application d'un test ARCH sur les séries résiduelles montre que celles-ci ne présentent pas un phénomène d'hétéroscédasticité conditionnelle. Nous pouvons, ainsi, affirmer que toute la corrélation entre les rendements des taux de change est complètement expliquée par les facteurs communs et spécifiques. Par conséquent, l'utilisation de notre modèle semble bien justifiée malgré les approximations que nous avons effectué au niveau des calculs.

5.8 Conclusion

Dans ce chapitre nous avons développé une nouvelle approche dans le cadre des modèles d'évaluation des actifs financiers permettant de tenir compte de deux aspects fondamentaux qui caractérisent la volatilité financière : co-mouvement des rendements conditionnellement hétéroscédastiques et transition entre différents régimes inobservables. En combinant les modèles à facteurs latents avec les modèles de chaîne de Markov cachés nous avons abouti à un modèle multivarié localement linéaire et dynamique pour la segmentation et la prévision des séries financières conditionnellement hétéroscédastiques. En particulier, nous avons considéré le cas où les facteurs communs suivent des processus GQARCH univariés.

L'inférence des structures cachées et l'estimation des paramètres ont été aussi discuté en adoptant deux approches différentes fondées sur le principe de l'algorithme EM généralisé et l'approximation du modèle par une spécification espace-état multi-régime. La première approche est basée sur une méthode pseudo-bayésienne généralisée, et la deuxième sur une approximation de Viterbi. La précision de cet algorithme a été illustrée par une étude sur des données simulées. En utilisant deux critères d'information basés sur la vraisemblance, nous avons démontré que cette méthode est capable de discriminer correctement les différentes classes de volatilité. L'analyse du jeu de données réelles a aussi confirmé la pertinence de cette nouvelle approche.

Une étude rigoureuse des propriétés statistiques de la méthode proposée semble difficile, étant donné que le processus étudié n'est pas généralement Markovien et homogène dans le temps. Cependant, notre modèle semble être pertinent pour l'étude des processus localement homogènes. Les idées présentées dans ce document peuvent ainsi être appliquées dans plusieurs domaines de recherche, faisant intervenir des techniques de réduction de la dimension de l'espace des paramètres, afin de pouvoir en obtenir une représentation factorielle dans un cadre non linéaire permettant de conserver la majeure partie de l'information analysée. Dans des recherches futures, il serait intéressant d'élargir notre modèle en introduisant une certaine dynamique au niveau des variances spécifiques. Nous pouvons aussi considérer le cas où les probabilités de transition ne sont pas homogènes, mais dépendent des états passés ou bien de certaines variables observées.

Les applications potentielles de ces modèles relèvent essentiellement du domaine financier, et notamment les analyses reposant sur les études d'événements en vue de tester l'efficience informationnelle des marchés. Ces modèles peuvent aussi être appliqués pour la prévision et fournir, ainsi, un instrument d'aide à la décision pour la gestion et la construction séquentielle de portefeuilles d'actifs qui nécessite la connaissance du rendement moyen, du risque de chaque classe d'actifs, aussi bien que du degré de corrélation existant entre chaque paire d'actifs.

Annexe : Optimisation des Paramètres

Le schéma d'optimisation des paramètres du modèle à facteurs conditionnellement hétéroscédastiques et à structure Markovienne cachée, basé sur l'algorithme EM, sera présenté dans cet annexe. Toutes les statistiques exhaustives seront évaluées en utilisant les paramètres de l'itération précédente $\Theta^{(i)}$. Nous supposons ici que le premier état discret est toujours l'état initial, et que tous les états sont émetteurs. La généralisation de cet algorithme pour le cas d'états non-émetteurs est directe.

1- Mise à jour des probabilités de l'état initial

En éliminant tous les termes qui ne sont pas liés directement aux probabilités de l'état initial π_j dans l'espérance conditionnelle de la log-vraisemblance complétée (5.31), la fonction auxiliaire qu'on cherche à maximiser sera donnée par :

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \sum_{j=1}^{m} M_{1/n}(j) \log(p(S_1))$$

Si on suppose que la chaîne a commencé à l'état j, l'utilisation du multiplicateur de Lagrange λ, sous la contrainte $\sum\limits_{j=1}^{m} \pi_j = 1$, où $\pi_j = p(S_1 = j)$ et $p(S_1) = \pi = [\pi_1, \pi_2, ..., \pi_m]'$, nous conduit à la maximisation de la fonction suivante :

$$g(\pi_j) = \sum_{i=1}^{m} M_{1/n}(i) \log(\pi_j) + \lambda \left(1 - \sum_{i=1}^{m} \pi_i\right)$$

