Factor Analysed Hidden Markov Models for Conditionally Heteroscedastic Financial Time Series

Christian Lavergne, Mohamed Saidane

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Factor Analysed Hidden Markov Models for Conditionally Heteroscedastic Financial Time Series

Christian Lavergne — Mohamed Saidane

N° 5862

March 2006

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--5862--FR+ENG

Factor Analysed Hidden Markov Models for Conditionally Heteroscedastic Financial Time Series

Christian Lavergne, Mohamed Saidane

Thème COG — Systèmes cognitifs
Projet MISTIS

Rapport de recherche n° 5862 — March 2006 — 48 pages

Abstract: In this article we develop a new approach within the framework of asset pricing models that incorporates two key features of the latent volatility: co-movement among conditionally heteroscedastic financial returns and switching between different unobservable regimes. By combining latent factor models with hidden Markov chain models (HMM) we derive a dynamical local model for segmentation and prediction of multivariate conditionally heteroscedastic financial time series. We concentrate, more precisely on situations where the factor variances are modeled by univariate GQARCH processes. The intuition behind our approach is the use a piece-wise multivariate and linear process – which we can also regard as a mixed-state dynamic linear system – for modeling the regime switches. In particular, we supposed that the observed series can be modeled using a time varying parameter model with the assumption that the evolution of these parameters is governed by a first-order hidden Markov process. The EM algorithm that we have developed for the maximum likelihood estimation, is based on a quasi-optimal switching Kalman filter approach combined with a Viterbi approximation which yield inferences about the unobservable path of the common factors, their variances and the latent variable of the state process. Extensive Monte Carlo simulations and preliminary experiments obtained with daily foreign exchange rate returns of eight currencies show promising results.

Key-words: Dynamic Factor Analysis, GQARCH Processes, HMM, EM Algorithm, Switching Kalman filter, Viterbi Approximation, Finance

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

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

Résumé : Dans cet article nous proposons une nouvelle approche dans le cadre des modèles d'évaluation des actifs financiers permettant de tenir compte de deux aspects fondamentaux qui caractérisent la volatilité latente: co-mouvement des rendements financiers conditionnellement hétéroscédastiques et transition entre différents régimes inobservables. En combinant les modèles à facteurs latents avec les modèles de chaîne de Markov cachés (HMM) nous dérivons un modèle multivarié localement linéaire et dynamique pour la segmentation et la prévision des séries financières conditionnellement hétéroscédastiques. Nous nous concentrons, plus précisément sur le cas où les facteurs communs suivent des processus GQARCH univariés. L'idée originale de ce travail est la modélisation de cette non stationnarité à l'aide d'un processus multivarié et linéaire par morceaux que l'on peut considérer aussi comme un système linéaire et dynamique à états mixtes. En particulier, nous avons supposé que les séries observées peuvent être approchées à l'aide d'un modèle dont les paramètres évoluent au cours du temps. Nous avons émis, aussi, l'hypothèse que l'évolution de ces paramètres est gouvernée par une variable inobservable que l'on peut modéliser à l'aide d'un processus Markovien caché d'ordre un. L'algorithme EM que nous avons développé pour l'estimation de maximum de vraisemblance et l'inférence des structures cachées est basé sur une version quasi-optimale du filtre de Kalman combinée avec une approximation de viterbi. Les résultats obtenus sur des simulations, aussi bien que sur des séries de rendements de taux de change de huit pays sont prometteurs.

Mots-clés : Modèles à Facteurs Dynamiques, Processus GQARCH, Modèles HMM, Algorithme EM, Algorithme de Viterbi, Filtrage de Kalman, Finance

1 Introduction

In the recent two decades, the multifactor analysis has become more and more attractive in the economic literature. In a factor model, the dynamics of multivariate time series can be parsimoniously determined by a small number of factors. Factor analysis was shown useful to understanding the dynamics of financial markets, macroeconomic business cycles and the structure of the consumer demand system. It has been used in finance and econometrics as an alternative to the Capital Asset Pricing Model (CAPM) since the early 1960s. In this context, factor models have been used as a parsimonious means of describing the covariance matrix of returns since the single-index model of Sharpe [43]. The arbitrage pricing theory (APT) of Ross [35, 36] assumes that asset returns follow a linear factor model in which factors are latent (unobservable) and both factor and idiosyncratic shocks are independent and identically distributed. The central empirical implication of the APT is that risk premia should be linear functions of loadings on systematic factors (see Connor and Korajczyk [10] for a review).

Traditionally, these issues were considered in a static framework, but recently, the emphasis has shifted toward inter-temporal asset pricing models in which agents decisions are based on the distribution of returns conditional on the available information, which is obviously changing. This is partly motivated by the fact that financial markets volatility changes over time. However, it was not until Engle's [14] work on Autoregressive Conditional Heteroscedasticity (ARCH) and Bollerslev's [5] Generalized ARCH (GARCH) that researchers were able to take into account the time variation in first and second moments of returns. For comprehensive surveys of the models in the ARCH family, one can refer to Bollerslev, Chou, and Kroner [7] and Bollerslev, Engle, and Nelson [8]. In parallel with these theoretical developments, numerous applications have appeared. By and large, though, most applied work pertains to univariate financial time series, as the application of these models in a multivariate context has been hampered by the large number of parameters involved. To avoid this problem, Diebold and Nerlove [13] propose a multivariate approach based on the same idea as traditional (i.e. conditionally homoscedastic) factor analysis. That is, it is assumed that each of q observed variables is a linear combination of k ($k < q$) common factors plus an idiosyncratic term, but allowing for ARCH type effects in the underlying factors.

Several researchers have used Factor-ARCH models to provide a plausible and parsimonious parametrization of the time varying variance-covariance structure of asset returns. Engle, Ng and Rothschild [15] apply such structures to model the pricing of Treasury bills. A similar model is used by Engle and Ng [16] to study the dynamic behavior of the term structure of interest rates. Diebold and Nerlove [13] use a latent factor ARCH model to describe the dynamics of exchange rate volatility. Engle and Susmel [17] use the factor ARCH to test for common volatility in international equity markets. Recently, an EM algorithm has been proposed by Demos and Sentana [11] to approximate the maximum likelihood estimates in the presence of ARCH effects in the common factors. Alternative estimation procedures for such models are, also, investigated by Lin [29] on the basis of Monte Carlo comparisons.

An assumption of these models is that the relationships between variables has not changed over time, but recent empirical works have shown that this assumption of structural stability is invalid for many financial and economic data sets. For example, the break in volatility in the United States, documented by McConnell and Perez-Quiros [31], suggests a change in the nature of US business cycles. In general, financial time series are often characterized by a non constant (conditional or unconditional) volatility. The volatility shocks tend to persist through time and can affect certain fundamental behaviors of the financial returns, such as the leptokurtic aspect and the mean reverting phenomenon. For example, Chou [9] reports that persistence of shocks to the stock-market volatility was high in the U.S. stock-market during 1962-1985. Similar results has been observed in the work of French, Schwert, and Stambaugh [18]. Lamoureux and Lastrapes [27] suggest that the apparent persistence of the variance may be overestimated because the possible structural shift in the model had not been taken into account. In the same article, the authors point out that means of identifying occasional switching in the parameter values, like the Markov switching model of Hamilton [22], may provide more appropriate modeling of volatility.

This paper extends the different models proposed in the above literature to a multi-state model by allowing for model transitions that are governed by a Markov chain on a set of possible models describing the different states of volatility. The main feature that characterizes this model is that the unobserved state variables are not real physical variables that happen to be missing. Instead, these variables represent underlying factors without precise physical definition, but which often, and desirably so, turn to have a meaningful physical interpretation. More specifically this model is based on the assumption that the data generating process changes over time, and there is a latent model selection procedure dependent on a discret state variable which randomly picks a parametric model each time. This procedure is characterized by defining a set or a subset of the model parameters to be mutually dependent on the state variable.

This article is organized as follows. In section 2, we introduce the general form of the model in its simplest version. It is in fact a standard factor analysis model combined with a first-order hidden Markov process. Using its likelihood function we estimate its parameters by using an exact EM algorithm inspired by the Baum and Welch algorithm for HMMs. In the third section we extend the standard model to study the co-movements of financial time series characterized by a dynamic heteroscedasticity in the variances. In section 4 we study it in a switching space-state structure, in order to estimate common factors by using an extended version of the Kalman filter based on the "moment matching" technique. The complete likelihood function and the conditional EM algorithm are presented in section 5, where we discuss with much more details the estimate of the parameters of the conditionally heteroscedastic component based on the restoration of the discrete and continuous hidden states by using, either posterior probabilities already provided by the smoothing algorithm, or an approximated version of the Viterbi algorithm. We also present an alternative approach for inference about the latent structures and estimation of the parameters of these models, based on a Viterbi approximation. In the last section we study and evaluate the performance of the maximum likelihood approach on both synthetic and financial data.

2 Factor Analysed Hidden Markov Models

The factor analysed hidden Markov model (FAHMM) is a dynamic state-space generalization of a multiple component factor analysis system. The k-dimensional state vectors are generated by a standard diagonal covariance Gaussian HMM. The q-dimensional observation vectors are generated by a multiple noise component factor analysis observation process. A generative model for FAHMM can be described as follows:

$$\boxed{\begin{array}{c} S_t \sim P(S_t = j/S_{t-1} = i) \\[1em] \text{for } \; t = 1, ..., n \; \text{ and } \; i, j = 1, ..., m \\[1em] \mathbf{y}_t = \mathbf{X}_{s_t}\mathbf{f}_t + \varepsilon_t \quad \text{with} \quad \left\{ \begin{array}{c} \varepsilon_t \sim \mathcal{N}\left(\theta_{s_t}, \boldsymbol{\Psi}_{s_t}\right) \\[1em] \mathbf{f}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{H}_{s_t}\right) \end{array} \right. \end{array}}$$

where $S_t \sim P(S_t = j/S_{t-1} = i)$ [1] is an homogenous hidden Markov chain indicating the state or the regime at the date t, and $\mathbf{y}_t$ is a $(q \times 1)$ random vector of observable variables. The HMM state transition probabilities from state i to state j are represented by p_{ij}. In an unspecified state $S_t = j$ $(j = 1, ..., m)$, $\mathbf{0}$ and $\mathbf{H}_j$ are, respectively, the $(k \times 1)$ mean vectors and $(k \times k)$ diagonal and definite-positive covariance matrices of the latent common factors $\mathbf{f}_t$; θ_j and $\boldsymbol{\Psi}_j$ are, respectively, the $(q \times 1)$ mean vectors and $(q \times q)$ diagonal and definite-positive covariance matrices of the $(q \times 1)$ vectors of idiosyncratic noises ε_t; $\mathbf{X}_j$ are the $(q \times k)$ factor loadings matrices, with $q \geq k$ and $rank(\mathbf{X}_j) = k$. Here we suppose that the common and specific (idiosyncratic) factors are uncorrelated. We suppose also that $\mathbf{f}_t$ and $\varepsilon_{t'}$ are mutually independent for all t, t'.

A dynamic bayesian network describing a FAHMM is shown in figure 1. The square nodes represent discrete random variables such as the HMM state $\{S_t\}$. Continuous random variables such as the state vectors, $\mathbf{f}_t$, are represented by round nodes. Shaded nodes depict observable variables, $\mathbf{y}_t$, leaving all the other FAHMM variables hidden. A conditional independence assumption is made between variables that are not connected by directed arcs. The state conditional independence assumption between the output densities of a standard HMM is also used in a FAHMM.

An important aspect of any generative model is the complexity of the likelihood calculations. The generative model above can be expressed by the two following Gaussian distributions:

$$\begin{aligned} p(\mathbf{f}_t/S_t = j) &= \mathcal{N}\left(\mathbf{0}, \mathbf{H}_j\right) \\ p(\mathbf{y}_t/\mathbf{f}_t, S_t = j) &= \mathcal{N}\left(\theta_j + \mathbf{X}_j\mathbf{f}_t, \boldsymbol{\Psi}_j\right) \end{aligned}$$

[1] The $\sim$ symbol in $S_t \sim P(S_t/S_{t-1})$ is used to represent a discrete Markov chain. Normally it means the variable on the left hand side is distributed according to the probability density function on the right hand side as in $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{H}_t)$.

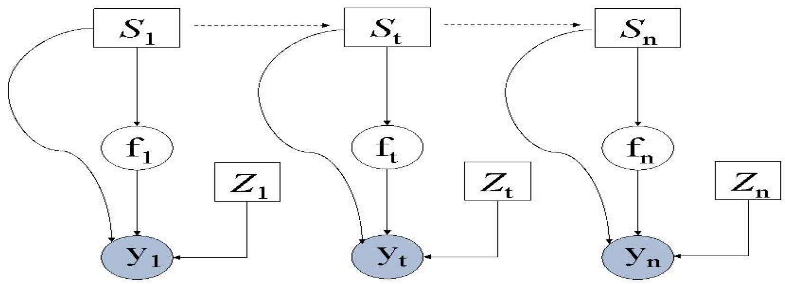

Figure 1: Dynamic Bayesian network representing a factor analysed hidden Markov model. $\mathbf{Z}_t$'s are eventual exogenous variables that can be introduced in the model as explanatory variables.

The likelihood of an observation $\mathbf{y}_t$ given the state $S_t = j$ can be obtained by integrating the state vector $\mathbf{f}_t$ out of the product of the above Gaussians. The resulting likelihood is also a Gaussian and can be written as:

$$b_j(\mathbf{y}_t) = p(\mathbf{y}_t / S_t = j) = \mathcal{N}\left(\theta_j, \mathbf{\Sigma}_j\right)$$

where $\mathbf{\Sigma}_j = \mathbf{X}_j \mathbf{H}_j \mathbf{X}_j' + \mathbf{\Psi}_j$. The likelihood calculation requires inverting m full $(q \times q)$ covariance matrices. If the amount of memory is not an issue, the inverses and the corresponding determinants for all the discrete states in the system can be computed prior to starting off with the training and recognition. However, this can rapidly become impractical for a large system. A more memory efficient implementation requires the computation of the inverses and determinants for each time instant. These can be efficiently obtained using the following equality for matrix inverses:

$$\left[\mathbf{X}_j \mathbf{H}_j \mathbf{X}_j' + \mathbf{\Psi}_j\right]^{-1} = \mathbf{\Psi}_j^{-1} - \mathbf{\Psi}_j^{-1} \mathbf{X}_j \left[\mathbf{X}_j' \mathbf{\Psi}_j^{-1} \mathbf{X}_j + \mathbf{H}_j^{-1}\right]^{-1} \mathbf{X}_j' \mathbf{\Psi}_j^{-1}$$

2.1 Optimizing FAHMM Parameters

The maximum likelihood (ML) criterion may be used to optimize the FAHMM parameters. It is also possible to find a discriminative training scheme such as minimum classification error (Saul and Rahim [38]). However, in the present work only ML training is considered. In conjunction with standard HMM training, we use the expectation maximization (EM) algorithm. For a sequence of observation vectors $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$, a sequence of continuous state vectors $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ and a sequence of discrete HMM states $\mathcal{S} = \{S_1, S_2, ..., S_n\}$, the complete likelihood function can be written as:

$$p(\mathcal{Y},\mathcal{F},\mathcal{S}) = p(S_1)\prod_{t=2}^{n} p(S_t/S_{t-1})\prod_{t=1}^{n} p(\mathbf{f}_t/S_t)p(\mathbf{y}_t/\mathbf{f}_t,S_t)$$

where $p(S_1) = \pi_{s_1}$ and $p(S_t/S_{t-1}) = p_{s_{t-1}s_t}$ are the initial state and the discrete state transition probabilities. The auxiliary function that will be maximized is given by:

$$\mathcal{Q}(\Theta,\Theta^{(i)}) = \mathbb{E}\left[\log p(\mathcal{Y},\mathcal{F},\mathcal{S}/\Theta^{(i)})/\mathcal{Y},\Theta\right]$$

here the set of current model parameters is represented by $\Theta^{(i)}$. A set of parameters, $\widehat{\Theta}$, that maximize the auxiliary function is found during the maximization step: $\widehat{\Theta} = \arg\max_{\Theta} \mathcal{Q}(\Theta,\widehat{\Theta})$. These parameters will be used as the set of old parameters in the following iteration, $\widehat{\Theta} \longrightarrow \Theta^{(i+1)}$.

For the M step, only the first and second-order statistics of $p(\mathcal{F}/\mathcal{Y},\mathcal{S},\Theta)$ are required since this distribution is conditionally Gaussian given the state. Sufficient statistics for $p(\mathcal{S}/\mathcal{Y},\Theta)$ can be obtained using the forward-backward algorithm described below.

