Université de Montréal

Estimation du modèle GARCH à changement de régimes et son utilité pour quantifier le risque de modèle dans les applications financières en actuariat

par

Maciej Augustyniak

Département de mathématiques et de statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Statistique

décembre 2013

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Estimation du modèle GARCH à changement de régimes et son utilité pour quantifier le risque de modèle dans les applications financières en actuariat

présentée par

Maciej Augustyniak

a été évaluée par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

Manuel Morales

(directeur de recherche)

Mathieu Boudreault

(co-directeur)

Alejandro Murua

(membres du jury)

Alexandru Badescu

(examinateur externe)

Benoît Perron

(représentant du doyen de la FES)

Thèse acceptée le:

5 décembre 2013

SOMMAIRE

Le modèle GARCH à changement de régimes est le fondement de cette thèse. Ce modèle offre de riches dynamiques pour modéliser les données financières en combinant une structure GARCH avec des paramètres qui varient dans le temps. Cette flexibilité donne malheureusement lieu à un problème de *path dependence*, qui a empêché l'estimation du modèle par le maximum de vraisemblance depuis son introduction, il y a déjà près de 20 ans. La première moitié de cette thèse procure une solution à ce problème en développant deux méthodologies permettant de calculer l'estimateur du maximum de vraisemblance du modèle GARCH à changement de régimes. La première technique d'estimation proposée est basée sur l'algorithme Monte Carlo EM et sur l'échantillonnage préférentiel, tandis que la deuxième consiste en la généralisation des approximations du modèle introduites dans les deux dernières décennies, connues sous le nom de *collapsing procedures*. Cette généralisation permet d'établir un lien méthodologique entre ces approximations et le filtre particulaire. La découverte de cette relation est importante, car elle permet de justifier la validité de l'approche dite par *collapsing* pour estimer le modèle GARCH à changement de régimes. La deuxième moitié de cette thèse tire sa motivation de la crise financière de la fin des années 2000 pendant laquelle une mauvaise évaluation des risques au sein de plusieurs compagnies financières a entraîné de nombreux échecs institutionnels. À l'aide d'un large éventail de 78 modèles économétriques, dont plusieurs généralisations du modèle GARCH à changement de régimes, il est démontré que le risque de modèle joue un rôle très important dans l'évaluation et la gestion du risque d'investissement à long terme dans le cadre des fonds distincts. Bien que la littérature financière a dévoué beaucoup de recherche pour faire progresser les modèles économétriques dans le but d'améliorer la tarification et la couverture des produits financiers, les approches permettant de mesurer l'efficacité d'une stratégie de couverture dynamique ont peu évolué. Cette thèse offre une contribution méthodologique dans ce domaine en proposant un cadre statistique, basé sur la régression, permettant de mieux mesurer cette efficacité.

Mots clés : économétrie financière, changement de régimes, GARCH, maximum de vraisemblance, filtre particulaire, algorithme EM, risque de modèle, couverture dynamique, efficacité de la couverture, fonds distincts

SUMMARY

The Markov-switching GARCH model is the foundation of this thesis. This model offers rich dynamics to model financial data by allowing for a GARCH structure with time-varying parameters. This flexibility is unfortunately undermined by a path dependence problem which has prevented maximum likelihood estimation of this model since its introduction, almost 20 years ago. The first half of this thesis provides a solution to this problem by developing two original estimation approaches allowing us to calculate the maximum likelihood estimator of the Markov-switching GARCH model. The first method is based on both the Monte Carlo expectation-maximization algorithm and importance sampling, while the second consists of a generalization of previously proposed approximations of the model, known as collapsing procedures. This generalization establishes a novel relationship in the econometric literature between particle filtering and collapsing procedures. The discovery of this relationship is important because it provides the missing link needed to justify the validity of the collapsing approach for estimating the Markov-switching GARCH model. The second half of this thesis is motivated by the events of the financial crisis of the late 2000s during which numerous institutional failures occurred because risk exposures were inappropriately measured. Using 78 different econometric models, including many generalizations of the Markov-switching GARCH model, it is shown that model risk plays an important role in the measurement and management of long-term investment risk in the context of variable annuities. Although the finance literature has devoted a lot of research into the development of advanced models for improving pricing and hedging performance, the approaches for measuring dynamic hedging effectiveness have evolved little. This thesis offers a methodological contribution in this area by proposing a statistical framework, based on regression analysis, for measuring the effectiveness of dynamic hedges for long-term investment guarantees.

Keywords : financial econometrics, regime-switching, GARCH, maximum likelihood, particle filtering, EM algorithm, model risk, dynamic hedging, hedging effectiveness, variable annuities

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Cette thèse est le fruit d'un effort collectif que je partage avec les membres de ma famille, mes professeurs, mes amis ainsi que plusieurs organismes et fondations encourageant les étudiants.

Premièrement, je tiens à dédier cette thèse à mes parents, Bożena et Krzysztof, car c'est grâce à leur éducation et leur soutien inconditionnel que j'ai pu m'épanouir et réussir tout au long de mon cheminement scolaire. *Dziękuję za wszystko, ta praca doktorska jest dla Was !* Merci à Jacek, mon frère et meilleur ami. Merci à Eden, l'amour de ma vie et meilleure amie.

Je remercie également tous les professeurs qui ont contribué à mon apprentissage dès mon premier jour en classe. Je vous en suis reconnaissant. Plus particulièrement, je remercie mes directeurs de thèse, Mathieu Boudreault et Manuel Morales, deux professeurs sur lesquels j'ai toujours pu compter, qui m'ont encouragé à chacune des étapes de mon doctorat et qui m'ont fourni d'innombrables conseils. Merci également à Louis Doray, mon directeur de maîtrise, avec qui j'ai eu mon premier contact avec la recherche. Merci à Ewa Duma et José Garrido, deux professeurs de l'Université Concordia qui ont été très influents dans ma décision de poursuivre des études supérieures.

Finalement, je tiens à remercier plusieurs organismes qui m'ont offert un soutien financier pendant mon doctorat : Autorité des marchés financiers (subvention de recherche), Conseil de recherches en sciences naturelles et en génie du Canada (Bourse d'études supérieures du Canada Alexander-Graham-Bell), Fonds de recherche du Québec – Nature et technologies (Bourse de doctorat en recherche), Institut de finance mathématique de Montréal (Bourse d'excellence Jean Turmel), Society of Actuaries (Society of Actuaries James C. Hickman Scholar), Université de Montréal (Bourse Serge Tardif).

INTRODUCTION

Cette thèse étudie deux thèmes de recherche et contient cinq chapitres, dont quatre sont sous un format d'article. Deux de ces articles sont publiés dans des revues scientifiques et un troisième est soumis pour publication. Premièrement, cette thèse développe l'estimation du modèle GARCH à changement de régimes, puis, deuxièmement, elle traite de l'importance du risque de modèle dans les applications financières en actuariat, notamment dans les fonds distincts (*variable annuities*). Un fonds distinct est un produit d'investissement vendu par des compagnies d'assurances au Canada. Il est semblable à un fonds commun de placement (*mutual fund*) assorti de garanties protégeant le capital investi contre une baisse des marchés boursiers.

Le modèle GARCH à changement de régimes est le fondement de cette thèse. Il intègre deux approches de modélisation en économétrie financière, soit GARCH et changement de régimes, dans le but d'offrir un meilleur ajustement aux données financières et de mieux refléter leurs dynamiques. Cette combinaison donne malheureusement lieu à un problème de *path dependence* qui complique son estimation. Par exemple, plusieurs auteurs ont affirmé que l'estimation des paramètres du modèle avec le principe du maximum de vraisemblance est très complexe, voire impossible, à effectuer. Ce problème, non résolu depuis l'introduction du modèle il y a déjà près de 20 ans, représente la motivation des trois premiers chapitres de cette thèse où deux approches originales sont développées permettant de calculer l'estimateur du maximum de vraisemblance du modèle GARCH à changement de régimes.

Le chapitre 1 introduit des notions en économétrie financière qui sont nécessaires à la compréhension des deux chapitres suivants. Premièrement, il définit les modèles GARCH et à changement de régimes, ainsi que les techniques statistiques couramment utilisées pour les estimer. Par exemple, le filtre d'Hamilton et l'algorithme espérance-maximisation (EM) sont étudiés. Deuxièmement, des méthodes en statistique computationnelle sont présentées, telles que l'échantillonnage préférentiel, le filtre particulaire et l'échantillonnage de Gibbs. Elles sont utiles aux

chapitres 2 et 3 pour le développement de l'estimation du modèle GARCH à changement de régimes. De plus, il est expliqué comment approximer l'algorithme EM avec des méthodes Monte Carlo dans les situations où il ne peut pas être implanté.

Le chapitre 2 inclut un article dont je suis le seul auteur et qui est publié dans la revue *Computational Statistics & Data Analysis* [9]. Sa contribution principale est de proposer une approche originale, basée sur l'algorithme Monte Carlo EM et sur l'échantillonnage préférentiel, permettant de calculer l'estimateur du maximum de vraisemblance du modèle GARCH à changement de régimes. Une contribution secondaire est de démontrer comment estimer la matrice de variance-covariance asymptotique de cet estimateur. Jusqu'à présent, aucune méthode n'existe pouvant atteindre ces objectifs sans recourir à une simplification du modèle. Par conséquent, la technique d'estimation proposée est la première méthode du maximum de vraisemblance dans le cadre du modèle GARCH à changement de régimes. Son efficacité est démontrée sur des données simulées et empiriques.

Le chapitre 3 contient un article réalisé avec la collaboration des coauteurs Mathieu Boudreault et Manuel Morales. Il a été soumis en octobre 2013 à la revue *Journal of Econometrics*. Sa contribution principale est de démontrer comment les approximations du modèle GARCH à changement de régimes proposées dans les deux dernières décennies, connues sous le nom de *collapsing procedures*, peuvent être généralisées, puis reformulées comme un filtre particulaire déterministe. Ce lien entre le filtre particulaire et l'approche dite par *collapsing* est une découverte importante, car il permet d'expliquer pourquoi cette approche parvient à calculer l'estimateur du maximum de vraisemblance du modèle. Le principal avantage de la méthodologie proposée par rapport aux techniques bayésiennes et à l'algorithme Monte Carlo EM est sa rapidité d'exécution. Elle permet donc de réduire le fossé existant entre l'attrait du modèle GARCH à changement de régimes et sa mise en pratique. La robustesse de l'algorithme est vérifiée avec des simulations et son efficacité est démontrée sur les mêmes jeux de données considérés dans le chapitre 2.

Les deux derniers chapitres développent le deuxième volet de recherche de cette thèse, soit l'importance du risque de modèle dans la gestion des risques en actuariat. Ce thème de recherche tire sa motivation de la crise financière de la fin des années 2000 qui a remis en question les modèles et stratégies de gestion des risques utilisés dans les deux dernières décennies par les institutions financières. En effet, plusieurs institutions ont mal évalué leur exposition aux risques pendant cette crise, résultant en de nombreux échecs institutionnels et causant des répercussions dommageables sur toutes les sphères de l'économie. Les deux derniers chapitres tentent, en premier lieu, de sensibiliser les actuaires à l'importance de

prendre en compte le risque de modèle dans l'évaluation du risque d'investissement à long terme (chapitre 4) et, en second lieu, d'expliquer comment mesurer plus adéquatement l'efficacité des stratégies de couverture dynamique utilisées par les compagnies d'assurances (chapitre 5).

Le chapitre 4 inclut un article accompli avec la collaboration du coauteur Mathieu Boudreault et publié dans la revue *North American Actuarial Journal* [10]. Bien que cet article fasse l'objet du quatrième chapitre de ma thèse, il représente le premier projet sur lequel j'ai travaillé. Il compare 78 modèles économétriques, dont plusieurs généralisations du modèle GARCH à changement de régimes, et examine l'efficacité des stratégies utilisées par les compagnies d'assurances pour gérer les risques liés aux fonds distincts dans le contexte de la crise financière de la fin des années 2000. L'utilisation d'un large éventail de modèles permet de démontrer l'importance que joue le risque de modèle sur l'évaluation et la gestion du risque d'investissement à long terme.

Le chapitre 5 contient un article dont le contenu est basé sur trois projets subventionnés par l'Autorité des marchés financiers et supervisés par Mathieu Boudreault. L'Autorité des marchés financiers est l'organisme mandaté par le gouvernement du Québec pour encadrer les marchés financiers québécois et prêter assistance aux consommateurs de produits et services financiers. La littérature financière a dévoué beaucoup de recherche pour faire progresser les modèles économétriques dans le but d'améliorer la tarification et la couverture des produits financiers. Toutefois, les approches permettant de mesurer l'efficacité d'une stratégie de couverture dynamique ont peu évolué. Le chapitre 5 offre une contribution méthodologique dans ce domaine en proposant un cadre statistique, basé sur la régression, permettant de mieux évaluer cette efficacité. L'importance de prendre en compte le risque de modèle est soulignée. De plus, ce chapitre démontre qu'une garantie d'investissement simple peut être reformulée comme un ensemble d'options à barrière lorsque l'investisseur a l'option de résilier son contrat de fonds distinct. Ceci nous permet de tarifer et gérer le risque de ce produit avec des formules analytiques et, ainsi, étudier l'impact de la déchéance dynamique (*dynamic lapsation*) sur l'efficacité de la couverture.

Chapitre 1

NOTIONS EN ÉCONOMÉTRIE FINANCIÈRE

L'évolution du prix d'un actif financier (par exemple, une action ou un indice boursier) dans le temps est généralement modélisée à l'aide d'un processus stochastique, symbolisé par $\{x_t\}_{t\in A}$, où A est un ensemble d'indices faisant référence au temps et $x_t \in \mathbb{R}_{\geq 0}$ est une variable aléatoire représentant le prix de l'actif au temps t. Si $A = \mathbb{R}$ ou $\mathbb{R}_{\geq 0}$, le processus stochastique est dit à temps continu et si $A = \mathbb{Z}$ ou $\mathbb{N}$, il est dit à temps discret. Cette thèse considère le cas discret où l'évolution de x_{t-1} vers x_t est modélisée par l'entremise d'une variable aléatoire $y_t \in \mathbb{R}$, satisfaisant la relation suivante :

$$x_t = x_{t-1}e^{y_t} \quad \Leftrightarrow \quad y_t = \log x_t - \log x_{t-1}.$$

Cette variable aléatoire représente le rendement composé de manière continue (*continuously compounded return*), aussi appelé le log-rendement (*log-return*), sur l'actif pour la période allant de $t-1$ à t.

De manière très générale, le processus stochastique $\{y_t\}$ peut être modélisé avec une équation prenant la forme :

$$y_t = \mu_t + \epsilon_t = \mu_t + \sigma_t \eta_t, \tag{1.0.1}$$

où $\{\eta_t\}$ est une suite de variables aléatoires indépendantes et identiquement distribuées de moyenne zéro et de variance un. On suppose généralement que $\{y_t\}$ est stationnaire au second ordre [1] (*second-order stationary*) et qu'on peut donc écrire $\mathrm{E}[y_t] = \mu$ et $\mathrm{Var}[y_t] = \sigma^2$, $\forall t$, où $\mu \in (-\infty, \infty)$ et $\sigma \in (0, \infty)$ sont des constantes. Les symboles μ_t et σ_t dénotent des fonctions mesurables par rapport

1. Le processus $\{y_t\}$ est dit stationnaire au second ordre si :

(i) $\mathrm{E}[y_t^2] < \infty, \quad \forall t \in \mathbb{Z}$,

(ii) $\mathrm{E}[y_t] = \mu, \quad \forall t \in \mathbb{Z}$,

(iii) $\mathrm{Cov}[y_{t+h}, y_t] = \mathrm{E}[(y_{t+h} - \mu)(y_t - \mu)] = \gamma(h), \quad \forall t, h \in \mathbb{Z}$.

La fonction $\gamma(\cdot)$ est appelée la fonction d'autocovariance de $\{y_t\}$.

à une tribu engendrée par le vecteur aléatoire $y_{1:t-1} = (y_1, \ldots, y_{t-1})$. Par conséquent, la moyenne conditionnelle du rendement au temps t est $\mathrm{E}[y_t \mid y_{1:t-1}] = \mu_t$ et sa variance conditionnelle est $\mathrm{Var}[y_t \mid y_{1:t-1}] = \sigma_t^2$. La moyenne et la variance conditionnelles sont donc déterministes étant donné l'information qui peut être observée sur les marchés.

Il est pratique courante en économétrie financière de modéliser les processus $\{\mu_t\}$ et $\{\sigma_t^2\}$ séparément dans le modèle (1.0.1). Par exemple, le modèle autorégressif, connu sous l'acronyme AR [voir 156], est populaire pour modéliser les dynamiques de la moyenne conditionnelle, tandis que la classe des modèles GARCH (voir la section 1.2) est souvent utilisée pour représenter celles de la variance conditionnelle.

1.1. Faits stylisés des rendements financiers

Les rendements financiers exposent des dynamiques complexes qui sont difficiles à reproduire avec des modèles économétriques. Plusieurs de ces propriétés statistiques sont appelées les « faits stylisés des rendements financiers » (*stylized facts*) puisqu'elles sont communes à un large éventail de séries chronologiques étudiées en finance. Un modèle économétrique adéquat devrait avoir la capacité de reproduire les faits stylisés suivants [voir 41] :

(1) Absence d'autocorrélation dans les rendements : La fonction d'autocorrélation échantillonnale (*sample autocorrelation function*) associée à $\mathrm{Corr}[y_{t+h}, y_t]$ est approximativement égale à zéro pour $h \geq 1$. [2]

(2) Présence d'autocorrélation positive dans les processus $\{\epsilon_t^2\}$ et $\{|\epsilon_t|\}$: Ceci implique une forme de dépendance entre les rendements et la présence d'hétéroscédasticité conditionnelle (*conditional heteroscedasticity*), c'est-à-dire, $\sigma_t^2 \neq \sigma^2$. La fonction d'autocorrélation échantillonnale associée à $\mathrm{Corr}[\epsilon_{t+h}^2, \epsilon_t^2]$ est généralement positive à $h = 1$ et décroît lentement lorsque h augmente.

(3) Clustering de volatilité (*volatility clustering*) : Les rendements de faible et de forte amplitudes tendent à être regroupés. Ce phénomène est lié au point (2).

(4) Queues épaisses et asymétrie négative (*heavy tails and negative skewness*) : La densité inconditionnelle du rendement y_t a des queues plus épaisses

2. La corrélation entre deux variables aléatoires X et Y est définit par

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}} = \frac{\mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}.$$

que la loi normale, c'est-à-dire, le coefficient d'aplatissement (*kurtosis*) de cette densité et plus élevé que celui de la loi normale qui est de trois. Le coefficient d'asymétrie (*skewness*) de la densité des rendements est généralement inférieur à zéro signifiant que les rendements négatifs sont moins nombreux que les rendements positifs, mais ils sont plus extrêmes.

(5) Effet de levier (*leverage effect*) : Un choc négatif important sur le rendement au temps t tend à accroître la volatilité des rendements subséquents par un plus grand écart comparativement à un choc positif de la même amplitude. Mathématiquement, cela s'exprime par la relation

$$\text{Corr}\left[|\epsilon_{t+h}|, \max(-\epsilon_t, 0)\right] > \text{Corr}\left[|\epsilon_{t+h}|, \max(\epsilon_t, 0)\right].$$

Ces faits stylisés indiquent que le plus grand défi de la modélisation des rendements financiers consiste à décrire les dynamiques de la variance conditionnelle. En effet, ces dynamiques sont généralement beaucoup plus complexes que celles de la moyenne conditionnelle. Par conséquent, il est souvent supposé que $\mu_t = \mu$, $\forall t$, une hypothèse qui concorde avec le fait stylisé (1).

1.2. Modèles GARCH

Les modèles GARCH (*generalized autoregressive conditional heteroscedasticity*) ont été introduit par Engle [62] et Bollerslev [21] dans les années 1980. Cette classe de modèles est rapidement devenue très populaire pour expliquer les dynamiques des rendements financiers puisqu'elle est en mesure de reproduire une multitude de faits stylisés. Le modèle GARCH(p, q) de Bollerslev [21] modélise la variance conditionnelle dans l'équation (1.0.1) par l'entremise de la relation paramétrique suivante :

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2.$$

Afin de garantir $\sigma_t^2 > 0$, les paramètres $\{\omega, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p\}$ doivent respecter les contraintes : $\omega > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, q$, et $\beta_j \geq 0$, $j = 1, \ldots, p$. Lorsque $\beta_j = 0$, $j = 1, \ldots, p$, on obtient le modèle ARCH(q) de Engle [62]. Cette thèse se restreint aux modèles GARCH du type GARCH$(1, 1)$, soit la formulation la plus populaire en économétrie financière, où la variance conditionnelle admet la forme suivante :

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{1.2.1}$$

Dans ce qui suit, le modèle dit GARCH correspondra au modèle (1.0.1) avec $\mu_t = \mu$, $\forall t$, et où l'évolution de σ_t^2 est donnée par l'équation (1.2.1).

Lorsque la condition $\alpha+\beta < 1$ est satisfaite dans le modèle GARCH, on obtient un processus $\{y_t\}$ qui est stationnaire au sens strict [3] (*strictly stationary*) ainsi qu'au second ordre [voir 71, chapitre 2]. Dans ce cas, la variance inconditionnelle est donnée par l'expression :

$$\mathrm{Var}[y_t] = \frac{\omega}{1-\alpha-\beta}.$$

Le modèle GARCH n'induit pas de corrélation dans les rendements, mais il génère une autocorrélation positive dans le processus $\{\epsilon_t^2\}$ et permet donc de reproduire des périodes de faible et de forte volatilités. De plus, le coefficient d'aplatissement de la distribution inconditionnelle de y_t est supérieur ou égal à celui du terme d'erreur η_t [voir 71, section 2.4.2]. Ceci signifie que le modèle GARCH peut générer une distribution inconditionnelle avec des queues plus épaisses que celles de la loi normale, même si le terme d'erreur suit cette loi. Par contre, il n'est pas capable de reproduire l'effet de levier. Par conséquent, plusieurs généralisations du modèle GARCH ont été proposées au début des années 1990, incluant :

(1) GJR GARCH : Le modèle Glosten-Jagannathan-Runkle GARCH, proposé par Glosten et al. [80], généralise l'équation (1.2.1) en lui ajoutant un terme permettant de refléter l'effet de levier :

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \lambda\epsilon_{t-1}^2\mathbf{1}_{\{\epsilon_{t-1}<0\}} + \beta\sigma_{t-1}^2.$$

Lorsque λ est positif, un choc négatif, $\epsilon_{t-1} < 0$, accroît la volatilité au temps t par un plus grand écart qu'un choc positif, $\epsilon_{t-1} > 0$, de la même amplitude.

(2) APARCH : Le modèle *asymmetric power* ARCH, introduit par Ding et al. [49], généralise le modèle GJR GARCH avec l'équation suivante :

$$\sigma_t^\delta = \omega + \alpha(|\epsilon_{t-1}| - \lambda\epsilon_{t-1})^\delta + \beta\sigma_{t-1}^\delta.$$

Lorsque $\delta = 2$, ce modèle est équivalent au GJR GARCH, bien que la paramétrisation n'est pas la même. De plus, lorsque $\delta = 1$, ce modèle prend la forme du *threshold* GARCH proposé par Zakoïan [168].

(3) EGARCH : Le modèle *exponential* GARCH, développé par Nelson [142], modélise le logarithme de la variance conditionnelle et permet également de tenir en compte l'effet de levier :

$$\log\sigma_t^2 = \omega + \alpha\left(|\eta_{t-1}| - \mathrm{E}[|\eta_{t-1}|]\right) + \lambda\eta_{t-1} + \beta\log\sigma_{t-1}^2.$$ [4]

3. Le processus $\{y_t\}$ est dit stationnaire au sens strict si les vecteurs $y_{1:t}$ et $y_{(1+h):(t+h)}$ possèdent la même distribution jointe pour tout $t \in \mathbb{N}$ et tout $h \in \mathbb{Z}$.

4. Lorsque η_{t-1} est de loi normale centrée réduite, nous avons $\mathrm{E}[|\eta_{t-1}|] = 2/\sqrt{2\pi}$.

L'avantage de modéliser le logarithme de la variance conditionnelle est qu'il n'est pas nécessaire d'imposer des contraintes sur les paramètres pour assurer $\sigma_t^2 > 0$.

Hentschel [104] a proposé une formulation du modèle GARCH qui regroupe les modèles GARCH, GJR GARCH, APARCH et EGARCH sous une même famille de modèles.

Bien que les modèles énumérés ci-dessus permettent d'incorporer l'effet de levier dans le modèle GARCH, ils ne génèrent pas nécessairement une asymétrie négative dans la distribution inconditionnelle des rendements. He et al. [101] ont démontré que cette distribution est symétrique si la moyenne est constante et si la distribution du terme d'erreur η_t est symétrique, peu importe la forme de l'équation utilisée pour modéliser la variance conditionnelle. Par conséquent, pour générer une asymétrie négative, il faut incorporer une moyenne qui varie dans le temps ou une densité asymétrique pour le terme d'erreur.

Un excellent ouvrage récent sur les modèles GARCH a été réalisé par Francq et Zakoïan [71] et une liste des modèles GARCH développés dans les trois dernières décennies a été composée par Bollerslev [23].

1.2.1. Estimation

L'estimation des paramètres d'un modèle, contenus dans un vecteur que l'on dénote par θ, s'effectue après avoir observé une série chronologique de T rendements que l'on dénote par $y_{1:T}$. Pour ne pas alourdir la notation, il n'y aura pas de distinction entre « variable aléatoire » et « observation » dans la notation puisque le contexte permettra d'identifier cette différence. L'estimation d'un modèle est souvent effectuée avec la méthode du maximum de vraisemblance (*maximum likelihood*) qui consiste à trouver le vecteur de paramètres maximisant la fonction de vraisemblance du modèle. Pour symboliser cette vraisemblance, nous utilisons la notation, $f(y_{1:T} \mid \theta)$, qui représente également la densité de probabilité jointe (*joint probability density function*) du vecteur aléatoire $y_{1:T}$, évaluée aux valeurs observées des rendements et à la valeur choisie du paramètre θ. L'estimateur du maximum de vraisemblance (*maximum likelihood estimator*), dénoté par $\hat{\theta}$, satisfait la relation suivante :

$$\hat{\theta} = \arg\max_{\theta} \ \log f(y_{1:T} \mid \theta).$$

La log-vraisemblance (*log-likelihood*), $\log f(y_{1:T} \mid \theta)$, se décompose de la manière suivante :

$$\log f(y_{1:T} \mid \theta) = f(y_1 \mid \theta) + \sum_{t=2}^{T} \log f(y_t \mid y_{1:t-1}, \theta), \tag{1.2.2}$$

où le symbole $f(y_t \mid y_{1:t-1}, \theta)$ dénote la densité de probabilité conditionnelle de la variable aléatoire y_t étant donné $y_{1:t-1}$, évaluée aux valeurs observées des rendements et à la valeur choisie du paramètre θ. Pour donner une forme plus explicite à cette log-vraisemblance, nous pouvons supposer que η_t est de loi normale centrée réduite dans le modèle (1.0.1). Sous cette hypothèse, nous avons

$$f(y_t \mid y_{1:t-1}, \theta) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left(-\frac{(y_t - \mu_t)^2}{2\sigma_t^2}\right),$$

et, donc, nous obtenons

$$\log f(y_{1:T} \mid \theta) = -\frac{T \log(2\pi)}{2} - \frac{1}{2} \sum_{t=1}^{T} \left[\log(\sigma_t^2) + \frac{(y_t - \mu_t)^2}{\sigma_t^2}\right]. \tag{1.2.3}$$

La popularité de la technique du maximum de vraisemblance dans le cadre des modèles GARCH est due à la facilité de calculer cette log-vraisemblance. Si l'on connaît σ_1^2, il est simple de calculer récursivement σ_t^2, $t = 2, \ldots, T$, par l'entremise de la relation (1.2.1) et, donc, d'obtenir la log-vraisemblance donnée en (1.2.3). La variance σ_1^2 dépend de σ_0^2 et ϵ_0^2 qui sont inconnus et il faut donc faire une hypothèse sur ces valeurs pour calculer σ_1^2. Dans les applications pratiques, elles sont généralement remplacées par la variance échantillonnale (*sample variance*). Les librairies fGARCH [165] et rugarch [78] disponibles avec le logiciel R [150] permettent d'estimer plusieurs types de modèles GARCH.

1.2.2. Persistance de la variance conditionnelle

Le clustering de volatilité, mentionné parmi les fait stylisés dans la section 1.1, est une propriété très importante des rendements financiers. Afin de quantifier ce phénomène dans un modèle, il faut établir une mesure de la persistance de la variance conditionnelle. Lamoureux et Lastrapes [125] définissent la notion de persistance comme suit : « Persistence in variance of a random variable, evolving through time, refers to the property of momentum in conditional variance ; past volatility explains current volatility. »

La persistance de la variance conditionnelle est fortement liée à la présence d'autocorrélation positive dans le processus $\{\epsilon_t^2\}$. Pour le modèle GARCH, si $\alpha + \beta < 1$, il est possible de démontrer que [voir 71, chapitre 2] :

$$\mathrm{Corr}[\epsilon_{t+h}^2, \epsilon_t^2] \propto (\alpha + \beta)^h, \quad h \geq 2.$$

Par conséquent, plus que la somme $\alpha + \beta$ s'approche de 1, plus que le modèle GARCH tend à générer des regroupements de rendements avec de faible et de forte amplitudes, c'est-à-dire, un clustering de volatilité. La littérature économétrique a donc adopté la somme $\alpha + \beta$ pour mesurer la persistance de la variance conditionnelle dans le modèle GARCH.

Lorsqu'on considère des séries financières quotidiennes ou hebdomadaires couvrant un horizon de plusieurs années, la somme $\alpha + \beta$ est fréquemment estimée tout juste au-dessous de un [voir 71, section 7.3]. Ceci indique un fort degré de persistance dans la variance des rendements financiers et a conduit Bollerslev [22] à proposer le modèle *integrated* GARCH (IGARCH) qui impose la contrainte $\alpha + \beta = 1$ au modèle GARCH. Lamoureux et Lastrapes [125] affirment : « One potential problem of IGARCH is that, unlike the random walk in mean for asset prices, it lacks theoretical motivation. » Cette affirmation a motivé Diebold [48] à postuler que le phénomène de la persistance élevée estimée dans le modèle GARCH est dû à des variations dans le paramètre ω qui sont négligées. Les études empirique et Monte Carlo menées par Lamoureux et Lastrapes [125] ont évoqué la plausibilité d'une telle hypothèse. Hillebrand [106] et Mikosch et Starica [135] ont par la suite démontré rigoureusement qu'ignorer des changements structurels dans les paramètres du modèle GARCH tend à biaiser l'estimation de $\alpha + \beta$ vers un, appuyant ainsi l'hypothèse de Diebold [48].

Une approche permettant de faire varier les paramètres du modèle GARCH dans le temps est de le combiner avec un modèle à changement de régimes. Cette combinaison, initialement proposée par Cai [29], Gray [83] et Hamilton et Susmel [93], donne naissance au modèle GARCH à changement de régimes qui fait l'objet de cette thèse et qui sera introduit au chapitre 2. La raison d'être de ce modèle se justifie donc par l'incapacité des modèles GARCH à pouvoir refléter des variations dans les paramètres qui sont souvent observées dans les séries financières.

1.3. Modèles à changement de régimes

Les modèles à changement de régimes (*regime-switching models*) supposent l'existence d'un processus stochastique non-observé régissant la distribution du rendement y_t. Ce processus est généralement une chaîne de Markov à temps discret et à espace d'états discret de dimension finie (*discrete-time Markov chain with finite state-space*). Dans ce cas, le modèle est dénommé *Markov-switching model* ou modèle de Markov caché (*hidden Markov model*). Lorsque la chaîne de Markov a un espace d'états continu, le modèle est généralement appelé modèle en variables d'état (*state-space model*) dans la littérature. L'attrait des modèles

à changement de régimes pour modéliser des séries économétriques a débuté avec l'article de Hamilton [90] qui les a proposés pour refléter l'influence de l'état de l'économie sur le produit national brut des États-Unis. Dans un contexte économétrique, la valeur prise par la chaîne de Markov au temps t est appelée état ou régime et a pour utilité de régir les changements structurels liés à la variable financière étudiée.

1.3.1. Chaîne de Markov à temps discret et à espace d'états discret

Une chaîne de Markov à temps discret et à espace d'états discret de dimension finie M est un processus stochastique à temps discret, dénoté par $\{S_t\}$, prenant des valeurs dans l'ensemble $\{1, 2, \ldots, M\}$, appelé espace d'états (*state-space*), et possédant la propriété de Markov :

$$\Pr[S_t = s_t \mid S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, \ldots] = \Pr[S_t = s_t \mid S_{t-1} = s_{t-1}].$$

Cette thèse se restreint aux chaînes de Markov homogènes (*time-homogeneous*) où les probabilités de transition d'un état $S_{t-1} = i$ à $S_t = j$ sont homogènes en fonction du temps, c'est-à-dire,

$$\forall t : \Pr[S_t = j \mid S_{t-1} = i] = p_{ij}, \quad i, j = 1, \ldots, M.$$

Ces probabilités de transition sont généralement rassemblées dans une matrice, $[\, p_{ij} \,]_{i,j=1}^{M}$, appelée matrice de transition de la chaîne de Markov. On suppose généralement que la chaîne de Markov est irréductible et que tous ses états sont apériodiques. [5] Ceci évite des situations dans lesquelles la chaîne pourrait rester enfermée dans un sous-ensemble de l'espace d'états et a pour conséquence que la chaîne de Markov est ergodique. Une chaîne de Markov ergodique satisfait la relation :

$$\lim_{t \to \infty} \Pr[S_t = j \mid S_0 = i] = \pi_j, \quad \forall i, j = 1, \ldots, M,$$

où $\{\pi_j\}_{j=1}^{M}$ est la solution unique du système d'équations :

$$0 < \pi_j < 1, \quad j = 1, \ldots, M,$$

$$\sum_{j=1}^{M} \pi_j = 1,$$

$$\pi_j = \sum_{i=1}^{M} \pi_i p_{ij}, \quad j = 1, \ldots, M.$$

5. Une chaîne de Markov est dite irréductible si chaque état de la chaîne est accessible (en un ou plusieurs pas) à partir de tous les états de la chaîne. Un état i est dit apériodique si

$$\text{pgdc}\{t : \Pr[S_t = i \mid S_0 = i] > 0\} = 1,$$

où « pgdc » signifie « plus grand diviseur commun ».

La solution $\{\pi_j\}_{j=1}^M$ est un vecteur de probabilités correspondant à la loi stationnaire de la chaîne de Markov.

1.3.2. Modèle à changement de régimes général

Un modèle à changement de régimes très général peut être construit en incorporant cette chaîne de Markov dans le modèle (1.0.1) comme suit :

$$y_t = \mu_t(S_{1:t}) + \sigma_t(S_{1:t})\eta_t, \tag{1.3.1}$$

où $\mu_t(S_{1:t})$ et $\sigma_t(S_{1:t})$ dénotent des fonctions mesurables par rapport à une tribu engendrée par le vecteur aléatoire $(y_{1:t-1}, S_{1:t})$ et où les processus $\{\eta_t\}$ et $\{S_t\}$ sont indépendants. Par conséquent, la moyenne conditionnelle du rendement au temps t est $\mathrm{E}[y_t \mid y_{1:t-1}, S_{1:t}] = \mu_t(S_{1:t})$ et sa variance conditionnelle est $\mathrm{Var}[y_t \mid y_{1:t-1}, S_{1:t}] = \sigma_t^2(S_{1:t})$. La moyenne et la variance conditionnelles dépendent maintenant de la chaîne de Markov non-observée, et ne sont donc plus déterministes étant donné l'information qui peut être observée sur les marchés.

Le modèle (1.3.1) a pour particularité que la distribution conditionnelle de y_t dépend de toute la trajectoire des régimes de 1 à t, rendant ainsi le calcul de la log-vraisemblance du modèle impossible en pratique et l'estimation des paramètres très difficile (voir la section 1.3.5 et les chapitres 2 et 3). Par conséquent, la presque totalité des modèles à changement de régimes considérés dans les applications empiriques font partie d'une sous-classe du modèle (1.3.1), soit :

$$y_t = \mu_t(S_t) + \sigma_t(S_t)\eta_t, \tag{1.3.2}$$

où $\mu_t(S_t)$ et $\sigma_t(S_t)$ dénotent des fonctions mesurables par rapport à une tribu engendrée par le vecteur aléatoire $(y_{1:t-1}, S_t)$. Dans ce modèle, la densité de probabilité conditionnelle de y_t dépend seulement de l'état au temps t plutôt que de l'historique de tous les régimes, c'est-à-dire,

$$f(y_t \mid y_{1:t-1}, S_{1:t}, \theta) = f(y_t \mid y_{1:t-1}, S_t, \theta). \tag{1.3.3}$$

Lorsque la condition (1.3.3) est satisfaite, les techniques existantes permettent d'estimer le modèle à changement de régimes par le maximum de vraisemblance. Cette optimisation peut être effectuée avec deux approches qui seront introduites dans les sections 1.3.3 et 1.3.4, soit (i) le filtre d'Hamilton et (ii) l'algorithme espérance-maximisation (EM). De plus, la section 1.3.5 expliquera pourquoi ces méthodes ne sont plus applicables dans le cadre du modèle général en (1.3.1), où la simplification (1.3.3) n'est pas possible.

La contribution principale de cette thèse est de proposer deux méthodologies d'estimation permettant de calculer l'estimateur du maximum de vraisemblance

des modèles du type général (1.3.1), pour lesquels il n'y a pas de technique d'estimation bien établie. Ces deux approches seront présentées aux chapitres 2 et 3 dans le cadre du modèle GARCH à changement de régimes qui est un cas particulier du modèle général en (1.3.1). Cependant, il est important de souligner que les techniques proposées peuvent être étendues à d'autres modèles du type (1.3.1), ainsi qu'à des variantes de ce modèle, telles que le modèle ARMA (*autoregressive moving-average*) à changement de régimes introduit par Francq et Zakoïan [68] et le modèle ARMA-GARCH à changement de régimes présenté par Henneke et al. [102].

Finalement, les approches bayésiennes sont également populaires pour estimer les modèles à changement de régimes [voir 74] et se basent sur un concept commun avec l'algorithme EM, soit la technique du *data augmentation* [159]. Cette thèse n'élaborera pas sur l'approche bayésienne puisque qu'elle considère principalement l'estimation par le maximum de vraisemblance dans un cadre fréquentiste.

1.3.2.1. *Exemple d'un modèle à changement de régimes*

Pour clore cette section, nous présentons en exemple un modèle à changement de régimes populaire en économétrie financière et en actuariat :

$$y_t = \mu_{S_t} + \sigma_{S_t}\eta_t, \tag{1.3.4}$$

où μ_{S_t} et $\sigma^2_{S_t}$ dénotent des fonctions mesurables par rapport à une tribu engendrée par la variable aléatoire $S_t \in \{1, 2, \ldots, M\}$ et représentent, respectivement, la moyenne et la variance conditionnelles du rendement y_t dans le régime S_t, c'est-à-dire, $\mu_{S_t} = \mathrm{E}[y_t \mid S_t]$ et $\sigma^2_{S_t} = \mathrm{Var}[y_t \mid S_t]$. Ce modèle est donc plus simple que celui présenté à l'équation (1.3.2), car la moyenne et la variance conditionnelles ne dépendent pas de l'historique des rendements.

Lorsque le terme d'erreur η_t est de loi normale, la densité de probabilité conditionnelle de y_t est :

$$f(y_t \mid S_t = i, \theta) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(y_t - \mu_i)^2}{2\sigma_i^2}\right), \quad i = 1, \ldots, M.$$

Dans ce cas, le vecteur des paramètres du modèle est $\theta = (\{\mu_i, \sigma_i^2\}_{i=1}^M, \{p_{ij}\}_{i,j=1}^M)$. Compte tenu de la contrainte $\sum_{j=1}^M p_{ij} = 1$, pour $i = 1, \ldots, M$, θ contient $(2M + M(M-1))$ paramètres qui doivent être estimés. Bien que ce modèle est simple, Hardy [94, 96] et Hardy et al. [97] ont démontré son utilité dans certaines applications actuarielles et l'évaluation des risques financiers. Par exemple, il permet de modéliser le rendement avec une distribution ayant une moyenne négative et une volatilité élevée en temps de crise, et une autre avec une moyenne positive

et une volatilité faible en période de prospérité économique. Ce mélange de distributions permet de créer le phénomène recherché du clustering de volatilité et de générer une distribution inconditionnelle du rendement avec une queue gauche épaisse et une asymétrie négative. Dans le domaine de l'actuariat, Hardy [94] a baptisé ce modèle « RSLN » (*regime-switching lognormal*).

1.3.3. Filtre d'Hamilton

Hamilton [90, 91, 92] a proposé un algorithme permettant de calculer la log-vraisemblance des modèles à changement de régimes respectant la condition (1.3.3). Cet algorithme, dénommé le filtre d'Hamilton (*Hamilton filter*) ou tout simplement l'algorithme de filtrage avant (*forward filtering algorithm*), est applicable aux modèles à changement de régimes (1.3.2) et (1.3.4), mais pas au modèle général (1.3.1).

La première étape de l'algorithme consiste à calculer :

$$\begin{aligned} f(y_1 \mid \theta) &= \sum_{j=1}^{M} f(y_1, S_1 = j \mid \theta) \\ &= \sum_{j=1}^{M} p(S_1 = j \mid \theta) \cdot f(y_1 \mid S_1 = j, \theta), \end{aligned}$$

et

$$p(S_1 = j \mid y_1, \theta) = \frac{f(y_1, S_1 = j \mid \theta)}{f(y_1 \mid \theta)}, \quad j = 1, \ldots, M,$$

où $p(\cdot)$ dénote une fonction de masse[6] (*probability mass function*). Par la suite, pour $t = 2, \ldots, T$, on calcule récursivement :

$$\begin{aligned} f(y_t \mid y_{1:t-1}, \theta) &= \sum_{j=1}^{M} f(y_t, S_t = j \mid y_{1:t-1}, \theta) \\ &= \sum_{i,j=1}^{M} p(S_{t-1} = i \mid y_{1:t-1}, \theta) \cdot p_{ij} \cdot f(y_t \mid y_{1:t-1}, S_t = j, \theta), \end{aligned}$$

et

$$p(S_t = j \mid y_{1:t}, \theta) = \frac{f(y_t, S_t = j \mid y_{1:t-1}, \theta)}{f(y_t \mid y_{1:t-1}, \theta)}, \quad j = 1, \ldots, M. \tag{1.3.5}$$

6. L'usage que l'on fait de $p(\cdot)$ est similaire à celui de la densité de probabilité $f(\cdot)$, c'est-à-dire, les mêmes abus de notation sont utilisés. Par rapport à $f(\cdot)$, le symbole $p(\cdot)$ précise que la variable (ou vecteur) aléatoire sous-jacente est discrète. Par exemple, $p(S_1 = j \mid \theta)$ est la fonction de masse de la variable aléatoire S_1 évaluée en j et à la valeur choisie du paramètre θ. Parfois, un argument explicite à la fonction n'a pas besoin d'être précisé et, dans ce cas, l'argument j est omis dans la fonction $p(\cdot)$. Par exemple, $p(S_1 \mid \theta)$ est la fonction de masse de la variable aléatoire S_1 évaluée en une réalisation arbitraire de S_1.

Finalement, la log-vraisemblance est obtenue avec la relation (1.2.2), soit,

$$\log f(y_{1:T} \mid \theta) = f(y_1 \mid \theta) + \sum_{t=2}^{T} \log f(y_t \mid y_{1:t-1}, \theta),$$

et peut donc être maximisée numériquement.

Notons qu'à la première étape de l'algorithme, la probabilité $p(S_1 \mid \theta)$ est requise et il faut donc établir une hypothèse sur la distribution de l'état au temps 0, S_0. D'une part, on peut supposer que la chaîne de Markov $\{S_t\}$ est stationnaire, c'est-à-dire, la loi de S_0 correspond à la loi stationnaire de la chaîne (voir la section 1.3.1). Dans ce cas, $p(S_1 \mid \theta)$ est obtenu directement à partir de cette loi stationnaire. D'autre part, on peut traiter la distribution initiale de S_0, dénotée par le vecteur de probabilités δ de dimension M, comme un paramètre du modèle et l'estimer. Cependant, maximiser la vraisemblance en fonction de δ génère un vecteur prenant la forme $(0, \ldots, 0, 1, 0, \ldots, 0)$ [voir 169, section 4.2.4]. Par conséquent, sans perte de généralité, on peut supposer $S_0 = 1$ puisque la vraisemblance des modèles à changement de régimes est invariante sous une permutation des régimes. Donc, si on suppose que la distribution de S_0 est inconnue, il suffit de remplacer $p(S_1 = j \mid \theta)$ par $p(S_1 = j \mid S_0 = 1, \theta) = p_{1j}$.

En effet, l'étiquette que l'on donne à chacun des régimes est arbitraire et il est possible d'interchanger les paramètres de deux régimes sans changer la valeur de la vraisemblance. Cette caractéristique du modèle cause des difficultés quand une approche bayésienne est utilisée pour estimer les paramètres du modèle, un problème connu sous le nom de *label switching problem*. Toutefois, « label switching is of no concern for maximum likelihood estimation, where the goal is to find one of the equivalent modes of the likelihood function » [74, Section 3.5.5]. En théorie, l'estimateur du maximum de vraisemblance n'est pas unique, mais, en pratique, les estimateurs du maximum de vraisemblance associés à chaque possible étiquetage des régimes sont tous équivalents.

1.3.3.1. *Probabilités filtrée et lissée*

L'estimation des paramètres n'est pas la seule inférence d'intérêt dans un modèle à changement de régimes. Puisque les régimes de la chaîne de Markov ne sont pas observés, il est également pertinent de les décoder à partir de l'information disponible, soit les rendements observés. Dans plusieurs modèles économétriques, ces régimes sont utilisés pour représenter différents états de l'économie et une telle inférence a donc un intérêt économique. Un bénéfice du filtre d'Hamilton est le calcul de la probabilité $p(S_t \mid y_{1:t}, \theta)$ à l'équation (1.3.5), dite probabilité filtrée (*filtered probability*), qui nous procure une inférence du régime au temps t

étant donné les rendements observés jusqu'à ce moment. Cette inférence est donc obtenue lors du calcul de la log-vraisemblance du modèle, sans effort additionnel.

Une seconde inférence d'intérêt est la probabilité $p(S_t \mid y_{1:T}, \theta)$, dite probabilité lissée (*smoothed probability*), qui est basée sur l'échantillon complet des données. Pour calculer les probabilités lissées, il faut faire suivre le filtre d'Hamilton d'un algorithme de lissage arrière (*backward smoothing algorithm*) [voir 117]. Il consiste à calculer les probabilités, $p(S_t \mid y_{1:T}, \theta)$, récursivement de $t = T-1$ à 1 en utilisant les deux relations suivantes :

$$\begin{aligned} p(S_t = i, S_{t+1} = j \mid y_{1:T}, \theta) &= p(S_{t+1} = j \mid y_{1:T}, \theta) \cdot p(S_t = i \mid S_{t+1} = j, y_{1:T}, \theta) \\ &= p(S_{t+1} = j \mid y_{1:T}, \theta) \cdot p(S_t = i \mid S_{t+1} = j, y_{1:t}, \theta) \\ &= p(S_{t+1} = j \mid y_{1:T}, \theta) \cdot \frac{p(S_t = i, S_{t+1} = j \mid y_{1:t}, \theta)}{p(S_{t+1} = j \mid y_{1:t}, \theta)} \\ &= \frac{p(S_{t+1} = j \mid y_{1:T}, \theta) \cdot p(S_t = i \mid y_{1:t}, \theta) \cdot p_{ij}}{p(S_{t+1} = j \mid y_{1:t}, \theta)}, \quad i, j = 1, \ldots, M, \end{aligned}$$

et

$$p(S_t = i \mid y_{1:T}, \theta) = \sum_{j=1}^{M} p(S_t = i, S_{t+1} = j \mid y_{1:T}, \theta), \quad i = 1, \ldots, M.$$

La condition (1.3.3) doit être satisfaite pour justifier le passage de la première ligne à la deuxième ligne dans la décomposition de $p(S_t = i, S_{t+1} = j \mid y_{1:T}, \theta)$. Le filtre d'Hamilton et l'algorithme de lissage arrière requièrent, chacun, un nombre d'opérations arithmétiques d'ordre $\mathcal{O}(TM^2)$. La combinaison de ces deux algorithmes est communément appelée *forward filtering–backward smoothing algorithm*.

1.3.4. Algorithme espérance-maximisation (EM)

L'algorithme espérance-maximisation (EM) est une technique d'estimation itérative permettant d'obtenir l'estimateur du maximum de vraisemblance sans avoir à maximiser ni à calculer la log-vraisemblance du modèle. Il est utile dans les situations où la log-vraisemblance est difficile à évaluer compte tenu la présence de données manquantes ou de variables non-observées. Le principal avantage de cet algorithme est que pour certains modèles, on obtient un algorithme d'estimation analytique qui ne requiert pas de routine d'optimisation. Ceci se produit lorsque l'algorithme EM est dérivé pour le modèle RSLN (voir la section 1.3.4.1). Par contre, dans des modèles à changement de régimes plus généraux, le filtre d'Hamilton décrit à la section 1.3.3 est d'habitude plus efficace d'un point de vue computationnel.

L'algorithme EM débute avec un choix initial du vecteur des paramètres, dénoté par $\theta^{(0)}$, et permet de générer une suite, $\{\theta^{(r)}\}_{r\geq 1}$, en alternant les deux étapes suivantes :

Algorithme EM

1. Étape espérance : Calculer $Q(\theta \mid \theta^{(r-1)})$, où,

$$\begin{aligned} Q(\theta \mid \theta^{(r-1)}) &= \mathrm{E}[\log f(y, S \mid \theta) \mid y, \theta^{(r-1)}] \\ &= \sum_{S} \log f(y, S \mid \theta) \cdot p(S \mid y, \theta^{(r-1)}). \end{aligned} \tag{1.3.6}$$

2. Étape maximisation : Effectuer l'optimisation

$$\theta^{(r)} = \arg\max_{\theta} \; Q(\theta \mid \theta^{(r-1)}).$$

L'exposant r utilisé sur un symbole indique qu'il se rattache à l'itération r de l'algorithme. Les symboles y et S sont utilisés pour représenter $y_{1:T}$ et $S_{1:T}$, respectivement.

L'utilité de cet algorithme repose sur une propriété de monotonie premièrement prouvée par Dempster et al. [47] : $f(y \mid \theta^{(r)}) \geq f(y \mid \theta^{(r-1)})$, c'est-à-dire, chaque itération nous procure un *meilleur* paramètre. Wu [163] a formellement étudié la convergence de l'algorithme EM et le Théorème 1.3.1 expose un de ses résultats les plus importants [voir également 134, chapitre 3].

Théorème 1.3.1 (Convergence de l'algorithme EM)**.** *Supposons les conditions suivantes satisfaites :*

(1) Conditions de régularité :

- *(i) L'espace des paramètres, dénoté par Ω, est un sous-ensemble de $\mathbb{R}^n$.*
- *(ii) La vraisemblance, $f(y \mid \theta)$, est continue dans Ω et différentiable dans l'intérieur de Ω.*
- *(iii) Le paramètre de départ, $\theta^{(0)}$, satisfait $f(y \mid \theta^{(0)}) > -\infty$.*
- *(iv) $\Omega_{\theta_0} = \{\theta \in \Omega : f(y \mid \theta) \geq f(y \mid \theta_0)\}$ est compact pour tout $f(y \mid \theta_0) > -\infty$.*
- *(v) Ω_{θ_0} est dans l'intérieur de Ω pour tout $\theta_0 \in \Omega$.*

(2) Condition de Wu [163] : La fonction $Q(\theta \mid \theta')$ est continue en θ et θ', pour tout $\theta, \theta' \in \Omega$.

Alors, tous les points limites de la suite générée avec l'algorithme EM, $\{\theta^{(r)}\}_{r\geq 1}$, sont des points stationnaires de la vraisemblance. De plus, $f(y \mid \theta^{(r)})$ converge de manière monotone croissante vers $f(y \mid \hat{\theta})$, où $\hat{\theta}$ est l'un de ces points stationnaires.

DÉMONSTRATION. Pour rendre la preuve de ce théorème accessible, nous supposons, additionnellement, que $Q(\theta \mid \theta')$ est différentiable en θ, pour tout θ et θ' dans l'intérieur de Ω.

On définit :

$$\begin{aligned} H(\theta \mid \theta^{(r)}) &= \mathrm{E}\left[\log p(S \mid y, \theta) \mid y, \theta^{(r)}\right] \\ &= \sum_S \log p(S \mid y, \theta) \cdot p(S \mid y, \theta^{(r)}). \end{aligned}$$

Pour tout $\theta^{(r)}$ et $S \in \{1, \ldots, M\}^T$, nous avons l'identité suivante :

$$\begin{aligned} \log f(y \mid \theta) &= \log f(y, S \mid \theta) - \log p(S \mid y, \theta) \\ &= \mathrm{E}\left[\log f(y, S \mid \theta) \mid y, \theta^{(r)}\right] - \mathrm{E}\left[\log p(S \mid y, \theta) \mid y, \theta^{(r)}\right] \\ &= Q(\theta \mid \theta^{(r)}) - H(\theta \mid \theta^{(r)}), \end{aligned} \tag{1.3.7}$$

et, donc,

$$\log f(y \mid \theta) - \log f(y \mid \theta^{(r)}) = [Q(\theta \mid \theta^{(r)}) - Q(\theta^{(r)} \mid \theta^{(r)})] - [H(\theta \mid \theta^{(r)}) - H(\theta^{(r)} \mid \theta^{(r)})].$$

De plus, si $\theta \neq \theta^{(r)}$, nous obtenons :

$$\begin{aligned} H(\theta \mid \theta^{(r)}) - H(\theta^{(r)} \mid \theta^{(r)}) &= \sum_S \log\left(\frac{p(S \mid y, \theta)}{p(S \mid y, \theta^{(r)})}\right) \cdot p(S \mid y, \theta^{(r)}) \\ &= \mathrm{E}\left[\log\left(\frac{p(S \mid y, \theta)}{p(S \mid y, \theta^{(r)})}\right) \,\middle|\, y, \theta^{(r)}\right] \\ &< \log \mathrm{E}\left[\frac{p(S \mid y, \theta)}{p(S \mid y, \theta^{(r)})} \,\middle|\, y, \theta^{(r)}\right] \qquad (1.3.8) \\ &< \log \mathrm{E}\left[\frac{p(S \mid y, \theta)}{p(S \mid y, \theta^{(r)})} \,\middle|\, y, \theta^{(r)}\right] \\ &< \log 1 = 0, \qquad (1.3.9) \end{aligned}$$

en appliquant l'inégalité de Jensen à la ligne (1.3.8) sur la fonction log qui est strictement concave. Pour $\theta \neq \theta^{(r)}$, ceci implique

$$\log f(y \mid \theta) - \log f(y \mid \theta^{(r)}) > Q(\theta \mid \theta^{(r)}) - Q(\theta^{(r)} \mid \theta^{(r)}).$$

Si on choisit $\theta = \theta^{(r+1)}$, où,

$$\theta^{(r+1)} = \arg\max_{\theta} \; Q(\theta \mid \theta^{(r)}),$$

on obtient

$$\log f(y \mid \theta^{(r+1)}) - \log f(y \mid \theta^{(r)}) \geq 0,$$

avec égalité si et seulement si $\theta^{(r+1)} = \theta^{(r)}$. Par conséquent, la suite $\{\log f(y \mid \theta^{(r)})\}_{r \geq 1}$ est monotone croissante. Les conditions de régularité *(ii)–(iv)* impliquent

qu'elle est également bornée supérieurement (*bounded above*) et, donc, convergente. De plus, la condition de régularité *(v)* implique que chaque itération du paramètre, $\theta^{(r)}$, est dans l'intérieur de Ω, plutôt qu'aux bords.

