Statistica Sinica **26** (2016), 1427-1451
doi:http://dx.doi.org/10.5705/ss.2014.265t

REGIME-SWITCHING FACTOR MODELS FOR HIGH-DIMENSIONAL TIME SERIES

Xialu Liu and Rong Chen

San Diego State University and Rutgers University

Abstract: We consider a factor model for high-dimensional time series with regime-switching dynamics. The switching is assumed to be driven by an unobserved Markov chain; the mean, factor loading matrix, and covariance matrix of the noise process are different among the regimes. The model is an extension of the traditional factor models for time series and provides flexibility in dealing with applications in which underlying states may be changing over time. We propose an iterative approach to estimating the loading space of each regime and clustering the data points, combining eigenanalysis and the Viterbi algorithm. The theoretical properties of the procedure are investigated. Simulation results and the analysis of a data example are presented.

Key words and phrases: Factor model, hidden Markov process, high-dimensional time series, nonstationary process, regime switch, Viterbi algorithm.

1. Introduction

Multivariate time series data are observed in such fields as finance, economics, and computational biology, and various models and methods, generalized from univariate cases, have been discussed in the literature. Vector ARMA models were first proposed by Quenouille (1957); the parameter estimation and model specification were investigated by Tiao and Box (1981), Tsay and Tiao (1983), Lütkepohl (1985), Tiao and Tsay (1989), and others. Models for nonstationary time series were also introduced, such as vector error correction models Engle and Granger (1987), multivariate stochastic variance models Harvey, Ruiz, and Shephard (1994) and MGARCH models Engle and Kroner (1995); Bauwens, Laurent, and Rombouts (2006). Lütkepohl (2005) provides a comprehensive introduction to the multivariate time series models and methods.

These models are often confronted with computational challenges, overparametrization, and overfitting issues when dealing with high-dimensional time series. Factor analysis is considered an effective way to alleviate these problems by dimension reduction, starting with Anderson (1963) and Priestley, Rao, and Tong (1974) who applied it to multivariate time series. In the last decades, much attention has been paid to the high-dimensional case. Chamberlain and Rothschild (1983) and Forni et al. (2000) studied the factor model consisting of

common factors and an idiosyncratic component with weak cross-sectional and serial dependence. Bai and Ng (2002) and Hallin and Liška (2007) proved that the number of factors can be estimated consistently and established the convergence rate of factor estimators. Peña and Box (1987) and Pan and Yao (2008) decomposed the time series into two parts, a latent factor process and a vector white noise process, in which strong cross-sectional dependence is allowed. Lam, Yao, and Bathia (2011) and Lam and Yao (2012) developed an approach that takes advantage of information from autocovariance matrices at nonzero lags via eigendecomposition to estimate the factor loading space, and they established the asymptotic properties as the dimension goes to infinity with sample size. This innovative method is applicable to nonstationary processes and processes with uncorrelated or endogenous regressors Chang, Guo, and Yao (2013).

Regime switching Hamilton (1989) has been introduced in different models, including threshold models Tong and Lim (1980); Tong (1983) and ARCH models Hamilton and Susmel (1994); Hamilton (1996), and has various applications in economics, including analyzing business cycles Kim and Nelson (1998), GNP Hansen (1992), interest rates Gray (1996) and monetary policy Bernanke and Gertler (2000); Sims and Zha (2006). Factor models with regime switching can be tracked back to Diebold and Rudebusch (1994). In this paper we generalize the factor models of Pan and Yao (2008), and introduce a factor model with an unobserved state variable switching between several regimes in which the mean, factor loadings, and covariance matrices of noise process are all different. By allowing these parameters to switch across regimes, one enhances flexibility in modeling multivariate time series, and provides an effective tool to distinguish and identify the dynamics over time.

For factor models, switching mechanisms can be found in many cases. For example, CAPM theory indicates that the expected market return is an important factor for the expected return of an asset, and it is expected that its impact (loadings) on any individual asset may be different depending on whether a stock market is volatile or stable. In economics, risk-free rate, unemployment, and economic growth are crucial factors of all economic activities and their performance indicators. Again, the loadings of these factors may vary under different fiscal policies (neutral, expansionary, or contractionary) or in different stages of the economic cycle (expansion, peak, contraction, or trough); see Kim and Nelson (1998).

In this paper, we develop an iterative algorithm for the estimation of model parameters and unobserved time-varying states based on eigendecomposition and the Viterbi algorithm. The theoretical properties of the estimators are investigated. As in Lam, Yao, and Bathia (2011), whose model is essentially a one-regime model in our case, the convergence rate of estimated loading space depends on the 'strength' of the state. We find that, with multiple states of different

'strength', the convergence rate of the loading space estimator for strong states is the same as the one-regime case, while the rate improves for weak states, gaining extra information from the strong states. Empirical results confirm such observations.

The rest of the paper is organized as follows. In Sections 2 and 3, detailed model setting and estimation procedure are introduced. The theoretical properties are investigated in Section 4. Simulation results are presented in Section 5 and a data example is analyzed in Section 6. All proofs are in the Supplementary Material.

2. Switching Factor Models

We introduce some notation. For any matrix $\mathbf{H}$, $\|\mathbf{H}\|_F$, and $\|\mathbf{H}\|_2$ denote the Frobenius and L-2 norms of $\mathbf{H}$; $\mathrm{tr}(\mathbf{H})$ and $\lambda_{\max}(\mathbf{H})$ are the trace and the largest nonzero eigenvalue of a square matrix $\mathbf{H}$, respectively, and $\|\mathbf{H}\|_{\min}$ is the square root of minimum nonzero eigenvalue of $\mathbf{H}'\mathbf{H}$. We write $a \asymp b$, if $a = O(b)$ and $b = O(a)$.

Let $\mathbf{y}_t$ be a $p \times 1$ observed time series and z_t be a homogenous and stationary hidden Markov chain taking values in $\{1, 2, \ldots, m\}$ with transition probabilities

$$\pi_{k,j} = P(z_{t+1} = j \mid z_t = k) \qquad k, j = 1, \cdots, m, \tag{2.1}$$

where the number of states m is known. We assume that, for $t = 1, \ldots, n$, when $z_t = k$,

$$\mathbf{y}_t = \boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)} \quad \text{and} \quad \boldsymbol{\varepsilon}_t^{(k)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{2.2}$$

where $\mathbf{x}_t$ is a $d \times 1$ latent factor process with d fixed and (much) smaller than p, independent of $\mathbf{z} = \{z_1, \ldots, z_n\}$, $\mathrm{E}(\mathbf{x}_t) = 0$. Here $\boldsymbol{\mu}_k$ is the mean of the process, $\mathbf{A}_k$ is the unknown loading matrix, and $\boldsymbol{\Sigma}_k$ is the covariance matrix of the noise process for state k. We assume that $\{\boldsymbol{\varepsilon}_t^{(1)}\}, \cdots, \{\boldsymbol{\varepsilon}_t^{(m)}\}$ are m uncorrelated white noise processes, independent of $\{(\mathbf{x}_t, z_t), t \in \mathbb{Z}\}$. Our model is a generalization of the factor models of Lam, Yao, and Bathia (2011). The dynamics of $\mathbf{y}_t$ are driven by the factor process $\mathbf{x}_t$ according to m states controlled by the state variable z_t.

As noted in Lam and Yao (2012), $\mathbf{A}_k$ is not uniquely defined since $(\mathbf{A}_k, \mathbf{x}_t)$ in (2.2) can be replaced by $(\mathbf{A}_k\mathbf{U}_k, \mathbf{U}_k^{-1}\mathbf{x}_t)$ for any $d \times d$ non-singular matrix $\mathbf{U}_k$. Denote the linear space spanned by the columns of a matrix $\mathbf{A}$ as $\mathcal{M}(\mathbf{A})$. It is easily seen that $\mathcal{M}(\mathbf{A}_k)$, the factor loading space for state k, is uniquely defined by (2.2). Hence, we can find a $p \times d$ matrix $\mathbf{Q}_k$ and a $d \times d$ non-singular matrix $\boldsymbol{\Gamma}_k$ satisfying

$$\mathbf{Q}_k'\mathbf{Q}_k = \mathbf{I}_d, \text{ and } \mathbf{A}_k = \mathbf{Q}_k\boldsymbol{\Gamma}_k, \quad k = 1, \cdots, m. \tag{2.3}$$

It follows that $\mathcal{M}(\mathbf{Q}_k) = \mathcal{M}(\mathbf{A}_k)$. The columns of $\mathbf{Q}_k$ are d orthonormal vectors, and the column space spanned by $\mathbf{Q}_k$ is the same as the column space spanned

by $\mathbf{A}_k$. In addition, let $\mathbf{B}_k = (\mathbf{b}_{k,1}, \ldots, \mathbf{b}_{k,p-d})$ be an orthnormal basis such that $\mathcal{M}(\mathbf{B}_k)$ is the orthogonal complement space of $\mathcal{M}(\mathbf{Q}_k)$. Hence $(\mathbf{Q}_k, \mathbf{B}_k)$ forms a $p \times p$ matrix with orthogonal columns, $\mathbf{Q}_k'\mathbf{B}_k = \mathbf{0}$, and $\mathbf{B}_k'\mathbf{B}_k = \mathbf{I}_{p-d}$. In practice,

$$\mathbf{A}_k'\mathbf{B}_k = \mathbf{0}. \tag{2.4}$$

Assume that the loading spaces are different across regimes, our goal is to cluster the data by regimes, and estimate d and $\mathcal{M}(\mathbf{Q}_k)$, for $k = 1, \ldots, m$.

Remark 1. The $(\mathbf{A}_k,\ \mathbf{x}_t)$ in (2.2) can be replaced by $(\mathbf{A}_k\mathbf{U}_k,\ \mathbf{U}_k^{-1}\mathbf{x}_t)$ for any $d \times d$ non-singular matrix $\mathbf{U}_k$. Hence, the factor process may not be stationary after such nonsingular transformations across regimes, if $\{\mathbf{U}_k, k = 1, \ldots, m\}$ are different. However, it does not directly affect the underlying process or the estimation procedure since we do not impose the stationarity on the latent process $\mathbf{x}_t$.

