Estimation and inference for high dimensional factor model with regime switching

Urga, Giovanni and Wang, Fa

7 May 2022

Estimation and Inference for High Dimensional Factor Model with Regime Switching

Giovanni Urga[*] Fa Wang[†]

May 7, 2022

Abstract

This paper proposes maximum (quasi)likelihood estimation for high dimensional factor models with regime switching in the loadings. The model parameters are estimated jointly by EM algorithm, which in the current context only requires iteratively calculating regime probabilities and principal components of the weighted sample covariance matrix. When regime dynamics are taken into account, smoothed regime probabilities are calculated using a recursive algorithm. Consistency, convergence rates and limit distributions of the estimated loadings and the estimated factors are established under weak cross-sectional and temporal dependence as well as heteroscedasticity. It is worth noting that due to high dimension, regime switching can be identified consistently right after the switching point with only one observation. Simulation results show good performance of the proposed method. An application to the FRED-MD dataset demonstrates the potential of the proposed method for quick detection of business cycle turning points.

Keywords: Factor model, Regime switching, Maximum likelihood, High dimension, EM algorithm, Turning points

JEL Classification: C13, C38, C55

[*]Faculty of Finance, Bayes Business School (formerly Cass), 106 Bunhill Row, London, EC1Y 8TZ, U.K.

[†]School of Economics, Peking University, 5 Yiheyuan Road, Beijing, 100871, China.

1 Introduction

A great deal of attention has focused on the loading instability issue in high dimensional factor models[1]. Various procedures are proposed to detect and/or estimate common abrupt breaks in loadings, including Cheng, Liao and Shorfheide (2016), Baltagi, Kao and Wang (2017, 2021) and Bai, Han and Shi (2020), to mention a few. Other models of time varying loadings, such as i.i.d./random walk, smooth change, vector autoregression and threshold type, are studied in Bates, Plagborg-Moller, Stock and Watson (2013), Su and Wang (2017), Mikkelsen, Hillebrand and Urga (2019) and Massacci (2017), respectively.

An alternative approach of modeling loading instability is common regime switching. In business cycle analysis, several unobservable factors summarize the comovements of many economic variables and loadings measure the importance of factors for each economic variable. The importance of each factor may be different depending on fiscal policy (expansionary, contractionary, neutral), or monetary policy (expansionary, contractionary), or stage of the business cycle (peak, trough, expansion, contraction), hence loadings may switch synchronously between several states under different scenarios. In stock return analysis, loadings measure the impact of the factor return on the expected return of each individual stock, hence loadings may switch synchronously depending on the stock market scenarios (bull versus bear markets, high versus low volatility), see for example Gu (2005) and Guidolin and Timmermann (2008) for related discussions. In bond return analysis, the yields of bonds with different maturities are well captured by the level factor, the slope factor and the curvature factor, see for example Cochrane and Piazzesi (2005) and Diebold and Li (2006). The importance of each factor could be different depending on stock market volatility, or stage of the business cycle, or unemployment rate, hence loadings may also switch synchronously according to these state variables. In general, large factor model with regime switching in loadings could also be useful for other topics, such as tracking labor productivity.

[1]For empirical evidences of parameter instability in macroeconomic and financial time series, see Banerjee, Marcellino and Masten (2008), Stock and Watson (2009) and Korobilis (2013), to mention a few.

There are only a few related results on large factor model with regime switching in loadings. Liu and Chen (2016) proposes an iterative algorithm for estimating model parameters and the hidden states based on eigen-decomposition and the Viterbi algorithm, however, the asymptotic properties of the estimated parameters are established only when the true states are known. Considering loadings as general functions of some recurrent states, Pelger and Xiong (2021) develops nonparametric kernel estimator for the loadings and factors, and establishes the relevant asymptotic theory. However, Pelger and Xiong (2021) requires observable state variables. In general, state variables may be misspecified or unobservable.

This paper proposes maximum (quasi)likelihood estimation for high dimensional factor model with regime switching in loadings when the state variables are unobservable. The model parameters are estimated jointly by EM algorithm, which in the current context only requires calculating principal components iteratively. Asymptotic properties of both the estimated parameters and the posterior probabilities of each regime are established under general assumptions.

More specifically, in the E-step, posterior probabilities of each regime are calculated based on the observed data and the parameter values at the current iteration, and the log joint likelihood of the observed data and the unobserved states are averaged with respect to the calculated regime probabilities. When state dynamics are ignored, regime probabilities at time t are inferred only from x_t (the observed time series at time t). When state dynamics are taken into account, regime probabilities at time t are inferred from all data using a recursive algorithm modified from Hamilton (1990). In the M-step, the estimated loadings for each regime are principal components of the weighted sample covariance matrix of the observed time series, where the weight on x_t equals the probability of that regime at time t. Since principal components can be easily calculated even when N (the dimension of time series) is large, our method is very easy to implement.

Ignoring state dynamics, this paper establishes the convergence rates of the estimated loading space and the estimated factor space, the limit distributions of the estimated loadings and the estimated factors, and consistency of the estimated probabilities for each regime. This paper then show that all these results are still valid when state dynamics are modeled as a Markov process and regime probabilities at each time t are inferred from all data using the proposed smoother algorithm. Consistency of the transition probability matrix and the unconditional regime probabilities are also proved. Note that asymptotic analysis under the regime switching setup is more difficult than under the structural break setup, because the pattern of regimes for the latter is much simpler.

These asymptotic results are essential in many empirical contexts. First, the limit distributions of the estimated factors allow us to construct confidence intervals for the true factors, which represent economic indices in many applications. The result on the estimated factor space implies that if estimated factors are used in factor-augmented forecasting (or factor-augmented VAR), the forecasting equation (or the VAR equation) would have induced regime switching in model parameters. Second, for asset management, the estimated loadings of each regime allow us to construct portfolios according to each specific market scenario. For structural dynamic factor analysis, consistently estimated loadings are also crucial for recovering the impulse responses. Third, consistency of the estimated probabilities for each regime implies that for each x_t, we can consistently identify which regime x_t belongs to as $N \to \infty$. For asset management, this allows us to consistently identify the current market scenario. For business cycle analysis, this allows us to consistently date turning points of business cycle and quickly detect new recessions or expansions, especially when high frequency (weekly, daily) data is utilized.

For cases with small N, various methods have been proposed for estimating factor models with regime switching. Kim (1994) proposes approximate Kalman filter for likelihood evaluation and uses nonlinear optimization for likelihood maximization. Kim and Yoo (1995) and Chauvet (1998) apply Kim (1994)'s method to a small number of economic series and obtain recession probabilities and turning points very close to the official NBER dates. Kim (1994) allows for regime switching in both factor mean and factor loadings, but when N is large, Kim (1994)'s method would be very time consuming and may have convergence problems[2]. Other methods, such as

[2]This is because the number of parameters grows proportionally to N and the likelihood function is calculated numerically and maximized by nonlinear optimization algorithm.

Diebold and Rudebusch (1996) and Kim and Nelson (1998), assume stable loadings and only focus on regime switching in factor mean. If loadings are unstable, these methods are not applicable. More importantly, if there is only regime switching in factor mean, we can not consistently identify each regime even when N is large.

In contrast with Kim (1994), our method is fast and easy to implement even when N is very large. The crucial point behind our EM algorithm is to ignore factor dynamics[3] and integrate out the factors in the likelihood function. If factors dynamics are taken into account or factors are treated as parameters in the likelihood function, the estimated loadings would not be the principal components of the weighted sample covariance matrix, and consequently both the algorithm and the asymptotic analysis would become infeasible. On the other hand, the efficiency loss of ignoring factor dynamics is small when N is large.

This paper may also contribute to the literature on dating turning points of business cycle. Currently there are two main approaches for dating business cycle using multiple time series. The first approach, aggregating then dating, is to date business cycle by focusing on a few highly aggregated time series such as GDP, industrial production and nonfarm employment. The second approach, dating then aggregating, is to date turning point in each disaggregated series and then aggregate these turning points in some appropriate way, see Burns and Mitchell (1946), Harding and Pagan (2006) and Chauvet and Piger (2008). These papers only use a small number of time series. Stock and Watson (2010, 2014) studies this issue using many time series. This paper shows that it is possible to consistently identify turning points if regime switching is synchronous and N is large enough. If N is small, consistency is not possible no matter how large T is. This paper also shows that if N is large, it is possible to consistently detect regime switching right after the turning point with only one observation, thus the speed of detection could be improved significantly. If N is small, we have to wait for enough observations from the new regime.

The rest of the paper is organized as follows. Section 2 introduces the model

[3] Factor dynamics are still allowed for the data generating process.

setup and estimation procedures ignoring state dynamics. Section 3 presents the assumptions and asymptotic results ignoring state dynamics. Section 4 introduces the estimation procedures and asymptotic results taking into account state dynamics. Section 5 presents simulation results. Section 6 presents an empirical application of the proposed method to the FRED-MD dataset. Section 7 concludes. All proofs are relegated to the appendix.

Through out the paper, $(N,T) \to \infty$ denotes N and T going to infinity jointly, $\delta_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. $\xrightarrow{p}$ and $\xrightarrow{d}$ denotes convergence in probability and convergence in distribution, respectively. For matrix A, let $\|A\|$, $\|A\|_F$, $\rho_{\max}(A)$ and $\rho_{\min}(A)$ denote its spectral norm, Frobenius norm, largest eigenvalue and smallest eigenvalue, respectively. Let $P_A = A(A'A)^{-1}A'$ denote the projection matrix and $M_A = I - P_A$. "w.p.a.1" denotes with probability approaching one.

2 Identification and Estimation

Consider the following factor model with regime switching: for $i = 1, ..., N$ and $t = 1, ..., T$,

$$x_{it} = f_t^{0\prime}\lambda_{ji}^0 + e_{it} \text{ if } z_t = j, \tag{1}$$

where λ_{ji}^0 is an r_j^0 dimensional vector of loadings for regime j, f_t^0 is an $r_{z_t}^0$ dimensional vector of factors, z_t is the state variable indicating which regime x_{it} belongs to, and e_{it} is the error term allowed to have cross-sectional and temporal dependence as well as heteroskedasticity. x_{it} is observable and all right hand side variables are unobservable. The number of regimes J^0 and the number of factors in each regime r_j^0 are fixed as $(N,T) \to \infty$ (N and T go to infinity jointly) and assumed to be known. r_j^0 is allowed to be different for different j. How to determine r_j^0 and J^0 will be studied in a separate paper.

The factor process $\{f_t^0, t = 1, ..., T\}$ is allowed to be dynamic, and similar to the principal component estimator (PCE) in Stock and Watson (2002) and Bai (2003) and the maximum likelihood estimator (MLE) in Bai and Li (2012, 2016), factor dynamics are ignored when estimating model parameters. Thus there is no need to model factor dynamics. The state process $\{z_t, t = 1, ..., T\}$ is independent with f_s^0 and e_{is} for all i and s, but is not required to be a Markov process. The estimation procedure in this section ignores the dynamics of z_t, and the corresponding asymptotic results in Section 3 are valid no matter whether z_t is dynamic or not. In Section 4, we shall assume $\{z_t, t = 1, ..., T\}$ to be Markov and take into account state dynamics in parameters estimation and smoothed inference for regime probabilities of each t. Let $q^0 = (q_1^0, ..., q_{J^0}^0)'$ denote the $(J^0 \times 1)$ vector of unconditional regime probabilities, i.e., $q_j^0 = \Pr(z_t = j)$. When $\{z_t, t = 1, ..., T\}$ is a Markov process, let Q^0 denote the $(J^0 \times J^0)$ matrix of transition probabilities and Q_{jk}^0 denote the probability of switching from state k to state j.

In vector form, the model can be written as:

$$x_t = \Lambda_j^0 f_t^0 + e_t \text{ if } z_t = j, \text{ for } t = 1, ..., T, \tag{2}$$

where $\Lambda_j^0 = (\lambda_{j1}^0, ..., \lambda_{jN}^0)'$, $x_t = (x_{1t}, ..., x_{Nt})'$ and $e_t = (e_{1t}, ..., e_{Nt})'$. Let $\Lambda^0 = (\Lambda_1^0, ..., \Lambda_{J^0}^0)$ and let $E = (e_1, ..., e_T)'$ be the $T \times N$ matrix of errors. When there are no superscripts, q, Q, Λ_j and Λ denote parameters as variables.

2.1 Identification

Since factors are unobservable, regimes are defined in terms of the linear spaces spanned by loadings. Two regimes are different if their loading spaces are different, and vice versa. More specifically, the identification condition is: for any j and k,

$$\min_t \frac{1}{N} \left\| M_{\Lambda_k^0} \Lambda_j^0 f_t^0 \right\|^2 = \min_t \frac{1}{N} f_t^{0\prime} \Lambda_j^{0\prime} M_{\Lambda_k^0} \Lambda_j^0 f_t^0 \geq C \text{ for some } C > 0. \tag{3}$$

A sufficient condition for (3) is:

$$\begin{aligned} &\lim_{N \to \infty} \frac{1}{N} \Lambda_k^{0\prime} M_{\Lambda_j^0} \Lambda_k^0 \text{ is positive definite for any } j \text{ and } k, \\ &\text{and } \min_t \|f_t\| \text{ is nonzero.} \end{aligned} \tag{4}$$

Condition (4) requires $\lim_{N \to \infty} \frac{1}{N} (\Lambda_j^0, \Lambda_k^0)'(\Lambda_j^0, \Lambda_k^0)$ to be full rank for any j and k. Thus Λ_j^0 and Λ_k^0 are not allowed to share some columns, and columns of Λ_j^0 could not be linear combination of Λ_k^0 and vice versa. An alternative sufficient condition for (3) is:

$$\lim_{N\to\infty}\frac{1}{N}\Lambda_k^{0\prime}M_{\Lambda_j^0}\Lambda_k^0 \neq 0 \text{ for any } j \text{ and } k, \qquad (5)$$
$$\text{and } \min_t \left|g_{jk}^{\prime} f_t\right| \text{ is nonzero,}$$

where g_{jk} is the eigenvector of $\lim_{N\to\infty}\frac{1}{N}\Lambda_k^{0\prime}M_{\Lambda_j^0}\Lambda_k^0$ corresponding to nonzero eigenvalue. Condition (5) only requires that the linear spaces spanned by Λ_j^0 and Λ_k^0 are different, thus Λ_j^0 and Λ_k^0 are allowed to share some columns, and some columns of Λ_j^0 are allowed to be linear combinations of Λ_k^0 and vice versa.

Note that condition (4) does not rule out the possibility that any regime j can be further decomposed into multiple regimes. Suppose the true model contains three regimes and Λ_1^0, Λ_2^0 and Λ_3^0 are linearly independent with each other. If we consider Λ_1^0 as the first regime and $(\Lambda_2^0, \Lambda_3^0)$ as the second regime, this misspecified model also satisfies condition (4). To rule out this possibility, we require that

$$plim\frac{1}{|A_j|}\sum\nolimits_{t\in A_j} f_t^0 f_t^{0\prime} \text{ is positive definite,} \qquad (6)$$

where A_j denotes any subset of $\{t : z_t = j\}$ with cardinality $|A_j|$ and $\lim\frac{|A_j|}{Tq_j^0} > 0$. If $\frac{1}{|A_j|}\sum_{t\in A_j} f_t^0 f_t^{0\prime}$ is not positive definite as $T \to \infty$ for some A_j, then A_j and $\{t : z_t = j, t \notin A_j\}$ are considered as two separate regimes.

2.2 First Order Conditions and EM Algorithm

Consider the following log-likelihood function for Gaussian mixture in covariance:

$$l(\Lambda, \sigma^2, q) = \sum\nolimits_{t=1}^{T}\log(\sum\nolimits_{j=1}^{J^0} q_j L(x_t\left|z_t = j; \Lambda_j, \sigma^2\right.)), \qquad (7)$$

where $L(x_t\left|z_t = j; \Lambda_j, \sigma^2\right.)$ is the density of x_t conditional on $z_t = j$ and evaluated at (Λ_j, σ^2), and

$$L(x_t\left|z_t = j; \Lambda_j, \sigma^2\right.) = (2\pi)^{-\frac{N}{2}}\left|\Sigma_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}x_t^{\prime}\Sigma_j{}^{-1}x_t}. \qquad (8)$$

Σ_j is the covariance matrix of x_t for regime j, and

$$\Sigma_j = \Lambda_j \Lambda_j' + \sigma^2 I_N. \tag{9}$$

The above log-likelihood function avoids estimating the factors. If factors are estimated jointly with loadings, we would not have analytical first order conditions as presented below, and consequently EM algorithm would become infeasible.

Equation (7) is a misspecified (quasi) log-likelihood function. First, similar to the principal component estimator in Stock and Watson (2002) and Bai (2003), cross-sectional and serial dependence and heteroscedasticity of the error term are ignored. We may also take into account heteroscedasticity as Doz, Giannone and Reichlin (2012) and Bai and Li (2012, 2016). With regime switching, the algorithm and asymptotic analysis would be much more complicated, but the results should be conceptually similar. Thus for simplicity, we do not pursue this direction.

Second, state dynamics are ignored. When N is large, x_t itself contains large information for z_t, thus the information in $z_1, ..., z_{t-1}$ is marginal. We shall show in Section 4.2 that asymptotic results remain the same when state dynamics are taken into account.

Third, factor dynamics are ignored. As shown in Bai (2003) for PCE and in Bai and Li (2012, 2016) for MLE, when there is no regime switching, the asymptotic properties of the estimated factors and loadings are robust to the presence of factor dynamics if both N and T are large. When there is regime switching, the asymptotic results in Section 3 are also robust to the presence of factor dynamics. More importantly, ignoring factor dynamics greatly simplifies the computation algorithm for regime switching factor models. As shown below, with factor dynamics ignored, we just need to calculate principal components iteratively. If factor dynamics are not ignored, Kim (1994)'s method would be very time consuming and may have convergence problems if N is large[4].

[4] When there is no regime switching, as suggested by Doz et al. (2012), large N factor model with factor dynamics can be calculated by the EM algorithm. However, when there are both regime switching and factor dynamics, the EM algorithm also fails. This is because in the E-step we need to calculate the likelihood for each possible state chain $z_1, ..., z_T$ and there are $(J^0)^T$ possibilities, and in the M-step numerical optimization is still needed.

Fourth, equation (7) implicitly assumes that $\mathbb{E}(f_t^0) = 0$ and $\mathbb{E}(f_t^0 f_t^{0\prime})$ is stable within each regime, and $\mathbb{E}(f_t^0 f_t^{0\prime})$ is absorbed into $\Lambda_j \Lambda_j'$ in equation (9). This is not a big issue since all results of this paper still hold when $\mathbb{E}(f_t^0) \neq 0$ and $\mathbb{E}(f_t^0 f_t^{0\prime})$ is unstable within regime, as long as Assumption 1 is satisfied.

First order conditions for Λ and σ^2

The parameters Λ, σ^2, q are estimated by maximizing $l(\Lambda, \sigma^2, q)$. The derivative of $q_j (2\pi)^{-\frac{N}{2}} |\Sigma_j|^{-\frac{1}{2}} e^{-\frac{1}{2} x_t' \Sigma_j^{-1} x_t}$ with respect to Λ_j equals itself multiplied by the derivative of $-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2} x_t' \Sigma_j^{-1} x_t$ with respect to Λ_j, and

$$\begin{aligned} \frac{\partial \log|\Sigma_j|}{\partial \Lambda_j} &= 2\Sigma_j^{-1}\Lambda_j, & (10) \\ \frac{\partial x_t' \Sigma_j^{-1} x_t}{\partial \Lambda_j} &= -2\Sigma_j^{-1} x_t x_t' \Sigma_j^{-1} \Lambda_j, & (11) \end{aligned}$$

see Chapter 14.3 in Andersen (2003) for details on calculating these derivatives. The probability of $z_t = j$ conditional on x_t and evaluated at (Λ, σ^2, q) is

$$\Pr(z_t = j \,|x_t; \Lambda, q, \sigma^2) = \frac{q_j L(x_t \,|z_t = j; \Lambda_j, \sigma^2)}{\sum_{k=1}^{J^0} q_k L(x_t \,|z_t = k; \Lambda_k, \sigma^2)}. \tag{12}$$

For simplicity, we just use p_{tj} to denote $\Pr(z_t = j \,|x_t; \Lambda, q, \sigma^2)$. It follows that

$$\frac{\partial l(\Lambda, \sigma^2, q)}{\partial \Lambda_j} = \sum\nolimits_{t=1}^{T} p_{tj}(-\Sigma_j^{-1}\Lambda_j + \Sigma_j^{-1} x_t x_t' \Sigma_j^{-1} \Lambda_j).$$

Set $\frac{\partial l(\Lambda, \sigma^2, q)}{\partial \Lambda_j}$ to 0, we have

$$\Sigma_j^{-1}\Lambda_j = \Sigma_j^{-1} S_j \Sigma_j^{-1} \Lambda_j, \tag{13}$$

where $S_j = \sum_{t=1}^T p_{tj} x_t x_t' / \sum_{t=1}^T p_{tj}$. S_j can be considered as sample covariance matrix for Σ_j based on importance sampling. The weights $p_{tj} / \sum_{t=1}^T p_{tj}$ depend on the importance of the sample x_t for regime j, the larger p_{tj} is, the more important x_t is for regime j. If true values of Λ, q, σ^2 are plugged into p_{tj}, then $\mathbb{E}(\sum_{t=1}^T p_{tj} x_t x_t') = T q_j^0 \Sigma_j$ and $\mathbb{E}(\sum_{t=1}^T p_{tj}) = T q_j^0$.

From equation (9), we have $\Sigma_j \Lambda_j = \Lambda_j(\Lambda_j'\Lambda_j + \sigma^2 I_{r_j^0})$. Left multiply $S_j \Sigma_j^{-1}$ on both sides, we have $S_j \Lambda_j = S_j \Sigma_j^{-1} \Lambda_j (\Lambda_j'\Lambda_j + \sigma^2 I_{r_j^0})$. From equation (13), we have

$\Lambda_j = S_j \Sigma_j^{-1} \Lambda_j$, thus

$$S_j \Lambda_j = \Lambda_j (\Lambda_j' \Lambda_j + \sigma^2 I_{r_j^0}). \tag{14}$$

If Λ_j is a solution for equation (14) and Λ_j^* equals post-multiplying Λ_j by the eigenvector matrix of $\Lambda_j'\Lambda_j$, then Λ_j^* is also a solution for equation (14) and $\Lambda_j^{*\prime}\Lambda_j^*$ is diagonal. Thus we can directly choose solution Λ_j with $\Lambda_j'\Lambda_j$ being diagonal. It follows that solution Λ_j are eigenvectors of S_j and $\Lambda_j'\Lambda_j + \sigma^2 I_{r_j^0}$ are the corresponding eigenvalues. We show in Appendix F that σ^2 satisfy the following condition:

$$\sigma^2 = \frac{1}{N} tr(\frac{1}{T} \sum_{t=1}^{T} x_t x_t' - \sum_{j=1}^{J^0} q_j \Lambda_j \Lambda_j'). \tag{15}$$

From the first order conditions, we can estimate the parameters as follows:

EM algorithm

Choose any q such that $q_j > 0$ for all j, e.g., $q_j = 1/J^0$. Start from randomly generated initial values of $\hat{\Lambda}^{(0)}$ and $\hat{\sigma}^{2(0)} = 1$. For $h = 0, 1, ...,$

(E-step): calculate $\hat{p}_{tj}^{(h)} = \Pr(z_t = j \,|\, x_t; \hat{\Lambda}^{(h)}, \hat{\sigma}^{2(h)}, q)$ from equation (12), and calculate $\hat{S}_j^{(h)} = \sum_{t=1}^{T} \hat{p}_{tj}^{(h)} x_t x_t' / \sum_{t=1}^{T} \hat{p}_{tj}^{(h)}$;

(M-step): given $\hat{p}_{tj}^{(h)}$ and $\hat{S}_j^{(h)}$, $\hat{\Lambda}_j^{(h+1)}$ are eigenvectors of $\hat{S}_j^{(h)}$ and normalize $\hat{\Lambda}_j^{(h+1)}$ such that $\hat{\Lambda}_j^{(h+1)\prime}\hat{\Lambda}_j^{(h+1)} + \hat{\sigma}^{2(h+1)} I_{r_j^0}$ are the corresponding eigenvalues, and $\hat{\sigma}^{2(h+1)} = \frac{1}{N} tr(\frac{1}{T}\sum_{t=1}^{T} x_t x_t' - \sum_{j=1}^{J^0} q_j \hat{\Lambda}_j^{(h+1)} \hat{\Lambda}_j^{(h+1)\prime})$. Note that computation of $\hat{\Lambda}_j^{(h+1)}$ and $\hat{\sigma}^{2(h+1)}$ requires iteration because $\hat{\Lambda}_j^{(h+1)}$ depends on $\hat{\sigma}^{2(h+1)}$ and vice versa.

Iterate the E-step and the M-step until converge. Let $\hat{\Lambda}_j = (\hat{\lambda}_{j1}, ..., \hat{\lambda}_{jN})'$, $\hat{\Lambda} = (\hat{\Lambda}_1, ..., \hat{\Lambda}_{J^0})$ and $\hat{\sigma}^2$ denote the estimated parameters, and $\hat{p}_{tj} = \Pr(z_t = j \,|\, x_t; \hat{\Lambda}, \hat{\sigma}^2, q)$ denote the estimated regime probabilities.

The asymptotic results in Section 3 hold for any q as long as $q_j > 0$ for all j. Here we choose $q_j = 1/J^0$ for all j. Once we have $\hat{\Lambda}$ and $\hat{\sigma}^2$, q_j^0 can be estimated by

$$\hat{q}_j = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_{tj}. \tag{16}$$

We may also replace $q_j = 1/J^0$ by $q_j = \hat{q}_j$ and repeat the EM algorithm. If we iterate between $(\hat{\Lambda}, \hat{\sigma}^2)$ and $\hat{q}_j$ until convergence, this turns out to be the maximum likelihood estimator when q is estimated jointly with Λ and σ^2 because equation (16) is the first order condition for q, see Appendix F.

The asymptotic results in Section 3 also holds for any $\hat{\sigma}^2$ as long as $\hat{\sigma}^2$ is bounded and bounded away from zero in probability. Consistency of $\hat{\sigma}^2$ is not needed. When implementing the above algorithm, $\hat{\sigma}^2$ will be restricted in $[\frac{1}{C^2}, C^2]$ for some C large enough such that the true σ^2 lies in $[\frac{1}{C^2}, C^2]$. We may also simply fix down $\hat{\sigma}^2 = 1$ to avoid the iteration between $\hat{\Lambda}_j^{(h+1)}$ and $\hat{\sigma}^{2(h+1)}$. This only affects the Euclidean norm of $\hat{\Lambda}_j^{(h+1)}$.

Remark 1 *Since S_j is a weighted average of $x_t x_t'$, Λ_j can be considered as a GLS estimator. This is related to the GLS estimation for factors in Breitung and Tenhofen (2011) and Choi (2012). Here the weights are $p_{tj} = \Pr(z_t = j \,|x_t; \Lambda, q, \sigma^2)$, while in those papers the weights are inverse of error variances.*

2.3 Estimate the Factors

Let $1_{z_t=j}$ denote the indicator function, i.e., $1_{z_t=j} = 1$ if $z_t = j$ and 0 otherwise. Ignoring factor dynamics, conditioning on x_t, the conditional expectation of f_t is

$$\hat{f}_t = \sum\nolimits_{j=1}^{J^0} \mathbb{E}(f_t \,\Big|\, x_t, z_t = j; \hat{\Lambda}_j, \hat{\sigma}^2)\hat{p}_{tj} = \sum\nolimits_{j=1}^{J^0} \hat{\Lambda}_j'(\hat{\Lambda}_j\hat{\Lambda}_j' + \hat{\sigma}^2 I_N)^{-1} x_t \hat{p}_{tj}. \tag{17}$$

If factor dynamics are taken into account, conditioning on $x_{1:t} \equiv (x_1, ..., x_t)$,

$$\hat{f}_t = \sum\nolimits_{z_1=1}^{J^0} ... \sum\nolimits_{z_t=1}^{J^0} \mathbb{E}(f_t \,\Big|\, x_{1:t}, z_{1:t}; \hat{\Lambda}, \hat{\sigma}^2) \Pr(z_{1:t} \,\Big|\, x_{1:t}; \hat{\Lambda}, \hat{\sigma}^2, q),$$

which is formidable since we need to calculate $(J^0)^t$ probabilities. For large N, the benefit of considering factor dynamics is marginal and is outweighed by the computational simplicity of ignoring factor dynamics.

3 Asymptotic Results

We assume the following conditions hold as $(N, T) \to \infty$.

Assumption 1 *(1) For* $j = 1, ..., J^0$*,* $\frac{1}{Tq_j^0}\sum_{t=1}^{T} f_t^0 f_t^{0\prime} 1_{z_t=j} \xrightarrow{p} \Sigma_{F_j}$ *for some positive definite* Σ_{F_j}*, and* $plim \frac{1}{|A_j|}\sum_{t\in A_j} f_t^0 f_t^{0\prime}$ *is also positive definite, where* A_j *is defined in section 2.1.*

(2) For some $\alpha > 16$*, there exists* $M > 0$ *such that* $\mathbb{E}(\|f_t^0\|^{\alpha}) \leq M$ *for all* t*.*

Assumption 1(1) rules out the possibility that for regime j, the subsample $\{t : z_t = j\}$ can be further decomposed into multiple regimes, see the discussion in Section 2.1. Assumption 1(1) allows the factor process to be dynamic such that $C(L)f_t = \epsilon_t$. Assumption 1(2) assumes that factors have bounded moments.

