INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/16

SHRINKAGE ESTIMATION FOR MULTIVARIATE HIDDEN MARKOV MIXTURE MODELS

FIECAS, M., FRANKE, J., von SACHS, R. and J. TADJUIDJE

Shrinkage Estimation for Multivariate Hidden Markov Mixture Models

Mark Fiecas[1] Jürgen Franke[2] Rainer von Sachs[3]
Joseph Tadjuidje[4]

Abstract

Motivated from a changing market environment over time, we consider high-dimensional data such as financial returns, generated by a hidden Markov model which 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 which is small compared to the dimension, we apply shrinkage and combine it with an EM-type algorithm. This approach will yield better estimates a more stable estimates of the covariance matrix, which allows for improved reconstruction of the hidden Markov chain. In addition to a simulation study and the analysis of a portfolio data set, we present a series of theoretical results which include a dimensionality asymptotics and which provide the motivation and theoretical foundation for certain techniques used by our method.

Keywords: High-dimensional time series, High-dimensional covariance matrices, Portfolio analysis, EM Algorithm

1 Introduction

Hidden Markov Models (HMM) are a popular class of models for time series data which, locally within a state, behave like i.i.d. data, but their statistical properties repeatedly change between states. A prominent example which motivates our approach is portfolio

[1]Corresponding author, Department of Psychiatry, University of California at San Diego, 9500 Gilman Drive #0738 La Jolla, CA 92093-0738, mfiecas@ucsd.edu

[2]Universität Kaiserslautern, Department of Mathematics, Erwin-Schrödinger-Straβe, 67653 Kaiserslautern, Germany, franke@mathematik.uni-kl.de

[3]ISBA, Universite Catholique de Louvain, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium, rvs@uclouvain.be

[4]Universität Kaiserslautern, Department of Mathematics, Erwin-Schrödinger-Straβe, 67653 Kaiserslautern, Germany, tadjuidj@mathematik.uni-kl.de

analysis which consists of analyzing high-dimensional time series data. A simple but widespread 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 $\mathbf{\Sigma}$ and volatility matrix $\mathbf{\Sigma}^{1/2}$, respectively. However, if the market environment changes, e.g., if it moves to a more volatile state, the covariance matrix changes too. This behavior can be modeled by a HMM with a finite number, say K, of states represented by the different covariance matrices $\mathbf{\Sigma}_k, k = 1, \ldots, K$. As such, the HMM model is particularly interesting for analyzing financial returns which are known to exhibit some particularities ("stylized facts", see, e.g., Rydén et al. [17]) such as departure from the normality assumption and existence of dependence between the data. Indeed, it is well known that a hidden Markov mixture model can circumvent both the problem of normality violation as well as that of dependence. Although locally within a state the data could behave as, e.g., i.i.d. Gaussian, the mixture is not necessarily Gaussian. Moreover, the dependence between the data is inherent to the Markovian structure of the hidden switching mechanism. Hence, in this paper we consider a multivariate time series model that can be regarded as a hidden Markov mixture of K different high-dimensional i.i.d. processes that are not necessarily Gaussian.

A prime goal in this given context of HMM is to estimate the model parameters represented by all the $\mathbf{\Sigma}_k, k = 1, \ldots, K$, as well as the transition probability matrix of the hidden Markov chain and, additionally, a filter which allows for the reconstruction of the values of the hidden Markov chain. For applying the fitted model to financial tasks like the calculation of risk measures, stable estimates of the inverse covariance matrices are also needed which, in view of the typical high dimension of the data, is no trivial task.

In this paper we address the aforementioned questions, studying the afore-motivated framework in order to give a clear-cut presentation of our new approach which combines EM-type estimation of HMM by shrinkage of covariance matrices of dimension p which tend to be high compared to the available ("effective") sample size.

To set up our 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 $\{e_1, \ldots, e_K\}$, where e_i is a unit vector in $\mathbb{R}^K$ and having the i-th entry equal to 1. Moreover, for $i, j = 1, \ldots, K$, we denote the transition probabilities $P(S_t = e_j \,|\, S_{t-1} = e_i, S_{t-2}, \cdots) = P(S_t = e_j \,|\, S_{t-1} = e_i) = a_{ij}$. Then, our model for the data generating process is the following:

$$X_t = \sum_{k=1}^{K} S_{t,k} \left(\mu_k + \mathbf{\Sigma}_k^{1/2} \varepsilon_t \right) \ , \tag{1}$$

where the ε_t are i.i.d. $(0, \mathbf{I}_p)$, and $S_{t,k} = 1$ iff $S_t = e_k$, 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 assume that 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:

A0) 1. S_t is aperiodic, irreducible and stationary,
2. S_t is α-mixing with exponentially decreasing rate,

3. $\varepsilon_t, -\infty < t < \infty$, are independent of S_t, $-\infty < t < \infty$,
4. There exists a finite constant κ_ε such that $\mathbb{E}\varepsilon_{t,i}^4 - 3 \leq \kappa_\varepsilon, i = 1, \ldots, p$; furthermore $\mathbb{E}\|\varepsilon_t\|^8 < \infty$.

For discrete Markov chains with a finite number of states, in particular for the hidden process S_t, aperiodicity and irreducibility already imply the existence of a unique stationary hidden process with stationary distribution $\pi = (\pi_1, \cdots, \pi_k)$ where $\pi_k = P(S_t = e_k)$. 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 data generating mechanism defined in (1) can be regarded as a special case of the vector autoregressive (VAR) model with Markov switching coefficient introduced in Yang [21], with the emphasis here on the switching covariance structure, or alternatively of the multivariate Markov switching ARMA by Francq and Zarkoïan [11]. Furthermore, it can also be considered as the multivariate version of the white noises driven by hidden Markov chains analyzed by Francq and Roussignol [10]. While Yang [21] and Francq and Zarkoïan [11] give conditions for the existence of a stationary solution as well as for geometric ergodicity, Francq and Roussignol [10] investigate the asymptotic behavior of maximum likelihood estimators of the model parameters.

In particular, it is obvious that X_t defined in equation (1) inherits its stationarity as well as mixing properties from those of the process S_t and the assumptions made on ε_t. Further, both properties are needed as key ingredients to derive the consistency and asymptotic normality of the maximum likelihood estimates as it can be adapted, for example, from Douc et al. [7]. Additionally, existence of the m-th moment of ε_t implies the existence of the m-th moment of X_t.

Besides the stability and geometric ergodicity properties of the model and the asymptotic validity of likelihood inference on the model parameters, in high-dimensional situations it is not guaranteed that straight-forward implementations of the estimates lead to reasonable results for realistic sample sizes. For multivariate data, it is well known that a large-scale sample covariance matrix is not guaranteed to be invertible if the dimensionality of the data is large with respect to the amount of data available. In particular, if the dimension of the process is large compared to the available data set it is guaranteed that the sample covariance will not be invertible as one can see from Section 8.4 in Härdle and Simar [13]. This is a problem for calculating the likelihood following a (pseudo-) Gaussian approach because an invertible estimate of the covariance matrix is necessary. This problem is natural and 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 will be very small. In this paper we take this important particularity into account by extending the approach proposed by Sancetta [18] to the situation of multivariate hidden Markov mixture models. Sancetta [18] introduced shrinkage estimator for covariance matrices for dependent data, which extended the work by Ledoit and Wolf [15] for i.i.d. multivariate data. The shrinkage estimator for large-scale covariance matrices developed by Ledoit and Wolf [15] guarantees an estimate of

the covariance matrix that is invertible, positive-definite, more numerically stable and have smaller mean-squared error than the sample covariance matrix. 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 [9]). When used in our context of a hidden Markov mixture model, the shrinkage estimator will also be beneficial for improving on the covariance matrix, especially when very few observations occur in a given state.

In case the data come from a time series, the assumption of independence within a state seems to be somewhat restrictive. The goal of this paper is to illustrate the interest in combining switching for high-dimensional time series data and shrinkage estimation of variance-covariance matrices. Therefore, for the sake of simplicity we will focus on this basic model that has the advantage of illustrating all the major challenges in this context. Note that shrinkage for dependent data has also been widely investigated in the frequency domain of time series, where the technology has been successfully applied to shrinkage estimation of the spectral density matrix (Boehm et al [3], Fiecas et al [9], and Fiecas and Ombao [8]).

This paper is organized as follows. In Section 2 we treat the shrinkage estimator for our model and deliver some theoretical results on how to optimally choose and estimate its parameters, the shrinkage weights. We include the case of both sample size T and dimensionality p tend to infinity in order to cover the situation of particular interest for shrinkage, i.e., when the dimensionality is of the order of the sample size or even (slightly) larger. In Section 3 the likelihood inference is studied with a focus on the EM algorithm for parameter estimation and the Viterbi algorithm for the hidden path retrieval. A simulation study is carried out in Section 4, completed by a portfolio analysis of a real data set in Section 5. All of the proofs are in the supplementary material for this paper.

2 Shrinkage Estimation for Covariance Matrices

2.1 Defining the oracle estimators

In this paper, our focus is the estimation of $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$, where we additionally have to estimate the means $\mu_1, \ldots, \mu_K$. As laid out in our model in Section 1, S_t are the values of a *hidden* Markov chain, but for the moment let us assume that an *oracle* tells us the values of S_t. In the following, an upper index $^{\mathbf{o}}$ designates oracle estimates which assume the state variables to be known.

Natural oracle estimates for μ_k and $\mathbf{\Sigma}_k$ would be the sample mean and the covariance matrix of all observations generated from the k-th regime, namely, for $\sum_{t=1}^{T} S_{t,k} \neq 0$, we have

$$\mu_k^{\mathbf{o}} = \frac{1}{\sum_{t=1}^{T} S_{t,k}} \sum_{t=1}^{T} S_{t,k} X_t,$$

and

$$\widetilde{\mathbf{\Sigma}}_k^{\mathbf{o}} = \frac{1}{\sum_{t=1}^{T} S_{t,k}} \sum_{t=1}^{T} S_{t,k} (X_t - \mu_k^{\mathbf{o}})(X_t - \mu_k^{\mathbf{o}})'.$$

If $\sum_{t=1}^{T} S_{t,k} = 0$, we set $\mu_k^{\mathbf{o}} = 0$ and $\overset{\sim}{\mathbf{\Sigma}}{}_k^{\mathbf{o}} = 0$ for convenience. This happens with exponentially decreasing probability and will be asymptotically negligible. In practice, however, we may already run into numerical problems if we have states which are rarely visited such that the effective sample size $\sum_{t=1}^{T} S_{t,k}$ is small, though not 0. Therefore, we consider the biased estimator of the variance-covariance matrix

$$\mathbf{\Sigma}_k^{\mathbf{o}} = \frac{1}{T} \sum_{t=1}^{T} S_{t,k} (X_t - \mu_k^{\mathbf{o}})(X_t - \mu_k^{\mathbf{o}})' = \pi_k^{\mathbf{o}} \overset{\sim}{\mathbf{\Sigma}}{}_k^{\mathbf{o}}$$

as the starting point for shrinkage, where

$$\pi_k^{\mathbf{o}} = \frac{1}{T} \sum_{t=1}^{T} S_{t,k}.$$

is the oracle estimate of π_k. Our approach is to define the shrinkage estimate in analogy to Ledoit et al. [15] and Sancetta [18], respectively. We follow the philosophy that the sample covariance matrix, possibly close to singular in high dimensions p, is regularized via a convex combination with a multiple $\alpha_k \mathbf{I}_p$ of the identity matrix, namely, we consider the shrinkage estimate

$$\mathbf{\Sigma}_k^s = (1 - W_k)\mathbf{\Sigma}_k^{\mathbf{o}} + W_k \alpha_k \mathbf{I}_p$$

with some shrinkage weight $0 \leq W_k \leq 1$. Note that, in the case of no shrinkage so that $W_k = 0$, we have $\mathbf{\Sigma}_k^s = \mathbf{\Sigma}_k^{\mathbf{o}}$. The proportionality factor α_k is chosen such that $\operatorname{tr}(\alpha_k \mathbf{I}_p) = \mathbb{E} \operatorname{tr}(\mathbf{\Sigma}_k^{\mathbf{o}})$, i.e., $\alpha_k \mathbf{I}_p$ is in a certain sense of the same size as $\mathbf{\Sigma}_k^{\mathbf{o}}$ but it has a much simpler structure and a high regularity.

