Learning Latent Factors from Diversified Projections and its Applications to Over-Estimated and Weak Factors

Jianqing Fan* Yuan Liao†

Abstract

Estimations and applications of factor models often rely on the crucial condition that the number of latent factors is consistently estimated, which in turn also requires that factors be relatively strong, data are stationary and weak serial dependence, and the sample size be fairly large, although in practical applications, one or several of these conditions may fail. In these cases it is difficult to analyze the eigenvectors of the data matrix. To address this issue, we propose simple estimators of the latent factors using cross-sectional projections of the panel data, by weighted averages with pre-determined weights. These weights are chosen to diversify away the idiosyncratic components, resulting in "diversified factors". Because the projections are conducted cross-sectionally, they are robust to serial conditions, easy to analyze and work even for finite length of time series. We formally prove that this procedure is robust to over-estimating the number of factors, and illustrate it in several applications, including post-selection inference, big data forecasts, large covariance estimation and factor specification tests. We also recommend several choices for the diversified weights.

Key words: Large dimensions, random projections, over-estimating the number of factors, principal components, factor-augmented regression

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. jqfan@princeton.edu. His research is supported by NSF grants DMS-1662139 and DMS-1712591.

†Department of Economics, Rutgers University, 75 Hamilton St., New Brunswick, NJ 08901, USA. yuan.liao@rutgers.edu

1 Introduction

Consider the following high-dimensional factor model:

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, \cdots, T, \tag{1.1}$$

where $\mathbf{x}_t = (x_{1t}, \cdots, x_{NT})'$ is an N-dimensional outcome. In addition, the model contains $\mathbf{f}_t$ as r-dimensional latent factors, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_N)'$ as $N \times r$ matrix of loadings, and $\mathbf{u}_t = (u_{1t}, \cdots, u_{Nt})'$ as idiosyncratic terms. Theoretical studies of the model have been crucially depending on the assumption that the number of factors, r, should be consistently estimated. This in turn, requires the factors be relatively strong, data have weak serial dependence, and length of time series T is long. But in practical applications, one or several of these conditions may fail to hold due to weak signal-noise ratios and nonstationary or noisy data, making the first r eigenvalues of the sample covariance of $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ empirically be not so-well separated from the remaining ones.

A promising remedy is to over-estimate the number of factors. But this approach has been quite challenging. Let R be the "working number of factors" that are empirically estimated. When $R > r$, it is often difficult to analyze the behavior of the $(R - r)$ eigenvalues/eigenvectors. As shown in Johnstone and Lu (2009), these eigenvectors can be inconsistent because their eigenvalues are not so "spiked". This creates challenges to many factor estimators, such as the popular principal components (PC)-estimator (Connor and Korajczyk, 1986; Stock and Watson, 2002), and therefore brings obstacles to applications when the number of factors is over-estimated. Another difficulty is to handle the serial dependence. As shown by Bai (2003), the PC-estimator is inconsistent under finite-T in the presence of serial correlations and heteroskedasticity, but many forecast applications using estimated factors favor relatively short time series, due to the concern of nonstationarity.

This paper proposes a new method to address issues of over-estimating the number of factors, small T, and strong serial conditions. We propose a simple factor estimator that does not rely on eigenvectors. Let $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_R)$ be a given exogenous (or deterministic) $N \times R$ matrix, where each of its R columns $\mathbf{w}_k$ is an $N \times 1$ vector of "diversified weights", in the sense that its strength should be approximately equally distributed on most of its components. We propose to estimate $\mathbf{f}_t$ by simply

$$\widehat{\mathbf{f}}_t = \frac{1}{N}\mathbf{W}'\mathbf{x}_t,$$

or more precisely, the linear space spanned by $\{\mathbf{f}_t\}_{t=1}^T$ is estimated by that spanned by $\{\widehat{\mathbf{f}}_t\}_{t=1}^T$. By substituting (1.1) into the definition, we have

$$\widehat{\mathbf{f}}_t = \underbrace{(\frac{1}{N}\mathbf{W}'\mathbf{B})}_{\text{affine transform}} \mathbf{f}_t + \frac{1}{N}\mathbf{W}'\mathbf{u}_t. \tag{1.2}$$

Thus $\widehat{\mathbf{f}}_t$ (consistently) estimates $\mathbf{f}_t$ up to an $R \times r$ affine transform, with $\mathbf{e}_t := \frac{1}{N}\mathbf{W}'\mathbf{u}_t$ as the estimation error. The assumption that $\mathbf{W}$ should be diversified ensures that as $N \to \infty$, $\mathbf{e}_t$ is "diversified away" (converging to zero in probability).

We call the new factor estimator as "diversified factors", which reduces the dimension of $\mathbf{x}_t$ through diversified projections. Because of the clean expansion (1.2), the mathematics for theoretical analysis is much simpler than most benchmark estimators. We show that $\widehat{\mathbf{f}}_t$ leads to valid inferences in several factor-augmented models so long as $R \geq r$. Therefore, we formally justify that the use of factor models is robust to over-estimating the number of factors. In particular, we admit $r = 0$ but $R \geq 1$ as a special case. That is, the inference is still valid even if there are no common factors present, but we nevertheless take out estimated factors (for insurance). Furthermore, the projection is conducted on cross-sections, so is not sensitive to serial conditions. We study several applications in detail, including the post-selection inference, big data forecasts, high-dimensional covariance estimation, and factor specification tests.

One of the key assumptions imposed is that while $\mathbf{W}$ diversifies away $\mathbf{u}_t$, we have

$$\text{rank}\left(\frac{1}{N}\mathbf{W}'\mathbf{B}\right) = r,$$

and the r th smallest singular value of $\frac{1}{N}\mathbf{W}'\mathbf{B}$ does not decay too fast. That is, $\mathbf{W}$ should not diversify away the factor components in the time series. This condition *does not* hold if $\mathbf{W}$ has more than $R-r$ columns that are nearly orthogonal to $\mathbf{B}$. This is another motivation of using over-estimated factors: if random weights are used the probability that more than $R-r$ columns of $\mathbf{W}$ are nearly orthogonal to the space of $\mathbf{B}$ should be very small.

To satisfy the above conditions on the weights, we rely on external information on the factor loadings, and recommend four choices for the weight matrix. The first choice is the individual-specific characteristics. As documented in semi-parametric factor models, Connor et al. (2012); Park et al. (2009); Fan et al. (2016), factor loadings are often driven by observed characteristics. When these variables are available, they can be naturally used as diversified weights. The second choice is based on rolling window estimations. Consider time series forecasts. To pertain the stationarity assumption, we divide the sampling periods into (I) $t = 1, ..., T_0$ and (II) $t = T_0 + 1, ..., T_0 + T$, and only use the most recent T observations from period (II) to learn the latent factors for forecasts. Or consider a time series where a structural break occurs at time T_0, so the most recent period (II) is of major interest. Assume that the loadings are correlated between the two periods, then the PC-estimated loadings from periods (I) would be a good choice of the diversified weights for period (II). For the third recommendation, when the time series is independent of the initial observation, we can use transformations of $\mathbf{x}_0$ as the weights. The fourth recommended choice is to use columns of the Walsh-Hadamard matrix from the statistical experimental design to form the diversified weights.

The idea of approximating factors by weighted averages of observations has been applied previously in the literature. In the asset pricing literature, factors are created by weighted averages of a large number of asset returns. There, the weights are also pre-determined, adapted to the filtration up to the last observation time. In the common correlated effects (CCE) literature (Pesaran, 2006; Chudik et al., 2011), factors are created using a set of random weights to estimate the effect of observables. There, R equals the dimensions of additionally observed regressors and the outcome variable, and certain rank conditions about the regressors are required. In the same setting, Westerlund and Urbain (2015) and Karabiyik et al. (2019) compared the cross-sectional average and the PC estimators, and also showed the validity of using $R > r$ number of cross-sectional averages. Moreover, Barigozzi and Cho (2018) proposed a different method to address the issue of over-estimating factors. One of our recommended weights is inspired by their approach. Moon and Weidner (2015) studied the problem in a panel data framework and showed that the inference about the parameter of interest is robust to over-estimating the number of factors. Finally, there is a large literature on estimating the number of factors. See Bai and Ng (2002); Hallin and Liška (2007); Ahn and Horenstein (2013); Li et al. (2017).

The rest of the paper is organized as follows. Section 2 explains the key ideas and intuitions in details. Section 3 presents several applications of the diversified factors. Section 4 recommends several choices of the weight matrix. Section 5 conducts extensive simulation studies using various models. All technical proofs are presented in the appendix.

We use the following notation. For a matrix $\mathbf{A}$, we use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote its smallest and largest eigenvalues. We define the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\operatorname{tr}(\mathbf{A}'\mathbf{A})}$ and the operator norm $\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}'\mathbf{A})}$. In addition, define projection matrices $\mathbf{M}_{\mathbf{A}} = \mathbf{I} - \mathbf{P}_{\mathbf{A}}$ and $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}$ when $\mathbf{A}'\mathbf{A}$ is invertible. Finally, for two (random) sequences a_T and b_T, we write $a_T \ll b_T$ (or $b_T \gg a_T$) if $a_T = o_P(b_T)$.

2 Factor Estimation Using Diversified Projections

2.1 The estimator

Let $R \geq r$ be a pre-determined bounded integer that does not grow with N, which we call "the working number of factors". As in practice we do not know the true number of factors r, we often take a slightly large R so that $R \geq r$ is likely to hold. Let $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_R)$ be a user-specified $N \times R$ matrix, either deterministic or random but independent of the σ-algebra generated by $\{\mathbf{u}_t : t = 1, 2, ...\}$. Each of its R columns $\mathbf{w}_k = (w_{k,1}, \cdots, w_{k,N})'$ ($k \leq R$) is an $N \times 1$ vector satisfying the following:

Assumption 2.1 (Diversified weights)**.** *There are constants $0 < c < C$, so that (almost surely if $\mathbf{W}$ is random) as $N \to \infty$,*
(i) $\max_{i \leq N} |w_{k,i}| < C$.
(ii) The $R \times R$ matrix $\frac{1}{N}\mathbf{W}'\mathbf{W}$ satisfies $\lambda_{\min}(\frac{1}{N}\mathbf{W}'\mathbf{W}) > c$.
(iii) $\mathbf{W}$ is independent of $\{\mathbf{u}_t : t \leq T\}$.

Construct a factor estimator as an $R \times 1$ vector at each time t:

$$\widehat{\mathbf{f}}_t := \frac{1}{N}\mathbf{W}'\mathbf{x}_t.$$

In financial economics applications where $\mathbf{x}_t$ is a vector of asset returns, then each component of $\widehat{\mathbf{f}}_t$ is essentially a diversified portfolio return at time t due to its linear form. The behavior of $\widehat{\mathbf{f}}_t$ is strikingly simple and clean. Define an $R \times r$ matrix

$$\mathbf{H} := \frac{1}{N}\mathbf{W}'\mathbf{B}.$$

Then, it follows from the definition and (1.1) that

$$\widehat{\mathbf{f}}_t = \mathbf{H}\mathbf{f}_t + \frac{1}{N}\mathbf{W}'\mathbf{u}_t. \tag{2.1}$$

Therefore, $\widehat{\mathbf{f}}_t$ estimates an affine transformation of $\mathbf{f}_t$, where $\mathbf{H}$ is the $R \times r$ transformation matrix. The estimation error equals the diversified idiosyncratic noise $\frac{1}{N}\mathbf{w}_k'\mathbf{u}_t = \frac{1}{N}\sum_{i=1}^N w_{k,i}u_{it}$ for each $k \leq R$. When $(u_{1t}, \cdots, u_{Nt})$ are cross-sectionally weakly dependent, Assumption 2.1 ensures that $\frac{1}{N}\mathbf{w}_k'\mathbf{u}_t$ admits a cross-sectional central limit theorem. For instance, in the special case of cross-sectional independence, it is straightforward to verify the Lindeberg's condition under Assumption 2.1, and therefore as $N \to \infty$,

$$\frac{1}{\sqrt{N}}\mathbf{W}'\mathbf{u}_t \xrightarrow{d} \mathcal{N}(0, \mathbf{V}), \tag{2.2}$$

where $\mathbf{V} = \lim_{N\to\infty} \frac{1}{N}\mathbf{W}' \operatorname{var}(\mathbf{u}_t)\mathbf{W}$ which is assumed to exist.

The convergence (2.2) shows that $\sqrt{N}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)$ is asymptotically normal for each $t \leq T$. Importantly, it holds regardless of whether $T \to \infty$, $R = r$, or not. It requires only that $N \to \infty$ and that the weights should be chosen to satisfy Assumption 2.1. This fact is particularly useful for analyzing short time series.

In addition, the factor components should not be diversified away. This gives rise to the following condition on the transformation matrix. Let $\nu_{\min}(\mathbf{H})$ and $\nu_{\max}(\mathbf{H})$ respectively denote the minimum and maximum *nonzero* singular values of $\mathbf{H}$.

Assumption 2.2. *Suppose $R \geq r$. Almost surely (i)* $\operatorname{rank}(\mathbf{H}) = r$.
(ii) There is $C > 0$,

$$\nu^2_{\min}(\mathbf{H}) \gg \frac{1}{N}, \quad \nu_{\max}(\mathbf{H}) \leq C\nu_{\min}(\mathbf{H}).$$

Assumption 2.2 requires that $\mathbf{W}$ have at least r columns that are not orthogonal to $\mathbf{B}$ so that $\mathbf{B}$ is not diversified away. This is the key assumption, but is not stringent in the context of over-estimating factors. In the current setting the factor strength is measured by $\nu_{\min}(\mathbf{H})$, which is required not to decay very fast by condition (ii). This quantity determines the rate of convergence in recovering the space spanned by the factors.

Given $\widehat{\mathbf{f}}_t$, it is straightforward to estimate the loading matrix by using the least squares:

$$\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_N)' = \sum_{t=1}^{T} \mathbf{x}_t \widehat{\mathbf{f}}_t' (\sum_{t=1}^{T} \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t')^{-1}.$$

We show that the $R \times R$ matrix $\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t'$ is nonsingular with probability approaching one even when $R > r$. So $\widehat{\mathbf{B}}$ is well defined. Finally, $\mathbf{u}_t$ can be estimated by

$$\widehat{\mathbf{u}}_t = (\widehat{u}_{1t}, \cdots, \widehat{u}_{Nt}) = \mathbf{x}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t. \tag{2.3}$$

Just like the PC-estimator, the diversified projection can estimate dynamic factor models by treating dynamic factors as static factors. In addition, it is straightforward to extend the model to allowing time-varying factor loadings, by time-domain local smoothing before applying the diversified projection. While these extensions are straightforward, here we focus on static and time invariant models.

2.2 Over-estimating the number of factors

The consistent estimation for the number of factors r often requires strong conditions that may be violated in finite sample. An advantage of the diversified factors is being robust to over-estimating the number of factors in many inference problems.

We start with a heuristic discussion of the main issue in this subsection. Recall that $\mathbf{H} = \frac{1}{N}\mathbf{W}'\mathbf{B}$ is the $R \times r$ matrix, which is no longer a square matrix when $R > r$. In this case $\widehat{\mathbf{B}}$ is essentially estimating $\mathbf{B}\mathbf{H}^+$, with the $r \times R$ transformation matrix $\mathbf{H}^+$ being the Moore-Penrose generalized inverse of $\mathbf{H}$, defined as follows. Suppose $\mathbf{H}'$ has the following singular value decomposition:

$$\mathbf{H}' = \mathbf{U}_H(\mathbf{D}_H, 0)\mathbf{E}'_H, \quad r \times R$$

where 0 in the above singular value matrix is present whenever $R > r$, and $\mathbf{D}_H$ is an $r \times r$ diagonal matrix of the nonzero singular values. Then $\mathbf{H}^+$ is an $r \times R$ matrix:

$$\mathbf{H}^+ = \mathbf{U}_H(\mathbf{D}_H^{-1}, 0)\mathbf{E}'_H.$$

It is straightforward to verify that $\mathbf{H}^+\mathbf{H} = \mathbf{I}_r$ holds and that for estimating the common component $\mathbf{B}\mathbf{f}_t$ using over-estimated number of factors, we have

$$\widehat{\mathbf{B}}\widehat{\mathbf{f}}_t = \mathbf{B}\mathbf{H}^+\mathbf{H}\mathbf{f}_t + o_P(1) = \mathbf{B}\mathbf{f}_t + o_P(1). \tag{2.4}$$

where $o_P(1)$ in the above approximation can be made uniformly across elements.

However, a key challenge of formalizing the intuition behind (2.4) is to analyze the invertibility of the gram matrix $\frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$, which appears in the definition of $\widehat{\mathbf{B}}$. It is also a key ingredient in most applications of factor-augmented models wherever the estimated factors are used as regressors. Define

$$\widehat{\mathbf{S}}_f = \frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t', \quad \mathbf{S}_f = \mathbf{H}\frac{1}{T}\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t'\mathbf{H}',$$

where $\mathbf{S}_f$ is the population analogue of $\widehat{\mathbf{S}}_f$. The following three bounds when $R > r$, proved in Proposition A.1, play a fundamental role in the asymptotic analysis throughout the paper:

(i) With probability approaching one, $\widehat{\mathbf{S}}_f$ is invertible, but its eigenvalues may decay quickly so that

$$\|\widehat{\mathbf{S}}_f^{-1}\| = O_P(N). \tag{2.5}$$

On the other hand, $\mathbf{S}_f$ is degenerate when $R > r$, whose rank equals r. Also note that we still have $\|\widehat{\mathbf{S}}_f^{-1}\| = O_P(1)$ when $R = r$ holds.

(ii) Even if $R > r$, $\|\mathbf{H}'\widehat{\mathbf{S}}_f^{-1}\|$ is much smaller:

$$\|\mathbf{H}'\widehat{\mathbf{S}}_f^{-1}\| = O_P\left(\sqrt{\frac{\max\{N,T\}}{T}}\right).$$

(iii) When $R > r$, $\|\widehat{\mathbf{S}}_f^{-1} - \mathbf{S}_f^{+}\| \neq o_P(1)$ but we have

$$\|\mathbf{H}'(\widehat{\mathbf{S}}_f^{-1} - \mathbf{S}_f^{+})\mathbf{H}\| = O_P\left(\frac{1}{T} + \frac{1}{N}\right).$$

Therefore, $\widehat{\mathbf{S}}_f$ is invertible, and when weighted by the transformation matrix $\mathbf{H}'$, its inverse is well behaved and fast converges to the generalized inverse of $\mathbf{S}_f$, even though $\mathbf{S}_f$ is singular when $R > r$. It is sufficient to consider $\mathbf{H}'\widehat{\mathbf{S}}_f^{-1}$ in most factor-augmented inference problems, because in regression models $\widehat{\mathbf{S}}_f^{-1}$ often appears in the projection matrix $\mathbf{P}_{\widehat{\mathbf{F}}} = \widehat{\mathbf{F}}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'$ through $\mathbf{H}'\widehat{\mathbf{S}}_f^{-1}$ asymptotically, where $\widehat{\mathbf{F}} := (\widehat{\mathbf{f}}_1, \cdots, \widehat{\mathbf{f}}_T)'$ and $\mathbf{F} := (\mathbf{f}_1, \cdots, \mathbf{f}_T)'$ denote the estimated and true factor matrices.

Remark 2.1. In the CCE literature, (e.g., Pesaran (2006); Chudik et al. (2011)), it has also been claimed that estimating the factors using cross-sectional averages does not require consistently estimating the number of factors. While the claim is true, its proof is not straightforward as $\|\widehat{\mathbf{S}}_f^{-1} - \mathbf{S}_f^{+}\| \neq O_P(1)\|\widehat{\mathbf{S}}_f - \mathbf{S}_f\|$ when $R > r$. Also see Karabiyik et al. (2017, 2019) for more discussions on the related issue. Our method therefore also potentially contributes to this literature as an alternative rigorous approach.

2.3 Estimating the factor space

Throughout the paper, the loading matrix $\mathbf{B}$ can be either deterministic or random. When they are random, it is assumed that it is independent of $\mathbf{u}_t$, and all the expectations throughout the paper is taken conditionally on $\mathbf{B}$.

We make the following conditions.

Assumption 2.3. *(i)* $\{(\mathbf{f}_t, \mathbf{u}_t) : t \leq T\}$ *is a stationary process, satisfying* $\mathbb{E}(\mathbf{u}_t|\mathbf{f}_t) = 0$.

(ii) There are constants $c, C > 0$, *so that* $\max_{i\leq N} \|\mathbf{b}_i\| < C$, *and almost surely*

$$c < \lambda_{\min}(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t') \leq \lambda_{\max}(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t') < C.$$

Assumption 2.4 (Weak dependence)**.** *There is a constant* $C > 0$,

(i) $\max_{j,i\leq N} \frac{1}{NT}\sum_{q,v\leq N}\sum_{t,s\leq T} |\operatorname{Cov}(u_{it}u_{qt}, u_{js}u_{vs}|\mathbf{F})| < C$ *almost surely in* $\mathbf{F}$,

(ii) $\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}\|\mathbf{f}_t\|\|\mathbf{f}_s\|\|\mathbb{E}(\mathbf{u}_t\mathbf{u}_s'|\mathbf{F})\| < C$ *and* $\mathbb{E}\|\mathbb{E}(\mathbf{u}_t\mathbf{u}_t'|\mathbf{F})\| < C$.

Theorem 2.1. *Suppose Assumptions 2.1 - 2.4 hold. Also* $N \to \infty$ *and* T *is either finite or grows. Then for all bounded* $R \geq r$,

$$\|\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{P}_{\mathbf{F}} - \mathbf{P}_{\mathbf{F}}\| = O_P\left(\frac{1}{\sqrt{N}}\nu_{\min}^{-1}(\mathbf{H})\right), \tag{2.6}$$

$$\|\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} - \mathbf{P}_{\mathbf{F}}\| = O_P\left(\frac{1}{\sqrt{N}}\nu_{\min}^{-1}(\mathbf{H})\right), \tag{2.7}$$

where $\mathbf{M} = (\mathbf{H}\mathbf{H}')^{+}\mathbf{H}$ *is an* $R \times r$ *matrix* [1].

Equation (2.6) shows that when $R \geq r$, the linear space spanned by $\widehat{\mathbf{F}}$ asymptotically covers the linear space spanned by $\mathbf{F}$. To understand the intuition, note that (2.6) implies $\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{P}_{\mathbf{F}}\mathbf{Y} \approx \mathbf{P}_{\mathbf{F}}\mathbf{Y}$ for an arbitrary random matrix $\mathbf{Y}$. Meanwhile, if we heuristically regard $\mathbf{P}_{\mathbf{F}}$ and $\mathbf{P}_{\widehat{\mathbf{F}}}$ as conditional expectations given $\mathbf{F}$ and $\widehat{\mathbf{F}}$, then approximately,

$$\mathbb{E}\left(\mathbb{E}(\mathbf{Y}|\mathbf{F})\middle|\widehat{\mathbf{F}}\right) \approx \mathbb{E}(\mathbf{Y}|\mathbf{F}). \tag{2.8}$$

[1] We show in the proof that $(\mathbf{M}'\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M})$ and $\widehat{\mathbf{F}}'\widehat{\mathbf{F}}$ are both invertible with probability approaching one. So $\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}}$ and $\mathbf{P}_{\widehat{\mathbf{F}}}$ are well defined asymptotically.

Let span($\mathbf{A}$) denote the linear space spanned by the columns of $\mathbf{A}$. The approximation (2.8) is well known to be the "tower property", which heuristically means span($\mathbf{F}$) $\subseteq$ span($\widehat{\mathbf{F}}$).

Equation (2.7) shows that a particular subspace of span($\widehat{\mathbf{F}}$) is consistent for span($\mathbf{F}$). In the special case $R = r$, we have $\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} = \mathbf{P}_{\widehat{\mathbf{F}}}$ since $\mathbf{M}$ in (2.7) is invertible. It then reduces to the usual space consistency. Importantly, we allow T to be finite.

To gain more insights of these results, let us compare with the usual methods based on estimating the number of factors, e.g., the eigenvalue-ratio method of Ahn and Horenstein (2013). There are two key quantities in this comparison: the strength of the spiked eigenvalues of $\mathbf{S}_x := \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$, and the largest eigenvalue of $\mathbf{S}_u := \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{u}_t'$.

We consider a setting where we can easily quantify the signal-noise ratio, as given in the following example.

Example 2.1. This example presents a *pervasive factor model* that satisfies Assumption 2.2. Suppose each individual loading satisfies $\mathbf{b}_i = \nu_N\boldsymbol{\lambda}_i$ for some sequence $\nu_N \asymp N^{-(1-\alpha)/2}$ and $\alpha \in (0, 1]$, where $\{\boldsymbol{\lambda}_i : i \leq N\}$ is a sequence of $r \times 1$ vectors such that:

(i) $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i' \to \mathbf{C}$ (or converges in probability if $\boldsymbol{\lambda}_i$ is random) for some positive definite matrix $\mathbf{C}$;

(ii) $\nu_{\min}(\frac{1}{N}\mathbf{W}'\boldsymbol{\Lambda})$ is bounded away from zero, where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N)'$.

Then Assumption 2.2 holds for $\nu_{\min}(\mathbf{H}) \asymp \nu_N$ and any $\alpha \in (0, 1]$. It is straightforward to verify that the r th spiked eigenvalue satisfies:

$$\lambda_r(\mathbf{S}_x) \asymp N^{\alpha}, \quad \alpha \in (0, 1].$$

Theorem 2.1 then shows that $\|\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} - \mathbf{P}_{\mathbf{F}}\| = o_P(1)$ for any $\alpha > 0$. To verify condition (ii) in this example, consider a "characteristic based" model described in Section 4, where the baseline loading can be decomposed as $\boldsymbol{\lambda}_i = \mathbf{g}(\mathbf{z}_i) + \boldsymbol{\gamma}_i$, with $\mathbb{E}(\boldsymbol{\gamma}_i|\mathbf{z}_i) = 0$; $\mathbf{g}(\mathbf{z}_i)$ is a nonparametric function of some observable characteristic $\mathbf{z}_i$, and $\boldsymbol{\gamma}_i$ is the loading components that is orthogonal to the characteristic effects. (See more detailed motivations of this model in Section 4.1) Now take $\mathbf{w}_i = \phi(\mathbf{z}_i)$ as an R-dimensional transformation using R predetermined basis functions ϕ. Then a sufficient condition for (ii) is that $\nu_{\min}(\mathbf{A})$ is bounded away from zero, where $\mathbf{A} := \frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{z}_i)\mathbf{g}(\mathbf{z}_i)'$. In addition, Assumption 2.2(ii) holds as long as $\nu_{\max}(\mathbf{A}) < C$. [2]

The key implication of Example 2.1 is that the strength of the spiked eigenvalues can grow at an arbitrarily slow polynomial rate in N, and T is allowed to be finite. In applications where $T \to \infty$ is required, the growth requirement of T can be very mild. For instance, as we shall show in the high-dimensional factor-augmented regression (Section 3.2), it is only required that $\log^2 N = o(T)$ if the number of "important" control variables (corresponding to nonzero coefficients) is finite. The relative flexibility on the growth of T is achieved thanks to the fact that the diversified projection does not demand strong eigenvalues of the population covariance matrix.

