

Contributions to The Estimation of Mixed-State Conditionally Heteroscedastic Latent Factor Models: A Comparative Study

Mohamed Saidane*, The University of 7 November at Carthage
ISCC Bizerte, Zarzouna 7021 (Bizerte) Tunisia
E-mail: Mohamed.Saidane@isg.rnu.tn

Christian Lavergne, I3M, UMR-CNRS 5149, University of Montpellier 2
Place Eugène Bataillon CC 051 – 34095 – France
E-mail: lavergne@math.univ-montp2.fr

Current version, June 2008

Abstract

Mixed-State conditionally heteroscedastic latent factor models attempt to describe a complex nonlinear dynamic system with a succession of linear latent factor models indexed by a switching variable. Unfortunately, despite the framework's simplicity exact state and parameter estimation are still intractable because of the interdependency across the latent factor volatility processes. Recently, a broad class of learning and inference algorithms for time series models have been successfully cast in the framework of dynamic Bayesian networks (DBN). This paper describes a novel DBN-based switching conditionally heteroscedastic latent factor model. The key methodological contribution of this paper is the novel use of the Generalized Pseudo-Bayesian method GPB2, a structured variational learning approach and an approximated version of the Viterbi algorithm in conjunction with the EM algorithm for overcoming the intractability of exact inference in mixed-state latent factor model. The conditional EM algorithm that we have developed for the maximum likelihood estimation, is based on an extended switching Kalman filter approach which yields inferences about the unobservable path of the common factors and their variances, and the latent variable of the state process. Extensive Monte Carlo simulations show promising results for tracking, interpolation, synthesis, and classification using learned models.

AMS Subject Classification: 62H25, 62M05, 62M10 and 62P20

Keywords: Latent Factor Models; EM Algorithm; GQARCH Processes; HMM; Viterbi Approximation, GPB method, Variational approximation, Time series segmentation, Finance.

*Correspondence to: Dr. Mohamed Saidane, Université du 7 Novembre à Carthage, Institut Supérieur de Commerce et de Comptabilité de Bizerte Zarzouna 7021, Bizerte Tunisia. Tel.: +216-71-719-452, **E-mail** : Mohamed.Saidane@isg.rnu.tn

1 Introduction

Most traditional time series models are based on the assumption of stationarity: the underlying generator of the data is assumed to be globally time invariant. However, it is well known that for many financial time series this assumption breaks down. For instance, one of the obstacles to the effective forecasting of exchange rates is a nonconstant conditional variance, known as heteroscedasticity. GARCH models have been developed to estimate a time-dependent variance (see Bollerslev, 1986).

A local assumption of stationarity is nevertheless acceptable if the system switches between different regimes but each regime is (approximately) locally stationary. In fields from econometrics to control engineering, hybrid approaches have been developed in order to model this behavior. One example is the mixture of experts (see Jacobs et al., 1991), (Shi and Weigend, 1997) which decomposes the global model into several (linear or non-linear) local models (known as experts as each specialises in modeling a small region of input space). One limitation of these models is that the gating network which combines the local models has no dynamics. It is controlled only by the current value of the time series. One way to address this limitation is to use a hidden Markov model (which does have dynamics) to switch between local models. Autoregressive hidden Markov models (ARHMMs) switch between autoregressive models, where the predictions are a linear combination of past values. ARHMMs have been applied to financial engineering in order to model high frequency foreign exchange data and have shown promising results (Shi and Weigend, 1997).

Switching conditionally heteroscedastic latent factor models consist of multiple linear factor models controlled by a dynamic switch. These models assume that the behavior of the system can be characterized by a finite number of conditionally heteroscedastic latent factor models with hidden states, each of which tracks the dynamics in a different regime. The approach is motivated by the fact that market behavior at different time periods might be explained by different underlying regimes. Using a switching conditionally heteroscedastic latent factor model allows us both to create a predictive model and to discover at what times transitions occur between regimes (i.e. to segment the time series).

A long standing limitation of these models is that the complexity of the exact training algorithm grows exponentially with order m^n , where m is the number of models and n is the length of the time sequence. During the last decade, various approximations have been proposed and studied theoretically and numerically in order to overcome the complexity problems related to the inference of latent structures in switching state space models (see Ghahramani and Hinton, 1998 and others). Recently, Saidane and Lavergne (2007a, 2007b) introduced the switching conditionally heteroscedastic latent factor model and proposed two efficient and principled approximate algorithms for training these models in a maximum likelihood approach. In this paper a new expectation maximization (EM) algorithm combined with a mixed-state version of the Viterbi algorithm is derived for maximum likelihood estimation.

The remaining of this article is organized as follows. In section 2, we introduce the general form of the model. In sections 3 and 4, we show how the parameters can be learned by

using the generalized pseudo bayesian algorithm (Saidane and Lavergne, 2007a), the structured variational method (Saidane and Lavergne, 2007b) and the approximated version of the Viterbi algorithm, which constitutes the major contribution of this paper. In section 5, the model selection problem is considered and we derived possible penalized criteria for choosing among several specific models. In the last section we evaluate, through a simulation study, the performance of the new maximum likelihood approach. We demonstrate the application of these learned models to segmentation and tracking tasks.

2 The Mixed-State Latent Factor Model

The model that we propose supposes that excess returns depend both on unobservable factors that are common across the multivariate time series, and on unobservable different regimes that describe the different states of volatility. This new specification, called switching conditionally heteroscedastic latent factor model, is defined by:

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

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 (financial returns in our case). The HMM state transition probabilities from state i to state j are represented by p_{ij} . In an unspecified state $S_t = j$, $\mathbf{0}$ and \mathbf{H}_{j_t} 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 Ψ_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. In this framework the common variances (diagonal elements of \mathbf{H}_{j_t}) are supposed to be time varying and their parameters change according to the regime. In particular, we suppose that these variances follow switching Generalized Quadratic Autoregressive Conditionally Heteroscedastic processes GQARCH(1,1), the l -th diagonal element of the matrix \mathbf{H}_{j_t} under a particular regime $S_t = j$ since $S_{t-1} = i$ is given by:

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

To guarantee the positivity of the conditional common variances and the covariance stationarity, we impose the constraints $w_j^l, \alpha_j^l, \delta_j^l > 0$, $\gamma_j^{l2} \leq 4\omega_j^l \alpha_j^l$ and $\alpha_j^l + \delta_j^l < 1$, $\forall j, l$. For model identification we suppose that $q \geq k$ and $rank(\mathbf{X}_j) = k$, $\forall j$. We suppose also

¹ The \sim symbol in $S_t \sim P(S_t | S_{t-1})$ is used to represent a discrete Markov chain.

that the common and idiosyncratic factors are uncorrelated, and that \mathbf{f}_t and $\varepsilon_{t'}$ are mutually independent for all t, t' (for more details on the identification problem, the reader is referred to Sentana and Fiorentini, 2001 and Carneo et al., 2004).

3 Inference in Mixed-Sate Latent Factor Models

The goal of inference in complex mixed-sate latent factor models is to estimate the posterior probability of the hidden states of the system (S_t and \mathbf{f}_t) given some known sequence of observations $\mathcal{Y}_n = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ and the known model parameters. Specifically, we need to find the sufficient statistics of the posterior $p(\mathcal{F}_n, \mathcal{S}_n | \mathcal{Y}_n)$. Given the form of p it is easy to show that these are the first and the second order statistics: mean and covariance among hidden states $\mathbf{f}_t, \mathbf{f}_{t-1}, S_t, S_{t-1}$.

If there were no switching dynamics, the inference would be straightforward – we could infer \mathbf{f}_t from \mathcal{Y}_n using Rauch-Tung-Striebel (RTS) smoothing (Rauch, 1963; Rauch et al., 1965). However, the presence of switching dynamics embedded in the transition matrix \mathbf{P} makes exact inference more complicated. To see that, assume that the initial distribution of \mathbf{f}_0 at $t = 0$ is Gaussian, at $t = 1$ the probability density function of the physical system state becomes a mixture of m Gaussian densities since we need to marginalize over m possible but unknown plant models. At time t we will have a mixture of m^t Gaussians, which is clearly intractable for even moderate sequence lengths. It is therefore necessary to explore approximate inference techniques that will result in a tractable learning method.

A generalized pseudo-Bayesian inference method and a structured variational learning approach were presented in Saidane and Lavergne (2007a, 2007b) and evaluated experimentally. We briefly review them in sections 3.1 and 3.2. In Section 3.3 we present our new approximate Viterbi inference algorithm. We then present in section 6 an extensive comparative simulation study of the three proposed algorithms in order to verify the correctness and effectiveness of these methods.