Dénotons par $\hat{\theta}$, un point limite de la suite $\{\theta^{(r)}\}_{r\geq 1}$. À la limite, nous devons avoir la relation,

$$\hat{\theta} = \arg\max_{\theta} \; Q(\theta \mid \hat{\theta}),$$

car autrement $\hat{\theta}$ ne pourrait être un point limite. Puisque $\theta = \hat{\theta}$ maximise $Q(\theta \mid \hat{\theta})$ et se situe à l'intérieur de l'espace des paramètres, nous avons nécessairement

$$\frac{\partial}{\partial\theta} Q(\theta \mid \hat{\theta}) \bigg|_{\theta=\hat{\theta}} = 0.$$

De plus, la relation (1.3.9) implique que $H(\theta \mid \hat{\theta})$ est à son maximum en $\theta = \hat{\theta}$ et, donc,

$$\frac{\partial}{\partial\theta} H(\theta \mid \hat{\theta}) \bigg|_{\theta=\hat{\theta}} = 0.$$

Finalement, par l'équation (1.3.7), on obtient :

$$\frac{\partial}{\partial\theta} \log f(y \mid \theta) \bigg|_{\theta=\hat{\theta}} = 0,$$

signifiant que $\hat{\theta}$ est un point stationnaire de la vraisemblance. Il a donc été démontré que tous les points limites de la suite générée avec l'algorithme EM, $\{\theta^{(r)}\}_{r\geq 1}$, sont des points stationnaires de la vraisemblance et que $f(y \mid \theta^{(r)})$ converge de manière monotone croissante vers $f(y \mid \hat{\theta})$, où $\hat{\theta}$ est l'un de ces points stationnaires. □

Bien que le Théorème 1.3.1 conclue que la suite $\{\log f(y \mid \theta^{(r)})\}_{r\geq 1}$ est convergente, ceci n'implique pas automatiquement la convergence de la suite $\{\theta^{(r)}\}_{r\geq 1}$, expliquant pourquoi on parle de points limites de $\{\theta^{(r)}\}_{r\geq 1}$ dans le théorème. Toutefois, la non-convergence de $\{\theta^{(r)}\}_{r\geq 1}$ se produit très rarement en pratique et fait plutôt partie d'un cas pathologique [voir 134, chapitre 3].

Si la log-vraisemblance est unimodale en θ et possède un seul point stationnaire, le Théorème 1.3.1 assure que l'algorithme EM convergera vers l'estimateur du maximum de vraisemblance peu importe le choix de $\theta^{(0)}$. Par contre, si la log-vraisemblance a plusieurs modes, il ne garantit pas que le point stationnaire obtenu, $\hat{\theta}$, est un maximum global. Toutefois, dans la quasi-totalité des situations, $\hat{\theta}$ est un maximum local de la log-vraisemblance [voir 134, section 3.6, pour des exceptions]. Par conséquent, la convergence de l'algorithme EM vers l'estimateur du maximum de vraisemblance dépend du choix de $\theta^{(0)}$. Notons que lorsqu'on

maximise directement la log-vraisemblance avec une routine d'optimisation standard, on aboutit également à un maximum local qui dépend du choix initial d'un paramètre.

1.3.4.1. *Un exemple de l'algorithme EM*

Cette section illustre l'application de l'algorithme EM pour le modèle RSLN, soit le modèle (1.3.4) avec un terme d'erreur η_t de loi normale. Tout d'abord, on décompose

$$\begin{aligned}
\log f(y, S \mid \theta) &= \log f(y \mid S, \theta) + \log p(S \mid \theta) \\
&= \log \prod_{t=1}^{T} f(y_t \mid S_t, \theta) + \log \prod_{t=1}^{T} p(S_t \mid S_{t-1}, \theta) \\
&= \sum_{t=1}^{T} \log f(y_t \mid S_t, \theta) + \sum_{t=1}^{T} \log p_{S_t, S_{t-1}} \\
&= \sum_{j=1}^{M} \sum_{t=1}^{T} \mathbf{1}_{\{S_t = j\}} \log f(y_t \mid S_t = j, \theta) + \\
&\qquad \sum_{i,j=1}^{M} \log p_{ij} \left[\sum_{t=1}^{T} \mathbf{1}_{\{S_{t-1}=i, S_t=j\}} \right].
\end{aligned}$$

À l'étape maximisation de l'algorithme EM, on doit maximiser la fonction :

$$\begin{aligned}
Q(\theta \mid \theta^{(r-1)}) &= \mathrm{E}[\log f(y, S \mid \theta) \mid y, \theta^{(r-1)}] \\
&= \sum_{j=1}^{M} \sum_{t=1}^{T} p(S_t = j \mid y, \theta^{(r-1)}) \cdot \log f(y_t \mid S_t = j, \theta) + \\
&\qquad \sum_{i,j=1}^{M} \log p_{ij} \left[\sum_{t=1}^{T} p(S_{t-1} = i, S_t = j \mid y, \theta^{(r-1)}) \right] \\
&= \text{terme 1 } + \text{ terme 2.}
\end{aligned}$$

Le terme 1 inclut une somme de M termes qui dépendent chacun d'une paire de paramètres (μ_j, σ_j^2), où $j \in \{1, \ldots, M\}$. Par conséquent, on peut maximiser $Q(\theta \mid \theta^{(r-1)})$ en fonction de (μ_j, σ_j^2) indépendamment des autres paramètres. Cette optimisation a une solution analytique :

$$\mu_j^{(r)} = \frac{\sum_{t=1}^{T} p(S_t = j \mid y, \theta^{(r-1)}) \cdot y_t}{\sum_{t=1}^{T} p(S_t = j \mid y, \theta^{(r-1)})}, \quad j = 1, \ldots, M, \tag{1.3.10}$$

et

$$\sigma_j^{2(r)} = \frac{\sum_{t=1}^{T} p(S_t = j \mid y, \theta^{(r-1)}) \cdot (y_t - \mu_j^{(r)})^2}{\sum_{t=1}^{T} p(S_t = j \mid y, \theta^{(r-1)})}, \quad j = 1, \ldots, M. \tag{1.3.11}$$

Le terme 2 dépend uniquement des paramètres $\{p_{ij}\}_{i,j=1}^{M}$ et peut également être maximisé analytiquement :

$$p_{ij}^{(r)} = \frac{f_{ij}}{\sum_{k=1}^{M} f_{ik}}, \quad i,j = 1,\ldots,M, \tag{1.3.12}$$

où,

$$f_{ij} = \sum_{t=1}^{T} p(S_{t-1} = i, S_t = j \mid y, \theta^{(r-1)}).$$

En somme, l'algorithme EM pour le modèle RSLN consiste en l'application itérative des deux étapes suivantes :

(1) Effectuer le *forward filtering–backward smoothing algorithm* conditionnellement au paramètre $\theta^{(r-1)}$. Cet algorithme permet d'obtenir les probabilités $p(S_{t-1} = i, S_t = j \mid y, \theta^{(r-1)})$ et $p(S_t = j \mid y, \theta^{(r-1)})$, pour $i,j = 1,\ldots,M$ (voir la section 1.3.3.1).

(2) Calculer le paramètre $\theta^{(r)}$ en appliquant les formules (1.3.10)–(1.3.12).

Les itérations s'effectuent donc de manière analytique dans ce contexte et l'estimateur du maximum de vraisemblance est calculé sans avoir recours à une routine d'optimisation numérique. Dans de telles situations, il est souvent préférable d'utiliser l'algorithme EM pour estimer le modèle. Par contre, lorsqu'il est possible de calculer la log-vraisemblance et lorsqu'il n'y a pas de solution analytique à l'étape maximisation de l'algorithme EM, il est généralement plus efficace d'opter pour une optimisation directe de la log-vraisemblance, c'est-à-dire, le filtre d'Hamilton. Néanmoins, l'algorithme EM conserve son attrait quand le calcul de la log-vraisemblance est impraticable ou coûteux numériquement. La librairie mhsmm [143] disponible avec le logiciel R permet d'estimer des modèles à changement de régimes avec l'algorithme EM, tels que le modèle RSLN.

1.3.5. Limites du filtre d'Hamilton et de l'algorithme EM

L'utilisation du *forward filtering–backward smoothing algorithm* est restreint aux modèles à changement de régimes respectant la condition (1.3.3). Il est possible d'étendre l'application de cet algorithme dans le cas où la densité de y_t dépend des q derniers régimes (voir le chapitre 3) :

$$f(y_t \mid y_{1:t-1}, S_{1:t}, \theta) = f(y_t \mid y_{1:t-1}, S_{t-q+1:t}, \theta).$$

Cette généralisation de l'algorithme requiert un nombre d'opérations arithmétiques d'ordre $\mathcal{O}(TM^{q\vee 2})$, nous limitant à de petites valeurs de q.[7] Par conséquent, lorsque la distribution conditionnelle de y_t dépend de toute la trajectoire

7. Il est également possible de définir une nouvelle variable d'état, $\bar{S}_t = S_{t-q+1:t}$, et de reformuler le modèle en fonction de cette variable. Puisque la distribution conditionnelle de y_t

des régimes du temps 1 à t, il est impraticable de maximiser la log-vraisemblance avec le filtre d'Hamilton ou d'utiliser l'algorithme EM. Une telle situation se produit dans le cadre du modèle GARCH à changement de régimes. Ceci explique pourquoi l'estimation de ce modèle est difficile à effectuer et nécessite l'utilisation d'approximations ou de méthodes computationnelles pour évaluer la log-vraisemblance et appliquer l'algorithme EM.

1.3.6. Lien avec les modèles en variables d'état

Les modèles en variables d'état (*state-space models*) partagent plusieurs caractéristiques avec les modèles à changement de régimes. Par exemple, dans les deux modèles, les dynamiques de la variable observée dépendent de la réalisation d'un processus stochastique latent ayant une structure markovienne. Cependant, l'espace d'états de ce processus est continu dans un modèle en variables d'état, plutôt que discret. Lorsque le modèle en variables d'état est linéaire et gaussien, il existe une technique analogue au filtre d'Hamilton, soit le filtre de Kalman [113, 114]. Dans ce cas, il est également possible de dériver un algorithme EM ayant une étape maximisation analytique comme celui présenté dans la section 1.3.4.1 [voir 156, section 6.3]. Quand le modèle prend une forme non-linéaire ou n'est pas gaussien, le filtre de Kalman n'est pas exact et des méthodes computationnelles sont requises afin de procéder à une inférence sur le modèle.

1.4. Statistique computationnelle

Les deux méthodologies développées dans les chapitres 2 et 3 pour calculer l'estimateur du maximum de vraisemblance du modèle GARCH à changement de régimes se basent sur différentes techniques en statistique computationnelle qui sont introduites dans cette section.

1.4.1. Échantillonnage préférentiel (*Importance sampling*)

Supposons que x est un vecteur aléatoire de loi (jointe) $q(x)$ et que l'objectif est d'évaluer l'espérance :

$$\mathrm{E}[h(x)] = \int h(x)q(x)dx \quad \text{ou} \quad \mathrm{E}[h(x)] = \sum h(x)q(x), \tag{1.4.1}$$

où $h(x) \in \mathbb{R}$ est une fonction du vecteur x. Dans le cas où on peut simuler N réalisations indépendantes du vecteur x selon la loi $q(x)$, dénotées par $\{x^{(i)}\}_{i=1}^{N}$,

dépend seulement du (nouveau) régime au temps t, $\bar{S}_t$, cette formulation nous permet d'appliquer le *forward filtering–backward smoothing algorithm* directement. Cependant, cette façon de procéder nécessite un nombre d'opérations arithmétiques d'ordre $\mathcal{O}(TM^{2q})$ qui est coûteux numériquement.

un estimateur Monte Carlo de l'espérance en (1.4.1) est

$$\frac{1}{N}\sum_{i=1}^{N} h(x^{(i)}). \tag{1.4.2}$$

Lorsqu'il n'est pas possible de simuler x selon $q(x)$, l'échantillonnage préférentiel (*importance sampling*) permet de simuler x avec une loi alternative $q_0(x)$, appelée proposition, satisfaisant la condition :

$$q(x) > 0 \Rightarrow q_0(x) > 0.$$

L'estimateur obtenu avec le principe de l'échantillonnage préférentiel est

$$\frac{1}{N}\sum_{i=1}^{N} h(x^{(i)})w^{(i)}, \tag{1.4.3}$$

où,

$$w^{(i)} = \frac{q(x^{(i)})}{q_0(x^{(i)})}, \quad i = 1, \ldots, N,$$

et $\{x^{(i)}\}_{i=1}^{N}$ est un échantillon indépendant généré avec la loi $q_0(x)$. Cet estimateur est sans biais et la loi forte des grands nombres implique qu'il converge presque sûrement vers l'espérance en (1.4.1). Son efficacité dépend du choix de $q_0(x)$ et dans certaines situations il est possible de construire une loi $q_0(x)$ résultant en un estimateur (1.4.3) ayant une plus faible variance que l'estimateur Monte Carlo (1.4.2). Par conséquent, l'échantillonnage préférentiel est également utilisé comme technique de réduction de variance.

Dans plusieurs contextes il est impossible de calculer la constante de normalisation associée au poids $w^{(i)}$ et l'estimateur

$$\sum_{i=1}^{N} h(x^{(i)})\bar{w}^{(i)}, \tag{1.4.4}$$

où,

$$\bar{w}^{(i)} = \frac{w^{(i)}}{\sum_{j=1}^{N} w^{(j)}}, \quad i = 1, \ldots, N,$$

est utilisé au lieu de celui présenté à l'équation (1.4.3). Cet estimateur converge presque sûrement vers l'espérance en (1.4.1), mais il a un biais d'ordre $1/N$.

Une limite de l'échantillonnage préférentiel est que la variance des estimateurs (1.4.3) ou (1.4.4) peut croître exponentiellement avec la dimension du vecteur x [voir 53]. Par conséquent, lorsque cette dimension est grande, il est important de choisir une proposition $q_0(x)$ offrant une excellente approximation de $q(x)$. L'échantillonnage préférentiel est discuté en détails par Robert et Casella [152] et Cappé et al. [34].

1.4.2. Filtres particulaires (*Particle filters*)

Les filtres particulaires (*particle filters*), aussi connus sous le nom de méthodes de Monte Carlo séquentielles (*sequential Monte Carlo methods*), sont des techniques par simulations qui se basent sur une application séquentielle de l'échantillonnage préférentiel. Ils sont très pratiques dans les modèles en variables d'état qui ne sont pas linéaires et gaussiens, car ils permettent d'estimer les récursions du filtre de Kalman qui ne peuvent pas être effectuées analytiquement. Ils procurent une méthodologie pour approximer les probabilités filtrées et estimer la log-vraisemblance du modèle. Dans le cadre des modèles à changement de régimes, ils sont utiles quand la distribution conditionnelle de la variable observée dépend de toute la trajectoire des régimes parce que le filtre d'Hamilton est impraticable dans ce cas (voir la section 1.3). Le filtre particulaire est en quelque sorte l'analogue Monte Carlo (ou stochastique) du filtre d'Hamilton et du filtre de Kalman. Le contenu de cette section est inspiré de deux excellents ouvrages récents, réalisés par Creal [43] et Doucet et Johansen [53], offrant un aperçu du filtre particulaire.

Dans plusieurs applications (par exemple, le calcul de la log-vraisemblance ou le décodage en ligne des régimes de la chaîne de Markov), nous avons besoin d'une approximation de la loi jointe et conditionnelle $p(S_{1:t} \mid y_{1:t})$ pour $t = 1$ à T.[8] Dans ce cas, il est beaucoup plus efficace d'un point de vue computationnel de générer ces approximations séquentiellement de $t = 1$ à T. Un filtre particulaire est une méthode, se basant sur le principe de l'échantillonnage préférentiel, permettant de *prolonger* une approximation de $p(S_{1:t} \mid y_{1:t})$ obtenue au temps t vers le temps $t + 1$. Dans le cadre des modèles à changement de régimes, la loi $p(S_{1:t} \mid y_{1:t})$ est approximée par l'entremise de N trajectoires simulées de la chaîne de Markov $\{S_t\}$, appelées particules et dénotées par $\{S_{1:t}^{(i)}\}_{i=1}^N$, avec probabilités respectives $\{\bar{w}_t^{(i)}\}_{i=1}^N$, c'est-à-dire,

$$\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N \sim p(S_{1:t} \mid y_{1:t}).$$

En d'autres termes, la loi $p(S_{1:t} \mid y_{1:t})$ est approximée avec une distribution (simulée) discrète représentée par $\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N$. Puisqu'il y a plusieurs procédures permettant de prolonger cette approximation au temps $t+1$, il y a plusieurs filtres particulaires. La section 1.4.2.1 présente une approche dite « conditionnellement optimale » [voir 43, section 2.5.3] qui est très efficace, mais qui ne peut pas être

8. Pour raccourcir la notation, la dépendance d'une loi sur le paramètre θ est omis dans cette section.

implantée pour tous les modèles, tandis que la section 1.4.2.2 illustre une approche moins efficace, mais plus générale.

1.4.2.1. *Approche dite « conditionnellement optimale »*

Supposons que nous avons l'échantillon suivant au temps $t-1$:

$$\{S_{1:t-1}^{(i)}, \bar{w}_{t-1}^{(i)}\}_{i=1}^{N} \sim p(S_{1:t-1} \mid y_{1:t-1}).$$

Nous voulons prolonger cet échantillon au temps t afin d'obtenir un ensemble de particules

$$\{S_{1:t}^{(i)}, \bar{w}_{t}^{(i)}\}_{i=1}^{N} \sim p(S_{1:t} \mid y_{1:t}),$$

où,

$$\begin{aligned} p(S_{1:t} \mid y_{1:t}) &= p(S_t \mid y_{1:t}, S_{1:t-1}) \cdot p(S_{1:t-1} \mid y_{1:t}) \\ &= p(S_t \mid y_{1:t}, S_{1:t-1}) \cdot \frac{f(y_t, S_{1:t-1} \mid y_{1:t-1})}{f(y_t \mid y_{1:t-1})} \\ &= p(S_t \mid y_{1:t}, S_{1:t-1}) \cdot \frac{f(y_t \mid y_{1:t-1}, S_{1:t-1})}{f(y_t \mid y_{1:t-1})} \cdot p(S_{1:t-1} \mid y_{1:t-1}) \quad (1.4.5) \\ &= \frac{f(y_t \mid y_{1:t-1}, S_{1:t})}{f(y_t \mid y_{1:t-1})} \cdot p(S_t \mid S_{t-1}) \cdot p(S_{1:t-1} \mid y_{1:t-1}). \quad (1.4.6) \end{aligned}$$

En d'autres mots, nous voulons garder les trajectoires passées intact et y ajouter des valeurs simulées de S_t. En se basant sur l'expression (1.4.5), ceci peut être accompli en deux étapes. Premièrement, les poids des particules $\{S_{1:t-1}^{(i)}, \bar{w}_{t-1}^{(i)}\}_{i=1}^{N}$ sont ajustés avec le principe de l'échantillonnage préférentiel afin de refléter la connaissance de l'observation y_t :

$$\{S_{1:t-1}^{(i)}, \bar{w}_{t}^{(i)}\}_{i=1}^{N} \sim p(S_{1:t-1} \mid y_{1:t}),$$

où,

$$\bar{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}, \quad i = 1, \ldots, N,$$

et

$$\begin{aligned} w_t^{(i)} &= \frac{f(y_t \mid y_{1:t-1}, S_{1:t-1}^{(i)})}{f(y_t \mid y_{1:t-1})} \cdot w_{t-1}^{(i)} \\ &\propto f(y_t \mid y_{1:t-1}, S_{1:t-1}^{(i)}) \cdot w_{t-1}^{(i)}, \quad i = 1, \ldots, N. \quad (1.4.7) \end{aligned}$$

Deuxièmement, les états, $\{S_t^{(i)}\}_{i=1}^{N}$, sont générés selon

$$p(S_t \mid y_{1:t}, S_{1:t-1}^{(i)}) = \frac{f(y_t \mid y_{1:t-1}, S_{1:t-1}^{(i)}, S_t) \cdot p(S_t \mid S_{t-1}^{(i)})}{f(y_t \mid y_{1:t-1}, S_{1:t-1}^{(i)})}, \quad i = 1, \ldots, N, \quad (1.4.8)$$

ce qui nous permet d'obtenir l'échantillon désiré,

$$\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N \sim p(S_{1:t} \mid y_{1:t}).$$

Cette façon de procéder est dite « conditionnellement optimale » puisque chaque variable S_t est simulée selon sa loi exacte étant donné $\{y_{1:t}, S_{1:t-1}\}$. Lorsque la variable non-observée est discrète, il est possible d'implanter une telle approche parce que les expressions (1.4.7) et (1.4.8) peuvent être calculées analytiquement. Par exemple, nous avons

$$f(y_t \mid y_{1:t-1}, S_{1:t-1}^{(i)}) = \sum_{j=1}^{M} f(y_t \mid y_{1:t-1}, S_{1:t-1}^{(i)}, S_t = j) \cdot p(S_t = j \mid S_{t-1}^{(i)}),$$

et il est simple de simuler les états, $\{S_t^{(i)}\}_{i=1}^N$, car l'expression (1.4.8) peut être calculée explicitement pour chaque valeur de $S_t = 1, \ldots, M$. Cependant, lorsque la variable non-observée est continue, il faut généralement utiliser une approche plus générale pour étendre l'échantillon $\{S_{1:t-1}^{(i)}, \bar{w}_{t-1}^{(i)}\}_{i=1}^N$ au temps t.

1.4.2.2. *Approche générale*

Comme précédemment, supposons que nous détenons l'échantillon suivant au temps $t-1$:

$$\{S_{1:t-1}^{(i)}, \bar{w}_{t-1}^{(i)}\}_{i=1}^N \sim p(S_{1:t-1} \mid y_{1:t-1}).$$

Une approche générale pour prolonger cet échantillon au temps t consiste, en premier lieu, à se baser sur l'équation (1.4.6) et simuler les états, $\{S_t^{(i)}\}_{i=1}^N$, selon

$$p(S_t \mid S_{t-1}^{(i)}), \quad i = 1, \ldots, N, \tag{1.4.9}$$

qui est typiquement connu. Cette étape nous fournit les particules

$$\{S_{1:t}^{(i)}, \bar{w}_{t-1}^{(i)}\}_{i=1}^N \sim p(S_{1:t} \mid y_{1:t-1}).$$

En second lieu, les poids de ces particules sont ajustés avec le principe de l'échantillonnage préférentiel afin d'obtenir un échantillon

$$\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N \sim p(S_{1:t} \mid y_{1:t}),$$

où,

$$\bar{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}, \quad i = 1, \ldots, N,$$

et

$$\begin{aligned} w_t^{(i)} &= \frac{f(y_t \mid y_{1:t-1}, S_{1:t}^{(i)})}{f(y_t \mid y_{1:t-1})} \cdot w_{t-1}^{(i)} \\ &\propto f(y_t \mid y_{1:t-1}, S_{1:t}^{(i)}) \cdot w_{t-1}^{(i)}. \end{aligned}$$

Cette approche est moins efficace que celle dite « conditionnellement optimale », car la simulation des états, $\{S_t^{(i)}\}_{i=1}^N$, ne tient pas compte de l'observation y_t. Cependant, elle est très générale, car pour l'implanter il faut seulement être en mesure de (i) simuler S_t selon $p(S_t \mid S_{t-1})$ et (ii) calculer l'expression $f(y_t \mid y_{1:t-1}, S_{1:t})$. Ces deux étapes peuvent être effectuées sans difficulté dans la grande majorité des modèles à changement de régimes et en variables d'état.

1.4.2.3. *Rééchantillonnage* (Resampling)

Les algorithmes présentés dans les sections 1.4.2.1 et 1.4.2.2 sont voués à l'échec, car ils consistent en l'application séquentielle de l'échantillonnage préférentiel sur une variable dont la dimensionnalité croît à chaque pas de temps. La proposition utilisée pour obtenir un échantillon de $p(S_{1:t} \mid y_{1:t})$ se détériore avec t et la variance des poids, $\{\bar{w}_t^{(i)}\}_{i=1}^N$, croît exponentiellement. Éventuellement, cette distribution est estimée avec une unique particule puisque les poids des autres particules convergent vers zéro, un problème connu sous le nom de *weight degeneracy* [voir 34, section 7.3.1].

Dans un article influent, Gordon et al. [82] proposent une solution à ce problème de la dégénérescence. Elle consiste à rééchantillonner (*resample*) les particules, $\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N$, proportionnellement à leurs poids. Ceci a pour effet de dupliquer les particules les plus probables et éliminer celles qui le sont moins, résultant en une réorganisation des particules dans la région d'intérêt de $p(S_{1:t} \mid y_{1:t})$. Après l'étape de rééchantillonnage, on obtient un nouvel ensemble de particules, $\{S_{1:t}^{(i)}, \bar{w}_t^{(i)} = 1/N\}_{i=1}^N$, avec des poids égaux.

Gordon et al. [82] ont proposé le premier filtre particulaire, dénommé *bootstrap filter*, qui correspond à l'algorithme de la section 1.4.2.2 combiné à une étape de rééchantillonnage des particules, $\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N$, à chaque pas de temps. D'autre part, l'algorithme de la section 1.4.2.1 combiné à une étape de rééchantillonnage des particules, $\{S_{1:t-1}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N$, est équivalent au *fully adapted auxiliary particle filter* introduit par Pitt et Shephard [147]. Dans ce cas, il est possible de rééchantillonner les particules avant de simuler les états, $\{S_t^{(i)}\}_{i=1}^N$, parce que les poids, $\{\bar{w}_t^{(i)}\}_{i=1}^N$, sont indépendants des valeurs de S_t. Ceci améliore généralement l'approximation offerte par les particules, $\{S_{1:t}^{(i)}, \bar{w}_t^{(i)}\}_{i=1}^N$, car l'étape de rééchantillonnage a pour effet de présélectionner les particules les plus intéressantes.

Malheureusement, le mécanisme de rééchantillonnage, bien que nécessaire pour éviter la dégénérescence du filtre particulaire, appauvrit la diversité des trajectoires et a donc un effet nuisible sur l'approximation de $p(S_{1:t} \mid y_{1:t})$. Pour $k \ll t$, toutes les particules au temps t sont éventuellement associées à la même trajectoire $S_{1:k}$ et l'approximation de $p(S_{1:k} \mid y_{1:t})$ est inappropriée. En revanche,

le filtre particulaire permet généralement d'estimer adéquatement la probabilité $p(S_{t-k+1:t} \mid y_{1:t})$ pour de petites valeurs de k.

1.4.3. Approximer l'algorithme EM

Nous avons vu dans la section 1.3.4 que l'algorithme EM dépend de la distribution a posteriori des régimes, $p(S \mid y, \theta^{(r-1)})$, par l'entremise de l'équation (1.3.6) à l'étape espérance (rappel : $S = S_{1:T} \in \{1, 2, \ldots, M\}^T$ et $y = y_{1:T} \in \mathbb{R}^T$). Il est impraticable de calculer cette fonction analytiquement pour de grandes valeurs de T, car son support discret contient M^T points qui est énorme, même à $T = 50$. Dans l'exemple présenté à la section 1.3.4.1 où l'algorithme EM a été dérivé pour le modèle RSLN, l'équation (1.3.6) se simplifiait et il était seulement nécessaire de calculer les distributions marginales $p(S_t \mid y, \theta^{(r-1)})$ et $p(S_{t-1:t} \mid y, \theta^{(r-1)})$ pour implanter l'algorithme. Ce type de simplification est possible lorsque la condition (1.3.3) est respectée et, dans ce cas, le *forward filtering–backward smoothing algorithm* nous permet de calculer ces distributions marginales. Toutefois, lorsque la distribution conditionnelle de y_t dépend de toute la trajectoire des régimes, une telle simplification n'est plus possible et l'algorithme EM doit être approximé. Ceci peut être accompli en simulant n trajectoires de S, dénotées par $\{S^{(i)}\}_{i=1}^n$, distribuées selon $p(S \mid y, \theta^{(r-1)})$, et en posant :

$$\begin{aligned} Q(\theta \mid \theta^{(r-1)}) &= \mathrm{E}[\log f(y, S \mid \theta) \mid y, \theta^{(r-1)}] \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \log f(y, S^{(i)} \mid \theta). \end{aligned} \tag{1.4.10}$$

Cette approximation de l'algorithme EM correspond à l'algorithme Monte Carlo EM introduit par Wei et Tanner [161]. Les sections 1.4.3.1 à 1.4.3.3 présenteront trois méthodes permettant d'approximer l'algorithme EM dans le cas où la distribution conditionnelle de y_t dépend de toute la trajectoire des régimes. Dans ce qui suit, la dépendance d'une loi sur le paramètre θ est omis.

1.4.3.1. *Échantillonnage de Gibbs à un pas* (Single-move Gibbs sampler)

L'échantillonnage de Gibbs à un pas (*single-move Gibbs sampler*) est une méthode Monte Carlo par chaînes de Markov (*Markov chain Monte Carlo*) proposée par Albert et Chib [1] et Robert et al. [153] dans le cadre des modèles à changement de régimes. L'algorithme débute avec une trajectoire initiale, $S^{(0)}$, et génère un échantillon de trajectoires dépendantes, $\{S^{(i)}\}_{i=1}^n$, en simulant chaque régime séquentiellement selon la distribution, $p(S_t \mid y, S_{1:t-1}, S_{t+1:T})$, dénommée *full conditional distribution*. Ces trajectoires représentent des réalisations d'une chaîne de Markov ergodique admettant la distribution stationnaire $p(S \mid y)$ [voir

74, section 3.4.1].

Échantillonnage de Gibbs à un pas

Pour $i = 1, \ldots, n$:

I. Simuler $S_1^{(i)} \sim p(S_1 \mid y, S_{2:T}^{(i-1)})$.

II. Pour $t = 2, \ldots, T-1$: Simuler $S_t^{(i)} \sim p(S_t \mid y, S_{1:t-1}^{(i)}, S_{t+1:T}^{(i-1)})$.

III. Simuler $S_T^{(i)} \sim p(S_T \mid y, S_{1:T-1}^{(i)})$.

Le théorème ergodique [152, section 7.2] implique la convergence presque sûre,

$$\frac{1}{n}\sum_{i=1}^{n} h(S^{(i)}) \longrightarrow E[h(S) \mid y], \tag{1.4.11}$$

quand $n \to \infty$, où $h(\cdot) \in \mathbb{R}$. Par conséquent, les trajectoires obtenues avec l'échantillonnage de Gibbs à un pas, bien que dépendantes, peuvent être utilisées pour approximer l'étape espérance de l'algorithme EM (voir l'équation (1.4.10)).

Un avantage de cet algorithme est sa généralité, puisqu'il peut être implanté lorsque la distribution de y_t dépend de toute la trajectoire des régimes. Par exemple, il est simple de simuler S_t selon $p(S_t \mid y, S_{1:t-1}, S_{t+1:T})$, car l'expression,

$$\begin{aligned} p(S_t \mid y, S_{1:t-1}, S_{t+1:T}) &\propto f(y \mid S) \cdot p(S) \\ &\propto f(y_{t:T} \mid y_{1:t-1}, S) \cdot p(S_t \mid S_{t-1}) \cdot p(S_{t+1} \mid S_t) \\ &\propto p(S_t \mid S_{t-1}) \cdot p(S_{t+1} \mid S_t) \cdot \prod_{k=t}^{T} f(y_k \mid y_{1:k-1}, S_{1:k}), \end{aligned}$$

peut être calculée pour chaque valeur de $S_t = 1, \ldots, M$.

Un désavantage de l'échantillonnage de Gibbs à un pas est que les trajectoires simulées, $\{S^{(i)}\}_{i=1}^{n}$, peuvent être fortement corrélées. Par conséquent, la convergence de ces trajectoires vers la distribution stationnaire peut être lente, résultant en un estimateur imprécis à l'équation (1.4.11).

1.4.3.2. *Échantillonnage de Gibbs à pas multiples* (Multi-move Gibbs sampler)

L'échantillonnage de Gibbs à pas multiples (*multi-move Gibbs sampler*), aussi connu sous l'appellation *forward filtering–backward sampling algorithm*, a été développé par Carter et Kohn [35] et Frühwirth-Schnatter [73] dans le cadre des modèles en variables d'état linéaires et gaussiens et par Chib [38] pour les modèles à changement de régimes. Cet algorithme permet d'obtenir un échantillon indépendant de n trajectoires, $\{S^{(i)}\}_{i=1}^{n}$, distribuées selon $p(S \mid y)$ et repose sur

la factorisation suivante :

$$p(S \mid y) = p(S_T \mid y) \cdot p(S_{T-1} \mid y, S_T) \cdots p(S_t \mid y, S_{t+1:T}) \cdots p(S_1 \mid y, S_{2:T}).$$

Échantillonnage de Gibbs à pas multiples

Pour $i = 1, \ldots, n$:

I. Simuler $S_T^{(i)} \sim p(S_T \mid y)$.

II. Pour $t = T-1, \ldots, 1$: Simuler $S_t^{(i)} \sim p(S_t \mid y, S_{t+1:T}^{(i)})$.

L'avantage de l'échantillonnage de Gibbs à pas multiples vis-à-vis celui à un pas est que les trajectoires générées sont indépendantes. Cependant, son utilité est limitée aux situations où il est possible de simuler à partir de $p(S_t \mid y, S_{t+1:T})$, où

$$\begin{aligned} p(S_t \mid y, S_{t+1:T}) &= \frac{f(y, S_{t:T})}{f(y, S_{t+1:T})} \\ &= \frac{f(y_{1:t}) \cdot p(S_t \mid y_{1:t}) \cdot p(S_{t+1} \mid S_t) \cdot f(y_{t+1:T}, S_{t+2:T} \mid y_{1:t}, S_{t:t+1})}{f(y_{1:t}) \cdot p(S_{t+1} \mid y_{1:t}) \cdot f(y_{t+1:T}, S_{t+2:T} \mid y_{1:t}, S_{t+1})} \\ &= \frac{p(S_t \mid y_{1:t}) \cdot p(S_{t+1} \mid S_t)}{p(S_{t+1} \mid y_{1:t})} \cdot \frac{f(y_{t+1:T} \mid S_{t:T}, y_{1:t})}{f(y_{t+1:T} \mid S_{t+1:T}, y_{1:t})} \qquad (1.4.12) \\ &= \text{terme 1} \ \cdot \ \text{terme 2}. \end{aligned}$$

Quand la condition (1.3.3) est satisfaite, c'est-à-dire, la distribution conditionnelle de y_t dépend seulement du régime au temps t, le terme 2 dans l'équation (1.4.12) s'annule et on obtient l'expression,

$$\begin{aligned} p(S_t \mid y, S_{t+1:T}) &= p(S_t \mid y_{1:t}, S_{t+1}) \\ &= \frac{p(S_t \mid y_{1:t}) \cdot p(S_{t+1} \mid S_t)}{\sum_{j=1}^{M} p(S_t = j \mid y_{1:t}) \cdot p(S_{t+1} \mid S_t = j)}, \end{aligned}$$

qui peut être calculée explicitement pour chaque valeur de $S_t = 1, ..., M$. Dans ce cas, l'échantillonnage de Gibbs à pas multiples requiert le calcul des probabilités filtrées, $p(S_t \mid y_{1:t})$, $t = 1, \ldots, T$, et il doit donc être précédé d'un algorithme de filtrage avant, tel que le filtre d'Hamilton, justifiant ainsi l'appellation *forward filtering–backward sampling algorithm.*

Quand la distribution conditionnelle de y_t dépend de toute la trajectoire des régimes du temps 1 à t, il est plus difficile de simuler à partir de $p(S_t \mid y, S_{t+1:T})$. Dans de telles situations, il est possible d'utiliser un filtre particulaire pour faciliter cette simulation grâce à la décomposition :

$$
\begin{aligned}
p(S_t \mid y, S_{t+1:T}) &= \sum_{S_{1:t-1}} p(S_{1:t} \mid y, S_{t+1:T}) \\
&\propto \sum_{S_{1:t-1}} f(y, S) \\
&= \sum_{S_{1:t-1}} \Big\{ f(y_{1:t-1}) \cdot p(S_{1:t-1} \mid y_{1:t-1}) \cdot p(S_{t:T} \mid y_{1:t-1}, S_{1:t-1}) \cdot \\
&\qquad\qquad f(y_{t:T} \mid y_{1:t-1}, S_{1:T}) \Big\} \\
&\propto p(S_{t+1} \mid S_t) \cdot \sum_{S_{1:t-1}} \Big\{ \Big[\prod_{k=t}^{T} f(y_k \mid y_{1:k-1}, S_{1:k}) \Big] \cdot p(S_t \mid S_{t-1}) \cdot \\
&\qquad\qquad p(S_{1:t-1} \mid y_{1:t-1}) \Big\} \\
&\approx p(S_{t+1} \mid S_t) \cdot \sum_{i=1}^{N} \Big\{ \Big[\prod_{k=t}^{T} f(y_k \mid y_{1:k-1}, S^{(i)}_{1:t-1}, S_{t:k}) \Big] \cdot p(S_t \mid S^{(i)}_{t-1}) \cdot \bar{w}^{(i)}_{t-1} \Big\},
\end{aligned}
$$

où,

$$
\{S^{(i)}_{1:t-1}, \bar{w}^{(i)}_{t-1}\}_{i=1}^{N} \sim p(S_{1:t-1} \mid y_{1:t-1}).
$$

Bien que la simulation selon $p(S_t \mid y, S_{t+1:T})$ est possible à l'aide d'un filtre particulaire, elle est beaucoup plus coûteuse numériquement que celle selon $p(S_t \mid y, S_{1:t-1}, S_{t+1:T})$. Pour cette raison, l'échantillonnage de Gibbs à pas multiples n'est pas nécessairement plus efficace que celui à un pas.

1.4.3.3. *Lissage à délai fixe* (Fixed-lag smoothing)

Au lieu de simuler des trajectoires distribuées selon $p(S \mid y)$ pour approximer l'équation (1.3.6) de l'algorithme EM, Olsson et al. [144] proposent d'utiliser une méthode de lissage à délai fixe (*fixed-lag smoothing*). Cette section introduit une généralisation de cette technique à des modèles où la distribution de y_t dépend de toute la trajectoire des régimes du temps 1 à t.

L'expression (1.3.6) à l'étape espérance de l'algorithme EM se décompose de la manière suivante :

$$
\begin{aligned}
\sum_{S} \log f(y, S) \cdot p(S \mid y) &= \sum_{S} \log f(y \mid S) \cdot p(S \mid y) + \sum_{S} \log p(S) \cdot p(S \mid y) \\
&= \sum_{t=1}^{T} \sum_{S_{1:t}} \log f(y_{1:t} \mid y_{1:t-1}, S_{1:t}) \cdot p(S_{1:t} \mid y) +
\end{aligned}
$$

$$\sum_{t=1}^{T} \sum_{S_{t-1:t}} \log p(S_t \mid S_{t-1}) \cdot p(S_{t-1:t} \mid y),$$

et on est donc intéressé à approximer une fonction ayant la forme,

$$\sum_{t=1}^{T} \mathrm{E}[h(S_{1:t}) \mid y] = \sum_{t=1}^{T} \sum_{S_{1:t}} h(S_{1:t}) \cdot p(S_{1:t} \mid y).$$

Le principe de lissage à délai fixe consiste à estimer $\mathrm{E}[h(S_{1:t}) \mid y]$, séquentiellement pour $t = 1, \ldots, T$, à l'aide d'un filtre particulaire et la relation,

$$\begin{aligned} \mathrm{E}[h(S_{1:t}) \mid y] &\approx \mathrm{E}[h(S_{1:t}) \mid y_{1:t+k}] \qquad (1.4.13) \\ &= \sum_{S_{1:t}} h(S_{1:t}) \cdot p(S_{1:t} \mid y_{1:t+k}) \\ &= \sum_{S_{1:t+k}} h(S_{1:t}) \cdot p(S_{1:t+k} \mid y_{1:t+k}) \\ &\approx \sum_{i=1}^{N} h(S_{1:t}^{(i)}) \cdot \bar{w}_{t+k}^{(i)}, \qquad (1.4.14) \end{aligned}$$

où,

$$\{S_{1:t+k}^{(i)}, \bar{w}_{t+k}^{(i)}\}_{i=1}^{N} \sim p(S_{1:t+k} \mid y_{1:t+k}),$$

et k est un entier choisi de tel sorte que l'approximation à l'équation (1.4.14) est satisfaisante. Par conséquent, l'espérance, $\mathrm{E}[h(S_{1:t}) \mid y]$, est estimée avec les particules au temps $t+k$. La motivation derrière cette approximation est que dans la grande majorité des modèles en variables d'état et à changement de régimes, les observations recueillies après le temps $t + k$ contiennent peu d'information au sujet de l'état au temps t, justifiant ainsi la troncation à l'équation (1.4.13). Cette approximation s'améliore avec k, lorsque $k \ll T$, mais si k est trop grand, elle se détériore compte tenu du problème de la dégénérescence du filtre particulaire (voir la section 1.4.2.3). En effet, si k est grand, les trajectoires, $\{S_{1:t}^{(i)}\}_{i=1}^{N}$, associées aux particules, $\{S_{1:t+k}^{(i)}, \bar{w}_{t+k}^{(i)}\}_{i=1}^{N}$, auront peu de diversité. Le choix optimal de k découle donc d'un compromis entre la qualité de l'approximation à l'équation (1.4.13) et l'atténuation du problème de la dégénérescence.

Le principal désavantage de la méthode à délai fixe est qu'il est difficile de déterminer un k optimal et ce choix est donc généralement fait de manière arbitraire. De plus, même si $N \to \infty$, l'approximation à l'équation (1.4.14) est biaisée. Par contre, cette méthode est très efficace d'un point de vue computationnel puisque l'étape espérance de l'algorithme EM est estimée avec une seule itération du filtre particulaire et ne nécessite pas de simulations additionnelles.

1.4.3.4. *Une remarque*

Le chapitre 2 propose d'approximer l'algorithme EM en utilisant l'échantillonnage de Gibbs à un pas dans le but de calculer l'estimateur du maximum de vraisemblance du modèle GARCH à changement de régimes. L'échantillonnage de Gibbs à pas multiples utilisé conjointement avec un filtre particulaire ainsi que la technique du lissage à délai fixe décrite à la section 1.4.3.3 sont deux alternatives qui pourraient également être exploitées pour approximer l'algorithme EM dans ce cas. Au meilleur de ma connaissance, ces deux approches n'ont pas encore été proposées ni implantées dans le cadre de modèles à changement de régimes où la distribution conditionnelle de la variable observée dépend de toute la trajectoire de la chaîne de Markov. En effet, les sections 1.4.3.2 et 1.4.3.3 n'introduisent pas seulement deux techniques existantes, mais elles offrent également une contribution en expliquant comment ces deux méthodes peuvent être adaptées pour estimer des modèles dépendants de toute la trajectoire des régimes. Ces techniques constituent donc des alternatives pouvant permettre d'améliorer l'efficacité de l'algorithme proposé au chapitre 2.

Chapitre 2

MAXIMUM LIKELIHOOD ESTIMATION OF THE MARKOV-SWITCHING GARCH MODEL

Cet ouvrage fait l'objet d'un article dont je suis le seul auteur. Il est publié dans la revue *Computational Statistics & Data Analysis* [9]. Je remercie l'éditeur associé anonyme du journal, deux arbitres anonymes, Mathieu Boudreault et Manuel Morales pour leurs nombreux commentaires et suggestions qui m'ont permis d'améliorer la présentation et le contenu de cet article. Je suis également reconnaissant à Brian Hartman et Eden Tsang qui m'ont indiqué des fautes de composition, ainsi qu'à Arnaud Dufays pour m'avoir fourni le jeu de données utilisé par Bauwens et al. [16].

Résumé

L'intérêt du modèle GARCH à changement de régimes est de combiner deux approches en économétrie financière dans le but d'offrir un meilleur ajustement aux données financières et de mieux refléter leurs dynamiques. L'estimation de ce modèle est une tâche difficile, car la fonction de vraisemblance dépend de toute la trajectoire des régimes de la chaîne de Markov. Cette difficulté a mené à des méthodes d'estimation basées sur une simplification du modèle ou à des techniques qui ne dépendent pas de la fonction de vraisemblance. Il n'existe pas de méthode pouvant calculer l'estimateur du maximum de vraisemblance sans recourir à une simplification du modèle. Une nouvelle approche est développée basée sur l'algorithme Monte Carlo EM et sur l'échantillonnage préférentiel permettant d'obtenir l'estimateur du maximum de vraisemblance et sa matrice de variance-covariance asymptotique. L'efficacité de cette méthode est démontrée sur des données simulées et empiriques.

Abstract

The Markov-switching GARCH model offers rich dynamics to model financial data. Estimating this path dependent model is a challenging task because exact computation of the likelihood is infeasible in practice. This difficulty led to estimation procedures either based on a simplification of the model or not dependent on the likelihood. There is no method available to obtain the maximum likelihood estimator without resorting to a modification of the model. A novel approach is developed based on both the Monte Carlo expectation-maximization algorithm and importance sampling to calculate the maximum likelihood estimator and asymptotic variance-covariance matrix of the Markov-switching GARCH model. Practical implementation of the proposed algorithm is discussed and its effectiveness is demonstrated in simulation and empirical studies.

Keywords : Markov-switching, GARCH, EM algorithm, importance sampling

2.1. Introduction

Financial time series exhibit complex statistical dynamics which are difficult to reproduce with stochastic models. These dynamics are often referred to as the stylized facts of financial data and include, among others, the heavy-tailed nature of the return distribution and volatility clustering [see 41]. The generalized autoregressive conditional heteroscedasticity (GARCH) class of models [62, 21] has been extensively used to model financial data as it offers an explicit way to model volatility. Markov-switching (MS) or regime-switching models have also attracted a lot of attention in the econometric literature since the seminal paper of Hamilton [90]. In MS models the return distribution at a given time depends on the state (or regime) of an unobserved Markov chain. The states of the Markov chain are often given an economic interpretation. For example, a regime with a negative mean return and high volatility may be associated with a state of financial distress in the economy.

Due to the popularity of MS and GARCH models, it is natural to combine these two approaches and consider a MS-GARCH model. The MS-GARCH model can be simply understood as a GARCH model where parameters depend on the state of an unobserved Markov chain. One way to justify such a combination is given by Lamoureux and Lastrapes [125] and Mikosch and Starica [135] who show that the high persistence observed in the variance of financial returns can be explained by time-varying GARCH parameters.

Hamilton and Susmel [93] were among the first authors to discuss the MS-GARCH model. They noted that the estimation of this path dependent model

is a challenging task because exact computation of the likelihood is infeasible in practice. This led some authors [56, 83, 86, 120] to propose estimating modified versions of the MS-GARCH model that circumvent the path dependence problem by maximum likelihood. Other authors suggested alternative estimation methods such as a generalized method of moments (GMM) procedure [70] and a Bayesian Markov chain Monte Carlo (MCMC) algorithm [18, 16]. To this date, there is no method available to obtain the maximum likelihood estimator (MLE) of the MS-GARCH model without resorting to a simplification of the model.

The objective and main contribution of this article is to develop a novel approach based on the Monte Carlo expectation-maximization (MCEM) algorithm [161] and the Monte Carlo maximum likelihood (MCML) method [76, 77] to estimate the MLE of the MS-GARCH model. The proposed algorithm requires simulations from the posterior distribution of the state vector. For this reason, it can be seen as a frequentist counterpart of the Bayesian MCMC method proposed by Bauwens et al. [18] in the sense that both algorithms build on the data augmentation technique [159]. A secondary contribution of this article is to show how the asymptotic variance-covariance matrix of the MLE can be estimated. This is relevant since Francq and Zakoïan [70] were not able to obtain the asymptotic standard errors of their GMM estimates due to numerical difficulties.

This paper is organized as follows. Section 2.2 defines the MS-GARCH model. Section 2.3 introduces the novel approach to calculate the MLE, proposes a procedure to approximate the asymptotic variance-covariance matrix of the MLE and discusses practical implementation of the algorithm. Section 2.4 demonstrates the effectiveness of the proposed method in a simulation study. Section 2.5 applies the estimation technique to daily and weekly log-returns on the S&P 500 index. Section 2.6 concludes and proposes avenues for further research. Moreover, Appendix A justifies the validity of the expectation-maximization (EM) algorithm when applied to the MS-GARCH model. Appendices B and C include a proof and some technical details related to the implementation of the algorithm.

2.2. The MS-GARCH model

2.2.1. Definition

Following Bauwens et al. [18] and Francq et al. [67], the MS-GARCH model can be defined by the following equations :

$$y_t = \mu_{S_t} + \sigma_t(S_{1:t})\eta_t, \tag{2.2.1}$$

$$\sigma_t^2(S_{1:t}) = \omega_{S_t} + \alpha_{S_t}\epsilon_{t-1}^2(S_{t-1}) + \beta_{S_t}\sigma_{t-1}^2(S_{1:t-1}), \tag{2.2.2}$$

$$\epsilon_{t-1}(S_{t-1}) = y_{t-1} - \mu_{S_{t-1}}, \tag{2.2.3}$$

where $t = 1, \ldots, T$. The return process $\{y_t\}$ is a discrete-time stochastic process taking values in $\mathbb{R}$ and η_t, $t = 1, \ldots, T$, are independent and identically distributed normal innovations with zero mean and unit variance. The return dynamics depend on an unobserved discrete-time process, $\{S_t\}$, which is an ergodic time-homogeneous Markov chain with M-dimensional discrete state space (i.e., S_t can take integer values from 1 to M). The $M \times M$ transition matrix of this Markov chain is defined by the transition probabilities $\{p_{ij} = \Pr(S_t = j \mid S_{t-1} = i)\}_{i,j=1}^{M}$. The processes $\{S_t\}$ and $\{\eta_t\}$ are assumed independent. The notations $y_{1:t}$ and $S_{1:t}$ are used to represent the vectors $(y_1, \ldots, y_t)$ and $(S_1, \ldots, S_t)$, respectively. The symbols μ_{S_t} and $\sigma_t(S_{1:t})$ denote measurable functions with respect to a σ-field generated by the random variable S_t and the random vector $(y_{1:t-1}, S_{1:t})$, respectively. Therefore, μ_{S_t} symbolizes the conditional mean of the return y_t, $\mathrm{E}(y_t \mid y_{1:t-1}, S_{1:t}) = \mathrm{E}(y_t \mid S_t)$, and $\sigma_t^2(S_{1:t})$ represents its conditional variance $\mathrm{Var}(y_t \mid y_{1:t-1}, S_{1:t})$. The vector $\theta = (\{\mu_i, \omega_i, \alpha_i, \beta_i\}_{i=1}^{M}, \{p_{ij}\}_{i,j=1}^{M})$ denotes the parameters of the model. To ensure positivity of the variance, the following constraints are required : $\omega_i > 0$, $\alpha_i \geq 0$ and $\beta_i \geq 0$, $i = 1, \ldots, M$. Since $\sum_{j=1}^{M} p_{ij} = 1$ for $i = 1, \ldots, M$, θ contains $(4M + M(M-1))$ free parameters. Conditions for stationarity and the existence of moments were studied by Bauwens et al. [18], Francq et al. [67] and Francq and Zakoïan [69].

2.2.2. Path dependence problem

The specification (2.2.1)–(2.2.3) causes difficulties in estimation since the conditional variance at time t depends on the entire regime path $S_{1:t}$. To emphasize this dependence, the notation $\sigma_t^2(S_{1:t})$ is used in equations (2.2.1)–(2.2.3), but to simplify it in what follows, σ_t^2 will be used to represent $\sigma_t^2(S_{1:t})$. Moreover, let y and S denote $y_{1:T}$ and $S_{1:T}$, respectively, and f (p) stand for a probability density (mass) function. The calculation of the likelihood of the observed returns, denoted by $f(y \mid \theta)$, can be accomplished by integrating out all possible regime

paths :

$$f(y \mid \theta) = \sum_S f(y, S \mid \theta) = \sum_S f(y \mid S, \theta) p(S \mid \theta)$$
$$= \sum_S \left[\prod_{t=1}^{T} \sigma_t^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_t - \mu_{S_t})^2}{2\sigma_t^2} \right) \right] p(S \mid \theta). \tag{2.2.4}$$

For large T, this integration is infeasible numerically as the summation in equation (2.2.4) contains M^T terms and quickly becomes very large. Even the estimation of the likelihood by brute force Monte Carlo (i.e., by simulating independent sequences of states from the underlying Markov chain) will fail since such estimators exhibit prohibitively large variances [see 45]. Nevertheless, as shown by Bauwens et al. [16], it is possible to obtain an accurate estimate of the log-likelihood by writing

$$\log f(y \mid \theta) = \log f(y_1 \mid \theta) + \sum_{t=1}^{T-1} \log f(y_{t+1} \mid y_{1:t}, \theta),$$

and estimating $f(y_{t+1} \mid y_{1:t}, \theta)$, $t = 1, \ldots, T-1$, sequentially with the aid of particle filters. Unfortunately, the estimate of the log-likelihood obtained with particle filters is not a continuous function of θ [see 145]. Therefore, this simulated log-likelihood is difficult to maximize with standard optimization routines.