For factor models in high-dimensional cases, it is common to assume that the squared L-2 norm of the $p \times d$ loading matrix grows with the dimension p Bai and Ng (2002); Doz, Giannone, and Reichlin (2011), with the growth rate defined as the strength of the factors in Lam, Yao, and Bathia (2011). In our multi-regime factor model in (2.2), the strength of the factors may be different across regimes. Assume that

$$\|\mathbf{A}_k\|_2^2 \asymp \|\mathbf{A}_k\|_{\min}^2 \asymp p^{1-\delta_k}, \qquad 0 \le \delta_k \le 1,$$

where $\|\mathbf{A}_k\|_{\min}^2$ is the minimum nonzero eigenvalue of $\mathbf{A}_k'\mathbf{A}_k$. If $\delta_k = 0$, the factors are 'strong' for state k and we call state k a *strong state* and $\mathbf{A}_k$ a dense loading matrix. If $\delta_k > 0$, the factors are 'weak' for state k and we call state k a *weak state* and $\mathbf{A}_k$ a sparse loading matrix. The strength of the state is an indicator of signal-to-noise ratio. It measures the relative growth rate of the amount of information which the observed process $\mathbf{y}_t$ carries about the common factors $\mathbf{x}_t$ as p increases, with respect to the growth rate of the amount of noise process. When the state is weak, the information contained in $\mathbf{y}_t$ about the factors grows more slowly than the noises introduced as p increases, hence the proportion of information is diluted by the noise. When the state is strong, the signal-to-noise ratio remains constant.

Setting

$$\mathbf{R}_t = \sum_{k=1}^{m} \mathbf{\Gamma}_k \mathbf{x}_t I(z_t = k), \tag{2.5}$$

the switching factor model can be written as

$$\mathbf{y}_t = \sum_{k=1}^{m} I(z_t = k)\left(\boldsymbol{\mu}_k + \mathbf{A}_k\mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)}\right) = \sum_{k=1}^{m} I(z_t = k)\left(\boldsymbol{\mu}_k + \mathbf{Q}_k\mathbf{\Gamma}_k\mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)}\right) \tag{2.6}$$

$$= \sum_{k=1}^{m} I(z_t = k) \left(\boldsymbol{\mu}_k + \mathbf{Q}_k \mathbf{R}_t + \boldsymbol{\varepsilon}_t^{(k)} \right). \quad (2.7)$$

This reveals different ways to decompose the dynamic part of the process. In (2.6), $\mathbf{Q}_k$ is the standardized loadings, $\mathbf{x}_t$ is the factor latent process, and $\boldsymbol{\Gamma}_k$ reflects the strength of the state. When the dynamic part is divided as in (2.7), $\mathbf{R}_t$ can be regarded as another latent factor process but with standardized loadings, and the L-2 norm of its variance matrix increases with p at different rates across regimes.

3. Estimation Procedure

In Section 3.1 we introduce a method that takes advantage of the autocovariance matrices to estimate the loading spaces when the state variable $\mathbf{z}$ is known; in Section 3.2 we propose a method using the Viterbi algorithm to estimate the hidden state variable when the loading spaces are known. Combining the two methods, we propose an iterative algorithm to estimate all the model parameters in Section 3.3.

3.1. Estimation of $\mathbf{B}_k$, $\boldsymbol{\mu}_k$, d and the transition probabilities given state indicator z

If the states $z_1, \ldots, z_n$ are given, transition probabilities can be estimated by

$$\hat{\pi}_{k,j} = \frac{\sum_{t=1}^{n-1} I(z_t = k, z_{t+1} = j)}{\sum_{t=1}^{n-1} I(z_t = k)}, \quad \text{for} \quad k, j = 1, \ldots, m,$$

and

$$\hat{\pi}_k = \frac{\sum_{t=1}^{n} I(z_t = k)}{n}, \quad \text{for} \quad k = 1, \ldots, m.$$

For the estimation of factor loading spaces, we adopt the procedure proposed by Lam, Yao, and Bathia (2011), Lam and Yao (2012), and Chang, Guo, and Yao (2013). It is based on the observation that, since the idiosyncratic noise $\boldsymbol{\varepsilon}_t^{(k)}$ is white, the dynamics of $\mathbf{y}_t$ (autocovariance) only come from the dynamics of the factor $\mathbf{x}_t$. Hence we can retrieve the factor loading space through an analysis of the autocovariance structure of $\mathbf{y}_t$. If

$$\boldsymbol{\Sigma}_x(l) = \frac{1}{n-l} \sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_{t+l}),$$

$$\boldsymbol{\Sigma}_{y,k}(l) = \frac{1}{n-l} \sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{y}_t, \mathbf{y}_{t+l} \mid z_t = k),$$

it follows that

$$\begin{aligned}\boldsymbol{\Sigma}_{y,k}(l) &= \frac{1}{n-l}\sum_{t=1}^{n-l}\sum_{j=1}^{m}\pi_{k,j}^{(l)}\mathrm{Cov}(\mathbf{y}_t,\, \mathbf{y}_{t+l}I(z_{t+l}=j) \mid z_t=k)\\ &= \mathbf{A}_k\boldsymbol{\Sigma}_x(l)\sum_{j=1}^{m}\pi_{k,j}^{(l)}\mathbf{A}_j', \end{aligned} \tag{3.1}$$

where $\pi_{k,j}^{(l)} = P(z_{t+l}=j \mid z_t=k)$, the transition probability from state k to state j in l steps. If $\mathbf{x}_t$ is stationary, then $\boldsymbol{\Sigma}_x(l)$ is the autocovariance matrix of $\mathbf{x}_t$ at lead l.

For a fixed prescribed integer l_0, define

$$\mathbf{M}_k = \sum_{l=1}^{l_0}\mathbf{M}_{k,l}, \tag{3.2}$$

where $\mathbf{M}_{k,l} = \boldsymbol{\Sigma}_{y,k}(l)\boldsymbol{\Sigma}_{y,k}(l)'$ is a quadratic version of the autocovariance matrix $\boldsymbol{\Sigma}_{y,k}(l)$. Because of (2.4) and (3.1), we have $\mathbf{M}_{k,l}\mathbf{B}_k = \mathbf{0}$ for all k. If $\sum_{j=1}^{m}\pi_{k,j}^{(l)}\mathbf{A}_j'$ is of full rank, $\mathbf{M}_k$ is a non-negative definite matrix sandwiched by $\mathbf{A}_k$ and $\mathbf{A}_k'$ with rank d. Then the d unit eigenvectors of $\mathbf{M}_k$ corresponding to its d non-zero eigenvalues form the space $\mathcal{M}(\mathbf{A}_k)$, the space spanned by the columns of $\mathbf{A}_k$.

We write the sample versions of these statistics, given $\mathbf{z} = \{z_1, \ldots, z_n\}$, for $k = 1, \ldots, m$,

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{y,k}(l) &= \frac{\sum_{t=1}^{n-l}\sum_{j=1}^{m}(\mathbf{y}_t - \hat{\boldsymbol{\mu}}_k)(\mathbf{y}_{t+l} - \hat{\boldsymbol{\mu}}_j)'I(z_t=k, z_{t+l}=j)}{\sum_{t=1}^{n-l}I(z_t=k)},\\ \hat{\boldsymbol{\mu}}_k &= \frac{\sum_{t=1}^{n}\mathbf{y}_t I(z_t=k)}{\sum_{t=1}^{n}I(z_t=k)},\\ \hat{\mathbf{M}}_{k,l} &= \hat{\boldsymbol{\Sigma}}_{y,k}(l)\hat{\boldsymbol{\Sigma}}_{y,k}(l)', \quad \hat{\mathbf{M}}_k = \sum_{l=1}^{l_0}\hat{\mathbf{M}}_{k,l}. \end{aligned} \tag{3.3}$$

Let $\hat{\lambda}_{k,1} \geq \hat{\lambda}_{k,2} \geq \ldots \geq \hat{\lambda}_{k,p}$ be the p eigenvalues of $\hat{\mathbf{M}}_k$ and $\hat{\mathbf{q}}_{k,1}, \ldots, \hat{\mathbf{q}}_{k,p}$ be the set of corresponding orthonormal eigenvectors. If

$$\hat{\mathbf{Q}}_k = (\hat{\mathbf{q}}_{k,1}, \ldots, \hat{\mathbf{q}}_{k,d}), \qquad \text{and} \quad \hat{\mathbf{B}}_k = (\hat{\mathbf{q}}_{k,d+1}, \ldots, \hat{\mathbf{q}}_{k,p}), \tag{3.4}$$

then $\mathcal{M}(\mathbf{Q}_k)$ and $\mathcal{M}(\mathbf{B}_k)$ can be estimated by $\mathcal{M}(\hat{\mathbf{Q}}_k)$ and $\mathcal{M}(\hat{\mathbf{B}}_k)$, respectively. To estimate the number of factors with data in each regime, we use the eigenvalue-ratio method of Lam and Yao (2012). Specifically, let

$$\hat{d}_k = \underset{1\leq j\leq c}{\operatorname{argmin}}\,\frac{\hat{\lambda}_{k,j+1}}{\hat{\lambda}_{k,j}}. \tag{3.5}$$

We set c to $p/2$, since the minimum eigenvalues of $\hat{\mathbf{M}}_k$ may be close to 0, especially when n is small and p is large; see Lam and Yao (2012).

Corollary 1 in Section 4 shows that under some mild conditions, $\hat{d}_1, \ldots, \hat{d}_m$ are all reasonable estimates of the number of factors d. Since d is common to all regimes, we choose the one from the strongest state, as the theoretical results show that the estimated nonzero eigenvalues from a stronger state have a faster convergence rate. Hence, we use $\hat{d} = \hat{d}_{\tilde{k}}$ to estimate d, where $\tilde{k} = \operatorname{argmax} \|\hat{\mathbf{M}}_k\|_2$.

