Shrinkage Estimation for Multivariate Hidden Markov Models

Mark Fiecas\textsuperscript{a}, Jürgen Franke\textsuperscript{b}, Rainer von Sachs\textsuperscript{c}, and Joseph Tadjuidje Kamgaing\textsuperscript{b}

\textsuperscript{a}Department of Statistics, University of Warwick, Coventry, United Kingdom; \textsuperscript{b}Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany; \textsuperscript{c}Institute of Statistics, Biostatistics, and Actuarial Sciences (ISBA) Université Catholique de Louvain, Louvain la Neuve, Belgium

\textbf{ABSTRACT}

Motivated from a changing market environment over time, we consider high-dimensional data such as financial returns, generated by a hidden Markov model that allows for switching between different regimes or states. To get more stable estimates of the covariance matrices of the different states, potentially driven by a number of observations that are small compared to the dimension, we modify the expectation–maximization (EM) algorithm so that it yields the shrinkage estimators for the covariance matrices. The final algorithm turns out to reproduce better estimates not only for the covariance matrices but also for the transition matrix. It results into a more stable and reliable filter that allows for reconstructing the values of the hidden Markov chain. In addition to a simulation study performed in this article, we also present a series of theoretical results that include dimensionality asymptotics and provide the motivation for certain techniques used in the algorithm. Supplementary materials for this article are available online.

\section{Introduction}

Hidden Markov models (HMM) are a popular class of models for time series that locally (within a state) behave like iid data, but repeatedly change the data-generating mechanism. The dependence in time is generated by a Markov process with usually a finite state space that controls switching between states, but is not observable directly. The structure of and inference for HMMs are well understood (see, e.g., Cappé, Moulines, and Rydén 2005), and there are tried and tested algorithms around like, for example, the expectation–maximization (EM) algorithm for calculating model parameter estimates, or the Viterbi algorithm for filtering that allows us to approximate the values of the hidden Markov process. However, these methods may cause problems in high dimensions that we address in this work.

HMMs have been applied in many fields. For illustration, we pick out one particular area of application in this article: modeling high-dimensional time series from finance with a view toward portfolio analysis. A simple but wide-spread model for the vector of returns of all assets in a stock portfolio is based on the assumption that the data are independent random vectors with covariance matrix $\Sigma$ and volatility matrix $\Sigma^{1/2}$, respectively. If the data are assumed to be Gaussian, this would correspond to a discrete-time version of the Black–Scholes model for stock prices. However, if the market environment changes, for example, if it moves to a more volatile state, the covariance matrix changes too. This behavior can be modeled by an HMM with a finite number, say $K$, of states represented by the different covariance matrices $\Sigma_k$, $k = 1, \ldots, K$. Additionally, the HMM model is particularly interesting for analyzing financial returns as it is able to describe certain particularities ("stylized facts," see, e.g., Rydén, Terasvirta, and Asbrink 1998) such as departures from the normality assumption and existence of dependence between the data, leading to volatility clusters. The dependence between the data is inherent to the Markovian structure of the hidden switching mechanism. Even if the observations behave locally (within a state) as iid Gaussian random vectors, the distribution is not normal, but a mixture of normals. In the following, we do not, however, assume normality of the conditional distributions of the returns given the hidden states to allow for more flexibility in modeling the tail distribution of the data.

A prime task in fitting HMMs to data is the estimation of model parameters like the covariance matrices $\Sigma_k$, $k = 1, \ldots, K$, their inverses, or the transition probability matrix of the hidden Markov chain. For high-dimensional data, which are common in portfolio management applications, this is no trivial task and requires some particular considerations. In this article, we illustrate with some applications to simulated and real data the instability of parameter estimates that work well in low-dimensional situations. For covariance matrix estimates, this does not come as a surprise, as it is well known that a large-scale sample covariance matrix is not stable and is even not guaranteed to be invertible if the dimensionality of the data is large with respect to the amount of data available (see, e.g., sec. 8.4 in Härdle and Simar 2007). This problem is even more pronounced in the context of switching regimes. Indeed, some of the states could be seldom visited, in which case the effective sample size for estimating the covariance matrix of that regime will be very small.

For HMMs, this instability of covariance estimates in particular influences the M-step of the EM-algorithm where Gaussian (pseudo-) likelihoods have to be calculated involving the
inverses. Indeed, if the estimate of the covariance matrix is numerically unstable, then a slight perturbation in the data will result in large changes in the inverse of this matrix (Fiecas et al. 2010). This has an unexpectedly dramatic effect on estimates of the transition matrix and on the reconstruction of the hidden process, which may break down more or less completely in high dimensions (compare Figure 2) if the standard estimation and filtering algorithms are applied in a straightforward manner.

We propose a remedy for that kind of problem by combining EM-type estimation of HMM with stable estimates of covariance or alternatively precision matrices in the M-step of the algorithm. In this article, we illustrate this idea with a particular approach based on shrinkage. Note that shrinkage for dependent data has been widely investigated, for example, in the frequency domain of time series, where the technology has been successfully applied to shrinkage estimation of the spectral density matrix (Boehm and von Sachs 2008; Fiecas and Ombao 2011). We extend the work of Sancetta (2008) to the situation of multivariate hidden Markov models. Sancetta (2008) introduced a shrinkage estimator for covariance matrices based on dependent data, which extended the popular shrinkage method of Ledoit and Wolf (2004) for iid multivariate data. This shrinkage estimator for large-scale covariance matrices guarantees an estimate that is invertible, positive-definite, more numerically stable, and has smaller mean squared error than the sample covariance matrix (Boehm and von Sachs 2008; Fiecas and Ombao 2011). We make the following assumptions, where in what follows we use $\| \cdot \|$ to denote the Euclidean norm of a $p$-dimensional vector:

1. $\epsilon_t$ is aperiodic, irreducible, and stationary,
2. $\epsilon_t$, $-\infty < t < \infty$, are independent of $S_t$, $-\infty < t < \infty$,
3. $\mathbb{E}\epsilon_t^4 - 3 \kappa_s$, $i = 1, \ldots, p$, for some constant $\kappa_s < \infty$, and $\mathbb{E}\epsilon_t^4, \kappa_s < \infty$, $i = 1, \ldots, p$.

Note that the first part of (A)3 requires uniformity with respect to $i$ even in case $p \rightarrow \infty$, and, therefore, does not follow from the second part. For discrete Markov chains with a finite number of states, in particular for our hidden process, aperiodicity and irreducibility already imply the existence of a unique stationary version $S_t$ with stationary distribution $\pi = (\pi_1, \ldots, \pi_K)$, where $\pi_k = P(S_t = \epsilon_k)$. Moreover, from Lemma 1 of Francq and Roussignol (1997), the stochastic process $S_t$ is also $\alpha$-mixing with exponentially decreasing rate. We have included stationarity in the assumptions above for convenience, but all results hold also for hidden processes starting in an arbitrary state, as they are always asymptotically stationary.

The observed process $X_t$ defined in Equation (1) inherits stationarity as well as $\alpha$-mixing with exponentially decreasing rate from the process $S_t$ which follows from (A)1 and (A)2 as in Lemma 1 of Francq and Roussignol (1997). Both properties are needed as key ingredients to derive the consistency and asymptotic normality of the maximum likelihood estimates by a straightforward adaptation of standard arguments given, for example, by Douc, Moulines, and Rydén (2004), compare also Tadjuidje Kamgaing (2013). Additionally, the existence of the $m$th moment of $\epsilon_t$ implies the existence of the $m$th moment of $X_t$.

The data-generating mechanism defined in (1) can be considered as multivariate generalization of the white noise driven by hidden Markov chains as introduced by Francq and Roussignol (1997). It can further be regarded as a multivariate version of the CHARMe model introduced by Stockis, Franke, and Tadjuidje Kamgaing (2010), with constant state means and covariances. However, it differs from the approach of Yang (2000), that focused on (trend) VAR hidden Markov models as opposed to the hidden Markov volatility model we are considering. Nevertheless, we point out that Yang (2000) and Francq and Zarkoian (2001) gave conditions for the existence of a stationary solution as well as for geometric ergodicity in these more complex models. Additionally, Francq and Roussignol (1997) investigated the asymptotic behavior of maximum likelihood estimators of the model parameters. However, in high-dimensional situations, the nice asymptotic properties of the estimates do not

section 2. Hidden Markov Models and the EM Algorithm

In this section, we first introduce the model for the data including the basic assumptions needed. Then, we give a brief review of the common EM algorithm for numerical approximation of the maximum Gaussian pseudo likelihood estimates, which also serves to introduce notation. In Section 3, we discuss the relationship between a penalized likelihood approach and shrinkage. Then, we generalize the results of Sancetta (2008) to HMMs, and we describe how to modify the EM algorithm to get more stable estimates of covariance matrices, transition probabilities, and other characteristics of the model. A simulation study is carried out in Section 4, completed by a numerical (portfolio) analysis of a real dataset in Section 5. All of the proofs are in the supplementary material for this article.