2.2.3. The solution of Gray [83]

Gray [83] was the first to suggest a method to estimate the MS-GARCH model. Recognizing that the likelihood cannot be computed exactly in practice, he proposed to replace equations (2.2.2) and (2.2.3) with :

$$\sigma_t^2 = \omega_{S_t} + \alpha_{S_t} \epsilon_{t-1}^2 + \beta_{S_t} h_{t-1},$$
$$\epsilon_{t-1} = y_{t-1} - \mathrm{E}(y_{t-1} \mid y_{1:t-2}),$$

where $h_{t-1} = \mathrm{Var}(y_{t-1} \mid y_{1:t-2})$. The expression h_{t-1} has the effect of collapsing all of the possible conditional variances at time $t-1$ into a single value that does not depend on the regime path. As a consequence, the conditional distribution of y_t, $f(y_t \mid y_{1:t-1}, S_{1:t}, \theta)$, is now independent of $S_{1:t-1}$ and maximum likelihood estimation is tractable [see 92]. Dueker [56] and Klaassen [120] expanded on Gray's idea by using broader information sets to collapse variances. The quality of these approximations to estimate the MLE of the MS-GARCH model has not been investigated. Intuitively, Gray's method will be more reliable if regimes can be inferred accurately from the data. For example, this occurs when regimes are persistent (i.e., p_{ii}, $i = 1, \ldots, M$, are close to 1) and well differentiated in their parameters. However, even in this case, the simulation study in Section

2.4 suggests that Gray's method does not generate consistent estimates for the MS-GARCH model. Therefore, an alternative approach to compute the MLE is needed.

2.3. A novel approach to estimate the MS-GARCH model

2.3.1. MCEM algorithm

There are many situations in statistical inference where it is difficult to maximize the likelihood of the observed data directly. The EM algorithm is a technique designed to obtain the MLE of the observed data likelihood through an iterative procedure that does not require the computation of the likelihood. Instead, we must be able to calculate and maximize

$$\begin{aligned} Q(\theta \mid \theta') &= \mathrm{E}\left(\log[f(y, S \mid \theta)] \mid y, \theta'\right) \\ &= \sum_{S} \log[f(y, S \mid \theta)] p(S \mid y, \theta'). \end{aligned} \tag{2.3.1}$$

The complete data likelihood of the MS-GARCH model, $f(y, S \mid \theta)$, admits a simple expression (see equation (2.2.4)) but it is not possible to calculate (2.3.1) exactly because of the path dependence problem. Nonetheless, we can simulate from $p(S \mid y, \theta')$ using Gibbs sampling and obtain a Monte Carlo approximation of $Q(\theta \mid \theta')$. When the expectation step of the EM algorithm is approximated with Monte Carlo methods, we obtain the MCEM algorithm introduced by Wei and Tanner [161]. In such cases, the monotonicity property of the EM algorithm (i.e., the likelihood of the observed data is never decreased at each iteration) is not guaranteed to hold because of the Monte Carlo error introduced. Consequently, the specification of the number of simulated state vectors at each iteration of the algorithm is of central importance. Wei and Tanner [161] recommend that small values be used in the initial stages and that these values be increased as the algorithm moves closer to convergence. The validity of the EM algorithm in the context of the MS-GARCH model is discussed in Appendix A.

2.3.2. MCML algorithm

A common criticism of the EM algorithm is that, although it can reach the neighborhood of the MLE quickly, it exhibits slow linear convergence in the neighborhood itself [see 134, Section 3.9]. For this reason, it is often suggested to combine the EM algorithm with a Newton-Raphson method or to simply switch to a faster method after a few EM iterations. The latter was suggested by McCulloch

[132] who proposed to follow the MCEM algorithm with the MCML approach of Geyer [76, 77]. Suppose that $\{S^{(i)}\}_{i=1}^{n^\star}$ are simulated state vectors from $p(S \mid y, \theta^\star)$ and define

$$w_{\theta|\theta^\star}^{(i)} = \frac{f(y, S^{(i)} \mid \theta)}{f(y, S^{(i)} \mid \theta^\star)}, \quad i = 1, \ldots, n^\star. \tag{2.3.2}$$

The MCML algorithm makes use of importance sampling to directly maximize the log-likelihood through the following relation :

$$\log f(y \mid \theta) - \log f(y \mid \theta^\star) = \log \mathrm{E}\left[\frac{f(y, S \mid \theta)}{f(y, S \mid \theta^\star)} \,\middle|\, y, \theta^\star \right] \approx \log \frac{1}{n^\star} \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)}. \tag{2.3.3}$$

Due to the well-known problems related to importance sampling, Cappé et al. [33] explain that this method does not work well unless $\theta^\star$ is in a close neighborhood of the MLE. Similarly to McCulloch [132], they suggest using it only with another consistent maximum likelihood estimation method.

2.3.3. MCEM-MCML algorithm for the MS-GARCH model

The discussion in Section 2.3.2 suggests a hybrid MCEM-MCML algorithm. First, iterations of the MCEM algorithm can be performed to get a good estimate, $\theta^\star$, of the MLE. This estimate is then used to generate the importance sample in the MCML algorithm. Both algorithms complement each other : the MCEM algorithm addresses the flaw of the MCML algorithm relating to the choice of $\theta^\star$ while the MCML method replaces many potential MCEM iterations with a single iteration, leading to a faster convergence.

Given an initial guess of the parameter vector, $\theta^{(0)}$, the following algorithm started at $r = 1$ produces a sequence of parameter iterates, $\{\theta^{(r)}\}_{r \geq 1}$, allowing us to compute the MLE of model (2.2.1)–(2.2.3) :

Algorithm 1 : MCEM-MCML algorithm

1. Simulate n_r samples of the state vector S from $p(S \mid y, \theta^{(r-1)})$ using a single-move Gibbs sampler. The states are simulated sequentially for $t = 1, \ldots, T$ based on the following full conditional distribution :

$$p(S_t \mid S_{1:t-1}^{(i)}, S_{t+1:T}^{(i-1)}, y, \theta^{(r-1)}) \propto p_{S_{t-1}^{(i)}, S_t} \, p_{S_t, S_{t+1}^{(i-1)}} \prod_{j=t}^{T} \sigma_j^{-1} \exp\left(-\frac{(y_j - \mu_{S_j})^2}{2\sigma_j^2} \right). \tag{2.3.4}$$

   To ease notation, the expression $\sigma_j(S_{1:t})$ was reduced to σ_j. In the context of equation (2.3.4), σ_j represents $\sigma_j(S_{1:t-1}^{(i)}, S_t, S_{t+1:j}^{(i-1)})$. It is straightforward to sample S_t from (2.3.4) since S_t can only take integer values from 1 to M. However, it should be noted that it is not possible to compute expression (2.3.4) numerically for each value of S_t since this will result in underflow. To

avoid underflow, we can calculate the ratios of these expressions and then recover the probabilities for $S_t = 1, \ldots, M$ from them. The n_r simulations of the state vector S that are obtained are denoted by $\{S^{(i)}\}_{i=1}^{n_r}$. These draws form a Markov chain with $p(S \mid y, \theta^{(r-1)})$ as its stationary distribution [see 74, Section 3.4.1].

2. Monte Carlo E-step : Calculate $\widehat{Q}(\theta \mid \theta^{(r-1)})$, an approximation of the conventional E-step $Q(\theta \mid \theta^{(r-1)})$, where

$$\widehat{Q}(\theta \mid \theta^{(r-1)}) = \frac{1}{n_r} \sum_{i=1}^{n_r} \log[f(y, S^{(i)} \mid \theta)] \tag{2.3.5}$$

$$= -\frac{T \log(2\pi)}{2} - \frac{1}{2n_r} \sum_{t=1}^{T} \sum_{i=1}^{n_r} \left[\log(\sigma_t^{(i)})^2 + \frac{(y_t - \mu_{S_t^{(i)}})^2}{(\sigma_t^{(i)})^2} \right] + \frac{1}{n_r} \sum_{t=1}^{T} \sum_{i=1}^{n_r} \log(p_{S_{t-1}^{(i)}, S_t^{(i)}}) \tag{2.3.6}$$

$$= \text{term 1} + \text{term 2}.$$

In the previous expressions, $\sigma_t^{(i)}$ is shorthand for $\sigma_t(S_{1:t}^{(i)})$.

3. M-step : Perform the following maximization :

$$\theta^{(r)} = \arg\max_{\theta} \ \widehat{Q}(\theta \mid \theta^{(r-1)}).$$

This optimization can be split into two independent steps since terms 1 and 2 of equation (2.3.6) involve different subsets of the parameters. Term 1 includes the mean and GARCH parameters while term 2 only contains transition probabilities. Maximization of term 1 must be performed numerically and is similar to a standard GARCH optimization to calculate the MLE. To improve the performance of that optimization, the gradient of term 1 with respect to the mean and GARCH parameters should be provided to the optimization routine (see Appendix C). Maximization of term 2 can be done analytically. Term 2 is at its maximum when the transition probabilities take the values

$$p_{jk} = \frac{f_{jk}}{\sum_{l=1}^{M} f_{jl}}, \quad j, k = 1, \ldots, M,$$

where f_{jk} denotes the total number of transitions from state j to state k in all of the n_r simulated state vectors. A proof of this result is in Appendix B.

4. Apply a decision rule to determine whether to switch to the MCML algorithm (see Section 2.3.5.2). If the decision is to switch, go to step 5 and set $\theta^{\star} = \theta^{(r)}$. Otherwise, add 1 to r and go to step 1.

5. Simulate $n^\star$ samples of the state vector S from $p(S \mid y, \theta^\star)$ using the single-move Gibbs sampler described in step 1 of the algorithm to obtain the importance sample $\{S^{(i)}\}_{i=1}^{n^\star}$.

6. MCML-step : Perform the following maximization to obtain the MLE :

$$\hat{\theta} = \arg\max_{\theta} \left[\log \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)} \right]. \tag{2.3.7}$$

   In contrast to the M-step, this optimization cannot be split into two steps. Appendix C provides some details related to its implementation.

Using importance sampling, the final sample, $\{S^{(i)}\}_{i=1}^{n^\star}$, generated at step 5 of the algorithm can be transformed into a weighted sample, $\{S^{(i)}, \bar{w}_{\hat{\theta}|\theta^\star}^{(i)}\}_{i=1}^{n^\star}$, from $p(S \mid y, \hat{\theta})$, where $\bar{w}_{\hat{\theta}|\theta^\star}^{(i)} = w_{\hat{\theta}|\theta^\star}^{(i)} / \sum_{i=1}^{n^\star} w_{\hat{\theta}|\theta^\star}^{(i)}$, $i = 1, \ldots, n^\star$. This sample can be used to obtain an estimate of the smoothed inference of the state at time t, $p(S_t = j \mid y, \hat{\theta})$, $j = 1, \ldots, M$, with $\sum_{i=1}^{n^\star} \bar{w}_{\hat{\theta}|\theta^\star}^{(i)} \mathbf{1}_{\{S_t^{(i)}=j\}}$ or to compute the asymptotic variance-covariance matrix of the MLE.

2.3.4. Asymptotic variance-covariance matrix of the MLE

The asymptotic variance-covariance matrix of the MLE for the MS-GARCH model can be obtained from the inverse of the Fisher information matrix, denoted by $\mathcal{F}(\theta) = \mathrm{E}\left[\mathcal{S}(\theta)\mathcal{S}(\theta)^T\right]$, where $\mathcal{S}(\theta) = \partial \log f(y \mid \theta)/\partial\theta$ is the score related to the observed data log-likelihood. The expectation is taken over y and $\mathcal{F}(\theta)$ must be evaluated at the true parameter values (not the MLE). We may approximate $\mathcal{F}(\theta)$ by generating independent realizations of y and averaging $\mathcal{S}(\theta)\mathcal{S}(\theta)^T$ over all of these realizations. Unfortunately, it is not possible to calculate the score exactly, but we may approximate it using the following relation [see 134, Section 3.7] :

$$\mathcal{S}(\theta) = \left[\frac{\partial}{\partial\theta'} Q(\theta' \mid \theta) \right]_{\theta'=\theta} \approx \left[\frac{\partial}{\partial\theta'} \widehat{Q}(\theta' \mid \theta) \right]_{\theta'=\theta}, \tag{2.3.8}$$

where $\partial\widehat{Q}(\theta' \mid \theta)/\partial\theta'$ is the gradient of the Monte Carlo E-step which is available in closed form (see Appendix C). An alternative to approximate the score is to use the MCML relation in expression (2.3.3). If $\theta^\star = \theta$, this alternative is equivalent to using expression (2.3.8). The procedure just described generates a valid approximation of the asymptotic variance-covariance matrix of the MLE when the true parameters of the MS-GARCH model are known. This is the method that is used to obtain the asymptotic standard errors of the MLE in the simulation study of Section 2.4.

When a real data set is fitted to the MS-GARCH model, the true parameters are not known and we have two choices to approximate the variance-covariance

matrix of the MLE. First, we may replace the true parameters with the MLE in the method just mentioned. However, it is common practice to estimate the variance-covariance matrix with the inverse of $\mathcal{I}(\hat{\theta})$, the observed information matrix evaluated at the MLE. For the MS-GARCH model, $\mathcal{I}(\hat{\theta})$ can be approximated based on the decomposition of the observed information matrix presented by Louis [129] [see also 134, Sections 4.2.2 and 6.3.5] :

$$\begin{aligned}\mathcal{I}(\hat{\theta}) &= \left[-\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(y \mid \theta)\right]_{\theta=\hat{\theta}} \\ &= \left[-\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta \mid \hat{\theta})\right]_{\theta=\hat{\theta}} - \mathrm{E}\left[\mathcal{S}_c(\hat{\theta})\mathcal{S}_c(\hat{\theta})^T \mid y, \hat{\theta}\right] + \mathcal{S}(\hat{\theta})\mathcal{S}(\hat{\theta})^T, \qquad (2.3.9)\end{aligned}$$

where $\mathcal{S}_c(\hat{\theta}) = [\partial \log f(y, S \mid \theta)/\partial\theta]_{\theta=\hat{\theta}}$ is the score related to the complete data log-likelihood evaluated at the MLE. Since $\mathcal{S}(\hat{\theta})$ will be very close to the zero vector in practice, it is only the first two terms in expression (2.3.9) that need to be estimated. This approximation can be performed using the final state vectors generated at step 5 of the MCEM-MCML algorithm. As mentioned at the end of Section 2.3.3, $\{S^{(i)}, \bar{w}^{(i)}_{\hat{\theta}|\theta^\star}\}_{i=1}^{n^\star}$ is a weighted sample from $p(S \mid y, \hat{\theta})$. Hence, we can write :

$$\begin{aligned}\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta \mid \hat{\theta}) &= \sum_S \left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log[f(y, S \mid \theta)]\right] p(S \mid y, \hat{\theta}) \\ &\approx \sum_{i=1}^{n^\star}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log[f(y, S^{(i)} \mid \theta)]\right]\bar{w}^{(i)}_{\hat{\theta}|\theta^\star}. \qquad (2.3.10)\end{aligned}$$

Expression (2.3.10) can be computed by numerical differentiation of

$$\sum_{i=1}^{n^\star}\left[\frac{\partial}{\partial\theta}\log[f(y, S^{(i)} \mid \theta)]\right]\bar{w}^{(i)}_{\hat{\theta}|\theta^\star},$$

which is available in closed form (see Appendix C). Moreover, the second term in equation (2.3.9) can be estimated with

$$\mathrm{E}\left[\mathcal{S}_c(\theta)\mathcal{S}_c(\theta)^T \mid y, \hat{\theta}\right] \approx \sum_{i=1}^{n^\star}\left[\frac{\partial}{\partial\theta}\log[f(y, S^{(i)} \mid \theta)]\right]\left[\frac{\partial}{\partial\theta}\log[f(y, S^{(i)} \mid \theta)]\right]^T \bar{w}^{(i)}_{\hat{\theta}|\theta^\star}.$$

Therefore, the estimation of the variance-covariance matrix of the MLE with $[\mathcal{I}(\hat{\theta})]^{-1}$ can be obtained as a by-product of the MCEM-MCML algorithm.

2.3.5. Practical considerations with regard to the MCEM-MCML algorithm

Section 2.3.3 shows how the MCEM and MCML algorithms can be used to estimate the MLE of the MS-GARCH model but it does not convey any guidance

on how to choose the number of simulated state vectors at each iteration, nor does it tell us when to switch to the MCML algorithm. Starting values for the algorithm and the Gibbs sampler are also of concern.

2.3.5.1. *Starting values*

For the MCEM algorithm and the Gibbs sampler, starting values are important since they can influence convergence. Section 2.4 demonstrates that the MLE of Gray's model (see Section 2.2.3) provides a good initial value for the MCEM-MCML algorithm. Nevertheless, it is always prudent to try a few additional starting points to avoid suboptimal convergence (see Appendix A). For example, the transition probabilities and mean parameters can be set equal to those estimated for a basic MS model while the volatility parameters can be derived from a fitted standard GARCH model. The approximations of Dueker [56] or Klaassen [120] can also be considered, but they will generally yield initial values in the same range as Gray's model.

To initialize the Gibbs sampler at the first iteration of the MCEM algorithm we require a vector of states. This can be obtained from smoothed inferences of the states using Gray's model which can be calculated recursively [see 91, p. 694]. For instance, for each $t = 1, \ldots, T$, we can compute the smoothed inference $p(S_t \mid y, \theta^{(0)})$ and pick the value of S_t which is most likely. This collection of states can form the initial state vector. Moreover, from one iteration of the MCEM algorithm to the next we may reuse the last state vector generated on a given iteration as the starting state vector on the next iteration. Since a reasonable starting value for the state vector is available, it is not strictly necessary to use a burn-in sample [see 111].

A final technical detail is the choice of S_0. To generate the first state of the Markov chain S_1, we need to make an assumption about S_0. We have at least two choices. First, we may assume that the Markov chain $\{S_t\}$ is stationary and therefore S_1 can be generated according to the stationary distribution of the Markov chain. Second, we may treat the initial distribution of S_0, say the vector δ, as a parameter of the model and estimate it. However, maximizing the likelihood over δ is equivalent to maximizing the likelihood conditionally on starting from the M possible different states. In other words, the MLE of δ is a unit vector of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$ [see 169, Section 4.2.4]. For simplicity and since this assumption does not play a material role in estimation, I will suppose that the initial state S_0 is given and fixed instead of needing estimation.

2.3.5.2. *Simulation schedule*

A strategy for increasing the sample size throughout the MCEM-MCML algorithm is proposed in this section and referred to as the simulation schedule of the algorithm. To determine an appropriate simulation schedule, I applied the MCEM-MCML algorithm with an excessive amount of simulations and iterations on the MS-GARCH model considered in Section 2.4 and on the empirical data studied in Section 2.5. Afterwards, I reduced the number of simulations and compared the ending parameter vectors to the ones obtained by brute force.

TABLE 2.1. Simulation schedule 1

n_1	n_2	n_3	n_4	$n^\star$
500	1000	2500	5000	10000

TABLE 2.2. Simulation schedule 2

n_1 to n_{10}	n_{11} to n_{28}	n_{29}	n_{30}	$n^\star$
500	1000	2500	5000	40000

I recommend two simulation schedules for the MCEM-MCML algorithm (see Tables 2.1 and 2.2). Simulation schedule 1 permits a fast estimation (see Section 2.4.4 for computational times) while simulation schedule 2 puts more emphasis on accuracy and is more robust with respect to the choice of starting values. In the simulation study (Section 2.4), simulation schedule 1 is used since simulation schedule 2 does not offer significant improvements in the estimation process. However, when considering empirical data in Section 2.5, Simulation schedule 2 is preferred because it yields important gains in precision. For example, different runs of the MCEM-MCML algorithm conducted with this strategy brought the parameter iterates to a close neighborhood of the ones obtained by brute force (see Section 2.5.1).

The effectiveness of a given simulation schedule depends on starting values and on the data set of observed returns. Consequently, it may be worthwhile to automate the MCEM-MCML algorithm, i.e., dictate rules that automatically select the sample size at each iteration and determine when to perform the final MCML iteration. For example, we may want to switch to the MCML algorithm if the relative differences between successive MCEM parameter iterates are below a certain threshold. There are at least two good papers dealing with the automation of the MCEM algorithm when MCMC samples are used. On one hand, Levine and Fan [127] propose to select the sample size at each iteration based on a

confidence ellipsoid created around the parameter iterates. On the other hand, Caffo et al. [28] base that decision on whether the Monte Carlo E-step of the algorithm actually increased the likelihood with high probability. Both of these approaches rely on asymptotic results and require subsampling approximately independent subsets of the MCMC sample of state vectors. As a consequence, it is not guaranteed that they can be used with high reliability. Additionally, a certain amount of manual adjustments will always be necessary. For these reasons, automated strategies were not implemented. Nevertheless, it would be interesting to see how these strategies compare to the ones that were recommended.

2.4. Simulation study

This section evaluates the effectiveness of the proposed MCEM-MCML algorithm to compute the MLE of the MS-GARCH model. The algorithm was coded to work with version 2.15.0 of the R software [150]. The Gibbs sampler involves nested sequential loops which render the generation of states a slow process with R. Consequently, I used the Rcpp package [59] available with R to implement the generation of states in C++ (the Rcpp package simplifies the integration of C++ code with R). All results presented in this section are reproducible since I seeded the random number generator in R. The code to reproduce these results is available on my website.

2.4.1. Description

I simulated 200 independent trajectories of the MS-GARCH model with sizes of $T = 500$, 1500 and 5000 based on the parameter set presented in the third column of Table 2.3. All of the simulations were started assuming an initial state value of $S_0 = 1$ and an initial variance of $\sigma_0^2 = 2.56$ (this value is approximately equal to the unconditional variance of the process). This parameter set (denoted by BPR) was considered by Bauwens et al. [18] to assess their Bayesian MCMC algorithm. Bauwens et al. [18] state that their choice of parameters is inspired by empirical studies. It is thus a reasonable representation of a process for generating financial returns. In particular, the first regime defines a positive mean return–low volatility environment while the second regime pertains to a negative mean return–high volatility state.

For each of the simulated trajectories of the MS-GARCH model, the MLE was estimated by applying the MCEM-MCML algorithm with simulation schedule 1 (see Table 2.1). To examine the effect of starting values, the iterations were started from the true parameter values and from the MLE of the model of Gray [83]. It is

not possible to compare the MCEM-MCML algorithm to the GMM algorithm of Francq and Zakoïan [70] on parameter set BPR because the GMM algorithm was developed on a restricted version of model (2.2.1)–(2.2.3) with zero means in all regimes. It seems difficult to extend this algorithm to the case of regime-dependent means or to more general versions of the MS-GARCH model.

2.4.2. Identifiability issues

MS models are generally not identifiable since the parameters of two states may be permuted without changing the likelihood. When Bayesian MCMC techniques are used to estimate MS models, the label-switching problem [see 74, Section 3.5.5] can seriously complicate statistical inference. This problem occurs because the posterior distribution of the parameters is invariant under a permutation of state indices when exchangeable priors are used. Therefore, the MCMC output is difficult to interpret when one needs to draw inferences about parameters in specific regimes. A similar problem occurs in a frequentist setting when we want to evaluate the sampling distribution of estimators in a simulation study. We must associate each estimated parameter to a state index since the objective is to assess biases and standard errors of the estimated parameters in a given regime. In contrast to the label-switching problem, this problem received very little attention [see 166]. The approach followed here to identify regimes is to impose an identifiability constraint of the form $\omega_1 < \omega_2 < \ldots < \omega_M$. This is a simple solution that works well here since the parameters ω_1 and ω_2 differ significantly between regimes in parameter set BPR. An alternative would be to impose a similar constraint on the means, but there is less of a clear-cut difference between the two regimes based on that choice. If the regimes are not well separated, such an approach may not correctly identify the estimated parameters with their corresponding regimes. More sophisticated approaches have been proposed to deal with the label-switching problem, but they are not always applicable in a frequentist setting.

It must be emphasized that when a data set is fitted with the MCEM-MCML algorithm, the label-switching problem is not encountered since the goal is to find a point estimate corresponding to one of the equivalent modes of the likelihood function. It also does not cause a problem for the estimation of the asymptotic variance-covariance matrix. This problem is only encountered when evaluating the sampling distribution of estimators.

2.4.3. Discussion of results

Table 2.3 displays the summary statistics for 200 estimates of the MLE obtained with Gray's model (denoted by Gray), the MCEM-MCML algorithm started from the MLE of Gray's model (denoted by MCEM_G) and the MCEM-MCML algorithm started from the true parameter values (denoted by MCEM_0) for each of the sample sizes considered ($T = 500$, 1500 and 5000). RMSE denotes the root-mean-square error and A-StErr stands for asymptotic standard error.

We can observe that the estimates of ω_1, ω_2, β_1 and β_2 based on Gray's model display significant biases which do not necessarily decrease as the sample size is increased. This suggests that Gray's model does not generate consistent estimates for the MS-GARCH model. However, it does provide a good starting value for the MCEM-MCML algorithm as results under columns MCEM_G and MCEM_0 are very close to each other. In fact, the MCEM-MCML algorithm started from the MLE of Gray's model or from the true parameter values generally converged to the same mode of the likelihood function.

The results shown in Table 2.3 corroborate that the MCEM-MCML algorithm started from the MLE of Gray's model is an effective method to estimate the MS-GARCH model. The following remarks support this assertion.

1. The RMSE and the A-StErr of the estimators always decrease when the sample size is increased.
2. The RMSE are, in general, close to the A-StErr, especially for a sample size of 5000.
3. When some bias is observed (e.g., parameters ω_2 and β_2), this bias always decreases when the sample size is increased.
4. The transition probabilities are estimated with high accuracy.

Even though the results seem satisfactory overall for the MCEM-MCML algorithm, the GARCH parameters in the second regime ω_2 and β_2 display some bias. To investigate this element, I generated 1000 trajectories with sizes of $T = 500$, 1500, 5000, 10000 and 50000 from a GARCH model with parameters matching those in the second regime. For each of these trajectories, the MLE of the GARCH model was estimated and the mean values of these estimates for each sample size are displayed in Table 2.4.

In Table 2.4, we can note a pattern that is analogous to that observed in Table 2.3 for the parameters of the second regime : the ω and β parameters are estimated with large bias for small sample sizes, but this bias is gradually decreased as the sample size is increased. This example demonstrates that the

TABLE 2.3. Mean and RMSE based on 200 estimates of the MLE

T		Value	A-StErr	Mean			RMSE		
				Gray	MCEM_G	MCEM_0	Gray	MCEM_G	MCEM_0
500	μ_1	0.06	0.039	0.061	0.059	0.060	0.059	0.075	0.070
	μ_2	−0.09	0.211	−0.080	−0.091	−0.093	0.216	0.204	0.212
	ω_1	0.30	0.096	0.329	0.298	0.300	0.092	0.091	0.091
	α_1	0.35	0.110	0.356	0.336	0.342	0.146	0.150	0.146
	β_1	0.20	0.162	0.106	0.211	0.201	0.184	0.197	0.182
	ω_2	2.00	1.539	3.684	2.786	2.630	2.324	1.852	1.720
	α_2	0.10	0.106	0.118	0.121	0.113	0.152	0.152	0.135
	β_2	0.60	0.279	0.354	0.425	0.469	0.391	0.365	0.339
	p_{11}	0.98	0.010	0.976	0.976	0.977	0.017	0.017	0.013
	p_{22}	0.96	0.023	0.952	0.954	0.954	0.028	0.029	0.026
1500	μ_1	0.06	0.025	0.062	0.061	0.061	0.024	0.023	0.023
	μ_2	−0.09	0.114	−0.087	−0.091	−0.093	0.111	0.114	0.115
	ω_1	0.30	0.054	0.345	0.300	0.301	0.069	0.058	0.059
	α_1	0.35	0.071	0.362	0.350	0.351	0.069	0.066	0.065
	β_1	0.20	0.091	0.081	0.201	0.199	0.131	0.103	0.103
	ω_2	2.00	1.006	3.733	2.304	2.325	1.987	1.193	1.171
	α_2	0.10	0.060	0.101	0.088	0.089	0.061	0.057	0.058
	β_2	0.60	0.187	0.376	0.562	0.559	0.269	0.198	0.192
	p_{11}	0.98	0.006	0.979	0.979	0.979	0.007	0.007	0.007
	p_{22}	0.96	0.011	0.957	0.958	0.958	0.015	0.014	0.015
5000	μ_1	0.06	0.012	0.062	0.061	0.061	0.014	0.013	0.013
	μ_2	−0.09	0.065	−0.081	−0.090	−0.090	0.061	0.061	0.061
	ω_1	0.30	0.032	0.351	0.300	0.301	0.057	0.028	0.027
	α_1	0.35	0.037	0.362	0.354	0.352	0.041	0.038	0.038
	β_1	0.20	0.052	0.072	0.198	0.196	0.131	0.050	0.049
	ω_2	2.00	0.563	3.638	2.051	2.060	1.721	0.554	0.532
	α_2	0.10	0.030	0.112	0.093	0.094	0.038	0.035	0.034
	β_2	0.60	0.101	0.385	0.599	0.597	0.234	0.103	0.098
	p_{11}	0.98	0.003	0.980	0.980	0.980	0.003	0.003	0.003
	p_{22}	0.96	0.006	0.958	0.959	0.959	0.007	0.007	0.006

TABLE 2.4. Mean values of 1000 estimates of the MLE for a GARCH model

T	μ	ω	α	β
500	-0.090	2.357	0.104	0.541
1500	-0.091	2.220	0.102	0.565
5000	-0.091	2.051	0.100	0.592
10000	-0.089	2.024	0.100	0.596
50000	-0.090	1.999	0.100	0.600
True value	-0.090	2.000	0.100	0.600

MLE of GARCH models can be biased in small sample sizes. It seems that the biases that are observed with regard to ω_2 and β_2 in Table 2.3 are normal and they are not due to a fault of the MCEM-MCML algorithm. In fact, an examination of the correlation matrix of the estimated parameters reveals that the estimates of ω_2 and β_2 are highly negatively correlated which undoubtedly complicates estimation.

2.4.4. Computational times

Although the MCEM-MCML algorithm is an iterative simulation-based algorithm, the estimation of a data set can be accomplished in a reasonable amount of time. For instance, it takes approximately 2, 6 and 30 minutes to estimate a data set with sample sizes of $T = 500$, 1500 and 5000, respectively, using simulation schedule 1 (see Table 2.1) on a 3.40 GHz Intel Core i7-2600 processor. The single-move Gibbs sampler implementation in C++ is efficient as it takes approximately 15, 70 and 500 seconds to generate 10000 draws from the posterior distribution of the state vector at sample sizes of $T = 500$, 1500 and 5000, respectively. The maximization step of the algorithm is computationally demanding because the conditional variance at time t, $\sigma_t^2(S_{1:t})$, must be recalculated for each simulated state sequence $S_{1:t}$ and each $t = 1, \ldots, T$ whenever the parameters change. This element justifies the use of the MCML algorithm to help reduce the number of iterations when convergence of the MCEM algorithm is slow.

2.5. EMPIRICAL STUDY

In this section, the MS-GARCH model is fitted to (i) weekly percentage log-returns on the S&P 500 price index from October 28, 1987 to October 31, 2012 and (ii) daily percentage log-returns on the S&P 500 price index from May 20, 1999 to April 25, 2011. Weekly data is from Wednesday's close to the following

Wednesday's close to avoid most holidays and includes 1305 observations. The daily data set contains 3000 observations and allows for a comparison of the estimation results obtained with the MCEM-MCML algorithm to the Bayesian MCMC approach developed by Bauwens et al. [16] since these authors considered the same data set. Descriptive statistics for the two financial time series are provided in Table 2.5 (the mean and standard deviation, abbreviated StDev, are given on an annualized basis). Similarly to Section 2.4, the estimation results presented in this section are reproducible with the code available on my website.

TABLE 2.5. Descriptive statistics

	Mean	StDev	Skewness	Kurtosis	Minimum	Maximum
Weekly S&P 500	7.2	16.6	−0.59	7.2	−16.5	10.2
Daily S&P 500	−0.1	21.5	−0.12	10.5	−9.5	11.0

2.5.1. Estimation results and effectiveness of simulation schedule 2

To demonstrate the effectiveness of simulation schedule 2 (see Table 2.2), the parameter estimation for each data set was repeated 40 times using that strategy and compared to a brute force implementation of the MCEM-MCML algorithm (with $n_1, \ldots, n_{40} = 10000$, $n_{41}, \ldots, n_{50} = 50000$ and $n^\star = 100000$). The mean, median and standard deviation of the estimated parameters and log-likelihood values (denoted by log-lik) obtained over these repeated estimations are given in Table 2.6. The log-likelihood of the MS-GARCH model was approximated using the particle filter methodology with 100000 particles [see 16, 43], which is accurate to the first decimal place.

Unconstrained estimation of MS-GARCH models with empirical data can lead to parameters being estimated on the boundary of the parameter space and result in slow convergence of the MCEM-MCML algorithm. For example, Bauwens et al. [18] and Francq and Zakoïan [70] fitted the MS-GARCH model to daily S&P 500 data : Bauwens et al. [18] used the constraint $\alpha_1 = \beta_1 = 0$ in the estimation process while Francq and Zakoïan [70] reported an estimated value of α_1 very close to zero. To obtain convergence in the interior of the parameter space, I fitted a constrained MS-GARCH model by imposing $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ in the estimation process. For weekly data, both the constrained and unconstrained versions were estimated but due to slow convergence simulation schedule 2 was not effective at estimating the unconstrained version. For daily data, the unconstrained MS-GARCH model was fitted with zero means (i.e., $\mu_1 = \mu_2 = 0$) to match the specification considered by Bauwens et al. [16]. This restriction on the

TABLE 2.6. Estimation results and effectiveness of simulation schedule 2

	μ_1	μ_2	ω_1	α_1	β_1	ω_2	α_2	β_2	p_{11}	p_{22}	log-lik
Weekly S&P 500 : unconstrained model											
Brute	0.547	−2.01	0.111	0.0277	0.807	0.098	$< 10^{-8}$	1.82	0.833	0.109	−2740.3
Weekly S&P 500 : constrained model with $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$											
Brute	0.343	−2.80	0.0446	0.0429	0.902	2.51	0.0429	0.902	0.946	0.308	−2757.5
Median	0.346	−2.77	0.0445	0.0433	0.901	2.50	0.0433	0.901	0.945	0.308	−2757.5
Mean	0.346	−2.77	0.0446	0.0432	0.901	2.50	0.0432	0.901	0.945	0.306	−2757.5
StDev	0.007	0.12	0.0012	0.0009	0.003	0.04	0.0009	0.003	0.004	0.013	0.06
Daily S&P 500 : unconstrained model with zero means											
Brute	–	–	0.0532	0.0946	0.885	0.0128	0.0194	0.953	0.999	0.999	−4476.6
Median	–	–	0.0538	0.0943	0.884	0.0125	0.0216	0.952	0.999	0.999	−4476.7
Mean	–	–	0.0609	0.0938	0.883	0.0131	0.0237	0.948	0.997	0.998	−4477.0
StDev	–	–	0.0151	0.0033	0.004	0.0035	0.0065	0.012	0.003	0.002	0.53
Daily S&P 500 : constrained model with $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$											
Brute	0.0682	−1.05	0.00698	0.0337	0.942	0.527	0.0337	0.942	0.980	0.638	−4450.9
Median	0.0683	−1.06	0.00696	0.0339	0.942	0.528	0.0339	0.942	0.980	0.635	−4450.9
Mean	0.0683	−1.06	0.00695	0.0340	0.942	0.527	0.0340	0.942	0.980	0.634	−4450.9
StDev	0.0007	0.04	0.00007	0.0008	0.001	0.027	0.0008	0.001	0.000	0.025	0.05

means caused the MCEM-MCML algorithm to converge in the interior of the parameter space. The results of this estimation are consistent with those presented by Bauwens et al. [16].

The results displayed in Table 2.6 suggest that simulation schedule 2 can be used reliably to estimate financial data when the MLE is not on the boundary of the parameter space. The estimation of the constrained model is very accurate, but more variability is observed for the unconstrained model fitted with zero means to daily data. This is due to the single-move Gibbs sampler struggling to move between highly persistent regimes. Ways to improve this sampler are discussed in the conclusion. To estimate the unconstrained model with weekly data, over 100 iterations were needed to be confident that the algorithm converged as the parameter α_2 was attracted to 0. Even though $\beta_2 > 1$, the estimated model

is covariance stationary as it satisfies the conditions given by Francq et al. [67] [see also 18].

Finally, the computational time required to complete a single estimation with simulation schedule 2 on a 3.40 GHz Intel Core i7-2600 processor is approximately 35 minutes for the weekly data set and 110 minutes for the daily data set.

2.5.2. Comparison of fit

The fit of the constrained MS-GARCH model for the weekly and daily data sets is compared to the model of Gray [83] and to standard MS and GARCH models in Tables 2.7 and 2.8. The MS model is a special case of the MS-GARCH model with $\alpha_1 = \alpha_2 = 0$ and $\beta_1 = \beta_2 = 0$. The Bayesian information criterion (BIC) adds a penalty of $0.5k \log T$ to the negative of the log-likelihood, where k is the number of parameters in the model and T is the number of observations. The preferred model is the one with the lowest BIC. The asymptotic standard errors of the MLE are given in parentheses.

The fit of the MS-GARCH model is superior to Gray's model according to the BIC for both data sets considered. The BIC reported for Gray's model is based on the log-likelihood of Gray's model. The log-likelihood of the MS-GARCH model evaluated at the MLE of Gray's model was generally below that of the GARCH model. This implies that Gray's model can only generate a crude estimate of the MLE for the MS-GARCH model. These results are in line with those obtained in Section 2.4.3, where it was shown that Gray's model does not generate consistent estimates for the MS-GARCH model.

For the two time series considered, the fitted standard MS models include two persistent regimes : the first regime defines a positive mean return–low volatility state while the second regime pertains to a negative mean return–high volatility environment. When GARCH dynamics are incorporated into the MS model, the regime associated with a negative mean return becomes much less persistent (i.e., p_{22} is significantly reduced). To illustrate the difference in the role played by regime two in the MS and MS-GARCH models, we can compare the smoothed inferences of the states given by these models (see Figure 2.1). For example, consider the global financial crisis that emerged in September of 2008 with the bankruptcy of Lehman Brothers and the collapse of large financial institutions around the world. From September of 2008 to August of 2009, the MS model infers that the return process is in regime two, i.e., the negative mean return–high volatility regime. During that same period, the MS-GARCH model infers that this process enters regime two at the beginning of September of 2008 and returns to regime one five weeks later. Consequently, state two in the MS-GARCH

TABLE 2.7. Weekly S&P 500 : Estimated parameters and asymptotic standard errors

	μ_1	μ_2	ω_1	ω_2	α	β	p_{11}	p_{22}	BIC
GARCH	0.209		0.176		0.131	0.841			2822.4
	(0.050)		(0.058)		(0.024)	(0.029)			
MS	0.281	−0.141	2.19	11.2			0.977	0.953	2815.5
	(0.056)	(0.167)	(0.18)	(1.0)			(0.009)	(0.017)	
Gray	0.236	−2.37	$< 10^{-8}$	4.34	0.0698	0.848	0.984	0.487	2805.8
	(0.058)	(0.80)	($< 10^{-8}$)	(2.08)	(0.0260)	(0.040)	(0.008)	(0.122)	
MS-GARCH	0.343	−2.80	0.0446	2.51	0.0429	0.902	0.946	0.308	2786.2
	(0.060)	(0.63)	(0.0222)	(0.52)	(0.0205)	(0.028)	(0.022)	(0.179)	

TABLE 2.8. Daily S&P 500 : Estimated parameters and asymptotic standard errors

	μ_1	μ_2	ω_1	ω_2	α	β	p_{11}	p_{22}	BIC
GARCH	0.00223		0.0125		0.0759	0.916			4510.2
	(0.0167)		(0.0030)		(0.0086)	(0.009)			
MS	0.0571	−0.110	0.631	4.10			0.989	0.979	4661.8
	(0.0190)	(0.064)	(0.033)	(0.25)			(0.003)	(0.006)	
Gray	0.0422	−1.73	$< 10^{-8}$	0.769	0.0680	0.908	0.996	0.507	4495.5
	(0.0164)	(0.65)	($< 10^{-8}$)	(0.511)	(0.0080)	(0.006)	(0.002)	(0.151)	
MS-GARCH	0.0682	−1.05	0.00698	0.527	0.0337	0.942	0.980	0.638	4483.0
	(0.0177)	(0.34)	(0.00222)	(0.211)	(0.0127)	(0.012)	(0.006)	(0.164)	

model represents a shock regime which induces a jump in the volatility process. When the model reverts back to regime one, the effect of this shock still persists in the volatility due to the GARCH dynamics. This example demonstrates that volatility persistence is captured differently in the MS and MS-GARCH models. For the MS model, it is directly tied to regime persistence, i.e., long periods of high

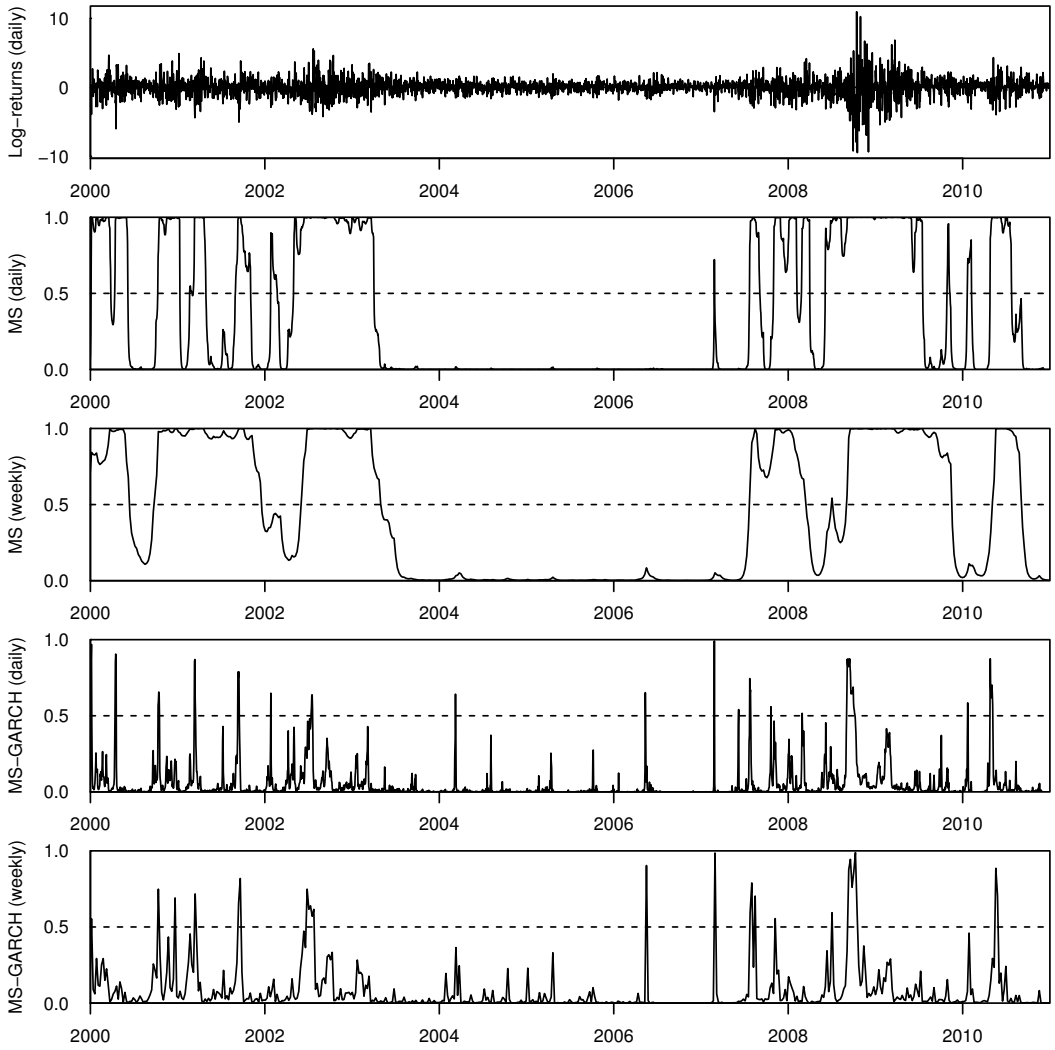

FIGURE 2.1. S&P 500 : Smoothed probabilities of being in regime two

volatility can only occur when the return process remains in regime two. For the MS-GARCH model, it is better explained by the GARCH dynamics of the model since the role of the MS process is now to allow for jumps in volatility. Therefore, it is not surprising that enriching the GARCH model with a MS process offers an improved fit as the presence of these jumps is well documented in the econometric literature [e.g., 63].

2.6. Conclusion

A novel approach was introduced based on the MCEM and MCML algorithms to calculate the MLE of the MS-GARCH model and its effectiveness was demonstrated with a simulation study. The main contribution of this method is

that it allows us to estimate the MS-GARCH model by maximum likelihood without resorting to a simplification of the model like the one used by Gray [83]. It was shown that Gray's model does not generate consistent estimates for model (2.2.1)–(2.2.3) and that the MCEM-MCML algorithm can significantly improve these estimates. The practical implementation of the algorithm was discussed and a simulation schedule was suggested to fit financial data. Finally, it was explained how to compute the variance-covariance matrix of the MLE. This paper offers many opportunities for further research.

First, the proposed algorithm can be extended to the case of asymmetric power GARCH regimes [see 49] with skewed or non-normal innovations. Haas [84] generalized the model of Haas et al. [86] in a similar way and showed that such a specification is sometimes preferred when it is fitted to financial data. As mentioned in Section 2.3.5.2, automated rules for selecting the sample size at each iteration of the algorithm can be investigated. Furthermore, the single-move Gibbs sampler that was proposed to generate the states may be poorly mixing in some instances [see 74, Section 11.5.6]. To improve this Gibbs sampler, states can be simulated in blocks instead of individually or in a random order instead of sequentially [see 126]. Simulation in blocks would increase the computational expense of the sampler by a factor of (M^b/b), where b is the block size. Consequently, it is not guaranteed that this will improve the performance of the algorithm for a fixed computational time. Recently, Bauwens et al. [16] proposed a multi-move Gibbs sampler (also known as a forward filtering–backward sampling algorithm) for the MS-GARCH model using particle filters. This algorithm can significantly enhance the mixing properties of the single-move sampler and reduce its sensitivity to initial values.

Second, other methods to estimate the MS-GARCH model can be investigated. For example, Jacquier et al. [108] proposed a MCMC maximum likelihood approach for latent state models that shares some similarities with the MCEM algorithm [see 52, 75, 110]. It involves a stochastic implementation of the M-step that can help reduce computational time and prevent convergence to a local mode of the log-likelihood. Moreover, since we can approximate the gradient and Hessian of the log-likelihood (see Section 2.3.4), gradient-based algorithms [see 34, Sections 10.1.3, 11.1.3 and 11.1.5] can be investigated.

Finally, the proposed algorithm applies to a univariate MS-GARCH model. The academic literature on multivariate MS-GARCH models is scarce, especially with regard to estimation. For example, Bauwens et al. [17] and Haas et al. [87] considered multivariate versions of the model of Haas et al. [86]. Since that model is not path dependent, estimation of these multivariate models can be done by a

direct maximization of the log-likelihood. In the context of a multivariate generalization of the (path dependent) MS-GARCH model, more sophisticated methods are needed. It would thus be interesting to extend the proposed algorithm to a multivariate setting or to perform estimation using a Bayesian approach.

Acknowledgments

I thank the Associate Editor, two referees, Mathieu Boudreault and Manuel Morales for their constructive comments which helped me improve this paper. I am also grateful to Brian Hartman and Eden Tsang who proofread this article and Arnaud Dufays for providing me with the data set used by Bauwens et al. [16]. Finally, I would like to acknowledge the financial support provided by the Natural Sciences and Engineering Research Council of Canada and the Society of Actuaries.

Appendix A. Validity of the EM algorithm for the MS-GARCH model

Strictly speaking, the MLE in mixture or MS models may not exist because of an unbounded likelihood. For example, consider a mixture of two normal distributions in which one of the normal densities has a mean exactly equal to one of the observations with a variance in the vicinity of 0. In this case, the likelihood is unbounded on the boundary of the parameter space so that a global maximum of the likelihood does not exist [see 74, Sections 6.1.2 and 6.1.3]. However, from a practical perspective, the area of the parameter space which creates this anomaly is of no interest since it occurs when one of the states is essentially trying to give discrete mass to one of the observations. Therefore, we may constrain the parameter space to exclude these pathological cases by, for example, posing a strictly positive lower bound on the variances. With this restriction, Kiefer [116] proves that there exists a bounded local maximizer of the likelihood which is consistent, efficient and asymptotically normal [see also 100, 107].

The convergence of the EM algorithm is discussed in great detail by Wu [163] who proves that if $Q(\theta \mid \theta')$ is continuous in both θ and θ' then (under regularity conditions) all limit points of the EM algorithm are stationary points of the likelihood. The following regularity conditions are needed (Ω denotes the parameter space) :

$$\Omega \text{ is a subset in the } d\text{-dimensional Euclidean space } \mathbb{R}^d,$$
$$\Omega_{\theta_0} = \{\theta \in \Omega : f(y \mid \theta) \geq f(y \mid \theta_0)\} \text{ is compact for any } f(y \mid \theta_0) > -\infty, \quad (2.6.1)$$

each parameter iterate $\theta^{(r)}$ is in the interior of Ω,

$f(y \mid \theta)$ is continuous in Ω and differentiable in the interior of Ω. (2.6.2)

A consequence of these conditions is that any sequence of likelihood values generated by the EM algorithm is bounded above. The compactness assumption (2.6.1) is difficult to verify but if we constrain the parameter space to exclude cases related to an unbounded likelihood, this assumption should hold for the MS-GARCH model (we can require that the GARCH parameters in all of the states be greater than some lower bound strictly greater than 0). Moreover, condition (2.6.2) is satisfied for the MS-GARCH model since (see equation (2.2.4))

$$f(y \mid \theta) = \sum_{S} \prod_{t=1}^{T} \left[\sigma_t^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_t - \mu_{S_t})^2}{2\sigma_t^2} \right) p_{S_{t-1},S_t} \right],$$

is continuous in Ω and differentiable in the interior of Ω. Finally,

$$\begin{aligned} Q(\theta \mid \theta') &= \sum_{S} \log[f(y, S \mid \theta)] p(S \mid y, \theta') \\ &= \frac{1}{f(y \mid \theta')} \sum_{S} \log[f(y, S \mid \theta)] f(y, S \mid \theta'), \end{aligned}$$

is continuous in both θ and θ' since $f(y, S \mid \theta)$ and $f(y \mid \theta)$ are positive continuous functions of θ.

This implies that it is valid to use the EM algorithm to estimate the MS-GARCH model since all limit points generated by the algorithm are stationary points of the likelihood. There is no guarantee that the stationary point is a local maximum, but convergence to a local minimum or a saddle point only occurs for some pathological cases so that in almost all instances, the stationary point is a local maximum [see 134, Section 3.6]. If the likelihood function is unimodal in θ and there is only one stationary point then the EM algorithm will converge to the global (unique) maximizer of the likelihood. However, the likelihood in mixture and MS models can have many modes and convergence of the EM sequence to a particular mode may depend on the choice of starting values. Hence, it is best to start the EM algorithm with different sets of starting values to avoid a suboptimal local maximum.

Appendix B. Analytical maximization of term 2 in expression (2.3.6)

Maximization of term 2 in expression (2.3.6) with respect to the transition probability parameters $\{p_{jk}\}_{j,k=1}^{M}$ subject to the constraint $\sum_{k=1}^{M} p_{jk} = 1, j = 1, \ldots, M$ can be done analytically. Let $f_{jk}^{(i)}$ denote the number of transitions from

state j to state k in the ith simulated state vector $S^{(i)}$ and let $f_{jk} = \sum_{i=1}^{n_r} f_{jk}^{(i)}$ (i.e., f_{jk} is the total number of transitions from state j to state k in all of the n_r simulated state vectors). Then, we may rewrite term 2 in expression (2.3.6) as

$$\begin{aligned}
\text{term 2} &= \frac{1}{n_r} \sum_{t=1}^{T} \sum_{i=1}^{n_r} \log(p_{S_{t-1}^{(i)}, S_t^{(i)}}) \\
&= \frac{1}{n_r} \sum_{j=1}^{M} \sum_{k=1}^{M} \sum_{i=1}^{n_r} f_{jk}^{(i)} \log(p_{jk}) \\
&= \frac{1}{n_r} \sum_{j=1}^{M} \sum_{k=1}^{M} f_{jk} \log(p_{jk}) \\
&= \frac{1}{n_r} \sum_{j=1}^{M} \left[\sum_{k=1}^{M-1} f_{jk} \log(p_{jk}) + f_{jM} \log(1 - \sum_{k=1}^{M-1} p_{jk}) \right]. \qquad (2.6.3)
\end{aligned}$$

For each j, the expression inside the summation of equation (2.6.3) only involves transition probabilities of the jth row of the transition matrix. Therefore, we may find the optimal values of the transition probabilities for each row independently. This implies that for row j we must find the values of $\{p_{jk}\}_{k=1}^{M-1}$ which maximize the expression

$$\sum_{k=1}^{M-1} f_{jk} \log(p_{jk}) + f_{jM} \log(1 - \sum_{k=1}^{M-1} p_{jk}).$$

This can be done by using straightforward calculus steps and the closed-form expressions for the maximizers of the transition probabilities are

$$p_{jk} = \frac{f_{jk}}{\sum_{l=1}^{M} f_{jl}}, \quad j, k = 1, \ldots, M.$$

Appendix C. Technical details related to the MCML-step of Algorithm 1

To obtain $\hat{\theta}$, we must calculate $w_{\theta|\theta^\star}^{(i)}$, $i = 1, \ldots, n^\star$, which are defined in equation (2.3.2). From a numerical perspective, it is best to calculate $\log(w_{\theta|\theta^\star}^{(i)})$, where

$$\begin{aligned}
\log(w_{\theta|\theta^\star}^{(i)}) &= \log f(y, S^{(i)} \mid \theta) - \log f(y, S^{(i)} \mid \theta^\star) \\
&= \frac{1}{2} \sum_{t=1}^{T} \left[\log(\sigma_t^{\star(i)})^2 - \log(\sigma_t^{(i)})^2 + \frac{(y_t - \mu^\star_{S_t^{(i)}})^2}{(\sigma_t^{\star(i)})^2} - \frac{(y_t - \mu_{S_t^{(i)}})^2}{(\sigma_t^{(i)})^2} \right] \\
&\qquad + \sum_{j=1}^{M} \sum_{k=1}^{M} f_{jk}^{(i)} \log(p_{jk}/p_{jk}^\star). \qquad (2.6.4)
\end{aligned}$$

The starred quantities in expression (2.6.4) must be calculated based on $\theta^\star$ and these that are not starred are based on θ. As in Appendix B, $f_{jk}^{(i)}$ denotes the

number of transitions from state j to state k in the ith simulated state vector $S^{(i)}$.

