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

December 17, 2012

Abstract

This paper proposes a Group Bridge estimator to select the correct number of factors in approximate factor models. It contributes to both shrinkage estimation and factor model literature. We extend 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 probability approaching one. Hence, it provides a consistent estimate for the number of factors. We also propose an algorithm for the new estimator and Monte Carlo experiments show that our algorithm converges reasonably fast and that our estimator has very good performance in small samples. An empirical example based on a commonly used US macroeconomic data set is provided in the paper as well.

Key words: bridge estimation, common factors, selection consistency

*We thank the participants of High Dimension Reduction (December 2010) Conference in London, Panel Data Conference (July 2011) in Montreal, and Robin Sickles, James Stock for their comments about the paper.

†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 gained importance in both finance and macroeconomics. They 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 variations of major macroeconomic variables (Sargent and Sims, 1977; Stock and Watson, 1989; Giannone, Reichlin, and Sala, 2004). Cross country differences can be explained through factors models by Gregory and Head (1999) and Forni, Hallin, Lippi, and Reichlin (2000). Other applications of factor models include, for example, forecasting (Stock and Watson, 2002), dynamic stochastic general equilibrium macroeconomic models (Bovine and Giannoni, 2006), structural VAR analysis (Bernanke, Boivin and Eliasz, 2005; Stock and Watson, 2005; Forni and Gembetti, 2010), and consumer demand and micro behaviors (Forni and Lippi, 1997; Lewbel, 1991).

Since 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), a wrong 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 will be misleading and provide wrong policy suggestions. Recent papers by Stock and Watson (2009) and Breitung and Eickermeier (2011) also found that the number of factors in subsamples could be applied to detect structural changes in factor models. This implies that correctly determining the number of factors can help us avoid using samples with structural breaks in forecasting.

In this paper, we propose a Group Bridge estimator to determine the number of factors in approximate factor models. One advantage of Bridge estimator is that, compared to conventional information criteria, Bridge can conduct both estimation and model selection in one step. Hence, it avoids the instability caused by subset selection or stepwise deletion (Breiman, 1996). The instability of any model selection tool means when new data are added up, 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 nice properties in the high dimensional factor models.

This paper builds a connection between shrinkage estimation and factor models. Despite the large literature in 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 the forecasts based on factor models. However, their application of shrinkage estimation is still restricted to a single equation instead of large panels. This paper contributes

to these two strands of literature in the following ways. First, we extend the conventional Bridge estimator from a single equation (such as Knight and Fu, 2000) 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. Also, 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 existing literature. It is common that many papers allow the number of parameters to increase with the sample size (for example, Zou and Zhang, 2009), 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 section dimension.

Second, 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 nonsingular rotation) of relevant factors and shrink the loadings of irrelevant factors to zero with 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 the cells in the zero factor loading vectors are shrunk simultaneously. Thus, our estimation has a group structure.

Additionally, most existing methods in the literature are information criteria that are roughly equivalent to finding some threshold to distinguish large and small eigenvalues of the data covariance matrix (for example, Bai and Ng, 2002, 2007; Hallin and Liska, 2007; Onatski, 2009, 2010; among others). Since the empirical distribution of the eigenvalues is determined by how the error terms are generated, the performance of this type of estimators can be largely affected by the unknown correlation structure in the error terms. For example, it is commonly noted and confirmed by our simulations that Bai and Ng's (2002) estimators tend to overestimate the number of factors when the errors are correlated in cross section or time dimension. Hallin and Liska (2007)'s and Onatski's (2010) estimators are more robust to correlated error terms, but simulations show that they tend to underestimate the number of factors in small samples when the correlation is relatively strong. In contrast, we solve the problem from another angle: the Group Bridge estimator directly penalizes the factor loading matrix. Simulations show that our estimator is robust to cross-sectional and serial correlation in the error terms and outperforms existing methods in certain cases.

Section 2 introduces the model, assumptions, and provides 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 covers all the proofs.