2.1. The Model

Let $S_t$ be a finite-state Markov chain assuming $K$ different values. To simplify notation later on, we choose as the set of states $\{\epsilon_1, \ldots, \epsilon_K\}$, where $\epsilon_i$ is the unit vector in $\mathbb{R}^K$ having the $i$th entry equal to 1. For $i, j = 1, \ldots, K$, we denote the transition probabilities $P(S_t = \epsilon_j | S_{t-1} = \epsilon_i, S_{t-2} = \ldots) = P(S_t = e_j | S_{t-1} = e_i) = a_{ij}$. Then, our model for the data-generating process is the following:

$$X_t = \sum_{k=1}^{K} S_{t,k} \left( \mu_k + \Sigma_k^{1/2} \epsilon_t \right),$$

where the $\epsilon_t$ are iid with mean 0 and unit covariance matrix $I_p$, and $S_{t,k} = 1$ iff $S_t = \epsilon_k$, and $= 0$, else. At each time instant $t$, exactly one of the components of $S_t$ is equal to 1 and the other components are 0. We observe $X_1, \ldots, X_T$, but not the hidden state process $S_1, \ldots, S_T$. We make the following assumptions, where in what follows we use $\| \cdot \|$ to denote the Euclidean norm of a $p$-dimensional vector:

(A) 1. $S_t$ is aperiodic, irreducible, and stationary,
2. $\epsilon_t$, $-\infty < t < \infty$, are independent of $S_t$, $-\infty < t < \infty$,
3. $\mathbb{E}\epsilon_t^4 - 3 \kappa_s$, $i = 1, \ldots, p$, for some constant $\kappa_s < \infty$, and $\mathbb{E}\epsilon_t^4, \kappa_s < \infty$, $i = 1, \ldots, p$.

Note that the first part of (A)3 requires uniformity with respect to $i$ even in case $p \rightarrow \infty$, and, therefore, does not follow from the second part. For discrete Markov chains with a finite number of states, in particular for our hidden process, aperiodicity and irreducibility already imply the existence of a unique stationary version $S_t$ with stationary distribution $\pi = (\pi_1, \ldots, \pi_K)$, where $\pi_k = P(S_t = \epsilon_k)$. Moreover, from Lemma 1 of Francq and Roussignol (1997), the stochastic process $S_t$ is also $\alpha$-mixing with exponentially decreasing rate. We have included stationarity in the assumptions above for convenience, but all results hold also for hidden processes starting in an arbitrary state, as they are always asymptotically stationary.

The observed process $X_t$ defined in Equation (1) inherits stationarity as well as $\alpha$-mixing with exponentially decreasing rate from the process $S_t$, which follows from (A)1 and (A)2 as in Lemma 1 of Francq and Roussignol (1997). Both properties are needed as key ingredients to derive the consistency and asymptotic normality of the maximum likelihood estimates by a straightforward adaptation of standard arguments given, for example, by Douc, Moulines, and Rydén (2004), compare also Tadjuidje Kamgaing (2013). Additionally, the existence of the $m$th moment of $\epsilon_t$ implies the existence of the $m$th moment of $X_t$.
help for realistic sample sizes as it is not guaranteed that straightforward implementations lead to reasonable results.

We propose to apply model (1) and the following estimation algorithms to high-dimensional time series data where one possible area of application is portfolio management in finance. On the first glance, the assumption of conditional independence of the observations given the states, which follows from the independence of the \( \epsilon_t \) and (A)2, seems to be somewhat restrictive. However, as mentioned above, Rydén, Teräsvirta, and Asbrink (1998) had shown that the dependence structure of the data induced by model (1) is already rich enough to cover most of the well-known stylized facts of financial time series.

### 2.2. The EM Algorithm

To estimate the parameters of model (1), we adopt a Gaussian pseudo maximum likelihood approach. Let \( \theta \) denote the vector of all parameters determining means and covariance matrices of the \( \mu_k, \Sigma_k, k = 1, \ldots, K \), of the different regimes as well as the transition probabilities of the Markov chain. Let

\[
\hat{f}_k(x) \equiv f(x \mid \epsilon, \theta) = f(x, \mu_k, \Sigma_k)
\]

\[
= \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) \right)
\]

(2)

\[
denote the conditional probability density of \( X_t \) given \( S_t = e_k \) if it would be Gaussian. Correspondingly, let

\[
L(X_1, \ldots, X_T \mid S, \theta) = \prod_{t=1}^{T} f(X_t \mid S_t, \theta)
\]

\[
= \prod_{t=1}^{T} \left( \sum_{k=1}^{K} \frac{1}{\sum_{k=1}^{K} S_{t,k}} f(X_t, \mu_k, \Sigma_k) \right)
\]

denote the conditional likelihood given the whole hidden sample path \( S = (S_1, \ldots, S_T) \), and let \( L(S \mid \theta) \) denote the probability of observing this particular path. Then, by the law of total probability,

\[
L(X_1, \ldots, X_T \mid \theta) = \sum_{\text{all possible } S} L(X_1, \ldots, X_T \mid S, \theta) L(S \mid \theta)
\]

(3)

is the Gaussian pseudo likelihood that we have to maximize.

In the classical context of HMMs, consistency and asymptotic normality of maximum likelihood estimates can be derived by a straightforward adaptation of standard arguments given, for example, by Douc, Moulines, and Rydén (2004) and Tadjudjeid Kamgaing (2013), as the strict stationarity of the process \( X_t \), which is the main ingredient of those arguments, is here trivially guaranteed by that of the hidden process \( S_t \) and the assumptions on the \( \epsilon_t \). In this article, however, we are interested in the \( p \sim T^p \) context, where we have to deal with finite sample problems caused by the instability of high-dimensional sample covariance matrices. These problems vanish for \( T \to \infty \) and fixed \( p \) and are not covered by standard asymptotic theory that we do not discuss here. In the next sections, we develop an estimation algorithm that is able to cope with those problems for large \( p \) and relatively small \( T \).

The direct numerical maximization of the pseudo maximum likelihood in Equation (3) would be quite cumbersome. We instead follow the popular approach of an expectation–maximization (EM) algorithm where the hidden process \( S_t \) is interpreted as missing data. The EM algorithm can be traced back to Hartley (1958) and has been presented by Baum et al. (1970) in a general framework. Finally, it was made popular through the seminal work of Dempster, Laird, and Rubin (1977), and since then various variants of this algorithm have been proposed in the literature. The motivation for the EM algorithm is the observation that the implementation of maximum likelihood becomes much easier if an oracle uncovers the hidden variables \( S_1, \ldots, S_T \). In that case, we could maximize the complete likelihood \( L_c \), that is, the likelihood given the observations \( X_1, \ldots, X_T \) and the hidden data \( S_1, \ldots, S_T \). \( L_c \) frequently has a much simpler structure than the incomplete likelihood based on the observations only, allowing for fast numerical calculations or even explicit analytical formulas for the maximum likelihood estimates. For model (1), the complete likelihood is given by

\[
L_c(X, S \mid \theta) = L(X_1, \ldots, X_T \mid S, \theta) L(S \mid \theta),
\]

and the corresponding log-likelihood \( \ell_c = \log L_c \) is

\[
\ell_c(X, S \mid \theta) = \sum_{k=1}^{K} \sum_{t=1}^{T} S_{t,k} \log f(X_t, \mu_k, \Sigma_k) + \log L(S \mid \theta).
\]

(4)

Note that \( \ell_c \) decomposes into \( K + 1 \) different terms depending only on \( \mu_k, \Sigma_k, k = 1, \ldots, K \), respectively, on the transition probabilities of the hidden Markov chain such that they may be maximized separately. For the means and covariance matrices, we get as explicit oracle estimates the sample mean and the covariance matrix of all observations generated from the \( k \)th regime, that is,

\[
\hat{\mu}_k = \frac{1}{\sum_{t=1}^{T} S_{t,k}} \sum_{t=1}^{T} S_{t,k} X_t,
\]

(5)

and

\[
\hat{\Sigma}_k = \frac{1}{\sum_{t=1}^{T} S_{t,k}} \sum_{t=1}^{T} S_{t,k} (X_t - \mu_k)(X_t - \mu_k)' \]

(6)

for \( \sum_{t=1}^{T} S_{t,k} \neq 0 \). Otherwise, we set \( \mu_k = 0 \) and \( \Sigma_k = 0 \) for convenience. This happens with exponentially decreasing probability and will be asymptotically negligible. In practice, however, we may run into numerical problems if we have states that are rarely visited such that the effective sample size \( \sum_{t=1}^{T} S_{t,k} \) is small, though not 0. Therefore, in the following algorithms we prefer to use the biased estimator of the variance-covariance matrix

\[
\hat{\Sigma}_k = \frac{1}{\sum_{t=1}^{T} S_{t,k}} \sum_{t=1}^{T} S_{t,k} (X_t - \mu_k)(X_t - \mu_k)' = \pi_k \hat{\Sigma}_k
\]

