Multiple Chains Hidden Markov Models for Bivariate Dynamical Systems

Leopoldo Catania

Aarhus BSS and CREATES

Abstract

We present a new modelling framework for the bivariate hidden Markov model. The proposed specification is composed by five latent Markovian chains which drive the evolution of the parameters of a bivariate Gaussian distribution. The maximum likelihood estimator is computed via an expectation conditional maximization algorithm with closed form conditional maximization steps, specifically developed for our model. Identification of model parameters, as well as consistency and asymptotic Normality of the maximum likelihood estimator are discussed. Finite sample properties of the estimator are investigated in an extensive simulation study. An empirical application with the bivariate series of US stocks and bond returns illustrates the benefits of the new specification with respect to the standard hidden Markov model.

Keywords: Hidden Markov models, Multiple Markov chains, Expectation conditional maximization

1. Introduction

Hidden Markov models describe the relationship between an observed process, $\{\mathbf{Y}_t\}$, and a latent (hidden) process, $\{S_t\}$, $t = 1, \ldots$. The distributional properties of $\mathbf{Y}_t$ are defined conditional on S_t, giving rise to a mixture representation of the observed process and providing great flexibility in empirical settings. One of the most appealing specification is the one where S_t is first order Markov

[*] Department of Economics and Business Economics, Aarhus BSS and CREATES, email: leopoldo.catania@econ.au.dk

Preprint submitted to *March 18, 2021*

and takes values on a discrete bounded support, and $\mathbf{Y}_t|S_t$ is multivariate Gaussian distributed. Through this paper, we refer to this specification as the hidden Markov model (HMM). The forward filtering backward smoothing (FFBS) algorithm paired with the expectation maximization (EM) algorithm of Dempster et al. (1977) allow us to easily estimate the parameters of the distribution of $\mathbf{Y}_t|S_t$ via the maximum likelihood (ML) estimator, as well as to infer about the distribution of S_t conditional on a sequence of observations, i.e. $S_t|\mathbf{Y}_1, \dots, \mathbf{Y}_T$, see Frühwirth-Schnatter (2006).

The underlying assumption in the HMM specification is that of a single Markovian chain driving the whole dynamic system. For example, in a bivariate setting, the two means, two variances, and the correlation parameter of the distribution of $\mathbf{Y}_t|S_t$ all depend on S_t. Of course, by allowing S_t to be "large", in the sense that S_t is allowed to take values on a large support, the single chain constraint is less stringent given the fact that many combinations of parameters are estimated. Unfortunately, increasing the state space of S_t leads to a proliferation of parameters, which can be harmful when the sample size is like the ones usually available in economics and (non-high-frequency) finance. The standard solution is to impose an a priori structure in the system, like assuming that some of the parameters of $\mathbf{Y}_t|S_t$ do no depend on S_t, or that transitions of S_t are somehow restricted. While alleviating the proliferation of parameters problem, these solutions are ad-hoc, and usually driven by expert knowledge instead of statistical arguments. Also, when statistical arguments are in place, the set of possible alternatives has already been narrowed to a large extent.

A more natural solution is to assume that multiple chains determine the evolution of the dynamic system. This modelling strategy has been used extensively in econometrics. For example, Phillips (1991) exploited the specification of Hamilton (1989) and studied the international transmission of the business cycles using a bivariate model where the means of the growth of output of two countries are determined by the realization of two independent Markov chains. Ravn and Sola (1995) extended the model of Phillips (1991) by assuming that also the variance parameters are affected by the same Markov chains. A similar specification, thought in the univariate framework, is considered by Doornik (2013). Ang and Bekaert (2002a), Sims and Zha (2006), and Sims et al. (2008) also discussed the use of independent multiple chains in their analyses.

In this paper, we provide a more flexible modelling framework for the bivariate HMM by introducing five first order Markovian chains. Each chain is independent from the others and it is dedicated to one of the parameters of the conditional distribution of the observed process. Specifically, two chains move the two conditional means, two chains the standard deviations, and the last chain the correlation parameter. Compared to previous approaches, the new specification allows for a complete separation of all parameters of the joint distribution, instead of only the means or the variances. As it will be discussed in the paper, disentangling the two variances and the correlation parameter results in a more involved estimation procedure.

We show that the proposed specification has an equivalent representation as HMM with a (usually) large state space, and a particular structure imposed to the parameters of $\mathbf{Y}_t|S_t$ and S_t, which prevents from the proliferation of parameters. Note that, this structure emerges from the specification of the model, and not from ad-hoc considerations. We label this new specification the multiple chains hidden Markov model (MCHMM).[1] Results from Leroux (1992), Bickel et al. (1998), and Gassiat et al. (2016) allow us to derive conditions that ensure identification of the parameters, as well as the consistency and asymptotic Normality of the ML estimator. For ML estimation, we develop an expectation conditional maximization (ECM) algorithm with closed form conditional maximization steps, see Meng and Rubin (1993). The ECM paired with the FFBS algorithm allows us to easily estimate the model parameters. Simulation results indicate that the resulting ML estimator has good finite sample properties.

The conditional maximization step for the correlation parameters turns out to be particularly involved, and usually wrongly reported in the literature for related models. This is generally acknowledged in the literature (Pelletier, 2006; Paolella et al., 2019), and it seems that its implications in terms of estimation accuracy of the model parameters have been underrated. We report a

[1] Berchtold (1999) proposes a specification that he labels the “double chain Markov model”, which is different from ours. Colombi and Giordano (2015) propose a model they refer to as “Multiple hidden Markov model” which is however different from the one we discuss here.

simulation analysis showing that the commonly employed wrong M-step leads to sizable root mean squared errors increases in finite samples.

We also address the independence assumption of the five Markov chains. This assumption is common in the literature and prevents from a proliferation of parameters. Its main implication is that, when dynamic parameters are viewed as random processes in the unobserved component representation of the model, they are independent as well. We propose a modelling framework that preserves the independence assumption, but allows for correlated dynamic parameters. The resulting specification has the previous one as a special case.

We conclude the article with an application in financial econometrics. We start from a finding of Guidolin and Timmermann (2006), who show that the joint process of US stocks and bond returns follows a rich and complex dynamic pattern. Specifically, these two series are characterized by well-defined regimes. Though, since there is very little coherence between these regimes, econometricians usually end up estimating very large systems. Similar results are reported by Ang and Bekaert (2002b) and Andersen et al. (2007). Our results indicate that, the selected MCHMM specification is less parameterized than then selected HMM specification. Also, MCHMM leads to estimated parameters which are easier to interpret than those of HMM.

The remainder of the paper is organized as follows. Section 2 introduces the MCHMM specification, and discusses its statistical properties. Section 3 details the ECM algorithm. Finite sample properties of the ML estimator computed via the ECM algorithm are discussed in Section 4. The specification with correlated dynamic parameters is discussed in Section 5. Section 6 reports the empirical illustration. Section 7 concludes and discusses avenues for future research. In the Appendix we report the derivation of the ECM algorithm for the specification of Section 5.

2. The model

Let $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})'$ be a bivariate stochastic vector whose realizations are observed. We assume that:

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} \left| \begin{matrix} S_t^{\mu_1} = a & S_t^{\mu_2} = b \\ S_t^{\sigma_1} = c & S_t^{\sigma_2} = d \\ S_t^{\rho} = e & \end{matrix} \right. \sim N\left(\begin{bmatrix} \mu_{1,a} \\ \mu_{2,b} \end{bmatrix} ; \begin{bmatrix} \sigma^2_{1,c} & \sigma_{1,c}\sigma_{2,d}\rho_e \\ \sigma_{1,c}\sigma_{2,d}\rho_e & \sigma^2_{2,d} \end{bmatrix} \right), \qquad (1)$$

where S_t^x is a first order ergodic Markov chain with state space $\{1, \ldots, D^x\}$, for $x \in \mathcal{X}$, where $\mathcal{X} = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$, whose realizations are not observed. We assume that $S_t^x \perp\!\!\!\perp S_t^{x'}$ for all $x \neq x'$, and denote by $\mathbf{\Gamma}^x = \left[\gamma_{ij}^x\right]$ the $D^x \times D^x$ transition probability matrix of S_t^x, with $\gamma_{ij}^x = P(S_t^x = j | S_{t-1}^x = i) > 0$, $\sum_{j=1}^{D^x} \gamma_{ij}^x = 1$, and by $\boldsymbol{\delta}^x$ the vector representing the initial distribution of S_t^x, with generic element $\delta_j^x = P(S_1^x = j) > 0$, with $\sum_{j=1}^{D^x} \delta_j^x = 1$, for $j = 1, \ldots, D_x$.

Conditional on S_t^x, $x \in \mathcal{X}$, the distribution of $\mathbf{Y}_t$ is assumed to be bivariate Normal with individual means, variances, and correlation determined by the (unobserved) realizations of the S_t^x variables. A graphical representation of the model is reported in the path diagram in Figure 1.

Note that, it is assumed that all Markov chains are independent among them, i.e. $S_t^x \perp\!\!\!\perp S_t^{x'}$ for all $x \neq x'$. This independence assumption can only be relaxed at the cost of a proliferation of parameters.[2] For instance, we might assume that, say, $P(S_t^{\mu_1} = j | S_t^{\mu_2} = l, S_{t-1}^{\mu_1} = i) = \gamma_{ijl}^{\mu_1}$ for all $l = 1, \ldots, D^{\mu_2}$, which would imply $D^{\mu_1}(D^{\mu_1} - 1)(D^{\mu_2} - 1)$ extra parameters to estimate. Besides the problem of parameter proliferation, this modelling strategy would also require to define an a priori dependence structure on the five Markovian chains. A closer look at the problem reveals that it is not the independence assumption between the Markov chains the main source of concerns. Indeed, we can write the model as follows:

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \mathbf{\Sigma}_t^{1/2} \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I}) \qquad (2)$$

where $\mathbf{I}$ is the 2×2 identity matrix, $\boldsymbol{\mu}_t = (\mu_{1,t}, \mu_{2,t})'$, $\mathbf{\Sigma}_t = \{\sigma_{ij,t}\}_{i,j=1,2}$ and $\sigma_{ii,t} = \sigma^2_{i,t}$,

[2] A similar assumption is that of independence of the shocks in the transition equation of a state space model.

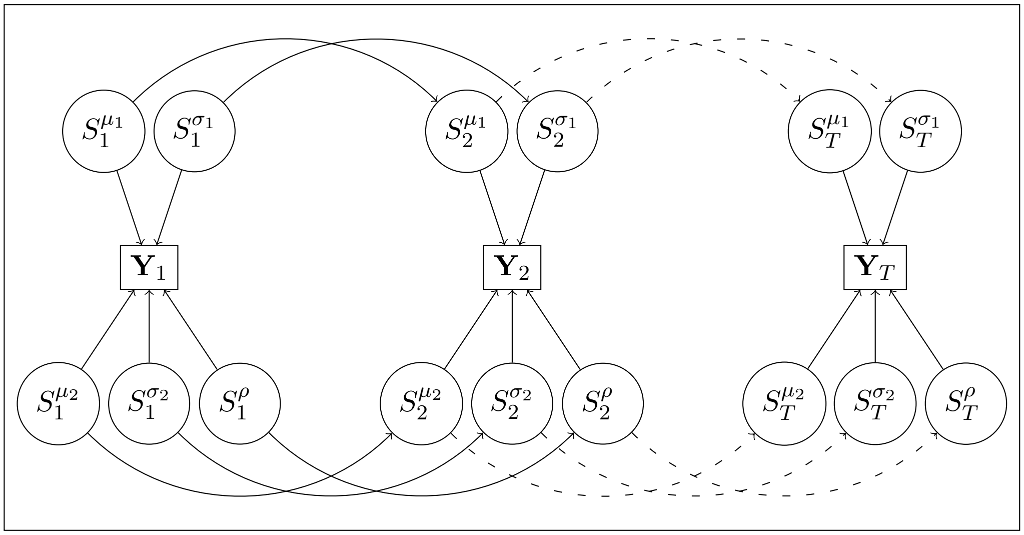


Figure 1: The path diagram of the model in Equation (1).

$\sigma_{ij,t} = \rho_t \sigma_{1,t} \sigma_{2,t}$ for $i, j = 1, 2$, and we indicate, for instance, $\mu_{1,t} = \sum_{i=1}^{D^{\mu_1}} \mu_{1,i} \mathbb{1}_{(S_t^{\mu_1}=i)}$. It follows that, what one should be interested in is the dependence between the dynamic parameters of the model, $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$, and not the Markov chains. Clearly, if the Markov chains are independent, then the dynamic parameters of the model are independent as well, however, the reverse is not true. In Section 5 we present a modelling strategy that preserves the independence of the Markov chains and allows for correlated dynamic parameters, namely $cor(\mu_{1,t}, \mu_{2,t}) \neq 0$, by introducing an additional latent variable. Before then, we continue with the previously introduced parameterization, which is more tractable and of simpler interpretation.