2.1.1 Forward-Backward Algorithm

The likelihood of being in discrete state j and the observations up to time instant t is represented by the forward variable, $\alpha_j(t) = p(S_t = j,\, \mathcal{Y}_{1:t})$. Assuming that the first observation is generated by the first discrete state, the forward variable is initialized as:

$$\begin{cases} b_1(\mathbf{y}_1) & , \quad j = 1 \\ 0 & , \quad j \neq 1 \end{cases}$$

Using the conditional independence assumption in HMMs, the forward variable at time instant t is defined by the following recursion:

$$\alpha_j(t) = p(S_t = j,\mathcal{Y}_{1:t}) = b_j(\mathbf{y}_t)\sum_{i=1}^{m} p_{ij}\alpha_i(t-1)$$

The likelihood of the observations from $t+1$ to n given being in state i at time instant t is represented by the backward variable, $\beta_i(t) = p(\mathcal{Y}_{t+1:n}/S_t = i)$. This backward variable is initialized as $\beta_i(n) = 1$ for all $i \in [1,m]$. Using the same conditional independence assumptions, the backward variable at time instant $t-1$ is defined by the following recursion:

$$\beta_i(t-1) = p(\mathcal{Y}_{t:n}/S_{t-1} = i) = \sum_{j=1}^{m} p_{ij}b_j(\mathbf{y}_t)\beta_j(t)$$

The likelihood of the observation sequence, $\mathcal{Y}$, can be represented in terms of the forward and backward variables as follows:

$$p(\mathcal{Y}) = \sum_{i=1}^{m} p(S_t = i, \mathcal{Y}_{1:t}) p(\mathcal{Y}_{t+1:n}/S_t = i) = \sum_{i=1}^{m} \alpha_i(t)\beta_i(t)$$

The probability of being in state j at time t given the observation sequence is needed in the parameter update formulae. This likelihood can be expressed in terms of the forward and backward variables as follows:

$$\gamma_j(t) = p(S_t = j/\mathcal{Y}) = \frac{\alpha_j(t)\beta_j(t)}{\sum\limits_{i=1}^{m} \alpha_i(t)\beta_i(t)}$$

The joint probability of being in state i at time instant $t-1$ and in state j at time instant t given the observation sequence is needed in the transition parameter update formulae. This likelihood can be expressed in terms of the forward and backward variables as follows:

$$\xi_{ij}(t) = p(S_{t-1} = i, S_t = j/\mathcal{Y}) = \frac{\alpha_i(t-1)p_{ij}b_j(\mathbf{y}_t)\beta_j(t)}{\sum\limits_{i=1}^{m} \alpha_i(t)\beta_i(t)}$$

2.1.2 Continuous State Posterior Statistics

Given the current discrete state, $S_t = j$, the joint likelihood of the current observation and continuous state vector is Gaussian:

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{f}_t \end{pmatrix} /S_t = j \sim \mathcal{N}\left[\begin{pmatrix} \theta_j \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}_j\mathbf{H}_j\mathbf{X}'_j + \mathbf{\Psi}_j & \mathbf{X}_j\mathbf{H}_j \\ \mathbf{H}_j\mathbf{X}'_j & \mathbf{H}_j \end{pmatrix} \right]$$

The posterior distribution is also Gaussian and can be written as:

$$p(\mathbf{f}_t/\mathbf{y}_t, S_t = j) = \mathcal{N}\left[\mathbf{K}_j(\mathbf{y}_t - \theta_j), \mathbf{H}_j - \mathbf{K}_j\mathbf{X}_j\mathbf{H}_j\right]$$

where $\mathbf{K}_j = \mathbf{H}_j\mathbf{X}'_j\left[\mathbf{X}_j\mathbf{H}_j\mathbf{X}'_j + \mathbf{\Psi}_j\right]^{-1}$. For parameter update formulae, the statistics $\widetilde{\mathbf{f}}_{jt} = \mathbf{K}_j\left(\mathbf{y}_t - \theta_j\right)$ and $\widetilde{\mathbf{R}}_j = \mathbf{H}_j - \mathbf{K}_j\mathbf{X}_j\mathbf{H}_j$ are also needed.

3 Conditionally Heteroscedastic FAHMMs

Our empirical model consists of:

- a hidden Markov structure for the model parameters in order to take into account different states of the world that can affect the evolution of the time series. In this case, the dynamic properties of the different series depend on the present regime, with the regimes being realizations of a hidden Markov chain with a finite state space.

- a linear factor model with constant regime parameters for excess returns, and
- univariate Generalized Quadratic Autoregressive Conditionally Heteroscedastic processes (GQARCH) for modeling the time-varying volatility clustering phenomenon of the common latent factors.

Let $\mathbf{y}_t$ denote the q-vector of excess asset returns and $\mathbf{f}_t$ denote the k-vector of latent factor shocks in period t. In our switching factor model, the realized excess return on an asset is the sum of its expected return, k systematic shocks and an idiosyncratic shock. In matrix notation, the mixed-state factor model for the excess return vector is

$$\boxed{\begin{array}{c} S_t \sim P(S_t = j / S_{t-1} = i) \\ t = 1, ..., n \quad \text{and} \quad i, j = 1, ..., m \\ \mathbf{f}_{s_t} = \mathbf{H}_{s_t}^{1/2} \mathbf{f}_t^* \qquad \text{where} \qquad \mathbf{f}_t^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k) \\ \mathbf{y}_t = \mathbf{X}_{s_t} \mathbf{f}_{s_t} + \varepsilon_{s_t} \quad \text{with} \quad \varepsilon_{s_t} \sim \mathcal{N}(\theta_{s_t}, \mathbf{\Psi}_{s_t}) \end{array}}$$

The same notations are used in this extended specification. $\mathbf{y}_t$ is always a $(q \times 1)$ random vector of observable variables. However, in this framework the common variances (diagonal elements of $\mathbf{H}_{jt}$) are supposed to be time varying and their parameters change according to the regime. In particular, we suppose that these variances follow switching GQARCH(1,1) processes, the *l-th* diagonal element of the matrix $\mathbf{H}_{jt}$ under a particular regime $S_t = j$ since $S_{t-1} = i$ is given by:

$$h_{lt}^{(j)} = w_j^l + \gamma_j^l f_{lt-1}^{(i)} + \alpha_j^l f_{lt-1}^{(i)\,2} + \delta_j^l h_{lt-1}^{(i)} \quad \text{for} \quad l = 1, ..., k$$

where the dynamic asymmetry parameter γ_j^l is usually different from 0, allowing for the possibility of a leverage effect (see Sentana [40]). We can see from this model that if $f_{lt-1} > 0$, its impact on h_{lt} is greater than in the case of $f_{lt-1} < 0$ (assuming that γ_j^l and α_j^l are positive). The special case of $\gamma_j^l = 0$ gives the GARCH(1,1) model, while the additional assumption of $\delta_j^l = 0$ gives ARCH(1). The stationarity used in GQARCH is a covariance stationarity, which is satisfied whenever the sum of α_j^l and δ_j^l is less than one. This sum of estimated parameters also provides a measure of the persistence of shocks to the variance process. Additionally, Sentana [40] shows that many properties of the GQARCH model are similar to the standard GARCH(1,1) model. In particular, the unconditional variance of f_{lt} as implied by the GQARCH(1,1) model is the same as that implied by the GARCH(1,1) model. Moreover, the condition for covariance-stationarity of the GQARCH(1,1) model and the condition for existence of the unconditional fourth moment are the same as the corresponding conditions in the GARCH(1,1) model. Furthermore, the autocorrelations of the squares of the GQARCH(1,1) model are exactly the same as those of the GARCH(1,1) model. Nonetheless, compared with the GARCH(1,1) model, the GQARCH(1,1) model has the main advantage of capturing the features of asymmetry and higher excess kurtosis. As demonstrated in Sentana [40], the kurtosis of the GQARCH(1,1) model is an increasing function of the absolute value of γ_j^l. For fixed values of the parameters w_j^l, α_j^l, δ_j^l in GQARCH(1,1), the kurtosis for the GQARCH(1,1) model is larger than the kurtosis for the corresponding GARCH(1,1) model. Thus, the GQARCH(1,1) model go in the right direction towards capturing some of the stylized facts (excess kurtosis and volatility asymmetry) pertaining to financial time series.

To guarantee the identification of the model, we suppose that $q \geq k$ and $rank(\mathbf{X}_j) = k$, $\forall\, j$. We suppose also that the common and idiosyncratic factors are uncorrelated, and that $\mathbf{f}_t$ and $\varepsilon_{t'}$ are mutually independent for all t, t'.

Our model is general enough that it allows for changing relationships among variables in the data set without imposing that these changes have occurred or assuming a date for the changes. It takes into account, simultaneously, the usual changing behavior of the common volatility due to common economic forces, as well as the sudden discrete shift in common and idiosyncratic volatilities that can be due to sudden abnormal events. This new specification allows us to pose a variety of interesting new questions. Can we distinguish distinct regimes in stock market returns? How do the regimes differ? How frequent are regime switches and when do they occur? Are returns predictable, even after accounting for regime switches? Are regime switches predictable? The answers to these questions give us a new set of stylized facts about stock market returns.

4 A Switching State-Space Representation

The model developed above can be regarded as a random field with indices $i = 1, ..., q$, $t = 1, ..., n$ and $j = 1, ..., m$. Therefore, it is not surprising that it has a switching time-series state-space representation, with $\mathbf{f}_t$ as the continuous state variables. The measurement and transition equations are given by:

$$\begin{array}{ll} \text{[Measurement Equation]} & \mathbf{y}_t = \theta_{s_t} + \mathbf{X}_{s_t}\mathbf{f}_{s_t} + \varepsilon_{s_t} \\ \text{[Transition Equation]} & \mathbf{f}_{s_t} = \mathbf{0}.\mathbf{f}_{s_{t-1}} + \mathbf{f}_{s_t} \end{array}$$

For the derivation of the filtering and smoothing equations, we have used the first order generalized pseudo-bayesian method (GPB(1)) based on the moment matching technique. These statistics will thereafter be introduced into a conditional EM algorithm in order to estimate all the parameters of the model. For the implementation of the filtering and smoothing algorithms, we start by introducing some notation.

$$\begin{aligned} \mathbf{f}_{t/\tau}^{i(j)} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_{t-1} = i, S_t = j\right] \\ \mathbf{f}_{t/\tau}^{(j)k} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t = j, S_{t+1} = k\right] \\ \mathbf{f}_{t/\tau}^{j} &= \mathbb{E}\left[\mathbf{f}_t/\mathcal{Y}_{1:\tau}, S_t = j\right] \end{aligned}$$

If $\tau = t$, these are called filtered statistics; if $\tau > t$, they are called smoothed statistics; and if $\tau < t$, we will refer to them as predicted statistics. Notice how the superscript inside the brackets gives the value of the switch node at time t; the superscript to the left describes the value of S_{t-1}, and to the right, S_{t+1}. We need these subtle distinctions to handle the cross-variance terms correctly. We also define the following:

$$
\begin{aligned}
h^{j}_{lt/\tau} &= Var\left(f_{lt}/\mathcal{Y}_{1:\tau}, S_t = j\right) \\
h^{i(j)}_{lt/t-1} &= Var\left(f_{lt}/\mathcal{Y}_{1:t-1}, S_{t-1} = i, S_t = j\right) \\
M_{t-1,t/\tau}(i,j) &= p\left(S_{t-1} = i, S_t = j/\mathcal{Y}_{1:\tau}\right) \\
M_{t/\tau}(j) &= p(S_t = j/\mathcal{Y}_{1:\tau}) \\
L_t(i,j) &= p(\mathbf{y}_t/\mathcal{Y}_{1:t-1}, S_{t-1} = i, S_t = j)
\end{aligned}
$$

where $L_t(i,j)$ is the likelihood of the innovation at time t, given that the current regime is j and the precedent is i.

4.1 Filtering Algorithm

We perform the following steps in sequence.

$$
\begin{aligned}
\mathbf{f}^{i(j)}_{t/t-1} &= \mathbf{0}.\mathbf{f}^{i}_{t-1/t-1} = \mathbf{0} \qquad \forall \quad i,j = 1,...,m \quad \text{and} && (1)\\
h^{i(j)}_{lt/t-1} &= w_{lj} + \gamma_{lj} f^{i}_{lt-1/t-1} + \alpha_{lj}\left[f^{i\,2}_{lt-1/t-1} + h^{i}_{lt-1/t-1}\right] + \delta_{lj} h^{i}_{lt-1/t-2} && (2)\\
\mathbf{H}^{i(j)}_{t/t-1} &= diag\left[h^{i(j)}_{lt/t-1}\right] \quad \text{with} \quad l = 1,2,...,k
\end{aligned}
$$

Then we compute the error in the prediction (the innovation), the variance of the error, the Kalman gain matrix and the likelihood of this observation.

$$\mathbf{e}_t(i,j) = \mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}^{i(j)}_{t/t-1}$$

$$\mathbf{\Sigma}^{i(j)}_{t/t-1} = \mathbf{X}_j \mathbf{H}^{i(j)}_{t/t-1} \mathbf{X}'_j + \mathbf{\Psi}_j$$

$$K_t(i,j) = \mathbf{H}^{i(j)}_{t/t-1} \mathbf{X}'_j \mathbf{\Sigma}^{i(j)\,-1}_{t/t-1}$$

$$L_t(i,j) = \mathcal{N}\left[\mathbf{0}\,, \mathbf{\Sigma}^{i(j)}_{t/t-1}\right]$$

Finally, we update our estimates of the mean and variance:

$$
\begin{aligned}
\mathbf{f}^{i(j)}_{t/t} &= \mathbf{f}^{i(j)}_{t/t-1} + K_t(i,j)\mathbf{e}_t(i,j) && (3)\\
\mathbf{H}^{i(j)}_{t/t} &= \left[\mathbf{I}_k - K_t(i,j)\mathbf{X}_j\right]\mathbf{H}^{i(j)}_{t/t-1} = \mathbf{H}^{i(j)}_{t/t-1} - K_t(i,j)\mathbf{\Sigma}^{i(j)}_{t/t-1} K_t(i,j)' && (4)
\end{aligned}
$$

The fundamental problem with switching Kalman Filters is that the belief state grows exponentially with time. To see this, suppose that the initial distribution $p(\mathbf{f}_1)$ is a mixture of m Gaussians, one for each value of S_1. Then each of these must be propagated through m different equations (one for each value of S_2), so that $p(\mathbf{f}_2)$ will be a mixture of m^2 Gaussians. In general, at time t, the belief state $p(\mathbf{f}_t/\mathcal{Y}_{1:t})$ will be a mixture of m^t Gaussians, one for each possible model history $S_1, ..., S_t$. To dealing with this exponential growth we have used the collapsing technique. This method consists in approximating the mixture of m^t Gaussians with a mixture of r Gaussians. This is called the Generalized Pseudo Bayesian algorithm of order r (GPB(r)) (see e.g., Bar-Shalom and Li [3], Kim [26]). When $r = 1$, we approximate a mixture of Gaussians with a single Gaussian using moment matching; this can be shown (e.g., Lauritzen [28]) to be the best (in the Kullback-Leibler sense) single Gaussian approximation. When $r = 2$, we "collapse" Gaussians which differ in their history two steps ago; in general, these will be more similar than Gaussians that differ in their more recent history.