Now let us revisit the conditions required by the eigenvalue-ratio method by Ahn and Horenstein (2013). If $\mathbf{u}_t$ is sub-Gaussian, under weak dependence conditions,

$$\lambda_{\max}(\mathbf{S}_u) = O_P\left(\frac{\max\{T, N\}}{T}\right).$$

The selection consistency requires $\lambda_r(\mathbf{S}_x) \gg \lambda_{\max}(\mathbf{S}_u)$, which in this context, becomes $T \gg N^{1-\alpha}$. In the case that the spiked eigenvalues are not so strong ($\alpha < 0.5$), it requires a considerably longer time series to override the effect of the idiosyncratic noise.

[2]Suppose $\{\boldsymbol{\gamma}_i : i \leq N\}$ are cross-sectionally conditionally weakly dependent given $\mathbf{Z} = (\mathbf{z}_i : i \leq N)$. Then $\|\frac{1}{N}\sum_i \phi(\mathbf{z}_i)\boldsymbol{\gamma}_i\|^2 = O_P(X)\frac{1}{N}\sum_i \|\phi(\mathbf{z}_i)\|^2 = o_P(1)$ given the assumption that $X := \max_i \frac{1}{N}\sum_j \|\mathbb{E}[\boldsymbol{\gamma}_i\boldsymbol{\gamma}_j'|\mathbf{Z}]\| = o_P(1)$, which holds if $\boldsymbol{\gamma}_i$ are conditionally weakly correlated. Then $\nu_{\min}(\frac{1}{N}\mathbf{W}'\boldsymbol{\Lambda}) \geq \nu_{\min}(\mathbf{A}) - \|\frac{1}{N}\sum_i \phi(\mathbf{z}_i)\boldsymbol{\gamma}_i\| \geq c - o_P(1)$. In addition, $\nu_{\max}(\mathbf{H}) = \nu_{\max}(\frac{1}{N}\mathbf{W}'\boldsymbol{\Lambda})\nu_N \leq [\nu_{\max}(\mathbf{A}) + o_P(1)]\nu_N \leq C\nu_N \leq Cc^{-1}\nu_{\min}(\mathbf{A})\nu_N \leq Cc^{-1}[\nu_{\min}(\frac{1}{N}\mathbf{W}'\boldsymbol{\Lambda}) + o_P(1)]\nu_N \leq 2Cc^{-1}\nu_{\min}(\mathbf{H})$.

2.4 Summary of advantages

Below we summarize key advantages of the use of diversified projection.

1. It is computationally and mathematically simple.

2. When the true number of factors is over estimated ($R \geq r$), inferences about transformation invariant parameters are still asymptotically valid. This leads to important implications on factor-augmented inferences and out-of sample forecasts.

3. It admits an interesting special case, where $r = 0$ and $R \geq 1$. That is, $\mathbf{x}_t$ is in fact weakly dependent, but we nevertheless estimate "factors". The resulting inference is still asymptotically valid in this case. We shall formally prove this in the high-dimensional factor-augmented inference in the next section. This shows that extracting estimated factors is a robust inference procedure.

4. As the diversified projections are applied cross-sectionally, some conditions that are needed for the PC-estimator can be weakened. For instance, the space spanned by the latent factors can be consistently estimated even if T is finite. It is also a good choice under weak signal-noise ratios where the consistent selection of the number of factors is hard to achieve.

5. After applying the diversified projection to $\mathbf{x}_t$ to reduce to a lower dimensional space, one can continue to employ the PCA on $\widehat{\mathbf{f}}_t$ to estimate the factor space and the number of factors. This becomes a low-dimensional PCA problem, and potentially much easier than benchmark methods dealing with large dimensional datasets.

3 Applications

We present several applications of the new diversified factors. Besides those imposed in Section 2, additional assumptions are required in each of these examples. These assumptions are application-specific and are required even if the oracle number of factors were available.

3.1 Forecasts using augmented factor regression

Consider forecasting time series using a large panel of augmented factor regression:

$$
\begin{aligned}
y_{t+h} &= \boldsymbol{\alpha}'\mathbf{f}_t + \boldsymbol{\beta}'\mathbf{g}_t + \varepsilon_{t+h}, \quad t = 1, \cdots ., T \\
\mathbf{x}_t &= \mathbf{B}\mathbf{f}_t + \mathbf{u}_t
\end{aligned}
$$

with observed data $\{(y_t, \mathbf{x}_t) : t \leq T\}$. Here $h \geq 0$ is the lead time and $\mathbf{g}_t$ is a vector of observed predictors including lagged outcome variables. The goal is the mean forecast:

$$
y_{T+h|T} := \boldsymbol{\alpha}'\mathbf{f}_T + \boldsymbol{\beta}'\mathbf{g}_T := \boldsymbol{\delta}'\mathbf{z}_T,
$$

where $\mathbf{z}_t = (\mathbf{f}_t'\mathbf{H}', \mathbf{g}_t')'$ and $\boldsymbol{\delta}' = (\boldsymbol{\alpha}'\mathbf{H}^+, \boldsymbol{\beta}')$. The prediction also depends on unobservable factors $\mathbf{f}_t$ whose information is contained in a high-dimensional panel of data. This model has been studied extensively in the literature, see e.g., Stock and Watson (2002); Bai and Ng (2006); Ludvigson and Ng (2007), where $\mathbf{f}_T$ is replaced by a consistent estimator. Once estimated factors $\widehat{\mathbf{f}}_t$ is obtained, the forecast of $y_{T+h|T}$ is straightforward:

$$
\widehat{y}_{T+h|T} = \widehat{\boldsymbol{\delta}}'\widehat{\mathbf{z}}_T, \quad \widehat{\boldsymbol{\delta}} = (\sum_{t=1}^{T-h} \widehat{\mathbf{z}}_t\widehat{\mathbf{z}}_t')^{-1} \sum_{t=1}^{T-h} \widehat{\mathbf{z}}_t y_{t+h}
$$

where $\widehat{\mathbf{z}}_t = (\widehat{\mathbf{f}}_t', \mathbf{g}_t')'$ denotes the estimated regressors. Note that $(\sum_{t=1}^{T-h} \widehat{\mathbf{z}}_t\widehat{\mathbf{z}}_t')^{-1}$ is well defined even if $R > r$ with high probability. This follows from the invertibility of $\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}}$, a claim to be proved (the definition of $\mathbf{G}$ is clear below, and the notation $\mathbf{M}_{\mathbf{G}}$ is defined in Introduction).

Our study is motivated by two important yet unsolved issues. First, the study of prediction rates has been crucially relying on the assumption that the number of latent factors is correctly estimated. Secondly, the time series that are being studied are often relatively short, to preserve the stationarity. As we explained in Section 2, this leads to strong conditions on the strength of factors of using the PC estimator.

We show below that by allowing $R > r$, the diversified projection does not require a consistent estimator of the number of factors. In addition to the assumptions in Section 2, we impose the following conditions on the forecast equation for y_{t+h}. Let $\mathbf{G}$ be the matrix of $\{\mathbf{g}_t : t \leq T - h\}$.

Assumption 3.1. *(i)* $\{\varepsilon_t, \mathbf{f}_t, \mathbf{g}_t, \mathbf{u}_t : t = 1, \cdots, T+h\}$ *is stationary with* $\mathbb{E}(\mathbf{u}_t|\mathbf{f}_t, \mathbf{g}_t) = 0$ *and* $\mathbb{E}(\varepsilon_t|\mathbf{f}_t, \mathbf{g}_t, \mathbf{u}_t, \mathbf{W}) = 0$.
(ii) Weak dependence: there is $C > 0$, $\max_{s\leq T} \sum_{t\leq T} |\mathbb{E}(\varepsilon_t\varepsilon_s|\mathbf{F}, \mathbf{G}, \mathbf{W})| < C$ *almost surely.*
(iii) Moment bounds: there are $c, C > 0$, $\lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{M}_{\mathbf{G}}\mathbf{F}) > c$, $\lambda_{\min}(\frac{1}{T}\mathbf{G}'\mathbf{M}_{\mathbf{FH}'}\mathbf{G}) > c$, *and* $c < \lambda_{\min}(\frac{1}{T}\mathbf{G}'\mathbf{G}) \leq \lambda_{\max}(\frac{1}{T}\mathbf{G}'\mathbf{G}) < C$.

Our theory *does not* follow from the standard theory of linear models of Bai and Ng (2006). A new technical phenomenon arises when $R > r$ due to the degeneracy of the gram matrices. Define $\widehat{\mathbf{Z}} = (\widehat{\mathbf{z}}_1', ..., \widehat{\mathbf{z}}_{T-h}')'$, $\mathbf{Z} = (\mathbf{z}_1', ..., \mathbf{z}_{T-h}')'$ and consider two gram matrices

$$\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}} = \begin{pmatrix} \widehat{\mathbf{F}}'\widehat{\mathbf{F}} & \widehat{\mathbf{F}}'\mathbf{G} \\ \mathbf{G}'\widehat{\mathbf{F}} & \mathbf{G}'\mathbf{G} \end{pmatrix}, \quad \mathbf{Z}'\mathbf{Z} = \begin{pmatrix} \mathbf{HF}'\mathbf{FH}' & \mathbf{HF}'\mathbf{G} \\ \mathbf{G}'\mathbf{FH}' & \mathbf{G}'\mathbf{G} \end{pmatrix}.$$

The linear regression theory crucially depends on the inverse of $\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}}$, whose population version $\mathbf{Z}'\mathbf{Z}$, in this context, becomes degenerate when $R > r$. The full rank matrix $\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}}$

converges to a degenerate matrix $\mathbf{H}\frac{1}{T}\mathbf{F}'\mathbf{M}_{\mathbf{G}}\mathbf{F}\mathbf{H}'$, and therefore in general

$$\left\| \left(\frac{1}{T}\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}}\right)^{-1} - \left(\frac{1}{T}\mathbf{Z}'\mathbf{Z}\right)^{+} \right\| \neq o_P(1).$$

We develop a new theory that takes advantage of $\mathbf{H}$, which allows to establish the three claims in Section 2.2. They imply that the convergence holds when weighted by $\widetilde{\mathbf{H}}$:

$$\left\| \widetilde{\mathbf{H}}' \left((\frac{1}{T}\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}})^{-1} - (\frac{1}{T}\mathbf{Z}'\mathbf{Z})^{+} \right) \widetilde{\mathbf{H}} \right\| = O_P\left(\frac{1}{T} + \frac{1}{N}\right), \text{ where } \quad \widetilde{\mathbf{H}} = \begin{pmatrix} \mathbf{H} & \\ & \mathbf{I} \end{pmatrix}.$$

The weighted convergence is sufficient to derive the prediction rate of $\widehat{y}_{T+h|T}$.

Theorem 3.1. *Suppose Assumptions 2.1 - 2.4, 3.1 hold. As $T, N \to \infty$, h is bounded, and for all bounded $R \geq r$,*

$$\widehat{y}_{T+h|T} - y_{T+h|T} = O_P(\frac{1}{\sqrt{T}} + \frac{1}{\nu_{\min}\sqrt{N}}).$$

3.2 High-dimensional inference in factor augmented models

3.2.1 Factor-augmented post-selection inference

Consider a high-dimensional regression model

$$\begin{aligned} y_t &= \boldsymbol{\beta}\mathbf{g}_t + \boldsymbol{\nu}'\mathbf{x}_t + \eta_t, \\ \mathbf{g}_t &= \boldsymbol{\theta}'\mathbf{x}_t + \boldsymbol{\varepsilon}_{g,t} \end{aligned} \tag{3.1}$$

where $\mathbf{g}_t$ is a treatment variable whose effect $\boldsymbol{\beta}$ is the main interest. The model contains high-dimensional control variables $\mathbf{x}_t = (x_{1t}, \cdots, x_{Nt})$ that determine both the outcome and treatment variables. Having many control variables creates challenges for statistical inferences, as such, we assume that $(\boldsymbol{\nu}, \boldsymbol{\theta})$ are sparse vectors. Belloni et al. (2014) proposed to make inference using Robinson (1988)'s residual-regression, by first selecting among the high-dimensional controls in both the y_t and $\mathbf{g}_t$ equations.

Often, the control variables are strongly correlated due to the presence of confounding factors

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t. \tag{3.2}$$

This invalidates the conditions of using penalized regressions to directly select among $\mathbf{x}_t$. Instead, if we substitute (3.2) to (3.1), we reach factor-augmented regression model:

$$\begin{aligned} y_t &= \boldsymbol{\alpha}_y' \mathbf{f}_t + \boldsymbol{\gamma}' \mathbf{u}_t + \varepsilon_{y,t}, \\ \mathbf{g}_t &= \boldsymbol{\alpha}_g' \mathbf{f}_t + \boldsymbol{\theta}' \mathbf{u}_t + \varepsilon_{g,t}, \\ \varepsilon_{y,t} &= \boldsymbol{\beta}' \varepsilon_{g,t} + \eta_t \end{aligned} \tag{3.3}$$

where $\boldsymbol{\alpha}_g' = \boldsymbol{\theta}'\mathbf{B}$, $\boldsymbol{\alpha}_y' = \boldsymbol{\beta}\boldsymbol{\alpha}_g' + \boldsymbol{\nu}'\mathbf{B}$, and $\boldsymbol{\gamma}' = \boldsymbol{\beta}\boldsymbol{\theta}' + \boldsymbol{\nu}'$. The model contains high-dimensional latent controls $\mathbf{u}_t$. Here $(\boldsymbol{\alpha}_y, \boldsymbol{\alpha}_g, \boldsymbol{\beta})$ are low -dimensional coefficient vectors while $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ are high-dimensional sparse vectors. Fan et al. (2020) and Hansen and Liao (2018) showed that the penalized regression can be successfully applied to (3.3) to select components in $\mathbf{u}_t$, which are cross-sectionally weakly correlated. They require strong conditions so that we can consistently estimate the number of factors $r = \dim(\mathbf{f}_t)$ first.

The main result of this section is to show that the factor-augmented post-selection inference is valid for any $R \geq r$. Therefore, we have addressed an important question in empirical applications, where the evidence of the number of factors is not so strong and one may use a slightly larger number of "working factors". The theoretical intuition, again, is that the model depends on $\mathbf{f}_t$ only through transformation invariant terms, so that $\widehat{\boldsymbol{\alpha}}_y' \widehat{\mathbf{f}}_t = \boldsymbol{\alpha}_y' \mathbf{H}^{+}\mathbf{H}\mathbf{f}_t + o_P(1) = \boldsymbol{\alpha}_y' \mathbf{f}_t + o_P(1)$. In addition, $\mathbf{u}_t$ can also be well estimated with over-identified number of factors.

Importantly, we admit the special case $r = 0$, and $R \geq 1$, leading to $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_g$ both being zero in (3.3). That is, there are no factors, so $\mathbf{x}_t = \mathbf{u}_t$ itself is cross-sectionally weakly dependent, but nevertheless we estimate $R \geq 1$ number of factors to run post-selection inference. This setting is empirically relevant as it allows to avoid pre-testing the presence of common factors for inference. The simulation in Section 5 shows that with $R \geq r$, this procedure works well even if $r = 0$; but when r ($r \geq 1$) factors are present, directly selecting $\mathbf{x}_t$ leads to severely biased estimations. Therefore as a practical guidance, we recommend that one should always run factor-augmented post-selection inference, with $R \geq 1$, to guard against confounding factors among the control variables.

Below we present the factor-augmented algorithm as in Hansen and Liao (2018) for estimating (3.1). For notational simplicity, we focus on the univariate case $\dim(\boldsymbol{\beta}) = 1$.

Algorithm 3.1. Estimate $\boldsymbol{\beta}$ as follows.

Step 1 Fix the working number of factors R. Estimate $\{(\mathbf{f}_t, \mathbf{u}_t) : t \leq T\}$ as in Section 2.

Step 2 (1) Estimate coefficients: $\widehat{\boldsymbol{\alpha}}_y = (\sum_{t=1}^T \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t')^{-1} \sum_{t=1}^T \widehat{\mathbf{f}}_t y_t$, and $\widehat{\boldsymbol{\alpha}}_g = (\sum_{t=1}^T \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t')^{-1} \sum_{t=1}^T \widehat{\mathbf{f}}_t \mathbf{g}_t$.

(2) Run penalized regression:

$$
\begin{aligned}
\widetilde{\boldsymbol{\gamma}} &= \arg\min_{\boldsymbol{\gamma}} \frac{1}{T} \sum_{t=1}^T (y_t - \widehat{\boldsymbol{\alpha}}_y' \widehat{\mathbf{f}}_t - \boldsymbol{\gamma}' \widehat{\mathbf{u}}_t)^2 + P_\tau(\boldsymbol{\gamma}), \\
\widetilde{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^T (\mathbf{g}_t - \widehat{\boldsymbol{\alpha}}_g' \widehat{\mathbf{f}}_t - \boldsymbol{\theta}' \widehat{\mathbf{u}}_t)^2 + P_\tau(\boldsymbol{\theta}).
\end{aligned}
$$

(3) Run post-selection refitting: let $\widehat{J} = \{j \leq p : \widetilde{\gamma}_j \neq 0\} \cup \{j \leq p : \widetilde{\theta}_j \neq 0\}$.

$$
\begin{aligned}
\widehat{\boldsymbol{\gamma}} &= \arg\min_{\boldsymbol{\gamma}} \frac{1}{T} \sum_{t=1}^T (y_t - \widehat{\boldsymbol{\alpha}}_y' \widehat{\mathbf{f}}_t - \boldsymbol{\gamma}' \widehat{\mathbf{u}}_t)^2, \quad \text{such that } \widehat{\gamma}_j = 0 \text{ if } j \notin \widehat{J}. \\
\widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^T (\mathbf{g}_t - \widehat{\boldsymbol{\alpha}}_g' \widehat{\mathbf{f}}_t - \boldsymbol{\theta}' \widehat{\mathbf{u}}_t)^2, \quad \text{such that } \widehat{\theta}_j = 0 \text{ if } j \notin \widehat{J}.
\end{aligned}
$$

Step 3 Estimate residuals: $\widehat{\boldsymbol{\varepsilon}}_{y,t} = y_t - (\widehat{\boldsymbol{\alpha}}_y'\widehat{\mathbf{f}}_t + \widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{u}}_t)$, and $\widehat{\boldsymbol{\varepsilon}}_{g,t} = \mathbf{g}_t - (\widehat{\boldsymbol{\alpha}}_g'\widehat{\mathbf{f}}_t + \widehat{\boldsymbol{\theta}}'\widehat{\mathbf{u}}_t)$.

Step 4 Estimate $\boldsymbol{\beta}$ by residual-regression:

$$\widehat{\boldsymbol{\beta}} = (\sum_{t=1}^{T} \widehat{\boldsymbol{\varepsilon}}_{g,t}^2)^{-1} \sum_{t=1}^{T} \widehat{\boldsymbol{\varepsilon}}_{g,t}\widehat{\boldsymbol{\varepsilon}}_{y,t}.$$

One can also simplify step 2 following Fan et al. (2020): finding $(\widehat{\boldsymbol{\alpha}}_y, \widehat{\boldsymbol{\gamma}})$ by minimizing $\frac{1}{T}\sum_{t=1}^{T}(y_t - \boldsymbol{\alpha}_y'\widehat{\mathbf{f}}_t - \boldsymbol{\gamma}'\widehat{\mathbf{u}}_t)^2 + P_\tau(\boldsymbol{\gamma})$ and defining $(\widehat{\boldsymbol{\alpha}}_g, \widehat{\boldsymbol{\theta}})$ similarly. Note that $\boldsymbol{\gamma} :\to P_\tau(\boldsymbol{\gamma})$ is a sparse-induced penalty function with a tuning parameter τ. In the main theorem below, we prove for the lasso $P_\tau(\boldsymbol{\gamma}) = \tau\|\boldsymbol{\gamma}\|_1$, where $\|\boldsymbol{\gamma}\|_1 = \sum_{j=1}^{N}|\gamma_j|$. Following Bickel et al. (2009), set

$$\tau = C\sqrt{\frac{\sigma^2 \log N}{T}}$$

for some constant $C > 4$, where $\sigma^2 = \mathrm{var}(\boldsymbol{\varepsilon}_{y,t})$ for estimating $\boldsymbol{\gamma}$, and $\sigma^2 = \mathrm{var}(\boldsymbol{\varepsilon}_{g,t})$ for estimating $\boldsymbol{\theta}$. We refer to Belloni et al. (2014) for feasible tunings so that σ^2 is estimated iteratively.

3.2.2 The main result

We impose the following assumptions.

Assumption 3.2. *(i)* $\mathbb{E}(\boldsymbol{\varepsilon}_{g,t}|\mathbf{u}_t, \mathbf{f}_t, \mathbf{W}) = 0$ *and* $\mathbb{E}(\boldsymbol{\varepsilon}_{y,t}|\mathbf{u}_t, \mathbf{f}_t, \mathbf{W}) = 0$,

(ii) Coefficients: there is $C > 0$, *so that* $\|\boldsymbol{\alpha}_y\|$, $\|\boldsymbol{\alpha}_g\|$, $\|\boldsymbol{\beta}\|$ *are all bounded by* C.

(iii) Weak dependence: There is $C > 0$, *almost surely,*

$\max_{s\le T}\sum_{t\le T}|\,\mathbb{E}(\boldsymbol{\varepsilon}_{y,t}\boldsymbol{\varepsilon}_{y,s}|\mathbf{F}, \mathbf{U}, \mathbf{W})| + \max_{s\le T}\sum_{t\le T}|\,\mathbb{E}(\boldsymbol{\varepsilon}_{g,t}\boldsymbol{\varepsilon}_{g,s}|\mathbf{F}, \mathbf{U}, \mathbf{W})| < C.$

(iv) Uniform bounds:

$\max_{i\le N}\|\frac{1}{T}\sum_{t=1}^{T} u_{it}\mathbf{v}_t\| = O_P(\sqrt{\frac{\log N}{T}})$ *for all* $\mathbf{v}_t \in \{\boldsymbol{\varepsilon}_{g,t}, \boldsymbol{\varepsilon}_{y,t}, \mathbf{f}_t\}$. *In addition,*

$\max_{i\le N}|\frac{1}{T}\sum_{t=1}^{T}(u_{it}u_{jt} - \mathbb{E}\,u_{it}u_{jt})| = O_P(\sqrt{\frac{\log N}{T}})$, *and*

$\max_{i\le N}|\frac{1}{TN}\sum_{t=1}^{T}\sum_{j=1}^{N}(u_{it}u_{jt} - \mathbb{E}\,u_{it}u_{jt})w_{k,j}| = O_P(\sqrt{\frac{\log N}{TN}})$ *for all* $k \le R$.

Assumption 3.2 (iv) holds generally under weak time-series dependent conditions for $\{(\mathbf{v}_t, \mathbf{u}_t) : t \leq N\}$ with sub-Gaussian tails.

Suppose the high-dimensional coefficients $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are strictly sparse. Let J denote the nonzero index set:

$$J = \{j \leq N : \theta_j \neq 0\} \cup \{j \leq N : \gamma_j \neq 0\}.$$

The following *sparse eigenvalue condition* is standard for the post-selection inference. Note that it is imposed on the covariance of $\mathbf{u}_t$ rather than $\mathbf{x}_t$, because $\mathbf{u}_t$ is weakly dependent.

Assumption 3.3 (Sparse eigenvalue condition)**.** *For any* $\mathbf{v} \in \mathbb{R}^N \backslash \{0\}$*, define:*

$$\phi_{\min}(m) = \inf_{\mathbf{v} \in \mathbb{R}^N : 1 \leq \|\mathbf{v}\|_0 \leq m} \mathcal{R}(\mathbf{v}), \qquad \text{and } \phi_{\max}(m) = \sup_{\mathbf{v} \in \mathbb{R}^N : 1 \leq \|\mathbf{v}\|_0 \leq m} \mathcal{R}(\mathbf{v}),$$

where $\mathcal{R}(\mathbf{v}) := \|\mathbf{v}\|^{-2} \mathbf{v}' \frac{1}{T} \sum_{t=1}^T \mathbf{u}_t \mathbf{u}_t' \mathbf{v}$*. Then there is a sequence* $l_T \to \infty$ *and* $c_1, c_2 > 0$ *so that with probability approaching one,*

$$c_1 < \phi_{\min}(l_T |J|_0) \leq \phi_{\max}(l_T |J|_0) < c_2.$$

Assumption 3.4. *(i)* $\frac{1}{T} \sum_{t=1}^T \boldsymbol{\varepsilon}_{g,t}^2 \xrightarrow{P} \sigma_g^2$ *for some* $\sigma_g^2 > 0$.
(ii) $\frac{1}{\sqrt{T}} \sum_{t=1}^T \eta_t \boldsymbol{\varepsilon}_{g,t} \xrightarrow{d} \mathcal{N}(0, \sigma_{\eta g}^2)$ *for some* $\sigma_{\eta g}^2 > 0$*. In addition, there is a consistent variance estimator* $\widehat{\sigma}_{\eta g}^2 \xrightarrow{P} \sigma_{\eta g}^2$.
(iii) The rates $(N, T, |J|_0)$ *satisfy:*

$$|J|_0^4 \log^2 N = o(T), \qquad \text{and } T|J|_0^4 = o(N^2 \min\{1, |J|_0^4 \nu_{\min}^4(\mathbf{H})\}).$$

Condition 3.4 (iii) requires the "effective dimension" $N\nu_{\min}^2(\mathbf{H})$ be relatively large in order to accurately estimate the latent factors.

Theorem 3.2. *Suppose* $\widehat{\mathbf{f}}_t$ *contains* $R \geq r \geq 0$ *number of diversified weighted averages of*

$\mathbf{x}_t$. If $r \geq 1$ (there are factors in $\mathbf{x}_t$), Assumptions 2.1 - 2.4, 3.2-3.4 hold. If $r = 0$ (there are no factors in $\mathbf{x}_t$), Assumption 2.2 is relaxed, and all $\mathbf{f}_t$ involved in the above assumptions can be removed. Then as $T, N \to \infty$, for all bounded $R \geq r \geq 0$,

$$\sigma_{\eta,g}^{-1}\sigma_g^2\sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0,1).$$

Fix a significant level τ, let ζ_τ be the $(1-\tau/2)$ quantile of standard normal distribution. In addition, let $\widehat{\sigma}_g^2 = \frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{\varepsilon}}_{g,t}^2$. Immediately, we have the following uniform coverage.

Corollary 3.1. *Suppose the assumptions of Theorem 3.2 hold. Let $\bar{R} > 0$ be a fixed upper bound for R. Then uniformly for all $0 \leq r \leq R \leq \bar{R}$,*

$$\mathbb{P}\left(\boldsymbol{\beta} \in [\widehat{\boldsymbol{\beta}} \pm \frac{1}{\sqrt{T}}\widehat{\sigma}_{\eta,g}\widehat{\sigma}_g^{-2}\zeta_\tau]\right) \to 1-\tau.$$

The novelty of the above uniformity is that the coverage is valid uniformly for all bounded r as the true number of factors, and all over-estimated R as the working number of factors. In particular, it also admits the weak-dependence $r = 0$ while $R \geq 1$ as a special case.