The tasks of filtering, prediction, and smoothing of the unobserved Markov chains are accomplished via the forward filtering backward smoothing (FFBS) algorithm relying of the equivalent representation of the model with a single Markov chain. See Frühwirth-Schnatter (2006) for details on the FFBS algorithm. Specifically, we define a new Markov chain, $\mathcal{S}_t$, with state space $\{(1, 1, 1, 1, 1), \dots, (D^{\mu_1}, D^{\mu_2}, D^{\sigma_1}, D^{\sigma_2}, D^{\rho})\}$, transition probability matrix $\boldsymbol{\Gamma} = \bigotimes_{x \in \mathcal{X}} \boldsymbol{\Gamma}^x$, and initial distribution $\boldsymbol{\delta} = \bigotimes_{x \in \mathcal{X}} \boldsymbol{\delta}^x$.[3] $\mathcal{S}_t$ it is still first order Markov and ergodic, and has $D = \prod_{x \in \mathcal{X}} D^x$ states.

Identification of the model is proven under the following classic set of assumptions: (A1) all Markov chains S_t^x for $x \in \mathcal{X}$ are irreducible, (A2) $\mu_{n,i_n} \neq \mu_{n,l_n}$, $\sigma_{n,j_n} \neq \sigma_{n,q_n}$, and $\rho_k \neq \rho_m$, for all $i_n \neq l_n$, $j_n \neq q_n$, for $n = 1, 2$, and $k \neq m$. The following proposition establishes the identification of the model in Equation (1).

Proposition 2.1 (Identification). *Given Assumptions (A1) and (A2) the model in Equation* (1) *is identified up to label swapping.*

The proof of Proposition 2.1 is done exploiting the single chain representation of the model and by an application of Theorem 1 of Gassiat et al. (2016). Assumption (A1) ensures that the single chain $\mathcal{S}_t$ is irreducible. Assumption (A2) is sufficient to ensures that all state dependent densities in the single chain representation are distinct. Hence, Theorem 1 of Gassiat et al. (2016) can be applied. Given that the identifiability of the model in Equation (1) holds under assumptions (A1) and (A2), the ML estimator obtained via the ECM algorithm outlined in the next section is guaranteed to be the solution of a well posed problem (up to label swapping). Having established identification of the model, and by imposing a unique order of the latent the states (for instance by setting $\mu_{n,j} < \mu_{n,i}$, $\sigma_{n,l} < \sigma_{n,h}$, $\rho_k < \rho_g$ for all $j < i$, $l < h$, $k < g$, and $n = 1, 2$) consistency and asymptotic Normality of the ML estimator follow from Theorem 3 of Leroux (1992) and Theorem 1 of Bickel et al. (1998), respectively.[4] Also, provided that (A1) holds true, straightforward calculations revel that the limiting distribution of

[3] For C matrices or vectors, $\mathbf{A}_c, c = 1, \dots, C$, we use the notation $\bigotimes_{c=1}^{C} \mathbf{A}_c$ to indicate $\mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_C$, where $\otimes$ is the Kronecker product.

[4] For consistency, we must check conditions 1-6 (L1-L6) of Leroux (1992). We note that: L1 is implied by our assumption (A1), L2 is ensured by Proposition 2.1, L3-L6 are straightforwardly satisfied in our context since $\mathbf{Y}_t | \mathcal{S}_t$ is bivariate Normal. For asymptotic Normality, we must check the six assumptions of Bickel et al. (1998) (BRR1-BRR6). We note that: BRR1 is implied by our assumption (A1), BRR2-BRR5 are straightforwardly satisfied in our context since $\mathbf{Y}_t | \mathcal{S}_t$ is bivariate Normal, and BRR6 is consistency of the ML estimator which is ensured by Theorem 3 of Leroux (1992).

$\mathbf{Y}_t$ is that of a mixture of bivariate Normal distributions with mixing probabilities determined by the limiting distribution of the Markov chains. Thus, all moments of $\mathbf{Y}_t$ exist. Furthermore, if $\boldsymbol{\delta}^x = \boldsymbol{\pi}^x$, where $\boldsymbol{\pi}^{x\prime} = \boldsymbol{\pi}^{x\prime}\boldsymbol{\Gamma}^x$ is the limiting distribution of S_t^x, $\mathbf{Y}_t$ is weakly stationary, with an autocorrelation function that converges to zero at the geometric rate λ^q, $q = 1, 2, \ldots$, where $\lambda = \prod_{x\in\mathcal{X}} \lambda^x < 1$, and λ^x is the second largest eigenvalue of $\boldsymbol{\Gamma}^x$, for all $x \in \mathcal{X}$.

3. Estimation

Let $\boldsymbol{\theta} = \left[\boldsymbol{\rho}', \operatorname{vec}(\boldsymbol{\Gamma}^{\rho})', \boldsymbol{\delta}^{\rho\prime}, \boldsymbol{\mu}_j', \boldsymbol{\sigma}_j', \operatorname{vec}(\boldsymbol{\Gamma}^{\mu_j})', \operatorname{vec}\left(\boldsymbol{\Gamma}^{\boldsymbol{\sigma}_j}\right)', \boldsymbol{\delta}^{\mu_j\prime}, \boldsymbol{\delta}^{\sigma_j\prime}, j = 1, 2\right]'$, where $\boldsymbol{\mu}_j = (\mu_{j,1}, \ldots, \mu_{j,D^{\mu_j}})'$, $\boldsymbol{\sigma}_j = (\sigma_{j,1}, \ldots, \sigma_{j,D^{\sigma_j}})'$, for $j = 1, 2$, and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{D^{\rho}})'$, be the vector of parameters. In total there are $K = \sum_{n=1}^{2}\left(D^{\mu_n}(D^{\mu_n} + 1) + D^{\sigma_n}(D^{\sigma_n} + 1)\right) + D^{\rho}(D^{\rho} + 1) - 5$ free parameters to be estimated. Maximum likelihood estimation of $\boldsymbol{\theta}$ can be performed via the expectation conditional maximization algorithm relying on the incomplete data representation of the model. Specifically, we treat the realizations of the variables S_t^x, for $x \in \mathcal{X}$, as missing and proceed to iteratively increase the expected value of the complete data log likelihood (CDLL). The CDLL for a vector of T observations is defined as $\mathcal{L}_c(\boldsymbol{\theta}) = \log P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}, \mathbf{S}_{1:T}^{\rho} = \mathbf{s}_{1:T}^{\rho}, \mathbf{S}_{1:T}^{\mu_j} = \mathbf{s}_{1:T}^{\mu_j}, \mathbf{S}_{1:T}^{\sigma_j} = \mathbf{s}_{1:T}^{\sigma_j}, j = 1, 2; \boldsymbol{\theta})$ and can be written as:

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\theta}) \propto & \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\mu_j}} u_{i_j,1}^{\mu_j} \log \delta_{i_j}^{\mu_j} + \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\sigma_j}} u_{i_j,1}^{\sigma_j} \log \delta_{i_j}^{\sigma_j} + \sum_{k=1}^{D^{\rho}} u_{k,1}^{\rho} \log \delta_k^{\rho} \\
& + \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\mu_j}}\sum_{l_j=1}^{D^{\mu_j}}\sum_{t=2}^{T} v_{i_j,l_j,t}^{\mu_j} \log \gamma_{i_j,l_j}^{\mu_j} + \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\sigma_j}}\sum_{l_j=1}^{D^{\sigma_j}}\sum_{t=2}^{T} v_{i_j,l_j,t}^{\sigma_j} \log \gamma_{i_j,l_j}^{\sigma_j} + \sum_{k=1}^{D^{\rho}}\sum_{q=1}^{D^{\rho}}\sum_{t=2}^{T} v_{k,q,t}^{\rho} \log \gamma_{k,q}^{\rho} \\
& + \sum_{t=1}^{T}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{k=1}^{D^{\rho}} u_{i_1,i_2,j_1,j_2,k,t} \Bigg\{ -\log \sigma_{j_1} - \log \sigma_{j_2} - \frac{1}{2}\log(1-\rho_k^2) \\
& - \frac{1}{2(1-\rho_k^2)} \left[\left(\frac{y_{1,t} - \mu_{1,i_1}}{\sigma_{j_1}} \right)^2 + \left(\frac{y_{2,t} - \mu_{2,i_2}}{\sigma_{j_2}} \right)^2 - 2\rho_k \left(\frac{y_{1,t} - \mu_{1,i_1}}{\sigma_{j_1}} \right) \left(\frac{y_{2,t} - \mu_{2,i_2}}{\sigma_{j_2}} \right) \right] \Bigg\} \qquad (3)
\end{aligned}$$

where $u_{i_1,i_2,j_1,j_2,k,t} = \mathbb{1}_{(S_t^{\mu_1}=i_1, S_t^{\mu_2}=i_2, S_t^{\sigma_1}=j_1, S_t^{\sigma_2}=j_2, S_t^{\rho}=k)}$, and $u^x_{i,1} = \mathbb{1}_{(S^x_1=i)}$, $v^x_{i,j,t} = \mathbb{1}_{(S^x_t=j, S^x_{t-1}=i)}$ for $x \in \mathcal{X}$.[5]

Let $\boldsymbol{\theta}^{(m)}$ be the value of the parameters at iteration m of the algorithm, the value at iteration $m+1$ is selected in a way that ensures

$$\mathcal{Q}(\boldsymbol{\theta}^{(m+1)}; \boldsymbol{\theta}^{(m)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(m)}; \boldsymbol{\theta}^{(m)}), \tag{4}$$

where $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)}) = E[\mathcal{L}_c(\boldsymbol{\theta}) | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}; \boldsymbol{\theta}^{(m)}]$, is the expected value of the CDLL evaluated in $\boldsymbol{\theta}$, where the expectation is taken with respect to the distribution of the unobserved random variables conditional on the observed ones, evaluated with parameters at iteration m. The space over which $\mathcal{Q}$ is maximized, $\boldsymbol{\Theta}$, is constructed to ensure that the constraints on the transition probability matrices and initial distributions of the Markov chains discussed in Section 2 are satisfied, standard deviations are positive, and correlation parameters are in $(-1, 1)$.

The ECM algorithm consists of two steps: *i*) The expectation step (E-step) where $\mathcal{Q}(\cdot;\cdot)$ is computed, and *ii*) a series of conditional maximization (CM) steps leading to $\boldsymbol{\theta}^{(m+1)}$.

The E-step coincides with the evaluation of $\widehat{u}_{i_1,i_2,j_1,j_2,k,t} = E[u_{i_1,i_2,j_1,j_2,k,t} | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}]$, $\widehat{u}^x_{j,1} = E[u^x_{j,1} | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}]$, and $\widehat{v}^x_{i,j,t} = E[v^x_{i,j,t} | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}]$, for $x \in \mathcal{X}$. One run of the FFBS algorithm using the single chain parameterization provides us with $P(\mathcal{S}_t = (i_1, i_2, j_1, j_2, k) | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}) = P(S_t^{\mu_1} = i_1, S_t^{\mu_2} = i_2, S_t^{\sigma_1} = j_1, S_t^{\sigma_2} = j_2, S_t^{\rho} = k | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}) = \widehat{u}_{i_1,i_2,j_1,j_2,k,t}$ from which $\widehat{u}^{\rho}_{k,1}$, and $\widehat{u}^{\mu_j}_{i_j,1}, \widehat{u}^{\sigma_j}_{i_j,1}$, for $j = 1, 2$ are obtained by marginalization setting $t = 1$. Joint probabilities like $E[v^x_{i,j,t} | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}] = P(S^x_t = j, S^x_{t-1} = i | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$ are also computed by marginalization from the output of the FFBS, starting from $P(\mathcal{S}_t = h, \mathcal{S}_{t-1} = l | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$.

For the CM step we split $\boldsymbol{\theta}$ in 6 blocks $\boldsymbol{\theta} = (\boldsymbol{\theta}_i', i = 1, \ldots, 6)'$, where: $\boldsymbol{\theta}_1 = \left[\text{vec}\,(\boldsymbol{\Gamma}^{\rho})', \boldsymbol{\delta}^{\rho\prime}, \text{vec}\,(\boldsymbol{\Gamma}^{\mu_j})', \text{vec}\left(\boldsymbol{\Gamma}^{\boldsymbol{\sigma}_j}\right)', \boldsymbol{\delta}^{\mu_j\prime}, \boldsymbol{\delta}^{\sigma_j\prime}, j = 1, 2\right]'$, $\boldsymbol{\theta}_2 = \boldsymbol{\mu}_1$, $\boldsymbol{\theta}_3 = \boldsymbol{\mu}_2$, $\boldsymbol{\theta}_4 = \boldsymbol{\rho}$, $\boldsymbol{\theta}_5 = \boldsymbol{\sigma}_1$, and $\boldsymbol{\theta}_6 = \boldsymbol{\sigma}_2$.

