Projected Principal Component Analysis in Factor Models

Jianqing Fan *,†, Yuan Liao ‡ and Weichen Wang *

*Department of Operations Research and Financial Engineering, Princeton University
† Bendheim Center for Finance, Princeton University
‡ Department of Mathematics, University of Maryland

Abstract

This paper introduces a Projected Principal Component Analysis (Projected-PCA), which is based on the projection of the data matrix onto a given linear space before performing the principal component analysis. When it applies to high-dimensional factor analysis, the projection removes idiosyncratic noisy components. We show that the unobserved latent factors can be more accurately estimated than the conventional PCA if the projection is genuine, or more precisely the factor loading matrices are related to the projected linear space, and that they can be estimated accurately when the dimensionality is large, even when the sample size is finite. In an effort to more accurately estimating factor loadings, we propose a flexible semi-parametric factor model, which decomposes the factor loading matrix into the component that can be explained by subject-specific covariates and the orthogonal residual component. The covariates effect on the factor loadings are further modeled by the additive model via sieve approximations. By using the newly proposed Projected-PCA, the rates of convergence of the smooth factor loading matrices are obtained, which are much faster than those of the conventional factor analysis. The convergence is achieved even when the sample size is finite and is particularly appealing in the high-dimension-low-sample-size situation. This leads us to developing nonparametric tests on whether observed covariates have explaining powers on the loadings and whether they fully explain the loadings. Finally, the proposed method is illustrated by both simulated data and the returns of the components of the S&P 500 index.

*Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: jqfan@princeton.edu, yuanliao@umd.edu, weichenw@princeton.edu. This project was supported by National Science Foundation grants DMS-1206464 and 2R01-GM072611-9.
Keywords: Additive models; approximate factor models; high-dimensionality; loading matrix modeling; semi-parametric efficiency; sieve approximation

1 Introduction

Factor analysis is one of the most useful tools for understanding common dependence among multivariate outputs. Suppose that we observe data \( \{y_{it}\}_{i \leq N, t \leq T} \) that can be decomposed as

\[
y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{kt} + u_{it}, \quad i = 1, \cdots, p, \quad t = 1, \cdots, T
\]

where \( \{f_{t1}, \cdots, f_{tK}\} \) are unobservable common factors; \( \{\lambda_{i1}, \cdots, \lambda_{iK}\} \) are corresponding factor loadings for variable \( i \), and \( u_{it} \) denotes the idiosyncratic error in the model. Here \( p \) and \( T \) respectively denote the dimension and sample size of the data. Model (1.1) has been extensively studied in the econometric literature, in which \( y_t = (y_{1t}, \cdots, y_{pt}) \) is the vector of economic outputs at time \( t \) such as the macro-economic variables at quarter \( t \) or excessive returns for individual assets on day \( t \). The unknown factors and loadings are typically estimated by the principal components (PC) method and the separations between the common factors and idiosyncratic components are characterized via pervasiveness assumptions. See, for instance, Stock and Watson (2002), Bai (2003), Bai and Ng (2002), Breitung and Tenhofen (2011), Choi (2012), Lam and Yao (2012), among others.

The factor model (1.1) has also been extensively studied in the statistics literature. In this case, \( t \) can represent the \( t^{th} \) microarray, proteomic or fMRI-image, whereas \( i \) represents a gene or protein or a voxel and \( y_t \) is expression profiles or blood oxygenation level dependent (BOLD) measurements. See, for example, Leek and Storey (2008); Friguet et al. (2009); Efron (2010); Desai and Storey (2012); Fan et al. (2012). The separations between the common factors and idiosyncratic components are carried out by the low-rank plus sparsity decomposition. See, for example, Candès and Recht (2009); Fan et al. (2011); Koltchinskii et al. (2011); Negahban and Wainwright (2011); Fan et al. (2013); Cai et al. (2013); Mā (2013). In the absence of the common factors, the problem becomes estimating a large sparse covariance matrix, which gains significant attentions in recent statistics literature (Bickel and Levina, 2008; Lam and Fan, 2009; Cai et al., 2010; Cai and Yuan, 2012; Rigollet and Tsybakov, 2012). These have further expanded into robust estimation based on Kendall’s tau estimation (Liu et al., 2012; Xue and Zou, 2012).

Accurately estimating the factor loadings and unobserved factors are very important in statistical and financial applications. In calculating the false-discovery proportion for large-scale hypothesis testing, one needs to adjust accurately the common dependence via
subtracting it from the data in (1.1) (Leek and Storey 2008; Friguet et al. 2009; Efron 2010; Desai and Storey 2012; Fan et al. 2012). In financial applications, we would like to understand accurately how each individual stock depends on unobserved common factors in order to appreciate its relative performance and risks. In the aforementioned applications, dimensionality is much higher than sample-size. However, the existing asymptotic analysis shows that the consistent estimation of the parameters in model (1.1) requires a relatively large $T$. In particular, the individual loadings can be estimated no faster than $O_p(T^{-1/2})$. But large sample sizes are not always available. Even with the availability of “Big Data”, heterogeneity and other issues make directly application of (1.1) with large $T$ infeasible. For instance, in financial applications, to pertain the stationarity in model (1.1) with loading coefficients time-independent, a relatively short time series is often used. To make observed data more independently, monthly returns are frequently used to reduce the serial correlations, yet a monthly data over three consecutive years contain merely 36 observations.

To overcome the aforementioned problems, it is necessary to model the loading matrix via additional explanatory variables. Let $X_i = (X_{i1}, \cdots, X_{id})$ be the $d$-dimensional covariates associated with $i^{th}$ variables. In the seminal works of Connor and Linton (2007) and Connor et al. (2012), the authors pioneered the study of the following semi-parametric factor model:

$$y_{it} = \sum_{k=1}^{K} g_k(X_i) f_{kt} + u_{it}, \quad i = 1, \cdots, p, t = 1, \cdots, T,$$  \hspace{1cm} (1.2)

where loading coefficients in (1.1) are modeled as $\lambda_{ik} = g_k(X_i)$ for some functions $g_k(\cdot)$. For instance, in health studies, $X_i$ can be individual characteristics (e.g. age, weight, clinical and genetic information); in financial applications $X_i$ can be a vector of firm-specific characteristics (market capitalization, price-earning ratio, etc).

The semiparametric model (1.2), however, can be restrictive in many cases, as it requires that the loading matrix be fully explained by the covariates. A natural relaxation is the following semiparametric model

$$\lambda_{ik} = g_k(X_i) + \gamma_{ik}, \quad i = 1, \cdots, p, k = 1, \cdots, K,$$  \hspace{1cm} (1.3)

where $\gamma_{ik}$ is the component of loading coefficient that can not be explained by the covariates $X_i$. Let $\gamma_i = (\gamma_{i1}, \cdots, \gamma_{iK})'$. Following the idea of mixed-effect models, we assume that $\{\gamma_i\}_{1 \leq p}$ are i.i.d. realizations from a population with mean zero, independent of covariates
and idiosyncratic $U$. In other words, we impose the following factor structure

$$y_{it} = \sum_{k=1}^{K} \{g_k(X_i) + \gamma_{ik}\} f_{kt} + u_{it}, \quad i = 1, \cdots, p; t = 1, \cdots, T,$$

which reduces to model (1.2) when $\gamma_{ik} = 0$ and model (1.1) when $g_k(\cdot) = 0$. When $X_i$ genuinely explains a part of loading coefficients $\lambda_{ik}$, the variability of $\gamma_{ik}$ is smaller than that of $\lambda_{ik}$. Hence, the coefficient $\gamma_{ik}$ can be more accurately estimated by using regression model (1.3), as long as the functions $g_k(\cdot)$ can be accurately estimated. Let $Y$ be the $p \times T$ matrix of $y_{it}$, $F$ be the $T \times K$ matrix of $f_{kt}$, $G(X)$ be the $p \times K$ matrix of $g_k(X_i)$, $\Gamma$ be the $p \times K$ matrix of $\gamma_{ik}$, and $U$ be $p \times T$ matrix of $u_{it}$. Then model (1.4) can be written in a more compact matrix form:

$$Y = \{G(X) + \Gamma\}F' + U.$$

We propose a Projected-PC estimator for both the loading functions and the factors. Our estimator differs from the traditional methods for factor analysis (e.g., [Stock and Watson 2002], [Bai and Ng 2002]) in that, it is constructed by first projecting $Y$ onto the sieve space spanned by $\{X_i\}_{i \leq p}$, then applying PCA (principal component analysis) on the projected data. Due to the approximate orthogonality condition of $X$, $U$ and $\Gamma$, the projection of $Y$ is approximately $G(X)F'$, as smoothing projection reduces noise in $\Gamma$ and $U$ substantially. Therefore, applying PCA to the projected data allows us to work directly on the sample covariance of $G(X)F'$, which is $G(X)G(X)'$ under normalization conditions. This substantially improves the estimation accuracy, and also facilitates the theoretical analysis. In contrast, the traditional PC method is no longer suitable in the current context.