Remark 3.1 (Case $r = 0, R \geq 1$)**.** We now explain the intuition of the case $\mathbf{x}_t = \mathbf{u}_t$ (no presence of confounding factors), but we nevertheless extract $R \geq 1$ "factors". In this case $\boldsymbol{\alpha}_y = \boldsymbol{\alpha}_g = 0$ in the system (3.3). Then $\widehat{\mathbf{f}}_t = \frac{1}{N}\sum_{i=1}^N \mathbf{w}_i u_{it} := \mathbf{e}_t$ degenerates to zero. Both $\mathbf{u}_t$ and $\boldsymbol{\alpha}_y'\mathbf{f}_t$ (which is zero) are still estimated well in the following sense:

$$\begin{aligned}
\max_{i\leq N}\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})^2 &= O_P\left(\frac{1}{N} + \frac{\log N}{T}\right) \\
\frac{1}{T}\sum_{t=1}^T(\widehat{\boldsymbol{\alpha}}_y'\widehat{\mathbf{f}}_t)^2 &= O_P\left(\frac{|J|_0^2}{N} + \frac{|J|_0^2}{T}\right).
\end{aligned}$$

Remark 3.2 (Case $R = 0$)**.** For completeness of the theorem, we define the estimator for the case $R = 0$. In this case we do not extract any factor estimators, and simply set $\widehat{\boldsymbol{\alpha}}_y = \widehat{\boldsymbol{\alpha}}_g = 0$, and $\widehat{\mathbf{u}}_t = \mathbf{x}_t$ in Algorithm 3.1. This is then the same setting as in Belloni et al. (2014).

3.3 Estimating the idiosyncratic covariance

The estimation of the $N \times N$ idiosyncratic covariance matrix $\mathbf{\Sigma}_u := \mathbb{E}\,\mathbf{u}_t\mathbf{u}_t'$ is of general interest in many applications. Examples include the efficient estimation of factor models (Bai and Li, 2012), high-dimensional testing (Fan et al., 2015), and bootstrapping latent factors (Goncalves and Perron, 2018), among many others. While this problem has been studied by Fan et al. (2013), they require that the true number of factors r has to be either known or consistently estimated, and the factors are estimated through PCA. Here we show that using the diversified factors, their conclusion holds for all fixed $R \geq r$.

A key assumption is that $\mathbf{\Sigma}_u = (\sigma_{u,ij})$ is sparse: As in Bickel and Levina (2008) the sparsity of $\mathbf{\Sigma}_u$ is measured by the following quantity:

$$m_N = \max_{i \leq N} \sum_{j \leq N} |\sigma_{u,ij}|^q, \qquad \text{for some } q \in [0,1].$$

Given the estimated residual $\widehat{u}_{it}$ that is obtained using a working number of factors R, we estimate $\mathbb{E}\,u_{it}u_{jt}$ by applying a generalized thresholding: define $s_{u,ij} := \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt}$,

$$\widehat{\sigma}_{u,ij} = \begin{cases} s_{u,ij}, & \text{if } i = j \\ h(s_{u,ij}, \tau_{ij}), & \text{if } i \neq j \end{cases}$$

where $h(s,\tau)$ is a thresholding function with threshold value τ. Then the sparse idiosyncratic covariance estimator is defined as $\widehat{\mathbf{\Sigma}}_u = (\widehat{\sigma}_{u,ij})_{N \times N}$. The threshold value τ_{ij} is chosen as

$$\tau_{ij} = C\sqrt{s_{u,ii}s_{u,jj}}\,\omega_{NT}, \quad \omega_{NT} := \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$$

for some large constant $C > 0$, which applies a constant thresholding to correlations.

In general, the thresholding function should satisfy:

(i) $h(s, \tau) = 0$ if $|s| < \tau$,

(ii) $|h(s, \tau) - s| \leq \tau$.

(iii) there are constants $a > 0$ and $b > 1$ such that $|h(s, \tau) - s| \leq a\tau^2$ if $|s| > b\tau$.

Note that condition (iii) requires that the thresholding bias should be of higher order. It is not necessary for consistent estimations, but we recommend using nearly unbiased thresholding (Antoniadis and Fan, 2001) for inference applications. One such example is known as SCAD. As noted in Fan et al. (2015), the unbiased thresholding is required to avoid size distortions in a large class of high-dimensional testing problems involving a "plug-in" estimator of $\mathbf{\Sigma}_u$. In particular, this rules out the popular *soft-thresholding* function, which does not satisfy (iii) due to its first-order shrinkage bias.

Theorem 3.3. *Let $\widehat{\mathbf{u}}_t$ be constructed using $R \geq r$ number of diversified weighted averages of $\mathbf{x}_t$. Suppose that Assumptions 2.1 - 2.4 hold and that $\log N = o(T)$. In addition, either $\nu^2_{\min}(\mathbf{H}) \gg \frac{1}{\sqrt{N}}$ or $\nu^2_{\min}(\mathbf{H}) \gg \frac{1}{N}\sqrt{\frac{T}{\log N}}$. Then as $N, T \to \infty$, for any $R \geq r \geq 0$,*

(i)

$$\max_{i \leq N} \frac{1}{T} \sum_{t=1}^{T} (\widehat{\mathbf{b}}_i' \widehat{\mathbf{f}}_t - \mathbf{b}_i' \mathbf{f}_t)^2 = O_P(\omega_{NT}).$$

(ii) For a sufficiently large constant $C > 0$ in the threshold τ_{ij},

$$\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\| = O_P(\omega_{NT}^{1-q} m_N).$$

(iii) If in addition, $\lambda_{\min}(\mathbf{\Sigma}_u) > c_0$ for some $c_0 > 0$ and $\omega_{NT}^{1-q} m_N = o(1)$, then

$$\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_P(\omega_{NT}^{1-q} m_N).$$

3.4 Testing Specification of Factors

In practical applications, many "observed factors" $\mathbf{g}_t$ have been proposed to approximate the true latent factors. For example, in asset pricing, popular choices of $\mathbf{g}_t$ are proposed and discussed in seminal works by Fama and French (1992); Carhart (1997), which are known as the Fama-French factors and Carhart four factor models.

We test the (linear) specification of a given set of empirical factors $\mathbf{g}_t$. That is, we test:

$$H_0 : \text{there is a } r \times r \text{ invertible matrix } \boldsymbol{\theta} \text{ so that } \mathbf{g}_t = \boldsymbol{\theta}\mathbf{f}_t, \quad \forall t \leq T.$$

Under the null hypothesis, $\mathbf{g}_t$ and $\mathbf{f}_t$ are linear functions of each other. We propose a simple statistic:

$$\|\mathbf{P}_{\mathbf{G}} - \mathbf{P}_{\widehat{\mathbf{F}}}\|_F^2$$

where $\mathbf{G} = (\mathbf{g}_1, \cdots, \mathbf{g}_T)'$ and recall that $\mathbf{P}_{(\cdot)}$ denotes the projection matrix. Here we still use the diversified factor estimator $\widehat{\mathbf{F}}$. The test statistic measures the distance between (linear) spaces respectively spanned by $\mathbf{g}_t$ and $\widehat{\mathbf{f}}_t$. To derive the asymptotic null distribution, we naturally set the working number of factors $R = \dim(\mathbf{g}_t)$, which is known and equals $\dim(\mathbf{f}_t) = r$ under the null. Then $\|\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{F}}\|_F = o_P(1)$, followed from Theorem 2.1.

3.4.1 Asymptotic null distribution

With the diversified factor estimators, the null distribution of the statistic is very easy to derive, and satisfies:

$$\frac{N\sqrt{T}(\|\mathbf{P}_{\mathbf{G}} - \mathbf{P}_{\widehat{\mathbf{F}}}\|_F^2 - \text{MEAN})}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1),$$

where for $\mathbf{A} = 2\mathbf{H}'^{-1}(\frac{1}{T}\mathbf{F}'\mathbf{F})^{-1}\mathbf{H}^{-1}$,

$$\text{MEAN} = \frac{1}{N^2}\operatorname{tr}\mathbf{A}\mathbf{W}'\,\mathbb{E}(\mathbf{u}_t\mathbf{u}_t'|\mathbf{F})\mathbf{W}, \qquad \sigma^2 = \operatorname{Var}(\frac{1}{N}\operatorname{tr}\mathbf{A}\mathbf{W}'\mathbf{u}_t\mathbf{u}_t'\mathbf{W}|\mathbf{F},\mathbf{W}) > 0.$$

Here we assume $\sigma^2 > 0$ to be bounded away from zero. To avoid nonparametrically estimating high-dimensional covariances, we shall assume the conditional covariances in both bias and variance are independent of $\mathbf{F}$ almost surely. Nevertheless, the bias depends on a high-dimensional matrix $\mathbf{\Sigma}_u = \mathbb{E}(\mathbf{u}_t\mathbf{u}_t')$. We employ the sparse covariance $\widehat{\mathbf{\Sigma}}_u$ as defined in Section 3.3 and replace the bias by

$$\widehat{\text{MEAN}} := \frac{1}{N^2}\operatorname{tr}\widehat{\mathbf{A}}\mathbf{W}'\widehat{\mathbf{\Sigma}}_u\mathbf{W} \quad \text{with} \quad \widehat{\mathbf{A}} := 2(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}.$$

Further suppose σ can be consistently estimated by some $\widehat{\sigma}$, then together, we have the feasible standardized statistic:

$$\frac{N\sqrt{T}(\|\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{G}}\|_F^2 - \widehat{\text{MEAN}})}{\widehat{\sigma}}. \tag{3.4}$$

The problem, however, is not as straightforward as it looks by far. The use of $\widehat{\text{MEAN}}$ and $\widehat{\sigma}$ both come with issues, as we now explain.

The issue of $\widehat{\text{MEAN}}$**.**

When deriving the asymptotic null distribution, we need to address the effect of $\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u$, which is to show

$$\frac{N\sqrt{T}(\widehat{\text{MEAN}} - \text{MEAN})}{\sigma} \approx \frac{N\sqrt{T}}{\sigma}\frac{1}{N^2}\operatorname{tr}\mathbf{A}\mathbf{W}'(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{W} \xrightarrow{P} 0. \tag{3.5}$$

But simply applying the rate of convergence of $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|$ in Theorem 3.3 fails to show the above convergence, even though the rate is minimax optimal [3]. Similar phenomena also arise in Fan et al. (2015); Bai and Liao (2017), where a plug-in estimator for $\mathbf{\Sigma}_u$ is used for inferences. Proving (3.5) requires a dedicated technical argument to address the accumulation of high-dimensional estimation errors. It requires a strengthened condition on the weak cross-sectional dependence, in Assumption 3.8 below.

The issue of $\widehat{\sigma}$.

It is difficult to estimate σ through residuals $\widehat{\mathbf{u}}_t$ since $\mathbf{W}'\widehat{\mathbf{u}}_t = 0$ almost surely. In fact, estimated $\mathbf{u}_t$ constructed based on any factor estimator would lead to *inconsistent* estimator for σ^2. Therefore, we propose to estimate σ^2 by parametric bootstrap. Observe that $\frac{1}{\sqrt{N}}\mathbf{W}'\mathbf{u}_t$ is asymptotically normal, whose variance is given by $\mathbf{V} = \frac{1}{N}\mathbf{W}'\mathbf{\Sigma}_u\mathbf{W}$. Hence σ^2 should be approximately equal to

$$f(\mathbf{A}, \mathbf{V}) := \mathrm{Var}(\frac{1}{N}\operatorname{tr}\mathbf{A}\mathbf{W}'\mathbf{Z}_t\mathbf{Z}_t'\mathbf{W}), \tag{3.6}$$

where $\mathbf{Z}_t$ is distributed as $\mathcal{N}(0, \mathbf{V})$. Therefore we estimate σ^2 by

$$\widehat{\sigma}^2 = f(\widehat{\mathbf{A}}, \widehat{\mathbf{V}}), \quad \text{with } \widehat{\mathbf{V}} = \frac{1}{N}\mathbf{W}'\widehat{\mathbf{\Sigma}}_u\mathbf{W},$$

which can be calculated by simulating from $\mathcal{N}(0, \widehat{\mathbf{V}})$.

Above all, despite of the simple construction of $\widehat{\mathbf{F}}$, the technical problem is still challenging. Therefore, this subsection calls for relatively stronger conditions, as we now impose.

Assumption 3.5. *(i) $\{\mathbf{u}_t : t \leq T\}$ are stationary and conditionally serially independent, given $\mathbf{F}$ and $\mathbf{G}$.*
(ii) There is $C > 0$, $\mathbb{E}[\|\frac{1}{\sqrt{N}}\mathbf{W}'\mathbf{u}_t\|^4|\mathbf{W}] < C$.
(iii) $\nu_{\min}(\mathbf{H}) > c$ for some $c > 0$.

The next assumption ensures that σ^2 can be estimated by simulating from the Gaussian

[3] A simple calculation would only yield $\frac{N\sqrt{T}}{\sigma}\frac{1}{N^2}\|\mathbf{A}\mathbf{W}'\|\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|\|\mathbf{W}\| \leq O_P(1)$ but not necessarily $o_P(1)$.

distribution.

Assumption 3.6. *(i) There is $c > 0$ so that $\sigma^2 > c$.*

(ii) As $N \to \infty$, $|\sigma^2 - f(\mathbf{A}, \mathbf{V})| \to 0$ almost surely in $\mathbf{F}$, where $f(\mathbf{A}, \mathbf{V})$ is given in (3.6).

Next, we shall require $\mathbf{\Sigma}_u$ be strictly sparse, in the sense that the "small" off-diagonal entries are exactly zero. In this case, we use the following measurement for the total sparsity:

$$D_N := \sum_{i,j \leq N} 1\{\mathbb{E}\, u_{it}u_{jt} \neq 0\}.$$

Recall that $\omega_{NT} := \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$. We assume:

Assumption 3.7 (Strict sparsity)**.** *(i)* $(\frac{\omega_{NT}^2\sqrt{T}}{N})D_N \to 0.$

(ii) $\min\{|\mathbb{E}\, u_{it}u_{jt}| : \mathbb{E}\, u_{it}u_{jt} \neq 0\} \gg \omega_{NT}$.

For block-diagonal matrices with finite block sizes, $D_N = O(N)$; for banded matrices with band size l_N, $D_N = O(l_N N)$. In general, suppose $D_N = l_N N$ with some slowly growing $l_N \to \infty$. Then condition (i) reduces to requiring $l_N^2 \log N \ll l_N\sqrt{T} \ll N$. This requires an upper bound for l_N; in addition, the lower bound for N arises from the requirement of estimating factors. Condition (ii) requires that the nonzero entries are well-separated from the statistical errors.

Assumption 3.8. *Write $\sigma_{u,ij} := \mathbb{E}\, u_{it}u_{jt}$. There is $C > 0$ so that*

$$\frac{1}{N} \sum_{(m,n):\sigma_{u,mn}\neq 0,} \sum_{(i,j):\sigma_{u,ij}\neq 0} |\operatorname{Cov}(u_{it}u_{jt}, u_{mt}u_{nt})| < C.$$

The above assumption is the key condition to argue for (3.5). It requires further conditions on the weak cross-sectional dependence, in addition to the sparsity. Fan et al. (2015)

proved that if u_{it} is Gaussian, then a sufficient condition for Assumption 3.8 is as follows:

$$D_N = O(N), \text{ and } \max_{i\leq N} \sum_{j\leq N} 1\{\mathbb{E}\, u_{it}u_{jt} \neq 0\} = O(1),$$

which is the case for block diagonal matrices with finite members in each block and banded matrices with $l_N = O(1)$.

Theorem 3.4. *Suppose $R = \dim(\mathbf{g}_t)$, and Assumptions 2.1 - 2.4, 3.5- 3.8 hold. As $N, T \to \infty$, under H_0,*

$$\frac{N\sqrt{T}(\|\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{G}}\|_F^2 - \widehat{\text{MEAN}})}{\widehat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

3.5 Factor-adjusted false discovery control for multiple testing.

Controlling the false discovery rate (FDR) in large-scale hypothesis testing based on strongly correlated testing series has been an important problem. Suppose the data are generated from:

$$\mathbf{x}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t,$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)'$ is the mean vector. This model allows strong cross-sectional dependences among $\mathbf{x}_t$. We are interested in testing N number of hypotheses:

$$H_0^i : \alpha_i = 0, \quad i = 1, ..., N.$$

The FDR control aims to develop test statistics Z_i and threshold values so that the overall false discovery rate is controlled at certain value. A crucial requirement is that these test statistics should be weakly dependent. However, for $\bar{\mathbf{f}} = \frac{1}{T}\sum_t \bar{\mathbf{f}}_t$ and $\bar{\mathbf{u}} = \frac{1}{T}\sum_t \bar{\mathbf{u}}_t$, we have $\bar{\mathbf{x}} = \frac{1}{T}\sum_t \bar{\mathbf{x}}_t = \boldsymbol{\alpha} + \mathbf{B}\bar{\mathbf{f}} + \bar{\mathbf{u}}$, so the presence of $\mathbf{B}\mathbf{f}_t$ makes the mean vector be cross-sectionally strongly dependent, failing usual FDR procedures based on the simple sample average. This is the well known confounding factor problem. While several methods have been proposed to remove the effect of confounding factors (Wang et al., 2017; Fan et al., 2019), again, it has been assumed that the number of factors should be consistently estimable.

The diversified projection can be applied directly as a simple implementation for the FDR control, valid for all $R \geq r$. Let the diversified projection be $\widehat{\mathbf{f}}_t = \frac{1}{N}\mathbf{W}'\mathbf{x}_t$, and let $\widehat{\mathbf{b}}_i$ be the OLS estimator for the slope vector by regressing x_{it} on $\widehat{\mathbf{f}}_t$ with intercept. Then we can define the factor-adjusted regularized multiple test (Fan et al., 2019) statistics $Z_i = \widehat{\alpha}_i / se(\widehat{\alpha}_i)$ where

$$\widehat{\alpha}_i = \bar{x}_i - \widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}, \quad \widehat{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t,$$

and $se(\widehat{\alpha}_i)$ is the associated standard error. Our theories imply the following expansion, uniformly for $i = 1, ..., N$ and all $R \geq r$,

$$\widehat{\alpha}_i - \alpha_i \;\; = \;\; \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_t u_{it} + o_P(T^{-1/2}),$$

where $\mathbf{g}_t = 1 - \bar{\mathbf{f}}'\mathbf{S}_f^{-1}(\mathbf{f}_t - \bar{\mathbf{f}})$, and $\mathbf{S}_f = \frac{1}{T}\sum_t(\mathbf{f}_t - \bar{\mathbf{f}})(\mathbf{f}_t - \bar{\mathbf{f}})'$. This gives rise to the desired expansion so that Z_i are weakly dependent. Therefore, we can apply standard procedures to Z_i for the false discovery control.

4 Choices of Diversified Weights

We discuss some specific examples to choose $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_R) = (w_{k,i} : k \leq R, i \leq N)$, the weight matrix.

4.1 Loading characteristics

Factor loadings are often driven by observed characteristics. For example, in genetic studies, single-nucleotide polymorphism (SNP) data are often collected with the gene expression data on the same group of subjects. The SNPs drive underlying structure in the gene expressions, clinical and demographics data, through affecting their loadings on the biological factors. In asset pricing studies, it has been well documented that factor loadings are driven by firm specific characteristics, which are independent of the model noise, but have strong explanatory powers on the loadings.

Motivated by the presence of characteristics, "characteristic based" factor models have been extensively studied in the literature, e.g., Gagliardini et al. (2016); Li et al. (2016); Connor et al. (2012). The general form of this model assumes the loadings have the following decomposition (Fan et al., 2016):

$$\mathbf{b}_i = \mathbf{g}(\mathbf{z}_i) + \boldsymbol{\gamma}_i, \quad \mathbb{E}(\boldsymbol{\gamma}_i|\mathbf{z}_i) = 0, \quad i \leq N, \tag{4.1}$$

where $\mathbf{z}_i$ is a vector of characteristics that are observed on each subject and $\mathbf{g}(\cdot)$ is a non-parametric mean function. It is assumed that $\{\mathbf{z}_i : i \leq N\}$ is independent of $\mathbf{u}_t$ and that $\mathbf{g}(\mathbf{z}_i)$ is not degenerate so that $\mathbf{z}_i$ has explanatory power. In addition, $\boldsymbol{\gamma}_i$ is the remaining loading components, after conditioning on $\mathbf{z}_i$. The decomposition of $\mathbf{b}_i$ in (4.1) is motivated from the asset pricing literature, where factor "betas" are known to be partially explained by individual-specific observables $\mathbf{z}_i$, which represent a set of time-invariant characteristics such as individual stocks' size, momentum, values. When $\mathbf{z}_i$ is available, we can employ them as a natural choice of the weights for the diversified factors. Fix an R-component of sieve basis functions: $(\phi_1(\cdot), ..., \phi_R(\cdot))$ such as the Fourier basis or B splines. Then define

$$\mathbf{W} := (w_{i,k})_{N\times R}, \qquad \text{where } w_{i,k} = \phi_k(\mathbf{z}_i).$$

The diversified projection using the so-constructed $\mathbf{W}$ is related to the "projected PCA" of Fan et al. (2016), but the latter is more complicated and requires stronger conditions than the diversified projection, because it is still PCA based.

4.2 Moving window estimations

This method is useful when $\mathbf{u}_t$ is serially independent, and related ideas have been used recently by Barigozzi and Cho (2018). Consider out-of-sample forecasts using moving windows. Suppose $\mathbf{x}_t$ is observed for $T + T_0$ periods in total, but to pertain the stationarity assumption, we only use the most recent T observations to learn the latent factors, where T may be potentially small. Divide the sample into two periods:

$$\begin{aligned}
&\text{periods (I) of learning weights:} \quad \mathbf{x}_t = \mathbf{B}_1\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, ..., T_0 \\
&\text{periods (II) of interest:} \quad \mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t = T_0 + 1, ..., T_0 + T.
\end{aligned}$$

While $\mathbf{B}_1$ and $\mathbf{B}$ can be different (e.g., presence of structural breaks), they are assumed to be closely related between two sampling periods. As such, we can learn about the diversified weights from periods (I) to estimate the latent factors for the periods of estimation interests (II). Specifically, apply PCA on periods (I) to extract R number of factor loadings: $\widehat{\mathbf{B}}_1 = (\widehat{b}_{i,k})_{N\times R}$. Now for a pre-determined constant $\epsilon > 0$, define $\mathbf{W} = (w_{i,k})_{N\times R}$ where

$$w_{i,k} = \frac{\widehat{b}_{i,k}}{\max\{1, \epsilon \max_{i\leq N} |\widehat{b}_{i,k}|\}}, \quad k \leq R, \quad i \leq N.$$

Barigozzi and Cho (2018) suggested a specific choice for ϵ that work well in their simulation studies. As discussed by these authors, the trimming constant ϵ ensures that with a large probability most of the corresponding $\widehat{b}_{i,k}$ are "preserved" by $w_{i,k}$. On the other hand, if a few elements of $\widehat{b}_{i,k}$ are spuriously large in finite sample, the trimming shrinks the corresponding

$\widehat{b}_{i,k}$ downwards to $1/\epsilon$.

In addition, if $\mathbf{u}_t$ is serially independent, then $\mathbf{W}$ is also independent of $\mathbf{u}_t$ for $t = m+1, ..., m+T$. As such, the conditions on the diversified weights are satisfied. It is straightforward to extending this idea to multi-periods rolling window forecasts, where weights are sequentially updated for rolling windows.

The aforementioned method uses the idea that sample splitting creates serial independences. In the presence of mixing-type serial dependences, Barigozzi and Cho (2018) proposed to split the data into blocks and estimate factor loadings using subsamples omitting the current block as well as its immediate neighbors. Their method can be also applied in the current context to create the weighting matrix.

4.3 Initial Transformation

A related idea is to use transformations of the initial observation $\mathbf{x}_t$ for $t = 0$. Suppose $(\mathbf{f}_0, \mathbf{u}_0)$ is independent of $\{\mathbf{u}_t : t \geq 1\}$, and let $\{\phi_k : k = 1, ..., R\}$ be a set of sieve transformations. Then we can apply $w_{i,k} = \phi_k(x_{i,0})$. These weights are correlated with $\mathbf{B}$ through $\mathbf{x}_0 = \mathbf{B}\mathbf{f}_0 + \mathbf{u}_0$ so that the rank condition is satisfied. The initial transformation method only requires $\{\mathbf{u}_t\}$ be independent of its initial value. The similar idea has been used recently by Juodis and Sarafidis (2020).

4.4 Hadamard projection

We can set deterministic weights as in the statistical experimental designs:

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 & 1 & \dots \\ 1 & -1 & 1 & 1 & \\ 1 & 1 & -1 & 1 & \\ 1 & -1 & -1 & -1 & \dots \\ 1 & 1 & 1 & -1 & \\ 1 & -1 & 1 & -1 & \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

So for each $2 \leq k \leq R$, the k th column of $\mathbf{W}$ equals $(1'_{k-1}, -1'_{k-1}, 1'_{k-1}, -1'_{k-1}, \dots)$, where 1_m denotes the m-dimensional vector of ones. Closely related types of matrices are known as the Walsh-Hadamard matrices, formed by rearranging the columns so that the number of sign changes in a column is in an increasing order, and the columns are orthogonal. Therefore, we can also set $\mathbf{W}$ as the $N \times R$ upper-left corner submatrix of a Hadamard matrix of dimension 2^K with $K = \lceil \log_2 N \rceil$, where $\lceil . \rceil$ denotes the ceiling function.

5 Monte Carlo Experiments

In this section we illustrate the finite sample properties of the forecasting and inference methods based on diversified factors, and use four types of weight matrices:

(i) Hadamard weight: $\mathbf{w}_1 = \mathbf{1}$ and $\mathbf{w}_k = (\mathbf{1}'_{k-1}, -\mathbf{1}'_{k-1}, \mathbf{1}'_{k-1}, -\mathbf{1}'_{k-1}, \dots)$ for $2 \leq k \leq R$, where $\mathbf{1}_{k-1}$ is a vector of one's of length $k-1$.

(ii) Loading characteristics: loadings depend on some characteristics $\mathbf{z}_i$, and we apply the polynomial transformations so that the i th row of $\mathbf{W}$ is $(g_1(\mathbf{z}_i), g_2(\mathbf{z}_i), ..., g_R(\mathbf{z}_i))$ for

$i \leq N$. In our numerical work, we take one characteristic and set $g_j(\mathbf{z}_i) = \mathbf{z}_i^j$.

(iii) Rolling windows: when conducting simulations for out-of-sample forecasts, we use the trimmed PCA as described in Section 4.2.

(iv) Initial transformations: we use the initial transformation so that the i th row of $\mathbf{W}$ is $(x_{i,0}, x_{i,0}^2, ..., x_{i,0}^R)$ for $i \leq N$.