For the implementation of this algorithm we calculate the following probabilities:

$$M_{t-1,t/t}(i,j) = \frac{L_t(i,j)p_{ij}M_{t-1/t-1}(i)}{\sum_{i=1}^{m}\sum_{j=1}^{m} L_t(i,j)p_{ij}M_{t-1/t-1}(i)}$$

The derivation of the mode update equation is as follows:

$$\begin{aligned}
M_{t-1,t/t}(i,j) &= p(S_{t-1}=i, S_t=j/\mathcal{Y}_{1:t}) \\
&= p(S_{t-1}=i, S_t=j/\mathbf{y}_t, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(S_{t-1}=i, S_t=j, \mathbf{y}_t/\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(\mathbf{y}_t/S_{t-1}=i, S_t=j, \mathcal{Y}_{1:t-1}) p(S_{t-1}=i, S_t=j/\mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} p(\mathbf{y}_t/S_{t-1}=i, S_t=j, \mathcal{Y}_{1:t-1}) p(S_{t-1}=i/\mathcal{Y}_{1:t-1}) \times \\
&\quad\; p(S_t=j/S_{t-1}=i, \mathcal{Y}_{1:t-1}) \\
&= \frac{1}{c} L_t(i,j)p_{ij}M_{t-1/t-1}(i)
\end{aligned}$$

where c is the normalization constant given by: $c = \sum_{i=1}^{m}\sum_{j=1}^{m} L_t(i,j)p_{ij}M_{t-1/t-1}(i)$. We calculate also the probabilities:

$$M_{t/t}(j) = \sum_{i=1}^{m} M_{t-1,t/t}(i,j)$$

$$Z_{i/j}(t) = p(S_{t-1}=i/S_t=j, \mathcal{Y}_{1:t}) = M_{t-1,t/t}(i,j)/M_{t/t}(j)$$

Finally, we update our estimates of the mean, volatility, and predicted volatility.

$$
\begin{aligned}
\mathbf{f}^{j}_{t/t} &= \sum_{i=1}^{m} Z_{i/j}(t)\mathbf{f}^{i(j)}_{t/t} \\
h^{j}_{lt/t} &= \sum_{i=1}^{m} Z_{i/j}(t)h^{i(j)}_{lt/t} + \sum_{i=1}^{m} Z_{i/j}(t)\left[f^{i(j)}_{lt/t} - f^{j}_{lt/t}\right]\left[f^{i(j)}_{lt/t} - f^{j}_{lt/t}\right]' \\
h^{j}_{lt/t-1} &= \sum_{i=1}^{m} Z_{i/j}(t)h^{i(j)}_{lt/t-1} + \sum_{i=1}^{m} Z_{i/j}(t)\left[f^{i(j)}_{lt/t-1} - f^{j}_{lt/t-1}\right]\left[f^{i(j)}_{lt/t-1} - f^{j}_{lt/t-1}\right]' \\
\mathbf{H}^{j}_{t/t} &= diag\left[h^{j}_{lt/t}\right] \quad \text{and} \quad \mathbf{H}^{j}_{t/t-1} = diag\left[h^{j}_{lt/t-1}\right] \quad \text{for} \quad l = 1, 2, ..., k
\end{aligned}
$$

4.2 Smoothing Algorithm

Given the degenerate nature of the (time-series) transition equation, the smoother gain matrix $J_t^{(j)k}$ is always null,

$$
J_t^{(j)k} = \mathbf{H}^{j}_{t/t}\mathbf{0}'_k\mathbf{H}^{(j)k\,-1}_{t+1/t} = \mathbf{0}
$$

hence, smoothing is unnecessary in this case.

$$
\begin{aligned}
\mathbf{f}^{(j)k}_{t/n} &= \mathbf{f}^{j}_{t/t} + J_t^{(j)k}\left[\mathbf{f}^{k}_{t+1/n} - \mathbf{f}^{j(k)}_{t+1/t}\right] = \mathbf{f}^{j}_{t/t} \\
\mathbf{H}^{(j)k}_{t/n} &= \mathbf{H}^{j}_{t/t} + J_t^{(j)k}\left[\mathbf{H}^{k}_{t+1/n} - \mathbf{H}^{j(k)}_{t+1/t}\right]J_t^{(j)k\prime} = \mathbf{H}^{j}_{t/t}
\end{aligned}
$$

Thereafter we calculate the probabilities,

$$
U^{j/k}_{t/t+1} = p(S_t = j/S_{t+1} = k, \mathcal{Y}_{1:n}) \simeq \frac{M_{t/t}(j)p_{jk}}{\sum\limits_{j'=1}^{m} M_{t/t}(j')p_{j'k}}
$$

where the approximation arises because S_t is not conditionally independent of the future evidence $\mathbf{y}_{t+1}, ..., \mathbf{y}_n$, given S_{t+1}. This approximation will not be too bad provided future evidence does not contain much information about S_t beyond what is contained in S_{t+1}.

For updating the parameters, we also have need of the probabilities:

$$
\begin{aligned}
M_{t,t+1/n}(j,k) &= U^{j/k}_{t/t+1}M_{t+1/n}(k) \\
M_{t/n}(j) &= \sum_{k=1}^{m} M_{t,t+1/n}(j,k)
\end{aligned}
$$

5 Viterbi Approximation For Latent Structure Inference

The task of the Viterbi approximation approach is to find the best sequence of switching states S_t and common factors $\mathbf{f}_t$ that minimizes the Hamiltonian cost in equation (5) for a given observation sequence $\mathcal{Y}_{1:n}$. The application of Viterbi inference to discrete state hidden Markov models (Rabiner and Juang [33]) and continuous state Gauss-Markov models (Kalman and Bucy [25]) is well known. We present here an algorithm for Viterbi inference in our switching conditionally heteroscedastic factor model. Let

$$\begin{aligned}\mathcal{H}(\mathcal{F}_{1:n},\mathcal{S}_{1:n},\mathcal{Y}_{1:n}) &\simeq Constant+\sum_{t=2}^{n}S_t'(-\log\mathbf{P})S_{t-1}+S_1'(-\log\pi)\\ &+\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m}\left[(\mathbf{y}_t-\mathbf{X}_j\mathbf{f}_{jt}-\theta_j)'\mathbf{\Psi}_j^{-1}(\mathbf{y}_t-\mathbf{X}_j\mathbf{f}_{jt}-\theta_j)+\log|\mathbf{\Psi}_j|\right]S_t(j)\\ &+\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m}\left[\mathbf{f}_{jt}'\mathbf{H}_{jt}^{-1}\mathbf{f}_{jt}+\log|\mathbf{H}_{jt}|\right]S_t(j) \qquad (5)\end{aligned}$$

be the Hamiltonean cost, where π is the vector of initial state probabilities and $\mathbf{P}$ the HMM transition matrix so that its *i-th* row is equal to $[p_{i1}\dots p_{im}]$ for $i=1,...,m$ and $S_t=[S_t(1),...,S_t(m)]'$, with $S_t(j)=1$ if $S_t=j$ and 0 otherwise.

Now, if the best sequence of switching states is denoted $\mathcal{S}^*_{1:n}$ we can approximate the desired posterior $p(\mathcal{F}_{1:n},\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})$ as [2]:

$$\begin{aligned}p(\mathcal{F}_{1:n},\mathcal{S}_{1:n}/\mathcal{Y}_{1:n}) &= p(\mathcal{F}_{1:n}/\mathcal{S}_{1:n},\mathcal{Y}_{1:n})p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})\\ &\simeq p(\mathcal{F}_{1:n}/\mathcal{S}_{1:n},\mathcal{Y}_{1:n})\mu(\mathcal{S}_{1:n}-\mathcal{S}^*_{1:n})\end{aligned}$$

i.e. the switching sequence posterior $p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})$ is approximated by its mode. More formally, we are looking for the switching sequence $\mathcal{S}^*_{1:n}$ such that

$$\mathcal{S}^*_{1:n}=\arg\max_{\mathcal{S}_{1:n}}p(\mathcal{S}_{1:n}/\mathcal{Y}_{1:n})$$

It is easy to shown that a (suboptimal) solution to this problem can be obtained by recursive optimization of the probability of the best sequence at time t:

$$\begin{aligned}J_{t,j} &= \max_{\mathcal{S}_{1:t-1}}p(\mathcal{S}_{1:t-1},S_t=j,\mathcal{Y}_{1:t})\\ &\simeq \max_i\Big\{p(\mathbf{y}_t/S_t=j,S_{t-1}=i,\mathcal{S}^*_{1:t-2}(i),\mathcal{Y}_{1:t-1})p(S_t=j/S_{t-1}=i)\\ &\qquad\times\max_{\mathcal{S}_{1:t-2}}p(\mathcal{S}_{1:t-2},S_{t-1}=i,\mathcal{Y}_{1:t-1})\Big\}\end{aligned}$$

[2] $\mu(x)=1$ for $x=\emptyset$ and zero otherwise.

where $\mathcal{S}^*_{1:t-2}(i) = \underset{\mathcal{S}_{1:t-2}}{\arg\max} J_{t-1,i}$ is the "best" switching sequence up to time $t-1$ when the system is in state i at time $t-1$.

Define first the "best" partial cost up to time t of the measurement sequence $\mathcal{Y}_{1:t}$ when the switch is in state j at time t:

$$J_{t,j} = \min_{\mathcal{S}_{1:t-1},\mathcal{F}_{1:t}} \mathcal{H}\left[\mathcal{F}_{1:t}, \{\mathcal{S}_{1:t-1}, S_t = j\}, \mathcal{Y}_{1:t}\right] \tag{6}$$

Namely, this cost is the least cost over all possible sequences of switching states $\mathcal{S}_{1:t-1}$ and corresponding factor model states $\mathcal{F}_{1:t}$. This partial cost is essential in Viterbi like total cost minimization. For a given switch state transition $i \rightarrow j$ it is now easy to establish relationship between the filtered and the predicted estimates (equations (1-2)). From the theory of Kalman estimation (Anderson and Moore [2]) and given a new observation $\mathbf{y}_t$ at time t each of these predicted estimates can now be filtered using Kalman measurement update framework (equations (3-4)). Hence, each of these $i \rightarrow j$ transitions has a certain innovation cost $J_{t,t-1,i,j}$ associated with it, as defined in Equation (7).

$$\begin{aligned} J_{t,t-1,i,j} &= \frac{1}{2}\left[\mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}^{i(j)}_{t/t-1}\right]' \mathbf{\Sigma}^{i(j)\,-1}_{t/t-1} \left[\mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}^{i(j)}_{t/t-1}\right] \\ &+ \frac{1}{2}\log\left|\mathbf{\Sigma}^{i(j)}_{t/t-1}\right| - \log p_{ij} \end{aligned} \tag{7}$$

One portion of this innovation cost reflects the continuous state transition, as indicated by the innovation terms in equation (3). The remaining cost ($-\log p_{ij}$) is due to switching from state i to state j.

Obviously, for every current switching state j there are m possible previous switching states from which the system could have originated from. To minimize the overall cost at every time step t and for every switching state j, one "best" previous state i is selected:

$$\begin{aligned} J_{t,j} &= \min_i \{J_{t,t-1,i,j} + J_{t-1,i}\} \\ \delta_{t-1,j} &= \arg\min_i \{J_{t,t-1,i,j} + J_{t-1,i}\} \end{aligned}$$

The index of this state is kept in the state transition record $\delta_{t-1,j}$. Consequently, we now obtain a set of m best filtered continuous states and their variances at time t: $\mathbf{f}^j_{t/t} = \mathbf{f}^{\delta_{t-1,j}\,(j)}_{t/t}$ and $\mathbf{H}^j_{t/t} = \mathbf{H}^{\delta_{t-1,j}\,(j)}_{t/t}$ with $h^j_{lt/t-1} = h^{\delta_{t-1,j}\,(j)}_{lt/t-1}$ for $l = 1, ..., k$. Once all n observations $\mathcal{Y}_{1:n}$ have been fused, the best overall cost is obtained as

$$J^*_n = \min_j J_{n,j}$$

To decode the "best" switching state sequence, one uses the index of the best final state, $j_n^* = \arg\min_j J_{n,j}$, then traces back through the state transition record $\delta_{t-1,j}$ in order to obtain the optimal state at each time step:

$$j_t^* = \delta_{t,j_{t+1}^*}$$

The Switching model's sufficient statistics are now simply given by $\mathbb{E}(S_t/.) = S_t(j^*)$ and $\mathbb{E}(S_t S_{t-1}'/.) = S_t(j^*)S_{t-1}(j^*)'$. [3] Given the "best" switching state sequence, the sufficient conditionally heteroscedastic factor model statistics can be easily obtained using the Rauch-Tung-Streibel [34] smoothing (for a review see also Rosti and Gales [37]). For example,

$$\mathbb{E}(\mathbf{f}_t, S_t(j)/.) = \begin{cases} \mathbf{f}_{t/n}^{j_t^*} & j = j_t^* \\ \mathbf{0} & \text{otherwise} \end{cases}$$

6 The EM Algorithm

An efficient learning algorithm for the parameters of our model can be derived by generalizing the Expectation Maximization (EM) algorithm (Dempster et al., [12]). EM is an iterative procedure to find the maximum likelihood estimates of parameters or the posterior mode of parameters in a model. The idea behind the algorithm is to augment the observed data with latent data, which can be either missing data or parameter values, so that the likelihood conditioned on the complete data has a form that is easy to analyze.

The algorithm can be broken down into three steps: the expectation step (E) and two conditional maximization steps. We assume that the data can separated into two components, $\mathcal{Y}$ and ($\mathcal{F}$, $\mathcal{S}$) (observed and latent variables). The E step finds $\mathcal{Q}(\Theta, \Theta^{(i)})$, the expected value of the log-likelihood of Θ, $\mathcal{L}(\Theta/\mathcal{Y}, \mathcal{F}, \mathcal{S})$, where the expectation is taken with respect to $\mathcal{F}$ and $\mathcal{S}$ conditioned on $\mathcal{Y}$ and $\Theta^{(i)}$, the current guess of Θ. For a sequence of observations $\mathcal{Y}$, a sequence of continuous states $\mathcal{F}$ and a sequence of discrete HMM states $\mathcal{S}$, $\mathcal{L}(\Theta/\mathcal{Y}, \mathcal{F}, \mathcal{S})$ can be written as:

$$\mathcal{L}(\Theta/\mathcal{Y}, \mathcal{F}, \mathcal{S}) = \log\left[p(S_1)\prod_{t=2}^{n} p(S_t/S_{t-1})\prod_{t=1}^{n} p(\mathbf{f}_t/S_t, \mathcal{D}_{1:t-1})p(\mathbf{y}_t/\mathbf{f}_t, S_t, \mathcal{D}_{1:t-1})\right]$$

where $\mathcal{D}_{1:t-1} = \{\mathcal{Y}_{1:t-1}, \mathcal{F}_{1:t-1}, \mathcal{S}_{1:t-1}\}$, is the information set at time $t-1$. The auxiliary function that will be maximized is given by:

$$\begin{aligned} \mathcal{Q}(\Theta, \Theta^{(i)}) &= \mathbb{E}\left[\log p(\mathcal{Y}, \mathcal{F}, \mathcal{S}/\Theta^{(i)})/\mathcal{Y}, \Theta\right] \\ &= \sum_{\forall \mathcal{S}} \int p(\mathcal{F}/\mathcal{Y}, \mathcal{S}, \Theta)p(\mathcal{S}/\mathcal{Y}, \Theta)\log p(\mathcal{Y}, \mathcal{F}, \mathcal{S}/\Theta^{(i)})\, d\mathcal{F} \end{aligned}$$

[3] The operator $\mathbb{E}(/.)$ denotes conditional expectation with respect to the posterior distribution, e.g. $\mathbb{E}(\mathbf{f}_t/.) = \sum_{\mathcal{S}} \int_{\mathcal{F}} \mathbf{f}_t p(\mathcal{F}, \mathcal{S}/\mathcal{Y})$.

The maximization steps then find $\Theta^{(i+1)}$, the value of Θ that maximizes $\mathcal{Q}(\Theta, \Theta^{(i)})$ over all values possible values of Θ. $\Theta^{(i+1)}$ replaces $\Theta^{(i)}$ in the E-step and $\Theta^{(i+2)}$ is chosen to maximize $\mathcal{Q}(\Theta, \Theta^{(i+1)})$. This procedure is repeated until the sequence $\Theta^{(0)}$, $\Theta^{(1)}$, $\Theta^{(2)}$, ... converges. The EM algorithm is constructed in such a way that the sequence of $\Theta^{(i)}$'s converges to the maximum likelihood estimate of Θ.

E Step:

Let $\mathcal{D}_n^{(i)} = \left\{\mathcal{Y}_{1:n}, \Theta^{(i)}\right\}$ and $\widetilde{\mathbf{y}}_{jt} = \mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j$. The conditional expectation of the complete log-likelihood function is given by:

$$\begin{aligned}\mathcal{Q}(\Theta/\Theta^{(i)}) \quad \simeq \quad & \sum_{j=1}^{m} M_{1/n}(j) \log p(S_1) - \sum_{t=2}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m} M_{t-1,t/n}(i,j) \log p_{ij} \\ - \quad & \frac{1}{2}\sum_{j=1}^{m}\sum_{t=1}^{n} M_{t/n}(j) \left[\log|\mathbf{\Psi}_j| + \mathbb{E}\left\{(\widetilde{\mathbf{y}}_{jt} - \theta_j)' \mathbf{\Psi}_j^{-1} (\widetilde{\mathbf{y}}_{jt} - \theta_j)/\mathcal{D}_n^{(i)}\right\}\right] \\ - \quad & \frac{1}{2}\sum_{j=1}^{m}\sum_{l=1}^{k}\sum_{t=1}^{n} M_{t/n}(j) \mathbb{E}\left[\log(h_{lt}^j) + \frac{f_{lt}^2}{h_{lt}^j} / \mathcal{D}_n^{(i)}\right] \end{aligned} \tag{8}$$

CM1 Step:

Given the two sets of sufficient statistics above, the model parameters can be optimized by maximizing the conditional expectation of the complete log-likelihood function (8) with respect to the initial state probabilities π_j, transition probabilities p_{ij}, observation noise mean vectors θ_j, factor loadings $\mathbf{X}_j$ and idiosyncratic variances $\mathbf{\Psi}_j$. Detailed derivation of the parameter optimization can be found in the Appendix.

$$\widehat{\pi}_j = \frac{M_{1/n}(j)}{\sum\limits_{i=1}^{m} M_{1/n}(i)}$$

$$\widehat{p}_{ij} = \frac{\sum\limits_{t=2}^{n} M_{t-1,t/n}(i,j)}{\sum\limits_{t=2}^{n} M_{t-1/n}(i)}$$

$$\widehat{\theta}_j = \frac{1}{\sum\limits_{t=1}^{n} M_{t/n}(j)} \sum_{t=1}^{n} M_{t/n}(j) \left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{t/n}^j\right)$$

$$\widehat{\mathbf{x}}_{jl}=\left[\sum_{t=1}^{n}M_{t/n}(j)(y_{tl}-\theta_{jl})\mathbf{f}^{j}_{t/n}\right]'\left[\sum_{t=1}^{n}M_{t/n}(j)\left[\mathbf{H}^{j}_{t/n}+\mathbf{f}^{j}_{t/n}\mathbf{f}^{j\prime}_{t/n}\right]\right]^{-1}$$

$$\widehat{\mathbf{\Psi}}_{j}=\frac{1}{\sum\limits_{t=1}^{n}M_{t/n}(j)}\sum_{t=1}^{n}M_{t/n}(j)diag\left\{\mathbf{y}_t\mathbf{y}'_t-\left[\begin{array}{cc}\mathbf{X}_j & \theta_j\end{array}\right]\left[\begin{array}{c}\mathbf{f}^{j}_{t/n}\mathbf{y}'_t\\ \mathbf{y}'_t\end{array}\right]-\left[\begin{array}{cc}\mathbf{y}_t\mathbf{f}^{j\prime}_{t/n} & \mathbf{y}_t\end{array}\right]\right.$$
$$\left.\times\left[\begin{array}{c}\mathbf{X}'_j\\ \theta'_j\end{array}\right]+\left[\begin{array}{cc}\mathbf{X}_j & \theta_j\end{array}\right]\left[\begin{array}{cc}\mathbf{H}^{j}_{t/n}+\mathbf{f}^{j}_{t/n}\mathbf{f}^{j\prime}_{t/n} & \mathbf{f}^{j}_{t/n}\\ \mathbf{f}^{j\prime}_{t/n} & 1\end{array}\right]\left[\begin{array}{c}\mathbf{X}'_j\\ \theta'_j\end{array}\right]\right\}$$

where $\mathbf{x}_{jl}$ is the *l-th* row vector of $\mathbf{X}_j$, y_{tl} and θ_{jl} are, respectively, the l-th elements of the current observation and the observation noise mean vectors under regime j.