Les dérivées par rapport à $g(\pi_j)$ seront données par :

$$\begin{cases} \frac{\partial g(\pi_j)}{\partial \pi_j} = \frac{M_{1/n}(j)}{\pi_j} - \lambda \\ \frac{\partial g(\pi_j)}{\partial \lambda} = 1 - \sum\limits_{i=1}^{m} \pi_i \end{cases}$$

La résolution des conditions du premier ordre nous donne la nouvelle valeur $\widehat{\pi}_j$ qui maximise la fonction $g(\pi_j)$ étant donné que la dérivée seconde de cette dernière en ce point est négative.

$$\widehat{\pi}_j = \frac{M_{1/n}(j)}{\sum\limits_{i=1}^{m} M_{1/n}(i)}$$

2- Mise à jour des probabilités de transition

Dans ce cas la fonction auxiliaire qu'on cherche à maximiser par rapport aux p_{ij} sous la contrainte $\sum_{j=1}^{m} p_{ij} = 1$ est donnée par :

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \sum_{t=2}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} M_{t-1,t/n}(i,j) \log(p_{ij})$$

L'utilisation du multiplicateur de Lagrange λ, nous conduit à la maximisation de la fonction suivante :

$$g(p_{ij}) = \lambda(1 - \sum_{j=1}^{m} p_{ij}) + \sum_{t=2}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} M_{t-1,t/n}(i,j) \log(p_{ij})$$

La différenciation de $g(p_{ij})$ donne :

$$\frac{\partial g(p_{ij})}{\partial p_{ij}} = -\lambda + \sum_{t=2}^{n} \frac{M_{t-1,t/n}(i,j)}{p_{ij}}$$

Les conditions du premier ordre seront donc données par :

$$\left\{ \begin{array}{r} -\lambda + \sum_{t=2}^{n} \frac{M_{t-1,t/n}(i,j)}{p_{ij}} = 0 \\ 1 - \sum_{j=1}^{m} p_{ij} = 0 \end{array} \right.$$

Enfin, la résolution de ces conditions nous permettra de trouver les nouvelles probabilités de transition, soient

$$\widehat{p}_{ij} = \frac{\sum_{t=2}^{n} M_{t-1,t/n}(i,j)}{\sum_{t=2}^{n} M_{t-1/n}(i)}$$

ceci est un maximum de $g(p_{ij})$ étant donné que la dérivée seconde de cette dernière en ce point est négative.

3- Mise à jour des Matrices de Pondérations

Soit $\mathbf{x}_{jl}$ le *l-ème* vecteur ligne de $\mathbf{X}_j$. La maximisation de l'équation (5.31) est équivalente à la maximisation de

$$g(\mathbf{x}_{jl}) = -\frac{1}{2} \sum_{l=1}^{q} \left[\mathbf{x}_{ji} \mathbf{G}_{jl} \mathbf{x}'_{jl} - \mathbf{x}_{jl} \mathbf{k}_{jl} \right]$$

où les matrices $\mathbf{G}_{jl}$ de dimension $(k \times k)$ et les vecteurs colonnes $\mathbf{k}_{jl}$ de dimension $(k \times 1)$ sont définis par :

$$\begin{aligned}\mathbf{G}_{jl} &= \frac{1}{\psi_{jl}} \sum_{t=1}^{n} M_{t/n}(j) \left[\mathbf{H}^{j}_{t/n} + \mathbf{f}^{j}_{t/n}\mathbf{f}^{j\prime}_{t/n}\right] \\ \mathbf{k}_{jl} &= \frac{1}{\psi_{jl}} \sum_{t=1}^{n} M_{t/n}(j)(y_{tl} - \theta_{jl})\mathbf{f}^{j}_{t/n}\end{aligned}$$

ici ψ_{jl} représente le *l-ème* élément de la diagonale de la matrice des variances idiosyncratiques $\mathbf{\Psi}_j$; y_{tl} et θ_{jl} sont, respectivement, les *l-èmes* éléments du vecteur d'observations à la date t, $\mathbf{y}_t$ et du vecteur des moyennes spécifiques θ_j.

La différenciation de $g(\mathbf{x}_{jl})$ donne

$$\frac{\partial g(\mathbf{x}_{jl})}{\partial \mathbf{x}_{jl}} = -\mathbf{G}_{jl}\mathbf{x}'_{jl} + \mathbf{k}_{jl}$$

La résolution des conditions du premier ordre nous permettra de trouver la nouvelle valeur de $\mathbf{x}_{jl}$, soit

$$\widehat{\mathbf{x}}_{jl} = \mathbf{k}'_{jl}\mathbf{G}^{-1}_{jl}$$

Une telle valeur maximise la fonction g étant donné que la dérivée de cette dernière en ce point est négative.