We generate the data from the following model motivated from Section 4.1:

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad \mathbf{B} = (b_{i,k}) * N^{-(1-\alpha)/2}, \quad \text{with } b_{i,k} = (\mathbf{z}_i^k + 0.5\gamma_{i,k}).$$

We set $\mathbf{z}_i = \sin(h_i)$ where both h_i and $\gamma_{i,k}$ are independent scalar standard normal variables. Here we use the polynomial transformation $\mathbf{z}_i^k$ to represent the effect of characteristics. In addition, the $\gamma_{i,k}$-component captures the unobservable beta components that are not explainable by the characteristics. With the identification condition $\mathbb{E}(\gamma_{i,k}|\mathbf{z}_i) = 0$, both components in $b_{i,k}$ can be consistently estimated. See more motivations of this model in Fan et al. (2016) and Kim et al. (2018). The multiplier $N^{-(1-\alpha)/2}$ measures the strength of the factors, whereas the spiked eigenvalue of the sample covariance grows at rate N^α. Hence larger α indicates stronger factors.

The factors are multivariate standard normal. To generate the idioscyncratic term, we set the $N \times T$ matrix $\mathbf{U} = \mathbf{\Sigma}_N^{1/2}\bar{\mathbf{U}}\mathbf{\Sigma}_T^{1/2}$; here $\bar{\mathbf{U}}$ is an $N \times T$ matrix, whose entries independent standard normal. The $N \times N$ matrix $\mathbf{\Sigma}_N$ and the $T \times T$ matrix $\mathbf{\Sigma}_T$ respectively govern the cross-sectional and serial correlations of u_{it}. We set $\mathbf{\Sigma}_T = (\rho_T^{|t-s|})_{st}$, and use a sparse cross-sectional covariance:

$$\mathbf{\Sigma}_N = \mathrm{diag}\{\underbrace{\mathbf{A}, \cdots, \mathbf{A}}_{n \text{ of them}}, \mathbf{I}\}, \quad \mathbf{A} = (\rho_N^{|i-j|}) \tag{5.1}$$

where $\mathbf{A}$ is a small four-dimensional block matrix and $\mathbf{I}$ is $(N - 4n) \times (N - 4n)$ identity matrix so that $\mathbf{\Sigma}_N$ has a block-diagonal structure. We fix $n = 3$ and $\rho_N = 0.7$. The numerical performances are studied in the following subsection with various choice of ρ_T to test about the sensitivity against serial correlations.

5.1 Covariance estimation

We first study the performance of estimating $\mathbf{\Sigma}_u$. To do so, we set $r = 1$ and respectively calculate $\widehat{\mathbf{\Sigma}}_u$ using $R = r, \cdots, r + 3$. As estimating $\mathbf{\Sigma}_u$ is particularly important in asset pricing models, we use the loading characteristic weights $w_{i,k} = \mathbf{z}_i^k$, $k = 1, ..., R$, as the characteristic $\mathbf{z}_i$ is often directly observable along with the return data.

For comparison purposes, we also estimate $\mathbf{\Sigma}_u$ using two benchmark estimators:

(i) The PC-estimator for factors with $R = r$ (the POET method by Fan et al. (2013)). So the PC-estimator in this simulation assumes the true number of factors $r = 1$ to be known;

(ii) The known-factor method. We use the true factors, and estimate loadings and u_{it} by OLS, followed by SCAD-thresholding.

We set two serial dependence scenarios: $\rho_T = 0.1$ (weak serial dependence) and $\rho_T = 0.7$ (strong serial dependence), as well as two factor-strength scenarios: $\alpha = 1$ and $\alpha = 0.5$.

Figure 1 plots $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|$ and $\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|$, averaged over 100 replications, as $N = T$ grows. While all estimators perform similarly, the POET-estimator is not always better than the diversifying projection (DP). For estimating $\mathbf{\Sigma}_u$, both the DP with $R = r$ and the known factor method are overall better than the POET estimator, followed by DP with other choices of R. This comparison is reasonable, reflecting the robustness of DP to the serial conditions and strength of factors. Perhaps what is surprising is the comparison for estimating the inverse covariance. In all four scenarios of the factor strength and serial correlations, the DP with $R = r$ performs the worst among the six estimators, and DP with over estimated R is in general better than both the known factor method and the POET. Our interpretation of this is that we set relatively strong cross-sectional correlations in the data generating process, making $\mathbf{\Sigma}_u^{-1}$ more unstable. The use of more diversified weights provides extra information to help stabilizing the inverse covariance estimator.

5.2 Out-of-sample forecast

We assess the performance of the proposed factor estimators on out-of-sample forecasts. Consider the following forecast model

$$y_{t+1} = \beta_0 + \beta y_t + \boldsymbol{\alpha}'\mathbf{f}_t + \varepsilon_{t+1}$$

where we set $r = \dim(\mathbf{f}_t) = 2$, $(\beta_0, \beta) = (1.5, 0.5)$, and $\boldsymbol{\alpha} = (1, 1)'$. In addition, ε_t are independent standard normal. The data generating process for $\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ is the same as before, in the presence of both serial and cross-sectional correlations. We conduct one-step-ahead out-of-sample forecast m times using a moving window of size T. Here T is also the sample size for estimations. We simulate $m + T$ observations in total. For each $t = 0, \cdots, m-1$, we use the data $\{(\mathbf{x}_{t+1}, y_{t+1}), \cdots, (\mathbf{x}_{t+T}, y_{t+T})\}$ to conduct one-step-ahead forecast of y_{t+T+1}. Specifically, we estimate the factors using $\{\mathbf{x}_{t+1}, \cdots, \mathbf{x}_{t+T}\}$, and obtain $\{\widehat{\mathbf{f}}_{t+1}, \cdots, \widehat{\mathbf{f}}_{t+T}\}$. The coefficients in the forecasting regression is then estimated by the OLS, denoted by $(\widehat{\beta}_{0,t+T}, \widehat{\beta}_{t+T}, \widehat{\boldsymbol{\alpha}}_{t+T})$. We then forecast y_{t+T+1} by

$$\widehat{y}_{t+T+1|t+T} = \widehat{\beta}_{0,t+T} + \widehat{\beta}_{t+T} y_{t+T} + \widehat{\boldsymbol{\alpha}}_{t+T}' \widehat{\mathbf{f}}_{t+T}.$$

Such a procedure continues for $t = 0, \cdots, m-1$.

We compute the diversified factor estimators using the two types of weights, with $R = r, r+1, r+3$ as the working number of factors. As for the moving windows weight, we

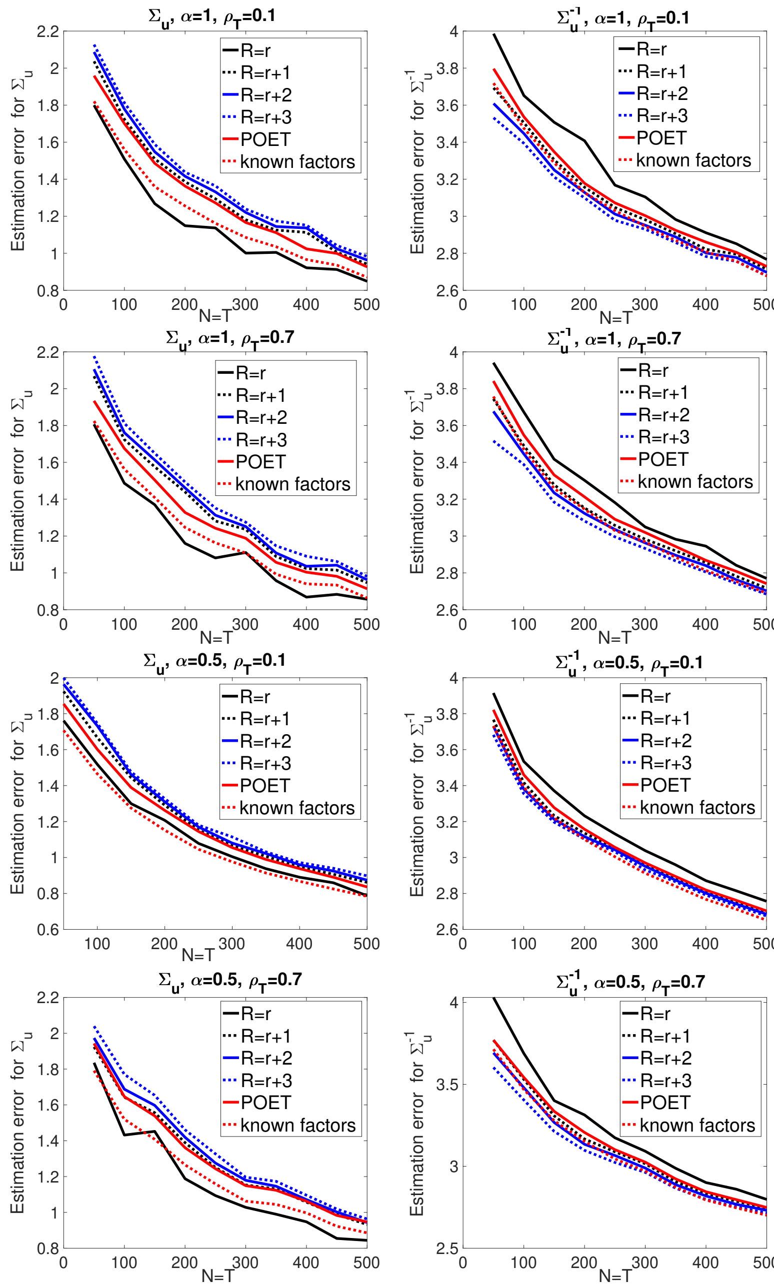

Figure 1: The estimation errors in operator-norm $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|$ (left) and $\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|$ (right) as the dimension increases, averaged over 100 replications. We set $N = T$. Here $R = r, \cdots, r+3$ correspond to the diversified factor estimators using R number of working factors. Characteristic weights are used. Here α measures the factor strength and ρ_T is the serial correlation.

Table 1: Out-of-Sample MSE(M)/MSE(PC) for three types of estimators.

			Characteristic weights			Rolling window weights			GDF		KF
							R				
ρ_T	N	T	r	$r+1$	$r+3$	r	$r+1$	$r+3$	3	4	r
							$\alpha = 1$				
0	100	50	1.141	1.090	1.109	0.968	1.001	1.010	0.991	1.016	1.007
		100	0.998	0.980	1.035	0.979	1.039	1.046	1.008	1.009	1.002
0.5		50	0.996	1.008	0.965	0.993	1.018	1.055	1.000	0.996	0.986
		100	0.885	0.886	0.917	0.937	0.922	0.939	0.995	0.997	1.005
0.9		50	0.602	0.621	0.637	0.608	0.620	0.680	0.763	0.772	1.023
		100	0.434	0.458	0.482	0.422	0.419	0.450	0.863	0.578	0.985
							$\alpha = 0.2$				
0		50	0.876	0.913	0.987	1.072	1.059	1.071	0.991	0.985	1.003
		100	0.931	0.906	0.966	1.065	1.114	1.156	0.996	1.012	0.992
0.5		50	0.891	0.897	1.044	1.059	1.082	1.149	1.002	0.981	0.958
		100	0.972	0.963	0.970	0.868	0.793	0.817	0.968	0.981	1.007
0.9		50	0.478	0.513	0.647	0.713	0.731	0.688	0.953	0.745	0.966
		100	0.762	0.765	0.767	0.788	0.806	0.849	0.927	0.851	0.951

Reported are the out-of-sample relative MSEs. The benckmark PC-estimator uses the true number of factors. The dimension $N = 100$ is fixed. The diversified projection uses R estimated factors with two types of weights: characteristic weights and rolling window weights. In addition, the columns of GDF estimates factors from the generalized dynamic factor model of Forni et al. (2005), with R number of dynamic factors. The Matlab codes for implementing Forni et al. (2005) and Hallin and Liška (2007) are downloaded from Matteo Barigozzi's website `www.barigozzi.eu/codes.html.` The column of KF refers to the Kalman filtering developed by Doz et al. (2011), which uses the true r number of factors. Both GDF and KF specifically estimate dynamic factors.

assume there is a historical time series $\mathbf{x}_t = \mathbf{B}_1\mathbf{f}_t + \mathbf{u}_t$, for $t = -T, ..., 0$, and the loadings $\mathbf{B}_1$ is correlated with $\mathbf{B}$ in the sense that $\mathbf{B}_1 = 0.8\mathbf{B} + 0.5\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal. We then apply the moving window method to create $\mathbf{W}$ as outlined in Section 4.2. Though the theory for the moving window weights requires serial correlation $\rho_T = 0$, we nevertheless set $\rho_T = 0, 0.5$ and 0.9 to examine the performance under serially correlated series.

The benchmark method is the PC-estimator, which uses the true number of factors. In addition, we also consider two well known methods that specifically estimate factor dynamics:

(i) GDF: the generalized dynamic factor model of Forni et al. (2005). The selection criterion of Hallin and Liška (2007) recommended using, on average, three dynamic factors, so we use $R = 3, 4$ numbers of dynamic factors.

(ii) KF: the two-step Kalman filtering of Doz et al. (2011). In the first step factors are preliminarily estimated and fit a VAR model; in the second step, Kalman smoother is applied to calculate the projection onto the observations. For this approach, we use $R = 2$, the true number of factors.

For each method M, we calculate the mean squared out-of-sample forecasting error:

$$\mathrm{MSE}(\mathrm{M}) = \frac{1}{m}\sum_{t=0}^{m-1}(y_{t+T+1} - \widehat{y}_{t+T+1|t+T})^2,$$

and report the relative MSE to the PC method: MSE(M)/MSE(PC). It is worthwhile to emphasize that this study does not aim to beat the PC-method. In fact, the PC-estimator yields the optimal rank r-estimation of the low-rank structure, in the sense that the estimated low-rank component $\mathbf{BF}'$ satisfies: $\widehat{\mathbf{B}}_{pc}\widehat{\mathbf{F}}'_{pc} = \arg\min_{\mathrm{rank}(\mathbf{A})=r} \|\mathbf{X}-\mathbf{A}\|_F^2$. So when the number of factors r is correctly specified and the time series dependence is not strong, the PC-estimator enjoys some optimal property. Nevertheless we use PC as the benchmark as it is the most commonly used in this literature. We aim to see how well the proposed DP method performs relative to the benchmark.

The results are reported in Table 1 for $m = 50$, and is computed based on one set of simulation replications. We see that the DP with various R and Generalized DF are in most scenarios similar to the PC-estimator, and DP outperforms under the strong serial correlations. In all cases, Kalman filtering is comparable with PC, including the case of strong serial correlations.

5.3 Post-selection inference

We now study the inference for the effect of $\mathbf{g}_t$ in the following factor-augmented model

$$
\begin{aligned}
y_t &= \boldsymbol{\beta}\mathbf{g}_t + \boldsymbol{\nu}'\mathbf{x}_t + \eta_t, \\
\mathbf{g}_t &= \boldsymbol{\theta}'\mathbf{x}_t + \boldsymbol{\varepsilon}_{g,t} \\
\mathbf{x}_t &= \mathbf{B}\mathbf{f}_t + \mathbf{u}_t,
\end{aligned}
$$

where both $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ are set to high-dimensional sparse vectors. The goal is to make inference about $\boldsymbol{\beta}$, using the factor-augmented post-selection inference. We generate $\mathbf{u}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_u)$, $(\eta_t, \boldsymbol{\varepsilon}_{g,t}) \sim \mathcal{N}(0, \mathbf{I})$. We set $(\mathbf{u}_t, \boldsymbol{\varepsilon}_{g,t}, \eta_t)$ be serially independent, but still allow the same cross-sectional dependence among $\mathbf{u}_t$. This allows us to focus on the effect of over-estimating factors. The r-dimensional $\mathbf{f}_t$ are independent standard normal. We set the true $\boldsymbol{\beta} = 1$, $\boldsymbol{\theta} = \boldsymbol{\nu} = (1, -1.5, 0.5, 0, ..., 0)$ and $T = N = 200$.

We employ the diversified factor estimator described in Section 3.2 with various working number of factors R, and compare with the benchmark "double-selection" method of Belloni et al. (2014). In particular, we consider two settings:

(i) $r = 0$: there are no factors so $\mathbf{x}_t$ itself is weakly dependent.

(ii) $r = 2$: there are two factors driving $\mathbf{x}_t$. Set $\alpha = 1$ so both factors are strong.

We calculate the standardized estimates: $z := \widehat{\sigma}_{\eta,g}^{-1}\widehat{\sigma}_g^2\sqrt{T}(\widehat{\beta} - \beta)$, where the standard error is the estimated feasible one. Our theory shows that the sampling distribution of z should be approximately standard normal.

Figures 5.2 and 5.3 plot the histograms of the standardized estimates over 200 replications, superimposed with the standard normal density. The histogram is scaled to be a density function. We present the results when the initial transformation are used as weights for the diversified factors. The results from characteristics and Hadamard weights are very similar. When $r = 0$, while it is expected that the double selection performs very well, as is shown in Figure 5.3, using $R \geq 1$ factors also produces z-statistics whose distribution is also close to the standard normality. This shows that the factor-augmented method is robust to the absence of factor structures. On the other hand, when $r = 2$, the factor-augmented method continues to perform well. In contrast, the double selection is severely biased, and the distribution of its z-statistic is far off from the standard normality.

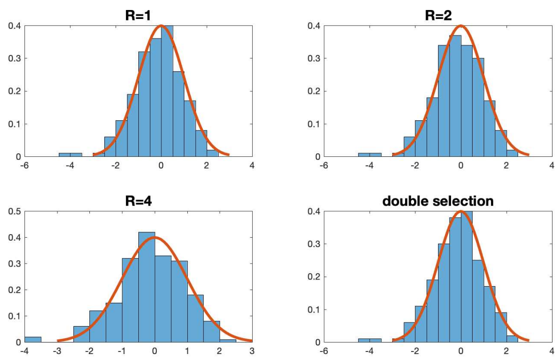

Figure 2: true $r = 0$

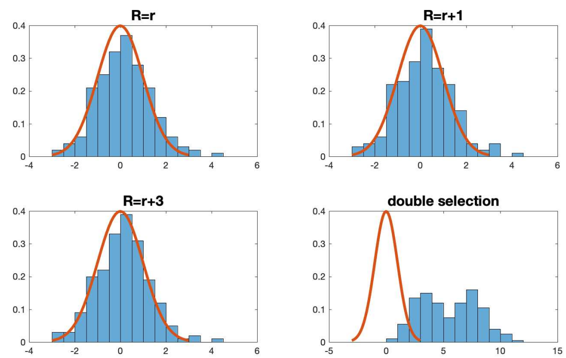

Figure 3: true $r = 2$

The first three panels employ the diversified factor estimator with R number of working factors. The last panel uses the double selection, which directly selects among $\mathbf{x}_t$. The weights used are the initial transformations ($t = 0$) so that the i th row of $\mathbf{W}$ is $(x_{i,0}, x_{i,0}^2, ..., x_{i,0}^R)$ for $i \leq N$.

5.4 Testing the specification of empirical factors

In the last simulation study, we study the size and power of the test statistic for $H_0 : \mathbf{g}_t = \boldsymbol{\theta}\mathbf{f}_t$ for some $r \times r$ invertible matrix $\boldsymbol{\theta}$. Here $\mathbf{g}_t$ is a vector of known "empirical factors" that applied researchers propose to approximate the true factors. We generate

$$\mathbf{g}_t = \boldsymbol{\theta}\mathbf{f}_t + \gamma\mathbf{h}_t, \quad t \leq T,$$

where $\boldsymbol{\theta}$ is an r-dimensional identity matrix, and $(\mathbf{f}_t, \mathbf{h}_t) \sim \mathcal{N}(0, \mathbf{I})$. Here γ governs the strength of the alternatives. We assume that $\mathbf{u}_t$ is serially independent normal generated from $\mathcal{N}(0, \mathbf{\Sigma}_N)$, with $\mathbf{\Sigma}_N$ as in (5.1), pertaining the same cross-sectional dependence. We set $R = r = 2$ and fix $N = 200$. In each of the simulations, we calculate the test statistic as defined in Section 3.4, and set the significance level to 0.05. We use the SCAD-thresholding to estimate $\mathbf{\Sigma}_u$ for both $\widehat{\text{MEAN}}$ and $\widehat{\sigma}$.

Table 2 presents the rejection probability over 1000 replications, with $\gamma = 0$ representing the size of the test. Above all, the results look satisfactory with controlled size and reasonable powers, while weights using initial transformations have some size distortions.

Table 2: Probability of rejection at level 0.05. γ represents the strength of alternatives.

γ	T	Characteristic weights	Hadamard weights	Initial transformation
0	100	0.054	0.046	0.065
	200	0.052	0.047	0.074
0.2	100	1.000	0.998	1.000
	200	0.975	1.000	1.000

6 Conclusion

We propose simple estimators of the latent factors using cross-sectional projections of the panel data, by weighted averages. These weights are chosen to diversify away the idiosyncratic components, resulting in "diversified factors". Because the projections are conducted cross-sectionally, they are robust to serial conditions, easy to analyze due to data-independent weights, and work even for finite length of time series. We formally prove that this procedure is robust to over-estimating the number of factors, and illustrate it in several applications. We also recommend several choices for the diversified weights.

References

Ahn, S. and Horenstein, A. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations. *Journal of the American Statistical Association* **96** 939–967.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40** 436–465.

Bai, J. and Liao, Y. (2017). Inferences in panel data with interactive effects using large covariance matrices. *Journal of Econometrics* **200** 59–78.

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

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150.

Barigozzi, M. and Cho, H. (2018). Consistent estimation of high-dimensional factor models when the factor number is over-estimated. *arXiv preprint arXiv:1811.00306* .

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37** 1705–1732.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance* **52** 57–82.

Chudik, A., Pesaran, M. H. and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal* **14** C45–C90.

Connor, G. and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* **15** 373–394.

Connor, G., Matthias, H. and Linton, O. (2012). Efficient semiparametric estimation of the fama-french model and extensions. *Econometrica* **80** 713–754.

Doz, C., Giannone, D. and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics* **164** 188–205.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance* **47** 427–465.

Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2019). Farmtest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association* 1–29.

Fan, J., Ke, Y. and Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics* .

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B* **75** 603–680.

Fan, J., Liao, Y. and Wang, W. (2016). Projected principal component analysis in factor models. *Annals of Statistics* **44** 219–254.

Fan, J., Liao, Y. and Yao, J. (2015). Power enhancement in high dimensional cross-sectional tests. *Econometrica* **83** 1497–1541.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association* **100** 830–840.

Gagliardini, P., Ossola, E. and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica* **84** 985–1046.

Goncalves, S. and Perron, B. (2018). Bootstrapping factor models with cross sectional dependence .

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

Hansen, C. and Liao, Y. (2018). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory* 1–45.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.

Juodis, A. and Sarafidis, V. (2020). A linear estimator for factoraugmented fixed-t panels with endogenous regressors. Tech. rep., Monash University, Department of Econometrics and Business Statistics.

Karabiyik, H., Reese, S. and Westerlund, J. (2017). On the role of the rank condition in cce estimation of factor-augmented panel regressions. *Journal of Econometrics* **197** 60–64.

Karabiyik, H., Urbain, J.-P. and Westerlund, J. (2019). Cce estimation of factor-augmented regression models with more factors than observables. *Journal of Applied Econometrics* **34** 268–284.

Kim, S., Korajczyk, R. A. and Neuhierl, A. (2018). Arbitrage portfolios in large panels. *Available at SSRN* .

Li, G., Yang, D., Nobel, A. B. and Shen, H. (2016). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis* **146** 7–17.

Li, H., Li, Q. and Shi, Y. (2017). Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics* **197** 76–86.

Ludvigson, S. and Ng, S. (2007). The empirical risk–return relation: a factor analysis approach. *Journal of Financial Economics* **83** 171–222.

Moon, R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* **83** 1543–1579.

Park, B. U., Mammen, E., Härdle, W. and Borak, S. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association* **104** 284–298.

Pesaran, H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74** 967–1012.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* **56** 931–954.

Stock, J. and Watson, M. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97** 1167–1179.

Wang, J., Zhao, Q., Hastie, T. and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics* **45** 1863.

Westerlund, J. and Urbain, J.-P. (2015). Cross-sectional averages versus principal components. *Journal of Econometrics* **185** 372–377.

Supplement to "Learning Latent Factors from Diversified Projections and its Applications to Over-Estimated and Weak Factors"

Jianqing Fan* Yuan Liao†

Abstract

This supplement contains all the technical proofs of the main paper.

Contents

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. jqfan@princeton.edu. His research is supported by NSF grants DMS-1662139 and DMS-1712591.

†Department of Economics, Rutgers University, 75 Hamilton St., New Brunswick, NJ 08901, USA. yuan.liao@rutgers.edu

A Technical Proofs

Throughout the proofs, we use C to denote a generic positive constant. Recall that $\nu_{\min}(\mathbf{H})$ and $\nu_{\max}(\mathbf{H})$ respectively denote the minimum and maximum nonzero singular values of $\mathbf{H}$. In addition, $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ and $\mathbf{M}_{\mathbf{A}} = \mathbf{I} - \mathbf{P}_{\mathbf{A}}$ denote the projection matrices of a matrix $\mathbf{A}$. If $\mathbf{A}'\mathbf{A}$ is singular, $(\mathbf{A}'\mathbf{A})^{-1}$ is replaced with its Moore-Penrose generalized inverse $(\mathbf{A}'\mathbf{A})^{+}$. Let $\mathbf{U}$ be the $N \times T$ matrix of u_{it}. Recall that $R = \dim(\widehat{\mathbf{f}}_t)$ and $r = \dim(\mathbf{f}_t)$.

We use $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$ to respectively denote the operator norm and Frobinus norm. Finally, we define $\|\mathbf{A}\|_\infty$ as follows: if $\mathbf{A}$ is an $N \times K$ matrix with $K = R$ or r, then $\|\mathbf{A}\|_\infty = \max_{i \le N} \|\mathbf{A}_i\|$ where $\mathbf{A}_i$ denotes the i th row of $\mathbf{A}$; if $\mathbf{A}$ is a $K \times N$ matrix with $K = R$ or r, then $\|\mathbf{A}\|_\infty = \max_{i \le N} \|\mathbf{A}_i\|$ where $\mathbf{A}_i$ denotes the i th column of $\mathbf{A}$; if $\mathbf{A}$ is an $N \times N$ matrix, then $\|\mathbf{A}\|_\infty = \max_{i,j \le N} |A_{ij}|$ where A_{ij} denotes the (i,j) th element of $\mathbf{A}$.

Throughout the proof, all $\mathbb{E}(.)$, $\mathbb{E}(.|.)$ and $\mathrm{Var}(.)$ are calculated conditionally on $\mathbf{W}$.