<u>*CM2 Step:*</u>

Now, being given the new values of π_j, p_{ij}, θ_j, $\mathbf{X}_j$ and $\mathbf{\Psi}_j$, if the factors and the discrete states were observed we would have:

$$\left(\begin{array}{c}\mathbf{y}_t\\ \mathbf{f}_t\end{array}\right)/\mathcal{D}_{1:t-1},S_t=j\ \sim\mathcal{N}\left[\left(\begin{array}{c}\theta_j\\ \mathbf{0}\end{array}\right),\left(\begin{array}{cc}\mathbf{X}_j\mathbf{H}_{jt}\mathbf{X}'_j+\mathbf{\Psi}_j & \mathbf{X}_j\mathbf{H}_{jt}\\ \mathbf{H}_{jt}\mathbf{X}'_j & \mathbf{H}_{jt}\end{array}\right)\right]$$

However, the $\mathbf{f}'_t s$ and $S'_t s$ are unobserved, but in a such situation and for the estimation of the parameters of the model, we can approximate the distribution of the $\mathbf{y}_t$'s, conditional on the information actually available at time $t-1$, by the following distribution (Harvey, Ruiz and Sentana [23]):

$$\mathbf{y}_t/\mathcal{Y}_{1:t-1},S_t=j,\mathcal{S}_{1:t-1}\approx\mathcal{N}\left[\theta_j,\mathbf{\Sigma}^{(j)}_{t/t-1}\right]$$

where " $\approx$ " stands for "approximately distributed", $\mathbf{\Sigma}^{(j)}_{t/t-1}=\mathbf{X}_j\mathbf{H}^{(j)}_{t/t-1}\mathbf{X}'_j+\mathbf{\Psi}_j$ and $\mathbf{H}^{(j)}_{t/t-1}$ is the expectation of $\mathbf{H}_t$, conditional on $\mathcal{Y}_{1:t-1}$ and $\mathcal{S}_{1:t}$, obtained via the quasi-optimal version of the Kalman filter. Here, the *l-th* diagonal element of the covariance matrix $\mathbf{H}^{(j)}_{t/t-1}$ is given by $h^{j}_{lt/t-1}=h^{\delta_{t-1,j}(j)}_{lt/t-1}$. Therefore, ignoring initial conditions, the pseudo log-likelihood function is given by:

$$\mathcal{L}^*=c-\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m}S_t(j)\left[\log|\mathbf{\Sigma}^{(j)}_{t/t-1}|+(\mathbf{y}_t-\theta_j)'\mathbf{\Sigma}^{(j)-1}_{t/t-1}(\mathbf{y}_t-\theta_j)\right]\qquad(9)$$

In the second maximization step, using the θ_j, $\mathbf{X}_j$ and $\mathbf{\Psi}_j$ parameter values founded in the first step, we maximize (9) with respect to the conditional variance parameters, w_j, γ_j, α_j and δ_j, and so on until convergence. However, for the implementation of the optimization algorithm it is necessary to identify the optimal sequence of the Markovian hidden states, which can be carried out by using the approximated version of the Viterbi algorithm or the posterior probabilities $M_{t/n}(j)$ given by the smoothing algorithm. Once this sequence is known, on each segment of data the function $\mathcal{L}^*$ is maximized through a quasi-Newton algorithm. The final parameter estimates obtained in this way will be the maximum likelihood estimates of our model.

The standard FAHMM is a particular case of the dynamic system presented in section 3. In this case the parameter update formulae are similar to those given above, except for the common factor covariance matrix $\mathbf{H}_j$ which is given by:

$$\widehat{\mathbf{H}}_j = \frac{1}{\sum\limits_{t=1}^{n} S_t(j)} diag \left\{ \sum_{t=1}^{n} S_t(j) \left[\widetilde{\mathbf{R}}_j + \widetilde{\mathbf{f}}_{jt} \widetilde{\mathbf{f}}'_{jt} \right] \right\}$$

The other parameters i.e. the factor loadings matrices $\mathbf{X}_j$, the mean vectors θ_j and the idiosyncratic variances $\mathbf{\Psi}_j$ are obtained by replacing $\mathbf{f}^j_{t/n}$ and $\mathbf{H}^j_{t/n}$ by $\widetilde{\mathbf{f}}_{jt}$ and $\widetilde{\mathbf{R}}_j$.

7 Monte Carlo Simulations

In this section extensive simulations have been carried out to verify the correctness and effectiveness of the proposed algorithms. We report on three experiments that were designed to address the following issues:

1. The most important question about the estimator is whether it is a consistent estimator for Θ. Once assured of this consistency, the next natural question is what are reasonable sequence sizes required to obtain accurate and stable estimates?

2. A crucial question for inference is whether the estimates are approximately normally distributed, and what sequence size is needed for such a behavior.

3. An other important question when using a multi-class model, is the choice of a reliable model, containing enough parameters to ensure a realistic fit to the learning data set, but not too much to avoid overfitting and poor performances for future use. To address this the two selection criteria: AIC and BIC will be explored.

7.1 Accuracy and Stability of the Estimates

The example used here has $q = 6$ series, three discrete hidden states and one GQARCH(1,1) latent factor. The regime switching dates are $t_1^* = n/3 + 1$ and $t_2^* = 2n/3 + 1$. The iterations

Table 1: Simulation parameters.

	θ	$\mathbf{X}$	$diag(\mathbf{\Psi})$	ϕ
State 1	1.0000 (0.0000)	1.0000 (0.5000)	1.0000 (0.5000)	0.5000 (0.1200)
	1.0000 (1.0000)	2.0000 (1.0000)	1.0000 (0.5000)	0.1000 (0.1800)
	1.0000 (0.5000)	3.0000 (1.0000)	1.0000 (0.5000)	0.8000 (0.3800)
	2.0000 (1.0000)	4.0000 (1.5000)	1.0000 (0.5000)	
	2.0000 (0.0000)	5.0000 (1.5000)	1.0000 (0.5000)	
	2.0000 (0.5000)	6.0000 (2.5000)	1.0000 (0.5000)	
State 2	1.0000 (1.0000)	2.0000 (1.0000)	2.0000 (0.5000)	0.1000 (0.2900)
	2.0000 (1.0000)	2.0000 (0.5000)	2.0000 (0.5000)	0.3000 (0.1200)
	1.0000 (1.0000)	2.0000 (0.5000)	2.0000 (0.5000)	0.4000 (0.7800)
	2.0000 (1.0000)	3.0000 (1.0000)	2.0000 (0.5000)	
	1.0000 (1.0000)	3.0000 (0.5000)	2.0000 (0.5000)	
	2.0000 (1.0000)	3.0000 (0.5000)	2.0000 (0.5000)	
State 3	2.0000 (1.0000)	1.0000 (1.0000)	3.0000 (0.5000)	0.2000 (0.6000)
	3.0000 (1.0000)	3.0000 (0.5000)	3.0000 (0.5000)	0.2000 (0.5400)
	2.0000 (1.0000)	1.0000 (0.5000)	3.0000 (0.5000)	0.6000 (0.2000)
	3.0000 (1.0000)	2.0000 (1.0000)	3.0000 (0.5000)	
	2.0000 (1.0000)	4.0000 (0.5000)	3.0000 (0.5000)	
	3.0000 (1.0000)	4.0000 (0.5000)	3.0000 (0.5000)	

. Parameter values for the true model, (.) Initial values for the EM algorithm.

of the EM algorithm stop when the relative change in the likelihood function between two subsequent iterations is smaller than a threshold value $= 10^{-4}$. In this experiment we try to estimate the parameters of a switched dynamic model and to study the behavior of the estimates when the size of the sequence n increases. With this intention, we generated sequences of observations of sizes $n = 600$, 900, 1200 and 1500 (with a hundred replications per simulation). Here the number of parameters is assumed to be known. The constant term of the the conditionally heteroscedastic component is assumed also to be known ($w_j = 1$ $\forall\, j = 1, 2, 3$ and the initializations given in table 1 were used.

The goal is to estimate the different dynamics and to measure the distance between estimates $\widetilde{\Theta}$ and true parameters Θ_0 through the Kullback-Leibler divergence. This distance measure was effectively used in earlier studies (see Juang and Rabiner [24]). For a finite sequence of length n, we define the sample Kullback-Leibler divergence between two parameter points as:

$$K_n(\Theta_0, \Theta) \stackrel{def}{=} \frac{1}{n}\left\{\log \mathcal{L}(\mathbf{y}_1, ..., \mathbf{y}_n; \Theta_0) - \log\, \mathcal{L}(\mathbf{y}_1, ..., \mathbf{y}_n; \Theta)\right\}$$

For each value of n, the estimation procedure was carried out a hundred times, and the distances $\widetilde{K}_n(\Theta_0, \widetilde{\Theta}_n)$ between each of the hundred estimators and the true parameter Θ_0 were evaluated on a new sequence, independent of the first hundred sequences used to obtain the estimators. This prevents the potential underestimation of the distance as a result of estimating the parameters and allows to evaluate its performance on the same sequence.

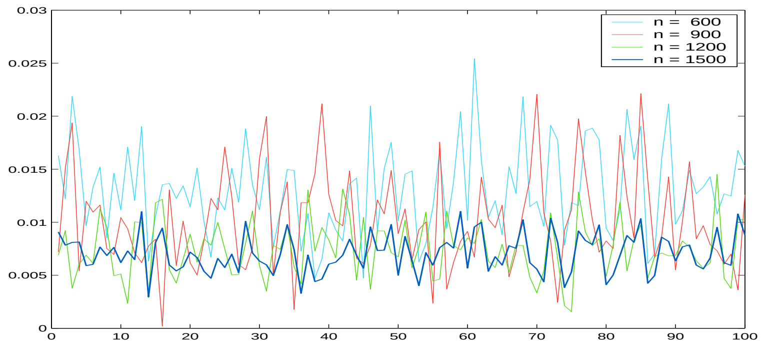

Figure 2: Box plots of $\widetilde{K}(\Theta_0, \widetilde{\Theta}_n)$.

Table 2: Averages and standard deviations (.) for the EM parameter estimates from the simulated data with $n = 1500$.

	θ	$\mathbf{X}$	$diag(\mathbf{\Psi})$	ϕ
	0.9833 (0.0983)	1.9980 (0.0682)	0.9989 (0.0472)	0.4988 (0.0736)
	1.0284 (0.0974)	1.9973 (0.0667)	1.0016 (0.0457)	0.1073 (0.0496)
State 1	1.0197 (0.0857)	2.9952 (0.0589)	0.9878 (0.0593)	0.7924 (0.0371)
	1.9875 (0.0861)	3.9940 (0.0571)	0.9958 (0.0607)	
	1.9914 (0.0973)	4.9945 (0.0577)	0.9980 (0.0572)	
	2.0841 (0.0866)	5.9952 (0.0604)	1.0006 (0.0486)	
	0.9932 (0.0718)	1.9961 (0.0583)	2.0162 (0.0615)	0.1017 (0.0765)
	1.9917 (0.0745)	2.0014 (0.0618)	2.0256 (0.0592)	0.3022 (0.0483)
State 2	1.0754 (0.0773)	2.0108 (0.0579)	1.9914 (0.0622)	0.3977 (0.0366)
	1.9886 (0.0852)	2.9972 (0.0564)	1.9947 (0.0638)	
	1.0381 (0.0836)	3.0127 (0.0591)	2.0082 (0.0676)	
	1.9914 (0.0794)	3.0194 (0.0538)	1.9928 (0.0606)	
	1.9726 (0.0833)	1.0134 (0.0475)	2.9988 (0.0584)	0.2046 (0.0776)
	2.9759 (0.0872)	3.0097 (0.0481)	3.0047 (0.0561)	0.1992 (0.0377)
State 3	1.9681 (0.0867)	1.0099 (0.0463)	2.9797 (0.0692)	0.5982 (0.0281)
	2.9726 (0.0954)	2.0092 (0.0454)	2.9783 (0.0667)	
	1.9718 (0.0988)	4.0154 (0.0508)	3.0046 (0.0689)	
	2.9690 (0.0826)	4.0105 (0.0511)	2.9992 (0.0712)	

Table 2 shows the mean and standard deviation of the estimates with $n = 1500$. Box plots of the sets of distances for the various values of n are presented under a unified scale in figure 2. This figure clearly shows a general decrease in average and spread of the distances with increasing n. Given that small values of $\widetilde{K}_n$ imply similarity between Θ_0 and $\widetilde{\Theta}_n$, the results of this experiment suggest an increasing accuracy and stability of the estimators as n increases.

7.2 Asymptotic Normality of the Estimates

To investigate the asymptotic distribution of $\widetilde{\Theta}_n$, we have used the Shapiro-Francia [41] statistic in order to test the univariate normality of each component of $\widetilde{\Theta}_n$. This is an omnibus test, and is generally considered relatively powerful against a variety of alternatives, and better than the Shapiro-Wilk [42] test for Leptokurtic Samples. The Shapiro-Francia test is based on an idea suggested (without proof) in Gupta [21] (see also Stephens [44]) according to which we obtain the statistic

$$\mathcal{W} = \frac{\left(m'\widetilde{\Theta}^{(v)}\right)^2}{(m'm)\sum_{i=1}^{v}(\widetilde{\Theta}_{(i)} - \overline{\Theta})^2} \quad \text{where} \quad m_i = \left(\frac{i-3/8}{n+1/4}\right)^{-1}, \quad i = 1, ..., v$$

Here $m' = [m_1, m_2, ..., m_v]$, $\widetilde{\Theta}^{(v)} = \left(\widetilde{\Theta}_{(1)}, ..., \widetilde{\Theta}_{(v)}\right)$ the corresponding ordered statistic of $\widetilde{\Theta} = \left(\widetilde{\Theta}_1, ..., \widetilde{\Theta}_v\right)$ and v is the number of replications. All the results presented in table 3 show that the Shapiro-Francia test fails to reject the null hypothesis (the Θ_i are a random sample from $\mathcal{N}(\mu, \sigma)$, with μ and σ unknown) at the significance level $\alpha = 5\%$.

7.3 Model Selection

The problem of model identification is to choose one among a set of candidate models to describe a given dataset. We often have candidates of a series of models with different number of parameters. It is evident that when the number of parameters in the model is increased, the likelihood of the training data is also increased; however, when the number of parameters is too large, this might cause the problem of overtraining. Several criteria for model selection have been introduced in the statistics literature, ranging from nonparametric methods such as cross-validation, to parametric methods such as the Bayesian Information Criterion BIC.

The BIC proposed by Schwarz [39], is a likelihood criterion penalized by the model complexity: the number of parameters in the model. In detail, let $\mathcal{Y}$ be the dataset we are modeling; let $\mathcal{M} = \{\mathcal{M}_i, \, i = 1, ..., I\}$ be the candidates of desired parametric models. Assuming we maximize the likelihood function separately for each model $\mathcal{M}$, obtaining the maximum likelihood, say $\mathcal{L}(\mathcal{Y}, \mathcal{M})$. Denote $v_{\mathcal{M}}$ as the number of parameters in the model $\mathcal{M}$. The BIC is defined as:

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

This criterion is closely related to other penalized likelihood criteria such as AIC proposed by Akaike [1]. This criterion makes use of the less stringent penalty term $2\,v_{\mathcal{M}}$ and, in many circumstances, is expected to select too complex models.

Table 3: Summary statistics for the Shapiro-Francia test (simulation with $n = 900$).