Assumption 2 *(1) For* $j = 1, ..., J^0$*,* $\frac{1}{N}\Lambda_j^{0\prime}\Lambda_j^0 \to \Sigma_{\Lambda_j}$ *for some positive definite* Σ_{Λ_j} *and* $\left\|\lambda_{ji}^0\right\| \leq M$ *for any* $i = 1, ..., N$*.*

(2) For any $j = 1, ..., J^0$ *and* $k = 1, ..., J^0$*,* $\min_t \frac{1}{N} f_t^{0\prime}\Lambda_j^{0\prime} M_{\Lambda_k^0}\Lambda_j^0 f_t^0 \geq C$ *for some* $C > 0$*.*

Assumption 2(1) ensures that within each regime, each factor has a nontrivial contribution. $\left\|\lambda_{ji}^0\right\|$ is assumed to be uniformly bounded over i. Assumption 2(2) is the identification condition for determining which regime each x_t belongs to, see Section 2.1 for more details.

Assumption 3 *(1)* $\mathbb{E}(e_{it}) = 0$*,* $\mathbb{E}(e_{it}^{\alpha}) \leq M$ *for some* $\alpha > 16$*.*

(2) $\sum_{k=1}^{N} \tau_{ik} \leq M$ *for any* i*, where* $\mathbb{E}(e_{it}e_{kt}) = \tau_{ik,t}$ *with* $|\tau_{ik,t}| \leq \tau_{ik}$ *for some* $\tau_{ik} > 0$ *and for all* t*.*

(3) $\sum_{s=1}^{T} \gamma_{ts} \leq M$ *for all* t*, where* $E(e_{it}e_{is}) = \gamma_{i,ts}$ *with* $\left|\gamma_{i,ts}\right| \leq \gamma_{ts}$ *for some* $\gamma_{ts} > 0$ *and for all* i*.*

(4) $\mathbb{E}(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(e_{it}e_{kt} - \mathbb{E}(e_{it}e_{kt}))1_{z_t=j}\right\|^2) \leq M$ *for all* $i = 1, ..., N$*,* $k = 1, ..., N$ *and* $j = 1, ..., J^0$*.*

Assumption 3 allows the error term to have limited cross-sectional and serial dependence as well as heteroscedasticity. These conditions are conventional in the literature for approximate factor model.

Assumption 4 *For* $j = 1, ..., J^0$*,* $\frac{1}{T}\sum_{t=1}^{T} 1_{z_t=j} \xrightarrow{p} q_j^0$ *and* $0 < q_j^0 < 1$*.*

Note that Assumption 4 does not require z_t to follow a Markov process, but if z_t is Markov, we can calculate smoothed estimates of probabilities of each regime for each x_t.

Assumption 5 *(1)For some* $\beta \geq 2$, $\mathbb{E}(\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_{ji}^{0} e_{it} \right\|^{\beta}) \leq M$ *for all* $j = 1, ..., J^0$ *and all* t.

(2) $\mathbb{E}(\left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_t^0 e_{it} 1_{z_t=j} \right\|^2) \leq M$ *for all* $j = 1, ..., J^0$ *and all* i.

Assumption 5(1) assumes that errors are weakly correlated across i for each t. When $\beta = 2$, Assumption 5(1) is implied by Assumptions 2(1), 3(1) and 3(2). Assumption 5(2) assumes that errors are weakly correlated across t for each i. Assumption 5(2) is implied by Assumptions 1(2), 3(1) and 3(4) if we further assume factors are nonrandom or independent with errors.

Assumption 6 *For each* $j = 1, ..., J^0$*, the eigenvalues of* $\Sigma_{\Lambda_j}^{\frac{1}{2}} \Sigma_{F_j} \Sigma_{\Lambda_j}^{\frac{1}{2}}$ *are different.*

With Assumption 6, loadings and factors are identifiable up to a rotation. For identification of the loading space and factor space, Assumption 6 is not needed.

Assumption 7 *(1)* $\mathbb{E}(\left\| \frac{1}{\sqrt{NT}} \sum_{k=1}^{N} \sum_{t=1}^{T} \lambda_i^0 (e_{it} e_{kt} - \mathbb{E}(e_{it} e_{kt})) 1_{z_t=j} \right\|^2) \leq M$ *for all* $i = 1, ..., N$ *and* $j = 1, ..., J^0$*; and* $\mathbb{E}(\left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (e_{it} e_{is} - \mathbb{E}(e_{it} e_{is})) f_t^0 1_{z_t=j} \right\|^2) \leq M$ *for all* $s = 1, ..., T$ *and* $j = 1, ..., J^0$.

(2) $\mathbb{E}(\left\| \frac{1}{\sqrt{NT}} \sum_{k=1}^{N} \sum_{t=1}^{T} \lambda_k^0 f_t^{0\prime} e_{kt} 1_{z_t=j} \right\|^2) \leq M$ *for* $j = 1, ..., J^0$.

(3) Define $\Phi_{ji} = plim \frac{1}{T} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{E}(f_t^0 f_s^{0\prime} e_{is} e_{it} 1_{z_s=j} 1_{z_t=j})$*. For* $j = 1, ..., J^0$, $\frac{1}{\sqrt{T q_j^0}} \sum_{t=1}^{T} f_t^0 e_{it} 1_{z_t=j} \xrightarrow{d} \mathcal{N}(0, \Phi_{ji})$.

(4) Define $\Gamma_{jt} = lim \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_{ji}^0 \lambda_{jk}^0 \mathbb{E}(e_{it} e_{kt})$*. For* $j = 1, ..., J^0$, $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_{ji}^0 e_{it} \xrightarrow{d} \mathcal{N}(0, \Gamma_{jt})$.

Assumption 7 is conventional, part (3) and part (4) are just central limit theorems and will be used for deriving the limit distributions of the estimated factors and loadings.

Consistency of estimated loading space

Theorem 1 *Under Assumptions 1, 2(1), 3 and 4,* $\frac{1}{N}\left\|M_{\hat{\Lambda}_j}\Lambda_j^0\right\|_F^2 = O_p(\frac{1}{\sqrt{\delta_{NT}}})$ *for each* j *as* $(N,T)\to\infty$.

Theorem 1 shows that the estimated loading space is consistent without observing the state variable z_t. Note that the estimated loadings $\hat{\Lambda}_j$ and the estimated regime probabilities $\hat{p}_{tj}$ depend on each other, thus we can not use standard technique in the literature, e.g., Bai (2003), for analyzing $\hat{\Lambda}_j$. The crucial point is to utilize large N, see the Appendix for more details. Based on Theorem 1, we can show that the estimated regime probabilities are consistent.

Consistency of estimated regime probabilities

Theorem 2 *Under Assumptions 1-4 and 5(1), as* $(N,T)\to\infty$, *for each* j *and for any* $\eta>0$,

(1) $\sup_t|\hat{p}_{tj}-1_{z_t=j}| = o_p(\frac{1}{N^\eta})$ *if* $T^{\frac{16}{\alpha}}/N\to 0$ *and* $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N\to 0$,

(2) $|\hat{p}_{tj}-1_{z_t=j}| = o_p(\frac{1}{N^\eta})$.

Note that η could be arbitrarily large, and α and β could also be large as long as Assumptions 1(2), 3(1) and 5(1) are satisfied. Theorem 2 shows that $\hat{p}_{tj}$ is consistent as $N\to\infty$ and is uniformly consistent if T is relatively small compared to N. The proof utilizes the exponential likelihood ratio.

Theorem 2 implies that we can consistently identify which regime x_t belongs to for all t, if there is common regime switching in loadings and the dimension of x_t tends to infinity. Theorem 2 also implies that we can consistently detect regime switching right after the turning point with only one observation, so that we do not need to wait for many observations of the time series from the new regime. This could improve the speed of detection of new turning points, especially when high frequency data is used.

Convergence rate of estimated loading space

If the true states z_t were known, asymptotic properties of the estimated loadings and factors are straightforward. Based on Theorem 2, we shall show that using estimated regime probabilities does not affect the asymptotic results. First, define $W_{jNT} = (\frac{\hat{\Lambda}_j'\hat{\Lambda}_j}{N} + \frac{\hat{\sigma}^2}{N} I_{r_j^0})(\frac{1}{T}\sum_{t=1}^T \hat{p}_{tj})$ and $H_j = \frac{\sum_{t=1}^T f_t^0 f_t^{0\prime} 1_{z_t=j}}{T} \frac{\Lambda_j^{0\prime}\hat{\Lambda}_j}{N} W_{jNT}^{-1}$, then we have:

Proposition 1 *Let V_j be a $r_j^0 \times r_j^0$ diagonal matrix consisting of eigenvalues of $\Sigma_{\Lambda_j}^{\frac{1}{2}} \Sigma_{F_j} \Sigma_{\Lambda_j}^{\frac{1}{2}}$ in descending order and Υ_j be the corresponding eigenvectors. Under Assumptions 1-6, and assume $T^{\frac{16}{\alpha}}/N \to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0$, as $(N,T) \to \infty$,*

(1) $W_{jNT} \xrightarrow{p} q_j^0 V_j$ for each j,

(2) $H_j \xrightarrow{p} \Sigma_{\Lambda_j}^{-\frac{1}{2}} \Upsilon_j V_j^{\frac{1}{2}}$ for each j.[5]

Proposition 1 is an important auxiliary result. Note that the proof strategy of Proposition 1 is slightly different from Bai (2003)'s framework, because here the first order condition (14) does not tells us whether $\hat{\Lambda}_j$ is eigenvectors corresponding to the largest r_j^0 eigenvalues of S_j or not[6].

Theorem 3 *Under Assumptions 1-6, and assume $T^{\frac{16}{\alpha}}/N \to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0$, as $(N,T) \to \infty$, $\frac{1}{N}\left\|\hat{\Lambda}_j - \Lambda_j^0 H_j\right\|_F^2 = O_p(\frac{1}{\delta_{NT}^2})$ for each j.*

Theorem 3 establishes the convergence rate of the estimated loading space for each regime. This could help us study the effect of using estimated loadings on subsequent applications. For example, if estimated loadings are used to construct portfolios, Theorem 3 could help us calculate how the estimation error contained in $\hat{\Lambda}_j$ would affect the performance of these portfolios.

Limit distributions of estimated loadings

Theorem 4 *Under Assumptions 1-7, and assume $\sqrt{T}/N \to 0$, $T^{\frac{16}{\alpha}}/N \to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0$, as $(N,T) \to \infty$, $\sqrt{Tq_j^0}(\hat{\lambda}_{ji} - H_j'\lambda_{ji}^0) \xrightarrow{d} \mathcal{N}(0, V_j^{-\frac{1}{2}} \Upsilon_j' \Sigma_{\Lambda_j}^{\frac{1}{2}} \Phi_{ji} \Sigma_{\Lambda_j}^{\frac{1}{2}} \Upsilon_j V_j^{-\frac{1}{2}})$ for each j.*

Theorem 4 shows that for each j and i, $\hat{\lambda}_{ji}$ has a limiting normal distribution. This allows us to construct confidence interval for the estimated loadings. Also note that the rotation matrix H_j is different for different regime.

[5] H_j corresponds to $(H^{-1})'$ for the rotation matrix H in Bai (2003).

[6] Because of this, we can not use Weyl's inequality. Bai (2003)'s proof of Proposition 1 relies on his Lemma A.3(i), which relies on Weyl's inequality. Our proof utilizes Theorem 1 and makes some modifications on Bai (2003)'s framework.

Remark 2 *$\hat{q}_j \xrightarrow{p} q_j^0$ is proved in the proof of Proposition 1(2). We can also prove consistency and limit distribution of $\hat{\sigma}^2$ (the probability limit of $\hat{\sigma}^2$ is $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2$), we omit it since this is not our focus.*

Asymptotic properties of estimated factors

Theorem 5 *Under Assumptions 1-7, and assume $\sqrt{N}/T \to 0$, $T^{\frac{16}{\alpha}}/N \to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0$, as $(N,T) \to \infty$,*

(1) $\frac{1}{T}\sum_{t=1}^{T} \left\| \hat{f}_t - H_{z_t}^{-1} f_t^0 \right\|^2 = O_p(\frac{1}{\delta_{NT}^2})$,

(2) $\sqrt{N}(\hat{f}_t - H_{z_t}^{-1} f_t^0) \xrightarrow{d} \mathcal{N}(0, V_{z_t}^{-\frac{1}{2}} \Upsilon_{z_t}' \Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}} \Gamma_{z_t t} \Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}} \Upsilon_{z_t} V_{z_t}^{-\frac{1}{2}})$.

Theorem 5(2) shows that the limit distribution of $\hat{f}_t$ is mixed normal, since the rotation matrix $H_{z_t}^{-1}$ and the asymptotic variance depend on the state variable z_t. Theorem 5(1) establishes the convergence rate of the estimated factor space. Note that if $\{\hat{f}_t, t = 1, ...T\}$ is used as proxies for the true factors in factor-augmented forecasting (or factor-augmented VAR), the forecasting equation (or the VAR equation) would have induced regime switching in model parameters, because $H_{z_t}^{-1}$ depends on z_t. For illustration, consider the following h-period ahead forecasting model using factors and some other observable variables W_t:

$$y_{t+h} = a' f_t^0 + b' W_t + u_{t+h}.$$

If $\hat{f}_t$ is used as proxies for f_t^0, the model can be written as

$$y_{t+h} = -a' H_{z_t}(\hat{f}_t - H_{z_t}^{-1} f_t^0) + a' H_{z_t} \hat{f}_t + b' W_t + u_{t+h}.$$

The first term on the right hand side is asymptotically negligible. It is easy to see that the coefficient $a' H_{z_t}$ depends on z_t and this need to be taken into account when we estimate the forecasting equation.

4 MLE with State Dynamics

In this section we assume state process z_t to be Markov and take into account the dynamics of z_t in the EM algorithm and the asymptotic analysis. Since the state process of the business cycle/stock market is highly persistent, regime switching models that capture the persistence should perform significantly better. In addition, taking into account state dynamics would also be advantageous for mixed frequency data or ragged edge data (data released at non-synchronized dates), since the number of series available is small at early times.

4.1 Algorithm for MLE with State Dynamics

Let $\phi = (\phi_1, ..., \phi_{J^0})'$ denote the initial probabilities of z_1. With state dynamics taken into account, the log-likelihood function is:

$$l(\Lambda, \sigma^2, Q, \phi) = \log[\sum_{z_T=1}^{J^0} ... \sum_{z_1=1}^{J^0} \prod_{t=1}^{T} L(x_t \left| z_t; \Lambda, \sigma^2 \right.) \Pr(z_1, ..., z_T \left| Q, \phi \right.)], \quad (18)$$

where $\prod_{t=1}^{T} L(x_t \left| z_t; \Lambda, \sigma^2 \right.)$ is the density of $(x_1, ..., x_T)$ conditioning on $(z_1, ..., z_T)$ and $\Pr(z_1, ..., z_T \left| Q, \phi \right.)$ is the probability of $(z_1, ..., z_T)$.

First order conditions for Λ and σ^2

Let $p_{tj|T} = \Pr(z_t = j \left| x_{1:T}; \Lambda, \sigma^2, Q, \phi \right.)$.

$$\frac{\partial l(\Lambda, \sigma^2, Q, \phi)}{\partial \Lambda_j} = \sum_{t=1}^{T} \frac{\partial \log L(x_t \left| z_t; \Lambda, \sigma^2 \right.)}{\partial \Lambda_j} p_{tj|T} = 0, \quad (19)$$

$$\frac{\partial l(\Lambda, \sigma^2, Q, \phi)}{\partial \sigma^2} = \sum_{t=1}^{T} \sum_{j=1}^{J^0} \frac{\partial \log L(x_t \left| z_t; \Lambda, \sigma^2 \right.)}{\partial \sigma^2} p_{tj|T} = 0. \quad (20)$$

Recall that $x_{1:t} = (x_1, ..., x_t)$ and $z_{1:t} = (z_1, ..., z_t)$. For any given Q and ϕ, let $\Pr(z_1, ..., z_T \left| x_{1:T}; \tilde{\theta}^{(h)} \right.)$ be the probability of $z_{1:T}$ conditioning on $x_{1:T}$ evaluated at $\tilde{\theta}^{(h)} = (\tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)}, Q, \phi)$. At the h-th iteration, the EM algorithm maximizes the expectation of the log-likelihood of $(x_{1:T}, z_{1:T})$ with respect to $\Pr(z_1, ..., z_T \left| x_{1:T}; \tilde{\theta}^{(h)} \right.)$, i.e.,

$$\begin{aligned} l^{(h)}(\Lambda, \sigma^2, Q, \phi) \equiv & \sum_{z_T=1}^{J^0} ... \sum_{z_1=1}^{J^0} \log[\prod_{t=1}^{T} L(x_t \left| z_t; \Lambda, \sigma^2 \right.) \Pr(z_1, ..., z_T \left| Q, \phi \right.)] \\ & \Pr(z_1, ..., z_T \left| x_{1:T}; \tilde{\theta}^{(h)} \right.). \end{aligned}$$

Since z_t is a Markov process, $\Pr(z_1, ..., z_T \,|Q, \phi) = \Pr(z_1 \,|\phi) \prod_{t=2}^{T} \Pr(z_t \,|z_{t-1}; Q)$. Thus

$$\begin{aligned} l^{(h)}(\Lambda, \sigma^2, Q, \phi) &= \sum_{z_T=1}^{J^0} ... \sum_{z_1=1}^{J^0} [\sum_{t=1}^{T} \log L(x_t \,\big|z_t; \Lambda, \sigma^2) \\ &\quad + \sum_{t=2}^{T} \log \Pr(z_t \,|z_{t-1}; Q) + \log \Pr(z_1 \,|\phi)] \Pr(z_1, ..., z_T \,\Big| x_{1:T}; \tilde{\theta}^{(h)}) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{J^0} \log L(x_t \,\big|z_t = j; \Lambda_j, \sigma^2) \tilde{p}_{tj|T}^{(h)} \\ &\quad + \sum_{t=2}^{T} \sum_{j=1}^{J^0} \sum_{k=1}^{J^0} \log Q_{jk} \tilde{p}_{tjk|T}^{(h)} + \sum_{k=1}^{J^0} \log \phi_k \tilde{p}_{1k|T}^{(h)}, \end{aligned} \tag{21}$$

where $\tilde{p}_{tjk|T}^{(h)} = \Pr(z_t = j, z_{t-1} = k \,\Big| x_{1:T}; \tilde{\theta}^{(h)})$ and $\tilde{p}_{tj|T}^{(h)} = \Pr(z_t = j \,\Big| x_{1:T}; \tilde{\theta}^{(h)}) = \sum_{k=1}^{J^0} \tilde{p}_{tjk|T}^{(h)}$ are smoothed probabilities based on $x_{1:T}$ and $\tilde{\theta}^{(h)}$. Appendix G presents a recursive algorithm for calculating $\tilde{p}_{tjk|T}^{(h)}$.

From equations (10) and (11), $\frac{\partial \log L(x_t \big| z_t = j; \Lambda_j, \sigma^2)}{\partial \Lambda_j} = -\Sigma_j^{-1}\Lambda_j + \Sigma_j^{-1} x_t x_t' \Sigma_j^{-1} \Lambda_j$. Thus

$$\frac{\partial \sum_{t=1}^{T} \log L(x_t \,|z_t = j; \Lambda_j, \sigma^2) \tilde{p}_{tj|T}^{(h)}}{\partial \Lambda_j} = \sum_{t=1}^{T} (-\Sigma_j^{-1}\Lambda_j + \Sigma_j^{-1} x_t x_t' \Sigma_j^{-1} \Lambda_j) \tilde{p}_{tj|T}^{(h)} = 0.$$

Let $\tilde{S}_j^{(h)} = \sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)} x_t x_t' / \sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)}$, then we have

$$\Sigma_j^{-1}\Lambda_j = \Sigma_j^{-1} \tilde{S}_j^{(h)} \Sigma_j^{-1} \Lambda_j. \tag{22}$$

Compare $\tilde{S}_j^{(h)}$ with S_j in Section 2.2, we can see the difference is that p_{tj} is replaced by the smoothed estimates $\tilde{p}_{tj|T}^{(h)}$. Similar to equation (14), equation (22) implies that

$$\tilde{S}_j^{(h)} \tilde{\Lambda}_j^{(h+1)} = \tilde{\Lambda}_j^{(h+1)} (\tilde{\Lambda}_j^{(h+1)\prime} \tilde{\Lambda}_j^{(h+1)} + \tilde{\sigma}^{2(h+1)} I_{r_j^0}), \tag{23}$$

thus $\tilde{\Lambda}_j^{(h+1)}$ are eigenvectors of $\tilde{S}_j^{(h)}$ and $\tilde{\Lambda}_j^{(h+1)\prime} \tilde{\Lambda}_j^{(h+1)} + \tilde{\sigma}^{2(h+1)} I_{r_j^0}$ are the corresponding eigenvalues. To save space, we show in Appendix F that

$$\tilde{\sigma}^{2(h+1)} = \frac{1}{N} tr(\frac{1}{T} \sum_{t=1}^{T} x_t x_t' - \sum_{j=1}^{J^0} \frac{1}{T} \sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)} \tilde{\Lambda}_j^{(h+1)} \tilde{\Lambda}_j^{(h+1)\prime}). \tag{24}$$

Remark 3 *The second equality of equation (21) is crucial. Since factor dynamics are ignored,* $L(x_{1:T} \,|z_{1:T}; \Lambda, \sigma^2) = \prod_{t=1}^{T} L(x_t \,|z_t; \Lambda, \sigma^2)$*, thus we only need to calculate* $\tilde{p}_{tj|T}^{(h)}$ *rather than the probability of the whole chain* $\Pr(z_1, ..., z_T \,\Big| x_{1:T}; \tilde{\theta}^{(h)})$*. The latter* *requires $(J^0)^T$ calculations, which is hopeless when T is large. If factor dynamics are not ignored, then $L(x_{1:T}\,|z_{1:T};\Lambda,\sigma^2) = L(x_1\,|z_{1:T};\Lambda,\sigma^2)\prod_{t=2}^T L(x_t\,|x_{1:t-1},z_{1:T};\Lambda,\sigma^2)$. $L(x_t\,|x_{1:t-1},z_{1:T};\Lambda,\sigma^2)$ depends on the chain $(z_1,...,z_T)$ through $z_{1:t}$, thus we need to calculate $\Pr(z_{1:t}\left|x_{1:T};\tilde{\theta}^{(h)}\right.)$. This requires $(J^0)^t$ calculations, which is hopeless when t is large.*

EM algorithm with state dynamics

Choose any Q and ϕ such that $Q_{jk} > 0$ for any j and k and $\phi_k > 0$ for all k, e.g., $Q_{jk} = 1/J^0$ and $\phi_k = 1/J^0$. Start from randomly generated initial values of $\tilde{\Lambda}^{(0)}$ and $\tilde{\sigma}^{2(0)} = 1$. For $h = 0, 1, ...,$

(E-step): calculate $\tilde{p}^{(h)}_{tjk|T}$ using the algorithm in Appendix G, and calculate $\tilde{S}^{(h)}_j = \sum_{t=1}^T \tilde{p}^{(h)}_{tj|T}\, x_t x_t' / \sum_{t=1}^T \tilde{p}^{(h)}_{tj|T}$ with $\tilde{p}^{(h)}_{tj|T} = \sum_{k=1}^{J^0} \tilde{p}^{(h)}_{tjk|T}$;

(M-step): given $\tilde{p}^{(h)}_{tjk|T}$ and $\tilde{S}^{(h)}_j$, calculate $\tilde{\Lambda}^{(h+1)}_j$ as *eigenvectors of $\tilde{S}^{(h)}_j$ and normalize $\tilde{\Lambda}^{(h+1)}_j$ such that $\tilde{\Lambda}^{(h+1)\prime}_j \tilde{\Lambda}^{(h+1)}_j + \tilde{\sigma}^{2(h+1)} I_{r_j^0}$ are the corresponding eigenvalues and equation (24) is satisfied. Note that the computation of $\tilde{\Lambda}^{(h+1)}_j$ and $\tilde{\sigma}^{2(h+1)}$ requires iteration between equations (23) and (24).*

Iterate the E-step and the M-step until converge. Let $\tilde{\Lambda}_j = (\tilde{\lambda}_{j1}, ..., \tilde{\lambda}_{jN})'$, $\tilde{\Lambda} = (\tilde{\Lambda}_1, ..., \tilde{\Lambda}_{J^0})$ and $\tilde{\sigma}^2$ denote the estimated parameters.

The asymptotic results in Section 4.2 hold for any Q and ϕ as long as $q_j > 0$ for any j and $Q_{jk} > 0$ for any j and k. Here we choose $Q_{jk} = 1/J^0$ and $\phi_k = 1/J^0$. Based on $\tilde{\Lambda}$ and $\tilde{\sigma}^2$, Q^0_{jk} and ϕ^0_k can be estimated by

$$\tilde{Q}_{jk} = \sum\nolimits_{t=2}^T \tilde{p}_{tjk|T} \Big/ \sum\nolimits_{j=1}^{J^0} \sum\nolimits_{t=2}^T \tilde{p}_{tjk|T}\,, \tag{25}$$

$$\tilde{\phi}_k = \tilde{p}_{1k|T} = \sum\nolimits_{j=1}^{J^0} \tilde{p}_{2jk|T}\,, \tag{26}$$

where $\tilde{p}_{tjk|T} = \Pr(z_t = j, z_{t-1} = k\left|x_{1:T};\tilde{\Lambda},\tilde{\sigma}^2,Q,\phi\right.)$. We may also plug $\tilde{Q}_{jk}$ and $\tilde{\phi}_k$ back in the EM algorithm and iterate between $(\tilde{\Lambda},\tilde{\sigma}^2)$ and $(\tilde{Q},\tilde{\phi})$ until convergence. This is the maximum likelihood estimator when (Q,ϕ) is estimated jointly with (Λ,σ^2), see Appendix F for details. The results in Section 4.2 are for $(\tilde{\Lambda},\tilde{\sigma}^2)$ with no iteration.

The asymptotic results in Section 4.2 are also valid as long as $\tilde{\sigma}^2$ is bounded and bounded away from zero in probability. Thus similar to the EM algorithm in Section 2.2, we will restrict $\tilde{\sigma}^2$ in $[\frac{1}{C^2}, C^2]$ for some large C or simply fix down $\tilde{\sigma}^2 = 1$ to avoid the iteration between $\tilde{\Lambda}_j^{(h+1)}$ and $\tilde{\sigma}^{2(h+1)}$.

Estimate the factors

The factors can be estimated by the expectation of f_t conditioning on $x_{1:T}$ and ignoring factor dynamics:

$$\begin{aligned}
\tilde{f}_t &= \sum\nolimits_{j=1}^{J^0} \mathbb{E}(f_t \,\Big|\, x_{1:T}, z_t = j; \tilde{\Lambda}_j, \tilde{\sigma}^2)\tilde{p}_{tj|T} = \sum\nolimits_{j=1}^{J^0} \mathbb{E}(f_t \,\Big|\, x_t, z_t = j; \tilde{\Lambda}_j, \tilde{\sigma}^2)\tilde{p}_{tj|T} \\
&= \sum\nolimits_{j=1}^{J^0} \tilde{\Lambda}_j'(\tilde{\Lambda}_j\tilde{\Lambda}_j' + \tilde{\sigma}^2 I_N)^{-1} x_t \tilde{p}_{tj|T}. \qquad (27)
\end{aligned}$$

4.2 Asymptotic Results for MLE with State Dynamics

For $\tilde{\Lambda}_j$, $\tilde{p}_{tj|T}$ and $\tilde{f}_t$, the asymptotic results in Section 3 hold under the same assumptions. We summarize these results in the following theorem.