(7)

where \( \pi_k = T^{-1} \sum_{t=1}^{T} S_{t,k} \) is the oracle estimate of \( \pi_k = P(S_t = e_k) \). Here and in the following, an upper index \( ^* \) always designates oracle estimates where we assume the state variables \( S_1, \ldots, S_T \) to be known.
In the iterations of the EM algorithm, the complete Gaussian pseudo log-likelihood as a function of the parameter of interest $\theta$ is replaced by its conditional expectation given only the observations $(X_1, \ldots, X_T)$ where the conditional expectation is calculated with respect to a dummy parameter value $\theta'$ (see, e.g., Baum et al. 1970 for more details on this issue). Neglecting constants that do not depend on the model parameters, we get

$$Q(\theta; \theta') = \mathbb{E}_{\theta'} \left( \ell_c(X, S | \theta) \mid X_1, \ldots, X_T \right)$$

$$= \sum_t \mathbb{E}_{\theta'} \left( \sum_{i,j=1}^{K} S_{t,j} \log a_{ij} \mid X_1, \ldots, X_T \right) + \mathbb{E}_{\theta'} \left( \sum_{k=1}^{K} S_{1,k} \log P(S_{1,k} = 1) \mid X_1, \ldots, X_T \right) - \frac{1}{2} \sum_t \mathbb{E}_{\theta'} \left( \sum_{k=1}^{K} S_{t,k} \log |\Sigma_k| + (X_t - \mu_k)' \right. \Sigma_k^{-1}(X_t - \mu_k) \left. \right) \mid X_1, \ldots, X_T \right).$$

The basic idea of the algorithm is now the following: In the E-step we calculate $Q(\theta; \theta')$ as a function of $\theta$, pretending that the model parameter $\theta'$, which is used for calculating the conditional expectation, is known and coincides with the parameter estimate $\hat{\theta}(n)$ from the previous iteration step. In the course of this step, we derive estimates $\hat{S}_t$ of the hidden state variables, where we make use of the forward and backward variables (Rabiner 1989).

In the M-step, we replace the hidden states in the definition of $Q(\theta; \theta')$, again with $\theta' = \hat{\theta}(n)$, by their estimates from the E-step. Then, we maximize with respect to $\theta$ to get an update $\hat{\theta}(n+1)$ of the model parameter estimate. The outline of the algorithm is as follows.

**Algorithm 1. EM Algorithm**

1. Choose an adequate starting value $\hat{\theta}(1)$ of the parameter
2. E-Step: at the $n$th step, compute $Q(\theta, \hat{\theta}(n))$ as a function of $\theta$. As part of this calculation, approximations $\hat{S}_{t,k} = P(S_{t,k} = 1 | X_1, \ldots, X_T, \hat{\theta}(n))$ are derived.
3. M-Step: Determine $\hat{\theta}(n+1)$ as the maximizer of $Q(\theta, \hat{\theta}(n))$ over $\theta$.
4. Iterate the E-step and M-step until a stopping criterion is satisfied.

The details of the E- and M-step are discussed in the following sections.

**2.2.1. E-Step**

The first part of the EM algorithm provides approximations $\hat{S}_t$ of the hidden state variables $S_t$. The procedure is quite standard (see, e.g., Rabiner 1989) and given here only for sake of completeness. For readability, we write $\hat{\theta}$ instead of the actually used parameter value $\hat{\theta}(n)$ of the previous $n$th iteration. Recall that $\theta$ contains all the information about the means $\mu_k$ and covariance matrices $\Sigma_k$ of the conditional distributions of $X_t$ given $S_{t,k} = 1, k = 1, \ldots, K,$ as well as the transition probabilities $a_{ij}$ and the corresponding stationary distribution $\pi_1, \ldots, \pi_K$ of the state variables. Moreover, $f_k(x) = f(x, \mu_k, \Sigma_k)$ given by (2) denotes the density of $X$ in state $k$, where the parameters $\mu_k, \Sigma_k$ of those Gaussian densities are part of $\theta$. We introduce the forward-backward variables as auxiliary quantities that may be calculated from $\theta$ and the data using the following recursions.

**Forward variables:** For $i = 1, \ldots, K$, we define $\alpha_i(t) = P(X_1, \ldots, X_t, S_t = i | \theta)$, for which we have the following relations: (1) $\alpha_i(1) = \pi_i f_i(X_1)$, (2) $\alpha_i(t+1) = \left( \sum_{k=1}^{K} \alpha_k(t) \nu_{ki} \right) f_i(X_{t+1})$, and (3) $L_i(X_1, \ldots, X_T | \theta) = \sum_{i=1}^{K} \alpha_i(T)$.

**Backward variables:** For $i = 1, \ldots, K$, we define $\beta_i(t) = P(X_{t+1}, \ldots, X_T | S_t = i = 1, \theta)$, for which we have the following: (1) $\beta_i(T) = 1$ and (2) $\beta_i(t) = \sum_{k=1}^{K} a_{ik} \nu_{kj} \beta_k(t+1)$, $T - 1 \geq t \geq 1$.

**State variables:** From the forward and backward variables given above, we calculate $\gamma_i(t) = \alpha_i(t) \beta_i(t) / \sum_{k=1}^{K} \alpha_k(t) \beta_k(t)$ and $\xi_{ij}(t) = \gamma_i(t) \nu_{ij} f_j(X_{t+1}) \beta_k(t+1) / \beta_i(t)$.

Note that $\xi_{ij}(t)$ provide approximations of the bivariate conditional distribution of adjacent state variables given the data and parameter $\theta$: $\xi_{ij}(t) = P(S_t = i, S_{t+1} = j | X_1, \ldots, X_T, \theta)$. In particular, we have $\sum_t \gamma_i(t) = \sum_t \mathbb{E}_{\theta}(S_{t,i} | X_1, \ldots, X_T)$ and $\sum_t \xi_{ij}(t) = \sum_t \mathbb{E}_{\theta}(S_{t,i} S_{t+1,j} | X_1, \ldots, X_T)$.

**2.2.2. M-Step**

In the M-step, we calculate an update $\hat{\theta}(n+1)$ of the parameter value $\hat{\theta}(n)$ used in the E-step. Using the state variables from the E-step, we first get updated estimates of the transition probabilities

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} = \frac{\sum_{t=1}^{T-1} \hat{S}_{t,i} \hat{S}_{t+1,j}}{\sum_{i=1}^{T} \hat{S}_{t,i}}, \quad i, j = 1, \ldots, K,$

and correspondingly, of the stationary state probabilities $\hat{\pi}_k = \frac{1}{T} \sum_{t=1}^{T} S_{t,k}$, for $k = 1, \ldots, K$. The state means are updated as

$$\hat{\mu}_k = \frac{\sum_{t=1}^{T} \hat{S}_{t,k} X_t}{\sum_{t=1}^{T} \hat{S}_{t,k}}, \quad k = 1, \ldots, K,$$

and the estimates of the covariance matrices are updated as $\hat{\Sigma}_k = \hat{\pi}_k^{-1} \hat{\Sigma}_k$, and

$$\hat{\Sigma}_k = \frac{1}{T} \sum_{t=1}^{T} \hat{S}_{t,k}(X_t - \hat{\mu}_k)(X_t - \hat{\mu}_k)'$$

Finally, for the hidden path reconstructions like those shown in Figure 2 we make use of the Viterbi algorithm, described in Appendix A, to compute the optimal state sequence in a hidden Markov model from a sequence of observed outputs. It is based on the maximization of the single best state sequence and uses the dynamic programming methodology, see, for example, Rabiner (1989) for more details.
3. A Modified EM Algorithm Using Shrinkage

In this section, we propose a modification of the EM algorithm where we replace the likelihood function with a penalized likelihood. With a properly chosen penalty term, in the M-step we will obtain well-conditioned estimates of the covariance matrix. In particular, the resulting estimators will have the form of the linear shrinkage estimator by Ledoit and Wolf (2004) and Sancetta (2008), which are proven and tested estimators that are capable of handling the instability of large covariance matrices. Indeed, these maximum penalized likelihood estimates will be obtained using a data-adaptive regularization parameter, as pointed out by Yuan and Huang (2009) in the iid case. We first review this kind of relationship between shrinkage and penalized likelihood as the basis for further discussion. In the following, we use as a matrix norm $||C||$ the (scaled) Frobenius norm: $||C||^2 = p^{-1} \text{tr}(CC')$.

3.1. Penalized Maximum Likelihood and Shrinkage

For the moment, let us consider the case $K = 1$ in model (1), that is, the observed data $X_1, \ldots, X_T$ are iid, and there are no hidden state variables. Then $\theta$ just parameterizes the common mean $\mu$ and covariance matrix $\Sigma$ of the data, and the complete and incomplete log likelihoods both coincide with $\ell(X \mid \theta) = \sum_{t=1}^T \log f(X_t, \mu, \Sigma)$. Let $\theta_0$, $\mu_0$, $\Sigma_0$ denote the true unknown parameter values. Instead of the log-likelihood we consider the penalized log-likelihood

$$\ell_p(X \mid \theta) = \frac{2}{T} \ell(X \mid \theta) - p \alpha(\theta)$$

with $\alpha(\theta) = \log ||\Sigma|| + \alpha \text{tr}(\Sigma^{-1})$, where $\alpha = \alpha(\theta_0)$ is up to a factor the expectation of the trace of the sample covariance matrix $\hat{\Sigma}$. Thus $\alpha = p^{-1} \text{tr}(\Sigma_0) = p^{-1} \text{tr}(\Sigma)$. We have chosen the penalty term $\ell_p(X \mid \theta_0)$ according to Yuan and Huang (2009) with the special choice $\alpha = \alpha I_p$, that is, the target matrix is a multiple of the $p \times p$ identity matrix. Up to a constant independent of $\Sigma$, $\ell_p(X \mid \theta_0)$ can be interpreted as the entropy loss between $\Sigma$ and $\alpha I_p$.

