Selecting the Correct Number of Factors in Approximate Factor Models: The Large Panel Case with Group Bridge Estimators*

Mehmet Caner †
North Carolina State University

Xu Han ‡
City University of Hong Kong

November 5, 2013

Abstract

This paper proposes a group bridge estimator to select the correct number of factors in approximate factor models. It contributes to the literature on shrinkage estimation and factor models by extending the conventional bridge estimator from a single equation to a large panel context. The proposed estimator can consistently estimate the factor loadings of relevant factors and shrink the loadings of irrelevant factors to zero with a probability approaching one. Hence, it provides a consistent estimate for the number of factors. We also propose an algorithm for the new estimator; Monte Carlo experiments show that our algorithm converges reasonably fast and that our estimator has very good performance in small samples. An empirical example is also presented based on a commonly used US macroeconomic data set.

Key words: bridge estimation, common factors, selection consistency

*We thank the co-editor Shakeeb Khan, an associate editor and two anonymous referees who made this paper better. We also thank Robin Sickles, James Stock, and the participants at the High Dimension Reduction Conference (December 2010) in London, the Panel Data Conference (July 2011) in Montreal, the CIREQ Econometrics Conference (May 2013), and the North American Summer Meeting of the Econometric Society (June 2013) for their comments.

†Department of Economics, 4168 Nelson Hall, Raleigh, NC 27695. email: mcaner@ncsu.edu.

‡Department of Economics and Finance, City University of Hong Kong. E-mail: xuhan25@cityu.edu.hk

1 Introduction

In recent years, factor models have become increasingly important in both finance and macroeconomics. These models use a small number of common factors to explain the co-movements of a large number of time series. In finance, factor models are the foundation for the extension of the arbitrage pricing theory (Chamberlain and Rothschild, 1983). In macroeconomics, empirical evidence shows that a few factors can explain a substantial amount of the variation in major macroeconomic variables (Sargent and Sims, 1977; Stock and Watson, 1989; Giannone, Reichlin, and Sala, 2004). Cross country differences have been explained using factor models by Gregory and Head (1999) and Forni, Hallin, Lippi, and Reichlin (2000). Other applications of factor models include forecasting (Stock and Watson, 2002), dynamic stochastic general equilibrium macroeconomic models (Boivin and Giannoni, 2006), structural VAR analysis (Bernanke, Boivin, and Eliasz, 2005; Stock and Watson, 2005; Forni and Gambetti, 2010), and consumer demand and micro-behavior (Forni and Lippi, 1997; Lewbel, 1991).

As the common factors are often unobserved, it is natural to ask how many factors should be included in practice. For example, in the factor-augmented VAR models of Bernanke, Boivin, and Eliasz (2005), an incorrect number of common factors can lead to an inconsistent estimate of the space spanned by the structural shocks, and impulse responses based on such estimates may be misleading and result in bad policy suggestions. Recent papers by Stock and Watson (2009) and Breitung and Eickmeier (2011) also find that structural breaks in factor loadings can lead to a change in the number of factors in subsamples. This implies that correctly determining the number of factors in subsamples can help us detect structural changes in factor models.

In this paper, we propose a group bridge estimator to determine the number of factors in approximate factor models. Compared to conventional information criteria, the bridge estimator has the advantage that it can conduct both estimation and model selection in one step. It avoids the instability caused by subset selection or stepwise deletion (Breiman, 1996). The instability associated with model selection shows up in the following way: when new data is added, the estimated model changes dramatically. Hence, instability in a factor model context means that the estimated number of factors fluctuates widely around the true number as new data are introduced. Also, Horowitz, Huang, and Ma (2008, HHM hereafter) show that the bridge estimator possesses the oracle property, i.e., it will provide efficient estimates as if the true model is given in advance. Moreover, as discussed in De Mol *et al.* (2008), shrinkage based estimators such as bridge have certain optimal risk properties as seen in Donoho and Johnstone (1994). Thus, we expect that the bridge estimator will preserve these good properties in the high dimensional factor models.

This paper builds a connection between shrinkage estimation and factor models. Despite the large literature on both factor models and shrinkage estimation, the interaction between these two fields is rather small. Bai and Ng (2008) apply the shrinkage estimation to select the relevant variables and refine forecasts based on factor models. However, their application of shrinkage estimation is still restricted to a single equation and does not apply to large panels. This paper

contributes to these two strands of the literature in the following ways. First, we extend the conventional bridge estimator from a single equation (such as Knight and Fu, 2000; Caner and Knight, 2013) to a large panel context. This extension is not trivial because shrinkage estimators in the literature (such as HHM) usually assume independent error terms. In contrast, our estimator allows the error terms to have correlations in both cross section and time dimensions. In addition, the regressors are not observed but estimated in factor models, so the estimation errors in factors bring another layer of difficulty to our extension. Moreover, the sparsity condition in factor models is different from the one that is common in the literature. Many papers allow the number of parameters to increase with the sample size (for example, Zou and Zhang, 2009; Caner and Zhang, 2013), but most of them rely on the sparsity condition that the number of nonzero coefficients has to diverge at a slower rate than the sample size. In factor models, however, the number of nonzero factor loadings is proportional to the cross-sectional dimension.

Additionally, our estimator provides a new way to determine the number of factors. We prove that our estimator can consistently estimate the factor loadings (up to a rotation) of relevant factors and shrink the loadings of irrelevant factors to zero with a probability approaching one. Hence, given the estimated factors, this new estimator can select the correct model specification and conduct the estimation of factor loadings in one step. We penalize the Euclidean norm of the factor loading vectors. Specifically, all cells in the zero factor loading vectors are shrunk simultaneously. Thus, our estimation has a group structure.

In recent years, much attention has been drawn to the estimation of the number of factors. Bai and Ng (2002) and Onatski (2010) propose statistics to determine the number of static factors in approximate factor models.[1] Onatski (2009) constructs tests for the number of factors based on the empirical distribution of eigenvalues of the data covariance matrix. Amengual and Watson (2007), Bai and Ng (2007), and Hallin and Liska (2007) develop estimators for the number of dynamic factors. These methods are roughly equivalent to finding some threshold to distinguish large and small eigenvalues of the data covariance matrix. Some other related methods are developed by Kapetanios (2010), Han (2012), and Seung and Horenstein (2013). We solve the problem from a different but related angle: the group bridge estimator directly focuses on the shrinkage estimation of the factor loading matrix. Simulations show that our estimator is robust to cross-sectional and serial correlations in the error terms and outperforms existing methods in certain cases.

Section 2 introduces the model, the assumptions, and presents the theoretical results of our estimator. Section 3 conducts simulations to explore the finite sample performance of our estimator. Section 4 provides an empirical example using a commonly used macroeconomic data set. Section 5 concludes. The appendix contains all proofs.

[1]The definitions of static factors, dynamic factors, and approximate factor models are discussed in the second paragraph of Section 2.

2 The Model

We use the following representation for the factor model:

$$X_{it} = \lambda_i^{0\prime} F_t^0 + e_{it}, \tag{2.1}$$

where X_{it} is the observed data for the i^{th} cross section at time t, for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$, F_t^0 is an $r \times 1$ vector of common factors, and λ_i^0 is an $r \times 1$ vector of factor loadings associated with F_t^0. The true number of factors is "r". $\lambda_i^{0\prime} F_t^0$ is the common component of X_{it} and e_{it} is the idiosyncratic error. F_t^0, λ_i^0, e_{it} and r are unobserved.

Here we explain several terms used in our discussion of factor models. We call (2.1) an *approximate factor model* because our assumptions (see Section 2.1) allow some cross-sectional correlations in the idiosyncratic errors. It is more general than a *strict factor model*, which assumes e_{it} to be uncorrelated across i. Also, we refer to (2.1) as a *static factor model* because X_{it} has a contemporaneous relationship with the factors. This is different from the *dynamic factor model* in Forni *et al.* (2000). The dynamic factor model has the representation $X_{it} = \sum_{j=1}^{q} \varphi_{ij}(L) f_{jt} + e_{it}$, where the $q-$dimensional dynamic factors are orthonormal white noises and the one-sided filters $\varphi_{ij}(L)$ are dynamic factor loadings. Hence, the factor model $X_{it} = a_i f_t + b_i f_{t-1} + e_{it}$ has one dynamic factor but two static factors. Our focus is the static factor model with cross-sectionally correlated e_{it}. For conciseness, we use the term "factor model" to refer to the "static factor model" and "approximate factor model" in the rest of the paper, unless otherwise specified.

Now, we rewrite model (2.1) in the vector form:

$$X_i = F^0 \lambda_i^0 + e_i \quad \text{for } i = 1, 2, ..., N, \tag{2.2}$$

where $X_i = (X_{i1}, X_{i2}, ..., X_{iT})'$ is a T-dimensional vector of observations on the i^{th} cross section, $F^0 = (F_1^0, F_2^0, ..., F_T^0)'$ is a $T \times r$ matrix of the unobserved factors, and $e_i = (e_{i1}, e_{i2}, ..., e_{iT})'$ is a T-dimensional vector of the idiosyncratic shock of the i^{th} cross section. (2.1) and (2.2) can also be represented in the matrix form:

$$X = F^0 \Lambda^{0\prime} + e, \tag{2.3}$$

where $X = (X_1, X_2, ..., X_N)$ is a $T \times N$ matrix of observed data, $\Lambda^0 = (\lambda_1^0, \lambda_2^0, ..., \lambda_N^0)'$ is the $N \times r$ factor loading matrix, and $e = (e_1, e_2, ..., e_N)$ is a $T \times N$ matrix of idiosyncratic shocks.

We use Principal Components Analysis (PCA) to estimate unknown factors, and thereafter we use bridge estimation to get the correct number of factors. The conventional PCA minimizes the following objective function:

$$V(k) = \min_{\Lambda^k, F^k} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \lambda_i^{k\prime} F_t^k)^2, \tag{2.4}$$

where the superscript k denotes the fact that the number of factors is to be determined, and both λ_i^k and F_t^k are $k \times 1$ vectors. We consider $0 \leq k \leq p$, where p is some predetermined upper bound such that $p \geq r$. For a given k, the solution to the above objective function (2.4) is that $\hat{F}^k(T \times k)$ is equal to $\sqrt{T}$ times the eigenvectors corresponding to the k largest eigenvalues of XX' and $\hat{\Lambda}_{OLS}^k = X'\hat{F}^k \left(\hat{F}^{k'}\hat{F}^k\right)^{-1} = X'\hat{F}^k/T$ (as eigenvectors are orthonormal, $\hat{F}^{k'}\hat{F}^k/T = I_k$).

Let $C_{NT} = \min(\sqrt{N}, \sqrt{T})$. We define our group bridge estimator as $\hat{\Lambda}$ that minimizes the following objective function:

$$\hat{\Lambda} = \text{argmin}_{\Lambda} L(\Lambda) \tag{2.5}$$

$$L(\Lambda) = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \lambda_i^{'}\hat{F}_t^p)^2 + \frac{\gamma}{C_{NT}^2}\sum_{j=1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_{ij}^2\right)^{\alpha}\right], \tag{2.6}$$

where $\hat{F}^p$ is the $T \times p$ principal component estimate of the factors, $\hat{F}_t^{\,p}$ is the transpose of the t^{th} row of $\hat{F}^p$, $\lambda_i = (\lambda_{i1}, ..., \lambda_{ip})'$ is a $p \times 1$ vector in a compact subset of $\mathbb{R}^p$, $\Lambda \equiv (\lambda_1, \lambda_2, ..., \lambda_N)'$ is $N \times p$, α is a constant with $0 < \alpha < 1/2$, and γ is a tuning parameter.

The penalty term in (2.6) is different from that of a conventional bridge estimator such as the one in HHM (2008). First, unlike the single equation shrinkage estimation, our penalty term is divided by C_{NT}^2, which indicates that the tuning parameter will depend on both N and T. Second, we use $\sum_{j=1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_{ij}^2\right)^{\alpha}$ instead of the HHM (2008) type of penalty $\sum_{i=1}^{N}\sum_{j=1}^{p}|\lambda_{ij}|^{\delta}$ $(0 < \delta < 1)$. A necessary condition for $\sum_{i=1}^{N}\sum_{j=1}^{p}|\lambda_{ij}|^{\delta}$ to shrink the last $p - r$ columns of Λ to zero is that the corresponding non-penalized estimates for these loadings have to converge in probability to zero for all i's and $j = r + 1, ..., p$. This necessary condition fails because the last $p - r$ columns of $\hat{F}^p$ are correlated with some X_i's, due to the limited cross-sectional correlation among idiosyncratic errors. Hence, using $\sum_{i=1}^{N}\sum_{j=1}^{p}|\lambda_{ij}|^{\delta}$ as the penalty will yield many zero and a few nonzero estimates in the last $p - r$ columns of Λ, and will cause overestimation if we use the number of nonzero columns in Λ as an estimator for r.

To shrink the entire j^{th} $(j > r)$ column of Λ to zero, we apply the penalty term in (2.6) to penalize the norm of the entire vector $\Lambda^j \equiv (\lambda_{1j}, ..., \lambda_{Nj})'$ rather than each single element λ_{ij}. The idea is related to the group LASSO (Yuan and Lin, 2006), which is designed for single equation models. In a factor model setup, the analog of the group LASSO's penalty term will be proportional to $\sum_{j=1}^{p}\sqrt{\sum_{i=1}^{N}\lambda_{ij}^2}$, which provides an intuition to explain why we have $0 < \alpha < 1/2$. If $N = 1$, then the model only involves a single equation, and the group LASSO's penalty reduces to $\sum_{j=1}^{p}|\lambda_j|$. Accordingly, our penalty becomes $\sum_{j=1}^{p}|\lambda_j|^{2\alpha}$ with $0 < 2\alpha < 1$, which is the familiar penalty term in a single equation bridge estimation.