Let $\mathbf{f}_t$ be the dynamic part of $\mathbf{y}_t$, $\mathbf{f}_t = \sum_{k=1}^m \mathbf{A}_k \mathbf{x}_t I(z_t = k)$. Since the column space of $\mathbf{A}_k$ is identifiable only up to a nonsingular transformation across regimes, we cannot recover $\mathbf{x}_t$ directly, but we have natural estimators

$$\hat{\mathbf{R}}_t = \sum_{k=1}^m \hat{\mathbf{Q}}_k'(\mathbf{y}_t - \hat{\boldsymbol{\mu}}_k) I(z_t = k), \quad \hat{\mathbf{f}}_t = \sum_{k=1}^m \hat{\mathbf{Q}}_k \hat{\mathbf{Q}}_k'(\mathbf{y}_t - \hat{\boldsymbol{\mu}}_k) I(z_t = k), \tag{3.6}$$

with the residuals

$$\hat{\boldsymbol{\varepsilon}}_t = \sum_{k=1}^m (\mathbf{I}_p - \hat{\mathbf{Q}}_k \hat{\mathbf{Q}}_k')(\mathbf{y}_t - \hat{\boldsymbol{\mu}}_k) I(z_t = k). \tag{3.7}$$

Remark 2. Our method works under weaker assumptions that dependence between $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_s$ when $t > s$ is allowed. If $\mathrm{E}(\boldsymbol{\varepsilon}_s^{(k)} \mathbf{x}_t') = 0$ only for $t \le s$, we can still follow the same procedure to estimate $\mathcal{M}(\mathbf{Q}_k)$, but take

$$\boldsymbol{\Sigma}_{y,k}(l) = \frac{1}{n-l} \sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{y}_{t+l},\, \mathbf{y}_t \mid z_{t+l} = k)$$

and $\mathbf{M}_k = \sum_{l=1}^{l_0} \boldsymbol{\Sigma}_{y,k}(l) \boldsymbol{\Sigma}_{y,k}(l)'$. Here we assume that $\{\boldsymbol{\varepsilon}_t^{(k)}, t \in \mathbb{Z}\}$ and $\{\mathbf{x}_t, t \in \mathbb{Z}\}$ for all $k = 1, \ldots, m$, are independent for simplicity.

Remark 3. This estimation procedure has been used for one-regime factor models with stationary processes in Tao et al. (2011) and Lam, Yao, and Bathia (2011), and with nonstationary processes in Chang, Guo, and Yao (2013). Many numerical results show that the estimation of the loading space is not sensitive to the choice of l_0; see Lam, Yao, and Bathia (2011), Lam and Yao (2012), and Chang, Guo, and Yao (2013). Although the estimator works with any $l_0 \ge 1$ both theoretically and numerically, the extra terms in $\mathbf{M}_k$ of (3.2) are very useful when the sample size is small and the variability in the estimation of the autocovariance matrices is large. Nevertheless, as the autocorrelation is often at its strongest at small time lags, a relatively small l_0 is usually adopted.

3.2. Estimation of the hidden state z given loading spaces and other model parameters

Although $\{\mathbf{B}_k, k = 1, \ldots, m\}$ are only uniquely identifiable up to orthogonal transformations, the density function of $\mathbf{B}_{z_t}' \mathbf{y}_t$ is invariant to such transformations. Thus, given $\pi_{k,j}$, π_k, $\boldsymbol{\mu}_k$, and $\mathbf{B}_k$, $k, j = 1, \ldots, m$, the state variables

$z_1, \ldots, z_n$ can be estimated by maximizing $G_n(\mathbf{z})$, the logarithm of the probability density function of $\{\mathbf{B}'_{z_t}\mathbf{y}_t, t = 1, \ldots, n\}$. Specifically, under the assumption that the noise process is normally distributed,

$$G_n(\mathbf{z}) = \log\left(\pi_{z_1} f(\mathbf{B}'_{z_1}\mathbf{y}_1)\right) + \sum_{t=2}^{n} \log\left(\pi_{z_{t-1},z_t} f(\mathbf{B}'_{z_t}\mathbf{y}_t)\right), \tag{3.8}$$

where

$$f(\mathbf{B}'_{z_t}\mathbf{y}_t) = -\frac{1}{\sqrt{(2\pi)^{p-d}|\mathbf{\Sigma}_{B,z_t}|}} - \exp\left[-\frac{(\mathbf{B}'_{z_t}(\mathbf{y}_t - \boldsymbol{\mu}_{z_t}))'\mathbf{\Sigma}^{-1}_{B,z_t}\mathbf{B}'_{z_t}(\mathbf{y}_t - \boldsymbol{\mu}_{z_t})}{2}\right]. \tag{3.9}$$

Here $G_n(\mathbf{z})$ is a sum of n functions in the form of

$$G_n(\mathbf{z}) = g_1(z_1) + \sum_{t=2}^{n} g_t(z_{t-1}, z_t),$$

due to the Markovian structure of $\mathbf{z}$, where

$$g_1(z_1) = \log\left(\pi_{z_1} f(\mathbf{B}'_{z_1}\mathbf{y}_1)\right),$$

and

$$g_t(z_{t-1}, z_t) = \log\left(\pi_{z_{t-1},z_t} f(\mathbf{B}'_{z_t}\mathbf{y}_t)\right), \quad \text{for } t = 2, \ldots, n.$$

Hence $G_n(\mathbf{z})$ can be maximized by the Viterbi algorithm Viterbi (1967) and Forney (1973). The maximizer of the state sequence $z_1, \ldots, z_n$ is given by the recurrence relations,

$$\begin{aligned} S_{1,k} &= k, \\ x_{t,k} &= \operatorname*{argmax}_{1\le j\le m} \left[g_t(z_{t-1} = j, z_t = k) + G_{t-1}(S_{t-1,j})\right], \\ \mathbf{S}_{t,k} &= (\mathbf{S}_{t-1,x_{t,k}}, k), \end{aligned}$$

where $\mathbf{S}_{t,k}$ is a $t \times 1$ vector and the maximizer of $G_t(z_1, \ldots, z_{t-1}, z_t = k)$ for the first t observations that has k as its final state. In each iteration there are m evaluations $g_t(\cdot, k)$ to update $G_t(\mathbf{S}_{t,k})$ for $k = 1, \ldots, m$, and m possible paths $\{\mathbf{S}_{t,1}, \ldots, \mathbf{S}_{t,m}\}$ to be compared. So the complexity of Viterbi algorithm is $O(m^2 n)$. The state variable can be estimated by

$$\hat{\mathbf{z}} = \operatorname*{argmax}_{1\le k\le m} G_n(\mathbf{S}_{n,k}).$$

The covariance matrix of $\mathbf{B}'_k\mathbf{y}_t$ given $z_t = k$, $\mathbf{\Sigma}_{B,k} = \mathbf{B}'_k\mathbf{\Sigma}_k\mathbf{B}_k$ can be estimated by

$$\hat{\mathbf{\Sigma}}_{B,k} = \sum_{t=1}^{n} \mathbf{B}'_k(\mathbf{y}_t - \boldsymbol{\mu}_k)(\mathbf{y}_t - \boldsymbol{\mu}_k)'\mathbf{B}_k I(\hat{z}_t = k) / [\sum_{t=1}^{n} I(\hat{z}_t = k) - 1]. \tag{3.10}$$

Remark 4. One would prefer to construct the density of $\mathbf{y}_t$ given z_t with (3.7). However, since $\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}_k' = \mathbf{B}_k\mathbf{B}_k'$, it follows that

$$(\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}_k')(\mathbf{y}_t - \boldsymbol{\mu}_k) \sim N(\mathbf{0}, \mathbf{B}_k\mathbf{B}_k'\boldsymbol{\Sigma}_k\mathbf{B}_k\mathbf{B}_k').$$

This p-variate normal distribution lives on a $(p-d)$-dimensional space, while

$$\mathbf{B}_k'(\mathbf{y}_t - \boldsymbol{\mu}_k) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{B,k}), \text{ where } \boldsymbol{\Sigma}_{B,k} = \mathbf{B}_k'\boldsymbol{\Sigma}_k\mathbf{B}_k, \tag{3.11}$$

is a non-degenerated representation of the distribution in (3.11) restricted on the $(p-d)$-dimensional space.

Remark 5. There are several advantages to using the density function of $\{\mathbf{B}_{z_t}'\mathbf{y}_t, t = 1, \ldots, n\}$, instead of $\{\mathbf{y}_t, t = 1, \ldots, n\}$. First, we do not need to estimate $\mathbf{A}_k$, since we can only estimate the space it spans. Second, in order to compute the density of $\mathbf{y}_t$, we would need to to assume a specific model for the latent process $\mathbf{x}_t$. Although there is a vast literature in dynamic factor models Forni et al. (2000); Bai and Ng (2002); Hallin and Liška (2007), here we choose to avoid the difficulty.

3.3. An iterative algorithm

We adopt the distance measure used in Chang, Guo, and Yao (2013). For any $p \times d_1$ orthonromal matrix $\mathbf{H}_1$ and $p \times d_2$ orthonormal $\mathbf{H}_2$,

$$\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) = \Big\{1 - \frac{1}{\max\{d_1, d_2\}}\mathrm{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2')\Big\}^{1/2}. \tag{3.12}$$

Here $\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) \in [0,1]$, $\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) = 0$ if and only if $d_1 = d_2$, $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$, and $\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) = 1$ if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$. We are ready to state the algorithm.

Step 1. Begin with some initial values $\hat{\mathbf{z}}$.

Step 2. Given $\hat{\mathbf{z}}$, obtain $\hat{d}$, $\hat{\pi}_k$, $\hat{\pi}_{j,k}$, $\hat{\boldsymbol{\mu}}_k$, $\hat{\mathbf{Q}}_k$, and $\hat{\mathbf{B}}_k$ based on the methods in Section 3.1, for $j, k = 1, \ldots, m$.

Step 3. Given the estimates obtained in Step 2, estimate $\mathbf{z}$ by maximizing $G_n(\mathbf{z})$ using the Viterbi algorithm in Section 3.2.

Step 4. Repeat Step 2 and Step 3 until either a maximum number of iterations is reached, or

$$\frac{1}{m}\sum_{k=1}^{m}\mathcal{D}(\hat{\mathbf{Q}}_k^{(1)}, \hat{\mathbf{Q}}_k^{(2)}) < c_1, \quad \text{and} \quad \Big|\frac{G_n(\hat{\mathbf{z}}^{(1)}) - G_n(\hat{\mathbf{z}}^{(2)})}{G_n(\hat{\mathbf{z}}^{(1)})}\Big| < c_2,$$

where $c_1, c_2 \in (0,1)$ are prescribed small constants, and $\hat{\mathbf{Q}}_k^{(1)}$, $\hat{\mathbf{Q}}_k^{(2)}$, $\hat{\mathbf{z}}^{(1)}$, and $\hat{\mathbf{z}}^{(2)}$ are successive estimates for $\mathbf{Q}_k$ and $\mathbf{z}$, respectively.