One possibility to measure the degree of regularity of a variance-covariance matrix is its condition number. Recall that the condition number of a matrix $\mathbf{\Sigma}$ is the ratio of its largest to its smallest eigenvalue, i.e., if we let $\lambda_1 \geq \ldots \geq \lambda_p$ be the ordered eigenvalues of $\mathbf{\Sigma}$, then its condition number is $\operatorname{cond}(\mathbf{\Sigma}) = \frac{\lambda_1}{\lambda_p}$. The condition number for a diagonal matrix is the ratio between the largest and the smallest diagonal element. The closer this ratio is to one, the better conditioned is the matrix, and hence the "more regular". Shrinkage is meant to reduce the dispersion between largest and smallest eigenvalues, roughly speaking shrinkage pushes all eigenvalues towards their"grand mean". Evidently a multiple of the identity matrix has eigenvalues with the smallest possible dispersion, and hence the best condition number. In Section 4 we use this concept to assess the performance of our shrinkage method.

2.2 Asymptotic considerations for oracle estimation

An oracle estimator for the factor α_k will be given by $\alpha_k^{\mathbf{o}} = \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_k^{\mathbf{o}}$. In the following lemma we show that this is a consistent estimator of $\alpha_k = \frac{1}{p} \operatorname{tr}(\pi_k \mathbf{\Sigma}_k)$ with an asymptotically negligible bias.

Here and in the following, $\lambda_{\max}(\mathbf{\Sigma}_k)$ and $\lambda_{\min}(\mathbf{\Sigma}_k)$ denote the largest and smallest eigenvalue of the covariance matrix $\mathbf{\Sigma}_k$, respectively. Also, we define the matrix norm $||\cdot||$ to be the (scaled) Frobenius norm: $||\mathbf{A}||^2 = \frac{1}{p} \operatorname{tr}(\mathbf{A}\mathbf{A}')$.

Lemma 1 *Let the data be generated from model (1), either for fixed p or for with T increasing $p \to \infty$, satisfying assumption A0). Consider the assumptions (which are automatically satisfied for fixed p)*

B1) $\max_{1\le i\le p} |\mu_{k,i}|$ *as well as* $||\mathbf{\Sigma}_k||^2$ *and* $\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}_k$ *are uniformly bounded in* p.

B2) $\pi_k > 0$, $\operatorname{tr}(\mathbf{\Sigma}_k) \ge cp^r$ *for some* $c > 0, 0 \le r \le 1$ *and all* p.

Then, it follows for some $0 < \beta < 1$ *and all* $k = 1, \dots, K$,

a) $\mathbb{E}\ \boldsymbol{\mu}_k^{\mathbf{o}} = \mu_k + \mu_k O(\beta^T) = \mu_k + O(p^{1/2}\beta^T)$,
$\mathbb{E}\ \pi_k^{\mathbf{o}}||\boldsymbol{\mu}_k^{\mathbf{o}} - \mu_k||^2 = \frac{1}{T}\pi_k \operatorname{tr}\mathbf{\Sigma}_k = O(\frac{p}{T})$ *and*
$\mathbb{E}\ \mathbf{\Sigma}_k^{\mathbf{o}} = (\pi_k - \frac{1}{T})\mathbf{\Sigma}_k + O(\frac{\beta^T}{T}) = \pi_k\mathbf{\Sigma}_k + O\left(\frac{1}{T}\right)$.

b) If B1) holds, the normalized trace estimator $\alpha_k^{\mathbf{o}}$ *is mean-square consistent,*

$$\mathbb{E}(\alpha_k^{\mathbf{o}} - \alpha_k)^2 = O\left(\frac{1}{T}\right),$$

where $\alpha_k^{\mathbf{o}} = \frac{1}{p}\operatorname{tr}\mathbf{\Sigma}_k^{\mathbf{o}}$, $\alpha_k = \frac{1}{p}\operatorname{tr}(\pi_k\mathbf{\Sigma}_k)$, *and* $\mathbb{E}\ \alpha_k^{\mathbf{o}} = \alpha_k + O\left(\frac{1}{T}\right)$.

c) If B1), B2) hold,

$$\frac{\operatorname{var}(\operatorname{tr}(\mathbf{\Sigma}_k^{\mathbf{o}}))}{|\operatorname{tr}(\pi_k\mathbf{\Sigma}_k)|^2} = O\left(\frac{p^{2(1-r)}}{T}\right),$$

i.e., the relative error of $\operatorname{tr}(\mathbf{\Sigma}_k^{\mathbf{o}})$ *as an estimate of* $\operatorname{tr}(\pi_k\mathbf{\Sigma}_k)$ *converges to 0 if* $p^{1-r} = o\left(\sqrt{T}\right)$.

Assumption B1) implies in particular

$$\frac{1}{p}||\mu_k||^2 = O(1),\ ||\mathbf{\Sigma}_k||^2 = O(1), \tag{2}$$

i.e., $\mu_k, \mathbf{\Sigma}_k$ do not grow with p intrinsically, but only due to the increasing dimension. Both assumptions on $\mathbf{\Sigma}_k$ are in particular satisfied if $\lambda_{\max}(\mathbf{\Sigma}_k)$ is bounded uniformly in p. However, that is not a necessary condition. B1) includes also, e.g., the situation where $\lambda_{\max}(\mathbf{\Sigma}_k)$ is growing with a rate up to $p^{1/2}$, as long as all except finitely many of the eigenvalues are uniformly bounded with p.

Note also that $\operatorname{tr}(\mathbf{\Sigma}_k) \ge p\lambda_{\min}(\mathbf{\Sigma}_k)$, and, therefore, the second part of assumption B2) is automatically satisfied if

$$\lambda_{\min}(\mathbf{\Sigma}_k) \ge \frac{c}{p^{1-r}},$$

i.e. the smallest eigenvalue does not converge to 0 too fast with increasing dimension p. However, B2) is weaker and allows even for singular $\mathbf{\Sigma}_k$.

2.3 Optimizing the shrinkage weights and properties of the resulting shrinkage estimator

In order to optimize the choice of the shrinkage weights, we look for weights $W_k^{\mathbf{o}}$ which minimize the mean-squared error of $\mathbf{\Sigma}_k^s$ as an estimate of $\pi_k \mathbf{\Sigma}_k$:

$$\begin{aligned} W_k^{\mathbf{o}} &= \arg \min_{W_k \in [0,1]} \mathbb{E}||\mathbf{\Sigma}_k^s - \pi_k \mathbf{\Sigma}_k||^2 \\ &= \arg \min_{W_k \in [0,1]} \mathbb{E}||(1 - W_k)\ \mathbf{\Sigma}_k^{\mathbf{o}} + W_k \alpha_k \mathbf{I}_p - \pi_k \mathbf{\Sigma}_k||^2. \end{aligned}$$

Assuming the existence of all necessary moments, we get the following Lemma directly from Proposition 1 of Sancetta [18].

Lemma 2 *Under the model (1), for $k = 1, \ldots, K$,*

$$W_k^{\mathbf{o}} = \frac{\mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}} - \pi_k \mathbf{\Sigma}_k||^2}{\mathbb{E}||\alpha_k I_p - \mathbf{\Sigma}_k^{\mathbf{o}}||^2} \wedge 1. \tag{3}$$

The shrinkage estimate $\mathbf{\Sigma}_k^{s,\mathbf{o}}$, which uses the optimal weights of Lemma 2, is, however, not a feasible estimate because $\alpha_k = \frac{1}{p} \operatorname{tr}(\pi_k \mathbf{\Sigma}_k)$ is not known. Referring to Lemma 1 we replace it in the denominator by its oracle estimate $\alpha_k^{\mathbf{o}} = \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_k^{\mathbf{o}}$. This choice for an estimator of the denominator of $W_k^{\mathbf{o}}$ has been discussed by Ledoit et al. [15] in detail; it leads to the sample estimator of the denominator where the expectation (risk) is just replaced by the sample analogue (loss), see equation 5 further below.

To get a direct estimate of the numerator in (3), we follow Sancetta's approach. From Lemma 5 (in the supplementary material), we have

$$\mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}} - \pi_k \mathbf{\Sigma}_k||^2 = \frac{1}{p} \sum_{i,j=1}^{p} \operatorname{var} \left(\frac{1}{T} \sum_{t=1}^{T} S_{t,k} (X_{ti} - \mu_{ki})(X_{tj} - \mu_{kj}) \right) + O\left(\frac{p}{T^2} \right),$$

where, similar to the arguments in Proposition 7.3.3 and 7.3.4. of Brockwell and Davis [4], replacing the estimates $\mu_k^{\mathbf{o}}$ in the definition of $\mathbf{\Sigma}_k^{\mathbf{o}}$ by their true values μ_k, has an asymptotically neglible effect. Note that the first term on the right-hand side is of order $O(\frac{p}{T})$ by the discussion preceeding Lemma 5.

Setting $y_{t,ij} = S_{t,k}(X_{ti} - \mu_{ki})(X_{tj} - \mu_{kj})$, the above expression is given by

$$\frac{1}{p} \sum_{i,j=1}^{p} \frac{1}{T} \operatorname{var} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t,ij} \right), \tag{4}$$

where asymptotically, for any fixed k, $\operatorname{var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t,ij} \right)$ is well-known as the long run variance, see, e.g., Hall [12], of the time series $y_{t,ij}$. 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

$$\bar{c}^{\mathbf{o}}_{ij,k}(s) = \frac{1}{T}\sum_{t=1}^{T-s}\left(y_{t,ij} - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,ij}\right)\left(y_{t+s,ij} - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,ij}\right)$$

for $0 \le s \le T-1,\ \bar{c}^{\mathbf{o}}_{ij,k}(-s) = \bar{c}^{\mathbf{o}}_{ij,k}(s)$.

From Lemma 3, the variance of the sample mean $\frac{1}{T}\sum_{t=1}^{T} y_{t,ij}$ is asymptotically $\frac{1}{T}f_{ij,k}(0)$. Following Sancetta [18], 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 [1]). For some kernel $K(u) \ge 0,\ K(u) = K(-u)$ and $K(0) = 1$ and bandwidth $b > 0$

$$\bar{f}^{b}_{ij,k}(0) = \sum_{s=-T+1}^{T-1} K\left(\frac{s}{b}\right)\bar{c}^{\mathbf{o}}_{ij,k}(s),$$

so that we estimate (4) with $\frac{1}{Tp}\sum_{i,j=1}^{p}\bar{f}^{b}_{ij,k}(0)$.

By the same kind of arguments used in showing Lemma 5, replacing the unknown means μ_k has an asymptotically neglible effect such that we may instead consider the following feasible estimates:

$$\begin{aligned}
\hat{y}_{t,ij} &= S_{t,k}(X_{ti} - \mu^{\mathbf{o}}_{ki})(X_{tj} - \mu^{\mathbf{o}}_{kj}),\\
\hat{c}^{\mathbf{o}}_{ij,k}(s) &= \frac{1}{T}\sum_{t=1}^{T-s}\left(\hat{y}_{t,ij} - \frac{1}{T}\sum_{\tau=1}^{T}\hat{y}_{\tau,ij}\right)\left(\hat{y}_{t+s,ij} - \frac{1}{T}\sum_{\tau=1}^{T}\hat{y}_{\tau,ij}\right),\\
\text{and } \hat{f}^{b}_{ij,k}(0) &= \sum_{s=-T+1}^{T-1} K\left(\frac{s}{b}\right)\hat{c}^{\mathbf{o}}_{ij,k}(s)\ .
\end{aligned}$$

We finally get an estimator of the optimal shrinkage weights $W^{\mathbf{o}}_k$ as follows:

$$\hat{W}^{\mathbf{o}}_k = \frac{\frac{1}{p}\frac{1}{T}\sum_{i,j=1}^{p}\hat{f}^{b}_{ij,k}(0)}{\left\|\mathbf{\Sigma}^{\mathbf{o}}_k - \alpha^{\mathbf{o}}_k\mathbf{I}_p\right\|^2} \wedge 1. \tag{5}$$

Now we can construct the optimal shrinkage estimator, still under the oracle of observed values of S_t:

$$\hat{\mathbf{\Sigma}}^{s}_k = (1 - \hat{W}^{\mathbf{o}}_k)\mathbf{\Sigma}^{\mathbf{o}}_k + \hat{W}^{\mathbf{o}}_k\alpha^{\mathbf{o}}_k I_p.$$

We are now in the position to state our main result in the following theorem.