[5] We use the notation $\mathbf{Y}_{1:T} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$.

In the first CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\theta}_1$. The solution is:

$$\gamma_{i,j,t}^{x\ (m+1)} = \frac{\sum_{t=2}^{T} \widehat{v}_{i,j,t}^{x}}{\sum_{k=1}^{D^x} \sum_{t=2}^{T} \widehat{v}_{i,k,t}^{x}}, \qquad \delta_j^{x\,(m+1)} = \widehat{u}_{j,1}^{x},$$

for all $i, j = 1, \ldots, D^x$ and $x \in \mathcal{X}$. In the second CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\mu}_1$ and obtain:

$$\mu_{1,j_1}^{(m+1)} = \frac{\sum_{t=1}^{T} \sum_{j_1=1}^{D^{\sigma_1}} \sum_{j_2=1}^{D^{\sigma_2}} \sum_{i_2=1}^{D^{\mu_2}} \sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} \left[\frac{y_{1,t} - \rho_k^{(m)} \frac{\sigma_{1,j_1}^{(m)}}{\sigma_{2,j_2}^{(m)}} \left(y_{2,t} - \mu_{2,i_2}^{(m)} \right)}{(\sigma_{1,j_1}^{(m)})^2 \left[1 - (\rho_k^{(m)})^2 \right]} \right]}{\sum_{t=1}^{T} \sum_{j_1=1}^{D^{\sigma_1}} \sum_{j_2=1}^{D^{\sigma_2}} \sum_{i_2=1}^{D^{\mu_2}} \sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} \frac{1}{(\sigma_{1,j_1}^{(m)})^2 \left[1 - (\rho_k^{(m)})^2 \right]}},$$

for $j_1 = 1, \ldots, D^{\mu_1}$. In the third CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\mu}_2$ and obtain:

$$\mu_{2,j_2}^{(m+1)} = \frac{\sum_{t=1}^{T} \sum_{j_1=1}^{D^{\sigma_1}} \sum_{j_2=1}^{D^{\sigma_2}} \sum_{i_1=1}^{D^{\mu_1}} \sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} \left[\frac{y_{2,t} - \rho_k^{(m)} \frac{\sigma_{2,j_2}^{(m)}}{\sigma_{1,j_1}^{(m)}} \left(y_{1,t} - \mu_{1,i_1}^{(m+1)} \right)}{(\sigma_{2,j_2}^{(m)})^2 \left[1 - (\rho_k^{(m)})^2 \right]} \right]}{\sum_{t=1}^{T} \sum_{j_1=1}^{D^{\sigma_1}} \sum_{j_2=1}^{D^{\sigma_2}} \sum_{i_2=1}^{D^{\mu_2}} \sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} \frac{1}{(\sigma_{2,j_2}^{(m)})^2 \left[1 - (\rho_k^{(m)})^2 \right]}},$$

for $j_2 = 1, \ldots, D^{\mu_2}$. In the fourth CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\rho}$. Maximization with respect to $\boldsymbol{\rho}$ is subject to the constraint $\rho_k \in (-1, 1)$ for $k = 1, \ldots, D^{\rho}$. Usually, the maximization step in regime switching correlation models like the ones in Pelletier (2006) and Paolella et al. (2019) is done without restricting $\rho_k \in (-1, 1)$ and subsequently "normalizing" it. Pelletier (2006) argues that this approach leads to estimates that are "very close" to the correct ones. In Section 4 we report a simulation experiment which shows that sizable biases might be obtained, leading to a root mean squared error increase ranging between 30% and almost 400%. More important, the normalizing approach also implies non-increasing values of the log likelihood that might lead to wrong conclusions about the convergence of the algorithm. We now report the correct CM step for ρ_k generalizing the result of Madansky (1965), see also Johnson (1967), Sampson (1978), and Fosdick and Raftery (2012).

We define:

$$\eta_k = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t}, \qquad \xi_k = \eta_k^{-1}\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} z_{1,i_1,j_1,t} z_{2,i_2,j_2,t}$$

$$\nu_{n,k} = \eta_k^{-1}\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} z^2_{n,i_n,j_n,t}, \text{ for } n = 1,2, \qquad c_k = \frac{36\xi_k + 2\xi_k^3 - 9\xi_k(\nu_{1,k}+\nu_{2,k})}{2|3(\nu_{1,k}+\nu_{2,k}-1) - \xi_k^2|^{3/2}},$$

where $z_{n,i_n,j_n,t} = (y_{n,t} - \mu^{(m+1)}_{n,i_n})/(\sigma^{(m)}_{n,j_n})$. Then, if $\eta_k^2 < 3(\nu_{1,k}+\nu_{2,k}-1)$

$$\rho_k^{(m+1)} = \frac{2}{3}\sqrt{3(\nu_{1,k}+\nu_{2,k}-1) - \eta_k^2}\,\sinh\left(\frac{1}{3}\sinh^{-1}(c_k)\right) + \frac{1}{3}\xi_k$$

if $\xi_k^2 \geq 3(\nu_{1,k}+\nu_{2,k}-1)$ we have:

$$\rho_k^{(m+1)} = \begin{cases} \frac{2}{3}\sqrt{\xi_k^2 - 3(\nu_{1,k}+\nu_{2,k}-1)}\cosh\left(\frac{1}{3}\cosh^{-1}(c_k)\right) + \frac{1}{3}\xi_k, & \text{if } c_k \geq 1 \\ \frac{2}{3}\sqrt{\xi_k^2 - 3(\nu_{1,k}+\xi_{2,k}-1)}\cos\left(\frac{4\pi}{3} + \frac{1}{3}\cos^{-1}(c_k)\right) + \frac{1}{3}\xi_k, & \text{if } c_k < 1. \end{cases}$$

In the fifth CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\sigma}_1$, subject to $\boldsymbol{\sigma}_1 > 0$.[6] We first define

$$a_{1,j_1} = \sum_{t=1}^{T}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t}, \quad b_{1,j_1} = \sum_{t=1}^{T}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t}\frac{\rho_k^{(m+1)} z_{1,i_1,j_1,t} z_{2,i_2,j_2,t}}{\sigma^{(m)}_{2,j_2}\left(1 - (\rho_k^{(m+1)})^2\right)}$$

$$c_{1,j_1} = \sum_{t=1}^{T}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} z^2_{1,i_1,j_1,t}\left(1 + \frac{(\rho_k^{(m+1)})^2}{1 - (\rho_k^{(m+1)})^2}\right)$$

then

$$\sigma^{(m+1)}_{1,j_1} = \frac{\sqrt{b^2_{1,j_1} - 4a_{1,j_1}c_{1,j_1}} - b_{1,j_1}}{2a_{1,j_1}},$$

where $z_{n,i_n,j_n,t}$, $n = 1,2$ are defined as in the CM step of $\boldsymbol{\rho}$. In the last CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\sigma}_2$ subject to $\boldsymbol{\sigma}_2 > 0$. To this end, we define $\widetilde{z}_{1,i_1,j_1,t} = (y_{1,t} - \mu^{(m+1)}_{1,i_1})/(\sigma^{(m+1)}_{1,j_1})$, and

$$a_{2,j_2} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t}, \quad b_{2,j_2} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t}\frac{\rho_k^{(m+1)} \widetilde{z}_{1,i_1,j_1,t} z_{2,i_2,j_2,t}}{\sigma^{(m+1)}_{1,j_1}\left(1 - (\rho_k^{(m+1)})^2\right)}$$

$$c_{2,j_2} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} z^2_{2,i_2,j_2,t}\left(1 + \frac{(\rho_k^{(m+1)})^2}{1 - (\rho_k^{(m+1)})^2}\right),$$

[6] For a vector $\mathbf{x}$, the notation $\mathbf{x} > 0$ is intended element-wise.

then

$$\sigma_{2,j_2}^{(m+1)} = \frac{\sqrt{b_{2,j_2}^2 - 4a_{2,j_2}c_{2,j_2}} - b_{2,j_2}}{2a_{2,j_2}}.$$

Starting from an initial guess, $\boldsymbol{\theta}_0$, the algorithm iterates between the E- and CM- steps until convergence which is defined as a relative increase in the log likelihood value of 10^{-8}. The log likelihood, $\mathcal{L}(\boldsymbol{\theta}) = \log P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}; \boldsymbol{\theta})$, is computed by one run of the FFBS algorithm using the single chain parameterization. Convergence to a local optimum is guaranteed since each CM step increases the value of the log likelihood, see Meng and Rubin (1993).

4. Simulation results

This section investigates the finite sample properties of the ML estimator computed via the ECM algorithm previously discussed, by means of an extensive Monte Carlo simulation analysis. We consider a model with two states for all Markov chains, i.e. $D^x = 2$ for all $x \in \mathcal{X}$, corresponding to a single chain parameterization with 32 states and 20 free parameters to estimate. We consider two experiments. In the first experiment, the states of all Markov chains but S_t^ρ are well separated with true parameters set to: $\mu_{1,1} = -1$, $\mu_{1,2} = 1$, $\sigma_{1,1}^2 = 0.5$, $\sigma_{1,2}^2 = 2$, $\mu_{2,1} = -2$, $\mu_{2,2} = 3$, $\sigma_{2,1}^2 = 0.2$, and $\sigma_{2,2}^2 = 1.4$. Correlation parameters are independently sampled from a uniform distribution defined on $(-1, 1)$, i.e. $\rho_k \sim \mathcal{U}(-1, 1)$ for $k = 1, 2$, which means that S_t^ρ might have states which are difficult to separate from a likelihood perspective. In the second experiment we consider all chains that might be not well separated by setting: $\mu_{n,i} \sim \mathcal{U}(-10, 10)$, $\sigma_{n,i}^2 \sim \mathcal{U}(0.01, 5)$, for $n = 1, 2$ and $i = 1, 2$, $\rho_k \sim \mathcal{U}(-1, 1)$ for $k = 1, 2$. For both experiments, self-transition probabilities are all set to 0.99, i.e. $\gamma_{i,i}^x = 0.99$ for all $x \in \mathcal{X}$ with $\gamma_{i,j}^x = 0.01$ for $i \neq j$. Persistent Markov chains like the one we consider help in the estimation process, especially for medium to large samples. We focus on persistent Markov chains because this is the most common situation for empirical applications. Less persistent Markov chains would imply a loss of efficiency and the need for larger sample sizes.

Both experiments proceed as follows: *i*) simulate T observations from the model, and *ii*) estimate parameters using the ECM algorithm. We replicate steps *i*) and *ii*) 10000 times for small ($T = 500$),