4- Mise à jour des Moyennes θ_j

La dérivée de la fonction auxiliaire (5.31) par rapport à θ_j donne :

$$\frac{\partial \mathcal{Q}(\Theta, \Theta^{(i)})}{\partial \theta_j} = \mathbf{\Psi}_j^{-1} \sum_{t=1}^{n} M_{t/n}(j) \left(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}^{j}_{t/n} - \theta_j\right)$$

et la résolution des conditions du premier ordre donne :

$$\widehat{\theta}_j = \frac{\sum_{t=1}^{n} M_{t/n}(j) \left(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}^{j}_{t/n}\right)}{\sum_{t=1}^{n} M_{t/n}(j)}$$

Ceci est un maximum étant donné que la dérivée seconde en ce point est négative.

5- Mise à jour des Variances Idiosyncratiques

En éliminant les termes qui ne dépendent pas directement de la matrice $\mathbf{\Psi}_j$, la fonction auxiliaire (5.31) peut être écrite sous la forme suivante :

$$\mathcal{Q}(\Theta,\Theta^{(i)}) = -\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m} M_{t/n}(j)\left(\log|\mathbf{\Psi}_j| + tr\left\{\mathbf{\Psi}_j^{-1}(\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{f}_{t/n}^{j}\mathbf{y}_t' \\ \mathbf{y}_t' \end{bmatrix}\right.\right.$$
$$\left.\left. - \begin{bmatrix} \mathbf{y}_t\mathbf{f}_{t/n}^{j\prime} & \mathbf{y}_t \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix} + \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{H}_{t/n}^{j} + \mathbf{f}_{t/n}^{j}\mathbf{f}_{t/n}^{j\prime} & \mathbf{f}_{t/n}^{j} \\ \mathbf{f}_{t/n}^{j\prime} & 1 \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix}\right\}\right)$$

Afin de trouver les nouvelles variances idiosyncratiques, la fonction auxiliaire ci-dessus sera maximisée par rapport à l'inverse de $\mathbf{\Psi}_j$. La résolution des conditions du premier ordre et l'annulation des éléments hors diagonale nous donnent la nouvelle valeur de $\mathbf{\Psi}_j$ suivante :

$$\widehat{\mathbf{\Psi}}_j = \frac{1}{\sum\limits_{t=1}^{n} M_{t/n}(j)} \sum_{t=1}^{n} M_{t/n}(j) diag\left\{\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{f}_{t/n}^{j}\mathbf{y}_t' \\ \mathbf{y}_t' \end{bmatrix} - \begin{bmatrix} \mathbf{y}_t\mathbf{f}_{t/n}^{j\prime} & \mathbf{y}_t \end{bmatrix}\right.$$
$$\left. \times \begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix} + \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{H}_{t/n}^{j} + \mathbf{f}_{t/n}^{j}\mathbf{f}_{t/n}^{j\prime} & \mathbf{f}_{t/n}^{j} \\ \mathbf{f}_{t/n}^{j\prime} & 1 \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix}\right\}$$

La dérivée seconde de $\mathcal{Q}(\Theta/\Theta^{(i)})$ par rapport à cette matrice est aussi négative.

Bibliographie

[1] Aggarval R., Inclan, C. et Leal, R. (1999). Volatility in emerging markets. *Journal of Financial and Quantitative Analysis* **34** (1), 33–55.

[2] Aguilar, O. et West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics* **18** (3), 338–357.

[3] Akaike, H. (1974). A new look at the statististical identification model. *IEEE Transactions on Automatic Control* **19** (6), 716–723.

[4] Anderson, B.O. et Moore, J.B. (1979). Optimal Filtering, Englewood Cliffs, NJ : Prentice Hall.

[5] Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis, Third Edition. Wiley Series in Probability and Statistics, Series Volume 107-338.

[6] Baillie, R., et Bollerslev T. (1991). Intraday and Intermarket Volatility in Foreign Exchange Rates. *Review of Economic Studies* **58** (3), 565–585.

[7] Bargmann, R.A. (1957). A study of independence and dependence in multivariate normal analysis. University of North Carolina, Institute of Statistics Mimeo Series N° 186.

[8] Bar-Shalom, Y., et Li, X-R. (1993). Estimation and Tracking : Principles, Techniques and Software. Artech House.