Note that, in the case of no shrinkage, that is, $W = 0$, we have $\Sigma' = \hat{\Sigma}$, so that the shrinkage estimate with weight $W_0$ is at least as good as the sample covariance matrix. The proportionality factor $\alpha$ is chosen such that the expectation of the trace of $\Sigma'$ for all $W$ coincides with the expectation of the trace of $\hat{\Sigma}$, that is, the shrunken matrix is in a certain sense of the same size as the sample covariance matrix, but more regular. In particular, the condition number of the regularized estimator, which is the ratio of the largest to the smallest eigenvalue, is closer to one as a result of a reduction in the dispersion of the eigenvalues of the estimator.

3.2. Shrinkage With Several Regimes

We now return to our hidden Markov model given by Equation (1). Here, the complete log-likelihood $\ell_c$ given by Equation (4) is an important feature of the EM algorithm described in Section 2.2. It is based on the assumption that an oracle provides the hidden variables $S_1, \ldots, S_T$. To get more stable estimates of the covariance matrix, we modify the log-likelihood by adding regularization terms in the spirit of the Section 3.1 that enforce a better condition of the covariance estimates. As $\ell_c$ decomposes into $K + 1$ summands, we may add separate penalties for each of the $K$ covariance matrix estimates, that is, we consider the analog to Equation (11) the penalized log-likelihood

$$p_c(X, S \mid \theta) = \frac{2}{T} \ell_c(X, S \mid \theta) - \sum_{k=1}^K \lambda_k J_k(\theta \mid \theta_0)$$

$$= \sum_{k=1}^K \left( \frac{2}{T} \sum_{t=1}^T S_{t,k} \log f(X_t, \mu_k, \Sigma_k) - \lambda_k J_k(\theta \mid \theta_0) \right)$$

$$+ \text{log} L(S \mid \theta),$$

where $J_k(\theta \mid \theta_0) = \log ||\Sigma_k|| + \tilde{\alpha}_k \text{tr}(\Sigma_k^{-1})$ with $\tilde{\alpha}_k = p^{-1} \text{tr}(\Sigma_{0,k})$. Here, $\theta_0$ again denotes the true underlying parameter, and $\Sigma_{0,k}$, $k = 1, \ldots, K$, are the corresponding covariance matrices. As in the case of one regime ($K = 1$) and iid data, we get from Yuan and Huang’s arguments (Yuan and Huang 2009) that, analogous to (12), the estimates of the covariance matrices maximizing $p_c$ are shrunken versions of the sample covariance matrices $\Sigma_k^o$ based on the data from the $K$ individual regimes introduced in Equation (6), that is, $\hat{\Sigma}_k^c = (1 - W_k)\hat{\Sigma}_k^o + W_k \tilde{\alpha}_k I_p$ with $W_k = \lambda_k/(1 + \lambda_k) \in [0, 1]$, $k = 1, \ldots, K$.

As discussed in Section 2.2, we prefer to work with $\Sigma_k^o = \pi_k^o \Sigma_k^o$ instead of $\Sigma_k^o$, compare Equation (7). For the corresponding shrinkage estimator, we have to modify the scaling factor $\tilde{\alpha}_k$ accordingly, and, using the optimal weights $W_{0,k}$ in the sense of Ledoit and Wolf (2004), we get as shrinkage estimates $\Sigma_k^c = (1 - W_{0,k})\Sigma_k^o + W_{0,k} \tilde{\alpha}_k I_p$ with $\tilde{\alpha}_k = p^{-1} \text{tr}(\pi_k^o \Sigma_{0,k})$, $k = 1, \ldots, K$. As the optimal weights $W_{0,k}$ are not known, we have to estimate them from the data. In the rest of this section, we show that and how the estimation method of Sancetta (2008), which generalizes the approach of Ledoit and Wolf (2004) to dependent data, can be applied to hidden Markov models. Note that we are considering a complete penalized likelihood for later use in the EM-algorithm, that is, we are looking at the situation where an oracle uncovers the hidden variables.
First, we discuss how to replace the scaling factor $\alpha_k$, depending on the underlying distribution, by an estimate. A suitable choice is $\alpha_k^o = p^{-1} \text{tr} \Sigma_k^o$, as the following lemma shows.

**Lemma 1.** Let the data be generated from model (1), either for fixed $p$ or for with $T$ increasing $p \to \infty$, satisfying assumption (A). Consider the assumptions (B1) $\max_{l \leq p} |\mu_{0,k,l}|$ as well as $|\Sigma_0^1|^2 = p^{-1} \text{tr} \Sigma_0^1$ are uniformly bounded in $p$, and (B2) $\pi_k > 0$, $\text{tr}(\Sigma_0^1) \geq cp^r$ for some $c > 0$, $0 \leq r \leq 1$ and all $p$. Then it follows for some $0 < \beta < 1$ and all $k = 1, \ldots, K$,

(a) \[ \mathbb{E}(\alpha_k^o - \alpha_k)^2 = O(T^{-1}) \]
(b) If (B1) holds, then the normalized trace estimator $\alpha_k^o$ is mean-square consistent, that is, $\mathbb{E}(\alpha_k^o - \alpha_k)^2 = O(T^{-1})$ and $\mathbb{E}(\alpha_k^o - \alpha_k) = O(T^{-1})$, where $\alpha_k^o = p^{-1} \text{tr} \Sigma_k^o$, $\alpha_k = p^{-1} \text{tr}(\pi_k \Sigma_0^1)$;
(c) If (B1) and (B2) hold, then
\[ \frac{\text{var}(\text{tr}(\Sigma_k^o))}{\text{var}(\text{tr}(\pi_k \Sigma_0^1))^2} = O\left( \frac{p^{2(1-r)}}{T} \right), \]
that is, the relative error of $\text{tr}(\Sigma_k^o)$ as an estimate of $\text{tr}(\pi_k \Sigma_0^1)$ converges to 0 if $p^{1-r} = o(\sqrt{T})$.

Note that the assumptions (B1) and (B2) are automatically satisfied for fixed $p$ and nonsingular $\Sigma_0,k$. In general, the second part of (B1) is satisfied if the largest eigenvalue $\lambda_{\text{max}}(\Sigma_0,k)$ is bounded uniformly in $p$, though this is not a necessary condition. (B1) includes also, for the situation where $\lambda_{\text{max}}(\Sigma_0,k)$ is growing with a rate up to $p$, as long as all except finitely many of the eigenvalues are uniformly bounded in $p$. As $\text{tr}(\Sigma_0,k) \geq \rho_{\text{min}}(\Sigma_0,k)$, the second part of assumption (B2) is automatically satisfied if the smallest eigenvalue satisfies $\lambda_{\text{min}}(\Sigma_0,k) \geq c/p^{1-r}$, that is, it does not converge to 0 too fast with increasing dimension $p$. However, (B2) is weaker and allows even for singular $\Sigma_0,k$.

The optimal shrinkage weight $W_0,k$ minimizes the mean squared error
\[ W_{0,k} = \arg \min_{W_{0,k} \in [0,1]} \mathbb{E}[(1 - W_{0,k}) \Sigma_0^o + W_{0,k} \alpha_k I_p - \pi_k \Sigma_0^1]^2. \]
Assuming the existence of all necessary moments, we get an explicit formula for $W_{0,k}$ directly from Proposition 1 of Sancetta (2008).

**Lemma 2.** Under the model (1), for $k = 1, \ldots, K$,
\[ W_{0,k} = \frac{\mathbb{E}[\Sigma_k^o - \pi_k \Sigma_0^1]^2}{\mathbb{E}[\alpha_k I_p - \pi_k \Sigma_0^1]^2} \wedge 1. \]

The optimal shrinkage estimate $\Sigma_k^o$, that uses the optimal weights of Lemma 2, is however, not a feasible estimate because $\alpha_k$ is not known. Referring to Lemma 1 we replace it in the denominator by its oracle estimate $\alpha_k^o$. This choice for an estimator of the denominator of $W_{0,k}$ has been discussed by Ledoit and Wolf (2004) in detail; it leads to the sample estimator of the denominator where the expectation (risk) is just replaced by the sample analog (loss), see Equation (16) further below.