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. The model can be rewritten 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 to estimate unknown factors, and thereafter we use Bridge estimation to get the correct number of factors. The conventional Principal Component Analysis (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 $\hat{\Lambda}_{OLS}^k = X'\hat{F}^k \left(\hat{F}^{k\prime}\hat{F}^k\right)^{-1} = X'\hat{F}^k/T$ (since eigenvectors are orthonormal, $\hat{F}^{k\prime}\hat{F}^k/T = I_k$).

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 the conventional Bridge estimator such as 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 $\sum_{i=1}^{N} \sum_{j=1}^{p} |\lambda_{ij}|^{\delta}$ $(0 < \delta < 1)$. Since approximate factor models allow weak cross sectional correlation in the idiosyncratic errors, the last $p - r$ columns of $\hat{F}^p$ will capture nontrivial fractions of X_i's variations for some i's. If we regress such an X_i on the $(p-r)^{th}$ columns of $\hat{F}^p$, the estimated coefficients will converge to some nonzero constants at the limit, and it will be impossible for $\sum_{i=1}^{N} \sum_{j=1}^{p} |\lambda_{ij}|^{\delta}$ to shrink these coefficients to zero. As a result, the last $p - r$ columns of Λ will contain many zero and a few nonzero estimates. Hence, the conventional Bridge penalty will cause overestimation if we use the number of nonzero columns in Λ as an estimator for r.

In order 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 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 the single equation Bridge estimation.

Huang *et al* (2009) also propose a group Bridge estimator in the literature of 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 corresponds 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 that meaningful under the current setup. Moreover, since our penalty term uses square rather than absolute values, the computation is much easier than

that of the penalty by Huang *et al* (2009).

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.*
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. 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 the assumption on the tuning parameter. Compared with 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 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, since 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 remarkable that the there is some 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 still hardly discussed in both factor model and random matrix theory literature to the best of our knowledge.[1] 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[2] converging to some nonsingular limit H_0. Since the factors are estimated up to the nonsingular rotation matrix H, we rewrite (2.2) as

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

[1] 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. Since 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.

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

Replacing the unobserved F^0H by 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

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

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 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 in the literature.

Given 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 $\hat{\Lambda}$ in (2.5) converges to Λ^*. 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 remarkable 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 in 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 on 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 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^*. It also implies that the Bridge estimation may be valuable in other applications of high dimensional factor models. In this paper's setup, the last $p-r$ columns of Λ^* are zeros. In a scenario other than determining r, the Λ^* can contain a zero submatrix in a different manner. For example, Stock and Watson (2005) investigate a factor augmented VAR model in which a fraction of the factor loading matrix is zero by the identification conditions on the structural shocks. With a proper modification of the penalty term, we expect that Bridge estimation can be applied in Stock and Watson's (2005) model. The result in Theorem 2 will be useful to check if their high dimension identification restrictions are satisfied or not.

Based on Theorem 2, we obtain the result that 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 factor 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 p factors via PCA in (2.4). Next, we show how to solve the optimization in (2.5) and (2.6). Since 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 $\mathscr{T} = 5\times10^{-4}$ 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 $g_1 = (X - \hat{F}\hat{\Lambda}^{(m)})'\hat{F}^p/N$ 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 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 similar to that of HHM (2008). The main modifications are the following: we use the OLS estimate instead of zero as the initial value to accelerate the algorithm. Also, the way to compute the gradient in step (2) is different from HHM (2008) but similar to Fan and Li (2001). The estimated loading will remain unchanged after being shrunk to zero. This modification substantially accelerates the convergence in our high dimensional setup. Additionally, note in step (4) that we gradually decrease the increment as m increases by dividing $1 + m/750$. This ensures the convergence of the algorithm. A larger constant than 750 will slow down the convergence but improve the accuracy. 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.