Theorem 6 *(i) Under the same assumptions as Theorem 1,* $\frac{1}{N}\left\|M_{\tilde{\Lambda}_j}\Lambda_j^0\right\|_F^2 = O_p(\frac{1}{\sqrt{\delta_{NT}}})$ *as* $(N,T)\to\infty$*,*

(ii) Under the same assumptions as Theorem 2, as $(N,T)\to\infty$*,*

$$\begin{aligned}
\left|\tilde{p}_{tj|T} - 1_{z_t=j}\right| &= o_p(\frac{1}{N^\eta}), \\
\sup_t \left|\tilde{p}_{tj|T} - 1_{z_t=j}\right| &= o_p(\frac{1}{N^\eta}) \text{ if } T^{\frac{16}{\alpha}}/N \to 0 \text{ and } T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0,
\end{aligned}$$

(iii) Under the same assumptions as Proposition 1, as $(N,T)\to\infty$*,*

$$\begin{aligned}
\bar{W}_{jNT} &= (\frac{\tilde{\Lambda}_j'\tilde{\Lambda}_j}{N} + \frac{\tilde{\sigma}^2}{N} I_{r_j^0})(\frac{1}{T}\sum\nolimits_{t=1}^{T}\tilde{p}_{tj|T}) \xrightarrow{p} q_j^0 V_j, \\
\bar{H}_j &= \frac{\sum_{t=1}^T f_t^0 f_t^{0\prime} 1_{z_t=j}}{T}\frac{\Lambda_j^{0\prime}\tilde{\Lambda}_j}{N}\bar{W}_{jNT}^{-1} \xrightarrow{p} \Sigma_{\Lambda_j}^{-\frac{1}{2}}\Upsilon_j V_j^{\frac{1}{2}},
\end{aligned}$$

(iv) Under the same assumptions as Theorem 3, as $(N,T)\to\infty$*,*

$$\frac{1}{N}\left\|\tilde{\Lambda}_j - \Lambda_j^0\bar{H}_j\right\|_F^2 = O_p(\frac{1}{\delta_{NT}^2}),$$

(v) Under the same assumptions as Theorem 4, as $(N,T)\to\infty$,

$$\sqrt{Tq_j^0}(\tilde{\lambda}_{ji}-\bar{H}_j'\lambda_{ji}^0)\xrightarrow{d}\mathcal{N}(0,V_j^{-\frac{1}{2}}\Upsilon_j'\Sigma_{\Lambda_j}^{\frac{1}{2}}\Phi_{ji}\Sigma_{\Lambda_j}^{\frac{1}{2}}\Upsilon_jV_j^{-\frac{1}{2}}),$$

(vi) Under the same assumptions as Theorem 5, as $(N,T)\to\infty$,

$$\frac{1}{T}\sum\nolimits_{t=1}^{T}\left\|\tilde{f}_t-\bar{H}_{z_t}^{-1}f_t^0\right\|^2=O_p(\frac{1}{\delta_{NT}^2}),$$

$$\sqrt{N}(\tilde{f}_t-\bar{H}_{z_t}^{-1}f_t^0)\xrightarrow{d}\mathcal{N}(0,V_{z_t}^{-\frac{1}{2}}\Upsilon_{z_t}'\Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}}\Gamma_{z_tt}\Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}}\Upsilon_{z_t}V_{z_t}^{-\frac{1}{2}}).$$

See the Appendix for the proof of Theorem 6. Theorem 6 shows that taking into account dynamics of z_t does not affect the asymptotic properties. This is because when N is large, the information contained in state dynamics for estimating regime probabilities becomes marginal.

Note that although the design of the smoothed algorithm utilizes the Markov property of the state process, the asymptotic results in Theorem 6 do not require the state process to be Markov. In fact, Theorem 6 allows for arbitrary regime pattern, as long as Assumption 4 is satisfied. An interesting special case is when the smoothed algorithm is applied to factor model with common structural breaks in the loadings. Various methods are proposed recently for estimating the break points, e.g., Cheng et al. (2016), Baltagi et al. (2017, 2021), Bai et al. (2020) and Ma and Su (2018). Theorem 6(ii) implies that we can also consistently estimate the break points using the smoothed algorithm. Finally, we show that the estimated transition probability matrix is also consistent.

Theorem 7 *Assume z_t follows a Markov process, under Assumptions 1-4 and 5(1), as $(N,T)\to\infty$, $\tilde{Q}_{jk}\xrightarrow{p}Q_{jk}^0$ for each j and k if $T^{\frac{16}{\alpha}}/N\to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N\to 0$.*

5 Simulations

In this section, we perform simulations to confirm the theoretical results and examine the finite sample performance of our methods under various empirically relevant scenarios.

5.1 Simulation Design

The data is generated as follows:

$$x_{it} = \begin{cases} f_t^{0\prime}\lambda_{1i}^0 + e_{it} \text{ if } z_t = 1, \\ f_t^{0\prime}\lambda_{2i}^0 + e_{it} \text{ if } z_t = 2, \end{cases} \text{ for } i = 1, ..., N \text{ and } t = 1, ..., T,$$

i.e., we consider two regimes. For the factors and the loadings, we consider three data generating processes (DGP) as listed below:

DGP 1: There are two factors in both regimes and the loadings of both factors have regime switching.

DGP 2: There are two factors in both regimes and only the loadings of the second factor have regime switching.

DGP 3: There is one factor in both regimes and its loadings have regime switching.

For all these DGPs, the factors are generated as follows:

$$f_{t,p}^0 = \rho f_{t-1,p}^0 + \epsilon_{t,p} \text{ for } t = 2, ..., T \text{ and } p = 1, ..., r^0.$$

$\epsilon_{t,p}$ is i.i.d. $N(0,1)$, and $f_{1,p}^0$ is i.i.d. $N(0, \frac{1}{1-\rho^2})$ so that the distributions of the factors are stationary. Serial correlation of the factors is controlled by the scalar ρ.

The errors are generated as follows:

$$e_{it} = \alpha e_{i,t-1} + v_{it} \text{ for } i = 1, ..., N \text{ and } t = 2, ..., T,$$

where $v_t = (v_{1,t}, ..., v_{N,t})'$ is i.i.d. $N(0, \Omega)$ for $t = 2, ..., T$ and $(e_{1,1}, ..., e_{N,1})'$ is $N(0, \frac{1}{1-\alpha^2}\Omega)$ so that the distributions of the errors are stationary. Serial correlation of the errors is controlled by the scalar α. For Ω, we set $\Omega_{ij} = \beta^{|i-j|}$ for some β between 0 and 1, thus cross-sectional dependence of the errors is controlled by β. In addition, the processes $\{\epsilon_{t,p}\}$ and $\{v_{it}\}$ are mutually independent for all p and i.

The loadings are generated as follows: For DGP1, both λ_{1i}^0 and λ_{2i}^0 are generated as i.i.d. $N(0, \frac{1-\rho^2}{1-\alpha^2}\frac{2R^2}{1-R^2}I_2)$ across i, and λ_{1i}^0 and λ_{2i}^0 are also independent with each other. For DGP2, λ_{1i}^0 and the second element of λ_{2i}^0 are generated as i.i.d. $N(0, \frac{1-\rho^2}{1-\alpha^2}\frac{2R^2}{1-R^2}I_3)$ across i. For DGP3, both λ_{1i}^0 and λ_{2i}^0 are generated as i.i.d. $N(0, \frac{1-\rho^2}{1-\alpha^2}\frac{R^2}{1-R^2})$ across i, and λ_{1i}^0 and λ_{2i}^0 are also independent with each other. All loadings are independent of the factors and the errors. The variance $\frac{1-\rho^2}{1-\alpha^2}\frac{2R^2}{1-R^2}$ guarantees that the regression R-square of each series i is equal to R^2, this controls the signal-noise ratio. Following the literature, we set $R^2 = 0.5$.

For the state process $\{z_t, t = 1, ..., T\}$, we consider four cases as listed below:

Regime Pattern 1: US business cycle 1945Q2-2020Q1

Regime Pattern 2: single common break at $t = T/2$

Regime Pattern 3: two common breaks at $t = T/3$ and $t = 2T/3$, *and the loadings switch back after the second break*

Regime Pattern 4: a randomly generated Markov process

Regime pattern 1 is based on the US business cycle from 1945 Quarter 2 to 2020 Quarter 1, as determined by the NBER business cycle dating committee. There are 75 years (300 quarters) in total, thus we have $T = 300$. For $t = 1, ..., 300$, $z_t = 1$ if the US economy at time t is in expansion and $z_t = 2$ if the US economy at time t is in recession. The transition probabilities of the state process calibrated to the US business cycle is $Q_{11}^0 = 0.95$ and $Q_{22}^0 = 0.72$ (average duration of expansion is $1/(1 - Q_{11}^0) = 20$ and average duration of recession is $1/(1 - Q_{22}^0) \approx 3.5$).

Regime patterns 2 and 3 correspond to the case where loadings have single common break and multiple common breaks, respectively. Regime patterns 3 is especially interesting since the case where there are multiple breaks and the loadings switch back to their original values after the second break is rarely studied in the literature. Various methods are proposed in the literature recently for estimating the break points, here we perform simulations for regime patterns 2 and 3 to evaluate the finite sample performance of our method when it is applied to these interesting cases.

Regime pattern 4 is a Markov process randomly generated with transition probabilities $Q_{11}^0 = 0.95$ and $Q_{22}^0 = 0.72$, and $\{z_t, t = 1, ..., T\}$ is independent with f_s^0 and e_{is} for all i and s. Regime patterns 1-3 are prespecified and are not necessarily Markov processes, thus here we consider regime pattern 4 to evaluate the performance of our method when applied to a Markov state process.

We study both the unsmoothed algorithm in Section 2.2 and the smoothed algorithm in Section 4.1. The key difference is that in the E-step, the former uses unsmoothed regime probabilities while the latter uses smoothed regime probabilities. Both algorithms start from randomly generated initial values of the loadings and iterate between the E-step and the M-step until convergence. To search for global maximum of the likelihood function, we generate initial values randomly for many times and take the one with the largest likelihood. For other parameters, we set $\sigma^2 = 1$, $q_j = 0.5$ for $j = 1, 2$, $\phi_k = 0.5$ for $k = 1, 2$, $Q_{11} = 0.95$ and $Q_{22} = 0.72$. Q_{11} and Q_{22} are calibrated to regime pattern 1. Once we get the estimated regime probabilities and the estimated loadings, $\tilde{Q}_{11}$ and $\tilde{Q}_{22}$ are estimated by equation (25), and factors are estimated by equations (17) and (27).

5.2 Simulation Results

Figure 1 displays the smoothed probabilities of regime 2 for DGP 1 with $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$. Subfigures 1-4 of Figure 1 correspond to regime patterns 1-4, respectively. It is easy to see that in all subfigures when the true regime is regime 1, the smoothed probabilities stay at zero with only a few short and mild spikes. At the beginning of each shaded region, the smoothed probabilities increase to one instantly, and at the end of each shaded region, the smoothed probabilities instantly decrease to zero. Figure 2 displays the unsmoothed probabilities of regime 2 for DGP1 under the four regime patterns with $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$. The estimated probabilities still stay at zero when it's regime 1 and instantly increase to one (decrease to zero) when there is regime switching, but compared to Figure 1, Figure 2 shows more and sharper spikes (upward or downward). These spikes are false positives in detecting regime switching. Figure 3 and Figure 4 display the smoothed and the unsmoothed probabilities of regime 2 for DGP2, respectively. The performance of the estimated probabilities deteriorates since for DGP2 only one factor has regime switching in its loadings. Overall, Figures 1-4 confirm the theoretical results that turning points (break points) can be identified consistently if N is large.

Figure 5 focuses on regime pattern 1 and displays the estimated probabilities of regime 2 for DGP1 and DGP2 with $N = 200$. Comparing the subfigure 3 and subfigure 4 of Figure 5 to subfigure 1 of Figure 3 and subfigure 1 of Figure 4, it is easy to see that $N = 200$ improves the performance of the estimated probabilities. Figure 6 also focuses on regime pattern 1 and displays the smoothed probabilities of regime 2 for DGP1 and DGP2 with $(\rho, \alpha, \beta) = (0.5, 0, 0)$ or $(0, 0.5, 0.5)$. Comparing to subfigure 1 of Figure 1 and subfigure 1 of Figure 3, it seems that the value of (ρ, α, β) does not affect the performance too much if they were far away from 1.

Comparing Figure 2 to Figure 1 and Figure 4 to Figure 3, it is obvious that the smoothed probabilities performs much better than the unsmoothed probabilities. Many false positives in Figure 3 and Figure 4 are eliminated by the smoother. This is because for each t, regimes at $t-1$ and $t+1$ contains information for detecting the regime at period t. Comparing subfigures 2-3 to subfigures 1 and 4 in Figures 1-4, we can see that the performance of the estimated probabilities under regime patterns 2-3 is better than the performance under patterns 1 and 4. This is also because regimes at neighborhood periods provide information for current regime. Roughly speaking, the performance is better when the regime pattern is relatively simple. In addition, we can also see that the performance under regime pattern 1 is slightly better than the performance under regime pattern 4. This is because the subsample size of regime 2 under pattern 4 is larger than the subsample size under pattern 1 (72 vs 45). In general, we find that to guarantee good performance, subsample size for each regime should be not less than 40.

The number of initial value trials also significantly affect the performance. We find that for regime pattern 1, normally 5 trials are enough, but to guarantee good performance in all of 1000 replications, 30 trials are needed. For regime pattern 4, normally 2 trials are enough and 15 trials are needed to guarantee good performance in all replications. For regime patterns 2-3, 5 trials are enough for all replications. In general, more trials are needed when the regime pattern is complex and subsample size is small.

Now we consider determining the turning points (break points for regime patterns 2-3) using the estimated probabilities of regime 2. A straightforward way is to determine the regime at each t by comparing the estimated probability of regime 2 to a prespecified threshold, i.e, if the estimated probability is above 0.9 (below 0.1), period t is classified as regime 2 (1). Since there are quite a few false positives in Figures

2-6, a more robust way is to use the moving average of the estimated probabilities, $p_t^{ma} \equiv (\tilde{p}_{t2|T} + ... + \tilde{p}_{t-d,2|T})/(d+1)$, where d is the order of the moving average. If regime switches from 2 to 1 (1 to 2) at the previous turning point, $t-d$ would be considered as the turning point from regime 1 to regime 2 (regime 2 to regime 1) when p_t^{ma} increases to above the threshold 0.9 (decreases to below 0.1) for the first time after the previous turning point.

Subfigures 1-2 of Figure 7 display the moving average of the smoothed probabilities under DGP2 and regime pattern 1 for $d = 1$ and $d = 2$, and subfigure 3 displays the moving average of the unsmoothed probabilities under DGP2 and regime pattern 3 for $d = 3$. Comparing to subfigure 1 of Figure 3 and subfigure 3 of Figure 4, it is obvious that almost all false positives are eliminated by the moving average, although true regime 2 are also eliminated three times in subfigures 1-2. In general, the cost of using moving average is that true regime 2 may be eliminated if it only lasts for a short period, and we need to wait for d periods to determine the turning point ($t-d$ is considered as the turning point when $p_t^{ma} > 0.9$). Therefore, $d = 1$ is a good choice if regime switches relatively frequently and we want to quickly detect regime switching. For offline estimation of the break points of regime patterns 2-3, large d is a better choice.

Finally, to access the adequacy of the asymptotic distributions of the estimated loadings and factors in approximating their finite sample counterparts, we display in Figures 8-11 the histograms of the standardized estimated factors for $t = T/2$ and the standardized estimated loadings for $i = N/2$ under DGP3. The number of simulations is 1000. The histograms are normalized to be a density function and the standard normal density curve is overlaid on them for comparison. It is easy to see that in all subfigures of Figures 8-11, the standard normal density curve provides good approximation to the normalized histograms. The histograms of the estimated factors in Figure 8 are slightly fat-tailed because of bad initial values. Comparing the four rows in each of Figures 8-11, we can see that the estimated loadings and factors using the smoothed algorithm perform better than using the unsmoothed algorithm, $(\rho, \alpha, \beta) = (0.5, 0.5, 0.5)$ does not matter too much, and $N = 200$ significantly improves the performance.

In addition, we also present in Table 1 the average R^2 of the estimated loadings of regime 1 and regime 2 projecting on the true loadings, the average R^2 of the estimated factors, and the average absolute error of the estimated transition probabilities. It is easy to see that in Table 1, R^2_{l1} and R^2_{l2} are always close to one. R^2_{Hf} is always close to one but R^2_f is much smaller than R^2_{Hf}. This is consistent with Theorem 5(1) and Theorem 6(vi). In summary, results in Figures 1-11 and Table 1 lend strong support to the theoretical results and illustrate the usefulness of the proposed EM algorithms.

6 Empirical Application

In this section we apply the proposed method to detect turning points of US business cycle from 02/1980 to 02/2020 using the FRED-MD (Federal Reserve Economic Data - Monthly Data) data set. The FRED database is maintained by the Research division of the Federal Reserve Bank of St. Louis, and is publicly accessible and updated in real-time. The 03/2020 vintage of the FRED-MD data set contains 128 unbalanced monthly time series from 01/1959 to 02/2020, including eight groups (output and income, labor market, housing, consumption and inventories, money and credit, prices, stock market). After removing those series with missing values and data transformation[7], we have 106 balanced monthly series ranging from 03/1959 to 02/2020. Finally, the data is demeaned and standardized.

For each month from 02/1980 to 02/2020 (481 months in total), we use the data from 03/1959 to that month for calculating the probability of recession of that month, i.e., we behave as if we were standing at that month[8]. More specifically, we apply the algorithm in Section 4.1 to the data from 03/1959 to the previous month to estimate the model parameters[9], and then use the estimated parameters and the data of that month to calculate the filtered probability of recession for that month. Since the data

[7] See the Appendix of McCracken and Ng (2016) for the details of data description and transformation.

[8] For simplicity, we do not use the vintage data of that month. Compared to the vintage data, the data we use contains revision in some series if more accurate observations were available after that month, but previous studies on business cycle dating show that data revisions have little effects on the results.

[9] US business cycle from 03/1959 to the previous month as determined by NBER is used as the initial values for probabilities.

of that month is available at the end of that month or the beginning of the next month, new recession or expansion starting from the beginning of that month could only be detected with at least one month delay.

To convert the recession probability of each month into a binary variable that indicates the state of the economy in that month, we compare the estimated recession probability to a prespecified threshold. More specifically, if the previous turning point is a trough and the recession probability of month t exceeds 0.8 for the first time after the previous turning point, month t would be considered as a new turning point from expansion to recession. Similarly, if the previous turning point is a peak and the recession probability of month t falls below 0.2 for the first time after the previous turning point, month t would be considered as a new turning point from recession to expansion. For robustness check, we also consider $(0.9, 0.1)$ as the threshold, the results are quite similar.

We consider the turning points determined by the NBER BCDC (business cycle dating committee) as the benchmark for comparison and we mainly focus on the accuracy and speed of the proposed method in detecting turning points. The proposed method is applied to both the whole panel and a subset of the whole panel which consists of only the first 50 series among all 106 series. The results of using only the first 50 series are better. We conjecture that this is mainly because not all 106 series had regime switching in the factor loadings at each turning point determined by the NBER BCDC[10], or some series had regime switching in their loadings at time periods that are different from the NBER BCDC turning points. Thus we may further improve the performance of the proposed method by selecting series that are most relevant to and synchronous with business cycle. A careful selection is out of the scope of this paper.

Table 2 presents the results of using the first 50 series. The number of factors in each regime is set to be six. mm/yyyy in the second and the seventh row indicate the starting month of each recession and expansion. The row corresponds to "NBER

[10]The NBER BCDC mainly focuses on four series, (1) non-farm payroll employment, (2) industrial production, (3) real manufacturing and trade sales, and (4) real personal income excluding transfer payments.

BCDC", "Chauvet Piger" and "This paper" shows the number of months it takes the NBER BCDC, Chauvet and Piger (2008) and this paper to detect each recession and expansion, respectively. For example, the recession starting from the beginning of February 1980 would be detected by the NBER BCDC at the beginning of June 1980, by Chauvet and Piger (2008) at the beginning of August 1980, and by this paper at the beginning of May 1980, respectively. Overall, it is easy to see that this paper detects turning points much faster than NBER BCDC and slightly faster than Chauvet and Piger (2008). On average, this paper detects recessions with 6.25 months delay and expansions with 5.4 months delay, NBER BCDC detects recessions with 7.4 months delay and expansions with 14.8 months delay, and Chauvet and Piger (2008) detects recessions with 8.6 months delay and expansions with 6.2 months delay.

Table 2 shows that using more series could improve the speed of turning points detection. However, using more series could also bring in false positives (turning points detected by the proposed method using many series but not detected by NBER BCDC), because the extra series may contain different turning points. Here we detect eight false recessions: 09/1983-11/1983, 10/1986-02/1987, 07/1989-10/1989, 01/1993-02/1993, 01/1995-03/1995, 08/1998, 05/2000-08/2000, 06/2010-10/2010, and one false expansion: 02/1982. While these false positives should not be ignored, most of them only last for a short periods and would have little effect on macroeconomic policy. Overall, our results demonstrate the potential of using a large number of series and factor model with common loading switching for quick real-time detection of business cycle turning points.

7 Conclusions

The exposure of economic time series to common factors may switch depending on state variables such as fiscal policy, monetary policy, business cycle stage, stock market volatility, technology and so on. For consistent estimation of the factor structure, it is crucial to take into account such regime switching phenomena. This paper considers maximum likelihood estimation for large factor models with common regime switching in the loadings and proposes EM algorithm for computation, which is easy to implement and runs fast even when N is large. Convergence rates and limit distributions of the estimated loadings and the estimated factors are established under the approximate factor model setup. This paper also shows that when N is large, regime switching can be identified consistently and only one observation after the switching point is needed. This allows us to detect regime switching at very early times. Monte Carlo simulations confirm the theoretical results and good performance of our method. An application to the FRED-MD dataset demonstrates the potential of using many time series with our method for quick detection of business cycle turning points.

Some related topics are worth further study. First, it would be interesting to see the performance of the portfolio constructed using regime specific loadings, and how the identified regime is related to exogenous variables such as market volatility and money growth. Second, our results imply that the forecasting equation would have induced regime switching if estimated factors are used for forecasting, so we want to know whether it indeed matters. Finally, a selection of time series that are most synchronous with or related to business cycle could improve the speed and accuracy of our method for turning points detection, so we would like to see how much we can achieve after careful selection.

References

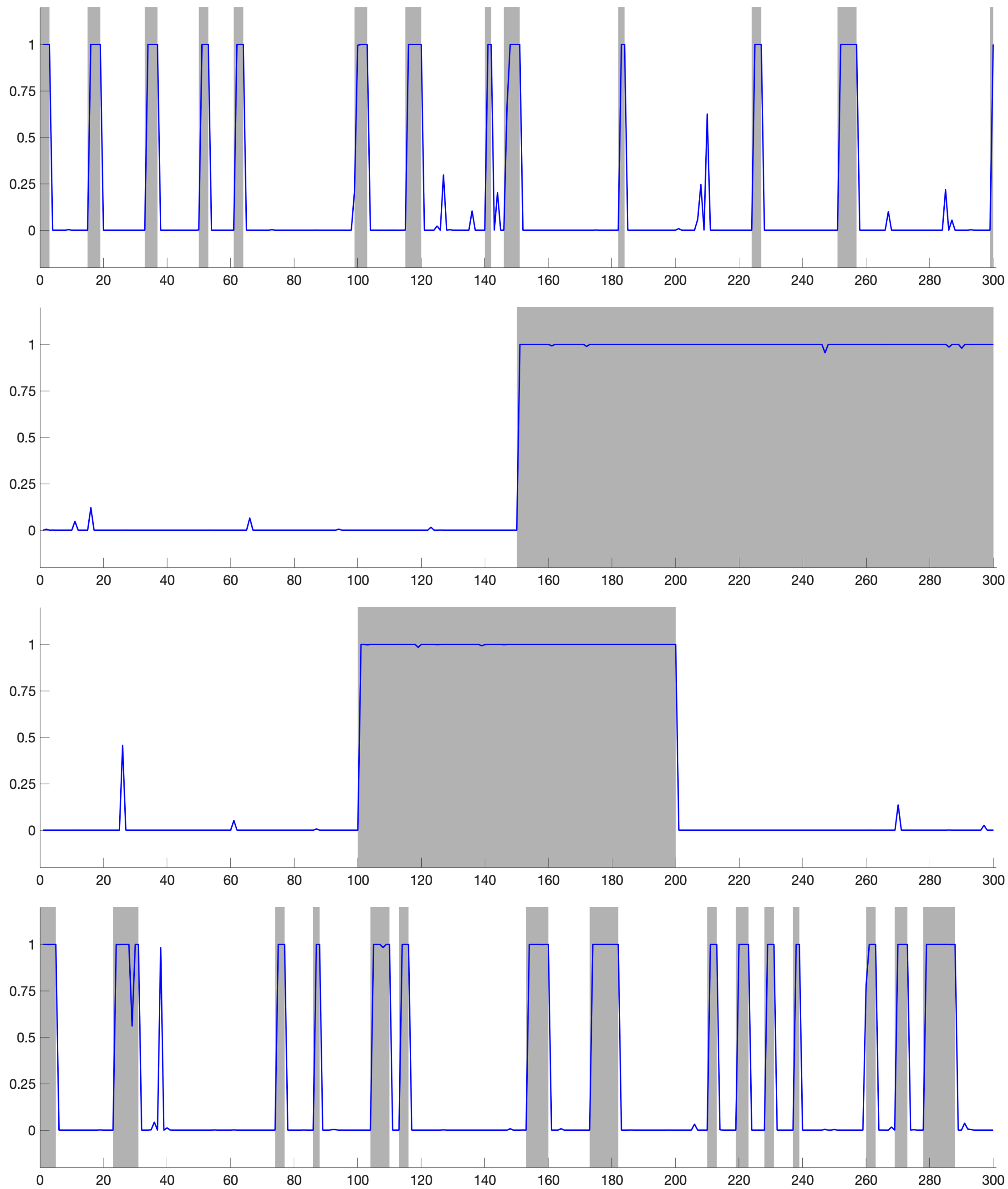
[1] Andersen, T.W., 2003. An introduction to multivariate statistical analysis. New York: Wiley.

[2] Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

[3] Bai, j., Han, X., Shi, Y., 2020. Estimation and inference of change points in high-dimensional factor models. Journal of Econometrics 219, 66-100.

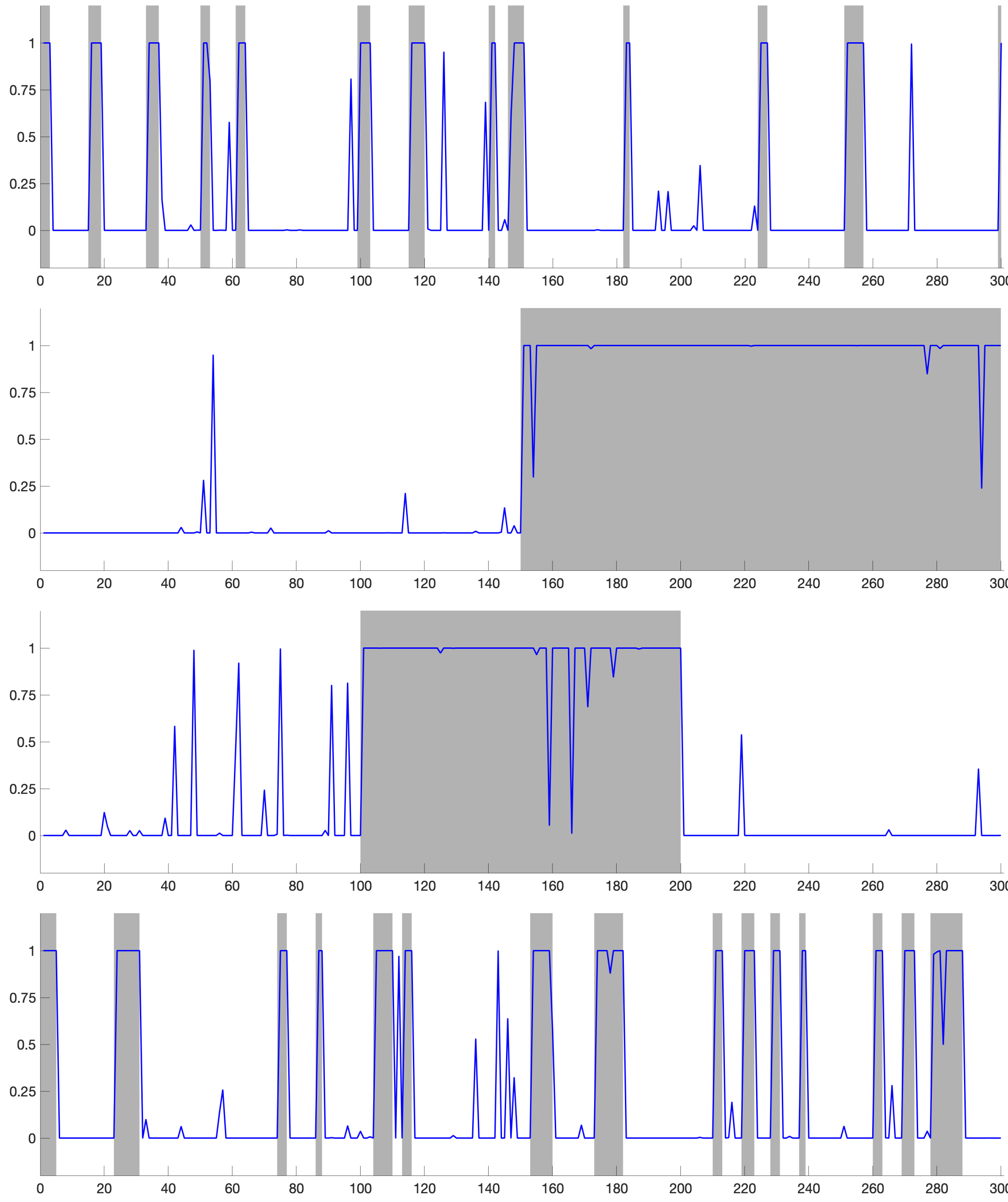
[4] Bai, J., Li, K., 2012. Statistical analysis of factor models of high dimension. Annals of Statistics 40, 436–465.