The asymptotic properties of the estimators are carefully studied. We demonstrate that as long as the projection is genuine, the consistency of the proposed estimator for latent factors and loading matrices requires only $p \to \infty$, and $T$ does not need to grow, which is attractive in the typical high-dimension-low-sample-size situations. In addition, if both $p$ and $T$ grow simultaneously, then with sufficiently smooth $g_k(\cdot)$, using the sieve approximation, the rate of convergence for the estimators is much faster than those of the existing results for model (1.1). Typically, the loading functions can be estimated at a convergence rate $O_p((pT)^{-1/2})$, and the factor can be estimated at $O_p(p^{-1})$.

Model (1.3) implies a decomposition of the loading matrix:

$$\Lambda = G(X) + \Gamma, \quad E(\Gamma|X) = 0,$$

where $G(X)$ and $\Gamma$ are orthogonal loading components in the sense that $EG(X)\Gamma' = 0$. We
conduct two specification tests for the hypotheses:

$$H_0^1 : G(X) = 0, \quad \text{and} \quad H_0^2 : \Gamma = 0.$$ 

The first testing problem tests whether the observed covariates have explaining power on the loadings. If the null hypothesis is rejected, it gives us the theoretical basis to employ the projected PCA, as the projection is now genuine. Our empirical study on the asset returns shows that firm market characteristics do have explaining power of the factor loadings, which lends further support to our projected-PCA method. The second tests whether covariates fully explain the loadings. Our aforementioned empirical study also shows that model (1.2) used in the financial econometrics literature is inadequate and more generalized model (1.5) is necessary. As claimed earlier, even if $H_0^2$ does not hold, as long as $G(X) \neq 0$, the Projected-PCA can still consistently estimate the factors as $p \to \infty$, and $T$ may or may not grow. Our simulated experiments confirm that the estimation accuracy is gained more significantly for small $T$’s. This shows one of the benefits of using our projected-PC method over the traditional methods in the literature.

In addition, as a further illustration of the benefits of using projected data, we apply the projected-PCA to consistently estimate the number of factors, which is a tuning-parameter-free method similar to those in Ahn and Horenstein (2013) and Lam and Yao (2012). Our contribution also lies on the allowance of semi-weak factors, which have caused growing interests in the recent literature on high-dimensional factor analysis (e.g., Chudik and Pesaran 2013, Onatski 2012).

The rest of the paper is organized as follows. Section 2 introduces the new projected-PC method and defines the corresponding estimators for the loadings and factors. Sections 3 and 4 provide asymptotic analysis of the introduced estimators. Section 5 introduces new specification tests for the orthogonal decomposition of the semi-parametric loadings. Section 6 concerns about estimating the number of factors. Section 7 presents numerical results. Finally, Section 8 concludes. All the proofs are given in the appendix and the supplementary material.
2 Projected Principal Component Analysis

2.1 Projected principal components for factor analysis

In the high-dimensional factor model, let $\Lambda$ be the $p \times K$ matrix of loadings. Then the general model (1.1) can be written as

$$ Y = \Lambda F' + U. \quad (2.1) $$

Assuming that the mean of each row of $Y$ has been centered to zero, the classical PC method estimates the factors and loadings by taking the first $K$ principal components of $\frac{1}{T}YY'$ (or the $T \times T$ matrix $\frac{1}{T}Y'Y$).

Suppose now that we have a known linear space $\mathcal{X}$ that is orthogonal to the error matrix $U$, in the sense that, if $\mathcal{P}$ denotes the projection operator onto $\mathcal{X}$, then $\mathcal{P}U = 0$ (At the population level, if $\mathcal{X}$ is the space spanned by a random vector $X$, then $\mathcal{P} = E(\cdot|X)$ has the property that $\mathcal{P}U = 0$. Later on, we will apply a sampling smoothing operator $\mathcal{P}$ to the noise vector $U$, so that $\mathcal{P}U \approx 0$). Applying $\mathcal{P}$ to $Y$, we have $\mathcal{P}Y = \mathcal{P}\Lambda F'$. Suppose the following normalization condition holds:

$$ \frac{1}{T}F'F = I_k, \quad \Lambda'\mathcal{P}\Lambda \text{ is a diagonal matrix with distinct entries}. \quad (2.2) $$

The covariance matrix of the projected data matrix is

$$ \frac{1}{T}Y'\mathcal{P}Y = \frac{1}{T}F\Lambda'\mathcal{P}\Lambda F'. $$

By the normalization condition, $\frac{1}{T}Y'\mathcal{P}Y = \frac{1}{T}F\Lambda'\mathcal{P}\Lambda F' = F\Lambda'\mathcal{P}\Lambda$, so the first $K$ eigenvectors of $\frac{1}{T}Y'\mathcal{P}Y$ are actually the columns of $F$. Therefore, the Projected-PCA is to use the first $K$ principal components of $\frac{1}{T}Y'\mathcal{P}Y$ to estimate $F$. In addition, since

$$ \frac{1}{T}\mathcal{P}YF = \frac{1}{T}\mathcal{P}\Lambda F' = \mathcal{P}\Lambda, $$

the projected loading matrix $\mathcal{P}\Lambda$ can also be recovered from the projected data $\mathcal{P}Y$. If in addition, $\Lambda = \mathcal{P}\Lambda$, that is, the loading matrix belongs to the space $\mathcal{X}$, then $\Lambda$ can also be recovered from the projected data. In other words, the effects of the idiosyncratic error term $U$ and residual loadings $\Gamma$ can be completely removed by the projection step.
2.2 Semiparametric Factor Model

As one of the useful examples of forming the space $\mathcal{X}$ and the projection operator, this paper considers model (1.4), where $X_i$'s and $y_t$'s are the only observable data, and \{g_k(\cdot)\}_{k=1}^{K} are unknown nonparametric functions. The specific case (1.2) (with $\gamma_{ik} = 0$) was used extensively in the financial studies by Connor and Linton (2007), Connor et al. (2012) and Park et al. (2009), with $X_i$'s being the observed “market characteristic variables”. We shall assume $K$ to be known for now. A more practical situation arises with unknown number of factors, so we will propose a projected-eigenvalue-ratio method to consistently estimate $K$ later.

To non-parametrically estimate $g_k(X_i)$ without the curse of dimensionality from the multi-dimensionality of $X_i$, we assume $g_k(\cdot)$ to be additive: for each $k \leq K, i \leq p$, there are $(g_{k1}, \cdots, g_{kd})$ nonparametric functions such that

$$g_k(X_i) = \sum_{l=1}^{d} g_{kl}(X_{il}), \quad d = \dim(X_i). \quad (2.3)$$

Each additive component of $g_k$ is estimated by the sieve method. Let \{$\phi_1(x), \phi_2(x), \cdots$\} be a set of basis functions (e.g., B-spline, Fourier series, wavelets, polynomial series), which spans a dense linear space of the functional space for \{g_{kl}\}. Then for each $l \leq d$,

$$g_{kl}(X_{il}) = \sum_{j=1}^{J} b_{j,kl} \phi_j(X_{il}) + R_{kl}(X_{il}), \quad k \leq K, i \leq p, l \leq d. \quad (2.4)$$

Here \{b_{j,kl}\}_{j=1}^{J} are the sieve coefficients of the $l$th additive component of $g_k(X_i)$, corresponding to the $k$th factor loading; $R_{kl}$ is a “remaining function”; $J$ denotes the number of sieve terms which grows slowly as $p \to \infty$. The basic assumption for sieve approximation is that $\sup_{x} |R_{kl}(x)| \to 0$ as $J \to \infty$. We here take the same basis functions in (2.4) purely for simplicity of notation.

Define, for each $k \leq K$ and for each $i \leq p$,

$$b_k' = (b_{1,k1}, \cdots, b_{J,k1}, \cdots, b_{1,kd}, \cdots, b_{J,kd}) \in \mathbb{R}^{Jd},$$

$$\phi(X_i)' = (\phi_1(X_{i1}), \cdots, \phi_J(X_{i1}), \cdots, \phi_1(X_{id}), \cdots, \phi_J(X_{id})) \in \mathbb{R}^{Jd}. $$

Then, we can write

$$g_k(X_i) = \phi(X_i)'b_k + \sum_{l=1}^{d} R_{kl}(X_{il}),$$

where $R_{kl}$ is the approximation error. Let $B = (b_1, \cdots, b_K)$ be a $(Jd) \times K$ matrix of sieve
coefficients, $\Phi(X) = (\phi(X_1), \cdots, \phi(X_p))'$ be a $p \times (Jd)$ matrix of basis functions, and $R(X)$ be $p \times K$ matrix with the $(i,k)$th element $\sum_{l=1}^{d} R_{kl}(X_{il})$. Then the matrix form of (2.3) and (2.4) is

$$G(X) = \Phi(X)B + R(X).$$

Substituting this into (1.5), we write

$$Y = (\Phi(X)B + \Gamma)F' + R(X)F' + U.$$

We see that the residual term consists of two parts: the sieve approximation error $R(X)F'$ and the idiosyncratic error $U$. Furthermore, the random effect assumption on the coefficients $\Gamma$ makes it also behave like noise when smoothing operators are applied.