Statistics						
						Mean Vectors
pval	*0.4964	0.4103	0.3976	0.4184	0.1838	0.4413
	**0.3114	0.4601	0.4819	0.3187	0.3489	0.2653
	***0.2500	0.1668	0.3310	0.1834	0.3513	0.2108
$\mathcal{W}$ statistic	0.0089	0.2267	-0.2594	0.2061	-0.9010	0.1478
	0.4920	0.1001	0.0454	0.4713	0.3882	0.6271
	0.6745	0.9667	0.4372	0.9027	0.3819	0.8035
						Factor Loadings
	0.3450	0.2838	0.1668	0.3760	0.3877	0.3819
	0.1997	0.4091	0.2831	0.3023	0.3190	0.2546
	0.4767	0.3227	0.2908	0.3868	0.1926	0.1367
	-0.3988	-0.5717	-0.9669	-0.3159	0.2853	0.3006
	0.8427	0.2299	0.5737	-0.5179	0.4705	-0.6601
	0.0586	-0.4602	-0.5509	0.2878	0.8683	-1.0954
						Idiosyncratic Variances
	0.4870	0.2921	0.2474	0.2510	0.2269	0.4519
	0.4104	0.4838	0.4740	0.3962	0.3703	0.2766
	0.3707	0.3742	0.2888	0.2860	0.2778	0.3131
	0.0326	0.5473	-0.6828	0.6714	0.7491	-0.1208
	-0.2265	0.0406	0.0652	0.2632	-0.3311	0.5930
	-0.3299	0.3208	-0.5570	0.5652	-0.5895	-0.4871
						GQARCH Parameters
	0.3224	0.1812	0.2117			
	0.3138	0.4761	0.3436			
	0.4356	0.4967	0.4386			
	0.4611	0.9109	0.8006			
	-0.4852	-0.0599	-0.4027			
	-0.1622	0.0084	-0.1545			

* First regime, ** second regime, ** third regime.

In this experiment we consider two different situations with factor models which differ by their dynamic hidden structures. In the first case the true model is the one used in 7.1. In the second case the true model is a GQARCH(1,1) factor model with $n = 800$, $m = k = 2$ and the regime switching date $t^* = n/2 + 1$ (the simulation parameters are given in table 4).

The steps for the model selection procedure are as follows. For each selection criterion, we train various model configurations (obtained by varying the number of states and the number of factors), using the ML criterion on the training dataset. In the second example random initialization was used for the implementation of the learning algorithm[4]. Minimizing the selection criteria – computed after each EM running – allows as to find the best model among the I models. Table 5 shows the results of the two examples. In the first example the BIC criterion chooses 3 states and one factor. This is the best classification, since the use of one or two states is not enough to represent the data, and choosing two factors corresponds to

[4]The initial parameters for the EM algorithm, were obtained by randomly perturbing the true parameter values by up to 20% of their true value.

Table 4: Simulation parameters (Example 2).

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$	ϕ	
State 1	0.5000	1.0000	1.0000	0.1000	0.1000	0.5000
	0.5000	2.0000	1.0000	0.1000	0.3000	0.1000
	0.7000	3.0000	2.0000	0.1000	0.4000	0.8000
	1.0000	4.0000	2.0000	0.1000		
	0.5000	5.0000	3.0000	0.1000		
	0.7000	6.0000	3.0000	0.1000		
State 2	1.0000	1.0000	1.0000	0.4000	0.3000	0.2000
	0.9000	1.0000	2.0000	0.4000	0.2000	0.1000
	1.0000	4.0000	3.0000	0.4000	0.7000	0.8000
	0.7000	4.0000	3.0000	0.4000		
	1.1000	2.0000	2.0000	0.4000		
	1.5000	2.0000	1.0000	0.4000		

an overfitting. In the second example, the BIC chooses also the true specification with two states and two conditionally heteroscedastic factors.

The mean square error criterion given by

$$\widehat{e} = \frac{1}{n} \sum_{i=1}^{q} \sum_{t=1}^{n} \|y_{it} - \widehat{y}_{it}\|^2$$

where $\widehat{\mathbf{y}}_t = \sum_{j=1}^{m} S_t(j) \left[\widehat{\theta}_j + \widehat{\mathbf{X}}_j \mathbf{f}^j_{t/n} \right]$ show also that $k = 1$ and $m = 3$ is strongly favored in the first example (figure 3).

To illustrate the evolution of the model estimates obtained by the EM method, figure 4 shows the HMM hidden states estimates at iteration 2, 5, 10 and 15. Each figure depicts the regime path process of the correct model. It can be concluded that a good segmentation is achieved after 15 iterations. Using the initial guesses given in table 1, the EM algorithm converged to estimates of the GQARCH processes after approximately 50 iterations as shown in figure 5. Figures 6 and 7 show that, except for the true model, all other models lead either to over-estimation or under-estimation.

To assess the previous results Monte Carlo experiments were performed. We have generated 100 different data experiments according to the true model for each example. The best number of common factors and hidden states according to BIC criterion was chosen. Figure 8 shows the choice frequencies for each specification. In the two examples, BIC prefers the true model most of the time.

8 Financial Data

The conditionally heteroscedastic factor model is now applied to modeling the interrelationships between the currencies of eight countries during the European Monetary System

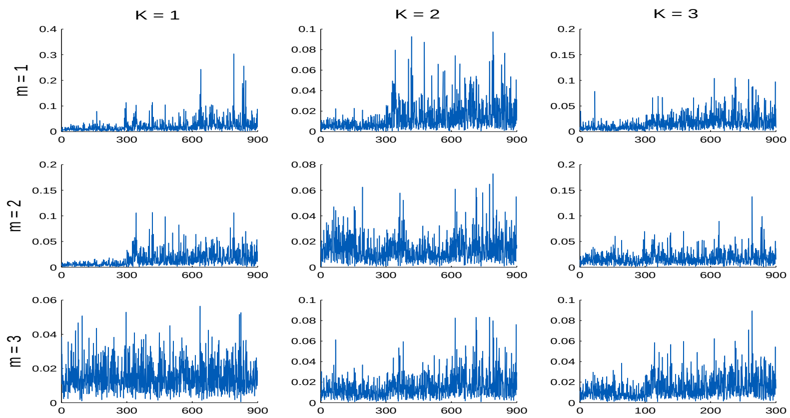

Figure 3: Computation of the estimation error (first example) for 9 different configurations with conditional heteroscedasticity.

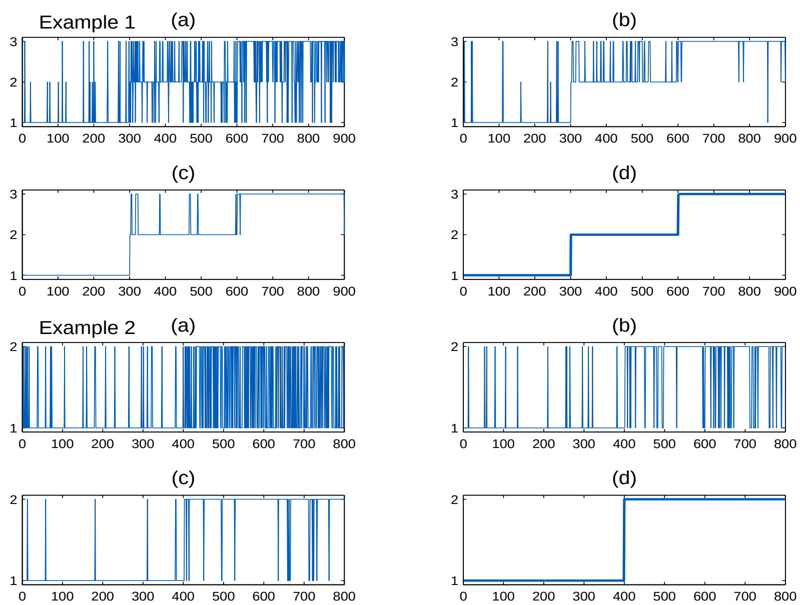

Figure 4: Evolution of the HMM state estimates using the true model: (a)iteration 2, (b) iteration 5, (c) iteration 10, (d) iteration (15).

Table 5: Values of the AIC and BIC statistics for the chosen factor models estimated on the same database. The values into brackets are the selection criteria of the second example.

Criterion	$m=1$		
	$k=1$	$k=2$	$k=3$
AIC	24244 (25262)	26746 (22479)	24178 (23337)
BIC	24456 (25412)	26832 (28911)	24409 (23688)
		$m=2$	
	24384 (22482)	23205 (22229)	24981 (23951)
	24605 (22744)	23565 (**22510**)	25403 (24120)
		$m=3$	
	23214 (22222)	24094 (24278)	24960 (22758)
	23550 (22634)	24267 (24493)	25061 (23049)

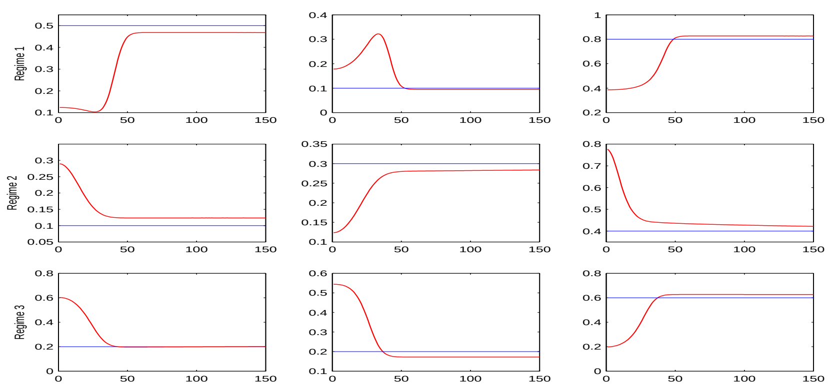

Figure 5: Evolution of the conditionally heteroscedastic parameter estimates during the EM iterations in the first example: γ_j (first column), α_j (second column) and δ_j (third column).

currency crisis of 1992-1993. In that period the European exchange rate mechanism has experienced a succession of crisis which reached its first culmination at the end of August 1992, when the dollar fell to its historically low value of 1.4 German marks. This led to the fall of the pound, which traded slightly above the bottom range of the Exchange Rate Mechanism and was soon followed by the lira. In 1992 most currencies fell to their lowest value of 4.5% (2.25% on each side from central parity). In 1993 the range expanded to 30%. Most currencies moved differently from the German mark in the first year after the range was expanded and started fluctuating within their former margins.

What has been the impact of these changes on the nature of volatility? Has the degree of co-movement increased or decreased? Have common fluctuations become more or less volatile? Has the impact of crises on individual countries evolved over time? These questions are of interest to both policy makers and academic economists. The questions of whether the common volatility has increased or declined, and whether countries have become more

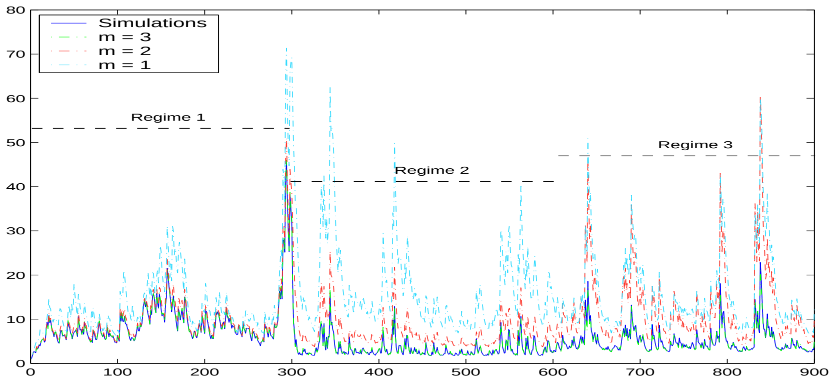

Figure 6: Example 1: Volatility of the common factor using different specifications.

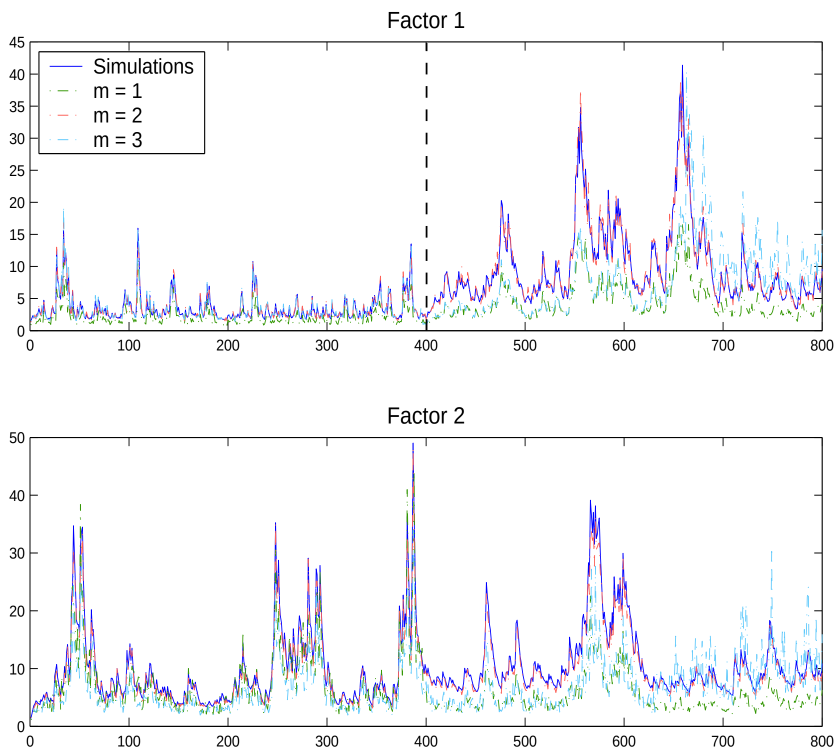

Figure 7: Example 2: Volatility of the two common factors using different specifications.

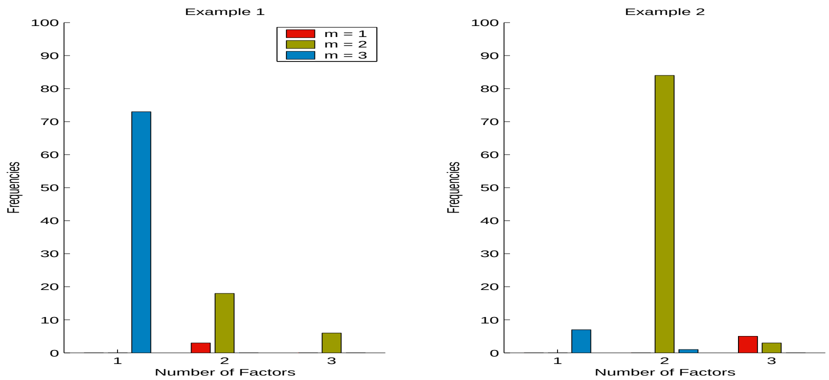

Figure 8: Frequencies of choosing each model with BIC.

or less symmetric, are central to monetary and fiscal policy issues. These questions are also of interest to academics, who have been debating the effects of trade, monetary and financial integration on cross-country exchange rate synchronization.

8.1 The Data

We analyzed a dataset on daily returns of closing spot prices for eight currencies relative to the British pound in price notations.[5] The dataset contains 601 observations that range from 03/05/1991 to 07/05/1993. The 601 observations are transformed in order to use one-day-ahead returns, resulting in the lost of the first observation:

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

where p_t is the daily closing exchange rate at time t. This quantity can be seen as the logarithm of the geometric growth and is known in finance as continuous compounded returns. Figure 9 shows the plot of time series in the order: United States Dollar (USD), Canadian Dollar (CAD), French Franc (FRF), Swiss Franc (CHF), Italian Lira (ITL), German Mark (DEM), Japanese Yen (JPY), and the Hong Kong Dollar (HKD). This figure presents also the 600 returns.

8.2 Exploratory Analysis

In order to assess the distributional properties of the data, various descriptive statistics are reported in table 6, including mean, standard deviation, skewness, kurtosis and other statistics. In particular, the hypothesis of normality is rejected for each exchange rate, using

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

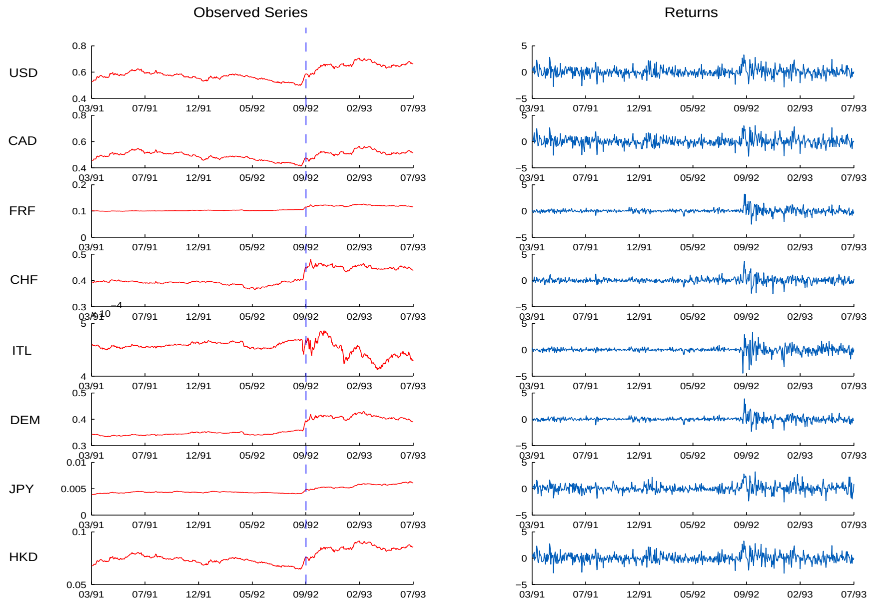

Figure 9: Real daily observed exchange rates and their returns from 03/05/1991 to 07/05/1993 (600 observations). The vertical line represent 31/08/1992.

the Bera and Jarque [4] joint test (BJ test). Further evidence on the nature of deviations from normality may be derived from the sample skewness and kurtosis, measures. The skewness of each series is always very close to zero, while the kurtosis is very large. Visual inspection of each series (figure 9) revealed no evidence of serial correlation, although there seems to be persistence in the conditional variances.