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, and simulations (see Table 7) show that our

estimator is very robust to the choice of α. The tuning parameter γ is set equal to $\phi \cdot [\min(N,T)]^{0.45}$ since Assumption 6 requires that $\gamma/C_{NT} \to 0$. The constant ϕ is varying 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 accelerates our algorithm. We follow Wang *et al*'s (2009) method to determine the optimal choice of γ. The optimal γ is the one that minimizes the following criterion:

$$\log[V(k(\gamma), \hat{F}^{k(\gamma)})] + k(\gamma)\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)\cdot \ln[\ln(N)]$$

where $k(\gamma)$ denotes the fact that the estimated loading matrix 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*'s (2009).

3 Simulations

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

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 PC_{p1} and IC_{p1} estimators are defined as the k that minimizes the following criteria:

$$PC_{p1}(k) = V(k, \hat{F}^k) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)$$

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

These two criteria can be considered as the analogs of the conventional BIC in the factor model context. They penalize the integer k to prevent the model from overfitting.

Hallin and Liska's (2007) estimator is developed 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 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 (2002) show that $[0, \bar{\mu}]$ 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 $IC^T_{1;n}$ 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 $\rho_k(.)$ denote 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[3], so it is also a robust estimator to correlated error terms. Thus, it is expected that competing with $IC^T_{1:n}$ and ED 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 tuning parameter follow section 2.3. We consider the data generating process (DGP) similar to that of Bai and Ng (2002):

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

where the factors F_{tj}'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 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 of Bai and Ng (2002). The parameters β and ρ control the cross-sectional and serial correlation 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 of Breitung and Eickmeier (2011). Note that $E(\lambda_{ij}F_{tj})^2 = 5/4$ 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% variation in the data. In experiment 6,

[3]See Onatski (2010) for the details about the computation of ζ.

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 for 500 times. The upper bound of 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 number of factors, while 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 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. However, as N and T increase, our estimator can detect the correct number of factors almost 100% of the times. 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 both cross-sectional and serial correlations coexist ($\beta = 0.1$ and $\rho = 0.6$). The results are reported in Tables 2 – 4, respectively. The results of these three experiments demonstrate very similar patterns. Both IC_{p1} and PC_{p1} of Bai and Ng (2002) tend to overestimate the number of factors. When $r = 1$, the estimates by IC_{p1} and PC_{p1} change dramatically as more data are introduced. This confirms Breiman's (1996) finding that integer based penalty and selection are unstable. Both $IC^T_{1;n}$ and ED tend to underestimate the number of factors when $r = 3$ or 5 in small samples (N or $T = 50$). In contrast, our estimator is more accurate than $IC^T_{1;n}$ and ED especially when $r = 5$ or when errors are serially correlated (see Tables 3 and 4).

In our fifth experiment, we consider conditionally heteroskedastic errors using DGP E2. The results are reported in Table 5. When $r = 3$ or 5, IC_{p1} performs much better than PC_{p1}, but it substantially overestimate 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, neither ED nor our estimator dominates each other: our estimator is more accurate when $r = 1$, while ED is more accurate in small samples ($T = 50$) when $r = 5$.

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 variation of 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) estimators are still severely upward biased, but IC_{p1} performs better than PC_{p1} when $N = T = 200$. In addition, our estimator tends to perform better than $IC^T_{1;n}$ and ED when $r = 1$ or 3. When $r = 5$, $IC^T_{1;n}$ seems to be more robust to the weak factor structure than our estimator and ED except for the case with $N = T = 200$.

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 $N = T = 100$. 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 3.37 when the errors are correlated. However, our simulation (not reported in the table) confirms that the bias vanishes as N and T increase.