Huang *et al.* (2009) also propose a group bridge estimator for single equation shrinkage estimation. In our large panel setup, the analog of Huang *et al.*'s (2009) penalty term will be proportional to $\sum_{j=1}^{p}\left(\sum_{i=1}^{N}|\lambda_{ij}|\right)^{\delta}$ $(0 < \delta < 1)$. Their penalty is designed to achieve selection consistency both within and among groups. Under this paper's framework, a group of coefficients correspond to a

column in the factor loading matrix. The context of this paper is different from that of Huang *et al.* (2009), and we do not need to consider the within group selection for three reasons. First, within group selection means distinguishing zero from nonzero elements in the first r columns of $\Lambda(N \times p)$, but this is not necessary because our purpose is to distinguish the first r nonzero columns from the last $p-r$ zero columns of Λ, i.e., we only need to select groups in order to determine r. Second, it is well known that both factors and factor loadings estimated by principal components are consistent estimates of their original counterparts up to some rotations. The rotations can be so generic that the economic meanings of the principal component estimators are completely unclear. The within group selection is based on the post-rotation factor loadings. Hence, unless there is a specific reason to detect zero and nonzero elements in the post-rotation factor loading matrix, within group selection is not very meaningful under the current setup. Moreover, as our penalty term uses square rather than absolute values, the computation is much easier than that of the penalty used by Huang *et al.* (2009).

Hirose and Konishi (2012) apply the group LASSO estimator in the context of strict factor models. However, their estimator is substantially different from (2.5), primarily because they focus on selecting variables that are not affected by any of the factors, whereas this paper focuses on selecting factors that do not affect any of the variables. In other words, they are interested in which rows of Λ are zero, whereas we care about which columns of Λ are zero, so Hirose and Konishi's (2012) penalty term cannot determine the number of factors by design. One could in principle construct a group LASSO estimator to select the zero columns of Λ. However, such an estimator is different from Hirose and Konishi (2012), and its theoretical property is an open question and beyond the scope of this paper.

2.1 Assumptions

Let $\|A\| \equiv [trace(A'A)]^{1/2}$ denote the norm of matrix A and $\to_p$ denote convergence in probability. Our theoretical results are derived based on the following assumptions.

1. $E\|F_t^0\|^4 < \infty$, $T^{-1}\sum_{t=1}^{T} F_t^0 F_t^{0'} \to_p \Sigma_F$ *($r \times r$) as* $T \to \infty$, *and* Σ_F *is finite, and positive definite.*
2. $\|\lambda_i^0\| \leq \bar{\lambda} < \infty$, $\|\Lambda^{0'}\Lambda^0/N - \Sigma_\Lambda\| \to 0$ *as* $N \to \infty$, *and* Σ_Λ *is finite, and positive definite. The* j^{th} *column of* Λ, *denoted as* Λ^j, *is inside a compact subset of* $\mathbb{R}^N$ *for all* N *and* $j = 1, ..., p$.
3. *There exists a positive and finite constant* M *that does not depend on* N *or* T, *such that for all* N *and* T,

(i). $Ee_{it} = 0$, $E|e_{it}|^8 < M$.

(ii). $E(e_s'e_t/N) = E[N^{-1}\sum_{i=1}^{N} e_{is}e_{it}] = \iota_N(s,t)$, $|\iota_N(s,s)| \leq M$ *for all* s, *and*

$$T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\iota_N(s,t)| \leq M.$$

(iii). $E(e_{it}e_{jt}) = \tau_{ij,t}$ *where* $|\tau_{ij,t}| \leq |\tau_{ij}|$ *for some* τ_{ij} *and for all* t. *In addition*

$$N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}| \leq M.$$

(iv). $E(e_{it}e_{js}) = \tau_{ij,ts}$ *and*

$$(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}|\tau_{ij,ts}| \leq M.$$

(v). *For every* (t, s),

$$E|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]|^4 \leq M.$$

4.

$$E[\frac{1}{N}\sum_{i=1}^{N}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^0 e_{it}\|^2] \leq M.$$

5. *The eigenvalues of the matrix* $(\Sigma_\Lambda\Sigma_F)$ *are distinct.*

6. (i). *As* N *and* $T \to \infty$,

$$\frac{\gamma}{C_{NT}} \to 0,$$

where $C_{NT} = \min(\sqrt{N}, \sqrt{T})$.

(ii). *Also, as* $N \to \infty$,

$$\frac{\gamma}{C_{NT}^{2\alpha}} \to \infty,$$

where $0 < \alpha < 1/2$.

Assumptions 1-4 are Assumptions A-D in Bai and Ng (2002). Assumption 1 is standard for factor models. With Assumption 2, we consider only nonrandom factor loadings. The results hold when the factor loadings are random but independent of factors and errors. The divergence rate of $\Lambda'\Lambda$ is assumed to be N, which is the same as Bai and Ng (2002). It is possible to allow a divergence rate slower than N, which indicates a weaker factor structure. A slower divergence rate of $\Lambda'\Lambda$ will slow down the convergence rate of F_t^k and change the divergence rate of γ specified in Assumption 6. We conjecture that our bridge estimator still works when $\Lambda'\Lambda$ diverges at a slower rate and we leave that as a future research topic. Assumption 3 allows cross-sectional as well as serial dependence in the idiosyncratic errors. Hence, the model follows an approximate factor structure. Assumption 4 allows for some dependency between factors and errors. Assumption 5 is Assumption G in Bai (2003). This assumption ensures the existence of the limit of the rotation matrix H, which will be introduced in the next subsection.

Assumption 6 is an assumption about the tuning parameter. Compared with the single equation bridge estimation where the tuning parameter only depends on N or T, γ's divergence rate depends

on both N and T. Hence, we extend the application of the bridge estimation from single equation models to large panels. Also, note that a conventional bridge penalty will involve the sum of $|\lambda_{ij}|^{\delta}$ with $0 < \delta < 1$. However, as we use λ_{ij}^2 instead of $|\lambda_{ij}|$ in the penalty term, the restriction on the power has to be adjusted accordingly, i.e., $0 < 2\alpha < 1$.

2.2 Theoretical Results

Before discussing the main theoretical results, it is useful to describe some notations and transformations. We partition $\hat{F}^p$ in the following way:

$$\hat{F}^p = (\hat{F}^{1:r} \vdots \hat{F}^{(r+1):p}), \tag{2.7}$$

where $\hat{F}^{1:r}$ corresponds to the first r columns of $\hat{F}^p$, and $\hat{F}^{(r+1):p}$ corresponds to the last $p-r$ columns of $\hat{F}^p$. It is worth noting that there is an important difference between $\hat{F}^{1:r}$ and $\hat{F}^{(r+1):p}$. It is well known that $\hat{F}^{1:r}$ consistently estimates F^0 (for each t) up to some rotation (Bai, 2003), but $\hat{F}^{(r+1):p}$ is just some vectors containing noisy information from the idiosyncratic errors. In fact, the property of $\hat{F}^{(r+1):p}$ is to the best of our knowledge still hardly discussed in either the factor model or random matrix theory literature.[2] We will treat $\hat{F}^{1:r}$ and $\hat{F}^{(r+1):p}$ using different techniques.

Let $\hat{F}_t^{1:r}$ denote the transpose of the t^{th} row of $\hat{F}^{1:r}$. Bai (2003) shows that $\hat{F}_t^{1:r} - H'F_t^0 \to_p 0$, where H is some $r \times r$ matrix[3] converging to some nonsingular limit H_0. Since the factors are estimated up to the nonsingular rotation matrix H, we rewrite (2.2) as

$$X_i = F^0 H \cdot H^{-1}\lambda_i^0 + e_i. \tag{2.8}$$

Replacing the unobserved F^0H with the principal component estimates $\hat{F}^p$, we obtain the following transformation

$$\begin{aligned} X_i &= \hat{F}^{1:r}H^{-1}\lambda_i^0 + e_i + (F^0H - \hat{F}^{1:r})H^{-1}\lambda_i^0 \\ &= \hat{F}^p\lambda_i^* + u_i, \end{aligned} \tag{2.9}$$

where u_i and λ_i^* are defined as

$$u_i \equiv e_i - (\hat{F}^{1:r} - F^0H)H^{-1}\lambda_i^0, \qquad \lambda_i^* \equiv \begin{bmatrix} H^{-1}\lambda_i^0 \\ 0_{(p-r)\times 1} \end{bmatrix}. \tag{2.10}$$

We set the last $p-r$ entries of λ_i^* equal to zero because the model has r factors. Note that $\hat{F}^p$ is

[2] This is why Bai and Ng (2002) define their estimator for factors as $\hat{F}^k V^k$, where V^k is a diagonal matrix consisting of the first k largest eigenvalues of XX'/NT in a descending order. Since the k^{th} $(k > r)$ eigenvalue of XX'/NT is $O_p(C_{NT}^{-2})$, the last $p-r$ columns of their factor matrix are asymptotically zeros and only the first r columns of the factor matrix matter. As their criteria focus on the sum of squared errors, this asymptotically not-of-full-column-rank design does not affect their result. Since we focus on estimating and penalizing the factor loadings, we use $\hat{F}^k$ to ensure full column rank instead of Bai and Ng's (2002) $\hat{F}^k V^k$ as the estimator for F^0.

[3] The definition of H is given by Lemma 1 in the Appendix.

an estimated regressor and the transformed error term u_i involves two components: the true error term, which is heteroskedastic and serially correlated, and an additional term depending on $\hat{F}^{1:r}$, F^0 and λ_i^0. Compared with the i.i.d. errors in HHM (2008) or the independent errors in Huang *et al* (2009), our error term e_i is much less restricted (see Assumption 3). Also, note that the second term in u_i involves estimated factors via principal components and that (2.6) uses estimated factors instead of the unobserved F^0. These are new in the shrinkage literature and bring an extra layer of difficulty. Hence, our estimator in large panels is a nontrivial extension of existing methods.

Given the λ_i^* defined above, we can obtain the transformed $N \times p$ loading matrix:

$$\Lambda^* \equiv (\lambda_1^*, ..., \lambda_N^*)' \equiv \left[\Lambda^0 H^{-1\prime} \vdots 0_{N\times(p-r)}\right].$$

Thus, the goal is to prove that $\hat{\Lambda} - \Lambda^*$ converges to zero. The last $p-r$ columns $\hat{\Lambda}$ should be zero and its first r columns should be nonzero, so that we can consistently determine the number of factors using the number of nonzero columns in $\hat{\Lambda}$. Let $\hat{\lambda}_i$ denote the transpose of the i^{th} row of $\hat{\Lambda}$, the solution to (2.6). The following theorem establishes the convergence rate of $\hat{\lambda}_i$.

Theorem 1: *Under Assumptions 1 - 6,*

$$N^{-1}\sum_{i=1}^{N} \|\hat{\lambda}_i - \lambda_i^*\|^2 = O_p(C_{NT}^{-2}).$$

It is noteworthy that this rate is not affected by γ, as long as the divergence rate of γ satisfies Assumption 6. This is an extension of Horowitz, Huang, and Ma's (2008) result to a high dimensional factor model context. Under some appropriate assumptions, HHM (2008) show that the bridge estimator (denoted by $\hat{\beta}$) has the familiar OLS convergence rate, i.e., $\|\hat{\beta} - \beta_0\|^2 = O_p(N^{-1})$, in a single equation model. In this paper, the rate depends on both N and T and is similar to Bai's (2003) result for the OLS estimator of factor loadings (given the principal components estimates of factors). Hence, Theorem 1 shows that the group bridge estimator defined in (2.5) is as good as conventional estimators in terms of convergence rate. The next theorem shows that our estimator can estimate the last $p-r$ columns of $\hat{\Lambda}$ exactly as zeros with a probability converging to one.

Theorem 2: *Under Assumptions 1 - 6,*
(i)$\hat{\Lambda}^{(r+1):p} = 0_{N\times(p-r)}$ *with probability converging to one* as $N,\ T \to \infty$, *where* $\hat{\Lambda}^{(r+1):p}$ *denotes the last* $p-r$ *columns of* $\hat{\Lambda}$.
(ii) $P(\hat{\Lambda}^j = 0) \to 0$ *as* $N,\ T \to \infty$ *for any* $j = 1, \cdots, r$, *where* $\hat{\Lambda}^j$ *represents the* j^{th} *column of the* $\hat{\Lambda}$.

This theorem shows that our estimator achieves the selection consistency for the zero elements in Λ^*: the first r columns in $\hat{\Lambda}$ will be nonzero and last $p-r$ columns of $\hat{\Lambda}$ will be zero as N and T

diverge. Hence, it is natural to define the estimator for the number of factors as

$$\hat{r} \equiv \text{the number of nonzero columns of } \hat{\Lambda}. \tag{2.11}$$

The following corollary establishes the consistency of $\hat{r}$, and the proof is straightforward given Theorem 2.

Corollary 1*: Under Assumptions 1 - 6, the number of factors in (2.3) can be consistently determined by* $\hat{r}$.