	First experiment								Second experiment							
	Bias×1000				RMSE				Bias×1000				RMSE			
$T =$	500	1000	2500	5000	500	1000	2500	5000	500	1000	2500	5000	500	1000	2500	5000
$\mu_{1,1}$	3.712	-0.551	-0.568	0.008	0.090	0.037	0.022	0.015	3.521	0.589	1.228	-0.080	0.217	0.069	0.039	0.028
$\mu_{1,2}$	-1.968	-0.317	0.062	0.247	0.087	0.037	0.022	0.015	2.647	0.609	-0.380	-0.152	0.183	0.067	0.040	0.028
$\gamma^{\mu_1}_{1,1}$	-5.829	-1.349	-0.489	-0.207	0.045	0.006	0.003	0.002	-6.718	-1.605	-0.537	-0.302	0.051	0.008	0.004	0.003
$\gamma^{\mu_1}_{2,2}$	-5.650	-1.418	-0.451	-0.223	0.042	0.006	0.003	0.002	-6.183	-1.450	-0.542	-0.283	0.047	0.007	0.004	0.003
$\sigma_{1,1}$	0.200	-1.723	-0.842	-0.178	0.082	0.034	0.020	0.014	-2.156	7.105	4.405	3.722	0.508	0.380	0.250	0.161
$\sigma_{1,2}$	4.603	-0.494	-0.599	0.220	0.227	0.134	0.080	0.056	13.669	11.054	3.946	1.893	0.558	0.367	0.239	0.155
$\gamma^{\sigma_1}_{1,1}$	-7.231	-1.486	-0.522	-0.279	0.039	0.007	0.003	0.002	-9.214	-4.556	-1.752	-0.816	0.042	0.020	0.010	0.006
$\gamma^{\sigma_1}_{2,2}$	-5.839	-1.628	-0.642	-0.310	0.027	0.007	0.004	0.002	-9.957	-4.131	-1.687	-0.817	0.044	0.019	0.010	0.006
$\mu_{2,1}$	12.802	0.050	-0.013	-0.063	0.221	0.025	0.014	0.010	-3.845	0.046	-0.930	-0.270	0.256	0.070	0.040	0.028
$\mu_{2,2}$	-8.438	0.116	-0.152	-0.030	0.185	0.025	0.014	0.010	-2.267	-0.627	-0.266	-0.526	0.202	0.065	0.041	0.028
$\gamma^{\mu_2}_{1,1}$	-6.523	-1.034	-0.370	-0.181	0.058	0.005	0.003	0.002	-7.075	-1.700	-0.562	-0.208	0.055	0.008	0.004	0.002
$\gamma^{\mu_2}_{2,2}$	-5.861	-1.144	-0.395	-0.225	0.051	0.006	0.003	0.002	-5.870	-1.510	-0.557	-0.201	0.044	0.007	0.004	0.002
$\sigma_{2,1}$	-0.204	-0.670	-0.198	-0.088	0.031	0.013	0.008	0.005	20.800	7.216	3.269	1.993	0.557	0.391	0.244	0.162
$\sigma_{2,2}$	-0.461	1.544	0.540	0.520	0.152	0.088	0.053	0.037	16.977	10.014	3.717	1.672	0.536	0.364	0.265	0.157
$\gamma^{\sigma_2}_{1,1}$	-6.710	-1.419	-0.490	-0.255	0.043	0.006	0.003	0.002	-10.570	-4.423	-1.586	-0.779	0.047	0.020	0.010	0.006
$\gamma^{\sigma_2}_{2,2}$	-5.171	-1.441	-0.475	-0.260	0.027	0.006	0.003	0.002	-10.129	-4.450	-1.859	-0.688	0.043	0.021	0.011	0.006
ρ_1	1.168	0.700	0.239	0.170	0.095	0.066	0.040	0.028	-0.684	0.874	0.208	0.034	0.103	0.065	0.039	0.028
ρ_2	0.353	0.524	-0.665	0.145	0.101	0.065	0.040	0.028	-0.919	-0.592	0.079	-0.525	0.101	0.066	0.040	0.029
$\gamma^{\rho}_{1,1}$	-8.650	-3.610	-1.256	-0.696	0.041	0.016	0.008	0.005	-9.207	-3.657	-1.255	-0.646	0.044	0.019	0.008	0.005
$\gamma^{\rho}_{2,2}$	-8.839	-3.498	-1.309	-0.648	0.041	0.015	0.008	0.005	-9.339	-3.440	-1.366	-0.735	0.040	0.016	0.009	0.005

Table 1: This table reports the bias multiplied by 1000 and the root mean squared error for the maximum likelihood estimator computed with the EM algorithm. Statistics are based on 10000 samples for different sample sizes, T, with $D^x = 2$ for all $x \in \mathcal{X}$. In the first experiment true parameters are: $\mu_{1,1} = -1$, $\mu_{1,2} = 1$, $\sigma^2_{1,1} = 0.5$, $\sigma^2_{1,2} = 2$, $\mu_{2,1} = -2$, $\mu_{2,2} = 3$, $\sigma^2_{2,1} = 0.2$, $\sigma^2_{2,2} = 1.4$, $\rho_k \sim \mathcal{U}(-1, 1)$ for $k = 1, 2$, and $\gamma^x_{i,i} = 0.99$ for all $x \in \mathcal{X}$. In the second experiment true parameters are: $\mu_{n,i} \sim \mathcal{U}(-10, 10)$, $\sigma^2_{n,i} \sim \mathcal{U}(0.01, 5)$, for $n = 1, 2$ and $i = 1, 2$, $\rho_k \sim \mathcal{U}(-1, 1)$ for $k = 1, 2$, and $\gamma^x_{i,i} = 0.99$ for all $x \in \mathcal{X}$.

medium small ($T = 1000$), medium large ($T = 2500$), and large ($T = 5000$) sample sizes. Table 1 reports the bias multiplied by 1000 and root mean squared error (RMSE) between true and estimated parameters over all replications. Results are good and indicate that, even when the sample is small ($T = 500$), the bias and RMSE are generally small. The bias increases substantially moving from the first to the second experiment, especially for small samples. For instance, the bias for $\sigma_{2,1}$ when $T = 500$ is -0.0002 in the first experiment and 0.0208 in the second one. Overall, the bias and RMSE decrease with the increase of the sample size, indicating a proper convergence of the estimator.

4.1. The effect of implementing the wrong M-step for ρ_m

It is common practice in the literature on switching correlation models to implement a wrong M step for the state dependent correlation parameters. The idea is to compute the M step of the covariance matrix of a conditionally Gaussian model, and then "normalizing" it. This is done for example in Equation (3.10) of Pelletier (2006) where it is reported that: "By doing this transformation, the estimates obtained with these equations will not exactly be the numerical maximum of the likelihood, but very close to it", where "transformation" refers to the aforementioned normalization step. Pelletier (2006) indicates that few steps of numerical maximization starting from the "normalized" estimates are enough to converge to the correct optimal solution. A similar approach is employed in Equation (13) of Paolella et al. (2019).[7]

We now report a simulation analysis aimed at quantifying the effect of implementing the wrong conditional maximization step. We consider a model where only the correlation parameter switches between $D^{\rho} = 2$ states, by setting $D^{\mu_1} = D^{\mu_2} = D^{\sigma_1} = D^{\sigma_2} = 1$ and $\mu_{1,1} = \mu_{2,1} = 0$, and $\sigma^2_{1,1} = \sigma^2_{2,1} = 1$. So, the random variable $\mathbf{Y}_t | S^{\rho}_t = k$ is bivariate standard Normal with correlation ρ_k. In the experiment, we fix $\mu_{i,1}$ and $\sigma^2_{i,1}$ for $i = 1, 2$ to their true values. The transition probability matrix, Γ^{ρ}, has entries $\gamma^{\rho}_{ii} = 0.95$ for $i = 1, 2$ and $\gamma_{ij} = 0.05$ for $i \neq j$. At each repetition of the experiment, correlation parameters are selected randomly from a Uniform distribution, $\rho_k \sim \mathcal{U}(-1, 1)$.

We modify the ECM algorithm of Section 3 by implementing the maximization step implied by

[7] Note that, the proof of Theorem 2 of Paolella et al. (2019) requires the correct conditional maximization step for the correlation parameters, like the one we report in Section 3, rather than the "normalized" one.

$T =$	500	1000	2500	5000	10000
ρ_1	1.350	1.521	1.733	1.923	1.890
ρ_2	1.305	1.456	1.697	1.870	1.947
$\gamma^{\rho}_{1,1}$	2.571	2.858	3.613	3.659	3.134
$\gamma^{\rho}_{2,2}$	2.311	2.808	3.555	3.841	3.289

Table 2: This table reports ratio between the root mean squared errors of estimates computed with the wrong and correct M steps for ρ_k $k = 1, 2$. Statistics are based on 10000 samples for different sample sizes, T.

Equation (3.10) of Pelletier (2006). Specifically, we set $\rho_k^{(m+1)}$ as the element $(1, 2)$ of the 2×2 correlation matrix $\mathbf{R}_k^{(m+1)} = \mathbf{D}_k^{-1/2}\tilde{\mathbf{R}}_k\mathbf{D}_k^{-1/2}$ where

$$\tilde{\mathbf{R}}_k = \frac{\sum_{t=1}^{T}(\mathbf{y}_t\mathbf{y}_t')\widehat{u}^{\rho}_{k,t}}{\sum_{t=1}^{T}\widehat{u}^{\rho}_{k,t}},$$

and $\mathbf{y}_t = (y_{1,t}, y_{2,t})'$, and $\mathbf{D}_k$ is a 2×2 diagonal matrix with entries equal to the diagonal entries of $\tilde{\mathbf{R}}_k$. The computation of $\tilde{\mathbf{R}}_k$ is the M step of an EM algorithm for the covariance matrix of a conditional Gaussian model with mean parameters equal to zero, while computing $\mathbf{R}_k^{(m+1)}$ coincides with the aforementioned normalization step.

The experiment proceeds as follows: *i*) simulate T observations from the model, and *ii*) estimate the correlation parameters ρ_k, $k = 1, 2$, and the transition probability matrix, Γ^{ρ}, with the wrong and correct M steps. We replicate steps *i*) and *ii*) 10000 times for $T \in \{500, 1000, 2500, 5000, 10000\}$.

Table 2 reports the ratio between the RMSEs of estimates computed with the wrong and correct M steps. A value higher than one indicates that the wrong M step leads to higher RMSE compared to the correct one. For instance, in the case of ρ_1 with $T = 2500$, the increase in RMSE is 73.3%. We note that, as expected, the ratio between the two RMSEs increases with T, suggesting that the wrong M step leads to inconsistent estimates of the correlation parameters and the transition probability matrix. Estimates of the transition probabilities are affected the most by the wrong M step, indicating that misleading inference about the evolution of S_t^{ρ} can result from a wrong implementation of the

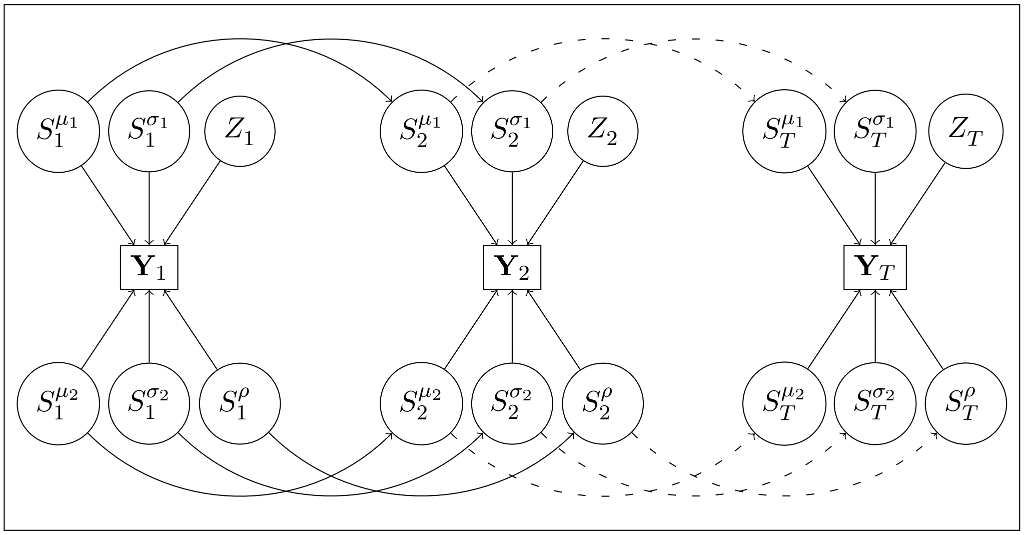


Figure 2: The path diagram of the model in Equation (5).

EM algorithm. Overall, our results indicate that implementing the correct M step is of primary importance for a proper estimation of regime switching correlation models.

5. Introducing correlated dynamic parameters

The independence assumption of the five Markovian chains might seems at a first glance overly restrictive. In this section, we argue that this is not necessarily the case, and provide a modelling strategy aimed at mitigating its implications.

Consider for example the unobserved components representation of the model provided in Equation (2). If the Markov chains are independent, the dynamic parameters of the model, $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$, are independent as well. However, for applications with moderate sample sizes, assuming independence for the unobserved dynamic parameters of a (possibly nonlinear) state space model is actually a quite common strategy to reduce the number of parameters of the model. For example, in the time varying vector autoregression with stochastic volatility of Primiceri (2005), unobserved components are assumed independent among them.

We now discuss a modelling strategy that induce correlation between the dynamic parameters of the model, but retains the independence assumption of the Markov chains. The idea is to introduce a new latent variable, Z_t, defined on the discrete state space $\{1, \dots, M\}$, and let all the dynamic parameters to depend on its realization by setting, for instance, $\mu_{1,t} = \sum_{i=1}^{D^{\mu_1}} \sum_{m=1}^{M} \mu_{1,i,m} \mathbb{1}_{(S_t^{\mu_1}=i, Z_t=m)}$, and similarly for the other parameters. The new variable is independent from the Markov chains, $Z_t \perp\!\!\!\perp S_t^x$ for all $x \in \mathcal{X}$, and it is assumed to be independently distributed as a Categorical distribution with $P(Z_t = m) = \omega_m > 0$ for $m = 1, \dots, M$ and $\sum_{m=1}^{M} \omega_m = 1$. Thus, the new model is defined as

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} \left| \begin{matrix} S_t^{\mu_1} = a & S_t^{\mu_2} = b \\ S_t^{\sigma_1} = c & S_t^{\sigma_2} = d \\ S_t^{\rho} = e & Z_t = m \end{matrix} \right. \sim N \left(\begin{bmatrix} \mu_{1,a,m} \\ \mu_{2,b,m} \end{bmatrix} ; \begin{bmatrix} \sigma_{1,c,m}^2 & \sigma_{1,c,m}\sigma_{2,d,m}\rho_{e,m} \\ \sigma_{1,c,m}\sigma_{2,d,m}\rho_{e,m} & \sigma_{2,d,m}^2 \end{bmatrix} \right), \tag{5}$$

and its path diagram is represented in Figure 2.