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 from 1960.1-2003.12. The variables are transformed to achieve stationarity and then outlier adjusted in the same way described in Appendix A of Stock and Watson (2005). We set the upper bound of 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 the same as static factors by design, the number of static and dynamic factors are not necessarily the same in practice. Hence, $IC^T_{1;n}$ is inconclusive about the number of static factors in the data. It finds 5 dynamic factors, which only implies that the number of static factors is no less than 5. Second, Bai and Ng's (2002) IC_{p1} and PC_{p1} find 7 and 9 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 section 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. This point is also made by Uhlig (2009) that these high number of factors found by IC_{p1} and PC_{p1} may be due to high temporal persistence of the data. In addition, we implement our estimator and find 2 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 1 static factors when $\alpha = 0.35$ and 2 static factors for all other values of α. This result is close to that by Onatski's (2010) ED estimator, which detects 1 static factor in this set. Finally, as in the literature, where the factors are estimated by PCA, the first factor is related to production and employment, and the second factor is related to the stock market, interest rates and yield spreads, which are similar to the findings of Stock and Watson (2002).

5 Conclusion

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 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 good performance in small samples.

In this paper, we only consider the Bridge type of estimator, and 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 estimator should not be just limited to selecting the number of factors. We expect that it can be applied in other context as well, such as selecting the identification scheme in a 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 $H = (\Lambda^{0'}\Lambda^0/N)(F^{0'}\hat{F}^{1:r}/T)V^{r^{-1}}$ 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 $\|H_0\| < \infty$, $Q \equiv \text{plim}_{N,T\to\infty}\hat{F}^{1:r'}F^0/T$ 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 $\hat{F}^{1:r'}F^0/T$ 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,

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

where we use Assumptions 3(i) and 4, the fact that $\|H\| = O_p(1)$ by Lemma 1(ii) and Lemma 1(i)

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

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

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 $XX'\hat{F}^p/NT = \hat{F}^pV^p$ 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

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

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

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

where we use the fact that

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

because $\hat{F}^{p'}\hat{F}^p/T = I_p$. By Lemma 2 and the definition of $D = T^{-1/2}\hat{F}^p(\hat{F}^{p'}\hat{F}^p/T)^{-1/2} = \hat{F}^p/T^{1/2}$, 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:

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

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 $S_{j,N,T} = \{\hat\Lambda : 2^{j-1} \le C_{NT}^2 N^{-1}\|\hat\Lambda - \Lambda^*\|^2 < 2^j\}$ 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 $j \ge M$. We want to show that this event happens with probability approaching zero. Since $\hat\Lambda$ minimizes (2.6), it follows that for any $\varepsilon > 0$,

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

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

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

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

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

where the equality uses the fact that $\hat{F}^{p'}\hat{F}^p/T=I_p$.

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,

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

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.

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

Since $\hat{F}^{p'}\hat{F}^p/T = I_p$, 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

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

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

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

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

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

[5] 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.

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

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

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

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

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

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

[21] 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.

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

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

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

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

[26] 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).

[27] 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.

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

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

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

[31] 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.

[32] 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.

[33] 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.

[34] van Der Vaart, A., and J. Wellner (1996), "Weak Convergence and Empirical Processes: with Applications to Statistics", Springer Verlag.