3.1 Approximate Generalized Pseudo Bayesian Inference

The Generalized Psuedo Bayesian (Bar-Shalom and Li, 1993; Kim, 1994) GPB approximation scheme is based on the general idea of "collapsing", i.e. representing a mixture of m^t Gaussians with a mixture of m^r Gaussians, where $r < t$ (see Murphy, 1998 for a detailed review). While there are several variations on this idea, our focus is the GPB2 algorithm (Saidane and Lavergne, 2007a), which maintains a mixture of m^2 Gaussians over time and can be reformulated to include smoothing as well as filtering.²

GPB2 is closely related to the Viterbi approximation of section 3.3. It differs in that instead of picking the most likely previous switching state i at every time step t and

²Other similar pseudo Bayesian algorithms of Bar-Shalom and Li (1993), GBP1 and *IMM*, do not have an obvious smoothing reformulation. The interacting multiple models *IMM*, can be obtained by first collapsing the prior to a single Gaussian (by moment matching), and then updating it using M different Kalman filters, one per value of S_t . Unfortunately, it is hard to extend *IMM* to the smoothing case, unlike GPB2.

switching state j , we collapse the m Gaussians (one for each possible value of i) down into a single Gaussian.

3.1.1 A Switching State-Space Representation

Our switching conditionally heteroscedastic factor model developed in section 2 can be regarded as a random field with indices $i = 1, \dots, q$, $t = 1, \dots, n$ and $j = 1, \dots, m$. The idiosyncratic covariance matrix is assumed diagonal, and the variances of the factors are parameterized as univariate ARCH models, but taking into account that the values of the factors are unobserved. In particular for the GQARCH(1,1) formulation of Sentana (1995) the state-space representation of our model, with continuous state variable \mathbf{f}_t , is given by:

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

For the implementation of the filtering and smoothing algorithms, we start by introducing some notation.

$$\begin{aligned} \mathbf{f}_{t|\tau}^{i(j)} &= \mathbb{E}[\mathbf{f}_t | \mathcal{Y}_{1:\tau}, S_{t-1} = i, S_t = j] \\ \mathbf{f}_{t|\tau}^{(j)k} &= \mathbb{E}[\mathbf{f}_t | \mathcal{Y}_{1:\tau}, S_t = j, S_{t+1} = k] \\ \mathbf{f}_{t|\tau}^j &= \mathbb{E}[\mathbf{f}_t | \mathcal{Y}_{1:\tau}, S_t = j] \\ h_{t|\tau}^j &= \text{Var}(f_{jt} | \mathcal{Y}_{1:\tau}, S_t = j) \\ h_{t|\tau}^{i(j)} &= \text{Var}(f_{jt} | \mathcal{Y}_{1:t-1}, S_{t-1} = i, S_t = j) \\ M_{t-1,t|\tau}(i, j) &= p(S_{t-1} = i, S_t = j | \mathcal{Y}_{1:\tau}) \\ M_{t|\tau}(j) &= p(S_t = j | \mathcal{Y}_{1:\tau}) \end{aligned}$$

3.1.2 Filtering Algorithm

We perform the following steps in sequence.

$$\mathbf{f}_{t|t-1}^{i(j)} = \mathbf{0} \cdot \mathbf{f}_{t-1|t-1}^i = \mathbf{0} \quad \forall \quad i, j = 1, \dots, m \quad \text{and} \quad (1)$$

$$h_{t|t-1}^{i(j)} = w_{lj} + \gamma_{lj} f_{lt-1|t-1}^i + \alpha_{lj} \left[f_{lt-1|t-1}^{i2} + h_{lt-1|t-1}^i \right] + \delta_{lj} h_{lt-1|t-2}^i \quad (2)$$

Then we compute the prediction error $\mathbf{e}_t(i, j) = \mathbf{y}_t - \theta_j - \mathbf{X}_j \mathbf{f}_{t|t-1}^{i(j)}$, the variance of the error $\Sigma_{t|t-1}^{i(j)} = \mathbf{X}_j \mathbf{H}_{t|t-1}^{i(j)} \mathbf{X}_j' + \Psi_j$, the Kalman gain matrix $K_t(i, j) = \mathbf{H}_{t|t-1}^{i(j)} \mathbf{X}_j' \Sigma_{t|t-1}^{i(j)-1}$, the likelihood of this observation $L_t(i, j) = \mathcal{N}[\mathbf{0}, \Sigma_{t|t-1}^{i(j)}]$ and we update our estimates of the mean and variance:

$$\mathbf{f}_{t|t}^{i(j)} = \mathbf{f}_{t|t-1}^{i(j)} + K_t(i, j) \mathbf{e}_t(i, j) \quad (3)$$

$$\mathbf{H}_{t|t}^{i(j)} = \mathbf{H}_{t|t-1}^{i(j)} - K_t(i, j) \Sigma_{t|t-1}^{i(j)} K_t(i, j)' \quad (4)$$

The fundamental problem with switching Kalman filters is that the belief state grows exponentially with time. To dealing with this problem 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 (GPBR). When $r = 1$, we approximate a mixture of Gaussians with a single Gaussian using moment matching; this can be shown (e.g., Lauritzen, 1996) to be the best (in the Kullback-Leibler sense) single Gaussian approximation. For the implementation of this algorithm we calculate the probabilities

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

where

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

and

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

Finally, we update our estimates of the mean and volatilities.

$$\begin{aligned} \mathbf{f}_{t|t}^j &= \sum_{i=1}^m Z_{i|j}(t) \mathbf{f}_{t|t}^{i(j)} \\ h_{t|t}^j &= \sum_{i=1}^m Z_{i|j}(t) h_{t|t}^{i(j)} + \sum_{i=1}^m Z_{i|j}(t) \left[f_{t|t}^{i(j)} - f_{t|t}^j \right] \left[f_{t|t}^{i(j)} - f_{t|t}^j \right]' \\ h_{t|t-1}^j &= \sum_{i=1}^m Z_{i|j}(t) h_{t|t-1}^{i(j)} + \sum_{i=1}^m Z_{i|j}(t) \left[f_{t|t-1}^{i(j)} - f_{t|t-1}^j \right] \left[f_{t|t-1}^{i(j)} - f_{t|t-1}^j \right]' \end{aligned}$$

3.1.3 Smoothing Algorithm

Given the degenerate nature of the (time-series) transition equation, the smoother gain matrix is always null, hence smoothing is unnecessary in this case: $\mathbf{f}_{t|n}^{(j)k} = \mathbf{f}_{t|t}^j$ and $\mathbf{H}_{t|n}^{(j)k} = \mathbf{H}_{t|t}^j$. For updating the parameters, we have need of the probabilities: $M_{t,t+1|n}(j, k) = U_{t|t+1}^{j|k} M_{t+1|n}(k)$ and $M_{t|n}(j) = \sum_{k=1}^m M_{t,t+1|n}(j, k)$, where

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

the approximation arises because S_t is not conditionally independent of the future evidence $\mathbf{y}_{t+1}, \dots, \mathbf{y}_n$, 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} .

3.2 Approximate Variational Inference

An other efficient learning algorithm for the parameters of our switching factor model can be derived by generalizing the Expectation Maximization (EM) algorithm (Dempster, et al., 1977). EM alternates between optimizing a distribution over the hidden states (the E-step) and optimizing the parameters given the distribution over hidden states (the M-step). Any distribution over the complete sequence of hidden states, $Q(\mathcal{S}, \mathcal{F})$, can be used to define a lower bound, \mathcal{B} , on the log-probability of the observed data:

$$\begin{aligned} \log p(\mathcal{Y}|\Theta) &= \log \left[\sum_{\mathcal{S}} \int p(\mathcal{S}, \mathcal{F}, \mathcal{Y}|\Theta) d\mathcal{F} \right] \\ &= \log \left[\sum_{\mathcal{S}} \int Q(\mathcal{S}, \mathcal{F}) \left\{ \frac{p(\mathcal{S}, \mathcal{F}, \mathcal{Y}|\Theta)}{Q(\mathcal{S}, \mathcal{F})} \right\} d\mathcal{F} \right] \\ &\geq \sum_{\mathcal{S}} \int Q(\mathcal{S}, \mathcal{F}) \log \left\{ \frac{p(\mathcal{S}, \mathcal{F}, \mathcal{Y}|\Theta)}{Q(\mathcal{S}, \mathcal{F})} \right\} d\mathcal{F} = \mathcal{B}(Q, \Theta) \end{aligned} \quad (5)$$

where Θ denotes the parameters of the model and we have made use of Jensen's inequality to establish (5). The E-step holds the parameters fixed and sets Q to be the posterior distribution over the hidden states given the parameters,

$$Q(\mathcal{S}, \mathcal{F}) = P(\mathcal{S}, \mathcal{F}|\mathcal{Y}, \Theta)$$

This maximizes \mathcal{B} with respect to the distribution, turning the lower bound into an equality, which can be easily seen by substitution. The M-step holds the distribution fixed and computes the parameters that maximize \mathcal{B} for that distribution. Given the change in the parameters produced by the M-step, the distribution produced by the previous E-step is typically no longer optimal, so the whole procedure must be iterated.