To improve the performance of the numerical optimization in expression (2.3.7), the gradient of $\log \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)}$ with respect to θ can be calculated in closed form using the following relation :

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)}\right] &= \sum_{i=1}^{n^\star} \frac{\partial f(y, S^{(i)} \mid \theta)/\partial\theta}{f(y, S^{(i)} \mid \theta^\star)} = \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)} \frac{\partial}{\partial\theta} \log f(y, S^{(i)} \mid \theta) \\ &= \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)} \frac{\partial}{\partial\theta} \log f(y \mid S^{(i)}, \theta) + \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)} \frac{\partial}{\partial\theta} \log p(S^{(i)} \mid \theta).\end{aligned}$$

Letting $\bar{w}_{\theta|\theta^\star}^{(i)} = w_{\theta|\theta^\star}^{(i)} / \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)}$, we obtain

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\log \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)}\right] &= \sum_{i=1}^{n^\star} \bar{w}_{\theta|\theta^\star}^{(i)} \frac{\partial}{\partial\theta} \log f(y \mid S^{(i)}, \theta) + \sum_{i=1}^{n^\star} \bar{w}_{\theta|\theta^\star}^{(i)} \frac{\partial}{\partial\theta} \log p(S^{(i)} \mid \theta) \\ &= \text{gradient 1} + \text{gradient 2}.\end{aligned} \tag{2.6.5}$$

Gradients 1 and 2 simplify to

$$\text{gradient 1} = -\frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n^\star} \bar{w}_{\theta|\theta^\star}^{(i)} \frac{\partial}{\partial\theta} \left[\log(\sigma_t^{(i)})^2 + \frac{(y_t - \mu_{S_t^{(i)}})^2}{(\sigma_t^{(i)})^2}\right], \tag{2.6.6}$$

$$\text{gradient 2} = \sum_{j=1}^{M} \sum_{k=1}^{M} \left[\left(\sum_{i=1}^{n^\star} \bar{w}_{\theta|\theta^\star}^{(i)} f_{jk}^{(i)}\right) \frac{\partial}{\partial\theta} \log(p_{jk})\right]. \tag{2.6.7}$$

Gradient 1 can be calculated recursively without difficulty while gradient 2 can be computed directly.

Note that the gradient of $\log \sum_{i=1}^{n^\star} w_{\theta|\theta^\star}^{(i)}$ is approximately equal to the gradient of the log-likelihood (see equation (2.3.3)). It is also closely related to the gradient of $\widehat{Q}(\theta \mid \theta^\star)$ which can be obtained from expressions (2.6.6) and (2.6.7) by replacing $\bar{w}_{\theta|\theta^\star}^{(i)}$, $i = 1, \ldots, n^\star$, with $1/n^\star$. In this case, gradient 1 is the gradient of term 1 in expression (2.3.6).

Chapitre 3

ESTIMATING THE MARKOV-SWITCHING GARCH MODEL WITH A DETERMINISTIC PARTICLE FILTER

Cet article a été réalisé avec la collaboration des coauteurs Mathieu Boudreault et Manuel Morales. Il a été soumis en octobre 2013 à la revue *Journal of Econometrics*.

Résumé

Le modèle GARCH à changement de régimes combine une structure GARCH avec des paramètres qui varient dans le temps. Son introduction dans les années 1990 tire sa motivation de l'incapacité du modèle GARCH à refléter des changements structurels dans les données financières, souvent observés sur de longues périodes de temps. La combinaison des processus GARCH et à changement de régimes donne lieu à un problème de *path dependence* qui complique l'estimation du modèle. Cette difficulté a mené à l'introduction de méthodes computationnelles intensives et à des techniques plus simples basées sur une approximation du modèle, connues sous le nom de *collapsing procedures*. Une méthodologie originale est développée permettant d'estimer le modèle GARCH à changement de régimes par le maximum de vraisemblance, généralisant et améliorant les approches dites par *collapsing* développées dans les deux dernières décennies. Cette généralisation permet d'établir un lien méthodologique entre ces approches et le filtre particulaire, démontrant ainsi qu'elles sont équivalentes à un filtre particulaire déterministe. La découverte de cette relation est importante, car elle permet de justifier la validité de l'approche dite par *collapsing* pour estimer le modèle GARCH à changement de régimes. Des études par simulations et empirique révèlent que la méthodologie proposée parvient à estimer le modèle avec précision et rapidité.

Abstract

The Markov-switching GARCH model allows for a GARCH structure with time-varying parameters. Its introduction in the 1990s was motivated by the inability of the standard GARCH model to reflect structural changes in financial time series which are usually found to occur over many years of data. The combination of Markov-switching and GARCH processes complicates estimation as exact computation of the log-likelihood is infeasible due to a path dependence problem. This difficulty led to computationally intensive estimation methods and to simpler techniques based on an approximation of the model, known as collapsing procedures. An original framework is developed to conduct maximum likelihood inference in the Markov-switching GARCH model, generalizing and improving previously proposed collapsing approaches. This method is demonstrated to correspond to a deterministic particle filter, establishing a novel relationship between particle filtering and collapsing procedures. The discovery of this relationship is important because it provides the missing link needed to justify the validity of the collapsing approach for estimating the Markov-switching GARCH model. Simulation and empirical studies show that it allows for a fast and accurate estimation of the model.

Keywords : Markov-switching, regime-switching, GARCH, maximum likelihood, particle filtering, collapsing

3.1. Introduction

Markov-switching GARCH (MS-GARCH) models allow for a GARCH structure with time-varying parameters. They were introduced in response to the findings of Lamoureux and Lastrapes [125] who suggested that a GARCH model with constant parameters may inaccurately measure volatility persistence in long time series of financial returns. This is due to its inability to reflect structural changes which are usually found to occur over many years of data. Hillebrand [106] and Mikosch and Starica [135] rigorously proved that neglecting parameter changes in a GARCH model creates an upward bias on the measure of volatility persistence. It is therefore not surprising that many authors [e.g., 8, 30, 86, 84, 120, 131, 155] were able to show that MS-GARCH models can provide a better fit to empirical data and improved volatility and Value-at-Risk forecasts.

The econometric literature has proposed two approaches for combining Markov-switching and GARCH frameworks, resulting in two different MS-GARCH models. The first model dates back to Cai [29], Gray [83] and Hamilton and Susmel [93] and can be referred to as the path dependent MS-GARCH model, which we

abbreviate by MSG. The second model was introduced by Haas et al. [86] and can be termed the path independent MS-GARCH model [see also 4, 85]. In the path dependent model the conditional variance at any given time depends on the entire history of regimes generated by the hidden Markov chain, while in the path independent model it only depends on the current regime. In applications, there is a clear advantage to using the path independent methodology because standard Markov-switching maximum likelihood inference is applicable without any complications [see 92]. For these standard tools to be suitable for the MSG model, one must employ the approximations proposed by Dueker [56], Gray [83] or Klaassen [120], known as collapsing procedures, because without them, exact calculation of the log-likelihood is infeasible [refer to 83, for a very good account of this problem]. Although these collapsing approaches are popular to fit financial data [e.g., 8, 10, 19, 103, 131, 155] and simulate from the posterior distribution of the state vector [e.g., 20, 57], their validity and accuracy have not been investigated. Using a simulation study, Augustyniak [9] exposed that maximum likelihood estimation based on the collapsing procedure proposed by Gray [83], the most popular in the econometric literature, results in highly biased parameter estimates. Therefore, there is a need to better understand these approximations.

This article demonstrates how the collapsing procedures proposed by Dueker [56], Gray [83] and Klaassen [120] can be generalized and recast as a deterministic particle filter (DPF). We establish a strong connection between this DPF and the optimal particle filter (OPF) developed by Fearnhead and Clifford [64] by showing that these two algorithms only differ in the particle selection (or resampling) stage, the DPF performing this selection deterministically rather than stochastically. This novel relationship between particle filtering and collapsing procedures provides the missing link needed to justify the validity of the collapsing approach for estimating the MSG model. The discovery of this link is an important contribution of this article as collapsing procedures were historically justified with intuitive arguments.

Traditional particle filters (e.g., bootstrap filter or auxiliary particle filter) provide a way to unbiasedly estimate the MSG likelihood, but the resulting approximation is very difficult to maximize as it is a discontinuous function of the parameters [see 43, 115]. Even though the OPF proposed by Fearnhead and Clifford [64] is shown in this article to be significantly more accurate for the MSG model than traditional particle filters, its use for maximum likelihood estimation is still unreliable and very slow due to the discontinuity problem. In contrast, the DPF that we develop yields an approximation of the MSG log-likelihood which is a continuous and differentiable function of the parameters, at the cost

of introducing a bias in its evaluation. By comparing log-likelihood estimates generated with the DPF and OPF algorithms in a simulation study, this bias is found to be very small in practical situations. Since the DPF achieves a smooth approximation of the log-likelihood with little bias, we propose its use to conduct maximum likelihood inference in the MSG model. An empirical study with S&P 500 data shows that it can generate accurate parameter estimates in a matter of minutes. Therefore, the algorithm introduced can offer a computational benefit over simulation-based inference techniques proposed by Augustyniak [9], Bauwens et al. [18, 16], and Dufays [57].

This article is organized as follows. Section 3.2 defines the MSG model. Section 3.3 develops the estimation methodology and establishes its relationship with particle filtering. Section 3.4 demonstrates its effectiveness for computing the log-likelihood and maximum likelihood estimator (MLE) of the MSG model. Section 3.5 concludes and proposes avenues for further research. Derivations and supplementary algorithms are provided in the appendix.

3.2. The MSG model

Consider the standard GARCH$(1,1)$ model :

$$\begin{aligned} y_t &= \mu + \sigma_t \eta_t, \\ \sigma_t^2 &= \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \\ \epsilon_{t-1} &= y_{t-1} - \mu, \end{aligned}$$

where $t = 1, \ldots, T$ and $y_{1:T} = (y_1, \ldots, y_T)$ is a vector of returns. The innovation η_t is assumed to follow a normal distribution with zero mean and unit variance, i.e., $\eta_t \sim \mathcal{N}(0,1)$, and to be independent over the index t. The conditional distribution of y_t given $y_{1:t-1}$ is $\mathcal{N}(\mu, \sigma_t^2)$, where σ_t^2 represents the conditional variance of the return at time t, i.e., $\sigma_t^2 = \mathrm{Var}(y_t \mid y_{1:t-1})$, while μ is the conditional, as well as unconditional, mean of that return. The remaining symbols $\{\omega, \alpha, \beta\}$ represent parameters of the GARCH process.

Furthermore, let $\{S_t\}$ be an ergodic time-homogeneous Markov chain with finite state space $\{1, \ldots, M\}$ and $M \times M$ transition matrix defined by the probabilities $\{p_{ij} = \Pr(S_t = j \mid S_{t-1} = i)\}_{i,j=1}^{M}$. This Markov chain is assumed to be unobserved and independent of $\{\eta_t\}$. The MSG model [29, 83, 93] integrates this Markov chain into the GARCH$(1,1)$ model as follows :

$$y_t = \mu + \sigma_t(S_{1:t})\eta_t, \tag{3.2.1}$$

$$\sigma_t^2(S_{1:t}) = \omega_{S_t} + \alpha_{S_t}\epsilon_{t-1}^2 + \beta_{S_t}\sigma_{t-1}^2(S_{1:t-1}) \tag{3.2.2}$$

$$= \sum_{j=1}^{t} \left[\left(\omega_{S_j} + \alpha_{S_j} \epsilon_{j-1}^2 \right) \prod_{k=0}^{t-j-1} \beta_{S_{t-k}} \right] + \sigma_0^2 \prod_{k=0}^{t-1} \beta_{S_{t-k}}, \tag{3.2.3}$$

where $S_{1:t} = (S_1, \ldots, S_t)$ and σ_0^2 is a known constant representing the variance at time $t = 0$ (for simplicity, we set $\epsilon_0^2 = \sigma_0^2$). This model can be simply understood as a GARCH model with time-varying parameters, $\{\omega_i, \alpha_i, \beta_i\}_{i=1}^M$, dependent upon the states of an unobserved Markov chain. The conditional distribution of y_t given $(y_{1:t-1}, S_{1:t})$ is $\mathcal{N}(\mu, \sigma_t^2(S_{1:t}))$, where $\sigma_t^2(S_{1:t}) = \mathrm{Var}(y_t \mid y_{1:t-1}, S_{1:t})$. Equation (3.2.3) demonstrates that this conditional variance depends on the entire regime path, which renders exact calculation of the log-likelihood infeasible in practice [see 83]. Consequently, parameter estimation for the MSG model is feasible by maximizing an approximated log-likelihood [56, 83, 120], by resorting to simulation-based techniques [9, 18, 16, 57] or by making use of the generalized method of moments [70]. The objective of this article is to generalize and improve the accuracy of approaches based on an approximated log-likelihood, known as collapsing procedures, and to relate them to a DPF. This relationship allows us to justify why they can be relied upon to approximate the MLE of the MSG model.

A switching mean can be introduced into the model, but we postpone this extension until Section 3.4, as it would unnecessarily complicate the exposition of our proposed methodology in Section 3.3. This article centers on the MSG model, but a comparison with the path independent MS-GARCH model of Haas et al. [86] is provided in Appendix A.

3.3. Relating deterministic and stochastic approaches for approximating the MSG log-likelihood

Let $f_t(\cdot)$ (respectively, $p_t(\cdot)$) denote a probability density (respectively, mass) function where the subscript t, if present, indicates that the density is conditioned on the returns up to time t. As mentioned in Section 3.2, exact calculation of the MSG log-likelihood is infeasible due to a path dependence problem [see 83] : the conditional density of y_t, $f_{t-1}(y_t \mid S_{1:t}) = f(y_t \mid y_{1:t-1}, S_{1:t})$, depends on the entire regime path $S_{1:t}$.

Hamilton [90, 91, 92] developed a recursive filter allowing us to calculate the log-likelihood of Markov-switching models when the conditional density of y_t depends on a finite (and not too large) number of lags, say q, of the unobserved

process $\{S_t\}$. Hence, we must have $f_{t-1}(y_t \mid S_{1:t}) = f_{t-1}(y_t \mid S_{t-q+1:t})$. The approach of Hamilton is based on the following decomposition of the log-likelihood :

$$\log f(y_{1:T}) = \log f(y_1) + \sum_{t=2}^{T} \log f_{t-1}(y_t), \text{ where} \tag{3.3.1}$$

$$f_{t-1}(y_t) = \sum_{S_{t-q+1:t}} \left[\sum_{S_{t-q}} p_{t-1}(S_{t-q:t-1}) p(S_t \mid S_{t-1}) f_{t-1}(y_t \mid S_{t-q+1:t}) \right], \tag{3.3.2}$$

and consists in calculating the M^q conditional probabilities $p_t(S_{t-q+1:t})$ recursively for $t = 1$ to $T-1$ from the relation

$$p_t(S_{t-q+1:t}) = \frac{\sum_{S_{t-q}} p_{t-1}(S_{t-q:t-1}) p(S_t \mid S_{t-1}) f_{t-1}(y_t \mid S_{t-q+1:t})}{f_{t-1}(y_t)}.$$

The computational cost of the Hamilton filter at each iteration is $\mathcal{O}(M^{q \vee 2})$. When q is large or, as is the case in path dependent models, grows linearly with the index t, computational limitations prevent the use of this algorithm. In this situation, it is still possible to approximate the recursions of the Hamilton filter with particle filters, or more generally sequential Monte Carlo (SMC) methods, and obtain an unbiased estimate of the likelihood [see 46, 130, 146]. However, the estimate of the likelihood obtained in this way is not a continuous function of the parameters, making it very difficult to maximize [see 43, 115].

Alternatively, we may adopt a deterministic approach for approximating the Hamilton filter. This idea, inspired by the works of Harrison and Stevens [98] and Harvey et al. [99], was proposed by Kim [117] who treated maximum likelihood estimation of dynamic linear models with Markov-switching. Similarly to the MSG model, these models suffer from the path dependence problem. When the Hamilton filter is applied on a path dependent model, each iteration results in an M-fold increase in the number of regime paths to consider inside the summation of equation (3.3.2). The solution proposed by Kim [117] was to limit this exponential growth with a collapsing procedure at the cost of introducing an approximation into the filter, or equivalently, into the model.

The methods of Dueker [56], Gray [83] and Klaassen [120], which consist in approximating the MSG log-likelihood, are all based on the collapsing idea of Kim [117]. Section 3.3.1 unifies these collapsing procedures under a general framework so that they can be significantly improved. Section 3.3.2 explains how the MSG log-likelihood can be estimated with particle filters and introduces the OPF of Fearnhead and Clifford [64] in the context of the MSG model. Finally, Section 3.3.3 establishes the relationship between the collapsing procedure and the OPF, demonstrating that the collapsing procedure is a DPF. This relationship is important because it helps justify the validity of the collapsing approach for

estimating the MSG model. Although we do not elaborate on this subject, our work also suggests that the collapsing procedure of Kim [117] is a DPF. To our knowledge, this connection has not been made in the econometric literature and is open for future research.

3.3.1. Deterministic approach : Collapsing procedure

Our contribution in this section is to unify the approaches of Dueker [56], Gray [83] and Klaassen [120] under a general framework so that they can be better understood and improved. For this purpose, we introduce a simplified (or approximated) MSG model where the conditional distribution of y_t depends on only q lags of the Markov-switching process for $t \geq q$:

$$y_t = \mu + \tilde{\sigma}_t(S_{t-q+1:t})\eta_t, \tag{3.3.3}$$

$$\tilde{\sigma}_t^2(S_{t-q+1:t}) = \omega_{S_t} + \alpha_{S_t}\epsilon_{t-1}^2 + \beta_{S_t}h_{t-1}(S_{t-q+1:t}), \tag{3.3.4}$$

$$h_t(S_{t-q+2:t+1}) = \widetilde{\mathrm{E}}\left[\tilde{\sigma}_t^2(S_{t-q+1:t}) \mid y_{1:t}, S_{t-q+2:t+1}\right] \tag{3.3.5}$$

$$= \sum_{S_{t-q+1}} \tilde{\sigma}_t^2(S_{t-q+1:t})\tilde{p}_t(S_{t-q+1} \mid S_{t-q+2:t+1}). \tag{3.3.6}$$

The tilde mark over a symbol indicates that this quantity is calculated with respect to the simplified MSG model. For $t \leq q$, this model is equivalent to the MSG model, and we have $\tilde{\sigma}_t(S_{1:t}) = \sigma_t(S_{1:t})$, $\tilde{p}_t(S_{1:t}) = p_t(S_{1:t})$, and $\tilde{f}_{t-1}(y_t \mid S_{1:t}) = f_{t-1}(y_t \mid S_{1:t})$. For $t > q$, the conditional distribution of y_t, $\tilde{f}_{t-1}(y_t \mid S_{1:t})$, reduces to $\tilde{f}_{t-1}(y_t \mid S_{t-q+1:t})$, which implies that the Hamilton filter can be used to calculate the log-likelihood of model (3.3.3)–(3.3.5) provided q is not too large. The probability, $\tilde{p}_t(S_{t-q+1} \mid S_{t-q+2:t+1})$, in equation (3.3.6) can be obtained as a by-product of this filter since

$$\tilde{p}_t(S_{t-q+1} \mid S_{t-q+2:t+1}) = \begin{cases} \dfrac{\tilde{p}_t(S_t)\tilde{p}(S_{t+1} \mid S_t)}{\sum_{S_t} \tilde{p}_t(S_t)\tilde{p}(S_{t+1} \mid S_t)}, & q = 1; \\ \dfrac{\tilde{p}_t(S_{t-q+1:t})}{\sum_{S_{t-q+1}} \tilde{p}_t(S_{t-q+1:t})}, & q \geq 2. \end{cases}$$

When $q \geq 2$, this probability is independent of S_{t+1} and equation (3.3.5) simplifies to

$$\widetilde{\mathrm{E}}\left[\tilde{\sigma}_t^2(S_{t-q+1:t}) \mid y_{1:t}, S_{t-q+2:t}\right].$$

Therefore, as long as $q \geq 2$, $h_t(S_{t-q+2:t+1})$ is independent of S_{t+1}, and can take up to M^{q-1} different values. To simplify notation, we write $\sigma_t = \sigma_t(S_{1:t})$, $\tilde{\sigma}_t = \tilde{\sigma}_t(S_{t-q+1:t})$ and $h_t = h_t(S_{t-q+2:t+1})$.

In the MSG model, there is an M-fold increase in the number of regime paths to consider each time t increases. The purpose of the simplified MSG model is

to limit this exponential growth to never exceed M^q. Therefore, at $t = q$, the M^q possible conditional variances, σ_q^2, are *collapsed* into M^{q-1} values of h_q with equation (3.3.5). At $t = q+1$, M^q new conditional variances, $\tilde{\sigma}_{q+1}^2$, are obtained based on equation (3.3.4) and the values of h_q. These conditional variances are then again *collapsed* into M^{q-1} values of h_{q+1}, and the process continues. When $q = 1$, equation (3.3.5) does not result in an M-fold decrease in the number of possibilities, but it still prevents them to multiply. This collapsing procedure ensures that the conditional variance and distribution of y_t depend on a moving window of q regimes instead of the whole regime path.

The approximations proposed by Dueker [56] and Klaassen [120] are special cases of model (3.3.3)–(3.3.5) with $q = 2$ and 1, respectively. The approximation of Gray [83] corresponds to $q = 1$, but with equation (3.3.5) substituted by

$$h_t = \widetilde{\mathrm{Var}}[y_t \mid y_{1:t-1}] = \widetilde{\mathrm{E}}[\tilde{\sigma}_t^2 \mid y_{1:t-1}].$$

Therefore, Gray's collapsing procedure does not take advantage of all available information in the collapsing process.

Algorithm 1 explains how to construct the collapsing procedure when $q \geq 2$. The case $q = 1$ requires a small modification and is thus treated in Appendix C (Algorithm 3). There are two ways to interpret this algorithm : (i) the exact Hamilton filter is executed on the simplified MSG model or (ii) an approximate (deterministic) filter is applied on the MSG model. Our exposition in this section accentuates the first of these interpretations, but in Section 3.3.3 we will emphasize the second one and establish that Algorithm 1 is a DPF.

Algorithm 1 : Collapsing procedure for the MSG model ($q \geq 2$)

Exact filtering stage

1. Calculate $a_1^{(j)} = f(y_1 \mid S_1 = j) \cdot p(S_1 = j \mid S_0)$ for $j = 1, \ldots, M$, and $a_1 = \sum_{j=1}^{M} a_1^{(j)}$.
2. For each $i = 1, \ldots, M$, set $S_1^{(i)} = i$ and $w_1^{(i)} = a_1^{(i)}/a_1$ to obtain $\{S_1^{(i)}, w_1^{(i)}\}_{i=1}^{M}$. This weighted sample is an exact representation of $p_1(S_1)$.

For $t = 2, \ldots, q$:

3. For $i = 1, \ldots, M^{t-1}$ and $j = 1, \ldots, M$, calculate :

$$a_t^{(i,j)} = f_{t-1}(y_t \mid S_{1:t-1}^{(i)}, S_t = j) \cdot p(S_t = j \mid S_{t-1}^{(i)}) \cdot w_{t-1}^{(i)} \text{ and } a_t = \sum_{i=1}^{M^{t-1}} \sum_{j=1}^{M} a_t^{(i,j)}.$$

4. For each $i = 1, \ldots, M^t$ and $j = 1, \ldots, M$, set $S_{1:t}^{(i,j)} = (S_{1:t-1}^{(i)}, j)$ and $w_t^{(i,j)} = a_t^{(i,j)}/a_t$. Combine indices (i, j) into a single index i to obtain $\{S_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^{M^t}$. This weighted sample is an exact representation of $p_t(S_{1:t})$ and $\tilde{p}_t(S_{1:t})$.

Approximate filtering stage

Set $N = M^{q-1}$. For $t = q+1, \ldots, T$:

5. Collapse the NM values of $\tilde{\sigma}^2_{t-1}$ into N values of h_{t-1} using equations (3.3.5)–(3.3.7).
6. Aggregate $\{S^{(k)}_{t-q:t-1}, w^{(k)}_{t-1}\}_{k=1}^{NM}$ over S_{t-q}. Let $S^{(j,k)}_{t-q:t-1} = (j, S^{(k)}_{t-q+1:t-1})$ and denote the corresponding weights by $w^{(j,k)}_{t-1}$. Set $w^{(k)}_{t-1} = \sum_{j=1}^{M} w^{(j,k)}_{t-1}$, and extract the N distinct elements from $\{S^{(k)}_{t-q+1:t-1}, w^{(k)}_{t-1}\}_{k=1}^{NM}$ to obtain the weighted sample $\{S^{(i)}_{t-q+1:t-1}, w^{(i)}_{t-1}\}_{i=1}^{N}$, an exact representation of $\tilde{p}_{t-1}(S_{t-q+1:t-1})$.
7. For $i = 1, \ldots, N$ and $j = 1, \ldots, M$, calculate :

$$a^{(i,j)}_t = \tilde{f}_{t-1}(y_t \mid S^{(i)}_{1:t-1}, S_t = j) \cdot p(S_t = j \mid S^{(i)}_{t-1}) \cdot w^{(i)}_{t-1} \text{ and } a_t = \sum_{i=1}^{N}\sum_{j=1}^{M} a^{(i,j)}_t.$$

8. For each $i = 1, \ldots, N$ and $j = 1, \ldots, M$, set $S^{(i,j)}_{t-q+1:t} = (S^{(i)}_{t-q+1:t-1}, j)$ and $w^{(i,j)}_t = a^{(i,j)}_t / a_t$. Combine indices (i, j) into a single index i to obtain $\{S^{(i)}_{t-q+1:t}, w^{(i)}_t\}_{i=1}^{N}$. This weighted sample is an exact representation of $\tilde{p}_t(S_{t-q+1:t})$ and an approximation to $p_t(S_{t-q+1:t})$.

After executing Algorithm 1, the log-likelihood of the simplified MSG model is obtained with $\sum_{t=1}^{T} \log a_t$.

3.3.1.1. *Understanding the approximations*

Intuitively, the quality of the approximation provided by Algorithm 1 should increase with q. If $q = T$, the approximation is exact, but infeasible to implement in practice as the computational cost of executing the Hamilton filter on the simplified MSG model is $\mathcal{O}(TM^{q\vee 2})$. Consequently, there is a tradeoff between accuracy and computational cost which requires investigation. In the context of the MSG model, it is not known how good of an approximation the simplified MSG model can offer. Augustyniak [9] showed in a simulation study that the estimation procedure of Gray [83] can result in highly biased parameter estimates. This is not surprising as Gray's approximation corresponds to the case $q = 1$ and does not take advantage of all available information in the collapsing process. To date, the accuracy of approximations proposed by Dueker [56] and Klaassen [120] have not been examined, not even by means of a simulation study. Even outside of the MSG context, there is only limited simulation evidence that collapsing procedures similar to Algorithm 1 can be effective [see 99, 117, 118]. These approximations were often applied in the context of switching state space models [see 74, Section 13.3.5], but seem to have lost their popularity in the econometric literature in favor of Bayesian Markov Chain Monte Carlo (MCMC) and SMC methods at

the turn of the century. To our knowledge, the validity and accuracy of these approximations have not been investigated from a theoretical point of view.

To better understand the nature of the approximation introduced by the simplified MSG model, we can examine the relationship between σ_t^2 and $\tilde{\sigma}_t^2$. At time $t \geq q$ and conditional on $y_{1:t-1}$, σ_t^2 and $\tilde{\sigma}_t^2$ can take M^t and M^q different values with associated probabilities $p_{t-1}(S_{1:t})$ and $\tilde{p}_{t-1}(S_{t-q+1:t})$, respectively. Therefore, the conditional distribution of y_t given $y_{1:t-1}$ in the MSG model,

$$f_{t-1}(y_t) = \sum_{S_{1:t}} f_{t-1}(y_t \mid S_{1:t}) p_{t-1}(S_{1:t}),$$

can be represented as a mixture of M^t distributions,

$$y_t \mid y_{1:t-1} \sim \mathcal{N}(\mu, \sigma_t^2(S_{1:t})) \quad \text{w.p.} \quad p_{t-1}(S_{1:t}), \quad S_{1:t} \in \{1, \ldots, M\}^t. \tag{3.3.8}$$

The motivation behind the simplified MSG model is to approximate this mixture with a lesser amount of M^q components :

$$y_t \mid y_{1:t-1} \sim \mathcal{N}(\mu, \tilde{\sigma}_t^2(S_{t-q+1:t})) \quad \text{w.p.} \quad \tilde{p}_{t-1}(S_{t-q+1:t}), \quad S_{t-q+1:t} \in \{1, \ldots, M\}^q.$$

Each component of this mixture serves to represent M^{t-q} elements in the mixture (3.3.8) with common states $S_{1:t-q}$. In other words, the simplified MSG model approximates M^{t-q} conditional variances in the MSG model with common states $S_{1:t-q}$ with a single conditional variance. Appendix B derives the relation

$$\tilde{\sigma}_t^2 = \widetilde{\mathrm{E}}\left[\sigma_t^2 \;\middle|\; y_{1:t-1}, S_{t-q+1:t}\right],$$

which implies that in the calculation of $\tilde{\sigma}_t^2$, the states $S_{1:t-q}$ are integrated out of σ_t^2. Consequently, $\tilde{\sigma}_t^2$ is the minimum mean square error predictor of σ_t^2 among functions of $(y_{1:t-1}, S_{t-q+1:t})$. This suggests that if an accurate inference of $S_{1:t-q}$ can be made based on $(y_{1:t-1}, S_{t-q+1:t})$, the loss of information due to collapsing will be small as values of σ_t^2 furthest from $\tilde{\sigma}_t^2$ will have negligible probability mass. Intuitively, this occurs when regimes are persistent and well differentiated in their parameters.

Having established a connection between σ_t^2 and $\tilde{\sigma}_t^2$, we can investigate the difference $\delta_t = \sigma_t^2 - \tilde{\sigma}_t^2$ which measures the global approximation error arising from sequentially applying the collapsing procedure from time 1 to t. Let

$$\tilde{\delta}_t = \tilde{\sigma}_{t-q}^2 - \widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-q}^2 \;\middle|\; y_{1:t-1}, S_{t-q+1:t}\right], \text{ and}$$
$$B_j = \prod_{k=0}^{j-1} \beta_{S_{t-k}}.$$

The symbol $\tilde{\delta}_t$ only involves quantities calculated under the simplified MSG model and measures the local approximation error due to collapsing at time t. Appendix B demonstrates that

$$\delta_t = \begin{cases} 0, & t = 0, ..., q; \\ \sum_{j=0}^{\lfloor t/q \rfloor - 1} \tilde{\delta}_{t-jq} B_{(j+1)q}, & t \geq q+1. \end{cases}$$

The global error δ_t depends on $S_{1:t}$ and $y_{1:t-1}$, but expression (3.3.9) shows that it can be decomposed into a weighted sum of previous $\tilde{\delta}_t$ at intervals of length q. For the simplified MSG model to offer a good approximation to the MSG model, the local errors $\{\tilde{\delta}_{t-jq}\}$ should not accumulate into δ_t as t increases. Let $\kappa = \sum_{i=1}^{M} \pi_i \log(\beta_i)$, where $\{\pi_i = \Pr(S_t = i)\}_{i=1}^{M}$ is the stationary distribution of the Markov chain $\{S_t\}$, and assume $\kappa < 0$. Similarly to Bauwens et al. [18, p. 222] and Nelson [141, Theorem 2], it can be shown that with probability one, $B_j = \mathcal{O}(\exp(\bar{\kappa} j))$ for some $\bar{\kappa} \in (\kappa, 0)$. Therefore, the weights $\{B_j\}$ in equation (3.3.9) decay exponentially with time and the errors $\{\tilde{\delta}_{t-j}\}$ do not accumulate into δ_t. The condition $\kappa < 0$ is typically always satisfied in practical situations as it is implied by the strict or second-order stationarity of the MSG model. Francq et al. [67] proved that a strictly stationary solution exists if and only if $\gamma = \sum_{i=1}^{M} \pi_i \, \mathrm{E}[\log(\alpha_i \eta_t^2 + \beta_i)] < 0$ [see also 18], and that this condition is necessary for second-order stationarity.

3.3.2. Stochastic approach : Particle filters

Particle filters provide approximate recursions of the Hamilton filter as they sequentially estimate $p_t(S_{1:t})$ with a set of N particles, $\{S_{1:t}^{(i)}\}_{i=1}^{N}$, having associated probability weights $\{w_t^{(i)}\}_{i=1}^{N}$:

$$\{S_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^{N} \sim p_t(S_{1:t}).$$

Traditional particle filters, namely the bootstrap filter, the sequential importance sampling with resampling algorithm and the auxiliary particle filter (APF), have a computational cost that is linear in the number of particles and can be applied to a large variety of models with discrete or continuous state spaces [see 43, 53, for an overview of particle filtering].

Using the output of particle filters, $f_{t-1}(y_t)$ can be approximated with

$$f_{t-1}(y_t) \approx \sum_{i=1}^{N} f_{t-1}(y_t \mid S_{1:t-1}^{(i)}) w_{t-1}^{(i)},$$

and the log-likelihood of the MSG model can be estimated with expression (3.3.1). The applicability of particle filters to the MSG model was only noted recently by Bauwens et al. [16] who proposed an APF. A slight improvement can be made to this particle filter by considering the so-called *fully adapted* APF [see 147]. [1] However, even in this case, the particle filter is not tailored to the MSG model as it does not take advantage of the discrete nature of the state space.

The *fully adapted* APF starts with a weighted sample, $\{S_{1:t-1}^{(i)}, w_{t-1}^{(i)} = 1/N\}_{i=1}^{N}$, from $p_{t-1}(S_{1:t-1})$ and has for objective to propagate these particles one step forward. To accomplish this, it first resamples the particles $\{S_{1:t-1}^{(i)}\}_{i=1}^{N}$ in proportion to the importance weights,

$$w_t^{(i)} = \frac{f_{t-1}(y_t \mid S_{1:t-1}^{(i)})}{\sum_{i=1}^{N} f_{t-1}(y_t \mid S_{1:t-1}^{(i)})}, \quad i = 1, \ldots, N.$$

This resampling step has the effect of duplicating the most desirable particles given the observed value of y_t and of resetting the corresponding weights to $\{w_t^{(i)} = 1/N\}_{i=1}^{N}$. Second, it samples a value of S_t for each of the particles from the proposal

$$p_t(S_t \mid S_{1:t-1}) \propto f_{t-1}(y_t \mid S_{1:t}) p(S_t \mid S_{t-1}).$$

The new set of particles $\{S_{1:t}^{(i)}, 1/N\}_{i=1}^{N}$ offers an approximation to $p_t(S_{1:t})$. The *fully adapted* APF for the MSG model is presented in Appendix C (Algorithm 4).

In the context of the MSG model, the *fully adapted* APF can be improved by exploiting the discreteness of the state space and the methodology introduced by Fearnhead and Clifford [64]. For instance, each particle $\{S_{1:t-1}^{(i)}, w_{t-1}^{(i)}\}$ has M possible descendants at time t : $\{S_{1:t}^{(i,j)} = (S_{1:t-1}^{(i)}, j), w_t^{(i,j)}\}_{j=1}^{M}$, where

$$w_t^{(i,j)} \propto f_{t-1}(y_t \mid S_{1:t}^{(i,j)}) p(S_t = j \mid S_{t-1}^{(i)}) w_{t-1}^{(i)}.$$

When selecting particles at time t, it is therefore preferable to list exhaustively the NM possible descendants, $\{\{S_{1:t}^{(i,j)}, w_t^{(i,j)}\}_{j=1}^{M}\}_{i=1}^{N}$, and directly choose among them. This allows for a full exploration of the state space and implies that having multiple copies of the same particle should be excluded. A very effective way to sample N particles that unbiasedly estimate the target distribution with NM support points, $\{\{S_{1:t}^{(i,j)}, w_t^{(i,j)}\}_{j=1}^{M}\}_{i=1}^{N}$, is the optimal sampling procedure introduced by Fearnhead and Clifford [64]. They eloquently describe the problem as "that of approximating a discrete probability mass function with finite support by a *stochastic* probability mass function, with fewer support points." Their method

1. The APF was originally introduced by Pitt and Shephard [147] using auxiliary variables. Our exposition of this algorithm follows the alternative interpretation given by Creal [43], Doucet and Johansen [53] and Johansen and Doucet [109].

consists in selecting all particles with weights greater than or equal to a cutoff value $w^\star$ and in using systematic resampling [see 64, 50] to sample from the remaining particles to obtain a total of N particles. They demonstrate that this procedure is optimal among unbiased sampling schemes in the sense of minimizing the expected squared error with respect to the target distribution [see also 34, Section 8.2.2]. Algorithm 2 outlines the OPF for the MSG model.

Algorithm 2 : OPF for the MSG model

Exact filtering stage

Calculate $q = \lfloor \log(NM)/\log(M) \rfloor$, where $\lfloor \cdot \rfloor$ is the integer part function, and execute the exact filtering stage of Algorithm 1 (steps 1 to 4).

Approximate filtering stage

For $t = q+1, \ldots, T$:

5. Calculate $w^\star$, the unique solution to $N = \sum_i \min(w_{t-1}^{(i)}/w^\star, 1)$.

6. Keep all particles with weights greater than or equal to $w^\star$. Assume that there are $N^\star$ such particles. Use systematic resampling to sample $N - N^\star$ particles from the remaining particles (those with weights less than $w^\star$) and assign each resampled particle a weight of $w^\star$. The new set of N particles is denoted by $\{S_{1:t-1}^{(i)}, w_{1:t-1}^{(i)}\}_{i=1}^N$.

7. For $i = 1, \ldots, N$ and $j = 1, \ldots, M$, calculate :

$$a_t^{(i,j)} = f_{t-1}(y_t \mid S_{1:t-1}^{(i)}, S_t = j) \cdot p(S_t = j \mid S_{t-1}^{(i)}) \cdot w_{t-1}^{(i)} \text{ and } a_t = \sum_{i=1}^{N} \sum_{j=1}^{M} a_t^{(i,j)}.$$

8. For each $i = 1, \ldots, N$ and $j = 1, \ldots, M$, set $S_{1:t}^{(i,j)} = (S_{1:t-1}^{(i)}, j)$ and $w_t^{(i,j)} = a_t^{(i,j)}/a_t$. Combine indices (i, j) into a single index i to obtain $\{S_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^{NM}$. This weighted sample is an approximation to $p_t(S_{1:t})$.

After executing Algorithm 2, the log-likelihood of the MSG model is estimated with $\sum_{t=1}^T \log a_t$. The estimate of the likelihood obtained with particle filters is unbiased. Taking the logarithm of this estimate yields a biased, but consistent estimate of the log-likelihood [see 130]. The bias is of order $1/N$ and for small N, it is preferable to make an adjustment to the estimated log-likelihood to reduce it [see 145]. Algorithm 2 differs from traditional particle filters in at least two areas. First, it combines the resampling and sampling steps into a single step (step 6) by directly choosing among all possible descendants of previous particles. Second, it excludes the possibility of having multiple copies of the same particle, a property guaranteed through the use of systematic resampling at step 6. The

computational cost for Algorithm 2 is of the same order as for traditional particle filters, i.e., linear in the number of particles, since Fearnhead and Clifford [64] proved that step 5 of Algorithm 2 can be solved analytically through $\mathcal{O}(NM)$ operations.

3.3.2.1. *Understanding the particle filter output*

The theory underlying particle filtering was studied in detail by Del Moral [46] and sharp convergence results are now available [see 43, 53, 130, for an overview]. In our specific setting we are not concerned with convergence as $N \to \infty$ since Algorithm 2 provides exact filtering recursions as soon as $N = M^{T-1}$. The same remark can be made for Algorithm 1 (collapsing procedure) when $q = T$. However, since these exact recursions cannot be implemented in practice due to computational limitations, we are interested in the accuracy of these two algorithms for estimating the log-likelihood of the MSG model when $N \ll M^{T-1}$ or $q \ll T$.

Step 8 of Algorithm 2 generates a weighted sample, $\{S_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^{NM}$, offering an approximation to $p_t(S_{1:t})$. Unfortunately, due to the well-known degeneracy problem in particle filters, this approximation is poor as "it is inherently impossible to accurately represent a distribution on a space of arbitrarily high dimension with a sample of fixed, finite size" [53]. Nevertheless, when the model has good forgetting properties [see 51, 158], particle filters generally do well in approximating the log-likelihood and $p_t(S_{t-L+1:t})$ for some integer $L \geq 1$ [see 53, 115]. For example, the variance of the log-likelihood estimator is typically finite and increases only linearly with T. In practice we do not know L, but since $p_t(S_{t-L+1:t})$ is a discrete distribution with M^L support points, it is reasonable to suppose that it can be estimated with a number M^L of particles. In particular, if we use NM particles with $N = M^{q-1}$, then $L = q$.

To understand why a good approximation to $p_t(S_{1:t})$ is not necessarily needed to estimate the likelihood, consider the extended state variable, $\xi_t = (\sigma_t^2, S_t)$, taking values in $\mathbb{R}_{>0} \times \{1, \ldots, M\}$. The process $\{\xi_t\}$ is a homogeneous Markov chain and the conditional distribution of y_t can be written only as a function of ξ_t : $f_{t-1}(y_t \mid S_{1:t}) = f(y_t \mid \xi_t)$. [2] As a consequence, when calculating the likelihood

2. The MSG model can be formulated as a (hierarchical) hidden Markov model [see 34], albeit with nonstandard properties. For example, in a standard hidden Markov model the conditional distribution of y_t given $\xi_{1:T}$ only depends on ξ_t. Although we have $f(y_t \mid \xi_{1:t}) = f(y_t \mid \xi_t)$ for the MSG model, the conditional distribution $f(y_t \mid \xi_t, \xi_{t+1})$ is degenerate as the knowledge of (ξ_t, ξ_{t+1}) implies that of y_t.

component,

$$f_{t-1}(y_t) = \sum_{S_{1:t}} f_{t-1}(y_t \mid S_{1:t}) p_{t-1}(S_{1:t}) = \sum_{\xi_t} f(y_t \mid \xi_t) p_{t-1}(\xi_t),$$

we are most concerned in estimating $p_{t-1}(\xi_t)$. When $k < t$, the influence of ξ_{t-k} or $S_{1:k}$ on ξ_t progressively weakens as t increases due to the recursive nature of σ_t^2. In fact, Bauwens et al. [18] proved that the process $\{y_t, \xi_{t+1}\}$ is a geometrically ergodic Markov chain provided that $\gamma < 0$. This suggests that past values of the chain are *forgotten* at an exponential rate and that particle filtering should yield a reasonable approximation to $p_t(S_{t-L+1:t})$ and $p_t(\xi_t)$. Therefore, in practice, the NM particles, $\{S_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^{NM}$, should really be understood as representing the weighted sample $\{(\sigma_t^{2(i)}, S_{t-L+1:t}^{(i)}), w_t^{(i)}\}_{i=1}^{NM}$, where $\sigma_t^{2(i)} = \sigma_t^2(S_{1:t}^{(i)})$.

3.3.3. Comparison of Algorithms 1 and 2

In Section 3.3.1, we mentioned that there are two ways to interpret Algorithm 1 : (i) the exact Hamilton filter is executed on the simplified MSG model or (ii) an approximate (deterministic) filter is applied on the MSG model. The first of these interpretations was emphasized in Section 3.3.1, but we now explain the second one. Algorithm 2 was presented in Section 3.3.2 as an approximate (stochastic) filter for the MSG model. By comparing Algorithms 1 and 2 and supposing $N = M^{q-1}$, we observe that these two algorithms have an equivalent computational cost and only differ in the particle selection (or collapsing) stage : steps 5 and 6. Therefore, although these algorithms were derived with different approaches, they are based on the same filtering mechanism.

For instance, each iteration $t > q$ of these two algorithms starts with NM particles approximating the support of σ_t^2 and the probabilities $p_{t-1}(S_{t-q:t-1})$ and $p_{t-1}(\xi_{t-1})$. For Algorithm 2, this approximation is rendered with the weighted sample $\{(\sigma_{t-1}^{2(i)}, S_{t-q:t-1}^{(i)}), w_{t-1}^{(i)}\}_{i=1}^{NM}$. In Algorithm 1, the approximating set has this exact same structure with $\sigma_{t-1}^{2(i)} = \tilde{\sigma}_{t-1}^{2(i)}$ and $w_{t-1}^{(i)} = \tilde{p}_{t-1}(S_{t-q:t-1}^{(i)})$. Steps 5 and 6 have the role of collapsing these NM particles to avoid their exponential growth. Algorithm 2 selects a subset of N particles stochastically from the target distribution in such a way that it estimates it unbiasedly. On the other hand, Algorithm 1 merges each set of M particles with common states $S_{t-q+1:t-1}^{(i)}$, say $\{(\sigma_{t-1}^{2(j,i)}, S_{t-q:t-1}^{(j,i)}), w_{t-1}^{(j,i)}\}_{j=1}^{M}$, into a single particle, $\{(h_{t-1}^{(i)}, S_{t-q+1:t-1}^{(i)}), \sum_{j=1}^{M} w_{t-1}^{(j,i)}\}$, where

$$h_{t-1}^{(i)} = \sum_{j=1}^{M} \left[\sigma_{t-1}^{2(j,i)} \frac{w_{t-1}^{(j,i)}}{\sum_{j=1}^{M} w_{t-1}^{(j,i)}} \right].$$

In opposition to Algorithm 2, this new set of N particles is not a subset of the original NM particles nor an unbiased estimator. Therefore, the collapsing procedure can be characterized as a particle filter with a deterministic selection stage, or in short, a DPF.

The relationship established between the collapsing procedure and the OPF helps us understand why the application of the Hamilton filter on the simplified MSG model constitutes a valid approach for estimating the MSG log-likelihood : the collapsing procedure is an approximate (deterministic) filter for the MSG model. The main advantage of this filter over the OPF is that it generates a log-likelihood which is a smooth function of the parameters (provided, of course, that the innovation has a smooth density). This benefit does, however, come at the cost of introducing bias into the filter. This bias corresponds to a deterministic error for which it is difficult to derive an upper bound since similar collapsing procedures were not studied theoretically. In contrast, the bias and variance of the log-likelihood estimator obtained with the OPF can be studied with its sampling distribution. Therefore, the accuracy of Algorithm 1 can be investigated by comparing it to Algorithm 2.

3.4. Applications

Our applications of the algorithms developed in Section 3.3 are divided into three subsections. First, Section 3.4.1 demonstrates the effectiveness of the collapsing procedure for computing the MLE of the MSG model. Second, Section 3.4.2 compares computational times for Algorithms 1 (DPF), 2 (OPF), and 4 (*fully adapted* APF), and shows that the OPF is much more effective than the *fully adapted* APF for calculating the MSG log-likelihood. Finally, Section 3.4.3 investigates the robustness of the DPF by evaluating the bias in the collapsed log-likelihood for different parameter choices. All algorithms presented in this article were coded to work with version 2.15.3 of the R software [150] and are available on the website of the corresponding author. Computational times are reported based on a 3.40 GHz Intel Core i7-2600 processor and numerical optimizations were carried out using the solnp function available with the Rsolnp package [79, 167] in R.

In our applications, we allow for a switching mean in the MSG model. This extension does not require any modification to Algorithm 1 other than the redefinition of y_t to $y_t = \mu_{S_t} + \tilde{\sigma}_t \eta_t$, where $\{\mu_i\}_{i=1}^M$ are regime-dependent conditional means. This implies that the conditional distribution of y_t given $(y_{1:t-1}, S_{1:t})$ is

$\mathcal{N}(\mu_{S_t}, \tilde{\sigma}_t^2)$, and $\epsilon_{t-1} = y_{t-1} - \mu_{S_{t-1}}$. Analogous adjustments are needed for Algorithms 2 and 4. The dependence of ϵ_{t-1} on regime S_{t-1} requires a slight modification to Algorithm 3 (collapsing procedure with $q = 1$), as $\tilde{\sigma}_t^2$ is only allowed to depend on S_t. Consequently, for $q = 1$, the conditional variance must be defined as

$$\tilde{\sigma}_t^2 = \omega_{S_t} + \alpha_{S_t} \widetilde{\mathrm{E}} \left[\epsilon_{t-1}^2 \mid y_{1:t-1}, S_t \right] + \beta_{S_t} h_{t-1}.$$

3.4.1. Parameter estimation

The MSG model is estimated to two data sets : (i) weekly percentage log-returns on the S&P 500 price index from October 28, 1987 to October 31, 2012 (1305 observations), and (ii) daily percentage log-returns on the S&P 500 price index from May 20, 1999 to April 25, 2011 (3000 observations). These data sets were considered by Augustyniak [9] who developed a Monte Carlo expectation-maximization (MCEM) algorithm for maximum likelihood estimation of the MSG model. In addition, the daily data set was investigated by Bauwens et al. [16] and Dufays [57] who estimated the model with Bayesian MCMC techniques.

Table 3.1 presents maximum likelihood estimation results for the weekly data set. To match the specification considered by Augustyniak [9], we consider a MSG model with two regimes, and apply the constraints $\alpha = \alpha_1 = \alpha_2$ and $\beta = \beta_1 = \beta_2$ on the parameters. The computation of the MLE was accomplished by maximizing the simplified MSG log-likelihood (Algorithm 1) for increasing values of q. The starting value used in these optimizations, which we denote by θ_0 (see third row of Table 3.1), is based on an estimation of basic MS and GARCH model. The third to last column of Table 3.1 displays the maximized log-likelihood of the simplified model used for estimation which we denote by $\ell(q)$. The second to last column, denoted by ℓ, gives the log-likelihood of the MSG model computed with the OPF and $N = 131072$ (2^{17}) particles. The values of this log-likelihood are accurate to the first decimal place (see Section 3.4.2). The last column reports computational times for carrying out numerical optimizations.

A priori, we expect that as q increases the MLE based on Algorithm 1 will come close to the one reported by Augustyniak [9] (see last row of Table 3.1), which is exactly what occurs. The main advantage of Algorithm 1 over a MCEM algorithm [9] and Bayesian MCMC techniques [18, 16, 57] is that it allows for a fast estimation of the MSG model. For example, when $q = 10$ it takes less than two minutes to estimate the model. Considering higher values of q will improve

3. The accuracy of the MCEM algorithm depends on the number of iterations and simulations used. The results presented are based on a brute force implementation of the algorithm as in Augustyniak [9, Section 5.1] which is time consuming, but very precise.

TABLE 3.1. Maximum likelihood estimation of the MSG model for weekly data

	μ_1	μ_2	ω_1	ω_2	α	β	p_{11}	p_{22}	$\ell(q)$	ℓ	time
MS	0.28	−0.14	2.188	11.20			0.977	0.95	−2794.0	−2794.0	3s
GARCH	0.21		0.176		0.131	0.841			−2808.0	−2808.0	0.8s
θ_0	0.50	−0.50	0.100	0.50	0.100	0.850	0.980	0.95		−2799.4	
Gray	0.24	−2.37	0.000	4.34	0.070	0.848	0.984	0.49	−2777.1	−2823.0	6s
Klaassen	0.36	−2.79	0.000	2.71	0.066	0.875	0.926	0.16	−2758.9	−2762.6	10s
Dueker	0.37	−2.67	0.003	2.51	0.060	0.881	0.923	0.19	−2758.3	−2760.6	14s
$q = 4$	0.34	−2.90	0.033	2.54	0.047	0.899	0.946	0.29	−2758.1	−2758.0	16s
$q = 6$	0.33	−2.91	0.037	2.62	0.039	0.908	0.950	0.32	−2757.2	−2757.8	18s
$q = 8$	0.34	−2.81	0.038	2.59	0.040	0.905	0.947	0.31	−2757.0	−2757.6	32s
$q = 10$	0.34	−2.79	0.040	2.56	0.041	0.904	0.945	0.30	−2757.0	−2757.6	91s
$q = 12$	0.34	−2.80	0.042	2.54	0.042	0.903	0.946	0.31	−2757.2	−2757.5	5.1m
$q = 16$	0.34	−2.80	0.043	2.53	0.041	0.905	0.948	0.32	−2757.3	−2757.5	65m
$q = 20$	0.34	−2.81	0.044	2.52	0.042	0.904	0.947	0.31	−2757.4	−2757.5	33h
MCEM [3]	0.34	−2.80	0.045	2.51	0.043	0.902	0.946	0.31		−2757.5	≈ 10h

accuracy of the MLE, but not significantly enough to justify the large increase in computational time. In general, we recommend to first maximize the simplified log-likelihood for different initial values and a low value of q. This allows for a quick exploration of the log-likelihood modes and prevents convergence of the algorithm to a sub-optimal local maximum. This preliminary step provides a good starting point to the final optimization with a larger value of q.

Table 3.2 repeats the previous exercise for the daily data set. The MSG model is fitted with two regimes, and zero means to match the specification considered by Augustyniak [9], Bauwens et al. [16] and Dufays [57]. Our results indicate that the simplified MSG model offers an almost exact approximation to the MSG model, even when $q = 1$ (Klaassen). This is due to the presence of highly persistent regimes which enable us to accurately infer them after having observed the data. In fact, the dynamics of the estimated model resemble those of a structural break GARCH model as Bauwens et al. [16] identified only three regime switches for this data set. Consequently, only a negligible fraction of the possible 2^{3000} regime paths need to be considered in practice, which explains why the collapsing procedure works very well in this context.

TABLE 3.2. Maximum likelihood estimation of the MSG model for daily data

	ω_1	ω_2	α_1	α_2	β_1	β_2	p_{11}	p_{22}	$\ell(q)$	ℓ	time
MS	0.647	4.196					0.9895	0.9780	−4643.5	−4643.5	4s
GARCH	0.013		0.076		0.916				−4494.5	−4494.5	1s
θ_0	0.010	0.080	0.010	0.150	0.900	0.800	0.9900	0.9900		−4543.5	
Gray	0.041	0.048	0.012	0.100	0.856	0.887	0.9975	0.9999	−4480.0	−4505.1	14s
Klaassen	0.013	0.053	0.020	0.095	0.952	0.884	0.9990	0.9987	−4476.5	−4476.6	16s
Dueker	0.013	0.053	0.019	0.095	0.953	0.884	0.9990	0.9987	−4476.5	−4476.6	41s
$q = 4$	0.013	0.053	0.020	0.095	0.953	0.884	0.9990	0.9987	−4476.5	−4476.6	46s
$q = 6$	0.013	0.053	0.020	0.095	0.953	0.885	0.9990	0.9987	−4476.5	−4476.6	53s
$q = 8$	0.013	0.053	0.020	0.095	0.953	0.885	0.9989	0.9987	−4476.5	−4476.6	79s
$q = 10$	0.013	0.053	0.019	0.095	0.954	0.885	0.9989	0.9987	−4476.5	−4476.6	5.3m
MCEM [3]	0.013	0.053	0.019	0.095	0.953	0.885	0.9990	0.9987		−4476.6	≈ 30h

Finally, Tables 3.1 and 3.2 show that Gray's estimation procedure is not reliable for computing the MLE of the MSG model. Augustyniak [9] made this observation based a simulation study, but our work provides a theoretical justification for this assertion. For instance, the collapsed conditional variance at time t in Gray's model, h_t, does not take into account the observation y_t, which is very informative for identifying its most credible value. Moreover, Gray's estimation approach significantly deviates from the collapsing procedure with $q = 1$ in the presence of a switching mean as he defines

$$h_t = \widetilde{\mathrm{Var}}[y_t \mid y_{1:t-1}] = \widetilde{\mathrm{E}}[\tilde{\sigma}_t^2 \mid y_{1:t-1}] + \sum_{i=1}^{M} \mu_i^2 \tilde{p}_{t-1}(S_t) - \left(\sum_{i=1}^{M} \mu_i \tilde{p}_{t-1}(S_t) \right)^2, \quad \text{and} \tag{3.4.1}$$

$$\tilde{\sigma}_t^2 = \omega_{S_t} + \alpha_{S_t} \left(\widetilde{\mathrm{E}} \left[\epsilon_{t-1} \mid y_{1:t-1}, S_t \right] \right)^2 + \beta_{S_t} h_{t-1}. \tag{3.4.2}$$

For these reasons, Gray's model fails in providing a good approximation to the MSG model. This remark is important as many empirical studies employ Gray's approach [e.g., 8, 103, 128, 155] or incorrectly adapt Klaassen's model in the presence of a switching mean based on equations (3.4.1) and (3.4.2) [e.g., 131].