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

[36] 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}	PC_{p1}	$IC^T_{1;n}$	ED
	50	50	1.00 (0|0)	1.00 (0|0)	5.33 (100|0)	1.01 (1|0)	1.03 (2|0)
	100	50	1.00 (0|0)	1.00 (0|0)	1.70 (61|0)	1.00 (0|0)	1.01 (1|0)
	100	100	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.02 (2|0)
1	100	200	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.04 (2|0)
	200	100	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.01 (1|0)
	200	200	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.00 (0|0)	1.01 (1|0)
	50	50	2.84 (0|13)	2.99 (0|1)	5.67 (100|0)	2.58 (1|27)	3.02 (1|0)
	100	50	2.93 (0|6)	3.00 (0|0)	3.09 (9|0)	2.85 (0|8)	3.00 (0|0)
3	100	100	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)	2.99 (0|1)	3.01 (1|0)
	100	200	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)	3.01 (1|0)
	200	100	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)	2.93 (0|4)	3.00 (0|0)
	200	200	3.00 (0|0)	3.00 (0|0)	3.00 (0|0)	2.45 (0|38)	3.01 (1|0)
	50	50	1.84 (0|99)	4.53 (0|40)	6.25 (89|0)	1.48 (0|93)	3.98 (0|29)
	100	50	1.88 (0|98)	4.90 (0|10)	5.00 (0|0)	1.46 (0|95)	4.98 (0|1)
5	100	100	4.79 (0|9)	5.00 (0|0)	5.00 (0|0)	4.57 (0|18)	5.00 (0|0)
	100	200	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)	4.99 (0|1)	5.00 (0|0)
	200	100	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)	4.13 (0|30)	5.00 (0|0)
	200	200	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)	3.33 (0|75)	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% variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1} and PC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, and Onatski's (2010) ED. 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 500 replications, while 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 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}	PC_{p1}	$IC^T_{1;n}$	ED
	50	50	1.42 (29|0)	7.13 (100|0)	8.95 (100|0)	1.62 (44|0)	1.82 (19|0)
	100	50	1.00 (0|0)	5.11 (96|0)	8.92 (100|0)	1.14 (9|0)	1.10 (4|0)
	100	100	1.00 (0|0)	9.15 (100|0)	9.74 (100|0)	2.11 (45|0)	1.05 (3|0)
1	100	200	1.00 (0|0)	10.00 (100|0)	10.00 (100|0)	2.66 (41|0)	1.03 (1|0)
	200	100	1.00 (0|0)	1.38 (32|0)	7.16 (100|0)	1.08 (5|0)	1.04 (4|0)
	200	200	1.00 (0|0)	7.14 (100|0)	9.52 (100|0)	1.01 (0|0)	1.02 (2|0)
	50	50	3.18 (28|13)	8.91 (100|0)	9.59 (100|0)	1.91 (11|66)	2.11 (11|60)
	100	50	2.93 (0|6)	6.81 (100|0)	9.37 (100|0)	2.28 (4|43)	2.96 (2|6)
3	100	100	3.00 (0|0)	9.81 (100|0)	9.96 (100|0)	3.59 (39|3)	3.01 (1|0)
	100	200	3.01 (0|0)	10.00 (100|0)	10.00 (100|0)	3.84 (34|0)	3.00 (0|0)
	200	100	3.00 (0|0)	3.46 (100|0)	7.65 (100|0)	3.04 (4|0)	3.02 (1|0)
	200	200	3.00 (0|0)	8.26 (100|0)	9.65 (100|0)	3.01 (1|0)	3.00 (0|0)
	50	50	2.66 (0|89)	9.71 (100|0)	9.93 (100|0)	1.19 (0|99)	1.01 (3|95)
	100	50	2.36 (0|93)	8.36 (100|0)	9.68 (100|0)	1.21 (0|98)	1.57 (0|83)
5	100	100	4.66 (3|21)	9.95 (100|0)	9.99 (100|0)	4.04 (21|50)	2.56 (0|58)
	100	200	5.33 (29|1)	10.00 (100|0)	10.00 (100|0)	4.90 (31|39)	4.18 (0|20)
	200	100	4.99 (0|1)	5.47 (100|0)	8.15 (100|0)	3.57 (0|52)	5.01 (0|0)
	200	200	5.00 (0|0)	9.18 (100|0)	9.85 (100|0)	4.95 (0|4)	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% variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1} and PC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, and Onatski's (2010) ED. 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 500 replications, while 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 correct estimation of the number of factors.