[9] Bartholomew, D. (1987). Latent Variable Models and Factor Analysis. Charles Griffin & Co. Ltd, London.

[10] Bauer F.L., et Reinsch C. (1971). Inversion of Positive Definite Matrices by the Gauss-Jordan Method. in Wilkinson, J.H. and Reinsch, C. eds., *Handbook for Automatic Computation* vol. **2** : Linear Algebra, Springer-Verlag, Berlin.

[11] Baum, L.E., et Eagon, J.A. (1967). An inequality with application to statistical estimation for probabilistic function of Markov processes and to a model for ecology. *Bulletin of the American Mathematicians Society* **73** (3), 360–363.

[12] Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities* **3** (1),1–8.

[13] Bera A. K. et Jarque C. M. (1982). Model specification tests : A simultaneous approach. *Journal of Econometrics* **20** (1) 59–82.

[14] Billio, M., Monfort, A., et Robert, C. (1998). The simulated likelihood method. Technical report DT-9821, CREST, INSEE, Paris.

[15] Bishop, C. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

[16] Black, F., et Scholes M.S. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* **81** (3), 637–654.

[17] Blanchard, O., et Watson, M. (1982). Bubbles, Rational Expectations, and Financial Markets. in Paul Wachtel, ed., Crises in the Economic and Financial Structure (Lexington Books), 295–315.

[18] Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31** (3), 307–327.

[19] Bollerslev T. (1987). A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics* **69** (3), 542–547.

[20] Bollerslev, T., Engle, R., et Wooldridge, M. (1988). A Capital Asset Pricing Model with Time-varying Covariances. *Journal of Political Economy* **96** (1), 116–131.

[21] Bollerslev T., et Wooldridge J.M. (1992). Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Variances. *Econometric Reviews* **11** (2), 143–172.

[22] Bollerslev, T., et Engle, R. (1994). Common Persistence in Conditional Variances. *Econometrica* **61** (1), 167–186.

[23] Boyen, X., et Koller, D. (1998). Tractable inference for complex stochastic processes. *Proceedings of the 14-th Conference on Uncertainty in Artificial Intelligence*, 33–42.

[24] Bozdogan, H., et Ramirez, D.E. (1987). An Expert Model Selection Approach to Determine the "Best" Pattern Structure in Factor Analysis Models. *Multivariate Statistical Modeling and Data Analysis* (eds H. Bozdogan and A.K. Gupta).

[25] Bozdogan, H., et Shigemasu, K. (1998). Bayesian factor analysis model and choosing the number of factors using a new informational complexity criterion. *Technical Report, Department of Statistics*, University of Tennessee.

[26] Brown, S., et Weinstein, M. (1983). A new approach to testing asset pricing models : The bilinear paradigm. *Journal of Finance* **38** (3), 711–743.

[27] Burnham, K.P., et Anderson, D.R. (1998). Model Selection and Inference. Springer-Verlag.

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

[29] Carter, C.K., et Kohn, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika* **83** (3), 589–601.

[30] Cecchetti, S.G., Lam, P-S., et Mark, N.C. (1990). Mean Reversion in Equilibrium Asset Prices. *American Economic Review* **80** (3), 398–418.

[31] Chow, G. (1960). Test of equality between sets of coefficients in two linear regressions. *Econometrica* **28** (3), 591–605.

[32] Clark, P. (1973). Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica* **41** (1), 135–156.

[33] Cutler, D., Poterba, J., et Summers, L.H. (1991). Speculative Dynamics. *Review of Economic Studies* **58** (3), 529–46.

[34] Demos A., et Parissi S. (1998). Testing Asset Pricing Models : The case of the Athens Stock Exchange. *Multinational Finance Journal* **2** (3), 189–223.

[35] Demos A., et Sentana E. (1998). An EM Algorithm for Conditionally Heteroscedastic Factor Models. *Journal of Business & Economic Statistics* **16** (3), 357–361.

[36] Dempster A., Laird N., et Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B* **39** (1), 1–38.

[37] Diebold F., et Nerlove M. (1989). The Dynamics of Exchange Rate Volatility : A Multivariate Latent Factor ARCH Model. *Journal of Applied Econometrics* **4** (1), 1–21.

[38] Doucet, A., et Andrieu, C. (2001). Iterative algorithms for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing* **49** (6), 1216–1227.

[39] Emmett, W.G. (1949). Factor analysis by Lawley's method of maximum likelihood. *British Journal of Psychology, Statistical Section* **2** (1), 90–97.

[40] Engle, R. et Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association* **76** (376), 774–781.