Theorem 1 *Under the above assumptions on the Markov chain S_t and the residuals ε_t, i.e. A0), 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, consider the bandwidth $b = b_T \to \infty$ such that $\frac{b_T}{\sqrt{T}} \to 0$. Moreover, assume either*

A1) p fixed, $\alpha_k \mathbf{I}_p \neq \pi_k \mathbf{\Sigma}_k$

or

A2) $p \to \infty$, $p^{1-\gamma} \left\| \alpha_k \mathbf{I}_p - \pi_k \mathbf{\Sigma}_k \right\|^2 \to c > 0$ *for some* $0 < \gamma < 2$ *such that* $\frac{p^{2-\gamma}}{T} \to 0$.

Then, with $a_T = T$ *in case A1) and* $a_T = \frac{T}{p^{2-\gamma}}$ *in case A2)*

a) $W_k^{\mathbf{o}} \asymp \frac{1}{a_T}$

b) $a_T \left(\hat{W}_k^{\mathbf{o}} - W_k^{\mathbf{o}} \right) = o_p(1)$

c) $\left\| \hat{\mathbf{\Sigma}}_k^s - \pi_k \mathbf{\Sigma}_k \right\| = \left\| \mathbf{\Sigma}_k^{s,\mathbf{o}} - \pi_k \mathbf{\Sigma}_k \right\| \left(1 + o_p \left(\frac{1}{\sqrt{a_T}} \right) \right)$

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 $\alpha_k \mathbf{I}_p$ were equal to the "truth" $\pi_k \mathbf{\Sigma}_k$. We just recall that the target has been chosen as a constrained version of the truth in order to regularize the latter one - some sort of deliberate "model misspecification" to achieve regularization. Quite naturally, as in Sancetta [18], the larger the parameter γ (< 2) of this model misspecifcation, the faster is p allowed to grow with T, and thus, the more important it is to use shrinkage. (To avoid misunderstandings, we recall the different scale of the norm used in A2), meaning that even asymptotically, the shrinkage target has to remain different from the truth.)

The above theorem states that $\hat{W}_k^{\mathbf{o}}$ is asymptotically equivalent to $W_k^{\mathbf{o}}$, and that the error of $\hat{\mathbf{\Sigma}}_k^s$ is in probability asymptotically as small as the corresponding error of the shrinkage estimator

$$\mathbf{\Sigma}_k^{s,\mathbf{o}} = (1 - W_k^{\mathbf{o}}) \mathbf{\Sigma}_k^{\mathbf{o}} + W_k^{\mathbf{o}} \alpha_k I_p,$$

which is based on the "true" optimal weights, i.e., with high probability, $\hat{\mathbf{\Sigma}}_k^s$ will be as much an improvement over $\mathbf{\Sigma}_k^{\mathbf{o}}$ as $\mathbf{\Sigma}_k^{s,\mathbf{o}}$ 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 in the order of the sample size T. Instead, we show via Theorem 1, c) what is appropriate in this general framework, namely, that shrinkage provides an improvement compared to the original estimate $\mathbf{\Sigma}_k^{\mathbf{o}}$. If however (for example, for fixed p) the latter is mean-square consistent (compare Lemma 1, c)), then $\hat{\mathbf{\Sigma}}_k^s$ will inherit this property.

3 Likelihood Inference

As we allow our switching regime process to have means $\mu_k \neq 0$, we have to include estimation of those state means μ_k into the parametrization of the now following likelihood inference.

Given $S_{t,k} = 1$, and using a Gaussian pseudo maximum likelihood approach, we start from the probability density

$$\begin{aligned} f_k(x) &\equiv f(x, \mu_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)'\mathbf{\Sigma}_k^{-1}(x-\mu_k)\right) \\ &= f\left(x \mid S_t = e_k, \theta\right) \end{aligned} \tag{6}$$

for X_t given $S_t = e_k$, where θ denotes the vector of all parameters of the K different probability density functions and, additionally, the transition probabilities of the Markov chain. Hence, if we were given the hidden sample path $S = (S_1, \ldots, S_T)$, the conditional likelihood is given by

$$L(X_1, \ldots, X_T \mid S, \theta) = \prod_{t=1}^{T} f(X_t \mid S_t, \theta) = \prod_{t=1}^{T} \left(\sum_{k=1}^{K} S_{t,k} f(X_t, \mu_k, \mathbf{\Sigma}_k) \right)$$

and therefore the complete likelihood is given by

$$P(X, S \mid \theta) = L(X_1, \ldots, X_T \mid S, \theta) L(S \mid \theta).$$

Unfortunately the hidden process is unknown. Therefore, we have to deal with the incomplete likelihood function where we have to take into account all possible hidden sample paths

$$L(X_1, \ldots, X_T \mid \theta) = \sum_{\text{all possible } S} P(X, S \mid \theta). \tag{7}$$

Adapting the results in Douc et al. [7], consistency and asymptotic normality of the parameter estimates, which maximize (7), could be shown, 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 ε_t.

The direct numerical calculation of those pseudo maximum likelihood estimates, however, would be obviously quite cumbersome. We prefer to interpret the hidden process S_t as missing data and to use one of the well known stochastically motivated procedures designed for calculating maximum likelihood estimates when the data exhibit some missing values. One of the most popular procedures, the expectation maximization (EM) algorithm, is discussed in the next section.

3.1 EM algorithm - the general idea

Before we present the algorithm in our special context, let us recall that the EM algorithm can be traced back to Hartley [14] and presented by Baum et al. [2] in a general framework. Finally, it was made popular through the now famous work of Dempster et al. [6], and since then various variants of this algorithm have been proposed in the literature.

First, we consider the conditional (pseudo) log likelihood of $(X_1, \ldots, X_T, S_1, \ldots, S_T)$ given the observations $(X_1, \ldots, X_T)$ as a function of the parameter of interest θ where the conditional expectation is calculated w.r.t. to a dummy parameter value θ' (see, e.g., Baum et al.

[2] for more details on this issue):

$$\begin{aligned}\mathcal{Q}(\theta;\theta') &= \sum_t \mathbb{E}_{\theta'}\left(\sum_{k=1}^{K} S_{t,k}\log f(X_t,\mu_k,\mathbf{\Sigma}_k) \mid X_1,\ldots,X_T\right)\\ &+\sum_t \mathbb{E}_{\theta'}\left(\sum_{i,j=1}^{K} S_{t,j}S_{t-1,i}\log a_{ij} \mid X_1,\ldots,X_T\right)\\ &+\mathbb{E}_{\theta'}\left(\sum_{k=1}^{K} S_{1,k}\log P(S_{1,k}=1) \mid X_1,\ldots,X_T\right).\end{aligned}$$

Using that we consider a Gaussian pseudo likelihood and neglecting constants which do not depend on the model parameters, we can alternatively set

$$\begin{aligned}\mathcal{Q}(\theta;\theta') &= \sum_t \mathbb{E}_{\theta'}\left(\sum_{i,j=1}^{K} S_{t,j}S_{t-1,i}\log a_{ij} \mid X_1,\ldots,X_T\right)\\ &+ \mathbb{E}_{\theta'}\left(\sum_{k=1}^{K} S_{1,k}\log P(S_{1,k}=1) \mid X_1,\ldots,X_T\right)\\ &- \frac{1}{2}\sum_t \mathbb{E}_{\theta'}\left(\sum_{k=1}^{K} S_{t,k}(\log|\mathbf{\Sigma}_k| + (X_t-\mu_k)'\mathbf{\Sigma}_K^{-1}(X_t-\mu_k)) \mid X_1,\ldots,X_T\right),\end{aligned}$$

The basic idea of the algorithm is now the following: In the E-step we assume the parameters of the model to be known and derive estimates of the hidden state variables, where we make use of the forward and backward variables [16]. As an alternative, we could use the recursive approach proposed in Cappé et al. [5] to calculate the estimated state variables $\hat{S}_{t,i} = \gamma_i(t), i, = 1,\ldots,K$. The latter may be interpreted as the conditional probabilities for the hidden process to be in state i at time t given the data $X_t,\ldots,X_T$. In the M-step, we assume the hidden states to be known and update the parameter estimates.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

$$\mathcal{Q}(\theta, \hat{\theta}^{(n)})$$

   as a function of θ. As part if this calculation, approximations $\hat{S}_{t,k} = P(S_{t,k} = 1|X_1, \ldots, X_T, \theta)$ are derived.
3. M-Step: Determine $\hat{\theta}^{(n+1)}$ as the maximizer of $\mathcal{Q}(\theta, \hat{\theta}^{(n)})$ over θ.
4. Iterate the E-step and M-Step until a stopping criterion is satisfied.

3.2 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, compare, e.g., Rabiner [16], and given here only for sake of completeness. For readability, we write θ instead of the actually used parameter value $\hat{\theta}^{(n)}$ of the n-th iteration. Recall that θ contains all the information about the means μ_k and covariance matrices $\mathbf{\Sigma}_k$ 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, $f_k(x) = f(x, \mu_k, \mathbf{\Sigma}_k)$ given by (6) denotes the density of X_t in state k, where the parameters $\mu_k, \mathbf{\Sigma}_k$ of those Gaussian densities are part of θ. We introduce the forward-backward variables as auxiliary quantities which may be calculated from θ and the data using the following recursions.

Forward variables: For $i = 1, \ldots, K$, we define

$$\alpha_i(t) = P(X_1, \ldots, X_t, S_{t,i} = 1 \mid \theta),$$

for which we have the following relations

1. $\alpha_i(1) = \pi_i f_i(X_1)$
2. $\alpha_i(t+1) = \left(\sum_{k=1}^K \alpha_k(t) a_{ki}\right) f_i(X_{t+1})$
3. $L(X_1, \ldots, X_T \mid \theta) = \sum_{k=1}^K \alpha_k(T)$

Backward variables: For $i = 1, \ldots, K$, we define

$$\beta_i(t) = P(X_{t+1}, \ldots, X_T \mid S_{t,i} = 1,\ \theta)$$

for which we have

1. $\beta_i(T) = 1$,

2. $\beta_i(t) = \sum_{k=1}^{K} a_{ik} f_k(X_{t+1}) \beta_k(t+1),\ T-1 \leq t \leq 1.$

State variables: From forward and backward variables, we calculate

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{k=1}^{K} \alpha_k(t)\beta_k(t)} \qquad \text{and} \qquad \xi_{ij}(t) = \frac{\gamma_i(t) a_{ij} f_j(X_{t+1}) \beta_j(t+1)}{\beta_i(t)}.$$

The former provides approximations of the hidden state variables, which are the conditional probabilities for the hidden process to be in state i at time t given the data $X_1, \ldots, X_T$ and parameter θ:

$$\hat{S}_{t,i} = \gamma_i(t) = P(S_{t,i} = 1 \mid X_1, \ldots, X_T, \theta) = \mathbb{E}(S_{t,i} \mid X_1, \ldots, X_T, \theta).$$