3.4.2. Effectiveness in computing the MSG log-likelihood

This section compares the effectiveness of Algorithms 1, 2 and 4 for computing the MSG log-likelihood. Algorithm 1 (DPF) is deterministic and yields a unique value given a choice of parameters. Algorithms 2 (OPF) and 4 (*fully adapted* APF) are stochastic and allow us to study the sampling distribution of the log-likelihood estimator. For a fair comparison between these two algorithms, systematic resampling is used at step 5 of Algorithm 4 to decrease the variance due to resampling. When $N = M^{q-1}$, all three algorithms are based on the same number of particles and have an equivalent computational cost of $\mathcal{O}(TN^q)$. Table 3.3 presents results for the computation of the MSG log-likelihood when it is evaluated at the MLE provided in Tables 3.1 ($q = 20$) and 3.2 ($q = 10$), and for N ranging from 32 to 131072. The mean and standard error (abbreviated StErr) of the log-likelihood estimator obtained with Algorithms 2 and 4 are based on 1000 estimates for $N \leq 512$, and 100 estimates for $N \geq 2048$. Algorithms 1, 2 and 4 are abbreviated DPF, OPF and APF, respectively, in Table 3.3.

First, the OPF is much more accurate than the *fully adapted* APF as it generates considerably lower standard errors for the log-likelihood estimator. For example, it offers variance reduction factors between 10 and 25 for weekly results ($N \geq 512$), and between 150 and 4000 for daily results. The improvement is much greater in the context of the daily MLE because in the presence of highly persistent regimes, only a few sequences of states are associated with large probabilities. The resampling step in Algorithm 4 unnecessarily replicates these particles many times, exacerbating the problem of sample impoverishment. This problem cannot be avoided as particle filters estimate a state space of growing dimension with a fixed number of particles. However, the OPF is able to reduce it to a minimum by avoiding particle duplication and by never discarding the most important regimes paths. Most importantly, the superior performance of the OPF over the *fully adapted* APF does not come at the cost of an increase in computational time.

The DPF is the fastest algorithm for evaluating the log-likelihood, especially at low values of N, but it is also the one for which bias decreases the slowest. This was to be expected as it uses a biased selection stage to choose among particles. In contrast, the bias in the log-likelihood estimator obtained with stochastic particle filters is of order $1/N$ and only arises because of Jensen's inequality : $\mathrm{E}(\log \widehat{f}) \leq \log \mathrm{E}(\widehat{f})$, where $\widehat{f}$ is an estimator of the likelihood. As a result, it can be approximated based on a second order Taylor expansion of $\log \widehat{f}$ at $\mathrm{E}(\widehat{f})$ [see 145] and is negligible when $N \geq 8192$. Although the DPF yields a larger

TABLE 3.3. Effectiveness in computing the MSG log-likelihood

	q	6	8	10	12	14	16	18
	$N = 2^{q-1}$	32	128	512	2048	8192	32768	131072
Log-likelihood evaluated at the MLE computed for $q = 20$ in Table 3.1								
Mean	DPF	−2757.50	−2757.09	−2757.08	−2757.23	−2757.27	−2757.29	−2757.35
	OPF	−2757.82	−2757.52	−2757.49	−2757.48	−2757.47	−2757.47	−2757.47
	APF	−2759.01	−2757.81	−2757.54	−2757.50	−2757.48	−2757.47	−2757.47
StErr	OPF	0.819	0.333	0.130	0.051	0.026	0.013	0.007
	APF	1.806	0.846	0.406	0.249	0.099	0.058	0.028
Time (s)	DPF	0.05	0.1	0.2	0.8	3.5	16	125
	OPF	0.9	0.9	1.4	3.2	12	62	259
	APF	0.8	0.9	1.3	2.9	11	66	261
Log-likelihood evaluated at the MLE computed for $q = 10$ in Table 3.2								
Mean	DPF	−4476.53	−4476.50	−4476.50	−4476.51	−4476.50	−4476.49	−4476.48
	OPF	−4476.58	−4476.55	−4476.55	−4476.56	−4476.56	−4476.55	−4476.55
	APF	−4482.96	−4477.06	−4476.61	−4476.56	−4476.57	−4476.56	−4476.56
StErr	OPF	0.236	0.083	0.028	0.009	0.003	0.001	0.0003
	APF	7.751	1.491	0.351	0.163	0.081	0.034	0.0181
Time (s)	DPF	0.1	0.2	0.6	2.3	10	83	315
	OPF	2.0	2.2	3.1	7.0	25	138	510
	APF	1.7	1.9	2.8	6.4	22	138	519

bias, it offers an important tradeoff : a smooth log-likelihood that can be maximized with gradient-based optimizers. It is possible to maximize the non-smooth log-likelihood computed with stochastic particle filters by using common random numbers and an optimizer for non-differentiable objective functions. However, we were not able to obtain accurate estimates with this alternative because for the estimation to complete in a reasonable amount of time, we are restricted to $N \leq 512$. Figure 3.1 illustrates the log-likelihood computed with the DPF ($q = 10$) and the OPF ($N = 512$ with common random numbers) as a function of β_1 in the vicinity of the daily MLE. Even though the OPF is very efficient for this particular case, the resulting log-likelihood approximation is still far from being smooth. Consequently, the DPF is best suited for likelihood optimization

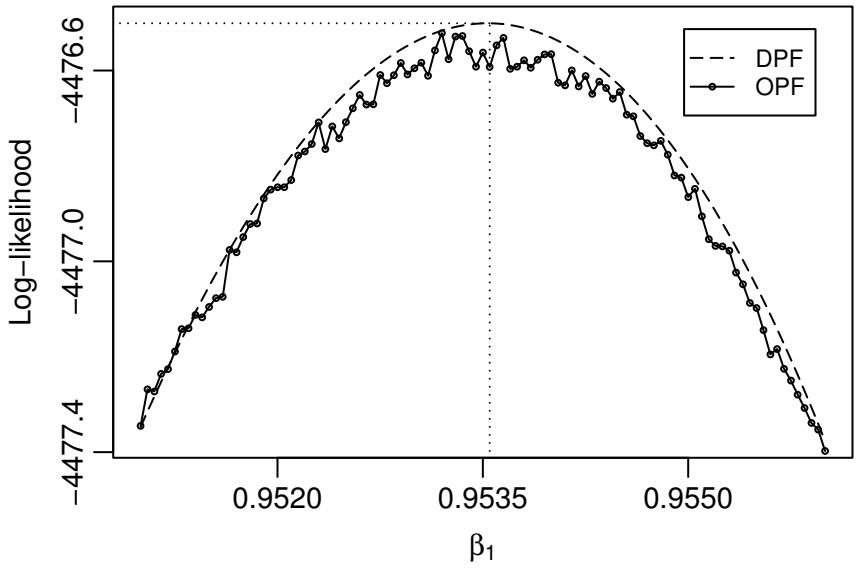

FIGURE 3.1. Log-likelihood computed with the DPF ($q = 10$) and the OPF ($N = 512$ with common random numbers) as a function of β_1 in the vicinity of the daily MLE

while the OPF is the most precise algorithm for calculating a point estimate of the MSG log-likelihood as it has negligible bias.

3.4.3. Robustness of the DPF

The robustness of the DPF is investigated by evaluating the bias in the collapsed log-likelihood for different parameter choices. To generate realistic parameters, we examined several studies that fitted MS-GARCH models to stock market index or exchange rate data [8, 9, 16, 56, 86, 84, 120, 131, 155]. Two types of models emerged. The first type is composed of two persistent regimes, one associated with low volatility and the other with high volatility. The second type includes a non-persistent regime which has the role of generating shocks in the return and volatility processes. This shock regime generally has a large negative mean return as well as a large value for the intercept in the GARCH equation. The daily MSG model estimated in Section 3.4.1 is of the first type (two persistent regimes), while the weekly MSG model is of the second type (one persistent and one shock regime).

We simulated 1000 parameter vectors for each of these two types of models based on a uniform distribution with lower and upper bounds given in Table 3.4. We separated these two cases because the bias of the DPF was observed to be smaller when regimes are highly persistent (see Sections 3.4.1 and 3.4.2). The parameters generated not only span the range of the parameter space where past

TABLE 3.4. Lower and upper bounds used for simulating parameters

	μ_1	μ_2	ω_1	α_1	β_1	ω_2	α_2	β_2	p_{11}	p_{22}
Two persistent regimes										
Lower	0.00	−0.30	0.00	0.01	0.90	0.0	0.02	0.70	0.990	0.950
Upper	0.15	0.05	0.05	0.05	0.98	1.0	0.15	0.95	0.999	0.999
One persistent and one shock regime										
Lower	0.00	−3.0	0.00	0.01	0.90	0.0	0.02	0.70	0.80	0.20
Upper	0.50	0.0	0.05	0.05	0.98	15.0	0.15	0.95	0.99	0.65

studies have estimated MS-GARCH models, but also produce MSG models which are not second-order stationary in approximately 10% of draws. This allows for a stronger test of robustness. For each parameter vector, we simulated a return path of length $T = 1500$ and 5000, and calculated the associated log-likelihood. This log-likelihood was computed with the OPF and $N = 131072$, as well as with the DPF and $q = 1, 2, 4, 8$, and 12. We considered these values of q because they represent reasonable choices for log-likelihood optimization. Afterwards, we calculated absolute differences between the log-likelihoods obtained with both algorithms and normalized them based on the highly accurate value computed with the OPF. Table 3.5 displays summary statistics for these relative errors, expressed in percentage. The standard deviation is abbreviated StDev.

As q increases, relative errors always decrease which was to be expected. However, this decrease is much more significant for parameter choices generating a shock regime. For example, from $q = 1$ (Klaassen) to $q = 12$, summary statistics are reduced by a factor of 4–5 for models with two persistent regimes, and by a factor of 15–30 for those with a shock regime. For low values of q, the DPF is more accurate in the presence of two persistent regimes, but for $q = 12$, the larger improvement in accuracy observed in the context of a shock regime results in comparable effectiveness for both ranges of parameters considered. Moreover, this effectiveness does not deteriorate with the sample size. We even observe a reduction in relative error from $T = 1500$ to 5000, implying that local approximations errors made at each time point in the collapsing procedure are more likely to offset each other than to accumulate over time. Finally, the maximum relative error committed in our simulations for $q = 12$ is 0.38%, which is very small. These results indicate that the collapsing procedure with a value from $q = 8$ to $q = 12$ should enable us to accurately approximate the MLE of the MSG model.

TABLE 3.5. Relative errors between log-likelihoods computed with the DPF and OPF (in %)

	$T = 1500$					$T = 5000$				
	Klaassen	Dueker	$q = 4$	$q = 8$	$q = 12$	Klaassen	Dueker	$q = 4$	$q = 8$	$q = 12$
Two persistent regimes										
Mean	0.042	0.035	0.027	0.017	0.011	0.034	0.028	0.019	0.011	0.007
StDev	0.068	0.058	0.044	0.028	0.019	0.053	0.043	0.029	0.017	0.011
90^{th} perc.	0.101	0.083	0.064	0.040	0.026	0.094	0.075	0.050	0.029	0.017
99^{th} perc.	0.316	0.264	0.223	0.114	0.079	0.257	0.212	0.145	0.096	0.054
Max.	0.843	0.705	0.662	0.546	0.382	0.630	0.519	0.323	0.181	0.130
One persistent and one shock regime										
Mean	0.192	0.122	0.060	0.023	0.011	0.178	0.110	0.050	0.016	0.008
StDev	0.402	0.230	0.102	0.038	0.019	0.360	0.202	0.085	0.027	0.013
90^{th} perc.	0.578	0.353	0.172	0.058	0.028	0.515	0.314	0.142	0.043	0.018
99^{th} perc.	1.814	1.142	0.470	0.198	0.092	1.815	1.054	0.427	0.146	0.076
Max.	4.085	2.221	0.992	0.481	0.258	3.572	1.600	0.661	0.236	0.107

3.5. CONCLUSION

We demonstrated how previously proposed approximations of the MSG model [56, 83, 120], based on collapsing ideas similar to Kim [117], can be generalized and improved to conduct fast and accurate maximum likelihood inference. Our general collapsing procedure yields an approximation of the MSG log-likelihood which is a continuous and differentiable function of the parameters, at the cost of introducing a small bias in its evaluation. A relationship between this method and particle filtering was established, allowing us to justify its validity. In fact, it was shown to be equivalent to a particle filter with a deterministic selection stage, i.e., the collapsing procedure can be characterized as a DPF. We demonstrated that the log-likelihood estimator obtained with the particle filter developed by Fearnhead and Clifford [64] is very precise in the context of the MSG model and generates considerably lower standards errors than traditional particle filters. However, its use for maximum likelihood estimation is unreliable and very slow because the log-likelihood approximation is not a continuous function of the parameters. Therefore, the smoothness property of the proposed DPF represents its main advantage over stochastic particle filters. Moreover, we explained why the model of Gray [83] fails in providing a good approximation to the MSG model.

In conclusion, we recommend to use the collapsing procedure with a value from $q = 8$ to 12 to approximate the MLE of the MSG model efficiently.

The proposed methodology can be extended without difficulty to more general GARCH specifications in the regimes, such as asymmetric power GARCH [49], with skewed or non-normal innovations. For example, Haas [84] generalized the path independent MS-GARCH model in this way and showed that asymmetric GARCH specifications with skewed normal innovations can provide a better fit to financial returns on equity markets. The estimation framework introduced thus provides a leeway for such an empirical analysis in the context of the MSG model, as well as for a comparison between path dependent and independent approaches to MS-GARCH models. It would also be interesting to study in greater depth the bias of the collapsing procedure, derive error bounds, and examine alternative ways to perform the collapsing process or deterministically select particles.

Acknowledgments

Financial support from the following sources is acknowledged : Fonds de recherche du Québec - Nature et technologies, Institut de finance mathématique de Montréal, Natural Sciences and Engineering Research Council of Canada and Society of Actuaries.

Appendix A. Comparison of MS-GARCH models

The path independent MS-GARCH model was proposed by Haas et al. [86] in response to the difficulties in estimating the MSG model. They hypothesized M separate GARCH processes evolving simultaneously in time ; at each time period one of these GARCH processes is chosen by the Markov chain to determine the conditional variance of the return. Their model is the following :

$$\begin{aligned} y_t &= \mu + \sigma_{t,S_t}\eta_t, \\ \sigma^2_{t,i} &= \omega_i + \alpha_i\epsilon^2_{t-1} + \beta_i\sigma^2_{t-1,i}, \qquad (3.5.1) \\ &= \sum_{j=1}^{t}\left[\left(\omega_i + \alpha_i\epsilon^2_{j-1}\right)\beta_i^{t-j}\right] + \sigma^2_{0,i}\beta_i^t, \quad i = 1,\dots,M. \qquad (3.5.2) \end{aligned}$$

where $\{\sigma^2_{t,i}\}_{t=0}^T$ denotes the separate GARCH process in state i. The conditional variance at time t is independent of $S_{1:t-1}$, i.e., $\mathrm{Var}(y_t \mid y_{1:t-1}, S_{1:t}) = \mathrm{Var}(y_t \mid y_{1:t-1}, S_t) = \sigma^2_{t,S_t}$. As a consequence, the log-likelihood can be calculated exactly and maximum likelihood estimation is tractable [see 92].

Given a sequence of regimes $S_{1:t}$, equations (3.2.3) and (3.5.2) show that in both, path dependent and independent models, the conditional variance at time

t is a linear function of past shocks $\{\epsilon_{j-1}^2\}_{j=1}^t$. The weight attributed to ϵ_{j-1}^2 is $\alpha_{S_j}\prod_{k=0}^{t-j-1}\beta_{S_{t-k}}$ in equation (3.2.3) [4] and $\alpha_{S_t}\beta_{S_t}^{t-j}$ in equation (3.5.2). To better understand this difference, consider a market participant who infers volatility based on a MS-GARCH model. If he uses equation (3.2.3), he never reevaluates the previous weights attributed to $\{\epsilon_{j-1}^2\}_{j=1}^t$ and simply updates them as new information arrives. In contrast, if he uses equation (3.5.2), the impact of previous shocks on the conditional variance are revised when there is a regime shift. This can reflect a transformation of the market's perception of previous shocks when the state of the economy changes.

In the MSG model, the feedback between successive conditional variances occurs more directly than in the path independent model. Consequently, this model shares the property of the standard GARCH model in that the variance employed to generate the observation at time $t-1$ is used to determine the variance of the return at time t. When there is a regime change from time $t-1$ to t, this property is not conserved in the model of Haas et al. [86]. This represents a difference between the two models, but is neither an advantage nor a disadvantage. Haas et al. [86] state that "economic intuition suggests that it is shocks that drive volatility," and, as explained in the previous paragraph, this main insight is preserved in both models.

Appendix B. Derivations

When $q \geq 2$, we obtain

$$\begin{aligned} h_t &= \widetilde{\mathrm{E}}\left[\tilde{\sigma}_t^2 \mid y_{1:t}, S_{t-q+2:t}\right] \\ &= \omega_{S_t} + \alpha_{S_t}\epsilon_{t-1}^2 + \beta_{S_t}\widetilde{\mathrm{E}}\left[\widetilde{\mathrm{E}}(\tilde{\sigma}_{t-1}^2 \mid y_{1:t-1}, S_{t-q+1:t-1}) \,\middle|\, y_{1:t}, S_{t-q+2:t}\right] \\ &= \omega_{S_t} + \alpha_{S_t}\epsilon_{t-1}^2 + \beta_{S_t}\widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-1}^2 \,\middle|\, y_{1:t}, S_{t-q+2:t}\right], \end{aligned} \tag{3.5.3}$$

by the law of iterated expectations since

$$\begin{aligned} \widetilde{\mathrm{E}}(\tilde{\sigma}_{t-1}^2 \mid y_{1:t-1}, S_{t-q+1:t-1}) &= \sum_{S_{t-q}} \tilde{\sigma}_{t-1}^2(S_{t-q:t-1})\tilde{p}_{t-1}(S_{t-q} \mid S_{t-q+1:t-1}) \\ &= \sum_{S_{t-q}} \tilde{\sigma}_{t-1}^2(S_{t-q:t-1})\tilde{p}_t(S_{t-q} \mid S_{t-q+1:t}) \\ &= \widetilde{\mathrm{E}}(\tilde{\sigma}_{t-1}^2 \mid y_{1:t}, S_{t-q+1:t}). \end{aligned}$$

4. To better associate the impact of a shock on the variance with its corresponding regime, a more natural definition of the MSG model is perhaps the one proposed by Dueker [56] :

$$\sigma_t^2(S_{1:t}) = \omega_{S_t} + \alpha_{S_{t-1}}\epsilon_{t-1}^2 + \beta_{S_{t-1}}\sigma_{t-1}^2(S_{1:t}),$$

which results in a weight of $\alpha_{S_{j-1}}\prod_{k=1}^{t-j}\beta_{S_{t-k}}$ for the shock term ϵ_{j-1}^2 in the conditional variance at time t.

Given $(y_{1:t-1}, S_{t-q+1:t-1})$, the variables (y_t, S_t) are independent of S_{t-q} and can thus be introduced into the conditioning set. Assuming $t \geq q \geq 2$, recursive substitution of expression (3.5.3) into

$$\tilde{\sigma}_t^2 = \omega_{S_t} + \alpha_{S_t}\epsilon_{t-1}^2 + \beta_{S_t}h_{t-1},$$

yields

$$\tilde{\sigma}_t^2 = \sum_{j=0}^{q-1}\left[\left(\omega_{S_{t-j}} + \alpha_{S_{t-j}}\epsilon_{t-1-j}^2\right)\prod_{k=0}^{j-1}\beta_{S_{t-k}}\right] + \widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-q}^2 \;\middle|\; y_{1:t-1}, S_{t-q+1:t-1}\right]\prod_{k=0}^{q-1}\beta_{S_{t-k}}.$$

Therefore, whenever $q \geq 1$, δ_t can be written as

$$\delta_t = \sigma_t^2 - \tilde{\sigma}_t^2 = \begin{cases} 0, & t = 0, ..., q; \\ \left(\sigma_{t-q}^2 - \widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-q}^2 \;\middle|\; y_{1:t-1}, S_{t-q+1:t}\right]\right)\prod_{k=0}^{q-1}\beta_{S_{t-k}}, & t \geq q+1. \end{cases}$$

Expression (3.3.9) is obtained by recursive substitution of (3.5.4).

We now demonstrate $\tilde{\sigma}_t^2 = \widetilde{\mathrm{E}}\left[\sigma_t^2 \mid y_{1:t-1}, S_{t-q+1:t}\right]$. Taking the conditional expectation of expression (3.3.9), for $t \geq q+1$, we obtain

$$\begin{aligned}\widetilde{\mathrm{E}}\left[\delta_t \mid y_{1:t-1}, S_{t-q+1:t}\right] &= \sum_{j=0}^{\lfloor t/q \rfloor - 1}\widetilde{\mathrm{E}}\left[\tilde{\delta}_{t-jq}B_{(j+1)q} \;\middle|\; y_{1:t-1}, S_{t-q+1:t}\right] \\ &= \sum_{j=0}^{\lfloor t/q \rfloor - 1} B_q\widetilde{\mathrm{E}}\left[\tilde{\delta}_{t-jq}\prod_{k=q}^{(j+1)q-1}\beta_{S_{t-k}} \;\middle|\; y_{1:t-1}, S_{t-q+1:t}\right]. \quad (3.5.5)\end{aligned}$$

By definition, we have

$$\tilde{\delta}_{t-jq} = \tilde{\sigma}_{t-(j+1)q}^2 - \widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-(j+1)q}^2 \;\middle|\; y_{1:t-jq-1}, S_{t-(j+1)q+1:t-jq}\right]. \quad (3.5.6)$$

The variables $(y_{t-jq:t-1}, S_{t-jq+1:t})$ can be introduced into the conditioning set of the second term on the right-hand side of equation (3.5.6) :

$$\widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-(j+1)q}^2 \;\middle|\; y_{1:t-jq-1}, S_{t-(j+1)q+1:t-jq}\right] = \widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-(j+1)q}^2 \;\middle|\; y_{1:t-1}, S_{t-(j+1)q+1:t}\right].$$

Therefore, we obtain

$$\tilde{\delta}_{t-jq}\prod_{k=q}^{(j+1)q-1}\beta_{S_{t-k}} = \tilde{\sigma}_{t-(j+1)q}^2\prod_{k=q}^{(j+1)q-1}\beta_{S_{t-k}} - \widetilde{\mathrm{E}}\left[\tilde{\sigma}_{t-(j+1)q}^2\prod_{k=q}^{(j+1)q-1}\beta_{S_{t-k}} \;\middle|\; y_{1:t-1}, S_{t-(j+1)q+1:t}\right]. \quad (3.5.7)$$

Substituting expression (3.5.7) into (3.5.5) yields $\widetilde{\mathrm{E}}\left[\delta_t \mid y_{1:t-1}, S_{t-q+1:t}\right] = 0$, which in turn implies $\tilde{\sigma}_t^2 = \widetilde{\mathrm{E}}\left[\sigma_t^2 \mid y_{1:t-1}, S_{t-q+1:t}\right]$. This result also allows us to relate h_t with σ_t^2 : $h_t = \widetilde{\mathrm{E}}\left[\sigma_t^2 \mid y_{1:t}, S_{t-q+2:t+1}\right]$.

Appendix C. Algorithms

Algorithm 3 : Collapsing procedure for the MSG model (q = 1)

Exact filtering stage

1. Calculate $a_1^{(j)} = f(y_1 \mid S_1 = j) \cdot p(S_1 = j \mid S_0)$ for $j = 1, \ldots, M$, and $a_1 = \sum_{j=1}^{M} a_1^{(j)}$.
2. For each $i = 1, \ldots, M$, set $S_1^{(i)} = i$ and $w_1^{(i)} = a_1^{(i)}/a_1$ to obtain $\{S_1^{(i)}, w_1^{(i)}\}_{i=1}^{M}$.

Approximate filtering stage

For $t = 2, \ldots, T$:

5. Collapse $\tilde{\sigma}_{t-1}^2$ into h_{t-1} using equations (3.3.5)–(3.3.7).
6. For $i, j = 1, \ldots, M$, calculate :

$$a_t^{(i,j)} = \tilde{f}_{t-1}(y_t \mid S_t = j) \cdot p(S_t = j \mid S_{t-1}^{(i)}) \cdot w_{t-1}^{(i)} \text{ and } a_t = \sum_{i=1}^{M} \sum_{j=1}^{M} a_t^{(i,j)}.$$

7. For each $j = 1, \ldots, M$, set $S_t^{(j)} = j$ and $w_t^{(j)} = \sum_{i=1}^{M} a_t^{(i,j)}/a_t$.

Algorithm 4 : *Fully adapted* APF for the MSG model

1. Calculate $a_1^{(j)} = f(y_1 \mid S_1 = j) \cdot p(S_1 = j \mid S_0)$ for $j = 1, \ldots, M$, and $a_1 = \sum_{j=1}^{M} a_1^{(j)}$.
2. For each $i = 1, \ldots, N$, simulate state $S_1^{(i)} = j$ proportionally to $a_1^{(j)}$ to obtain the weighted sample $\{S_1^{(i)}, 1/N\}_{i=1}^{N}$.

For $t = 2, \ldots, T$:

3. For $i = 1, \ldots, N$ and $j = 1, \ldots, M$, calculate :

$$a_t^{(i,j)} = f_{t-1}(y_t \mid S_{1:t-1}^{(i)}, S_t = j) \cdot p(S_t = j \mid S_{t-1}^{(i)}) \cdot 1/N \text{ and } a_t = \sum_{i=1}^{N} \sum_{j=1}^{M} a_t^{(i,j)}.$$

4. For each $i = 1, \ldots, N$, set $w_t^{(i)} = \sum_{j=1}^{M} a_t^{(i,j)}/a_t$.
5. Resample $\{S_{1:t-1}^{(i)}, w_t^{(i)}\}_{i=1}^{N}$ in proportion to the importance weights to obtain a new set of particles $\{S_{1:t-1}^{(i)}, 1/N\}_{i=1}^{N}$.
6. For each $i = 1, \ldots, N$, simulate state $S_t^{(i)} = j$ proportionally to $a_t^{(i,j)}$ to obtain $\{S_{1:t}^{(i)}, 1/N\}_{i=1}^{N}$.

Chapitre 4

AN OUT-OF-SAMPLE ANALYSIS OF INVESTMENT GUARANTEES FOR EQUITY-LINKED PRODUCTS : LESSONS FROM THE FINANCIAL CRISIS OF THE LATE 2000S

Ce chapitre inclut un article réalisé avec la collaboration du coauteur Mathieu Boudreault. Il débute avec un avant-propos permettant de faire le lien avec les deux derniers chapitres. L'article de ce chapitre est publié dans la revue *North American Actuarial Journal* [10]. Nous remercions un arbitre anonyme pour nous avoir donné des commentaires qui ont permis d'améliorer son contenu.

Avant-propos

La science actuarielle est une discipline qui applique des méthodes mathématiques et statistiques dans les domaines de la finance et des assurances. Elle a surtout pour objectif l'évaluation des risques à long terme, tels que le risque de mortalité, le risque de longévité, le risque d'investissement et la pérennité des régimes de retraite. Ce chapitre focalise sur l'évaluation et la gestion du risque d'investissement à long terme dans le cadre de produits financiers offerts avec des garanties, connus sous l'appellation fonds distincts au Canada (*segregated funds*, *variable annuities*, *equity-linked insurance*). Un fonds distinct est un produit d'investissement vendu par des compagnies d'assurances au Canada. Il est semblable à un fonds commun de placement (*mutual fund*) assorti de garanties protégeant le capital investi contre une baisse des marchés boursiers.

Un exemple d'un produit financier offert avec une garantie d'investissement est le *guaranteed minimum maturity benefit*, connu sous l'acronyme GMMB. Un

investisseur dépose un montant de $A_0 > 0$ dans un fonds qui suit la performance d'un indice boursier tel que le S&P 500. La valeur du fonds au temps t est représentée par A_t. Le fonds garanti à l'investisseur qu'il pourra récupérer son investissement initial après T années, peu importe la performance du fonds. À l'échéance T du contrat, l'investisseur a droit à $\max(A_T, A_0)$. Si la valeur du fonds au temps T est supérieure à A_0, on dit que la garantie est hors du cours (*out-of-the-money*) et l'assureur ne doit pas faire de versement à l'assuré. Dans le cas contraire, la garantie est dans le cours (*in-the-money*) et il a l'obligation de verser $A_0 - A_T$ à l'assuré pour que ce dernier puisse récupérer son investissement initial. Par conséquent, le versement de l'assureur à l'échéance du contrat correspond à celui d'une option de vente (*put option*), soit $\max(0, A_0 - A_T)$. Le produit GMMB est donc équivalent à un investissement dans un fonds commun de placement (*mutual fund*) combiné à l'achat d'une option de vente. Cependant, l'assuré ne paie pas pour cette option à la signature du contrat, mais plutôt, des frais sont déduits en proportion de la valeur du fonds permettant ainsi de compenser l'assureur pour la protection offerte.

Pour couvrir son risque de perte, l'assureur pourrait en principe acheter l'option de vente sous-jacente au contrat sur le marché. Cependant, les garanties offertes ont généralement de longues échéances, typiquement de $T = 3$ à 20 années, et des options avec de telles échéances ne sont pas disponibles sur le marché ou ne sont pas liquides. Il est donc difficile pour la compagnie d'assurances de transférer ce risque à une contrepartie, ce qui implique qu'elle doit gérer elle-même le risque lié aux protections offertes. Par conséquent, les assureurs doivent faire appel aux modèles économétriques pour modéliser les variables financières dont dépendent leurs produits.

La crise financière de la fin des années 2000 a remis en question les modèles et stratégies de gestion des risques utilisés dans les deux dernières décennies, car les compagnies d'assurances ont subi des pertes significatives sur leurs produits. Elle a souligné l'importance de considérer des modèles permettant de répliquer les dynamiques des actifs financiers observées sur de longues périodes de temps, d'où l'intérêt des modèles GARCH à changement de régimes. Les chapitres 2 et 3 ont focalisé sur l'estimation du modèle GARCH à changement de régimes. L'article contenu dans ce chapitre considère l'application de cette classe de modèles aux applications financières en actuariat. Il compare 78 modèles économétriques, dont plusieurs généralisations du modèle GARCH à changement de régimes, et examine l'efficacité des stratégies de gestion des risques utilisées dans le contexte de la crise financière. L'utilisation d'un large éventail de modèles permet de démontrer

l'importance que joue le risque de modèle sur l'évaluation et la gestion du risque d'investissement à long terme.

Résumé

Le risque lié aux garanties d'investissement est analysé à l'aide de 78 modèles économétriques : modèles GARCH, modèles à changement de régimes, mélanges, ainsi que des combinaisons de ces méthodes. Les rendements projetés avec ce vaste éventail de modèles sont comparés aux rendements extrêmes observés pendant la crise financière de la fin des années 2000 dans le cadre d'une analyse hors-échantillon. Malgré l'ajustement adéquat aux données des modèles récents, peu d'entre eux sont capables de générer des rendements négatifs d'une ampleur comparable à ceux observés pendant la crise financière. De plus, les mesures de risque varient considérablement d'un modèle à l'autre soulignant l'importance du risque de modèle. Puisque les compagnies d'assurances implantent maintenant des stratégies de couverture dynamique pour gérer le risque de leurs produits vendus avec des garanties, la robustesse de la couverture en delta de Black-Scholes est également étudiée. Les erreurs de couverture peuvent être importantes signifiant que le risque de modèle doit être tenu en compte dans la couverture des garanties d'investissement.

Abstract

The risk underlying investment guarantees is analyzed with 78 different econometric models, namely GARCH, regime-switching, mixtures, and combinations of these approaches. The returns generated under this extensive set of models are compared with extreme returns observed during the financial crisis of the late 2000s in an out-of-sample analysis. Despite the very good fit of recent models, few are capable of generating negative returns similar in magnitude to those observed during the financial crisis. Moreover, tail risk measures vary significantly across models, emphasizing the importance of model risk. As many insurance companies are now dynamically hedging their investment guarantees, the robustness of the Black-Scholes delta hedging strategy is also investigated. Hedging errors can be very large among top fitting models, implying that model risk must be taken into consideration when hedging investment guarantees.

Keywords : investment guarantees, out-of-sample, actuarial approach, delta hedging, model risk

4.1. Introduction

Investment guarantees are very popular features in life insurance policies because in addition to paying a death benefit, these policies are tied to the return of an underlying asset or an actively managed portfolio. Thus, the policy also acts as an investment because the investor's capital is credited a minimum return. In exchange for this protection, the policyholder pays a higher premium, reflecting the market risk assumed by the insurance company. Because the investment guarantee is essentially a non-standard long-term put option, it is very difficult for insurance companies to completely match this liability with a similar put option on the market. Thus, insurers need econometric models to forecast the potential loss on this guarantee.

The Canadian Institute of Actuaries (CIA) (through the Task Force on Segregated Fund Investment Guarantees), the American Academy of Actuaries (AAA) (through the RBC C3 Phase II report) and Hardy [94, 96] have strongly recommended the use of stochastic (econometric) models for reserving the loss on investment guarantees. The industry, backed by professional organizations, has been mainly managing these products using the traditional actuarial approach. This consists in projecting the loss on investment guarantees using multiple scenarios of the underlying asset, and reserving a sufficient amount based upon tail risk measures (say the Value-at-Risk (VaR) or the Conditional Tail Expectation (CTE)). However, the recent financial crisis has forced many banks and insurers to revise their risk management policies and many life insurance companies are now dynamically hedging their investment guarantees.

In this paper, we conduct an extensive analysis of basic and advanced univariate (single asset) econometric models for the purpose of managing investment guarantees [for a study of multivariate (multiple assets) models, refer to 24]. There are mainly three reasons why we decided to make this investigation. First, the univariate econometric literature has evolved significantly since Hardy [94, 95, 96], Wong and Chan [162] and Hardy et al. [97]. Second, the year 2008 has seen very large investment banks (almost) default, leading to the worst recession since the Great Depression. Many insurers suffered important losses on their segregated funds or variable annuities because of this financial crisis. As a consequence, insurance companies are now focusing more and more on dynamic hedging. However, and this is our third motivation for this analysis, the empirical literature regarding hedging investment guarantees is scarce.

The purposes of our paper are three-fold. First, we would like to complete the empirical analysis of Hardy et al. [97] by analyzing the fit of 78 models (likelihood-based criteria, quantile-quantile plots, normality and heteroscedasticity tests). These models differ in their volatility dynamics (GARCH-type, regime-switching (RS) or combinations of these) and the use of different error distributions. Rather than recommending a specific model, these models form the basis of a thorough robustness analysis. Second, we analyze the capability of these models to consistently generate low returns over long periods of time. This enables us to assess if reserving approaches, based upon traditional actuarial techniques, would have been appropriate to cover losses of various investment guarantees that matured during the financial crisis. Third, we apply the Black-Scholes delta hedging strategy with scenarios generated from these 78 models to analyze the distribution of hedging errors. Results are compared with similar contracts that could have been issued in the past to check the validity of the outcomes.

Even if it is possible to find a model with an excellent fit, this model will always be a simplification of the true market dynamics, which are unknown. The fact that 78 models are analyzed helps understand an important risk that is often overlooked : model risk, i.e., the uncertainty regarding the appropriateness of a chosen model with respect to the true dynamics of the underlying process. Within a subset of models that fit the market data very well, we find significant variations across models, i.e., important model risk. This is true for tail risk measures of the loss on investment guarantees as well as for hedging errors.

This paper is structured as follows. Section 4.2 briefly discusses the various models that were fitted and the results of numerous statistical tests that were conducted on these models. Section 4.3 focuses on the capability of the models to generate low returns over long periods of time. In Section 4.4, we analyze the robustness of the Black-Scholes delta hedging strategy and Section 4.5 concludes. Appendix A shows our results for all of the 78 models since the body of the paper focuses on a subset of these models.

4.2. Models and their adequacy

This section summarizes the models used in our analysis along with the results of numerous statistical tests that assess the quality of the fit of these models. The data that we consider is the set of monthly log-returns on the S&P 500 total return index between February 1956 and December 2010. This data set comprises the financial crisis of the late 2000s. Note that financial markets showed an initial sign of weakness in July 2007 when investors lost confidence in mortgage-backed securities but the debacle started in September 2008 with the bankruptcy of

Lehman Brothers. Financial markets generally reached their lows in the beginning of March 2009.

4.2.1. Overview of models

In this section, we briefly present and describe the different classes of models that are considered in our paper. These models are widely used in the econometric literature to model data and aim to replicate the broadly accepted stylized facts of financial data. These include fat tails and negative skewness of the returns' distribution, jumps in the volatility, volatility clustering and the leverage effect which suggests that past returns and future volatilities are negatively correlated [see 31, 160].

4.2.1.1. *GARCH models*

The GARCH family of models allows for the variance in the returns to be time-varying. Given past information, the current variance depends on previous innovations and variances, and is thus conditionally deterministic. There exists many different variations of GARCH models : (1) the standard GARCH model of Bollerslev [21] ; (2) the Glosten-Jagannathan-Runkle GARCH (GJR GARCH) of Glosten et al. [80] ; (3) the asymmetric power ARCH (APARCH) of Ding et al. [49] ; (4) the exponential GARCH (EGARCH) of Nelson [142]. The GJR GARCH model is an extension of the standard GARCH model as it allows the variance equation to respond asymmetrically to shocks to take into account the leverage effect. The APARCH model further generalizes that model by allowing the volatility exponent in the variance equation to be different than two. The EGARCH model is similar to the GJR GARCH model in the sense that it allows asymmetric shocks but it parametrizes the logarithm of the variance instead of the variance. For a survey of GARCH models, refer to Bollerslev [23] or Tsay [160].

4.2.1.2. *RS models*

Models in the RS family involve a process that switches between two or more distributions according to a Markov chain. This Markov chain is usually unobserved and is often given an economic interpretation such as representing the states of an economy (e.g., recession and expansion). For instance, a process that switches between two normal distributions with different means and variances is a RS model if the switching is governed by a Markov chain. In that example, given past observed information, the conditional variance is stochastic since the current

state of the economy is unknown. The term RS is not the only term used to describe such a model in the academic literature. Alternative terms include hidden Markov model (HMM), hidden Markov process, Markov-dependent mixture and Markov-switching model. Hamilton [90] is generally cited as the one to have introduced and popularized RS models in the economic and econometric literature although some of his ideas were already present in Goldfeld and Quandt [81]. RS models became popular in actuarial science since Hardy [94] used them in the context of investment guarantees. The RS model of Hardy [94] has been shown to provide a very good fit to the left tail of the returns' distribution in addition to feature a closed-form solution for the computation of the CTE reserve.

4.2.1.3. *RS-GARCH models*

It is not difficult to combine a GARCH model with a RS framework to build a RS-GARCH model. One way to justify such a combination is given by Lamoureux and Lastrapes [125] who show that the high persistence observed in the variance of financial returns can be due to time-varying GARCH parameters. Estimating parameters of a RS-GARCH model is a challenging task because at each time point, the conditional variance depends on the entire history of regimes. This path dependence problem renders exact computation of the likelihood infeasible in practice [see 83, 93, for discussions]. To circumvent this problem, many authors have proposed simplifications, either in the model or in its estimation. In this paper, we consider three popular methods in the econometric literature to estimate RS-GARCH models : Gray [83], Klaassen [120] and Haas et al. [86]. We note that Gray's RS-GARCH model was fitted in Hardy et al. [97].

4.2.1.4. *Distribution of the error term*

In all of the previous models, the error distribution is usually normally distributed but other assumptions are possible as well. In this work, the following distributions were used for the error term : (1) the normal distribution (abbreviated NORM) ; (2) the Student's t-distribution (abbreviated STD) ; (3) the generalized error distribution (abbreviated GED) as defined by Nelson [142] ; (4) the normal inverse Gaussian distribution (abbreviated NIG). Skewed versions of these distributions are also considered where skewness is introduced by the method of Fernandez and Steel [65]. These distributions and their skewed counterparts are available from the R software [150] as part of the fGarch package [165] and the fBasics package [164].

We note that it is common to combine GARCH models with heavy tailed error distributions. For example, when Nelson [142] introduced the EGARCH model,

he proposed to use the GED distribution. In addition, Bollerslev [22] supports the use of the GARCH model with a Student's t-distribution based on its fit to five different monthly stock price indices for the U.S. economy. Moreover, the distribution of the error term in RS models is not restricted to the normal law and estimation of the parameters is not more complicated when distributions other than the normal are considered [see 92]. For example, Elliott and Miao [60] advocate the use of a RS model with a Student's t-distribution for the error term.

4.2.1.5. *Models used*

All of the models used in this work were estimated by maximum likelihood with the R software. The GARCH models enumerated in Section 4.2.1.1 were considered [1] with the different error distributions of Section 4.2.1.4. These models are available in the rugarch package [78]. To refer to these models in our tables, we use the model name followed by the error distribution in parentheses. We add an S to the abbreviated name of the error distribution to denote the skewed version of the error distribution. For example, the EGARCH model with a skewed Student's t-distribution is denoted by EGARCH (SSTD).

We considered both RS and mixture (abbreviated MIX) models. Mixture models can be seen as RS models where the transition probability matrix has identical rows. These models were combined with all of the error distributions [2] of Section 4.2.1.4. We refer to these models in our tables in a manner analogous to GARCH models. For example, we denote the RS model with a normal distribution by RS (NORM). Note that the RS (NORM) model in this paper is the same model as the RS lognormal (RSLN) model of Hardy [94]. Mixture models were used to determine whether regime persistence is important in the RS models. In all but one of the RS models that we fitted, we used two regimes, the exception being the RS3 (NORM) which includes three normal regimes. We note that independent models with the distributions of Section 4.2.1.4 were also considered.

Moreover, we used many models in the RS-GARCH class but we limited the error distribution in these models to the normal and the skewed normal. The models of Gray [83], Klaassen [120] and Haas et al. [86] were fitted as described in the original papers. We also exploited Haas [84]'s generalization of the model of Haas et al. [86] which involves an APARCH structure in each regime and a skewed normal distribution for the error term. Similarly as in that article, we did not estimate the power coefficient of the APARCH equations but we rather set it

1. We restricted ourselves to GARCH(1,1) representations, i.e., we allowed only one lag of the shock and variance terms in the GARCH equation.

2. We let the kurtosis parameter switch across regimes but the skewness parameter was held constant in all regimes.

to one or two. For example, that model with a power coefficient of 1 is denoted by RS-APARCH-Haas (SNORM ; pow=1) in our tables. We also extended Klaassen's RS-GARCH model in an equivalent way and considered many variants of Klaassen's model such as a RS-EGARCH. Whenever we do not specify which approach is used to estimate a RS-GARCH model, it is assumed that it is Klaassen's approach that is employed. We also considered a restricted version of Klaassen's model which is a MIX-GARCH model where the α parameter in the GARCH equation is set to zero. With a skewed normal distribution for the error term, this model is denoted by MIX-GARCH (SNORM ; $\alpha = 0$).

Moreover, we want to note that almost all of the RS-GARCH models that we fitted only allow the constant term in the GARCH equations to switch across regimes [3]. When this is not the case, we added the mention *full* in parentheses to indicate that all parameters are allowed to switch across regimes. This restriction was adopted due to parsimony concerns ; all but one of the models considered have 10 parameters or less.

Finally, a total of 78 models are considered in this article. Regarding the desired features of an econometric model, it should be noted that the very large majority of these models have time-varying volatility and jumps in the volatility process are present in RS models. Classic continuous-time stochastic volatility and jump-diffusion models were not considered in our article because once discretized, they behave very similarly to the previous models [see 54, 55, 141] with an added estimation complexity.

4.2.2. Overview of tests

The capability of the models to fit the data and to replicate patterns of time-varying volatility was checked with various criteria and statistical tests. First, the global fit among models was compared with log-likelihood values, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). More details on these criteria can be found in Klugman et al. [123].

The Ljung-Box test [see 156] with 10 lags was applied to residuals and to squared residuals to check if there is autocorrelation in the residuals and if there is still some heteroscedasticity that has not been accounted for in the model. Only results for squared residuals are shown in our tables since, for all models, the test never rejects the null hypothesis of no autocorrelation in the residuals (based on a 5% significance level). We also used the ARCH Lagrange multiplier test [see 160,

3. Haas [84] finds that RS-GARCH models with only the constant term switching in the GARCH equations are preferred according to fit criteria.

p. 114] with 10 lags as an additional heteroscedasticity test. Finally, the normality of the residuals was assessed with the Shapiro-Wilk and Jarque-Bera tests.

The tests mentioned in the previous paragraph rely on the assumption that residuals are normally distributed. In many instances in our models, the error distribution is not normal and these tests cannot be directly applied on the residuals. Moreover, the indicator or weighted residuals, which are often used in the context of RS models, are not normally distributed even when the error distribution is normal [see 72]. Freeland et al. [72] propose a way to obtain residuals in RS models which are normally distributed but the drawback of their approach is that we do not obtain a single set of residuals.

To compute residuals which are normally distributed using the same method for all models, we use the approach presented in Haas [84] which is based on the Rosenblatt transform [154]. To this end, we calculate the quantities $u_t = \widehat{F}(y_t \mid y_{t-1}, y_{t-2}, \ldots)$ for each t, where $\widehat{F}(y_t \mid y_{t-1}, y_{t-2}, \ldots)$ denotes the conditional cumulative distribution function (CDF) of the observed return at time t, y_t, under the estimated model. We then transform these values using the inverse of the standard normal CDF and this resulting batch of residuals should behave as a sequence of i.i.d standard normal variates if the model is well specified.

4.2.3. Results

In this section, we analyze the quality of the fit of the models. Tables 4.1 and 4.2 present our results for a subset of the models that we considered. These are either important benchmarks or models that fit the data well. Results for all of the 78 models are given in Tables 4.8 and 4.9 of Appendix A. A first observation that we can draw from Table 4.1 is that the inclusion of the financial crisis did not affect the relative ranking of benchmark models ; the RS (NORM) still has a better fit than the GARCH (NORM) and many other simpler models. This confirms the results of Hardy [94, 96], even 10 years later. However, more recent econometric models that are also parsimonious do better with respect to the BIC. In Hardy et al. [97], Gray's RS-GARCH had a better global fit, and it is still the case here. Klassen's RS-GARCH does even better. APARCH and EGARCH models with skewed error distributions are among the top models in terms of the BIC but it is the RS-EGARCH (SNORM) that achieves the highest score. The gain in fit over the RS (NORM) is approximately 20 points, which is roughly equivalent to the gain in fit of the RS (NORM) when compared to the NORM model. This can provide an idea of how important that improvement is. Moreover, when APARCH models are combined with RS, it is generally the mixture version of these models that is preferred. This entails that the role of RS

TABLE 4.1. Fit summary

Model	Params	Log-Lik	AIC	BIC
RS-EGARCH (SNORM)	10	1210.2	1200.2	1177.7
EGARCH (SSTD)	7	1198.1	1191.1	1175.4
MIX-APARCH-Haas (SNORM ; pow=1)	8	1200.0	1192.0	1174.1
MIX-GARCH (SNORM ; $\alpha = 0$)	7	1194.9	1187.9	1172.2
APARCH (SNIG)	8	1195.9	1187.9	1169.9
MIX-APARCH (SNORM ; pow=1)	9	1198.3	1189.3	1169.1
MIX-GARCH (NORM)	7	1191.7	1184.7	1169.0
RS-GARCH-Klaassen (NORM)	8	1192.3	1184.3	1166.3
RS-APARCH (NORM ; pow=2)	9	1192.8	1183.8	1163.6
RS-GARCH-Gray (NORM)	8	1188.4	1180.4	1162.5
RS (SGED)	9	1184.8	1175.8	1155.6
RS (NORM)	6	1174.7	1168.7	1155.2
SNIG	4	1164.2	1160.2	1151.2
GARCH (NORM)	4	1163.6	1159.6	1150.6
MIX (NORM)	5	1161.8	1156.8	1145.6
NORM	2	1138.7	1136.7	1132.2

is mainly to provide a possibility for the volatility to jump and that persistence in volatility may be better explained by GARCH-type dynamics than solely by regime persistence. We now take a look at the residual analysis for these models.

Table 4.2 contains the p-values of the statistical tests which were discussed in Section 4.2.2. Results from that table indicate that there is no model that performs best overall since the best models (with respect to the BIC) do not necessarily pass all heteroscedasticity and normality tests. Nevertheless, there are a few models, such as the MIX-GARCH (SNORM ; $\alpha = 0$) and the MIX-APARCH (SNORM ; pow=1), that have a good global fit and that do pass these tests at a 5% significance level. The fact that there is not a single model or a class of models that dominates other candidates in all aspects will have to be taken into consideration when evaluating different risks.

The log-likelihood values as well as their penalized versions (AIC and BIC), normality and heteroscedasticity tests only measure how a model globally fits the data. This can be very interesting if one wants to draw inferences on the dynamics of a financial asset or market. When issuing equity-linked insurance, the left tail of the returns' distribution is most important because of the asymmetry in the payoff. Moreover, given that these insurance products typically have long

TABLE 4.2. Residual analysis

Model	ARCH-LM	Ljung-Box	Jarque-Bera	Shapiro-Wilk
RS-EGARCH (SNORM)	0.469	0.431	0.046	0.022
EGARCH (SSTD)	0.046	0.035	0.980	0.445
MIX-APARCH-Haas (SNORM ; pow=1)	0.027	0.020	0.607	0.300
MIX-GARCH (SNORM ; $\alpha = 0$)	0.183	0.135	0.836	0.084
APARCH (SNIG)	0.083	0.074	0.909	0.295
MIX-APARCH (SNORM ; pow=1)	0.134	0.095	0.671	0.154
MIX-GARCH (NORM)	0.236	0.180	0.518	0.015
RS-GARCH-Klaassen (NORM)	0.224	0.164	0.355	0.015
RS-APARCH (NORM ; pow=2)	0.272	0.196	0.194	0.009
RS-GARCH-Gray (NORM)	0.247	0.178	0.142	0.020
RS (SGED)	0.062	0.053	0.566	0.541
RS (NORM)	0.028	0.014	0.186	0.072
SNIG	0.000	0.000	0.981	0.985
GARCH (NORM)	0.864	0.843	0.000	0.000
MIX (NORM)	0.000	0.000	0.437	0.422
NORM	0.000	0.000	0.000	0.000

maturities (three to 20 years), models need to be able to replicate stock market crashes that are long-lasting. This will be the focus of the next section.

4.3. Left tail analysis

Over the course of financial history, there have been many stock market crashes that took between six months to five years to fully recover. At the time of this writing (January 2012), the S&P 500 still has not fully recovered from its low of March 2009. Models need to have the capability to generate accumulation factors with a sufficiently fat left tail over long periods of time for an investment guarantee to mature in-the-money. In this section, we intend to measure this element by comparing the cost of the investment guarantee generated by our various models to the actual cost incurred during the financial crisis.

4.3.1. Assumptions

We assume that an insurer writes an investment guarantee on an asset or portfolio that tracks the S&P 500 total return index. The contract is issued T years (where $T = 3, 5, 7$ or 10) prior to the maturity date of February 28, 2009. We selected this maturity date because it corresponds to the lowest end-of-month index value during the financial crisis. The initial investment and the guarantee

at maturity are both 100$ and we assume that the policyholder will not die or lapse its contract until the end. Fees, expressed as a percentage of the fund value, are deducted monthly from the fund. They comprise a 0.5% annual management expense ratio (MER) and an additional fee reserved to fund the guarantee that depends on the investment horizon T. The total fees that were used for maturities of $T = 3, 5, 7$ and 10 are 5.0%, 3.5%, 2.5% and 2.0%, respectively [4].

In order to set up an appropriate reserve, the insurance company uses data available from February 1956 to February $(2009 - T)$ to estimate its model. It can then stochastically project the cost of the guarantee at maturity, calculate a risk measure and discount it to obtain the reserve amount needed at inception of the contract. This way of managing risk within investment guarantees is commonly called the actuarial approach. In our analysis, the discounting part is not necessary because we will be comparing the cost of the guarantee at maturity. The risk measures that we employed are the 99% VaR and the 95% CTE.

The interest in this out-of-sample analysis is to check whether the reserve forecasted by the models is large enough to meet the insurer's obligations in February 2009. It is important to note that our objective is not to point out the appropriateness of the reserve or the risk measure that should be used by actuaries. It is for actuaries and financial analysts to decide of the severity of the financial crisis and to argue whether it is a one in 20 event or a one in 200 event. Rather, we want to investigate if simple and complex models are capable of generating low returns over long-term periods.

4.3.2. Results

Based on the assumptions presented in the previous section, we calculated the 99% VaR and the 95% CTE of the guarantee cost based on 400,000 simulated paths for maturities of $T = 3, 5, 7$ and 10 for a total of 62 models [5]. Tables 4.3 and 4.4 show our results for the same subset of models that was previously analyzed in Section 4.2.3. Results for all models can be found in Tables 4.10 and 4.11 of Appendix A. The first line of each panel gives the out-of-sample provision that would have been necessary in February 2009 to break even given the observed

4. The fee reserved to fund the guarantee was established based on the approach of Hardy [96, p. 142]. The method consists in equating the arbitrage-free valuation of the guarantee income to that of the value of the underlying put option sold under the Black-Merton-Scholes framework. For that purpose, we used a risk-free interest rate of 3% and assumed a volatility of the underlying asset corresponding to 1.1 times the empirical in-sample volatility. The factor of 1.1 is arbitrary and reflects a small margin of conservatism with respect to pricing.

5. A small portion of the 78 models that were fitted in Section 4.2 was discarded in our simulation analyses since they did not provide a good fit to the data and they were not important benchmarks.