2.3 Computation

In this subsection we show how to implement our method. The initial step is estimating the p factors via PCA in (2.4). Next, we show how to solve the optimization in (2.5) and (2.6). As the bridge penalty is not differentiable at zero for $0 < \alpha < 1/2$, standard gradient based methods are not applicable to our estimator. We develop an algorithm to compute the solution for (2.6). Let $\hat{\Lambda}^{(m)}$ be the value of the m^{th} iteration from the optimization algorithm, $m = 0, 1, ...$ Let $\mathscr{T}$ denote the convergence tolerance and ν denote some small positive number. In this paper, we set $\mathscr{T} = 5 \times 10^{-4}$ and $\nu = 10^{-4}$.
Set the initial value equal to the OLS solution, $\hat{\Lambda}^{(0)} = X'\hat{F}^p/T$. For $m = 1, 2, ...,$
(1) Let $\hat{\Lambda}^{(m),j}$ denote the j^{th} column of $\hat{\Lambda}^{(m)}$. Compute $g_1 = (X - \hat{F}^p\hat{\Lambda}^{(m)})'\hat{F}^p/N$ and

$$g_2(j,\nu) = -\frac{\alpha\gamma\hat{\Lambda}^{(m),j}}{N\left[\left(\hat{\Lambda}^{(m),j'}\hat{\Lambda}^{(m),j}/N\right)^{1-\alpha} + \nu\right]} \cdot \frac{2T}{C_{NT}^2},$$

where ν avoids the zero denominator when $\hat{\Lambda}^{(m),j} = 0$. Since the value of the tuning parameter will be selected as well, the constant term $2T/C_{NT}^2$ (for a given sample) can be dropped in the algorithm. Set $g_2(\nu) = [g_2(1,\nu), g_2(2,\nu), ..., g_2(p,\nu)]$.
(2) Define the $N \times p$ gradient matrix $g = [g^1,\ g^2, ..., g^p]$ $(N \times p)$, and the i^{th} element of the j^{th}column g^j $(j = 1, ..., p)$ is defined as

$$\begin{aligned} g_i^j &= g_{1,ij} + g_2(j,\nu)_i \quad \text{if } |\lambda_i^{(m),j}| > \mathscr{T} \\ g_i^j &= 0 \quad \text{if } |\lambda_i^{(m),j}| \leq \mathscr{T}, \end{aligned}$$

where $g_{1,ij}$ is the $(i,j)^{th}$ element of g_1, $g_2(j,\nu)_i$ is the i^{th} element of $g_2(j,\nu)$, and $\lambda_i^{(m),j}$ is the i^{th} element of $\hat{\Lambda}^{(m),j}$.
(3) Let $\max|g|$ denote the largest element in g in terms of absolute value. Re-scale $g = g/\max|g|$ if $\max|g| > 0$, otherwise set $g = 0$.
(4) $\hat{\Lambda}^{(m+1)} = \hat{\Lambda}^{(m)} + \Delta \times g/(1 + m/750)$, where Δ is the increment for this iteration algorithm. We set $\Delta = 2 \times 10^{-3}$, which follows the value used by HHM (2008).

(5) Replace m by $m+1$ and repeat steps (1) - (5) until

$$\max_{i=1,...,N;\ j=1,...,p} |\lambda_i^{(m),j} - \lambda_i^{(m+1),j}| \leq \mathscr{T}.$$

After convergence, we truncate $\lambda_i^{(m),j}$ to zero if $|\lambda_i^{(m),j}| \leq \mathscr{T}$.
(6) Set $\hat{r}$ = the number of nonzero columns in $\hat{\Lambda}^{(m)}$.

This algorithm is a modified version of the one proposed by HHM (2008). We use the OLS estimate instead of zero as the initial value to accelerate the algorithm. Also, we modify HHM's computation of the gradient in step (2) so that the estimated loading will remain unchanged after being shrunk to zero. This is suggested by Fan and Li (2001) and can substantially accelerate the convergence in our high dimensional setup. Additionally, note that in step (4) we gradually decrease the increment, as m increases by dividing $1+m/750$.[4] This ensures the convergence of the algorithm.

To see that the algorithm yields an estimator converging to the true value, first note that the algorithm approximates the bridge penalty function by

$$\frac{2\alpha\gamma}{C_{NT}^2 N} \cdot \sum_{j=1}^{p} \int^{\Lambda^j} \frac{1}{(u'u/N)^{1-\alpha}+\nu} u'du, \tag{2.12}$$

where the variable of integration u is an $N \times 1$ vector and $\int^{\Lambda^j}$ is a line integral through the gradient of the approximated penalty term with the endpoint Λ^j. (2.12) is continuously differentiable with respect to the $N \times 1$ vector Λ^j, so the computed estimator will converge to the minimum of the approximated objective function. Also, the approximated penalty function and its gradient converge to the bridge penalty and its gradient as $\nu \to 0$, respectively. Since all factor loadings are defined in a compact set by Assumption 2 and the bridge estimator is a global minimizer by HHM (2008), the minimum of the approximated objective function will converge to the minimum of (2.6) as $\nu \to 0$. Hence, our algorithm can generate a consistent estimator. To ensure that the algorithm is not trapped in a local minimum, we use the OLS estimate as the initial value, which is close to the global minimum of (2.6) and reduces the computation load. One could also use multiple starting values to avoid obtaining a local minimum.

To implement our estimator, we need to determine the values of α and γ. We set $\alpha = 0.25$ as our benchmark value. We also vary α between 0.1 and 0.4; simulations (see Table 7) show that our estimator is very robust to the choice of α. The tuning parameter γ is set equal to $\phi \cdot [\min(N,T)]^{0.45}$ because Assumption 6 requires that $\gamma/C_{NT} \to 0$. The constant ϕ varies on the grid [1.5:0.25:8]. Instead of searching on a grid such as $\gamma = 1, 5, 10, ..., 1000$ for all sample sizes, this setup allows the grid to change as the sample size changes, so that it avoids unnecessary computation and thus

[4]Our Monte Carlo experiments show that the algorithm is reasonably fast and performs very well in terms of selecting the true value of r. We also tried 500 and 1000 instead of 750 to adjust the increment, and the results are very similar and hence not reported.

accelerates our algorithm. We follow the method developed by Wang *et al.* (2009) to determine the optimal choice of γ. The optimal γ is the one that minimizes the following criterion:

$$\log[(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\hat{\lambda}_i'\hat{F}_t^p)^2]+\hat{r}(\gamma)\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)\cdot\ln[\ln(N)],$$

where $\hat{r}(\gamma)$ denotes the fact that the estimated number of factors is affected by the choice of γ. Note that the number of parameters in our model is proportional to N, so we use $\ln[\ln(N)]$ in the penalty, which has the same rate as suggested by Wang *et al.* (2009).

3 Simulations

In this section, we explore the finite sample properties of our estimator. We also compare our estimator with some of the estimators in the literature: IC_{p1} proposed by Bai and Ng (2002), $IC_{1;n}^T$ proposed by Hallin and Liska (2007), ED proposed by Onatski (2010), and ER and GR proposed by Seung and Horenstein (2013).

Given the principal component estimator $\hat{F}^k$, Bai and Ng (2002) use OLS to estimate the factor loadings by minimizing the sum of squared residuals:

$$V(k,\hat{F}^k)=\min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\lambda_i^{k'}\hat{F}_t^k)^2.$$

Then the IC_{p1} estimator is defined as the k that minimizes the following criterion:

$$IC_{p1}(k)=\log[V(k,\hat{F}^k)]+k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right).$$

The criterion can be considered the analog of the conventional BIC in the factor model context. It penalizes the integer k to prevent the model from overfitting.

Hallin and Liska's (2007) estimator is designed to determine the number of dynamic factors (usually denoted by q in the literature), but it is applicable in our Monte Carlo experiments where static and dynamic factors coincide by design, i.e. $r=q$. Hallin and Liska's (2007) estimator follows the idea of Bai and Ng (2002), but their penalty term involves an additional scale parameter c. In theory, the choice of $c>0$ does not matter as N and $T\to\infty$. For a given N and T, they propose an algorithm to select the scale parameter: c is varied on a predetermined grid ranging from zero to some positive number $\bar{c}$, and their statistic is computed for each value of c using J nested subsamples ($i=1,...,N_j;\ t=1,...,T_j$) for $j=1,...,J$, where J is a fixed positive integer, $0<N_1<...<N_j<...<N_J=N$ and $0<T_1<...<T_j<...<T_J=T$. Hallin and Liska (2007) show that $[0,\bar{c}]$ will contain several intervals that generate estimates not varying across subsamples. They call these "stability intervals" and suggest that the c belonging to the second stability interval is the optimal one (the first stability intervals contains zero). Since the scale parameter in $IC_{1;n}^T$

can adjust its penalty, it is expected that $IC^T_{1;n}$ should be more robust to the correlation in the idiosyncratic errors than Bai and Ng's (2002) estimators, whose penalty solely depends on N and T.

Onatski's (2010) ED estimator is designed to select the number of static factors, so it is directly comparable with our estimator. ED chooses the largest k that belongs to the following set:

$$\{k \leq p : \ \rho_k(X'X/T) - \rho_{k+1}(X'X/T) > \zeta\},$$

where ζ is a fixed positive constant, and $\rho_k(.)$ denotes the k^{th} largest eigenvalue of a matrix. The choice of ζ takes into account the empirical distribution of the eigenvalues of the data sample covariance matrix[5], so it is also a robust estimator to correlated error terms.

Recently, Seung and Horenstein's (2013) propose two estimators based on the eigenvalues of $X'X$. They define

$$ER(k) = \frac{\rho_k(X'X/NT)}{\rho_{k+1}(X'X/NT)}, \qquad GR(k) = \frac{\ln[V(k-1)/V(k)]}{\ln[V(k)/V(k+1)]}$$

where $V(k) = \sum_{j=k+1}^{\min(N,T)} \rho_j(X'X/NT)$. The ER and GR estimators are the integers that maximize $ER(k)$ and $GR(k)$, respectively[6]. Seung and Horenstein's (2013) results show that their estimators can outperform the ED estimator in many cases. Thus, it is expected that competing with $IC^T_{1:n}$, ED, ER, and GR will not be an easy task; however, simulations show that our estimator is still more accurate in certain cases, so it will be useful and valuable in practice.

We conduct seven experiments in our Monte Carlo simulations. In the first six experiments, we set $\alpha = 0.25$. In the last experiment, we vary α between 0.1 and 0.4 to see the stability of our estimator. The computation of our estimator and the selection of the tuning parameter follow the steps outlined in section 2.3. We consider the data generating process (DGP) similar to that of Bai and Ng (2002):

$$X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{jt} + \sqrt{\theta} e_{it}$$

where the factors F_{jt}'s are drawn from $N(0,1)$ and the factor loadings λ_{ij}'s are drawn from $N(0.5,1)$. θ will be selected to control the signal-to-noise ratio in the factor model. The following DGPs are applied to generate the error term e_{it} in the seven different experiments:

E1: $e_{it} = \sigma_i u_{it}$, $u_{it} = \rho u_{i,t-1} + v_{it} + \sum_{1 \leq |h| \leq 5} \beta v_{i-h,t}$, where $v_{it} \sim i.i.d.\ N(0,\ 1)$ and σ_i is $i.i.d.$ and uniformly distributed between 0.5 and 1.5.
E2: $e_{it} = v_{it}\|F_t\|$, where $v_{it} \sim i.i.d.\ N(0,1)$.

DGP E1 is similar to the setup in Bai and Ng (2002). The parameters β and ρ control the cross-

[5]See Onatski (2010) for the details of the computation of ζ.
[6]Ideas similar to the ER estimator have also been considered by Luo *et al.* (2009) and Wang (2012)

sectional and serial correlations of the error terms, respectively. We also consider cross-sectional heteroskedasticity by introducing σ_i. The setup of σ_i is the same as that in Breitung and Eickmeier (2011). Note that $E(\lambda_{ij}F_{jt})^2 = 5/4$ and $E(\sigma_i^2) = 13/12$. We set $\theta = \theta_0 = 15r(1-\rho^2)/13(1+10\beta^2)$ in experiments 1–4 and 7 so that the factors explain 50% of the variation in the data. In experiment 6, we set $\theta = 2\theta_0$ to explore how the estimators perform in data with a lower signal-to-noise ratio. DGP E2 is applied in experiment 5, where we explore the case of conditional heteroskedasticity. In this experiment, we set $\theta = 5/4$ so that the R^2 is also 50%. All experiments are replicated 1000 times. The upper bound on the number of factors is 10, i.e., $p = 10$.

In experiments 1 – 4, we vary the values of β and ρ. Table 1 summarizes the results of our first experiment with no correlation in the errors, i.e., $\beta = \rho = 0$. The numbers outside the parentheses are the means of different estimators for the number of factors, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors. It is not surprising that IC_{p1} is rather accurate with no correlation in e. When $r = 5$ and $T = 50$, our estimator is better than $IC^T_{1;n}$ but less accurate than ED, ER, and GR. However, as N and T increase, our estimator can detect the correct number of factors almost 100% of the time. In contrast, $IC^T_{1;n}$ still has a downward bias with $N = T = 200$, when $r = 3$ or 5.

In experiments 2 – 4, we consider three correlation structures in the error terms: cross-sectional correlation only ($\beta = 0.2$ and $\rho = 0$), serial correlation only ($\beta = 0$ and $\rho = 0.7$), and co-existing cross-sectional and serial correlations ($\beta = 0.1$ and $\rho = 0.6$). The results are reported in Tables 2 – 4, respectively. The results of these three experiments demonstrate very similar patterns. Bai and Ng's (2002) IC_{p1} tends to overestimate the number of factors. When $r = 1$, the estimates by IC_{p1} change dramatically as more data are introduced. All other estimators tend to underestimate the number of factors when $r = 3$ or 5 in small samples (N or $T = 50$). However, our estimator is more accurate than $IC^T_{1;n}$ and ED, especially when $r = 5$. Compared with ER and GR, our estimator also has some advantages when errors are serially correlated. For example, our estimator has uniformly smaller biases than GR except when $r = 1$ and $N = T = 50$ in Table 4.