To get a direct estimate of the numerator in Equation (14), we follow Sancetta’s approach. From Lemma 5 in the supplementary material on “Proofs and Technical Lemmas,” we have
\[ \mathbb{E}[(\Sigma_k^o - \pi_k \Sigma_0^1)^2] = \frac{1}{p} \sum_{i,j=1}^p \text{var} \left( \frac{1}{T} \sum_{t=1}^T S_{i,k}(X_{ti} - \mu_{0,k,i})(X_{tj} - \mu_{0,k,j}) \right) + O\left( \frac{p}{T^2} \right), \]
where, similar to the arguments in Proposition 7.3.3 and 7.3.4. of Brockwell and Davis (1991), replacing the estimates $\mu_k^o$ in the definition of $\Sigma_k^o$ by their true values $\mu_{0,k}$, has an asymptotically negligible effect. Note that the first term on the right-hand side is of order $O(p/T)$ by the discussion preceding Lemma 5.

Setting $y_{t,i,j} = S_{i,k}(X_{ti} - \mu_{0,k,i})(X_{tj} - \mu_{0,k,j})$, the above expression is given by
\[ \frac{1}{p} \sum_{i,j=1}^p \frac{1}{T} \text{var} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T y_{t,i,j} \right), \]
where asymptotically, for any fixed $k$, $\text{var}(T^{-1/2} \sum_{t=1}^T y_{t,i,j})$ is well-known as the long run variance, see, for example, Hall (2005), of the time series $y_{t,i,j}$. We note that this time series is stationary with, by Lemma 3 in the supplementary material, exponentially decaying autocovariances $c_{ij,k}(t)$, which implies the existence of a continuous spectral density $f_{ij,k}(\omega)$. Let the corresponding sample autocovariances be
\[ c_{ij,k}(s) = \frac{1}{T} \sum_{t=1}^{T-s} \left( y_{t,i,j} - \frac{1}{T} \sum_{t=1}^T y_{t,i,j} \right) \left( y_{t+s,i,j} - \frac{1}{T} \sum_{t=1}^T y_{t,i,j} \right) \]
for $0 \leq s \leq T - 1$, $c_{ij,k}(-s) = c_{ij,k}(s)$.

From Lemma 3, the variance of the sample mean $T^{-1} \sum_{t=1}^T y_{t,i,j}$ is asymptotically $T^{-1} f_{ij,k}(0)$. Following Sancetta (2008), we approximate the spectral densities $f_{ij,k}(0) = \sum_{t=\infty}^{-\infty} c_{ij,k}(t)$ at the frequency 0 by applying a popular nonparametric estimator, namely, via windowing in the time domain (see, e.g., Andrews 1991). For some kernel $K(u) \geq 0$, $K(u) = K(-u)$, and $K(0) = 1$ and bandwidth $b > 0$, $\hat{f}_{ij,k}(0) = \sum_{s=-T+1}^{T-1} K(s/b) c_{ij,k}(s)$, so that we estimate Equation (15) with $(pT)^{-1} \sum_{i,j=1}^p \hat{f}_{ij,k}(0)$.

By the same kind of arguments used in showing Lemma 5, replacing the unknown means $\mu_k$ has an asymptotically negligible effect such that we may instead consider the following feasible estimates:
\[ \hat{y}_{t,i,j} = S_{i,k}(X_{ti} - \mu_{k,i}^o)(X_{tj} - \mu_{k,j}^o), \]
\[ \hat{c}_{ij,k}(s) = \frac{1}{T} \sum_{t=1}^{T-s} \left( \hat{y}_{t,i,j} - \frac{1}{T} \sum_{t=1}^T \hat{y}_{t,i,j} \right) \left( \hat{y}_{t+s,i,j} - \frac{1}{T} \sum_{t=1}^T \hat{y}_{t,i,j} \right) \]
and
\[ \hat{f}_{ij,k}(0) = \sum_{s=-T+1}^{T-1} K \left( \frac{s}{b} \right) \hat{c}_{ij,k}(s). \]
We finally get as an oracle estimator of the optimal shrinkage weights $W_{0:k}$:

$$W_{0:k}^o = \frac{(pT)^{-1} \sum_{i,j=1}^{p} \hat{b}_{k,i,j}(0)}{\left\| \hat{\Sigma}_k - o_k^a I_p \right\|^2} \land 1,$$  \hspace{1cm} (16)

and the optimal shrinkage estimator is given by

$$\hat{\Sigma}_k = (1 - W_{0:k}^o) \Sigma_k^o + W_{0:k}^o o_k^a I_p.$$

We are now in the position to state our main result (whose proof is in the supplementary material).

**Theorem 1.** Under assumption (A) on the Markov chain $S_t$ and the residuals $e_t$, and under assumption (B1) of Lemma 1, let the time window $K(u)$ be continuous, symmetric, nonnegative, and, for $u > 0$, decreasing with $K(0) = 1$ and $\int_0^\infty K^2(u)du < \infty$. Additionally, choose the bandwidth $b = [\sqrt{T}] \to \infty$ such that $b_T/\sqrt{T} \to 0$. Moreover, assume either (A1) fixed, $o_k^a I_p \neq \pi_k^a \Sigma_0$ or (A2) $p \to \infty$, $p^{1-\gamma} \|o_k^a I_p - \pi_k^a \Sigma_0\|_2 \to \epsilon > 0$ for some $0 < \gamma \leq 2$ such that $p^{2-\gamma}/T \to 0$. Then, with $\alpha_T = T$ in case (A1) or $\alpha_T = T/p^{2-\gamma}$ in case (A2), we have the following results:

(a) $W_{0:k} \approx \frac{1}{\alpha_T}$,

(b) $\alpha_T(W_{0:k} - W_{0:k}) = o_p(1)$, and

(c) $\left\| \hat{\Sigma}_k - \pi_k \Sigma_0 \right\| = \left\| \Sigma_k^o - \pi_k \Sigma_0 \right\| (1 + o_p(\sqrt{\alpha_T})).$

Assumptions (A1) and (A2), respectively, derive from classical assumptions of shrinkage theory, which would quite obviously cease to be of interest if the shrinkage targets $o_k^a I_p$ were equal to the “truth” $\pi_k^a \Sigma_0$. We just recall that the target has been chosen as a constrained version of the truth to regularize the latter one—some sort of deliberate “model misspecification” to achieve regularization. Quite naturally, as in Sancetta (2008), the larger the parameter $\gamma$ ($< 2$) of this model misspecification, the faster is $p$ allowed to grow with $T$, and thus, the more important it is to use shrinkage. To avoid misunderstandings, recall that we use a scaled Frobenius norm in (A2), such that, even asymptotically, the shrinkage target has to remain different from the truth.

The above theorem states that $W_{0:k}^o$ is asymptotically equivalent to $W_{0:k}$ and that the error of $\hat{\Sigma}_k^o$ is in probability asymptotically as small as the corresponding error of the shrinkage estimator $\Sigma_k^o = (1 - W_{0:k}) \Sigma_k^o + W_{0:k} o_k^a I_p$, which is based on the “true” optimal weights, that is, with high probability, $\Sigma_k^o$ will be as much an improvement over $\Sigma_k^o$ as $\Sigma_k^o$ improves the non-shrunken estimator. Note that we do not state an assertion about consistency of the covariance estimates—which is not necessarily of interest, and in fact not possible, if the dimensionality $p$ is of the same order as the sample size $T$. Instead, we show via Theorem (1c) what is appropriate in this general framework, namely, that shrinkage provides an improvement compared to the original estimate $\Sigma_k^0$. If however (e.g., for fixed $p$) the latter is mean-square consistent (compare Lemma 1c)), then $\hat{\Sigma}_k$ will inherit this property.

### 3.3. Shrinkage as Part of an EM Algorithm

The common covariance matrix estimators $\tilde{\Sigma}_k$ are an essential part of the EM algorithm of Section 2.2, but they face the aforementioned numerical problems. Even in a low-dimensional setting, some of these matrices may be numerically very close to singular or even not invertible at all, in particular if, by chance, there are only few observations from a given regime $k$.

As a remedy, we use our developments of the shrinkage estimator of Section 3 to modify the EM algorithm in the following manner. First, in the definition of $Q$ given by Equation (8) we replace the complete log-likelihood $l$, by its penalized version such that now

$$Q(\theta; \theta') = E_{\theta'} \left( \ln p(X, S | \theta) \mid X_1, \ldots, X_T \right).$$

Then, in the E-step again we calculate $Q(\theta; \hat{\theta}^{(n)})$ where $\hat{\theta}^{(n)}$ is the parameter approximation from the previous iteration. Note that the penalty term is independent of $\hat{\theta}^{(n)}$ except for the factors $\alpha_k$ that are linear in the corresponding covariance matrices. Using exactly the same formulas as in Section 2.2.1, the E-step again provides updates of $\hat{S}_t$ that replace $S_t$ in the complete penalized log-likelihood.