Unfortunately, the exact E-step for our switching conditionally heteroscedastic factor model is intractable, because the posterior probability of the real-valued states is a Gaussian mixture with m^n terms. In order to derive an efficient learning algorithm for this system, we relax the EM algorithm by approximating the posterior probability of the hidden states. The basic idea is that, since expectations with respect to P are intractable, rather than setting $Q(\mathcal{S}, \mathcal{F}) = P(\mathcal{S}, \mathcal{F}|\mathcal{Y})$ in the E-step, a tractable distribution Q is used to approximate P . The difference between the bound \mathcal{B} and the log likelihood is given by the Kullback-Liebler (KL) divergence between Q and P :

$$KL(Q||P) = \sum_{\mathcal{S}} \int Q(\mathcal{S}, \mathcal{F}) \log \left[\frac{Q(\mathcal{S}, \mathcal{F})}{P(\mathcal{S}, \mathcal{F}|\mathcal{Y})} \right] d\mathcal{F}$$

While there are many possible approximations to the posterior distribution of the hidden variables that one could use for learning and inference in switching factor models, we focus on the following:

$$Q(\mathcal{S}, \mathcal{F}) = \frac{1}{Z_Q} \left[\mathcal{P}(S_1) \prod_{t=2}^n \mathcal{P}(S_t | S_{t-1}) \right] \left[\mathcal{P}(\mathbf{f}_1) \prod_{t=2}^n \mathcal{P}(\mathbf{f}_t | \mathbf{f}_{t-1}) \right]$$

where the \mathcal{P} are unnormalized probabilities, which we will call potential functions and define soon, and Z_Q is a normalization constant ensuring that Q integrates to one. The terms involving the switch variables S_t define a discrete Markov chain and the terms involving the state vectors \mathbf{f}_t define m uncoupled factor models. Like in mean field approximations we have removed the stochastic coupling between the chains that results from the fact that the observation at time t depends on all the hidden variables at time t . However, we retain the coupling between the hidden variables at successive time steps since these couplings can be handled exactly using the forward-backward and Kalman smoothing recursions. The discrete switching process is defined by

$$\begin{aligned} \mathcal{P}(S_1 = j) &= p(S_1 = j) q_1^{(j)} \\ \mathcal{P}(S_t = j | S_{t-1}) &= p(S_t = j | S_{t-1}) q_t^{(j)} \end{aligned}$$

where the $q_t^{(j)}$ are variational parameters of the Q distribution. These parameters scale the probabilities of each of the states of the switch variable at each time step, so that $q_t^{(j)}$ plays exactly the same role as the observation probability $p(\mathbf{y}_t | S_t = j)$ would play in a regular hidden Markov model (see Saidane and Lavergne, 2006).

The uncoupled factor models in the approximation Q are also defined by potential functions which are related to probabilities in the original system. These potentials are the prior and transition probabilities for \mathbf{f}_t multiplied by a factor that changes these potentials to try to account for the data:

$$\begin{aligned} \mathcal{P}(\mathbf{f}_1^j) &= [p(\mathbf{f}_1 | S_1 = j) p(\mathbf{y}_1 | \mathbf{f}_1, S_1 = j)]^{\xi_1^{(j)}} \\ \mathcal{P}(\mathbf{f}_t^j | \mathbf{f}_{t-1}) &= [p(\mathbf{f}_t | \mathbf{f}_{t-1}, S_t = j) p(\mathbf{y}_t | \mathbf{f}_t, S_t = j)]^{\xi_t^{(j)}} \end{aligned}$$

where the $\xi_t^{(j)}$ are variational parameters of Q . The vector ξ_t plays a role very similar to the switch variable S_t . Each component $\xi_t^{(j)}$ can range between 0 and 1. When $\xi_t^{(j)} = 0$ the posterior probability of \mathbf{f}_t^j under Q does not depend on the observation at time t . When $\xi_t^{(j)} = 1$, the posterior probability of \mathbf{f}_t^j under Q includes a term which assumes that factor model j generated \mathbf{y}_t . We call $\xi_t^{(j)}$ the responsibility assigned to factor model j for the observation vector \mathbf{y}_t .

In order to maximize the lower bound on the log-likelihood, $KL(Q||P)$ is minimized with respect to the variational parameters $\xi_t^{(j)}$ and $q_t^{(j)}$ separately for each sequence of observations. For convenience we will express the probability density P in the log domain, through its associated energy function or hamiltonian, \mathcal{H} . The probability density is related to the hamiltonian through the usual Boltzmann distribution (at a temperature of 1), $P(\cdot) = \frac{1}{Z} \exp\{-\mathcal{H}(\cdot)\}$, where Z is a normalization constant required such that P

integrates to unity. We then similarly express the approximating distribution Q through its hamiltonian \mathcal{H}_Q .

Comparing \mathcal{H}_Q with \mathcal{H} we see that the interaction between the $S_t^{(j)}$ and the \mathbf{f}_t^j variables has been eliminated, while introducing two sets of variational parameters.³ In order to obtain the approximation Q which maximizes the lower bound on the log-likelihood, we minimize $KL(Q\|P)$ as a function of these variational parameters,

$$KL(Q\|P) = \mathbb{E}_Q [\mathcal{H} - \mathcal{H}_Q] - \log Z_Q + \log Z$$

where \mathbb{E}_Q denotes expectation over the approximating distribution Q . Both Q and P define distributions in the exponential family. As a consequence, the zeros of the derivatives of KL with respect to the variational parameters can be obtained simply by equating derivatives of $\mathbb{E}_Q(\mathcal{H})$ and $\mathbb{E}_Q(\mathcal{H}_Q)$ with respect to corresponding sufficient statistics $S_t^{(j)}$, \mathbf{f}_t^j and R_t^j , where $R_t^j = \mathbb{E}_Q[\mathbf{f}_t^j \mathbf{f}_t^{j'}] - \mathbb{E}_Q[\mathbf{f}_t^j] \mathbb{E}_Q[\mathbf{f}_t^j]'$ is the covariance of \mathbf{f}_t^j under Q . Many terms cancel when we subtract the two hamiltonians

$$\begin{aligned} \mathcal{H}_Q - \mathcal{H} = & - \sum_{t=1}^n \sum_{j=1}^m S_t^{(j)} \log q_t^{(j)} \\ & + \frac{1}{2} \sum_{j=1}^m \sum_{t=1}^n \left(\xi_t^{(j)} - S_t^{(j)} \right) \left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j - \theta_j \right)' \boldsymbol{\Psi}_j^{-1} \left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j - \theta_j \right) \\ & + \frac{1}{2} \sum_{j=1}^m \sum_{t=1}^n \left(\xi_t^{(j)} - S_t^{(j)} \right) \left[\mathbf{f}_t^{j'} \mathbf{H}_t^j{}^{-1} \mathbf{f}_t^j + \log |\boldsymbol{\Psi}_j| + \log |\mathbf{H}_t^j| \right] \end{aligned}$$

Taking derivatives and equating to zero, we get the fixed-point equations for $q_t^{(j)}$ and $\xi_t^{(j)}$:⁴

$$\begin{aligned} \xi_t^{(j)} &= Q(S_t = j) \\ q_t^{(j)} &= \exp \left\{ -\frac{1}{2} \mathbb{E}_Q \left[\left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j - \theta_j \right)' \boldsymbol{\Psi}_j^{-1} \left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j - \theta_j \right) \right] \right. \\ & \quad \left. - \frac{1}{2} \log |\boldsymbol{\Psi}_j| - \frac{1}{2} \mathbb{E}_Q \left[\mathbf{f}_t^{j'} \mathbf{H}_t^j{}^{-1} \mathbf{f}_t^j \right] - \frac{1}{2} \log |\mathbf{H}_t^j| \right\} \end{aligned}$$

To compute $\xi_t^{(j)}$ it is necessary to sum Q over all the S_τ variables not including S_t . This can be done efficiently using the forward-backward algorithm on the switch state variables, with $q_t^{(j)}$ playing exactly the same role as an observation probability associated with each setting of the switch variable. To compute $q_t^{(j)}$ it is necessary to calculate the expectations of \mathbf{f}_t^j and $\mathbf{f}_t^j \mathbf{f}_t^{j'}$ under Q . These expectations can be computed efficiently using the Kalman smoothing algorithm on each state-space approximation of the factor model (see Saidane and Lavergne, 2007b for more details), where for model j at time t , the data is weighted by the responsibilities $\xi_t^{(j)}$.

³where $S_1^{(j)} = 1$ if the switch state is in state j , and 0 otherwise.

⁴The equations are satisfied when $\xi_t^{(j)} = \mathbb{E}_Q[S_t^{(j)}]$. Using the fact that $\mathbb{E}_Q[S_t^{(j)}] = Q(S_t = j)$, we get the fixed-point equation for $\xi_t^{(j)}$.