### 2.3 The estimator

Using the idea described in Section 2.1, we propose a Projected-PCA method, where $X$ is the sieve space spanned by the basis functions of $X$, and $P$ is chosen as the projection matrix onto $X$, defined by the $p \times p$ projection matrix

$$P = \Phi(X)(\Phi(X)'\Phi(X))^{-1}\Phi(X)'.$$

The estimators of the model parameters in (1.5) are defined as follows. The columns of $F/\sqrt{T}$ are defined as the eigenvectors corresponding to the first $K$ largest eigenvalues of the $T \times T$ matrix $Y'PY$, and

$$\widehat{G}(X) = \frac{1}{T}PY\widehat{F}. \quad (2.5)$$

is the estimator of $G(X)$.

The weight matrix $P$ projects the original data matrix onto the sieve space spanned by $X$. Therefore, unlike the traditional PCA method for usual factor models (e.g., Bai (2003), Stock and Watson (2002)), the projected-PC takes the principal components of the projected data $PY$. The estimator is thus invariant to the rotation-transformations of the sieve bases. The intuition can be readily seen from the discussions in Section 2.1. The projection of $Y$ onto the space spanned by $X$ is approximately $G(X)F'$, since $X$ are approximately orthogonal to $U$ and $\Gamma$. Therefore, the projected-PCs on the projected data $PY$ are essentially equivalent to taking the principal components of $\frac{1}{T}FG(X)'PG(X)F'$, which is approximately $\frac{1}{T}FG(X)'G(X)F'$ because in this case the loading matrix belongs to the sieve space spanned by $X$. Hence the effect of the error term $U$ and $\Gamma$ is naturally “projected off”, which enables us to achieve a more accurate estimation.
The estimation of the loading component $\Gamma$ that can not be explained by the covariates can be estimated as follows. With the estimated factors $\hat{F}$, the least-squares estimator of loading matrix is $\hat{\Lambda} = Y\hat{F}/T$, by using (2.1) and (2.2). Therefore, by (1.5), a natural estimator of $\Gamma$ is

$$\hat{\Gamma} = \hat{\Lambda} - \hat{G}(X) = \frac{1}{T}(I - P)Y\hat{F}. \quad (2.6)$$

We will show that the projected-PC gives a faster rate of convergence than the traditional PC under some mild conditions that require the projection be genuine. As a result, $\hat{\Lambda}$ is estimated faster than the conventional PC and hence $\hat{\Gamma}$ can also be estimated more precisely, as the errors in estimating the smooth function $G$ are negligible.

### 2.4 Connection with panel data models with time-varying coefficients

Consider a panel data model with time-varying coefficients as follows:

$$y_{it} = X_i'\beta_t + \mu_t + u_{it}, \quad i \leq N, t \leq T, \quad (2.7)$$

where $X_i$ is a $d$-dimensional vector of time-invariant regressors for individual $i$; $\mu_t$ denotes the unobservable random time effect; $u_{it}$ is the regression error term. The regression coefficient $\beta_t$ is also assumed to be random and time-varying, but is common across the cross-sectional individuals.

The semi-parametric factor model includes (2.7) as a special case. Note that (2.7) can be rewritten as $y_{it} = g(X_i)'f_t + u_{it}$ with $K = d + 1$ unobservable “factors” $f_t = (\mu_t, \beta_t)'$ and “loading” $g(X_i) = (1, X_i)'$. The model (1.4) being considered, on the other hand, allows more general nonparametric loading functions.

### 3 Projected-PCA in Conventional Factor Models

Let us first consider the asymptotic performance of the projected-PCA in the conventional factor model:

$$Y = \Lambda F' + U. \quad (3.1)$$

Assume that the projection is genuine so that the covariates explain genuinely a part of the loading matrix $\Lambda$. More precisely, $\Lambda'PA\Lambda$ can be normalized to a diagonal matrix with spiked distinct entries. We now show that latent factors $F$ can be estimated at a faster rate of convergence by Projected-PCA than the conventional PCA and that they can be consistently estimated even when sample size $T$ is finite.
Example 3.1. To appreciate the intuition, let us consider a specific case in which $K = 1$ so that model (1.4) reduces to

$$y_{it} = g(X_i)f_t + \gamma_i f_t + u_{it}.$$ 

Assume that $g(\cdot)$ is so smooth that it is in fact a constant $\beta$ (otherwise, we can use a local constant approximation). Then, the model reduces further to

$$y_{it} = \beta f_t + \gamma_i f_t + u_{it}.$$ 

The projection in this case is the averaging over $i$, which yields

$$\bar{y}_t = \beta f_t + \bar{\gamma}_f + \bar{u}_t,$$

where $\bar{y}_t$, $\bar{\gamma}_t$, and $\bar{u}_t$ denote the averages of their corresponding quantities over $i$. For identification purpose, suppose $E \gamma_i = Eu_{it} = 0$, and $\sum_{t=1}^{T} f_t^2 = T$. Ignoring the last two terms, we obtain estimators

$$\hat{\beta} = \left( \frac{1}{T} \sum_{t=1}^{T} \bar{y}_t^2 \right)^{1/2}, \quad \text{and} \quad \hat{f}_t = \bar{y}_t / \hat{\beta}.$$ 

Since the ignored two terms are of order $O_P(p^{-1/2})$, $\hat{\beta}$ and $\hat{f}_t$ converge whether or not $T$ is large. Taking the average over $t$ into account, a more careful heuristic argument shows that the rate of convergence for $\beta$ should be $O_P((Tp)^{-1/2})$.

3.1 Asymptotic Properties of Projected-PCA

We now state the conditions and results more formally in the more general factor model (3.1). Let us begin with some assumptions. Recall that the projection matrix is defined as

$$P = \Phi(X)(\Phi(X)\Phi(X))^{-1}\Phi(X)'.$$ 

Note that $F$ and $\Lambda$ are not separately identified, because for any nonsingular $H$, $\Lambda F' = \Lambda H^{-1}HF'$. Therefore, for the identification purpose, we assume:

Assumption 3.1 (Identification). The true factor and loading matrices satisfy, almost surely:

(i) $T^{-1}F'F = I_K$, and $\Lambda'P\Lambda$ is a $K \times K$ diagonal matrix with distinct entries.

(ii) There are positive constants $c_{\min}$ and $c_{\max}$ such that

$$c_{\min} < \lambda_{\min}(p^{-1}\Lambda'P\Lambda) < \lambda_{\max}(p^{-1}\Lambda'P\Lambda) < c_{\max}.$$
This condition guarantees that the factors and loadings are separately identified. Condition (i) means that columns of factors and loadings can be orthogonalized. It corresponds to the PC1 condition of Bai and Ng (2013), and is often used in the classical factor analysis for identification. Condition (ii) is often known as the pervasive condition on the factor loadings (Stock and Watson (2002)). In our context, this condition requires the covariates $X$ have non-vanishing explaining power on the loading matrix, so that the projection matrix $\Lambda'P\Lambda$ has spiked eigenvalues.

**Assumption 3.2** (Basis function). There are $d_{\text{min}}$ and $d_{\text{max}} > 0$ so that almost surely,

$$d_{\text{min}} < \lambda_{\text{min}}(p^{-1}\Phi(X)'\Phi(X)) < \lambda_{\text{max}}(p^{-1}\Phi(X)'\Phi(X)) < d_{\text{max}}.$$ 

Note that $p^{-1}\Phi(X)'\Phi(X) = p^{-1}\sum_{i=1}^{p}\phi(X_i)'\phi(X_i)$ and $\phi(X_i)$ is a vector of dimensionality $Jd \ll p$. Thus, Assumption 3.2 follows from a strong law of large numbers, given that in the population level, $E\phi(X_i)'\phi(X_i)$ is well-conditioned. This condition can be satisfied through proper normalizations of commonly used basis functions such as B-splines, wavelets, Fourier basis, etc.

We impose the strong mixing condition. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the $\sigma$-algebras generated by $\{(f_t,u_t): t \leq 0\}$ and $\{(f_t,u_t): t \geq T\}$ respectively. Define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |P(A)P(B) - P(AB)|. \quad (3.2)$$

**Assumption 3.3** (Data generating process). (i) $\{u_t, f_t\}_{t \geq 1}$ is strictly stationary, and $X_1, \cdots, X_p$ are independent and identically distributed. In addition, $Eu_{it} = 0$ for all $i \leq p, j \leq K$; $\{u_t\}_{t \leq T}$ is independent of $\{X_i, f_i\}_{i \leq p, t \leq T}$.