Figure 1: Smoothed Probabilities of Regime 2 for DGP 1



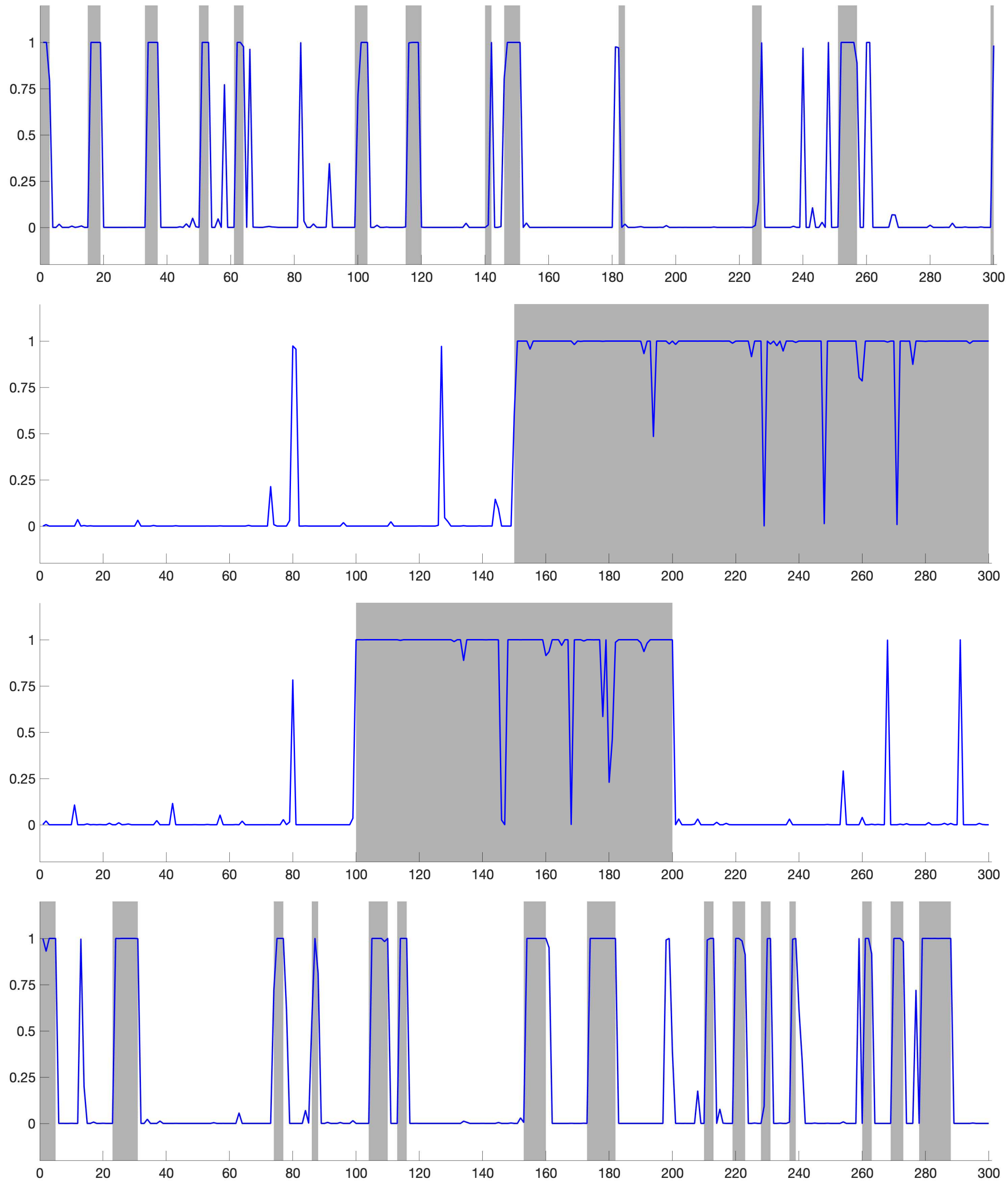
Notes: Subfigures 1-4 correspond to regime pattern 1-4, respectively. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2. $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$.

Figure 2: Unsmoothed Probabilities of Regime 2 for DGP 1



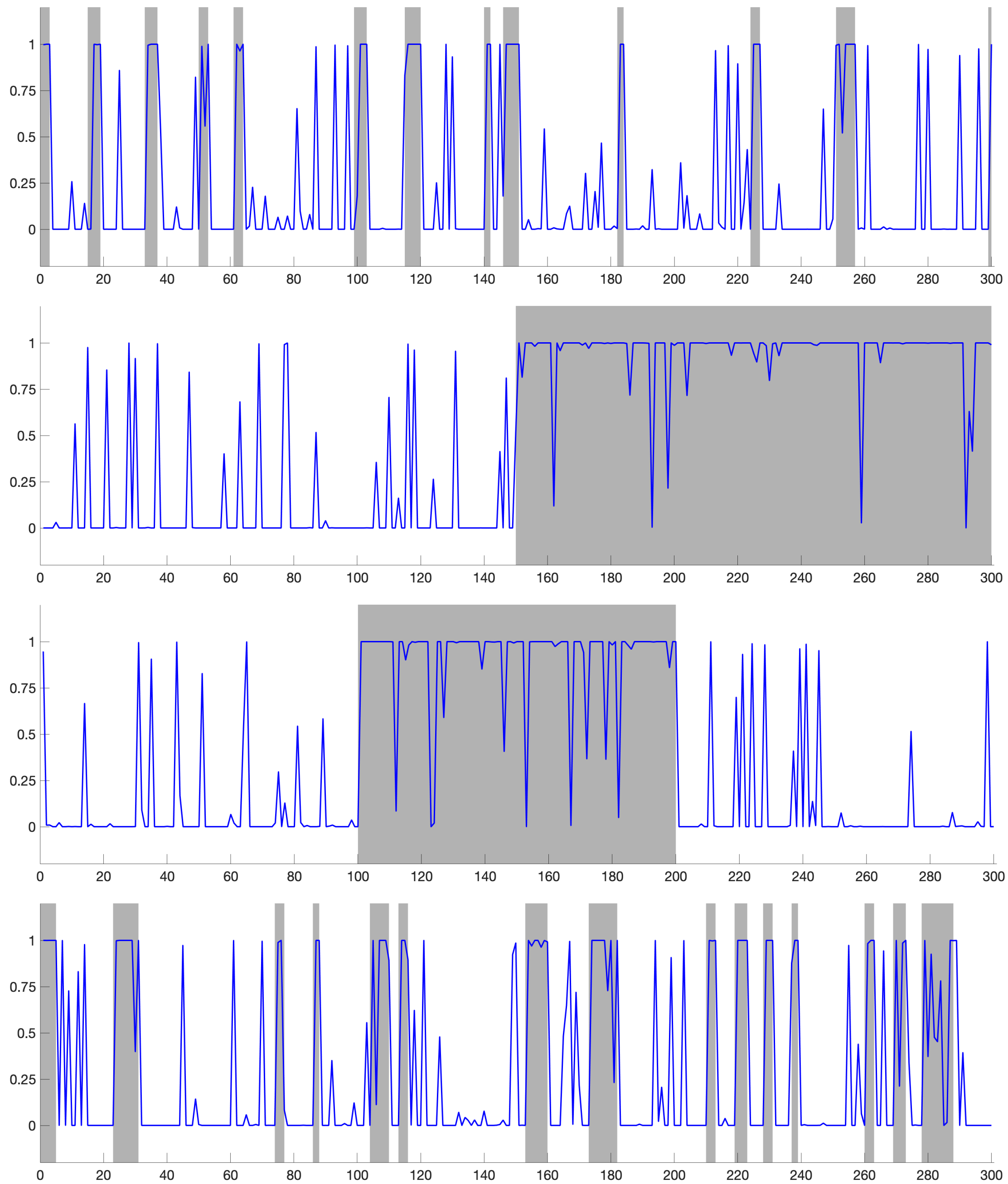
Notes: Subfigures 1-4 correspond to regime pattern 1-4, respectively. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2. $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$.

Figure 3: Smoothed Probabilities of Regime 2 for DGP 2



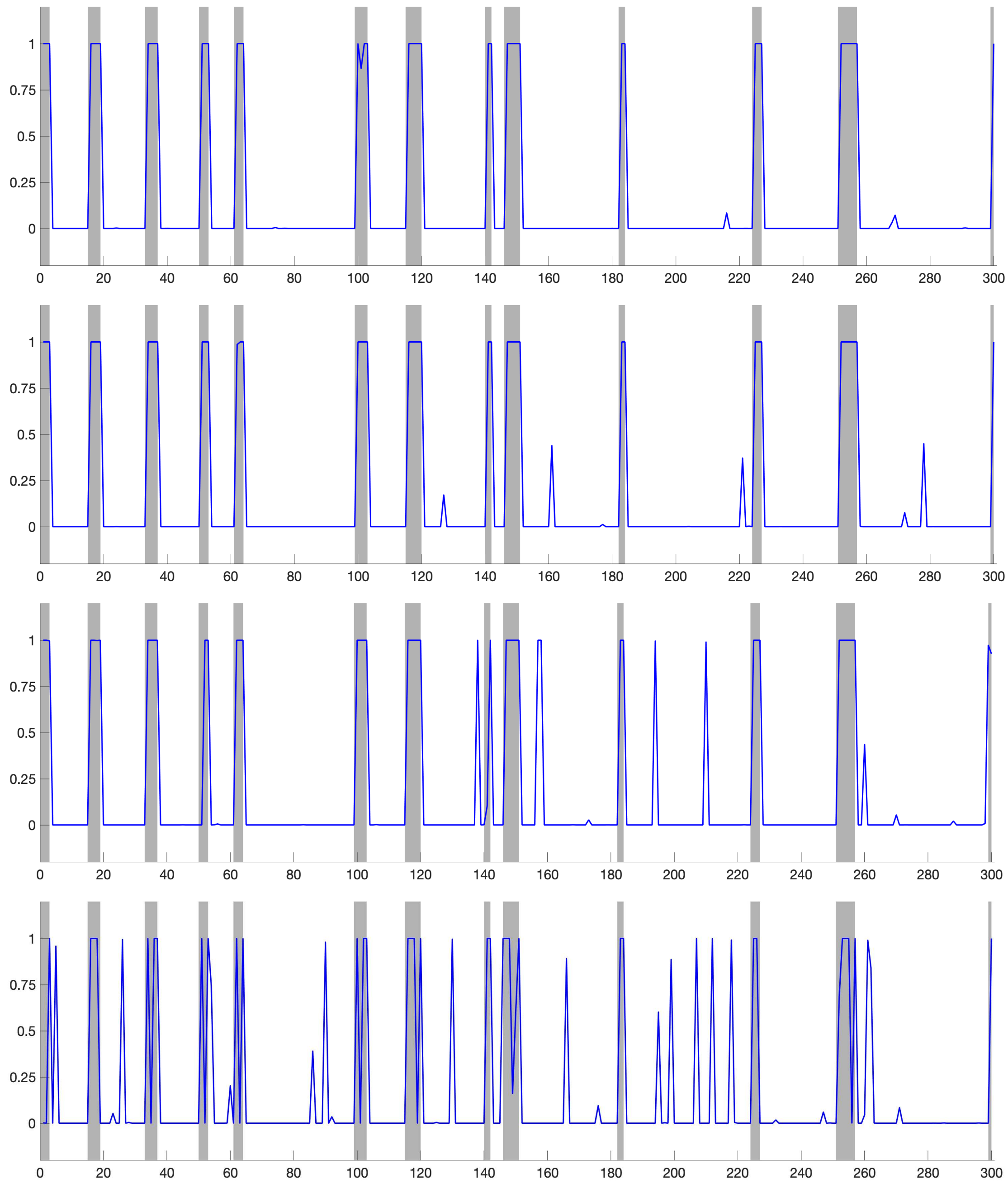
Notes: Subfigures 1-4 correspond to regime pattern 1-4, respectively. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2. $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$.

Figure 4: Unsmoothed Probabilities of Regime 2 for DGP 2



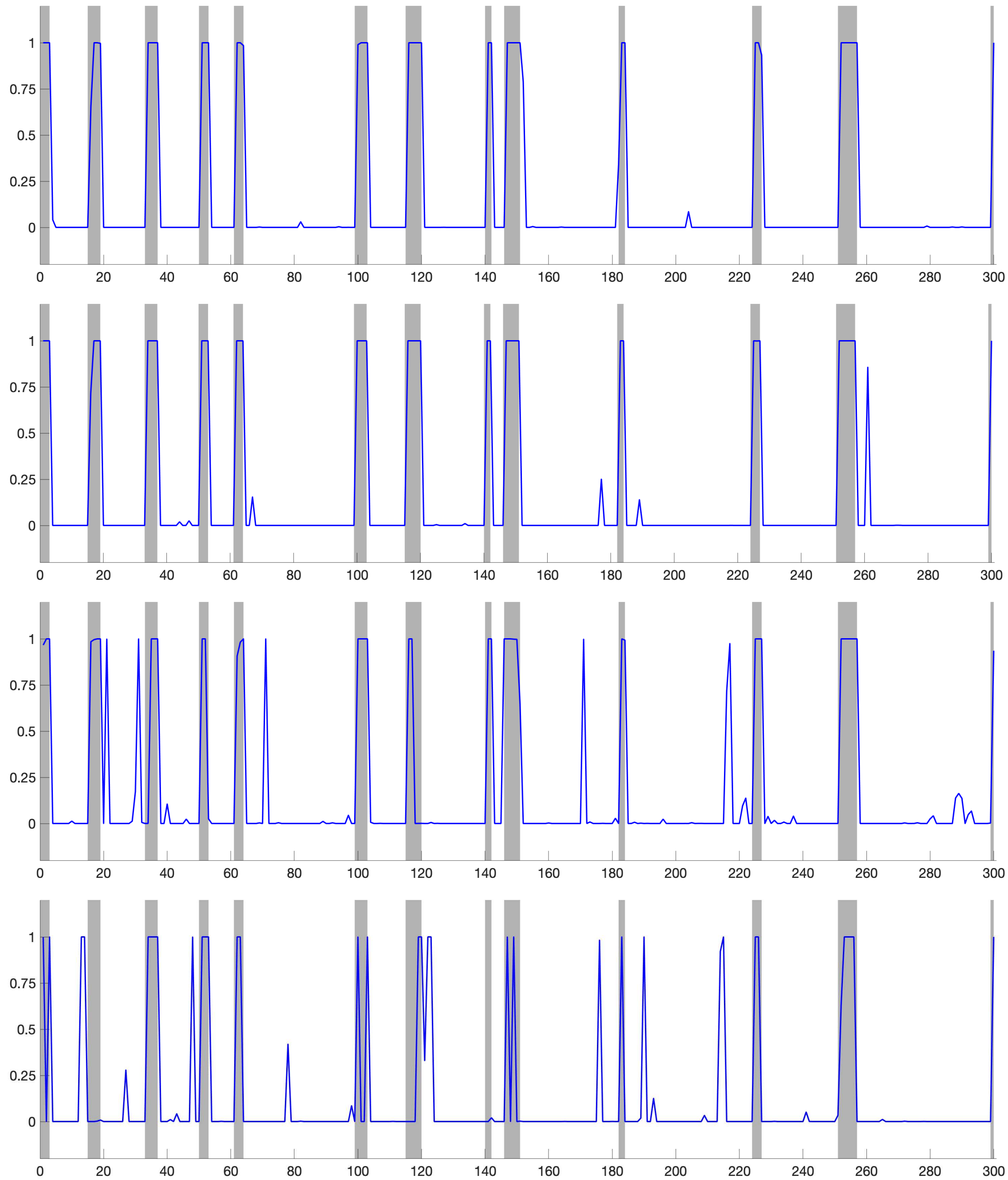
Notes: Subfigures 1-4 correspond to regime pattern 1-4, respectively. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2. $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$.

Figure 5: Smoothed and Unsmoothed Probabilities of Regime 2 for Regime Pattern 1, $(N,T) = (200,300)$ and $(\rho,\alpha,\beta) = (0,0,0)$



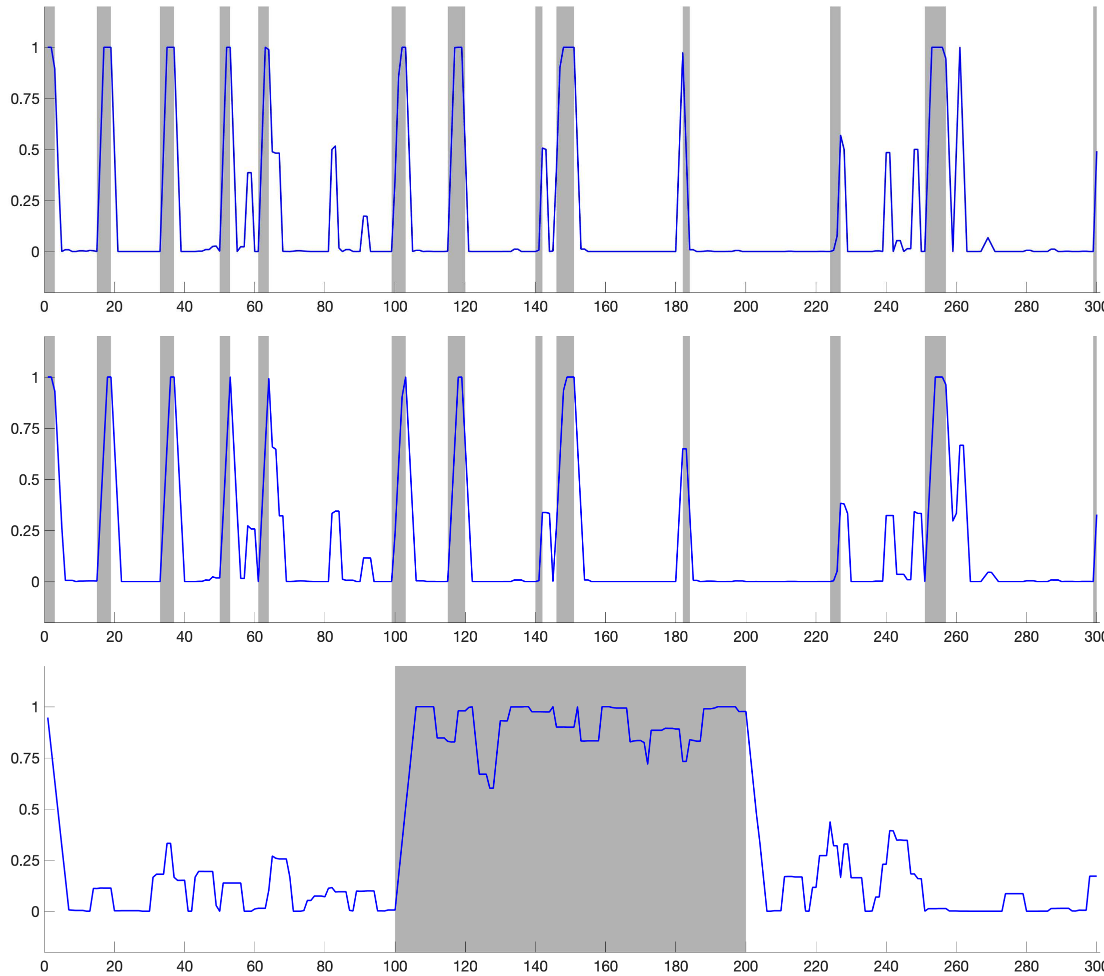
Notes: Subfigures 1-4 correspond to smoothed probabilities for DGP1, unsmoothed probabilities for DGP1, smoothed probabilities for DGP2 and unsmoothed probabilities for DGP2, respectively. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2.

Figure 6: Smoothed Probabilities of Regime 2 for Regime Pattern 1, $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0.5, 0, 0)$ or $(\rho, \alpha, \beta) = (0, 0.5, 0.5)$



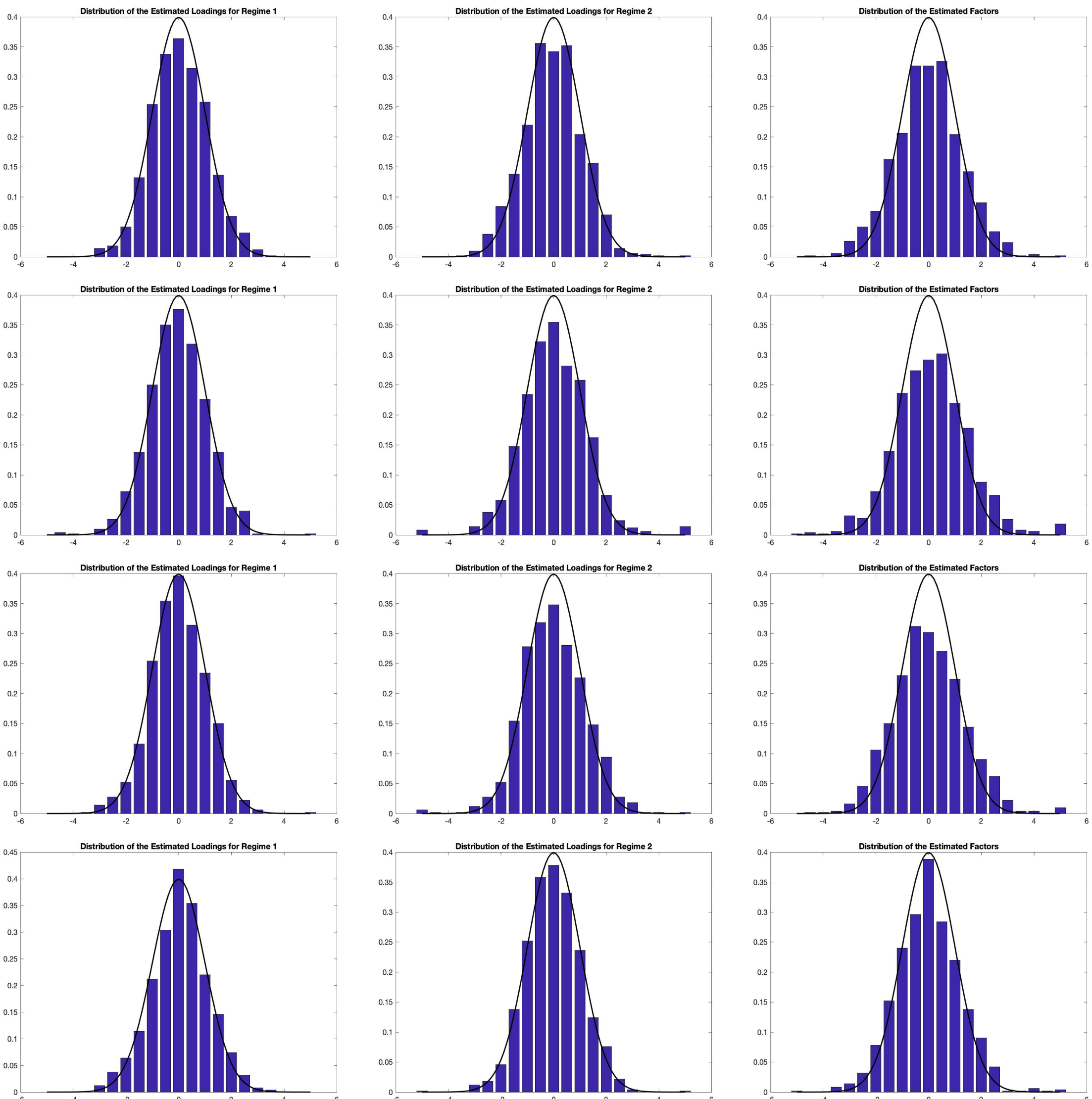
Notes: Subfigures 1-4 correspond to smoothed probabilities for DGP1 with $(\rho, \alpha, \beta) = (0.5, 0, 0)$, DGP1 with $(\rho, \alpha, \beta) = (0, 0.5, 0.5)$, DGP2 with $(\rho, \alpha, \beta) = (0.5, 0, 0)$ and DGP2 with $(\rho, \alpha, \beta) = (0, 0.5, 0.5)$, respectively. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2.

Figure 7: Moving Average of the Estimated Probabilities of Regime 2



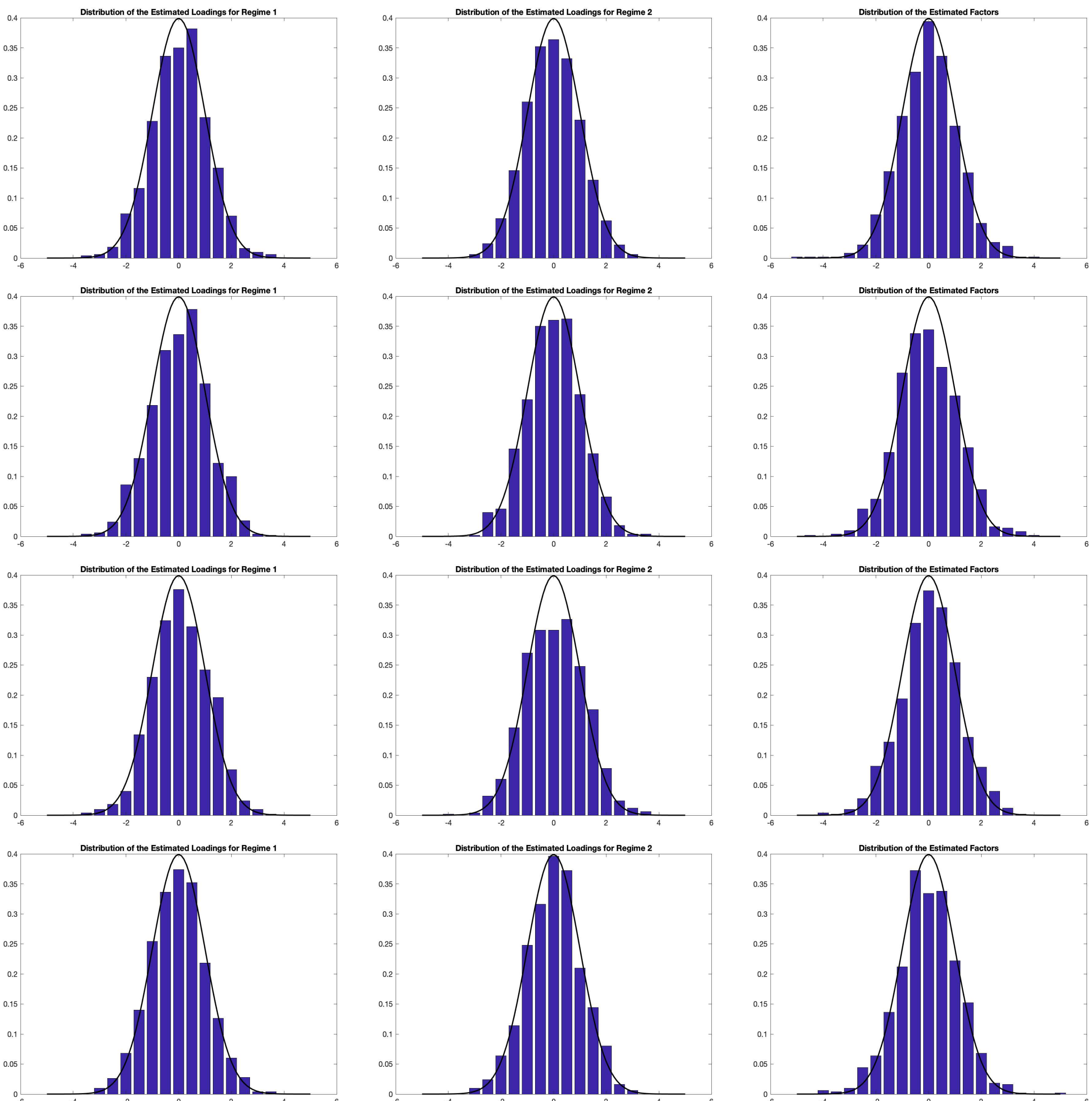
Notes: Subfigures 1-2 correspond to MA(1) and MA(2) of smoothed probabilities for DGP2 under regime pattern 1, respectively. Subfigure 3 corresponds to MA(3) of unsmoothed probabilities for DGP2 under regime pattern 3. The x-axis is time and the y-axis is the probability. The shaded regions correspond to regime 2. $(N, T) = (100, 300)$ and $(\rho, \alpha, \beta) = (0, 0, 0)$.