3.3 Viterbi Approximation for Latent Structure Inference

This section presents a powerful new approximation to the mixed-state latent factor model. The approximation is based on a Viterbi technique which finds the best sequence of switching states S_t and common factors \mathbf{f}_t that minimizes the Hamiltonian cost in equation (6) for a given observation sequence $\mathcal{Y}_{1:n}$.

$$\begin{aligned} \mathcal{H}(\mathcal{F}_{1:n}, \mathcal{S}_{1:n}, \mathcal{Y}_{1:n}) &\simeq \text{Constant} + \sum_{t=2}^n S'_t(-\log \mathbf{P})S_{t-1} + S'_1(-\log \pi) \\ &+ \frac{1}{2} \sum_{t=1}^n \sum_{j=1}^m \left[(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{jt} - \theta_j)' \boldsymbol{\Psi}_j^{-1} (\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{jt} - \theta_j) + \log |\boldsymbol{\Psi}_j| \right] S_t(j) \\ &+ \frac{1}{2} \sum_{t=1}^n \sum_{j=1}^m \left[\mathbf{f}'_{jt} \mathbf{H}_{jt}^{-1} \mathbf{f}_{jt} + \log |\mathbf{H}_{jt}| \right] S_t(j) \end{aligned} \quad (6)$$

where $\mathcal{F}_{1:\tau} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_\tau\}$, π the vector of initial state probabilities, \mathbf{P} the HMM transition matrix so that its i -th row is equal to $[p_{i1} \dots p_{im}]$ for $i = 1, \dots, m$ and $S_t = [S_t(1), \dots, S_t(m)]'$, where $S_t(j) = 1$ if $S_t = j$ and 0 otherwise.

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:

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

i.e. the switching sequence posterior $p(\mathcal{S}_{1:n} | \mathcal{Y}_{1:n})$ is approximated by its mode, where $\mu(x) = 1$ for $x = \emptyset$ and zero otherwise. 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 show that a (suboptimal) solution to this problem can be obtained by recursive optimization of the probability of the best sequence at time t :

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

where $\mathcal{S}_{1:t-2}^*(i) = \arg \max_{\mathcal{S}_{1:t-2}} J_{t-1,i}$ 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 :

$$J_{t,j} = \min_{\mathcal{S}_{1:t-1}, \mathcal{F}_{1:t}} \mathcal{H} \left[\mathcal{F}_{1:t}, \{\mathcal{S}_{1:t-1}, S_t = j\}, \mathcal{Y}_{1:t} \right] \quad (7)$$

Namely, this cost is the least cost over all possible sequences of switching states $\mathcal{S}_{1:t-1}$ and corresponding factor model states $\mathcal{F}_{1:t}$. For a given switch state transition $i \rightarrow j$, the associated innovation cost $J_{t,t-1,i,j}$ is given by:

$$J_{t,t-1,i,j} = \frac{1}{2} \mathbf{e}_t(i,j)' \boldsymbol{\Sigma}_{t|t-1}^{i(j)-1} \mathbf{e}_t(i,j) + \frac{1}{2} \log \left| \boldsymbol{\Sigma}_{t|t-1}^{i(j)} \right| - \log p_{ij} \quad (8)$$

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 : $\mathbf{f}_{t|t}^j = \mathbf{f}_{t|t}^{\delta_{t-1,j}(j)}$ and $\mathbf{H}_{t|t}^j = \mathbf{H}_{t|t}^{\delta_{t-1,j}(j)}$ with $h_{t|t-1}^j = h_{t|t-1}^{\delta_{t-1,j}(j)}$ for $l = 1, \dots, k$. 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: $j_t^* = \delta_{t,j_{t+1}^*}$.

Given the degenerate nature of the transition equation, the smoother gain matrix $J_t^{(j)k}$ is always zero, $J_t^{(j)k} = \mathbf{H}_{t|t}^j \mathbf{0}'_k \mathbf{H}_{t+1|t}^{(j)k-1} = \mathbf{0}$. Hence, smoothing is unnecessary in this case because there are no dynamics in the mean specification of the factors. The smoothing equations are simply:

$$\begin{aligned} \mathbf{f}_{t|n}^{(j)k} &= \mathbf{f}_{t|t}^j + J_t^{(j)k} \left[\mathbf{f}_{t+1|n}^k - \mathbf{f}_{t+1|t}^{j(k)} \right] = \mathbf{f}_{t|t}^j \\ \mathbf{H}_{t|n}^{(j)k} &= \mathbf{H}_{t|t}^j + J_t^{(j)k} \left[\mathbf{H}_{t+1|n}^k - \mathbf{H}_{t+1|t}^{j(k)} \right] J_t^{(j)k'} = \mathbf{H}_{t|t}^j \end{aligned}$$

The Switching model's sufficient statistics are now simply given by $\mathbb{E}(S_t|\cdot) = S_t(j^*)$ and $\mathbb{E}(S_t S_{t-1}'|\cdot) = S_t(j^*) S_{t-1}(j^*)'$. The operator $\mathbb{E}(\cdot|\cdot)$ denotes conditional expectation with respect to the posterior distribution, e.g. $\mathbb{E}(\mathbf{f}_t|\cdot) = \sum_{\mathcal{S}} \int_{\mathcal{F}} \mathbf{f}_t p(\mathcal{F}, \mathcal{S}|\mathcal{Y})$. Given the "best" switching state sequence, the sufficient conditionally heteroscedastic factor model statistics can be easily obtained using the Rauch-Tung-Streiber smoothing (for a review see Rosti and Gales, 2001). For example

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

4 The EM Algorithm

An efficient learning algorithm for the parameters of our model can be derived by generalizing the EM algorithm (Dempster et al., 1977). The algorithm can be broken down into three steps: the expectation step (E) and two conditional maximization steps. We assume that the data can be 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 observation vectors $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$, a sequence of continuous state vectors $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ and a sequence of discrete HMM states $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$, 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 $t-1$, $p(S_1) = \pi_{s_1}$: the initial state probability and $p(S_t|S_{t-1}) = p_{s_{t-1}s_t}$: the transition probabilities. The auxiliary function that will be maximized is given by:

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

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 $\Theta^{(0)}, \Theta^{(1)}, \Theta^{(2)}, \dots$ 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 .

For $\mathcal{D}_n^{(i)} = \{\mathcal{Y}_{1:n}, \Theta^{(i)}\}$ and $\tilde{\mathbf{y}}_{jt} = \mathbf{y}_t - \mathbf{X}_j \mathbf{f}_t^j$, the conditional expectation of the complete log-likelihood function $\mathcal{L}(\Theta|\mathcal{Y}, \mathcal{F}, \mathcal{S})$ can be written as:

$$\begin{aligned} \mathcal{Q}(\Theta, \Theta^{(i)}) &\simeq \sum_{j=1}^m S_1(j) \log p(S_1) - \sum_{t=2}^n \sum_{i=1}^m \sum_{j=1}^m S_t(j) S_{t-1}(i) \log p_{ij} \\ &- \frac{1}{2} \sum_{j=1}^m \sum_{t=1}^n S_t(j) \left[\log |\Psi_j| + \mathbb{E} \left\{ (\tilde{\mathbf{y}}_{jt} - \theta_j)' \Psi_j^{-1} (\tilde{\mathbf{y}}_{jt} - \theta_j) | \mathcal{D}_n^{(i)} \right\} \right] \\ &- \frac{1}{2} \sum_{j=1}^m \sum_{l=1}^k \sum_{t=1}^n S_t(j) \mathbb{E} \left[\log(h_{lt}^j) + \frac{f_{lt}^2}{h_{lt}^j} | \mathcal{D}_n^{(i)} \right] \end{aligned} \quad (9)$$

The first maximization step is defined by substituting the expectations computed in the E-step for the complete-data sufficient statistics on the right-hand side of the above expressions to obtain expressions for the new iterates of the initial state probabilities π_j , transition probabilities p_{ij} , observation noise mean vectors θ_j , factor loadings \mathbf{X}_j and idiosyncratic variances Ψ_j .

Maximizing this function with respect to the discrete initial state probabilities, π_j , can be carried out using the Lagrange multiplier together with the sum to unity constraint $\sum_{j=1}^m \pi_j = 1$. The new discrete initial state probabilities can be written as

$$\hat{\pi}_j = \frac{S_1(j)}{\sum_{i=1}^m S_1(i)}$$

Maximizing the function (9) with respect to the discrete state transition probabilities, p_{ij} , can also be carried out using the Lagrange multiplier together with the sum to unity constraint $\sum_{j=1}^m p_{ij} = 1$. The new discrete state transition probabilities can be written as

$$\hat{p}_{ij} = \frac{\sum_{t=2}^n S_t(j)S_{t-1}(i)}{\sum_{t=2}^n S_{t-1}(i)}$$

Maximizing the auxiliary function in equation (9) with respect to the observation noise mean vector, θ_j , yields

$$\hat{\theta}_j = \frac{1}{\sum_{t=1}^n S_t(j)} \sum_{t=1}^n S_t(j) (\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{t|n}^j)$$

The new factor loadings matrix, \mathbf{X}_j , has to be optimized row by row. The l -th row vector $\hat{\mathbf{x}}_{jl}$ of the new factor loadings matrix can be written as

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

where y_{tl} and θ_{jl} are, respectively, the l -th elements of the current observation and the observation noise mean vectors under regime j .