(ii) **Strong mixing**: There exist $r_1, C_1 > 0$ such that for all $T > 0$,

$$\alpha(T) < \exp(-C_1 T^{-r_1}).$$

(iii) **Weak dependence**: There is $C_2 > 0$ so that

$$\max_{j \leq p} \sum_{i=1}^{p} |E u_{it} u_{jt}| < C_2, \quad \frac{1}{pT} \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{t=1}^{T} \sum_{s=1}^{T} |E u_{it} u_{js}| < C_2,$$

$$\max_{i \leq p} \frac{1}{pT} \sum_{k=1}^{p} \sum_{m=1}^{p} \sum_{t=1}^{T} \sum_{s=1}^{T} |\text{cov}(u_{it} u_{kt}, u_{is} u_{ms})| < C_2.$$ 

(v) **Exponential tail**: There exist $r_2, r_3 > 0$ satisfying $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$ and $b_1, b_2 > 0$, such
that for any $s > 0$, $i \leq p$ and $j \leq K$,

$$P(|u_{it}| > s) \leq \exp(-(s/b_1)^2), \quad P(|f_{jt}| > s) \leq \exp(-(s/b_2)^3).$$

Assumption [3.3] is standard, especially condition (iii) is commonly imposed for high-dimensional factor analysis, which requires $\{u_{it}\}_{i=1,t=2}^p$ be weakly dependent both serially and cross-sectionally. It is often satisfied when the covariance matrix $E(u_iu_i')$ is sufficiently sparse under the strong mixing condition.

Formally, we have the following theorem:

**Theorem 3.1.** Consider the conventional factor model (3.1) with Assumptions [3.1-3.3]. The Projected-PCA estimators $\hat{F}$ and $\hat{G}(X)$ defined in Section [2.3] satisfy, as $p \to \infty$ ($J,T$ may either grow simultaneously with $p$ satisfying $J = o(\sqrt{p})$ or stay constant),

$$\frac{1}{T} \|\hat{F} - F\|_F^2 = O_p\left(\frac{J}{p}\right),$$

$$\frac{1}{p} \|\hat{G}(X) - P\Lambda\|_F^2 = O_p\left(\frac{J}{pT} + \frac{J^2}{p^2}\right).$$

To compare with the regular PC method, note that the convergence rate for the estimated factors is improved for small $T$. In particular, the projected-PC does not require $T \to \infty$, and also has a good rate of convergence for the loading matrix up to a projection transformation. In contrast, the conventional PC method achieves a rate of convergence to be $O_p(1/p + 1/T^2)$ for factors, and $O_p(1/T + 1/p)$ for loadings. See Remarks 4.1, 4.2 below for additional details.

## 4 Projected-PCA in Semi-parametric Factor Models

### 4.1 Sieve approximations

In the semi-parametric factor model, it is assumed that $\lambda_{ik} = g_k(X_i) + \gamma_{ik}$, where $g_k(X_i)$ is a nonparametric smooth function for the observed covariates, and $\gamma_{ik}$ is the unobserved random loading component that is independent of $X_i$. Hence the model is written as

$$y_{it} = \sum_{k=1}^K \{g_k(X_i) + \gamma_{ik}\}f_{kt} + u_{it}, \quad i = 1, \cdots, p, \quad t = 1, \cdots, T. \quad (4.1)$$

In the matrix form,

$$Y = \{G(X) + \Gamma\}F' + U. \quad (4.2)$$
Suppose that $G(X)$ does not vanish (pervasive condition, see Assumption 4.1 below). Because $X$ has non-vanishing explaining power of the loading matrix, all the unknown quantities can be estimated more accurately by the PCA on the projected data matrix $Y'PY$. In particular, the factors and nonparametric loading functions $\{g_k(\cdot)\}$ can be consistently estimated with a finite $T$.

The estimators $\hat{F}$ and $\hat{G}(X)$ are the projected-PC estimators as defined in Section 2.3. We now define the estimator of the nonparametric function $g_k(\cdot)$, $k = 1, \ldots, K$. We assume $g_k(\cdot) \in G$ to be additive, and estimate it using sieve approximation: in the matrix form, the projected data has the following sieve approximated representation:

$$PY = \Phi(X)BF' + E,$$  \hspace{1cm} (4.3)

where $E = P\Gamma F' + PR(X)F' + PU$ is “small” because $\Gamma$ and $U$ are orthogonal to the function space spanned by $X$, and $R(X)$ is the sieve approximation error. The sieve coefficient matrix $B = (b_1, \ldots, b_K)$ can be estimated by least squares from the projected model (4.3): Ignore $E$, replace $F$ with $\hat{F}$, and solve (4.3) to obtain

$$\hat{B} = (\hat{b}_1, \ldots, \hat{b}_K) = \frac{1}{T} [\Phi(X)'\Phi(X)]^{-1}\Phi(X)'Y\hat{F}.$$  

We then estimate $g_k(\cdot)$ by

$$\hat{g}_k(x) = \phi(x)'\hat{b}_k, \hspace{0.5cm} \forall x \in \mathcal{X}, \hspace{0.5cm} k = 1, \ldots, K,$$

where $\mathcal{X}$ denotes the support of $X_i$. Also note that $\hat{G}(X) = PY\hat{F}/T = \Phi(X)\hat{B}$.

### 4.2 Asymptotic analysis

Note that even with the condition $E(T|X) = 0$, we still have the identifiability issue for any $K \times K$ invertible matrix $H$, $(G(X) + \Gamma)F' = (G(X) + \Gamma)H^{-1}HF'$. As a result, we assume the following:

**Assumption 4.1 (Identification).** (i) $\{\gamma_i, X_i\}$ are independent and identically distributed; $E\gamma_{ik} = 0$ and $\{X_i\}_{i \geq 1}$ is independent of $\{\gamma_{ik}\}_{i \leq 1}$.

(ii) $T^{-1}FF' = I_K$, and $G(X)'G(X)$ is a $K \times K$ diagonal matrix with distinct entries.

(iii) There are two positive constants $c_{\min}$ and $c_{\max}$ so that

$$c_{\min} < \lambda_{\min}(p^{-1}G(X)'G(X)) < \lambda_{\max}(p^{-1}G(X)'G(X)) < c_{\max}.$$
In particular, Condition (iii) guarantees that the common factors are \textit{strong}, and can be estimated $\sqrt{p}$-consistently. This condition can be weakened so that the eigenvalues of $p^{-\alpha} \mathbf{G}(\mathbf{X})' \mathbf{G}(\mathbf{X})$ are bounded for some $\alpha \in (0, 1]$. This then allows for \textit{semi-weak} factors (see detailed discussions in Section 6).

The following set of conditions is concerned about the accuracy of the sieve approximation.

**Assumption 4.2 (Accuracy of sieve approximation).** For each $l \leq d, k \leq K$,

(i) the loading component $g_{kl}(\cdot)$ belongs to a H"older class $\mathcal{G}$ defined by

$$\mathcal{G} = \{ g : |g^{(r)}(s) - g^{(r)}(t)| \leq L|s - t|^\alpha \}$$

for some $L > 0$; (ii) the sieve coefficients $\{ b_{k,jl} \}_{j=1}^J$ satisfy: for $\kappa = 2(r + \alpha) \geq 4$, as $J \to \infty$,

$$\sup_{x \in \mathcal{X}_l} |g_{kl}(x) - \sum_{j=1}^J b_{k,jl} \phi_j(x)|^2 = O(J^{-\kappa}),$$

where $\mathcal{X}_l$ is the support of the $l$th element of $\mathbf{X}_l$, and $J$ is the sieve dimension.

(iii) The eigenvalues of the $K \times K$ matrix $\mathbf{B}' \mathbf{B}$ are bounded away from zero.

Condition (ii) is satisfied by common basis. For example, when $\{ \phi_j \}$ is polynomial basis or B-splines, condition (ii) is implied by condition (i) (see e.g., Lorentz (1986) and Chen (2007)).