Remark 6. Since the two iterative steps do not minimize the same objective function, the algorithm is not guaranteed to reach a fixed point solution given a finite sample. However, the objectives of the two steps are consistent. In Step 2, we try to extract common factors when estimating the loading spaces, hence try to reduce the remainder error terms in the factor models. In Step 3, maximizing of the density function is equivalent to minimizing the errors in the factor models for normally distributed errors. Such issues are common to estimation procedures of dynamic factor models, and an iterative algorithm is widely used Watson and Engle (1983); Stock and Watson (2005); Doz, Giannone, and Reichlin (2011).

Remark 7. We estimate the state variables instead of the transition probabilities in Step 3, because finding the maximizer of $\{\pi_{k,j}\}$ of density function of $\{\mathbf{B}'_{z_t}\mathbf{y}_t\}$ is more computationally expensive than the estimation of $\mathbf{z}$ by the Viterbi algorithm due to the dependence of the state variables. Although misclassification occurs because of the nature of hard clustering Kearns, Mansour, and Ng (1998); Hastie, Tibshirani, and Friedman (2009), this does not have much influence on the estimation of the loading spaces, since it often occurs when the data points lie near the intersection of the loading spaces. Numerical results show that our algorithm is able to cluster the data by regimes efficiently, and can estimate the loading spaces effectively. To obtain consistent estimators of transition probabilities and avoid misclassification, a discarding algorithm can be applied Chen (1995), in which only the data points that can be 'clearly separated' by the objective function, e.g. $f(\mathbf{B}'_k\mathbf{y})/f(\mathbf{B}'_j\mathbf{y}) > c_0$, are used for estimation. Since a small number of observations are used, the consistency of the estimators can be achieved with lower efficiency, as c_0 goes to infinity together with the sample size.

3.4. Initial values of z

Our experience suggests that the initial values of the state variable $\mathbf{z}$ are crucial for the estimation procedure. Here we provide a method for finding reasonable initial values of $\mathbf{z}$. Let

$$\begin{aligned}\mathbf{\Sigma}_y(l) &= \frac{1}{n-l}\sum_{t=1}^{n-l}\mathrm{Cov}(\mathbf{y}_t,\mathbf{y}_{t+l}) = \frac{1}{n-l}\sum_{t=1}^{n-l}\sum_{k=1}^{m}\pi_k\mathrm{Cov}(\mathbf{y}_t,\mathbf{y}_{t+l}\mid z_t=k)\\ &= \sum_{k=1}^{m}\pi_k\mathbf{\Sigma}_{y,k}(l).\end{aligned}$$

As a sum of m rank-d matrices, the matrix $\mathbf{\Sigma}_y(l)$ has a rank less than or equal to dm. If $\pi_{j,k} = 1/m$ for $j,k = 1,\ldots,m$, we have

$$\mathbf{\Sigma}_y(l) = \frac{1}{m^2}\sum_{k=1}^{m}\mathbf{A}_k\mathbf{\Sigma}_x(l)\sum_{j=1}^{m}\mathbf{A}'_j.$$

Hence $\mathbf{M} = \sum_{l=1}^{l_0} \mathbf{\Sigma}_y(l)\mathbf{\Sigma}_y(l)'$ is a matrix sandwiched by $\sum_{k=1}^m \mathbf{A}_k$ and $\sum_{k=1}^m \mathbf{A}_k'$ with rank smaller or equal to d. We find the eigenvalues of $\mathbf{M}$ and use the ratio estimator in (3.5) for the initial estimate of d. Specifically, let

$$\hat{\mathbf{\Sigma}}_y(l) = \frac{1}{n-l}\sum_{t=1}^{n-l}(\mathbf{y}_t - \hat{\boldsymbol{\mu}})(\mathbf{y}_{t+l} - \hat{\boldsymbol{\mu}}'), \quad \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t, \quad \hat{\mathbf{M}} = \sum_{l=1}^{l_0}\hat{\mathbf{\Sigma}}_y(l)\hat{\mathbf{\Sigma}}_y(l)'.$$

Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ be the eigenvalues of $\hat{\mathbf{M}}$ in descending order. We use $d_0 = \operatorname{argmin}_{1\le j\le p/2} \hat{\lambda}_{j+1}/\hat{\lambda}_j$ as the initial value of d.

The dynamic part of the observed process at time t lies in the column space of $\mathbf{A}_k$ if $z_t = k$. Therefore, $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ should be located near the m d-dimensional subspaces $\mathcal{M}(\mathbf{A}_1), \ldots, \mathcal{M}(\mathbf{A}_m)$. With d_0, we perform a principal component analysis on $\mathbf{y}_t$ to find the $d_0 m$ directions, $\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_{d_0 m}$ in descending order that account for the most variation of $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. We can then construct the m subspaces, $\mathcal{S}_1, \ldots, \mathcal{S}_m$ by dividing the set $\{\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_{d_0 m}\}$ into m groups that minimize the squared distance between $\mathbf{y}_t$ and its closest subspace. Specifically, let $\mathbf{s}_t = \{s_{1,t}, \ldots, s_{d_0 m,t}, \ldots, s_{p,t}\}$ be the principal component scores of $\mathbf{y}_t$, and $\{\mathcal{K}_1, \ldots, \mathcal{K}_m\}$ be a partition of the index set of $\{1, \ldots, d_0 m\}$, each $\mathcal{K}_i$ containing d_0 elements. Let

$$W(\mathcal{K}_1, \ldots, \mathcal{K}_m) = \sum_{t=1}^{n} \max_{1\le i\le m} \sum_{j\in\mathcal{K}_i} s_{j,t}^2, \tag{3.13}$$

and select the partition $\mathcal{K}_1^*, \ldots, \mathcal{K}_m^*$ that maximizes W. Here $\sum_{j\in\mathcal{K}_i}(s_{j,t})^2$ is the squared norm of the projection of $\mathbf{y}_t$ onto the space $\mathcal{S}_i$, where $\mathcal{S}_i = \mathcal{M}(\mathbf{q}_j,\, j \in \mathcal{K}_i)$, and it is maximized by the index corresponding to the subspace to which $\mathbf{y}_t$ is the closest, from $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$. Hence, the initial values of state variables can be set as

$$\hat{z}_t = \operatorname*{argmax}_{1\le i\le m} \sum_{j\in\mathcal{K}_i^*} s_{j,t}^2.$$

Finding the optimal partition is computationally extensive, unless $d_0 m$ is small. With large $d_0 m$, one can use a procedure similar to K-mean clustering to find a tentative solution, as the procedure is only for searching a good set of initial values.

Since the directions obtained by principal component analysis are orthogonal to each other, the constructed subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_m$ are also orthogonal. However, we do not assume that for $\mathcal{M}(\mathbf{A}_1), \ldots, \mathcal{M}(\mathbf{A}_m)$, so the constructed subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_m$ are not necessarily good estimates for loading spaces. It follows that the $d_0 m$ orthogonal directions are often more than needed and there may be states to which only a few observations are assigned. If it happens, a smaller d_0 can be used.

4. Theoretical Properties

In this section, we first investigate the convergence rates of the proposed estimator $\mathcal{M}(\hat{\mathbf{Q}}_k)$ and $\hat{d}$ as n and p go to infinity, given true state classification $\mathbf{z}$, under the setting of Section 3.1. Second, we introduce a theorem regarding misclassification under the setting of Section 3.2.

Some regularity conditions are needed.

Condition 1. The process $\mathbf{x}_t$ is α-mixing with mixing coefficients satisfying $\sum_{t=1}^{\infty} \alpha(t)^{1-2/\gamma} < \infty$, for some $\gamma > 2$, where

$$\alpha(t) = \sup_i \sup_{A \in \mathcal{F}_{-\infty}^i, B \in \mathcal{F}_{i+t}^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

and $\mathcal{F}_i^j$ is the σ-field generated by $\{\mathbf{x}_t : i \le t \le j\}$.

Condition 2. For any $j = 1, \ldots, d$, and $t = 1, \ldots, n$, $\mathrm{E}(|x_{j,t}|^{2\gamma}) \le C$, where $x_{j,t}$ is the j-th element of $\mathbf{x}_t$, $C > 0$ is a constant, and γ is given in Condition 1. For $l = 1, \ldots, l_0$, $\mathbf{\Sigma}_x(l)$ is of full rank, and $\|\mathbf{\Sigma}_x(l)\|_2 \asymp O(1) \asymp \|\mathbf{\Sigma}_x(l)\|_{\min}$.

Condition 3. Each element of $\mathbf{\Sigma}_k$, for $k = 1, \ldots, m$, remains bounded as p increases to infinity.

Condition 4. For each k, $k = 1, \ldots, m$, there exists a constant $\delta_k \in [0, 1]$ such that $\|\mathbf{A}_k\|_2^2 \asymp p^{1-\delta_k} \asymp \|\mathbf{A}_k\|_{\min}^2$, as p goes to infinity.

Condition 5. The Markov chain $\mathbf{z}$ is irreducible, positive recurrent and aperiodic.

Condition 6. For each $k = 1, \ldots, m$, let $\mathcal{C} = \{j \mid \delta_j = \min_{1 \le k \le m} \delta_k\}$ contain all the indices of the strongest states, and for each state k there exists an integer l_k, satisfying that $l_k \le l_0$, $\sum_{j \in \mathcal{C}} \pi_{k,j}^{(l_k)} \mathbf{A}_j$ is of rank d, and

$$\Big\| \sum_{j \in \mathcal{C}} \pi_{k,j}^{(l_k)} \mathbf{A}_j \Big\|_{\min}^2 \asymp p^{1-\delta_{\min}}. \tag{4.1}$$

Condition 7. For $k = 1, \ldots, m$, $\mathbf{M}_k$ in (3.2) has d distinct positive eigenvalues. For $j \ne k$, $\mathcal{D}(\mathbf{Q}_j, \mathbf{Q}_k) \ne 0$, where $\mathcal{D}(\cdot, \cdot)$ is defined in (3.12).

Remark 8. The stationarity of the latent process is not required, though we do require the mixing conditions stated in Condition 1.

Remark 9. $\mathbf{M}_k$, the quadratic form of autocovariance matrices of $\mathbf{y}_t$ depends on observations from other regimes, including the strongest regime. Hence the most dense loading matrices influence the estimation of the loading space for each state. Condition (4.1) requires that at one of the nonzero lags, the impact of the dense loading matrices do not cancel each other out. It is also used to bound $\|\mathbf{M}_k\|_{\min}$.