In our fifth experiment, we consider conditionally heteroskedastic errors using DGP E2. The results are reported in Table 5. IC_{p1} substantially overestimates the number of factors when $r = 1$. Compared with $IC^T_{1;n}$, our estimator tends to be more accurate, especially in small samples. Also, none of ED, GR, and our estimator dominates the other two. Our estimator is more accurate than ED when $r = 1$, and more accurate than ER and GR when $r = 3$, whereas ER and GR tend to have smaller biases when $r = 1$ and $r = 5$.

Table 6 reports the results of our sixth experiment, where we consider a weaker factor structure by setting $\theta = 2\theta_0$, i.e., factors only explain 1/3 of the variation in the data. The error terms are generated using DGP E2 with $\beta = 0.1$ and $\rho = 0.6$. It is expected that it is more difficult to estimate the correct number of factors in this setup. Bai and Ng's (2002) IC_{p1} is still severely upwardly biased. In addition, our estimator tends to perform better than $IC^T_{1;n}$ and ED when $r = 1$ or 3. Compared with ER and GR, our estimator tends to have smaller biases when $N = 200$ or

$T = 200$.

As mentioned previously, it is very difficult for our estimator to outperform the competitors uniformly. However, it can be seen from Tables 1 - 6 that our estimator performs better than the competing estimators under many circumstances, especially when the idiosyncratic shocks have substantial serial and cross-sectional correlations.

In the last experiment, we vary the value of α to check the stability of our estimator. We set $\alpha \in \{0.1, 0.15, 0.25, 0.35, 0.4\}$, $(\beta, \rho) \in \{(0,0),\ (0.1, 0.6)\}$ and $N = T = 100$. When $r = 0$, θ is set equal to $15(1-\rho^2)/13(1+10\beta^2)$. Generally speaking, our estimator is rather stable to the choice of α. When $r = 1$ or 3, our estimator almost always gives the correct result. When $r = 5$, our estimator performs well for all values of α except for $\alpha = 0.4$: the estimate is 3.36 when the errors are correlated. When $r = 0$, our estimator performs well for all values of α except for $\alpha = 0.1$: it has an upward bias when the errors are correlated. However, our simulation (not reported in the table) confirms that these biases vanish as N and T increase. These results indicate that it is better to avoid very small (close to zero) or very large (close to 0.5) α for small samples in practice.

4 Empirical Application

In this section, we apply our method to the data set used by Stock and Watson (2005). The data set consists of 132 US macroeconomic time series, spanning the period of 1960.1-2003.12. The variables are transformed to achieve stationarity and then outliers are adjusted as described in Appendix A of Stock and Watson (2005). We set the upper bound on the number of factors equal to 10 and apply various methods to the data. First, one should be careful about the interpretation of the result by Hallin and Liska's (2007) $IC^T_{1;n}$. Unlike in our simulations, where the dynamic factors are by design the same as static factors, the number of static and dynamic factors are not necessarily the same in practice. Hence, $IC^T_{1;n}$ is inconclusive with regards to the number of static factors in the data. It finds five dynamic factors; however, this only implies that the number of static factors is no less than five. Alessi, Barigozzi, and Capasso's (2010) estimator, which is the static factor version of Hallin and Liska's estimator, suggests that there are two static factors. Second, Bai and Ng's (2002) IC_{p1} and PC_{p1} find seven and nine static factors, respectively. Since the data set consists of monthly observations on macroeconomic variables from several categories (industrial production indexes, price indexes, employment, housing start, asset returns, etc.), the error terms are likely to be correlated in both time and cross-sectional dimensions. Our simulations have already shown that IC_{p1} and PC_{p1} tend to overestimate the correct number of factors when the errors are correlated. The point is also made by Uhlig (2009) that the high number of factors found by IC_{p1} and PC_{p1} may be due to the high temporal persistence of the data. In contrast, Onatski's (2010) ED and Seung and Horenstein's (2013) ER and GR detect only one static factor. Additionally, we implement our estimator and find two static factors when $\alpha = 0.25$. We also vary α on the grid $[0.1 : 0.05 : 0.4]$ and the result is very stable. Our estimator finds one static factors when $\alpha = 0.35$ and two static factors for all other values of α.

Finally, we compare the accuracy of forecasts based on different numbers of factors using the same data set. We use rolling a window scheme and the last 120 periods are used for out-of-sample evaluation based on the mean square forecast errors (MSFE). We consider the one-step ahead forecasting model:

$$X_{it+1} = c_{i0} + c_{i1}(L)\hat{F}_t^k + c_{i2}(L)X_{it} + \epsilon_{it+1}, \tag{4.1}$$

where $\hat{F}_t^k$ denotes the first k factors estimated by principal components using rolling windows, and $c_{i1}(L)$ and $c_{i2}(L)$ are lag polynomials. The lag orders of $c_{i1}(L)$ and $c_{i2}(L)$, denoted by ℓ_{i1} and ℓ_{i2}, respectively, are selected by minimizing the MSFE for the given k and i. We set $0 \leq k \leq 10$, $0 \leq \ell_{i1} \leq 6$, and $0 \leq \ell_{i2} \leq 6$. We use the AR model as the benchmark where no factor is included in (4.1), i.e., $k = 0$. We compute the ratios of MSFEs of (4.1) for $k = 1, ..., 10$ to the MSFE of the benchmark model, so 10 relative MSFEs are obtained for each of the 132 series. It should be noted that a factor informative for one series is not necessarily useful for the prediction of another series. Hence, it is almost impossible for a specific value of k to generate the best forecasts for all 132 time series. Here we are only interested in which value of k can generate better forecasts on average. Each column of Table 8 reports the mean and median of the 132 relative MSFEs. The mean of MSFEs is minimized when $k = 2$, and the median of MSFEs is minimized when $k = 3$, so the overall performance based on two or three factors is the best. This is close to the finding of our new bridge estimator.

5 Conclusions

In this paper, we develop a group bridge estimator to determine the correct number of factors in approximate factor models. This extends the conventional bridge estimator from a single equation to a large panel context. The proposed estimator can consistently estimate the factor loadings of relevant factors and shrink the loadings of irrelevant factors to zero with a probability approaching one. Hence, the new estimator can select the correct model specification and conduct the estimation of factor loadings in one step. Monte Carlo simulations confirm our theoretical results and show that the new estimator has a good performance in small samples.

In this paper, we only consider the bridge type of estimator, but it is possible to generalize the result and find the whole family of shrinkage estimators that can achieve the oracle property in factor models. Also, the application of bridge estimators should not be limited to selecting the number of factors. We expect that it can be applied in other contexts, such as selecting the identification scheme in the factor-augmented VAR model of Stock and Watson (2005), which involves a large number of zero restrictions. We leave these issues for future research.

Appendix

Section A of the appendix provides several useful lemmas and Section B provides the proofs of Theorems 1 – 2.

A Proofs of Lemmas

In this section, we prove four lemmas that are useful for the proofs of the theorems. Lemma 1 cites some useful results of Bai (2003). Lemma 2 obtains the correlation between u_i and $\hat{F}^p$. Lemma 3 shows that $N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 = o_p(1)$, which is useful to prove the convergence rate in Theorem 1. Lemma 4 is derived based on Lemma 3, and it ensures that the mean value theorem is applicable in the proof of Theorem 1.

Lemma 1:
(i) Under Assumptions 1 - 4,

$$T^{-1}\sum_{t=1}^{T}\|\hat{F}_t^{1:r} - H'F_t^0\|^2 = O_p(C_{NT}^{-2}),$$

where $H = (\Lambda^{0\prime}\Lambda^0/N)(F^{0\prime}\hat{F}^{1:r}/T)V^{r^{-1}}$ and V^r is the diagonal matrix consisting of the first r largest eigenvalues of XX'/NT in descending order.
(ii) Under Assumptions 1 - 5,

$$H \to_p H_0$$

where $H_0 \equiv \Sigma_\Lambda Q' V_0^{-1}$ is nonsingular and $\|H_0\| < \infty$, $Q \equiv \text{plim}_{N,T\to\infty}\hat{F}^{1:r\prime}F^0/T$ is nonsingular, and V_0 is the diagonal matrix consisting of the r positive eigenvalues of $\Sigma_\Lambda\Sigma_F$ in descending order.
Proof:
(i) This is Lemma A.1 of Bai (2003).
(ii) Proposition 1 of Bai (2003) proves that $\hat{F}^{1:r\prime}F^0/T$ converges in probability to a nonsingular limit Q. Also, Lemma A.3 of Bai (2003) proves that $V^r \to_p V_0$ as $N, T \to \infty$. Combining these two results, the definition of H and Assumption 2, it follows that $H \to_p \Sigma_\Lambda Q' V_0^{-1}$. Additionally, Assumptions 1 and 2 imply that $\Sigma_\Lambda\Sigma_F$ is positive definite and finite, so V_0 is nonsingular and finite. Hence, V_0^{-1} is also nonsingular and finite. Since Q is nonsingular and finite by Proposition 1 of Bai (2003), it follows that H_0 is nonsingular and $\|H_0\| < \infty$. ■

Lemma 2: Under Assumptions 1 - 4,

$$\frac{1}{NT}\sum_{i=1}^{N}\left\|\frac{u_i'\hat{F}^p}{\sqrt{T}}\right\|^2 = O_p(C_{NT}^{-2})$$

Proof: We use (2.10) and (2.7), so

$$\frac{1}{\sqrt{T}}u_i'\hat{F}^p = \frac{1}{\sqrt{T}}[e_i - (\hat{F}^{1:r}H^{-1} - F^0)\lambda_i^0]'(\hat{F}^{1:r}\vdots\hat{F}^{(r+1):p})$$

It will be sufficient to show that $(NT)^{-1}\sum_{i=1}^{N}\|a_{i,j}\|^2 = O_p(C_{NT}^{-2})$ for $j = 1, 2, 3, 4$, where

$$\begin{aligned}
a_{i,1} &= \frac{1}{\sqrt{T}} e_i' \hat{F}^{1:r} \\
a_{i,2} &= \frac{1}{\sqrt{T}} \lambda_i^{0\prime} (\hat{F}^{1:r} H^{-1} - F^0)' \hat{F}^{1:r} \\
a_{i,3} &= \frac{1}{\sqrt{T}} e_i' \hat{F}^{(r+1):p} \\
a_{i,4} &= \frac{1}{\sqrt{T}} \lambda_i^{0\prime} (\hat{F}^{1:r} H^{-1} - F^0)' \hat{F}^{(r+1):p}
\end{aligned}$$

Now, for the first term,

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\|a_{i,1}\|^2 &= \frac{1}{NT^2}\sum_{i=1}^{N}\|e_i'\hat{F}^{1:r}\|^2 \\
&\leq \frac{2}{NT^2}\sum_{i=1}^{N}\left(\|e_i'(\hat{F}^{1:r} - F^0 H)\|^2 + \|e_i' F^0 H\|^2\right) \\
&\leq \frac{2}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} e_{it}^2\right)\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t^{1:r} - H' F_t^0\|^2 + \frac{2}{NT}\sum_{i=1}^{N}\left\|\frac{e_i' F^0}{\sqrt{T}}\right\|^2 \|H\|^2 \\
&= O_p(C_{NT}^{-2}) + O_p(T^{-1})
\end{aligned}$$

where we use Assumptions 3(i) and 4, the facts that $\|H\| = O_p(1)$ by Lemma 1(ii) and that $\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t^{1:r} - H' F_t^0\|^2 = O_p(C_{NT}^{-2})$ by Lemma 1(i). For the second term,

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\|a_{i,2}\|^2 &= \frac{1}{NT^2}\sum_{i=1}^{N}\|\lambda_i^{0\prime}(\hat{F}^{1:r} H^{-1} - F^0)'\hat{F}^{1:r}\|^2 \\
&\leq \frac{2}{NT^2}\sum_{i=1}^{N}\|\lambda_i^{0\prime} H^{-1\prime}\|^2\left(\|(\hat{F}^{1:r} - F^0 H)'(\hat{F}^{1:r} - F^0 H)\|^2 + \|(\hat{F}^{1:r} - F^0 H)' F^0 H\|^2\right) \\
&\leq \frac{2}{N}\sum_{i=1}^{N}\|\lambda_i^{0\prime}\|^2\|H^{-1\prime}\|^2\left[\left(\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t^{1:r} - H' F_t^0\|^2\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t^{1:r} - H' F_t^0\|^2 \frac{1}{T}\sum_{t=1}^{T}\|H' F_t^0\|^2\right] \\
&= O_p(1)[O_p(C_{NT}^{-4}) + O_p(C_{NT}^{-2})]
\end{aligned}$$

where the $O_p(1)$ term follows from Assumption 2 and the fact that H^{-1} is nonsingular, the $O_p(C_{NT}^{-4})$ term follows from Lemma 1(i) and $O_p(C_{NT}^{-2})$ follows from Lemma 1(i) and Assumption 1.

For the third term, first note that the OLS estimate

$$\hat{\lambda}_{i,OLS}^{(r+1):p} \equiv \hat{F}^{(r+1):p\prime} X_i / T. \tag{A.1}$$

Let V^p be the diagonal matrix of the first largest p eigenvalues of XX'/NT in decreasing order. Note that $XX'\hat{F}^p/NT = \hat{F}^p V^p$ and $\hat{F}^{p\prime}\hat{F}^p/T = I_p$, so it follows that $\hat{F}^{p\prime} XX'\hat{F}^p/(NT^2) = V^p$.