Define

$$\nu_p = \max_{k \leq K} \frac{1}{p} \sum_{i=1}^p \text{var}(\gamma_{ik}).$$

**Theorem 4.1.** Suppose $J = O(\sqrt{p})$. Under Assumptions 3.2, 3.3, 4.1, 4.2, as $p, J \to \infty$, $T$ can be either divergent or bounded and note that $\kappa \geq 4$,

$$\frac{1}{T} \| \hat{\mathbf{F}} - \mathbf{F} \|_F^2 = O_P \left( \frac{1}{p} + \frac{1}{J^{-\kappa}} \right),$$

$$\frac{1}{p} \| \hat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X}) \|_F^2 = O_P \left( \frac{J}{p^2} + \frac{J}{pT} + \frac{J}{J^\kappa} + \frac{J \nu_p}{p} \right),$$

$$\max_{k \leq K} \sup_{x \in \mathcal{X}} |\hat{g}_k(x) - g_k(x)| = O_P \left( \frac{J}{p} + \frac{J}{\sqrt{pT}} + \frac{J}{J^{\kappa/2}} + \sqrt{\frac{\nu_p}{p}} \right) \max_{j \leq T} \sup_x |\phi_j(x)|$$

In addition, if $T \to \infty$ simultaneously with $p$ and $J$, then

$$\frac{1}{p} \| \hat{\mathbf{G}} - \mathbf{G} \|^2 = O_P \left( \frac{J}{p^2} + \frac{1}{T} + \frac{1}{J^\kappa} + \frac{J \nu_p}{p} \right).$$
The optimal \( J^* = (p \min\{T, p, \nu^{-1}\})^{1/\kappa} \) simultaneously minimizes the convergence rates of the factors and nonparametric loading function \( g_k(\cdot) \). It also satisfies the constraint \( J^* = O(\sqrt{p}) \) as \( \kappa \geq 4 \). With \( J = J^* \),

\[
\frac{1}{T} \sum_{t=1}^{T} \| \tilde{f}_t - f_t \|^2 = O_P \left( \frac{1}{p} \right),
\]

\[
\frac{1}{p} \sum_{i=1}^{p} \| \tilde{g}_k(X_i) - g_k(X_i) \|^2 = O_P \left( \frac{1}{(p \min\{T, p, \nu^{-1}\})^{1-1/\kappa}} \right), \quad \forall k
\]

\[
\max_{k \leq K} \sup_{x \in \mathcal{X}} |\tilde{g}_k(x) - g_k(x)| = O_P \left( \frac{1}{(p \min\{T, p, \nu^{-1}\})^{1/2-1/\kappa}} \right)
\]

and

\[
\frac{1}{p} \sum_{i=1}^{p} \| \tilde{\gamma}_i - \gamma_i \|^2 = O_P \left( \frac{1}{(p \min\{T, p, \nu^{-1}\})^{1-1/\kappa} + \frac{1}{T}} \right).
\]

Some remarks about these rates of convergence compared with those of the conventional factor analysis are in order.

**Remark 4.1.** The rates of convergence for factors and nonparametric functions do not require \( T \to \infty \). Hence consistency can still be achieved even if \( T \) is bounded. When \( T = O(1) \),

\[
\frac{1}{T} \sum_{t=1}^{T} \| \tilde{f}_t - f_t \|^2 = O_P \left( \frac{1}{p} \right), \quad \frac{1}{p} \sum_{i=1}^{p} \| \tilde{g}_k(X_i) - g_k(X_i) \|^2 = O_P \left( \frac{1}{p^{1-1/\kappa}} \right).
\]

The rates still converge fast when \( p \) is large, demonstrating the blessing of dimensionality. This is an attractive feature of the projected-PCA in high-dimension-low-sample-size situations. In particular, for many practical applications, the stationarity of a time series holds only for a short period of time. In contrast, in the usual factor analysis, consistency is granted only when \( T \to \infty \). For example, according to Stock and Watson (2002), Fan et al. (2014), the regular PC method has the following convergence rate

\[
\frac{1}{T} \sum_{t=1}^{T} \| \tilde{f}_t - f_t \|^2 = O_P(\frac{1}{p} + \frac{1}{T^2}),
\]

which does not go to zero when \( T \) is bounded.

**Remark 4.2.** As noted in the previous remark, the accuracy of estimated factors by projected-PCA is no worst than that based on the regular PCA. The loading matrix is
estimated by $\hat{\Lambda} = \hat{G}(X) + \hat{\Gamma}$, which performs no worse than that from the ordinary PCA, even when $T$ is large. When $\gamma_{ik}$ vanishes, $\lambda_{ik} = g_{k}(X_i)$. In this case, we have

$$\frac{1}{p} \sum_{i=1}^{p} |\hat{\lambda}_{ik} - \lambda_{ik}|^2 = \frac{1}{p} \sum_{i=1}^{p} |\hat{g}_{k}(X_i) - g_{k}(X_i)|^2 = O_P\left(\frac{1}{(pT)^{1-1/\kappa}} + \frac{1}{p^{2-2/\kappa}}\right)$$

In contrast, using the regular PC method as in Stock and Watson (2002), we have

$$\frac{1}{p} \sum_{i=1}^{p} |\tilde{\lambda}_{ik} - \lambda_{ik}|^2 = O_P\left(\frac{1}{T} + \frac{1}{p}\right).$$

Comparing the rates of the projected-PC with the regular PC method, we see that besides an improved rate for the estimated factors when $T$ is small, when $g_{k}(\cdot)$’s are sufficiently smooth (larger $\kappa$), the rate of convergence for the estimated loadings is also improved even under a relatively large $T$.

5 Semiparametric Specification Test

The loading matrix always have the following orthogonal decomposition:

$$\Lambda = G(X) + \Gamma,$$

where $\Gamma$ is interpreted as the loading component that cannot be explained by $X$. We consider two types of specification tests: testing $H_0^1 : G(X) = 0$, and $H_0^2 : \Gamma = 0$. The former tests whether the observed covariates have explaining powers on the loadings, while the latter tests whether the covariates fully explain the loadings. The former provides a diagnostic tool as to whether or not to employ the projected PCA. The latter test the adequacy of the semiparametric factor models in the literature.

5.1 Testing $G(X) = 0$

Testing whether the observed covariates have explaining powers on the factor loadings can be formulated as the following null hypothesis: almost surely

$$H_0^1 : G(X) = 0.$$

Applying the projection matrix $P$ to $\Lambda = G(X) + \Gamma$, due to the orthogonality of $X$ and $\Gamma$, approximately we have $P\Lambda \approx G(X)$. Hence, the null hypothesis is approximately equivalent
to

\[ H_0 : P\Lambda = 0, \text{almost surely.} \]

This motivates the test statistic \( \|P\tilde{\Lambda}\|_F^2 = \text{tr}(\tilde{\Lambda}'P\tilde{\Lambda}) \) for a consistent loading estimator \( \tilde{\Lambda} \). Normalizing the test statistic by its asymptotic variance leads to the test statistic

\[ S_G = \frac{1}{p} \text{tr}(W_1\tilde{\Lambda}'P\tilde{\Lambda}), \quad W_1 = (\frac{1}{p}\tilde{\Lambda}'\tilde{\Lambda})^{-1}, \]

where the \( K \times K \) matrix \( W_1 \) is the weight matrix. The null hypothesis is rejected when \( S_G \) is large.

The least squares estimator of \( \Lambda \) based on the estimated factors is \( \tilde{\Lambda} = Y\tilde{F}/T \). Hence the test statistic can be equivalently written as

\[ S_G = \frac{1}{T^2p} \text{tr}(W_1\tilde{F}'Y'PY\tilde{F}). \]

Here \( \tilde{\Phi} \) should be a consistent estimator of realized factors. Under the null hypothesis, \( X \) has no explaining power on the loadings, and the projected-PC estimator is inappropriate. Therefore, we take \( \tilde{\Phi} \) as the regular PC estimator: the columns of \( \tilde{\Phi}/\sqrt{T} \) are the first \( K \) eigenvectors of the \( T \times T \) data matrix \( Y'Y \).

5.2 Testing \( \Gamma = 0 \)

Connor and Linton (2007), Connor et al. (2012) applied the semi-parametric factor model to analyzing financial returns, who assumed that \( \Gamma = 0 \), that is, the loading matrix can be fully explained by the observed covariates. It is therefore natural to question such an assumption and to test the following null hypothesis of specification: almost surely,

\[ H^2_0 : \Gamma = 0. \]

Recall that \( G(X) \approx PA \) so that \( \Lambda \approx PA + \Gamma \). Therefore essentially the specification testing problem is equivalent to testing:

\[ H_0 : PA = \Lambda, \text{almost surely.} \]

That is, we are testing whether the loading matrix in the factor model belongs to the space spanned by the observed covariates.
A natural test statistic is thus based on the weighted quadratic form

\[ \text{tr}(\hat{\Gamma}'W_2\hat{\Gamma}) = \text{tr}((I - P)\hat{\Lambda}'(I - P)\hat{\Lambda}), \]

for some \( p \times p \) positive definite weight matrix \( W_2 \), where \( \hat{\Lambda} = Y\hat{F}/T \) and \( \hat{F} \) is the projected-PC estimator for factors. To control the size of the test, we take \( W_2 = \Sigma_u^{-1} \), where \( \Sigma_u \) is a diagonal covariance matrix of \( u_t \) under \( H_0 \), assuming that \((u_{1t}, \ldots, u_{pt})\) are uncorrelated.

In practice, we replace \( \Sigma_u^{-1} \) with its consistent estimator: let \( \hat{U} = Y - \hat{\Lambda}'\hat{F}' \). Define

\[ \hat{\Sigma}_u = T^{-1}\text{diag}\{\hat{U}\hat{U}'\} = T^{-1}\text{diag}\{Y(I - T^{-1}\hat{F}\hat{F}')Y'\}. \]

Then the operational test statistic is defined to be

\[ S_\Gamma = \text{tr}(\hat{\Lambda}'(I - P)'\hat{\Sigma}_u^{-1}(I - P)\hat{\Lambda}) \]

The null hypothesis is rejected for large values of \( S_\Gamma \).