The latter provides approximations of the bivariate conditional distribution of adjacent state variables given the data and parameter θ:

$$\xi_{ij}(t) = P(S_{t,i} = 1, S_{t+1,j} = 1 \mid X_1, \ldots, X_T, \theta).$$

In particular, we have

$$\sum_t \gamma_i(t) = \sum_t \mathbb{E}_\theta S_{t,i} \qquad \text{and} \qquad \sum_t \xi_{ij}(t) = \sum_t \mathbb{E}_\theta S_{t,i} S_{t+1,j}. \tag{8}$$

3.3 M-Step

In the M-step, we calculate an update $\hat{\theta}^{(n+1)}$ of the parameter value $\hat{\theta}^{(n)} = \theta$ used in the E-step. Using (8), we first get updated estimates of the *transition probabilities*

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} = \frac{\sum_{t=1}^{T-1} \hat{S}_{t,i} \hat{S}_{t+1,j}}{\sum_{t=1}^{T-1} \hat{S}_{t,i}}, \quad i, j = 1, \ldots, K, \tag{9}$$

and correspondingly, of the *stationary state probabilities*

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

The *state means* are updated as

$$\hat{\mu}_k = \frac{\sum_{t=1}^{T} \hat{S}_{t,k} X_t}{\sum_{t=1}^{T} \hat{S}_{t,k}}, \quad k =, 1, \ldots, K. \tag{10}$$

Those recursions are standard in this setting. However, at this stage we benefit from our developments of the shrinkage estimator of Section 2 as using the classical covariance estimators $\widetilde{\mathbf{\Sigma}}_k^{\mathbf{o}}$, one may face the aforementioned numerical problems. We recall that there is

no guarantee of obtaining estimates of the covariance matrices that are invertible because it might happen that very few observations enter into the covariance estimator for a given state k. In fact, it is clear that even in a low-dimensional setting some of these matrices may be numerically very close to singular. It will turn out that using our more numerically stable shrinkage-based estimators of the *state covariance matrices* will result in considerable improvements of estimates of transition probabilities and of reconstructions of the hidden state variables, too. Our shrinkage-based estimate of the state covariance matrices are

$$\hat{\mathbf{\Sigma}}_k = \frac{1}{\hat{\pi}_k}\hat{\mathbf{\Sigma}}_k^s = \frac{1}{\hat{\pi}_k}\left((1-\hat{W}_k^{\mathbf{o}})\hat{\mathbf{\Sigma}}_k^{\mathbf{o}} + \hat{W}_k^{\mathbf{o}}\alpha_k^{\mathbf{o}} I_p\right), \quad k = 1, \ldots, K, \tag{11}$$

where

$$\hat{\mathbf{\Sigma}}_k^{\mathbf{o}} = \frac{1}{T}\sum_t \hat{S}_{t,k}(X_t - \hat{\mu}_k)(X_t - \hat{\mu}_k)', \quad k = 1, \ldots, K.$$

Note that in the algorithm, we do not use the oracle estimate $\mathbf{\Sigma}_k^{\mathbf{o}}$ from Section 2, but its approximation based on replacing the unknown state variable $S_{t,k}$ by their approximations from the E-step. Equations (9), (10) and (11) provide the components of the updated parameter vector $\hat{\theta}^{(n+1)}$.

Finally, for the hidden path reconstruction we make use of the Viterbi algorithm, described in Appendix A, which computes 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, e.g., Rabiner [16] for more details.

4 Numerical Simulations

In this section we verify our proposed methodology using simulated data. For the sake of simplicity and conciseness we carry out the simulation study only for a symmetric transition probability matrix and a data of sample size $T = 256$ for a multivariate time series of dimension $p = 20$. This consideration seems 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 two states could be smaller than $T/2 = 128$. 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 instabilty may increase the difficulty in recovering the estimated hidden path, but will not affect the covariance matrices estimation if we increase the sample size.

4.1 Simulation Settings

In this section we consider the model defined in equation (1), with $\mu_k = 0$ for all states k, in order to focus on estimation of the state covariance matrices. Note, however that in our algorithms, we do not use this information from the true model. Throughout, let $\mathbf{0}_q$ be the

$q \times q$ matrix of 0s and $\mathbf{I}_q$ be the $q \times q$ identity matrix. Furthermore, we assume here a two state hidden Markov process $S_t \in \{0, 1\}, \varepsilon_t \sim \mathcal{N}(0, \mathbf{I}_p)$ and therefore define

$$X_t = (S_t \boldsymbol{\Sigma}_1^{1/2} + (1 - S_t)\boldsymbol{\Sigma}_2^{1/2})\varepsilon_t.$$

Additionally, the hidden process S_t has a transition probability matrix

$$\mathbf{A} = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}$$

and stationary distribution $\pi = (0.5, 0.5)$. Each of the two states has a particular variance-covariance structure, which is as follows.

The variance-covariance matrix for the first state will have a block-diagonal structure, which we construct as follows. Let $\mathbf{C} = 2\mathbf{I}_{20}$ be the diagonal matrix containing the variance for each dimension of the data. To construct the correlation structure of the data, first let

$$\mathbf{R}_1 = \begin{pmatrix} 1 & .3 & .3 & .3 & .3 \\ .3 & 1 & .3 & .3 & .3 \\ .3 & .3 & 1 & .3 & .3 \\ .3 & .3 & .3 & 1 & .3 \\ .3 & .3 & .3 & .3 & 1 \end{pmatrix},$$

which will serve as the correlation structure per block, and then construct the block-diagonal correlation matrix

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{R}_1 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{R}_1 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{R}_1 \end{pmatrix}.$$

The variance-covariance matrix for the first state is then $\boldsymbol{\Sigma}_1 = \mathbf{CRC}$.

The variance-covariance matrix of the data in the second state is a tridiagonal matrix with 1.0 on the diagonal and 0.15 on the first diagonal both above and below the main diagonal.

4.2 The Performance of the Oracle

To illustrate the behavior and the performance of the shrinkage estimator, we assume that the true S_t is known. Our goal here is to evaluate how well $\hat{\boldsymbol{\Sigma}}_k^s$ estimates $\boldsymbol{\Sigma}_k$ independent of the performance of the state reconstruction. To illustrate how the shrinkage estimator behaves, we show the distribution over the Monte Carlo samples of the shrinkage weights in Figure 1. The shrinkage estimator is a biased estimator, namely, biased towards the scaled identity matrix, and the shrinkage weight controls the amount of bias to introduce to the sample variance-covariance matrix in a manner such that mean-squared error is minimized. Note that, on the average, the shrinkage weight in State 2 is much larger than the shrinkage weight in State 1. This is because $\boldsymbol{\Sigma}_2$ is tridiagonal, whereas $\boldsymbol{\Sigma}_1$ is block-diagonal with

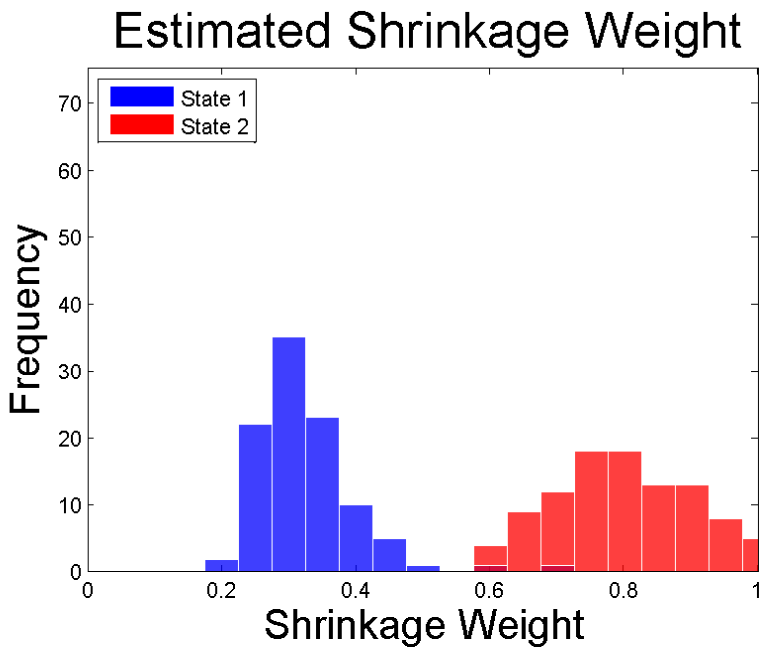

Figure 1: The distribution of shrinkage weights per state.

relatively large blocks, i.e., the closer the true variance-covariance matrix is to a scaled identity matrix, the larger the shrinkage weight.

We assessed performance of the estimators by looking at the percentage relative improvement in average loss (PRIAL), given by

$$\text{PRIAL}(\hat{\boldsymbol{\Sigma}}_k^s) = 100 \times \frac{\mathbb{E}||\boldsymbol{\Sigma}_k^{\text{o}} - \pi_k \boldsymbol{\Sigma}_k||^2 - \mathbb{E}||\hat{\boldsymbol{\Sigma}}_k^s - \pi_k \boldsymbol{\Sigma}_k||^2}{\mathbb{E}||\boldsymbol{\Sigma}_k^{\text{o}} - \pi_k \boldsymbol{\Sigma}_k||^2}. \tag{12}$$

We show the estimated PRIALs in Table 1, where the expectation is estimated by averaging over the Monte Carlo samples. The shrinkage estimator improves on the sample variance-covariance matrix by 24.490% in State 1, and in State 2, the shrinkage estimator improved on the sample variance-covariance matrix by 56.633%. The PRIALS behaved in parallel with the shrinkage weights. Indeed, whenever the shrinkage weight is equal to 0.0 so that $\hat{\boldsymbol{\Sigma}}_k^s = \boldsymbol{\Sigma}_k^{\text{o}}$, of course there is no improvement.

The numerical instability of high-dimensional variance-covariance matrices becomes a concern when we consider its inverse, also known as 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 use the PRIAL, similarly defined as that given by Equation (12) with all the matrices replaced by their inverses. The PRIALs for the shrinkage estimator when estimating the precision matrix yield a substantial improvement over the sample variance-covariance matrix for both states. This is very vital for evaluating the likelihood function. Finally, we point out that in some iterations of the simulation

study, the sample variance-covariance matrices $\hat{\mathbf{\Sigma}}_k$ were not invertible though the shrinkage estimates were always invertible, forcing us to restart some iterations. Thus, the PRIALs in estimating the precision matrix using the sample variance-covariance matrix that we report in Table 1 are smaller than they actually are.

State	Covariance Matrix	Precision Matrix
1	24.490	84.572
2	56.633	90.912

Table 1: PRIALs per state for the covariance matrix and the precision matrix when the true state is known.

To illustrate how much improvement we obtain in numerical stability of the estimated covariance matrix, we show the reduction in the condition number. The larger the condition number, the more numerically unstable the inversion of the matrix. In Figure 2 we see the condition number of the estimates per Monte Carlo iteration. We see that in all iterations of each simulation setting, the shrinkage estimator improves on the condition number of the sample covariance matrix. Moreover, the larger the condition number of the sample covariance matrix, the more substantial the reduction in the condition number of its shrunken estimate.