Remark 10. Condition 7 makes $\mathbf{Q}_k$ uniquely defined and identifiable, where $\mathbf{Q}_k = (\mathbf{q}_{k,1}, \ldots, \mathbf{q}_{k,d})$ with $\mathbf{q}_{k,1}, \ldots, \mathbf{q}_{k,d}$ the d orthonormal eigenvectors of $\mathbf{M}_k$ corresponding to the d nonzero eigenvalues $\lambda_{k,1} > \cdots > \lambda_{k,d}$.

Theorem 1. *If Conditions* 1–7 *hold with given observed state* $\mathbf{z}$ *and true* d *and, for* $k = 1, \ldots, m$, $p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$ *as* $n, p \to \infty$ *we have*

$$\|\hat{\mathbf{Q}}_k - \mathbf{Q}_k\|_2 = O_p(p^{\delta_k/2+\delta_{\min}/2} n^{-1/2}),$$

where $\delta_{\min} = \min_{1 \le k \le m} \delta_k$.

When $m = 1$, a special case of our setting, Lam, Yao, and Bathia (2011) proved that the convergence rate of the estimator of the loading space is $O_p(p^{\delta} n^{-1/2})$. For the regime switching model with $m > 1$, our results show that, except for the strongest states (with $\delta_{\min}$), the estimators of loading spaces for all the weaker states converge faster than $p^{\delta_k} n^{-1/2}$. Thus, the estimators of the loading spaces for the strongest states retain the same convergence rate, while these for other states gain some efficiency from regime switching mechanism. The main reason for this is that our approach depends on the autocovariance matrices of $\mathbf{y}_t$ given $z_t = k$ at leads $1, \ldots, l_0$. It is a linear combination of autocovariance matrices given current state k switching to all the states. The autocovariance matrices switching to the strongest states have the leading order and all other terms are of smaller order.

Theorem 2. *If Conditions* 1–7 *hold with observed state* $\mathbf{z}$ *and true* d *and, for* $k = 1, \ldots, m$, $p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$ *and* $\|\mathbf{\Sigma}_k\|_2$ *is bounded, as* $n, p \to \infty$ *we have*

$$p^{-1/2} \|\hat{\mathbf{f}}_t - \mathbf{f}_t\|_2 = O_p(p^{\delta_{\min}/2} n^{-1/2} + p^{-1/2}).$$

Theorem 2 provides the convergence of the extracted factor term, and the rate does not vary across regimes, free of δ_k. It shows that by introducing stronger states, the estimated dynamic part of the observed process shows an overall improvement.

If the distance measure in (3.12) is adopted for the loading space $\mathcal{M}(\mathbf{Q}_k)$, then we have a result about its estimation error.

Theorem 3. *If Conditions* 1–7 *hold with observed state* $\mathbf{z}$ *and true* d *and, for* $k = 1, \ldots, m$, $p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$ *as* $n, p \to \infty$, *we have*

$$\mathcal{D}(\hat{\mathbf{Q}}_k, \mathbf{Q}_k) = O_p(p^{\delta_k/2+\delta_{\min}/2} n^{-1/2}).$$

Theorem 3 shows that the error for estimated loading space is on the same order as that for the estimated $\mathbf{Q}_k$ when $\mathbf{Q}_k$ is uniquely defined as in Remark 10.

Theorem 4. *If Conditions* 1–7 *hold with observed states* $\mathbf{z}$ *and, for* $k = 1, \ldots, m$, $h_{n,k} = p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$ *as* n, $p \to \infty$, *the eigenvalues* $\{\hat{\lambda}_{k,1}, \ldots, \hat{\lambda}_{k,p}\}$ *of* $\hat{\mathbf{M}}_k$ *satisfy*

(1) $|\hat{\lambda}_{k,i} - \lambda_{k,i}| = O_p(p^{2-\delta_k/2-\delta_{\min}/2} n^{-1/2})$ *for* $i = 1, \ldots, d$, *and*
(2) $\hat{\lambda}_{k,j} = O_p(p^2 n^{-1})$ *for* $j = d+1, \ldots, p$.

Corollary 1. *Under the conditions of Theorem* 4, *we have* $\hat{\lambda}_{k,j+1}/\hat{\lambda}_{k,j} \asymp 1$ *for* $j = 1, \ldots, d$, *and* $\hat{\lambda}_{k,d+1}/\hat{\lambda}_{k,d} = O_p(p^{\delta_k+\delta_{\min}} n^{-1})$, *for* $k = 1, \ldots, m$.

Theorem 4 shows that the estimators for the d nonzero eigenvalues of $\mathbf{M}_k$ converge more slowly than those for the $p-d$ zero eigenvalues. Corollary 1 gives the order of ratio of the estimated eigenvalues, and provides partial theoretical support for the ratio estimator proposed in Section 3.1. Because of differences in δ_k, the stronger the state k is, the faster convergence rate $\hat{\lambda}_{k,d+1}/\hat{\lambda}_{k,d}$ has. Therefore, we choose $\hat{d}_{\tilde{k}}$ as the estimator of the number of factors using the state $\tilde{k}$ for maximizing $\|\hat{\mathbf{M}}_k\|_2$, since it is related through $\|\hat{\mathbf{M}}_k\|_2 = O_p(p^{2-\delta_k-\delta_{\min}})$, proved by Lemma 3 and 4 in Supplementary Material.

Remark 11. The asymptotics of $\hat{\lambda}_{k,i+1}/\hat{\lambda}_{k,i}$ with $i > d$ are difficult to obtain, even when $m = 1$; see Remark 2 in Lam and Yao (2012). Chang, Guo, and Yao (2013) adjusted the ratio estimator as

$$\hat{d} = \underset{1 \le j \le p/2}{\operatorname{argmin}} \Big\{ \frac{\hat{\lambda}_{j+1} + C_T}{\hat{\lambda}_j + C_T} \Big\}, \tag{4.2}$$

where $C_T = p^{2-\delta} n^{-1/2} \log n$ for a one-regime model, and proved it is a consistent estimator for d. However, the adjusted ratio estimator in (4.2) cannot be used for data analysis as δ is unknown. In practice, the ratio estimator in (3.5) is used; see Lam, Yao, and Bathia (2011), Lam and Yao (2012) and Chang, Guo, and Yao (2013).

Next we investigate the performance of the estimator of the state $\mathbf{z}$. To simplify matters, we assume the $\pi_{k,j}$ for $k, j = 1, \ldots, m$ are equal, hence estimating z_t can be done separately for each t, instead of relying on the Viterbi algorithm. This is also equivalent to pure classification without the Markov chain mechanism. The setting is not exactly what we assumed in Section 3.2, but the results reveal how misclassification occurs and its impact on the estimation of the rest of the parameters.

Let

$$w_{t,k,j} = \log[f(\mathbf{B}_k' \mathbf{y}_t)] - \log[f(\mathbf{B}_j' \mathbf{y}_t)] = l(k|\mathbf{y}_t) - l(j|\mathbf{y}_t), \tag{4.3}$$

where

$$l(k|\mathbf{y}_t) = \log(f(\mathbf{B}_k'\mathbf{y}_t)) \\ = -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,k}| - \frac{(\mathbf{B}_k'(\mathbf{y}_t-\boldsymbol{\mu}_k))'\mathbf{\Sigma}_{B,k}^{-1}\mathbf{B}_k'(\mathbf{y}_t-\boldsymbol{\mu}_k)}{2}. \quad (4.4)$$

The estimator of z_t under the equal transition probability assumption can be rewritten as $\hat{z}_t = k$ if $w_{t,k,j} > 0$ for all $j \neq k$. Hence misclassification occurs when there exists a j such that $w_{t,z_t,j} < 0$. Specifically, the probability of misclassification, when $z_t = k$, is $P(\hat{z}_t \neq z_t = k) = P(\min_{j\neq k} w_{t,k,j} < 0)$. When $z_t = k$,

$$l(k|\mathbf{y}_t) = -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,k}| - \frac{1}{2}(\mathbf{B}_k'(\mathbf{y}_t-\boldsymbol{\mu}_k))'\mathbf{\Sigma}_{B,k}{}^{-1}\mathbf{B}_k'(\mathbf{y}_t-\boldsymbol{\mu}_k) \\ = -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,k}| - \frac{1}{2}(\boldsymbol{\varepsilon}_t^{(k)\prime}\mathbf{B}_k\mathbf{\Sigma}_{B,k}{}^{-1}\mathbf{B}_k'\boldsymbol{\varepsilon}_t^{(k)}),$$

and

$$l(j|\mathbf{y}_t) = -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,j}| \\ -\frac{1}{2}\left(\mathbf{B}_j'(\mathbf{A}_k\mathbf{x}_t+\boldsymbol{\mu}_k-\boldsymbol{\mu}_j+\boldsymbol{\varepsilon}_t^{(k)})\right)'\mathbf{\Sigma}_{B,j}{}^{-1}\left(\mathbf{B}_j'(\mathbf{A}_k\mathbf{x}_t+\boldsymbol{\mu}_k-\boldsymbol{\mu}_j+\boldsymbol{\varepsilon}_t^{(k)})\right).$$

Hence,

$$w_{t,k,j} = \frac{1}{2}(\log|\mathbf{\Sigma}_{B,j}| - \log|\mathbf{\Sigma}_{B,k}|) + \frac{1}{2}\boldsymbol{\varepsilon}_t^{(k)\prime}(\mathbf{B}_j\mathbf{\Sigma}_{B,j}{}^{-1}\mathbf{B}_j' - \mathbf{B}_k\mathbf{\Sigma}_{B,k}{}^{-1}\mathbf{B}_k')\boldsymbol{\varepsilon}_t^{(k)} \\ +\frac{1}{2}\left(\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t + \mathbf{B}_j'(\boldsymbol{\mu}_k-\boldsymbol{\mu}_j)\right)'\mathbf{\Sigma}_{B,j}{}^{-1}\left(\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t + \mathbf{B}_j'(\boldsymbol{\mu}_k-\boldsymbol{\mu}_j)\right) \\ +(\mathbf{B}_j'(\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k-\boldsymbol{\mu}_j))'\mathbf{\Sigma}_{B,j}^{-1}\mathbf{B}_j'\boldsymbol{\varepsilon}_t^{(k)} \\ = I_1 + I_2 + I_3 + I_4. \quad (4.5)$$

Here I_1 is a given constant, measuring the differences in variation of the two states. I_2 reflects the impact of the noise after being projected into the space $\mathbf{B}_k$ and $\mathbf{B}_j$. The third term I_3 shows the size of the noises needed for misclassification. In addition, $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ is the projection of $\mathbf{x}_t$ on the intersection of $\mathcal{M}(\mathbf{B}_j)$ and $\mathcal{M}(\mathbf{A}_k)$. If $\mathcal{M}(\mathbf{A}_k)$ and $\mathcal{M}(\mathbf{A}_j)$ are less common, then $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ is larger (in magnitude), and the chance for misclassification is less. Of course, if the difference in the mean $\boldsymbol{\mu}_k - \boldsymbol{\mu}_j$ is larger, then the misclassification probability is smaller. Here I_4 is the cross term of I_2 and I_3.