### 5.3 Asymptotic null distributions

We assume \( T, p, J \to \infty \) simultaneously. The following assumption regulates the relation between \( T \) and \( p \).

**Assumption 5.1.** Suppose (i) \( T^{2/3} = o(p) \), and \( p(\log p)^4 = o(T^2) \).

(ii) \( J \) and \( \kappa \) satisfy: \( J = o(\min\{\sqrt{p}, \sqrt{T}\}) \), and \( \max\{T, \sqrt{p}, p\} = o(J^\kappa) \).

Condition (i) requires a balance of the dimensionality and the sample size. On one hand, a relatively large sample size is desired \( (p(\log p)^4 = o(T^2)) \) so that the effect of estimating \( \Sigma_u^{-1} \) is negligible asymptotically. On the other hand, as is common in high-dimensional factor analysis, a lower bound of the dimensionality is also required (condition \( T^{2/3} = o(p) \)) to ensure that the factors are estimated accurately enough. Such a required balance between the sample size and dimensionality is common for high-dimensional factor analysis (e.g., Bai (2003), Stock and Watson (2002)) and in the recent literature for PCA (e.g., Jung and Marron (2009), Shen et al. (2013)).

We focus on the Gaussian error case. It will be shown that under \( H_0^1 \),

\[ S_G = (1 + o_P(1))\frac{1}{p}\text{tr}(W_1\Gamma'\Pi\Gamma). \]

and \( H_0^2 \)

\[ S_\Gamma = (1 + o_P(1))\frac{1}{T^2}\text{tr}(F'U'\Sigma_u^{-1}UF), \]

18
whose conditional distributions (given $F$) under the null are $\chi^2$ with degree of freedom respectively $JdK$ and $pK$. We can derive their standardized limiting distribution as $J, T, p \to \infty$. This is given in the following result.

**Theorem 5.1.** Suppose Assumptions 3.2, 3.3, 4.1, 4.2 hold. Then under $H_0^1$,

$$\frac{pS_G - JdK}{\sqrt{2JdK}} \to^d N(0, 1),$$

where $K = \dim(f_t)$ and $d = \dim(X_i)$. In addition, suppose Assumption 5.1 holds, $\{u_t\}_{t \leq T}$ is i.i.d. $N(0, \Sigma_u)$ with a diagonal covariance matrix $\Sigma_u$ whose elements are bounded away from zero and infinity. Then under $H_0^2$,

$$\frac{T \Gamma - pK}{\sqrt{2pK}} \to^d N(0, 1).$$

In practical applications, when a relatively small sieve dimension $J$ is used, one can use the upper $\alpha$-quantile of the $\chi^2_{JdK}$ distribution for $pS_G$, which is expected to be more accurate.

**Remark 5.1.** We require $u_{it}$ be independent across $t$, which ensures that the covariance matrix of the leading term $\text{vec}(\frac{1}{\sqrt{T}}UF')$ to have a simple form $\Sigma_u^{-1} \otimes I_K$. This assumption can be relaxed to allow for weakly dependent time series errors, but many autocovariance terms will also be involved in the covariance matrix. In that case, one may regularize the standard autocovariance matrix estimators such as Newey and West (1987) and optimal kernel (Andrews (1991)) to account for the high dimensionality. Moreover, we assume $\Sigma_u$ be diagonal to facilitate estimating $\Sigma_u^{-1}$. This assumption can also be weakened to sparsity on $\Sigma_u$. Regularization methods such as thresholding (Bickel and Levina (2008)) can then be employed.

6 Estimating the Number of Factors with Semi-weak Factors

We now address the problem of estimating $K$ when it is unknown. Once consistent estimation of $K$ is obtained, all the results we have achieved carry over to the unknown $K$ case using a conditioning argument. In principle many consistent estimators of $K$ can be employed, e.g., Bai and Ng (2002), Alessi et al. (2010), Breitung and Pigorsch (2009), Hallin and Liška (2007). More recently, Ahn and Horenstein (2013) and Lam and Yao (2012) proposed a tuning parameter-free estimator, which selects the largest ratio of the adjacent
eigenvalues of $Y'Y$, based on the fact that the $K$ largest eigenvalues of the sample covariance matrix grow unboundedly as $p$ increases, while the remaining eigenvalues remain bounded.

However, as discussed before, when the loadings depend on the observable characteristics, it is more desirable to apply the eigenvalue-ratio procedure on the projected data $PY$. Due to the orthogonality condition of $U$ and $X$, the projected data matrix is approximately equal to $G(X)F'$. The projected matrix $PY(PY)'$ allows us to study the eigenvalues of the principal matrix component $G(X)G(X)'$, which directly connects with the strengths of those factors. Since the non-vanishing eigenvalues of $PY(PY)'$ and $(PY)PY = Y'PY$ are the same, we can work directly with the eigenvalues of the matrix $Y'PY$.

Let $\lambda_k(Y'PY)$ denote the $k$th largest eigenvalue of the projected data matrix $Y'PY$. Let $\lfloor\min\{p,T\}/2\rfloor$ denote the nearest integer of $\min\{p,T\}/2$. The estimator of $K = \dim(f_i)$ is defined as:

$$\hat{K} = \arg\max_{1 \leq k \leq \lfloor\min\{p,T\}/2\rfloor} \frac{\lambda_k(Y'PY)}{\lambda_{k+1}(Y'PY)}.$$ 

We extend Ahn and Horenstein (2013)’s theory in two ways, which broaden the applicability of the eigenvalue-ratio method for determining the number of factors. On one hand, we allow the presence of “semi-weak” factors, that is, there is $\alpha \in (0,1]$ so that almost surely, the eigenvalues of the $K \times K$ matrix

$$p^{-\alpha}G(X)'G(X)$$

are bounded away from both zero and infinity as $p \to \infty$. When $\alpha = 1$, it becomes the “pervasive condition” in Assumption 4.1 which has been frequently assumed in the existing literature (e.g., Ahn and Horenstein (2013), Fan et al. (2013)). When $\alpha < 1$, it reduces to the semi-weak factors (Chudik and Pesaran (2013)). Therefore we impose a weaker condition on the strength of the factors.

On the other hand, we allow $p/T \to \infty$. Hence the dimensionality can be much larger than the sample size. Specifically, the relationship among $(T, p, \alpha)$ is given by the following assumption.

**Assumption 6.1.** Suppose there is $\alpha \in (0,1]$ so that

(i) there is $C_1, C_2 > 0$, almost surely,

$$C_1 < \lambda_{\min}(p^{-\alpha}G(X)'G(X)) \leq \lambda_{\max}(p^{-\alpha}G(X)'G(X)) < C_2$$

(ii) as $T, p \to \infty$, $p^{1-\alpha} = o(T)$.

Condition (ii) clarifies the relationship between the strength of the factor and the required sample size. Given $p$, weaker factors (smaller $\alpha$) requires a relatively larger sample size. Note
that when \( p = O(T) \), it holds for any \( \alpha > 0 \). As shown in Johnston and Lu (2009), in the high-dimensional setting, the spectrum of the sample covariance matrix is not consistent. As a result, a relatively larger sample size is required here to reduce the estimation error of the eigenvalue-ratios. Moreover, in the case of \( \alpha = 1 \), the only requirement is \( p, T \to \infty \), and no specific relationship between \( p \) and \( T \) is required.

Note that we do require \( \alpha > 0 \) so that the factors are not too weak. As is shown in Onatski (2012), when \( \alpha = 0 \), the principal components estimator of the factors is inconsistent, because the strength of factors does not dominate the idiosyncratic errors.

In the assumption below, recall that \( U = (u_1, \ldots, u_T) \) is a \( p \times T \) matrix of the idiosyncratic errors, and \( \Sigma_u = E u_t u'_t \) denote the \( p \times p \) covariance matrix of \( u_t \).

**Assumption 6.2.** The error matrix \( U \) can be decomposed as

\[
U = \Sigma_u^{1/2} EM^{1/2},
\]

where,

(i) The eigenvalues of \( \Sigma_u \) are bounded away from both zero and infinity.

(ii) \( M \) is a \( T \) by \( T \) positive semi-definite non-stochastic matrix, whose eigenvalues are bounded away from 0 and infinity.

(iii) \( E = (e_{it})_{p \times T} \) is a \( p \times T \) stochastic matrix, where \( \{e_{it}\}_{t \leq p, t \leq T} \) are i.i.d. random variables, and \( E e_{it}^4 < \infty \).

This assumption allows the error matrix \( U \) to be both cross-sectionally and serially dependent. In particular, the \( T \times T \) matrix \( M \) captures the serial dependence across \( t \). In the special case of no-serial-dependence, the decomposition (6.1) is satisfied by taking \( M = I \).

The following theorem is the main result of this section.