Given the new factor loadings matrix, the idiosyncratic variances can be optimized using the following formulae

$$\hat{\Psi}_j = \frac{1}{\sum_{t=1}^n S_t(j)} \sum_{t=1}^n S_t(j) \text{diag} \left[\left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{t|n}^j - \theta_j \right) \left(\mathbf{y}_t - \mathbf{X}_j \mathbf{f}_{t|n}^j - \theta_j \right)' + \mathbf{X}_j \mathbf{H}_{t|n}^j \mathbf{X}_j' \right]$$

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

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{f}_t \end{pmatrix} | \mathcal{D}_{1:t-1}, S_t = j \sim \mathcal{N} \left[\begin{pmatrix} \theta_j \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}_j \mathbf{H}_{jt} \mathbf{X}'_j + \Psi_j & \mathbf{X}_j \mathbf{H}_{jt} \\ \mathbf{H}_{jt} \mathbf{X}'_j & \mathbf{H}_{jt} \end{pmatrix} \right]$$

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, 1992):

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

where " \approx " stands for "approximately distributed", $\Sigma_{t|t-1}^{(j)} = \mathbf{X}_j \mathbf{H}_{t|t-1}^{(j)} \mathbf{X}'_j + \Psi_j$ and $\mathbf{H}_{t|t-1}^{(j)}$ 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 $\mathbf{H}_{t|t-1}^{(j)}$ is given by $h_{lt|t-1}^j = h_{lt|t-1}^{\delta_{t-1,j}^{(j)}}$. Therefore, ignoring initial conditions, the pseudo log-likelihood function is given by:

$$\mathcal{L}^* = c - \frac{1}{2} \sum_{t=1}^n \sum_{j=1}^m S_t(j) \left[\log |\Sigma_{t|t-1}^{(j)}| + (\mathbf{y}_t - \theta_j)' \Sigma_{t|t-1}^{(j)-1} (\mathbf{y}_t - \theta_j) \right] \quad (10)$$

The maximization of the function (9) with respect to $(\pi_j, p_{ij}, \theta_j, \mathbf{X}_j$ and $\Psi_j)$ can be done ignoring the last two terms. However, if we were to assume that $\mathbf{f}_t = \mathbf{H}_{t|t-1}^{1/2} \mathbf{f}_t^*$, this would no longer be true because $h_{lt|t-1}^j$ indirectly depends on θ_j , \mathbf{X}_j and Ψ_j . In that case it is conceptually possible that the parameter values that maximize the first part of (9) might actually decrease the second part. Nevertheless, provided that these parameters increase the whole expression, the generalized EM principle still applies (see Demos and Sentana, 1998). Hence, we just need to find the conditional expectations in the first two lines of equation (9). These conditional expectations can be derived using the Kalman filter.

Unfortunately, the conditional variance parameters $\phi = \{\omega, \gamma, \alpha, \delta\}$ are in practice unknown. The most obvious possibility is to apply the EM algorithm to estimate these as well. However, as explained above, this is not easy because of the nonlinear dependence structure in the variances of the common latent factors. An alternative possibility is based on the following idea. In the first step, we maximize the log-likelihood function in (9) with respect to the parameters in $\pi_j, p_{ij}, \theta_j, \mathbf{X}_j$ and Ψ_j by means of the EM algorithm, holding the factor variances' parameters fixed at the value of the previous iteration. In the second maximization step, using $\pi_j, p_{ij}, \theta_j, \mathbf{X}_j$ and Ψ_j parameter values found in the first step, we maximize the observed log-likelihood function (10) with respect to the conditional variance parameters, and so on until convergence. The final parameter estimates obtained in this way will be the maximum likelihood estimates of our model.

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, the hidden markovian states posterior probabilities given by the smoothing algorithm (see, Saidane and Lavergne, 2007a), or the variational parameters $\xi_t^{(j)}$ (Saidane and Lavergne 2007b). Once this sequence is known, on each segment of data the function \mathcal{L}^* is maximized through the `fmincon` constrained optimization `Matlab` function. `fmincon` finds the constrained minimum of a scalar function of several variables starting at an initial estimate. This is generally referred to as constrained nonlinear optimization.

5 Choosing an Honest Model

The various model selection criteria such as AIC (Akaike, 1974) and BIC (Schwarz, 1978) implicitly assume that the sampling distribution belongs to, at least, one of the models in competition. This assumption is most often unrealistic and can lead to under-penalized complex models (see Burnham and Anderson, 1998). Taking into account the modeling purpose can counter this tendency efficiently. This approach is sensible for hidden structure models. In this setting, discovering the hidden structure to derive a reliable clustering of the dataset is often of primary interest to the user. Thus, we propose a model selection criterion favoring minimal missing information models. This criterion, the so called ICL criterion previously proposed in (Biernacki et al., 2000) for mixture models, can be generalized to any hidden structure model. In such a case though, it seems preferable to base the selection of a model on the maximization of the integrated complete log-likelihood function defined by:

$$p(\mathcal{Y}, \mathcal{Z}|\mathcal{M}) = \int p(\mathcal{Y}, \mathcal{Z}|\mathcal{M}, \Theta_{\mathcal{M}})\pi(\Theta_{\mathcal{M}})d\Theta_{\mathcal{M}}$$

where $\mathcal{M} = \{\mathcal{M}_i, i = 1, \dots, I\}$ be the candidates of desired parametric models and \mathcal{Z} indicates the hidden variables: the latent common factors and the discrete hidden state variables. An equivalent approximation for the integrated complete log-likelihood function is given by:

$$\text{ICL}(\mathcal{M}) = \text{BIC}(\mathcal{M}) + \log p(\mathcal{Z}|\mathcal{Y}, \hat{\Theta}_{\mathcal{M}})$$

$\log p(\mathcal{Z}|\mathcal{Y}, \hat{\Theta}_{\mathcal{M}})$ being a measurement of the missing information carried by the model \mathcal{M} . This expression highlights that the ICL criterion over-penalizes the models at significant missing information as compared to BIC.

As defined, the ICL criterion is not calculable since the states \mathcal{Z} are not observed. A natural approximation to $\log p(\mathcal{Z}|\mathcal{Y}, \hat{\Theta}_{\mathcal{M}})$ is:

$$\log p(\mathcal{Z}|\mathcal{Y}, \hat{\Theta}_{\mathcal{M}}) = \max_{\mathcal{Z}} \left[\log p(\mathcal{Z}|\mathcal{Y}, \hat{\Theta}_{\mathcal{M}}) \right]$$

In the case of our model, this problem can be resolved by the Viterbi algorithm.

6 Monte Carlo Experiments

There are two important empirical questions that should be addressed for the class of mixed-state conditionally heteroscedastic latent factor models:

1. Which approximation inference scheme in mixed-state factor models results in the best learning performance?
2. An other important question is the choice of a reliable model, containing enough parameters to ensure a realistic fit to the learning dataset.

In this section we report some early progress in addressing these questions.

6.1 Model Learning and Stability of the Estimates

The example used here has $q = 6$ observable variables and only one GQARCH(1,1) latent factor. We consider the case of three states model with the initial state $S_1 = 1$ and a transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.90 & 0.05 \\ 0 & 0.05 & 0.95 \end{pmatrix}$$

The iterations 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 . Here the constant term of the conditionally heteroscedastic component is assumed to be known ($\omega_j = 1 \forall j = 1, 2, 3$ and the initializations given in table 1 were used).

The goal is to estimate the different dynamics and to measure the distance between estimates $\tilde{\Theta}$ and true parameters Θ_0 through the empirical Kullback-Leibler divergence (see Juang and Rabiner, 1985). For each value of n , the estimation procedure was carried out a hundred times, and the KL distances between each of the hundred estimators and the true parameter were evaluated on a new sequence, independent of the first hundred sequences used to obtain the estimators. Table 2 shows the mean and standard deviation of the estimates with $n = 900$. In this case the estimated transition matrix $\tilde{\mathbf{P}}$ is given by

$$\tilde{\mathbf{P}} = \begin{bmatrix} 0.9481 & 0.0510 & 0.0009 \\ (0.0041) & (0.0063) & (0.0014) \\ 0.0435 & 0.9079 & 0.0486 \\ (0.0029) & (0.0066) & (0.0017) \\ 0.0006 & 0.0519 & 0.9475 \\ (0.0032) & (0.0041) & (0.0013) \end{bmatrix}$$

Table 1: Simulation parameters.