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

r	N	T	Bridge	IC_{p1}	PC_{p1}	$IC^T_{1;n}$	ED
	50	50	1.09 (8|0)	9.02 (100|0)	9.85 (100|0)	1.34 (24|0)	1.22 (10|0)
	100	50	1.02 (2|0)	9.50 (100|0)	9.88 (100|0)	1.21 (15|0)	1.12 (5|0)
	100	100	1.00 (0|0)	6.91 (99|0)	9.45 (100|0)	1.34 (17|0)	1.05 (4|0)
1	100	200	1.00 (0|0)	1.13 (12|0)	6.52 (100|0)	1.00 (0|0)	1.03 (3|0)
	200	100	1.00 (0|0)	9.77 (100|0)	9.98 (100|0)	1.13 (5|0)	1.01 (1|0)
	200	200	1.00 (0|0)	1.88 (59|0)	7.55 (100|0)	1.00 (0|0)	1.01 (1|0)
	50	50	3.09 (18|9)	9.68 (100|0)	9.96 (100|0)	2.35 (8|38)	2.18 (4|39)
	100	50	3.03 (8|4)	9.94 (100|0)	9.99 (100|0)	2.63 (6|23)	2.62 (3|21)
3	100	100	3.00 (0|0)	8.53 (100|0)	9.75 (100|0)	3.02 (4|2)	3.00 (1|0)
	100	200	3.00 (0|0)	3.13 (13|0)	6.95 (100|0)	3.00 (0|0)	3.03 (1|0)
	200	100	3.00 (0|0)	9.97 (100|0)	10.00 (100|0)	3.03 (3|0)	3.01 (0|0)
	200	200	3.00 (0|0)	3.89 (59|0)	8.04 (100|0)	3.00 (0|0)	3.01 (1|0)
	50	50	2.72 (0|88)	9.92 (100|0)	9.99 (100|0)	1.30 (1|93)	0.94 (2|96)
	100	50	2.98 (0|83)	9.99 (100|0)	10.00 (100|0)	1.21 (0|93)	0.95 (1|94)
5	100	100	4.85 (2|10)	9.40 (100|0)	9.90 (100|0)	3.50 (0|54)	3.95 (1|26)
	100	200	5.00 (0|1)	5.15 (14|0)	7.49 (100|0)	4.93 (0|4)	5.01 (1|0)
	200	100	5.00 (1|1)	9.99 (100|0)	10.00 (100|0)	3.44 (0|49)	4.88 (0|3)
	200	200	5.00 (0|0)	5.93 (61|0)	8.41 (100|0)	4.99 (0|1)	5.01 (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% variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1} and PC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, and Onatski's (2010) ED. 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 500 replications, while 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 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}	PC_{p1}	$IC^T_{1;n}$	ED
	50	50	1.11 (10|0)	7.90 (99|0)	9.79 (100|0)	1.37 (27|0)	1.24 (12|0)
	100	50	1.00 (0|0)	7.46 (99|0)	9.60 (100|0)	1.15 (13|0)	1.15 (10|0)
	100	100	1.00 (0|0)	3.95 (95|0)	8.53 (100|0)	1.39 (20|0)	1.08 (6|0)
1	100	200	1.00 (0|0)	2.62 (85|0)	7.21 (100|0)	1.05 (3|0)	1.07 (4|0)
	200	100	1.00 (0|0)	3.25 (89|0)	8.34 (100|0)	1.04 (3|0)	1.06 (4|0)
	200	200	1.00 (0|0)	1.74 (56|0)	6.37 (100|0)	1.00 (0|0)	1.05 (3|0)
	50	50	3.06 (17|10)	9.02 (100|0)	9.89 (100|0)	2.35 (7|42)	2.35 (8|35)
	100	50	2.97 (2|5)	9.01 (100|0)	9.86 (100|0)	2.69 (5|19)	2.86 (4|10)
3	100	100	3.00 (0|0)	5.99 (96|0)	8.93 (100|0)	3.09 (10|2)	3.03 (3|0)
	100	200	3.00 (0|0)	4.45 (84|0)	7.78 (100|0)	3.05 (5|0)	3.05 (3|0)
	200	100	3.00 (0|0)	5.42 (91|0)	8.81 (100|0)	2.99 (0|1)	3.03 (2|0)
	200	200	3.00 (0|0)	3.78 (57|0)	7.07 (100|0)	3.00 (0|0)	3.01 (1|0)
	50	50	2.67 (0|88)	9.59 (100|0)	9.96 (100|0)	1.35 (0|95)	1.05 (1|94)
	100	50	2.79 (0|86)	9.67 (100|0)	9.96 (100|0)	1.43 (1|89)	1.43 (2|85)
5	100	100	4.84 (1|11)	7.73 (95|0)	9.38 (100|0)	3.58 (1|56)	3.81 (2|28)
	100	200	5.05 (5|0)	6.52 (85|0)	8.47 (100|0)	4.88 (0|7)	4.95 (2|2)
	200	100	4.99 (0|1)	7.42 (91|0)	9.28 (100|0)	4.38 (0|23)	4.99 (1|0)
	200	200	5.00 (0|0)	5.69 (54|0)	7.77 (100|0)	4.98 (0|1)	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% variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1} and PC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, and Onatski's (2010) ED. 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 500 replications, while 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 correct estimation of the number of factors.