Figure 8: Histograms of the Estimated Loadings and the Estimated Factors for Regime Pattern 1



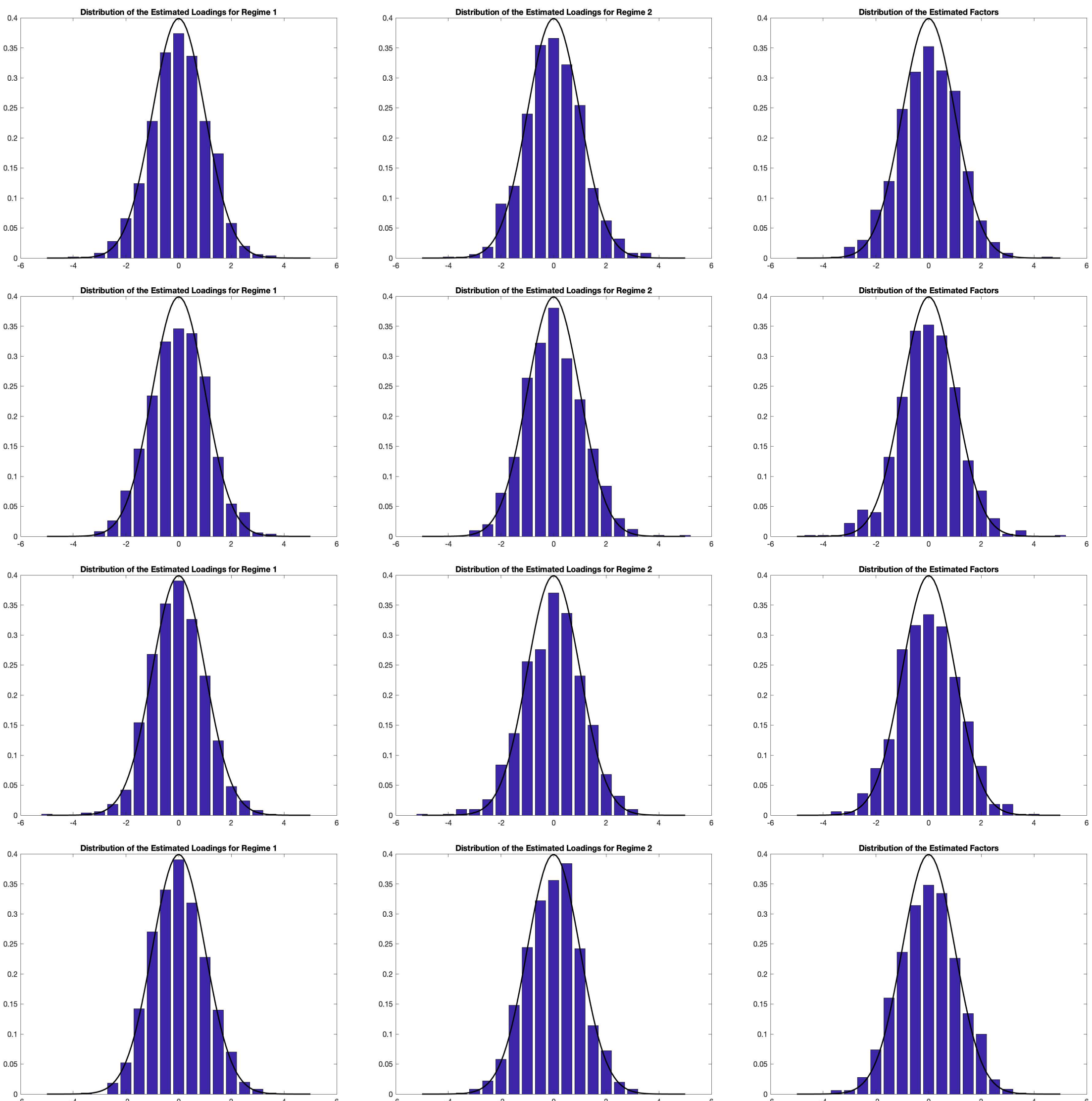
Notes: Subfigures in the first to the fourth row correspond to the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the unsmoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the smoothed algorithm with $\rho = \alpha = \beta = 0.5$ and $(N, T) = (100, 300)$, and the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (200, 300)$, respectively. Subfigures in the first to the third column correspond to the estimated loadings for regime 1, the estimated loadings for regime 2 and the estimated factors, respectively. The curve overlaid on the histograms is the standard normal density function.

Figure 9: Histograms of the Estimated Loadings and the Estimated Factors for Regime Pattern 2



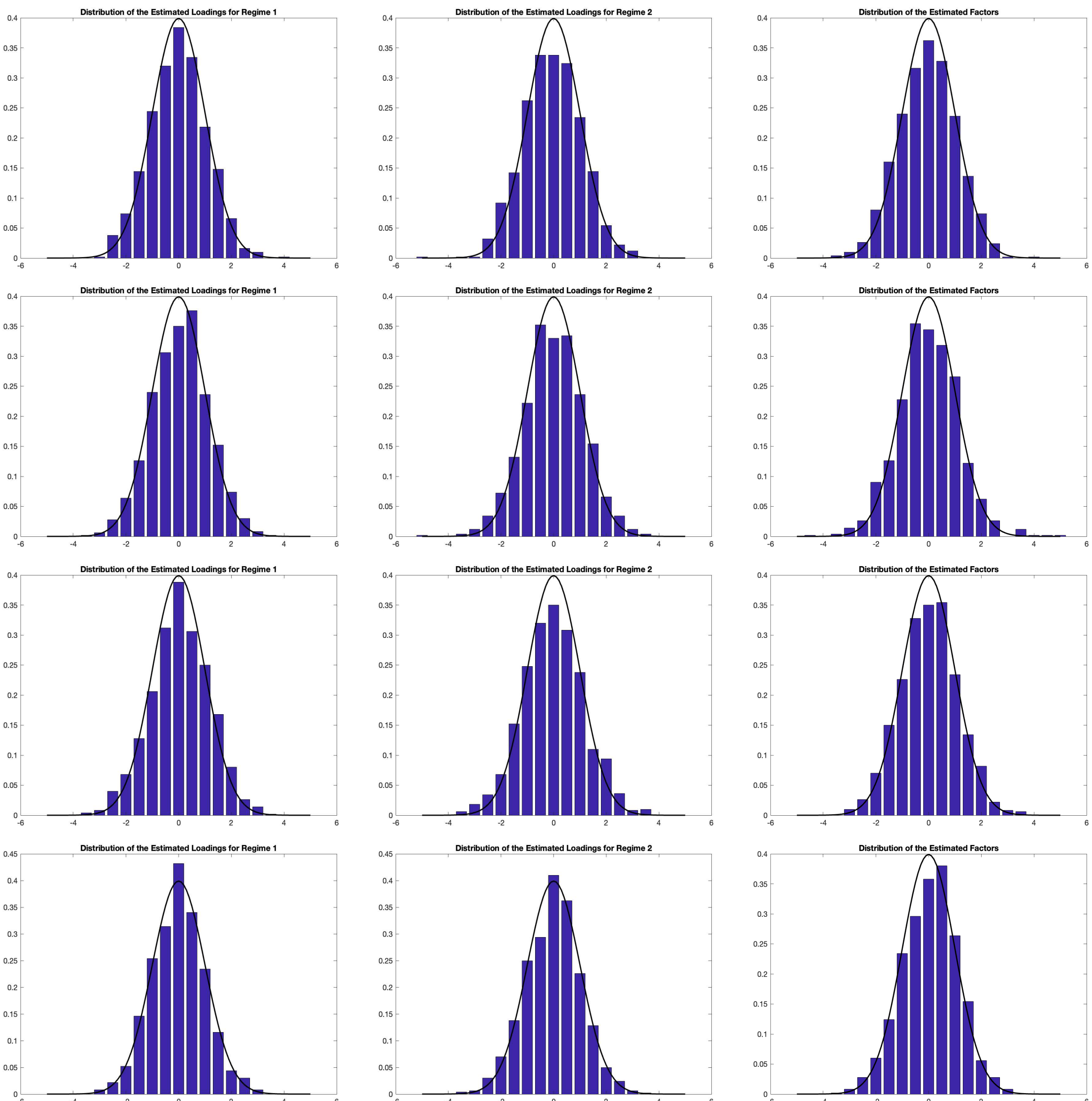
Notes: Subfigures in the first to the fourth row correspond to the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the unsmoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the smoothed algorithm with $\rho = \alpha = \beta = 0.5$ and $(N, T) = (100, 300)$, and the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (200, 300)$, respectively. Subfigures in the first to the third column correspond to the estimated loadings for regime 1, the estimated loadings for regime 2 and the estimated factors, respectively. The curve overlaid on the histograms is the standard normal density function.

Figure 10: Histograms of the Estimated Loadings and the Estimated Factors for Regime Pattern 3



Notes: Subfigures in the first to the fourth row correspond to the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the unsmoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the smoothed algorithm with $\rho = \alpha = \beta = 0.5$ and $(N, T) = (100, 300)$, and the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (200, 300)$, respectively. Subfigures in the first to the third column correspond to the estimated loadings for regime 1, the estimated loadings for regime 2 and the estimated factors, respectively. The curve overlaid on the histograms is the standard normal density function.

Figure 11: Histograms of the Estimated Loadings and the Estimated Factors for Regime Pattern 4



Notes: Subfigures in the first to the fourth row correspond to the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the unsmoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (100, 300)$, the smoothed algorithm with $\rho = \alpha = \beta = 0.5$ and $(N, T) = (100, 300)$, and the smoothed algorithm with $\rho = \alpha = \beta = 0$ and $(N, T) = (200, 300)$, respectively. Subfigures in the first to the third column correspond to the estimated loadings for regime 1, the estimated loadings for regime 2 and the estimated factors, respectively. The curve overlaid on the histograms is the standard normal density function.

Table 1: Average R^2 of the Estimated Loading Space, Average R^2 of the Estimated Factor Space, and Average Absolute Error of the Estimated Transition Probabilities

	R^2_{l1}	R^2_{l2}	R^2_f	R^2_{Hf}	$\tilde{Q}_{11}$	$\tilde{Q}_{22}$
Smoothed with $(\rho, \alpha, \beta) = (0, 0, 0)$ and $(N, T) = (100, 300)$						
Pattern 1	0.996	0.9762	0.7337	0.9889	0.0028	0.013
Pattern 2	0.9931	0.9932	0.5155	0.9896	N.A.	N.A.
Pattern 3	0.9949	0.9895	0.541	0.9894	N.A.	N.A.
Pattern 4	0.9955	0.9854	0.6256	0.9892	0.0216	0.0378
Unsmoothed with $(\rho, \alpha, \beta) = (0, 0, 0)$ and $(N, T) = (100, 300)$						
Pattern 1	0.9959	0.9678	0.6786	0.9782	N.A.	N.A.
Pattern 2	0.9931	0.9932	0.4855	0.9885	N.A.	N.A.
Pattern 3	0.9949	0.9892	0.5157	0.988	N.A.	N.A.
Pattern 4	0.9955	0.9853	0.6127	0.9875	N.A.	N.A.
Smoothed with $(\rho, \alpha, \beta) = (0.5, 0.5, 0.5)$ and $(N, T) = (100, 300)$						
Pattern 1	0.9933	0.9631	0.7255	0.9849	0.0053	0.0137
Pattern 2	0.9928	0.9929	0.4797	0.9891	N.A.	N.A.
Pattern 3	0.9915	0.9827	0.5458	0.9889	N.A.	N.A.
Pattern 4	0.9927	0.9782	0.6285	0.9886	0.0239	0.0328
Smoothed with $(\rho, \alpha, \beta) = (0, 0, 0)$ and $(N, T) = (200, 300)$						
Pattern 1	0.996	0.9756	0.7408	0.9936	0.0017	0.0151
Pattern 2	0.9933	0.9933	0.5183	0.9949	N.A.	N.A.
Pattern 3	0.995	0.9898	0.5611	0.9949	N.A.	N.A.
Pattern 4	0.9956	0.9856	0.6227	0.9947	0.019	0.024

Notes: The column under R^2_{l1} shows the average R^2 of the estimated loadings of regime 1 projecting on the true loadings of regime 1. The column under R^2_{l2} shows the average R^2 of the estimated loadings of regime 2 projecting on the true loadings of regime 2. The column under R^2_f shows the average R^2 of the estimated factors projecting on the true factors. The column under R^2_{Hf} shows the average R^2 of the estimated factors projecting on the factors rotated by the regime dependent rotation matrix $\bar{H}_{z_t}$.

Table 2: Out of Sample Turning Points Detection

	Recession 02/1980	Expansion 08/1980	Recession 08/1981	Expansion 12/1982	Recession 08/1990
NBER BCDC	4	11	5	7	9
Chauvet Piger	6	5	7	6	7
This paper	3	2	3	7	N.A.
	Expansion 04/1991	Recession 04/2001	Expansion 12/2001	Recession 01/2008	Expansion 07/2009
NBER BCDC	21	8	20	11	15
Chauvet Piger	6	10	7	13	7
This paper	1	8	7	11	10

Notes: mm/yyyy in the second and the seventh row indicate the starting month of each recession and expansion. The row corresponds to "NBER BCDC", "Chauvet Piger" and "This paper" shows the number of months it takes the NBER BCDC, Chauvet and Piger (2008) and this paper to detect each recession and expansion, respectively.

[5] Bai, J., Li, K., 2016. Maximum likelihood estimation and inference for approximate factor models of high dimension. Review of Economics and Statistics 98, 298-309.

[6] Bai, J., Ng, S., 2002. Determining the Number of Factors in Approximate Factor Models. Econometrica 70, 191–221.

[7] Baltagi, B.H., Kao, C., Wang, F., 2017. Identification and estimation of a large factor model with structural instability. Journal of Econometrics 197, 87–100.

[8] Baltagi, B.H., Kao, C., Wang, F., 2021. Estimating and testing high dimensional factor models with multiple structural changes. Journal of Econometrics 220, 349-365.

[9] Banerjee, A., Marcellino, M., Masten, I., 2008. Forecasting macroeconomic variables using diffusion indexes in short samples with structural change, Vol. 3. Emerald Group Publishing Limited, pp. 149–194.

[10] Bates, B., Plagborg-Moller, M., Stock, J.H., Watson, M.W., 2013. Consistent factor estimation in dynamic factor models with structural instability. Journal of Econometrics 177, 289–304.

[11] Breitung, J., Tenhofen, J., 2011. GLS estimation of dynamic factor models. Journal of the American Statistical Association 106, 1150–1156.

[12] Burns, A.F., Mitchell, W.C., 1946. Measuring business cycles. NBER.

[13] Chauvet, M., 1998. An econometric characterization of business cycle dynamics with factor structure and regime switching. International Economic Review 39, 969–996.

[14] Chauvet, M., Piger, J., 2008. A comparison of the real-time performance of business cycle dating methods. Journal of Business & Economic Statistics 26, 42–49.

[15] Cheng, X., Liao, Z., Shorfheide, F., 2016. Shrinkage estimation of high-dimensional factor models with structural instabilities. Review of Economic Studies 83, 1511–1543.

[16] Choi, I., 2012. Efficient estimation of factor models. Econometric Theory 28, 274–308.

[17] Cochrane, J.H., Piazzesi, M., 2005. Bond risk premia. American Economic Review 95, 138–160.

[18] Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. Journal of Econometrics 130, 337–364.

[19] Diebold, F.X., Rudebusch, G.D., 1996. Measuring business cycles: a modern perspective. Review of Economics and Statistics 78, 67–77.

[20] Doz, C., Giannone,.D., Reichlin, L., 2012. A quasi-maximum likelihood approach for large approximate dynamic factor models. Review of Economics and Statistics 94, 1014–1024.

[21] Gu, L., 2005. Asymmetric risk loadings in the cross section of stock returns", SSRN working paper 676845.

[22] Guidolin, M., Timmermann, A., 2008. Size and value anomalies under regime shifts. Journal of Financial Econometrics 6, 1-48.

[23] Hamilton, J.D., 1990. Analysis of time series subject to changes in regime. Journal of Econometrics 45, 39-70.

[24] Hamilton, J.D., 1994. Time series analysis. Princeton University Press.

[25] Harding, D., Pagan, A., 2006. Synchronization of cycles. Journal of Econometrics 132, 59–79.

[26] Kim, C.J., 1994. Dynamic linear models with Markov-switching. Journal of Econometrics 60, 1–22.

[27] Kim, C.J., 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. Review of Economics and Statistics 80, 188–201.

[28] Kim, M.J., Yoo, J.S., 1995. New index of coincident indicators: A multivariate Markov switching factor model approach. Journal of Monetary Economics 36, 607-630.

[29] Korobilis, D., 2013. Assessing the transmission of monetary policy using time-varying parameter dynamic factor models. Oxford Bulletin of Economics and Statistics 75, 157-179.

[30] Liu, X., Chen, R., 2016. Regime-switching factor models for high-dimensional time series. Statistica Sinica, 1427-1451.

[31] Ma, S., Su, L., 2018. Estimation of large dimensional factor models with an unknown number of breaks. Journal of Econometrics 207, 1–29.

[32] Massacci, D., 2017. Least squares estimation of large dimensional threshold factor models. Journal of Econometrics 197, 101-129.

[33] McCracken, M.W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34, 574-589.

[34] Mikkelsen, J.G., Hillebrand,.E., Urga, G., 2019. Consistent estimation of time-varying loadings in high-dimensional factor models. Journal of Econometrics 208, 535-562.

[35] Pelger, M., Xiong, R., 2021. State-varying factor models of large dimensions. Journal of Business & Economic Statistics, 1-19.

[36] Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. Journal of American Statistical Association 97, 1167–1179.

[37] Stock, J.H., Watson, M.W., 2009. Forecasting in dynamic factor models subject to structural instability. In: Hendry, D.F., Castle, J., Shephard, N. (Eds.), The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry. Oxford University Press, pp. 173–205.

[38] Stock, J.H., Watson, M.W., 2010. Indicators for dating business cycles: cross-history selection and comparisons. American Economic Review 100, 16–19.

[39] Stock, J.H., Watson, M.W., 2014. Estimating turning points using large data sets. Journal of Econometrics 178, 368-381.

[40] Su, L., Wang, X., 2017. On time-varying factor models: estimation and testing. Journal of Econometrics 198, 84-101.

APPENDIX

A Details for Theorem 1

Lemma 1 *Under Assumption 3(2) and 3(4),* $\|E\| = O_p(N^{\frac{1}{4}}T^{\frac{1}{2}} + N^{\frac{1}{2}}T^{\frac{1}{4}})$.

Proof. We shall show $\mathbb{E}\|E\|^4 = O(NT^2 + N^2T)$. First note that

$$\|E\|^4 = \|E'E\|^2 \leq \|E'E\|_F^2 = \sum_{i=1}^{N}\sum_{k=1}^{N}(\sum_{t=1}^{T} e_{it}e_{kt})^2.$$

It is easy to see that $\mathbb{E}(\sum_{t=1}^{T} e_{it}e_{kt})^2$ is not larger than the sum of $2\mathbb{E}(\sum_{t=1}^{T} e_{it}e_{kt} - \sum_{t=1}^{T}\mathbb{E}(e_{it}e_{kt}))^2$ and $2(\sum_{t=1}^{T}\mathbb{E}(e_{it}e_{kt}))^2$. The sum of the former over i and k is not larger than N^2TM since by Assumption 3(4), $\mathbb{E}(\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(e_{it}e_{kt} - \mathbb{E}(e_{it}e_{kt}))\right\|^2) \leq M$. The sum of the latter over i and k is not larger than NT^2M under Assumption 3(2). ■

Proof of Theorem 1

Proof. Step (1): For $(\hat{\Lambda}, \hat{\sigma}^2, q)$, let

$$m_t = \arg\max_j\{(2\pi)^{-\frac{N}{2}}\left|\hat{\Lambda}_j\hat{\Lambda}_j' + \hat{\sigma}^2 I_N\right|^{-\frac{1}{2}} e^{-\frac{1}{2}x_t'(\hat{\Lambda}_j\hat{\Lambda}_j'+\hat{\sigma}^2 I_N)^{-1}x_t}, j = 1, ..., J^0\}. \tag{28}$$

Since $\sum_{j=1}^{J^0} q_j = 1$,

$$\begin{aligned} l(\hat{\Lambda}, \hat{\sigma}^2, q) &\leq \sum_{t=1}^{T}\log[(2\pi)^{-\frac{N}{2}}\left|\hat{\Lambda}_{m_t}\hat{\Lambda}_{m_t}' + \hat{\sigma}^2 I_N\right|^{-\frac{1}{2}} e^{-\frac{1}{2}x_t'(\hat{\Lambda}_{m_t}\hat{\Lambda}_{m_t}'+\hat{\sigma}^2 I_N)^{-1}x_t}] \\ &= -\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left|\hat{\Lambda}_{m_t}\hat{\Lambda}_{m_t}' + \hat{\sigma}^2 I_N\right| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T} x_t'(\hat{\Lambda}_{m_t}\hat{\Lambda}_{m_t}' + \hat{\sigma}^2 I_N)^{-1}x_t. \end{aligned} \tag{29}$$

Consider the last term on the right hand side of equation (29). By Woodbury identity, $(\hat{\Lambda}_{m_t}\hat{\Lambda}_{m_t}' + \hat{\sigma}^2 I_N)^{-1} = \hat{\sigma}^{-2}I_N - \hat{\sigma}^{-2}\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r^0_{m_t}} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}'$. Thus

$$\sum_{t=1}^{T} x_t'(\hat{\Lambda}_{m_t}\hat{\Lambda}_{m_t}'+\hat{\sigma}^2 I_N)^{-1}x_t = \hat{\sigma}^{-2}\sum_{t=1}^{T} x_t'x_t - \hat{\sigma}^{-2}\sum_{t=1}^{T} x_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r^0_{m_t}} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}'x_t.$$

Since $(\hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1} - (\hat{\sigma}^2 I_{r^0_{m_t}} + \hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1} = (\hat{\sigma}^2 I_{r^0_{m_t}} + \hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1}\hat{\sigma}^2(\hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1}$,

$$\begin{aligned} & \sum\nolimits_{t=1}^{T} x'_t(\hat{\Lambda}_{m_t}\hat{\Lambda}'_{m_t} + \hat{\sigma}^2 I_N)^{-1}x_t \\ = \; & \hat{\sigma}^{-2}\sum\nolimits_{t=1}^{T}\left\|M_{\hat{\Lambda}_{m_t}}x_t\right\|^2 + \sum\nolimits_{t=1}^{T} x'_t\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r^0_{m_t}} + \hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}'_{m_t}x_t. \end{aligned} \tag{30}$$

Step (2): $l(\Lambda^0,\hat{\sigma}^2,q) = \sum_{t=1}^{T}\log[\sum_{j=1}^{J^0} q_j(2\pi)^{-\frac{N}{2}}\left|\Lambda^0_j\Lambda^{0\prime}_j + \hat{\sigma}^2 I_N\right|^{-\frac{1}{2}} e^{-\frac{1}{2}x'_t(\Lambda^0_j\Lambda^{0\prime}_j+\hat{\sigma}^2 I_N)^{-1}x_t}]$.

The summation in the bracket has J^0 terms. Throw away all the other terms and only keep the term for $j = z_t$, we have

$$\begin{aligned} l(\Lambda^0,\hat{\sigma}^2,q) \;\geq\; & \sum\nolimits_{t=1}^{T}\log[q_{z_t}(2\pi)^{-\frac{N}{2}}\left|\Lambda^0_{z_t}\Lambda^{0\prime}_{z_t} + \hat{\sigma}^2 I_N\right|^{-\frac{1}{2}} e^{-\frac{1}{2}x'_t(\Lambda^0_{z_t}\Lambda^{0\prime}_{z_t}+\hat{\sigma}^2 I_N)^{-1}x_t}] \\ = \; & \sum\nolimits_{t=1}^{T}\log q_{z_t} - \frac{NT}{2}\log 2\pi - \frac{1}{2}\sum\nolimits_{t=1}^{T}\log\left|\Lambda^0_{z_t}\Lambda^{0\prime}_{z_t} + \hat{\sigma}^2 I_N\right| \\ & -\frac{1}{2}\sum\nolimits_{t=1}^{T} x'_t(\Lambda^0_{z_t}\Lambda^{0\prime}_{z_t} + \hat{\sigma}^2 I_N)^{-1}x_t. \end{aligned} \tag{31}$$

Similar to equation (30),

$$\begin{aligned} & \sum\nolimits_{t=1}^{T} x'_t(\Lambda^0_{z_t}\Lambda^{0\prime}_{z_t} + \hat{\sigma}^2 I_N)^{-1}x_t \\ = \; & \hat{\sigma}^{-2}\sum\nolimits_{t=1}^{T}\left\|M_{\Lambda^0_{z_t}}x_t\right\|^2 + \sum\nolimits_{t=1}^{T} x'_t\Lambda^0_{z_t}(\hat{\sigma}^2 I_{r^0_{z_t}} + \Lambda^{0\prime}_{z_t}\Lambda^0_{z_t})^{-1}(\Lambda^{0\prime}_{z_t}\Lambda^0_{z_t})^{-1}\Lambda^{0\prime}_{z_t}x_t. \end{aligned} \tag{32}$$

Step (3): $l(\hat{\Lambda},\hat{\sigma}^2,q) - l(\Lambda^0,\hat{\sigma}^2,q) \geq 0$. Thus from equations (29)-(32), we have

$$\begin{aligned} & \frac{1}{2}[\hat{\sigma}^{-2}\sum\nolimits_{t=1}^{T}\left\|M_{\hat{\Lambda}_{m_t}}x_t\right\|^2 - \hat{\sigma}^{-2}\sum\nolimits_{t=1}^{T}\left\|M_{\Lambda^0_{z_t}}x_t\right\|^2] \\ \leq \; & -\sum\nolimits_{t=1}^{T}\log q_{z_t} - \frac{1}{2}\sum\nolimits_{t=1}^{T}\log\frac{\left|\hat{\Lambda}_{m_t}\hat{\Lambda}'_{m_t} + \hat{\sigma}^2 I_N\right|}{\left|\Lambda^0_{z_t}\Lambda^{0\prime}_{z_t} + \hat{\sigma}^2 I_N\right|} \\ & -\frac{1}{2}\sum\nolimits_{t=1}^{T} x'_t\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r^0_{m_t}} + \hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}'_{m_t}x_t \\ & +\frac{1}{2}\sum\nolimits_{t=1}^{T} x'_t\Lambda^0_{z_t}(\hat{\sigma}^2 I_{r^0_{z_t}} + \Lambda^{0\prime}_{z_t}\Lambda^0_{z_t})^{-1}(\Lambda^{0\prime}_{z_t}\Lambda^0_{z_t})^{-1}\Lambda^{0\prime}_{z_t}x_t. \end{aligned} \tag{33}$$

(3.1) The first term on the right hand side: $\left|-\sum_{t=1}^{T}\log q_{z_t}\right| \leq T\log\frac{1}{\min_j q_j} = O(T)$.

(3.2) The second term on the right hand side equals $-\frac{1}{2}\sum_{t=1}^{T}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}'_{m_t}\hat{\Lambda}_{m_t} + I_{r^0_{m_t}}\right| +$

$\frac{1}{2}\sum_{t=1}^{T}\log\left|\frac{1}{\hat\sigma^2}\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0+I_{r_{z_t}^0}\right|$ since

$$\left|\Lambda_{z_t}^0\Lambda_{z_t}^{0\prime}+\hat\sigma^2 I_N\right| = \hat\sigma^{2N}\left|\frac{1}{\hat\sigma^2}\Lambda_{z_t}^0\Lambda_{z_t}^{0\prime}+I_N\right| = \hat\sigma^{2N}\left|\frac{1}{\hat\sigma^2}\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0+I_{r_{z_t}^0}\right|, \quad (34)$$

$$\text{and similarly } \left|\hat\Lambda_{m_t}\hat\Lambda_{m_t}'+\hat\sigma^2 I_N\right| = \hat\sigma^{2N}\left|\frac{1}{\hat\sigma^2}\hat\Lambda_{m_t}'\hat\Lambda_{m_t}+I_{r_{m_t}^0}\right|. \quad (35)$$

$-\frac{1}{2}\sum_{t=1}^{T}\log\left|\frac{1}{\hat\sigma^2}\hat\Lambda_{m_t}'\hat\Lambda_{m_t}+I_{r_{m_t}^0}\right|$ is negative, thus inequality (33) still holds when this term is thrown away. By Assumption 2(1), $\left|\frac{1}{\hat\sigma^2}\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0+I_{r_{z_t}^0}\right|\le c(\frac{N}{\hat\sigma^2})^{r_{z_t}^0}$ for some $c>0$, thus $\frac{1}{2}\sum_{t=1}^{T}\log\left|\frac{1}{\hat\sigma^2}\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0+I_{r_{z_t}^0}\right|$ is $O_p(T\log N)$.

(3.3) The third term on the right hand side is negative, thus inequality (33) still holds when this term is thrown away.

(3.4) The fourth term is bounded by $\frac{1}{2}\sum_{t=1}^{T}\|x_t\|^2\left\|(\hat\sigma^2 I_{r_{z_t}^0}+\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{-1}\right\|$ since $\left\|(\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{-\frac{1}{2}}\Lambda_{z_t}^{0\prime}x_t\right\|=\left\|P_{\Lambda_{z_t}^0}x_t\right\|\le\|x_t\|$ and $(\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{\frac{1}{2}}(\hat\sigma^2 I_{r_{z_t}^0}+\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{-1}(\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{-\frac{1}{2}}=(\hat\sigma^2 I_{r_{z_t}^0}+\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{-1}$. The latter is because $\hat\sigma^2 I_{r_{z_t}^0}+\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0$ and $\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0$ have the same eigenvectors. By Assumption 2(1), $\left\|(\hat\sigma^2 I_{r_{z_t}^0}+\Lambda_{z_t}^{0\prime}\Lambda_{z_t}^0)^{-1}\right\|\le\sup_j\left\|(\Lambda_j^{0\prime}\Lambda_j^0)^{-1}\right\|=O_p(\frac{1}{N})$. By Assumptions 1(2), 2(1) and 3(1), $\sum_{t=1}^{T}\|x_t\|^2=O_p(NT)$. Thus the fourth term is $O_p(T)$.