In the M-step of Section 2.2, we used maximum likelihood estimates based on the complete log-likelihood where the hidden $S_{t,k}$ were replaced by their approximations from the previous E-step. Now, we use instead maximum penalized likelihood estimates with a data-adaptive regularization parameter. As the penalty term only influences the choice of the covariance matrices, the estimates given by Equations (9) and (10) of Section 2.2.2 for $\alpha_{ij}, \mu_{ik}$ remain unchanged. For estimating the covariance matrix $\Sigma_k$, we started from the rescaled sample covariance matrix (7) as maximizer of the complete log-likelihood with hidden state variables $S_{t,k}$ replaced by their conditional expectations from the E-step. In the modified version of the EM algorithm, we analogously start from the shrunken covariance matrix estimate given by Equation (17), which is motivated by maximization of the complete penalized log-likelihood including a data-adaptive choice of the optimal regularization parameter. Again, we replace the hidden state variables $S_{t,k}$ by their conditional expectations from the E-step. Hence, we consider in the M-step the covariance matrix estimates

$$\tilde{\Sigma}_k = \frac{1}{\pi_k} \hat{\Sigma}_k = \frac{1}{\pi_k} ((1 - \hat{W}_{0:k}) \hat{\Sigma}_k^o + \hat{W}_{0:k} o_k^a I_p), \quad k = 1, \ldots, K,$$  \hspace{1cm} (18)

with $\hat{\Sigma}_k^o = T^{-1} \sum_t \hat{S}_{t,k}(X_t - \hat{\mu}_{t,k})(X_t - \hat{\mu}_{t,k})'$, $o_k^a = p^{-1} \text{tr} (\hat{\Sigma}_k^o)$, for $k = 1, \ldots, K$, and with $\hat{W}_{0:k}$ defined as $\hat{W}_{0:k}^o$ in Equation (16) with $\hat{S}_{t,k}$ replacing $S_{t,k}$ everywhere in the definition. Equations (9), (10), and (18) together provide the components of the updated parameter vector $\hat{\theta}^{(n+1)}$ in the modified EM algorithm.

The modified EM algorithm turns out in simulations to provide numerically more stable covariance matrix estimates, which in turn results in considerable improvements of estimates of transition probabilities and of reconstructions of the hidden state variables, too. Note, however, that we lose the monotonicity of the original numerical algorithm for which the incomplete log-likelihood, that is, the logarithm of (3), evaluated at
4. Numerical Simulations

In this section, we verify our proposed methodology using simulated data. The data will be a multivariate time series of dimension $p = 20$ or 50 generated from either $K = 2$ states, and where stated, $K = 3$ states, with sample size $T = 128$ or 256. To circumvent the problem of identifiability of the states, the transition probability matrices were constructed to have equal entries along the diagonal and equal entries among the off-diagonal elements. This consideration may seem to be restrictive. However, considering the sample size in relation to the dimensionality and taking into account the hidden process, one can easily see that the effective sample size for one of the states could be smaller than $T/2$ or $T/3$. This gives a flavor of how well the procedure works under challenging conditions. Any other choice of the transition probability will make the hidden process either less stable (a lot of fluctuations of the hidden process) or unbalanced. The instability may increase the difficulty in recovering the estimated hidden path, but will not affect the estimation of the covariance matrices if we increase the sample size.

We considered the model defined in Equation (1) to simulate the data. We give the true covariance matrices for each state in Appendix B. The transition probability matrices for the hidden state variables $S_t$ are as follows:

\[
A = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}
\]

with stationary distribution $\pi = (0.5, 0.5)$ for the two-state simulation study, and

\[
A = \begin{pmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{pmatrix}
\]

with stationary distribution $\pi = (1/3, 1/3, 1/3)$ for the three-state simulation study.

We investigated multiple scenarios in this study: (1) we assumed that the state variables were known and fixed (performance of the oracle), (2) we estimated all of the parameters in the model, including the hidden state variables, (3) we let the means differ between states, and (4) we simulated data from the multivariate $t$ distribution. The oracle setting in Simulation 1 allowed us to assess our estimation of each state’s covariance and precision matrix without the added challenges of estimating the hidden states. In Simulations 2–4, we estimated the states using the (modified) EM algorithm. Simulation 2 is a copy of Simulation 1, except the hidden states must be estimated. Simulation 3 investigated the impact of different means between states. And Simulation 4 investigated the performance of the methodologies whenever the underlying distribution has heavy tails.

We compared the performance of our proposed method relative to the traditional EM algorithm, which uses the sample covariance matrix. We assessed performance of the estimators by looking at the percentage relative improvement in average loss (PRIAL), given by

\[
PRIAL(\hat{\Sigma}_k) = 100 \times \frac{E[||\hat{\Sigma}_k - \pi_k \Sigma_0, k||^2] - E[||\hat{\Sigma}_k - \pi_k \Sigma_0, k||^2]}{E[||\Sigma - \pi_k \Sigma_0, k||^2]}.
\]

The expectation is computed by averaging the corresponding statistics over 100 Monte Carlo samples.

Finally, note that we have a label-switching problem, and so we cannot evaluate the performance of estimating each state’s covariance matrix. To address this, we assessed the performance of the estimates by looking at (1) the estimates of the transition probability matrix (because it is symmetric by design, and hence, not affected by the label-switching problem) and (2) the marginal covariance matrix, given by $\Sigma = \sum_{k=1}^{K} \pi_k \Sigma_k$.

4.1. The Performance of the Oracle

We show the estimated PRIALs in Table 1. In all of the cases, we see that shrinkage improves on the estimates of the covariance matrices for each state, with the larger sample size corresponding to a decrease in the PRIALs and the higher dimensionality corresponding to an increase in the PRIALs. The simplest setting is when $T = 256$ and $p = 20$ for two states, as this would result in the largest effective sample size for estimating the parameters. Indeed, this is the only setting where our proposed method yielded the smallest relative improvement.

| Number of states | $p$ | $T$ | State | Covariance matrix | Precision matrix | Marginal covariance matrix |
|------------------|-----|-----|-------|-------------------|------------------|---------------------------|
| Two states       | 20  | 128 | 1     | 9.43             | —                | 62.88                     |
|                  | 2   | 63.02 | —        |                   |                  |                           |
|                  | 256 | 1    | 2.61   | 92.99            | 55.84            |                           |
|                  | 2   | 56.93 | 87.15   |                   |                  |                           |
|                  | 50  | 1    | 25.24  | —                | 78.54            |                           |
|                  | 2   | 79.87 | —       |                   |                  |                           |
|                  | 256 | 1    | 8.92   | 99.67            | 72.26            |                           |
|                  | 2   | 75.51 | —       |                   |                  |                           |
| Three states     | 20  | 128 | 1      | 8.74             | —                | 45.77                     |
|                  | 2   | 66.42 | —       |                   |                  |                           |
|                  | 3   | 11.92 | —       |                   |                  |                           |
|                  | 256 | 1    | 6.48   | 96.74            | 39.09            |                           |
|                  | 2   | 62.87 | 97.77   |                   |                  |                           |
|                  | 3   | 5.52  | 99.99   |                   |                  |                           |
|                  | 50  | 1    | 32.09  | —                | 55.13            |                           |
|                  | 2   | 85.21 | —       |                   |                  |                           |
|                  | 3   | 10.78 | —       |                   |                  |                           |
|                  | 256 | 1    | 17.48  | —                | 51.25            |                           |
|                  | 2   | 80.81 | —       |                   |                  |                           |
|                  | 3   | 8.55  | —       |                   |                  |                           |
The numerical instability of high-dimensional sample covariance matrices becomes a concern when we consider the precision matrix, which is necessary for evaluating the likelihood function. We invert the shrinkage estimator and the sample covariance matrix and investigated how well these inverses perform in estimating the true precision matrix. To assess performance, we again used the PRIAL, similarly defined as that given by Equation (19) with all the matrices replaced by their inverses. The PRIALs for the shrinkage estimator when estimating the precision matrix yielded a substantial improvement over the sample covariance matrix in all simulation settings. Finally, we point out that in many of the simulation settings, the sample covariance matrices \( \hat{\Sigma}_t \) were not always invertible so that PRIALs could not be computed, though the shrinkage estimates were always invertible.

4.2. Performance of the Modified EM Algorithm

In practice, the states are not known and must be estimated. We show these results in Tables 2–4. In Table 2, we see that, in the two-state scenario, the two estimators have comparable performance in estimating the transition probability matrix. However, in all dimensions and sample sizes, the estimate of the marginal covariance matrix was improved when we employed shrinkage via the modified EM algorithm. As expected, we see larger benefits of our proposed method at higher dimensions (for a fixed sample size). The benefits are least pronounced in the easiest considered case of \( p = 20 \) and \( T = 256 \). The benefits of our proposed method are most pronounced when we considered three states. Just as in the oracle scenario in Simulation 1, here the classical EM algorithm did not converge because of singular estimates of the covariance matrix.

In Table 3, we show the results whenever the means between the two states are also different. These results are comparable to the previous setting when the means were truly zero. Indeed, we see that in all cases, our proposed method yielded improved estimates of the covariance matrix marginal over the states, and improved estimates of the transition probability matrix, too. This is most pronounced at the most difficult setting of \( p = 50 \) and \( T = 128 \).

In the previous scenarios, data were simulated from the multivariate normal distribution. Results whenever the data were simulated from the multivariate \( t \) distribution are shown in Table 4. In this scenario, we again see the benefits of shrinkage in estimating the transition probability matrix, and this is most pronounced in the most difficult setting of \( p = 50 \) and \( T = 128 \).