Table 5: Conditionally heteroskedastic errors

r	N	T	Bridge	IC_{p1}	PC_{p1}	$IC^T_{1;n}$	ED
	50	50	1.21 (16|0)	8.37 (97|0)	9.92 (100|0)	1.33 (25|0)	1.26 (20|0)
	100	50	1.18 (15|0)	8.56 (98|0)	9.88 (100|0)	1.34 (25|0)	1.35 (23|0)
	100	100	1.03 (3|0)	4.22 (89|0)	8.48 (100|0)	1.51 (28|0)	1.30 (21|0)
1	100	200	1.00 (0|0)	1.69 (49|0)	5.56 (100|0)	1.16 (10|0)	1.18 (15|0)
	200	100	1.06 (6|0)	5.78 (95|0)	8.77 (100|0)	1.30 (21|0)	1.30 (22|0)
	200	200	1.02 (2|0)	2.99 (84|0)	6.30 (100|0)	1.06 (5|0)	1.31 (23|0)
	50	50	2.81 (0|16)	3.31 (25|0)	8.27 (100|0)	2.14 (3|48)	2.88 (5|9)
	100	50	2.96 (0|4)	3.34 (25|0)	7.25 (100|0)	2.58 (4|25)	3.11 (10|1)
3	100	100	3.00 (0|0)	3.02 (2|0)	4.65 (93|0)	3.06 (8|1)	3.08 (5|0)
	100	200	3.00 (0|0)	3.00 (0|0)	3.09 (9|0)	3.00 (0|0)	3.04 (4|0)
	200	100	3.00 (0|0)	3.08 (7|0)	4.16 (78|0)	3.01 (1|1)	3.10 (9|0)
	200	200	3.00 (0|0)	3.00 (0|0)	3.08 (7|0)	2.99 (0|1)	3.07 (6|0)
	50	50	1.56 (0|100)	4.63 (3|36)	7.66 (100|0)	1.28 (0|97)	2.21 (1|69)
	100	50	1.63 (0|100)	4.95 (1|6)	6.56 (95|0)	1.41 (0|93)	4.07 (1|23)
5	100	100	4.53 (0|20)	5.00 (0|0)	5.10 (10|0)	4.12 (0|35)	5.00 (1|0)
	100	200	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)	4.95 (0|3)	5.02 (2|0)
	200	100	4.99 (0|1)	5.00 (0|0)	5.04 (4|0)	4.69 (0|11)	5.04 (3|0)
	200	200	5.00 (0|0)	5.00 (0|0)	5.00 (0|0)	4.97 (0|1)	5.03 (3|0)