**Theorem 6.1.** Suppose \( J = o(p^\alpha) \) and assume the identifiability condition \( \frac{1}{T} \sum_{t=1}^{T} f_t f'_t = I_K \). Under assumptions 3.2, 3.3, 6.1, 6.2 as \( p, T \to \infty \),

\[
P(\hat{K} = K) \to 1.
\]

### 7 Numerical Studies

This section presents numerical results to demonstrate the performance of projected-PC method for estimating loading and factors using both real data and simulated data.
7.1 Estimating loading curves with real data

We collected stocks in S&P500 index constituents from CRSP which have complete daily closing prices from year 2005 to 2013, and their corresponding market capitalization and book value from Compustat. Stocks that have extreme values on the characteristics were removed. Hence there are 337 stocks in our data set, whose daily excess returns were calculated. We considered four characteristics $X$ as in Connor et al. (2012) for each stock: size, value, momentum and volatility. They were all calculated using the data before a certain data analyzing window so that characteristics are assumed to be given. Size effect is measured by logarithm of the market capitalization (in millions) on the day before the data analyzing window. Value of a firm is the ratio of the market value of equity to the book value of equity in the previous year. Momentum is calculated as cumulative half-year return of the previous 126 days and volatility is the standard deviation of the daily returns of the previous 126 days. All four characteristics are standardized to have mean zero and unit variance. Note that the construction of the four characteristics makes their values independent of the current data.

We fix the time window to be the first quarter of the year 2006, which contains $T = 63$ observations. Given the excess returns $\{y_{it}\}_{t=1}^{337}$ and characteristics $X_i$ as the input data and setting $K = 3$, we fit loading functions $g_k(X_i) = \alpha_{ik} + \sum_{l=1}^{4} g_{kl}(X_{il})$ for $k = 1, 2, 3$ using the projected-PC method. The four additive components $g_{kl}(\cdot)$ are fitted using the cubic spline in the R package “GAM” with sieve dimension $J = 4$. All the four loading functions for each factor are plotted in Figure 1. The contribution of each characteristic to each factor is quite nonlinear.

7.2 Calibrating the model with real data

We now treat the estimated functions $g_{kl}(\cdot)$ as the true loading functions, and calibrate a model for simulations. The “true model” is calibrated as follows:

1. Take the estimated $g_{kl}(\cdot)$ from the real data as the true loading functions.

2. For each $p$, generate $\{u_t\}_{t=1}^{T}$ from $N(0, D\Sigma_0 D)$ where $D$ is diagonal and $\Sigma_0$ sparse. Generate the diagonal elements of $D$ from Gamma($\alpha, \beta$) with $\alpha = 7.06$, $\beta = 536.93$ (calibrated from the real data), and generate the off-diagonal elements of $\Sigma_0$ from $N(\mu_u, \sigma_u^2)$ with $\mu_u = -0.0019$, $\sigma_u = 0.1499$. Then truncate $\Sigma_0$ by a threshold of correlation 0.03 to produce a sparse matrix and make it positive definite by R package “nearPD”.

3. Generate $\{\gamma_{ik}\}$ from iid Gaussian distribution with mean 0 and standard deviation 0.0027, which is calibrated with real data.
4. Generate $f_t$ from a stationary VAR model $f_t = A f_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0, \Sigma_\epsilon)$. The model parameters are calibrated with the market data and listed in Table 1.

| $\Sigma_\epsilon$ | $A$       |
|------------------|-----------|
| 0.9076 0.0049 0.0230 | -0.0371 -0.1226 -0.1130 |
| 0.0049 0.8737 0.0403 | -0.2339 0.1060 -0.2793 |
| 0.0230 0.0403 0.9266 | 0.2803 0.0755 -0.0529 |

Table 1: Calibration for factor generating process.

5. Finally, generate $X_t \sim N(0, \Sigma_X)$. Here $\Sigma_X$ is a $4 \times 4$ correlation matrix estimated from the real data.

Figure 1: Estimated additive loading functions $g_{kl}$, $l = 1, \cdots, 4$, from financial returns of 337 stocks in S&P 500 index. They are taken as the true functions in the simulation studies. In each panel (fixed $l$), the true and estimated curves for $k = 1, 2, 3$ are plotted and compared. The solid, dashed and dotted red curves are the true curves corresponding to the first, second and third factors respectively. The blue curves are their estimates from one simulation of the calibrated model with $T = 50, p = 300$.

We simulate the data from the calibrated model, and estimate the loadings and factors for $T = 10$ and $50$ with $p$ varying from 20 through 500. The “true” and estimated loading
curves are plotted in Figure 1 to demonstrate the performance of projected-PC. Note that the “true” loading curves in the simulation are taken from the estimates calibrated using the real data.

The estimates capture the shape of the true curve, though we also notice slight biases at boundaries. But in general, projected-PC fits the model well. We also compare our method with the regular PC method (e.g., Stock and Watson (2002)). The mean values of $\|\hat{G} - G_0\|_{\text{max}}$, $\|\hat{G} - G_0\|_F$, $\|\hat{\Gamma} - \Gamma\|_{\text{max}}$, $\|\hat{\Gamma} - \Gamma\|_F$, $\|\hat{F} - F_0\|_{\text{max}}$ and $\|\hat{F} - F_0\|_F$ are plotted in Figures 2 and 3. In comparison, projected-PC outperforms PC in estimating both factors and loadings including the nonparametric curves $G(X)$ and random noise $\Gamma$. The estimation errors of Projected-PC decrease as the dimension increases, which is consistent with our asymptotic theory.
Figure 3: Average estimation error of factors over 500 repetitions i.e. $\|\hat{F} - F_0\|_{\text{max}}$ and $\|\hat{F} - F_0\|_F$ by Projected-PC (PPC, solid red) and regular PC (dashed blue) for sample size $T = 10$ (up panel) and $T = 50$ (bottom panel).

7.3 Design 2

Consider a different design with only one observed covariate and three factors. The three characteristic functions are $g_1 = x, g_2 = x^2 - 1, g_3 = x^3 - 2x$ with the characteristic $X$ being standard normal. Generate $\{f_t\}_{t=1}^T$ from the stationary VAR(1) model, that is $f_t = Af_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0, I)$. Here for simplicity we considered $\Gamma = 0$.

We simulate the data for $T = 10$ or 50 and various $p$ ranging from 20 to 500. To ensure that the true factor and loading matrices satisfy the identifiability conditions, we calculate a $3 \times 3$ transformation matrix $H$ such that $\frac{1}{p}HF'FH = I_K$, $H^{-1}G'G'H^{-1}$ is diagonal. Let the final true factors and loadings be $F_0 = FH, G_0 = GH^{-1}$. For each $p$, we run the simulation for 50 times.

We estimate the loadings and factors using both projected-PC and PC. For projected-PC, as in our theorem, we choose $J = C(p \min(T, p))^{1/\kappa}$, with $\kappa = 4$ and $C = 3$. To estimate the loading matrix, we also compare with a third method: sieve-least-squares (SLS), assuming the factors are observable. In this case, the loading matrix is estimated by $PYF_0/T$, where
Figure 4: Average estimation error of loadings over 500 repetitions, i.e. \( \| \hat{G} - G_0 \|_{\text{max}} \) and \( \| \hat{G} - G_0 \|_F \). PPC, PC and SLS respectively represent projected-PC, regular PC and sieve least squares with known factors: Design 2

\( \mathbf{F}_0 \) is the true factor matrix of simulated data.

The estimation error measured in max and Frobenius norm for both loadings and factors are reported in Figures 4 and 5. The plots demonstrate the good performance of projected-PC in estimating both loadings and factors. In particular, the projected-PC works well when we encounter small \( T \) but an large \( p \). As a result \( \Gamma \) can also be more accurately estimated. It is also very interesting to compare projected-PC with SLS (Sieve Least-Squares with observed factors) in estimating the loadings, which corresponds to the cases of unobserved and observed factors. As we see from Figure 4, when \( p \) is small, the projected-PC is not as good as SLS. But the two methods behave similarly as \( p \) increases. This further demonstrates the intuition that as the dimension becomes larger, the effects of estimating the unknown factors are negligible. The projected-PCA method significantly outperforms the traditional PCA. Figure 5 shows that the factors are also better estimated by projected-PCA than the traditional one, particularly when \( T \) is small.
7.4 Estimating number of factors

We now demonstrate the effectiveness of estimating $K$ by the projected-PC’s eigenvalue-ratio method on the projected data. We simulated our data in the same way as in Design 2. $T = 10$ or $50$ and took the values of $p$ ranging from $20$ to $500$. This time, our goal is to compare our projected-PC based on the projected data matrix $Y'PY$ to the eigenvalue-ratio test (AH) of Ahn and Horenstein (2013) and Lam and Yao (2012), which works on the original data matrix $Y'Y$.