Note that, the model in Equation (5), as the one of Equation (1), has an equivalent stochastic representation in terms of a HMM with a single Markov chain, $\tilde{\mathcal{S}}_t = \{\mathcal{S}_t, Z_t\}$, with state space $\{(1, 1, 1, 1, 1, 1), \dots, (D^{\mu_1}, D^{\mu_2}, D^{\sigma_1}, D^{\sigma_2}, D^{\rho}, M)\}$, transition probability matrix $\tilde{\mathbf{\Gamma}} = (\mathbf{u}_D \mathbf{u}_D') \otimes \boldsymbol{\omega}' \otimes \mathbf{u}_M \odot \mathbf{\Gamma}$, and initial distribution $\tilde{\boldsymbol{\delta}} = \boldsymbol{\omega} \otimes \boldsymbol{\delta}$, where $\mathbf{u}_C$ is a vector of ones of length C, $\odot$ is the Hadamard product, and D, $\mathbf{\Gamma}$, and $\boldsymbol{\delta}$ are defined in Section 2. $\tilde{\mathcal{S}}_t$ it is still first order Markov and ergodic, and has $\tilde{D} = DM$ states. When $M = 1$, the models of Equations (1) and (5) coincides.

We now establish the identification of the model in Equation (5) under assumptions (A1), and the following additional sufficient condition (A2)$'$: $\mu_{n,i_n,m} \neq \mu_{n,l_n,g}$, $\sigma_{n,j_n,m} \neq \sigma_{n,q_n,g}$, and $\rho_{k,m} \neq \rho_{h,g}$, for all $i_n \neq l_n$, $j_n \neq q_n$, for $n = 1, 2$, and $m \neq g$.

Proposition 5.1 (Identification of the model with Z_t). *Given Assumptions (A1) and (A2)$'$ the model in Equation* (5) *is identified up to label swapping.*

The proof of Proposition 5.1 is done exploiting the single chain representation of the five chains, $\mathcal{S}_t$, and the fact that $\mathbf{Y}_t | \mathcal{S}_t$ is a mixture of M bivariate Normal distributions with mixing weights $\omega_m > 0$,

$m = 1, \dots, M$. Note that, by assumption (A2)$'$ these mixture of bivariate Normal distributions are distinct, such that the proof of Proposition 5.1 follows by an application of Theorem 1 of Gassiat et al. (2016). An EM algorithm for the model of Equation (5) to compute the maximum likelihood estimator of the model parameters is reported in Appendix A. Filtering and smoothing of all latent variables is done by a run of the FFBS algorithm exploiting the single chain representation of the model, $\tilde{\mathcal{S}}_t$. The model of Equation (5) has $K = \sum_{n=1}^{2} \left(D^{\mu_n}(D^{\mu_n} + M) + D^{\sigma_n}(D^{\sigma_n} + M)\right) + D^{\rho}(D^{\rho} + M) + M - 6$ free parameters, which are $\sum_{n=1}^{2} \left(D^{\mu_n}(M-1) + D^{\sigma_n}(M-1)\right) + D^{\rho}(M-1) + M - 1$ more than the one of Equation (1). When M is large, the model soon becomes highly parameterized. However, if we would have opt for correlated Markov chains employing, for example, the following factorization of their joint distribution $P(S_t^{\mu_1}|S_t^{\mu_2}, S_t^{\sigma_1}, S_t^{\sigma_2}, S_t^{\rho})P(S_t^{\mu_2}|S_t^{\sigma_1}, S_t^{\sigma_2}, S_t^{\rho})P(S_t^{\sigma_1}|S_t^{\sigma_2}, S_t^{\rho})P(S_t^{\sigma_2}|S_t^{\rho})P(S_t^{\rho})$, the number of parameters would have grown by $D^{\mu_1}(D^{\mu_1} - 1)(D^{\mu_2}D^{\sigma_1}D^{\sigma_2}D^{\rho} - 1) + D^{\mu_2}(D^{\mu_2} - 1)(D^{\sigma_1}D^{\sigma_2}D^{\rho} - 1) + D^{\sigma_1}(D^{\sigma_1} - 1)(D^{\sigma_2}D^{\rho} - 1) + D^{\sigma_2}(D^{\sigma_2} - 1)(D^{\rho} - 1)$, which would have implied an extremely parameterized specification.

The fact that Z_t affects all parameters induce the aforementioned correlation. To see this, consider the correlation between the mean and the variance of the first series, namely $cor(\mu_{1,t}, \sigma_{1,t}^2)$, which is non zero if $E[\mu_{1,t}\sigma_{1,t}^2] \neq E[\mu_{1,t}]E[\sigma_{1,t}^2]$. Clearly, if $M = 1$, $E[\mu_{1,t}\sigma_{1,t}^2] = E[\mu_{1,t}]E[\sigma_{1,t}^2]$ by construction. However, if $M > 1$ we have

$$E[\mu_{1,t}\sigma_{1,t}^2] = \sum_{i=1}^{D^{\mu_1}} \sum_{j=1}^{D^{\sigma_1}} \sum_{m=1}^{M} \mu_{1,i,m}\sigma_{1,j,m}^2 \delta_i^{\mu_1} \delta_j^{\sigma_1} \omega_m,$$

where we have assumed that the initial distributions of $S_t^{\mu_1}$ and $S_t^{\sigma_1}$ coincide with their limiting distributions, such that the chains are stationary. Note that, in the case $M > 1$ the result $E[\mu_{1,t}\sigma_{1,t}^2] \neq E[\mu_{1,t}]E[\sigma_{1,t}^2]$ is admissible. The same arguments hold for every pair of parameters, such that the covariance matrix of the random variable $\boldsymbol{\xi}_t = (\mu_{1,t}, \mu_{2,t}, \sigma_{1,t}^2, \sigma_{2,t}^2, \rho_t)'$ is constructed as

$cov(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t) = E[\boldsymbol{\xi}_t \boldsymbol{\xi}_t'] - E[\boldsymbol{\xi}_t]E[\boldsymbol{\xi}_t']$ where

$$\begin{aligned}
E[\boldsymbol{\xi}_t \boldsymbol{\xi}_t'] &= \sum_{j_1=1}^{D^{\sigma_1}} \sum_{j_2=1}^{D^{\sigma_2}} \sum_{i_1=1}^{D^{\mu_1}} \sum_{i_2=1}^{D^{\mu_2}} \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbf{b}_{i_1,i_2,j_2,j_2,k,m} \mathbf{b}'_{i_1,i_2,j_2,j_2,k,m} \delta_{i_1}^{\mu_1} \delta_{i_2}^{\mu_2} \delta_{j_1}^{\sigma_1} \delta_{j_2}^{\sigma_2} \delta_k^{\rho} \omega_m, \\
E[\boldsymbol{\xi}_t] &= \sum_{j_1=1}^{D^{\sigma_1}} \sum_{j_2=1}^{D^{\sigma_2}} \sum_{i_1=1}^{D^{\mu_1}} \sum_{i_2=1}^{D^{\mu_2}} \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbf{b}_{i_1,i_2,j_2,j_2,k,m} \delta_{i_1}^{\mu_1} \delta_{i_2}^{\mu_2} \delta_{j_1}^{\sigma_1} \delta_{j_2}^{\sigma_2} \delta_k^{\rho} \omega_m,
\end{aligned} \tag{6}$$

and $\mathbf{b}_{i_1,i_2,j_2,j_2,k,m} = (\mu_{1,i_1,m}, \mu_{2,i_2,m}, \sigma^2_{1,j_1,m}, \sigma^2_{2,j_2,m}, \rho^2_{k,m})'$.

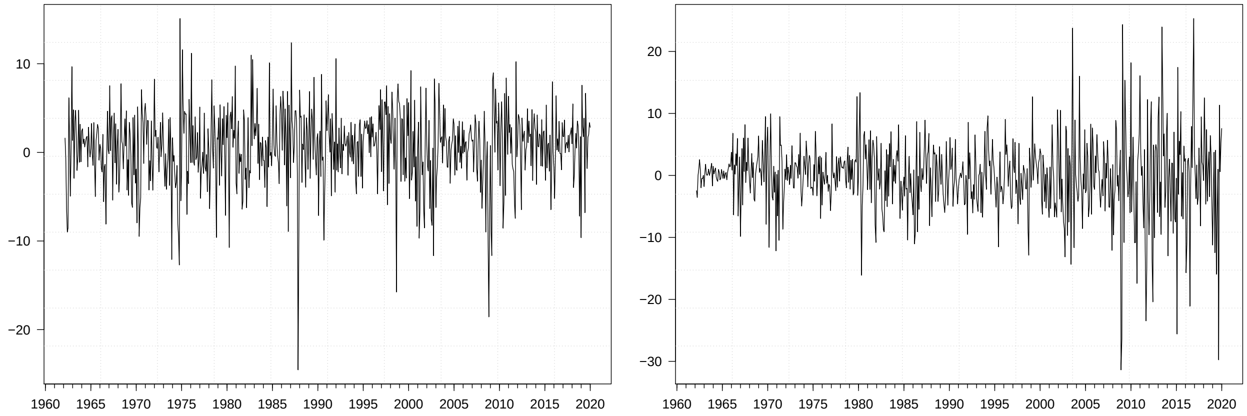
In the following empirical illustration, the choice of M, together with that of D^x for $x \in \mathcal{X}$ is done using penalized likelihood criteria.

6. Empirical illustration

Guidolin and Timmermann (2006) find that the marginal distributions of US stocks and bond returns are characterized by well-defined Markovian regimes. Though, since there is very little coherence between these regimes, a large system need to be defined in order to properly describe their joint distribution. They end up estimating a moderately large parameterized model with four regimes which they refer to as: *crash*, *slow growth*, *bull*, and *recovery*.[8] The MCHMM perfectly fits this setting and allows us to estimate a less parameterized model, leading to an increased interpretability of estimated parameters.

For our analysis, we consider the Standard & Poor's 500 index (SP500) and the ten years treasury constant maturity rate (DSG10), representing the US stocks and bond markets, respectively. Data points are collected at the monthly frequency from January 1962 to December 2019, SP500 is downloaded from Yahoo finance and DSG10 from the Federal Reserve Bank of St. Louis web site. The augmented Dickey-Fuller test indicates the presence of a stochastic trend for both series, which we remove by applying the usual logarithmic difference transformation. The resulting series are

[8] They also consider an autoregressive specification for the conditional mean of the process, but ended up estimating a model that do not include this feature. Also, they divide the stock market in large and small capitalized firms estimating a trivariate model. However, large and small capitalized firms essentially follow the same regimes.



(a) SP500

(b) DGS10

Figure 3: Standard & Poor's 500 index (SP500) and the 10-Years Treasury Constant Maturity Rate (DSG10) percentage logarithmic returns at the monthly frequency from January 1962 to December 2019.

multiplied by 100, consist of 695 data points, and are displayed in Figure 6. Usual financial returns stylized facts, such as the presence of heteroscedasticity and extreme observations (given actual volatility levels), clearly emerge from the figures. Standard hidden Markov models are well suited for each univariate series, as discussed for example by Rydén et al. (1998).

We estimate the MCHMM on the bivariate data set with $D^x \in \{1, \ldots, 4\}$ for all $x \in \mathcal{X}$ and $M \in \{1, \ldots, 5\}$ resulting in a total of 5120 specifications. Given the closed form of the CM steps in the EM algorithm, estimation of a single specification is not computationally expensive. We follow Guidolin and Timmermann (2006) and select the number of regimes using the Hannan-Quinn information criterion (HQC).[9]

We consider two models, one with $M = 1$ which assumes independence between the dynamic parameters of the model, and one with M selected via the HQC criterion. In the case $M = 1$, the specification that minimizes HQC is $D^{\mu_1} = D^{\mu_2} = 1$, $D^{\sigma_1} = 2$, $D^{\sigma_2} = 3$, and $D^{\rho} = 2$, where the

[9] HQC is defined as $\text{HQC} = -2L + 2K \log(\log(T))$, where L is the value of the log likelihood at its maximum and K is the number of parameters.