Note: The conditionally heteroskedastic error terms are generated using DGP E2. The factors explain 50% variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1} and PC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, and Onatski's (2010) ED. 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 500 replications, while 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 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}	PC_{p1}	$IC^T_{1;n}$	ED
	50	50	1.11 (10|0)	6.79 (99|0)	9.65 (100|0)	1.30 (23|0)	1.21 (11|1)
	100	50	1.00 (0|0)	6.82 (100|0)	9.51 (100|0)	1.17 (14|0)	1.17 (9|0)
	100	100	1.00 (0|0)	3.30 (91|0)	8.22 (100|0)	1.40 (19|0)	1.06 (5|0)
1	100	200	1.00 (0|0)	2.08 (71|0)	6.75 (100|0)	1.12 (5|0)	1.08 (5|0)
	200	100	1.00 (0|0)	2.70 (82|0)	8.11 (100|0)	1.10 (5|0)	1.04 (3|0)
	200	200	1.00 (0|0)	1.44 (38|0)	5.91 (100|0)	1.00 (0|0)	1.05 (4|0)
	50	50	1.88 (6|74)	8.24 (99|0)	9.79 (100|0)	1.41 (4|79)	1.09 (5|84)
	100	50	1.83 (0|72)	8.39 (100|0)	9.80 (100|0)	1.49 (2|75)	1.41 (3|71)
3	100	100	2.79 (0|15)	5.24 (90|0)	8.61 (100|0)	2.96 (8|14)	2.77 (4|13)
	100	200	3.00 (0|0)	3.97 (68|0)	7.39 (100|0)	3.06 (3|1)	3.04 (4|0)
	200	100	2.98 (0|1)	4.77 (82|0)	8.55 (100|0)	2.87 (2|10)	3.02 (2|0)
	200	200	3.00 (0|0)	3.44 (37|0)	6.68 (100|0)	3.00 (0|0)	3.02 (1|0)
	50	50	0.85 (0|100)	8.42 (88|6)	9.87 (100|0)	0.98 (0|99)	0.72 (0|99)
	100	50	0.64 (0|100)	8.79 (96|2)	9.87 (100|0)	0.85 (0|99)	0.74 (0|99)
5	100	100	0.96 (0|100)	6.69 (82|2)	9.04 (100|0)	2.16 (3|87)	1.04 (1|95)
	100	200	1.46 (0|97)	5.86 (62|0)	8.07 (100|0)	3.88 (2|50)	2.11 (1|69)
	200	100	1.13 (0|99)	6.67 (81|0)	9.00 (100|0)	2.18 (0|87)	2.42 (1|62)
	200	200	4.60 (0|12)	5.45 (39|0)	7.40 (100|0)	4.08 (0|35)	4.93 (0|2)

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% variation in the data. Our estimator is compared with Bai and Ng's (2002) IC_{p1} and PC_{p1}, Hallin and Liska's (2007) $IC^T_{1;n}$, and Onatski's (2010) ED. 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 500 replications, while 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 correct estimation of the number of factors.

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

β	ρ	α	$r = 1$	$r = 3$	$r = 5$
0	0	0.10	1.00 (0|0)	3.00 (0|0)	4.89 (0|8)
0	0	0.15	1.00 (0|0)	3.00 (0|0)	4.88 (0|10)
0	0	0.25	1.00 (0|0)	3.00 (0|0)	4.82 (0|9)
0	0	0.35	1.00 (0|0)	3.00 (0|0)	4.89 (0|3)
0	0	0.40	1.00 (0|0)	3.00 (0|0)	4.81 (0|4)
0.1	0.6	0.10	1.04 (4|0)	3.00 (0|0)	4.95 (1|5)
0.1	0.6	0.15	1.00 (0|0)	3.00 (0|0)	4.89 (1|10)
0.1	0.6	0.25	1.00 (0|0)	3.00 (0|0)	4.84 (1|11)
0.1	0.6	0.35	1.00 (0|0)	3.01 (1|0)	4.60 (2|14)
0.1	0.6	0.40	1.00 (0|0)	3.01 (1|0)	3.37 (7|45)

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% variation in the data. 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 500 replications, while 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 correct estimation of the number of factors.