[41] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** (4), 987–1006.

[42] Engle, R. (1987). Multivariate ARCH with factor structures : cointegration in variance. Unpublished working paper, University of California at San Diego.

[43] Engle, R., Lilien, R.M., et Robins, R.P. (1987). Estimating Time Varying Risk Premia in the Term Structure : The ARCH-M Model. *Econometrica* **55** (2), 391–407.

[44] Engle R., Ng, V., et Rothschild, M. (1990). Asset Pricing with a Factor-ARCH Structure : Empirical Estimates for Treasury Bills. *Journal of Econometrics* **45**, (1-2) 213–237.

[45] Engle R., et Ng, V. (1993). Time Varying Volatility and the Dynamic Behavior of the Term Structure. *Journal of Money Credit and Banking* **25** (3) 336–349.

[46] Engle R., et Susmel, R. (1993). Common Volatility in International Equity Markets. *Journal of Business & Economic Statistics* **11** (2), 167–176.

[47] Ephraim, Y., et Merhav, N. (2002). Hidden Markov Processes. *IEEE Transactions on Information Theory* **48** (6), 1518–1569.

[48] Everitt B.S. et Dunn G. (1991). Covariance Structure Models. In Applied Multivariate Data Analysis. Edward Arnold, London.

[49] Fama, E.F., et MacBeth J.D. (1973). Risk, return, and equilibrium : empirical tests. *Journal of Political Economy* **81** (3), 607–636.

[50] Fiorentini, G., Sentana, E., et Shephard, N. (2004). Likelihood-Based Estimation of Latent Generalized ARCH Structures. *Econometrica* **72** (5), 1481–1517.

[51] Forni, M., Hallin, M., Lippi, M., et Reichlin, L. (2004). The generalized dynamic factor model : consistency and rates. *Journal of Econometrics* **119** (2), 231–255.

[52] Fredkin, D.R., et Rice, J.A. (1992). Bayesian Restoration of Single-Channel Patch Clamp Recordings. *Biometrics* **48** (2), 427–448.

[53] French, K.R., Schwert, G.W., et Stambaugh, R.F. (1987). Expected Stock Returns and Volatility. *Journal of Financial Economics* **19** (1), 3–29.

[54] Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **38** (6), 897–906.

[55] Geweke, J. (1977). The dynamic factor analysis of economic time series models. In D.J. Aigner and A.S. Goldberger (eds.), Latent Variables in Socio-economic Models, pp. 365–383. North Holland, Amsterdam.

[56] Geweke J.F., et Singleton K.J. (1980). Interpreting the Likelihood Ratio Statistic in Factor Models when Sample Size is Small. *Journal of the American Statistical Association* **75** (369), 133–137.

[57] Geweke, J., et Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies* **9** (2), 557–587.

[58] Ghahramani, Z. et Hinton, G.E. (1996). Parameter estimation for linear dynamical systems. University of Toronto Technical Report, CRG-TR-96-2.

[59] Ghahramani, Z., et Hinton, G.E. (2000). Variational learning for switching state-space models. *Neural Computation* **12** (4), 963–996.

[60] Gouriéroux, C., Monfort, A., et Renault, E. (1995). Inference in Factor Models. *Advances in Econometrics and Quantitative Economics*, Essays in Honor of C. R. Rao, édité par G. S. Maddala, P.C.B. Phillips et T.N. Srinivasan, Basil Blackwell, 311–353.

[61] Granger, C. (2002). Some Comments on Risk. *Journal of Applied Econometrics* **17** (5), 447–456.

[62] Gray S.F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* **42** (1), 27–62.

[63] Gupta, A.K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika* **39** (3-4), 260–73.

[64] Hamilton, J. (1988). Rational expectations econometric analysis of changes in regime : an investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control* **12** (2-3), 385–423.

[65] Hamilton, J. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **57** (2), 357–384.

[66] Hamilton, J. (1990). Analysis of Time Series Subject to Changes in Regime. *Journal of Econometrics*, **45** (1), 39–70.

[67] Harvey, A. (1989). Forecasting structural time series models and the Kalman filter. Cambridge University Press.

[68] Harvey, A., Ruiz, E., et Sentana, E. (1992). Unobserved component time series models with ARCH disturbances. *Journal of Econometrics* **52** (1-2), 129–157.

[69] He, C., et Teräsvirta, T. (1999). Properties of Moments of a Family of GARCH Processes. *Journal of Econometrics* **92** (1), 173–192.

[70] Heywood H.B. (1931). On Finite Sequences of Real Numbers. *Proceedings of the Royal Society, Series A* **134** 486–510.