Since $\hat{\Lambda}_{OLS} = X'\hat{F}^p/T$, we have

$$\frac{1}{N}\hat{\Lambda}'_{OLS}\hat{\Lambda}_{OLS} = V^p$$

$$\frac{1}{N}\|\hat{\Lambda}_{OLS}\|^2 = \frac{1}{N}\text{trace}(\hat{\Lambda}'_{OLS}\hat{\Lambda}_{OLS}) = \sum_{j=1}^{p} v_j \tag{A.2}$$

where v_j is the j^{th} largest eigenvalue of XX'/NT.

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\|a_{i,3}\|^2 &= \frac{1}{NT^2}\sum_{i=1}^{N}\|e_i'\hat{F}^{(r+1):p}\|^2 \\
&= \frac{1}{NT^2}\sum_{i=1}^{N}\|(X_i - F^0\lambda_i^0)'\hat{F}^{(r+1):p}\|^2 \\
&\leq \frac{2}{NT^2}\sum_{i=1}^{N}\left(\|\hat{\lambda}_{i,OLS}^{(r+1):p\prime}T\|^2 + \|\lambda_i^{0\prime}F^{0\prime}\hat{F}^{(r+1):p}\|^2\right) \\
&= 2\sum_{j=r+1}^{p} v_j + \frac{2}{NT^2}\sum_{i=1}^{N}\|\lambda_i^{0\prime}F^{0\prime}\hat{F}^{(r+1):p}\|^2
\end{aligned} \tag{A.3}$$

where we use the definition of $\hat{\lambda}_{i,OLS}^{(r+1):p}$ in the third line of (A.3) and (A.2) in the last line of (A.3). For the first term, Lemma 4 of Bai and Ng (2002) shows that $\sum_{j=r+1}^{p} v_j = O_p(C_{NT}^{-2})$. For the second term in (A.3), since $\hat{F}^{1:r\prime}\hat{F}^{(r+1):p} \equiv 0_{r\times(p-r)}$, it can be rewritten as

$$\begin{aligned}
\frac{2}{NT^2}\sum_{i=1}^{N}\|\lambda_i^{0\prime}F^{0\prime}\hat{F}^{(r+1):p}\|^2 &= \frac{2}{NT^2}\sum_{i=1}^{N}\|\lambda_i'H^{-1\prime}(F^0H - \hat{F}^{1:r})'\hat{F}^{(r+1):p}\|^2 \\
&\leq \frac{2}{N}\sum_{i=1}^{N}\|\lambda_i^{0\prime}H^{-1\prime}\|^2\left(\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t^{1:r} - H'F_t^0\|^2\right)\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t^{(r+1):p}\|^2 \\
&= O_p(1)O_p(C_{NT}^{-2})O_p(1)
\end{aligned} \tag{A.4}$$

where we use Assumption 2, Lemma 1(i), and the facts that H is nonsingular and that $\hat{F}^{(r+1):p\prime}\hat{F}^{(r+1):p}/T = I_{p-r}$. Thus, $\frac{1}{NT}\sum_{i=1}^{N}\|a_{i,3}\|^2 = O_p(C_{NT}^{-2})$.

For the fourth term,

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\|a_{i,4}\|^2 &\leq \frac{2}{N}\sum_{i=1}^{N}\|\lambda_i^{0\prime}H^{-1\prime}\|^2\|\frac{1}{T}(\hat{F}^{1:r} - F^0H)'\hat{F}^{(r+1):p}\|^2 \\
&= O_p(1)O_p(C_{NT}^{-2})
\end{aligned}$$

where we use the same argument as the one to prove (A.4). To sum up, we have shown that $(NT)^{-1}\sum_{i=1}^{N}\|a_{i,j}\|^2 = O_p(C_{NT}^{-2})$ for $j = 1, 2, 3, 4$, which implies the desired result. ■

Lemma 3: Under Assumptions 1 - 6, $N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 = o_p(1)$.
Proof:

Recall that $\hat{\lambda}_i$ denotes the transpose of the i^{th} row of $\hat{\Lambda}$, the solution to (2.6). It follows that

$$\begin{aligned} & \frac{1}{NT}\sum_{i=1}^{N}(X_i - \hat{F}^p\hat{\lambda}_i)'(X_i - \hat{F}^p\hat{\lambda}_i) + \frac{\gamma}{C_{NT}^2}\sum_{j=1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}_{ij}^2\right)^{\alpha} \\ \leq & \frac{1}{NT}\sum_{i=1}^{N}(X_i - \hat{F}^p\lambda_i^*)'(X_i - \hat{F}^p\lambda_i^*) + \frac{\gamma}{C_{NT}^2}\sum_{j=1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_{ij}^{*2}\right)^{\alpha} \end{aligned}$$

Let $\eta \equiv C_{NT}^{-2}\gamma\sum_{j=1}^{p}\left(N^{-1}\sum_{i=1}^{N}\lambda_{ij}^{*2}\right)^{\alpha}$. Let $\psi_i \equiv (\hat{F}^{p\prime}\hat{F}^p)^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^*)$ and $D \equiv \hat{F}^p(\hat{F}^{p\prime}\hat{F}^p)^{-\frac{1}{2}}$. By (2.9), it is straightforward that

$$\begin{aligned} \eta \geq & \frac{1}{NT}\sum_{i=1}^{N}(X_i - \hat{F}^p\hat{\lambda}_i)'(X_i - \hat{F}^p\hat{\lambda}_i) - \frac{1}{NT}\sum_{i=1}^{N}(X_i - \hat{F}^p\lambda_i^*)'(X_i - \hat{F}^p\lambda_i^*) \\ = & \frac{1}{NT}\sum_{i=1}^{N}[(\lambda_i^* - \hat{\lambda}_i)'\hat{F}^{p\prime}\hat{F}^p(\lambda_i^* - \hat{\lambda}_i) - 2u_i'\hat{F}^p(\hat{\lambda}_i - \lambda_i^*)] \\ = & \frac{1}{NT}\sum_{i=1}^{N}[(\psi_i - D'u_i)'(\psi_i - D'u_i) - u_i'D'Du_i] \end{aligned}$$

Hence,

$$\frac{1}{NT}\sum_{i=1}^{N}\|\psi_i - D'u_i\|^2 \leq \frac{1}{NT}\sum_{i=1}^{N}\|D'u_i\|^2 + \eta \tag{A.5}$$

Since $\frac{1}{2}\|\psi_i\|^2 \leq \|\psi_i - D'u_i\|^2 + \|D'u_i\|^2$,

$$\frac{1}{2NT}\sum_{i=1}^{N}\|\psi_i\|^2 = \frac{1}{2N}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 \leq \frac{2}{NT}\sum_{i=1}^{N}\|D'u_i\|^2 + \eta$$

where we use the fact that

$$\|\psi_i\|^2 = (\lambda_i^* - \hat{\lambda}_i)'\hat{F}^{p\prime}\hat{F}^p(\lambda_i^* - \hat{\lambda}_i) = T\|\lambda_i^* - \hat{\lambda}_i\|^2$$

because $\hat{F}^{p\prime}\hat{F}^p/T = I_p$. By Lemma 2 and the definition of $D = T^{-1/2}\hat{F}^p(\hat{F}^{p\prime}\hat{F}^p/T)^{-1/2} = \hat{F}^p/T^{1/2}$, we obtain $(NT)^{-1}\sum_{i=1}^{N}\|D'u_i\|^2 = O_p(C_{NT}^{-2})$. Also, see that

$$\eta \leq C_{NT}^{-2}\gamma p\bar{\lambda}^{2\alpha} = C_{NT}^{-2}\gamma O_p(1).$$

By Assumption 6(i) that $\gamma/C_{NT} \to 0$, we obtain

$$\frac{1}{N}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 = o_p(1) \tag{A.6}$$

■

Lemma 4: Under Assumptions 1 - 4,

$$N^{-1}\sum_{i=1}^{N}\hat{\lambda}_{ij}^2 - N^{-1}\sum_{i=1}^{N}\lambda_{ij}^{*^2} = o_p(1).$$

Proof:

$$\begin{aligned}
\left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij}^2 - \lambda_{ij}^{*^2})\right| &= \left|\frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij} - \lambda_{ij}^*)(\hat{\lambda}_{ij} + \lambda_{ij}^*)\right| \\
&\leq \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij} - \lambda_{ij}^*)^2 \frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij} + \lambda_{ij}^*)^2\right]^{\frac{1}{2}} \\
&\leq \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij} - \lambda_{ij}^*)^2\right]^{\frac{1}{2}} \left[\frac{2}{N}\sum_{i=1}^{N}\left((\hat{\lambda}_{ij} - \lambda_{ij}^*)^2 + 4\lambda_{ij}^{*^2}\right)\right]^{\frac{1}{2}} \\
&\leq \frac{\sqrt{2}}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij} - \lambda_{ij}^*)^2 + \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij} - \lambda_{ij}^*)^2\right]^{\frac{1}{2}} \left(\frac{8}{N}\sum_{i=1}^{N}\|\lambda_i^0\|^2\right)^{\frac{1}{2}} \|H^{-1}\| \quad \text{(A.7)} \\
&= o_p(1)
\end{aligned}$$

where we use (A.6), Assumption 2 and the definition of λ_i^* in (2.10). ■

B Proofs of Theorems

Proof of Theorem 1:
Lemma 3 has proved that $N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 = o_p(1)$. Hence, we only need to show that

$$C_{NT}^2 N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 = O_p(1). \tag{B.1}$$

For each N and T, we partition the parameter space into the shells $S_{j,N,T} = \{\hat{\Lambda} : 2^{j-1} \leq C_{NT}^2 N^{-1}\|\hat{\Lambda} - \Lambda^*\|^2 < 2^j\}$ with j ranging over integers. If $C_{NT}^2 N^{-1}\|\hat{\Lambda} - \Lambda^*\|^2 > 2^M$ for some integer M, then $\hat{\Lambda}$ is in one of the shells with $j \geq M$. We want to show that this event happens

with probability approaching zero. Since $\hat{\Lambda}$ minimizes (2.6), it follows that for any $\varepsilon > 0$,

$$\begin{aligned}
& P(C_{NT}^2 N^{-1} \sum_{i=1}^{N} \|\hat{\lambda}_i - \lambda_i^*\|^2 > 2^M) \\
\leq & \sum_{j \geq M, 2^j \leq \varepsilon C_{NT}^2} P\left(\inf_{\hat{\Lambda} \in S_{j,N,T}} [L(\hat{\Lambda}) - L(\Lambda^*)] \leq 0 \right) + P\left(\frac{2}{N} \|\hat{\Lambda} - \Lambda^*\|^2 > \varepsilon \right) \\
\leq & \sum_{j \geq M,\ 2^j \leq \varepsilon C_{NT}^2} P\left(\inf_{\hat{\Lambda} \in S_{j,N,T}} [L(\hat{\Lambda}) - L(\Lambda^*)] \leq 0,\ \frac{1}{NT^2} \sum_{i=1}^{N} \|u_i' \hat{F}^p\|^2 \leq \frac{M}{C_{NT}^2} \right) \\
& + P\left(\frac{2}{N} \|\hat{\Lambda} - \Lambda^*\|^2 > \varepsilon \right) + P\left(\frac{1}{NT^2} \sum_{i=1}^{N} \|u_i' \hat{F}^p\|^2 > \frac{M}{C_{NT}^2} \right)
\end{aligned}$$

where the second term $P\left(\frac{2}{N}\|\hat{\Lambda} - \Lambda^*\|^2 > \varepsilon\right) \to 0$ by (A.6) and the third term is also $o(1)$ as $M \to \infty$ by Lemma 2. Then it will be sufficient to show that the first term $\to 0$ as $M \to \infty$. For the first term, note that

$$\begin{aligned}
& L(\hat{\Lambda}) - L(\Lambda^*) \\
= & \frac{1}{NT} \sum_{i=1}^{N} (X_i - \hat{F}^p \hat{\lambda}_i)'(X_i - \hat{F}^p \hat{\lambda}_i) + \frac{\gamma}{C_{NT}^2} \sum_{j=1}^{p} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{\lambda}_{ij}^2 \right)^{\alpha} \\
& - \frac{1}{NT} \sum_{i=1}^{N} (X_i - \hat{F}^p \lambda_i^*)'(X_i - \hat{F}^p \lambda_i^*) - \frac{\gamma}{C_{NT}^2} \sum_{j=1}^{p} \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_{ij}^{*2} \right)^{\alpha} \\
\geq & \frac{1}{NT} \sum_{i=1}^{N} (X_i - \hat{F}^p \hat{\lambda}_i)'(X_i - \hat{F}^p \hat{\lambda}_i) + \frac{\gamma}{C_{NT}^2} \sum_{j=1}^{r} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{\lambda}_{ij}^2 \right)^{\alpha} \\
& - \frac{1}{NT} \sum_{i=1}^{N} (X_i - \hat{F}^p \lambda_i^*)'(X_i - \hat{F}^p \lambda_i^*) - \frac{\gamma}{C_{NT}^2} \sum_{j=1}^{r} \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_{ij}^{*2} \right)^{\alpha} \qquad \text{(B.2)}
\end{aligned}$$

Now we can rewrite the right hand side of (B.2) as

$$\begin{aligned}
& \frac{1}{NT} \sum_{i=1}^{N} [(\hat{\lambda}_i - \lambda_i^*)' \hat{F}^{p\prime} \hat{F}^p (\hat{\lambda}_i - \lambda_i^*) - 2 u_i' \hat{F}^p (\hat{\lambda}_i - \lambda_i^*)] + \frac{\gamma}{C_{NT}^2} \sum_{j=1}^{r} \left[\left(\frac{1}{N} \sum_{i=1}^{N} \hat{\lambda}_{ij}^2 \right)^{\alpha} - \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_{ij}^{*2} \right)^{\alpha} \right] \\
= & \frac{1}{N} \sum_{i=1}^{N} \|\hat{\lambda}_i - \lambda_i^*\|^2 - \frac{2}{NT} \sum_{i=1}^{N} u_i' \hat{F}^p (\hat{\lambda}_i - \lambda_i^*) + \frac{\gamma}{C_{NT}^2} \sum_{j=1}^{r} \left[\left(\frac{1}{N} \sum_{i=1}^{N} \hat{\lambda}_{ij}^2 \right)^{\alpha} - \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_{ij}^{*2} \right)^{\alpha} \right] \\
= & v_1 + v_2 + v_3 \qquad \text{(B.3)}
\end{aligned}$$

where the equality uses the fact that $\hat{F}^{p\prime} \hat{F}^p / T = I_p$.