In summary, we have illustrated the benefits of our proposed method for obtaining stable estimates of the transition probability matrix and the covariance matrices under challenging scenarios. The main benefits are in the event when the dimensionality is large and when the effective sample size is small. Indeed, because our proposed method yields invertible estimates of covariance matrices, we can fit our model in cases when the classical EM algorithm would yield unstable estimates or would simply fail.

5. Analysis of the U.S. Portfolio Dataset

Portfolio datasets are inherently high-dimensional. The present dataset, a U.S. industry portfolio dataset publicly available (The dataset can be downloaded at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html), consists of monthly returns from \( p = 30 \) different industry sectors taken from NYSE, NASDAQ, and AMEX. A description of each of industry sector can be obtained from the website. We investigated the performance of our proposed method by looking at the data restricted from January 2000 to December 2011, yielding \( T = 144 \) time points. Thus, the number of parameters in the model greatly exceeds the amount of data available.

We took the log-transform of the data to induce a light tail in the distribution to satisfy the moment conditions of our model. The transformed log-returns were then mean-centered. We optimized the penalized likelihood via our modified EM algorithm at 1000 random initializations. We report results based on the initialization that yielded the largest penalized likelihood. We considered up to five states, and for each number of states we computed the Akaike information criterion (AIC); the model with \( K = 2 \) yielded the lowest AIC.

The estimated shrinkage weights were 0.123 for State 1 and 0.320 for State 2. Comparing the two states, State 1 represents a state of lower volatility and stronger correlations in the industry portfolios. In particular, the games and recreation industry

| Table 2. Results for simulation 2: Frobenius norm of the difference between the true transition probability matrix and the estimated transition probability matrix, and the PRIAL for the estimate of the marginal covariance matrix. |
|---|---|---|---|
| Number of states | \( p \) | \( T \) | Shrinkage | Sample covariance matrix | Marginal covariance matrix |
| 2 | 20 | 128 | 0.01 | 0.02 | 15.71 |
| | 256 | <0.01 | <0.01 | 3.25 |
| | 50 | 128 | 0.01 | — | — |
| | 256 | <0.01 | 0.04 | 34.17 |
| 3 | 20 | 128 | 0.03 | — | — |
| | 256 | 0.03 | 0.04 | 23.85 |
| | 50 | 128 | 0.03 | — | — |
| | 256 | 0.01 | — | — |

| Table 3. Results for simulation 3: Frobenius norm of the difference between the true transition probability matrix and the estimated transition probability matrix, and the PRIAL for the estimate of the marginal covariance matrix. |
|---|---|---|---|
| \( p \) | \( T \) | Shrinkage | Sample covariance matrix | Marginal covariance matrix |
| 20 | 128 | 0.01 | 0.03 | 47.74 |
| | 256 | <0.01 | <0.01 | 51.23 |
| 50 | 128 | 0.01 | 0.06 | 60.81 |
| | 256 | <0.01 | <0.01 | 60.85 |
has larger correlations with the other industries in State 1 than in State 2. Moreover, the chemicals, textiles, construction, steel, machinery, electrical equipment, automobiles, transportation equipment, metal mining, oil, and utility industries have portfolios that are highly correlated with one another; the latter three industries have stronger correlations within this block of industries in State 1 than in State 2. This block of industry portfolios is correlated with another block, namely, the industries regarding business supplies, transportation, wholesale, retail, restaurants and hotels, banking and trading, and an industry labeled as “other.”

The estimated transition probability matrix for this two-state model is

\[
\hat{A} = \begin{pmatrix}
0.953 & 0.047 \\
0.590 & 0.410
\end{pmatrix}.
\]

This suggests that the industry portfolios are more likely to move and stay in the less volatile State 1 than State 2. In Figure 1, we can see that State 1 occurred during important events in the history of the U.S. industry, such as the dot-com bubble collapse (early 2000s), 9/11, and the 2008 financial crisis. We point out that the grouping of companies into industry portfolios is a coarse classification. This aggregation into industry classifications, in addition to the data being observed monthly as opposed to daily, will not allow us to detect the precise date that specific companies are most affected by changes in the economy. Nevertheless, we were able to capture transitions in the U.S. industry between states of stability and high-risk in the market.

We also did the analysis without using our proposed modified EM algorithm. As before, we used 1000 random initializations for the EM algorithm, and used the initialization that yielded the largest log-likelihood. A comparison of the resulting path reconstructions is shown in Figure 2. We see that, though the two results had some agreement in the estimated states, our modified EM algorithm yielded parameter estimates that would result in a more interpretable and more stable reconstruction of the underlying state sequence.

We conclude with a remark that the number of states picked up by AIC should not be taken too seriously. While AIC is well known to asymptotically choose slightly too large models, for small sample sizes that is no longer true. Here, AIC is only able to get hold of major features of the data structure neglecting finer elements (compare, e.g., Steinberg, Gasser, and Franke 1985). In our case, the sample size \( T = 144 \) is small for multivariate time series data. There likely are more than two states, but the sample size does not suffice to discover them.

6. Outlook

For stabilizing the EM algorithm for high-dimensional HMMs, we have used a particular popular shrinkage approach, motivated as a maximum penalized likelihood approach with data-adaptive regularization parameter. Our goal was to develop one simple and feasible solution for the problems encountered with the direct implementation of the standard algorithm and to obtain some theoretical basis for it. For that purpose, we modified the EM algorithm to incorporate a penalty term that results in estimates of the covariance matrices that are, from the point of view of theory and practice, well-established for estimating high-dimensional covariance matrices.

Other methods for getting stable estimates for the covariance matrices or, depending on the intended applications, for the precision matrices could be used as well for replacing the plain empirical covariance matrices and their inverses in the standard version of the M-step. One possibility would be to incorporate additional knowledge into the covariance or precision matrices, such as sparsity or a graphical model (see Friedman, Hastie, and Tibshirani 2007; Bickel and Levina 2008; Fan, Feng, and Wu 2009; Cai, Liu, and Luo 2011, and the references therein). However, such additional knowledge is not always available or trustworthy. Recently, Ledoit and Wolf (2012) proposed...
Comparing Path Reconstructions

Figure 2. The reconstructed state sequence whenever the original EM algorithm was used (blue) and our modified EM algorithm that uses shrinkage (black).

nonlinear shrinkage estimation that is capable of estimating the precision matrix directly, but the computations required are much more involved.

Data-adaptive regularization is not new, and indeed, there have been works on, for example, adaptive lasso techniques (e.g., Fan, Feng, and Wu 2009; Leng, Tran, and Nott 2014). The data-adaptive regularization we have developed in this work led to the loss of the monotonicity property of the EM algorithm, which is a common consequence with modified EM algorithms (see, e.g., Green 1990). As suggested by one of the anonymous referees, one could potentially circumvent this problem by considering instead an iterative adaptive lasso approach such as the one proposed by Sun, Ibrahim, and Zou (2010) that uses the expectation conditional maximization (ECM) algorithm. This approach has a penalty term that changes per iteration, but it also presents new challenges such as the selection of hyperparameters, which is not done automatically as in our approach. The investigation of these alternatives is postponed to future research.

We restricted the discussion to HMMs that are special cases of more general Markov switching models, for example, Markov switching AR or GARCH processes. These more complex models that allow for additional structure due to direct dependence of the observations on each other provide more flexibility, but, of course, estimation and filtering suffers from the same problems as in the case of HMMs. However, the basic model (1) already exhibits the major challenges. A further discussion of the more complex situation of general Markov switching models would go beyond the scope of this article.

Appendix A: Hidden Path Reconstruction

The Viterbi algorithm computes the optimal (most likely) state sequence $S_1, \ldots, S_T$ in a Hidden Markov Model given a sequence of observed outputs. It is based on the maximization of the single best state sequence and uses the dynamic programming methodology—for details, compare, for example, Rabiner (1989). Mark that we assume the model parameter $\theta$ to be given. In practice, it is replaced by the estimate from the EM algorithm.

To find the single best state sequence $\{S_1, S_2, \ldots, S_T\}$ matching the observations $\{X_1, X_2, \ldots, X_T\}$, we define

$$
\delta_t(i) = \max_{S_1, \ldots, S_t} \log P(S_1, S_2, \ldots, S_t, i = 1, X_1, X_2, \ldots, X_t),
$$

that is, $\delta_t(i)$ is the highest log-probability along a single path, at time $t$, which accounts for the first $t$ observations and ends in state $i$. By induction we have

$$
\delta_{t+1}(j) = \max_{S_1, \ldots, S_{t+1}} \log P(S_1, S_2, \ldots, S_{t+1}, j = 1, X_1, X_2, \ldots, X_{t+1})
= \max_{S_1, \ldots, S_{t+1}} \log[P(S_1, \ldots, S_{t+1}, j = 1, X_1, \ldots, X_t) \times P(S_{t+1}, i = 1, S_{t+1}, i = 1) P(X_{t+1} | S_{t+1}, i = 1)],
= \max_{S_1, \ldots, S_{t+1}} \log(\delta_t(i) + \log a_{ij}) + \log f_j(X_{t+1}),
$$

where $f_j(x)$ is given by Equation (2).