[71] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24** (6-7), 417–441 ; 498–520.

[72] Householder A.S. (1964). The Theory of Matrices in Numerical Analysis. Blaisdell Publishing Company, London.

[73] Howe, W.G. (1955). Some contributions to factor analysis. Report N° ORNL-1919, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

[74] Isakov, D. (1999). Is beta still alive ? Conclusive evidence from the swiss stock market. *The European Journal of Finance* **5** (3), 202–212.

[75] Jelinek, F., Bahl, L.R., et Mercer, R.L. (1975). Design of a linguistic statistical decoder for the recognition on continuous speech. *IEEE Transactions on Information Theory* **IT-21** (3), 250–256.

[76] Jensen, M.C. (1972). Capital Markets : Theory and Evidence. *Bell Journal of Economics* **3** (2), 357–398.

[77] Jordan, M.I. (1998). Learning in Graphical Models. The MIT Press.

[78] Jordan, M.I., Ghahramani, Z., Jaakkola, T., et Saul, L. (1999). An introduction to variational methods in graphical models. *Machine Learning* **37** (2), 183–233.

[79] Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32** (4), 443–482.

[80] Jöreskog, K.G. (1969). A General Approach to Confimatory Maximum Likelihood Factor Analysis. *Psychometrika* **34** (2), 183–202.

[81] Jöreskog, K.G., et Sörbom, D. (1988). LISREL-7 : A guide to the program and applications (2nd edition). Chicago : SPSS.

[82] Juang, B.H., et Rabiner, L.R. (1985). A Probabilistic Distance Measure for Hidden Markov Models. *AT&T Technical Journal* **64** (2), 391–408.

[83] Kaiser, T. (1997). Factor-GARCH Models for German Stocks : A Model Comparison. *Operations Research Proceedings*, Springer-Verlag, Berlin u.a.

[84] Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME series D : Journal of Basic Engineering* **82** (1), 35–45.

[85] Kalman, R.E. et Bucy, R.S. (1961). New results in linear filtering and prediction. *Journal of Basic Engineering* **83** (3), 95–108.

[86] Kass, R.E. et Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association* **90** (430), 773–795.

[87] Kelley, T.L. (1928). Crossroads in the Mind of Mind. Stanford : Stanford University Press.

[88] Kim C.J. (1994). Dynamic linear models with Markov switching. *Journal of Econometrics* **60** (1), 1–22.

[89] King, M., Sentana E., et Wadhwani, S. (1994). Volatility and Links between National Stock Markets. *Econometrica* **62** (4), 901–933.

[90] Kroner, K.F. (1987). Estimating and testing for factor GARCH. University of California at San Diego, mimeo.

[91] Kulp. D., Haussler, D., Reese, M.G., et Eeckman, F.H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* **4**, 134–142.

[92] Laird, N., Lange, N., et Stram, D. (1987). Maximum likelihood computations with repeated measures : application of the EM algorithm. *Journal of the American Statistical Association* **82** (397), 97–105.

[93] Lamoureux, C., et Lastrapes, W. (1990). Persistence in Variance, Structural Change, and the GARCH Model. *Journal of Business & Economic Statistics* **8** (2), 225–234.

[94] Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* **5** (1), 1–18.

[95] Lastrapes, W. (1989). Exchange Rate Volatility and U.S. Monetary Policy : An ARCH Application. *Journal of Money, Credit and Banking* **21** (1), 66–77.

[96] Lauritzen S. (1996). Graphical Models. Claredon Press, Oxford, UK.

[97] Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, Section A* **60**, 64–82.

[98] Lawley, D.N. (1942). Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh, Section A* **61**, 176–185.

[99] Lawley, D.N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology* **33**, 172–175.

[100] Lawley, D.N. (1967). Some new results in maximum likelihood factor analysis. *Proceedings of the Royal Society of Edinburgh, Section A* **67**, 256–264.

[101] Lawley, D.N., et Maxwell, A.E. (1971). Factor Analysis as a statistical method, Second Edition. London : Butterworths.

[102] Lee, L.J., Attias, H., Deng, L. (2003). Variational inference and learning for segmental switching state space models of hidden speech dynamics. *Proceedings, IEEE ICASSP* **1**, 920–923.

[103] Lin, W., Engle, R., et Ito, T. (1991). Do Bulls and Bears Move Across Borders ? International Transmission of Stock Returns and Volatility as the World Turns. *NBER Working Papers* 3911, NBER, Inc.

[104] Lin, W. (1992). Alternative estimators for factor GARCH models : a Monte Carlo comparaison. *Journal of Applied Econometrics* **7** (3), 259–279.