(3.5) Now consider the left hand side of expression (33). Since $x_t=\Lambda_{z_t}^0 f_t^0+e_t$ and $M_{\Lambda_{z_t}^0}\Lambda_{z_t}^0 f_t^0=0$, it is easy to verify that the left hand side equals

$$\frac{1}{2}\hat\sigma^{-2}[\sum_{t=1}^{T}\left\|M_{\hat\Lambda_{m_t}}\Lambda_{z_t}^0 f_t^0\right\|^2+2\sum_{t=1}^{T}e_t'M_{\hat\Lambda_{m_t}}\Lambda_{z_t}^0 f_t^0+\sum_{t=1}^{T}\left\|P_{\Lambda_{z_t}^0}e_t\right\|^2-\sum_{t=1}^{T}\left\|P_{\hat\Lambda_{m_t}}e_t\right\|^2]. \quad (36)$$

For the fourth term of expression (36), we have

$$\begin{aligned}\sum_{t=1}^{T}\left\|P_{\hat\Lambda_{m_t}}e_t\right\|^2 &\le \sum_{j=1}^{J^0}\sum_{t=1}^{T}\left\|P_{\hat\Lambda_j}e_t\right\|^2=\sum_{j=1}^{J^0}\sum_{t=1}^{T}e_t'\hat\Lambda_j(\hat\Lambda_j'\hat\Lambda_j)^{-1}\hat\Lambda_j'e_t\\ &= \sum_{j=1}^{J^0}tr[(\hat\Lambda_j'\hat\Lambda_j)^{-\frac{1}{2}}\hat\Lambda_j'(\sum_{t=1}^{T}e_te_t')\hat\Lambda_j(\hat\Lambda_j'\hat\Lambda_j)^{-\frac{1}{2}}]\\ &\le \sum_{j=1}^{J^0}r_j^0\rho_{\max}(\sum_{t=1}^{T}e_te_t')=\sum_{j=1}^{J^0}r_j^0\|E'E\|=O_p(N^{\frac{1}{2}}T+NT^{\frac{1}{2}}).\end{aligned} \quad (37)$$

The last equality follows from Lemma 1. Similarly, the third term of expression (36) is $O_p(N^{\frac{1}{2}}T+NT^{\frac{1}{2}})$.

The second term of expression (36) equals $2\sum_{t=1}^{T} e_t'\Lambda_{z_t}^0 f_t^0 - 2\sum_{t=1}^{T} e_t' P_{\hat{\Lambda}_{m_t}} \Lambda_{z_t}^0 f_t^0$. $2\sum_{t=1}^{T} e_t'\Lambda_{z_t}^0 f_t^0$ is $O_p(N^{\frac{1}{2}}T)$ because $\left\|\sum_{t=1}^{T} e_t'\Lambda_{z_t}^0 f_t^0\right\| \leq (\sum_{t=1}^{T} \left\|e_t'\Lambda_{z_t}^0\right\|^2)^{\frac{1}{2}}(\sum_{t=1}^{T} \|f_t^0\|^2)^{\frac{1}{2}}$, $\sum_{t=1}^{T} \|f_t^0\|^2 = O_p(T)$ by Assumption 1(2), and by Assumptions 2(1), 3(1) and 3(2), $\mathbb{E}(\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_{ji}^0 e_{it}\right\|^2) \leq M$ for all $j = 1, ..., J^0$ and all t.

By expression (37), Assumption 1(2) and Assumption 2(1), $\left\|\sum_{t=1}^{T} e_t' P_{\hat{\Lambda}_{m_t}} \Lambda_{z_t}^0 f_t^0\right\| \leq (\sum_{t=1}^{T} \left\|P_{\hat{\Lambda}_{m_t}} e_t\right\|^2 \sum_{t=1}^{T} \|f_t^0\|^2)^{\frac{1}{2}} \sup_j \left\|\Lambda_j^0\right\| = O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}})$. Thus the second term of expression (36) is $O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}})$.

(3.6) Move the second to the fourth term of expression (36) to the right hand side of equation (33), and take the results (3.1)-(3.5) together, we have

$$\begin{aligned} 0 \leq & \frac{1}{2}\hat{\sigma}^{-2}\sum_{t=1}^{T}\left\|M_{\hat{\Lambda}_{m_t}}\Lambda_{z_t}^0 f_t^0\right\|^2 \leq O(T) + O_p(T\log N) \\ & + O(T) + O_p(N^{\frac{1}{2}}T + NT^{\frac{1}{2}}) + O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}}). \end{aligned}$$

Thus $\sum_{t=1}^{T}\left\|M_{\hat{\Lambda}_{m_t}}\Lambda_{z_t}^0 f_t^0\right\|^2$ is $O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}})$. In the summation, there are $q_1^0 T$ terms[11] with $\Lambda_{z_t}^0 = \Lambda_1^0$, since q_1^0 is the unconditional probability of $z_t = 1$. For each t with $z_t = 1$, $\Lambda_1^0 f_t^0$ are projected on one of $\hat{\Lambda}_j, j = 1, ..., J^0$, thus there exists one certain $\hat{\Lambda}_j$ such that $\Lambda_1^0 f_t^0$ is projected on $\hat{\Lambda}_j$ at least $\frac{q_1^0 T}{J^0}$ times. Define this $\hat{\Lambda}_j$ as $\hat{\Lambda}_1$, then $\sum_{t=1}^{T} 1_{m_t=1}1_{z_t=1} \geq \frac{q_1^0 T}{J^0}$. Thus by Assumption 1(1), $\rho_{\min}(\frac{1}{\sum_{t=1}^{T} 1_{m_t=1}1_{z_t=1}}\sum_{t=1}^{T} f_t^0 f_t^{0\prime} 1_{m_t=1}1_{z_t=1}) \geq c$ for some $c > 0$ w.p.a.1. Since $\left\|M_{\hat{\Lambda}_{m_t}}\Lambda_{z_t}^0 f_t^0\right\|$ is positive for any z_t and m_t, we have

$$\begin{aligned} O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}}) &= \sum_{t=1}^{T}\left\|M_{\hat{\Lambda}_{m_t}}\Lambda_{z_t}^0 f_t^0\right\|^2 \geq \sum_{t=1}^{T}\left\|M_{\hat{\Lambda}_1}\Lambda_1^0 f_t^0\right\|^2 1_{m_t=1}1_{z_t=1} \\ &= tr(\Lambda_1^{0\prime} M_{\hat{\Lambda}_1}\Lambda_1^0 \sum_{t=1}^{T} f_t^0 f_t^{0\prime} 1_{m_t=1}1_{z_t=1}) \\ &\geq tr(\Lambda_1^{0\prime} M_{\hat{\Lambda}_1}\Lambda_1^0)\rho_{\min}(\sum_{t=1}^{T} f_t^0 f_t^{0\prime} 1_{m_t=1}1_{z_t=1}) \\ &\geq tr(\Lambda_1^{0\prime} M_{\hat{\Lambda}_1}\Lambda_1^0) T\frac{q_1^0 c}{J^0} \text{ w.p.a.1.} \end{aligned}$$

Thus $\frac{1}{N}\left\|M_{\hat{\Lambda}_1}\Lambda_1^0\right\|_F^2 = \frac{1}{N}tr(\Lambda_1^{0\prime} M_{\hat{\Lambda}_1}\Lambda_1^0) = O_p(\frac{1}{\sqrt{\delta_{NT}}})$. Similarly, for $j = 2, ..., J^0$, we also have $\frac{1}{N}\left\|M_{\hat{\Lambda}_j}\Lambda_j^0\right\|_F^2 = O_p(\frac{1}{\sqrt{\delta_{NT}}})$. ∎

[11]Rigorously speaking, there are $\sum_{t=1}^{T} 1_{z_t=1}$ terms, but $\frac{1}{T}\sum_{t=1}^{T} 1_{z_t=1} \xrightarrow{p} q_1^0$ as $T \to \infty$.

B Details for Theorem 2

Lemma 2 *Under the assumptions of Theorem 1,* $\frac{1}{\hat{\sigma}^2+\hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}} = O_p(\frac{1}{\sqrt{\delta_{NT}}})$ *for each* $j = 1, ..., J^0$ *and each* $l = 1, ..., r_j^0$*, where* $\hat{\Lambda}_{jl}$ *denotes the* l*-th column of* $\hat{\Lambda}_j$*.*

Proof. (1) Consider expression (33). In step (3.1), (3.2) and (3.4) of proof of Theorem 1, we have shown that the first, the second, and the fourth term on the right hand side of expression (33) is $O_p(T)$, $O_p(T \log N)$ and $O_p(T)$ respectively. In step (3.5) we have shown that the left hand side of expression (33) equals expression (36), and the last three terms of expression (36) together is $O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}})$. Move the last three terms of expression (36) to the right hand side of expression (33), and move the third term on the right hand side of expression (33) to the left hand side, then we have

$$\begin{aligned}
& \frac{1}{2}\hat{\sigma}^{-2}[\sum\nolimits_{t=1}^{T} \left\| M_{\hat{\Lambda}_{m_t}} \Lambda_{z_t}^0 f_t^0 \right\|^2 + \frac{1}{2}\sum\nolimits_{t=1}^{T} x_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}' x_t \\
= \; & O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}}). \qquad (38)
\end{aligned}$$

The two terms on the left hand side of (38) are nonnegative, thus $\sum_{t=1}^{T} x_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}' x_t = O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}})$.

(2)

$$\begin{aligned}
& \left\| \sum\nolimits_{t=1}^{T} e_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}'\Lambda_{z_t}^0 f_t^0 \right\| \\
\leq \; & (\sum\nolimits_{t=1}^{T} e_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-2}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}' e_t)^{\frac{1}{2}} (\sum\nolimits_{t=1}^{T} \left\| \Lambda_{z_t}^0 f_t^0 \right\|^2)^{\frac{1}{2}} \\
\leq \; & \frac{1}{\hat{\sigma}^2}(\sum\nolimits_{t=1}^{T} e_t' P_{\hat{\Lambda}_{m_t}} e_t)^{\frac{1}{2}} (\sum\nolimits_{t=1}^{T} \left\| f_t^0 \right\|^2)^{\frac{1}{2}} \sup_j \left\| \Lambda_j^0 \right\| = O_p(N^{\frac{3}{4}}T + NT^{\frac{3}{4}}),
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second inequality follows from the fact that $\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t}$ is diagonal and all diagonal elements of $\hat{\sigma}^2 I_{r_{m_t}^0} + \hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t}$ are larger than $\hat{\sigma}^2$, and the equality follows from Assumption 1(2), Assumption 2(1) and expression (37) in step (3.5) of proof of Theorem 1.

(3) It follows from (1) and (2) that

$$\begin{aligned}
& \sum_{t=1}^{T} f_t^{0\prime}\Lambda_{z_t}^{0\prime}\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0}+\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}'\Lambda_{z_t}^0 f_t^0 \\
& +\sum_{t=1}^{T} e_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0}+\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}' e_t \\
= & \sum_{t=1}^{T} x_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0}+\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}' x_t \\
& -2\sum_{t=1}^{T} e_t'\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0}+\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}'\Lambda_{z_t}^0 f_t^0 \\
= & O_p(N^{\frac{3}{4}}T+NT^{\frac{3}{4}}). \qquad (39)
\end{aligned}$$

The two terms on the left hand side of (39) are nonnegative, thus $\sum_{t=1}^{T} f_t^{0\prime}\Lambda_{z_t}^{0\prime}\hat{\Lambda}_{m_t}(\hat{\sigma}^2 I_{r_{m_t}^0}+\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}(\hat{\Lambda}_{m_t}'\hat{\Lambda}_{m_t})^{-1}\hat{\Lambda}_{m_t}'\Lambda_{z_t}^0 f_t^0 = O_p(N^{\frac{3}{4}}T+NT^{\frac{3}{4}})$. Since each term in the summation is nonnegative, we have $\sum_{t=1}^{T} f_t^{0\prime}\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0 f_t^0 1_{z_t=j}1_{m_t=j} = O_p(N^{\frac{3}{4}}T+NT^{\frac{3}{4}})$ for each j.

As explained in step (3.6) of proof of Theorem 1, $\sum_{t=1}^{T} 1_{m_t=j}1_{z_t=j} \geq \frac{q_j^0 T}{J^0}$, and by Assumption 1(1), $\rho_{\min}(\frac{1}{\sum_{t=1}^{T} 1_{m_t=j}1_{z_t=j}}\sum_{t=1}^{T} f_t^0 f_t^{0\prime}1_{m_t=j}1_{z_t=j}) \geq c$ for some $c>0$ w.p.a.1. Thus we have

$$\begin{aligned}
O_p(N^{\frac{3}{4}}T+NT^{\frac{3}{4}}) &= \sum_{t=1}^{T} f_t^{0\prime}\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0 f_t^0 1_{z_t=j}1_{m_t=j} \\
&= tr(\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0 \sum_{t=1}^{T} f_t^0 f_t^{0\prime}1_{z_t=j}1_{m_t=j}) \\
&\geq tr(\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0)\rho_{\min}(\sum_{t=1}^{T} f_t^0 f_t^{0\prime}1_{m_t=j}1_{z_t=j}) \\
&\geq tr(\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0)T\frac{q_j^0 c}{J^0} \text{ w.p.a.1.}
\end{aligned}$$

Thus $tr(\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0) = O_p(\frac{N}{\sqrt{\delta_{NT}}})$ for each j.

(4) Noting that $\hat{\Lambda}_{jl}$ is orthogonal to $\hat{\Lambda}_{jl'}$ for $l \neq l'$, we have

$$\sum_{l=1}^{r_j^0}\left\|P_{\hat{\Lambda}_{jl}}\Lambda_j^0\right\|_F^2 = \left\|P_{\hat{\Lambda}_j}\Lambda_j^0\right\|_F^2 = \left\|\Lambda_j^0\right\|_F^2 - \left\|M_{\hat{\Lambda}_j}\Lambda_j^0\right\|_F^2, \qquad (40)$$

$$\sum_{l=1}^{r_j^0}\frac{1}{\hat{\sigma}^2+\hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}}\left\|P_{\hat{\Lambda}_{jl}}\Lambda_j^0\right\|_F^2 = tr(\Lambda_j^{0\prime}\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0}+\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'\Lambda_j^0). \qquad (41)$$

Each term in the summation on the left hand side of equation (41) is nonnegative, thus

$$\frac{1}{\hat{\sigma}^2+\hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}}\left\|P_{\hat{\Lambda}_{jl}}\Lambda_j^0\right\|_F^2=O_p(\frac{N}{\sqrt{\delta_{NT}}})\text{ for each } j \text{ and } l. \tag{42}$$

Now consider $\left\|P_{\hat{\Lambda}_{j1}}\Lambda_j^0\right\|_F^2$. Let $\hat{\Lambda}_{j,-1}=(\hat{\Lambda}_{j2},...,\hat{\Lambda}_{jr_j^0})$, we have

$$\begin{aligned}\sum\nolimits_{l\neq 1}\left\|P_{\hat{\Lambda}_{jl}}\Lambda_j^0\right\|_F^2 &= \left\|P_{\hat{\Lambda}_{j,-1}}\Lambda_j^0\right\|_F^2=tr(\Lambda_j^{0\prime}P_{\hat{\Lambda}_{j,-1}}\Lambda_j^0)\\ &= tr[(\hat{\Lambda}_{j,-1}'\hat{\Lambda}_{j,-1})^{-\frac{1}{2}}\hat{\Lambda}_{j,-1}'\Lambda_j^0\Lambda_j^{0\prime}\hat{\Lambda}_{j,-1}(\hat{\Lambda}_{j,-1}'\hat{\Lambda}_{j,-1})^{-\frac{1}{2}}]\\ &\leq \left\|\Lambda_j^0\right\|_F^2-\rho_{\min}(\Lambda_j^0\Lambda_j^{0\prime}). \end{aligned}\tag{43}$$

The inequality in expression (43) becomes equality when $\hat{\Lambda}_{j,-1}(\hat{\Lambda}_{j,-1}'\hat{\Lambda}_{j,-1})^{-\frac{1}{2}}$ are eigenvectors of $\Lambda_j^0\Lambda_j^{0\prime}$ corresponding to the largest r_j^0-1 eigenvalues. Expressions (40) and (43) together implies that $\left\|P_{\hat{\Lambda}_{j1}}\Lambda_j^0\right\|_F^2\geq\rho_{\min}(\Lambda_j^0\Lambda_j^{0\prime})-\left\|M_{\hat{\Lambda}_j}\Lambda_j^0\right\|_F^2$, thus by Assumption 2(1) and Theorem 1, $\frac{1}{N}\left\|P_{\hat{\Lambda}_{j1}}\Lambda_j^0\right\|_F^2$ is bounded away from zero in probability. This together with expression (42) implies that $\frac{1}{\hat{\sigma}^2+\hat{\Lambda}_{j1}'\hat{\Lambda}_{j1}}=O_p(\frac{1}{\sqrt{\delta_{NT}}})$. Similarly, $\frac{1}{\hat{\sigma}^2+\hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}}=O_p(\frac{1}{\sqrt{\delta_{NT}}})$ for $l=2,...,r_j^0$. ∎

Proof of Theorem 2

Proof. (1) From equation (12), $\hat{p}_{tj}=\frac{q_jL(x_t\left|z_t=j;\hat{\Lambda}_j,\hat{\sigma}^2\right.)}{\sum_{k=1}^{J^0}q_kL(x_t\left|z_t=k;\hat{\Lambda}_k,\hat{\sigma}^2\right.)}$. Thus when $z_t=j$,

$$|\hat{p}_{tj}-1_{z_t=j}|=\frac{\sum_{k\neq j}q_kL(x_t\left|z_t=k;\hat{\Lambda}_k,\hat{\sigma}^2\right.)}{\sum_{k=1}^{J^0}q_kL(x_t\left|z_t=k;\hat{\Lambda}_k,\hat{\sigma}^2\right.)}\leq\sum\nolimits_{k\neq j}\frac{q_k}{q_j}e^{\log L(x_t\left|z_t=k;\hat{\Lambda}_k,\hat{\sigma}^2\right.)-\log L(x_t\left|z_t=j;\hat{\Lambda}_j,\hat{\sigma}^2\right.)},$$

and when $z_t=h\neq j$,

$$|\hat{p}_{tj}-1_{z_t=j}|=\frac{q_jL(x_t\left|z_t=j;\hat{\Lambda}_j,\hat{\sigma}^2\right.)}{\sum_{k=1}^{J^0}q_kL(x_t\left|z_t=k;\hat{\Lambda}_k,\hat{\sigma}^2\right.)}\leq\frac{q_j}{q_h}e^{\log L(x_t\left|z_t=j;\hat{\Lambda}_j,\hat{\sigma}^2\right.)-\log L(x_t\left|z_t=h;\hat{\Lambda}_h,\hat{\sigma}^2\right.)}.$$

Since $\sum_{k=1}^{J^0}\hat{p}_{tk}=1$, for $\hat{p}_{tj}$ it suffices to consider the case $z_t=j$.

Since $q_j>0$, it suffices to show $\sup_t e^{\log L(x_t\left|z_t=k;\hat{\Lambda}_k,\hat{\sigma}^2\right.)-\log L(x_t\left|z_t=j;\hat{\Lambda}_j,\hat{\sigma}^2\right.)}=o_p(\frac{1}{N^\eta})$

for any $k \neq j$, i.e., it suffices to show for any $M > 0$,

$$
\begin{aligned}
&\Pr(\sup_t[\log L(x_t \Big| z_t = k; \hat{\Lambda}_k, \hat{\sigma}^2) - \log L(x_t \Big| z_t = j; \hat{\Lambda}_j, \hat{\sigma}^2)] \geq \log \frac{M}{N^\eta}) \to 0, \\
\text{or } &\Pr(\min_t[\log L(x_t \Big| z_t = j; \hat{\Lambda}_j, \hat{\sigma}^2) - \log L(x_t \Big| z_t = k; \hat{\Lambda}_k, \hat{\sigma}^2)] \leq \eta \log N - \log M) \to 0 \quad (44)
\end{aligned}
$$

Similar to equations (32),

$$
\begin{aligned}
\log L(x_t \Big| z_t = j; \hat{\Lambda}_j, \hat{\sigma}^2) &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\left|\hat{\Lambda}_j\hat{\Lambda}_j' + \hat{\sigma}^2 I_N\right| - \frac{1}{2}\hat{\sigma}^{-2}\left\|M_{\hat{\Lambda}_j}x_t\right\|^2 \\
&\quad -\frac{1}{2}x_t'\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' x_t, \\
\log L(x_t \Big| z_t = k; \hat{\Lambda}_k, \hat{\sigma}^2) &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\left|\hat{\Lambda}_k\hat{\Lambda}_k' + \hat{\sigma}^2 I_N\right| - \frac{1}{2}\hat{\sigma}^{-2}\left\|M_{\hat{\Lambda}_k}x_t\right\|^2 \\
&\quad -\frac{1}{2}x_t'\hat{\Lambda}_k(\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}(\hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' x_t,
\end{aligned}
$$

and similar to equations (34) and (35), $\frac{\left|\hat{\Lambda}_j\hat{\Lambda}_j' + \hat{\sigma}^2 I_N\right|}{\left|\hat{\Lambda}_k\hat{\Lambda}_k' + \hat{\sigma}^2 I_N\right|} = \frac{\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_j'\hat{\Lambda}_j + I_{r_j^0}\right|}{\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_k'\hat{\Lambda}_k + I_{r_k^0}\right|}$. Thus

$$
\begin{aligned}
&\log L(x_t \Big| z_t = j; \hat{\Lambda}_j, \hat{\sigma}^2) - \log L(x_t \Big| z_t = k; \hat{\Lambda}_k, \hat{\sigma}^2) \\
= &-\frac{1}{2}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_j'\hat{\Lambda}_j + I_{r_j^0}\right| + \frac{1}{2}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_k'\hat{\Lambda}_k + I_{r_k^0}\right| \\
&+\frac{1}{2}\hat{\sigma}^{-2}(\left\|M_{\hat{\Lambda}_k}x_t\right\|^2 - \left\|M_{\Lambda_k^0}x_t\right\|^2 + \left\|M_{\Lambda_j^0}x_t\right\|^2 - \left\|M_{\hat{\Lambda}_j}x_t\right\|^2 + \left\|M_{\Lambda_k^0}x_t\right\|^2 - \left\|M_{\Lambda_j^0}x_t\right\|^2) \\
&-\frac{1}{2}x_t'\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' x_t + \frac{1}{2}x_t'\hat{\Lambda}_k(\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}(\hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' x_t \\
\geq &-\frac{1}{2}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_j'\hat{\Lambda}_j + I_{r_j^0}\right| - \frac{1}{2}x_t'\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' x_t \\
&+\frac{1}{2}\hat{\sigma}^{-2}(\left\|M_{\hat{\Lambda}_k}x_t\right\|^2 - \left\|M_{\Lambda_k^0}x_t\right\|^2) + \frac{1}{2}\hat{\sigma}^{-2}(\left\|M_{\Lambda_j^0}x_t\right\|^2 - \left\|M_{\hat{\Lambda}_j}x_t\right\|^2) \\
&-\frac{1}{2}\hat{\sigma}^{-2}e_t'P_{\Lambda_k^0}e_t + \hat{\sigma}^{-2}e_t'M_{\Lambda_k^0}\Lambda_j^0 f_t^0 + \frac{1}{2}\hat{\sigma}^{-2}f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0, \qquad (45)
\end{aligned}
$$

where the inequality follows from throwing away $\frac{1}{2}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_k'\hat{\Lambda}_k + I_{r_k^0}\right|$, $\frac{1}{2}x_t'\hat{\Lambda}_k(\hat{\sigma}^2 I_{r_k^0} +$

$\hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}(\hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k'x_t$ and $e_t'P_{\Lambda_j^0}e_t$. It follows that

$$\begin{aligned}
& \min_t[\log L(x_t \left| z_t = j; \hat{\Lambda}_j, \hat{\sigma}^2) - \log L(x_t \right| z_t = k; \hat{\Lambda}_k, \hat{\sigma}^2)] \\
\geq & -\frac{1}{2}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_j'\hat{\Lambda}_j + I_{r_j^0}\right| - \frac{1}{2}\sup_t x_t'\hat{\Lambda}_j(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'x_t \\
& -\frac{1}{2}\hat{\sigma}^{-2}\sup_t\left|\left\|M_{\hat{\Lambda}_k}x_t\right\|^2 - \left\|M_{\Lambda_k^0}x_t\right\|^2\right| - \frac{1}{2}\hat{\sigma}^{-2}\sup_t\left|\left\|M_{\Lambda_j^0}x_t\right\|^2 - \left\|M_{\hat{\Lambda}_j}x_t\right\|^2\right| \\
& -\frac{1}{2}\hat{\sigma}^{-2}\sup_t e_t'P_{\Lambda_k^0}e_t - \hat{\sigma}^{-2}\sup_t\left|e_t'M_{\Lambda_k^0}\Lambda_j^0 f_t^0\right| + \frac{1}{2}\hat{\sigma}^{-2}\min_t f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0 \\
\equiv & -(A_1 + A_2 + A_3 + A_4 + A_5 + A_6) + \frac{1}{2}\hat{\sigma}^{-2}\min_t f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0. \qquad (46)
\end{aligned}$$

Thus for expression (44), it suffices to show

$$\Pr(\frac{1}{2}\hat{\sigma}^{-2}\min_t f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0 \leq A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + \eta\log N) \to 0. \quad (47)$$

By Assumption 2(2), $\min_t f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0 \geq NC$ for some $C > 0$. Thus it suffices to show that $A_1, ..., A_6$ are all $o_p(N)$ when $T^{\frac{16}{\alpha}}/N \to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0$.

Term A_1: As shown in equation (14), $\hat{\Lambda}_{jl}'\hat{\Lambda}_{jl} + \hat{\sigma}^2$ is eigenvalue of $S_j = \frac{\sum_{t=1}^T p_{tj}x_t x_t'}{\sum_{t=1}^T p_{tj}}$, which is bounded by $\sup_t \|x_t\|^2$. We next show that $\sup_t \|x_t\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{\alpha}})$. By Assumption 2(1) and 1(2), $\sup_t \left\|\Lambda_{z_t}^0 f_t^0\right\|^\alpha \leq \sup_j \left\|\Lambda_j^0\right\|^\alpha \sum_{t=1}^T \|f_t^0\|^\alpha = N^{\frac{\alpha}{2}}T$. By Holder inequality, $\|e_t\|^2 = \sum_{i=1}^N e_{it}^2 \leq (\sum_{i=1}^N e_{it}^\alpha)^{\frac{2}{\alpha}}N^{1-\frac{2}{\alpha}}$, thus $\sup_t \|e_t\|^\alpha \leq N^{\frac{\alpha}{2}-1}\sup_t(\sum_{i=1}^N e_{it}^\alpha) \leq N^{\frac{\alpha}{2}-1}\sum_{t=1}^T\sum_{i=1}^N e_{it}^\alpha = O_p(N^{\frac{\alpha}{2}}T)$ by Assumption 3(1). It follows that $\sup_t \|x_t\| \leq \sup_t \left\|\Lambda_{z_t}^0 f_t^0\right\| + \sup_t \|e_t\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{\alpha}})$. Thus

$$\begin{aligned}
A_1 &= \frac{1}{2}\log\left|\frac{1}{\hat{\sigma}^2}\hat{\Lambda}_j'\hat{\Lambda}_j + I_{r_j^0}\right| = \frac{1}{2}\sum_{l=1}^{r_j^0}\log\frac{\hat{\Lambda}_{jl}'\hat{\Lambda}_{jl} + \hat{\sigma}^2}{\hat{\sigma}^2} \\
&\leq \frac{1}{2}r_j^0\log\frac{\sup_t\|x_t\|^2}{\hat{\sigma}^2} = O_p(\log NT^{\frac{2}{\alpha}}) = o_p(N) \text{ when } \frac{\log T}{N} \to 0.
\end{aligned}$$

Term A_2: By Lemma 2, $\frac{1}{\hat{\sigma}^2 + \hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}} = O_p(\frac{1}{\sqrt{\delta_{NT}}})$ for each j and each l. We have shown for term A_1 that $\sup_t \|x_t\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{\alpha}})$. Thus

$$\begin{aligned}
A_2 &\leq \frac{1}{2}\sup_t(x_t' P_{\hat{\Lambda}_j} x_t \sup_l \frac{1}{\hat{\sigma}^2 + \hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}}) \leq \frac{1}{2}\sup_t \|x_t\|^2 \sup_l \frac{1}{\hat{\sigma}^2 + \hat{\Lambda}_{jl}'\hat{\Lambda}_{jl}} \\
&= O_p(NT^{\frac{2}{\alpha}})O_p(\frac{1}{\sqrt{\delta_{NT}}}) = o_p(N) \text{ when } T^{\frac{8}{\alpha}}/N \to 0 \text{ and } \alpha > 8.
\end{aligned}$$

Term A_3:

$$\begin{aligned}
\left\|P_{\Lambda_k^0} - P_{\hat{\Lambda}_k}\right\|^2 &\leq \left\|P_{\Lambda_k^0} - P_{\hat{\Lambda}_k}\right\|_F^2 = tr[(P_{\Lambda_k^0} - P_{\hat{\Lambda}_k})^2] \\
&= 2tr(I_{r_k^0} - P_{\Lambda_k^0}P_{\hat{\Lambda}_k}) = 2\left\|M_{\hat{\Lambda}_k}\Lambda_k^0(\Lambda_k^{0\prime}\Lambda_k^0)^{-\frac{1}{2}}\right\|_F^2 \\
&\leq 2\frac{1}{N}\left\|M_{\hat{\Lambda}_k}\Lambda_k^0\right\|_F^2 \left\|(\frac{1}{N}\Lambda_k^{0\prime}\Lambda_k^0)^{-\frac{1}{2}}\right\|_F^2 = O_p(\frac{1}{\sqrt{\delta_{NT}}}), \qquad (48)
\end{aligned}$$

where the last equality follows from Theorem 1 and Assumption 2(1). We have shown for term A_1 that $\sup_t \|x_t\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{\alpha}})$. Thus

$$\begin{aligned}
A_3 &= \frac{1}{2}\hat{\sigma}^{-2}\sup_t\left|x_t'(P_{\Lambda_k^0} - P_{\hat{\Lambda}_k})x_t\right| \leq \frac{1}{2}\hat{\sigma}^{-2}\left\|P_{\Lambda_k^0} - P_{\hat{\Lambda}_k}\right\|\sup_t\|x_t\|^2 \\
&= O_p(\delta_{NT}^{-\frac{1}{4}})NT^{\frac{2}{\alpha}} = o_p(N) \text{ when } T^{\frac{16}{\alpha}}/N \to 0 \text{ and } \alpha > 16.
\end{aligned}$$

Similar to term A_3, Term A_4 is also $o_p(N)$ when $T^{\frac{16}{\alpha}}/N \to 0$ and $\alpha > 16$.