	MCHMM					HMM				
State	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_1^2$	$\widehat{\sigma}_2^2$	$\widehat{\rho}$	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_1^2$	$\widehat{\sigma}_2^2$	$\widehat{\rho}$
1	0.73 (0.13)	0.16 (0.15)	6.09 (0.78)	1.06 (1.03)	-0.40 (0.07)	0.84 (0.79)	0.52 (0.42)	9.12 (3.26)	2.03 (1.21)	-0.21 (0.25)
2	-	-	26.41 (2.29)	18.73 (1.51)	0.29 (0.08)	0.86 (0.25)	0.02 (0.29)	14.05 (1.39)	19.10 (1.94)	-0.37 (0.06)
3	-	-	-	95.79 (13.71)	-	-2.15 (1.49)	-0.81 (1.24)	47.41 (12.56)	33.07 (8.97)	0.29 (0.19)
4	-	-	-	-	-	0.97 (0.65)	-0.41 (1.22)	14.69 (3.16)	97.91 (19.73)	0.32 (0.15)

Table 3: Maximum likelihood estimated parameters excluding transition probabilities for the MCHMM and HMM specifications. Standard errors computed via the parametric bootstrap (Tibshirani and Efron, 1993) are reported in brackets and are based on 10000 replicates.

first series is SP500, and the second one is DSG10. The selected specification has 23 parameters and reports a HQC value of 8143.8. If M is selected together with the number of regimes, the specification that minimizes HQC is the one with $D^{\mu_1} = 3$, $D^{\mu_2} = 3$, $D^{\sigma_1} = 1$, $D^{\sigma_2} = 3$, $D^{\rho} = 2$, and $M = 4$. The HQC value for this model is 7953.6, and the number of parameters is 71. Note that, this specification is much more parameterized than the one obtained by setting $M = 1$. For the HMM specification used by Guidolin and Timmermann (2006), HQC selects four regimes also with our data set. The estimated specification has 47 parameters and a HQC value of 8230.5. Remarkably, the selected MCHMM specification with $M = 1$ is less parameterized and reports a lower HQC value.[10] However, while reporting a sizable decrease in HQC, the specification with $M = 4$ is more parameterized than the one obtained by Guidolin and Timmermann (2006).

Estimated parameters for the MCHMM with $M = 1$ and HMM specifications are reported in Table 3. Standard errors are computed via the parametric bootstrap (Tibshirani and Efron, 1993) and are based on 10000 replicates. The estimated transition probability matrices for MCHMM with

[10] Results based on the Akaike and Bayes information criteria are analogous.

$M = 1$ are

$$\widehat{\mathbf{\Gamma}}^{\sigma_1} = \begin{pmatrix} \underset{(0.02)}{0.95} & \underset{(0.02)}{0.05} \\ \underset{(0.02)}{0.03} & \underset{(0.02)}{0.97} \end{pmatrix}, \qquad \widehat{\mathbf{\Gamma}}^{\sigma_2} = \begin{pmatrix} \underset{(0.25)}{0.96} & \underset{(0.25)}{0.04} & \underset{(0.00)}{0.00} \\ \underset{(0.02)}{0.00} & \underset{(0.02)}{0.99} & \underset{(0.01)}{0.01} \\ \underset{(0.00)}{0.00} & \underset{(0.04)}{0.02} & \underset{(0.04)}{0.98} \end{pmatrix}, \qquad \widehat{\mathbf{\Gamma}}^{\rho} = \begin{pmatrix} \underset{(0.02)}{0.99} & \underset{(0.02)}{0.01} \\ \underset{(0.02)}{0.01} & \underset{(0.02)}{0.99} \end{pmatrix},$$

while for HMM the estimated transition probability matrix is

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} \underset{(0.12)}{0.95} & \underset{(0.12)}{0.05} & \underset{(0.03)}{0.00} & \underset{(0.00)}{0.00} \\ \underset{(0.02)}{0.01} & \underset{(0.03)}{0.97} & \underset{(0.02)}{0.02} & \underset{(0.00)}{0.00} \\ \underset{(0.04)}{0.02} & \underset{(0.08)}{0.07} & \underset{(0.11)}{0.88} & \underset{(0.06)}{0.03} \\ \underset{(0.00)}{0.00} & \underset{(0.07)}{0.01} & \underset{(0.00)}{0.00} & \underset{(0.07)}{0.99} \end{pmatrix},$$

where "0.00" means a number lower than 0.005. Estimates of the MCHMM model with $M = 4$ are more difficult to interpret and are not reported to save space. The correlation matrix of the dynamic parameters implied by the model with $M = 4$ is reported in Table 4. We see that the correlation between the two means, $\mu_{1,t}$ and $\mu_{2,t}$, is negative and small (-0.030), while the correlation between the variance parameters is positive and large (0.275). We also note that the correlation between the mean and the variance of the SP500 (first series) is negative and sizable (-0.292) indicating the presence of contemporaneous leverage effect, meaning that the variance increases when returns are negative, see Black (1976) and Catania (2020) for a recent specification that includes this feature. For US bonds (second series) the correlation between $\mu_{2,t}$ and $\sigma^2_{2,t}$ is still negative, but smaller (-0.093). Overall, the correlation parameter, ρ_t, is the one most correlated with the other parameters.

Results from MCHMM with $M = 1$ indicate that both series are characterized by small positive means, with μ_2 very close to zero (t-test of 1.06). Volatility as well as correlation regimes are persistent and well separated for MCHMM, with $M = 1$. Regimes in the HMM specification are also moderately persistent but are more difficult to interpret. Standard errors for estimated means are large for DSG10, and for SP500 in regime 1. According to the definition of Guidolin and Timmermann

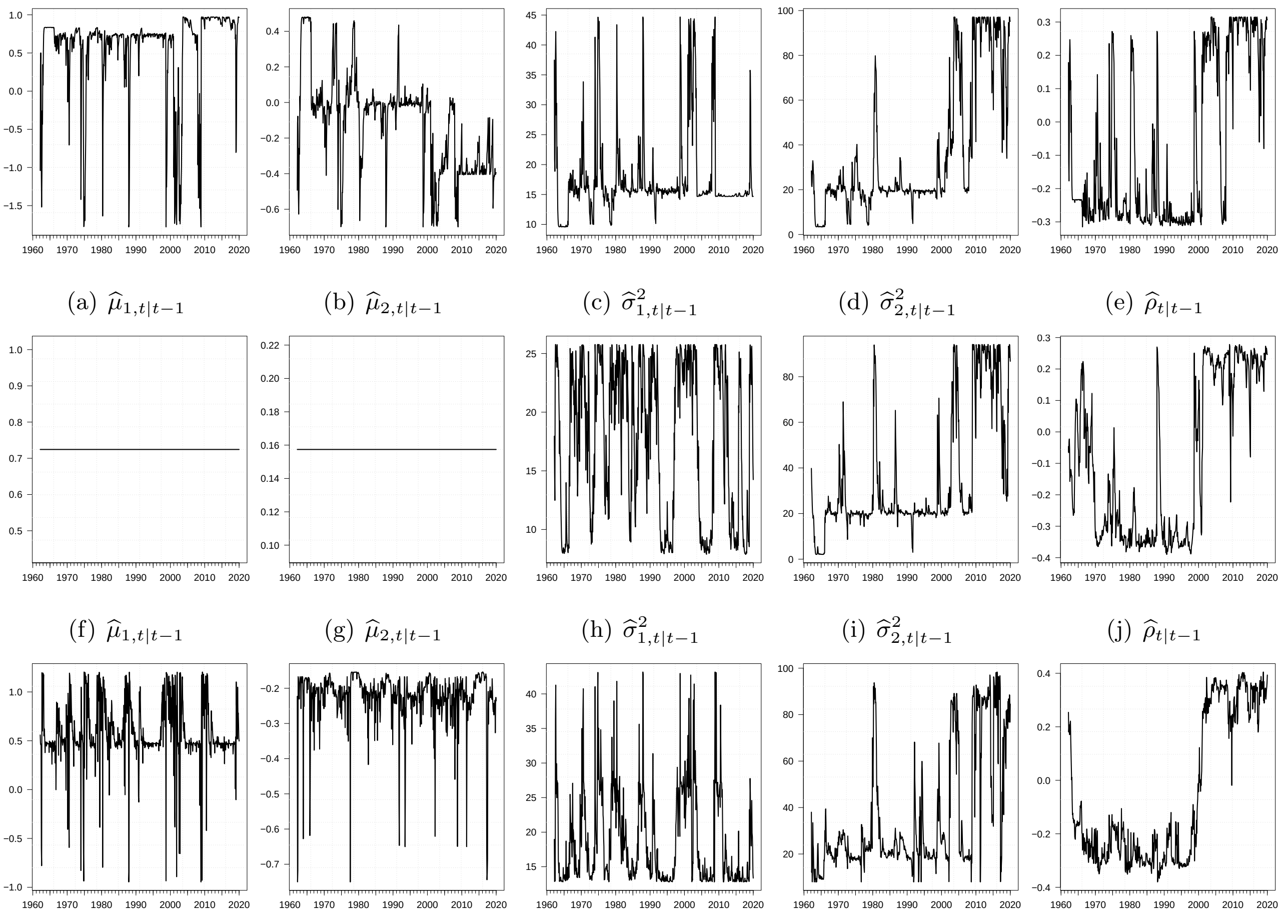
	$\mu_{1,t}$	$\mu_{2,t}$	$\sigma^2_{1,t}$	$\sigma^2_{2,t}$	ρ_t
$\mu_{1,t}$	1.000	-0.030	-0.292	-0.093	-0.158
$\mu_{2,t}$	-0.030	1.000	-0.147	-0.024	-0.243
$\sigma^2_{1,t}$	-0.292	-0.147	1.000	0.275	0.158
$\sigma^2_{2,t}$	-0.093	-0.024	0.275	1.000	0.037
ρ_t	-0.158	-0.243	0.158	0.037	1.000

Table 4: Estimated correlation matrix between the dynamic parameters of the MCHMM model with $M = 4$. The first series is SP500, and the second one is DSG10.

(2006), our regime 3 is the *crash* state characterized by large negative mean, high volatilities, and positive correlation. Regime 1 is the *slow growth* state, with low volatility and small positive mean. Their definition of the *bull* and *recovery* states does not really apply to our estimates. Differences are also present when comparing our estimated transition probability matrix and the one they report. The reason for these inconsistencies is probably related to the different estimation period, indeed, their analysis include data up to December 1999.

All specifications estimate a switch in the sign of the correlation between stocks and bond returns. This result is well known in the literature and has been investigated by Andersen et al. (2007) in their analysis of news response. Earlier evidences of this phenomenon are reported by Erb et al. (1994), King et al. (1994), Longin and Solnik (1995, 2001), and De Santis and Gerard (1997). Ang and Bekaert (2002a) exploit this result to show that the existence of a high-volatility bear market does not negate the benefits of international diversification if investments follow a regime switching asset allocation setting.

To compare the HMM model of Guidolin and Timmermann (2006) with the MCHMM with $M = 1$ and $M = 4$, in Figure 4 we report the estimated moments of the random variable $\mathbf{Y}_t|\mathcal{F}_{t-1}$, for $t = 1, \ldots, Y$. The distribution of $\mathbf{Y}_t|\mathcal{F}_{t-1}$ is a mixture of bivariate Normal distributions with mixing weights equal to the predictive probabilities obtained by a run of the forward filter. For the



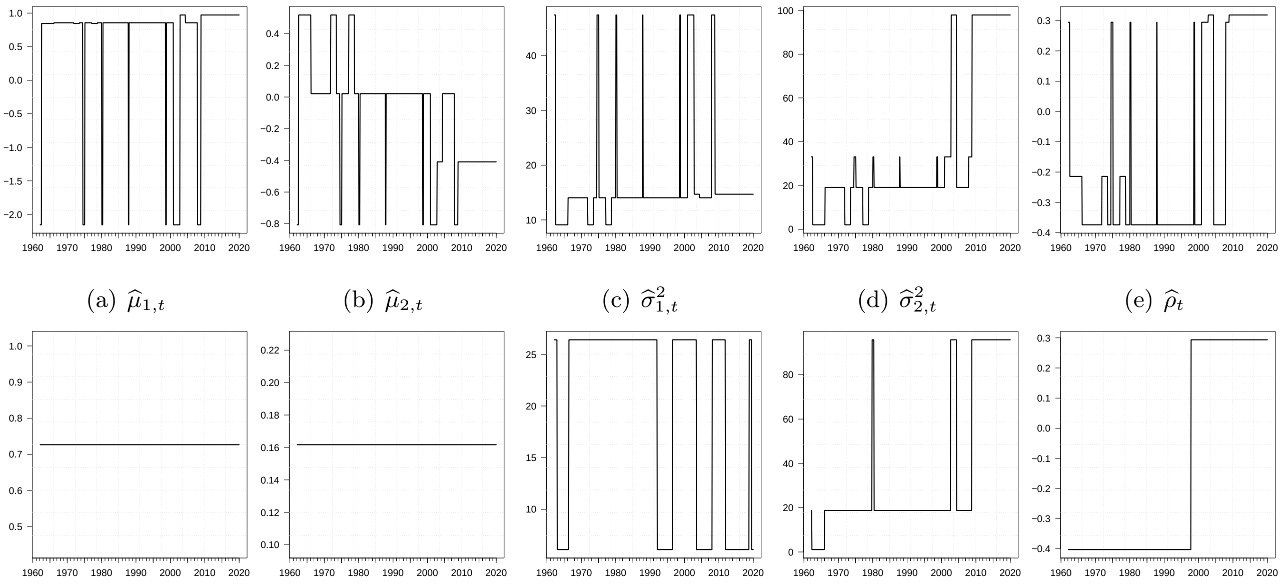
(a) $\widehat{\mu}_{1,t|t-1}$ (b) $\widehat{\mu}_{2,t|t-1}$ (c) $\widehat{\sigma}^2_{1,t|t-1}$ (d) $\widehat{\sigma}^2_{2,t|t-1}$ (e) $\widehat{\rho}_{t|t-1}$

(f) $\widehat{\mu}_{1,t|t-1}$ (g) $\widehat{\mu}_{2,t|t-1}$ (h) $\widehat{\sigma}^2_{1,t|t-1}$ (i) $\widehat{\sigma}^2_{2,t|t-1}$ (j) $\widehat{\rho}_{t|t-1}$

(k) $\widehat{\mu}_{1,t|t-1}$ (l) $\widehat{\mu}_{2,t|t-1}$ (m) $\widehat{\sigma}^2_{1,t|t-1}$ (n) $\widehat{\sigma}^2_{2,t|t-1}$ (o) $\widehat{\rho}_{t|t-1}$

Figure 4: Moments of $\mathbf{Y}_t|\mathcal{F}_{t-1}$ estimated by the HMM (panels (a)–(e)), the MCHMM with $M = 1$ (panels (f)–(j)), and the MCHMM with $M = 4$ (panels (k)–(o)).