[105] Liporace, L.R. (1982). Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory* **IT-28** (5), 729–734.

[106] Liu, C., et Rubin, D.B. (1994). Ehe ECME algorithm : A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81** (4), 633–648.

[107] Liu, C., et Rubin, D.B (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica* **8** (3), 729–747.

[108] Ljung G. et Box G. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika* **67** (2), 297–303.

[109] Lopes, H.F., et West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14** (1), 41–67.

[110] Lord, F.M. (1956). A study of speed factors in tests and academic grades. *Psychometrika* **21** (1), 31–50.

[111] Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44** (2), 226–233.

[112] Magnus J.R., et Neudecker H. (1988). Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester.

[113] Markowitz, H. (1952). Portfolio Selection. *Journal of Finance* **7** (1), 77–91.

[114] Maxwell, E.A. (1961). Recent trends in factor analysis. *Journal of the Royal Statistical Society, Series A* **124** (1), 49–59.

[115] McLachlan, G.J., et Krishnan, T. (1997). The EM Algorithm and Extensions. Wiley series in probability and statistics. John Wiley & Sons.

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

[117] Merton, R.C. (1980). On Estimating the Expected Return on the Market : An Exploratory Investigation. *Journal of Financial Economics* **8** (4), 323–361.

[118] Murphy, K.P.(2002). Dynamic Bayesian Networks : Representation, Inference and Learning. PhD thesis, University of California, Berkeley.

[119] Nelson, D. (1991). Conditional heteroskedasticity in asset returns : A new approach. *Econometrica* **59** (2), 347–370.

[120] Ng, L. (1991). Tests of the CAPM with Time-Varying Covariances : A Multivariate GARCH Approach. *The Journal of Finance* **46** (4), 1507–1521.

[121] Ng, V., Engle, R., et Rothschild, M. (1992). A multi-dynamic factor model for stock returns. *Journal of Econometrics* **52** (1-2), 245–266.

[122] Nijman, T., et Sentana, E. (1996). Marginalization and contemporaneous aggregation in multivariate GARCH processes. *Journal of Econometrics* **71** (1-2), 71–87.

[123] Pavlovic, V., Rehg, J.M., Cham, T-J., et Murphy, K.P. (1999). A dynamic Bayesian network approach to figure tracking using learned dynamic models. *Proceedings of the International Conference on Computer Vision*, 94–101.

[124] Pavlovic, V., Rehg, J.M., et MacCormick, J. (2000). Learning switching linear models of human motion. *Proceedings of the Neural Information Processing Systems Conference*, 981–987.

[125] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

[126] Press S.J. (1985). Applied Multivariate Analysis : Using Bayesian and Frequentist Methods of Inference. *California : Krieger.*

[127] Press, S.J., et Shigemasu, K. (1989). Bayesian inference in factor analysis. In Contributions to Probability and Statistics : Essays in Honor of Ingram Olkin (eds S.J. Press L.J Gleser M.D. Perlman and A.R. Sampson), pp. 271–287. New York, Springer-Verlag.

[128] Quah, D., et Sargent, T. (1993). A dynamic index model for large cross sections. In James H. Stock and Mark W. Watson, Eds, Business Cycles, Indicators, and Forecasting, NBER and University of Chicago Press, Chicago.

[129] Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** (2), 257–286.

[130] Rabiner, L.R. et Juang, B.H. (1993). Fundamentals of Speech Recognition, Englewood Cliffs, NJ : Prentice Hall.

[131] Rao, C.R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika* **20** (2), 93–111.

[132] Rauch, H.E., Tung, F., et Striebel, C.T. (1965). Maximum likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronautics Journal*, **3** (8), 1445–1450.

[133] Roeder, K. et Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association* **92** (439), 894–902.

[134] Roll, R. (1977). A critique of the asset pricing theory's tests ; part I : on past and potential testability of the theory. *Journal of Financial Economics* **4** (2), 129–176.

[135] Ross, S. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* **13** (3), 341–360.

[136] Rosti, A-V.I et Gales M.J.F. (2001). Generalised Linear Gaussian Models. Technical Report CUED/F-INFENG/TR.420, Cambridge University, Engineering Department.

[137] Rosti, A-V.I., et Gales, M.J.F. (2003). Switching Linear Dynamical Systems for Speech Recognition. Technical Report CUED/F-INFENG/TR.461, Cambridge University, Engineering Department.