To gain more knowledge about the data and to set the ground for further and deeper investigations with complex switching factor models, we first develop traditional exploratory analyses. Classical static factor analysis is performed and the estimates for the means, the factor loadings and idiosyncratic variances were found when fitting a $k = 1$, 2 and 3-factor models. All the results are given in table 7. These are crude estimates which do not take into consideration any time-varying structure for the time series covariances. Nevertheless, they point out some interesting directions, we summarize below.

1. The first factor weights (first column of the factor loading matrix) has basically the same structure when one, two or three factor models are fitted to the data.

2. The third factor seems to be less important.

Table 6: Summary statistics for the daily returns from 03/05/1991 to 07/05/1993. Q_1 and Q_3 are the first and the third quartile respectively and BJ is the joint test of normality that is based on skewness and kurtosis and follows Chi-square distribution with two degrees of freedom. LB(12) is the Ljung and Box test estimated for the 12-*th* serial correlation for the squared returns of our data.

Statistic	USD	CAD	FRF	CHF	ITL	DEM	JPY	HKD
Mean	0.0385	0.0205	0.0224	0.0179	-0.0110	0.0213	0.0758	0.0393
Std.Dev.	0.8236	0.8290	0.4651	0.5431	0.6038	0.4743	0.7315	0.8262
Skewness	0.2991	0.3298	0.9279	0.5182	-1.0302	1.0804	0.4624	0.2946
Kurtosis	4.4333	4.6104	13.4776	10.2500	15.8776	15.4744	4.9780	4.5005
BJ test 10^3	0.0601	0.0754	2.8212	1.3365	4.2378	3.9936	0.1188	0.0647
Maximum	3.2860	3.0359	3.2270	3.6562	3.3113	3.9079	3.2273	3.2676
Q_3	0.5021	0.4692	0.1946	0.2507	0.1893	0.1824	0.4159	0.4956
Median	0	-0.0140	0.0005	-0.0087	-0.0098	0.0029	0.0147	0.0060
Q_1	-0.4648	-0.4691	-0.1534	-0.2465	-0.1824	-0.1693	-0.3103	-0.4460
Minimum	-2.8506	-2.8345	-2.5251	-2.5592	-4.4431	-2.3295	-2.5374	-2.8564
LB(12)	33.916	35.982	54.356	37.223	37.125	58.206	48.727	30.562

Table 7: Standard Factor Models with different number of factors.

Number of Factors	θ	$diag(\boldsymbol{\Psi})$	$\mathbf{X}$		
	0.0385	0.0015	0.9229		
	0.0205	0.0634	0.8861		
	0.0224	0.1917	0.1748		
$k=1$	0.0179	0.2767	0.1497		
	-0.0110	0.3453	0.1535		
	0.0213	0.2076	0.1467		
	0.0758	0.2633	0.5844		
	0.0393	0.0043	0.9239		
	0.0385	0.0003	0.9876	0.0000	
	0.0205	0.0648	0.9070	0.0074	
	0.0224	0.0129	0.7809	1.4996	
$k=2$	0.0179	0.1037	0.7037	1.4434	
	-0.0110	0.2313	0.4536	1.1920	
	0.0213	0.0011	0.8289	1.7703	
	0.0758	0.2449	0.7096	0.1256	
	0.0393	0.0058	0.9938	0.0212	
	0.0385	0.0001	0.9967	0.0000	0.0000
	0.0205	0.0643	0.9612	0.0312	0.0000
	0.0224	0.0124	0.8923	1.6482	0.1207
$k=3$	0.0179	0.0996	0.8538	1.6833	0.2337
	-0.0110	0.1669	0.7279	1.1453	0.0364
	0.0213	0.0002	0.9163	1.7929	0.1072
	0.0758	0.2445	0.8571	0.5501	0.0841
	0.0393	0.0057	0.9998	0.0192	0.0015

3. Figure 10 provides information about the autocorrelation structure of the data. It presents the first 25 autocorrelation coefficients. The Ljung-Box [30] statistic for the serial correlation of the squared return of the exchange rate imply significant relationship. Bollerslev [6] interprets the high autocorrelation of the squared returns data as a sign of conditional heteroscedasticity. More precisely, "clustering" in returns is very common: this means that volatility changes over time depending on its past values and hence it is predictable.

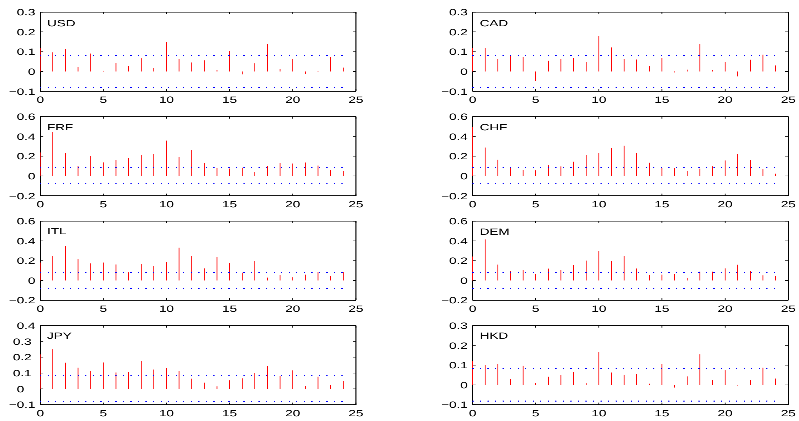

Figure 10: Squared returns autocorrelograms based on data from 03/05/1991 to 07/05/1993.

8.3 Dynamic Factor Analysis

In this section, we will evaluate the performance of the new mixed-state factor model developed in section 3 using objective measures and different specifications. We trained factor models which differ by their hidden volatility structure on the 600 observations of the international exchange rate returns data from 03/05/1991 to 07/05/1993. The key point for initialization is to start with a good segmentation of the data set where, by segmentation, we mean a partition of the data, such that each part is modeled by a particular dynamic factor model. For our model, the best strategy of initialization consists in implementing a standard EM algorithm by supposing that the factors are homoscedastic [6]. Thereafter and given the output of the EM algorithm, we can use the posterior probabilities $p(S_t = j/\mathcal{Y})$ in order to obtain the optimal sequence of hidden states. At the second step, a particular simple conditionally heteroscedastic factor model is initialized for each segment. For this, one can use the empirical covariance matrices as estimates of the idiosyncratic variance matrices $\mathbf{\Psi}_j$ and the empirical means as estimates of the means θ_j. The parameters of the

[6] In practice, 20 iterations of the EM algorithm are largely sufficient.

conditionally heteroscedastic variances are initialized by applying a GQARCH(1,1) model to each segment of data. Finally, the elements of the transition matrix $\mathbf{P}$, can be initialized by counting the number of transitions from state i to state j, $(i, j = 1, ..., m)$, and dividing by the number of transitions from state i to any other state.

In order to identify the number of common factors and discrete states in the model, we estimate different models, assuming that each state variable (discrete and continuous) can take one to three states and choose the model with the lowest criterion presented in section 7.3. For simplicity we have assumed that the coefficients of the conditionally heteroscedastic variances are constant for all t.

Our proposed model must also be further constrained to define a unique model free from identification problems. A first constraint is that $\mathbf{X}_t$ be of full rank k, $\forall\ t$, to avoid identification problems arising through invariance of the model under location shifts of the factor loading matrix (e.g., Geweke and Singleton [19]). Second, we must further constrain the factor loading matrix to avoid overparametrization - simply ensuring that the number of free parameters at time t in the factor representation does not exceed the $q(q+1)/2$ parameters in an unrestricted $\mathbf{\Sigma}_t$. Finally, we need to ensure invariance under invertible linear transformations of the factor vectors (Press [32], chapter 10). On this latter issue, our work follows Geweke and Zhou [20], among others, in adopting the "hierarchical" structural constraint in which the loadings matrix has the form:

$$\mathbf{X}_j = \begin{pmatrix} x_{11j} & 0 & 0 & \dots & 0 \\ x_{21j} & x_{22j} & 0 & \dots & 0 \\ x_{31j} & x_{32j} & x_{33j} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1j} & x_{k2j} & x_{k3j} & \dots & x_{kkj} \\ x_{k+1,1j} & x_{k+1,2j} & x_{k+1,3j} & \dots & x_{k+1,kj} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{q1j} & x_{q2j} & x_{q3j} & \dots & x_{qkj} \end{pmatrix}$$

where $x_{i,ij} > 0$ for $i = 1, ..., k;\ j = 1..., m$ and $x_{i,lj} = 0$ for $i < l,\ i, l = 1, ..., k$. This form immediately ensures that $\mathbf{X}_j$ is of full rank k.

Table 8 provides summaries of the various information criteria for assessment of the number of factors as well as the number of hidden markovian states. Figure 11 shows the evolution of the log-likelihood function of the different specifications during the EM iterations. The overall suggestions is that $k = 2$ and $m = 3$ is strongly favored. The summaries of the empirical analyses with $k = 2$ and 3 conditionally heteroscedastic common factors within a 3 state hidden markovian structure are, respectively, given in tables 9 and 10 and figures 12 to 19. In the first case (when $k = 2$), the transition matrix and the initial state probabilities are given by:

$$\mathbf{P} = \begin{pmatrix} 0.9773 & 0.0227 & 0.0000 \\ 0.0000 & 0.9698 & 0.0302 \\ 0.0834 & 0.2478 & 0.6688 \end{pmatrix} \quad \text{and} \quad \pi = \begin{pmatrix} 0.0000 \\ 1.0000 \\ 0.0000 \end{pmatrix}$$

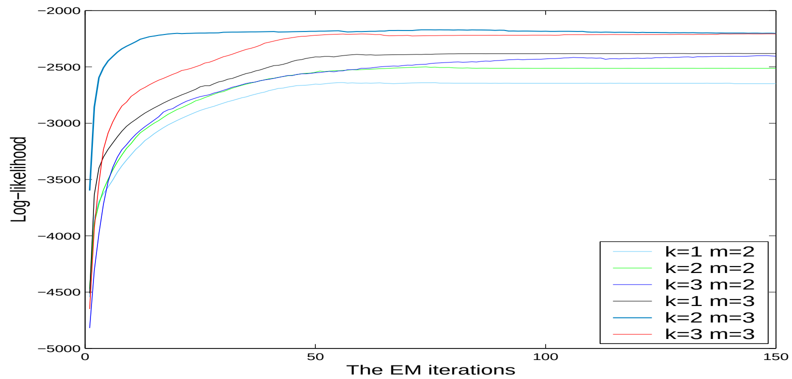

Figure 11: Likelihood of different specifications with conditional heteroscedasticity.

Table 8: Values of the log-likelihood, AIC and BIC statistic for the chosen factor models estimated on the period March 05, 1991-July 05, 1993.

Criterion	$k=1$	$k=2$	$k=3$
		$m=1$	
log-likelihood (-)	3904.5 (3921.2)	4321.6 (3755.9)	4279.4 (3759.6)
AIC	7865.0 (7890.3)	8699.2 (7575.9)	8638.8 (7599.2)
BIC	7988.1 (7995.8)	8822.3 (7716.6)	8814.7 (7775.1)
		$m=2$	
	2648.7 (2660.4)	2512.6 (2531.6)	2404.7 (2506.3)
	5413.5 (5428.8)	5145.1 (5203.2)	4961.3 (5184.6)
	5668.5 (5666.2)	5408.9 (5511.0)	5295.5 (5562.8)
		$m=3$	
	2381.2 (2400.7)	**2202.7** (2225.7)	2207.5 (2253.4)
	4938.4 (4969.3)	**4631.4** (4715.3)	4685.0 (4722.8)
	5325.3 (5338.7)	**5128.3** (5295.7)	5278.6 (5197.7)

. Models with conditional heteroscedasticity, (.) Standard Models

Using this specification, figure 12 shows how the model is capable of accurately detecting abrupt changes in the DEM time series structure and, in particular, the severe disruption by the violent storm which hit the European currency markets in September and October 1992, following the difficulties over ratifying the Maastricht Treaty in Denmark and France. We can clearly see from figure 13 that the third model is responsible for the high volatility segments, the second model is mainly responsible for the time period before August 1992, and the first one for the lower volatility segments after October 1992. Figure 12 shows also that the average duration stay in the first regime is about 37.8 weeks versus 76 in the second and 6.2 in the third. Some other interesting points arise from this analysis:

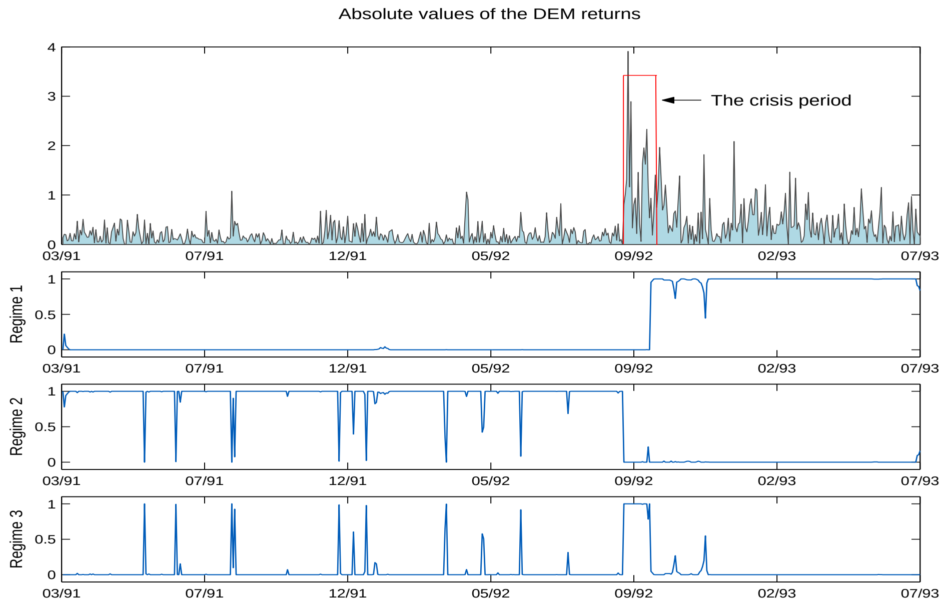

Figure 12: *Graphics 2,3,4:* Posterior probabilities of the three hidden states $M_{t/n}(j)$ given by the smoothing algorithm.

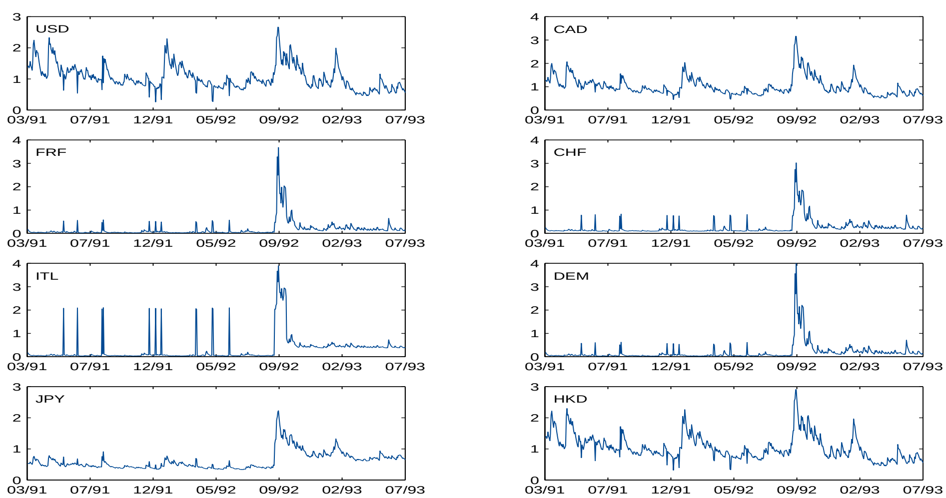

Figure 13: Volatility of the different series using three states and two conditionally heteroscedastic factors.