MCHMM predictive probabilities are computed using the single chain representation of the model. In Figure 4, we use the notation $\widehat{\mu}_{i,t|t-1} = E[Y_{i,t}|\mathcal{F}_{t-1}]$, $\widehat{\sigma}^2_{i,t|t-1} = Var[Y_{i,t}|\mathcal{F}_{t-1}]$ for $i = 1, 2$, and $\widehat{\rho}_{t|t-1} = cor(Y_{1,t}, Y_{2,t}|\mathcal{F}_{t-1})$. Results indicate that the three models generally agree about the evolution of the first two moments of $\mathbf{Y}_t|\mathcal{F}_{t-1}$. For all specifications, the switch in the sign of the correlation between SP500 and the US bonds is evident and takes place around the beginning of 2000. We note



(a) $\widehat{\mu}_{1,t}$ (b) $\widehat{\mu}_{2,t}$ (c) $\widehat{\sigma}^2_{1,t}$ (d) $\widehat{\sigma}^2_{2,t}$ (e) $\widehat{\rho}_t$

(f) $\widehat{\mu}_{1,t}$ (g) $\widehat{\mu}_{2,t}$ (h) $\widehat{\sigma}^2_{1,t}$ (i) $\widehat{\sigma}^2_{2,t}$ (j) $\widehat{\rho}_t$

Figure 5: Global decoded parameters values from the HMM and MCHMM with $M = 1$ specifications using the Viterbi (1967) algorithm. HMM parameters are reported in panels (a)–(e), while MCHMM parameters in panels (f)–(j).

that the MCHMM with $M = 4$ (panels (k)–(o)) exhibits a more regular evolution of the correlation parameter compared to MCHMM with $M = 1$ (panels (f)–(j)) and the HMM (panels (a)–(e)). This result directly follows from the increased parameterization of MCHMM with $M = 4$.

We conclude our analysis performing global decoding of the unobserved chains in the MCHMM with $M = 1$ and HMM specifications. Global decoding is achieved via a run of the Viterbi (1967) algorithm. For MCHMM with $M = 1$, we run the algorithm on the single chain representation of the model and subsequently recover the states of the individual chains. For MCHMM with $M = 4$ global decoding reports very irregular estimates due to the fact that Z_t is iid by assumption. For this specification global decoding does not provide insightful results.

Figure 5 reports the decoded values of the parameters of the MCHMM with $M = 1$ and HMM. For example, we write $\widehat{\rho}_t = \sum_{k=1}^{D^\rho} \widehat{\rho}_k \mathbb{1}_{(\widehat{s}^\rho_t = k)}$, where $\widehat{\rho}_k$ is the ML estimate of ρ_k, and $\widehat{s}^\rho_t$ is the globally decoded value of S_t^ρ. The figure shows that recovered parameters values from the MCHMM specification are not characterized by large spikes like those of the standard HMM, and are thus much easier to interpret.

7. Conclusion

We introduced the MCHMM specification and discussed its ML estimation via an ECM algorithm with closed form conditional maximization steps. By exploiting the single chain representation of the model, we discussed identification of the model parameters as well as the consistency and asymptotic Normality of the ML estimator. An extensive simulation analysis showed that the ML estimator computed by the ECM algorithm has good finite sample properties. The empirical illustration indicated that the MCHMM provides more insightful results than the standard HMM for the analysis of US stocks and bond returns.

We also discussed a modelling strategy that allows for correlated dynamic parameters. The resulting specification is more parameterized and includes the previous one as a special case. For the series we considered in our empirical analysis, we found that introducing correlated dynamic parameters does not substantially modifies the conclusion of the analysis.

Extensions of the MCHMM specification to the fully multivariate case is of course very attractive. Two main issues are in place. First, while the problem of parameter proliferation is attenuated in the bivariate case, this is not true for the fully multivariate case, where the model soon becomes overparametrized even for medium cross sections. In this case, having one single chain for the entire correlation matrix is the most natural approach. Second, the derivation of the conditional maximization step for the correlation matrix of a multivariate MCHMM is prohibitive. Initial analysis should be devoted to the derivation of the ML estimator of the correlation matrix for the multivariate Gaussian distribution when means and variances are known, thus extending the result of Madansky (1965) to the multivariate case. An additional extension which is worth investigating is the inclusion of regressors in the mean parameters via a linear specification. In this case, the conditional maximization step would be easy to derive, and the resulting model might be compared with the MS-VAR specifications of Ang and Bekaert (2002b) and Guidolin and Timmermann (2006).

From an applied point of view, the bivariate MCHMM specification already offers a large number of attractive applications. First, an application of the regression lemma for the bivariate Normal distribution allows us to write $Y_{1,t}$ as a function of $Y_{2,t}$ as follows: $Y_{1,t} = \mu_{1,t} + \rho_t \frac{\sigma_{2,t}}{\sigma_{1,t}}(Y_{2,t} - \mu_{2,t}) + \sigma_{1,t}\sqrt{1-\rho_t^2}X_{1,t}$, where the value of the time varying parameters is determined by the realization of the associated Markov chain, and $X_{1,t}$ is independent from the Markov chains and $Y_{2,t}$. By writing $\beta_t = \rho_t \frac{\sigma_{2,t}}{\sigma_{1,t}}$ we have a time-varying capital asset pricing model (CAPM) which might be useful to understand which sources of time variation leads to changes in β_t, see for instance Ferson (1990), Evans (1994), Ghysels (1998), or more recently Gagliardini et al. (2016).

Another attractive area of application is the study of financial contagion. Specifically, Forbes and Rigobon (2002) define contagion as a significant increase in the cross-market linkages after a shock to one series of the system. In our context, the linkage is represented by the correlation between the series, and the shock is represented by a switch of the Markov chain associated to the conditional variance (or mean) of one of the series. Note that, this approach would not suffer from the volatility bias described in Forbes and Rigobon (2002).

Appendix A. An ECM for the model of Equation (5)

For the specification of Equation (5), we collect model parameters in $\boldsymbol{\theta} = \left[\boldsymbol{\omega}', \boldsymbol{\rho}_m', \mathrm{vec}\left(\boldsymbol{\Gamma}^{\rho}\right)', \boldsymbol{\delta}^{\rho\prime}, \boldsymbol{\mu}_{j,m}', \boldsymbol{\sigma}_{j,m}', \mathrm{vec}\left(\boldsymbol{\Gamma}^{\mu_j}\right)', \mathrm{vec}\left(\boldsymbol{\Gamma}^{\boldsymbol{\sigma}_j}\right)', \boldsymbol{\delta}^{\mu_j\prime}, \boldsymbol{\delta}^{\sigma_j\prime}, j = 1, 2, m = 1, \ldots, M\right]'$, where $\boldsymbol{\mu}_{j,m} = (\mu_{j,1,m}, \ldots, \mu_{j,D^{\mu_j},m})'$, $\boldsymbol{\sigma}_{j,m} = (\sigma_{j,1,m}, \ldots, \sigma_{j,D^{\sigma_j},m})'$, for $j = 1, 2$, and $\boldsymbol{\rho}_m = (\rho_{1,m}, \ldots, \rho_{D^{\rho},m})'$, be the vector of parameters for $m = 1, \ldots, M$. The CDLL for a vector of T observations is defined as $\mathcal{L}_c(\boldsymbol{\theta}) = \log P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T}, \mathbf{S}_{1:T}^{\rho} = \mathbf{s}_{1:T}^{\rho}, \mathbf{Z}_{1:T} = \mathbf{z}_{1:T}, \mathbf{S}_{1:T}^{\mu_j} = \mathbf{s}_{1:T}^{\mu_j}, \mathbf{S}_{1:T}^{\sigma_j} =$

$\mathbf{s}^{\sigma_j}_{1:T}, j = 1, 2; \boldsymbol{\theta})$ and can be written as:

$$
\begin{aligned}
\mathcal{L}_c(\boldsymbol{\theta}) \propto{}& \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\mu_j}} u^{\mu_j}_{i_j,1}\log\delta^{\mu_j}_{i_j} + \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\sigma_j}} u^{\sigma_j}_{i_j,1}\log\delta^{\sigma_j}_{i_j} + \sum_{k=1}^{D^{\rho}} u^{\rho}_{k,1}\log\delta^{\rho}_{k} + \sum_{t=1}^{T}\sum_{m=1}^{M} u^{\omega}_{m,t}\log\omega_m \\
&+ \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\mu_j}}\sum_{l_j=1}^{D^{\mu_j}}\sum_{t=2}^{T} v^{\mu_j}_{i_j,l_j,t}\log\gamma^{\mu_j}_{i_j,l_j} + \sum_{j=1}^{2}\sum_{i_j=1}^{D^{\sigma_j}}\sum_{l_j=1}^{D^{\sigma_j}}\sum_{t=2}^{T} v^{\sigma_j}_{i_j,l_j,t}\log\gamma^{\sigma_j}_{i_j,l_j} + \sum_{k=1}^{D^{\rho}}\sum_{q=1}^{D^{\rho}}\sum_{t=2}^{T} v^{\rho}_{k,q,t}\log\gamma^{\rho}_{k,q} \\
&+ \sum_{t=1}^{T}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{k=1}^{D^{\rho}}\sum_{m=1}^{M} u_{i_1,i_2,j_1,j_2,k,m,t}\Bigg\{-\log\sigma_{j_1,m} - \log\sigma_{j_2,m} - \frac{1}{2}\log(1-\rho^2_{k,m}) \\
&- \frac{1}{2(1-\rho^2_{k,m})}\left[\left(\frac{y_{1,t}-\mu_{1,i_1,m}}{\sigma_{j_1,m}}\right)^2 + \left(\frac{y_{2,t}-\mu_{2,i_2,m}}{\sigma_{j_2,m}}\right)^2 - 2\rho_{k,m}\left(\frac{y_{1,t}-\mu_{1,i_1,m}}{\sigma_{j_1,m}}\right)\left(\frac{y_{2,t}-\mu_{2,i_2,m}}{\sigma_{j_2,m}}\right)\right]\Bigg\}
\end{aligned}
\tag{A.1}
$$

where $u_{i_1,i_2,j_1,j_2,k,m,t} = \mathbb{1}_{(S^{\mu_1}_t=i_1, S^{\mu_2}_t=i_2, S^{\sigma_1}_t=j_1, S^{\sigma_2}_t=j_2, S^{\rho}_t=k, Z_t=m)}$, and $u^x_{i,1} = \mathbb{1}_{(S^x_1=i)}$, $u^{\omega}_{m,t} = \mathbb{1}_{(Z_t=m)}$ $v^x_{i,j,t} = \mathbb{1}_{(S^x_t=j, S^x_{t-1}=i)}$ for $x \in \mathcal{X}$.

Given the value of the parameters at iteration h, $\boldsymbol{\theta}^{(h)}$, the function $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(h)}) = E[\mathcal{L}_c(\boldsymbol{\theta})|\mathbf{Y}_{1:T} = \mathbf{y}_{1:T};\boldsymbol{\theta}^{(h)}]$ is computed by a run of the FFBS exploiting the single chain representation of the model.