Misclassification of z_t may not have large impact on the estimation of the loading spaces. Small $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ leads to misclassifying the observation from state k to state j. It happens in two situations. For some observations, $\mathbf{A}_k\mathbf{x}_t$ are close to the column space of $\mathbf{A}_j$, which makes $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ small. Such observations hence lie close to the space $\mathcal{M}(\mathbf{A}_j)$ and do not have large impact in the estimation of $\mathcal{M}(\mathbf{A}_j)$. The other possibility is that these misclassified observations have a small

signal-to-noise ratio which makes $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ small, hence they are less influential for the estimation of $\mathcal{M}(\mathbf{A}_j)$.

Theorem 5. *If $\mathbf{x}_t$ is a normally distributed random process, given true $\mathbf{B}_k$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_{B,k}$, $k = 1, \ldots, m$, we have*

$$\begin{aligned} \mathrm{E}(w_{t,k,j}) &= \frac{1}{2}(\log|\boldsymbol{\Sigma}_{B,j}| - \log|\boldsymbol{\Sigma}_{B,k}|) \\ &\quad + \frac{1}{2}\mathrm{tr}\left((\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{f,k,t} + \mathbf{U}_{k,j})\mathbf{W}_j\right) - \frac{(p-d)}{2}, \qquad (4.6) \\ \mathrm{Var}(w_{t,k,j}) &= \frac{1}{2}\|\boldsymbol{\Sigma}_k^{1/2}(\mathbf{W}_j - \mathbf{W}_k)\boldsymbol{\Sigma}_k^{1/2}\|_F^2 + \frac{1}{2}\|\boldsymbol{\Sigma}_k^{1/2}\mathbf{A}_k'\mathbf{W}_j\mathbf{A}_k\boldsymbol{\Sigma}_k^{1/2}\|_F^2 \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{f,k,t}\mathbf{W}_j\mathbf{U}_{k,j}\mathbf{W}_j\right) + \mathrm{tr}\left((\boldsymbol{\Sigma}_{f,k,t} + \mathbf{U}_{k,j})\mathbf{W}_j\boldsymbol{\Sigma}_k\mathbf{W}_j\right), \qquad (4.7) \end{aligned}$$

where $\boldsymbol{\Sigma}_{f,k,t} = \mathrm{Var}(\mathbf{A}_k\mathbf{x}_t)$, and $\mathbf{U}_{k,j} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)'(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)$, $\mathbf{W}_j = \mathbf{B}_j\boldsymbol{\Sigma}_{B,j}^{-1}\mathbf{B}_j'$.

The mean and variance of $\omega_{t,k,j}$ increase with p. As expected, misclassification is unavoidable. For weak states, as p increases, the accumulated noise tends to overwhelm the difference in $\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k - \boldsymbol{\mu}_j$. Hence classification error may increase with p. On the other hand, for strong states, the signal remains strong and the misclassification rate is much better than those for weak states.

5. Simulations

In this section, we illustrate the performance of the proposed estimators with some numerical experiments, compare their convergence rates for states with different strength, and explore the interactions among states. The performance of the estimators of $\mathcal{M}(\mathbf{Q}_k)$ and d, and the performance of clustering are presented separately.

With two switching regimes $m = 2$, we considered three models. In Model 1, both states were strong, with $\delta_1 = \delta_2 = 0$. In Model 2, one of the states was strong, and one weak, with $\delta_1 = 0$, and $\delta_2 = 1$. In Model 3, both states were weak, with $\delta_1 = \delta_2 = 1$. The transition probabilities between the two states were set to 0.5. In the simulation, all $p \times d$ entries in $\mathbf{A}_k$ were generated independently from the uniform distribution on $[-p^{-\delta_k/2}, p^{-\delta_k/2}]$ with strength δ_k. The mean of observed process $\boldsymbol{\mu}_k$ was a $p \times 1$ vector with all entries zero, for $k = 1, 2$. Different values of d, and different structures of the latent process and noises were used. In all the examples, we used $l_0 = 1$. Estimation error of $\mathcal{M}(\hat{\mathbf{Q}}_k)$ is defined as $\mathcal{D}(\hat{\mathbf{Q}}_k, \mathbf{Q}_k)$.

5.1. The performance of $\mathcal{M}(\hat{\mathbf{Q}}_k)$

In this experiment d was set to 1 and we estimated the loading spaces using the true d. The factor process x_t was from an AR(1) process with AR coefficient

Table 1. Means of the estimation errors $\mathcal{D}(\hat{\mathbf{Q}}_k, \mathbf{Q}_k)$.

		$\mathbf{z}$ observed			$\mathbf{z}$ unobserved		
		$p = 20$	$p = 40$	$p = 80$	$p = 20$	$p = 40$	$p = 80$
Model 1	State 1 ($\delta_1 = 0$)	0.0159	0.0164	0.0161	0.0438	0.0606	0.1055
	State 2 ($\delta_2 = 0$)	0.0143	0.0155	0.0169	0.0445	0.0711	0.0958
Model 2	State 1 ($\delta_1 = 0$)	0.0203	0.0216	0.0207	0.0225	0.0274	0.0304
	State 2 ($\delta_2 = 1$)	0.0856	0.1274	0.2131	0.0977	0.1495	0.2689
Model 3	State 1 ($\delta_1 = 1$)	0.0796	0.1489	0.4149	0.2424	0.5563	0.6067
	State 2 ($\delta_2 = 1$)	0.0803	0.1453	0.4226	0.2626	0.5091	0.6614

0.9 and $N(0, 4)$ noises. The noise process $\{\boldsymbol{\varepsilon}_t^{(1)}, \ldots, \boldsymbol{\varepsilon}_t^{(m)}\}$ were m independent vector white noise processes whose covariance matrix had 1 on the diagonal and 0.95 for all off-diagonal entries. We set the pre-specified controls t_0, c_1, and c_2 in the iterative algorithm in Section 3.3 to 50, 0.001, and 0.001, respectively.

We repeated the simulation 100 times with sample size n = 1,000. Let $p = 20, 40, 80$. Table 1 and Figure 1 show the results for when $\mathbf{z}$ is observed and when $\mathbf{z}$ is unobserved.

When the state variable $\mathbf{z}$ is observed, by comparing the results of Model 2 to these of Model 1, we can see that the estimation for the strong state is slightly worse after a weak state is introduced to the model. However, by comparing the results of Model 2 to these of Model 3, the estimation for the weaker state is much better due to the existence of a strong state, especially when p is large. It shows that the estimation for weak states benefits from the stronger states as our theory indicates.

The top panels in Figure 1 display the boxplots of estimation errors for different p when $\mathbf{z}$ is observed in each model on the same scale. In addition to what we can see from Table 1, it shows that the estimation variation increases with p and is larger for the weak states.

When $\mathbf{z}$ is unobserved, the estimation errors of the loading spaces for each model with different p have similar pattern to those in the case when $\mathbf{z}$ is known, as seen in Table 1 and the bottom panels in Figure 1. The estimation of strong states still has good performance in absence of weak states; weak states benefit from the existence of strong states. The estimators are less accurate if the state variable is unobserved. As p increases, even the strong states suffer from unobserved $\mathbf{z}$. Because of lack of information on $\mathbf{z}$, it happens that our algorithm is trapped in a local maximum.

5.2. The clustering performance

In this experiment we used the settings in Section 5.1 for a comparison of the clustering performance among models with different strength. Results of

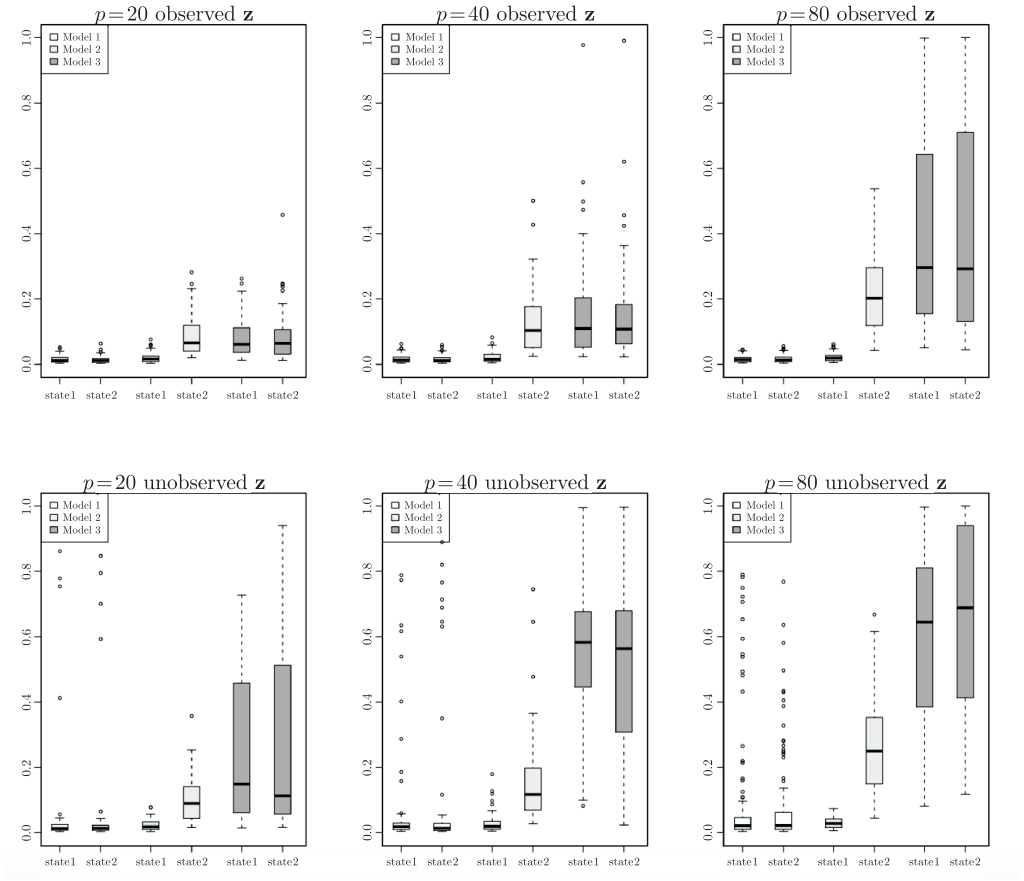

Figure 1. Boxplots of estimation errors of $\mathcal{M}(\hat{\mathbf{Q}}_k)$ for $p = 20, 40, 80$ when $\mathbf{z}$ is observed (top panels), and when $\mathbf{z}$ is unobserved (bottom panels), under true d for the model described in Section 5.1.