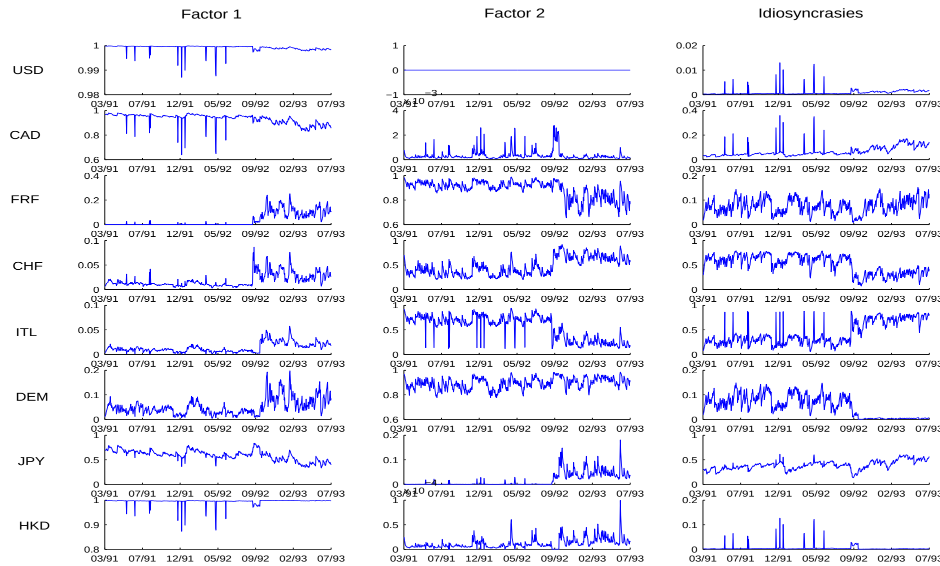

Figure 14: Two factor model: Proportion of the time series variances explained by each of the factors (common and specific), from 03/05/1991 to 07/05/1993.

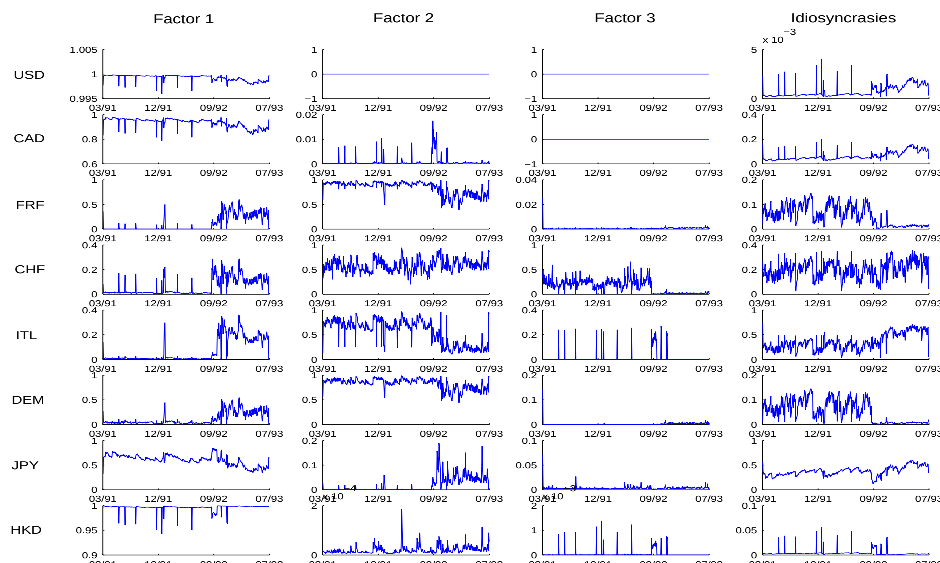

Figure 15: Three factor model: Proportion of the time series variances explained by each of the factors (common and specific), from 03/05/1991 to 07/05/1993.

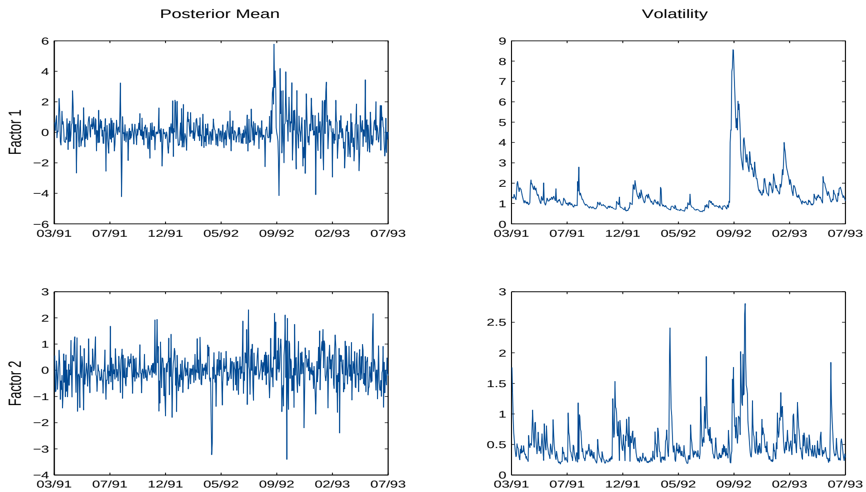

Figure 16: Two factor model: Common factors means and their volatilities, the diagonal elements of $\mathbf{H}_t$ (03/05/1991 to 07/05/1993).

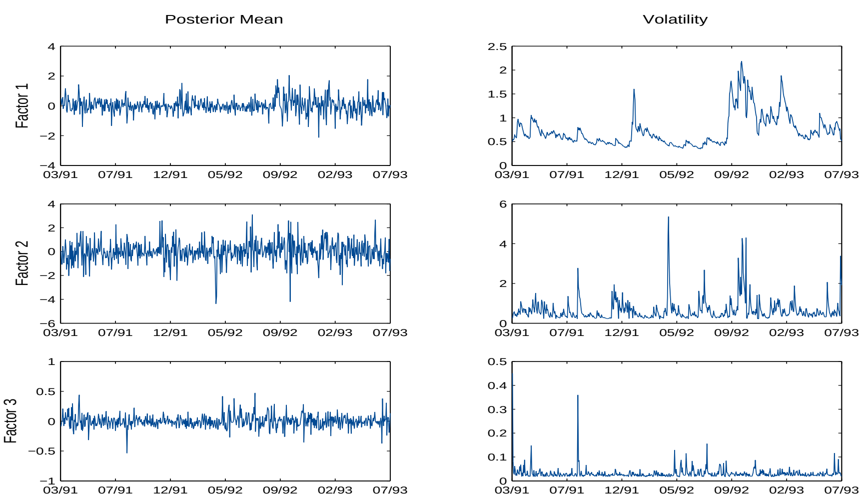

Figure 17: Three factor model: Common factors means and their volatilities, the diagonal elements of $\mathbf{H}_t$ (03/05/1991 to 07/05/1993).

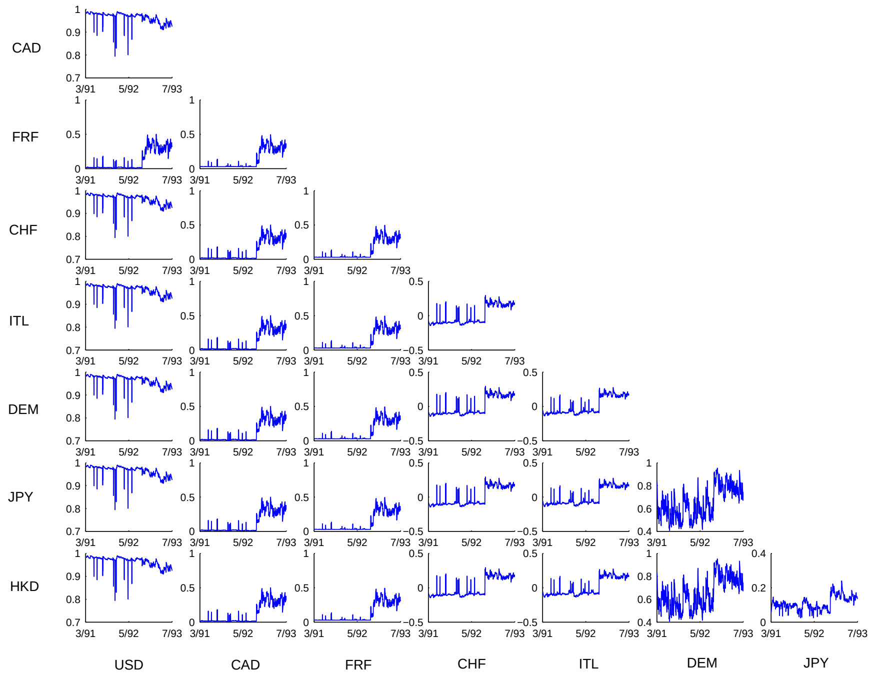

Figure 18: Two factor model: Time series co-dependence structure from 03/05/1991 to 07/05/1993.

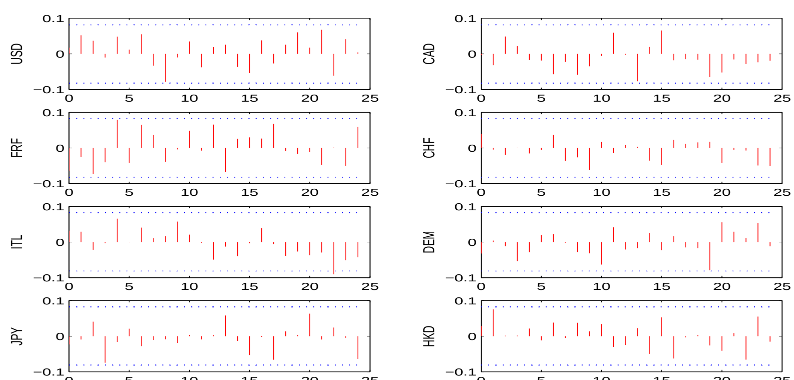

Figure 19: Two factor model: Autocorrelation functions of the residuals.

Table 9: Two factor model with conditional heteroscedasticity

	θ	$\mathbf{X}$		$diag(\mathbf{\Psi})$	ϕ_1	ϕ_2
	0.0127	0.7044	0.0000	0.0010	0.0860	0.0826
	-0.0082	0.6778	0.0178	0.0869	0.1071	0.1504
	-0.0414	0.1107	0.5790	0.0157	0.0826	0.1919
State 1	-0.0467	0.0693	0.6167	0.0817	0.8742	0.6332
	-0.0548	0.0790	0.4661	0.3194		
	-0.0457	0.0953	0.6159	0.0007		
	0.0915	0.4798	0.2731	0.3692		
	0.0120	0.6997	-0.0054	0.0015		
	-0.0035	1.0375	0.0000	0.0004	0.0860	0.0826
	-0.0159	0.9688	0.0237	0.0402	0.1071	0.1504
	0.0188	0.0030	0.3176	0.0031	0.0826	0.1919
State 2	0.0116	-0.0347	0.3163	0.0725	0.8742	0.6332
	0.0131	0.0211	0.3039	0.0134		
	0.0212	-0.0419	0.3289	0.0037		
	0.0129	0.5367	0.0277	0.1599		
	-0.0009	1.0302	0.0044	0.0032		
	0.0492	0.5576	0.0000	0.0034	0.0860	0.0826
	0.0042	0.5914	-0.0706	0.1615	0.1071	0.1504
	0.0290	0.0833	1.4265	0.0318	0.0826	0.1919
State 3	0.0562	0.1109	1.1904	0.4202	0.8742	0.6332
	-0.3031	0.0384	1.0897	1.7904		
	0.0067	0.0654	1.4910	0.0348		
	0.1477	0.4625	0.2418	0.2974		
	0.0315	0.5788	0.0009	0.0400		

1. From figure 14, It appears, that the second factor is responsible for time evolving movements in the variances of the European currencies, FRF, CHF, ITL and DEM. Figure 15 shows that the third factor seems to provide little or nocontribution to the fit of the model.

2. Looking at the common factor variances (figures 16 and 17), it can be argued that the first two factors are indeed more important than the third one. Figure 13 shows marked changes and increased volatility in the European currencies across the board towards the end of 1992 when Britain and Italy withdrew from the European exchange rate agreement. The impact of this event is evident in the estimated trajectories of both the posterior volatilities of currencies and in the corresponding variances of the common factors.

3. Notice also that the end-1992 volatility changes impact across all factors, highlighting the apparent dependencies in factor trajectories across the entire time period. This indicates the need for dependence structure in modeling latent volatility processes in dynamic factor analyses, as is allowed in the theoretical framework described in section 3. For the two common factors, the sum of the estimated α_i and δ_i is slightly less than

Table 10: Three factor model with conditional heteroscedasticity

	θ	$\mathbf{X}$			$diag(\mathbf{\Psi})$	ϕ_1	ϕ_2	ϕ_2
	0.0226	0.6805	0.0000	0.0000	0.0011	0.0692	0.5184	0.0788
	0.0001	0.6541	0.0090	0.0000	0.0875	0.0825	0.1122	0.1312
	-0.0379	0.2243	0.2386	-0.0894	0.0032	0.0605	0.2956	0.1586
State 1	-0.0381	0.1490	0.2396	0.4421	0.0732	0.9090	0.5884	0.1866
	-0.0514	0.2467	0.1897	0.0148	0.3175			
	-0.0396	0.2054	0.2434	0.2207	0.0018			
	0.1041	0.4463	0.1039	0.3130	0.3588			
	0.0219	0.6764	-0.0024	0.0008	0.0013			
	-0.0033	0.9921	0.0000	0.0000	0.0004	0.0692	0.5184	0.0788
	-0.0176	0.9312	0.0076	0.0000	0.0415	0.0825	0.1122	0.1312
	0.0202	0.0039	0.1181	0.0173	0.0032	0.0605	0.2956	0.1586
State 2	0.0106	-0.0260	0.1130	0.7365	0.0125	0.9090	0.5884	0.1866
	0.0151	0.0185	0.1139	-0.0002	0.0135			
	0.0229	-0.0400	0.1224	0.0121	0.0037			
	0.0110	0.5178	0.0059	0.2120	0.1573			
	-0.0006	0.9847	0.0022	-0.0155	0.0032			
	0.1452	0.8805	0.0000	0.0000	0.0024	0.0692	0.5184	0.0788
	0.1197	0.9175	-0.0775	0.0000	0.1673	0.0825	0.1122	0.1312
	0.0855	0.2714	0.6790	0.1739	0.0578	0.0605	0.2956	0.1586
State 3	0.1267	0.3729	0.6019	-0.6542	0.3393	0.9090	0.5884	0.1866
	-0.2990	0.1667	0.5007	3.4690	0.6658			
	0.0571	0.2660	0.7232	0.0800	0.0101			
	0.2352	0.7487	0.1151	0.2919	0.2972			
	0.1249	0.9215	-0.0026	-0.1860	0.0394			

one. It indicates strong GARCH effects and persistence in the volatility of exchange rates.

4. From figure 14 it can be observed that the first common factor explains at least 95% of the USD, CAD and HKD currency's variances at all times (and at least 99% before the 1992 crisis for the USD and HKD). This factor explains also 70% of the Japanese currency variances before August 1992 and 50% after this date. The contribution of the second factor in the variance of these currencies is negligible, except for the JPY where the contribution is about 10% after August 1992. As a matter of fact, the variance dynamics of the JPY are determined at 50% (at most) by its idiosyncratic variances after the 1992 crisis.

5. From this figure it can also be observed that the second factor is responsible for about 90% of the variances of the FRF, ITL and DEM currencies before the 1992 crisis. For the CHF, the contribution of the second factor is about 80% after August 1992 and 40% before this date. The contribution of the first factor in the variances of these European currencies is in particular meaningless before August 1992.

Broadly, the results show that all the correlations between the European currencies have increased just after August 1992 (figure 18). This is the effect of financial contagion that can be defined as a significant increase in comovement of financial prices such as the exchange rate, interest rates, and stock prices experienced by a group of countries, after controlling fundamentals (domestic news, ...) and common shocks, following a crisis elsewhere. The first factor represents the value of sterling relative to a basket of currencies in which the HKD, USD and CAD are dominant. Table 9 shows that the USD, CAD and HKD are roughly equally weighted, which is expected as CAD and HKD rates are heavily determined in international markets by USD rates. This first factor may be termed the North American factor. The second factor may be similarly termed the European Union factor. It represents a restricted basket of currencies dominated by the European currencies, with a relatively reduced weighting on JPY. USD, HKD and CAD are practically absent from this factor with $x_{1,2j}$, $x_{2,2j}$ and $x_{8,2j}$ for $j = 1, 2, 3$ indicating very small values. Inferences about idiosyncratic variances strengthen and extend these conclusions. Those of USD and DEM are very small, indicating that these two currencies play determining roles in defining their sector factor. CHF and ITL have larger idiosyncratic variances (during the crisis period), indicative of their departures from their sector factors. Finally, the sample autocorrelation functions of the residuals reported in figure 19 show no autocorrelation. The Ljung-Box statistic for the serial correlation of the squared residuals does not also reject the null hypothesis of uncorrelated squared residuals. Hence, all the covariance or correlations between the different exchange rate returns are explained by the common and specific factors.