For the CM steps we split $\boldsymbol{\theta}$ in 6 blocks $\boldsymbol{\theta} = (\boldsymbol{\theta}_i', i = 1,\ldots,6)'$, where: $\boldsymbol{\theta}_1 = \left[\boldsymbol{\omega}', \operatorname{vec}(\boldsymbol{\Gamma}^{\rho})', \boldsymbol{\delta}^{\rho\prime}, \operatorname{vec}(\boldsymbol{\Gamma}^{\mu_j})', \operatorname{vec}(\boldsymbol{\Gamma}^{\boldsymbol{\sigma}_j})', \boldsymbol{\delta}^{\mu_j\prime}, \boldsymbol{\delta}^{\sigma_j\prime}, j = 1, 2\right]'$, $\boldsymbol{\theta}_2 = \boldsymbol{\mu}_1$, $\boldsymbol{\theta}_3 = \boldsymbol{\mu}_2$, $\boldsymbol{\theta}_4 = \boldsymbol{\rho}$, $\boldsymbol{\theta}_5 = \boldsymbol{\sigma}_1$, and $\boldsymbol{\theta}_6 = \boldsymbol{\sigma}_2$.

In the first CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\theta}_1$. The solution for $\gamma^{x}_{i,j,t}{}^{(h+1)}$ and $\delta^{x\,(h+1)}_j$ for all $i, j = 1,\ldots,D^x$ and $x \in \mathcal{X}$ is analogous to that in Section 3, for $\omega^{(h+1)}_m$ we obtain $\omega^{(h+1)}_m = \frac{\sum_{t=1}^{T}\widehat{u}^{\omega}_{j,t}}{T}$. In the second CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\mu}_1$ and obtain:

$$
\mu^{(h+1)}_{1,j_1,m} = \frac{\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}}\widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}\left[\frac{y_{1,t}-\rho^{(h)}_{k,m}\frac{\sigma^{(h)}_{1,j_1,m}}{\sigma^{(h)}_{2,j_2,m}}\left(y_{2,t}-\mu^{(h)}_{2,i_2,m}\right)}{(\sigma^{(h)}_{1,j_1,m})^2\left[1-(\rho^{(h)}_{k,m})^2\right]}\right]}{\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}}\widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}\frac{1}{(\sigma^{(h)}_{1,j_1,m})^2\left[1-(\rho^{(h)}_{k,m})^2\right]}},
$$

for $j_1 = 1,\ldots,D^{\mu_1}$ and $m = 1,\ldots,M$. In the third CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\mu}_2$ and obtain:

$$\mu_{2,j_2,m}^{(h+1)} = \frac{\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t} \left[\frac{y_{2,t}-\rho_{k,m}^{(h)}\frac{\sigma_{2,j_2,m}^{(h)}}{\sigma_{1,j_1,m}^{(h)}}\left(y_{1,t}-\mu_{1,i_1,m}^{(h+1)}\right)}{(\sigma_{2,j_2,m}^{(h)})^2\left[1-(\rho_{k,m}^{(h)})^2\right]}\right]}{\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}\frac{1}{(\sigma_{2,j_2,m}^{(h)})^2\left[1-(\rho_{k,m}^{(h)})^2\right]}},$$

for $j_2 = 1, \dots, D^{\mu_2}$ and $m = 1, \dots, M$. In the fourth CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\rho}$. We define:

$$\eta_{k,m} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}, \qquad \xi_{k,m} = \eta_{k,m}^{-1}\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t} z_{1,i_1,j_1,m,t} z_{2,i_2,j_2,m,t}$$

$$\nu_{n,k,m} = \eta_{k,m}^{-1}\sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t} z_{n,i_n,j_n,m,t}^2, \text{ for } n = 1, 2, \qquad c_{k,m} = \frac{36\xi_{k,m} + 2\xi_{k,m}^3 - 9\xi_{k,m}(\nu_{1,k,m} + \nu_{2,k,m})}{2|3(\nu_{1,k,m} + \nu_{2,k,m} - 1) - \xi_{k,m}^2|^{3/2}}$$

where $z_{n,i_n,j_n,m,t} = (y_{n,t} - \mu_{n,i_n,m}^{(h+1)})/(\sigma_{n,j_n,m}^{(h)})$. Then, if $\eta_{k,m}^2 < 3(\nu_{1,k,m} + \nu_{2,k,m} - 1)$

$$\rho_{k,m}^{(h+1)} = \frac{2}{3}\sqrt{3(\nu_{1,k,m} + \nu_{2,k,m} - 1) - \eta_{k,m}^2}\,\sinh\left(\frac{1}{3}\sinh^{-1}(c_{k,m})\right) + \frac{1}{3}\xi_{k,m}$$

if $\xi_{k,m}^2 \geq 3(\nu_{1,k,m} + \nu_{2,k,m} - 1)$ we have:

$$\rho_{k,m}^{(h+1)} = \begin{cases} \frac{2}{3}\sqrt{\xi_{k,m}^2 - 3(\nu_{1,k,m} + \nu_{2,k,m} - 1)}\,\cosh\left(\frac{1}{3}\cosh^{-1}(c_{k,m})\right) + \frac{1}{3}\xi_{k,m}, & \text{if } c_{k,m} \geq 1 \\ \frac{2}{3}\sqrt{\xi_{k,m}^2 - 3(\nu_{1,k,m} + \xi_{2,k,m} - 1)}\,\cos\left(\frac{4\pi}{3} + \frac{1}{3}\cos^{-1}(c_{k,m})\right) + \frac{1}{3}\xi_{k,m}, & \text{if } c_{k,m} < 1. \end{cases}$$

In the fifth CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\sigma}_1$, subject to $\boldsymbol{\sigma}_1 > 0$. We first define

$$a_{1,j_1,m} = \sum_{t=1}^{T}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}, \quad b_{1,j_1,m} = \sum_{t=1}^{T}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}\frac{\rho_{k,m}^{(h+1)} z_{1,i_1,j_1,m,t} z_{2,i_2,j_2,m,t}}{\sigma_{2,j_2,m}^{(h)}\left(1 - (\rho_{k,m}^{(h+1)})^2\right)}$$

$$c_{1,j_1,m} = \sum_{t=1}^{T}\sum_{j_2=1}^{D^{\sigma_2}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,t} z_{1,i_1,j_1,m,t}^2\left(1 + \frac{(\rho_{k,m}^{(h+1)})^2}{1 - (\rho_{k,m}^{(h+1)})^2}\right)$$

then

$$\sigma_{1,j_1,m}^{(h+1)} = \frac{\sqrt{b_{1,j_1,m}^2 - 4a_{1,j_1,m}c_{1,j_1,m}} - b_{1,j_1,m}}{2a_{1,j_1,m}},$$

where $z_{n,i_n,j_n,m,t}$, $n = 1, 2$ are defined as in the CM step of $\boldsymbol{\rho}$. In the last CM step we maximize $\mathcal{Q}$ with respect to $\boldsymbol{\sigma}_2$ subject to $\boldsymbol{\sigma}_2 > 0$. To this end, we define $\widetilde{z}_{1,i_1,j_1,m,t} = (y_{1,t} - \mu^{(h+1)}_{1,i_1,m})/(\sigma^{(h+1)}_{1,j_1,m})$, and

$$a_{2,j_2,m} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t}, \quad b_{2,j_2,m} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t} \frac{\rho^{(h+1)}_{k,m}\widetilde{z}_{1,i_1,j_1,t} z_{2,i_2,j_2,m,t}}{\sigma^{(h+1)}_{1,j_1,m}\left(1 - (\rho^{(h+1)}_{k,m})^2\right)}$$

$$c_{2,j_2,m} = \sum_{t=1}^{T}\sum_{j_1=1}^{D^{\sigma_1}}\sum_{i_1=1}^{D^{\mu_1}}\sum_{i_2=1}^{D^{\mu_2}}\sum_{k=1}^{D^{\rho}} \widehat{u}_{i_1,i_2,j_1,j_2,k,m,t} z^2_{2,i_2,j_2,m,t}\left(1 + \frac{(\rho^{(h+1)}_{k,m})^2}{1 - (\rho^{(h+1)}_{k,m})^2}\right),$$

then

$$\sigma^{(h+1)}_{2,j_2,m} = \frac{\sqrt{b^2_{2,j_2,m} - 4a_{2,j_2,m}c_{2,j_2,m}} - b_{2,j_2,m}}{2a_{2,j_2,m}}.$$

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics*, 73(2):251–277.

Ang, A. and Bekaert, G. (2002a). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4):1137–1187.

Ang, A. and Bekaert, G. (2002b). Regime switches in interest rates. *Journal of Business & Economic Statistics*, 20(2):163–182.

Berchtold, A. (1999). The double chain markov model. *Communications in Statistics-Theory and Methods*, 28(11):2569–2589.

Bickel, P. J., Ritov, Y., Ryden, T., et al. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics*, 26(4):1614–1635.

Black, F. (1976). Studies of stock price volatility changes. In *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, pages 177–181.

Catania, L. (2020). A stochastic volatility model with a general leverage specification. *Journal of Business & Economic Statistics*, pages 1–31.

Colombi, R. and Giordano, S. (2015). Multiple hidden markov models for categorical time series. *Journal of Multivariate Analysis*, 140:19–30.

De Santis, G. and Gerard, B. (1997). International asset pricing and portfolio diversification with time-varying risk. *The Journal of Finance*, 52(5):1881–1912.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39(1):1–38.

Doornik, J. A. (2013). Markov-switching model with component structure for US GNP. *Economics Letters*, 118(2):265–268.

Erb, C. B., Harvey, C. R., and Viskanta, T. E. (1994). Forecasting international equity correlations. *Financial Analysts Journal*, 50(6):32–45.

Evans, M. D. (1994). Expected returns, time-varying risk, and risk premia. *The Journal of Finance*, 49(2):655–679.

Ferson, W. E. (1990). Are the latent variables in time-varying expected returns compensation for consumption risk? *The Journal of Finance*, 45(2):397–429.

Forbes, K. J. and Rigobon, R. (2002). No contagion, only interdependence: measuring stock market comovements. *The Journal of Finance*, 57(5):2223–2261.

Fosdick, B. K. and Raftery, A. E. (2012). Estimating the correlation in bivariate normal data with known variances and small sample sizes. *The American Statistician*, 66(1):34–41.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* Springer Science & Business Media.

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica*, 84(3):985–1046.

Gassiat, É., Cleynen, A., and Robin, S. (2016). Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26(1-2):61–71.

Ghysels, E. (1998). On stable factor structures in the pricing of risk: do time-varying betas help or hurt? *The Journal of Finance*, 53(2):549–573.

Guidolin, M. and Timmermann, A. (2006). An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *Journal of Applied Econometrics*, 21(1):1–22.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384.

Johnson, N. L. (1967). Bivariate samples with missing values. *Technometrics*, 9(4):679–682.

King, M., Sentana, E., and Wadhwani, S. (1994). Volatility and links between national stock markets. *Econometrica*, 62(4):901–33.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models. *Stochastic Processes and Their Applications*, 40(1):127–143.

Longin, F. and Solnik, B. (1995). Is the correlation in international equity returns constant: 1960–1990? *Journal of International Money and Finance*, 14(1):3–26.

Longin, F. and Solnik, B. (2001). Extreme correlation of international equity markets. *The Journal of Finance*, 56(2):649–676.

Madansky, A. (1965). On the maximum likelihood estimate of the correlation coefficient. Defense technical information center, RAND CORP, Santa Monica, California.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278.

Paolella, M. S., Polak, P., and Walker, P. S. (2019). Regime switching dynamic correlations for asymmetric and fat-tailed conditional returns. *Journal of Econometrics*, 213(2):493–515.

Pelletier, D. (2006). Regime switching for dynamic correlations. *Journal of Econometrics*, 131(1-2):445–473.

Phillips, K. L. (1991). A two-country model of stochastic output with changes in regime. *Journal of international economics*, 31(1-2):121–142.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Ravn, M. O. and Sola, M. (1995). Stylized facts and regime changes: Are prices procyclical? *Journal of Monetary Economics*, 36(3):497–526.

Rydén, T., Teräsvirta, T., and rAsbrink, S. (1998). Stylized facts of daily return series and the hidden markov model. *Journal of Applied Econometrics*, 13(3):217–244.

Sampson, A. R. (1978). Simple ban estimators of correlations for certain multivariate normal models with known variances. *Journal of the American Statistical Association*, 73(364):859–862.

Sims, C. A., Waggoner, D. F., and Zha, T. (2008). Methods for inference in large multiple-equation markov-switching models. *Journal of Econometrics*, 146(2):255–274.

Sims, C. A. and Zha, T. (2006). Were there regime switches in us monetary policy? *American Economic Review*, 96(1):54–81.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.