Table 2. Means(sd) of misclassification rates of the hidden states.

	$p = 20$	$p = 40$	$p = 80$
Model 1	0.0334(0.0909)	0.0485(0.1169)	0.1153(0.1418)
Model 2	0.0452(0.0128)	0.0635(0.0609)	0.1525(0.0723)
Model 3	0.2155(0.1905)	0.4439(0.0482)	0.4686(0.0434)

misclassification rates and transition probabilities are summarized in Tables 2 and 3, respectively.

Table 2 shows the misclassification rates for each model with different p. It is seen that misclassification occurs very often when all the states are weak, but less so in the presence of at least one strong state.

Table 3 shows the means and standard deviations of the estimated transi-

Table 3. Means(sd) of estimated transition matrices. The true values are all 0.5.

	$p = 20$		$p = 40$		$p = 80$	
Model 1	0.5110	0.4890 (0.1008)	0.6435	0.3565 (0.0689)	0.5273	0.4727 (0.1134)
	0.5000	0.5000 (0.0877)	0.6133	0.3867 (0.0582)	0.4918	0.5082 (0.1021)
Model 2	0.5322	0.4678 (0.0296)	0.5423	0.4577 (0.0431)	0.4996	0.5004 (0.0558)
	0.5219	0.4781 (0.0260)	0.5373	0.4627 (0.0422)	0.4739	0.5261 (0.0601)
Model 3	0.5153	0.4847 (0.2145)	0.5861	0.4139 (0.3114)	0.4240	0.5760(0.3339)
	0.4799	0.5201 (0.2201)	0.5200	0.4800 (0.3129)	0.3686	0.6314(0.3388)

Table 4. The relative frequency estimates of $\hat{d} = d$.

	n	50	100	200	500	1,000
Model 1	$p = 0.1n$	0.180	0.385	0.725	0.995	1
	$p = 0.5n$	0.380	0.610	0.850	0.995	1
	$p = 0.8n$	0.390	0.585	0.855	0.995	1
Model 2	$p = 0.1n$	0.200	0.405	0.820	1	1
	$p = 0.5n$	0.365	0.605	0.915	1	1
	$p = 0.8n$	0.380	0.620	0.905	1	1
Model 3	$p = 0.1n$	0.115	0.125	0.075	0.075	0.275
	$p = 0.5n$	0.055	0.100	0.200	0.065	0
	$p = 0.8n$	0.080	0.080	0.060	0	0

tion probabilities, where the true transition probabilities are all 0.5. For Model 3, because of random noises and lack of information, some observations are misclassified to each state with larger probability comparing to Model 1 and Model 2, since the standard deviations of estimates of transition probabilities for Model 3 are much larger than those for Model 1 and Model 3.

5.3. The performance of $\hat{d}$

In this experiment we set the number of factors to 3 ($d = 3$) and investigated the performance of the proposed estimator for d, under true $\mathbf{z}$. Here the latent process $\mathbf{x}_t$ was set to be three independent AR(1) processes with $N(0, 4)$ noises and AR coefficients 0.6, -0.5 and 0.8, respectively. $\{\boldsymbol{\varepsilon}_t^{(1)}, \ldots, \boldsymbol{\varepsilon}_t^{(m)}\}$ were m white noise process whose covariance matrix had 1 on the diagonal and 0.2 for the off-diagonal entries. We took $n = 50$, 100, 200, 500, 1,000, and $p = 0.1n$, $0.5n$, $0.8n$. We repeated the simulation 200 times for each (n, p) setting and the relative frequencies of correct estimates of d are reported in Table 4.

From Table 4 we see that the existence of a strong state, no matter whether or not there is a weaker state, produces much more accurate estimates for the number of factors d. As n increases, the estimations all improve in the presence of a strong state. Regarding the impact of p, it is seen that the estimation of d benefits from a 'blessing of dimensionality' when one or more strong states exist,

and perform better as p increases. However, when all states are extremely weak ($\delta_1 = \delta_2 = 1$), the number of correct estimation goes to 0 as n increases. The features do not change much with p, partially because the increase of information in n offsets the increase of noise introduced as p increases.

6. Data Analysis

We applied our approach to the daily returns of 123 stocks from January 2, 2002 to July 11, 2008. These stocks were selected among those included in the S&P 500 and traded every day during the period. The returns were calculated in percentages based on daily closing prices. This data was analyzed by Lam and Yao (2012) and Chang, Guo, and Yao (2013). We have the sample size n =1,642 and the dimension of observations $p = 123$. We assume that there are two regimes, $m = 2$, and set $l_0 = 1$, $t_0 = 200$, $c_0 = c_1 = 0.001$. Varying the value of l_0 does not change the estimation results significantly.

The proposed iterative procedure yields $\hat{d} = 1$, which differs from the number of factors estimated in Lam and Yao (2012). Hence allowing the loading matrix to change across two regimes reduces the number of factors needed. The one factor accounts for 25.92% of the total variation of stock returns. The residuals $\hat{\boldsymbol{\varepsilon}}_t$ are computed from (3.7). The sample cross-autocorrelations of $\hat{\boldsymbol{\varepsilon}}_t$ for the first 7 stocks are plotted in Figure 2. There are almost no significant nonzero autocorrelations for $\hat{\boldsymbol{\varepsilon}}_t$, showing that, after extracting the latent factor, little serial dependence is left in the data. Our results indicate that only one factor drives the 123 stocks, but the factor loadings switch between two states. Ignoring the switching structure as in Lam and Yao (2012), it would appear that there are two different factors.

Even with $d = 1$, $\mathbf{Q}_k$ is still not unique due to a trivial replacement of $(\mathbf{Q}_k, \mathbf{R}_t I(z_t = k))$ by $(-\mathbf{Q}_k, -\mathbf{R}_t I(z_t = k))$ for either $k = 1$ or $k = 2$, or both, in (2.7). According to (2.3) and (2.5), take $\mathbf{A}_k = \gamma_k \mathbf{Q}_k$ and $\mathbf{R}_t I(z_t = k) = \gamma_k \mathbf{x}_t$, where we could set γ_k to 1 or -1 when the dimension is fixed. Here we choose γ_k which makes the majority of the entries in $\mathbf{Q}_k$ positive, hence $\mathbf{y}_t$ is most positively correlated with the corresponding latent factor $\mathbf{x}_t$ and $\mathbf{R}_t$, since

$$\mathbf{y}_t = \boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)} = \boldsymbol{\mu}_k + \mathbf{Q}_k \mathbf{R}_t + \boldsymbol{\varepsilon}_t^{(k)}, \text{ when } z_t = k.$$

Specifically, let $\hat{\mathbf{Q}}_k$ and $\hat{\mathbf{R}}_t$ be the estimate of $\mathbf{Q}_k$ and $\mathbf{R}_t$ without consideration of signs. We adjust the sign of $\hat{\mathbf{Q}}_k$ as

$$\hat{\mathbf{Q}}_{\text{adj},k} = \begin{cases} -\hat{\mathbf{Q}}_k, & \text{if } \sum_{i=1}^{p} I(\hat{q}_{k,i} > 0) > \sum_{i=1}^{p} I(-\hat{q}_{k,i} > 0), \\ \hat{\mathbf{Q}}_k, & \text{otherwise,} \end{cases} \tag{5.1}$$

where $\hat{q}_{k,i}$ is the i-th entry in $\hat{\mathbf{Q}}_k$ for $k = 1, 2$. The adjusted $\hat{\mathbf{Q}}_k$ makes most of its entries positive. $\hat{\mathbf{R}}_{\text{adj},t}$ is obtained according to $\hat{\mathbf{Q}}_{\text{adj},1}$ and $\hat{\mathbf{Q}}_{\text{adj},2}$.