A.1 A key Proposition for asymptotic analysis when $R \ge r$

Proposition A.1. *Suppose $R \ge r$ and $T, N \to \infty$. Also suppose $\mathbf{G}$ is a $T \times d$ matrix so that $\mathbb{E}(\mathbf{U}|\mathbf{G}) = 0$, $\frac{1}{T}\|\mathbf{G}\|^2 = O_P(1)$, for some fixed dimension d, and Assumption 2.1 - 2.4 hold. In addition, for each $\mathbf{K} \in \{\mathbf{I}_T, \mathbf{M}_{\mathbf{G}}\}$, suppose $\lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}) > c > 0$. Then*
(i) $\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) \ge cN^{-1}$ with probability approaching one for some $c > 0$,
(ii) $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\| = O_P(\nu_{\min}^{-1} + \sqrt{\frac{N}{T}})$, and $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H}\| = O_P(1)$.
(iii) $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H} - \mathbf{H}'(\mathbf{H}\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}')^{+}\mathbf{H}\| = O_P(\frac{1}{N\nu_{\min}^2} + \frac{1}{T})$, and $\frac{1}{T}\mathbf{G}'(\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{F}\mathbf{H}'})\mathbf{G} = O_P(\frac{1}{N\nu_{\min}^2} + \frac{1}{T})$.

Proof. The proof applies for both $\mathbf{K} = \mathbf{I}_T$ and $\mathbf{K} = \mathbf{M}_{\mathbf{G}}$. In addition, the proof depends on results in the later Lemma A.1; the latter is proved independently which does not depend on this proposition. Write $\nu_{\min} := \nu_{\min}(\mathbf{H})$, and $\nu_{\max} := \nu_{\max}(\mathbf{H})$.

First, it is easy to see

$$\widehat{\mathbf{F}} = \mathbf{F}\mathbf{H}' + \mathbf{E}.$$

where $\mathbf{E} = (\mathbf{e}_1, \cdots, \mathbf{e}_T)' = \frac{1}{N}\mathbf{U}'\mathbf{W}$, which is $T \times R$. Write

$$\boldsymbol{\Delta} := \frac{1}{T}\mathbb{E}\,\mathbf{E}'\mathbf{E} + \frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{E} + \frac{1}{T}\mathbf{E}'\mathbf{K}\mathbf{F}\mathbf{H}' + \frac{1}{T}(\mathbf{E}'\mathbf{E} - \mathbb{E}\,\mathbf{E}'\mathbf{E}) + \boldsymbol{\Delta}_1$$

where $\boldsymbol{\Delta}_1 = 0$ if $\mathbf{K} = \mathbf{I}_T$ and $\boldsymbol{\Delta}_1 = -\frac{1}{T}\mathbf{E}'\mathbf{P}_{\mathbf{G}}\mathbf{E}$ if $\mathbf{K} = \mathbf{M}_{\mathbf{G}}$.

(i) We have

$$\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}} = \frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}' + \boldsymbol{\Delta}.$$

By assumption $\lambda_{\min}(\frac{1}{T}\mathbb{E}\,\mathbf{U}\mathbf{U}') \geq c_0$, so $\lambda_{\min}(\frac{1}{T}\mathbb{E}\,\mathbf{E}'\mathbf{E}) \geq \lambda_{\min}(\frac{1}{T}\mathbb{E}\,\mathbf{U}\mathbf{U}')\lambda_{\min}(\frac{1}{N^2}\mathbf{W}'\mathbf{W}) \geq c_0 N^{-1}$ for some $c_0 > 0$. In addition, Lemma A.1 shows $\frac{1}{T}(\mathbf{E}'\mathbf{E} - \mathbb{E}\,\mathbf{E}'\mathbf{E}) + \boldsymbol{\Delta}_1 = O_P(\frac{1}{N\sqrt{T}})$. Hence $\|\frac{1}{T}(\mathbf{E}'\mathbf{E} - \mathbb{E}\,\mathbf{E}'\mathbf{E}) + \boldsymbol{\Delta}_1\| \leq \frac{1}{2}\lambda_{\min}(\frac{1}{T}\mathbb{E}\,\mathbf{E}'\mathbf{E})$ with large probability. We now continue the argument conditioning on this event.

Now let $\mathbf{v}$ be the unit vector so that $\mathbf{v}'\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}\mathbf{v} = \lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})$ and let

$$\eta_v^2 := \frac{1}{T}\mathbf{v}'\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}'\mathbf{v}.$$

Because $\mathbf{v}'\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}\mathbf{v} = \eta_v^2 + \mathbf{v}'\boldsymbol{\Delta}\mathbf{v}$, we have

$$\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) \geq \eta_v^2 + 2\mathbf{v}'\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{E}\mathbf{v} + \frac{c_0}{2N}.$$

If $\mathbf{v}'\mathbf{H} = 0$ then $\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) \geq \frac{c_0}{2N}$. If $\mathbf{v}'\mathbf{H} \neq 0$ then $\eta_v^2 \neq 0$ with large probability because $\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}$ is positive definite. Now let

$$X := (\frac{\eta_v^2}{TN})^{-1/2} 2\mathbf{v}'\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{E}\mathbf{v}, \quad 2\mathbf{v}'\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{E}\mathbf{v} = X\sqrt{\frac{\eta_v^2}{TN}}.$$

Then

$$\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) \geq \eta_v^2 + X\sqrt{\frac{\eta_v^2}{TN}} + \frac{c_0}{2N}.$$

Suppose for now $X = O_P(1)$, a claim to be proved later. Then consider two cases.

In case 1, $\eta_v^2 \leq 4|X|\sqrt{\frac{\eta_v^2}{TN}}$. Then $|\eta_v| \leq 4|X|\frac{1}{\sqrt{TN}}$ and

$$\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) \geq \frac{c_0}{2N} - |X||\eta_v|\frac{1}{\sqrt{TN}} \geq \frac{c_0}{2N} - 4|X|^2\frac{1}{TN} \geq \frac{c_0}{4N}$$

where the last inequality holds for $X = O_P(1)$ and as $T \to \infty$, with probability approaching one.

In case 2, $\eta_v^2 > 4|X|\sqrt{\frac{\eta_v^2}{TN}}$, then

$$\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) \geq \eta_v^2 - |X|\sqrt{\frac{\eta_v^2}{TN}} + \frac{c_0}{2N} \geq \frac{3}{4}\eta_v^2 + \frac{c_0}{2N} \geq \frac{c_0}{2N}.$$

In both cases, $\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}) > c_0/N$ for some $c_0 > 0$ with overwhelming probability.

It remains to argue $X = O_P(1)$. By the assumption $\lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}) > c > 0$, we have

$$\eta_v^2 \geq \lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})\mathbf{v}'\mathbf{H}\mathbf{H}'\mathbf{v} > c\|\mathbf{v}'\mathbf{H}\|^2.$$

In addition, Lemma A.1 shows $\|\frac{1}{T}\mathbf{F}'\mathbf{E}\|^2 = O_P(\frac{1}{TN})$ and $\|\frac{1}{T}\mathbf{G}'\mathbf{E}\|^2 = O_P(\frac{1}{TN})$. With the condition $\frac{1}{T}\|\mathbf{G}\|^2 = O_P(1)$, we reach $\|\frac{1}{T}\mathbf{F}'\mathbf{M}_{\mathbf{G}}\mathbf{E}\|^2 \leq O_P(\frac{1}{TN}) + \|\mathbf{F}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\|^2\|\frac{1}{T}\mathbf{G}'\mathbf{E}\|^2 = O_P(\frac{1}{TN})$. Therefore $\|\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{E}\|^2 = O_P(\frac{1}{TN})$ and consequently,

$$|X|^2 \leq 4TN\eta_v^{-2}\|\mathbf{v}'\mathbf{H}\|^2\|\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{E}\|^2 \leq O_P(1)\eta_v^{-2}\|\mathbf{v}'\mathbf{H}\|^2 \leq O_P(1)c^{-1}\|\mathbf{v}'\mathbf{H}\|^{-2}\|\mathbf{v}'\mathbf{H}\|^2 = O_P(1).$$

(ii) Write $\bar{\mathbf{H}} := \mathbf{H}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{1/2}$ and $\mathbf{S} := \frac{N}{T}\mathbb{E}\,\mathbf{E}'\mathbf{E} = \frac{1}{N}\mathbf{W}'\mathbf{\Sigma}_u\mathbf{W}$. Then

$$\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}} = \bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S} + \frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{E} + \frac{1}{T}\mathbf{E}'\mathbf{K}\mathbf{F}\mathbf{H}' + \mathbf{\Delta}_2 \tag{A.1}$$

where we proved in (i) that $\|\mathbf{\Delta}_2\| = \|\frac{1}{T}(\mathbf{E}'\mathbf{E} - \mathbb{E}\,\mathbf{E}'\mathbf{E}) + \mathbf{\Delta}_1\| = O_P(\frac{1}{N\sqrt{T}})$. Also all eigenvalues of $\mathbf{S}$ are bounded away from both zero and infinity. In addition, $\bar{\mathbf{H}}$ is a $R \times r$ matrix with $R \geq r$, whose Moore-Penrose generalized inverse is $\bar{\mathbf{H}}^+ = (\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\mathbf{H}^+$. Also, $\bar{\mathbf{H}}$ is of rank r. Let

$$\bar{\mathbf{H}}' = \mathbf{U}_{\bar{H}}(\mathbf{D}_{\bar{H}}, 0)\mathbf{E}'_{\bar{H}}$$

be the singular value decomposition (SVD) of $\bar{\mathbf{H}}'$, where 0 is present when $R > r$. Since $\lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}) > c > 0$, we have $\lambda_{\min}(\mathbf{D}_{\bar{H}}) \geq c\nu_{\min}$ where $\nu_{\min} := \nu_{\min}(\mathbf{H})$.

The proof is divided into several steps.

Step 1. Show $\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{a}{N}\mathbf{I})^{-j}\bar{\mathbf{H}}\| = O_P(\nu_{\min}^{-(2j-2)})$ for any fixed $a > 0$ and $j = 1, 2$.

Because $\lambda_{\min}(\mathbf{D}_{\bar{H}}) \geq c\nu_{\min}$, for $j = 1, 2$,

$$\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{a}{N}\mathbf{I})^{-j}\bar{\mathbf{H}}\| = \|\mathbf{U}_{\bar{H}}(\mathbf{D}_{\bar{H}}^2(\mathbf{D}_{\bar{H}}^2 + \frac{a}{N}\mathbf{I})^{-j}, 0)\mathbf{U}'_{\bar{H}}\| = \|\mathbf{D}_{\bar{H}}^2(\mathbf{D}_{\bar{H}}^2 + \frac{a}{N}\mathbf{I})^{-j}\| \leq \|\mathbf{D}_{\bar{H}}^{-2j+2}\|.$$

Step 2. Show $\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}}\| = O_P(1)$.

Let $0 < a < \lambda_{\min}(\mathbf{S})$ be a constant. Then $(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{a}{N}\mathbf{I})^{-1} - (\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}$ is positive definite. (This is because, if both $\mathbf{A}_1$ and $\mathbf{A}_2 - \mathbf{A}_1$ are positive definite, then so is $\mathbf{A}_1^{-1} - \mathbf{A}_2^{-1}$.) Let $\mathbf{v}$ be a unit vector so that $\mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}}\mathbf{v} = \|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}}\|$. Then

$$\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}}\| \leq \mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{a}{N}\mathbf{I})^{-1}\bar{\mathbf{H}}\mathbf{v} \leq \|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{a}{N}\mathbf{I})^{-1}\bar{\mathbf{H}}\|.$$

The right hand side is $O_P(1)$ due to step 1.

Step 3. Show $\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\|=O_P(\nu_{\min}^{-1})$.

Fix any $a>0$. Let $\mathbf{M}=\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{a}{N}\mathbf{I})^{-1}$. By step 1, $\|\mathbf{M}\|=\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{a}{N}\mathbf{I})^{-2}\bar{\mathbf{H}}\|^{1/2}=O_P(\nu_{\min}^{-1})$. So

$$\begin{aligned}\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\| &\leq \|\mathbf{M}\|+\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}-\mathbf{M}\|\\ &=^{(1)} \|\mathbf{M}\|+\|\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{a}{N}\mathbf{I})^{-1}(\frac{1}{N}\mathbf{S}-\frac{a}{N}\mathbf{I})(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\|\\ &\leq \|\mathbf{M}\|+\frac{C}{N}\|\mathbf{M}\|\|(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\|\\ &\leq^{(2)} \|\mathbf{M}\|(1+O_P(1))=O_P(\nu_{\min}^{-1}).\end{aligned}$$

(1) used $\mathbf{A}_1^{-1}-\mathbf{A}_2^{-1}=\mathbf{A}_1^{-1}(\mathbf{A}_2-\mathbf{A}_1)\mathbf{A}_2^{-1}$; (2) is from: $\|(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\|\leq\lambda_{\min}^{-1}(\frac{1}{N}\mathbf{S})=O_P(N)$.

Step 4. Show $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|=O_P(\nu_{\min}^{-1}+\sqrt{\frac{N}{T}})$.

Let $\mathbf{A}:=\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S}$. By steps 2,3 $\|\bar{\mathbf{H}}\mathbf{A}^{-1}\|=O_P(\nu_{\min}^{-1})$ and $\|\bar{\mathbf{H}}\mathbf{A}^{-1}\bar{\mathbf{H}}\|=O_P(1)$. Now

$$\begin{aligned}&\|\bar{\mathbf{H}}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}-\bar{\mathbf{H}}'\mathbf{A}^{-1}\|=\|\bar{\mathbf{H}}'\mathbf{A}^{-1}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}-\mathbf{A})(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|\\ \leq^{(3)}\ & O_P(\frac{\nu_{\max}(\mathbf{H})}{\nu_{\min}(\mathbf{H})\sqrt{TN}})\|(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|=^{(4)}O_P(\frac{N}{\sqrt{NT}})=O_P(\sqrt{\frac{N}{T}}).\end{aligned}$$

In (3) we used $\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}-\mathbf{A}=O_P(\frac{1}{N\sqrt{T}}+\|\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{E}\|)=O_P(\frac{1}{N\sqrt{T}}+\frac{\nu_{\max}}{\sqrt{TN}})=O_P(\frac{\nu_{\max}}{\sqrt{TN}})$; in (4) we used $(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}=O_P(N)$ by part (i) and $\nu_{\max}\leq C\nu_{\min}$. Hence

$$\|\bar{\mathbf{H}}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|\leq O_P(\sqrt{\frac{N}{T}})+\|\bar{\mathbf{H}}\mathbf{A}^{-1}\|=O_P(\nu_{\min}^{-1}+\sqrt{\frac{N}{T}}).$$

Thus $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|\leq\|(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\|\|\bar{\mathbf{H}}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|$, which leads to the result for $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\|=O_P(\nu_{\min}^{-1}+\sqrt{\frac{N}{T}})$.

Step 5. show $\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H}=\mathbf{H}'(\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}'+\frac{1}{N}\mathbf{S})^{-1}\mathbf{H}+O_P(\frac{1}{\nu_{\min}\sqrt{NT}}+\frac{1}{T})$.

Because $\|\bar{\mathbf{H}}\mathbf{A}^{-1}\|=O_P(\nu_{\min}^{-1})$ and $\|\bar{\mathbf{H}}\mathbf{A}^{-1}\bar{\mathbf{H}}\|=O_P(1)$ by step 3, (A.1) implies

$$\begin{aligned}&\|\bar{\mathbf{H}}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\bar{\mathbf{H}}-\bar{\mathbf{H}}'\mathbf{A}^{-1}\bar{\mathbf{H}}\|=\|\bar{\mathbf{H}}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}}-\mathbf{A})\mathbf{A}^{-1}\bar{\mathbf{H}}\|\\ \leq\ & \|\bar{\mathbf{H}}'\mathbf{A}^{-1}\bar{\mathbf{H}}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{E}(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\bar{\mathbf{H}}\|+\|\bar{\mathbf{H}}'\mathbf{A}^{-1}\frac{1}{T}\mathbf{E}'\mathbf{K}\mathbf{F}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\bar{\mathbf{H}}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\bar{\mathbf{H}}\|\\ &+\|\bar{\mathbf{H}}'\mathbf{A}^{-1}\boldsymbol{\Delta}_1(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\bar{\mathbf{H}}\|\end{aligned}$$

$$\leq \quad O_P(\nu_{\min}^{-1}\frac{1}{N\sqrt{T}}+\frac{1}{\sqrt{NT}})\|(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\bar{\mathbf{H}}\| =^{(5)} O_P(\frac{1}{\sqrt{NT}})O_P(\nu_{\min}^{-1}+\sqrt{\frac{N}{T}}) = O_P(\frac{1}{\nu_{\min}\sqrt{NT}}+\frac{1}{T}).$$

(5) follows from step 4 and $\nu_{\min} \gg N^{-1/2}$. Then due to $\|(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\| = O_P(1)$,

$$\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H} = \mathbf{H}'(\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}'+\frac{1}{N}\mathbf{S})^{-1}\mathbf{H} + O_P(\frac{1}{\nu_{\min}\sqrt{NT}}+\frac{1}{T}).$$

In addition, step 3 implies $\|\mathbf{H}'(\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}'+\frac{1}{N}\mathbf{S})^{-1}\mathbf{H}\| \leq O_P(\nu_{\min}^{-1}\nu_{\max}) = O_P(1)$, so

$$\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H}\| = O_P(1+\frac{1}{\nu_{\min}\sqrt{NT}}+\frac{1}{T}) = O_P(1).$$

(iii) The proof still consists of several steps.

Step 1. $\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H} = \mathbf{H}'(\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}'+\frac{1}{N}\mathbf{S})^{-1}\mathbf{H} + O_P(\frac{1}{\nu_{\min}\sqrt{NT}}+\frac{1}{T})$.

It follows from step 5 of part (ii).

Step 2. show $\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}} = \bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}} + O_P(\frac{1}{N\nu_{\min}^2})$ where $\bar{\mathbf{H}} = \mathbf{H}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{1/2}$. Write $\mathbf{T} = \bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}} - \bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}}$. The goal is to show $\|\mathbf{T}\| = O_P(\frac{1}{N\nu_{\min}^2})$. Let $\mathbf{v}$ be the unit vector so that $|\mathbf{v}'\mathbf{T}\mathbf{v}| = \|\mathbf{T}\|$. Define a function, for $d>0$,

$$g(d) := \mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{d}{N}\mathbf{I})^{-1}\bar{\mathbf{H}}\mathbf{v}.$$

Note that there are constants $c, C > 0$ so that $\frac{c}{N} < \lambda_{\min}(\frac{1}{N}\mathbf{S}) \leq \lambda_{\max}(\frac{1}{N}\mathbf{S}) < \frac{C}{N}$. Then we have $g(C) < \mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}'+\frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}}\mathbf{v} < g(c)$. Hence

$$|\mathbf{v}'\mathbf{T}\mathbf{v}| \leq |g(c) - \mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}}\mathbf{v}| + |g(C) - \mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}}\mathbf{v}|.$$

Recall $\bar{\mathbf{H}}' = \mathbf{U}_{\bar{H}}(\mathbf{D}_{\bar{H}}, 0)\mathbf{E}_{\bar{H}}'$ is the SVD of $\bar{\mathbf{H}}'$ and $N^{-1}\lambda_{\min}^{-1}(\mathbf{D}_{\bar{H}}^2) = o_P(1)$. Then for any $d \in \{c, C\}$, as $N \to \infty$, $g(d) = \mathbf{v}'\mathbf{U}_{\bar{H}}\mathbf{D}_{\bar{H}}^2(\mathbf{D}_{\bar{H}}^2+\frac{d}{N}\mathbf{I})^{-1}\mathbf{U}_{\bar{H}}'\mathbf{v} \xrightarrow{P} \mathbf{v}'\mathbf{v} = \mathbf{v}'\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}}\mathbf{v}$, where we used $\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}} = \mathbf{I}$, easy to see from its SVD. The rate of convergence is

$$\|\mathbf{D}_{\bar{H}}^2(\mathbf{D}_{\bar{H}}^2+\frac{d}{N}\mathbf{I})^{-1} - \mathbf{I}\| \leq \|\mathbf{D}_{\bar{H}}^2(\mathbf{D}_{\bar{H}}^2+\frac{d}{N}\mathbf{I})^{-1}\frac{d}{N}\mathbf{D}_{\bar{H}}^{-2}\| = O_P(\frac{1}{N\nu_{\min}^2}).$$

Hence $|\mathbf{v}'\mathbf{T}\mathbf{v}| = O_P(\frac{1}{N\nu_{\min}^2})$.

Step 3. show $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H} - \mathbf{H}'(\mathbf{H}\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}\mathbf{H}')^{+}\mathbf{H}\| = O_P(\frac{1}{N\nu_{\min}^2}+\frac{1}{T})$. By steps 1 and

2,

$$
\begin{aligned}
\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\mathbf{H} &= \mathbf{H}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}\mathbf{H} + O_P(\frac{1}{\nu_{\min}\sqrt{NT}} + \frac{1}{T}) \\
&= (\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}' + \frac{1}{N}\mathbf{S})^{-1}\bar{\mathbf{H}}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2} + O_P(\frac{1}{\nu_{\min}\sqrt{NT}} + \frac{1}{T}) \\
&=^{(6)} (\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2}\bar{\mathbf{H}}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\bar{\mathbf{H}}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F})^{-1/2} + O_P(\frac{1}{N\nu_{\min}^2} + \frac{1}{\nu_{\min}\sqrt{NT}} + \frac{1}{T}) \\
&= \mathbf{H}'(\bar{\mathbf{H}}\bar{\mathbf{H}}')^{+}\mathbf{H} + O_P(\frac{1}{N\nu_{\min}^2} + \frac{1}{T}).
\end{aligned}
$$

where (6) is due to $\lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{K}\mathbf{F}) > c$ and step 2.

Step 4. show $\frac{1}{T}\mathbf{G}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{G} = \frac{1}{T}\mathbf{G}'\mathbf{P}_{\mathbf{FH}'}\mathbf{G} + O_P(\frac{1}{N\nu_{\min}^2} + \frac{1}{T})$.

By part (ii) $\|\mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{K}\widehat{\mathbf{F}})^{-1}\| = O_P(\nu_{\min}^{-1} + \sqrt{\frac{N}{T}})$, and that $\frac{1}{T}\mathbf{G}'\mathbf{E} = O_P(\frac{1}{\sqrt{NT}})$,

$$
\begin{aligned}
\frac{1}{T}\mathbf{G}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{G} &= \frac{1}{T}\mathbf{G}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{G} + \frac{1}{T}\mathbf{G}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{G} + \frac{1}{T}\mathbf{G}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{G} \\
&\quad + \frac{1}{T}\mathbf{G}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{G} \\
&= \frac{1}{T}\mathbf{G}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{G} + O_P(\frac{1}{T} + \frac{1}{\nu_{\min}\sqrt{NT}}) \\
&= \frac{1}{T}\mathbf{G}'\mathbf{F}\mathbf{H}'(\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{+}\mathbf{H}\mathbf{F}'\mathbf{G} + O_P(\frac{1}{N\nu_{\min}^2} + \frac{1}{T}),
\end{aligned}
$$

where the last equality follows from step 3. □

The proof of Lemma A.1 below does not rely on Proposition A.1, as it does not involve $\mathbf{H}$ or $\widehat{\mathbf{F}}$. Also, let $\mathbf{E} = (\mathbf{e}_1, \cdots, \mathbf{e}_T)' = \frac{1}{N}\mathbf{U}'\mathbf{W}$. In addition, we shall use the following inequality $\mathrm{tr}(\mathbf{W}'\mathbf{\Sigma}\mathbf{W}) \leq R\|\mathbf{W}\|^2\|\mathbf{\Sigma}\|$ for any semipositive definite matrix $\mathbf{\Sigma}$, whose simple proof is as follows: let $\mathbf{v}_i$ be the i th eigenvector of $\mathbf{W}'\mathbf{\Sigma}\mathbf{W}$. Then

$$
\mathrm{tr}(\mathbf{W}'\mathbf{\Sigma}\mathbf{W}) = \sum_{i=1}^{R}\mathbf{v}_i'\mathbf{W}'\mathbf{\Sigma}\mathbf{W}\mathbf{v}_i \leq \|\mathbf{\Sigma}\|\sum_{i=1}^{R}\|\mathbf{W}\mathbf{v}_i\|^2 \leq \|\mathbf{\Sigma}\|\|\mathbf{W}\|^2 R.
$$

Lemma A.1. *For any $R \geq 1$, (R can be either smaller, equal to or larger than r),*
(i) $\|\frac{1}{T}\mathbb{E}\,\mathbf{E}'\mathbf{E}\| \leq \frac{C}{N}$ *and* $\|\mathbf{E}\| = O_P(\sqrt{\frac{T}{N}})$.
(ii) $\mathbb{E}\,\|\frac{1}{T}\mathbf{F}'\mathbf{E}\|^2 \leq O(\frac{1}{TN})$, $\mathbb{E}\,\|\frac{1}{T}\mathbf{G}'\mathbf{E}\|^2 \leq O(\frac{1}{TN})$, *here* $\mathbf{G}$ *is defined as in Section 3.1*
(iii) $\|\frac{1}{T}(\mathbf{E}'\mathbf{E} - \mathbb{E}\,\mathbf{E}'\mathbf{E})\| \leq O_P(\frac{1}{N\sqrt{T}})$, $\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\mathbf{G}}\mathbf{E}\| = O_P(\frac{1}{NT})$.
(iv) $\|\frac{1}{N}\mathbf{U}'\mathbf{W}\| \leq O_P(\sqrt{\frac{T}{N}})$.

Proof. (i) By the assumption $\|\frac{1}{T}\mathbb{E}\,\mathbf{U}\mathbf{U}'\| = \|\mathbb{E}\,\mathbf{u}_t\mathbf{u}_t'\| \leq \mathbb{E}\,\|\mathbb{E}(\mathbf{u}_t\mathbf{u}_t'|\mathbf{F})\| < C$. Thus

$$\|\frac{1}{T}\mathbb{E}\,\mathbf{E}'\mathbf{E}\| = \frac{1}{N^2}\|\mathbf{W}'\frac{1}{T}\mathbb{E}\,\mathbf{U}\mathbf{U}'\mathbf{W}\| \leq \frac{1}{N^2}\|\mathbf{W}\|^2 \leq \frac{C}{N}.$$

Also, $\mathbb{E}\,\|\mathbf{E}\|^2 \leq \operatorname{tr}\mathbb{E}\,\mathbf{E}'\mathbf{E} \leq R\|\mathbb{E}\,\mathbf{E}'\mathbf{E}\| \leq \frac{CT}{N}$.