To retrieve the state sequence, we need to follow the trajectory delivered by the argument that maximized the previous equation for each $t$ and $j$. We will achieve it via an auxiliary variable $\psi_t(j)$ and the complete procedure is written as follows.

Algorithm 2. Viterbi Algorithm

1. Initialization: $\delta_1(j) = \log \pi_j f_j(X_1)$, $j = 1, \ldots, K$, and $\psi_1(j) = 0$.
2. Recursion:

$$
\delta_t(j) = \max_{1 \leq i \leq K} (\delta_{t-1}(i) + \log a_{ij}) + \log f_j(X_t),
\quad 2 \leq t \leq T,
\quad j = 1, \ldots, K
$$
$$
\psi_t(j) = \arg \max_{1 \leq i \leq K} (\delta_{t-1}(i) + \log a_{ij}),
\quad 2 \leq t \leq T,
\quad j = 1, \ldots, K.
$$
3. Termination: $q^*_T = \arg \max_{1 \leq i \leq K} \delta_T(i)$.
4. Path (State Sequence) Backtracking:

$$
q^*_t = \psi_{t+1}(q^*_{t+1}),
\quad t = T - 1, T - 2, \ldots, 1,
\quad S_{t, q^*_t} = 1, \quad S_{t, i} = 0, \quad i \neq q^*_t,
\quad t = 1, \ldots, T.
$$
Appendix B: Simulation Setup

We consider the following three covariance matrices. For the first state, the covariance matrix comes from an AR(1) model: \( \Sigma_{1,ij} = 0.7^{(|i-j|)} \).

The covariance matrix for the second state is block-diagonal constructed as follows. First, construct a 5 \times 5 correlation matrix \( R_5^{(2)} \) so that all off-diagonal elements are 0.25 and all diagonal elements are 1.0. Then construct a \( p \times p \) block-diagonal correlation matrix \( R_p^{(2)} \) along the diagonals, recalling that \( p = 20 \) or 50. Let \( C_p^{(2)} = 1.5I_p \). Then the covariance matrix for the second state is \( \Sigma_2 = C_p^{(2)}R_p^{(2)}C_p^{(2)} \).

In the three-state simulation study, we have a third-covariance matrix that is also block-diagonal. The correlation block has dimensions \( (p/2) \times (p/2) \) with all off-diagonal elements equal to 0.5 so that we have a \( p \times p \) block-diagonal correlation matrix \( R_p^{(3)} \) composed of only two blocks. Let \( C_p^{(3)} = 1.5I_p \). Then the covariance matrix for the third state is \( \Sigma_3 = C_p^{(3)}R_p^{(3)}C_p^{(3)} \).

For the simulation with different means, we used a mean constant across all dimensions. Specifically, for the \( k \)th state, \( k \in \{1, 2\} \), we set each entry of the mean vector to be \( 2(-1)^k/\sqrt{pk} \).

Finally, to generate the data \( X_t \), we obtained \( \Sigma_1^{1/2} \) using the Cholesky decomposition.

Supplementary Materials

The supplementary material contains the proofs and technical lemmas.

Funding

The authors acknowledge the following sources of financial support for this work: the DAAD Graduate Research Grant, and funding by contract “Projet d’Actions de Recherche Concertées” No. 12/17-045 of the “Communauté française de Belgique” and by IAP research network Grant P7/06 of the Belgian government (Belgian Science Policy).

References

Andrews, D. W. K. (1991), “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation,” *Econometrica*, 59, 817–858. [429]
Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), “A Maximization Technique Occurring in The Statistical Analysis of Probabilistic Functions of Markov Chains,” *The Annals of Mathematical Statistics*, 41, 164–171. [426]
Bickel, P., and Levina, E. (2008), “Covariance Regularization by Thresholding,” *Annals of Statistics*, 36, 2577–2604. [433]
Boehm, H., and von Sachs, R. (2008), “Shrinkage Estimation in the Frequency Domain of Multivariate Time Series,” *Journal of Multivariate Analysis*, 100, 913–935. [425]
Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods* (2nd ed.), New York: Springer. [429]
Cai, T., Liu, W., and Luo, X. (2011), “A Constrained \( l_1 \) Minimization Approach to Sparse Precision Matrix Estimation,” *Journal of the American Statistical Association*, 106, 594–607. [433]
Cappé, O., Moulines, E., and Rydén, T. (2005), *Inference in Hidden Markov Models* (Springer Series in Statistics), New York: Springer. [424]
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, 39, 1–38. [426]
Douc, R., Moulines, E., and Rydén, T. (2004), “Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models With Markov Regime,” *Annals of Statistics*, 32, 2254–2304. [425,426]
Fier, J., Feng, Y., and Wu, Y. (2009), “Network Exploration via the Adaptive Lasso and Scad Penalties,” *The Annals of Applied Statistics*, 3, 521–541. [433,434]
Fiecas, M., and Ombao, H. (2011), “The Generalized Shrinkage Estimator for the Analysis of Functional Connectivity of Brain Signals,” *Annals of Applied Statistics*, 5, 1102–1125. [425]
Fiecas, M., Ombao, H., Linkletter, C., Thompson, W., and Sanes, J. (2010), “Functional Connectivity: Shrinkage Estimation and Randomization Test,” *Neuroimage*, 5, 1102–1125. [425]
Francq, C., and Roussignol, M. (1997), “On White Noises Driven by Hidden Markov Chains,” *Journal of Time Series Analysis*, 18, 553–578. [425]
Francq, C., and Zakoian, J.-M. (2001), “Stationarity of Multivariate Markov-Switching ARMA Models,” *Journal of Econometrics*, 102, 339–364. [425]
Franke, J., Li, W. K., Stockis, J.-P., and Tadjuidje Kangaing, J. (2011), “Mixtures of Nonparametric Autoregressions,” *Journal of Nonparametric Statistics*, 23, 287–303. [431]
Friedman, J., Hastie, H., and Tibshirani, R. (2007), “Sparse Inverse Covariance Estimation With The Graphical Lasso,” *Biostatistics*, 9, 432–441. [433]
Green, P. (1990), “On Use of the EM Algorithm for Penalized Likelihood Estimation,” *Journal of the Royal Statistical Society*, Series B, 52, 443–452. [431,434]
Hall, A. R. (2005), *Generalized Methods of Moments*, Oxford, UK: Oxford University Press. [429]
Härdle, W., and Simar, L. (2007), *Applied Multivariate Statistical Analysis* (2nd ed.), Germany: Springer. [424]
Hartley, H. O. (1958), “Maximum Likelihood Estimation From Incomplete Data,” *Biometrics*, 14, 174–194. [426]
Ledoit, O., and Wolf, M. (2004), A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices,” *Journal of Multivariate Analysis*, 88, 365–411. [425,428,429]
——— (2012), “Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices,” *Annals of Statistics*, 40, 1024–1060. [433]
Leng, C., Tran, M.-N., and Nott, D. (2014), “Bayesian Adaptive Lasso,” *Annals of the Institute of Statistical Mathematics*, 66, 221–244. [434]
Rabiner, L. R. (1989), “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceeding of the IEEE*, 77, 257–286. [427,434]
Rydén, T., Terasvirta, T., and Asbrink, S. (1998), “Stylized Facts of Daily Return Series and the Hidden Markov Model of Absolute Returns,” *Journal of Applied Econometrics*, 13, 217–244. [424,426]
Sancetta, A. (2008), “Sample Covariance Shrinkage for High Dimensional Dependent Data,” *Journal of Multivariate Analysis*, 99, 949–967. [425,428,429,430]
——— (2012), “Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices,” *Annals of Statistics*, 40, 1024–1060. [433]
Steinberg, H. W., Gasser, T., and Franke, J. (1985), “Fitting Autoregressive Models to EEG Time Series: An Empirical Comparison of Estimates of the Order,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33, 143–150. [433]
Stockis, J.-P., Franke, J., and Tadjuidje Kangaing, J. (2010), “On Geometric Ergodicity of Charme Models,” *Journal of Time Series Analysis*, 31, 141–152. [425]
Sun, W., Ibrahim, J., and Zou, F. (2010), “Genomewide Multiple-Loci Mapping in Experimental Crosses by Iterative Adaptive Penalized Regression,” *Genetics*, 185, 349–359. [434]
Tadjuidje Kangaing, J. (2013), “Maximum Likelihood Estimators for Multivariate Hidden Markov Mixture Models,” Technical Report, Report in Wirtschaftsmathematik 146, Department of Mathematics, University of Kaiserslautern. [425,426]
Yang, M. (2000), Some Properties of Vector Autoregressive Processes With Markov Switching Coefficients,” *Econometric Theory*, 16, 23–43. [425]
Yuan, M., and Huang, J. (2009), “Regularized Parameter Estimation of High Dimensional \( \tau \) Distribution,” *Journal of Statistical Planning and Inference*, 139, 2284–2292. [428]