Term A_5: By Assumption 5(1), $\sup_t \left\|\frac{\Lambda_k^{0\prime}e_t}{\sqrt{N}}\right\|^\beta \leq \sum_{t=1}^T \left\|\frac{\Lambda_k^{0\prime}e_t}{\sqrt{N}}\right\|^\beta = O_p(T)$. Thus

$$A_5 \leq \frac{1}{2}\hat{\sigma}^{-2}\left\|(\frac{1}{N}\Lambda_k^{0\prime}\Lambda_k^0)^{-1}\right\|\sup_t\left\|\frac{\Lambda_k^{0\prime}e_t}{\sqrt{N}}\right\|^2 = O_p(T^{\frac{2}{\beta}}) = o_p(N) \text{ when } T^{\frac{2}{\beta}}/N \to 0.$$

Term A_6: By Assumption 1(2), $\sup_t \|f_t^0\|^\alpha \leq \sum_{t=1}^T \|f_t^0\|^\alpha = O_p(T)$, thus $\sup_t \|f_t^0\| = O_p(T^{\frac{1}{\alpha}})$. We have shown for term A_5 that $\sup_t \left\|\frac{\Lambda_j^{0\prime}e_t}{\sqrt{N}}\right\| = O_p(T^{\frac{1}{\beta}})$. Thus

$$\begin{aligned}
A_6 &\leq \hat{\sigma}^{-2}\sup_t\left|e_t'\Lambda_j^0 f_t^0\right| + \hat{\sigma}^{-2}\sup_t\left|e_t'\Lambda_k^0(\frac{\Lambda_k^{0\prime}\Lambda_k^0}{N})^{-1}\frac{\Lambda_k^{0\prime}\Lambda_j^0}{N}f_t^0\right| \\
&\leq \hat{\sigma}^{-2}\sup_t\left\|e_t'\Lambda_j^0\right\|\sup_t\left\|f_t^0\right\|(1 + \left\|(\frac{\Lambda_k^{0\prime}\Lambda_k^0}{N})^{-1}\right\|\left\|\frac{\Lambda_k^{0\prime}\Lambda_j^0}{N}\right\|) \\
&= O_p(N^{\frac{1}{2}}T^{\frac{1}{\alpha}+\frac{1}{\beta}}) = o_p(N) \text{ when } T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0.
\end{aligned}$$

(2) Similar to expression (47), it suffices to show

$$\Pr(\frac{1}{2}\hat{\sigma}^{-2}f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0 \leq A_1' + A_2' + A_3' + A_4' + A_5' + A_6' + \eta\log N) \rightarrow 0, \tag{49}$$

where $A_1', ..., A_6'$ equals $A_1, ..., A_6$ without taking supremum with respect to t. Given the calculation of terms $A_1, ..., A_6$, it is not difficult to see that without taking supremum, $A_1', ..., A_6'$ becomes $O_p(\log N)$, $O_p(\frac{N}{\sqrt{\delta_{NT}}})$, $O_p(\frac{N}{\delta_{NT}^{\frac{1}{4}}})$, $O_p(\frac{N}{\delta_{NT}^{\frac{1}{4}}})$, $O_p(1)$ and $O_p(N^{\frac{1}{2}})$ respectively. Since $f_t^{0\prime}\Lambda_j^{0\prime}M_{\Lambda_k^0}\Lambda_j^0 f_t^0 \geq NC$ for some $C > 0$, $A_1', ..., A_6'$ are all dominated by this term. ■

C Details for Theorem 3 and Theorem 4

Proof of Proposition 1

Proof. (1) Let V_{jNT} be an $r_j^0 \times r_j^0$ diagonal matrix consisting of eigenvalues of $\frac{(\Lambda_j^{0\prime}\Lambda_j^0)^{\frac{1}{2}}(\sum_{t=1}^T f_t^0 f_t^{0\prime}1_{z_t=j})(\Lambda_j^{0\prime}\Lambda_j^0)^{\frac{1}{2}}}{NTq_j^0}$ in descending order and Υ_{jNT} be the corresponding eigenvectors. Let $\bar{\Lambda}_j^0 = \Lambda_j^0(\Lambda_j^{0\prime}\Lambda_j^0)^{-\frac{1}{2}}\Upsilon_{jNT}$ be normalized version of Λ_j^0, then $\bar{\Lambda}_j^{0\prime}\bar{\Lambda}_j^0 = I_{r_j^0}$. Let $\check{\Lambda}_j = \hat{\Lambda}_j(\hat{\Lambda}_j'\hat{\Lambda}_j)^{-\frac{1}{2}}$ be normalized version of $\hat{\Lambda}_j$, then $\check{\Lambda}_j'\check{\Lambda}_j = I_{r_j^0}$.

From equation (14), we have $\check{\Lambda}_j W_{jNT} = (\frac{1}{NT}\sum_{t=1}^T \hat{p}_{tj}x_t x_t')\check{\Lambda}_j$. The left hand side equals $P_{\bar{\Lambda}_j^0}\check{\Lambda}_j W_{jNT} + M_{\bar{\Lambda}_j^0}\check{\Lambda}_j W_{jNT} = \bar{\Lambda}_j^0\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} + M_{\bar{\Lambda}_j^0}\check{\Lambda}_j W_{jNT}$. The right hand side equals

$$\begin{aligned}
& \Lambda_j^0\frac{(\sum_{t=1}^T f_t^0 f_t^{0\prime}1_{z_t=j})\Lambda_j^{0\prime}\check{\Lambda}_j}{NT} + \frac{\sum_{t=1}^T \mathbb{E}(e_t e_t')1_{z_t=j}\check{\Lambda}_j}{NT} + \frac{\sum_{t=1}^T(e_t e_t' - \mathbb{E}(e_t e_t'))1_{z_t=j}\check{\Lambda}_j}{NT} \\
& + \frac{\sum_{t=1}^T e_t f_t^{0\prime}1_{z_t=j}\Lambda_j^{0\prime}\check{\Lambda}_j}{NT} + \frac{\Lambda_j^0\sum_{t=1}^T f_t^0 e_t' 1_{z_t=j}\check{\Lambda}_j}{NT} + \frac{\sum_{t=1}^T(\hat{p}_{tj} - 1_{z_t=j})x_t x_t'}{NT}\check{\Lambda}_j \\
\equiv\ & \Lambda_j^0\frac{(\sum_{t=1}^T f_t^0 f_t^{0\prime}1_{z_t=j})\Lambda_j^{0\prime}\check{\Lambda}_j}{NT} + I + II + III + IV + D.
\end{aligned} \tag{50}$$

Note that $\Lambda_j^0\frac{(\sum_{t=1}^T f_t^0 f_t^{0\prime}1_{z_t=j})\Lambda_j^{0\prime}\check{\Lambda}_j}{NT} = \bar{\Lambda}_j^0 q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$, thus we have

$$\bar{\Lambda}_j^0(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} - q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j) + M_{\bar{\Lambda}_j^0}\check{\Lambda}_j W_{jNT} = I + II + III + IV + D \tag{51}$$

Terms $I, ..., IV$ correspond to the right hand of equation (A.1) in Bai (2003). By

Assumption 3(2), $\|I\|_F^2 = O_p(\frac{1}{N})$. By Assumption 3(4), $\|II\|_F^2 = O_p(\frac{1}{T})$. By Assumptions 5(2) and 2(1), $\|III\|_F^2$ and $\|IV\|_F^2$ are $O_p(\frac{1}{T})$. The detailed calculation is similar to the proof of Theorem 1 in Bai and Ng (2002), hence omitted here. Now consider term D. Since $\left\|\frac{\sum_{t=1}^T(\hat{p}_{tj}-1_{z_t=j})x_t x_t'}{NT}\right\| \leq \frac{\sum_{t=1}^T\left|\hat{p}_{tj}-1_{z_t=j}\right|\|x_t\|^2}{NT} \leq \sup_t\left|\hat{p}_{tj}-1_{z_t=j}\right|\frac{\sum_{t=1}^T\|x_t\|^2}{NT}$, we have

$$\|D\|_F \leq \left\|\frac{\sum_{t=1}^T(\hat{p}_{tj}-1_{z_t=j})x_t x_t'}{NT}\right\|\|\check{\Lambda}_j\|_F \leq \sqrt{r_j^0}\sup_t\left|\hat{p}_{tj}-1_{z_t=j}\right|\frac{\sum_{t=1}^T\|x_t\|^2}{NT} = o_p(\frac{1}{N^\eta}). \tag{52}$$

The last equality follows from Theorem 2 and $\frac{\sum_{t=1}^T\|x_t\|^2}{NT} = O_p(1)$, which can be easily shown using Assumptions 1(2), 2(1) and 3(1). In summary, the right hand side of equation (51) is $O_p(\frac{1}{\delta_{NT}})$. The two terms on the left hand side[12] are orthogonal to each other, thus both $\left\|M_{\bar{\Lambda}_j^0}\check{\Lambda}_j W_{jNT}\right\|_F$ and $\left\|\bar{\Lambda}_j^0(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} - q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)\right\|_F$ are $O_p(\frac{1}{\delta_{NT}})$. Since $\|A\|_F^2 = tr(A'A)$ for any matrix A and $\bar{\Lambda}_j^0$ is orthonormal,

$$\left\|\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} - q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j\right\|_F = \left\|\bar{\Lambda}_j^0(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} - q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)\right\|_F = o_p(1). \tag{53}$$

We next show that equation (53) implies that $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j \xrightarrow{p} I_{r_j^0}$ and $W_{jNT} \xrightarrow{p} q_j^0 V_j$.

First, the Euclidean norm of each column of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$ converges in probability to 1 and the inner product of different columns converges in probability to 0, because

$$\begin{aligned}
\left\|I_{r_j^0} - \check{\Lambda}_j'\bar{\Lambda}_j^0\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j\right\|_F &\leq \sqrt{r_j^0}\left\|I_{r_j^0} - \check{\Lambda}_j'\bar{\Lambda}_j^0\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j\right\| \leq \sqrt{r_j^0}tr(I_{r_j^0} - \check{\Lambda}_j'\bar{\Lambda}_j^0\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j) \\
&= \sqrt{r_j^0}\left\|M_{\bar{\Lambda}_j^0}\check{\Lambda}_j\right\|_F^2 = \sqrt{r_j^0}\left\|M_{\check{\Lambda}_j}\bar{\Lambda}_j^0\right\|_F^2 = o_p(1).
\end{aligned} \tag{54}$$

The second inequality follows from the fact that $I_{r_j^0} - \check{\Lambda}_j'\bar{\Lambda}_j^0\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$ is positive semi-definite. The second to last equality follows from the fact that both $\bar{\Lambda}_j^0$ and $\check{\Lambda}_j$ are orthonormal. The last equality follows from Theorem 1.

Let $V_{jNT,i}$, $W_{jNT,1}$ and $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}$ denote the i-th diagonal element of V_{jNT}, the 1st diagonal element of W_{jNT} and the $(i,1)$-th element of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$, then the first column of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} - q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$ is $(W_{jNT,1} - q_j^0 V_{jNT,i})(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}$, $i = 1, ..., r_j^0$. Equation

[12] The left hand side of equation (51) corresponds to a further decomposition of the left hand side of equation (A.1) in Bai (2003).

(53) implies that for all $i = 1, ..., r_j^0$, $(W_{jNT,1} - q_j^0 V_{jNT,i})(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}$ is $o_p(1)$. We have shown through expression (54) that $\sum_{i=1}^{r_j^0} (\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}{}^2 \xrightarrow{p} 1$, thus there exists at least one certain i such that $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}$ is bounded away from zero in probability. Without loss of generality, suppose $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{11}$ is bounded away from zero in probability. Since $(W_{jNT,1} - q_j^0 V_{jNT,1})(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{11}$ is $o_p(1)$, we must have $W_{jNT,1} - q_j^0 V_{jNT,1} = o_p(1)$. This implies that $W_{jNT,1} - q_j^0 V_{jNT,i}$ is bounded away from zero in probability for $i \neq 1$ because by Assumption 6, $V_{jNT,i} \neq V_{jNT,1}$ w.p.a.1. Since $(W_{jNT,1} - q_j^0 V_{jNT,i})(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}$ is $o_p(1)$ for all i, we must have $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1} = o_p(1)$ for $i \neq 1$. This together with $\sum_{i=1}^{r_j^0} (\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i1}{}^2 \xrightarrow{p} 1$ implies that $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{11} \xrightarrow{p} 1$. In summary, we have shown that the first column of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$ converges in probability to $(1, 0, ..., 0)$.

Similarly, for the second column of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j W_{jNT} - q_j^0 V_{jNT}\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$, we can also show that one element converges in probability to 1 and the other elements are $o_p(1)$. Since the inner product of the first column and the second column of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$ is $o_p(1)$, $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{12}$ must be $o_p(1)$. Thus $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i2} \xrightarrow{p} 1$ for certain $i \neq 1$ and $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i2} = o_p(1)$ for all the other i. Without loss of generality, suppose $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{22} \xrightarrow{p} 1$ and $(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i2} = o_p(1)$ for $i \neq 2$. Since $(W_{jNT,2} - q_j^0 V_{jNT,i})(\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j)_{i2}$ is $o_p(1)$ for all i, we must have $W_{jNT,2} - q_j^0 V_{jNT,2} = o_p(1)$.

Similarly, the third column of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$ converges in probability to $(0, 0, 1, ..., 0)$ and $W_{jNT,3} - q_j^0 V_{jNT,3} = o_p(1)$. Repeat the argument for all columns of $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j$, we have $\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j \xrightarrow{p} I_{r_j^0}$ and $W_{jNT} - q_j^0 V_{jNT} = o_p(1)$. Since $V_{jNT} \xrightarrow{p} V_j$, we have $W_{jNT} \xrightarrow{p} q_j^0 V_j$.

(2) By Theorem 2(1), $\left|\frac{1}{T}\sum_{t=1}^T (\hat{p}_{tj} - 1_{z_t=j})\right| \leq \sup_t |\hat{p}_{tj} - 1_{z_t=j}| = o_p(\frac{1}{N^\eta})$. By Assumption 4, $\frac{1}{T}\sum_{t=1}^T 1_{z_t=j} \xrightarrow{p} q_j^0$. Thus $\frac{1}{T}\sum_{t=1}^T \hat{p}_{tj} \xrightarrow{p} q_j^0$. We have shown that $W_{jNT} \xrightarrow{p} q_j^0 V_j$, thus $\frac{\hat{\Lambda}_j'\hat{\Lambda}_j}{N} = W_{jNT}/\frac{1}{T}\sum_{t=1}^T \hat{p}_{tj} - \frac{\hat{\sigma}^2}{N} I_{r_j^0} \xrightarrow{p} V_j$. It follows that

$$
\begin{aligned}
H_j &= \frac{\sum_{t=1}^T f_t^0 f_t^{0\prime} 1_{z_t=j}}{T} \frac{\Lambda_j^{0\prime}\hat{\Lambda}_j}{N} W_{jNT}^{-1} \\
&= (\Lambda_j^{0\prime}\Lambda_j^0)^{-\frac{1}{2}} \frac{(\Lambda_j^{0\prime}\Lambda_j^0)^{\frac{1}{2}} (\sum_{t=1}^T f_t^0 f_t^{0\prime} 1_{z_t=j}) (\Lambda_j^{0\prime}\Lambda_j^0)^{\frac{1}{2}}}{NT} (\Lambda_j^{0\prime}\Lambda_j^0)^{-\frac{1}{2}} \Lambda_j^{0\prime}\check{\Lambda}_j (\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} W_{jNT}^{-1} \\
&= (\frac{\Lambda_j^{0\prime}\Lambda_j^0}{N})^{-\frac{1}{2}} \Upsilon_{jNT} V_{jNT} (\bar{\Lambda}_j^{0\prime}\check{\Lambda}_j) (\frac{\hat{\Lambda}_j'\hat{\Lambda}_j}{N})^{\frac{1}{2}} W_{jNT}^{-1} q_j^0 \xrightarrow{p} \Sigma_{\Lambda_j}^{-\frac{1}{2}} \Upsilon_j V_j^{\frac{1}{2}}. \qquad (55)
\end{aligned}
$$

■

Proof of Theorem 3

Proof. From equation (50), we have $\hat{\Lambda}_j W_{jNT} = \Lambda_j^0 \frac{(\sum_{t=1}^T f_t^0 f_t^{0\prime} 1_{z_t=j})\Lambda_j^{0\prime}\hat{\Lambda}_j}{NT} + (I + II + III + IV + D)(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}$, i.e.,

$$\hat{\Lambda}_j - \Lambda_j^0 H_j = (I + II + III + IV + D)(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} W_{jNT}^{-1}. \tag{56}$$

We have shown in Proposition 1 that $\|I + II + III + IV + D\|_F^2 = O_p(\frac{1}{\delta_{NT}^2})$, $W_{jNT} \xrightarrow{p} q_j^0 V_j$ and $\frac{\hat{\Lambda}_j'\hat{\Lambda}_j}{N} \xrightarrow{p} V_j$. Thus $\frac{1}{N}\left\|\hat{\Lambda}_j - \Lambda_j^0 H_j\right\|_F^2 = O_p(\frac{1}{\delta_{NT}^2})$. ■

Proof of Theorem 4

Proof. Let I_i, II_i, III_i, IV_i and D_i denote the i-th row of I, II, III, IV and D respectively. From equation (56), we have

$$\hat{\lambda}_{ji}' - \lambda_{ji}^{0\prime} H_j = (I_i + II_i + III_i + IV_i + D_i)(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} W_{jNT}^{-1}.$$

By Assumptions 2(1) and 3(2) and Theorem 3, $I_i(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} = O_p(\frac{1}{\sqrt{N}\delta_{NT}})$. By Assumptions 3(4) and 7(1) and Theorem 3, $II_i(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} = O_p(\frac{1}{\sqrt{T}\delta_{NT}})$. By Assumption 3(2) and Theorem 3, $III_i(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} = \frac{\sum_{t=1}^T e_{it} f_t^{0\prime} 1_{z_t=j}\Lambda_j^{0\prime}\Lambda_j^0 H_j}{NT} + O_p(\frac{1}{\sqrt{T}\delta_{NT}})$. By Assumptions 3(2) and 7(2) and Theorem 3, $IV_i(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} = O_p(\frac{1}{\sqrt{T}\delta_{NT}})$. The detailed calculation of these four terms is similar to the proof of Lemma A.2 in Bai (2003), hence omitted here. For the term $D_i(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}$, we have

$$\begin{aligned}
\left\|D_i(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}\right\|^2 &= \left\|\frac{1}{NT}\sum_{t=1}^T (\hat{p}_{tj} - 1_{z_t=j}) x_{it} x_t' \hat{\Lambda}_j\right\|^2 \\
&\leq \frac{1}{N^2T^2}\sum_{t=1}^T (\hat{p}_{tj} - 1_{z_t=j})^2 x_{it}^2 \sum_{t=1}^T \|x_t\|^2 \left\|\hat{\Lambda}_j\right\|_F^2 \\
&\leq \sup_t |\hat{p}_{tj} - 1_{z_t=j}|^2 \frac{\sum_{t=1}^T x_{it}^2}{T} \frac{\sum_{t=1}^T \|x_t\|^2}{NT} \frac{\left\|\hat{\Lambda}_j\right\|_F^2}{N} = o_p(\frac{1}{N^{2\eta}}),
\end{aligned}$$

where the last equality follows from Theorem 2. We have shown in Proposition 1 that

$W_{jNT} \xrightarrow{p} q_j^0 V_j$, thus $W_{jNT}^{-1} = O_p(1)$. It follows that

$$\sqrt{Tq_j^0}(\hat{\lambda}_{ji} - H_j' \lambda_{ji}^0) = q_j^0 W_{jNT}^{-1} H_j' \frac{\Lambda_j^{0\prime}\Lambda_j^0}{N} \frac{\sum_{t=1}^T f_t^0 e_{it} 1_{z_t=j}}{\sqrt{Tq_j^0}} + O_p(\frac{\sqrt{T}}{N}) + o_p(1).$$

Thus by Proposition 1 and Assumption 7(3),

$$\sqrt{Tq_j^0}(\hat{\lambda}_{ji} - H_j' \lambda_{ji}^0) \xrightarrow{d} \mathcal{N}(0, V_j^{-\frac{1}{2}} \Upsilon_j' \Sigma_{\Lambda_j}^{\frac{1}{2}} \Phi_{ji} \Sigma_{\Lambda_j}^{\frac{1}{2}} \Upsilon_j V_j^{-\frac{1}{2}}) \text{ when } \sqrt{T}/N \to 0.$$

■

D Details for Theorem 5

Lemma 3 *Under Assumptions 1-7, and assume $T^{\frac{16}{\alpha}}/N \to 0$ and $T^{\frac{2}{\alpha}+\frac{2}{\beta}}/N \to 0$,*

(1) $\frac{1}{N} e_t'(\hat{\Lambda}_j - \Lambda_j^0 H_j) = O_p(\frac{1}{\delta_{NT}^2})$ for each j and t,

(2) $\frac{1}{N} \Lambda_j^{0\prime}(\hat{\Lambda}_j - \Lambda_j^0 H_j) = O_p(\frac{1}{\delta_{NT}^2})$ for each j.

Proof. Part (1): From equation (56), we have $\frac{1}{N} e_t'(\hat{\Lambda}_k - \Lambda_k^0 H_k) = \frac{1}{N} e_t'(I + II + III + IV + D)(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}} W_{jNT}^{-1}$. Consider each term one by one.

$$\frac{e_t' I (\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}}{N} = e_t' \frac{\sum_{t=1}^T \mathbb{E}(e_t e_t') 1_{z_t=j}}{N^2 T} \Lambda_j^0 H_j + e_t' \frac{\sum_{t=1}^T \mathbb{E}(e_t e_t') 1_{z_t=j}}{N^2 T} (\hat{\Lambda}_j - \Lambda_j^0 H_j).$$

By Assumption 3(1) and 3(2), the first term is $O_p(\frac{1}{N})$. By Assumption 3(2) and Theorem 3, the second term is $O_p(\frac{1}{\sqrt{N}\delta_{NT}})$.

$$\frac{e_t' II (\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}}{N} = e_t' \frac{\sum_{t=1}^T (e_t e_t' - \mathbb{E}(e_t e_t')) 1_{z_t=j}}{N^2 T} \Lambda_j^0 H_j + e_t' \frac{\sum_{t=1}^T (e_t e_t' - \mathbb{E}(e_t e_t')) 1_{z_t=j}}{N^2 T} (\hat{\Lambda}_j - \Lambda_j^0 H_j).$$

By Assumption 7(1), the first term is $O_p(\frac{1}{\sqrt{NT}})$. By Assumption 3(4) and Theorem 3, the second term is $O_p(\frac{1}{\sqrt{T}\delta_{NT}})$.

$$\frac{e_t' III (\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}}{N} = e_t' \frac{\sum_{t=1}^T e_t f_t^{0\prime} 1_{z_t=j} \Lambda_j^{0\prime}}{N^2 T} \Lambda_j^0 H_j + e_t' \frac{\sum_{t=1}^T e_t f_t^{0\prime} 1_{z_t=j} \Lambda_j^{0\prime}}{N^2 T} (\hat{\Lambda}_j - \Lambda_j^0 H_j).$$

The first term is $O_p(\frac{1}{\sqrt{NT}}) + O_p(\frac{1}{T})$ since $\frac{1}{NT} e_t' \sum_{t=1}^T e_t f_t^{0\prime} 1_{z_t=j} = O_p(\frac{1}{\sqrt{NT}}) + O_p(\frac{1}{T})$, which follows from Assumptions 3(3), 7(1) and $e_{it}e_{is} = \gamma_{i,ts} + (e_{it}e_{is} - \gamma_{i,ts})$. By Theorem 3, the second term is $O_p(\frac{1}{\sqrt{T}\delta_{NT}})$.

$$\frac{e_t' IV(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}}{N} = e_t' \frac{\Lambda_j^0 \sum_{t=1}^T f_t^0 e_t' 1_{z_t=j}}{N^2 T}\Lambda_j^0 H_j + e_t' \frac{\Lambda_j^0 \sum_{t=1}^T f_t^0 e_t' 1_{z_t=j}}{N^2 T}(\hat{\Lambda}_j - \Lambda_j^0 H_j).$$

The first term is $O_p(\frac{1}{\sqrt{NT}})$ since by Assumption 7(2), $\frac{1}{NT}\sum_{t=1}^T f_t^0 e_t' 1_{z_t=j}\Lambda_j^0 = O_p(\frac{1}{\sqrt{NT}})$. By Theorem 3, the second term is $O_p(\frac{1}{\sqrt{T}\delta_{NT}})$.

$$\left\| \frac{e_t' D(\hat{\Lambda}_j'\hat{\Lambda}_j)^{\frac{1}{2}}}{N} \right\| \leq \left\| \frac{e_t'}{\sqrt{N}} \right\| \|D\|_F \left\| \left(\frac{\hat{\Lambda}_j'\hat{\Lambda}_j}{N}\right)^{\frac{1}{2}} \right\|.$$

Thus from equation (52), this term is $o_p(\frac{1}{N^\eta})$. Finally, note that $W_{jNT}^{-1} \xrightarrow{d} \frac{1}{q_j^0} V_j^{-1}$, part (1) is proved.

Part (2) can be proved similarly. ■

Proof of Theorem 5

Proof. First, by Woodbury identity,

$$\begin{aligned}
\hat{f}_t &= \sum_{j=1}^{J^0} (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' x_t \hat{p}_{tj} \\
&= \sum_{k=1}^{J^0} \sum_{j=1}^{J^0} \hat{p}_{tj} (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' \Lambda_k^0 f_t^0 1_{z_t=k} + \sum_{j=1}^{J^0} \hat{p}_{tj} (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' e_t.
\end{aligned}$$

When $z_t = k$, we have

$$\begin{aligned}
& \sum_{j=1}^{J^0} \hat{p}_{tj} (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' \Lambda_k^0 f_t^0 \\
= \; & (\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' \Lambda_k^0 f_t^0 \\
& + (\hat{p}_{tk} - 1)(\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' \Lambda_k^0 f_t^0 + \sum_{j \neq k} \hat{p}_{tj} (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' \Lambda_k^0 f_t^0 \\
= \; & H_k^{-1} f_t^0 + (\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' (\Lambda_k^0 H_k - \hat{\Lambda}_k) H_k^{-1} f_t^0 - (\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\sigma}^2 H_k^{-1} f_t^0 \\
& + (\hat{p}_{tk} - 1)(\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' \Lambda_k^0 f_t^0 + \sum_{j \neq k} \hat{p}_{tj} (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' \Lambda_k^0 f_t^0 \\
\equiv \; & H_k^{-1} f_t^0 + B_{k1t} - B_{k2t} + B_{k3t} + B_{k4t},
\end{aligned}$$

and

$$
\begin{aligned}
& \sum\nolimits_{j=1}^{J^0} \hat{p}_{tj}(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' e_t \\
= \;& (\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}(\hat{\Lambda}_k - \Lambda_k^0 H_k)' e_t + (\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1} H_k' \Lambda_k^{0\prime} e_t \\
& + (\hat{p}_{tk} - 1)(\hat{\sigma}^2 I_{r_k^0} + \hat{\Lambda}_k'\hat{\Lambda}_k)^{-1}\hat{\Lambda}_k' e_t + \sum\nolimits_{j\neq k}^{J^0} \hat{p}_{tj}(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j' e_t \\
\equiv \;& C_{k1t} + C_{k2t} + C_{k3t} + C_{k4t}.
\end{aligned}
$$

It follows that $\hat{f}_t - H_{z_t}^{-1} f_t^0 = B_{z_t 1t} - B_{z_t 2t} + B_{z_t 3t} + B_{z_t 4t} + C_{z_t 1t} + C_{z_t 2t} + C_{z_t 3t} + C_{z_t 4t}$.

Proof of part (1): First consider $B_{z_t 1t}$.

$$
\frac{\sum_{t=1}^T \left\| B_{z_t 1t} \right\|^2}{T} \leq \sum\nolimits_{j=1}^{J^0} \left\| (\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j)^{-1}\hat{\Lambda}_j'(\Lambda_j^0 H_j - \hat{\Lambda}_j) H_j^{-1} \right\|^2 \frac{\sum_{t=1}^T \left\| f_t^0 \right\|^2}{T} = O_p(\frac{1}{\delta_{NT}^2}),
$$

where the equality is due to the following facts:

(1) By Proposition 1(1), $\frac{1}{N}(\hat{\sigma}^2 I_{r_j^0} + \hat{\Lambda}_j'\hat{\Lambda}_j) = W_{jNT} / \frac{1}{T}\sum_{t=1}^T \hat{p}_{tj} \xrightarrow{p} V_j$ for all j.