Now, we look at each of the three terms in (B.3). For the first term in (B.3), we obtain by the

definition of $S_{j,N,T}$:

$$v_1 = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 = \frac{1}{N}\|\hat{\Lambda} - \Lambda^*\|^2 \geq 2^{j-1}C_{NT}^{-2}$$

For the second term, by Cauchy-Schwarz inequality,

$$|v_2| \leq 2\left(\frac{1}{NT^2}\sum_{i=1}^{N}\|u_i'\hat{F}^p\|^2\right)^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^{n}\|\hat{\lambda}_i - \lambda_i^*\|^2\right)^{\frac{1}{2}} \tag{B.4}$$

By mean value theorem, v_3 in (B.3) reduces to:

$$\begin{aligned} v_3 &= \frac{\gamma}{C_{NT}^2}\sum_{j=1}^{r}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}_{ij}^2\right)^{\alpha} - \left(\frac{1}{N}\sum_{i=1}^{N}\lambda_{ij}^{*^2}\right)^{\alpha}\right] \\ &= \frac{\gamma\alpha}{C_{NT}^2}\sum_{j=1}^{r}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\check{\lambda}_{ij}^2\right)^{\alpha-1}\frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_{ij}^2 - \lambda_{ij}^{*^2})\right] \end{aligned}$$

where $N^{-1}\sum_{i=1}^{N}\check{\lambda}_{ij}^2$ is between $\frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}_{ij}^2$ and $\frac{1}{N}\sum_{i=1}^{N}\lambda_{ij}^{*^2}$ for $j = 1, ..., r$. Using Lemma 4, recall that H has a nonsingular limit H_0 by Lemma 1(ii) and that $\|\lambda_i^0\| \leq \bar{\lambda} < \infty$ for all i, so (A.7) implies that $N^{-1}|\sum_{i=1}^{N}(\hat{\lambda}_{ij}^2 - \lambda_{ij}^{*^2})|$ can be bounded by $[N^{-1}\sum_{i=1}^{N}(\hat{\lambda}_{ij} - \lambda_{ij}^*)^2]^{1/2}M_1$ for some $M_1 < \infty$ and sufficiently large N and T. There exists a constant $c_1 < \infty$ such that

$$|v_3| \leq \frac{\gamma c_1}{C_{NT}^2}\left(\frac{1}{N}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2\right)^{\frac{1}{2}} \leq \frac{\gamma c_1}{C_{NT}^2}\cdot\frac{2^{\frac{j}{2}}}{C_{NT}}.$$

Thus, on $S_{j,N,T}$

$$L(\hat{\Lambda}) - L(\Lambda^*) \geq -|v_2| + 2^{j-1}C_{NT}^{-2} - \frac{c_1\gamma}{C_{NT}^2}\left(\frac{2^{\frac{j}{2}}}{C_{NT}}\right)$$

It follows that

$$\begin{aligned} & P\left(\inf_{\hat{\Lambda}\in S_{j,N,T}}[L(\hat{\Lambda}) - L(\Lambda^*)] \leq 0,\ \frac{1}{NT^2}\sum_{i=1}^{N}\|u_i'\hat{F}^p\|^2 \leq \frac{M}{C_{NT}^2}\right) \\ \leq\ & P\left(\sup_{\hat{\Lambda}\in S_{j,N,T}}|v_2| \geq 2^{j-1}C_{NT}^{-2} - \frac{c_1\gamma}{C_{NT}^2}\left(\frac{2^{\frac{j}{2}}}{C_{NT}}\right),\ \frac{1}{NT^2}\sum_{i=1}^{N}\|u_i'\hat{F}^p\|^2 \leq \frac{M}{C_{NT}^2}\right) \\ \leq\ & \frac{2\sqrt{M}C_{NT}^{-2}\cdot 2^{\frac{j}{2}}}{2^{j-1}C_{NT}^{-2} - c_1\gamma 2^{\frac{j}{2}}C_{NT}^{-3}} = \frac{2\sqrt{M}}{2^{\frac{j}{2}-1} - c_1\gamma C_{NT}^{-1}} \end{aligned} \tag{B.5}$$

where we use Markov's inequality, (B.4), and the fact that $E(N^{-1}T^{-2}\sum_{i=1}^{N}\|u_i'\hat{F}^p\|^2) \leq MC_{NT}^{-2}$ and $N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 \leq 2^jC_{NT}^{-2}$ in the set $S_{j,N,T}$. By Assumption 6(i), $\gamma/C_{NT} \to 0$ as N and

$T \to \infty$. For sufficiently large N and T, $2^{\frac{j}{2}-1} - c_1\gamma C_{NT}^{-1} \geq 2^{\frac{j}{2}-2}$ for all $j \geq 4$. Thus,

$$\begin{aligned} & \sum_{j\geq M,\ 2^j \leq \varepsilon C_{NT}^2} P\left(\inf_{\hat{\Lambda}\in S_{j,N,T}} [L(\hat{\Lambda}) - L(\Lambda^*)] \leq 0,\ \frac{1}{NT^2}\sum_{i=1}^{N} \|u_i'\hat{F}^p\|^2 \leq \frac{M}{C_{NT}^2} \right) \\ \leq & \sum_{j\geq M} \frac{\sqrt{M}}{2^{\frac{j}{2}-3}} \leq \frac{2^{-(\frac{M}{2}-4)}}{2-\sqrt{2}}\sqrt{M} \end{aligned}$$

which converges to zero for $M \to \infty$. This completes the proof of (B.1). ■

Proof of Theorem 2:

Part (i): Let $\breve{\Lambda}$ $(N \times p)$ be an estimator for Λ^*. We will show that if the last $p-r$ columns of $\breve{\Lambda}$ are nonzero, then the objective function evaluated at $\breve{\Lambda}$ will be larger than the objective function evaluated at the correct estimate $\hat{\Lambda}$. Consider the ball

$$\{\breve{\Lambda} :\ N^{-1}\|\breve{\Lambda} - \Lambda^*\|^2 \leq C_{NT}^{-2} W\}, \tag{B.6}$$

where $0 < W < \infty$ is a constant, and $\breve{\Lambda}^{1:r}$ and $\breve{\Lambda}^{(r+1):p}$ denote the first r and last $p-r$ columns of $\breve{\Lambda}$, respectively. Theorem 1 has shown that for a sufficiently large W, $\hat{\Lambda}$ lies in the ball (B.6) with probability converging to 1, where $C_{NT}^2 = \min(N,T)$. To prove Theorem 2, it is sufficient to show that, for any $\breve{\Lambda}$ satisfying (B.6), if $\|\breve{\Lambda}^{(r+1):p}\| > 0$, then there exists $\tilde{\Lambda} \equiv [\breve{\Lambda}^{1:r} \vdots 0_{N\times(p-r)}]$ such that $L(\breve{\Lambda}) - L(\tilde{\Lambda}) > 0$ with probability converging to one. Namely, if $\breve{\Lambda}$ is the solution of (2.6), then it must follow that $P(\breve{\Lambda}^{(r+1):p} = 0) \to 1$.

Now, let $\breve{\lambda}_i$ and $\tilde{\lambda}_i$ denote the transpose of the i^{th} rows of $\breve{\Lambda}$ and $\tilde{\Lambda}$, respectively.

$$\begin{aligned} L(\breve{\Lambda}) - L(\tilde{\Lambda}) &= \frac{1}{NT}\sum_{i=1}^{N}(X_i - \hat{F}^p\breve{\lambda}_i)'(X_i - \hat{F}^p\breve{\lambda}_i) - \frac{1}{NT}\sum_{i=1}^{N}(X_i - \hat{F}^p\tilde{\lambda}_i)'(X_i - \hat{F}^p\tilde{\lambda}_i) \\ &\quad + \frac{\gamma}{C_{NT}^2}\sum_{j=r+1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\breve{\lambda}_{ij}^2\right)^{\alpha} \\ &= \frac{1}{NT}\sum_{i=1}^{N}(\breve{\lambda}_i'\hat{F}^{p\prime}\hat{F}^p\breve{\lambda}_i - \tilde{\lambda}_i'\hat{F}^{p\prime}\hat{F}^p\tilde{\lambda}_i) - \frac{2}{NT}\sum_{i=1}^{N}(\breve{\lambda}_i - \tilde{\lambda}_i)'\hat{F}^{p\prime}X_i \\ &\quad + \frac{\gamma}{C_{NT}^2}\sum_{j=r+1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\breve{\lambda}_{ij}^2\right)^{\alpha} = I + II + III \end{aligned}$$

Since $\hat{F}^{p\prime}\hat{F}^{p}/T = I_p$, the first term can be rewritten as

$$\begin{aligned} I &= \frac{1}{N}\sum_{i=1}^{N}(\breve{\lambda}_i'\breve{\lambda}_i - \tilde{\lambda}_i'\tilde{\lambda}_i) = \frac{1}{N}\sum_{i=1}^{N}\|\breve{\lambda}_i^{(r+1):p'}\|^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\|\breve{\lambda}_i^{(r+1):p} - \lambda_i^{*(r+1):p}\|^2 \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\|\breve{\lambda}_i - \lambda_i^*\|^2 = O_p(C_{NT}^{-2}) \end{aligned}$$

by (B.6). For the second term, by the definitions of $\breve{\lambda}_i$, $\tilde{\lambda}_i$ and $\hat{\lambda}_{i,OLS}^{(r+1):p}$, we have

$$\begin{aligned} \frac{1}{NT}\sum_{i=1}^{N}(\breve{\lambda}_i - \tilde{\lambda}_i)'\hat{F}^{p\prime}X_i &= \frac{1}{NT}\sum_{i=1}^{N}\breve{\lambda}_i^{(r+1):p'}\hat{F}^{(r+1):p'}X_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\breve{\lambda}_i^{(r+1):p'}\hat{\lambda}_{i,OLS}^{(r+1):p} \\ &\leq \left(\frac{1}{N}\sum_{i=1}^{N}\|\breve{\lambda}_i^{(r+1):p}\|^2\frac{1}{N}\sum_{i=1}^{N}\|\hat{\lambda}_{i,OLS}^{(r+1):p}\|^2\right)^{\frac{1}{2}} \\ &= O_p(C_{NT}^{-2}) \end{aligned}$$

where we use the fact that $N^{-1}\sum_{i=1}^{N}\|\breve{\lambda}_i^{(r+1):p}\|^2 = O_p(C_{NT}^{-2})$, which was proved in term I, and the fact that

$$N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_{i,OLS}^{(r+1):p}\|^2 = \sum_{j=r+1}^{p}v_j = V(r) - V(p) = O_p(C_{NT}^{-2})$$

by (A.2) and Lemma 4 of Bai and Ng (2002). For term III, note that

$$\sum_{j=r+1}^{p}\left(\frac{1}{N}\|\breve{\Lambda}^j\|^2\right)^{\alpha} \geq \left(\sum_{j=r+1}^{p}\frac{1}{N}\|\breve{\Lambda}^j\|^2\right)^{\alpha} = \left(\frac{1}{N}\|\breve{\Lambda}^{(r+1):p}\|^2\right)^{\alpha},$$

where we use Loeve's C_r inequality in (9.63) of Davidson (1994) and $\breve{\Lambda}^j$ denotes the j^{th} column of $\breve{\Lambda}$. It follows that by Assumption 6(ii) and (B.6)

$$III = \frac{\gamma}{C_{NT}^2}\sum_{j=r+1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\breve{\Lambda}_{ij}^2\right)^{\alpha} \geq \frac{\gamma}{C_{NT}^2}\cdot\left(\frac{1}{N}\|\breve{\Lambda}^{(r+1):p}\|^2\right)^{\alpha} = \frac{1}{C_{NT}^2}O_p\left(\frac{\gamma}{C_{NT}^{2\alpha}}\right) \equiv III^*$$

Thus, III^* converges to zero slower than I and II, which implies that $L(\breve{\Lambda}) > L(\tilde{\Lambda})$ with probability converging to 1. This is true since $C_{NT}^2 III \to \infty$ dominates the negative term $C_{NT}^2 II = O_p(1)$ in the limit.

Part (ii): The part proceeds using proof by contradiction. Suppose that the j^{th} column ($j < r$)

of $\hat{\Lambda}$ is zero with a positive probability as $N,\ T \to \infty$. It follows that

$$N^{-1}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i^*\|^2 \geq N^{-1}\sum_{i=1}^{N}|\hat{\lambda}_{ij} - \lambda_{ij}^*|^2 = N^{-1}\sum_{i=1}^{N}\lambda_{ij}^{*^2} \to_p c > 0 \tag{B.7}$$

because $\text{rank}(\Lambda^{*'}\Lambda^*/N) = \text{rank}(H^{-1}\Lambda'\Lambda H^{-1'}/N) = r$ as N and $T \to \infty$ by Assumption 2 and Lemma 1(ii). Also, since $N^{-1}\sum_{i=1}^{N}|\hat{\lambda}_{ij} - \lambda_{ij}^*|^2 = o_p(1)$ by Theorem 1 and $\hat{\lambda}_{ij} = 0$ for $i = 1, ..., N$, it follows that

$$N^{-1}\sum_{i=1}^{N}|\lambda_{ij}^*|^2 = o_p(1)$$

which contradicts (B.7). Hence, the j^{th} column ($j < r$) of $\hat{\Lambda}$ will be nonzero as N and $T \to \infty$.■

References

[1] Alessi, L., M. Barigozzi, and M. Capasso (2010), "Improved Penalization for Determining the Number of Factors in Approximate Factor Models," *Statistics and Probability Letters*, 80, 1806-1813.