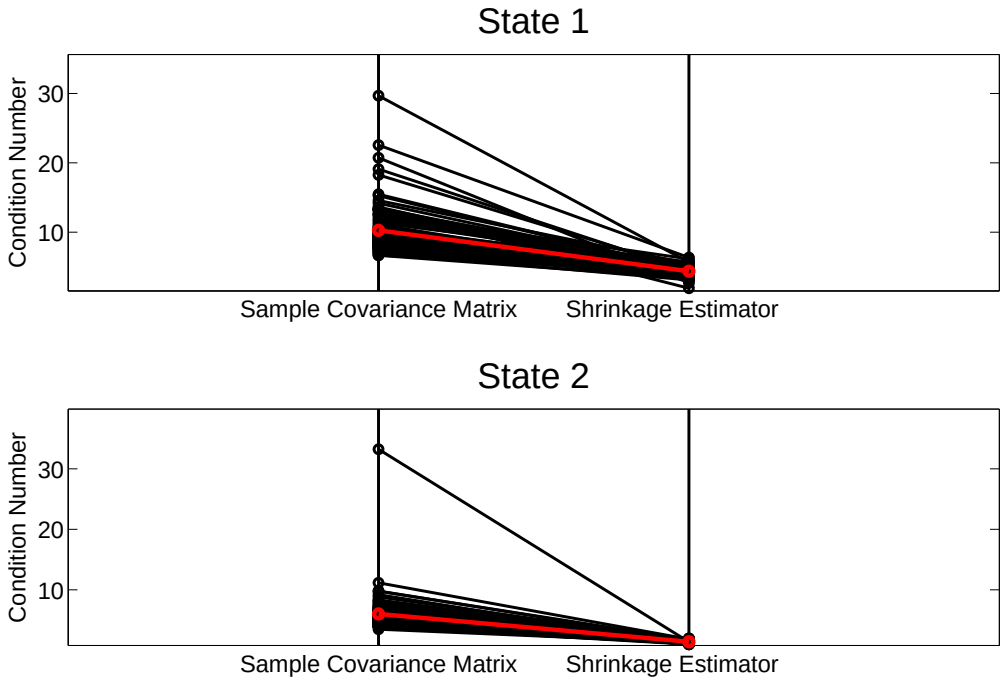

Figure 2: Comparison of condition numbers for each Monte Carlo iteration in Simulation 1 for (top) state 1 and (bottom) state 2. The mean decreasing trend is shown in red.

4.3 Transition Probabilities Estimation and Hidden Path Reconstruction

In practice, the states are not known and must be estimated. We now show that using the shrinkage estimator can yield much better reconstructions of the true state sequence using the Viterbi algorithm.

We show a representative result of a reconstructed state sequence in Figure 3. The performance of the reconstruction of the state sequence was clearly better if we used the shrinkage estimator. The oscillatory nature of the state sequence when we used the sample covariance matrix is due to the numerical instability of the sample variance-covariance matrix.

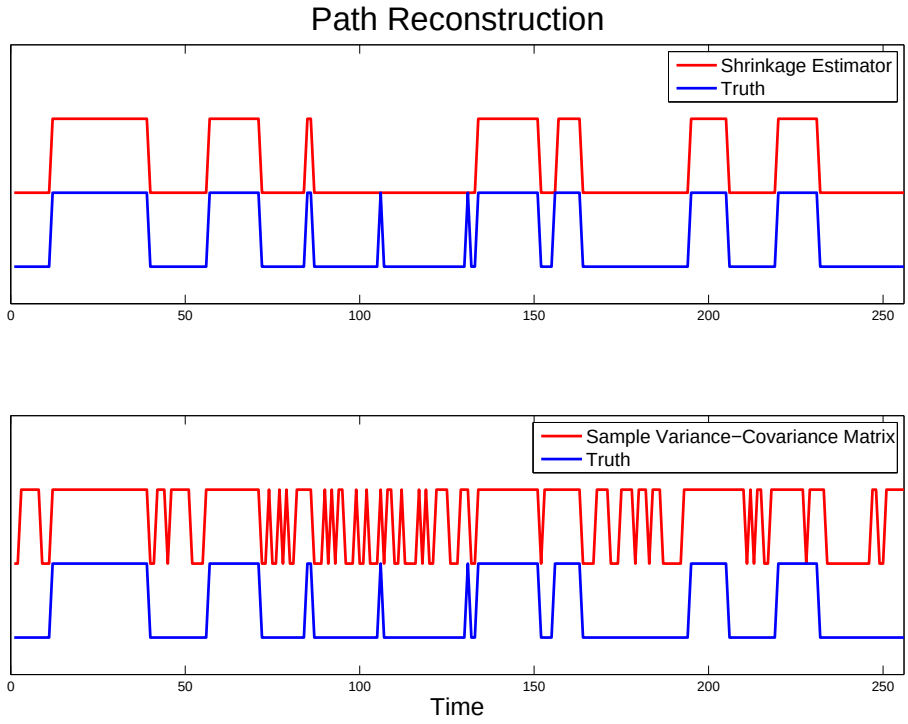

Figure 3: Reconstructed state sequence. (Top) Comparing the true state sequence with the estimated state sequence when the shrinkage estimator was used. (Bottom) Comparing the true state sequence with the estimated state sequence when the sample variance-covariance matrix was used.

We emphasize again that an estimate of the precision matrix is necessary for evaluating the likelihood function which is necessary for the hidden path reconstruction, and we thus need a numerically stable estimator of the covariance matrix. We have shown in this simulation study that the shrinkage estimator yields not only a more numerically stable estimator that is guaranteed to be invertible, but it also improves on the sample covariance matrix in terms of mean squared error.

K	1	2	3	4	5
AIC	-8914.3	-9354.8	-8366.8	-7462.5	-6448.4

Table 2: AIC values, used to decide the number of states in the model.

5 Analysis of the US Portfolio Data Set

Portfolio data sets are inherently high-dimensional. The present data set, a US industry portfolio data set publicly available[1], 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 in order to make the normality assumption more reasonable [19, 20], and the transformed log-returns were then mean-centered. We used our model to analyze this data set. The number of states K was picked so that AIC was minimized. Up to 5 states were considered and the AICs were computed and are given in Table 2, where we see that the model with $K = 2$ yielded the lowest AIC.

In Figure 4, we show the estimated correlation matrix for State 1 (upper triangle) and State 2 (lower triangle). Comparing the two states, we see that State 1 represents a state of greater volatility and stronger correlations in the industry portfolios. In particular, the fourth and fifth dimensions of the data, which corresponds to the games and recreation industry and printing and publishing industry, respectively, have larger correlations with the other industries in State 1 than in State 2. Moreover, dimensions 9-19, which correspond to the chemicals, textiles, construction, steel, machinery, electrical equipment, automobiles, transportation equipment, metal mining, oil, utility industries have portfolios which are correlated with one another. The latter three industries have stronger correlations within this block of industries in State 1 than in State 2. Note also that this block of industry portfolios is correlated with another block, namely dimensions 24-30, which correspond to the industries regarding business supplies, transportation, wholesale, retail, restaurants and hotels, banking and trading, and what the author of the data labeled as "other".

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

$$\widehat{\mathbf{A}} = \begin{pmatrix} 0.7203 & 0.2797 \\ 0.0593 & 0.9407 \end{pmatrix}.$$

This suggests that the industry portfolios are more likely to move and stay in the less volatile State 2 than State 1. In Figure 5, we can see that State 1 occurred during important events in the history of the US industry, such as the dot-com bubble collapse (early 2000s), 9/11, and the more recent financial crisis.

[1]The data set can be downloaded at
`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

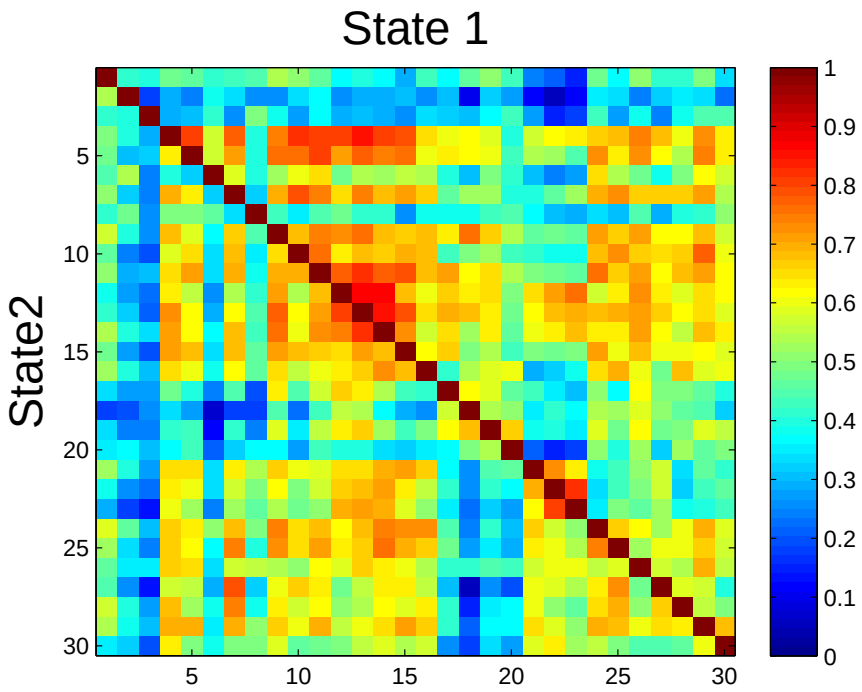

Figure 4: The correlation matrix for State 1 is in the upper triangle and for State 2 is in the lower triangle.

We also did the analysis without using our proposed shrinkage procedure. A comparison of the path reconstructions is shown in Figure 6. We see that if shrinkage was not used, then the reconstructed state sequence oscillates wildly between the two states. The estimated shrinkage weights were 0.010 for State 1 and 0.057 for State 2. Just as we had shown in our simulation study, because many of the dimensions of the data are correlated as shown in Figure 4, the shrinkage weight for each state is small. But, as we have just seen, even a small amount of regularization can make a substantial impact in the reconstruction of the hidden states.

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, e.g., Rabiner [16]. 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

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

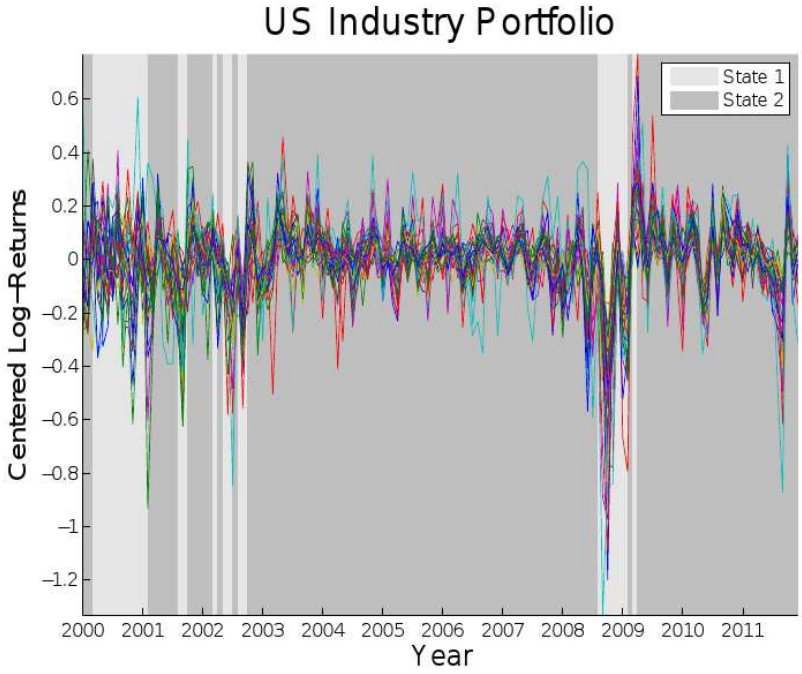

Figure 5: The centered log-returns is shown for each of the $P = 30$ industries. The state sequence, denoted by the two distinct shades of gray, is in the background.

i.e., $\delta_t(i)$ 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

$$\begin{aligned} \delta_{t+1}(j) &= \max_{S_1,\ldots,S_t} \log P(\, S_1, S_2, \ldots, S_{t+1,j} = 1, X_1, X_2, \ldots, X_{t+1}\,) \\ &= \max_{S_1,\ldots,S_t} \log\left[P(S_1, \ldots, S_{t,i} = 1, X_1, \ldots, X_t) P(S_{t+1,j} = 1 \mid S_{t,i} = 1) P(X_{t+1} \mid S_{t+1,j} = 1)\right], \end{aligned}$$

i.e.,

$$\delta_{t+1}(j) \;=\; \max_i(\, \delta_t(i) + \log a_{ij}\,) + \log f_j(X_{t+1}),$$

where $f_j(x)$ is given by (6).

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.