(ii) Let $f_{k,t}$ be the k th entry of $\mathbf{f}_t$. By the assumption $\frac{1}{T}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\,\|\mathbf{f}_t\|\|\mathbf{f}_s\|\|\mathbb{E}(\mathbf{u}_t\mathbf{u}_s'|\mathbf{F})\| < C$,

$$\begin{aligned}
\mathbb{E}\,\|\frac{1}{T}\mathbf{F}'\mathbf{E}\|^2 &= \frac{1}{T^2N^2}\mathbb{E}\,\|\sum_{t=1}^T \mathbf{W}'\mathbf{u}_t\mathbf{f}_t'\|^2 \leq \sum_{k=1}^r \frac{1}{T^2N^2}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\,f_{k,t}f_{k,s}\,\mathbb{E}(\mathbf{u}_s'\mathbf{W}\mathbf{W}'\mathbf{u}_t|\mathbf{F}) \\
&\leq \sum_{k=1}^r \frac{1}{T^2N^2}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\,f_{k,t}f_{k,s}\operatorname{tr}\mathbf{W}'\,\mathbb{E}(\mathbf{u}_t\mathbf{u}_s'|\mathbf{F})\mathbf{W} \\
&\leq \sum_{k=1}^r \frac{1}{T^2N^2}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\,|f_{k,t}f_{k,s}|\|\mathbf{W}\|_F^2\|\mathbb{E}(\mathbf{u}_t\mathbf{u}_s'|\mathbf{F})\| \\
&\leq \frac{C}{T^2N}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\,\|\mathbf{f}_t\|\|\mathbf{f}_s\|\|\mathbb{E}(\mathbf{u}_t\mathbf{u}_s'|\mathbf{F})\| \\
&\leq \frac{C}{TN}.
\end{aligned}$$

Similarly, $\mathbb{E}\,\|\frac{1}{T}\mathbf{G}'\mathbf{E}\|^2 \leq O(\frac{1}{TN})$.

(iii) By the assumption that $\frac{1}{TN^2}\sum_{t,s\leq T}\sum_{i,j,m,n\leq N}|\operatorname{Cov}(u_{it}u_{jt}, u_{ms}u_{ns})| < C$,

$$\begin{aligned}
&\mathbb{E}\,\|\frac{1}{T}(\mathbf{E}'\mathbf{E} - \mathbb{E}\,\mathbf{E}'\mathbf{E})\|^2 \leq \sum_{k,q\leq R}\mathbb{E}(\frac{1}{TN^2}\sum_{t=1}^T\sum_{i,j\leq N} w_{k,i}w_{q,j}(u_{it}u_{jt} - \mathbb{E}\,u_{it}u_{jt}))^2 \\
\leq\ & \frac{C}{TN^2}\frac{1}{TN^2}\sum_{t,s\leq T}\sum_{i,j,m,n\leq N}|\operatorname{Cov}(u_{it}u_{jt}, u_{ms}u_{ns})| \leq \frac{C}{TN^2}.
\end{aligned}$$

Next, by part (ii)

$$\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\mathbf{G}}\mathbf{E}\| \leq \|\frac{1}{T}\mathbf{E}'\mathbf{G}\|^2\|(\frac{1}{T}\mathbf{G}'\mathbf{G})^{-1}\| \leq O_P(\frac{1}{TN}).$$

(iv) $\mathbb{E}\,\|\frac{1}{N}\mathbf{U}'\mathbf{W}\|^2 \leq \frac{1}{N^2}\operatorname{tr}\mathbb{E}\,\mathbf{W}'\mathbf{U}\mathbf{U}'\mathbf{W} \leq \frac{CT}{N^2}\|\mathbf{W}\|_F^2 \leq \frac{CT}{N}$, where we used the assumption that $\|\mathbb{E}\,\mathbf{u}_t\mathbf{u}_t'\| < C$.

□

A.2 Proof of Theorem 2.1

Proof. We shall first show the convergence of $\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} - \mathbf{P}_{\mathbf{F}}$, and then the convergence of $\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{P}_{\mathbf{F}} - \mathbf{P}_{\mathbf{F}}$.

First, from the SVD $\mathbf{H}' = \mathbf{U}_H(\mathbf{D}_H, 0)\mathbf{E}_H'$, it is straightforward to verify that $\mathbf{M}' = \mathbf{U}_H(\mathbf{D}_H^{-1}, 0)\mathbf{E}_H'$. Then from Proposition A.1, $\lambda_{\min}(\frac{1}{T}\mathbf{M}'\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M}) \geq c_0 N^{-1}\lambda_{\min}(\mathbf{D}_H^{-2})$ with large probability. Hence $\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}}$ is well defined.

Next, it is easy to see $\mathbf{H}'(\mathbf{H}\mathbf{H}')^{+}\mathbf{H} = \mathbf{I}$ when $R \geq r$. Then $\widehat{\mathbf{F}} = \mathbf{F}\mathbf{H}' + \mathbf{E}$ implies $\widehat{\mathbf{F}}\mathbf{M} - \mathbf{F} = \mathbf{E}(\mathbf{H}\mathbf{H}')^{+}\mathbf{H}$ with $\mathbf{M} = (\mathbf{H}\mathbf{H}')^{+}\mathbf{H}$. Since $\|(\mathbf{H}\mathbf{H}')^{+}\mathbf{H}\| = O_P(\nu_{\min}^{-1})$, we have

$$\frac{1}{\sqrt{T}}\|\widehat{\mathbf{F}}\mathbf{M} - \mathbf{F}\| = O_P(\frac{1}{\sqrt{N}}\nu_{\min}^{-1}), \quad \frac{1}{T}\|\mathbf{F}'(\widehat{\mathbf{F}}\mathbf{M} - \mathbf{F})\| = O_P(\frac{1}{\sqrt{NT}}\nu_{\min}^{-1})$$

where the second statement uses Lemma A.1. Then $\|\frac{1}{T}\mathbf{M}'\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M} - \frac{1}{T}\mathbf{F}'\mathbf{F}\| = O_P(\frac{1}{\sqrt{NT}}\nu_{\min}^{-1} + \frac{1}{N}\nu_{\min}^{-2})$. Thus $(\frac{1}{T}\mathbf{M}'\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M})^{-1} = O_P(1)$ and

$$\|(\frac{1}{T}\mathbf{M}'\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M})^{-1} - (\frac{1}{T}\mathbf{F}'\mathbf{F})^{-1}\| = O_P(\frac{1}{\sqrt{NT}}\nu_{\min}^{-1} + \frac{1}{N}\nu_{\min}^{-2}). \tag{A.2}$$

The triangular inequality then implies $\|\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} - \mathbf{P}_{\mathbf{F}}\| \leq O_P(\frac{1}{\sqrt{N}}\nu_{\min}^{-1})$.

Finally, $\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} = \mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}}$ gives

$$\|\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{P}_{\mathbf{F}} - \mathbf{P}_{\mathbf{F}}\| \leq \|\mathbf{P}_{\widehat{\mathbf{F}}}(\mathbf{P}_{\mathbf{F}} - \mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}})\| + \|\mathbf{P}_{\widehat{\mathbf{F}}\mathbf{M}} - \mathbf{P}_{\mathbf{F}}\| \leq O_P(\frac{1}{\sqrt{N}}\nu_{\min}^{-1}).$$

□

A.3 Proof of Theorem 3.1

Proof. Here we assume $R \geq r$. We let $\mathbf{z}_t = (\mathbf{f}_t'\mathbf{H}', \mathbf{g}_t')'$ and $\boldsymbol{\delta} = (\boldsymbol{\alpha}'\mathbf{H}^{+}, \boldsymbol{\beta}')'$. Then $\boldsymbol{\delta}'\mathbf{z}_t = y_{t+h|t}$. First, we have the following expansion

$$\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{z}}_T - \boldsymbol{\delta}'\mathbf{z}_T = (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})'\widehat{\mathbf{z}}_T + \boldsymbol{\alpha}'\mathbf{H}^{+}(\widehat{\mathbf{f}}_T - \mathbf{H}\mathbf{f}_T).$$

Now $\widehat{\boldsymbol{\delta}} = (\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}})^{-1}\widehat{\mathbf{Z}}'\mathbf{Y}$, where $\mathbf{Y}$ is the $(T-h)\times 1$ vector of y_{t+h}, and $\widehat{\mathbf{Z}}$ is the $(T-h)\times \dim(\boldsymbol{\delta})$ matrix of $\widehat{\mathbf{z}}_t$, $t=1,\cdots,T-h$. Also recall that $\mathbf{e}_t = \widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t = \frac{1}{N}\mathbf{W}'\mathbf{u}_t$. Then

$$\begin{aligned}
\widehat{\mathbf{z}}_T'(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}) &= \widehat{\mathbf{z}}_T'(\frac{1}{T}\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}})^{-1}\sum_{d=1}^{4} a_d, \text{ where} \\
a_1 &= (\frac{1}{T}\sum_t \varepsilon_t\mathbf{e}_t', 0)', \quad a_2 = \frac{1}{T}\sum_t \mathbf{z}_t\varepsilon_t \\
a_3 &= (-\boldsymbol{\alpha}'\mathbf{H}^{+}\frac{1}{T}\sum_t \mathbf{e}_t\mathbf{e}_t', 0)', \quad a_4 = -\frac{1}{T}\sum_t \mathbf{z}_t\mathbf{e}_t'\mathbf{H}^{+'}\boldsymbol{\alpha}.
\end{aligned}$$

On the other hand, let $\mathbf{G}$ be the $(T-h)\times\dim(\mathbf{g}_t)$ matrix of $\{\mathbf{g}_t : g\leq T-h\}$. We have, by the matrix block inverse formula, for the operator $\mathbf{M}_{\mathbf{A}} := \mathbf{I}-\mathbf{P}_{\mathbf{A}}$,

$$(\frac{1}{T}\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}})^{-1} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_2' & \mathbf{A}_3 \end{pmatrix}, \quad \text{where } \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \mathbf{A}_3 \end{pmatrix} = \begin{pmatrix} (\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}})^{-1} \\ -\mathbf{A}_1\widehat{\mathbf{F}}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} \\ (\frac{1}{T}\mathbf{G}'\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{G})^{-1} \end{pmatrix}.$$

Then $\widehat{\mathbf{z}}_T'(\frac{1}{T}\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}})^{-1} = (\widehat{\mathbf{f}}_T'\mathbf{A}_1 + \mathbf{g}_T'\mathbf{A}_2', \widehat{\mathbf{f}}_T'\mathbf{A}_2 + \mathbf{g}_T'\mathbf{A}_3)$. This implies

$$\begin{aligned}
\widehat{\mathbf{z}}_T'(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}) &= (\widehat{\mathbf{f}}_T'\mathbf{A}_1 + \mathbf{g}_T'\mathbf{A}_2')\frac{1}{T}\sum_t[\mathbf{e}_t\varepsilon_t - \mathbf{e}_t\mathbf{e}_t'\mathbf{H}^{+'}\boldsymbol{\alpha}] \\
&\quad + (\widehat{\mathbf{f}}_T'\mathbf{A}_1\mathbf{H} + \mathbf{g}_T'\mathbf{A}_2'\mathbf{H})\frac{1}{T}\sum_t[\mathbf{f}_t\varepsilon_t - \mathbf{f}_t\mathbf{e}_t'\mathbf{H}^{+'}\boldsymbol{\alpha}] \\
&\quad + (\widehat{\mathbf{f}}_T'\mathbf{A}_2 + \mathbf{g}_T'\mathbf{A}_3)\frac{1}{T}\sum_t[\mathbf{g}_t\varepsilon_t - \mathbf{g}_t\mathbf{e}_t'\mathbf{H}^{+'}\boldsymbol{\alpha}].
\end{aligned}$$

It is easy to show $\|\frac{1}{T}\sum_t \mathbf{f}_t\varepsilon_t\| + \|\frac{1}{T}\sum_t \mathbf{g}_t\varepsilon_t\| = O_P(\frac{1}{\sqrt{T}})$ and $\|\frac{1}{T}\sum_t \mathbf{e}_t\varepsilon_t\| = O_P(\frac{1}{\sqrt{TN}})$. Also Lemma A.1 gives $\frac{1}{T}\sum_t \mathbf{e}_t\mathbf{e}_t' = \frac{1}{T}\mathbf{E}'\mathbf{E} = O_P(\frac{1}{N})$, $\frac{1}{T}\sum_t \mathbf{f}_t\mathbf{e}_t = \frac{1}{T}\mathbf{F}'\mathbf{E} = O_P(\frac{1}{\sqrt{TN}})$, and $\frac{1}{T}\sum_t \mathbf{g}_t\mathbf{e}_t = \frac{1}{T}\mathbf{F}'\mathbf{E} = O_P(\frac{1}{\sqrt{TN}})$. Together with Lemma A.2,

$$\begin{aligned}
\widehat{\mathbf{z}}_T'(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}) &= \|\widehat{\mathbf{f}}_T'\mathbf{A}_1 + \mathbf{g}_T'\mathbf{A}_2\|O_P(\frac{1}{\sqrt{TN}} + \frac{1}{N\nu_{\min}}) \\
&\quad + \|\widehat{\mathbf{f}}_T'\mathbf{A}_1\mathbf{H} + \mathbf{g}_T'\mathbf{A}_2'\mathbf{H}\|O_P(\frac{1}{\sqrt{T}}) + \|\widehat{\mathbf{f}}_T'\mathbf{A}_2 + \mathbf{g}_T'\mathbf{A}_3\|O_P(\frac{1}{\sqrt{T}}) \\
&= O_P(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}\nu_{\min}}).
\end{aligned}$$

Finally, as $\|\mathbf{H}^{+}\| = O_P(\nu_{\min}^{-1})$, $\boldsymbol{\alpha}'\mathbf{H}^{+}(\widehat{\mathbf{f}}_T - \mathbf{H}\mathbf{f}_T) = O_P(\nu_{\min}^{-1})\|\mathbf{e}_T\| = O_P(\nu_{\min}^{-1}N^{-1/2})$.

$\square$

Lemma A.2. *For all $R \geq r$, (i) $\|\mathbf{A}_1\widehat{\mathbf{f}}_T\| + \|\mathbf{A}_2\| = O_P(\sqrt{N})$, and*
$\|\mathbf{H}'\mathbf{A}_1\widehat{\mathbf{f}}_T\| + \|\mathbf{H}'\mathbf{A}_2\| + \|\mathbf{A}_2'\widehat{\mathbf{f}}_T\| + \|\mathbf{A}_3\| = O_P(1)$.

Proof. First, by Proposition A.1, $\|\mathbf{A}_1\| = O_P(N)$ and $\|\mathbf{A}_1\mathbf{H}\| = O_P(\nu_{\min}^{-1} + \sqrt{\frac{N}{T}})$, and $\frac{1}{T}\mathbf{E}'\mathbf{G} = O_P(\frac{1}{\sqrt{NT}})$

$$
\begin{aligned}
\mathbf{A}_1\widehat{\mathbf{f}}_T &= (\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}})^{-1}\mathbf{e}_T + (\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{f}_T = O_P(\sqrt{N}) \\
\mathbf{H}'\mathbf{A}_1\widehat{\mathbf{f}}_T &= \mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}})^{-1}\mathbf{e}_T + \mathbf{H}'(\frac{1}{T}\widehat{\mathbf{F}}'\mathbf{M}_{\mathbf{G}}\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{f}_T = O_P(1) \\
-\mathbf{A}_2 &= \mathbf{A}_1\widehat{\mathbf{F}}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} = \mathbf{A}_1\mathbf{E}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} + \mathbf{A}_1\mathbf{H}\mathbf{F}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} = O_P(\sqrt{\frac{N}{T}} + \nu_{\min}^{-1}) \\
-\mathbf{H}'\mathbf{A}_2 &= \mathbf{H}'\mathbf{A}_1\mathbf{E}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} + \mathbf{H}'\mathbf{A}_1\mathbf{H}\mathbf{F}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} = O_P(1) \\
\mathbf{A}_2'\widehat{\mathbf{f}}_T &= \mathbf{A}_2'\mathbf{H}\mathbf{f}_T + \mathbf{A}_2'\mathbf{e}_T = O_P(1).
\end{aligned}
$$

Finally, it follows from Proposition A.1 that $\frac{1}{T}\mathbf{G}'(\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{F}\mathbf{H}'})\mathbf{G} = O_P(\frac{1}{T} + \frac{1}{N\nu_{\min}^2})$. Hence $\|\mathbf{A}_3\| = O_P(1)$ since $\lambda_{\min}(\frac{1}{T}\mathbf{G}'\mathbf{M}_{\mathbf{F}\mathbf{H}'}\mathbf{G}) > c$.

$\square$

A.4 Proof of Theorem 3.2

Let $\widehat{\boldsymbol{\varepsilon}}_g$, $\widehat{\boldsymbol{\varepsilon}}_y$,$\boldsymbol{\varepsilon}_g$, $\boldsymbol{\varepsilon}_y$, $\mathbf{Y}$, $\mathbf{G}$ and $\boldsymbol{\eta}$ be $T \times 1$ vectors of $\widehat{\varepsilon}_{g,t}$, $\widehat{\varepsilon}_{y,t}$, $\varepsilon_{g,t}$, $\varepsilon_{y,t}$, y_t, $\mathbf{g}_t$ and η_t. Let $\widehat{J}$ denote the index set of components in $\widehat{\mathbf{u}}_t$ that are selected by *either* $\widehat{\boldsymbol{\gamma}}$ *or* $\widehat{\boldsymbol{\theta}}$. Let $\widehat{\mathbf{U}}_{\widehat{J}}$ denote the $N \times |J|_0$ matrix of rows of $\widehat{\mathbf{U}}$ selected by J. Then

$$
\widehat{\boldsymbol{\varepsilon}}_y = \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{Y}, \quad \widehat{\boldsymbol{\varepsilon}}_g = \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{G}.
$$

A.4.1 The case $r \geq 1$.

Proof. From Lemma A.7

$$
\begin{aligned}
\sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \sqrt{T}[(\widehat{\boldsymbol{\varepsilon}}_g'\widehat{\boldsymbol{\varepsilon}}_g)^{-1}\widehat{\boldsymbol{\varepsilon}}_g'(\widehat{\boldsymbol{\varepsilon}}_y - \boldsymbol{\varepsilon}_y) + (\widehat{\boldsymbol{\varepsilon}}_g'\widehat{\boldsymbol{\varepsilon}}_g)^{-1}\widehat{\boldsymbol{\varepsilon}}_g'\boldsymbol{\eta} + (\widehat{\boldsymbol{\varepsilon}}_g'\widehat{\boldsymbol{\varepsilon}}_g)^{-1}\widehat{\boldsymbol{\varepsilon}}_g'(\boldsymbol{\varepsilon}_g - \widehat{\boldsymbol{\varepsilon}}_g)\boldsymbol{\beta}] \\
&= O_P(1)\frac{1}{\sqrt{T}}\widehat{\boldsymbol{\varepsilon}}_g'(\widehat{\boldsymbol{\varepsilon}}_y - \boldsymbol{\varepsilon}_y) + O_P(1)\frac{1}{\sqrt{T}}\widehat{\boldsymbol{\varepsilon}}_g'(\boldsymbol{\varepsilon}_g - \widehat{\boldsymbol{\varepsilon}}_g) + O_P(1)\frac{1}{\sqrt{T}}\boldsymbol{\eta}'(\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g) \\
&\quad + (\frac{1}{T}\boldsymbol{\varepsilon}_g'\boldsymbol{\varepsilon}_g)^{-1}\frac{1}{\sqrt{T}}\boldsymbol{\varepsilon}_g'\boldsymbol{\eta} \\
&= \sigma_g^{-2}\frac{1}{\sqrt{T}}\boldsymbol{\varepsilon}_g'\boldsymbol{\eta} + o_P(1) \xrightarrow{d} \mathcal{N}(0, \sigma_g^{-4}\sigma_{\eta g}^2).
\end{aligned} \tag{A.3}
$$

In the above, we used the condition that $|J|_0^4+|J|_0^2\log^2 N=o(T)$, $T|J|_0^4=o(N^2\min\{1,\nu_{\min}^4|J|_0^4\})$ and $\sqrt{\log N}|J|_0^2=o(N\nu_{\min}^2)$, whose sufficient conditions are $T|J|_0^4=o(N^2\min\{1,\nu_{\min}^4|J|_0^4\})$ and $|J|_0^4\log^2 N=o(T)$.

In addition, $\widehat{\sigma}_{\eta,g}^{-1}\widehat{\sigma}_g^2\sqrt{T}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}\mathcal{N}(0,1)$, follows from $\widehat{\sigma}_g^2:=\frac{1}{T}\widehat{\boldsymbol{\varepsilon}}_g'\widehat{\boldsymbol{\varepsilon}}_g\xrightarrow{P}\sigma_g^2$.

□

Proposition A.2. *Suppose* $T=O(\nu_{\min}^4N^2\log N)$, $|J|_0^2T=O(\nu_{\min}^2N^2\log N)$, $|J|_0^2=O(N\nu_{\min}^2\log N)$ *and* $|J|_0^2\log N=O(T)$, $|J|_0^2=o(N)$ *For all* $R\geq r$,
(i) $\frac{1}{T}\|\widehat{\mathbf{U}}'\boldsymbol{\theta}-\widehat{\mathbf{U}}'\widetilde{\boldsymbol{\theta}}\|^2=O_P(|J|_0\frac{\log N}{T})$ *and* $\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_1=O_P(|J|_0\sqrt{\frac{\log N}{T}})$.
(ii) $|\widehat{J}|_0=O_P(|J|_0)$.

Proof. (i) Let $L(\boldsymbol{\theta}):=\frac{1}{T}\sum_{t=1}^T(\mathbf{g}_t-\widehat{\boldsymbol{\alpha}}_g'\widehat{\mathbf{f}}_t-\boldsymbol{\theta}'\widehat{\mathbf{u}}_t)^2+\tau\|\boldsymbol{\theta}\|_1$,

$$d_t=\boldsymbol{\alpha}_g'\mathbf{f}_t-\widehat{\boldsymbol{\alpha}}_g'\widehat{\mathbf{f}}_t+(\mathbf{u}_t-\widehat{\mathbf{u}}_t)'\boldsymbol{\theta},\quad \boldsymbol{\Delta}=\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}.$$

Then $\mathbf{g}_t=\boldsymbol{\alpha}_g'\mathbf{f}_t+\boldsymbol{\theta}'\mathbf{u}_t+\varepsilon_{g,t}$, and $L(\widetilde{\boldsymbol{\theta}})\leq L(\boldsymbol{\theta})$ imply

$$\frac{1}{T}\sum_{t=1}^T[(\widehat{\mathbf{u}}_t'\boldsymbol{\Delta})^2+2(\varepsilon_{g,t}+d_t)\widehat{\mathbf{u}}_t'\boldsymbol{\Delta}]+\tau\|\widetilde{\boldsymbol{\theta}}\|_1\leq\tau\|\boldsymbol{\theta}\|_1.$$

It follows from Lemma A.5 that $\|\frac{1}{T}\widehat{\mathbf{U}}\boldsymbol{\varepsilon}_g\|_\infty\leq O_P(\sqrt{\frac{\log N}{T}})$. Also Lemma A.4 implies that

$$\begin{aligned}\|\frac{1}{T}\sum_{t=1}^T d_t\widehat{\mathbf{u}}_t\|_\infty &\leq \|\frac{1}{T}\widehat{\mathbf{U}}\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}\|_\infty+\|\frac{1}{T}\widehat{\mathbf{U}}\mathbf{E}(\mathbf{H}^{+'}\boldsymbol{\alpha}_g-\widehat{\boldsymbol{\alpha}}_g)\|_\infty+\|\frac{1}{T}\widehat{\mathbf{U}}\mathbf{F}\mathbf{H}'(\mathbf{H}^{+'}\boldsymbol{\alpha}_g-\widehat{\boldsymbol{\alpha}}_g)\|_\infty\\ &\quad+\|\frac{1}{T}\boldsymbol{\theta}'(\widehat{\mathbf{U}}-\mathbf{U})\widehat{\mathbf{U}}'\|_\infty\\ &\leq O_P(|J|_0\sqrt{\frac{\log N}{TN}}+|J|_0\frac{\log N}{T}+\frac{1}{N\nu_{\min}^2}+\nu_{\min}^{-1}\sqrt{\frac{\log N}{TN}}+\frac{|J|_0}{N\nu_{\min}}+\frac{|J|_0}{\nu_{\min}\sqrt{NT}}).\end{aligned}$$

Thus the "score" satisfies $\|\frac{1}{T}\sum_{t=1}^T 2(\boldsymbol{\varepsilon}_{g,t}+d_t)\widehat{\mathbf{u}}_t'\|_\infty\leq\tau/2$ for sufficiently large $C>0$ in $\tau=C\sigma\sqrt{\frac{\log N}{T}}$ with probability arbitrarily close to one, given $T=O(\nu_{\min}^4N^2\log N)$, $|J|_0^2T=O(\nu_{\min}^2N^2\log N)$, $|J|_0^2=O(N\nu_{\min}^2\log N)$ and $|J|_0^2\log N=O(T)$. Then by the standard argument in the lasso literature,

$$\frac{1}{T}\sum_{t=1}^T(\widehat{\mathbf{u}}_t'\boldsymbol{\Delta})^2+\frac{\tau}{2}\|\boldsymbol{\Delta}_{J^c}\|_1\leq\frac{3\tau}{2}\|\boldsymbol{\Delta}_J\|_1.$$

Meanwhile, by the restricted eigenvalue condition and Lemma A.4,

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{u}}_t'\boldsymbol{\Delta})^2 \geq \frac{1}{T}\sum_{t=1}^{T}(\mathbf{u}_t'\boldsymbol{\Delta})^2 - \|\boldsymbol{\Delta}\|_1^2\|\frac{1}{T}\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \mathbf{U}\mathbf{U}'\|_\infty \geq \|\boldsymbol{\Delta}\|_2^2(\phi_{\min} - o_P(1))$$

where the last inequality follows from $|J|_0 O_P(\nu_{\min}^{-2}\frac{1}{N} + \frac{\log N}{T}) = o_P(1)$ (Lemma A.3). From here, the desired convergence results follow from the standard argument in the lasso literature, we omit details for brevity, and refer to, e.g., Hansen and Liao (2018).

(ii) The proof of $|\widehat{J}|_0 = O_P(|J|_0)$ also follows from the standard argument in the lasso literature, we omit details but refer to the proof of Proposition D.1 of Hansen and Liao (2018) and Belloni et al. (2014).

$\square$

Lemma A.3. *(i)* $\|\frac{1}{T}\mathbf{E}'\mathbf{U}'\|_\infty = O_P(\sqrt{\frac{\log N}{TN}} + \frac{1}{N})$
(ii) $\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\| = O_P(\frac{1}{N})$, $\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\|_\infty = O_P(\sqrt{\frac{\log N}{TN}} + \frac{1}{N})$,
(iii) $\|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathbf{U})'\|_\infty + 2\|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})\mathbf{U}'\|_\infty = O_P(\nu_{\min}^{-2}\frac{1}{N} + \frac{\log N}{T})$.
(iv) $\|\frac{1}{T}\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \frac{1}{T}\mathbf{U}\mathbf{U}'\|_\infty = O_P(\nu_{\min}^{-2}\frac{1}{N} + \frac{\log N}{T})$.