	θ	\mathbf{X}	$diag(\Psi)$	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.

Values into brackets represent standard deviation of the estimates. The sets of distances for the various values of n are presented under a unified scale in figure 1. This figure clearly shows that the average amplitude of fluctuations increases with the GPB2 and variational methods when the number of observations is relatively small (not more than 900). This figure shows also a general decrease in average and spread of the distances with increasing n . Given that small values of KL imply similarity between Θ_0 and $\tilde{\Theta}_n$, the results of this experiment suggest an increasing accuracy and stability of the estimators obtained with the EM-based Viterbi approximation algorithm as n increases.

To investigate the asymptotic distribution of $\tilde{\Theta}_n$, we have used the Shapiro-Francia (1972) statistic in order to test the univariate normality of each component of $\tilde{\Theta}_n$. This is an omnibus test, and is generally considered better than the Shapiro-Wilk (1965) test for Leptokurtic Samples. All the results presented in table 3, for the simulation with $n = 900$, 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 $\alpha = 5\%$.

6.2 Model Selection

In this experiment we illustrate the estimation of the numbers of the discrete hidden states and common latent factors which describe the trajectories best. For this purpose we consider two different situations with factor models which differ by their dynamic

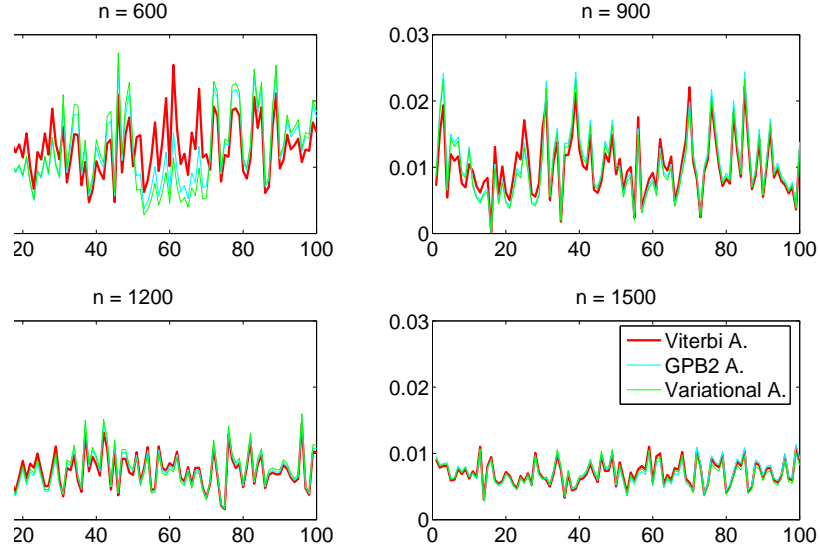


Figure 1: The KL distances for the different values of n .

Table 2: Averages and standard deviations (.) for the EM parameter estimates from the simulated data with $n = 900$.

	θ	\mathbf{X}	$diag(\Psi)$	ϕ
State 1	0.9952 (0.0817)	1.9921 (0.0631)	0.9938 (0.0451)	0.4982 (0.0731)
	1.0211 (0.0932)	1.9981 (0.0655)	1.0058 (0.0461)	0.1022 (0.0447)
	1.0136 (0.0866)	2.9947 (0.0591)	0.9931 (0.0582)	0.7931 (0.0352)
	1.9906 (0.0883)	3.9986 (0.0582)	0.9916 (0.0633)	
	1.9947 (0.0922)	4.9961 (0.0591)	0.9936 (0.0591)	
	2.0772 (0.0859)	5.9913 (0.0622)	1.0011 (0.0473)	
State 2	0.9914 (0.0732)	1.9977 (0.0552)	2.0117 (0.0622)	0.1026 (0.0752)
	1.9934 (0.0739)	2.0061 (0.0602)	2.0117 (0.0573)	0.3011 (0.0442)
	1.0610 (0.0758)	2.0094 (0.0588)	1.9937 (0.0654)	0.4011 (0.0361)
	1.9911 (0.0866)	2.9980 (0.0562)	1.9966 (0.0641)	
	1.0361 (0.0839)	3.0097 (0.0588)	2.0089 (0.0679)	
	1.9952 (0.0786)	3.0109 (0.0535)	1.9977 (0.0612)	
State 3	1.9733 (0.0837)	1.0141 (0.0468)	2.9971 (0.0576)	0.2039 (0.0771)
	2.9811 (0.0878)	3.0056 (0.0483)	3.0051 (0.0558)	0.1996 (0.0373)
	1.9718 (0.0855)	1.0072 (0.0467)	2.9819 (0.0687)	0.5989 (0.0277)
	2.9813 (0.0951)	2.0069 (0.0456)	2.9901 (0.0662)	
	1.9911 (0.0985)	4.0114 (0.0510)	3.0062 (0.0687)	
	2.9721 (0.0811)	4.0097 (0.0506)	2.9983 (0.0708)	

hidden structures. In the first case, parameters of the true model are given in table 1 (we consider here $n = 900$). In the second case the true model is a GQARCH(1,1) factor

Table 3: Summary statistics for the Shapiro-Francia test [*simulation with $n = 900$*].

	θ	\mathbf{X}	$diag(\Psi)$	ϕ
State 1	*0.3971 (0.2608)	0.4625 (0.0940)	0.1264 (1.1434)	0.1448 (1.0588)
	0.4815 (-0.0464)	0.3976 (0.2595)	0.2045 (0.8256)	0.2194 (-0.7741)
	0.3688 (0.3350)	0.3690 (-0.3344)	0.4469 (-0.1335)	0.1530 (-1.0238)
	0.4353 (-0.1630)	0.4901 (-0.0249)	0.2052 (-0.8233)	
	0.3975 (-0.2597)	0.2106 (-0.8045)	0.4572 (-0.1075)	
	0.4882 (0.0295)	0.4355 (0.1623)	0.4335 (-0.1674)	
State 2	0.2751 (0.5976)	0.3384 (0.4169)	0.2272 (0.4782)	0.2692 (-0.6152)
	0.1887 (0.8825)	0.2947 (-0.5398)	0.4210 (-0.1994)	0.4692 (0.0774)
	0.3528 (0.3779)	0.4636 (0.0914)	0.4228 (-0.1947)	0.4603 (0.0996)
	0.2029 (0.8315)	0.4955 (0.0112)	0.4019 (-0.2485)	
	0.1273 (1.1394)	0.3105 (0.4945)	0.4790 (0.0527)	
	0.2647 (0.6291)	0.2544 (0.6607)	0.3304 (-0.4389)	
State 3	0.3074 (0.5031)	0.0568 (1.5820)	0.1478 (-1.0459)	0.3845 (0.2937)
	0.2660 (0.6250)	0.3849 (-0.2927)	0.1741 (0.9379)	0.3516 (0.3810)
	0.3028 (0.5162)	0.4429 (-0.1437)	0.4273 (0.1833)	0.4844 (-0.0390)
	0.1981 (0.8484)	0.4324 (-0.1704)	0.1362 (-0.0978)	
	0.1270 (1.1405)	0.2751 (-0.5976)	0.3236 (-0.4577)	
	0.3187 (0.4713)	0.3495 (-0.3867)	0.4682 (0.0798)	

* pval, (.) \mathcal{W} statistic. All the pval are greater than 0.05, hence 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 $\alpha = 5\%$.

model with $n = 800$, $m = k = 2$ and the regime switching date $t^* = n/2 + 1$.

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 maximum likelihood criterion on the training dataset. In the second example random initialization was used for the implementation of the learning algorithm. In this case the initial parameters for the EM algorithm, were obtained by randomly perturbing the true parameter values by up to 20% of their true value. Minimizing the selection criteria – computed after each EM running – allows us to find the best model among the \mathcal{M} models. Table 4 shows the results of the two examples obtained with the EM-based Viterbi approximation algorithm. In the first example BIC and ICL criteria choose 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 an overfitting. In the second example, BIC and ICL choose also the true specification with two states and two conditionally heteroscedastic factors.

To illustrate the evolution of the model estimates obtained by the EM-based Viterbi approximation algorithm, figure 2 shows the HMM hidden states estimates at iteration 1, 3, 5 and 7. Each figure depicts the regime path process of the correct model. It can be concluded that a good segmentation is achieved after 7 iterations.