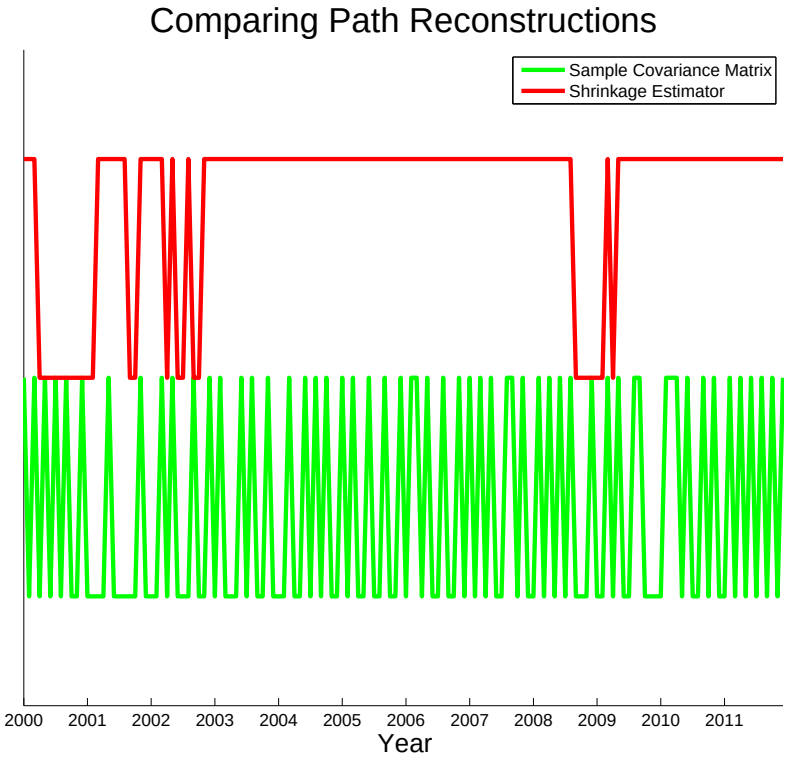

Figure 6: The reconstructed state sequence oscillates wildly between the two states if the sample variance-covariance matrix was used (green) as opposed to the shrinkage estimator.

Algorithm 2 Viterbi Algorithm

1. Initialization:

$$\delta_1(j) = \log \pi_j f_j(X_1) \quad j = 1, \ldots, K,$$
$$\psi_1(j) = 0,$$

2. Recursion:

$$\delta_t(j) = \max_{1 \le i \le K} (\, \delta_{t-1}(i) + \log a_{ij} \,) + \log f_j(X_t), \quad 2 \le t \le T, \quad j = 1, \ldots, K$$

$$\psi_t(j) = \arg \max_{1 \le i \le K} (\, \delta_{t-1}(i) + \log a_{ij} \,), \quad 2 \le t \le T, \quad j = 1, \ldots, K.$$

3. Termination:

$$q_T^* = \arg \max_{1 \le i \le K} \delta_T(i)$$

4. Path (State Sequence) Backtracking:

$$\begin{aligned} q_t^* &= \psi_{t+1}(\, q_{t+1}^* \,), \quad t = T-1, T-2, \ldots, 1, \\ S_{t,q_t^*} &= 1, \; S_{t,i} = 0, i \neq q_t^*, \quad t = 1, \ldots, T. \end{aligned}$$

References

[1] D. W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858, 1991.

[2] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41:164–171, 1970.

[3] H. Boehm and R. von Sachs. Shrinkage estimation in the frequency domain of multivariate time series. *Journal of Multivariate Analysis*, 100:913–935, 2008.

[4] Brockwell, P.J. and Davis, R.A. *Time Series: Theory and Methods, second edition.* Springer, New York, 1991.

[5] Cappé, O., Moulines, E., and Rydén, T. *Inference in hidden Markov models.* Springer Series in Statistics., New York, 2005.

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

[7] Douc, R., Moulines, É., and Rydén, T. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *Ann. Stat.*, 32:2254–2304, 2004.

[8] Fiecas, M. and Ombao, H. The generalized shrinkage estimator for the analysis of functional connectivity of brain signals . *Annals of Applied Statistics*, 5:1102–1125, 2011.

[9] Fiecas, M., Ombao, H., Linkletter, C., Thompson, W., and Sanes, J. Functional Connectivity: Shrinkage Estimation and Randomization Test. *NeuroImage*, 5:1102–1125, 2010.

[10] Francq, C. and Roussignol, M. On white noises driven by hidden markov chains. *Journal of Time Series Analysis*, 18:553–578, 1997.

[11] Francq, C. and Zarkoïan, J. -M. Stationarity of multivariate markov-switching arma models. *Journal of Econometrics*, 102:339–364, 2001.

[12] Hall, A.R. *Generalized Methods of Moments.* Oxford University Press, Oxford, 2005.

[13] Härdle, W. and Simar, L. *Applied Multivariate Statistical Analysis.* Springer, Germany, 2nd edition, 2007.

[14] Hartley, H. O. Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194, 1958.

[15] Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[16] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of the IEEE*, 77:257–286, 1989.

[17] Rydén, T., Terasvirta, T., and Asbrink, S. Stylized facts of daily return series and the hidden markov model of absolute returns. *Journal of Applied Econometrics*, 13:217–244, 1998.

[18] Sancetta, A. Sample covariance shrinkage for high dimensional dependent data. *Journal of Multivariate Analysis*, 99:949–967, 2008.

[19] Talih, M. and Hengartner, N. Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society, Series B*, 67:321–341, 2005.

[20] Xuan, X. and Murphy, K. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[21] Yang, M. Some properties of vector autoregressive processes with markov switching coefficients. *Econometric Theory*, 16:23–43, 2000.

Proofs and Technical Lemmas

To simplify notation, we sometimes use the following abbreviating notation in this section:

$$R_t = S_{t,k},\ \mathbf{Y}_t = S_{t,k}(X_t - \mu_k)(X_t - \mu_k)',$$

for any given fixed k. With a_{ij} denoting the transition probabilities matrix of the Markov chain $\{S_t\}$, the reduced Markov chain R_t has state space $\{0, 1\}$ and transition matrix

$$\mathbf{B} = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix} = \begin{pmatrix} 1 - \sum_{l \neq k} a_{lk} & \sum_{l \neq k} a_{lk} \\ 1 - a_{kk} & a_{kk} \end{pmatrix}$$

and $P(R_t = 1) = \pi_k$.

First we need the following auxiliary lemma.

Lemma 3 *For any $1 \leq k \leq K$, $1 \leq i, j \leq p$, the univariate time series $y_t = \mathbf{Y}_{t,ij} = S_{t,k}(X_{ti} - \mu_{k,i})(X_{tj} - \mu_{k,j})$ is stationary with autocovariances*

$$c_{ij,k}(n) = \mathrm{cov}(y_t, y_{t+n}) = \pi_k \mathbf{\Sigma}^2_{ij,k}\{(\mathbf{B}^n)_{11} - \pi_k\} \to 0$$

exponentially fast ($n \to \infty$), and having a continuous spectral density

$$f_{ij,k}(\omega) = \sum_{n=-\infty}^{\infty} c_{ij,k}(n) e^{in\omega}.$$

Proof: Let i, j, k be given and fixed. Writing $q_t = (\mathbf{\Sigma}_k^{1/2}\varepsilon_t)_i(\mathbf{\Sigma}_k^{1/2}\varepsilon_t)_j$, we have

$$y_t = R_t q_t.$$

The q_t are i.i.d. with $\mathbb{E}q_t = \mathbf{\Sigma}_{ij,k}$ and independent of R_s for all t, s. Therefore,

$$\begin{aligned} \mathbb{E}y_t y_{t+n} &= \mathbb{E}R_t R_{t+n} \mathbf{\Sigma}^2_{ij,k} \\ \mathbb{E}R_t R_{t+n} &= P(R_t = R_{t+n} = 1) \\ &= \pi_k P(R_{t+n} = 1 | R_t = 1) = \pi_k (\mathbf{B}^n)_{11} \end{aligned}$$

We get

$$\begin{aligned} c_{ij,k}(n) &= \mathbb{E}\ y_t y_{t+n} - \mathbb{E}y_t \mathbb{E}y_{t+n} \\ &= \pi_k \mathbf{\Sigma}^2_{ij,k}\{(\mathbf{B}^n)_{11} - \pi_k\} \end{aligned}$$

As $c_{ij,k}(n) = \mathbf{\Sigma}^2_{ij,k}\,\mathrm{cov}(S_{t,k}, S_{t+n,k})$, it decreases exponentially fast to 0 by b) i) in the proof of Lemma 4. Hence, the spectral density converges pointwise and is continuous. ∎

Our main result follows immediately from Theorem 1 of Sancetta [4] once we have checked that the conditions of that theorem are satisfied. We first formulate some auxiliary results.

Lemma 4 *Let $S_t, t \in \mathbb{Z}$, be stationary, aperiodic and irreducible. Then,*

a) $\{S_t\}$ is (γ, L^∞, ψ)-weakly dependent in the sense of Doukhan and Louhichi [3] with an exponentially decreasing sequence $\gamma = (\gamma_1, \gamma_2, \ldots)$ and $\psi(f,g,u,v) \leq 4||f||_\infty||g||_\infty$, i.e., more precisely,

$$|\operatorname{cov}(f(S_{t_1}, \ldots, S_{t_u}), g(S_{\tau_1}, \ldots, S_{\tau_v}))| \leq 4||f||_\infty||g||_\infty\gamma_r$$

for all $t_1 < \ldots < t_u \leq t_u + r \leq \tau_1 < \ldots < \tau_v$, $r, u, v \geq 1$, $f, g \in L^\infty$.

b) For some constant $c > 0$

$$|\operatorname{cov}(X_{t_1,i_1} \cdot \ldots \cdot X_{t_u,u_u}, X_{\tau_1,j_1} \cdot \ldots \cdot X_{\tau_v,j_v})| \leq c\ \gamma_r$$

for all $t_1 < \ldots < t_u \leq t_u + r \leq \tau_1 < \ldots < \tau_v$, $1 \leq u, v \leq 4$, $r \geq 1$.

Proof:

a) From Theorem 1 of Bradley [1] $\{S_t\}$ is α-mixing with exponentially decreasing mixing coefficients, say γ_r. Therefore, from section 3.2, Lemma 6 of Doukhan and Louhichi [3], the assertion a) follows.

b) i) Choosing $f(z_1, \ldots, z_v) = z_1 \cdot \ldots \cdot z_v$ for $|z_1|, \ldots, |z_v| \leq 1$, and 0 else and, correspondingly, $g(z_1, \ldots, z_u) = z_1 \cdot \ldots \cdot z_u$ for $|z_1|, \ldots, |z_v| \leq 1$, and 0 else, then $f, g \in L_\infty$, $||f||_\infty||g||_\infty \leq 1$, and we get from a)

$$|\operatorname{cov}(S_{t_1,k_1} \cdot \ldots \cdot S_{t_u,k_u}, S_{\tau_1,l_1} \cdot \ldots \cdot S_{\tau_v,l_v})| \leq 4\gamma_r$$

for all $t_1, < \ldots < t_u \leq t_u + r \leq \tau_1 < \ldots < \tau_v$, $1 \leq k_1, \ldots, k_u, l_1, \ldots, l_v \leq K$, $u, v, r \geq 1$.

ii) Let U_1, U_2, V_1, V_2 be real random variables such that the pair $(U_1, U_2), V_1$ and V_2 are independent. Then, a straightforward calculation shows that

$$\operatorname{cov}(U_1V_1, U_2V_2) = \operatorname{cov}(U_1, U_2)\ \mathbb{E}V_1\ \mathbb{E}V_2\ .$$

iii) Applying the preceding results i) and ii) on $X_{sm} = \sum_{k=1}^K S_{s,k}(\mathbf{\Sigma}_k^{1/2}\varepsilon_s)_m$, using bilinearity of the covariance and exploiting that the terms in the sum over k have all the corresponding factor structure as in ii), we have

$$|\operatorname{cov}(X_{t_1,i_1} \cdot \ldots \cdot X_{t_u,i_u}, X_{\tau_1,j_1} \cdot \ldots \cdot X_{\tau_v,j_v})| \leq$$
$$\sum_{\substack{k_1,\ldots,k_u,\\ l_1,\ldots,l_v=1}}^K 4\gamma_r\ \mathbb{E}\left\{\left(\mathbf{\Sigma}_{k_1}^{1/2}\varepsilon_{t_1}\right)_{i_1} \cdot \ldots \cdot \left(\mathbf{\Sigma}_{k_u}^{1/2}\varepsilon_{t_u}\right)_{i_u}\right\}\ \mathbb{E}\left\{\left(\mathbf{\Sigma}_{l_1}^{1/2}\varepsilon_{\tau_1}\right)_{j_1} \cdot \ldots \cdot \left(\mathbf{\Sigma}_{l_v}^{1/2}\varepsilon_{\tau_v}\right)_{j_v}\right\}$$

With this, assertion b) of the lemma follows as $K, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ are fixed and ε_t has bounded 4^{th} moments by assumption.