Proof. Let $\widehat{\mathbf{F}} = (\widehat{\mathbf{f}}_1, \cdots, \widehat{\mathbf{f}}_T)'$. In addition, $\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}^+ = -\mathbf{B}\mathbf{H}^+\mathbf{E}'\widehat{\mathbf{F}}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} + \mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} + \mathbf{U}\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}$. Therefore,

$$\begin{aligned}\mathbf{U} - \widehat{\mathbf{U}} &= \widehat{\mathbf{B}}\widehat{\mathbf{F}}' - \mathbf{B}\mathbf{F}' = (\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}^+)\widehat{\mathbf{F}}' + \mathbf{B}\mathbf{H}^+\mathbf{E}' \\ &= -\mathbf{B}\mathbf{H}^+\mathbf{E}'\widehat{\mathbf{F}}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}' + \mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}' + \mathbf{U}\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}' + \mathbf{B}\mathbf{H}^+\mathbf{E}'. \quad (A.4)\end{aligned}$$

(i) We have

$$\|\frac{1}{T}\mathbf{U}\mathbf{E}\|_\infty \leq \sum_{k\leq r}\max_{i\leq N}|\frac{1}{TN}\sum_t\sum_j(u_{it}u_{jt} - \mathbb{E}\,u_{it}u_{jt})w_{k,j}| + O(\frac{1}{N}) = O_P(\sqrt{\frac{\log N}{TN}} + \frac{1}{N})$$

(ii) By Proposition A.1 , Lemma A.1 , $\nu_{\min} \gg N^{-1/2}$, and $\|\frac{1}{T}\mathbf{F}'\mathbf{U}'\|_\infty = O_P(\sqrt{\frac{\log N}{T}})$

$$\begin{aligned}\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\| &\leq \|\frac{1}{T}\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{E}\| + \|\frac{2}{T}\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{E}\| + \|\frac{1}{T}\mathbf{E}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{E}\| \\ &\leq O_P(\frac{1}{N}) \\ \|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\|_\infty &\leq \|\frac{1}{T}\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\|_\infty + \|\frac{1}{T}\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{U}'\|_\infty \\ &\quad + \|\frac{1}{T}\mathbf{E}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\|_\infty + \|\frac{1}{T}\mathbf{E}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{U}'\|_\infty\end{aligned}$$

$$\leq \quad O_P(\sqrt{\frac{\log N}{TN}} + \frac{1}{N}).$$

(iii) We have $\|\mathbf{H}^+\| = O(\nu_{\min}^{-1})$. Also, $\|\widehat{\mathbf{F}}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'\| \leq 1$. In addition, by Lemma A.1, $\|(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'\|^2 = \|(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\| \leq O_P(\frac{N}{T})$ and that $\|\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'\|^2 = \|\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\| = O_P(\frac{1}{T})$. Next, by Lemma A.1, $\|\mathbf{E}\| = O_P(\sqrt{\frac{T}{N}})$, and $\max_i \|\mathbf{b}_i\| < C$. Substitute the expansion (A.4), and by Proposition A.1,

$$\begin{aligned}
& \|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathbf{U})'\|_\infty + 2\|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})\mathbf{U}'\|_\infty \\
\leq \; & \|\frac{2}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{U}'\|_\infty + \|\frac{1}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{E}\mathbf{H}^{+'}\mathbf{B}'\|_\infty + \|\frac{3}{T}\mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\|_\infty \\
& + \|\frac{4}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\|_\infty + \|\frac{4}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{U}'\|_\infty \\
& + \|(\frac{6}{T}\mathbf{U}\mathbf{E} + \frac{3}{T}\mathbf{U}\mathbf{F}\mathbf{H}')(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{U}'\|_\infty + \|\frac{4}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}(\mathbf{H}\mathbf{F}'\mathbf{U}' + \mathbf{E}'\mathbf{U}')\|_\infty \\
& + \|\frac{2}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\|_\infty + \|\frac{3}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\mathbf{H}^{+'}\mathbf{B}'\|_\infty \\
\leq \; & \|\frac{C}{T}\mathbf{E}'\mathbf{U}'\|_\infty O_P(\nu_{\min}^{-1}) + \|\frac{C}{T}\mathbf{E}'\mathbf{E}\|O_P(\nu_{\min}^{-2}) + N\|\frac{C}{T}\mathbf{U}\mathbf{E}\|_\infty^2 + N\|\frac{C}{T}\mathbf{E}'\mathbf{E}\|\|\frac{1}{T}\mathbf{E}'\mathbf{U}'\|_\infty O_P(\nu_{\min}^{-1}) \\
& + O_P(\nu_{\min}^{-1})\|\frac{C}{T}\mathbf{E}'\mathbf{E}\|\|(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\|\|\mathbf{F}'\mathbf{U}'\|_\infty + \|\frac{6}{T}\mathbf{U}\mathbf{E}\|_\infty\|(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\|\|\mathbf{F}'\mathbf{U}'\|_\infty \\
& + \|\frac{3}{T}\mathbf{U}\mathbf{F}\|_\infty\|\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\|\|\mathbf{F}'\mathbf{U}'\|_\infty + O_P(\nu_{\min}^{-1})\|\frac{4}{T}\mathbf{E}'\mathbf{F}\|\|\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\|\|\mathbf{F}'\mathbf{U}'\|_\infty \\
& + O_P(\nu_{\min}^{-1})\|\frac{4}{T}\mathbf{E}'\mathbf{F}\|\|\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\|\|\mathbf{E}'\mathbf{U}'\|_\infty + O_P(\nu_{\min}^{-1})\|\frac{C}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\|_\infty + O_P(\nu_{\min}^{-2})\|\frac{C}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\| \\
= \; & O_P(\nu_{\min}^{-2}\frac{1}{N} + \frac{\log N}{T}).
\end{aligned}$$

Also, $\|\frac{1}{T}\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \frac{1}{T}\mathbf{U}\mathbf{U}'\|_\infty \leq \|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathbf{U})'\|_\infty + 2\|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})\mathbf{U}'\|_\infty \leq O_P(\nu_{\min}^{-2}\frac{1}{N} + \frac{\log N}{T})$.

□

Lemma A.4. *For all $R \geq r$,*
(i) $\|\frac{1}{T}\boldsymbol{\theta}'(\widehat{\mathbf{U}} - \mathbf{U})\widehat{\mathbf{U}}'\|_\infty \leq O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2})|J|_0$.
(ii) $\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{F}\| = O_P(\frac{1}{N\nu_{\min}} + \frac{1}{\sqrt{NT}})$, $\|\frac{1}{T}\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{F}\|_\infty = O_P(\sqrt{\frac{\log N}{T}} + \frac{1}{N\nu_{\min}})$.
(iii) $\|\frac{1}{T}\mathbf{E}'\widehat{\mathbf{U}}'\|_\infty \leq O_P(\sqrt{\frac{\log N}{TN}} + \frac{1}{N\nu_{\min}})$, $\|\frac{1}{T}\mathbf{F}'\widehat{\mathbf{U}}'\|_\infty \leq O_P(\sqrt{\frac{\log N}{T}} + \frac{1}{N\nu_{\min}^2})$,
(iv) $\|\frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{E}\| = |J|_0 O_P(\frac{1}{N} + \frac{1}{\sqrt{NT}})$, $\|\frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{F}\| = O_P(\sqrt{\frac{|J|_0}{T}})$,
(v) $\widehat{\boldsymbol{\alpha}}_g - \mathbf{H}^{+'}\boldsymbol{\alpha}_g = |J|_0 O_P(1 + \sqrt{\frac{N}{T}}) + O_P(\nu_{\min}^{-1})$, $\mathbf{H}'(\widehat{\boldsymbol{\alpha}}_g - \mathbf{H}^{+'}\boldsymbol{\alpha}_g) = O_P(\nu_{\min}^{-1}\frac{|J|_0}{N} + \sqrt{\frac{|J|_0}{T}} + \nu_{\min}^{-2}\frac{1}{N})$.

Proof. (i) By Lemma A.3 $\|\frac{1}{T}\boldsymbol{\theta}'(\widehat{\mathbf{U}}-\mathbf{U})\widehat{\mathbf{U}}'\|_\infty \leq \|\boldsymbol{\theta}\|_1\|\frac{1}{T}(\widehat{\mathbf{U}}-\mathbf{U})\widehat{\mathbf{U}}'\|_\infty \leq O_P(\frac{\log N}{T}+\frac{1}{N\nu_{\min}^2})|J|_0$.
(ii) Note $\mathbf{H}'\mathbf{H}^{+'}=\mathbf{I}$, Lemma A.3 shows $\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\| = O_P(\frac{1}{N})$, $\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\|_\infty = O_P(\sqrt{\frac{\log N}{TN}}+\frac{1}{N})$,

$$
\begin{aligned}
\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{F}\| &\leq \|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\mathbf{H}^{+'}\| + \|\frac{1}{T}\mathbf{E}'\mathbf{E}\mathbf{H}^{+'}\| + \|\frac{1}{T}\mathbf{E}'\mathbf{F}\| = O_P(\frac{1}{N\nu_{\min}}+\frac{1}{\sqrt{NT}}) \\
\|\frac{1}{T}\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{F}\|_\infty &\leq \|\frac{1}{T}\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\mathbf{H}^{+'}\|_\infty + \|\frac{1}{T}\mathbf{U}\mathbf{E}\mathbf{H}^{+'}\|_\infty + \|\frac{1}{T}\mathbf{U}\mathbf{F}\|_\infty \\
&\leq O_P(\sqrt{\frac{\log N}{T}}+\frac{1}{N\nu_{\min}}).
\end{aligned}
$$

(iii) By Lemma A.3 $\|\frac{1}{T}\mathbf{E}'\mathbf{U}'\|_\infty = O_P(\sqrt{\frac{\log N}{TN}}+\frac{1}{N})$ and (ii)

$$
\begin{aligned}
\|\frac{1}{T}\widehat{\mathbf{U}}\mathbf{E}\|_\infty &\leq \|\frac{1}{T}\mathbf{U}\mathbf{E}\|_\infty + \|\frac{1}{T}(\widehat{\mathbf{U}}-\mathbf{U})\mathbf{E}\|_\infty \\
&\leq \|\frac{1}{T}\mathbf{U}\mathbf{E}\|_\infty + \|\frac{1}{T}\mathbf{B}\mathbf{H}^{+}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\|_\infty + \|\frac{1}{T}\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\|_\infty + \|\frac{1}{T}\mathbf{B}\mathbf{H}^{+}\mathbf{E}'\mathbf{E}\|_\infty \\
&\leq O_P(\sqrt{\frac{\log N}{TN}}+\frac{1}{N\nu_{\min}}) \\
\|\frac{1}{T}\widehat{\mathbf{U}}\mathbf{F}\|_\infty &\leq \|\frac{1}{T}\mathbf{U}\mathbf{F}\|_\infty + \|\frac{1}{T}(\widehat{\mathbf{U}}-\mathbf{U})\mathbf{F}\|_\infty \\
&\leq \|\frac{1}{T}\mathbf{U}\mathbf{F}\|_\infty + \|\frac{1}{T}\mathbf{B}\mathbf{H}^{+}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{F}\|_\infty + \|\frac{1}{T}\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{F}\|_\infty + \|\frac{1}{T}\mathbf{B}\mathbf{H}^{+}\mathbf{E}'\mathbf{F}\|_\infty \\
&\leq O_P(\sqrt{\frac{\log N}{T}}+\frac{1}{N\nu_{\min}^2}).
\end{aligned}
$$

(iv) $\frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{E} = \frac{1}{NT}\boldsymbol{\theta}'(\mathbf{U}\mathbf{U}'-\mathbb{E}\,\mathbf{U}\mathbf{U}')\mathbf{W} + \frac{1}{NT}\boldsymbol{\theta}'\,\mathbb{E}\,\mathbf{U}\mathbf{U}'\mathbf{W}$. So

$$
\begin{aligned}
\mathbb{E}\,\|\frac{1}{NT}\boldsymbol{\theta}'(\mathbf{U}\mathbf{U}'-\mathbb{E}\,\mathbf{U}\mathbf{U}')\mathbf{W}\|^2 = \sum_{k=1}^{R}\frac{1}{N^2T^2}\operatorname{Var}(\sum_{t=1}^{T}\boldsymbol{\theta}'\mathbf{u}_t\mathbf{u}_t'\mathbf{w}_k) \\
\leq \frac{C}{N^2T^2}\|\boldsymbol{\theta}\|_1^2\max_{j,i\leq N}\sum_{q,v\leq N}\sum_{t,s\leq T}|\operatorname{Cov}(u_{it}u_{qt},u_{js}u_{vs})| \leq \frac{C|J|_0^2}{NT}.
\end{aligned}
$$

Also, $\|\frac{1}{NT}\boldsymbol{\theta}'\,\mathbb{E}\,\mathbf{U}\mathbf{U}'\mathbf{W}\| \leq \max_{j\leq N}\sum_k|w_{k,j}|\|\boldsymbol{\theta}\|_1\|\frac{1}{TN}\,\mathbb{E}\,\mathbf{U}\mathbf{U}'\|_1 \leq O(\frac{|J|_0}{N})$. Also,

$$
\begin{aligned}
\mathbb{E}\,\|\frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{F}\|^2 = \frac{1}{T^2}\operatorname{tr}\mathbb{E}\,\mathbf{F}'\,\mathbb{E}(\mathbf{U}'\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{U}|\mathbf{F})\mathbf{F} \leq \frac{C}{T}\|\,\mathbb{E}(\mathbf{U}'\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{U}|\mathbf{F})\|_1 \\
\leq \frac{C}{T}\max_t\sum_{s=1}^{T}|\,\mathbb{E}(\boldsymbol{\theta}'\mathbf{u}_t\mathbf{u}_s'\boldsymbol{\theta}|\mathbf{F})| \leq \frac{C}{T}\max_t\sum_{s=1}^{T}\|\,\mathbb{E}(\mathbf{u}_t\mathbf{u}_s'|\mathbf{F})\|_1\|\boldsymbol{\theta}\|_1\|\boldsymbol{\theta}\|_\infty \leq \frac{C|J|_0}{T}.
\end{aligned}
$$

(v) Since $\widehat{\boldsymbol{\alpha}}_g = (\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'\mathbf{G}$, simple calculations using Proposition A.1 yield

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}}_g - \mathbf{H}^{+'}\boldsymbol{\alpha}_g &= (\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'\mathbf{G} - \mathbf{H}^{+'}\boldsymbol{\alpha}_g \\
&= (\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g - (\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}_g + (\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\boldsymbol{\theta} + O_P(\sqrt{\frac{|J|_0}{T}}) \\
&= |J|_0 O_P(1+\sqrt{\frac{N}{T}}) + O_P(\nu_{\min}^{-1}) \\
\mathbf{H}'(\widehat{\boldsymbol{\alpha}}_g - \mathbf{H}^{+'}\boldsymbol{\alpha}_g) &= \mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g - \mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}_g + \mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\boldsymbol{\theta} + O_P(\sqrt{\frac{|J|_0}{T}}) \\
&= O_P(\nu_{\min}^{-1}\frac{|J|_0}{N} + \sqrt{\frac{|J|_0}{T}} + \nu_{\min}^{-2}\frac{1}{N}).
\end{aligned}
$$

□

Lemma A.5. *Suppose* $|J|_0 = o(N\nu_{\min}^2)$. *For any* $R \geq r$
(i) $\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\boldsymbol{\theta}\|^2 = O_P(\frac{|J|_0^2}{N} + \frac{|J|_0^2}{T} + \frac{|J|_0^{3/2}}{\nu_{\min}N\sqrt{T}})$, $\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_g\|^2 = O_P(\frac{1}{T})$,
(ii) $\|\frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})\boldsymbol{\varepsilon}_g\|_\infty = O_P(\frac{\nu_{\min}^{-1}}{\sqrt{NT}} + \frac{\sqrt{\log N}}{T})$, *and* $\|\frac{1}{T}\widehat{\mathbf{U}}\boldsymbol{\varepsilon}_g\|_\infty = O_P(\sqrt{\frac{\log N}{T}}) = \|\frac{1}{T}\widehat{\mathbf{U}}\boldsymbol{\varepsilon}_y\|_\infty$
(iii) $\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{U}}_{\widehat{J}}\widehat{\mathbf{U}}_{\widehat{J}}') > c_0$ *with probability approaching one.* $\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_g\|^2 = O_P(\frac{|J|_0\log N}{T}) = \frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_y\|^2$.
(iv) $\frac{1}{T}\|(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta}\|^2 = O_P(\frac{|J|_0^2 + \nu_{\min}^{-2}}{N} + \frac{|J|_0^2}{T} + \frac{\nu_{\min}^{-1}|J|_0^{3/2}}{N\sqrt{T}})$, $\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y = O_P(\frac{1}{\sqrt{NT}})$,
$\frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y = O_P(\frac{|J|_0}{T} + \frac{|J|_0}{\sqrt{NT}} + \frac{\nu_{\min}^{-1/2}|J|_0^{3/4}}{\sqrt{N}T^{3/4}})$.

Proof. (i) By Lemma A.4 (vi) and Proposition A.1,

$$
\begin{aligned}
\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\boldsymbol{\theta}\|^2 &= \frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{U}'\boldsymbol{\theta} + \frac{2}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{U}'\boldsymbol{\theta} \\
&\quad + \frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{U}'\boldsymbol{\theta} \\
&\leq O_P(\frac{|J|_0^2}{N} + \frac{|J|_0^2}{T} + \frac{|J|_0^{3/2}}{\nu_{\min}N\sqrt{T}}), \\
\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_g\|^2 &= \frac{1}{T}\boldsymbol{\varepsilon}_g'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g + \frac{2}{T}\boldsymbol{\varepsilon}_g'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\boldsymbol{\varepsilon}_g + \frac{1}{T}\boldsymbol{\varepsilon}_g'\mathbf{F}\mathbf{H}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\boldsymbol{\varepsilon}_g \\
&\leq O_P(\frac{N}{NT}) + O_P(\frac{1}{\sqrt{NT}})\frac{\nu_{\min}^{-1}}{\sqrt{T}} + O_P(\frac{1}{T}) = O_P(\frac{1}{T}).
\end{aligned}
$$

(ii) By (A.4)

$$
\begin{aligned}
\frac{1}{T}(\mathbf{U} - \widehat{\mathbf{U}})\boldsymbol{\varepsilon}_g &= -\frac{1}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g - \frac{1}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g + \frac{1}{T}\mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g \\
&\quad -\frac{1}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\boldsymbol{\varepsilon}_g - \frac{1}{T}\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\boldsymbol{\varepsilon}_g + \frac{1}{T}\mathbf{U}\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\boldsymbol{\varepsilon}_g
\end{aligned}
$$

$$+\frac{1}{T}\mathbf{UFH}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_g+\frac{1}{T}\mathbf{UFH}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{HF}'\boldsymbol{\varepsilon}_g+\frac{1}{T}\mathbf{BH}^{+}\mathbf{E}'\boldsymbol{\varepsilon}_g.$$

So by Lemmas A.1 and $\|\frac{1}{T}\mathbf{UE}\|_\infty=O_P(\sqrt{\frac{\log N}{TN}}+\frac{1}{N})$, $\|\frac{1}{T}(\widehat{\mathbf{U}}-\mathbf{U})\boldsymbol{\varepsilon}_g\|_\infty=O_P(\frac{\nu_{\min}^{-1}}{\sqrt{NT}}+\frac{\sqrt{\log N}}{T})$.

Also, with $\|\frac{1}{T}\mathbf{U}\boldsymbol{\varepsilon}_g\|_\infty=O_P(\sqrt{\frac{\log N}{T}})$ we have $\|\frac{1}{T}\widehat{\mathbf{U}}\boldsymbol{\varepsilon}_g\|_\infty=O_P(\sqrt{\frac{\log N}{T}})$. The proof for $\|\frac{1}{T}\widehat{\mathbf{U}}\boldsymbol{\varepsilon}_y\|_\infty$ is the same.

(iii) First, it follows from Lemma A.4 that $\|\frac{1}{T}\widehat{\mathbf{U}}\widehat{\mathbf{U}}'-\frac{1}{T}\mathbf{UU}'\|_\infty\leq O_P(\frac{\log N}{T}+\frac{\nu_{\min}^{-2}}{N})$.

Also by Proposition A.2, $|\widehat{J}|_0=O_P(|J|_0)$. Then with probability approaching one,

$$\begin{aligned}
\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{U}}_{\widehat{J}}\widehat{\mathbf{U}}_{\widehat{J}}') &\geq \lambda_{\min}(\frac{1}{T}\mathbf{U}_{\widehat{J}}\mathbf{U}_{\widehat{J}}')-\|\frac{1}{T}\widehat{\mathbf{U}}\widehat{\mathbf{U}}'-\frac{1}{T}\mathbf{UU}'\|_\infty|\widehat{J}|_0\\
&\geq \phi_{\min}-O_P(\frac{\log N}{T}+\frac{\nu_{\min}^{-2}}{N})|J|_0\geq c\\
\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_g\|^2 &= \frac{1}{T}\boldsymbol{\varepsilon}_g'\widehat{\mathbf{U}}_{\widehat{J}}'(\widehat{\mathbf{U}}_{\widehat{J}}\widehat{\mathbf{U}}_{\widehat{J}}')^{-1}\widehat{\mathbf{U}}_{\widehat{J}}\boldsymbol{\varepsilon}_g\leq\|\frac{1}{T}\boldsymbol{\varepsilon}_g'\widehat{\mathbf{U}}_{\widehat{J}}'\|^2\lambda_{\min}^{-1}(\frac{1}{T}\widehat{\mathbf{U}}_{\widehat{J}}\widehat{\mathbf{U}}_{\widehat{J}}')\\
&\leq c\|\frac{1}{T}\boldsymbol{\varepsilon}_g'\widehat{\mathbf{U}}'\|_\infty^2|\widehat{J}|_0\leq O_P(\frac{|J|_0\log N}{T}).
\end{aligned}$$

$\frac{1}{T}\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_y\|^2$ follows from the same proof.

(iv) Recall that $\|\boldsymbol{\alpha}_g'\|=\|\boldsymbol{\theta}'\mathbf{B}\|<C$. By part (i) and Lemma A.4,

$$\begin{aligned}
\frac{1}{T}\|\boldsymbol{\theta}'(\widehat{\mathbf{U}}-\mathbf{U})\|^2 &\leq \frac{1}{T}\|\boldsymbol{\theta}'\mathbf{BH}^{+}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\|^2+\frac{1}{T}\|\boldsymbol{\theta}'\mathbf{UP}_{\widehat{\mathbf{F}}}\|^2+\frac{1}{T}\|\boldsymbol{\theta}'\mathbf{BH}^{+}\mathbf{E}'\|^2\\
&\leq O_P(\frac{|J|_0^2+\nu_{\min}^{-2}}{N}+\frac{|J|_0^2}{T}+\frac{\nu_{\min}^{-1}|J|_0^{3/2}}{N\sqrt{T}}).\\
\|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y\| &\leq \|\frac{1}{T}\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\|\|\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y\|=O_P(\frac{1}{\sqrt{NT}})\\
\frac{1}{T}\boldsymbol{\theta}'\mathbf{UP}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y &\leq \frac{1}{T}\|\boldsymbol{\theta}'\mathbf{UP}_{\widehat{\mathbf{F}}}\|\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y\|=O_P(\frac{|J|_0}{T}+\frac{|J|_0}{\sqrt{NT}}+\frac{\nu_{\min}^{-1/2}|J|_0^{3/4}}{\sqrt{N}T^{3/4}}).
\end{aligned}$$

□

Lemma A.6. *For any $R\geq r$*

(i) $\frac{1}{T}\|\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\widehat{\mathbf{U}}'\boldsymbol{\theta}\|^2=O_P(|J|_0\frac{\log N}{T})$, $\frac{1}{T}\|\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta}\|^2=O_P(\frac{|J|_0\log N}{T}+\frac{|J|_0^2+\nu_{\min}^{-2}}{N}+\frac{|J|_0^2}{T})$.

(ii) $\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}(\widehat{\mathbf{U}}-\mathbf{U})'\boldsymbol{\theta}=|J|_0^2\sqrt{\frac{\log N}{T}}O_P(\frac{\log N}{T}+\frac{1}{N\nu_{\min}^2})$,

$\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta}\leq O_P(\frac{|J|_0\log N}{T}+\frac{|J|_0+\nu_{\min}^{-1}}{\sqrt{NT}}+\frac{\nu_{\min}^{-1/2}|J|_0^{3/4}}{\sqrt{N}T^{3/4}}+\sqrt{\frac{\log N}{T}}\frac{|J|_0^2}{N\nu_{\min}^2})$,

(iii) $\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{E}\|=O_P(\sqrt{\frac{|J|_0\log N}{N}}+\frac{\sqrt{T|J|_0}}{N\nu_{\min}})$, $\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{E}=O_P(\frac{|J|_0\log N}{T\sqrt{N}}+\frac{|J|_0\sqrt{\log N}}{N\nu_{\min}\sqrt{T}})$.

Proof. (i) First note that $\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\widehat{\mathbf{U}}'\boldsymbol{\theta} = \widehat{\mathbf{U}}'\widehat{\mathbf{m}}$, where

$$\widehat{\mathbf{m}} = (\widehat{m}_1, \cdots, \widehat{m}_N)' = \arg\min_{\mathbf{m}} \|\widehat{\mathbf{U}}'(\boldsymbol{\theta} - \mathbf{m})\| : \quad m_j = 0, \text{ for } j \notin \widehat{J}.$$

Thus by the definition of $\widehat{\mathbf{m}}$, Proposition A.2 and Lemma A.5,

$$\begin{aligned}
\frac{1}{T}\|\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\widehat{\mathbf{U}}'\boldsymbol{\theta}\|^2 &= \frac{1}{T}\|\widehat{\mathbf{U}}'\boldsymbol{\theta} - \widehat{\mathbf{U}}'\widehat{\mathbf{m}}\|^2 \leq \frac{1}{T}\|\widehat{\mathbf{U}}'\boldsymbol{\theta} - \widehat{\mathbf{U}}'\widetilde{\boldsymbol{\theta}}\|^2 \leq O_P(|J|_0\frac{\log N}{T}) \\
\frac{1}{T}\|\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta}\|^2 &\leq O_P(\frac{|J|_0\log N}{T}) + \frac{1}{T}\|(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta}\|^2 = O_P(\frac{|J|_0\log N + |J|_0^2}{T} + \frac{|J|_0^2 + \nu_{\min}^{-2}}{N})
\end{aligned}$$

where we used $\frac{\nu_{\min}^{-1}|J|_0^{3/2}}{N\sqrt{T}} = O_P(\frac{|J|_0\log N}{T})$ by our assumption.