TABLE 4.3. Guarantee costs : 3 and 5-year contracts

Model	3-Year		5-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	47.5		40.6	
RS-GARCH-Gray (NORM)	40.3	46.7	41.1	48.8
RS (NORM)	36.8	43.3	37.4	45.4
RS-GARCH-Klaassen (NORM)	36.8	42.6	36.7	44.1
MIX-GARCH (SNORM ; $\alpha = 0$)	36.7	42.9	38.4	46.1
MIX-GARCH (NORM)	36.2	42.0	36.1	43.3
MIX-APARCH-Haas (SNORM ; pow=1)	35.9	42.8	35.6	44.2
RS-APARCH (NORM ; pow=2)	35.7	42.0	35.1	42.7
MIX-APARCH (SNORM ; pow=1)	35.3	42.3	34.8	43.6
APARCH (SNIG)	34.6	41.6	32.4	41.1
EGARCH (SSTD)	34.1	41.1	33.1	42.0
RS-EGARCH (SNORM)	34.0	41.2	36.8	46.0
SNIG	30.9	35.9	28.7	35.4
MIX (NORM)	30.9	35.8	29.0	35.7
NORM	30.0	34.7	27.8	34.2
RS (SGED)	27.8	33.6	31.3	38.5
GARCH (NORM)	27.1	32.2	28.1	35.2

market performance. For example, over the three-year period ending in February 2009, the S&P 500 total return index had a cumulative return of -38.8% so that the cost of the guarantee at maturity was 47.5% of the initial fund value (under assumptions of Section 4.3.1). Over the 10-year period that included the burst of the internet bubble, the observed cumulative return was -29.5% and the corresponding out-of-sample cost was 42.4%.

First, we focus our analysis on short-term investment horizons (i.e., three and five-year periods). For a contract of three years, all models fail to generate a 95% CTE or 99% VaR that is high enough to meet the requirements of the financial crisis. However, the gap between the risk measures and the out-of-sample result is not that large. For a contract of five years, most models fail at the 95% CTE level but not at the 99% VaR level. There is some similarity between results of three and five-year contracts. For instance, the RS (NORM) model is one of the most conservative, indicating an important potential to generate negative returns over three and five-year periods. This had to be expected since in many scenarios, the latent Markov chain stays in its high volatility–low return regime for extended periods of time. More recent and sophisticated models also

TABLE 4.4. Guarantee costs : 7 and 10-year contracts

Model	7-Year		10-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	36.4		42.4	
MIX-GARCH (SNORM ; $\alpha = 0$)	36.2	45.4	24.5	37.2
RS-EGARCH (SNORM)	36.1	47.3	20.3	35.1
RS-GARCH-Gray (NORM)	34.5	44.1	19.3	32.5
MIX-APARCH-Haas (SNORM ; pow=1)	34.2	44.7	20.8	34.5
MIX-APARCH (SNORM ; pow=1)	33.3	44.3	18.8	33.1
RS-GARCH-Klaassen (NORM)	32.5	41.5	19.0	31.9
MIX-GARCH (NORM)	32.2	41.0	18.9	31.3
RS-APARCH (NORM ; pow=2)	32.0	41.4	18.5	31.1
EGARCH (SSTD)	30.8	41.9	17.8	32.2
APARCH (SNIG)	30.6	41.5	18.6	33.3
RS (NORM)	27.6	37.5	10.2	20.8
GARCH (NORM)	23.9	33.0	11.4	22.4
RS (SGED)	20.9	29.6	5.9	12.0
MIX (NORM)	20.1	28.9	5.6	10.7
SNIG	19.8	28.5	5.3	10.3
NORM	18.9	27.3	4.9	9.3

do comparably well over these periods. Moreover, the range of the 95% CTE across models with a good fit varies between 34% and 40% for a three-year period and between 32% and 41% for a five-year period. This is relatively narrow when compared to periods of seven and 10 years.

We now switch our focus to longer-term horizons of seven and 10 years. For a seven-year period, there are some models in the RS-GARCH class that do generate a 95% CTE that is close to the out-of-sample value of 36.4%. However, for a 10-year period this is not the case ; most models generate a 95% CTE that is half or below half of the out-of-sample value of 42.4%. In other words, we would have needed to put aside twice as much money to meet the capital requirements of a 10-year contract issued in February 1999. During that period, the S&P 500 index experienced two important crashes : the internet bubble and the financial crisis of the late 2000s. Moreover, the range of possible reserves for models providing a good fit to the data is much larger than over shorter-term periods. For example, the 95% CTE of the RS-EGARCH (SNORM) for a 10-year period is double that of the RS (NORM) ; for periods of three and five years the relative difference was less than 10%. This highlights the growing uncertainty of future returns

and accumulation factors which implies that long-term investment risk is difficult to evaluate with accuracy. Finally, we note that recent econometric models do generally better than benchmark models in providing conservative figures but they still fail to generate very low returns over long-term periods.

It is worth pointing out that the risk measures generated by the RS (NORM) model over a 10-year period are much less conservative than those of more complex models that also provide an adequate fit to the data. A parameter that has a large influence on the capability of the RS (NORM) model to generate low returns over an extended period of time is p_{22}, which represents the probability of staying in the high volatility–low return regime. For the period going from February 1956 to February 1999, we estimated this parameter at 61% for the S&P 500 index [6]. To illustrate the importance of that parameter, we repeated our simulations by modifying it to 80% and leaving all other parameters unchanged. With that change, the RS (NORM) model generates a 95% CTE of 47.4%, which is almost a five-fold increase with respect to the original 95% CTE of 10.2%. It is now even sufficient to cover the out-of-sample value of 42.4%. We must stress that we are not suggesting to calibrate the value of p_{22} to better match the out-of-sample CTE. Rather, we want to point out the importance that parameter has on the left tail of the RS model, i.e., the calculation of the CTE is strongly influenced by its value. It is important to be aware of this element since Hardy [94] showed that p_{22} is the parameter which is estimated with the most uncertainty in the RS model.

The previous example illustrates that to evaluate long-term investment risk, it is more prudent to analyze the results given by various models instead of focusing on the output of a single model. We showed that many models having a good global fit can generate very different risk measures when we considered a period of 10 years. Hence, it is important to be aware that model risk is important in the context of investment guarantees and that it is difficult to accurately estimate the distribution of the guarantee cost over long periods of time.

4.4. Dynamic hedging

4.4.1. Background

The traditional actuarial approach to managing risks is to set up a reserve at the inception of the contract to meet future obligations with a high probability. This reserve is usually updated as time passes by and as new information becomes

6. This result is in line with the estimated parameters given in Hardy [94] who considered a similar estimation period.

available. However, no matter how well the reserve is managed, the insurance company always assumes the underlying risk of the contract. The reserve's role is to absorb the risk rather than to eliminate it.

Advances in financial theory due mainly to the work of Fischer Black, Robert Merton and Myron Scholes showed that it is theoretically possible to exactly replicate some financial risks (derivatives) with tradable assets under a certain set of assumptions. This is known as hedging. A growing number of insurance companies have now established hedging strategies with the objective of mitigating the financial risk on their equity-linked insurance products.

In the Black-Merton-Scholes (BMS) framework, one can replicate the payoff of an investment guarantee, which is essentially a put option, by trading in the underlying stock and the risk-free asset. For that replication to be perfect, rather stringent assumptions are required : the market model should follow a geometric Brownian motion (GBM), there should be no market frictions (no transaction costs and no constraints on trading) and trading in continuous time should be possible so that the replicating portfolio can be continuously rebalanced. Although market frictions are usually small for large investment banks and insurance companies, trading in continuous time is not possible, and, as shown in Section 4.2.3, the market model is far from being a GBM. Consequently, dynamically hedging the long-term investment guarantee using the basic hedging strategy of the BMS framework will not perfectly replicate the payoff of the guarantee. Thus, hedging under imperfect conditions entails a risk.

During the time of the contract, the company will incur gains or losses each time it will rebalance its hedging portfolio because the required investment in the updated portfolio will generally be different from the current portfolio value. These gains and losses are a stream of cash inflows and outflows which are called hedging errors. Hedging errors come mainly from two sources : discretization error and model error. Discretization error results from the fact that rebalancing cannot be done in continuous time. Model error stems from the fact that the true market model is different from a GBM. It is not possible to know at contract inception whether hedging errors will lead to a gain or to a profit. Hence, the insurance company must set up a reserve to protect itself from that uncertainty.

4.4.2. Objectives and assumptions

In this section, we aim to investigate to what extent model risk matters when a long-term investment guarantee is dynamically hedged based on the BMS framework. More precisely, we will restrict ourselves to the standard Black-Scholes (BS) delta hedge [see 96]. This analysis will enable us to evaluate the robustness of

that hedging strategy with respect to its underlying assumption that the market model is a GBM.

For each of the models previously estimated [7], we generated 400,000 market scenarios under the real-world probability measure based on a monthly frequency and investment horizons of three and 10 years. We then applied the BS delta hedge dynamically through time (at a monthly frequency) on each of these simulated paths. This way, we were able to calculate the present value of hedging errors (PVHE) for each simulation. We discounted the hedging errors with a 3% interest rate which corresponds roughly to the average 1-Month Treasury Constant Maturity rate for periods of three and 10 years prior to the financial crisis. We remain again in an out-of-sample context and assume that the investment guarantee matures in February 2009.

First, we may look at the 95% CTE of the PVHE generated by each model and analyze whether model specific dynamics influence that risk measure. If the 95% CTE is highly variable across models, then this implies that the BS delta hedge is not robust with respect to the GBM assumption, i.e., there is high model risk. Second, we can compare the 95% CTE generated by our models to the out-of-sample values observed during the financial crisis to check whether models underestimate hedging risk or not.

We will now state the assumptions used in the BS delta hedge. The risk-free rate was set to 3% for the whole length of the contract. The constant volatility parameter was estimated by the in-sample volatility at the inception of the contract. Whenever transactions costs were included, they were set to represent 0.2% of the change in the market value of the stock position used for hedging.

4.4.3. Results

Tables 4.5 and 4.6 show the standard deviation (abbreviated StDev) and the 95% CTE of the PVHE for an investment guarantee with a maturity of three and 10 years, respectively, for the same subset of models that were previously analyzed in Sections 4.2.3 and 4.3.2. Results for all models can be found in Tables 4.12 and 4.13 of Appendix A. Each panel also includes the mean volatility of the simulated paths. The mean volatility of the NORM model corresponds to the volatility input used in the BS delta hedge (i.e., the in-sample empirical volatility). Moreover, the *incl.* and *excl.* column titles denote values with and without transaction costs, respectively. We first focus on a maturity of three years.

7. A small portion of the 78 models that were fitted in Section 4.2 was discarded in our simulation analyses since they did not provide a good fit to the data and they were not important benchmarks. Hence, 62 models are considered here.

TABLE 4.5. PVHE : 3-year contract

Model	Mean Volatility of Returns	PV of Hedging Errors			
		StDev		95% CTE	
		Excl.	Incl.	Excl.	Incl.
MIX-APARCH (SNORM ; pow=1)	0.137	2.18	2.24	5.94	6.45
MIX-GARCH (SNORM ; $\alpha = 0$)	0.138	2.33	2.40	5.83	6.38
RS-APARCH (NORM ; pow=2)	0.138	2.31	2.37	5.75	6.30
EGARCH (SSTD)	0.134	2.16	2.22	5.62	6.15
MIX-GARCH (NORM)	0.140	2.24	2.30	5.61	6.16
MIX-APARCH-Haas (SNORM ; pow=1)	0.140	2.09	2.15	5.53	6.07
RS-GARCH-Klaassen (NORM)	0.140	2.17	2.23	5.38	5.92
APARCH (SNIG)	0.135	2.06	2.12	5.24	5.77
RS-EGARCH (SNORM)	0.131	2.08	2.14	5.07	5.61
RS-GARCH-Gray (NORM)	0.140	1.97	2.01	4.96	5.47
RS (NORM)	0.142	1.85	1.89	4.94	5.46
GARCH (NORM)	0.130	1.98	2.02	4.43	4.96
SNIG	0.144	1.57	1.60	4.12	4.61
MIX (NORM)	0.145	1.56	1.59	4.03	4.52
RS (SGED)	0.126	2.04	2.09	3.74	4.26
NORM	0.145	1.29	1.30	3.07	3.57

A first observation that we can make from Table 4.5 is that for most models the mean volatility of the simulated paths is generally below that of the in-sample volatility of 14.5%. This occurs because models are generally in a low volatility state at the start of the projection in February 2006. In such a situation, when the mean volatility of the simulated paths is lower than the volatility used for hedging, the distribution of hedging errors is shifted towards the left [see 96, p. 152]. Another remark that can be made is that the NORM model generates the smallest standard deviation and 95% CTE of the PVHE. This is of course expected as the BS delta hedge should perform best under its own assumption of a GBM. When this assumption is not valid, we see that model risk can more than double the 95% CTE of the PVHE. That risk measure varies between 3.0 and 7.5 (assuming no transaction costs) when all models are considered.

Moreover, the 95% CTE of the PVHE can be decomposed into two parts : one part accounts for the discretization error and the remaining part for model error. The discretization error corresponds to roughly 3.0 since it is the value associated with the NORM model. The excess over that figure represents model error. Hence, we note here that model error can account for more than half of the total hedging

error so that it must be taken into account. To better illustrate the influence of model risk, we must look at its impact in relative terms. The PVHE is not the only cost that the insurer has to bear in its hedging program ; rather, it is the uncertain part of its total cost. Its total cost also comprises the initial purchase of the replicating portfolio whose price is known *a priori* and corresponds to that of the underlying put option in the equity-linked product. Under the assumptions of Section 4.4.2, the initial hedge cost is 12.0 in this example. Therefore, under many sophisticated models the hedging errors can represent more than half of that initial cost which is substantial.

Lastly, we may look at the impact of transaction costs on the PVHE. At first, it may be reasonable to suppose that the proportion of transaction costs contained in the 95% CTE of the PVHE is roughly constant across models, however this is not what we observe. For all models, transaction costs in the 95% CTE of the PVHE are approximately constant in absolute terms (0.5) but not in relative terms. Moreover, the standard deviation of the PVHE does not increase by much when transaction costs are included. This implies that the main cause of the high 95% CTE observed in some models is not more variability in the movement of the stock position used in the hedge (since this would entail higher transaction costs) but rather the direction of that movement which is more one-sided (hedging errors tend to cancel each other out much less frequently than under the NORM model). We now repeat our example with a maturity of 10 years.

First, we must stress that it is difficult to directly compare the values in Tables 4.5 and 4.6 ; the models are estimated based on a different sample, they do not necessarily start from the same volatility state and the present value factor gives less weight to the hedging errors that occur later. However, conclusions that can be reached for a 10-year maturity go in the same direction as those that were made previously for a three-year maturity. For instance, the range of the 95% CTE of the PVHE across models is large (it is now even much wider than for a three-year maturity) and this entails that model risk is important. While that risk measure is under a value of 8.0 for most models, it can reach a figure as high as 15.1 (see Table 4.13). The discretization error given by the NORM model is 1.7 in this example so that model error now represents the vast majority of the total hedging error. For example, the model error component in the 95% CTE of the PVHE of the EGARCH (SSTD) model corresponds to 5.8 which is almost 80% of the total error. Therefore, model risk has an even larger influence for a maturity of 10 years. Moreover, if we put that same model error component in perspective to the initial hedge cost (which is 10.5), we deduce that model error alone can amount to more than half of that initial cost. Finally, concerning the impact of

TABLE 4.6. PVHE : 10-year contract

Model	Mean Volatility of Returns	PV of Hedging Errors			
		StDev		95% CTE	
		Excl.	Incl.	Excl.	Incl.
EGARCH (SSTD)	0.148	2.17	2.29	7.54	8.22
APARCH (SNIG)	0.148	2.16	2.29	7.54	8.24
RS-EGARCH (SNORM)	0.149	2.07	2.21	7.14	7.87
MIX-APARCH (SNORM ; pow=1)	0.148	2.04	2.17	7.06	7.78
MIX-GARCH (SNORM ; $\alpha = 0$)	0.151	1.97	2.11	6.63	7.38
MIX-APARCH-Haas (SNORM ; pow=1)	0.151	1.89	2.02	6.55	7.27
GARCH (NORM)	0.154	1.79	1.87	6.03	6.68
RS-APARCH (NORM ; pow=2)	0.147	1.69	1.80	5.45	6.15
MIX-GARCH (NORM)	0.147	1.62	1.73	5.19	5.90
RS-GARCH-Klaassen (NORM)	0.147	1.61	1.72	5.13	5.84
RS-GARCH-Gray (NORM)	0.143	1.40	1.49	4.28	4.93
RS (NORM)	0.144	1.18	1.26	3.65	4.24
MIX (NORM)	0.144	0.94	0.99	2.64	3.18
RS (SGED)	0.144	0.95	0.99	2.63	3.15
SNIG	0.143	0.91	0.97	2.59	3.14
NORM	0.144	0.70	0.72	1.72	2.23

transaction costs, we also observe that increases in the 95% CTE of the PVHE from one model to the other are not accompanied by proportional increases in transaction costs.

The results and discussions presented in this section illustrate that different volatility dynamics can lead to very different reserves for the PVHE. Moreover, we showed that model risk is very important to consider, especially for a maturity of 10 years. For that maturity, model risk represents the majority of the uncertainty related to the PVHE and can significantly increase hedge costs. Since the 95% CTE of the PVHE was very unstable across models (it was in the range of 1.7 to 15.1) for a maturity of 10 years, we must seriously question the robustness of the BS delta hedge with respect to the GBM assumption since its effectiveness was highly dependent on the underlying market model.

4.4.4. Bootstrap

Given the growing importance of dynamic hedging in the insurance industry, we further investigate the behavior of the PVHE using bootstrap. More precisely, we now wish to examine whether the PVHE reserves forecasted by our models are

consistent with those that can be obtained with bootstrapped data. Bootstrap is a technique that consists in creating samples by drawing with replacement from the original sample (this is commonly called resampling). Because asset returns are serially correlated in their squares, it is preferable to resample blocks of data to take that dependence into account. For this exercise, we employed all available data (i.e., February 1956 to December 2010) and generated 400,000 samples with blocks of different sizes. Table 4.7 shows the 95% CTE of the PVHE using bootstrapped data (with and without transaction costs, denoted respectively by *incl.* and *excl.* in the column titles) for horizons of three and 10 years.

TABLE 4.7. PVHE based on bootstrapped data

Bootstrap Blocks	3-Year 95% CTE		10-Year 95% CTE	
	Excl.	Incl.	Excl.	Incl.
Blocks of 12 (Annual)	5.20	5.69	4.53	5.18
Blocks of 6 (Semi-Annual)	5.29	5.77	4.25	4.91
Blocks of 4	5.17	5.65	4.04	4.71
Blocks of 3 (Quarter)	5.08	5.56	3.90	4.55
Independent (Monthly)	4.80	5.27	3.38	4.01

A first observation that can be made is that the 95% CTE of the PVHE generally increases with the block size which demonstrates that an independent bootstrap would underestimate the distribution of hedging errors. Moreover, for a three-year maturity, we obtain a reserve of 5.3 (without transaction costs) when the data is bootstrapped with semi-annual blocks. This is consistent with results obtained in Table 4.5 for models that provide a good fit to the data. However, for a 10-year maturity this statement does not hold because the reserve based on bootstrapped data is in the low range of those generated by most models. This may suggest that some models may overestimate the risk of hedging, but jumping to such a conclusion may not be very cautious since models that do generate high PVHE also provide a good fit to the data. We also need to stress that the data is based upon one sample path of the true market model so that bootstrap results may not give a complete picture of future possible dynamics.

4.4.5. Out-of-sample

In this final example, we investigate the out-of-sample PVHE using observed data from the S&P 500 total return index. The objective here is to determine whether there were moments in history where the BS delta hedge was not effective and generated high PVHE for contracts of three and 10 years. For example,

suppose that a contract is issued in March 1998 and matures three years later in March 2001. Applying the BS delta hedge to the S&P data for that period yields a PVHE of 4.4 (assuming no transaction costs) under the same assumptions used in the previous exercises (see Section 4.4.2). Figure 4.1 presents the out-of-sample PVHE (assuming no transaction costs) for three-year contracts maturing from January 1985 to December 2010.

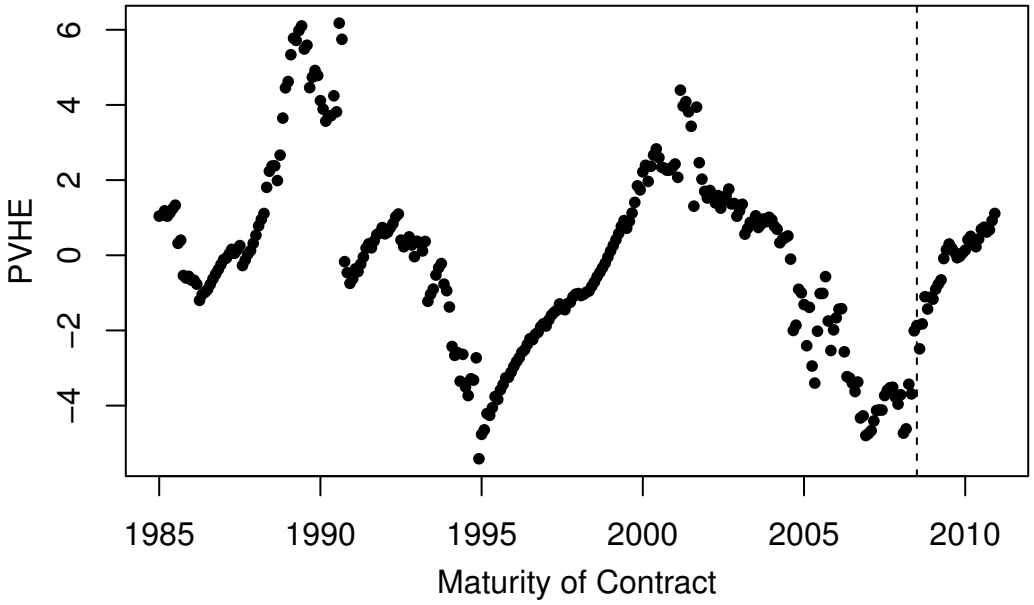

FIGURE 4.1. Empirical PVHE through time for a 3-year contract

The vertical dotted line in Figure 4.1 is simply a marker for June 2008. Hence, the points located to the right of that line represent the PVHE for contracts maturing when the market crashed significantly during the financial crisis. A first observation that can be made is that the PVHE for contracts maturing in the late 1980s reached over 6.0. For contracts maturing right after the internet bubble, the PVHE could go as high as 4.4. Therefore, there are periods in the data where the observed PVHE reach levels corresponding to the 95% CTE of the PVHE obtained under many models with a good fit to the data (see Table 4.5). Nevertheless, we note that the BS delta hedge did not do so badly for products maturing during the recent financial crisis as the PVHE were always under 2.0.

We now repeat that same exercise for 10-year contracts maturing from January 1985 to December 2010. Figure 4.2 shows our results. Barring one exception [8] the PVHE in Figure 4.2 are always under 3.0. This value is much lower than the 95% CTE obtained under many models and bootstrapped data. It is important

8. The PVHE for a 10-year contract maturing in October 2008 is 5.9.

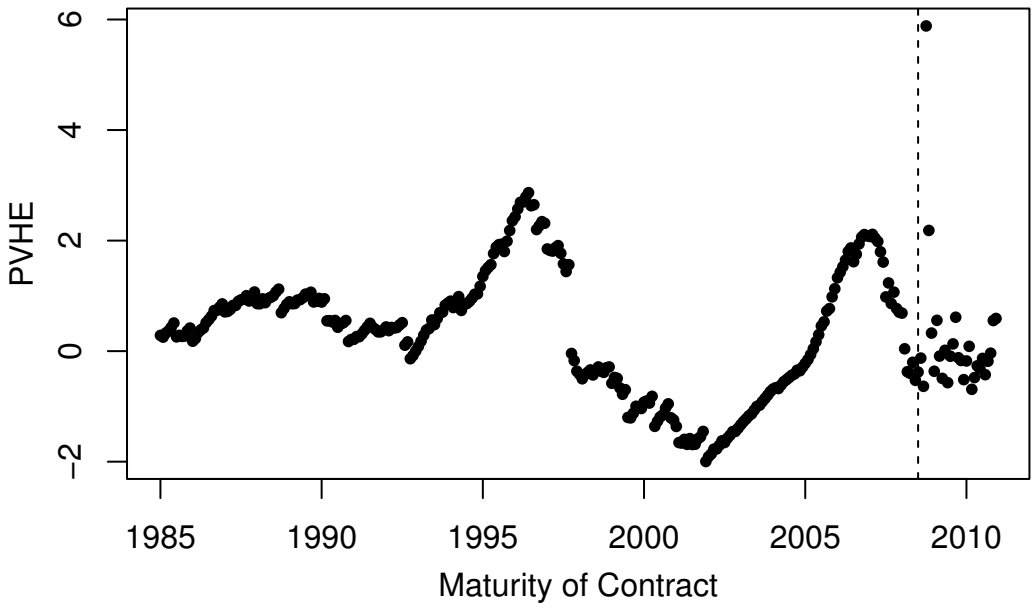

FIGURE 4.2. Empirical PVHE through time for a 10-year contract

to stress that there are not too many disjoint 10-year periods in the data. The fact that we did not observe many 10-year periods that generate high PVHE does not imply that this will be the case in the future. The risk of having much higher hedging errors exists as shown in Section 4.4.3.

4.5. CONCLUSION

In this paper, we analyzed the risk underlying investment guarantees using a very large set of financial econometric models. We found that despite the excellent fit of many models, too few of them are capable of generating low returns over long periods of time (say five to 10 years), i.e., the type of cumulative returns observed during the financial crisis of the late 2000s. This is a crucial element for insurance companies that issue equity-linked insurance. We generated scenarios under each of these models to check the robustness of the Black-Scholes delta hedging strategy. We found that hedging losses can be significant implying that large reserves for hedging errors are required. We also showed that results can be very variable across models and this highlights that model risk is important to take into consideration when managing the risk underlying investment guarantees.

How can we make the Black-Scholes delta hedging strategy more robust, i.e., less sensitive to model risk ? One may rebalance weekly, daily or even several times a day. However, it is well documented in the finance literature that daily or high frequency data has much fatter tails than monthly data [see 31, 160],

meaning even more deviations from the GBM. Thus, it is not clear how effective rebalancing more frequently can really be. Moreover, one may use other Greeks in the replicating portfolio, such as gamma, vega and rho, which are respectively the sensitivity of the derivative's price with respect to large price movements of the underlying asset, its volatility and the risk-free rate. Using these Greeks may help reduce hedging errors, but their effectiveness will depend on the true market model. Finally, one may turn to the financial engineering literature and use the true replicating portfolio of some of these 78 models that fit the data well. In addition to being more challenging to implement due to the incompleteness of the underlying market, their reliance on a risk premium parameter or process makes it difficult to predict as to how robust this approach will be. We conclude by quoting Rantala [151] who mentions that "in the face of model risk, rather than to base decisions on a single selected 'best' model, the modeller can base his inference on an entire set of models by using model averaging."

Acknowledgments

The authors acknowledge the financial support provided by the Natural Sciences and Engineering Research Council of Canada and the Society of Actuaries.

Appendix A. Results for all models

Table 4.8: Fit summary

Model	Params	Log-Lik	AIC	BIC
RS-EGARCH (SNORM)	10	1210.2	1200.2	1177.7
EGARCH (SSTD)	7	1198.1	1191.1	1175.4
EGARCH (SNIG)	7	1196.8	1189.8	1174.1
MIX-APARCH-Haas (SNORM ; pow=1)	8	1200.0	1192.0	1174.1
MIX-GARCH (NORM ; $\alpha = 0$)	6	1192.2	1186.2	1172.7
RS-APARCH-Haas (SNORM ; pow=1)	9	1201.7	1192.7	1172.4
MIX-GARCH (SNORM ; $\alpha = 0$)	7	1194.9	1187.9	1172.2
EGARCH (SGED)	7	1194.9	1187.9	1172.2
MIX-GJR GARCH-Scale (SNORM)	9	1201.3	1192.3	1172.1
GJR GARCH (SSTD)	7	1194.8	1187.8	1172.1
MIX-APARCH-Haas (SNORM ; pow=2)	8	1197.3	1189.3	1171.3
APARCH (SSTD)	8	1197.2	1189.2	1171.3
MIX-EGARCH (SNORM)	9	1200.2	1191.2	1171.0
GJR GARCH (SNIG)	7	1193.5	1186.5	1170.8
RS-GJR GARCH-Scale (SNORM)	10	1203.1	1193.1	1170.7
GARCH (SSTD)	6	1189.9	1183.9	1170.4
APARCH (SNIG)	8	1195.9	1187.9	1169.9
GARCH (SNIG)	6	1189.0	1183.0	1169.5
RS-APARCH-Haas (SNORM ; pow=2)	9	1198.5	1189.5	1169.3
MIX-APARCH (SNORM ; pow=1)	9	1198.3	1189.3	1169.1
MIX-GARCH (NORM)	7	1191.7	1184.7	1169.0
MIX-APARCH (SNORM ; pow=2)	9	1197.9	1188.9	1168.7
GJR GARCH (SGED)	7	1191.4	1184.4	1168.6
EGARCH (STD)	6	1187.7	1181.7	1168.2
APARCH (SGED)	8	1194.0	1186.0	1168.0
GARCH (SGED)	6	1187.4	1181.4	1167.9
MIX-APARCH (NORM ; pow=1)	8	1193.8	1185.8	1167.8
MIX-APARCH-Haas (NORM ; pow=1)	7	1190.4	1183.4	1167.7
MIX-EGARCH (NORM)	8	1193.6	1185.6	1167.6
MIX-APARCH (NORM ; pow=2)	8	1192.7	1184.7	1166.8
RS-GARCH-Klaassen (NORM)	8	1192.3	1184.3	1166.3
RS-APARCH (SNORM ; pow=1)	10	1198.5	1188.5	1166.0
RS-APARCH (SNORM ; pow=2)	10	1198.0	1188.0	1165.5
MIX-GJR GARCH-Scale (NORM)	8	1191.4	1183.4	1165.4
GJR GARCH (STD)	6	1184.6	1178.6	1165.1
RS-APARCH (NORM ; pow=1)	9	1193.9	1184.9	1164.7
RS-EGARCH (NORM)	9	1193.8	1184.8	1164.6
RS-APARCH-Haas (NORM ; pow=1)	8	1190.4	1182.4	1164.4
APARCH (STD)	7	1187.0	1180.0	1164.3
MIX-APARCH-Haas (NORM ; pow=2)	7	1186.8	1179.8	1164.1
RS-APARCH (NORM ; pow=2)	9	1192.8	1183.8	1163.6
MIX-GARCH (NORM ; full)	9	1192.3	1183.3	1163.1
RS-GJR GARCH-Scale (NORM)	9	1191.8	1182.8	1162.6
GARCH (STD)	5	1178.8	1173.8	1162.6
RS-GARCH-Gray (NORM)	8	1188.4	1180.4	1162.5
EGARCH (GED)	6	1181.7	1175.7	1162.2
RS-APARCH-Haas (NORM ; pow=2)	8	1186.8	1178.8	1160.9
RS (SNORM)	7	1183.2	1176.2	1160.5

Continued on next page...

Table 4.8: Fit summary

Model	Params	Log-Lik	AIC	BIC
RS-GARCH-Klaassen (NORM; full)	10	1192.4	1182.4	1160.0
GJR GARCH (GED)	6	1178.1	1172.1	1158.7
APARCH (GED)	7	1181.2	1174.2	1158.4
GARCH (GED)	5	1174.6	1169.6	1158.4
RS (SSTD)	9	1186.7	1177.7	1157.5
RS-GARCH-Gray (NORM; full)	10	1189.6	1179.6	1157.1
RS (SNIG)	9	1185.6	1176.6	1156.4
EGARCH (NORM)	5	1171.9	1166.9	1155.7
RS-GARCH-Haas (NORM)	7	1178.3	1171.3	1155.6
RS (SGED)	9	1184.8	1175.8	1155.6
RS (NORM)	6	1174.7	1168.7	1155.2
APARCH (NORM)	6	1171.9	1165.9	1152.4
RS (STD)	8	1177.7	1169.7	1151.7
SSTD	4	1164.3	1160.3	1151.3
SNIG	4	1164.2	1160.2	1151.2
GARCH (NORM)	4	1163.6	1159.6	1150.6
GJR GARCH (NORM)	5	1166.4	1161.4	1150.2
RS-GARCH-Haas (NORM; full)	9	1178.9	1169.9	1149.7
RS (GED)	8	1175.2	1167.2	1149.2
SGED	4	1162.1	1158.1	1149.2
STD	3	1157.8	1154.8	1148.1
RS3 (NORM)	12	1185.4	1173.4	1146.4
MIX (NORM)	5	1161.8	1156.8	1145.6
GED	3	1154.7	1151.7	1145.0
MIX (STD)	7	1164.5	1157.5	1141.8
MIX (SGED)	8	1167.0	1159.0	1141.1
MIX (GED)	7	1163.5	1156.5	1140.8
MIX (SNIG)	8	1166.6	1158.6	1140.7
MIX (SSTD)	8	1164.6	1156.6	1138.6
NORM	2	1138.7	1136.7	1132.2

Table 4.9: Residual analysis

Model	ARCH-LM	Ljung-Box	Jarque-Bera	Shapiro-Wilk
RS-EGARCH (SNORM)	0.469	0.431	0.046	0.022
EGARCH (SSTD)	0.046	0.035	0.980	0.445
EGARCH (SNIG)	0.060	0.046	0.906	0.236
MIX-APARCH-Haas (SNORM; pow=1)	0.027	0.020	0.607	0.300
MIX-GARCH (NORM; $\alpha = 0$)	0.335	0.287	0.303	0.011
RS-APARCH-Haas (SNORM; pow=1)	0.035	0.027	0.657	0.310
MIX-GARCH (SNORM; $\alpha = 0$)	0.183	0.135	0.836	0.084
EGARCH (SGED)	0.117	0.093	0.124	0.036
MIX-GJR GARCH-Scale (SNORM)	0.042	0.030	0.868	0.232
GJR GARCH (SSTD)	0.064	0.043	0.984	0.240
MIX-APARCH-Haas (SNORM; pow=2)	0.022	0.013	0.574	0.379
APARCH (SSTD)	0.068	0.060	0.981	0.529
MIX-EGARCH (SNORM)	0.052	0.041	0.722	0.398
GJR GARCH (SNIG)	0.090	0.062	0.917	0.119
RS-GJR GARCH-Scale (SNORM)	0.052	0.039	0.875	0.271
GARCH (SSTD)	0.225	0.165	0.983	0.550

Continued on next page...

Table 4.9: Residual analysis

Model	ARCH-LM	Ljung-Box	Jarque-Bera	Shapiro-Wilk
APARCH (SNIG)	0.083	0.074	0.909	0.295
GARCH (SNIG)	0.258	0.193	0.994	0.376
RS-APARCH-Haas (SNORM ; pow=2)	0.021	0.013	0.612	0.397
MIX-APARCH (SNORM ; pow=1)	0.134	0.095	0.671	0.154
MIX-GARCH (NORM)	0.236	0.180	0.518	0.015
MIX-APARCH (SNORM ; pow=2)	0.133	0.077	0.623	0.117
GJR GARCH (SGED)	0.169	0.124	0.128	0.022
EGARCH (STD)	0.356	0.342	0.000	0.000
APARCH (SGED)	0.157	0.141	0.133	0.053
GARCH (SGED)	0.321	0.250	0.384	0.166
MIX-APARCH (NORM ; pow=1)	0.279	0.220	0.150	0.006
MIX-APARCH-Haas (NORM ; pow=1)	0.253	0.246	0.002	0.000
MIX-EGARCH (NORM)	0.134	0.110	0.139	0.010
MIX-APARCH (NORM ; pow=2)	0.289	0.211	0.194	0.008
RS-GARCH-Klaassen (NORM)	0.224	0.164	0.355	0.015
RS-APARCH (SNORM ; pow=1)	0.144	0.102	0.733	0.165
RS-APARCH (SNORM ; pow=2)	0.137	0.080	0.644	0.120
MIX-GJR GARCH-Scale (NORM)	0.207	0.186	0.050	0.006
GJR GARCH (STD)	0.378	0.342	0.001	0.000
RS-APARCH (NORM ; pow=1)	0.308	0.246	0.156	0.006
RS-EGARCH (NORM)	0.155	0.125	0.155	0.011
RS-APARCH-Haas (NORM ; pow=1)	0.256	0.248	0.002	0.000
APARCH (STD)	0.431	0.436	0.001	0.000
MIX-APARCH-Haas (NORM ; pow=2)	0.277	0.247	0.002	0.000
RS-APARCH (NORM ; pow=2)	0.272	0.196	0.194	0.009
MIX-GARCH (NORM ; full)	0.315	0.260	0.340	0.011
RS-GJR GARCH-Scale (NORM)	0.142	0.122	0.062	0.008
GARCH (STD)	0.563	0.541	0.001	0.000
RS-GARCH-Gray (NORM)	0.247	0.178	0.142	0.020
EGARCH (GED)	0.672	0.647	0.000	0.000
RS-APARCH-Haas (NORM ; pow=2)	0.280	0.249	0.002	0.000
RS (SNORM)	0.083	0.073	0.014	0.090
RS-GARCH-Klaassen (NORM ; full)	0.277	0.218	0.276	0.011
GJR GARCH (GED)	0.712	0.673	0.000	0.000
APARCH (GED)	0.741	0.739	0.000	0.000
GARCH (GED)	0.729	0.705	0.000	0.000
RS (SSTD)	0.053	0.046	0.991	0.854
RS-GARCH-Gray (NORM ; full)	0.539	0.499	0.029	0.003
RS (SNIG)	0.040	0.030	0.971	0.595
EGARCH (NORM)	0.886	0.868	0.000	0.000
RS-GARCH-Haas (NORM)	0.476	0.460	0.001	0.000
RS (SGED)	0.062	0.053	0.566	0.541
RS (NORM)	0.028	0.014	0.186	0.072
APARCH (NORM)	0.923	0.919	0.000	0.000
RS (STD)	0.032	0.020	0.001	0.000
SSTD	0.000	0.000	0.992	0.992
SNIG	0.000	0.000	0.981	0.985
GARCH (NORM)	0.864	0.843	0.000	0.000
GJR GARCH (NORM)	0.883	0.862	0.000	0.000
RS-GARCH-Haas (NORM ; full)	0.508	0.488	0.000	0.000

Continued on next page...

TABLE 4.9: Residual analysis

Model	ARCH-LM	Ljung-Box	Jarque-Bera	Shapiro-Wilk
RS (GED)	0.034	0.017	0.220	0.069
SGED	0.000	0.000	0.390	0.584
STD	0.000	0.000	0.008	0.006
RS3 (NORM)	0.002	0.001	0.291	0.184
MIX (NORM)	0.000	0.000	0.437	0.422
GED	0.000	0.000	0.000	0.001
MIX (STD)	0.000	0.000	0.985	0.992
MIX (SGED)	0.000	0.000	0.000	0.000
MIX (GED)	0.000	0.000	0.879	0.850
MIX (SNIG)	0.000	0.000	0.990	0.995
MIX (SSTD)	0.000	0.000	0.989	0.990
NORM	0.000	0.000	0.000	0.000

TABLE 4.10: Guarantee costs : 3 and 5-year contracts

Model	3-Year		5-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	47.5		40.6	
RS-GARCH-Gray (NORM)	40.3	46.7	41.1	48.8
MIX-APARCH (SNORM ; pow=2)	38.6	45.4	39.8	48.1
RS-APARCH-Haas (SNORM ; pow=2)	38.3	45.2	38.9	47.3
RS-APARCH (SNORM ; pow=2)	38.0	44.8	39.1	47.6
MIX-APARCH-Haas (SNORM ; pow=2)	38.0	45.0	38.9	47.7
MIX (SGED)	37.2	43.1	37.3	44.5
RS (NORM)	36.8	43.3	37.4	45.4
RS-GARCH-Klaassen (NORM)	36.8	42.6	36.7	44.1
MIX-GARCH (SNORM ; $\alpha = 0$)	36.7	42.9	38.4	46.1
GJR GARCH (SNIG)	36.6	44.3	38.0	47.7
GJR GARCH (SSTD)	36.5	44.2	37.3	47.1
GJR GARCH (SGED)	36.4	44.0	38.9	48.4
MIX-GARCH (NORM)	36.2	42.0	36.1	43.3
MIX-APARCH (NORM ; pow=2)	36.2	42.3	36.0	43.7
MIX-GARCH (NORM ; full)	35.9	41.9	36.7	44.4
MIX-APARCH-Haas (SNORM ; pow=1)	35.9	42.8	35.6	44.2
RS-APARCH (NORM ; pow=2)	35.7	42.0	35.1	42.7
RS-APARCH-Haas (SNORM ; pow=1)	35.7	42.5	35.7	44.3
MIX-GARCH (NORM ; $\alpha = 0$)	35.5	41.5	36.3	43.9
MIX-EGARCH (SNORM)	35.5	42.4	35.0	43.6
RS-GARCH-Gray (NORM ; full)	35.4	42.2	40.4	48.2
MIX-APARCH (SNORM ; pow=1)	35.3	42.3	34.8	43.6
RS-APARCH (SNORM ; pow=1)	35.2	42.0	35.1	43.9
RS-GARCH-Klaassen (NORM ; full)	34.9	40.9	35.7	43.4
APARCH (SNIG)	34.6	41.6	32.4	41.1
RS-EGARCH (NORM)	34.6	41.1	32.6	40.7
EGARCH (SNIG)	34.5	41.5	33.9	42.7
EGARCH (SGED)	34.5	41.4	35.1	43.9
MIX-EGARCH (NORM)	34.4	40.9	32.4	40.8
MIX-GJR GARCH-Scale (SNORM)	34.4	41.2	34.4	43.2
APARCH (SGED)	34.4	41.4	34.6	43.1
RS-APARCH (NORM ; pow=1)	34.3	40.9	32.6	41.1

Continued on next page...

Table 4.10: Guarantee costs : 3 and 5-year contracts

Model	3-Year		5-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	47.5		40.6	
MIX-APARCH (NORM ; pow=1)	34.2	40.8	32.5	41.0
APARCH (SSTD)	34.2	41.3	32.1	40.7
EGARCH (SSTD)	34.1	41.1	33.1	42.0
RS-EGARCH (SNORM)	34.0	41.2	36.8	46.0
MIX-GJR GARCH-Scale (NORM)	32.6	39.0	30.7	38.8
APARCH (NORM)	32.4	38.6	31.9	39.1
EGARCH (NORM)	31.4	37.3	30.8	38.5
RS-APARCH-Haas (NORM ; pow=1)	31.1	37.7	27.4	36.0
SSTD	30.9	36.0	28.7	35.3
MIX-APARCH-Haas (NORM ; pow=1)	30.9	37.5	27.3	36.0
SNIG	30.9	35.9	28.7	35.4
MIX (NORM)	30.9	35.8	29.0	35.7
SGED	30.8	35.8	28.7	35.3
GJR GARCH (NORM)	30.7	36.1	29.6	36.9
MIX-APARCH-Haas (NORM ; pow=2)	30.5	37.2	26.9	35.5
RS-APARCH-Haas (NORM ; pow=2)	30.5	37.2	27.2	35.9
MIX (SSTD)	30.4	35.4	28.2	35.0
NORM	30.0	34.7	27.8	34.2
GARCH (SSTD)	29.5	36.0	30.8	39.7
GARCH (SNIG)	29.4	35.9	31.1	39.7
GARCH (SGED)	29.1	35.3	31.1	39.8
RS (SGED)	27.8	33.6	31.3	38.5
RS (SNORM)	27.3	32.9	29.8	37.0
GARCH (NORM)	27.1	32.2	28.1	35.2
RS (SSTD)	26.9	33.0	31.3	38.6
RS3 (NORM)	26.7	32.5	32.5	39.9
STD	26.0	31.1	20.5	27.7
RS-GARCH-Haas (NORM)	25.9	31.8	21.6	29.5
GED	25.6	30.7	19.5	26.7
RS-GARCH-Haas (NORM ; full)	24.2	29.8	29.3	37.8

Table 4.11: Guarantee costs : 7 and 10-year contracts

Model	7-Year		10-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	36.4		42.4	
RS-GARCH-Gray (NORM ; full)	41.0	52.3	30.9	45.9
GJR GARCH (SGED)	40.3	52.7	34.6	51.4
GJR GARCH (SNIG)	38.6	51.0	30.6	47.6
MIX-APARCH (SNORM ; pow=2)	37.9	48.0	27.7	41.2
RS-APARCH (SNORM ; pow=2)	37.2	47.2	26.5	39.8
GJR GARCH (SSTD)	36.5	48.7	27.8	44.5
RS-APARCH-Haas (SNORM ; pow=2)	36.3	46.5	25.3	38.3
MIX-GARCH (SNORM ; $\alpha = 0$)	36.2	45.4	24.5	37.2
MIX-APARCH-Haas (SNORM ; pow=2)	36.2	46.3	25.0	38.1
RS-EGARCH (SNORM)	36.1	47.3	20.3	35.1
RS-GARCH-Klaassen (NORM ; full)	35.3	45.1	22.3	36.2
MIX-GJR GARCH-Scale (SNORM)	35.2	46.7	22.1	37.6

Continued on next page. . .

TABLE 4.11: Guarantee costs : 7 and 10-year contracts

Model	7-Year		10-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	36.4		42.4	
RS-GARCH-Gray (NORM)	34.5	44.1	19.3	32.5
APARCH (SGED)	34.3	45.2	21.4	37.0
MIX-APARCH-Haas (SNORM ; pow=1)	34.2	44.7	20.8	34.5
MIX-GARCH (NORM ; full)	34.1	43.6	21.2	34.5
RS-APARCH (SNORM ; pow=1)	33.8	44.6	19.4	33.9
MIX-GARCH (NORM ; $\alpha = 0$)	33.7	43.1	20.5	33.4
EGARCH (SGED)	33.5	44.2	21.9	36.5
MIX-EGARCH (SNORM)	33.5	44.0	18.8	33.2
RS-APARCH-Haas (SNORM ; pow=1)	33.5	43.9	20.1	33.5
MIX-APARCH (SNORM ; pow=1)	33.3	44.3	18.8	33.1
MIX-APARCH (NORM ; pow=2)	32.8	42.2	19.2	32.0
RS-GARCH-Klaassen (NORM)	32.5	41.5	19.0	31.9
EGARCH (SNIG)	32.4	43.1	19.7	34.7
MIX-GARCH (NORM)	32.2	41.0	18.9	31.3
MIX (SGED)	32.1	41.2	18.5	31.0
RS-APARCH (NORM ; pow=2)	32.0	41.4	18.5	31.1
EGARCH (SSTD)	30.8	41.9	17.8	32.2
APARCH (SSTD)	30.6	41.4	17.2	31.8
APARCH (SNIG)	30.6	41.5	18.6	33.3
RS-APARCH (NORM ; pow=1)	29.8	40.4	14.4	27.6
MIX-EGARCH (NORM)	29.5	39.8	14.6	27.6
MIX-APARCH (NORM ; pow=1)	29.5	39.9	14.5	27.7
RS-EGARCH (NORM)	29.4	39.5	14.6	27.6
GARCH (SGED)	28.6	39.7	18.2	33.4
MIX-GJR GARCH-Scale (NORM)	28.1	38.5	12.1	24.2
GARCH (SNIG)	28.0	39.1	16.4	31.1
GJR GARCH (NORM)	27.7	37.0	14.5	26.8
RS (NORM)	27.6	37.5	10.2	20.8
EGARCH (NORM)	27.2	36.9	13.1	25.6
GARCH (SSTD)	27.1	38.2	15.9	30.2
RS (SSTD)	25.4	34.4	9.8	20.5
APARCH (NORM)	24.9	35.0	4.2	6.0
RS (SNORM)	24.1	33.0	6.6	13.2
GARCH (NORM)	23.9	33.0	11.4	22.4
RS3 (NORM)	23.2	32.6	6.4	12.9
MIX (SSTD)	22.9	31.9	4.5	7.7
MIX-APARCH-Haas (NORM ; pow=1)	22.0	33.2	7.1	13.6
RS-APARCH-Haas (NORM ; pow=1)	21.8	32.9	7.1	14.1
RS (SGED)	20.9	29.6	5.9	12.0
MIX-APARCH-Haas (NORM ; pow=2)	20.6	31.5	6.7	12.8
RS-APARCH-Haas (NORM ; pow=2)	20.4	31.4	6.9	13.6
MIX (NORM)	20.1	28.9	5.6	10.7
SSTD	19.8	28.6	5.7	11.3
SNIG	19.8	28.5	5.3	10.3
SGED	19.5	28.2	5.2	10.0
NORM	18.9	27.3	4.9	9.3
RS-GARCH-Haas (NORM ; full)	16.0	27.0	7.8	14.3
RS-GARCH-Haas (NORM)	12.3	22.1	2.9	0.0

Continued on next page. . .

Table 4.11: Guarantee costs : 7 and 10-year contracts

Model	7-Year		10-Year	
	95% CTE	99% VaR	95% CTE	99% VaR
Out-of-sample	36.4		42.4	
STD	9.1	17.6	1.4	0.0
GED	8.5	17.0	1.0	0.0

Table 4.12: PVHE : 3-year contract

Model	Mean Vol. of Simul.	PV of Hedging Errors			
		StDev		95% CTE	
		Excl.	Incl.	Excl.	Incl.
GJR GARCH (SSTD)	0.140	2.66	2.71	7.50	8.04
GJR GARCH (SNIG)	0.139	2.60	2.66	7.26	7.81
RS-GARCH-Haas (NORM)	0.138	2.59	2.63	7.26	7.75
GJR GARCH (SGED)	0.138	2.54	2.60	6.97	7.53
RS-APARCH (SNORM ; pow=2)	0.140	2.59	2.66	6.84	7.40
MIX-APARCH (SNORM ; pow=2)	0.140	2.58	2.65	6.82	7.38
MIX-APARCH-Haas (SNORM ; pow=2)	0.146	2.39	2.45	6.80	7.35
MIX-APARCH-Haas (NORM ; pow=2)	0.135	2.40	2.43	6.75	7.22
RS-APARCH-Haas (NORM ; pow=2)	0.135	2.39	2.43	6.73	7.19
RS-APARCH-Haas (SNORM ; pow=2)	0.146	2.35	2.41	6.66	7.20
GARCH (SSTD)	0.130	2.51	2.56	6.54	7.07
RS-GARCH-Haas (NORM ; full)	0.132	2.51	2.55	6.47	6.97
MIX-APARCH-Haas (NORM ; pow=1)	0.136	2.24	2.28	6.30	6.76
RS-GARCH-Gray (NORM ; full)	0.129	2.46	2.51	6.27	6.81
RS-APARCH-Haas (NORM ; pow=1)	0.136	2.23	2.26	6.25	6.71
GARCH (SNIG)	0.129	2.44	2.49	6.18	6.71
RS-GARCH-Klaassen (NORM ; full)	0.137	2.39	2.46	6.18	6.73
MIX-EGARCH (SNORM)	0.137	2.23	2.28	6.03	6.54
RS-APARCH (SNORM ; pow=1)	0.137	2.19	2.25	5.98	6.48
MIX-APARCH (SNORM ; pow=1)	0.137	2.18	2.24	5.94	6.45
MIX-GARCH (SNORM ; $\alpha = 0$)	0.138	2.33	2.40	5.83	6.38
RS-APARCH (NORM ; pow=2)	0.138	2.31	2.37	5.75	6.30
MIX-APARCH (NORM ; pow=1)	0.137	2.13	2.18	5.72	6.24
MIX-EGARCH (NORM)	0.137	2.15	2.20	5.72	6.23
MIX-APARCH (NORM ; pow=2)	0.138	2.30	2.36	5.72	6.26
RS-APARCH (NORM ; pow=1)	0.137	2.12	2.17	5.69	6.21
RS-EGARCH (NORM)	0.137	2.14	2.19	5.69	6.20
MIX-GARCH (NORM ; full)	0.137	2.26	2.32	5.64	6.19
MIX-GJR GARCH-Scale (SNORM)	0.134	2.18	2.23	5.64	6.16
GARCH (SGED)	0.128	2.30	2.35	5.63	6.17
EGARCH (SSTD)	0.134	2.16	2.22	5.62	6.15
MIX-GARCH (NORM)	0.140	2.24	2.30	5.61	6.16
MIX-APARCH-Haas (SNORM ; pow=1)	0.140	2.09	2.15	5.53	6.07
MIX (SGED)	0.140	2.20	2.26	5.49	6.03
EGARCH (SNIG)	0.134	2.14	2.19	5.46	5.99
RS-GARCH-Klaassen (NORM)	0.140	2.17	2.23	5.38	5.92
APARCH (SSTD)	0.135	2.09	2.14	5.37	5.91
RS-APARCH-Haas (SNORM ; pow=1)	0.139	2.05	2.11	5.35	5.89
EGARCH (SGED)	0.134	2.11	2.17	5.33	5.88
MIX-GARCH (NORM ; $\alpha = 0$)	0.135	2.20	2.26	5.28	5.82

Continued on next page...

Table 4.12: PVHE : 3-year contract

Model	Mean Vol. of Simul.	PV of Hedging Errors			
		StDev		95% CTE	
		Excl.	Incl.	Excl.	Incl.
APARCH (SNIG)	0.135	2.06	2.12	5.24	5.77
APARCH (SGED)	0.135	2.06	2.12	5.21	5.76
APARCH (NORM)	0.145	1.84	1.90	5.13	5.69
RS-EGARCH (SNORM)	0.131	2.08	2.14	5.07	5.61
RS-GARCH-Gray (NORM)	0.140	1.97	2.01	4.96	5.47
RS (NORM)	0.142	1.85	1.89	4.94	5.46
MIX-GJR GARCH-Scale (NORM)	0.134	1.97	2.01	4.92	5.43
STD	0.145	1.72	1.73	4.62	5.06
MIX (SSTD)	0.145	1.69	1.72	4.61	5.08
EGARCH (NORM)	0.135	1.88	1.93	4.58	5.13
SSTD	0.145	1.67	1.69	4.55	5.02
GARCH (NORM)	0.130	1.98	2.02	4.43	4.96
GJR GARCH (NORM)	0.132	1.89	1.93	4.37	4.90
SNIG	0.144	1.57	1.60	4.12	4.61
MIX (NORM)	0.145	1.56	1.59	4.03	4.52
RS (SSTD)	0.124	2.05	2.10	3.98	4.47
RS (SGED)	0.126	2.04	2.09	3.74	4.26
GED	0.144	1.50	1.52	3.74	4.19
RS (SNORM)	0.127	2.05	2.10	3.71	4.22
SGED	0.144	1.46	1.49	3.66	4.16
RS3 (NORM)	0.126	1.90	1.95	3.55	4.06
NORM	0.145	1.29	1.30	3.07	3.57

Table 4.13: PVHE : 10-year contract

Model	Mean Vol. of Simul.	PV of Hedging Errors			
		StDev		95% CTE	
		Excl.	Incl.	Excl.	Incl.
GJR GARCH (SGED)	0.167	4.06	4.21	15.07	15.87
RS-GARCH-Gray (NORM ; full)	0.160	4.05	4.20	14.45	15.25
GJR GARCH (SNIG)	0.162	3.74	3.87	13.72	14.46
GJR GARCH (SSTD)	0.159	3.62	3.74	13.16	13.87
RS-GARCH-Haas (NORM ; full)	0.159	3.72	3.79	12.29	12.87
GARCH (SGED)	0.158	3.13	3.25	11.33	12.06
GARCH (SSTD)	0.154	3.09	3.19	10.88	11.56
GARCH (SNIG)	0.155	3.02	3.14	10.81	11.51
MIX-GJR GARCH-Scale (SNORM)	0.148	2.53	2.66	8.85	9.58
APARCH (SGED)	0.152	2.35	2.49	8.36	9.10
RS-GARCH-Klaassen (NORM ; full)	0.150	2.36	2.50	8.14	8.90
EGARCH (SGED)	0.152	2.28	2.42	8.06	8.80
MIX-APARCH (SNORM ; pow=2)	0.154	2.26	2.41	7.88	8.65
RS-APARCH (SNORM ; pow=2)	0.154	2.25	2.39	7.85	8.62
MIX-APARCH-Haas (SNORM ; pow=2)	0.155	2.17	2.31	7.72	8.46
EGARCH (SSTD)	0.148	2.17	2.29	7.54	8.22
APARCH (SNIG)	0.148	2.16	2.29	7.54	8.24
EGARCH (SNIG)	0.149	2.16	2.28	7.52	8.21
RS-APARCH-Haas (SNORM ; pow=2)	0.154	2.12	2.26	7.51	8.25
APARCH (SSTD)	0.148	2.15	2.27	7.49	8.18

Continued on next page...