(2) $\left\| \frac{1}{\sqrt{N}} \hat{\Lambda}_j' \right\|_F = \sqrt{tr(\frac{1}{N}\hat{\Lambda}_j'\hat{\Lambda}_j)} \xrightarrow{p} \sqrt{tr(V_j)}$ for all j.

(3) By Theorem 3, $\left\| \frac{1}{\sqrt{N}} (\Lambda_j^0 H_j - \hat{\Lambda}_j) \right\| = O_p(\frac{1}{\delta_{NT}})$ for all j.

(4) By Proposition 1(1), $\left\| H_j^{-1} \right\| = O_p(1)$ for all j.

(5) $\frac{1}{T}\sum_{t=1}^T \left\| f_t^0 \right\|^2 = O_p(1)$ by Assumption 1.

It is easy to see that $\frac{1}{T}\sum_{t=1}^T \left\| B_{z_t 2t} \right\|^2 = O_p(\frac{1}{N^2})$. For $B_{z_t 3t}$, we have

$$
\frac{\sum_{t=1}^T \left\| B_{z_t 3t} \right\|^2}{T} \leq \sup_t \left\| \hat{p}_{t z_t} - 1 \right\|^2 \left\| (\hat{\sigma}^2 I_{r_{z_t}^0} + \hat{\Lambda}_{z_t}'\hat{\Lambda}_{z_t})^{-1}\hat{\Lambda}_{z_t}'\Lambda_{z_t}^0 \right\|^2 \frac{\sum_{t=1}^T \left\| f_t^0 \right\|^2}{T} = o_p(\frac{1}{N^{2\eta}}),
$$

where the equality is due to $\sup_t \left\| \hat{p}_{t z_t} - 1 \right\| \leq \sup_j \sup_t \left\| \hat{p}_{tj} - 1_{z_t = j} \right\| = o_p(\frac{1}{N^\eta})$ by Theorem 2. It is easy to see that $\frac{1}{T}\sum_{t=1}^T \left\| B_{z_t 4t} \right\|^2 = o_p(\frac{1}{N^{2\eta}})$. Similarly, we can show that $\frac{\sum_{t=1}^T \left\| C_{z_t 1t} \right\|^2}{T}$ is $O_p(\frac{1}{\delta_{NT}^2})$, $\frac{\sum_{t=1}^T \left\| C_{z_t 2t} \right\|^2}{T}$ is $O_p(\frac{1}{N})$, and both $\frac{\sum_{t=1}^T \left\| C_{z_t 3t} \right\|^2}{T}$ and $\frac{\sum_{t=1}^T \left\| C_{z_t 4t} \right\|^2}{T}$ are $o_p(\frac{1}{N^{2\eta}})(O_p(\frac{1}{N}) + O_p(\frac{1}{\delta_{NT}^2}))$. Thus $\frac{1}{T}\sum_{t=1}^T \left\| \hat{f}_t - H_{z_t}^{-1} f_t^0 \right\|^2 = O_p(\frac{1}{\delta_{NT}^2})$.

Proof of part (2): By Theorem 3 and Lemma 3(2), $\hat{\Lambda}_k'(\Lambda_k^0 H_k - \hat{\Lambda}_k) = O_p(\frac{1}{\delta_{NT}^2})$ for any k. This together with fact (1) and fact (4) listed above implies $B_{z_t 1t} = O_p(\frac{1}{\delta_{NT}^2})$. Similarly, it is easy to see that $B_{z_t 2t} = O_p(\frac{1}{N})$, $B_{z_t 3t} = o_p(\frac{1}{N^\eta})$, $B_{z_t 4t} = o_p(\frac{1}{N^\eta})$, $C_{z_t 1t} = O_p(\frac{1}{\delta_{NT}^2})$, $C_{z_t 2t} = O_p(\frac{1}{\sqrt{N}})$, $C_{z_t 3t} = o_p(\frac{1}{N^\eta})$ and $C_{z_t 4t} = o_p(\frac{1}{N^\eta})$. The leading term is $C_{z_t 2t}$. Since $\frac{\hat{\Lambda}'_{z_t}\hat{\Lambda}_{z_t}}{N} \xrightarrow{p} V_{z_t}$, $H_{z_t} \xrightarrow{p} \Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}} \Upsilon_{z_t} V_{z_t}^{\frac{1}{2}}$ and $\frac{1}{\sqrt{N}}\Lambda^{0\prime}_{z_t} e_t \xrightarrow{d} \mathcal{N}(0, \Gamma_{z_t t})$ by Assumption 7(4), we have $\sqrt{N}(\hat{f}_t - H_{z_t}^{-1} f_t^0) \xrightarrow{d} \mathcal{N}(0, V_{z_t}^{-\frac{1}{2}} \Upsilon'_{z_t} \Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}} \Gamma_{z_t t} \Sigma_{\Lambda_{z_t}}^{-\frac{1}{2}} \Upsilon_{z_t} V_{z_t}^{-\frac{1}{2}})$. ■

E Details for Results with State Dynamics

Proof of Theorem 6

Proof. Part (i): Since z_T follows Markov process,

$$l(\Lambda, \sigma^2, Q, \phi) = \log[\sum_{z_T=1}^{J^0} ... \sum_{z_1=1}^{J^0} \prod_{t=1}^{T} L(x_t \left| z_t; \Lambda, \sigma^2 \right.) \Pr(z_1 \left| \phi \right.) \prod_{t=2}^{T} \Pr(z_t \left| z_{t-1}; Q \right.)].$$

For $(\tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi)$, define m_t as equation (28) with $\hat{\Lambda}_j$ and $\hat{\sigma}^2$ replaced by $\tilde{\Lambda}$ and $\tilde{\sigma}^2$ respectively, i.e, $L(x_t \Big| z_t = j; \tilde{\Lambda}, \tilde{\sigma}^2)$ takes maximum when $j = m_t$. Since $\sum_{z_t=1}^{J^0} \Pr(z_t \left| z_{t-1}; Q \right.) = 1$ for any z_{t-1}, $\sum_{z_t=1}^{J^0} L(x_t \Big| z_t; \tilde{\Lambda}, \tilde{\sigma}^2) \Pr(z_t \left| z_{t-1}; Q \right.) \leq L(x_t \Big| z_t = m_t; \tilde{\Lambda}, \tilde{\sigma}^2)$. Thus

$$\begin{aligned}
& l(\tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi) \\
= \;& \log\{\sum_{z_{T-1}=1}^{J^0} ... \sum_{z_1=1}^{J^0} \prod_{t=1}^{T-1} L(x_t \Big| z_t; \tilde{\Lambda}, \tilde{\sigma}^2) \Pr(z_1 \left| \phi \right.) \prod_{t=2}^{T-1} \Pr(z_t \left| z_{t-1}; Q \right.) \\
& [\sum_{z_T=1}^{J^0} L(x_T \Big| z_T; \tilde{\Lambda}, \tilde{\sigma}^2) \Pr(z_T \left| z_{T-1}; Q \right.)]\} \\
\leq \;& \log\{\sum_{z_{T-1}=1}^{J^0} ... \sum_{z_1=1}^{J^0} \prod_{t=1}^{T-1} L(x_t \Big| z_t; \tilde{\Lambda}, \tilde{\sigma}^2) \Pr(z_1 \left| \phi \right.) \prod_{t=2}^{T-1} \Pr(z_t \left| z_{t-1}; Q \right.) \\
& L(x_T \Big| z_T = m_T; \tilde{\Lambda}, \tilde{\sigma}^2)\} \\
\leq \;& \log\{\sum_{z_{T-2}=1}^{J^0} ... \sum_{z_1=1}^{J^0} \prod_{t=1}^{T-2} L(x_t \Big| z_t; \tilde{\Lambda}, \tilde{\sigma}^2) \Pr(z_1 \left| \phi \right.) \prod_{t=2}^{T-2} \Pr(z_t \left| z_{t-1}; Q \right.) \\
& L(x_{T-1} \Big| z_{T-1} = m_{T-1}; \tilde{\Lambda}, \tilde{\sigma}^2) L(x_T \Big| z_T = m_T; \tilde{\Lambda}, \tilde{\sigma}^2)\} \\
\leq \;& ... \leq \sum_{t=1}^{T} \log L(x_t \Big| z_t = m_t; \tilde{\Lambda}, \tilde{\sigma}^2),
\end{aligned} \tag{57}$$

i.e, equation (29) in the proof of Theorem 1 is still valid when state dynamics are taken into account.

Now consider $l(\Lambda^0, \tilde{\sigma}^2, Q, \phi)$. Since $\Pr(z_t \left| z_{t-1}; Q \right.) \geq \min_{j,k} Q_{jk}$,

$$\sum_{z_t=1}^{J^0} L(x_t \left| z_t; \Lambda^0, \tilde{\sigma}^2 \right.) \Pr(z_t \left| z_{t-1}; Q \right.) \geq L(x_t \left| z_t; \Lambda^0, \tilde{\sigma}^2 \right.) \min_{j,k} Q_{jk}.$$

Note that on the right hand side of the inequality, z_t denotes the true state. Then similar to inequality (57),

$$l(\Lambda^0, \tilde{\sigma}^2, Q, \phi) \geq \sum_{t=1}^{T} \log L(x_t \left| z_t; \Lambda^0, \tilde{\sigma}^2 \right) + T \log \min_{j,k} Q_{jk},$$

i.e., equation (31) is still valid when state dynamics are taken into account.

The rest of the proof is the same as Theorem 1.

Part (ii):

Step (1): We first show $\left|\tilde{p}_{tj|t} - 1_{z_t=j}\right| = o_p(\frac{1}{N^\eta})$.

The proof is the same as the proof of Theorem 2, with slight modifications. First, $\hat{q}_k$ is replaced by $\tilde{p}_{tk|t-1}$, and $\hat{\Lambda}_j$ and $\hat{\sigma}^2$ are replaced by $\tilde{\Lambda}_j$ and $\tilde{\sigma}^2$. Second, the proof of Theorem 2 utilizes Theorem 1 while here the proof utilizes part (i). Third, the proof of Theorem 2 requires $\hat{q}_k$ to be bounded away from zero. Here we have $\tilde{p}_{tk|t-1} = Q_{k\cdot}\tilde{p}_{t-1|t-1} \geq \min_l Q_{kl} > 0$ for all k, where $Q_{k\cdot}$ denotes the k-th row of Q.

Step (2): We next prove $\tilde{p}_{tk|T} = o_p(\frac{1}{N^\eta})$ for $k \neq j$ when the true state is $z_t = j$. Let $Q_{\cdot k}$ denote the k-th column of Q and "$\div$" denotes element-wise division for two vectors.

$$\begin{aligned}
\tilde{p}_{tk|T} &= \tilde{p}_{tk|t} \times Q'_{\cdot k}(\tilde{p}_{t+1|T} \div \tilde{p}_{t+1|t}) = \tilde{p}_{tk|t} \times \tilde{p}'_{t+1|T}(Q_{\cdot k} \div \tilde{p}_{t+1|t}) \\
&\leq \tilde{p}_{tk|t} \max_l \frac{Q_{lk}}{Q_{lj}} \frac{1}{\tilde{p}_{tj|t}} = o_p(\frac{1}{N^\eta}),
\end{aligned}$$

where the inequality is due to the fact that each element of $\tilde{p}_{t+1|t} = Q\tilde{p}_{t|t}$ is not smaller than $Q_{\cdot j}\tilde{p}_{tj|t}$ and the last equality follows from step (1) and $\min_l Q_{lj} > 0$.

Part (iii): The proof is the same as the proof of Proposition 1, with slight modifications. First, $\hat{\Lambda}_j$, $\hat{\sigma}^2$, $\hat{p}_{tj}$, W_{jNT} and H_j is replaced by $\tilde{\Lambda}_j$, $\tilde{\sigma}^2$, $\tilde{p}_{tj|T}$, $\bar{W}_{jNT}$ and $\bar{H}_j$ respectively. Second, the proof of Proposition 1 utilizes Theorem 1 and Theorem 2, here we utilize part (i) and part (ii).

Part (iv): The proof is the same as the proof of Theorem 3, with $\hat{\Lambda}_j$, W_{jNT} and H_j replaced by $\tilde{\Lambda}_j$, $\bar{W}_{jNT}$ and $\bar{H}_j$ respectively.

Part (v): The proof is the same as the proof of Theorem 4, with $\hat{\lambda}_{ji}$, $\hat{\Lambda}_j$, W_{jNT}, H_j and $\hat{p}_{tj}$ replaced by $\tilde{\lambda}_{ji}$, $\tilde{\Lambda}_j$, $\bar{W}_{jNT}$, $\bar{H}_j$ and $\tilde{p}_{tj|T}$ respectively. The proof of Theorem

4 utilizes Theorem 2, Proposition 1 and Theorem 3, here we utilize parts (ii)-(iv).

Part (vi): The proof is the same as the proof of Theorem 5, with $\hat{f}_t$, $\hat{\sigma}^2$, $\hat{\Lambda}_j$, $\hat{p}_{tj}$ and H_{z_t} replaced by $\tilde{f}_t$, $\tilde{\sigma}^2$, $\tilde{\Lambda}_j$, $\tilde{p}_{tj|T}$ and $\bar{H}_{z_t}$. The proof of Theorem 5 utilizes Theorem 2, Proposition 1, Theorem 3 and Lemma 3. Here we utilize parts (ii)-(iv), and the results in Lemma 3 also can be proved for $\tilde{\Lambda}_j$. ■

Proof of Theorem 7

Proof. First, $\tilde{Q}_{jk} = \sum_{t=2}^{T} \tilde{p}_{tjk|T} / \sum_{j=1}^{J^0} \sum_{t=2}^{T} \tilde{p}_{tjk|T} = \frac{1}{T-1} \sum_{t=2}^{T} \tilde{p}_{tjk|T} / \frac{1}{T-1} \sum_{t=1}^{T-1} \tilde{p}_{tk|T}$. For the denominator, by Theorem 6(ii), we have

$$\frac{1}{T-1} \sum\nolimits_{t=1}^{T-1} \tilde{p}_{tk|T} = \frac{1}{T-1} \sum\nolimits_{t=1}^{T-1} 1_{z_t=k} + o_p(\frac{1}{N^\eta}) \xrightarrow{p} q_k^0. \tag{58}$$

For the numerator, we have

$$\begin{aligned} \frac{1}{T-1} \sum\nolimits_{t=2}^{T} \tilde{p}_{tjk|T} &= \frac{1}{T-1} \sum\nolimits_{t=2}^{T} \tilde{p}_{tj|T} \Pr(z_{t-1} = k \,\Big|\, z_t = j, x_{1:T}; \tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi) \\ &= \frac{1}{T-1} \sum\nolimits_{t=2}^{T} [1_{z_t=j} + o_p(\frac{1}{N^\eta})][1_{z_{t-1}=k} + o_p(\frac{1}{N^\eta})]. \end{aligned} \tag{59}$$

The second equality of (59) follows from: (1) $\tilde{p}_{tj|T} = 1_{z_t=j} + o_p(\frac{1}{N^\eta})$ by Theorem 6(ii),

$$\begin{aligned} (2)\ \Pr(z_{t-1} &= k \,\Big|\, z_t = j, x_{1:T}; \tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi) = \Pr(z_{t-1} = k \,\Big|\, z_t = j, x_{1:t-1}; \tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi) \\ &= \frac{\Pr(z_{t-1} = k, z_t = j \,\Big|\, x_{1:t-1}; \tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi)}{\Pr(z_t = j \,\Big|\, x_{1:t-1}; \tilde{\Lambda}, \tilde{\sigma}^2, Q, \phi)} = \frac{Q_{jk} \tilde{p}_{t-1,k|t-1}}{\sum_{h=1}^{J^0} Q_{jh} \tilde{p}_{t-1,h|t-1}} \\ &= 1_{z_{t-1}=k} + o_p(\frac{1}{N^\eta}), \end{aligned}$$

where the last equality follows from Theorem 6(ii). Since z_t follows a Markov process,

$$\frac{1}{T-1} \sum\nolimits_{t=2}^{T} 1_{z_t=j} 1_{z_{t-1}=k} \xrightarrow{p} \mathbb{E}(1_{z_t=j} 1_{z_{t-1}=k}) = \mathbb{E}[\mathbb{E}(1_{z_t=j} 1_{z_{t-1}=k} \,\big|\, 1_{z_{t-1}=k})] = q_k^0 Q_{jk}^0. \tag{60}$$

Take equations (58)-(60) together, we have shown $\tilde{Q}_{jk} \xrightarrow{p} Q_{jk}^0$. ■

F Details on First Order Conditions

First order condition of σ^2 with no state dynamics:

$$
\begin{aligned}
\frac{\partial l(\Lambda,\sigma^2,q)}{\partial\sigma^2} &= \sum\nolimits_{t=1}^{T}\sum\nolimits_{j=1}^{J^0} p_{tj}\frac{\partial(-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}x_t'\Sigma_j^{-1}x_t)}{\partial\sigma^2} \\
&= -\frac{1}{2}\sum\nolimits_{t=1}^{T}\sum\nolimits_{j=1}^{J^0} p_{tj}tr(\Sigma_j^{-1} - \Sigma_j^{-1}x_t x_t'\Sigma_j^{-1}) \\
&= -\frac{1}{2}\sum\nolimits_{j=1}^{J^0}\sum\nolimits_{t=1}^{T} p_{tj}tr(\Sigma_j^{-1} - \Sigma_j^{-1}S_j\Sigma_j^{-1}) \\
&= -\frac{1}{2\sigma^4}\sum\nolimits_{j=1}^{J^0}\sum\nolimits_{t=1}^{T} p_{tj}tr(\Sigma_j - S_j) \\
&= -\frac{1}{2\sigma^4}tr(\sum\nolimits_{j=1}^{J^0}\sum\nolimits_{t=1}^{T} p_{tj}\Lambda_j\Lambda_j' + T\sigma^2 I_N - \sum\nolimits_{t=1}^{T} x_t x_t'),
\end{aligned}
$$

where the second equality is due to

$$
\begin{aligned}
\frac{\partial\log|\Sigma_j|}{\partial\sigma^2} &= tr(\Sigma_j^{-1}), && (61) \\
\frac{\partial x_t'\Sigma_j^{-1}x_t}{\partial\sigma^2} &= -tr(\Sigma_j^{-1}x_t x_t'\Sigma_j^{-1}), && (62)
\end{aligned}
$$

and the second last equality is due to

$$
\Sigma_j(\Sigma_j - S_j)\Sigma_j = (\Lambda_j\Lambda_j' + \sigma^2 I_N)(\Sigma_j - S_j)(\Lambda_j\Lambda_j' + \sigma^2 I_N) = \sigma^4(\Sigma_j - S_j), \tag{63}
$$

since $(\Sigma_j - S_j)\Lambda_j\Lambda_j' = (\Lambda_j\Lambda_j' + \sigma^2 I_N - S_j)\Lambda_j\Lambda_j' = 0$ by equation (14). Set $\frac{\partial l(\Lambda,\sigma^2,q)}{\partial\sigma^2}$ to zero, we have $\sigma^2 = \frac{1}{N}tr(\frac{1}{T}\sum_{t=1}^{T} x_t x_t' - \sum_{j=1}^{J^0}\frac{1}{T}\sum_{t=1}^{T} p_{tj}\Lambda_j\Lambda_j')$.

First order condition of q_j with no state dynamics:

The Lagrangean is $l(\Lambda,\sigma^2,q) + w(1 - q_1 - q_1 - ... - q_{J^0})$. The derivative of the Lagrangean with respect to q_j is $\sum_{t=1}^{T}\frac{p_{tj}}{q_j} - w$. Set it to zero, we have $\sum_{t=1}^{T} p_{tj} = wq_j$. Take sum with respect to j, we have $w = T$. Thus $q_j = \frac{1}{T}\sum_{t=1}^{T} p_{tj}$.

First order condition of σ^2 with state dynamics:

From equations (61) and (62), $\frac{\partial \log L(x_t \left| z_t=j;\Lambda_j,\sigma^2 \right.)}{\partial\sigma^2} = -\frac{1}{2}tr(\Sigma_j^{-1} - \Sigma_j^{-1}x_t x_t' \Sigma_j^{-1})$. Thus

$$\begin{aligned}
& \frac{\partial \sum_{t=1}^{T}\sum_{j=1}^{J^0} \log L(x_t \left| z_t = j; \Lambda_j, \sigma^2 \right.)\tilde{p}_{tj|T}^{(h)}}{\partial \sigma^2} \\
= & -\frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{J^0} \tilde{p}_{tj|T}^{(h)} tr(\Sigma_j^{-1} - \Sigma_j^{-1}x_t x_t' \Sigma_j^{-1}) \\
= & -\frac{1}{2}\sum_{j=1}^{J^0} tr(\sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)} \Sigma_j^{-1} - \Sigma_j^{-1} \sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)} x_t x_t' \Sigma_j^{-1}) \\
= & -\frac{1}{2}\sum_{j=1}^{J^0} (\sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)}) tr(\Sigma_j^{-1} - \Sigma_j^{-1} \tilde{S}_j^{(h)} \Sigma_j^{-1}) \\
= & -\frac{1}{2\sigma^4}\sum_{j=1}^{J^0} (\sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)}) tr(\Sigma_j - \tilde{S}_j^{(h)}) \\
= & -\frac{1}{2\sigma^4} tr(\sum_{j=1}^{J^0}\sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)} \Lambda_j \Lambda_j' + T\sigma^2 I_N - \sum_{t=1}^{T} x_t x_t'),
\end{aligned}$$

where the second last equality is explained in equation (63). Set $\frac{\partial \log L(x_t \left| z_t=j;\Lambda_j,\sigma^2 \right.)}{\partial\sigma^2}$ to zero, we have $\tilde{\sigma}^{2(h+1)} = \frac{1}{N} tr(\frac{1}{T}\sum_{t=1}^{T} x_t x_t' - \sum_{j=1}^{J^0} \frac{1}{T} \sum_{t=1}^{T} \tilde{p}_{tj|T}^{(h)} \tilde{\Lambda}_j^{(h+1)} \tilde{\Lambda}_j^{(h+1)\prime})$.

First order condition of Q with state dynamics:

First, when (Q, ϕ) also enters the iteration, let $(\tilde{Q}^{(h)}, \tilde{\phi}^{(h)})$ denote the value of (Q, ϕ) for the h-th iteration, and let $\tilde{\theta}^{(h)} = (\tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)}, \tilde{Q}^{(h)}, \tilde{\phi}^{(h)})$.

Since $\sum_{j=1}^{J^0} Q_{jk} = 1$, the Lagrangean is $\sum_{t=2}^{T}\sum_{j=1}^{J^0}\sum_{k=1}^{J^0} \log Q_{jk} \tilde{p}_{tjk|T}^{(h)} + \sum_{k=1}^{J^0} w_k (1 - Q_{1k} - Q_{2k} - ... - Q_{J^0 k})$. The first order derivative of the Lagrangean with respect to Q_{jk} is $\frac{1}{Q_{jk}} \sum_{t=2}^{T} \tilde{p}_{tjk|T}^{(h)} - w_k$. Set it to zero, we have $\sum_{t=2}^{T} \tilde{p}_{tjk|T}^{(h)} = Q_{jk} w_k$. Take sum over j, we have $\sum_{j=1}^{J^0}\sum_{t=2}^{T} \tilde{p}_{tjk|T}^{(h)} = \sum_{j=1}^{J^0} Q_{jk} w_k = w_k$. Thus $\tilde{Q}_{jk}^{(h+1)} = \sum_{t=2}^{T} \tilde{p}_{tjk|T}^{(h)} / \sum_{j=1}^{J^0}\sum_{t=2}^{T} \tilde{p}_{tjk|T}^{(h)}$.

First order condition of ϕ with state dynamics:

Since $\sum_{k=1}^{J^0} \phi_k = 1$, the Lagrangean is $\sum_{k=1}^{J^0} \log \phi_k \tilde{p}_{1k|T}^{(h)} + w(1 - \phi_1 - \phi_2 - ... - \phi_{J^0})$. The first order derivative of the Lagrangean with respect to ϕ_k is $\frac{1}{\phi_k}\tilde{p}_{1k|T}^{(h)} - w$. Set it to zero, we have $\tilde{p}_{1k|T}^{(h)} = \phi_k w$. Take sum over k, we have $1 = \sum_{k=1}^{J^0} p_{1k|T}^{(h)} = \sum_{k=1}^{J^0} \phi_k w = w$, thus $\tilde{\phi}_k^{(h+1)} = \tilde{p}_{1k|T}^{(h)} = \sum_{j=1}^{J^0} \tilde{p}_{2jk|T}^{(h)}$.

G Smoother Algorithm for $\tilde{p}_{tjk|T}^{(h)}$

Step (1): Calculate conditional likelihoods $L(x_t \left| x_{1:t-1}; \tilde{\theta}^{(h)} \right.)$ and filtered estimates $\tilde{p}_{tjk|t}^{(h)}$ for $t = 2, ..., T$.

$$\begin{aligned}
\tilde{p}_{tjk|t}^{(h)} &= \Pr(z_t = j, z_{t-1} = k \left| x_{1:t}; \tilde{\theta}^{(h)} \right.) = L(x_t \left| z_t = j; \tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)} \right.)) \\
\times \Pr(z_t &= j \left| z_{t-1} = k; Q \right.) \Pr(z_{t-1} = k \left| x_{1:t-1}; \tilde{\theta}^{(h)} \right.) / L(x_t \left| x_{1:t-1}; \tilde{\theta}^{(h)} \right.),
\end{aligned}$$

where $\Pr(z_1 = k \left| x_1; \tilde{\theta}^{(h)} \right.) = \frac{L(x_1 \left| z_1 = k; \tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)} \right.)\phi_k}{\sum_{j=1}^{J^0} L(x_1 \left| z_1 = j; \tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)} \right.)\phi_j}$ and $\Pr(z_{t-1} = k \left| x_{1:t-1}; \tilde{\theta}^{(h)} \right.) = \sum_{z_{t-2}=1}^{J^0} \Pr(z_{t-1} = k, z_{t-2} \left| x_{1:t-1}; \tilde{\theta}^{(h)} \right.)$. The denominator $L(x_t \left| x_{1:t-1}; \tilde{\theta}^{(h)} \right.)$ equals the sum of the numerator with respect to z_t and z_{t-1}.

Step (2): Fix down $z_t = j, z_{t-1} = k$, for all z_{t+1},

$$\begin{aligned}
\Pr(z_{t+1}, z_t &= j, z_{t-1} = k \left| x_{1:t+1}; \tilde{\theta}^{(h)} \right.) = L(x_{t+1} \left| z_{t+1}; \tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)} \right.)) \Pr(z_{t+1} \left| z_t = j; Q \right.) \\
\times \Pr(z_t &= j, z_{t-1} = k \left| x_{1:t}; \tilde{\theta}^{(h)} \right.) / L(x_{t+1} \left| x_{1:t}; \tilde{\theta}^{(h)} \right.),
\end{aligned}$$

for all z_{t+1} and z_{t+2},

$$\begin{aligned}
\Pr(z_{t+2}, z_{t+1}, z_t &= j, z_{t-1} = k \left| x_{1:t+2}; \tilde{\theta}^{(h)} \right.) = L(x_{t+2} \left| z_{t+2}; \tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)} \right.)) \Pr(z_{t+2} \left| z_{t+1}; Q \right.) \\
\times \Pr(z_{t+1}, z_t &= j, z_{t-1} = k \left| x_{1:t+1}; \tilde{\theta}^{(h)} \right.) / L(x_{t+2} \left| x_{1:t+1}; \tilde{\theta}^{(h)} \right.),
\end{aligned}$$

and for $\tau = t + 3, ..., T$, for all z_τ and $z_{\tau-1}$,

$$\begin{aligned}
\Pr(z_\tau, z_{\tau-1}, z_t &= j, z_{t-1} = k \left| x_{1:\tau}; \tilde{\theta}^{(h)} \right.) = L(x_\tau \left| z_\tau; \tilde{\Lambda}^{(h)}, \tilde{\sigma}^{2(h)} \right.)) \Pr(z_\tau \left| z_{\tau-1}; Q \right.) \\
\times \Pr(z_{\tau-1}, z_t &= j, z_{t-1} = k \left| x_{1:\tau-1}; \tilde{\theta}^{(h)} \right.) / L(x_\tau \left| x_{1:\tau-1}; \tilde{\theta}^{(h)} \right.),
\end{aligned}$$

where $\Pr(z_{\tau-1}, z_t = j, z_{t-1} = k \left| x_{1:\tau-1}; \tilde{\theta}^{(h)} \right.) = \sum_{z_{\tau-2}=1}^{J^0} \Pr(z_{\tau-1}, z_{\tau-2}, z_t = j, z_{t-1} = k \left| x_{1:\tau-1}; \tilde{\theta}^{(h)} \right.)$.

Step (3): Calculate $\tilde{p}_{tjk|T}^{(h)} = \sum_{z_T=1}^{J^0} \sum_{z_{T-1}=1}^{J^0} \Pr(z_T, z_{T-1}, z_t = j, z_{t-1} = k \left| x_{1:T}; \tilde{\theta}^{(h)} \right.)$.

Repeat steps (1)-(3) for all j and k.