(ii) Let $\boldsymbol{\Delta} = \boldsymbol{\theta} - \widehat{\mathbf{m}}$. Then $\dim(\boldsymbol{\Delta}) = O_P(|J|_0)$. Also, by Lemma A.4,

$$\boldsymbol{\Delta}'\frac{1}{T}(\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \mathbf{U}\mathbf{U}')\boldsymbol{\Delta} \leq \|\boldsymbol{\Delta}\|_1^2\|\frac{1}{T}(\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \mathbf{U}\mathbf{U}')\|_\infty \leq O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2})\|\boldsymbol{\Delta}\|^2|J|_0.$$

Also, $\|\boldsymbol{\Delta}\|^2 \leq \frac{C}{T}\|\mathbf{U}'\boldsymbol{\Delta}\|^2$ due to the spare eigenvalue condition on $\frac{1}{T}\mathbf{U}\mathbf{U}'$. Then $\widetilde{\boldsymbol{\theta}}_j = 0$ for $j \notin \widehat{J}$ implies $\|\widehat{\mathbf{U}}'\boldsymbol{\Delta}\| \leq \|\widehat{\mathbf{U}}'(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})\|$ and Proposition A.2 implies

$$\begin{aligned}
\|\boldsymbol{\theta} - \widehat{\mathbf{m}}\|_1^2 &\leq |J|_0\|\boldsymbol{\Delta}\|^2 \leq |J|_0\frac{1}{T}\|\mathbf{U}'\boldsymbol{\Delta}\|^2 \leq |J|_0\frac{1}{T}\|\widehat{\mathbf{U}}'\boldsymbol{\Delta}\|^2 + O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2})\|\boldsymbol{\Delta}\|^2|J|_0 \\
&\leq |J|_0\frac{1}{T}\|\widehat{\mathbf{U}}'\boldsymbol{\theta} - \widehat{\mathbf{U}}'\widetilde{\boldsymbol{\theta}}\|^2 + O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2})\|\boldsymbol{\Delta}\|^2|J|_0 \\
&\leq \frac{|J|_0^2\log N}{T} + O_P(\frac{|J|_0\log N}{T} + \frac{|J|_0}{N\nu_{\min}^2})\|\boldsymbol{\Delta}\|^2.
\end{aligned}$$

The above implies $\|\boldsymbol{\theta} - \widehat{\mathbf{m}}\|_1^2 \leq O_P(|J|_0^2\frac{\log N}{T})$. Hence by Lemma A.5,

$$\begin{aligned}
\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta} &\leq \|\frac{1}{\sqrt{T}}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\|\|\widehat{\mathbf{U}}(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta}\|_\infty\frac{\sqrt{|J|_0}}{T}\lambda_{\min}^{-1/2}(\frac{1}{T}\widehat{\mathbf{U}}_{\widehat{J}}\widehat{\mathbf{U}}_{\widehat{J}}') \\
&\leq |J|_0^2\sqrt{\frac{\log N}{T}}O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2}). \\
\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\widehat{\mathbf{U}}'\boldsymbol{\theta} &= \frac{1}{T}\boldsymbol{\varepsilon}_y'\widehat{\mathbf{U}}'(\boldsymbol{\theta} - \widehat{\mathbf{m}}) \leq \|\frac{1}{T}\boldsymbol{\varepsilon}_y'\widehat{\mathbf{U}}'\|_\infty\|\boldsymbol{\theta} - \widehat{\mathbf{m}}\|_1 \leq O_P(\frac{|J|_0\log N}{T}). \\
\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta} &\leq \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\widehat{\mathbf{U}}'\boldsymbol{\theta} + \frac{1}{T}\boldsymbol{\varepsilon}_y'(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta} - \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta} \\
&\leq O_P(\frac{|J|_0\log N}{T}) + \frac{1}{T}\boldsymbol{\theta}'\mathbf{B}\mathbf{H}^+\mathbf{E}'\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y + \frac{1}{T}\boldsymbol{\theta}'\mathbf{U}\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_y + \frac{1}{T}\boldsymbol{\theta}'\mathbf{B}\mathbf{H}^+\mathbf{E}'\boldsymbol{\varepsilon}_y \\
&\quad -\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}(\widehat{\mathbf{U}} - \mathbf{U})'\boldsymbol{\theta}
\end{aligned}$$

$$\leq \quad O_P(\frac{|J|_0 \log N}{T} + \frac{|J|_0 + \nu_{\min}^{-1}}{\sqrt{NT}} + \frac{\nu_{\min}^{-1/2}|J|_0^{3/4}}{\sqrt{N}T^{3/4}} + \sqrt{\frac{\log N}{T}}\frac{|J|_0^2}{N\nu_{\min}^2}).$$

(iii) By Lemma A.4,

$$\begin{aligned}
\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{E}\| &\leq \|\widehat{\mathbf{U}}_{\widehat{J}}'(\frac{1}{T}\widehat{\mathbf{U}}_{\widehat{J}}\widehat{\mathbf{U}}_{\widehat{J}}')^{-1}\|\frac{1}{T}\|\widehat{\mathbf{U}}\mathbf{E}\|_\infty\sqrt{|J|_0} \leq O_P(\sqrt{\frac{|J|_0\log N}{N}} + \frac{\sqrt{T|J|_0}}{N\nu_{\min}}) \\
\|\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{E}\| &\leq \|\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\|\|\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{E}\| = O_P(\frac{|J|_0\log N}{T\sqrt{N}} + \frac{|J|_0\sqrt{\log N}}{N\nu_{\min}\sqrt{T}})
\end{aligned}$$

□

Lemma A.7. *For any $R \geq r$,*
(i) $\frac{1}{T}\|\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g\|^2 = O_P(\frac{|J|_0^2+|J|_0\log N}{T} + \frac{|J|_0^2+\nu_{\min}^{-2}}{N} + \frac{|J|_0^{3/2}}{\nu_{\min}N\sqrt{T}}) = \frac{1}{T}\|\widehat{\boldsymbol{\varepsilon}}_y - \boldsymbol{\varepsilon}_y\|^2$.
(ii) $\frac{1}{T}\boldsymbol{\varepsilon}_y'(\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g) = O_P(\frac{|J|_0\log N}{T} + \frac{|J|_0+\nu_{\min}^{-1}}{\sqrt{NT}} + \frac{\nu_{\min}^{-1/2}|J|_0^{3/4}}{\sqrt{N}T^{3/4}} + \sqrt{\frac{\log N}{T}}\frac{|J|_0^2}{N\nu_{\min}^2})$. *The same rate applies to* $\frac{1}{T}\boldsymbol{\varepsilon}_g'(\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g)$, $\frac{1}{T}\boldsymbol{\eta}'(\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g)$, $\frac{1}{T}\boldsymbol{\varepsilon}_g'(\widehat{\boldsymbol{\varepsilon}}_y - \boldsymbol{\varepsilon}_y)$ *and* $\frac{1}{T}\boldsymbol{\varepsilon}_y'(\widehat{\boldsymbol{\varepsilon}}_y - \boldsymbol{\varepsilon}_y)$.
(iii) $\frac{1}{T}\widehat{\boldsymbol{\varepsilon}}_g'\widehat{\boldsymbol{\varepsilon}}_g = \frac{1}{T}\boldsymbol{\varepsilon}_g'\boldsymbol{\varepsilon}_g + o_P(1)$.

Proof. Note that $\widehat{\boldsymbol{\varepsilon}}_g = \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{G}$ and $\mathbf{G} = \mathbf{F}\boldsymbol{\alpha}_g + \mathbf{U}'\boldsymbol{\theta} + \boldsymbol{\varepsilon}_g$. Also, $\widehat{\mathbf{U}} = \mathbf{X}\mathbf{M}_{\widehat{\mathbf{F}}}$ implies

$$\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{P}_{\widehat{\mathbf{F}}} = 0, \text{ and } \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{M}_{\widehat{\mathbf{F}}} = \mathbf{M}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}.$$

Recall that $\mathbf{H}^+\mathbf{H} = \mathbf{I}$ and $\widehat{\mathbf{F}} = \mathbf{F}\mathbf{H}' + \mathbf{E}$, hence straightforward calculations yield

$$\begin{aligned}
\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g &= \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta} - \mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\boldsymbol{\theta} + \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{F}\boldsymbol{\alpha}_g - \mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_g - \mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_g \\
&= \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta} - \mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\boldsymbol{\theta} - \mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_g - \mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_g - (\mathbf{I} - \mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}})\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}_g. \quad \text{(A.5)}
\end{aligned}$$

It follows from Lemmas A.5, A.6 that $\frac{1}{T}\|\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g\|^2 = O_P(\frac{|J|_0^2+|J|_0\log N}{T} + \frac{|J|_0^2+\nu_{\min}^{-2}}{N} + \frac{|J|_0^{3/2}}{\nu_{\min}N\sqrt{T}})$. The proof for $\frac{1}{T}\|\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g\|^2$ follows similarly.

(ii) It follows from (A.5) and Lemmas A.5 A.6 that

$$\begin{aligned}
\frac{1}{T}\boldsymbol{\varepsilon}_y'(\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g) &= \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta} - \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{U}'\boldsymbol{\theta} - \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_g - \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{F}}}\boldsymbol{\varepsilon}_g \\
&\quad -\frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}_g - \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{F}}}\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}_g - \frac{1}{T}\boldsymbol{\varepsilon}_y'\mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{E}\mathbf{H}^{+'}\boldsymbol{\alpha}_g \\
&\leq O_P(\frac{|J|_0\log N}{T} + \frac{|J|_0 + \nu_{\min}^{-1}}{\sqrt{NT}} + \frac{\nu_{\min}^{-1/2}|J|_0^{3/4}}{\sqrt{N}T^{3/4}} + \sqrt{\frac{\log N}{T}}\frac{|J|_0^2}{N\nu_{\min}^2}).
\end{aligned}$$

The same proof applies to other terms as well.

(iii) It follows from parts (i) that all these terms are $o_P(1)$, given that $|J|_0^2 = o(\min\{T, N\})$, $|J|_0 \log N = o(T)$.

$\square$

A.4.2 The case $r = 0$: there are no factors.

Proof. In this case $\mathbf{x}_t = \mathbf{u}_t$. And we have

$$\widehat{\mathbf{F}} = \frac{1}{N}\mathbf{X}'\mathbf{W} = \frac{1}{N}\mathbf{U}'\mathbf{W} := \mathbf{E}.$$

Then $\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}) = \lambda_{\min}(\frac{1}{T}\mathbf{E}'\mathbf{E}) \geq \frac{c}{N}$ with probability approaching one, still by Lemma A.1. Hence $\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}$ is still invertible. In addition, $\widehat{\mathbf{U}} = \mathbf{X}\mathbf{M}_{\widehat{\mathbf{F}}}$ implies $\mathbf{U} - \widehat{\mathbf{U}} = \mathbf{U}\mathbf{P}_{\mathbf{E}}$. Also,

$$\begin{aligned} y_t &= \boldsymbol{\gamma}'\mathbf{u}_t + \boldsymbol{\varepsilon}_{y,t} \\ \mathbf{g}_t &= \boldsymbol{\theta}'\mathbf{u}_t + \boldsymbol{\varepsilon}_{g,t} \\ \boldsymbol{\varepsilon}_{y,t} &= \boldsymbol{\beta}'\boldsymbol{\varepsilon}_{g,t} + \eta_t \end{aligned}$$

Hence $\boldsymbol{\alpha}_g = \boldsymbol{\alpha}_y = 0$. Then $\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}} = \frac{1}{T}\mathbf{E}'\mathbf{E} = \frac{1}{N^2}\mathbf{W}'\operatorname{Cov}(\mathbf{u}_t)\mathbf{W} + O_P(\frac{1}{N\sqrt{T}})$. Hence with probability approaching one $\lambda_{\min}(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}) \geq cN^{-1}$. In addition, $\widehat{\boldsymbol{\alpha}}_y = (\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'\mathbf{U}'\boldsymbol{\gamma} + (\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\varepsilon}_y$ implies $\frac{1}{T}\sum_{t=1}^T(\widehat{\boldsymbol{\alpha}}_y'\widehat{\mathbf{f}}_t)^2 = O_P(\frac{|J|_0^2}{N} + \frac{|J|_0^2}{T})$.

As for the "score" $\max_i |\frac{1}{T}\sum_t(\varepsilon_{g,t} + d_t)\widehat{u}_{it}|$ in the proof of Proposition A.2, note that

$$\begin{aligned} \max_{i,j\leq N} |\frac{1}{T}\sum_t(\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})| &\leq \frac{3}{T}\|\mathbf{U}\mathbf{P}_{\mathbf{E}}\mathbf{U}'\|_\infty = O_P(\frac{1}{N} + \frac{\log N}{T}) \\ \max_{i\leq N} |\frac{1}{T}\sum_t \widehat{\boldsymbol{\alpha}}_y'\widehat{\mathbf{f}}_t\widehat{u}_{it}| &= O_P(\frac{|J|_0}{N} + \frac{|J|_0\log N}{T}) \\ \max_{i\leq N} |\frac{1}{T}\sum_t \widehat{u}_{it}(\mathbf{u}_t - \widehat{\mathbf{u}}_t)'\boldsymbol{\theta}| &= \frac{1}{T}\|\mathbf{U}\mathbf{P}_{\mathbf{E}}\mathbf{U}'\|_\infty O_P(|J|_0) = O_P(\frac{|J|_0}{N} + \frac{|J|_0\log N}{T}) \\ \max_{i\leq N} |\frac{1}{T}\sum_t \widehat{u}_{it}\varepsilon_{g,t}| &= O_P(\frac{\sqrt{\log N}}{T} + \frac{1}{\sqrt{TN}}). \end{aligned} \tag{A.6}$$

As for the residual, note that $\widehat{\boldsymbol{\varepsilon}}_g = \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{M}_{\mathbf{E}}\mathbf{G}$ and $\mathbf{G} = \mathbf{U}'\boldsymbol{\theta} + \boldsymbol{\varepsilon}_g$. Then

$$\widehat{\boldsymbol{\varepsilon}}_g - \boldsymbol{\varepsilon}_g = \mathbf{M}_{\widehat{\mathbf{U}}_{\widehat{J}}}\mathbf{U}'\boldsymbol{\theta} - \mathbf{P}_{\mathbf{E}}\mathbf{U}'\boldsymbol{\theta} - \mathbf{P}_{\widehat{\mathbf{U}}_{\widehat{J}}}\boldsymbol{\varepsilon}_g - \mathbf{P}_{\mathbf{E}}\boldsymbol{\varepsilon}_g.$$

All the proofs in Section A.4.1 carry over. In fact, all terms involving $\boldsymbol{\alpha}_g, \mathbf{H}$ and $\mathbf{H}^+$ can be set to zero.

In addition, in the case $R = r = 0$, the setting/estimators are the same as in Belloni et al. (2014). $\square$

A.4.3 Proof of Corollary 3.1.

Proof. The corollary immediately follows from Theorem 3.2. If there exist a pair (r, R) that violate the conclusion of the corollary, then it also violates the conclusion of Theorem 3.2. This finishes the proof. $\square$

A.5 Proof of Theorem 3.3

Proof. In the proof of Theorem 3.3 we assume $R \geq r$.

(i) When $r > 0$, by Lemma A.3,

$$\max_{i,j\leq N} |\frac{1}{T}\sum_t (\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})| \leq \|\frac{1}{T}\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \frac{1}{T}\mathbf{U}\mathbf{U}'\|_\infty \leq O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2}).$$

When $r = 0$ and $R > 0$, by (A.6), $\max_{i,j\leq N} |\frac{1}{T}\sum_t (\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})| \leq O_P(\frac{\log N}{T} + \frac{1}{N\nu_{\min}^2})$.

In both cases, part (i) implies, for $\nu_{\min}^2 \gg \frac{1}{\sqrt{N}}$ or $\nu_{\min}^2 \gg \frac{1}{N}\sqrt{\frac{T}{\log N}}$,

$$\begin{aligned}\max_{i,j\leq N} |s_{u,ij} - \mathbb{E}\, u_{it}u_{jt}| &\leq \max_{i,j\leq N} |\frac{1}{T}\sum_t \widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt}| + \max_{i,j\leq N} |\frac{1}{T}\sum_t u_{it}u_{jt} - \mathbb{E}\, u_{it}u_{jt}| \\ &\leq O_P(\sqrt{\frac{\log N}{T}} + \frac{1}{N\nu_{\min}^2}) = O_P(\sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}).\end{aligned}$$

where $\max_{i,j\leq N} |\frac{1}{T}\sum_t u_{it}u_{jt} - \mathbb{E}\, u_{it}u_{jt}| = O_P(\sqrt{\frac{\log N}{T}})$.

Given this convergence, the convergence of $\widehat{\mathbf{\Sigma}}_u$ and $\widehat{\mathbf{\Sigma}}_u^{-1}$ in (ii)(iii) then follows from the same proof of Theorem A.1 of Fan et al. (2013). We thus omit it for brevity. Finally, the case $r = R = 0$ is the usual case of sparse thresholding as in Bickel and Levina (2008). $\square$

A.6 Proof of Theorem 3.4

Proof. First note that when $R = r$, by (A.2)

$$\|(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} - (\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}\| \leq O_P(\frac{1}{N} + \frac{\nu_{\max}(\mathbf{H})}{\sqrt{TN}})\frac{1}{\nu_{\min}^4(\mathbf{H})}.$$

Also by the proof of Theorem 2.1 for $\|(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\| + \|(\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}\| \leq \frac{c}{\nu^2_{\min}(\mathbf{H})}$. Because $\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{G}} = \mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}' + \mathbf{F}\mathbf{H}'[(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} - (\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}]\mathbf{H}\mathbf{F}' + \widehat{\mathbf{F}}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'$, we have

$$\begin{aligned}
\|\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{G}}\|_F^2 &= \operatorname{tr}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}'(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{E} + \operatorname{tr}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{E} \\
&\quad + 2\operatorname{tr}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}'[(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} - (\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}]\mathbf{H}\mathbf{F}'\mathbf{E} \\
&\quad + \operatorname{tr}[(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} - (\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}]\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}'[(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} - (\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}]\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}' \\
&\quad + 2\operatorname{tr}\mathbf{F}\mathbf{H}'[(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1} - (\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}')^{-1}]\mathbf{H}\mathbf{F}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}' \\
&\quad + 2\operatorname{tr}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{E}'\mathbf{E} \\
&\quad + 2\operatorname{tr}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{E}(\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\mathbf{H}\mathbf{F}'\mathbf{E} \\
&= 2\operatorname{tr}\mathbf{H}'^{-1}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{H}^{-1}\mathbf{E}'\mathbf{E} + O_P\left(\frac{1}{TN\nu^2_{\min}} + \frac{1}{N^2\nu^4_{\min}} + \frac{1}{N\sqrt{NT}\nu^3_{\min}}\right).
\end{aligned}$$

Write $X := 2\operatorname{tr}\mathbf{H}'^{-1}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{H}^{-1}\mathbf{E}'\mathbf{E} = \operatorname{tr}(\mathbf{A}\frac{1}{T}\mathbf{E}'\mathbf{E})$ and $\mathbf{A} := 2\mathbf{H}'^{-1}(\frac{1}{T}\mathbf{F}'\mathbf{F})^{-1}\mathbf{H}^{-1}$. Now

$$\text{MEAN} = \mathbb{E}(X|\mathbf{F}, \mathbf{W}) = \operatorname{tr}\mathbf{A}\frac{1}{N^2}\mathbf{W}'(\mathbb{E}\,\mathbf{u}_t\mathbf{u}_t'|\mathbf{F})\mathbf{W} = \operatorname{tr}\mathbf{A}\frac{1}{N^2}\mathbf{W}'\mathbf{\Sigma}_u\mathbf{W}.$$

We note that $\operatorname{Var}(X|\mathbf{F}) = \frac{1}{TN^2}\sigma^2$ and that $N\sqrt{T}\frac{(X-\text{MEAN})}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$ due to the serial indepence of $\mathbf{u}_t\mathbf{u}_t'$ conditionally on $\mathbf{F}$ and that $\mathbb{E}\|\frac{1}{\sqrt{N}}\mathbf{W}'\mathbf{u}_t\|^4 < C$. In addition, Lemma A.8 below shows that with $\widehat{\text{MEAN}} = \operatorname{tr}\widehat{\mathbf{A}}\frac{1}{N^2}\mathbf{W}'\widehat{\mathbf{\Sigma}}_u\mathbf{W}$, and $\widehat{\mathbf{A}} = 2(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}$, we have

$$(\widehat{\text{MEAN}} - \text{MEAN})N\sqrt{T} = o_P(1).$$

Also, the same lemma shows $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$. As a result

$$\frac{\|\mathbf{P}_{\widehat{\mathbf{F}}} - \mathbf{P}_{\mathbf{G}}\|_F^2 - \widehat{\text{MEAN}}}{\frac{1}{N\sqrt{T}}\widehat{\sigma}} = \frac{X - \text{MEAN}}{\frac{1}{N\sqrt{T}}\sigma} + o_P(1) \xrightarrow{d} \mathcal{N}(0,1).$$

given that $\sigma > 0$, $\sqrt{T} = o(N)$.

$\square$

Lemma A.8. *Suppose $R = r$. Let $g_{NT} := \nu^{-2}_{\min}\frac{1}{N} + \frac{\log N}{T}$.*
(i) $\widehat{\text{MEAN}} - \text{MEAN} = O_P(\frac{g^2_{NT}}{N^2\nu^2_{\min}})\sum_{\sigma_{u,ij}\neq 0} 1 + O_P(\frac{1}{N^2\nu^4_{\min}} + \frac{1}{N\sqrt{NT}\nu^3_{\min}})$.
(ii) $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$.

Proof. By lemma A.3,

$$\max_{ij}|\frac{1}{T}\sum_t u_{it}(\widehat{u}_{jt} - u_{jt})| \leq O_P(g_{NT}).$$

(i) Recall $\mathbf{A} := 2\mathbf{H}'^{-1}(\frac{1}{T}\mathbf{F}'\mathbf{F})^{-1}\mathbf{H}^{-1}$. Note that $\|\mathbf{A}\| = O_P(\frac{1}{\nu^2_{\min}(\mathbf{H})})$. We now bound $\frac{1}{N}\mathbf{W}'(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{W}$. For simplicity we focus on the case $r = R = 1$ and hard-thresholding estimator. The proof of SCAD thresholding follows from the same argument. We have

$$\frac{1}{N}\mathbf{W}'(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{W} = \frac{1}{N}\sum_{\sigma_{u,ij}=0} w_i w_j \widehat{\sigma}_{u,ij} + \frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} w_i w_j(\widehat{\sigma}_{u,ij} - \sigma_{u,ij}) := a_1 + a_2.$$

Term a_1 satisfies: for any $\epsilon > 0$, when C in the threshold is large enough,

$$\mathbb{P}(a_1 > (NT)^{-2}) \leq \mathbb{P}(\max_{\sigma_{u,ij}=0} |\widehat{\sigma}_{u,ij}| \neq 0) \leq \mathbb{P}(|s_{u,ij}| > \tau_{ij}, \text{ for some } \sigma_{u,ij} = 0) < \epsilon.$$

Thus $a_1 = O_P((NT)^{-2})$. The main task is to bound $a_2 = \frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} w_i w_j(\widehat{\sigma}_{u,ij} - \sigma_{u,ij})$.

$$\begin{aligned}
a_2 &= a_{21} + a_{22},\\
a_{21} &= \frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} w_i w_j \frac{1}{T}\sum_t (\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})\\
a_{22} &= \frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} w_i w_j \frac{1}{T}\sum_t (u_{it}u_{jt} - \mathbb{E}\, u_{it}u_{jt}).
\end{aligned}$$

Now for $\omega_{NT} := \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$, by part (i),

$$\begin{aligned}
a_{21} &= \frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} w_i w_j \frac{1}{T}\sum_t (\widehat{u}_{it} - u_{it})(\widehat{u}_{jt} - u_{jt}) + \frac{2}{N}\sum_{\sigma_{u,ij}\neq 0} w_i w_j \frac{1}{T}\sum_t u_{it}(\widehat{u}_{jt} - u_{jt})\\
&\leq [\max_i \frac{1}{T}\sum_t (\widehat{u}_{it} - u_{it})^2 + \max_{ij} |\frac{1}{T}\sum_t u_{it}(\widehat{u}_{jt} - u_{jt})|] \frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} 1\\
&\leq O_P(g^2_{NT})\frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} 1.
\end{aligned}$$

As for a_{22}, due to $\frac{1}{N}\sum_{\sigma_{u,mn}\neq 0}\sum_{\sigma_{u,ij}\neq 0} |\operatorname{Cov}(u_{it}u_{jt}, u_{mt}u_{nt})| < C$ and serial independence,

$$\begin{aligned}
\operatorname{Var}(a_{22}) &\leq \frac{1}{N^2T^2}\sum_{s,t\leq T}\sum_{\sigma_{u,mn}\neq 0}\sum_{\sigma_{u,ij}\neq 0} |\operatorname{Cov}(u_{it}u_{jt}, u_{ms}u_{ns})|\\
&\leq \frac{1}{N^2T}\sum_{\sigma_{u,mn}\neq 0}\sum_{\sigma_{u,ij}\neq 0} |\operatorname{Cov}(u_{it}u_{jt}, u_{mt}u_{nt})| \leq O(\frac{1}{NT}).
\end{aligned}$$

Together $a_2 = O_P(g_{NT}^2)\frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} 1 + O_P(\frac{1}{\sqrt{NT}})$. Therefore

$$\frac{1}{N}\mathbf{W}'(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{W} = O_P(g_{NT}^2)\frac{1}{N}\sum_{\sigma_{u,ij}\neq 0} 1 + O_P(\frac{1}{\sqrt{NT}}).$$

This implies

$$\begin{aligned} |\widehat{\mathrm{MEAN}} - \mathrm{MEAN}| &\leq \frac{C}{N}\|\mathbf{A}\|\|\frac{1}{N}\mathbf{W}'(\mathbf{\Sigma}_u - \widehat{\mathbf{\Sigma}}_u)\mathbf{W}\| + O_P(\frac{1}{N})\|\mathbf{A} - 2(\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}})^{-1}\| \\ &\leq O_P(\frac{g_{NT}^2}{N^2\nu_{\min}^2})\sum_{\sigma_{u,ij}\neq 0} 1 + O_P(\frac{1}{N^2\nu_{\min}^4} + \frac{1}{N\sqrt{NT}\nu_{\min}^3}). \end{aligned}$$

(ii) First, note that $|\sigma^2 - f(\mathbf{A},\mathbf{V})| \to 0$ by the assumption. In addition, it is easy to show that $\|\widehat{\mathbf{A}} - \mathbf{A}\| = o_P(1)$ and $\|\widehat{\mathbf{V}} - \mathbf{V}\| \leq \frac{1}{N}\|\mathbf{W}\|^2\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\| = o_P(1)$. Since $f(\mathbf{A},\mathbf{V})$ is continuous in $(\mathbf{A},\mathbf{V})$ due to the property of the normality of $\mathbf{Z}_t$, we have $|f(\mathbf{A},\mathbf{V}) - f(\widehat{\mathbf{A}},\widehat{\mathbf{V}})| = o_P(1)$. Hence $|f(\widehat{\mathbf{A}},\widehat{\mathbf{V}}) - \sigma^2| = o_P(1)$. This finishes the proof since $\widehat{\sigma}^2 := f(\widehat{\mathbf{A}},\widehat{\mathbf{V}})$. □

References

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B* **75** 603–680.

Hansen, C. and Liao, Y. (2018). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory* 1–45.