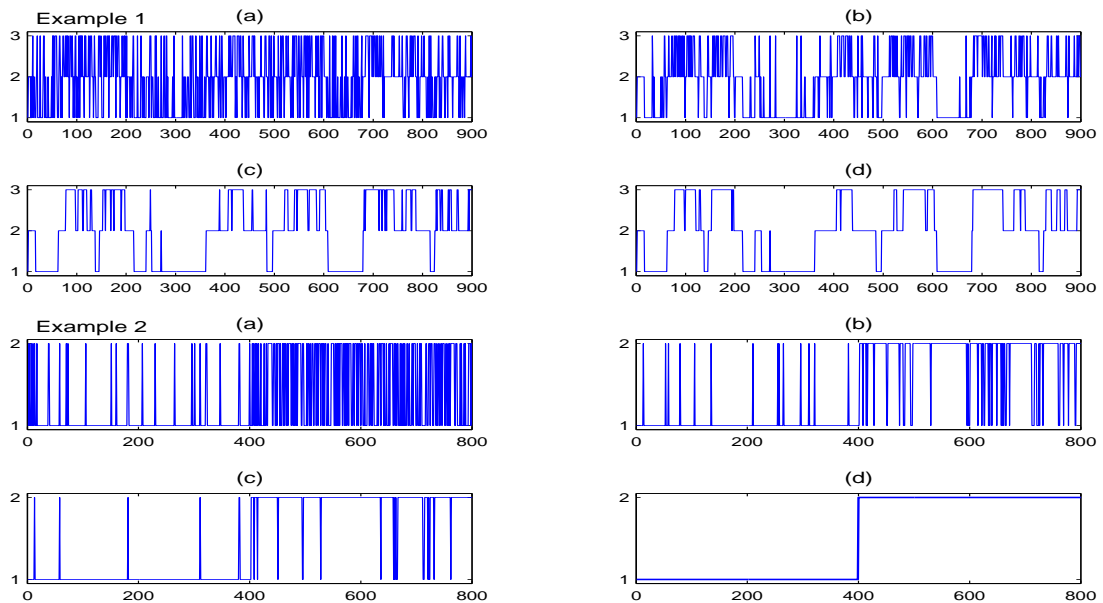


Figure 2: Evolution of the HMM state estimates using the true model: (a) iteration 1, (b) iteration 3, (c) iteration 5, (d) iteration (7).

Identification of the state sequence adds to the understanding of the process, since it enables the analyst to relate historical events to the state process. In previous works many authors have used the maximal a posteriori probability method, by which they estimate S_t by the state that maximizes the marginal a posteriori probability $p(S_t | \mathcal{Y}_{1:\tau}; \Theta)$, $1 \leq \tau \leq n$, where Θ is substituted by its maximum likelihood estimate. If $\tau = t$, these are given by the filtering algorithm; if $\tau > t$ they are given by the smoothing algorithm, and if $\tau < t$, we will refer to them as predicted probabilities. Clearly, the most useful state estimates from investors point of view is the one that enables the forecast of the next state S_{t+1} based on the information set available at time t , $\mathcal{Y}_{1:t}$. We can similarly derive h -step state predictions (see Saidane and Lavergne, 2007a for more details). Figure 3 compares the decoding errors of the most likely hidden-state sequence given by the prediction, smoothing (GPB2) and Viterbi algorithms obtained on the test set of simulated data with a 3 state and one factor-conditionally heteroscedastic model. It shows that the Viterbi approximation has a lower decoding error than the other three methods. From this figure it can be seen that the best segmentation path is provided by the Viterbi approximation. We see, also, that the two other methods are more globally sensitive to outliers and can not give the best segmentation performance. Hence, our results suggest that this new method is a promising technique which improves the segmentation quality and consistency and gives more attractive performance compared to the GPB2 method. This finding is also supported by the results of the model-selection exercise reported in figure 6.

Using the initial guesses given in table 1, the EM-based Viterbi approximation algorithm converged to estimates of the GQARCH processes after approximately 50 iterations as shown in figure 4. The same figure shows a weak convergence to the true parameter



Figure 3: Example 1: Decoding errors of the most likely hidden-state sequence given by the prediction, smoothing (GPB2) and Viterbi algorithms.

values when we use the GPB2 method. Figure 5 shows that the same comment applies to the convergence of the likelihood function. In this case the GPB2 method, compared to the Viterbi approximation one, leads to a relative slow convergence when the number of hidden states is greater than 2. As a consequence, our results show that the GPB2 and variational methods lead either to an apparent over-estimation or under-estimation of the mean and the volatility behavior of the common latent factor. Finally, the sample autocorrelation functions of the estimation errors obtained with the Viterbi approximation method 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 series are explained by the common and specific factors. Scatter plots of the residuals versus the factor estimates show that the residuals do not exhibit any systematic structure which indicate that the model fits the data well.

From these results we conclude that with the Viterbi approximation method the accuracy of classification increases as the order of the switching states and the conditionally heteroscedastic latent factor model order increase. Analysis of three different mixed-state latent factor model inference schemes indicates that Viterbi scheme do seem to yields appealing classifications. However, the GPB2 and variational methods does not considerably lack behind the mentioned scheme and sometimes even outperforms the first one. Moreover, inference process of GPB2 is clearly more involved than those of the Viterbi or the variational approximation. Unlike Viterbi, GPB2 provides "soft" estimates of switching states at each time t . Like Viterbi GPB2 is a local approximation scheme and as such does not guarantee global optimality inherent in the variational approximation. However,

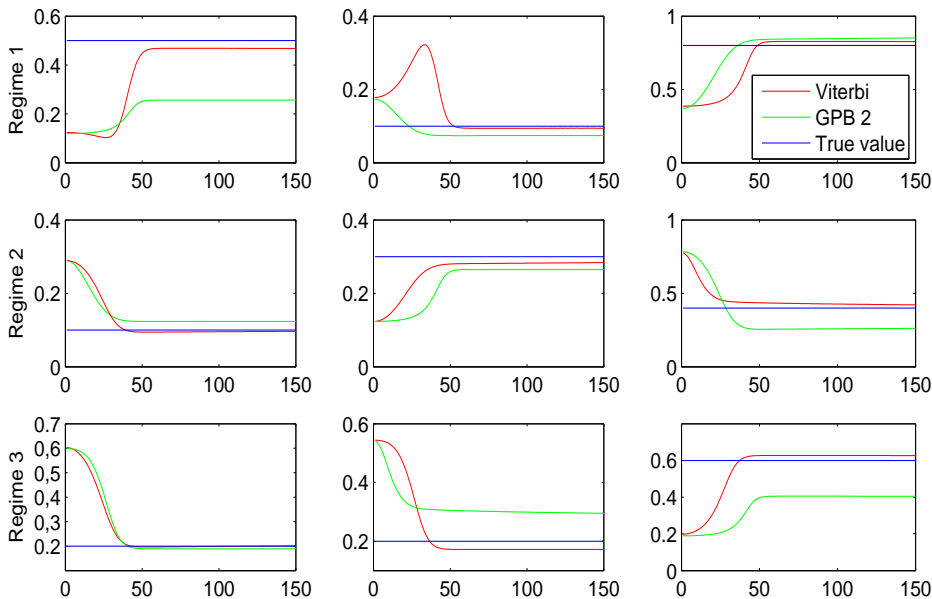


Figure 4: Example 1: Evolution of the conditionally heteroscedastic parameter estimates during the EM iterations: γ_j (first column), α_j (second column) and δ_j (third column).

some recent work (see Boyen and Friedman, 1999) on this type of local approximation in general DBNs has emerged that provides conditions for it to be globally optimal. In terms of computational complexity, Viterbi does seem to be the clear winner among the mixed-state latent factor model schemes.

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 our proposed ICL criterion was chosen. Figure 6 shows the choice frequencies for each specification. In the two examples, ICL prefers the true model most of the time. Our Monte Carlo simulation experiments show also that the EM-based Viterbi approximation algorithm increases the precision of the new proposed ICL criterion. Here we can argue that the improvement in accuracy is due entirely to the significant improvement in the identification of the most probable path through the states of the model given by the Viterbi decoding.