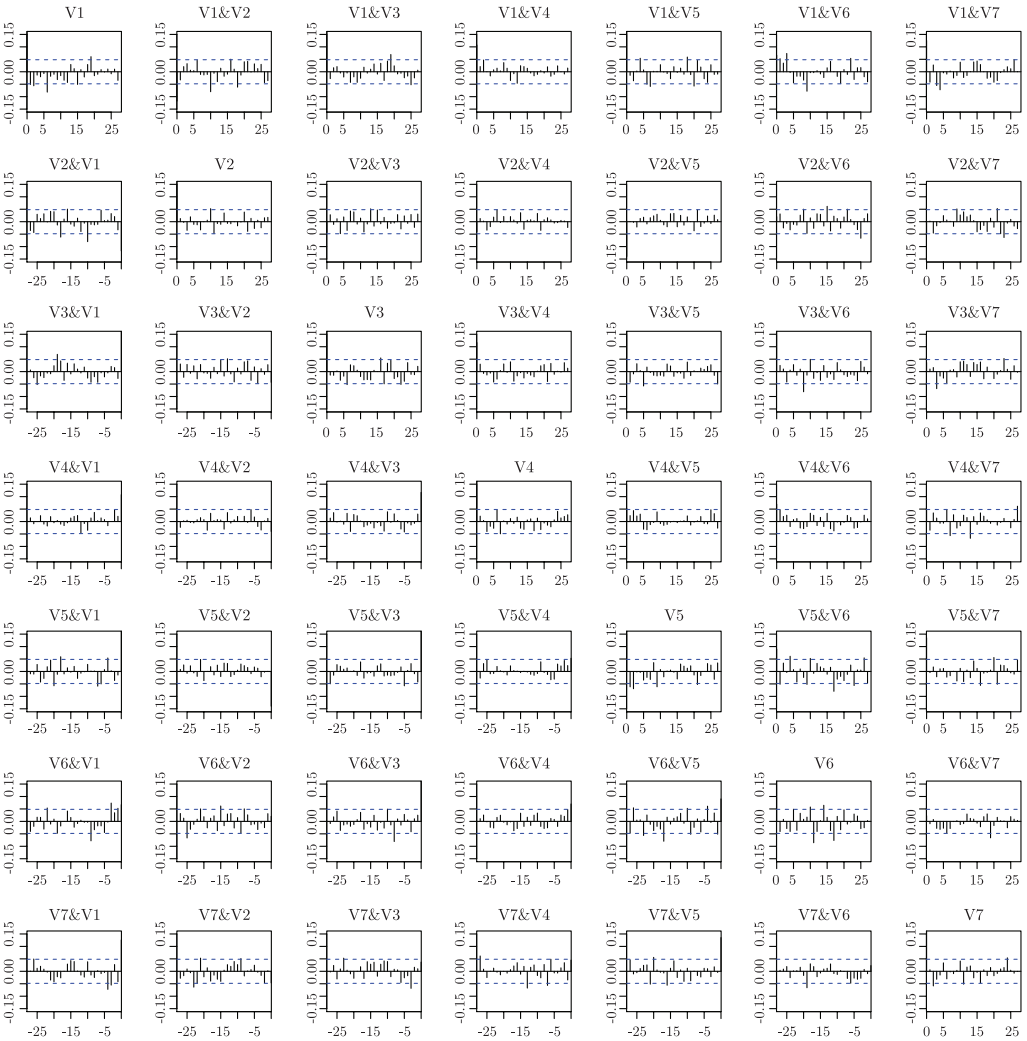

Figure 2. Plots of the sample cross-autocorrelations of $\hat{\boldsymbol{\varepsilon}}_t$ of the first 7 stocks with lag 0 autocorrelation removed.

Figure 3 displays the time series plots of $\hat{\mathbf{R}}_{\text{adj},t}$ in the top panel and returns of the S&P 500 index in the bottom panel. $\hat{\mathbf{R}}_{\text{adj},t}$ changes along with the S&P 500 index in this period, except for a few days around July 22, 2002, and it explains 76.27% of the total variation in the S&P 500 index. Hence, this factor can be regarded as a representation of market performance. Because index funds, which aim to replicate the movements of an index of a financial market, build their investment portfolio with all stocks in the index and trade them together, it causes synchronous oscillations between the market and the stocks. The popularity of index funds provides a reason that the market factor accounts for a

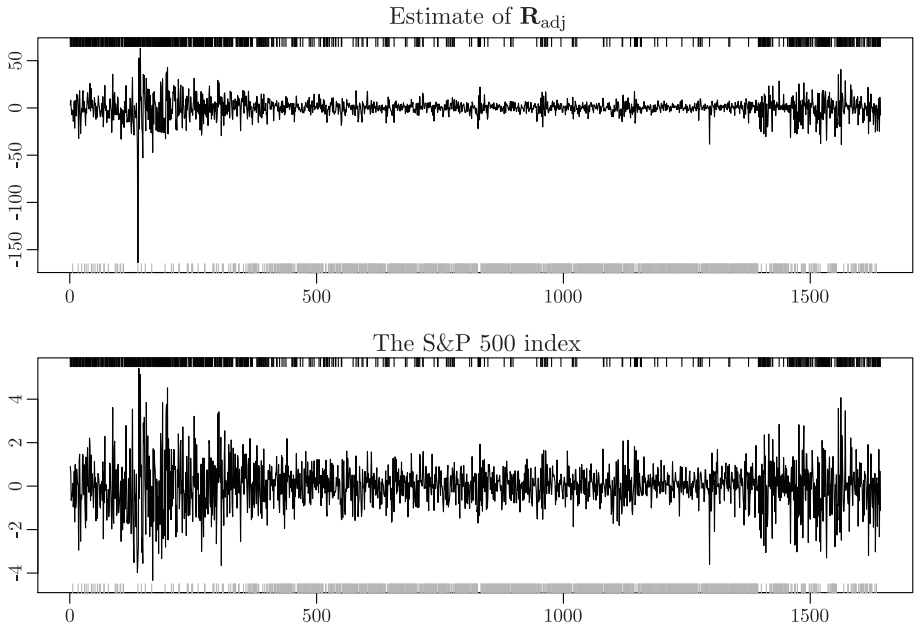

Figure 3. Time series plots of $\hat{\mathbf{R}}_{\text{adj},t}$ (top panel) and the return series of the S&P 500 index (bottom panel) in the same period. Indicators of the estimated states of the observations $I(\hat{z}_t = k)$ for $k = 1, 2$, are shown in the rug plots, on the top for State 1 and at the bottom for State 2.

large percentage of the total variation of stock returns.

The indicators of the estimated state variable $I(\hat{z}_t = k)$, for $k = 1, 2$, are shown in the rug plots of both panels in Figure 3, State 1 on the top and State 2 at the bottom. It is obvious that the state variable is strongly correlated with the volatility of the market. The standard deviation of the S&P 500 index is 1.4642 given $\hat{z}_t = 1$, while the standard deviation of the S&P 500 index is 0.6649 given $\hat{z}_t = 2$. When the S&P 500 index was volatile in 2002, 2003, and 2007 due to the internet bubble, the invasion of Iraq, and the subprime crisis, respectively, the observations were more likely to belong to State 1; when the S&P 500 index was stable in 2004-2006, the observations tend to be assigned to State 2.

For State 1, the factor accounts for 34.89% of the total variation in $\mathbf{y}_t$, while for State 2, it only accounts for 15.75%. A possible explanation is that investors may prefer passive management, such as index-tracking funds, to avoid nonsystematic risk when the market is volatile.

The estimated transition probabilities are shown in Table 5. During this period, about two-thirds of the time the system stays in State 2. The transition

Table 5. Estimated transition matrix and stationary probabilities.

	State 1	State 2	π_k
State 1	0.6758	0.3242	0.3782
State 2	0.1969	0.8031	0.6281

between the states are frequent, especially from State 1 to State 2.

Supplementary Material

The detailed proofs of Lemma 1−4, Theorem 1−5, and Corollary 1 are provided in the Supplementary Material.

Acknowledgements

Rong Chen is the corresponding author. The research is partially supported by NSF grants DMS-0905763 and DMS-1209085. The authors wish to thank the co-editors and two anonymous referees for their insightful comments which leads to significant improvement of the paper. We also thank Qiwei Yao for his insightful comments and providing the data set used in the example.

References

Anderson, T. W. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika* **28**, 1-25.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191-221.

Bauwens, L., Laurent, S. and Rombouts, J. V. (2006). Multivariate GARCH models: a survey. *J. Appl. Econom.* **21**, 79-109.

Bathia, N., Yao, Q. and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *Ann. Statist.* **38**, 3352-3386.

Bernanke, B. S. and Gertler, M. (2000). Monetary policy and asset price volatility. National Bureau of Economic Research, working paper.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51**, 1281-1304.

Chang, J., Guo, B. and Yao, Q. (2013). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. Technical report.

Chen, R. (1995). Threshold variable selection in open-loop threshold autoregressive models. *J. Time Series* **16**, 461-481.

Diebold, F. X. and Rudebusch, G. D. (1994). Measuring business cycles: a modern perspective. National Bureau of Economic Research, working paper.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.* **84**, 313-342.

Doz, C., Giannone, D. and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *J. Econom.* **164**, 188-205.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*. **55**, 251-276.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econom. Theory*. **11**, 122-150.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York.

Forney Jr, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE* **61**, 268-278.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *Rev. Econom. Statist.* **82**, 540-554.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *J. Econom.* **42**, 27-62.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*. **102**, 603-617.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357-384.

Hamilton, J. (1996). Specification testing in Markov-switching time-series models. *J. Econom.* **70**, 127-157.

Hamilton, J. and Susmel, R. (1994). Autoregressive conditional heteroscedasticity and changes in regime. *J. Econom.* **64**, 307-333.

Hansen, B. E. (1992). The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *J. Appl. Econom.* **7**, S61-S82.

Harvey, A., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *Rev. Econom. Studies* **61**, 247-264.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Kearns, M., Mansour, Y. and Ng, A. Y. (1998). An information-theoretic analysis of hard and soft assignment methods for clustering. *Learning in Graphical Models*. 495-520.

Kemeny, J. G. and Snell, J. L. (1960). *Finite Markov Chains*. Nostrand, Princeton, NJ.

Kim, C. J. and Nelson, C. R. (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *Rev. Econom. Statist.* **80**,188-201.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Statist.* **40**, 694-726.

Lam, C., Yao, Q. and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*. **98**, 901-918.

Lütkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. *Journal of Time Series Analysis*. **6**, 35-52.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Merikoski, J. K. and Kumar, R. (2004). Inequalities for spreads of matrix sums and products. *Appl. Math. E-notes* **4**, 150-159.

Meyne, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95**, 365-379.

Peña, D. and Box, G. E. P. (1987). Identifying a simplifying structure in time series. *J. Amer. Statist. Assoc.* **82**, 836-843.

Priestley, M. B., Rao, T. S. and Tong, H. (1974). Application of principal component analysis and factor analysis in the identification of multivariable systems. *IEEE Trans. Automat. Control* **19**, 730-734.

Quenouille, M. H. (1957). *The Analysis of Multiple Time Series*. Griffin, London.

Sims, C. A. and Zha, T. (2006). Were there regime switches in US monetary policy? *Amer. Econom. Rev.*, 54-81.

Stock, J. H. and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *National Bureau of Economic Research, Macroeconomics Annual*. **4**, 351-394.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indices. *J. Business Econom. Statist.* **20**, 147-162.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. National Bureau of Economic Research, working paper.

Tao, M., Wang, Y., Yao, Q. and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Amer. Statist. Assoc.* **106**, 1025-1040.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *J. Amer. Statist. Assoc.* **76**, 802-816.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series (with discussion). *J. Roy. Statist. Soc. Ser. B* **51**, 157-213.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. Springer-Verlag.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. Ser. B*, 245-292.

Tsay, R. S. (2010). *Analysis of Financial Time Series*. Wiley, New Jersey.

Tsay, R. S. and Tiao, G. C. (1983). Identification of multiplicative ARMA models for seasonal time series. Technical report; University of Chicago, Chicago.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* **13**, 260-269.

Watson, M. W. and Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *J. Econom.* **23**, 385-400.

Management Information Systems Department, San Diego State University, San Diego, CA 92182, USA.

E-mail: xialu.liu@mail.sdsu.edu

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.

E-mail: rongchen@stat.rutgers.edu

(Received August 2014; accepted March 2015)