[2] Amengual, D. and M. Watson (2007), "Consistent Estimation of the Number of Dynamic Factors in a Large N And T Panel," *Journal of Business and Economic Statistics*, 25(1): 91-96.

[3] Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-173.

[4] Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

[5] Bai, J. and S. Ng (2007), "Determining the Number of Primitive Shocks in Factor Models," *Journal of Business and Economic Statistics*, 25, 52-60.

[6] Bai, J. and S. Ng (2008), "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304-317.

[7] Bernanke, B.S., J. Boivin and P. Eliasz (2005), "Measuring the Effects of Monetary Policy: a Factor-augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics* 120, 387–422.

[8] Boivin, J. and M. Giannoni (2006), "DSGE Models in a Data-Rich Environment," *NBER Working Paper* No. 12772.

[9] Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, 24, 2350-2383.

[10] Breitung, J. and S. Eickmeier (2011), "Testing For Structural Breaks in Dynamic Factor Models," *Journal of Econometrics*, 163, 71–84.

[11] Caner, M. and K. Knight (2013), "An Alternative to Unit Root Tests: Bridge Estimators Differentiate between Nonstationary versus Stationary Models and Select Optimal Lag," *Journal of Statistical Planning and Inference*, 143, 691-715.

[12] Caner, M. and H. Zhang (2013), "Adaptive Elastic Net GMM with Diverging Number of Moments," *Journal of Business and Economics Statistics*, forthcoming.

[13] Chamberlain, G., and M. Rothschild (1983), "Arbitrage, Factor Structure, and Mean Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281-1304.

[14] Davidson, J. (1994), "Stochastic Limit Theory: An Introduction for Econometricians," Oxford University Press.

[15] De Mol, C., D. Giannone, and L. Reichlin (2008), "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components," *Journal of Econometrics*, 146, 318-328.

[16] Donoho D., and I. Johnstone (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425-455.

[17] Fan, J. and R. Li (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

[18] Forni, M. and L. Gambetti (2010), "The Dynamic Effects of Monetary Policy: A Structural Factor Model Approach," *Journal of Monetary Economics*, 57(2), 203-216.

[19] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The Generalized Dynamic Factor Model: Identification and Estimation," *Review of Economics Statistics*, 82, 540-554.

[20] Forni, M. and M. Lippi (1997), "Agregation and the Microfoundations of Dynamic Macroeconomics," Oxford, U.K.: Oxford University Press.

[21] Giannone, D., L. Reichlin, and L. Sala (2004), "Monetary Policy in Real Time," *NBER Macroeconomics Annual*, 2004.

[22] Gregory, A. and A. Head (1999), "Common and Country-Specific Fluctuations in Productivity, Investment, and the Current Account," *Journal of Monetary Economics*, 44, 423-452.

[23] Han, X. (2012), "Determining the Number of Factors with Potentially Strong Cross-sectional Correlation in Idiosyncratic Shocks," manuscript, North Carolina State University.

[24] Hallin, M. and Liska R. (2007), "The Generalized Dynamic Factor Model: Determining the Number of Factors," *Journal of the American Statistical Association*, 102, 603-617.

[25] Hirose, K. and S. Konishi (2012), "Variable Selection via the Weighted Group LASSO for Factor Analysis Models," *the Canadian Journal of Statistics*, 40(2), 345-361.

[26] Huang, J., S. Ma, H. Xie, and C. Zhang (2009), " A Group Bridge Approach for Variable Selection," *Biometrika*, 96, 339-355.

[27] Huang, J., J. Horowitz, and S. Ma (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *Annals of Statistics*, 36, 587-613.

[28] Knight. K. and W. Fu (2000), "Asymptotics for Lasso Type Estimators," *Annals of Statistics*, 28, 1356-1378.

[29] Kapetanios, G. (2010), "A Testing Procedure for Determining the Number of Factors in Approximate Factor Models with Large Datasets," *Journal of Business and Economic Statistics,* 28(3), 397-409.

[30] Lewbel, A. (1991), "The Rank of Demand Systems: Theory and Nonparametric Estimation," *Econometrica*, 59, 711-730.

[31] Luo, R., H. Wang, and C.L. Tsai (2009), "Contour Projected Dimension Reduction", *The Annals of Statistics*, 37, 3743-3778.

[32] Onatski, A. (2009), "Testing hypotheses about the number of factors in large factor models," *Econometrica*, 77, 1447-1479.

[33] Onatski, A. (2010), "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *The Review of Economics and Statistics*, 92(4), 1004-1016.

[34] Sargent, T.J. and C.A.Sims (1977), "Business Cycle Modelling without Pretending to Have Too Much a-priori Economic Theory," in: Sims et al., eds., New Methods in Business Cycle Research (Federal Reserve Bank of Minneapolis, Minneapolis).

[35] Seung, A. and Horenstein A. (2013), "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81(3), 1203-1227.

[36] Stock, J. and M. Watson (1989), "New Indexes of Coincident and Leading Economic Indications," in NBER Macroeconomics Annual 1989 ed. by O.J. Blanchard, and S.Fischer. Cambridge: MIT press.

[37] Stock, J. and M. Watson (1999), "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293-335.

[38] Stock, J. and M. Watson (2002), "Macroeconomic forecasting using diffusion indexes," *Journal of Business and Economic Statistics* 20, 147-162.

[39] Stock, J. and M. Watson (2005), "Implications of Dynamic Factor Models for VAR Analysis," NBER Working Paper, 11647.

[40] Stock, J. and M. Watson (2009), "Forecasting in Dynamic Factor Models Subject to Structural Instability," in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, Jennifer Castle and Neil Shephard (eds), Oxford: Oxford University Press.

[41] Uhlig, H. (2009), "Macroeconomic Dynamics in the Euro area. Discussion by Harald Uhlig," in NBER Macroeconomics Annual 23, ed, by D. Acemoglu, K. Rogoff, M.Woodford.

[42] Wang, H. (2012), "Factor Profiled Sure Independence Screening," *Biometrika*, 99, 15-28.

[43] Wang, H., Li, B., and Leng, C. (2009), "Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters," *Journal of the Royal Statistical Society*, 71, 671-683.

[44] Yuan, M. and Y. Lin (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, 68, 49-67.

[45] Zou, H. and H. Zhang (2009), "On the Adaptive Elastic Net With a Diverging Number of Parameters," *Annals of Statistics*, 37, 1733-1751.

Table 1: No cross-sectional or serial correlation, $\beta = 0$ and $\rho = 0$

r	N	T	Bridge	IC_{p1}	$IC^T_{1;n}$	ED	ER	GR
	50	50	1.00 (0|0)	1.00 (0|0)	1.01 (1|0)	1.02 (2|0)	1.00 (0|0)	1.00 (0|0)
	100	50	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.02 (1|0)	1.00 (0|0)	1.00 (0|0)
1	100	100	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.01 (1|0)	1.00 (0|0)	1.00 (0|0)
	100	200	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.03 (1|0)	1.00 (0|0)	1.00 (0|0)
	200	100	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.01 (1|0)	1.00 (0|0)	1.00 (0|0)
	200	200	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.02 (2|0)	1.00 (0|0)	1.00 (0|0)
	50	50	2.83 (0|13)	2.99 (0|1)	2.60 (1|25)	3.02 (1|0)	2.88 (0|8)	2.94 (0|4)
	100	50	2.94 (0|5)	3.00 (0|0)	2.87 (0|8)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
3	100	100	3.00 (0|0)	3.00 (0|0)	2.99 (0|1)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	100	200	3.00 (0|0)	3.00 (0|0)	2.99 (0|0)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	200	100	3.00 (0|0)	3.00 (0|0)	2.95 (0|4)	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)
	200	200	3.00 (0|0)	3.00 (0|0)	2.45 (0|37)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	50	50	1.81 (0|98)	4.54 (0|40)	1.43 (0|94)	4.06 (1|28)	3.15 (0|59)	3.75 (0|45)
	100	50	1.87 (0|98)	4.90 (0|10)	1.41 (0|94)	4.99 (0|1)	4.53 (0|15)	4.77 (0|8)
5	100	100	4.80 (0|9)	5.00 (0|0)	4.52 (0|20)	5.00 (0|0)	4.97 (0|1)	5.00 (0|0)
	100	200	5.00 (0|0)	5.00 (0|0)	4.98 (0|1)	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)
	200	100	5.00 (0|0)	5.00 (0|0)	4.13 (0|30)	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)
	200	200	5.00 (0|0)	5.00 (0|0)	3.29 (0|77)	5.01 (0|0)	5.00 (0|0)	5.00 (0|0)

Note: The error terms are generated using DGP E1. β controls the cross-sectional correlation, and ρ controls the serial correlation of the errors. The factors explain 50% of the variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, Onatski's (2010) ED, and Seung and Horenstein's (2013) ER and GR. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 2: Cross-sectional correlation only, $\beta = 0.2$ and $\rho = 0$

r	N	T	Bridge	IC_{p1}	$IC^T_{1;n}$	ED	ER	GR
	50	50	1.41 (28|0)	7.18 (100|0)	1.63 (44|0)	1.73 (17|0)	1.00 (0|0)	1.01 (0|0)
	100	50	1.00 (0|0)	5.06 (96|0)	1.13 (9|0)	1.11 (6|0)	1.00 (0|0)	1.00 (0|0)
	100	100	1.00 (0|0)	9.16 (100|0)	2.08 (44|0)	1.04 (3|0)	1.00 (0|0)	1.00 (0|0)
1	100	200	1.00 (0|0)	10.00 (100|0)	2.63 (40|0)	1.02 (1|0)	1.00 (0|0)	1.00 (0|0)
	200	100	1.00 (0|0)	1.41 (34|0)	1.08 (4|0)	1.04 (4|0)	1.00 (0|0)	1.00 (0|0)
	200	200	1.00 (0|0)	7.21 (100|0)	1.01 (1|0)	1.02 (1|0)	1.00 (0|0)	1.00 (0|0)
	50	50	3.21 (31|12)	8.92 (100|0)	1.95 (12|65)	2.08 (11|60)	2.38 (9|54)	3.11 (21|41)
	100	50	2.94 (0|6)	6.85 (97|0)	2.29 (4|43)	2.95 (3|7)	2.70 (0|19)	2.80 (0|13)
3	100	100	3.00 (0|0)	9.79 (100|0)	3.60 (40|2)	3.01 (1|0)	2.93 (0|5)	2.98 (0|2)
	100	200	3.00 (0|0)	10.00 (100|0)	3.78 (33|0)	3.00 (0|0)	2.97 (0|1)	2.99 (0|1)
	200	100	3.00 (0|0)	3.46 (37|0)	3.04 (5|0)	3.02 (1|0)	3.00 (0|0)	3.00 (0|0)
	200	200	3.00 (0|0)	8.31 (100|0)	3.00 (0|0)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	50	50	2.62 (1|90)	9.73 (100|0)	1.21 (0|99)	0.99 (3|96)	2.14 (11|86)	2.86 (20|76)
	100	50	2.35 (0|94)	8.39 (97|0)	1.19 (0|98)	1.55 (1|83)	1.88 (1|84)	2.42 (3|74)
5	100	100	4.69 (3|20)	9.96 (100|0)	3.87 (19|52)	2.66 (0|56)	2.36 (0|71)	2.80 (1|61)
	100	200	5.32 (28|2)	10.00 (100|0)	4.87 (30|39)	4.12 (0|21)	2.72 (0|59)	3.25 (0|45)
	200	100	4.99 (0|1)	5.50 (38|0)	3.48 (0|55)	5.02 (1|0)	4.68 (0|8)	4.89 (0|3)
	200	200	5.00 (0|0)	9.25 (100|0)	4.93 (0|4)	5.00 (0|0)	4.98 (0|0)	5.00 (0|0)