■

Proof of Theorem 1: We check the conditions of Theorem 1 of Sancetta [4] where we have to note that Sancetta uses the nonscaled version of the Frobenius norm. Condition 3 is just following from our conditions on $K(u)$ which are chosen to be slightly stronger for sake of convenience. Condition 2 follows from Lemma 4b). Condition 1 (1) from Lemma 1, b), as we shrink towards $\mathbf{F} = \alpha_k \mathbf{I}_p$. Condition 1 (2) follows immediately with $\beta = 1$ as $\mathbf{F}$ is diagonal in our case. Condition 1 (3), (4) follow from our assumptions A1) resp. A2). Finally, the stationarity condition in condition 4 follows from stationarity of S_t and the fact that X_t is a function of S_t and the uncorrelated random vectors ε_t; the moment condition of condition 4 follows from $\mathbb{E}||\varepsilon_t||^8 < \infty$ and the boundedness of S_t by a similar argument as in proving Lemma 4 b). ■

Proof of Lemma 1: We use the notation at the beginning of the section, i.e., in particular $R_t = S_{t,k}$ for fixed k and $b_{00} = P(R_{t+1} = 0 | R_t = 0)$.

a) By construction, writing $R_\cdot = \sum_{t=1}^T R_t$

$$\begin{aligned}
\mathbb{E}\ \mu_k^{\mathbf{o}} &= \mathbb{E}\ \mu_k^{\mathbf{o}} 1_{\{R_\cdot > 0\}} \\
&= \mathbb{E}\mathbb{E}\left(\frac{1_{\{R_\cdot > 0\}}}{R_\cdot} \sum_{t=1}^T R_t X_t \mid S_1, \ldots, S_T \right) = \mathbb{E}\left(\frac{1_{\{R_\cdot > 0\}}}{R_\cdot} \sum_{t=1}^T R_t \mathbb{E}\{X_t \mid S_t\} \right) \\
&= \mu_k P(R_\cdot > 0) = \mu_k \left(1 - P(R_1 = \ldots R_T = 0)\right) = \mu_k - \mu_k (1 - \pi_k) b_{00}^{T-1},
\end{aligned}$$

using $R_t \mathbb{E}\{X_t | S_t\} = R_t \mathbb{E}\{X_t | S_{t,k} = 1\} = R_t \mu_k$, as $R_t = S_{t,k}$ takes on the values 0 and 1 only, and $P(R_1 = 0) = P(S_{1,k} = 0) = 1 - \pi_k$. Note that, by our assumptions on the Markov chain, $0 < b_{00} < 1$. Choosing β as the maximum of those K values and applying (2), the first part of assertion a) follows.

Using the abbreviations $\eta_t = \mathbf{\Sigma}_k^{1/2} \varepsilon_t$ and

$$\delta_k = \mu_k^{\mathbf{o}} - \mu_k = \frac{1_{\{R_\cdot > 0\}}}{R_\cdot} \sum_{t=1}^T R_t \eta_t - \mu_k 1_{\{R_\cdot = 0\}}, \tag{13}$$

and remarking that $\eta_t, t = 1, \ldots, T$, are i.i.d. zero-mean random vectors, independent of the state variables $S_1, \ldots, S_T$, we have, as $\pi_k^{\mathbf{o}} = 0$ for $R_\cdot = 0$,

$$\begin{aligned}
\mathbb{E}\ \pi_k^{\mathbf{o}} ||\delta_k||^2 &= \mathbb{E}\ \pi_k^{\mathbf{o}} \left(\frac{1_{\{R_\cdot > 0\}}}{(R_\cdot)^2} \sum_{s,t=1}^T R_t R_s \mathbb{E}\{\eta_t' \eta_s \mid S_1, \ldots, S_T\} \right) \\
&= \frac{1}{T} \pi_k \mathbb{E} \eta_t' \eta_t = \frac{1}{T} \pi_k \operatorname{tr} \mathbf{\Sigma}_k = O\left(\frac{p}{T}\right).
\end{aligned}$$

as $\mathbb{E}\eta_t'\eta_t = \mathbb{E} \operatorname{tr} \eta_t \eta_t' = \operatorname{tr} \mathbf{\Sigma}_k = O(p)$. The second part of assertion a) follows.

Analogously, using

$$R_t\mathbb{E}\{X_t-\mu_k\,|\,S_1,\ldots,S_T\}=0,$$
$$R_t\mathbb{E}\{(X_t-\mu_k)(X_\tau-\mu_k)'\mid S_1,\ldots,S_T\}=R_t\boldsymbol{\Sigma}_k,\ \text{if } t=\tau, \qquad \text{and} \quad =0,\ \text{else},$$

and (13), we get after some lengthy but elementary calculations

$$\mathbb{E}\{\boldsymbol{\Sigma}_k^{\mathbf{o}}\mid S_1,\ldots,S_T\}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(R_t\mathbb{E}\{(X_t-\mu_k^{\mathbf{o}})(X_t-\mu_k^{\mathbf{o}})'\mid S_1,\ldots,S_T\}\right)$$
$$=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(R_t\mathbb{E}\{(X_t-\mu_k+\mu_k-\mu_k^{\mathbf{o}})(X_t-\mu_k+\mu_k-\mu_k^{\mathbf{o}})'\mid S_1,\ldots,S_T\}\right)$$
$$=\left(\pi_k^{\mathbf{o}}-\frac{1_{\{R.>0\}}}{T}\right)\boldsymbol{\Sigma}_k+\pi_k^{\mathbf{o}}\mu_k\mu_k'1_{\{R.=0\}}=\left(\pi_k^{\mathbf{o}}-\frac{1_{\{R.>0\}}}{T}\right)\boldsymbol{\Sigma}_k, \tag{14}$$

as, by definition, $\pi_k^{\mathbf{o}}1_{\{R.=0\}}=0$. Hence, as $\mathbb{E}\pi_k^{\mathbf{o}}=\pi_k$, and $\mathbb{E}1_{\{R.=0\}}=P(R.=0)=O\left(\beta^T\right)$ and as $||\boldsymbol{\Sigma}_k||$ is bounded by (2), we get $\mathbb{E}\ \boldsymbol{\Sigma}_k^{\mathbf{o}}=(\pi_k-\frac{1}{T})\boldsymbol{\Sigma}_k+O\left(\frac{\beta^T}{T}\right)$, and the third part of assertion a) follows.

b) Using a) and, due to assumption B1), $\operatorname{tr}\boldsymbol{\Sigma}_k=O(p)$, we get

$$\mathbb{E}\alpha_k^{\mathbf{o}}-\alpha_k=\frac{1}{p}\operatorname{tr}\mathbb{E}\boldsymbol{\Sigma}_k^{\mathbf{o}}-\alpha_k=\frac{1}{Tp}\operatorname{tr}\boldsymbol{\Sigma}_k+\frac{1}{Tp}O\left(\beta^T\right)=O\left(\frac{1}{T}\right).$$

and, consequently,

$$\mathbb{E}(\alpha_k^{\mathbf{o}}-\alpha_k)^2 = \frac{1}{p^2}\mathbb{E}\left(\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}-\pi_k\operatorname{tr}\boldsymbol{\Sigma}_k\right)^2$$
$$= \frac{1}{p^2}\operatorname{var}\left(\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}\right)+O(\frac{1}{T^2})$$

Therefore it suffices to show

$$\operatorname{var}\left(\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}\right)=\mathbb{E}\operatorname{var}\{\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}\mid S_1,\ldots,S_T\}+\operatorname{var}\left(\mathbb{E}\{\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}\mid S_1,\ldots,S_T\}\right)=O\left(\frac{p^2}{T}\right). \tag{15}$$

b1) From (14), we have

$$\operatorname{var}\left(\mathbb{E}\{\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}\mid S_1,\ldots,S_T\}\right)=\operatorname{var}\left(\pi_k^{\mathbf{o}}-\frac{1_{\{R.>0\}}}{T}\right)\left(\operatorname{tr}\boldsymbol{\Sigma}_k\right)^2$$

As $\pi_k^{\mathbf{o}}1_{\{R.>0\}}=\pi_k^{\mathbf{o}}$, $\mathbb{E}1_{\{R.>0\}}=1-P(R.=0)$, $\operatorname{var}1_{\{R.>0\}}=P(R.=0)\left(1-P(R.=0)\right)$ and $P(R.=0)=O(\beta^T)$ we get,

$$\operatorname{var}\left(\mathbb{E}\{\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}}\mid S_1,\ldots,S_T\}\right)=\left(\operatorname{tr}\boldsymbol{\Sigma}_k\right)^2\operatorname{var}\pi_k^{\mathbf{o}}+O\left(\frac{\beta^T}{T}\right)=O\left(\frac{p^2}{T}\right) \tag{16}$$

as $\operatorname{tr} \mathbf{\Sigma}_k = O(p)$ and $\operatorname{var} \pi_k^{\mathbf{o}} = \operatorname{var}\left(\frac{1}{T}\sum_{t=1}^{T} R_t\right) = O\left(\frac{1}{T}\right)$. The latter follows from Remark 1 to Theorem 7.1.1. of Brockwell and Davis [2], recalling the mixing assumption on $R_t = S_{t,k}$.

b2) For the first term of (15), we first remark that

$$\operatorname{tr} \mathbf{\Sigma}_k^{\mathbf{o}} = 0 \quad \text{if } S_{t,k} = R_t = 0, t = 1, \ldots, T. \tag{17}$$

Therefore, we assume for the moment that $R_\cdot > 0$ in calculating the conditional variance. Recalling again that $\eta_t, t = 1, \ldots, T$, are i.i.d. random vectors, independent of the state variables, with mean 0 and covariance matrix $\mathbf{\Sigma}_k$ we get, using (13),

$$\begin{aligned} \operatorname{tr} \mathbf{\Sigma}_k^{\mathbf{o}} &= \frac{1}{T}\sum_{t=1}^{T} R_t \operatorname{tr}(\eta_t - \delta_k)(\eta_t - \delta_k)' = \frac{1}{T}\sum_{t=1}^{T} R_t (\eta_t - \delta_k)'(\eta_t - \delta_k) \\ &= \frac{1}{T}\sum_{t=1}^{T} R_t \eta_t' \eta_t - \pi_k^{\mathbf{o}} \delta_k' \delta_k = \frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||^2 - \pi_k^{\mathbf{o}} ||\delta_k||^2. \end{aligned} \tag{18}$$

As $\eta_t, t = 1, \ldots, T$, are i.i.d. and independent of the state variables and as $R_t^2 = R_t$, we have

$$\begin{aligned} \operatorname{var}\left\{\frac{1}{T}\sum_{t=1}^{T} R_t ||\eta_t||^2 \mid S_1, \ldots, S_T\right\} &= \frac{1}{T^2}\sum_{t=1}^{T} R_t \operatorname{var}\left(||\eta_t||^2\right) \\ &\leq \frac{1}{T}\pi_k^{\mathbf{o}} \mathbb{E}||\eta_t||^4 = \pi_k^{\mathbf{o}} O\left(\frac{p^2}{T}\right), \end{aligned} \tag{19}$$