9 Conclusion

Two powerful and popular tools of modeling conditionally heteroscedastic financial data characterized by dynamic changes over time (which can be attributed to various reasons, such as: wars, changes in economic policy and business cycles), in a multivariate framework, are linear factor models and HMM's. In this article we proposed a model that combines these tools and we called it Conditionally Heteroscedastic Factor Analysed Hidden Markov Model. We formulated the model and developed maximum likelihood estimates for its parameter. An iterative method based on the generalized EM principle, was adopted to circumvent the statistical inference problem inherent to the modeling of regime switching within the conditionally heteroscedastic structure of the common latent factors.

Our simulation experiments have demonstrated promising results in classification of the volatility behavior. Using two model selection criteria, we demonstrated accurate discrimination between specifications characterized by different hidden structures. We showed that switching conditionally heteroscedastic factor models provide more robust tracking performance than simple standard models or conditionally heteroscedastic models without regime switching. The fact that these models can be learned from data may be an important advantage in financial applications, where accurate on-line predictions of the time varying covariance matrices are very useful for dynamic asset allocation, active portfolio management and the analysis of options prices.

From a broader viewpoint, this study illustrates the usefulness of Markov regime switching models in the analysis of process that exhibit only local homogeneity. Such complex process can be found in a variety of scientific fields, and we believe the ideas presented here can be successfully applied in many such contexts. The analysis in this paper can be also extended in several ways. First, our model can be generalized to one where we allow the idiosyncratic variances to be a stochastic function of time. Secondly, we can also think of the case where the state transition probabilities are not homogeneous in time, but depend on the previous state and the previously observed covariates levels. The study of such models would provide a further step in the extension of hidden Markov models to dynamic factor analysis and allow for further flexibility in applications.

Appendix: Parameter Optimization

The parameter optimization scheme for conditionally heteroscedastic factorial HMMs based on the generalized expectation maximization (EM) algorithm is presented in this appendix. All the sufficient statistics are evaluated using the parameters from the previous iteration and therefore writing $\Theta^{(i)}$ explicitly is omitted for clarity. This derivation assumes that the first discrete state is always the initial state and all states are emitting. It is easy to extend the derivation for use with explicit initial discrete state probabilities and to include non-emitting states.

1- Initial State Probability Update Formulae

Discarding terms independent of the discrete initial state probabilities, the auxiliary function can be written as

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \sum_{j=1}^{m} M_{1/n}(j) \log(p(S_1))$$

Maximizing this function with respect to the discrete initial state probabilities, π_j , can be carried out using the Lagrange multiplier λ together with the sum to unity constraint $\sum_{j=1}^{m} \pi_j = 1$. It is equivalent to maximising the following Lagrangian

$$g(\pi_j) = \sum_{i=1}^{m} M_{1/n}(i) \log(\pi_j) + \lambda \left(1 - \sum_{i=1}^{m} \pi_i\right)$$

Differentiating $g(\pi_j)$ yields

$$\begin{cases} \frac{\partial g(\pi_j)}{\partial \pi_j} = \frac{M_{1/n}(j)}{\pi_j} - \lambda \\ \frac{\partial g(\pi_j)}{\partial \lambda} = 1 - \sum_{i=1}^{m} \pi_i \end{cases}$$

Setting the derivative to zero together with the sum to unity constraint forms the following pair of equations and solving for π_j, the new discrete initial state probabilities can be written as

$$\widehat{\pi}_j = \frac{M_{1/n}(j)}{\sum\limits_{i=1}^{m} M_{1/n}(i)}$$

This is a maximum of $g(\pi_j)$ since its second derivative with respect to π_j is negative.

2- Transition State Probability Update Formulae

Discarding terms independent of the discrete state transition probabilities, the auxiliary function can be written as

$$\mathcal{Q}(\Theta, \Theta^{(i)}) = \sum_{t=2}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} M_{t-1,t/n}(i,j) \log(p_{ij})$$

Maximizing this function with respect to the discrete state transition probabilities, p_{ij} , can be carried out using the Lagrange multiplier λ together with the sum to unity constraint $\sum\limits_{j=1}^{m} p_{ij} = 1$. It is equivalent to maximising the following Lagrangian

$$g(p_{ij}) = \lambda(1 - \sum_{j=1}^{m} p_{ij}) + \sum_{t=2}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} M_{t-1,t/n}(i,j) \log(p_{ij})$$

Differentiating $g(p_{ij})$ yields

$$\frac{\partial g(p_{ij})}{\partial p_{ij}} = -\lambda + \sum_{t=2}^{n} \frac{M_{t-1,t/n}(i,j)}{p_{ij}}$$

Setting the derivative to zero together with the sum to unity constraint forms the following pair of equations

$$\begin{cases} -\lambda + \sum\limits_{t=2}^{n} \frac{M_{t-1,t/n}(i,j)}{p_{ij}} = 0 \\ 1 - \sum\limits_{j=1}^{m} p_{ij} = 0 \end{cases}$$

Solving for p_{ij}, the new discrete state transition probabilities can be written as

$$\widehat{p}_{ij} = \frac{\sum\limits_{t=2}^{n} M_{t-1,t/n}(i,j)}{\sum\limits_{t=2}^{n} M_{t-1/n}(i)}$$

This is a maximum of $g(p_{ij})$ since its second derivative with respect to p_{ij} is negative.

3- Factor Loadings Update Formulae

Let $\mathbf{x}_{jl}$ denote the *l-th* row vector of $\mathbf{X}_j$. Maximizing the Equation (7) is equivalent to maximizing

$$g(\mathbf{x}_{jl}) = -\frac{1}{2}\sum_{l=1}^{q}\Big[\mathbf{x}_{ji}\mathbf{G}_{jl}\mathbf{x}'_{jl} - \mathbf{x}_{jl}\mathbf{k}_{jl}\Big]$$

where the k by k matrices $\mathbf{G}_{jl}$ and k-dimensional column vectors $\mathbf{k}_{jl}$ are defined as follows

$$\begin{aligned}\mathbf{G}_{jl} &= \frac{1}{\psi_{jl}}\sum_{t=1}^{n} M_{t/n}(j)\Big[\mathbf{H}^j_{t/n} + \mathbf{f}^j_{t/n}\mathbf{f}^{j\prime}_{t/n}\Big] \\ \mathbf{k}_{jl} &= \frac{1}{\psi_{jl}}\sum_{t=1}^{n} M_{t/n}(j)(y_{tl} - \theta_{jl})\mathbf{f}^j_{t/n}\end{aligned}$$

where ψ_{jl} is the *l-th* diagonal element of the idiosyncratic covariance matrix $\mathbf{\Psi}_j$, y_{tl} and θ_{jl} are the *l-th* elements of the current observation and the idiosyncratic noise mean vectors, respectively.

Differentiating $g(\mathbf{x}_{jl})$ yields

$$\frac{\partial g(\mathbf{x}_{jl})}{\partial \mathbf{x}_{jl}} = -\mathbf{G}_{jl}\mathbf{x}'_{jl} + \mathbf{k}_{jl}$$

Setting the derivative to zero and solving for $\mathbf{x}_{jl}$ results in the updated row vector of the factor loading matrix

$$\widehat{\mathbf{x}}_{jl} = \mathbf{k}'_{jl}\mathbf{G}^{-1}_{jl}$$

This is a maximum since the second derivative of $g(\mathbf{x}_{jl})$ with respect $\mathbf{x}_{jl}$ is negative.

4- Observation Noise Mean Update Formulae

Differentiating the auxiliary function in equation (7) with respect to the observation noise mean vector, θ_j, yields

$$\frac{\partial \mathcal{Q}(\Theta,\Theta^{(i)})}{\partial \theta_j} = \mathbf{\Psi}^{-1}_j\sum_{t=1}^{n} M_{t/n}(j)\left(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}^j_{t/n} - \theta_j\right)$$

Equating this to zero and solving for θ_j result in the updated observation noise mean vector

$$\widehat{\theta}_j = \frac{\sum_{t=1}^{n} M_{t/n}(j)\left(\mathbf{y}_t - \mathbf{X}_j\mathbf{f}^j_{t/n}\right)}{\sum_{t=1}^{n} M_{t/n}(j)}$$

This is a maximum since the second derivative of $\mathcal{Q}(\Theta/\Theta^{(i)})$ with respect to θ_j is negative.

5- Idiosyncratic Variances Update Formulae

Applying some matrix manipulations and discarding terms independent of the idiosyncratic noise covariance matrix, $\mathbf{\Psi}_j$, the auxiliary function in equation (7) may be rewritten as

$$\begin{aligned}\mathcal{Q}(\Theta,\Theta^{(i)}) &= -\frac{1}{2}\sum_{t=1}^{n}\sum_{j=1}^{m} M_{t/n}(j)\Bigg(\log|\mathbf{\Psi}_j| + tr\Bigg\{\mathbf{\Psi}_j^{-1}(\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{f}_{t/n}^{j}\mathbf{y}_t' \\ \mathbf{y}_t' \end{bmatrix} \\ &- \begin{bmatrix} \mathbf{y}_t\mathbf{f}_{t/n}^{j\prime} & \mathbf{y}_t \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix} + \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{H}_{t/n}^{j} + \mathbf{f}_{t/n}^{j}\mathbf{f}_{t/n}^{j\prime} & \mathbf{f}_{t/n}^{j} \\ \mathbf{f}_{t/n}^{j\prime} & 1 \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix}\Bigg\}\Bigg)\end{aligned}$$

To find the new idiosyncratic noise covariance matrix, the auxiliary function above is differentiated with respect to its inverse, $\mathbf{\Psi}_j^{-1}$, and equated to zero. Solving for $\mathbf{\Psi}_j$ and setting the off-diagonal elements to zeroes result in the updated idiosyncratic noise covariance matrix

$$\begin{aligned}\widehat{\mathbf{\Psi}}_j = \frac{1}{\sum\limits_{t=1}^{n} M_{t/n}(j)}\sum_{t=1}^{n} M_{t/n}(j) diag\Bigg\{&\mathbf{y}_t\mathbf{y}_t' - \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{f}_{t/n}^{j}\mathbf{y}_t' \\ \mathbf{y}_t' \end{bmatrix} - \begin{bmatrix} \mathbf{y}_t\mathbf{f}_{t/n}^{j\prime} & \mathbf{y}_t \end{bmatrix} \\ &\times \begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix} + \begin{bmatrix} \mathbf{X}_j & \theta_j \end{bmatrix}\begin{bmatrix} \mathbf{H}_{t/n}^{j} + \mathbf{f}_{t/n}^{j}\mathbf{f}_{t/n}^{j\prime} & \mathbf{f}_{t/n}^{j} \\ \mathbf{f}_{t/n}^{j\prime} & 1 \end{bmatrix}\begin{bmatrix} \mathbf{X}_j' \\ \theta_j' \end{bmatrix}\Bigg\}\end{aligned}$$

This is a maximum since the second derivative of $\mathcal{Q}(\Theta/\Theta^{(i)})$ with respect to $\mathbf{\Psi}_j^{-1}$ is negative.

References

[1] H. Akaike, A new look at the statististical identification model, *IEEE Trans. Auto. Control*, 1974; **19**: 716–723.

[2] B.O. Anderson, J.B. Moore, Optimal Filtering, Englewood Cliffs, NJ: Prentice Hall, 1979.

[3] Bar-Shalom Y. and Li X.R, Estimation and tracking: principles, techniques and software. *Boston, London: Artech House Inc*, 1993.

[4] Bera A. K. et Jarque C. M. Model specification tests: A simultaneous approach. *Journal of Econometrics*, 1982; **20**: 59–82.

[5] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 1986; **31**: 307–327.

[6] Bollerslev T. A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics*, 1987; **69**: 542–547.

[7] Bollerslev T., Chou R. and Kroner K. ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence. *Journal of Econometrics*, 1992; **52**: 5–59.

[8] Bollerslev T., Engle R. and Nelson D. ARCH Models, *Journal of Econometrics*, 1994; **52**: 5–59.

[9] Chou R. Volatility Persistence and Stock Valuations: Some Empirical Evidence Using GARCH. *Journal of Applied Econometrics*, 1988; **3**: 279–294.

[10] Connor, G. and R.A. Korajczyk, The Arbitrage Pricing Theory and Multifactor Models of Asset Returns, in: R.A. Jarrow, V. Maksimovic and W.T. Ziemba, eds., Handbooks in Operations Research and Management Science (North-Holland, Amsterdam), 1995; **9**: 87–144.

[11] A. Demos, E. Sentana, An EM Algorithm for Conditionally Heteroscedastic Factor Models. *Journal of Business & Economic Statistics*, 1998; **16**: 357–361.

[12] A. Dempster, N. Laird, D. Rubin, Maximum Likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society (Series B)*, 1977; **39**: 1–38.

[13] F. Diebold, M. Nerlove, The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. *Journal of Applied Econometrics*, 1988; **4**: 1–21.

[14] R. Engle, Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 1982; **50**: 987–1008.

[15] R. Engle, V.K. Ng, M. Rothschild, A Multi-Dynamic Factor Model for Stock Returns. *Journal of Econometrics*, 1992; **52**: 245–266.

[16] Engle R. and Ng V. Time Varying Volatility and the Dynamic Behavior of the Term Structure. *Journal of Money Credit and Banking*, 1993; **25**: 336–349.

[17] R. Engle, R. Susmel, Common Volatility in International Equity Markets. *Journal of Business & Economic Statistics*, 1993; **11**: 369–380.

[18] French K., Schwert G. and Stambaugh R. Expected Stock Returns and Volatility. *Journal of Financial Economics*, 1987; **19**: 3–30.

[19] Geweke J.F., and Singleton K.J. Interpreting the Likelihood Ratio Statistic in Factor Models when Sample Size is Small. *Journal of the American Statistical Association*, 1980; **75**: 133–137.

[20] Geweke J.F., and Zhou G. Measuring the Pricing Error of the Arbitrage Pricing Theory. *The Review of Financial Studies*, 1996; **9**: 557–587.

[21] A.K. Gupta, Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika*, 1952; **39**: 260–73.

[22] Hamilton J. Analysis of Time Series Subject to Changes in Regime. *Journal of Econometrics*, 1990; **45**: 39–70.

[23] A. Harvey, E. Ruiz, E. Sentana, Unobserved component time series models with ARCH disturbances. *Journal of Econometrics*, 1992; **52**: 129–157.

[24] Juang B. H. and L. R. Rabiner A Probabilistic Distance Measure for Hidden Markov Models. *AT&T Technical Journal*, 1985; **64**: 391–408.

[25] R.E. Kalman, R.S. Bucy, New results in linear filtering and prediction. *Journal of Basic Engineering (ASME)*, 1961; **83**: 95–108.

[26] Kim C.J. Dynamic linear models with Markov switching. *Journal of Econometrics*, 1994; **60**: 1–22.

[27] C. Lamoureux, W. Lastrapes, Persistence in Variance, Structural Change, and the GARCH Model. *Journal of Business & Economic Statistics*, 1990; **8**: 225–234.

[28] Lauritzen S. Graphical Models. *OUP*, 1996.

[29] W. Lin, Alternative estimators for factor GARCH models: a Monte Carlo comparaison. *Journal of Applied Econometrics*, 1992; **7**: 259–279.

[30] Ljung G. and Box G. On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 1978; **67**: 297–303.

[31] McConnell M. and Perez-Quiros G. Output Fluctuations in the United States: What Has Changed since the 1980s? *American Economic Review*, 2000; **90**: 1464–1476.

[32] Press S.J. Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference. *California: Krieger*, 1985.

[33] L.R. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Prentice Hall, 1993.

[34] H.E. Rauch, F. Tung, C.T. Striebel, Maximum likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronautics Journal*, 1965; **8**: 1445–1450.

[35] S.A. Ross, The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 1976; **13**: 341–360.

[36] S.A. Ross, Risk, return and arbitrage, In I. Friend & J. Bicksler, editor, Risk, Return in Finance. Ballinger, Cambridge, Mass, 1977.

[37] A.V.I Rosti, M.J.F. Gales M.J.F, Generalised Linear Gaussian Models, CUED/F-INFENG/TR.420, Cambridge University, 2001.

[38] Saul, L. and M. Rahim, Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2000; **8**: 115–125.

[39] G. Schwarz, Estimating the dimension of a model. *Annals of Statistics*, 1978; **6**: 461–464.

[40] E. Sentana, Quadratic ARCH Models. *Review of Economic Studies*, 1995; **62**: 639–661.

[41] S.S. Shapiro, R.S. Francia, An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, 1972; **67**: 215–216.

[42] S.S. Shapiro, M.B. Wilk, An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 1965; **52**: 591–611.

[43] W.F. Sharpe, A simplified model for portfolio analysis. *Management Science*, 1963; **9**: 277–293.

[44] M.A. Stephens, Asymptotic properties for covariance matrices of order statistics. *Biometrika*, 1975; **62**: 23–28.

Contents

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399