TABLE 4.13: PVHE : 10-year contract

Model	Mean Vol. of Simul.	PV of Hedging Errors			
		StDev		95% CTE	
		Excl.	Incl.	Excl.	Incl.
RS-EGARCH (SNORM)	0.149	2.07	2.21	7.14	7.87
MIX-EGARCH (SNORM)	0.148	2.06	2.19	7.13	7.84
RS-APARCH (SNORM ; pow=1)	0.148	2.05	2.18	7.10	7.81
MIX-APARCH (SNORM ; pow=1)	0.148	2.04	2.17	7.06	7.78
MIX-GARCH (SNORM ; $\alpha = 0$)	0.151	1.97	2.11	6.63	7.38
MIX-APARCH-Haas (SNORM ; pow=1)	0.151	1.89	2.02	6.55	7.27
RS-APARCH-Haas (SNORM ; pow=1)	0.151	1.85	1.99	6.42	7.15
MIX-GARCH (NORM ; full)	0.147	1.90	2.03	6.27	7.00
GARCH (NORM)	0.154	1.79	1.87	6.03	6.68
MIX-GARCH (NORM ; $\alpha = 0$)	0.147	1.79	1.92	5.82	6.55
MIX-EGARCH (NORM)	0.145	1.73	1.84	5.80	6.47
MIX-APARCH (NORM ; pow=1)	0.146	1.72	1.83	5.78	6.45
RS-APARCH (NORM ; pow=1)	0.146	1.72	1.83	5.77	6.44
RS-EGARCH (NORM)	0.145	1.72	1.83	5.77	6.44
MIX-APARCH-Haas (NORM ; pow=2)	0.143	1.70	1.77	5.63	6.17
RS-APARCH-Haas (NORM ; pow=2)	0.143	1.67	1.74	5.51	6.05
GJR GARCH (NORM)	0.150	1.64	1.72	5.49	6.15
RS-APARCH (NORM ; pow=2)	0.147	1.69	1.80	5.45	6.15
MIX-APARCH (NORM ; pow=2)	0.147	1.67	1.79	5.40	6.11
RS-APARCH-Haas (NORM ; pow=1)	0.144	1.59	1.67	5.29	5.84
MIX-APARCH-Haas (NORM ; pow=1)	0.144	1.59	1.67	5.27	5.83
MIX (SGED)	0.148	1.63	1.74	5.26	5.97
EGARCH (NORM)	0.151	1.53	1.64	5.21	5.90
MIX-GARCH (NORM)	0.147	1.62	1.73	5.19	5.90
RS-GARCH-Klaassen (NORM)	0.147	1.61	1.72	5.13	5.84
MIX-GJR GARCH-Scale (NORM)	0.141	1.55	1.64	4.99	5.62
RS-GARCH-Haas (NORM)	0.146	1.43	1.47	4.38	4.85
RS-GARCH-Gray (NORM)	0.143	1.40	1.49	4.28	4.93
RS (SSTD)	0.144	1.33	1.41	4.27	4.85
RS (NORM)	0.144	1.18	1.26	3.65	4.24
RS3 (NORM)	0.144	1.13	1.20	3.42	3.99
RS (SNORM)	0.144	1.18	1.24	3.41	3.95
SSTD	0.144	1.04	1.10	3.10	3.63
MIX (SSTD)	0.144	1.02	1.07	3.06	3.59
APARCH (NORM)	0.146	0.86	0.93	2.71	3.26
MIX (NORM)	0.144	0.94	0.99	2.64	3.18
RS (SGED)	0.144	0.95	0.99	2.63	3.15
SNIG	0.143	0.91	0.97	2.59	3.14
STD	0.143	0.92	0.94	2.56	2.94
SGED	0.143	0.82	0.88	2.24	2.79
GED	0.143	0.74	0.76	1.89	2.26
NORM	0.144	0.70	0.72	1.72	2.23

Chapitre 5

MEASURING THE EFFECTIVENESS OF DYNAMIC HEDGES FOR LONG-TERM INVESTMENT GUARANTEES

Le contenu de ce chapitre est basé sur trois projets subventionnés par l'Autorité des marchés financiers et supervisés par Mathieu Boudreault. L'Autorité des marchés financiers est l'organisme mandaté par le gouvernement du Québec pour encadrer les marchés financiers québécois et prêter assistance aux consommateurs de produits et services financiers. Ce chapitre est écrit sous la forme d'un article pour uniformiser la présentation de cette thèse.

Résumé

La littérature financière a dévoué beaucoup de recherche pour faire progresser les modèles économétriques dans le but d'améliorer la tarification et la couverture des produits financiers. Toutefois, les approches permettant de mesurer l'efficacité d'une stratégie de couverture dynamique ont peu évolué. Cet article propose un cadre statistique, basé sur la régression, permettant de mieux évaluer l'efficacité de différentes stratégies de couverture dynamique dans le contexte de produits vendus avec des garanties d'investissement à long terme (ex. : fonds distincts). L'importance de prendre en compte le risque de modèle est soulignée. De plus, il est démontré qu'une garantie d'investissement simple, connue sous l'acronyme GMMB (*guaranteed minimum maturity benefit*), peut être reformulée comme un ensemble d'options à barrière lorsque l'investisseur a l'option de résilier son contrat. Ceci nous permet de tarifer et couvrir ce produit avec des formules analytiques et, ainsi, étudier l'impact de la déchéance dynamique (*dynamic lapsation*) sur l'efficacité de la couverture. Les difficultés associées à réduire le risque de modèle et celui du comportement de l'assuré (*policyholder behavior*)

lors d'une couverture dynamique nous amènent à proposer une nouvelle perspective sur le rôle de la couverture, soit celle d'un outil permettant de modifier le profil risque–récompense de la position non-couverte.

Abstract

Although the finance literature has devoted a lot of research into the development of advanced models for improving pricing and hedging performance, the approaches for measuring dynamic hedging effectiveness have evolved little. This article offers a methodological contribution in this area by proposing a statistical framework, based on regression analysis, for measuring the effectiveness of dynamic hedges for long-term investment guarantees (e.g., segregated funds or variable annuities). The importance of taking model risk into account is emphasized. Moreover, it is shown that a simple investment guarantee, the guaranteed minimum maturity benefit, can be decomposed into a basket of barrier options when the policyholder surrenders his contract based on a fixed moneyness level. This allows us to price and hedge this product with analytical formulas and investigate the impact of dynamic lapsation risk on hedging effectiveness. The difficulties in reducing model and policyholder behavior risks when hedging lead us to propose a new perspective on hedging, and recognize it as a tool to modify the risk–reward relationship of the unhedged position.

Keywords : hedging effectiveness, dynamic hedging, model risk, dynamic lapsation, investment guarantee, variable annuity

5.1. Introduction

Risk management practices in the financial industry were put to a test during the financial crisis of the late 2000s. Many corporations did not measure their risk exposures appropriately and numerous institutional failures occurred (e.g., Lehman Brothers). Some financial risks, such as market risk, have a systematic component that cannot be diversified. To avoid having a large exposure to these types of risks, the corporation can establish a hedging strategy by trading in financial derivatives. To correctly evaluate the residual risk that remains, it must assess the effectiveness of this strategy for mitigating the underlying risk exposure.

Suppose that an institution has sold a derivative and established a hedging strategy to manage its risk. Assume that the value of this derivative and hedge at some time in the future is X and Y, respectively. The hedged loss, or equivalently the hedging error, of the institution then corresponds to $X - Y$ (a positive value indicates a loss while a negative value indicates a profit). The standard approach

in the finance literature to measure hedging effectiveness is to calculate a statistic on the distribution of $X-Y$. Popular statistics include measures of dispersion such as standard deviation (StDev) and average absolute deviation from zero (AAD), as well as risk measures such as conditional tail expectation (CTE), and Value-at-Risk (VaR). For example, in Chapter 4, hedging effectiveness was evaluated by examining the StDev and CTE of hedging errors, which is in agreement with standard practices in finance and actuarial science.

Although the finance literature has devoted a lot of research into the development of advanced models for improving pricing and hedging performance, the approaches for measuring dynamic hedging effectiveness have evolved little. For example, Bakshi et al. [13] develop models admitting stochastic volatility, stochastic interest rates, and random jumps and test empirically the features which improve dynamic hedging performance of plain vanilla options. To measure hedging effectiveness, they calculate the AAD of hedging errors based on dynamic delta and delta-vega strategies. These hedging errors are calculated based on a single rebalancing of the hedge (usually daily or weekly) and do not reflect the cumulative hedging error from the sale of the option until maturity. Similar approaches are applied in many other articles [e.g., 2, 3, 5, 6, 14, 25, 58, 112, 119, 140]. However, Branger et al. [25] and Kaeck [112] recently improve on the methodology to measure hedging effectiveness by comparing the distributions of anticipated and realized hedging errors.[1] If the distributions of anticipated and realized hedging errors differ significantly, this suggests some form of model misspecification in the hedge.

The studies mentioned in the previous paragraph have two weaknesses when it comes to measuring hedging effectiveness. First, they ignore the relationship between X and Y and do not examine its strength. For instance, why is the StDev of $X - Y$ high ? Is it due to a systematic deviation from the ideal hedging relationship $Y = X$, or is it the consequence of a high residual standard error around an average relationship of $Y = X$? One contribution of this chapter is to propose a statistical framework based on regression analysis to more thoroughly examine the effectiveness of dynamic hedges. Statistics based on the distribution of $X - Y$ are good indicators of *relative* hedging performance and allow us to discriminate between competing hedging strategies. However, they are not good indicators of *absolute* hedging performance as they ignore the relationship between X and Y. Regression analysis allows us to assess hedging effectiveness in *absolute*

1. Anticipated hedging errors are those obtained when the model used to determine the hedge coincides with the data generating process. Realized hedging errors correspond to those that are observed when the hedging strategy is implemented on empirical data or on simulated data based on a data generating process that is not consistent with the model used for hedging.

terms and understand the sources of hedging ineffectiveness. To our knowledge, the use of regression analysis to measure the effectiveness of dynamic hedges has not been investigated in the finance or actuarial science literature. On the other hand, the accounting literature [e.g., 36, 66, 89] has investigated the use of regression tools to measure the strength of a hedging relationship, but only in the context of very simple hedges which are not dynamic and not model dependent (e.g., hedging a spot index with a futures contract). [2]

The second weakness of empirical studies measuring hedging effectiveness in the finance literature is that they examine hedging errors calculated based on a single rebalancing of the hedge (usually daily or weekly) and do not reflect the cumulative hedging error from the sale of the option until maturity. The rationale for following such an approach is that traders assess their profits and gains on a daily or weekly basis and do not always keep their positions open until maturity. Therefore, it is acceptable to measure hedging effectiveness based on intermediate, rather than cumulative, hedging errors. However, intermediate hedging errors are influenced by the time to maturity and the moneyness of the option. For example, the hedge may be less dependent on the chosen model far from maturity and when the option is deep out-the-money. Consequently, when we calculate statistics based on intermediate hedging errors, we are aggregating observations that are not identically distributed. To alleviate this problem, most of the studies display statistics across different moneyness ranges, but not all of them separate their results according to time to maturity. However, empirical studies in the finance literature aim to discriminate among competing hedging strategies, and their main objective is, thus, the evaluation of *relative* hedging effectiveness. For this purpose, it is not problematic to consider intermediate hedging errors. Nevertheless, these errors only provide partial information with respect to the capacity of the hedge to replicate the desired payoff exactly, until

2. The interest of the accounting literature for measuring hedging effectiveness is due to Statements of Financial Accounting Standards No. 133, Accounting for Derivative Instruments and Hedging Activities, commonly known as FAS 133. FAS 133, which came into effect in the U.S. at the turn of the century, requires that all derivatives entered into by a corporation must be marked to market and changes in their values reported in the income statement. This accounting treatment can create earnings volatility when derivatives are used for risk management purposes as the timing of gains and losses on the hedged items may not be matched with those on the corresponding hedging derivatives. To remedy this problem, FAS 133 allows corporations to match the timing of these gains and losses, provided that they demonstrate and document that the hedge is *highly effective* in offsetting changes in fair value for the risks being hedged. FAS 133 does not endorse any specific testing methodology, but recommends the use of statistical tests. A good overview of approaches proposed in the accounting literature for measuring hedging effectiveness is given by Charnes et al. [36], Finnerty and Grant [66], and Hailer and Rump [89].

maturity. Therefore, *absolute* hedging effectiveness is best assessed by examining the cumulative hedging error until maturity.

This concept is especially important in the context of variable annuities because guarantees sold have long-term maturities, typically from three to 20 years, and we are mainly interested in the cumulative performance of the hedge. This is in strong contrast to maturities of derivatives considered in the finance literature which rarely are above one year. Moreover, insurers are not traders and cannot close their positions by transferring them to a third party. In other words, investment guarantees are illiquid securities. Accordingly, all studies on dynamic hedging effectiveness of long-term investment guarantees in the actuarial literature have considered the cumulative hedging error until maturity [e.g., 10, 39, 40, 96, 122].

Section 5.3 proposes a statistical framework based on both, regression analysis and the standard practice of calculating risk measures, to measure the effectiveness of dynamic hedges on *absolute* as well as *relative* terms. This statistical framework is presented with a case study examining the effectiveness of the Black-Scholes delta hedging strategy under return path scenarios generated with the regime-switching GARCH model (RS-GARCH). We consider this hedging strategy for simplicity, but also because the majority of insurers in Canada establish Black-Scholes dynamic delta or delta-rho hedging strategies to manage risks in variable annuities. Hence, our case study reflects reality. We do not hedge the rho risk, i.e., interest rate risk, as this thesis is concerned with econometric modeling of equity returns. Although the focus of Section 5.3 is on the measurement of hedging effectiveness, rather than model risk, this risk is an integral part of the analysis as dynamic hedging strategies depend on model assumptions. The performance of the hedge depends on how well these assumptions represent reality. We consider an extended definition of model risk relative to the one presented in Chapter 4, and include parameter risk as a specific type of model uncertainty, see Cont [42, Remark 4.1] for a discussion. Model uncertainty refers to the combined effects of model and parameter risks. The purpose is not to distinguish between these two types of risks, but to examine the impact of particular types of model uncertainty on hedging effectiveness. In our case study, the product hedged is the same as the one considered in Chapter 4, a guaranteed minimum maturity benefit (GMMB). Although our application centers on options embedded in financial products with guarantees, the concepts introduced apply to a wide range of derivatives.

5.2. Preliminaries

Before describing the statistical framework for measuring the effectiveness of dynamic hedges, we review some concepts in relation to the GMMB, introduce the data, models and assumptions, and explain how to calculate the insurer's unhedged and cumulative hedged losses at maturity.

5.2.1. Pricing and delta hedging for the GMMB product

Suppose that an insured invests $A_0 > 0$ in a GMMB product that must be held to maturity T. The investment performance of the GMMB tracks a market index from time $t = 0$ to T, denoted by $\{S_t\}_{0 \le t \le T}$. Without loss of generality, let $S_0 = A_0$. As compensation for the guarantee offered, the insurer deducts fees in proportion to the account value of the insured at a continuous annual rate of δ. Therefore, the account value of the insured, $\{A_t\}_{0 \le t \le T}$, satisfies the relationship :

$$A_t = S_t e^{-\delta t}, \quad 0 \le t \le T.$$

At maturity T, the insured is entitled to $\max(A_T, G)$, where G denotes the amount of the guarantee. If $A_T < G$, the guarantee matures in-the-money and the insurer is responsible for the shortfall, i.e., his liability is the payoff of a put option : $\max(0, G - A_T)$. The value of this put option at any time t under the Black-Scholes model, denoted by $P_t(A_t, G, \delta)$, is obtained from the following formula :

$$\begin{aligned} P_t(A_t, G, \delta) &= G e^{-r(T-t)} N(-d_2) - A_t e^{-\delta(T-t)} N(-d_1), \\ d_1 &= \frac{\log(A_t/G) + (r - \delta + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \\ d_2 &= d_1 - \sigma\sqrt{T-t}, \end{aligned}$$

where r is the constant risk-free rate, σ is the constant annual volatility of the market index and $N(\cdot)$ denotes the standard normal cumulative distribution function. Moreover, the fair value of fees that will be collected by the insurer between times t and T is given by

$$A_t(1 - e^{-\delta(T-t)}).$$

To see why, we can interpret the charge δ as a dividend rate. The fair value of dividends to be received between times t and T is the difference between the fund value at time t (A_t) and the prepaid forward price for a claim paying A_T at time T ($A_t e^{-\delta(T-t)}$).

The net liability of the insurer at time t, i.e., the net obligation of the insurer towards the policyholder, is

$$L_t = \text{value of put option} \; - \; \text{fair value of fees that will be collected}$$

$$= P_t(A_t, G, \delta) - A_t(1 - e^{-\delta(T-t)}).$$

A fair value for δ can be determined by setting the net liability at inception of the contract to zero :

$$L_0 = P_0(A_0, G, \delta) - A_0(1 - e^{-\delta T}) = 0. \tag{5.2.1}$$

Solving equation (5.2.1) is not a difficult problem since the net liability is strictly decreasing in δ :

$$\frac{\partial}{\partial \delta}\left[P_0(A_0, G, \delta) - A_0(1 - e^{-\delta T})\right] = -TA_0e^{-\delta T}N(d_1) < 0.$$

Consequently, the solution is unique and easily computed numerically.

We can rewrite equation (5.2.1) as

$$\begin{aligned} A_0 &= S_0e^{-\delta T} + P_0(A_0, G, \delta), \\ &= S_0e^{-\delta T} + P_0(S_0, Ge^{\delta T}, 0)e^{-\delta T}, \end{aligned}$$

where $P_0(S_0, Ge^{\delta T}, 0)$ is the price at $t = 0$ of a put option on the market index with maturity T and strike $Ge^{\delta T}$. This suggests that an investment of A_0 in a GMMB product is equivalent to buying $e^{-\delta T}$ shares of the underlying market index and $e^{-\delta T}$ put options on that index.

To obtain an ideal hedge, the objective function that needs to be hedged is the net liability of the insurer. Unfortunately, this notion has not been emphasized in the actuarial literature. For example, Hardy [96] suggests to hedge the GMMB by replicating a long position in the underlying put position. However, the GMMB is not a standard put option as no premium is paid at inception of the contract. Since there is uncertainty in both the payoff and the premium, both components need to be hedged to eliminate risk. The uncertainty with respect to the fees is even greater when the policyholder is allowed to surrender his policy before maturity.

A delta hedge under the Black-Scholes framework can be established by holding a position of Δ_t in the underlying market index at time t, where

$$\begin{aligned} \Delta_t = \frac{\partial}{\partial S_t}L_t &= \frac{\partial L_t}{\partial A_t} \cdot \frac{\partial A_t}{\partial S_t} \\ &= [-e^{-\delta(T-t)}N(-d_1) - (1 - e^{-\delta(T-t)})] \cdot e^{-\delta t} \\ &= -e^{-\delta T}N(-d_1) - (e^{-\delta t} - e^{-\delta T}) \leq -e^{-\delta T}N(-d_1). \end{aligned}$$

This involves selling $(e^{-\delta t} - e^{-\delta T})$ more shares of the underlying market index than if only the put position was hedged as the delta of the put position is $-e^{-\delta T}N(-d_1)$. The intuition behind this result is the following : If the stock price

falls, the fair value of fees will be reduced (since fees are charged in proportion to the account value) and the additional short position in the stock will provide a hedge against this decrease in premium.

5.2.2. Data, parameter estimation and assumptions

To realistically model return dynamics on the market index $\{S_t\}$, we use the RS-GARCH model given in equations (2.2.1)–(2.2.3) with constraints $\alpha = \alpha_1 = \alpha_2$ and $\beta = \beta_1 = \beta_2$. In addition to modeling volatility persistence with GARCH dynamics, this model generates jumps in returns [3] (switching mean) and volatility (switching intercept in the GARCH equation). Therefore, it contains the essential ingredients of a good financial model as the presence of stochastic volatility, jumps in returns and jumps in volatility receive strong evidence in the finance literature [e.g., 26, 63].

We estimate the parameters of the model to two data sets : daily and weekly percentage log-returns on the S&P 500 index from October 28, 1987 to October 31, 2012. The daily data set includes 6305 observations. The weekly data set was considered in Chapters 2 and 3 and contains 1305 observations. The use of both data sets allows us to evaluate the consistency of results obtained with daily and weekly data. Descriptive statistics are given in Table 5.1 (the mean and StDev are given on an annualized basis). The parameter estimation was accomplished based on the collapsing technique developed in Chapter 3 with $q = 16$. Table 5.2 provides maximum likelihood estimates of the RS-GARCH model for daily and weekly data sets. Asymptotic standard errors are given in parentheses.

Table 5.1. Descriptive statistics

	Mean	StDev	Skewness	Kurtosis	Minimum	Maximum
Daily S&P 500	7.2	18.5	−0.27	11.4	−9.5	11.0
Weekly S&P 500	7.2	16.6	−0.59	7.2	−16.5	10.2

We assume that the insurer uses the Black-Scholes model to price the GMMB, i.e., to determine the annual fee assumption δ, and delta hedge its risk. Since the hedging strategy is not derived from the market model, we are able to analyze the robustness of the Black-Scholes delta hedge under RS-GARCH dynamics. The

3. The switching mean does not necessarily generate a jump in returns, but there is a high likelihood of observing a significant negative return when the model enters the crisis regime. When jump-diffusion models are discretized, the occurrence of a negative jump is equivalent to observing a large negative return in a given time interval.

TABLE 5.2. Maximum likelihood estimates for the S&P 500 data sets

μ_1	μ_2	ω_1	ω_2	α	β	p_{11}	p_{22}
Daily RS-GARCH							
0.081	−1.63	0.0058	0.544	0.042	0.936	0.980	0.339
(0.010)	(0.20)	(0.0013)	(0.087)	(0.006)	(0.006)	(0.004)	(0.083)
Weekly RS-GARCH							
0.339	−2.80	0.0431	2.527	0.041	0.905	0.948	0.316
(0.064)	(0.54)	(0.0185)	(0.518)	(0.018)	(0.024)	(0.022)	(0.105)

following assumptions are used :

$$A_0 = 100, \quad G = 100, \quad T = 10, \quad r = 0.03, \quad \sigma = 0.169 \quad \text{and} \quad \delta = 0.0112,$$

where σ is the volatility parameter used in the Black-Scholes model to determine the fair value of fees and to calculate the hedge position Δ_t in the market index. The value of $\delta = 0.0112$ satisfies equation (5.2.1) with these assumptions. The volatility assumption σ corresponds to the unconditional annualized volatility for the daily RS-GARCH model when parameter risk is taken into account (see Section 5.2.3).

5.2.3. Projection of the insurer's loss

There are three variables that must be considered when evaluating hedging effectiveness :

$$\begin{aligned} X &= \text{Net unhedged loss at maturity,} \\ Y &= \text{Cumulative mark-to-market gain on the hedge,} \\ X - Y &= \text{Net hedged loss at maturity.} \end{aligned}$$

When the insurer does not use a hedging strategy, his net loss at maturity is X. When he employs a hedging strategy, his net loss is $X - Y$. The losses are "net" because they take into account the fee income received by the insurer. To assess hedging effectiveness, we are interested in the relationship between X and Y and in the dispersion of $X - Y$ around zero. Consequently, we must project many realizations of these variables by simulating return path scenarios.

First, we generated 100000 scenarios of daily returns over $T = 10$ years with the Black-Scholes model. Log-returns under the Black-Scholes model are independent and identically distributed normal variables. We used an annualized mean

return of 7.2% (this value matches the empirical mean given in Table 5.1) and an annualized volatility of $\sigma = 16.9\%$. This simulation allows us to assess the effectiveness of the Black-Scholes delta hedge under an ideal situation where the hedging strategy is derived from the market model. In this context, hedging ineffectiveness arises solely from discrete rebalancing of the hedge portfolio as there is no model uncertainty. This type of hedging ineffectiveness is referred to as the discretization error of the hedge. Results obtained for the return projection under the Black-Scholes model serve as a benchmark and allow us to properly evaluate the impact of model uncertainty on the Black-Scholes delta hedging strategy.

Second, we generated 100000 scenarios of daily and weekly returns over $T = 10$ years with the two RS-GARCH models presented in Table 5.2. To evaluate the impact of parameter risk, we generated one set of scenarios with maximum likelihood estimates and another set of scenarios by drawing parameter vectors from the asymptotic normal distribution of the maximum likelihood estimator.

For each of the return path scenarios, the values of the market index, $\{S_t\}_{0\leq t\leq T}$, and the account value, $\{A_t\}_{0\leq t\leq T}$, are easily calculated. The net unhedged loss at maturity is

$$\begin{aligned} X &= \text{payoff to the insured } - \text{ accumulated value of fees} \\ &= \max(0, G - A_T) - \sum_{i=0}^{T/h-1} A_{ih}(1 - e^{-\delta h})e^{r(T-ih)}, \end{aligned}$$

where $h = 1/252$ for daily scenarios and $h = 1/52$ for weekly scenarios.

To calculate the net hedged loss at maturity, the cumulative mark-to-market gain on the hedge must be subtracted from the net unhedged loss. The mark-to-market gain at time $t + h$ of the delta hedge established at time t is

$$\Delta_t \cdot (S_{t+h} - S_t e^{rh}),$$

where h represents the rebalancing frequency of the hedge portfolio. The cumulative mark-to-market gain on the hedge corresponds to the accumulated values of these gains to maturity :

$$Y = \sum_{i=0}^{T/h-1} \Delta_{ih} \cdot (S_{(i+1)h} - S_{ih}e^{rh}) \cdot e^{r(T-(i+1)h)}.$$

5.3. Statistical framework for measuring the effectiveness of dynamic hedges

5.3.1. Regression analysis

This section explains how regression techniques can be used to investigate hedging effectiveness. The analysis is conducted for scenarios projected on a daily frequency and the hedge portfolio is assumed to be rebalanced daily.

5.3.1.1. *Relationship between Y and X*

Figure 5.1 illustrates the regression relationship between the cumulative mark-to-market gain on the hedge, Y, (denoted by hedge performance on the vertical axis) and the net unhedged loss at maturity, X, for 100000 scenarios projected under the Black-Scholes model, the RS-GARCH model and the RS-GARCH model with parameter risk. On each graph, we show (1) the fitted linear regression line of the form "$y = a + bx$," (2) the residual standard error (σ_ϵ), and (3) Pearson's and Spearman's correlation coefficients (ρ and ρ_{rank}, respectively). Spearman's correlation coefficient measures the relationship between the ranks of variables. These values are calculated based on all observations and based on only the part of the distribution for the net unhedged loss (X) contained between the 50^{th} and 95^{th} percentiles (between red lines). The reason for considering this area of the distribution is the following. When a put option is deep in-the-money or deep out-of-the money, its delta is very close to -1 or 0, respectively, regardless of the model for the underlying. In these situations, hedging the put option entails less model uncertainty as the delta under the *true* model is closer to the Black-Scholes delta. Hence, the Black-Scholes delta hedge is expected to be more effective in the lower half and the upper part of the distribution for the net unhedged loss.

For scenarios projected with the Black-Scholes model, the correlation coefficients between the hedge performance (Y) and the net unhedged loss (X) are close to one for all observations and between the red lines. The regression relationship corresponds almost exactly to the equation $Y = X$, which was to be expected as the hedging strategy is constructed from the market model.

For RS-GARCH scenarios, the correlation coefficients calculated on all observations misrepresent the strength of the relationship between the hedge performance and the net unhedged loss as they markedly decrease when we consider the area between the red lines. With parameter risk, Spearman's correlation coefficient is approximately 0.85 which is a sizable decrease relative to the value of 0.99 observed for the Black-Scholes scenarios. Moreover, the slope of the regression line

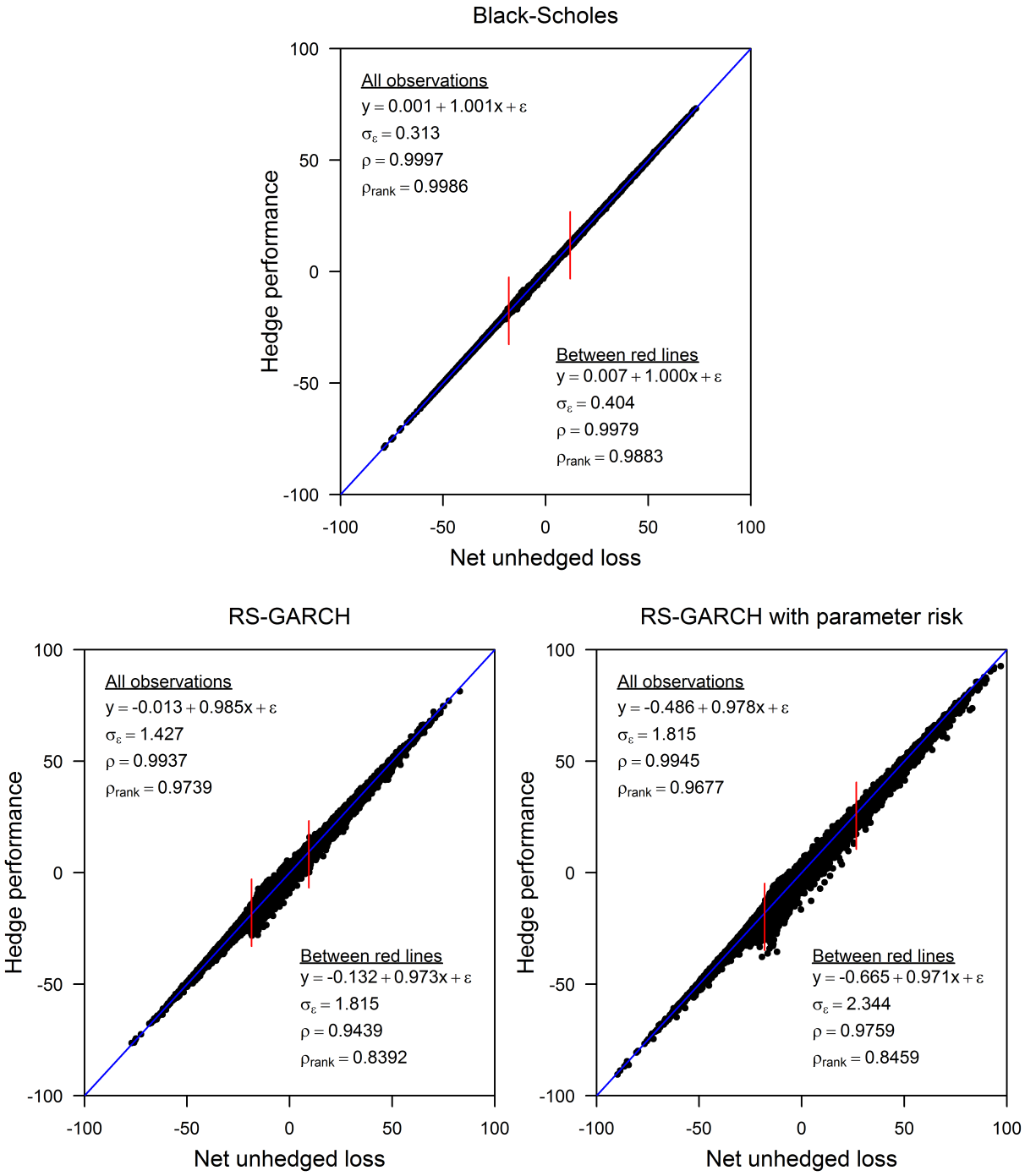

FIGURE 5.1. Relationship between the hedge performance and the net unhedged loss based on 100000 daily scenarios projected with the Black-Scholes model (top panel), the RS-GARCH model (bottom-left panel) and the RS-GARCH model with parameter risk (bottom-right panel).

is below one which indicates that the hedge portfolio falls short in offsetting the insurer's loss. This is one advantage of using regression analysis to investigate hedging effectiveness as it tells us what proportion of the risk is not replicated on average. With parameter risk, over 2% of the risk is not offset while in between red lines this value increases to 3%. The intercept in the regression is also informative as it can detect a structural loss for the insurer due to hedging. This can in turn suggest that the GMMB product is underpriced. For example, with parameter risk, the intercept term is negative indicating a structural loss. Finally, the residual standard error is magnified by a factor of almost six with respect to the Black-Scholes scenarios and parameter risk accounts for approximately 25% of that increase. Therefore, model uncertainty much more significantly impacts σ_ϵ than the average hedging relationship, $Y = X$.

The signs of hedging ineffectiveness observed under model uncertainty can be important because by hedging the net liability, the insurer protects himself against losses at the cost of limiting his upside potential. In fact, hedging can be viewed as a tool to modify the risk–reward relationship of the product. While it is clear that hedging reduces the overall risk, the resulting risk–reward relationship may be less attractive to the insurer if the strategy employed has inefficiencies. Therefore, an effective hedging strategy should not only reduce the risk of the insurer's position, but also offer a way to improve the risk–reward tradeoff for the product. For example, suppose that instead of hedging the net liability, the insurer only hedges its liability, i.e., the put option. Hence, the fee income is not hedged and the Black-Scholes delta at time t is $-e^{-\delta T}N(-d_1)$. Figure 5.2 illustrates the hedging relationship for this situation under the RS-GARCH model with parameter risk. We observe that in the first half of the distribution for the net unhedged loss, i.e., in 50% of the scenarios, the insurer is able to take advantage of a sizable upside because the fee income is not hedged. The tradeoff for this benefit is a reduced hedging effectiveness in the upper half of the distribution. On average, the hedging strategy falls short in offsetting 17% of the risk in between red lines and the negative intercept indicates a structural loss due to hedging. In the right tail of the net unhedged loss, most of the scatter points fall below the blue line, $Y = X$, implying that the net hedged loss of the insurer is positive. This example demonstrates that the effectiveness of a hedge should also be evaluated with respect to the risk–reward relationship it offers because a perfect hedge is infeasible in a practical context.

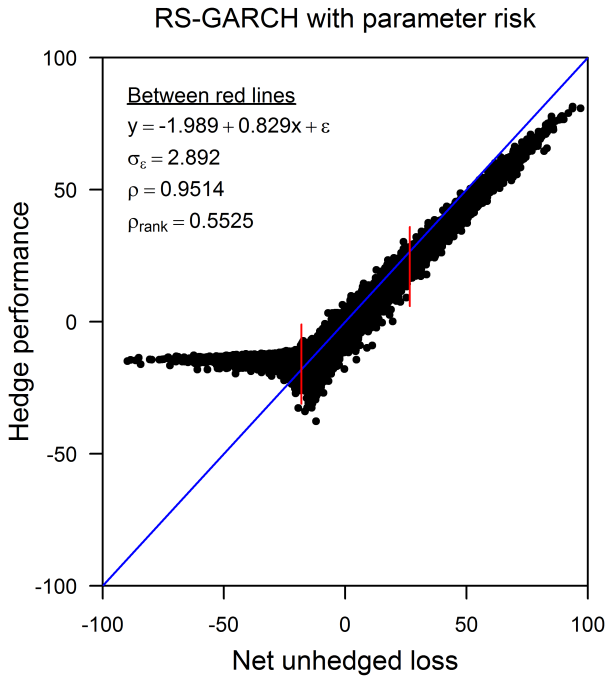

FIGURE 5.2. Relationship between the hedge performance and the net unhedged loss based on 100000 daily scenarios projected the RS-GARCH model with parameter risk. Only the underlying put option is delta hedged.

5.3.1.2. *Relationship between $X - Y$ and X*

Having studied the relationship between Y and X, we now examine the dependence between the net hedged loss, $X - Y$, and the net unhedged loss, X. These two variables inevitably exhibit a form of dependence because, as mentioned previously, hedging effectiveness tends to be higher in the lower half and the upper part of the distribution for X. However, they should ideally be uncorrelated since what is desired is a hedging strategy whose effectiveness is uniform under a wide variety of scenarios. This is simply because hedging must protect the insurer against an unknown return path scenario which is to occur in the real-world.

Figure 5.3 illustrates the relationship between the net hedged loss and the net unhedged loss for 100000 scenarios projected under the Black-Scholes model, the RS-GARCH model and the RS-GARCH model with parameter risk. On each graph, we show (1) Pearson's correlation coefficient between X and $X - Y$ (ρ), (2) the standard deviation of $X - Y$ (σ), and (3) the standard deviation of $X - Y$ in between red lines (σ_{red}). As before, the area in between red lines spans the 50^{th} to 95^{th} percentiles of the distribution for X.

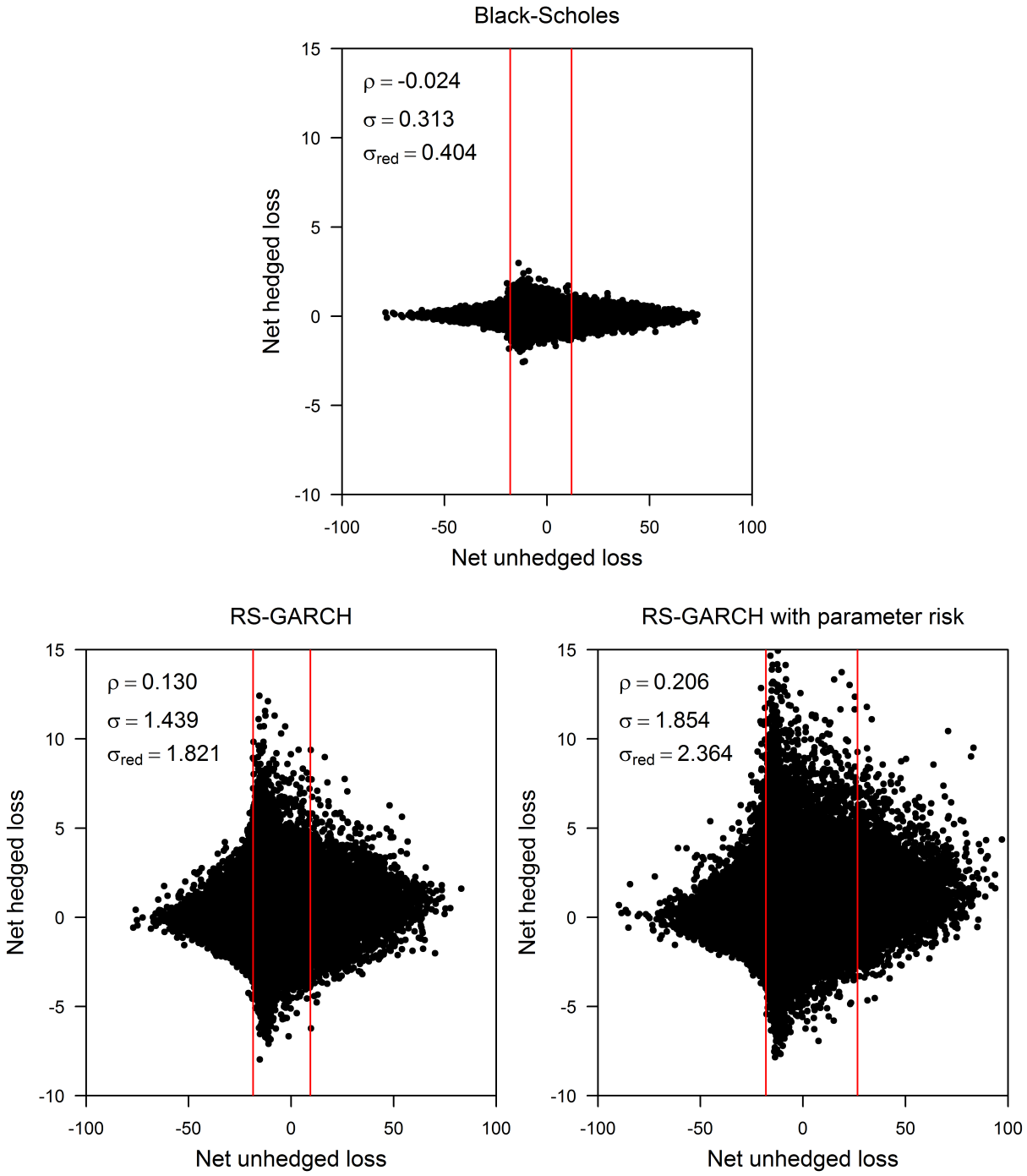

FIGURE 5.3. Relationship between the net hedged loss and the net unhedged loss based on 100000 daily scenarios projected with the Black-Scholes model (top panel), the RS-GARCH model (bottom-left panel) and the RS-GARCH model with parameter risk (bottom-right panel).

As expected, hedging errors are less dispersed in the tails of the distribution for the net unhedged loss and the scatter plots are organized in a diamond. Figure 5.3 clearly illustrates the impact of model and parameter risks. Model risk has a considerable influence on hedging effectiveness as the dispersion of scatter points under RS-GARCH scenarios is much greater than that observed under Black-Scholes scenarios. The impact of parameter risk is less than model risk, but it is still significant. In particular, parameter risk considerably widens the right tail of the distribution for the net unhedged loss, which implies that extreme investment losses are more likely.

Furthermore, the correlation coefficient between the net hedged loss and the net unhedged loss based on Black-Scholes scenarios is close to zero. This was to be expected since scenarios generating the largest hedging errors are generally those associated with a high volatility. As there is no dependence between returns and volatility in the Black-Scholes model, there should be no linear relationship between the net hedged loss and the net unhedged loss. In contrast, the RS-GARCH model entails a leverage effect, i.e., a negative relationship between returns and future volatility, since it simultaneously generates negative returns and positive jumps in volatility when it enters the crisis regime (regime two). Consequently, we expect a positive correlation between the net hedged loss and the net unhedged loss which is what is observed. This systematic trend could offer a way to improve the hedging strategy. For example, when returns are negative, the put option gains value directly through the price decrease of the underlying asset and indirectly through an expected higher future volatility. The latter part is not taken into account with delta hedging. To consider it, a quantity dependent upon the strength of the leverage effect should be subtracted from the Black-Scholes delta.

5.3.1.3. *Distributions of X and $X - Y$*

The left panel in Figure 5.4 illustrates the empirical density of the net unhedged loss (X) based on 100000 scenarios projected under the RS-GARCH model with and without parameter risk. The right panel shows the Q-Q plot of standardized quantiles between these two distributions. This figure confirms that parameter risk has a non-negligible effect on the tails of the net unhedged loss, and it should therefore not be ignored.

Figure 5.5 repeats the previous exercise for the net hedged loss ($X - Y$). This is analogous to examining the distribution of residuals from the perfect hedging relationship $Y = X$. As observed previously, the dispersion of hedging errors with model risk is much greater than without it (top-left panel), but the Q-Q

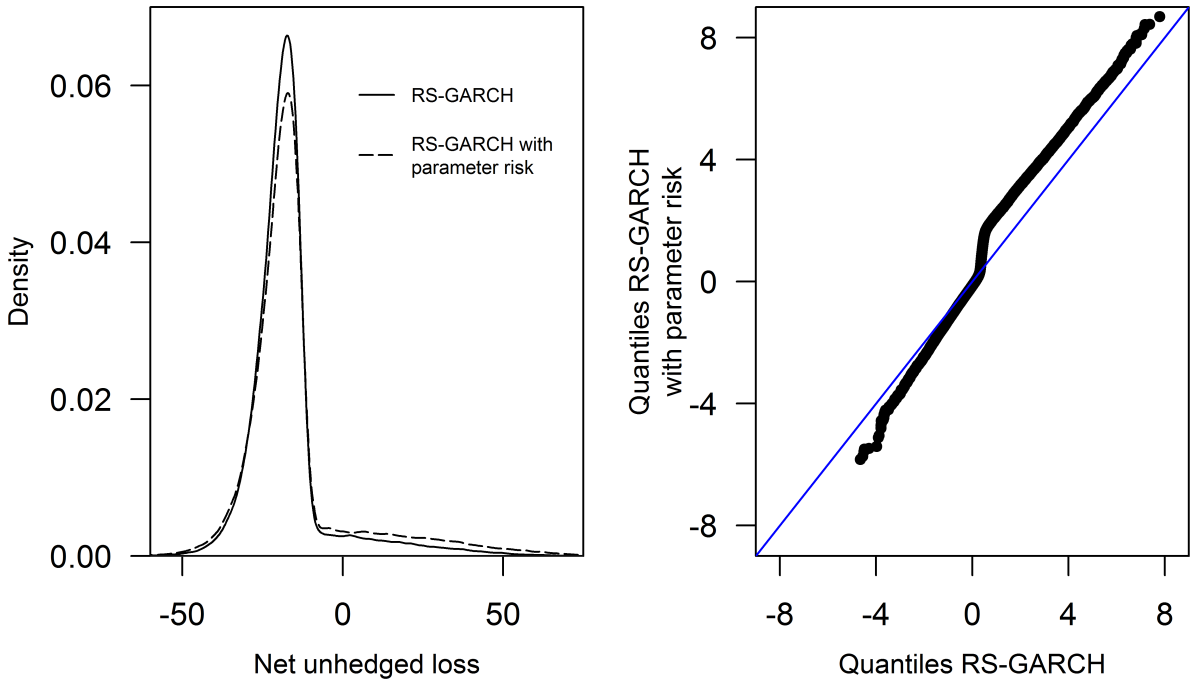

FIGURE 5.4. Left panel : Empirical density of the net unhedged loss (X) based on 100000 daily scenarios projected under the RS-GARCH model with and without parameter risk. Right panel : Q-Q plot of standardized quantiles between these two distributions.

plot (bottom-left panel) allows us to quantify the magnitude of this discrepancy. Similar to the net unhedged loss, parameter risk widens the tails of the net hedged loss (see bottom-right panel), but its impact on the right tail is even more pronounced here.

However, the most interesting result is perhaps derived from the Q-Q plot in the top-right panel. This plot compares standardized quantiles between the distribution of hedging errors under the Black-Scholes model and those of the standard normal distribution, denoted by $N(0,1)$. It clearly shows that hedging risk is a heavy-tailed risk, even in the absence of model uncertainty. In other words, the insurer is exposed to possibly large losses, five to 10 standard deviations away from the mean loss, when he employs the ideal hedging strategy in the context of a very simple data generating process (Black-Scholes model) that does not even allow for jumps or stochastic volatility. These extreme losses are due to the discretization error of the hedge. In theory, this type of error can be reduced by rebalancing the hedge portfolio more frequently, say many times a day. However, in practice, frequent rebalancing of the hedge portfolio entails large transaction costs and sudden price jumps can occur in between rebalancing times. Hedging at

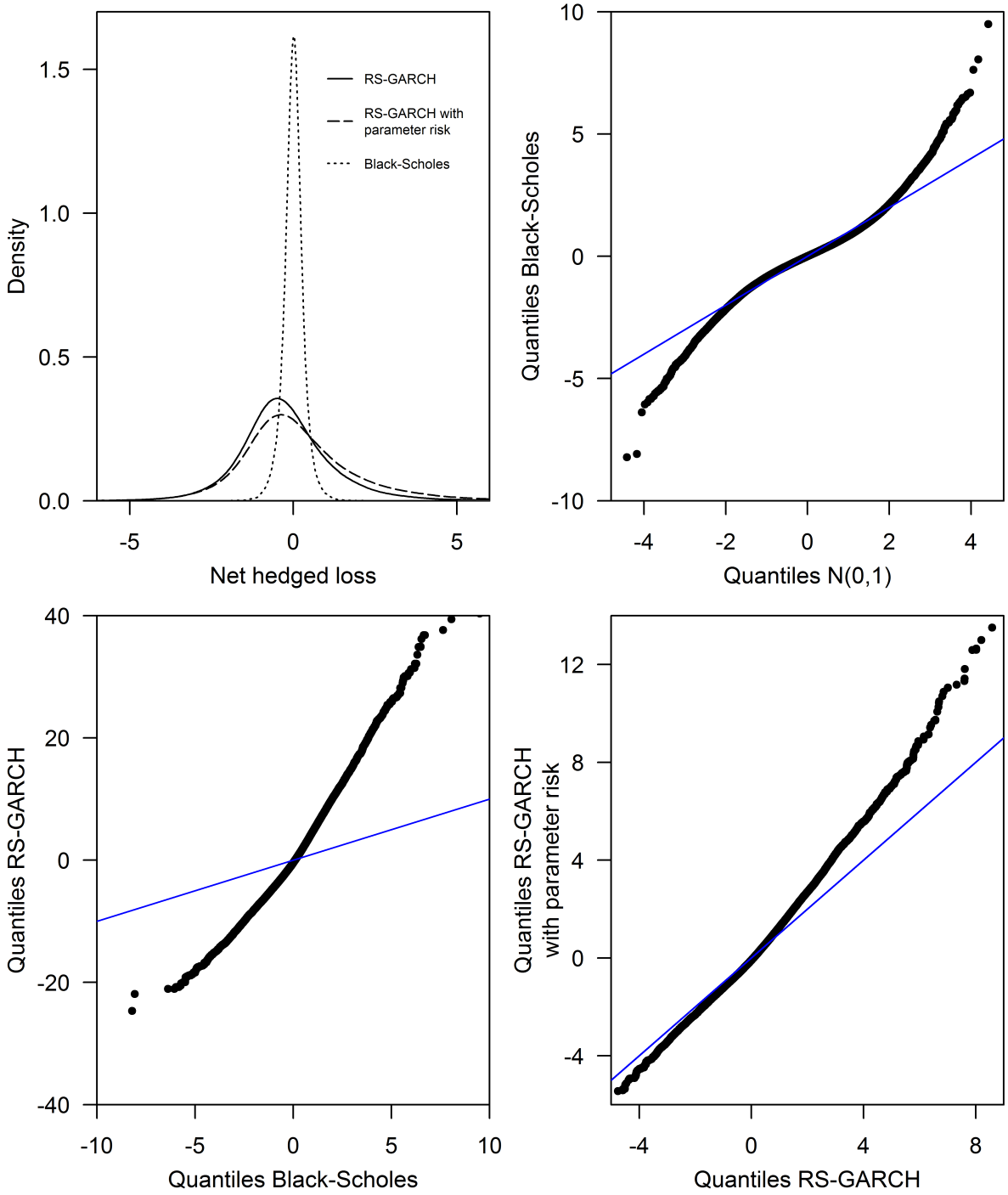

FIGURE 5.5. Top-left panel : Empirical density of the net hedged loss $(X - Y)$ based on 100000 daily scenarios projected under the Black-Scholes model, the RS-GARCH model and the RS-GARCH model with parameter risk. Other panels : Q-Q plots of standardized quantiles comparing distributions for the net hedged loss.

very high frequencies also implies additional model uncertainty (further deviations from the normal distribution) as the distribution of high frequency returns is more volatile and has heavier tails than the one associated with daily or weekly log-returns. This discussion emphasizes that the insurer should not aim for a perfect hedge, but for one that offers a desired risk–reward tradeoff. In practice, hedging significantly reduces the overall risk of loss, but a residual heavy-tailed hedging risk remains. For a hedging strategy to be effective, the insurer must be rewarded for taking this risk.

5.3.2. Risk measures

The conventional approach to measuring hedging effectiveness and discriminating among competing strategies is based on measures of dispersion and risk measures calculated on the distribution for the net unhedged loss $(X - Y)$. For example, we can consider the mean, StDev, AAD, CTE, and VaR. Given n sampled values from a random variable Z, denoted by $\{z_i\}_{i=1}^n$, the AAD is given by

$$\text{AAD} = \frac{1}{n}\sum_{i=1}^{n} |z_i|.$$

The StDev and the AAD measure the dispersion of hedging errors around the mean and zero, respectively. The CTE and VaR measure the heaviness of the right tail of the distribution for $X - Y$. These statistics can offer a good way to summarize important elements of the distribution for $X - Y$ and discriminate among competing hedging strategies. However, information about the relationship between X and Y is lost.

This section uses risk measures to determine how the rebalancing frequency of the hedge portfolio impacts hedging effectiveness. We also examine the consistency of results obtained with daily and weekly RS-GARCH models estimated in Table 5.2. Table 5.3 shows the results for the analysis of hedge rebalancing frequency based on 100000 scenarios projected under the Black-Scholes model and the RS-GARCH model with parameter risk. First, we observe that increasing this frequency lowers risk measures and thus improves hedging effectiveness of the Black-Scholes delta hedge with or without model uncertainty. However, this improvement is not as significant in the context of model uncertainty. For example, from a monthly to a daily rebalancing frequency, the 95% CTE is reduced by a factor of five for the Black-Scholes model and by a factor of only 1.6 with model uncertainty. This implies that the reduction in the discretization error is greater than the increase in model error (at frequencies studied here), but model error accounts for a greater proportion of the total hedging error at higher

TABLE 5.3. Risk measures for the insurer's net loss at maturity based on the Black-Scholes model (B-S) and the RS-GARCH model with parameter risk (RSG)

Rebalancing	Mean		StDev		AAD		95% CTE		99% VaR	
	B-S	RSG	B-S	RSG	B-S	RSG	B-S	RSG	B-S	RSG
Unhedged	−16.3	−14.7	13.0	17.6	19.4	20.9	27.4	42.9	37.2	53.7
Annual	1.5	2.4	5.5	5.9	4.4	4.7	14.5	17.4	16.8	20.3
Monthly	0.1	0.9	1.5	2.8	1.1	2.0	3.4	8.6	4.0	10.2
Weekly	0.0	0.6	0.7	2.2	0.5	1.6	1.7	6.8	2.0	8.0
Daily	0.0	0.2	0.3	1.9	0.2	1.3	0.7	5.2	0.9	6.3

rebalancing frequencies. For instance, daily rebalancing offers little improvement over weekly rebalancing and may even be less effective once transaction costs are taken into account.