For the last relation, we use A0), recall $\operatorname{var} \varepsilon_{t,i} = 1$, and denote by ρ_{ij} the (i,j)$^{\text{th}}$ entry of $\mathbf{\Sigma}_k^{1/2}$ and by $\rho(k) = (\rho_{1k}, \ldots, \rho_{pk})$ the k$^{\text{th}}$ row of this matrix,

$$\begin{aligned} \mathbb{E}||\eta_t||^4 &= \mathbb{E}||\mathbf{\Sigma}_k^{1/2}\varepsilon_t||^4 \\ &\leq \kappa_\varepsilon \sum_k \sum_{i,j} \rho_{ik}^2 \rho_{jk}^2 + \sum_{i,j,l,k} \rho_{ik}^2 \rho_{jl}^2 + 2\sum_{i,j,l,k} \rho_{ik}\rho_{jk}\rho_{il}\rho_{jl} \\ &= \kappa_\varepsilon \sum_k ||\rho(k)||^4 + p^2 ||\mathbf{\Sigma}_k^{1/2}||^4 + 2\sum_{l,k} (\rho(k)\rho'(l))^2 \\ &\leq \kappa_\varepsilon \sum_{l,k} ||\rho(k)||^2 ||\rho(l)||^2 + p^2 ||\mathbf{\Sigma}_k^{1/2}||^4 + 2\sum_{l,k} ||\rho(k)||^2 ||\rho(l)||^2 \\ &= (\kappa_\varepsilon + 3) p^2 ||\mathbf{\Sigma}_k^{1/2}||^4 = O(p^2). \end{aligned} \tag{20}$$

For the last relation, we use that, by assumption B1), $||\mathbf{\Sigma}_k^{1/2}||^2 \leq 1 + ||\mathbf{\Sigma}_k||^2 = O(1)$ where the inequality follows from the representation of the Frobenius norm in terms of eigenvalues.

Similarly, we have

$$\operatorname{var}\left\{\pi_k^{\mathbf{o}}||\delta_k||^2 \mid S_1,\ldots,S_T\right\} = \frac{(\pi_k^{\mathbf{o}})^2}{R_\cdot^4}\sum_{t,s,u,v=1}^{T} R_t R_s R_u R_v \operatorname{cov}(\eta_t'\eta_s, \eta_u'\eta_v) = O\left(\frac{p^2}{T^2}\right) \quad (21)$$

by the following argument. As the η_t are i.i.d. with mean 0, the covariances vanish except for the cases $t=s=u=v$, $t=u\neq s=v$, $t=v\neq s=u$. Using $R_t^2 = R_t$, the contribution of the first case to (21) is

$$\frac{(\pi_k^{\mathbf{o}})^2}{R_\cdot^4}\sum_{t=1}^{T} R_t \operatorname{var}(||\eta_t||^2) = \frac{(\pi_k^{\mathbf{o}})^2}{R_\cdot^3}O(p^2) \leq \frac{1}{T^2}O(p^2),$$

by (19), the definition of $\pi_k^{\mathbf{o}}$ and using $R_\cdot \geq 1$, if it is not vanishing. The contribution of the second and third case to (21) is of the form, using that $\eta_t'\eta_s$ are i.i.d. and that all terms in the sum are nonnegative,

$$\begin{aligned}(\pi_k^{\mathbf{o}})^2\frac{1}{R_\cdot^4}\sum_{t\neq s} R_t R_s \operatorname{var}(\eta_t'\eta_s) &\leq \frac{1}{T^2R_\cdot^2}\sum_{t,s=1}^{T} R_t R_s \mathbb{E}(\eta_1'\eta_2)^2 \leq \frac{1}{T^2}\mathbb{E}(||\eta_1||^2||\eta_2||^2) \\ &= \frac{1}{T^2}\left(\mathbb{E}(||\eta_1||^2\right)^2 \leq \frac{1}{T^2}\mathbb{E}(||\eta_1||^4 = O\left(\frac{p^2}{T^2}\right),\end{aligned}$$

again by (19). From (18)-(21) and (17) we get

$$\mathbb{E}\operatorname{var}\left\{\operatorname{tr}\boldsymbol{\Sigma}_k^{\mathbf{o}} \mid S_1,\ldots,S_T\right\} = O\left(\frac{p^2}{T}\right),$$

which, together with (16), implies (15)

c) In b), we have shown the rate condition (15). Combining it with assumption B2), c) follows immediately.

■

The following Lemma shows that the effect of replacing the oracle estimate $\mu_k^{\mathbf{o}}$ of the mean μ_k by its unknown true value in the definition of $\boldsymbol{\Sigma}_k^{\mathbf{o}}$, i.e. considering

$$\boldsymbol{\Sigma}_k^{\mathbf{o}*} = \frac{1}{T}\sum_{t=1}^{T} S_{t,k}(X_t-\mu_k)(X_t-\mu_k)',$$

instead, is asymptotically neglible. Note that $\boldsymbol{\Sigma}_{ij,k}^{\mathbf{o}*}$ is the sample mean of the time series $y_t = S_{t,k}(X_{ti}-\mu_{ki})(X_{tj}-\mu_{kj})$ which, by the discussion in Section 2 and Lemma 3, is stationary and satisfies the usual regularity conditions, such that

$$\mathbb{E}\left(\boldsymbol{\Sigma}_{ij,k}^{\mathbf{o}*} - \pi_k\boldsymbol{\Sigma}_{ij,k}\right)^2 = \operatorname{var}\left(\boldsymbol{\Sigma}_{ij,k}^{\mathbf{o}*}\right) = O\left(\frac{1}{T}\right), \quad \text{such that} \quad \mathbb{E}||\boldsymbol{\Sigma}_k^{\mathbf{o}*} - \pi_k\boldsymbol{\Sigma}_k||^2 = O\left(\frac{p}{T}\right)$$

by our definition of the scaled Frobenius matrix.

Lemma 5 *Under the conditions of Lemma 1*

$$a) \qquad \mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}*} - \mathbf{\Sigma}_k^{\mathbf{o}}||^2 = O\left(\frac{p}{T^2}\right)$$

$$b) \qquad \mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}} - \pi_k \mathbf{\Sigma}_k||^2 = \mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}*} - \pi_k \mathbf{\Sigma}_k||^2 + O\left(\frac{p}{T^2}\right)$$

Proof: a) Using the notation of the previous proof, we have

$$\mathbf{\Sigma}_k^{\mathbf{o}*} = \frac{1}{T}\sum_{t=1}^{T} R_t \eta_t \eta_t',$$

and a straightforward calculation shows, recalling that $\pi_k^{\mathbf{o}} > 0$ iff $R_{\cdot} > 0$,

$$\mathbf{\Sigma}_k^{\mathbf{o}*} - \mathbf{\Sigma}_k^{\mathbf{o}} = \pi_k^{\mathbf{o}} \delta_k \delta_k' = \frac{\pi_k^{\mathbf{o}}}{R_{\cdot}^2} \sum_{t,s=1}^{T} R_t R_s \eta_t \eta_s'.$$

As $\eta_t = \Sigma_k^{1/2} \varepsilon_t$ and as $\varepsilon_t, t = 1, \ldots, T$, are independent with mean 0 and unit covariance matrix and independent of the state variables, we get for all $i, j = 1, \ldots, p$

$$\begin{aligned}
& \mathbb{E}\left\{\left(\mathbf{\Sigma}_{ij,k}^{\mathbf{o}*} - \mathbf{\Sigma}_{ij,k}^{\mathbf{o}}\right)^2 \mid S_1, \ldots, S_T\right\} \\
= \; & \frac{(\pi_k^{\mathbf{o}})^2}{R_{\cdot}^4} \sum_{t,s,u,v=1}^{T} R_t R_s R_u R_v \mathbb{E}\eta_{t,i}\eta_{s,j}\eta_{u,i}\eta_{v,j} \\
= \; & \frac{(\pi_k^{\mathbf{o}})^2}{R_{\cdot}^3}\left(\mathbb{E}\eta_{t,i}^2\eta_{t,j}^2 - 2\mathbf{\Sigma}_{ij,k}^2 - \mathbf{\Sigma}_{ii,k}\mathbf{\Sigma}_{jj,k}\right) + \frac{(\pi_k^{\mathbf{o}})^2}{R_{\cdot}^2}\left(2\mathbf{\Sigma}_{ij,k}^2 + \mathbf{\Sigma}_{ii,k}\mathbf{\Sigma}_{jj,k}\right) \\
\leq \; & \frac{1}{T^2}\left(\mathbb{E}\eta_{t,i}^2\eta_{t,j}^2 + 4\mathbf{\Sigma}_{ij,k}^2 + 2\mathbf{\Sigma}_{ii,k}\mathbf{\Sigma}_{jj,k}\right),
\end{aligned}$$

by the same kind of argument as in showing (21), using $\mathbb{E}\eta_{t,i}\eta_{t,j} = \mathbf{\Sigma}_{ij,k}$, the definition of $\pi_k^{\mathbf{o}}$ and $R_{\cdot} \geq 1$ if $\pi_k^{\mathbf{o}} > 0$. Therefore, we have

$$\mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}*} - \mathbf{\Sigma}_k^{\mathbf{o}}||^2 \leq \frac{1}{T^2 p}\sum_{i,j=1}^{p}\left(\mathbb{E}\eta_{t,i}^2\eta_{t,j}^2 + 4\mathbf{\Sigma}_{ij,k}^2 + 2\mathbf{\Sigma}_{ii,k}\mathbf{\Sigma}_{jj,k}\right) = O\left(\frac{p}{T^2}\right)$$

using $\mathbb{E}||\eta_t||^4 = O(p^2)$ by (20), and, by assumption B1),

$$\frac{1}{p}\sum_{i,j=1}^{p}\mathbf{\Sigma}_{ii,k}\mathbf{\Sigma}_{jj,k} = \frac{1}{p}(\operatorname{tr}\mathbf{\Sigma}_k)^2 = O(p), \quad \frac{1}{p}\sum_{i,j=1}^{p}\mathbf{\Sigma}_{ij,k}^2 = ||\mathbf{\Sigma}_k||^2 = O(1) \quad \text{by (2)}.$$

b) Similarly, as $\mathbb{E}\mathbf{\Sigma}_k^{\mathbf{o}*} = \pi_k \mathbf{\Sigma}_k$,

$$\begin{aligned}
\mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}} - \pi_k\mathbf{\Sigma}_k||^2 &= \mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}} - \mathbf{\Sigma}_k^{\mathbf{o}*}||^2 + \mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}*} - \pi_k\mathbf{\Sigma}_k||^2 - \frac{2}{T^2 p}\sum_{i,j=1}^{p}\operatorname{var}(\eta_{t,i}\eta_{t,j}) \\
&= O\left(\frac{p}{T^2}\right) + \mathbb{E}||\mathbf{\Sigma}_k^{\mathbf{o}*} - \pi_k\mathbf{\Sigma}_k||^2 + O\left(\frac{p}{T^2}\right)
\end{aligned}$$

by a) and, using (20), by

$$\frac{1}{p}\sum_{i,j=1}^{p}\mathrm{var}(\eta_{t,i}\eta_{t,j}) \leq \frac{1}{p}\sum_{i,j=1}^{p}\mathbb{E}(\eta_{t,i}^2\eta_{t,j}^2) = \frac{1}{p}\mathbb{E}||\eta_t||^4 = O(p).$$

■

References

[1] Bradley, R.C. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[2] Brockwell, P.J. and Davis, R.A. *Time Series: Theory and Methods, second edition.* Springer, New York, 1991.

[3] Doukhan, P., and Louhichi, S. A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84:313–342, 1999.

[4] Sancetta, A. Sample covariance shrinkage for high dimensional dependent data. *Journal of Multivariate Analysis*, 99:949–967, 2008.