Note: The error terms are generated using DGP E1. β controls the cross-sectional correlation, and ρ controls the serial correlation of the errors. The factors explain 50% of the variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, Onatski's (2010) ED, and Seung and Horenstein's (2013) ER and GR. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 3: Serial correlation only, $\beta = 0$ and $\rho = 0.7$

r	N	T	Bridge	IC_{p1}	$IC^T_{1;n}$	ED	ER	GR
	50	50	1.09 (9|0)	9.03 (100|0)	1.33 (24|0)	1.19 (10|0)	1.00 (0|0)	1.00 (0|0)
	100	50	1.02 (2|0)	9.52 (100|0)	1.21 (15|0)	1.13 (5|0)	1.00 (0|0)	1.00 (0|0)
	100	100	1.00 (0|0)	6.90 (100|0)	1.34 (16|0)	1.05 (4|0)	1.00 (0|0)	1.00 (0|0)
1	100	200	1.00 (0|0)	1.13 (12|0)	1.00 (0|0)	1.04 (3|0)	1.00 (0|0)	1.00 (0|0)
	200	100	1.00 (0|0)	9.78 (100|0)	1.11 (5|0)	1.02 (2|0)	1.00 (0|0)	1.00 (0|0)
	200	200	1.00 (0|0)	1.91 (58|0)	1.00 (0|0)	1.03 (2|0)	1.00 (0|0)	1.00 (0|0)
	50	50	3.07 (17|9)	9.69 (100|0)	2.40 (8|37)	2.18 (5|41)	2.43 (3|40)	2.72 (8|31)
	100	50	3.03 (7|4)	9.92 (100|0)	2.61 (6|24)	2.61 (3|21)	2.58 (1|27)	2.76 (3|20)
3	100	100	3.00 (0|0)	8.47 (100|0)	3.04 (5|2)	3.03 (1|0)	2.98 (0|1)	2.99 (0|0)
	100	200	3.00 (0|0)	3.15 (14|0)	3.00 (0|0)	3.02 (1|0)	3.00 (0|0)	3.00 (0|0)
	200	100	3.00 (0|0)	9.96 (100|0)	3.02 (3|1)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	200	200	3.00 (0|0)	3.91 (59|0)	3.00 (0|0)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	50	50	2.74 (0|87)	9.92 (100|0)	1.32 (1|93)	0.89 (2|96)	2.20 (7|86)	2.79 (13|78)
	100	50	3.01 (0|82)	10.00 (100|0)	1.19 (1|93)	0.88 (1|95)	1.84 (4|86)	2.51 (9|77)
5	100	100	4.82 (1|11)	9.35 (100|0)	3.42 (1|56)	3.92 (1|26)	3.38 (0|45)	3.86 (0|33)
	100	200	5.00 (0|0)	5.14 (13|0)	4.91 (0|4)	5.01 (1|0)	4.82 (0|5)	4.96 (0|1)
	200	100	5.00 (0|1)	9.99 (100|0)	3.59 (0|46)	4.88 (0|3)	4.11 (0|23)	4.53 (0|12)
	200	200	5.00 (0|0)	5.91 (60|0)	4.98 (0|1)	5.01 (0|0)	5.00 (0|0)	5.00 (0|0)

Note: The error terms are generated using DGP E1. β controls the cross-sectional correlation, and ρ controls the serial correlation of the errors. The factors explain 50% of the variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, Onatski's (2010) ED, and Seung and Horenstein's (2013) ER and GR. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 4: Both cross-sectional and serial correlation, $\beta = 0.1$ and $\rho = 0.6$

r	N	T	Bridge	IC_{p1}	$IC^T_{1;n}$	ED	ER	GR
	50	50	1.11 (9|0)	7.90 (100|0)	1.37 (26|0)	1.23 (13|0)	1.00 (0|0)	1.00 (0|0)
	100	50	1.00 (0|0)	7.44 (100|0)	1.15 (12|0)	1.13 (9|0)	1.00 (0|0)	1.00 (0|0)
	100	100	1.00 (0|0)	3.96 (95|0)	1.37 (19|0)	1.07 (5|0)	1.00 (0|0)	1.00 (0|0)
1	100	200	1.00 (0|0)	2.61 (84|0)	1.06 (3|0)	1.08 (5|0)	1.00 (0|0)	1.00 (0|0)
	200	100	1.00 (0|0)	3.20 (89|0)	1.05 (3|0)	1.06 (4|0)	1.00 (0|0)	1.00 (0|0)
	200	200	1.00 (0|0)	1.75 (56|0)	1.00 (0|0)	1.05 (4|0)	1.00 (0|0)	1.00 (0|0)
	50	50	3.06 (18|11)	9.04 (100|0)	2.40 (8|39)	2.32 (8|37)	2.39 (4|42)	2.68 (9|32)
	100	50	2.97 (1|4)	9.02 (100|0)	2.71 (5|18)	2.87 (5|10)	2.65 (0|23)	2.80 (1|14)
3	100	100	3.00 (0|0)	5.98 (96|0)	3.09 (9|1)	3.04 (3|0)	2.98 (0|1)	2.99 (0|1)
	100	200	3.00 (0|0)	4.51 (84|0)	3.04 (4|0)	3.04 (3|0)	3.00 (0|0)	3.00 (0|0)
	200	100	3.00 (0|0)	5.37 (93|0)	3.00 (1|1)	3.02 (2|0)	3.00 (0|0)	3.00 (0|0)
	200	200	3.00 (0|0)	3.75 (56|0)	3.00 (0|0)	3.02 (1|0)	3.00 (0|0)	3.00 (0|0)
	50	50	2.65 (0|89)	9.58 (100|0)	1.35 (1|94)	1.02 (2|94)	2.08 (6|87)	2.59 (11|79)
	100	50	2.75 (0|86)	9.68 (100|0)	1.42 (0|89)	1.32 (1|86)	1.98 (2|82)	2.50 (4|74)
5	100	100	4.81 (1|12)	7.76 (96|0)	3.57 (1|54)	3.87 (2|28)	3.24 (0|47)	3.80 (1|34)
	100	200	5.03 (4|1)	6.50 (83|0)	4.88 (0|7)	4.95 (2|2)	4.45 (0|14)	4.71 (0|8)
	200	100	4.99 (0|1)	7.45 (92|0)	4.43 (0|21)	4.99 (1|1)	4.50 (0|13)	4.81 (0|5)
	200	200	5.00 (0|0)	5.74 (56|0)	4.99 (0|1)	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)

Note: The error terms are generated using DGP E1. β controls the cross-sectional correlation, and ρ controls the serial correlation of the errors. The factors explain 50% of the variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, Onatski's (2010) ED, and Seung and Horenstein's (2013) ER and GR. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 5: Conditionally heteroskedastic errors

r	N	T	Bridge	IC_{p1}	$IC^T_{1;n}$	ED	ER	GR
	50	50	1.22 (17|0)	8.34 (97|0)	1.34 (25|0)	1.28 (20|0)	1.00 (0|0)	1.01 (1|0)
	100	50	1.17 (15|0)	8.58 (98|0)	1.30 (22|0)	1.36 (23|0)	1.00 (0|0)	1.00 (0|0)
	100	100	1.04 (4|0)	4.22 (89|0)	1.50 (29|0)	1.29 (21|0)	1.00 (0|0)	1.00 (0|0)
1	100	200	1.00 (0|0)	1.68 (48|0)	1.13 (8|0)	1.19 (16|0)	1.00 (0|0)	1.00 (0|0)
	200	100	1.06 (5|0)	5.79 (96|0)	1.32 (21|0)	1.32 (22|0)	1.00 (0|0)	1.00 (0|0)
	200	200	1.01 (1|0)	2.94 (83|0)	1.07 (6|0)	1.29 (22|0)	1.00 (0|0)	1.00 (0|0)
	50	50	2.82 (0|16)	3.30 (23|0)	2.18 (3|48)	2.93 (5|8)	2.66 (0|23)	2.77 (1|16)
	100	50	2.96 (0|4)	3.30 (23|0)	2.54 (3|26)	3.08 (8|1)	2.89 (0|8)	2.94 (1|5)
3	100	100	3.00 (0|0)	3.02 (2|0)	3.07 (8|1)	3.06 (5|0)	3.00 (0|0)	3.00 (0|0)
	100	200	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)	3.04 (4|0)	3.00 (0|0)	3.00 (0|0)
	200	100	3.00 (0|0)	3.07 (7|0)	3.01 (1|0)	3.10 (9|0)	3.00 (0|0)	3.00 (0|0)
	200	200	3.00 (0|0)	3.00 (0|0)	2.99 (0|1)	3.09 (8|0)	3.00 (0|0)	3.00 (0|0)
	50	50	1.57 (0|100)	4.61 (3|37)	1.31 (0|97)	2.25 (1|69)	2.31 (0|79)	2.96 (1|67)
	100	50	1.62 (0|100)	4.94 (1|7)	1.39 (0|93)	4.05 (1|24)	3.34 (0|48)	3.98 (1|33)
5	100	100	4.53 (0|19)	5.00 (0|0)	4.08 (0|36)	5.01 (2|0)	4.75 (0|7)	4.89 (0|3)
	100	200	5.00 (0|0)	5.00 (0|0)	4.92 (0|4)	5.03 (2|0)	5.00 (0|0)	5.00 (0|0)
	200	100	4.98 (0|1)	5.00 (0|0)	4.63 (0|13)	5.04 (3|0)	4.99 (0|0)	5.00 (0|0)
	200	200	5.00 (0|0)	5.00 (0|0)	4.96 (0|2)	5.04 (4|0)	5.00 (0|0)	5.00 (0|0)

Note: The conditionally heteroskedastic error terms are generated using DGP E2. The factors explain 50% of the variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, Onatski's (2010) ED, and Seung and Horenstein's (2013) ER and GR. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 6: Weaker factor structure, $\beta = 0.1$, $\rho = 0.6$ and $R^2 = 33.3\%$

r	N	T	Bridge	IC_{p1}	$IC^T_{1;n}$	ED	ER	GR
	50	50	1.10 (10|0)	6.80 (99|0)	1.31 (24|0)	1.23 (14|0)	1.00 (0|0)	1.00 (0|0)
	100	50	1.00 (0|0)	6.76 (100|0)	1.17 (14|0)	1.13 (8|0)	1.00 (0|0)	1.00 (0|0)
	100	100	1.00 (0|0)	3.26 (91|0)	1.44 (19|0)	1.07 (5|0)	1.00 (0|0)	1.00 (0|0)
1	100	200	1.00 (0|0)	2.09 (71|0)	1.12 (4|0)	1.09 (5|0)	1.00 (0|0)	1.00 (0|0)
	200	100	1.00 (0|0)	2.66 (82|0)	1.08 (4|0)	1.04 (4|0)	1.00 (0|0)	1.00 (0|0)
	200	200	1.00 (0|0)	1.45 (39|0)	1.00 (0|0)	1.04 (3|0)	1.00 (0|0)	1.00 (0|0)
	50	50	1.85 (5|73)	8.22 (98|0)	1.40 (4|81)	1.10 (5|83)	1.66 (7|77)	2.08 (14|67)
	100	50	1.80 (0|73)	8.38 (99|0)	1.48 (2|75)	1.44 (3|70)	1.50 (1|76)	1.88 (5|65)
3	100	100	2.78 (0|16)	5.22 (90|0)	2.93 (7|14)	2.78 (4|13)	2.41 (0|33)	2.59 (0|24)
	100	200	2.99 (0|1)	4.01 (68|0)	3.07 (4|1)	3.04 (3|0)	2.84 (0|9)	2.92 (0|5)
	200	100	2.99 (0|1)	4.79 (83|0)	2.88 (2|10)	3.02 (2|0)	2.88 (0|7)	2.95 (0|3)
	200	200	3.00 (0|0)	3.42 (35|0)	3.00 (0|0)	3.01 (1|0)	3.00 (0|0)	3.00 (0|0)
	50	50	0.81 (0|100)	8.43 (89|5)	0.97 (0|99)	0.69 (1|99)	0.82 (1|99)	1.12 (3|96)
	100	50	0.66 (0|100)	8.83 (96|2)	0.83 (0|99)	0.75 (0|99)	0.50 (0|100)	0.61 (0|100)
5	100	100	0.95 (0|100)	6.66 (81|2)	2.12 (3|87)	1.07 (1|95)	0.78 (0|99)	0.98 (0|98)
	100	200	1.44 (0|98)	5.83 (61|0)	3.83 (2|51)	2.22 (1|67)	0.93 (0|95)	1.30 (0|89)
	200	100	1.15 (0|99)	6.68 (81|0)	2.21 (0|85)	2.43 (0|62)	0.84 (0|96)	1.24 (0|89)
	200	200	4.61 (0|11)	5.46 (39|0)	4.00 (0|37)	4.95 (1|2)	3.45 (0|38)	4.02 (0|25)

Note: The error terms are generated using DGP E1. β controls the cross-sectional correlation, and ρ controls the serial correlation of the errors. The factors explain 33.3% of the variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, Onatski's (2010) ED, and Seung and Horenstein's (2013) ER and GR. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 7: Stability to the choice of α ($N = T = 100$)

β	ρ	α	$r = 0$	$r = 1$	$r = 3$	$r = 5$
0	0	0.10	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.88 (0|9)
0	0	0.15	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.85 (0|11)
0	0	0.25	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.80 (0|9)
0	0	0.35	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.90 (0|3)
0	0	0.40	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.81 (0|5)
0.1	0.6	0.10	0.26 (22|0)	1.04 (4|0)	3.00 (0|0)	4.93 (1|7)
0.1	0.6	0.15	0.05 (5|0)	1.00 (0|0)	3.00 (0|0)	4.87 (0|11)
0.1	0.6	0.25	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.81 (1|12)
0.1	0.6	0.35	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	4.58 (2|15)
0.1	0.6	0.40	0.00 (0|0)	1.00 (0|0)	3.00 (0|0)	3.36 (8|46)

Note: The error terms are generated using DGP E1. β controls the cross-sectional correlation, and ρ controls the serial correlation of the errors. The factors explain 50% of the variation in the data when $r > 0$. The upper bound of the number of factors is set equal to 10, and r is the true number of factors. The numbers outside the parentheses are the means of different estimators over 1000 replications, and the numbers in $(a \mid b)$ mean that $a\%$ of the replications produce overestimation, $b\%$ of the replications produce underestimation, and $1 - a\% - b\%$ of the replications produce the correct estimation of the number of factors.

Table 8: Relative Mean Square Forecast Errors

k	1	2	3	4	5	6	7	8	9	10
mean	0.9727	0.9713	0.9743	0.9724	0.9811	0.9845	0.9906	1.0027	1.0083	1.0158
median	0.9844	0.9765	0.9727	0.9734	0.9811	0.9798	0.9837	0.9851	0.9972	1.0016

Note: The benchmark is the autoregressive model. k denotes the number of factors used as predictors.