[138] Rosti, A-V.I., et Gales, M.J.F. (2004). Rao-Blackwellised Gibbs sampling for switching linear dynamical systems. *Proceedings, IEEE ICASSP* **1**, 809–812.

[139] Rubin, D.B., et Thayer, D.T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** (1), 69–76.

[140] Rubin, D.B., et Thayer, D.T. (1983). More on EM for ML factor analysis. *Psychometrika* **48** (2), 253–257.

[141] Saul, L., et Jordan, M.I. (1996). Exploiting tractable substructures in intractable networks. *Proceedings of the Neural Information Processing Systems Conference*, 486–492.

[142] Saul, L., et Rahim, M. (2000). Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing* **8** (2), 115–125.

[143] Schaller, H., et van Norden, S. (1997). Regime Switching in Stock Market Returns. *Applied Financial Economics* **7** (2), 177–191.

[144] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics.* **6** (2), 461–464.

[145] Schwert, G.W. (1990). Stock Volatility and the Crash of 87. *Review of Financial Studies* **3** (1), 77–102.

[146] Sentana, E., et Shah, M., and Wadhwani, S. (1992). Factor representing portfolios in large asset markets. London School of Economics, Discussion paper 193, Financial Markets Group.

[147] Sentana, E., et Shah, M. (1994). An Index of Co-Movements in Financial Time Series. London School of Economics, Discussion paper 193, Financial Markets Group.

[148] Sentana, E. (1994). The Likelihood Function of a Conditionally Heteroskdastic Factor Model with Heywood Cases. Papers 9420, Centro de Estudios Monetarios Y Financieros.

[149] Sentana, E. (1995). Quadratic ARCH models. *Review of Economic Studies* **62** (4), 639–661.

[150] Sentana, E. (1997). Risk and Return in the Spanish Stock Market : Some Evidence from Individual Assets. *Investigaciones Económicas* **21** (2), 297–359.

[151] Sentana, E. (1998). The Relation Between Conditionally Heteroskedastic Factor Models and Factor GARCH Models. *Econometrics Journal* **1** (1), 1–9.

[152] Sentana, E. (2000). The Likehood Function of Conditionally Heteroskedastic Factor Models. *Anales d'économie et de Statistique* **58** (1), 1–19.

[153] Sentana, E. (2002). Did the EMS reduce the cost of capital ? *The Economic Journal* **112** (482), 786–809.

[154] Sentana, E. (2004). Factor representing portfolios in large asset markets. *Journal of Econometrics* **119** (2), 257–289.

[155] Shapiro, S.S., et Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52** (3-4), 591–611.

[156] Shapiro, S.S. et Francia, R.S. (1972). An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association* **67** (337), 215–216.

[157] Sharpe, W.F. (1963). A simplified model for portofolio analysis. *Management Science* **9** (2), 277–293.

[158] Sharpe, W.F. (1964). Capital asset prices : A theory of market equilibrium under conditions of risk. *Journal of Finance* **19** (3), 425-442.

[159] Shumway, R.H., et Stoffer, D.S. (1991). Dynamic linear models with switching. *Journal of the American Statistical Association* **86** (415), 763–769.

[160] Shumway, R.H. et Stoffer, D.S. (2000). Time Series Analysis and Its Applications. Springer, New York.

[161] Smith, A., et Markov, U. (1980). Bayesian detection and estimation of jumps in linear systems. In O. Jacobs, M. Davis, M. Dempster, C. Harris, and P. Parks, editors. Analysis and Optimization of Stochastic Systems.

[162] Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology* **15** (2), 201–293.

[163] Stephens, M.A. (1975). Asymptotic properties for covariance matrices of order statistics. *Biometrika* **62** (1), 23–28.

[164] Stock, J., et Watson, M. (1989). New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual*, Washington, D.C., 351-409.

[165] Stock, J., et Watson, M. (1993). A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems. *Econometrica* **61** (4), 783–820.

[166] Thurstone, L.L. (1931). Multiple factor analysis. *Psychological Review* **38** (4), 406–427.

[167] Treynor, J. (1961). Towards a theory of market value of risky assets, unpublished manuscript.

[168] Viterbi, A.J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Processing*, **13** (2), 260–269.

[169] Watson, M., et Engle, R. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics* **23** (3), 385–400.

[170] West, M., et Harrison, J. (1997). Bayesian Forecasting and Dynamic Models. Second Edition, Springer-Verlag, New York.

[171] Xu, L., et Jordan, M.I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation* **8** (2),129–151.

[172] Young, S.J., Russell, N.H. et Thornton, J.H.S. (1989). Token passing : a simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR.38, Cambridge University Engineering Department.