Although daily rebalancing generates the lowest risk measures, the mean loss of the insurer is approximately zero. This implies that he is not rewarded for assuming hedging risk. In contrast, when he does not hedge, his mean loss is negative, implying a profit, but he is exposed to very large losses. Since market risk is not diversifiable, the insurer may not tolerate such an exposure and prefer to hedge. However, when a classic risk-neutral pricing approach is used to price the product, hedging risk is not considered and the hedging strategy may result in an unattractive risk–reward relationship. To address this issue, insurers generally perform pricing with conservative assumptions and include a margin for profit. However, this methodology does not consider hedging risk directly. It seems preferable to incorporate this risk into pricing and construct a hedging strategy that offers a desirable risk–reward tradeoff. This entails that both, risk-neutral and real-world measures, must be considered in pricing as hedging risk can only be properly quantified under the real-world measure.

Table 5.4 examines the consistency of results obtained with daily and weekly RS-GARCH models estimated in Table 5.2. Both models are estimated on the same period, October 28, 1987 to October 31, 2012, and 100000 return path scenarios are projected taking parameter risk into account. First, we observe that the 95% CTE and the 99% VaR of the net unhedged loss differ markedly, suggesting that long-term investment risk is difficult to measure accurately. This observation was also made in Chapter 4. This is in part due to the difficulty in estimating the average mean return, i.e., the drift dynamics of the underlying asset. For example,

TABLE 5.4. Risk measures for the insurer's net loss at maturity based on daily (abbreviated D) and weekly (abbreviated W) RS-GARCH models with parameter risk.

Rebalancing	Mean		StDev		AAD		95% CTE		99% VaR	
	D	W	D	W	D	W	D	W	D	W
Unhedged	−14.7	−7.8	17.6	23.4	20.9	21.2	42.9	62.0	53.7	71.9
Annual	2.4	1.8	5.9	6.6	4.7	4.8	17.4	19.4	20.3	23.1
Monthly	0.9	0.7	2.8	3.5	2.0	2.4	8.6	10.7	10.2	12.7
Weekly	0.6	0.1	2.2	2.9	1.6	2.0	6.8	8.3	8.0	9.8

the daily RS-GARCH model generates an average annualized mean return of 7.1% which corresponds approximately to the empirical annualized mean of 7.2% (see Table 5.1). In contrast, we obtain an average annualized mean return of 4.6% for the weekly model. This explains, in part, why higher risk measures are computed for the net unhedged loss based on the weekly model. Moreover, the estimation period has an even greater influence on the drift assumption. For example, the daily S&P 500 data from May 20, 1999 to April 25, 2011, considered in Chapters 2 and 3, has an average annualized mean of −0.1 (see Table 2.5). Since future drift dynamics are difficult to infer from historical data, it is preferable to use a risk management strategy whose dependence on this assumption is weak. Fortunately, hedging generally provides such an outcome.

We observe that risk measures at different rebalancing frequencies generally fall in a comparable range for the daily and weekly models. However, they are a little higher for the weekly model. This is mainly due to increased parameter risk at a weekly frequency caused by a smaller historical sample. Although the choice of data frequency does influence results, it does not do so enough to alter conclusions.

5.4. Impact of dynamic lapsation on hedging effectiveness

The analyses performed in Chapter 4 and in Section 5.3 assumed the GMMB product is held to maturity, i.e., the policyholder does not surrender his contract. In reality, he generally has the option to lapse his policy, but to discourage him in doing so, the insurer charges a penalty, called surrender charge. For example, suppose that markets perform well and the guarantee falls far out-of-the-money. In this situation, the insurance is less valuable to the policyholder, but he is

still charged the original fees in proportion to the account value. Consequently, he has a strong incentive to surrender his policy because the insurance is now expensive. When lapsation depends on the level of the market index, it is referred to as dynamic lapsation. Therefore, the insurer's hedging strategy is exposed to both, model and policyholder behavior uncertainties. A second element that we investigated in the series of projects done in collaboration with the Autorité des marchés financiers is the impact of dynamic policyholder behavior on hedging effectiveness. In other words, how is hedging effectiveness affected when there is uncertainty with respect to the decision of the insured to surrender his contract ?

There are many ways to model lapse behavior, the two extremes being deterministic lapsation and optimal behavior. Deterministic lapsation is a diversifiable risk and can be treated in the same way as mortality risk for pricing and hedging purposes. Optimal behavior refers to an optimal choice of dynamic lapsation, but it is not supported by available data. Knoller et al. [124] [see also 138] make an empirical investigation to determine which factors influence surrenders and find that the moneyness of the guarantee is a key driver of lapse behavior. For example, surrender rates are higher when the guarantee is out-of-the-money and lower when it is in-the-money. Consequently, lapsation risk is not independent of the account value and, unlike mortality risk, it cannot be diversified away according to the principle of the law of large numbers. The Canadian Institute of Actuaries [32] and the American Academy of Actuaries [7] both recommended to take dynamic lapsation into account by varying the lapse rate depending on the moneyness of the guarantee. According to a report from the Society of Actuaries [157], approximately 60% of insurers follow this practice. Therefore, we will base the surrender decision of the policyholder on the moneyness of the guarantee.

To date, the actuarial literature centered on the impact of optimal policyholder behavior on pricing variable annuities. For example, Bacinello [11], Bacinello et al. [12], and Milevsky and Salisbury [136] investigate pricing of investment guarantees under the assumption of optimal dynamic lapsation. Chen et al. [37], Dai et al. [44], and Milevsky and Salisbury [137] examine pricing of guaranteed minimum withdrawal benefits (GMWB) [4] under optimal policyholder behavior. Even though policyholders do not behave optimally, pricing assuming that they do consists in a conservative approach. On the other hand, the impact of dynamic policyholder behavior (optimal or not) on hedging effectiveness has seldom been investigated. Kling et al. [121] consider different approaches to modeling policyholder behavior (deterministic, optimal, and based on moneyness) for the GMWB, and finds that

4. The GMWB product guarantees the insured that he will be able to make minimum periodic withdrawals from his account until a set maturity regardless of investment performance.

hedging effectiveness deteriorates significantly when actual policyholder behavior deviates from the one assumed in the hedge.

Therefore, there is a need to better understand how policyholder behavior uncertainty impacts hedging effectiveness, and this is precisely the objective of this section. Assuming that the policyholder bases his decision to surrender the GMMB product on a fixed moneyness level, we show that the GMMB product can be decomposed into a basket of barrier options. This allows us to price and hedge this product with analytical formulas under the Black-Scholes model, and investigate the impact of ignoring lapsation risk, or of not modeling it appropriately, on hedging effectiveness.

5.4.1. GMMB product with dynamic lapsation risk

This section explains how to calculate the net liability of the insurer for the GMMB product with dynamic lapsation risk. In case of surrender at time $0 < t < T$, the policyholder receives $A_t(1-\kappa(t))$ from the insurance company, where $\kappa(t)$ represents the surrender charge at time t ($0 \leq \kappa(t) \leq 1$). We assume the policyholder will surrender his contract once the moneyness ratio, A_t/G, hits a predetermined barrier $\xi(t)$, $0 < t < T$. The time at which surrender occurs is denoted by τ, where

$$\tau = \min\left\{0 < t < T : \frac{A_t}{G} \geq \xi(t)\right\},$$

with the convention that $\tau = T$ if the barrier is never reached during the term of the contract. For a specific choice of $\xi(t)$, τ will correspond to the time of optimal lapsation. To model lapse behavior realistically and for analytic tractability, we assume that $\xi(t)$ and $\kappa(t)$ do not depend on t and set $\xi(t) = \xi$ and $\kappa(t) = \kappa$. In this particular situation, the GMMB product is surrendered at the first time the fund value hits the constant barrier $H = G\xi$ and the policyholder receives $H(1-\kappa)$.

Table 5.5 shows that when $\xi(t)$ and $\kappa(t)$ are constant, the GMMB product with dynamic lapsation risk is equivalent to a basket of barrier options. This decomposition has not appeared in the literature previously and allows us to relate this product to well-known barrier options. Note that it includes no lapsation as a special case ($\xi = \infty$). The combination of options II and III in Table 5.5 is similar to an American capped call (with zero strike) which is a type of option discussed by Broadie and Detemple [27]. McDonald [133, Section 22] gives analytical expressions for the prices of barrier options presented in Table 5.5 under the Black-Scholes model. Consequently, the net liability of the insurer at time t can be calculated analytically as it corresponds to the sum of option prices in

TABLE 5.5. Decomposition of the payoff for the GMMB product with dynamic lapsation risk

	Barrier is hit at τ	Barrier is not hit
I. Up-and-out put	0	$\max(0, G - A_T)$
II. Rebate option	$H(1-\kappa)$ [paid at τ]	0
III. Up-and-out call with zero strike	0	A_T
Total : GMMB with lapsation	$H(1-\kappa)$ [paid at τ]	$\max(A_T, G)$

Table 5.5 less the account value A_t. In the more general case where $\xi(t)$ and $\kappa(t)$ are time-varying, finite difference methods must be used to determine the value of the GMMB.

5.4.2. Pricing

Given H and κ, the fair annual fee charge δ can be determined by setting the net liability to zero at inception of the contract. When the policyholder lapses, the income of the insurer comprises fees collected periodically until surrender and the surrender charge, $H\kappa$. This charge compensates the insurer for the lost fee income due to lapsation. Therefore, well designed surrender charges can help mitigate lapsation risk with respect to pricing.

We use the same assumptions as in Section 5.2.2, where we examined the case of no lapsation :

$$A_0 = 100, \quad G = 100, \quad T = 10, \quad r = 0.03 \quad \text{and} \quad \sigma = 0.169,$$

and set $\kappa = 0.04$. Figure 5.6 plots the fair charge δ as a function of the fund barrier inducing surrender, H, under these assumptions. The horizontal dotted line corresponds to $\delta = 0.0112$ which is the fair charge when the policyholder does not lapse. We observe that as H increases, the fair value of δ converges to this quantity which was to be expected. If the insurer prices the GMMB product assuming no lapsation, we must choose a fund barrier between 126.9 and 213.3 to gain advantage of the surrender option and the *best* choice of H is 143.5. However, the benefit gained is not substantial which implies that lapsation risk with respect to pricing is mitigated when $\kappa = 0.04$. For instance, if there were no surrender charges, i.e., $\kappa = 0$, the fair value of δ would be 0.0206 (instead of 0.0129) with $H = 143.5$, which represents almost a two-fold increase with respect to the case of no lapsation ($\delta = 0.0112$).

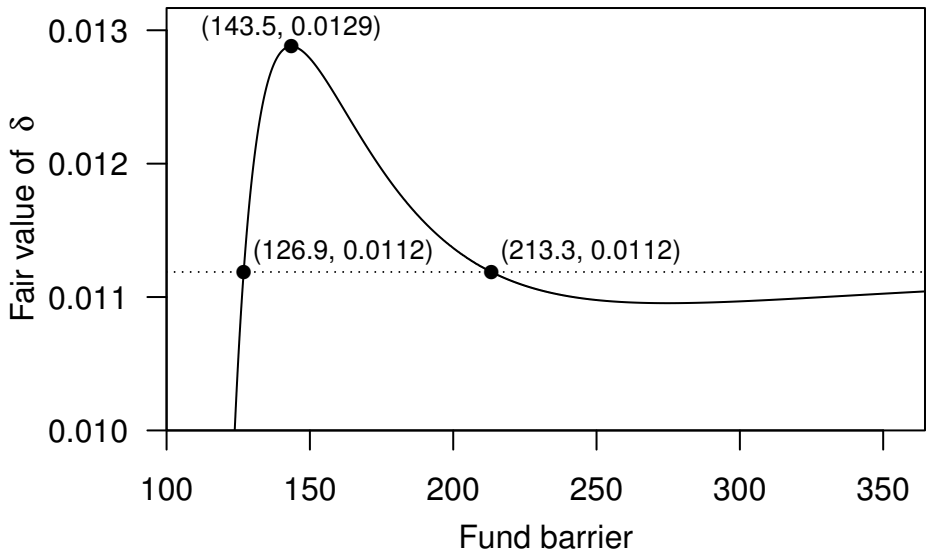

FIGURE 5.6. Fair value of δ for the GMMB product with dynamic lapsation risk as a function of the fund barrier inducing lapsation, H. The surrender charge assumption is $\kappa = 0.04$.

5.4.3. Projection of the insurer's loss

When the product is not surrendered, the insurer's unhedged and hedged losses are calculated as in Section 5.2.3. If the product is lapsed before maturity at time τ, the net unhedged loss at maturity is defined as

$$\begin{aligned} X &= -(\text{accumulated value of fees and surrender charges}) \\ &= -\sum_{i=0}^{\tau/h-1} A_{ih}(1-e^{-\delta h})e^{r(T-ih)} - H\kappa \cdot e^{r(T-\tau)}. \end{aligned}$$

In this situation, the hedging strategy is stopped at time τ and the cumulative mark-to-market gain on the hedge accumulated to maturity is

$$Y = \sum_{i=0}^{\tau/h-1} \Delta_{ih} \cdot (S_{(i+1)h} - S_{ih}e^{rh}) \cdot e^{r(T-(i+1)h)}.$$

As before, the net hedged loss at maturity is $X - Y$.

5.4.4. Hedging effectiveness under dynamic lapsation risk

This section examines how hedging errors are impacted when the possibility of dynamic lapsation is not taken into account, or not modeled appropriately, in the hedge. We assume $\delta = 0.0112$ which corresponds to the fair annual charge in the case of no lapsation. As before, the insurer delta hedges the net liability under the Black-Scholes model. Four scenarios are considered :

I. **Benchmark scenario** : The insurer hedges the GMMB product assuming no surrenders and the policyholder does not lapse.

II. **Appropriate hedge scenario** : The insurer hedges the GMMB product assuming the policyholder lapses his contract once the fund value reaches $H = 150$ and the policyholder conforms to this behavior.

III. **Wrong barrier hedge scenario** : The insurer hedges the GMMB product assuming the policyholder lapses his contract once the fund value reaches $H = 175$, but the policyholder lapses his contract once the fund value reaches $H = 150$. This situation allows us to assess the impact of dynamic lapsation on a hedging program when this risk is not modeled appropriately.

IV. **Inappropriate hedge scenario** : The insurer hedges the GMMB product assuming no surrenders and the policyholder lapses his contract once the fund value reaches $H = 150$. This situation allows us to assess the impact of dynamic lapsation on a hedging program when this risk is ignored.

Table 5.6 displays risk measures calculated for the net hedged loss based on 100000 daily return path scenarios projected under the Black-Scholes model and the RS-GARCH model with parameter risk. First, let us ignore model uncertainty and concentrate on the results for the Black-Scholes model. We observe a positive mean loss in scenario II because the insured surrenders his policy to his advantage on average, and this type of behavior is not priced by the insurer. Correcting for this mean loss, the risk measures for scenarios I and II are comparable. This suggests that the GMMB product with dynamic lapsation risk is not more difficult to hedge than without this risk in a context where the insurer has exact knowledge of this risk. Our finding is consistent with the one reported by Kling et al. [121] who concluded that the impact of policyholder behavior on hedging effectiveness "is rather small and very similar in all scenarios where assumed and

TABLE 5.6. Risk measures for the insurer's net loss at maturity based on the Black-Scholes model (B-S) and the RS-GARCH model with parameter risk (RSG)

Scenario	Mean		StDev		AAD		95% CTE		99% VaR	
	B-S	RSG	B-S	RSG	B-S	RSG	B-S	RSG	B-S	RSG
I	0.0	0.2	0.3	1.9	0.2	1.3	0.7	5.2	0.9	6.3
II	0.7	0.6	0.3	1.8	0.8	1.4	1.6	5.5	1.7	6.4
III	1.1	1.0	1.2	2.0	1.4	1.7	2.4	5.2	2.5	6.0
IV	1.7	1.8	3.2	3.4	2.9	3.2	7.1	7.5	7.4	7.9

actual policyholder behavior coincide, no matter how the considered policyholder strategy looks like."

However, when the insurer ignores dynamic lapsation risk in his delta hedge, we observe a significant negative impact on hedging effectiveness. For example, the 95% CTE of the net hedged loss increases from 1.6 to 7.1 for scenarios II to IV. Even after correcting for the mean loss, the differences are important. Nevertheless, the results in scenario III are encouraging. If the insurer's model for dynamic lapsation is a little off from the actual policyholder behavior, risk measures increase with respect to scenario II, but by a much smaller margin than in scenario IV.

We now use regression analysis to determine what proportion of the net unhedged loss is not being replicated on average when dynamic lapsation risk is not modeled properly or ignored. To isolate the impact of this risk, we continue our analysis under the Black-Scholes projection. Figure 5.7 illustrates the regression relationship between Y and X in scenarios II, III and IV, and displays regression equations computed for values falling below the 90^{th} percentile of the net unhedged loss (red line). The scenarios in the upper decile generate fund values that remain far below the barrier inducing lapsation and, consequently, the hedging relationship is close to perfect in this region as there is little lapsation risk and no model risk. In other words, if $A_t \ll H$, it is unlikely that the insured will surrender his policy and the net liability with and without dynamic lapsation risk are approximately equal. First, we observe that in scenario II, the hedging relationship to the left of the red line is excellent as the slope of the regression and correlation coefficients are close to one. The intercept is negative since we assumed no lapsation risk for the purpose of pricing, which results in a price that is too low, i.e., a structural loss for the insurer. In scenario III, this hedging relationship deteriorates considerably as 35% of the risk is not being replicated. This is surprising because the risk measures analysis suggested only a small decrease in hedging effectiveness. In the area to the left of the red line, the net unhedged loss is negative, implying a gain for the insurer. Consequently, offsetting only a portion of this gain can be seen positively. However, the intercept term drops significantly from -0.6 to -5.8 which reduces this advantage. The combination of these two effects results in a small increase in the risk measures from scenario II to III. This example demonstrates that by only examining risk measures, important information about the hedging relationship is lost. Finally, in scenario IV, the hedging relationship to the left of the red line is nonexistent. The risk measures analysis already detected a problem, but it is now clear that for at least 90% of the scenarios, the insurer is not hedging.

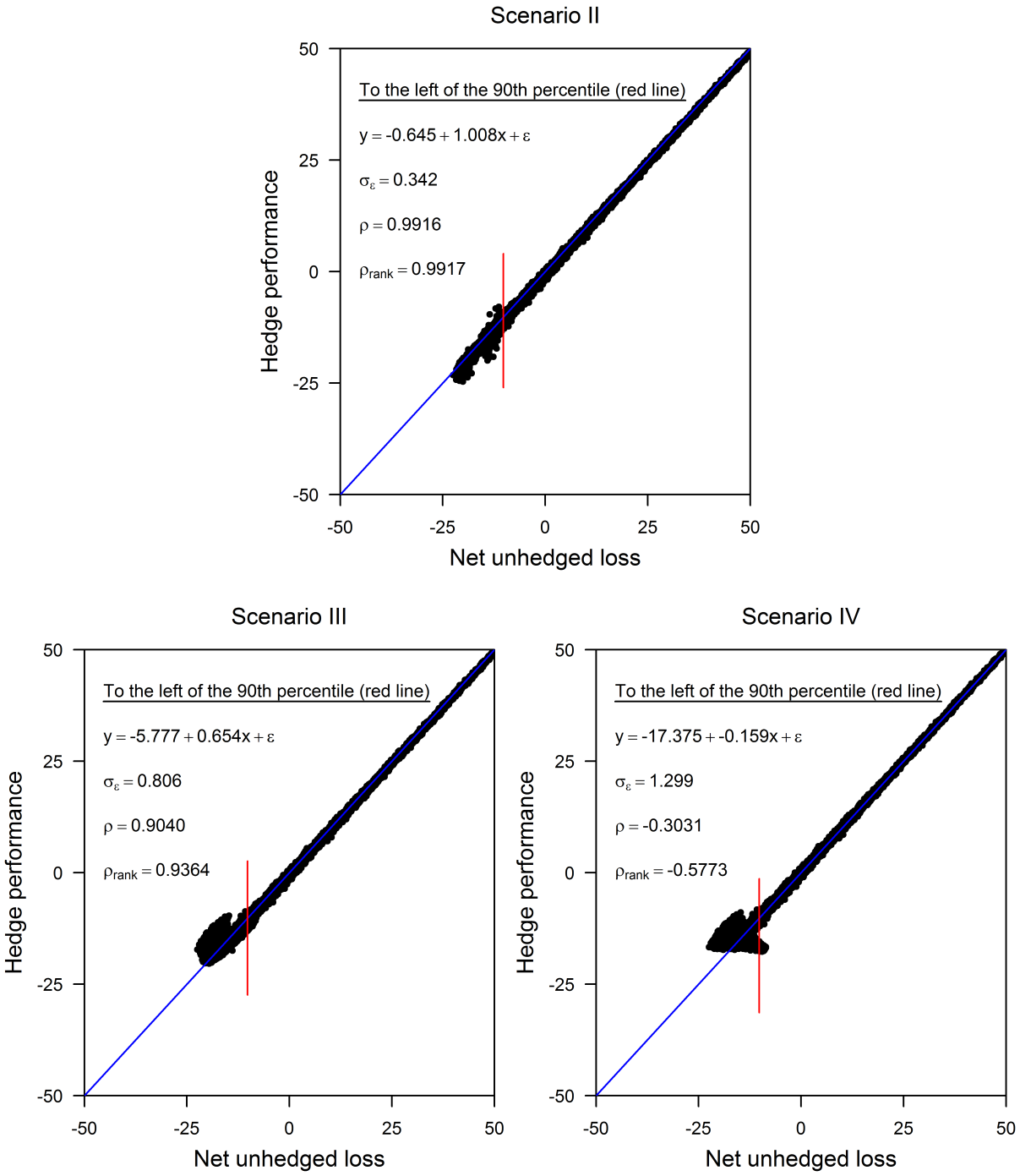

FIGURE 5.7. Relationship between the hedge performance and the net unhedged loss based on 100000 daily scenarios projected with the Black-Scholes model for scenarios II (top panel), III (bottom-left panel) and IV (bottom-right panel).

These examples emphasize the significant impact of dynamic lapsation risk on hedging effectiveness : Ignoring or not modeling this risk appropriately can destroy the hedging relationship. In contrast, model uncertainty had a limited impact on the hedging relationship as only a small portion of the net unhedged loss was not replicated (2–3%). This risk mainly impacted the regression relationship through an increase in the residual standard error.

We now turn our attention to the combined effects of model, parameter and dynamic lapsation risks on hedging effectiveness. For scenarios I and II, we obtain the same conclusions as in Section 5.3.2 where it was shown that model uncertainty significantly increases risk measures of the net hedged loss. Not hedging dynamic lapsation risk further increases these risk measures in scenario IV, but by a much smaller margin than under the Black-Scholes model. In fact, risk measures are comparable for scenario IV under the Black-Scholes and RS-GARCH models. This seems to indicate that not taking dynamic lapsation risk has the effect of offsetting a portion of the model error in the delta hedge. In Section 5.3.1.2, we suggested that one way to improve the Black-Scholes delta hedge is to adjust it, as to reflect the leverage effect, by subtracting a positive quantity from it. In our context, this involves selling more shares of the underlying market index than what is prescribed by the Black-Scholes model. However, dynamic lapsation risk has the opposite effect on the delta when the GMMB is out-of-the-money, i.e., it increases it. In this situation, there is a higher chance that the fund will reach the moneyness barrier, that the product will be lapsed and that the insurance will not be needed. Consequently, the insurer needs to sell less shares of the underlying market index in his delta hedge as he is only required to replicate the guarantee partially. This justifies why dynamic lapsation and model risks tend to offset each other in the context of the Black-Scholes delta hedge. In other words, model error is reduced by committing another error, ignoring lapsation risk. Of course, this is not a correct approach to reduce model error.

5.5. Conclusion and avenues for future research

We have proposed a statistical framework based on regression analysis to measure the effectiveness of dynamic hedging strategies and evaluated how model and policyholder behavior uncertainties can impact this effectiveness. We found that model uncertainty has a limited impact on the proportion of the risk that is not replicated with the Black-Scholes delta hedge, in the order of 2–3%. However, this risk considerably increases the residual standard error of the hedging relationship, $Y = X$. To reduce this standard error, we may increase the rebalancing frequency, and rebalance many times a day, but this approach has limitations as

observed price dynamics are not continuous and can exhibit jumps, even on very small time intervals. An alternative is to construct a hedging strategy based on a more sophisticated model than Black-Scholes, such as Heston's [105], and use the principle of local risk minimization [see 148].

For example, Poulsen et al. [148] show that the locally risk-minimizing delta hedge in Heston's model is given by

$$\Delta + \rho\sigma\frac{v}{S}, \tag{5.5.1}$$

where Δ is the price sensitivity of the derivative in Heston's model, ρ is the instantaneous correlation between price and volatility risks, σ is the volatility of volatility, v is the sensitivity of the derivative with respect to the instantaneous variance, and S is the asset price. When $\rho < 0$, Heston's model reproduces the leverage effect, and the locally risk-minimizing delta (5.5.1) incorporates a downward correction to the standard delta. This correction reflects that volatility risk can be hedged indirectly with the underlying asset as a drop in the asset price is generally accompanied with a rise in volatility. In Section 5.3.1.2, we suggested that such an adjustment would lead to an improvement in hedging effectiveness. It would be interesting to evaluate whether the correction implied by the locally risk-minimizing delta (5.5.1) is accurate under RS-GARCH dynamics, and whether it can bring the hedging relationship closer to $Y = X$. A priori, we expect an improvement, but it is unclear whether the residual standard error will also be reduced.

The discussion in the previous paragraph highlights the importance of evaluating the added value provided by sophisticated models for hedging purposes. The finance literature includes many empirical studies addressing this subject [e.g, 2, 3, 5, 6, 14, 25, 58, 112, 119, 140]. The consensus is that modeling stochastic volatility is of first order importance for hedging purposes, and the inclusion of jumps into the models does not result in any clear improvement. However, much of this research also finds that ad-hoc adjustments to Black-Scholes hedges, which are popular with traders, can do well [see 2, 6]. The comparison of hedging effectiveness in these studies is exclusively based on empirical data. This data generally comprises prices of plain vanilla options with time to maturities smaller than one year for which there is a developed liquid market. As we tested hedging effectiveness on simulated data, our approach deviates from the one followed in those studies. Testing hedging effectiveness on empirical data is equivalent to evaluating it on a single, but very credible, scenario. However, the performance of the hedging strategy is not tested against a wide enough range of dynamics that could occur in the future. In other words, simulated data can help us assess

the robustness of a hedging strategy in the presence of model uncertainty, while empirical data can confirm whether this strategy was effective in the past. Moreover, simulated data can sometimes be our only resort to make a meaningful analysis of hedging effectiveness, e.g., in the context of variable annuities that have long-term maturities and where a liquid market for the underlying options does not exist.

There are only a few studies that use simulated data to evaluate the robustness of dynamic hedges [see 25, 148, 149]. For example, Poulsen et al. [148] conclude that the locally risk-minimizing delta hedging strategy under Heston's model is robust to model misspecification. However, the robustness of this strategy was only evaluated with respect to simulated data from a single model, the SABR model of Hagan et al. [88], which lies in the same general class of stochastic volatility models as Heston's model. Branger et al. [25] also test the hedging effectiveness of Heston's model under simulated data generated with a stochastic volatility model with jumps, and find that it performs well. However, their study is based on intermediate (rather than cumulative) hedging errors, the hedging strategy is derived from a special case of the data generating process, and hedging risk is analyzed under the risk-neutral measure. Consequently, the impact of model uncertainty on hedging effectiveness is a subject that should be further researched. We hope that the statistical framework provided in this chapter can help enhance these types of analyses.

This chapter also evaluated how policyholder behavior uncertainty impacts hedging effectiveness. We found this uncertainty to be detrimental to the hedging relationship, considerably more so than model uncertainty. The difficulties in reducing these types of risks lead us to suggest a new perspective on the role of hedging. The traditionally accepted purpose of hedging is risk elimination, or maximum risk reduction. However, as a perfect hedge is unattainable and the capacity of risk reduction is limited in practice, hedging could be reinterpreted as a tool to modify the risk–reward relationship of the unhedged position. In this framework, the role of hedging is to reduce risk to an acceptable level (not necessarily to a maximum level), and offer the highest reward for bearing it. For instance, the risk–reward relationship of the insurer may improve if he decides not to hedge the fees in certain situations or hedge only a portion of the liability. This approach deviates from the traditional risk-neutral framework used for pricing by incorporating hedging risk directly into the pricing methodology, as opposed to indirectly with conservative assumptions. This implies that pricing should reflect both, risk-neutral and real-world measures. This new perspective on hedging constitutes an interesting avenue for future research in actuarial applications.

CONCLUSION

Cette thèse a premièrement traité de l'estimation du modèle GARCH à changement de régimes, puis, deuxièmement, de l'importance du risque de modèle dans les applications financières en actuariat, notamment dans les fonds distincts (*variable annuities*).

La contribution principale du premier thème de recherche consiste en le développement de deux approches originales permettant de calculer l'estimateur du maximum de vraisemblance du modèle GARCH à changement de régimes, un problème qui n'a pas été résolu depuis l'introduction de ce modèle il y a déjà près de 20 ans. D'une part, une approche basée sur l'algorithme Monte Carlo EM et sur l'échantillonnage préférentiel a été proposée pour estimer le modèle et calculer la matrice de variance-covariance asymptotique de l'estimateur du maximum de vraisemblance. D'autre part, il a été démontré comment construire un filtre particulaire déterministe permettant d'optimiser la log-vraisemblance du modèle très efficacement. Ceci a permis d'établir un lien méthodologique entre le filtre particulaire et l'approche dite par *collapsing*, validant ainsi, pour la première fois dans la littérature économétrique, la pertinence de cette approche comme technique d'estimation. Il serait donc intéressant d'étudier cette approche d'un point de vue purement théorique et de déterminer si elle peut également être efficace dans d'autres contextes, tels que dans les modèles avec un espace d'états continu qui sont souvent très difficiles à estimer. Les deux méthodologies introduites ne se limitent pas aux modèles GARCH à changement de régimes et peuvent facilement être étendues à d'autres modèles à changement de régimes souffrant du problème de *path dependence*. Le modèle ARMA à changement de régimes introduit par Francq et Zakoïan [68] et le modèle ARMA-GARCH à changement de régimes présenté par Henneke et al. [102] sont deux exemples de tels modèles. Finalement, cette thèse a considéré le cas univarié et il serait donc pertinent de généraliser les méthodes proposées au cas multivarié.

Le deuxième volet de recherche de cette thèse tire sa motivation de la crise financière de la fin des années 2000 qui a résulté en de nombreux échecs institutionnels causés par une mauvaise évaluation des risques au sein de plusieurs compagnies financières. Le chapitre 4 a révélé que l'évaluation des risques dans les fonds distincts est fortement dépendante du choix du modèle économétrique. Bien que la littérature financière a historiquement mis peu d'emphase sur le risque de modèle, ce risque est très important puisque les *vraies* dynamiques des variables financières sont inconnues et en constante évolution. Par exemple, Morini [139] affirme : « The fact that thousands of technical papers speak of very advanced models, and just a handful focus on model risk and how to manage it, is one of our problems ». Par conséquent, il serait intéressant de développer des méthodologies pouvant incorporer ce risque dans l'évaluation et la gestion des risques financiers. Barrieu et Scandolo [15], Cont [42] et Embrechts et al. [61] sont des articles récents qui abordent cette problématique.

La chapitre 4 a également souligné l'importance de développer des stratégies de couverture efficaces. La littérature financière a dévoué beaucoup de recherche pour faire progresser les modèles économétriques dans le but d'améliorer la construction de telles stratégies, mais les approches permettant d'évaluer leur efficacité ont peu évolué. Le chapitre 5 a offert une contribution méthodologique dans ce domaine en proposant un cadre statistique, basé sur la régression, permettant de mieux mesurer cette efficacité. Cette contribution ouvre la voie à plusieurs analyses pouvant permettre d'évaluer la valeur ajoutée offerte par des stratégies de couverture dynamique dérivées à partir de modèles plus sophistiqués que celui de Black-Scholes. Finalement, ce même chapitre a révélé la difficulté de réduire à zéro le risque de couverture dans le cadre des fonds distincts compte tenu (i) de la présence du risque de modèle, (ii) de l'incertitude liée au comportement de l'assuré (*policyholder behavior*) et (iii) de l'absence d'un marché liquide pour les options sous-jacentes. Cette difficulté nous a amené à proposer une nouvelle perspective sur le rôle de la couverture dynamique, soit celle d'un outil permettant de modifier le profil risque–récompense du produit vendu. Cette approche dévie de l'approche traditionnelle en finance, dite risque-neutre, en incorporant le risque de couverture directement dans la tarification des produits dérivés. Cette nouvelle perspective constitue une avenue de recherche très intéressante dans les applications actuarielles.

Bibliographie

[1] Albert, J. H. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics*, 11(1) :1–15.

[2] Alexander, C. and Kaeck, A. (2012). Does model fit matter for hedging ? Evidence from FTSE 100 options. *Journal of Futures Markets*, 32(7) :609–638.

[3] Alexander, C., Kaeck, A., and Nogueira, L. M. (2009). Model risk adjusted hedge ratios. *Journal of Futures Markets*, 29(11) :1021–1049.

[4] Alexander, C. and Lazar, E. (2006). Normal mixture GARCH(1,1) : Applications to exchange rate modelling. *Journal of Applied Econometrics*, 21(3) :307–336.

[5] Alexander, C. and Nogueira, L. M. (2007). Model-free hedge ratios and scale-invariant models. *Journal of Banking & Finance*, 31(6) :1839–1861.

[6] Alexander, C., Rubinov, A., Kalepky, M., and Leontsinis, S. (2012). Regime-dependent smile-adjusted delta hedging. *Journal of Futures Markets*, 32(3) :203–229.

[7] American Academy of Actuaries (2005). Recommended approach for setting regulatory risk-based capital requirements for variable annuities and similar products. Available at `http://www.actuary.org/pdf/life/c3_june05.pdf`.

[8] Ané, T. and Ureche-Rangau, L. (2006). Stock market dynamics in a regime-switching asymmetric power GARCH model. *International Review of Financial Analysis*, 15(2) :109–129.

[9] Augustyniak, M. (2013). Maximum likelihood estimation of the Markov-switching GARCH model. *Computational Statistics & Data Analysis*. In press, doi :10.1016/j.csda.2013.01.026.

[10] Augustyniak, M. and Boudreault, M. (2012). An out-of-sample analysis of investment guarantees for equity-linked products : Lessons from the financial crisis of the late 2000s. *North American Actuarial Journal*, 16(2) :183–206.

[11] Bacinello, A. R. (2003). Pricing guaranteed life insurance participating policies with annual premiums and surrender option. *North American Actuarial*

Journal, 7(3) :1–17.

[12] Bacinello, A. R., Biffis, E., and Millossovich, P. (2009). Pricing life insurance contracts with early exercise features. *Journal of Computational and Applied Mathematics*, 233(1) :27–35.

[13] Bakshi, G., Cao, C., and Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, 52(5) :2003–2049.

[14] Bakshi, G., Cao, C., and Chen, Z. (2000). Pricing and hedging long-term options. *Journal of Econometrics*, 94(1–2) :277–318.

[15] Barrieu, P. and Scandolo, G. (2013). Assessing financial model risk. Working paper, Cornell University Library. Available at `http://arxiv.org/abs/1307.0684`.

[16] Bauwens, L., Dufays, A., and Rombouts, J. V. (2013). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*. In press, doi :10.1016/j.jeconom.2013.08.017.

[17] Bauwens, L., Hafner, C. M., and Rombouts, J. V. K. (2007). Multivariate mixed normal conditional heteroskedasticity. *Computational Statistics & Data Analysis*, 51(7) :3551–3566.

[18] Bauwens, L., Preminger, A., and Rombouts, J. V. K. (2010). Theory and inference for a Markov switching GARCH model. *The Econometrics Journal*, 13(2) :218–244.

[19] Billio, M. and Caporin, M. (2005). Multivariate Markov switching dynamic conditional correlation GARCH representations for contagion analysis. *Statistical Methods & Applications*, 14(2) :145–161.

[20] Billio, M., Casarin, R., and Osuntuyi, A. (2012). Efficient Gibbs sampling for Markov switching GARCH models. Working Papers 35/WP/2012, University of Venice Ca' Foscari, Department of Economics. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2198837`.

[21] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3) :307–327.

[22] Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69(3) :542–547.

[23] Bollerslev, T. (2010). Glossary to ARCH (GARCH). In Bollerslev, T., Russell, J., and Watson, M., editors, *Volatility and Time Series Econometrics : Essays in Honor of Robert Engle*. Oxford University Press, USA.

[24] Boudreault, M. and Panneton, C.-M. (2009). Multivariate models of equity returns for investment guarantees valuation. *North American Actuarial Journal*, 13(1) :36–53.

[25] Branger, N., Krautheim, E., Schlag, C., and Seeger, N. (2012). Hedging under model misspecification : All risk factors are equal, but some are more equal than others... *Journal of Futures Markets*, 32(5) :397–430.

[26] Broadie, M., Chernov, M., and Johannes, M. (2007). Model specification and risk premia : Evidence from futures options. *The Journal of Finance*, 62(3) :1453–1490.

[27] Broadie, M. and Detemple, J. (1995). American capped call options on dividend-paying assets. *The Review of Financial Studies*, 8(1) :161–191.

[28] Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67(2) :235–251.

[29] Cai, J. (1994). A Markov model of switching-regime ARCH. *Journal of Business & Economic Statistics*, 12(3) :309–316.

[30] Calvet, L. E. and Fisher, A. J. (2004). How to forecast long-run volatility : Regime switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, 2(1) :49–83.

[31] Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, New Jersey, second printing, with corrections edition.

[32] Canadian Institute of Actuaries (2002). Report of the CIA task force on segregated fund investment guarantees. Available at `http://www.actuaries.ca/members/publications/2002/202012e.pdf`.

[33] Cappé, O., Douc, R., Moulines, E., and Robert, C. (2002). On the convergence of the Monte Carlo maximum likelihood method for latent variable models. *Scandinavian Journal of Statistics. Theory and Applications*, 29(4) :615–635.

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

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

[36] Charnes, J. M., Koch, P., and Berkman, H. (2003). Measuring hedge effectiveness for FAS 133 compliance. *Journal of Applied Corporate Finance*, 15(4) :95–103.

[37] Chen, Z., Vetzal, K., and Forsyth, P. (2008). The effect of modelling parameters on the value of GMWB guarantees. *Insurance : Mathematics and Economics*, 43(1) :165–173.

[38] Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75(1) :79–97.

[39] Coleman, T. F., Kim, Y., Li, Y., and Patron, M. (2007). Robustly hedging variable annuities with guarantees under jumps and volatility risks. *The Journal of Risk and Insurance*, 74(2) :347–376.

[40] Coleman, T. F., Li, Y., and Patron, M.-C. (2006). Hedging guarantees in variable annuities under both equity and interest rate risks. *Insurance : Mathematics & Economics*, 38(2) :215–228.

[41] Cont, R. (2001). Empirical properties of asset returns : stylized facts and statistical issues. *Quantitative Finance*, 1(2) :223–236.

[42] Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance*, 16(3) :519–547.

[43] Creal, D. (2012). A survey of sequential Monte Carlo methods for economics and finance. *Econometric Reviews*, 31(3) :245–296.

[44] Dai, M., Kuen Kwok, Y., and Zong, J. (2008). Guaranteed minimum withdrawal benefit in variable annuities. *Mathematical Finance*, 18(4) :595–611.

[45] Danielsson, J. and Richard, J.-F. (1993). Accelerated Gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics*, 8(S1) :S153–S173.

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

[47] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39(1) :1–38.

[48] Diebold, F. X. (1986). Modeling the persistence of conditional variances : A comment. *Econometric Reviews*, 5(1) :51–56.

[49] Ding, Z., Granger, C. W. J., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1) :83–106.

[50] Douc, R., Cappé, O., and Moulines, E. (2005). Comparison of resampling schemes for particle filtering. In *4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69.

[51] Douc, R., Moulines, E., and Olsson, J. (2009). Optimality of the auxiliary particle filter. *Probability and Mathematical Statistics*, 29(1) :1–28.

[52] Doucet, A., Godsill, S. J., and Robert, C. P. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12(1) :77–84.

[53] Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing : Fifteen years later. In Crisan, D. and Rozovskii, B., editors, *The*

Oxford Handbook of Nonlinear Filtering, Oxford Handbooks in Mathematics, pages 656–704. Oxford University Press, New York.

[54] Duan, J.-C. (1997). Augmented GARCH(p, q) process and its diffusion limit. *Journal of Econometrics*, 79(1) :97–127.

[55] Duan, J.-C., Ritchken, P., and Sun, Z. (2006). Approximating GARCH-jump models, jump-diffusion processes, and option pricing. *Mathematical Finance*, 16(1) :21–52.

[56] Dueker, M. J. (1997). Markov switching in GARCH processes and mean-reverting stock-market volatility. *Journal of Business & Economic Statistics*, 15(1) :26–34.

[57] Dufays, A. (2012). Infinite-state Markov-switching for dynamic volatility and correlation models. CORE Discussion Paper 2012/43, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE). Available at `http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2012_43web.pdf`.

[58] Dumas, B., Fleming, J., and Whaley, R. E. (1998). Implied volatility functions : Empirical tests. *The Journal of Finance*, 53(6) :2059–2106.

[59] Eddelbuettel, D. and François, R. (2011). Rcpp : Seamless R and C++ integration. *Journal of Statistical Software*, 40(8) :1–18.

[60] Elliott, R. J. and Miao, H. (2009). VaR and expected shortfall : a non-normal regime switching framework. *Quantitative Finance*, 9(6) :747–755.

[61] Embrechts, P., Puccetti, G., and Rüschendorf, L. (2013). Model uncertainty and VaR aggregation. *Journal of Banking & Finance*, 37(8) :2750–2764.

[62] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4) :987–1007.

[63] Eraker, B., Johannes, M., and Polson, N. (2003). The impact of jumps in volatility and returns. *The Journal of Finance*, 58(3) :1269–1300.

[64] Fearnhead, P. and Clifford, P. (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 65(4) :887–899.

[65] Fernandez, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441) :359–371.

[66] Finnerty, J. D. and Grant, D. (2002). Alternative approaches to testing hedge effectiveness under SFAS No. 133. *Accounting Horizons*, 16(2) :95–108.

[67] Francq, C., Roussignol, M., and Zakoïan, J.-M. (2001). Conditional heteroskedasticity driven by hidden Markov chains. *Journal of Time Series Analysis*, 22(2) :197–220.

[68] Francq, C. and Zakoïan, J.-M. (2001). Stationarity of multivariate Markov-switching ARMA models. *Journal of Econometrics*, 102(2) :339–364.

[69] Francq, C. and Zakoïan, J.-M. (2005). The L^2-structures of standard and switching-regime GARCH models. *Stochastic Processes and their Applications*, 115(9) :1557–1582.

[70] Francq, C. and Zakoïan, J.-M. (2008). Deriving the autocovariances of powers of Markov-switching GARCH models, with applications to statistical inference. *Computational Statistics & Data Analysis*, 52(6) :3027–3046.

[71] Francq, C. and Zakoïan, J.-M. (2010). *GARCH Models : Structure, Statistical Inference and Financial Applications*. John Wiley & Sons, Ltd, West Sussex.

[72] Freeland, R. K., Till, M. C., and Hardy, M. R. (2009). Residual based assessment of regime switching models. Working paper, University of Waterloo.

[73] Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2) :183–202.

[74] Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York.

[75] Gaetan, C. and Yao, J.-F. (2003). A multiple-imputation Metropolis version of the EM algorithm. *Biometrika*, 90(3) :643–654.

[76] Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B. Methodological*, 56(1) :261–274.

[77] Geyer, C. J. (1996). Estimation and optimization of functions. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice*, pages 241–258. Chapman & Hall, London.

[78] Ghalanos, A. (2013). *rugarch : Univariate GARCH models.* R package version 1.0-16.

[79] Ghalanos, A. and Theussl, S. (2012). *Rsolnp : General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.14.

[80] Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5) :1779–1801.

[81] Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1(1) :3–15.

[82] Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F, Radar and Signal Processing*, 140(2) :107–113.

[83] Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1) :27–62.

[84] Haas, M. (2010). Skew-normal mixture and Markov-switching GARCH processes. *Studies in Nonlinear Dynamics & Econometrics*, 14(4) :Article 1.

[85] Haas, M., Mittnik, S., and Paolella, M. S. (2004a). Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics*, 2(2) :211–250.

[86] Haas, M., Mittnik, S., and Paolella, M. S. (2004b). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics*, 2(4) :493–530.

[87] Haas, M., Mittnik, S., and Paolella, M. S. (2009). Asymmetric multivariate normal mixture GARCH. *Computational Statistics & Data Analysis*, 53(6) :2129–2154.

[88] Hagan, P. S., Kumar, D., Lesniewski, A. S., and Woodward, D. E. (2002). Managing smile risk. *Wilmott magazine*, pages 84–108.

[89] Hailer, A. C. and Rump, S. M. (2005). Evaluation of hedge effectiveness tests. *Journal of Derivatives Accounting*, 2(1) :31–51.

[90] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2) :357–384.

[91] Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press, New Jersey.

[92] Hamilton, J. D. (2008). Regime switching models. In Durlauf, S. N. and Blume, L. E., editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke.

[93] Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1-2) :307–333.

[94] Hardy, M. R. (2001). A regime-switching model of long-term stock returns. *North American Actuarial Journal*, 5(2) :41–53.

[95] Hardy, M. R. (2002). Bayesian risk management for equity-linked insurance. *Scandinavian Actuarial Journal*, 2002(3) :185–211.

[96] Hardy, M. R. (2003). *Investment guarantees : Modeling and risk management for equity-linked life insurance*. John Wiley & Sons, New Jersey.

[97] Hardy, M. R., Freeland, R. K., and Till, M. C. (2006). Validation of long-term equity return models for equity-linked guarantees. *North American Actuarial Journal*, 10(4) :28–47.

[98] Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society. Series B. Methodological*, 38(3) :205–247.

[99] Harvey, A., Ruiz, E., and Sentana, E. (1992). Unobserved component time series models with ARCH disturbances. *Journal of Econometrics*, 52(1–2) :129–157.

[100] Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2) :795–800.

[101] He, C., Silvennoinen, A., and Teräsvirta, T. (2008). Parameterizing unconditional skewness in models for financial time series. *Journal of Financial Econometrics*, 6(2) :208–230.

[102] Henneke, J., Rachev, S., Fabozzi, F., and Nikolov, M. (2011). MCMC-based estimation of Markov switching ARMA-GARCH models. *Applied Economics*, 43 :259–271.

[103] Henry, O. T. (2009). Regime switching in the relationship between equity returns and short-term interest rates in the UK. *Journal of Banking & Finance*, 33(2) :405–414.

[104] Hentschel, L. (1995). All in the family : Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, 39(1) :71–104.

[105] Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2) :327–343.

[106] Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. *Journal of Econometrics*, 129(1-2) :121–138.

[107] Ingrassia, S. and Rocci, R. (2007). Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis*, 51(11) :5339–5351.

[108] Jacquier, E., Johannes, M., and Polson, N. (2007). MCMC maximum likelihood for latent state models. *Journal of Econometrics*, 137(2) :615–640.

[109] Johansen, A. M. and Doucet, A. (2008). A note on auxiliary particle filters. *Statistics & Probability Letters*, 78(12) :1498–1504.

[110] Johansen, A. M., Doucet, A., and Davy, M. (2008). Particle methods for maximum likelihood estimation in latent variable models. *Statistics and Computing*, 18(1) :47–57.

[111] Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16(4) :312–334.

[112] Kaeck, A. (2012). Hedging surprises, jumps, and model misspecification : A risk management perspective on hedging S&P 500 options. *Review of Finance*. In press, doi :10.1093/rof/rfs027.

[113] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering. Series D*, 82 :35–45.

[114] Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME—Journal of Basic Engineering. Series D*, 83 :95–108.

[115] Kantas, N., Doucet, A., Singh, S. S., and Maciejowski, J. M. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state space models. In *15th IFAC Symposium on System Identification*, Saint-Malo, France.

[116] Kiefer, N. M. (1978). Discrete parameter variation : efficient estimation of a switching regression model. *Econometrica*, 46(2) :427–434.

[117] Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2) :1–22.

[118] Kim, C.-J. and Nelson, C. R. (1999). *State-Space Models with Regime Switching : Classical and Gibbs-Sampling Approaches with Applications*. The MIT Press, Massachusetts.

[119] Kim, I. J. and Kim, S. (2004). Empirical comparison of alternative stochastic volatility option pricing models : Evidence from Korean KOSPI 200 index options market. *Pacific-Basin Finance Journal*, 12(2) :117–142.

[120] Klaassen, F. (2002). Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Economics*, 27 :363–394.

[121] Kling, A., Ruez, F., and Ruß, J. (2011a). The impact of policyholder behavior on pricing, hedging and hedge efficiency of withdrawal benefit guarantees in variable annuities. Working paper. Available at `http://www.actuaris.com/site/pdf/4e8dd713e78a2.pdf`.

[122] Kling, A., Ruez, F., and Ruß, J. (2011b). The impact of stochastic volatility on pricing, hedging and hedge efficiency of withdrawal benefit guarantees in variable annuities. *Astin Bulletin*, 41(2) :511–545.

[123] Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2008). *Loss models*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., New Jersey, third edition. From data to decisions.

[124] Knoller, C., Kraut, G., and Schoenmaekers, P. (2012). On the propensity to surrender a variable annuity contract—An empirical analysis of dynamic policyholder behaviour. Working Paper 7, Munich Risk and Insurance Center. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1976203`.

[125] Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business & Economic Statistics*, 8(2) :225–234.

[126] Levine, R. A. (2005). A note on Markov chain Monte Carlo sweep strategies. *Journal of Statistical Computation and Simulation*, 75(4) :253–262.

[127] Levine, R. A. and Fan, J. (2004). An automated (Markov chain) Monte Carlo EM algorithm. *Journal of Statistical Computation and Simulation*, 74(5) :349–359.

[128] Lin, G. (1998). *Nesting regime-switching GARCH models and stock market volatility, returns and the business cycle*. PhD thesis, University of California, San Diego.

[129] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 44(2) :226–233.

[130] Malik, S. and Pitt, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, 165(2) :190–209.

[131] Marcucci, J. (2005). Forecasting stock market volatility with regime-switching GARCH models. *Studies in Nonlinear Dynamics & Econometrics*, 9(4) :Article 6.

[132] McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437) :162–170.

[133] McDonald, R. L. (2006). *Derivatives markets*. The Addison-Wesley Series in Finance. Addison Wesley, Massachusetts, second edition.

[134] McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience, New Jersey, second edition.

[135] Mikosch, T. and Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *The Review of Economics and Statistics*, 86(1) :378–390.

[136] Milevsky, M. A. and Salisbury, T. S. (2001). The real option to lapse a variable annuity : Can surrender charges complete the market ? In *Proceedings of the 11th Annual Actuarial Approach for Financial Risks International Colloquium*, volume 2.

[137] Milevsky, M. A. and Salisbury, T. S. (2006). Financial valuation of guaranteed minimum withdrawal benefits. *Insurance : Mathematics & Economics*, 38(1) :21–38.

[138] Milliman (2011). Variable annuity dynamic lapse study : A data mining approach. Research report, Milliman. Available at `http://insight.milliman.com/article.php?cntid=7656`.

[139] Morini, M. (2011). *Understanding and managing model risk : A practical guide for quants, traders and validators*. John Wiley & Sons, Ltd., West Sussex.

[140] Nandi, S. (1998). How important is the correlation between returns and volatility in a stochastic volatility model ? Empirical evidence from pricing and hedging in the S&P 500 index options market. *Journal of Banking & Finance*, 22(5) :589–610.

[141] Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory*, 6(3) :318–334.

[142] Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns : A new approach. *Econometrica*, 59(2) :347–370.

[143] O'Connell, J. and Højsgaard, S. (2011). Hidden semi Markov models for multiple observation sequences : The mhsmm package for R. *Journal of Statistical Software*, 39(4) :1–22.

[144] Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1) :155–179.

[145] Pitt, M. K. (2002). Smooth particle filters for likelihood evaluation and maximisation. Warwick Economic Research Papers 651, Department of Economics, University of Warwick. Available at `http://wrap.warwick.ac.uk/1536/1/WRAP_Pitt_twerp651.pdf`.

[146] Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2) :134–151.

[147] Pitt, M. K. and Shephard, N. (1999). Filtering via simulation : Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446) :590–599.

[148] Poulsen, R., Schenk-Hoppé, K. R., and Ewald, C.-O. (2009). Risk minimization in stochastic volatility models : model risk and empirical performance. *Quantitative Finance*, 9(6) :693–704.

[149] Psychoyios, D. and Skiadopoulos, G. (2006). Volatility options : Hedging effectiveness, pricing, and model error. *Journal of Futures Markets*, 26(1) :1–31.

[150] R Core Team (2013). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[151] Rantala, J. (2006). On joint and separate history of probability, statistics and actuarial science. In *Festschrift for Tarmo Pukkila on his 60th birthday*, pages 261–284. Dep. Math. Stat. Philos. Univ. Tampere, Tampere.

[152] Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

[153] Robert, C. P., Celeux, G., and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains : A stochastic implementation. *Statistics & Probability Letters*, 16(1) :77–83.

[154] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23 :470–472.

[155] Sajjad, R., Coakley, J., and Nankervis, J. C. (2008). Markov-switching GARCH modelling of Value-at-Risk. *Studies in Nonlinear Dynamics & Econometrics*, 12(3) :Article 7.

[156] Shumway, R. H. and Stoffer, D. S. (2011). *Time series analysis and its applications*. Springer Texts in Statistics. Springer, New York, third edition. With R examples.

[157] Society of Actuaries (2012). Policyholder behavior in the tail : Variable annuity guaranteed benefits—2011 survey results. Research report, Society of Actuaries. Available at `http://soa.org/files/research/projects/research-policy-behavior-tail-result-report.pdf`.

[158] Tadić, V. B. and Doucet, A. (2005). Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models. *Stochastic Processes and their Applications*, 115(8) :1408–1436.

[159] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398) :528–540.

[160] Tsay, R. S. (2010). *Analysis of financial time series*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., New Jersey, third edition.

[161] Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411) :699–704.

[162] Wong, A. C. S. and Chan, W.-S. (2005). Mixture Gaussian time series modeling of long-term market returns. *North American Actuarial Journal*, 9(4) :83–94.

[163] Wu, C.-F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1) :95–103.

[164] Wuertz, D. (2010). *fBasics : Rmetrics—Markets and Basic Statistics*. R package version 2160.85. With Rmetrics core team members.

[165] Wuertz, D. and Chalabi, Y. (2012). *fGarch : Rmetrics—Autoregressive Conditional Heteroskedastic Modelling*. R package version 2150.81. With contribution from Michal Miklovic, Chris Boudt, Pierre Chausse and others.

[166] Yao, W. (2011). Label switching and its simple solutions for frequentist mixture models. Manuscript submitted for publication. Available at `http://www-personal.ksu.edu/~wxyao/research.html`.

[167] Ye, Y. (1987). *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Stanford University, California.

[168] Zakoïan, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, 18(5) :931–955.

[169] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov models for time series : an introduction using R*, volume 110 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Florida.