7 Conclusion

We have introduced a new approach to dynamics learning based on switching conditionally heteroscedastic factor models. We have proposed a Viterbi approximation technique which overcomes the exponential complexity of exact inference. Our proposed model is general enough that it allows for changing relationships among variables in the dataset without imposing that these changes have occurred or assuming a date for the changes. It takes

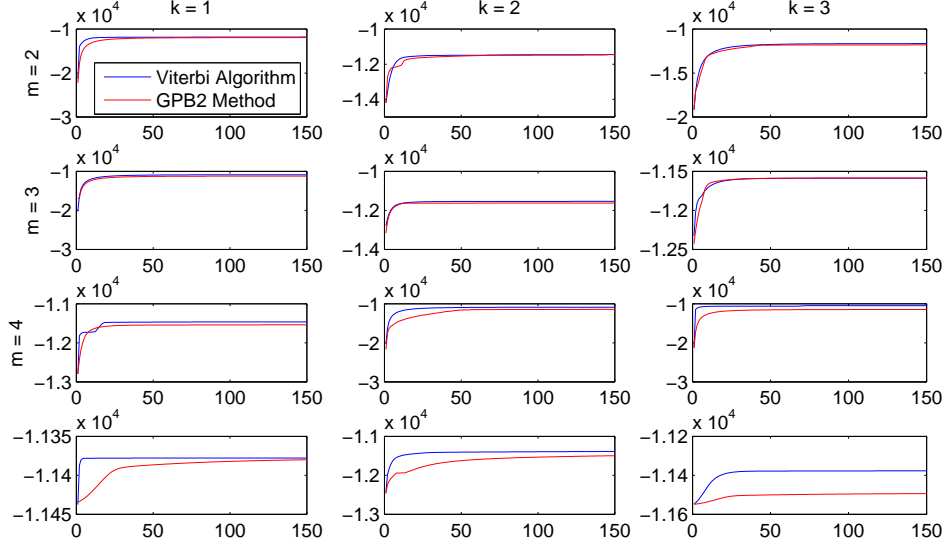


Figure 5: Log-likelihoods of the different specifications (with different values of m and k) using Viterbi and GPB2 methods on the same datasets.

Table 4: Values of the AIC, BIC and ICL 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	24310 (22610)	24082 (22494)	24016 (22414)
BIC	24411 (22708)	24226 (22635)	24203 (22597)
ICL	24409 (22704)	24223 (22635)	24201 (22594)
	$m = 2$		
	23398 (22332)	23248 (22240)	23160 (22312)
	23629 (22557)	23565 (22549)	23563 (22706)
	23629 (22557)	23564 (22544)	23563 (22700)
	$m = 3$		
	23190 (22324)	23412 (22248)	23544 (22380)
	23550 (22675)	23902 (22726)	24164 (22984)
	23548 (22675)	23902 (22724)	24161 (22982)

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.

Our proposed algorithm has been tested on simulated data and it showed very promising results compared to the GPB2 and variational methods. There are several benefits to use an EM-based Viterbi approximation algorithm for mixed-state latent factor models.

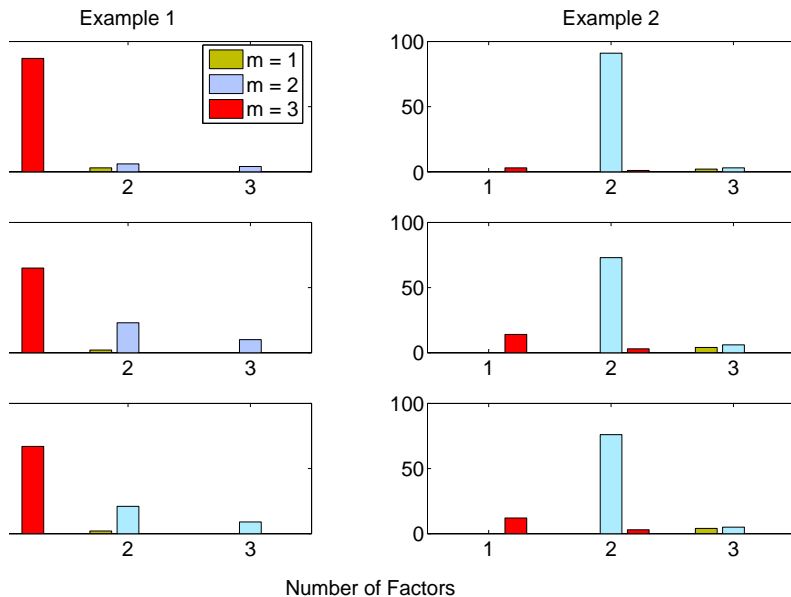


Figure 6: Frequencies of choosing each model with ICL. Results of the Viterbi approximation in the first line. Those of the GPB2 and variational methods in the second and third lines, respectively.

Most of these advantages revolve around the tractability of the learning and inference processes and the improvement in the classification of the volatility behavior. Another specific advantage of our new proposed algorithm is that it can accelerate the convergence of the EM iterations. Indeed, if there are a large number of hidden states, it may be too slow to perform m^2 or even m Kalman Filtering updates, as required by GPB2 and *IMM*.

Our numerical experiments on simulated data of the resulting ICL criterion show that it performs well both for choosing a mixed-state factor model and a relevant number of common conditionally heteroscedastic latent factors. With this new criterion we demonstrated accurate discrimination between specifications characterized by different hidden structures. In simulated events, this selection criterion with Viterbi approximation method gives the right answer, about 91% of the time. In particular, ICL appears to be more robust than AIC and BIC to violation of some of the mixed-state latent factor model assumptions and it can select a number of hidden states and common latent factors leading to a sensible partitioning of the data.

The fact that our proposed model 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. The analysis in this paper can be also extended in several ways. First, our model can be generalized to one where one allows 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.

Acknowledgements

This joint work has been made possible by cross-invitations at INRIA Rhône-Alpes – IS2 and MISTIS teams – and Institut de Mathématiques et de Modélisation de Montpellier I3M which are gratefully acknowledged.

References

- [1] Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Automatic Control* AC-19:716–723.
- [2] Bar-Shalom Y. and Li X.R. (1993). Estimation and tracking: principles, techniques and software. *Boston, London: Artech House Inc.*
- [3] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. on PAMI* 22:719–725.
- [4] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31:307–327.
- [5] Boyen, X., Firedman, N., and Koller, D. (1999). Discovering the hidden structure of complex dynamic systems. In Proc. Uncertainty in Artificial Intelligence, pages 91–100.
- [6] Burnham, K.P., and Anderson, D.R. (1998). *Model Selection and Inference: a Practical Information Theoretic Approach*. Springer-Verlag, New York.
- [7] Carnero, M.A., Peña, D., and Ruiz, E. (2004). Persistence and Kurtosis in GARCH and Stochastic Volatility Models. *Journal of Financial Econometrics* 2:319–342.
- [8] Demos, A., and Sentana, E. (1998). An EM Algorithm for Conditionally Heteroscedastic Factor Models. *Journal of Business & Economic Statistics*, 16:357–361.
- [9] Dempster A., Laird N., and Rubin D. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B* 39:1–38.
- [10] Ghahramani, Z. and Hinton, G.E. (1998). Switching statespace models. Technical report, Departement of Computer Science, University of Toronto. <ftp://ftp.cs.toronto.edu/pub/zoubin/switchftp.ps.gz>.
- [11] Harvey, A., Ruiz, E., and Sentana, E. (1992). Unobserved component time series models with ARCH disturbances. *Journal of Econometrics* 52:129–157.

- [12] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. (1991). Adaptive mixture of experts. *Neural Computation*, 3:79–87.
- [13] Juang, B.H., and Rabiner, L.R. (1985). A Probabilistic Distance Measure for Hidden Markov Models. *AT&T Technical Journal* 64:391–408.
- [14] Kim, C-J. (1994). Dynamic linear models with markov-switching. *Journal of Econometrics* 60:1–22.
- [15] Lauritzen S. (1996) Graphical Models. *OUP*.
- [16] Murphy, K.P. (1998). *Learning switching kalman filter models*. Technical Report 98-10, Compaq Cambridge Research Lab.
- [17] Rauch, H.E. (1963). Solutions to the linear smoothing problem. *IEEE Transactions on Automatic Control*, 8:371–372.
- [18] Rauch, H.E., Tung, F., and Striebel, C.T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3:1445–1450.
- [19] Rosti, A.V.I., and Gales M.J.F. (2001). *Generalised Linear Gaussian Models*. Tech. Rep. CUED/F-INFENG/TR.420, Cambridge University Engineering Department.
- [20] Saidane, M., and Lavergne, C. (2006). On Factorial HMMs for Time Series in Finance. *The Kyoto Economic Review* 75:63–90.
- [21] Saidane, M., and Lavergne, C. (2007a). Conditionally Heteroscedastic Factorial HMMs for Time Series in Finance. *Applied Stochastic Models in Business and Industry* 23:503–529.
- [22] Saidane, M., and Lavergne, C. (2007b). A Structured Variational Learning Approach for Switching Latent Factor Models. *Advances in Statistical Analysis - Journal of the German Statistical Society* 91:245–268.
- [23] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6:461–464.
- [24] Sentana, E., and Fiorentini, G. (2001). Identification, Estimation and Testing of Conditionally Heteroskedastic Factor Models. *Journal of Econometrics* 102:143–164.
- [25] Sentana, E. (1995). Quadratic ARCH Models. *Review of Economic Studies* 62:639–661.
- [26] Shapiro, S.S., and Francia, R.S. (1972). An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association* 67:215–216.
- [27] Shapiro, S.S., and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52:591–611.

- [28] Shi, S. and Weigend, A. S. (1997). Taking time seriously: Hidden Markov experts applied to financial engineering. In Conference on Computational Intelligence for Financial Engineering. CIFE, IEEE/IAFE.