For each pair of $T, p$, we repeat the simulation for $50$ times and report the mean and standard deviation of the estimated number of factors in Figure 6. The projected-PC outperforms AH after projection, which significantly reduces the impact of idiosyncratic errors. When $T = 50$, we can recover the number of factors almost all the time, especially for large dimensions ($p > 200$). On the other hand, even when $T = 10$, projected-PC still obtains a closer estimated number of factors.
Figure 6: Mean and standard deviation of the estimated number of factors over 50 repetitions. True $K = 3$. PPC and AH respectively represent the methods of projected-PC and Ahn and Horenstein (2013). Left panel: Mean; Right panel: standard deviation

7.5 Loading specification tests with real data

We test the loading specifications on the real data. We used the same data set as in Section 6.1, consisting of excess returns from 2005 through 2013. The tests were conducted based on rolling windows, with the length of windows spanning from 10 days, a month, a quarter, and half a year. For each fixed window-length ($T$), we computed the standardized test statistic of $S_G, S_\Gamma$, and plotted them along the rolling windows respectively in Figure 7. In almost all cases, the number of factors is estimated to be one in various combinations of $(T, p, J)$.

Figure 7 suggests that the semi-parametric factor model is strongly supported by the data. Judging from the upper panel (testing $H_0^1 : G(X) = 0$), we have very strong evidence of the existence of non-vanishing covariare effect, which demonstrates the dependence of the market beta’s on the covariates $X$. In other words, the market beta’s can be explained at least partially by the characteristics of assets. The results also provide the theoretical basis for using projected PCA to get more accurate estimation.

In the bottom panel of Figure 7 (testing $H_0^2 : \Gamma = 0$), we see for a majority of period, the null hypothesis is rejected. In other words, the characteristics of assets can not be fully explained the market beta as intuitively expected, and model (1.2) in the literature is inadequate. However, fully nonparametric loadings could be possible in certain time range mostly before financial crisis. During 2008-2010, the market’s behavior had much more complexities, which causes more rejections of the null hypothesis. The null hypothesis $\Gamma = 0$
is accepted more often since 2012. We also notice that larger $T$ tends to yield larger statistics in both tests, as the evidence against the null hypothesis is stronger with larger $T$. After all, the semi-parametric model being considered provides flexible ways of modeling equity markets and understanding the nonparametric loading curves.

Figure 7: Normalized test statistics $S_T$ from 2006/01/03 to 2012/11/30 for various $T$’s. The dotted lines are $\pm 1.96$ critical values.

8 Conclusions

This paper proposes and studies a high-dimensional semi-parametric factor model with nonparametric loading functions that depend on a few observed covariate variables. This model is motivated by the fact that observed variables can explain partially the factor loadings. We propose a projected principal components method to estimate the unknown factors,
loadings, and number of factors. It is shown that after projecting the response variable onto the sieve space spanned by the covariate variables, the projected-PCA yields a significant improvement on the rates of convergence than the regular methods in conventional factor analysis. In particular, consistency can be achieved without a diverging sample size, as long as the dimensionality grows. This demonstrates that the high dimensionality for semi-parametric factor analysis is actually a blessing, and thus the proposed method is useful in the typical high-dimension-low-sample-size situations. In addition, we also propose new specification tests for the orthogonal decomposition of the loadings, which fills the gap of the testing literature for semi-parametric factor models. Our empirical findings show that firm characteristics can explain partially the factor loadings, which provides theoretical basis for employing projected-PCA methods for more accurate estimation of latent factors and their associated loading matrices. On the other hand, our empirical study also shows that the firm characteristics can not fully explain the factor loadings so that the generalized model (1.4) is more appropriate.

A Proofs for Section 4

The proof of Theorem 3.1 is given in the supplementary material. In Appendix A, we prove Theorem 4.1.

Let $V$ denote the $K \times K$ diagonal matrix consisting of the first $K$ largest eigenvalues of $(pT)^{-1}Y'PY$ in descending order.

A.1 Preliminary analysis

**Lemma A.1.** There are constants $c_1, c_2 > 0$ so that with probability approaching one,

$$c_1 < \lambda_{\min}(p^{-1}B'\Phi(X)'\Phi(X)B) \leq \lambda_{\max}(p^{-1}B'\Phi(X)'\Phi(X)B) < c_2$$

**Proof.** Note that $G(X) = \Phi(X)B + R(X)$ implies $\|\Phi(X)B\|^2_2 \leq 2\|G(X)\|^2_2 + 2\|R(X)\|^2_2$. Since $p^{-1}\|R(X)\|^2_2 = o_p(1)$, we have, with probability approaching one (w.p.a.1), $p^{-1}\|R(X)\|^2_2 < 1$. Furthermore, $2\|G(X)\|^2_2/p \leq 2c_{\max}$ almost surely. Hence with probability approaching one,

$$\|\Phi(X)B\|^2_2/p \leq 2c_{\max} + 1.$$  

In addition, since $\|G(X)'G(X) - B'\Phi(X)'\Phi(X)B\|_2 \leq 2\|R(X)'G(X)\|_2 + \|R(X)\|^2_2$, hence w.p.a.1, $\frac{1}{p}\|G(X)'G(X) - B'\Phi(X)'\Phi(X)B\|_2 \leq c_{\min}/2$. This implies

$$\lambda_{\min}(p^{-1}B'\Phi(X)'\Phi(X)B) \geq \lambda_{\min}(p^{-1}G(X)'G(X)) - c_{\min}/2 \geq c_{\min}/2.$$
Hence the lemma follows with $c_1 = c_{\min}/2$ and $c_2 = 2c_{\max} + 1$.

**Lemma A.2.** $\|V\|_2 = O_P(1)$ and $\|V^{-1}\|_2 = O_P(1)$.

**Proof.** Note that $\|Y\|_F^2 = O_P(pT)$, and by Assumption 3.2 $\|P\|_2 = O_P(1)$. It then follows immediately that $\|V\|_2 = O_P(1)$. To prove $\|V^{-1}\|_2 = O_P(1)$, note that the eigenvalues of $V$ are the same as those of

$$W = \frac{1}{T} \left( \Phi(X)\Phi(X) \right)^{-1/2} \Phi(X)\Phi(X)^T \left( \Phi(X)\Phi(X) \right)^{-1/2}$$

Write $\Lambda = G(X) + \Gamma$. Substituting $Y = \Lambda F' + U$, and $F'F/T = I_K$, we have $W = \sum_{i=1}^4 W_i$, where

$$W_1 = \frac{1}{p} \left( \Phi(X)\Phi(X) \right)^{-1/2} \Phi(X)\Phi(X) \Lambda \Lambda' \Phi(X) \Phi(X)^T \left( \Phi(X)\Phi(X) \right)^{-1/2}$$

$$W_2 = \frac{1}{p} \left( \Phi(X)\Phi(X) \right)^{-1/2} \Phi(X)\Phi(X) \left( \frac{AF'U'}{T} \right) \Phi(X) \Phi(X)^T \left( \Phi(X)\Phi(X) \right)^{-1/2}$$

$$W_3 = \frac{1}{p} \left( \Phi(X)\Phi(X) \right)^{-1/2} \Phi(X)\Phi(X) \frac{UU'}{T} \Phi(X) \Phi(X)^T \left( \Phi(X)\Phi(X) \right)^{-1/2}$$

$$W_4 = \frac{1}{p} \left( \Phi(X)\Phi(X) \right)^{-1/2} \Phi(X)\Phi(X)^T \left( \Phi(X)\Phi(X) \right)^{-1/2}$$

$$\|W_2\|_2 = O_P\left( \frac{p}{p_T} \right) \|F'U'\Phi(X)\|_F = O_P\left( \frac{\sqrt{p}}{\sqrt{p_T}} \right).$$

$$\|W_4\|_2 = O_P\left( \frac{1}{p^2 T} \right) \|\Phi(X)\|_F^2 = O_P\left( \frac{1}{p} \right).$$

Therefore, for $k = 1, \ldots, K$,

$$|\lambda_k(W) - \lambda_k(W_1)| \leq \|W - W_1\|_2 = o_P(1).$$

Hence it suffices to prove that the first $K$ eigenvalues of

$$\frac{1}{p} \left( \Phi(X)\Phi(X) \right)^{-1/2} \Phi(X)\Phi(X) \Lambda \Lambda' \Phi(X) \Phi(X)^T \left( \Phi(X)\Phi(X) \right)^{-1/2}$$

are bounded away from zero, which are also the first $K$ eigenvalues of $\frac{1}{p} \Lambda'Pp = \frac{1}{p} \left( G(X) + \Gamma \right)' P \left( G(X) + \Gamma \right)$. By the technical Lemma ?? in the supplementary material, $\frac{1}{p} \left( G(X) + \Gamma \right)' P \left( G(X) + \Gamma \right) = \frac{1}{p} G(X)' PG(X) + o_P(1)$ Hence it suffices to show that the eigenvalues of $\frac{1}{p} G(X)' PG(X)$ are bounded away from zero.

Note that $p^{-2}G(X)' \Phi(X) \Phi(X)' G(X) = \sum_{i=1}^4 C_i$, where

$$C_1 = \frac{1}{p^2} B' \left( \Phi(X)\Phi(X) \right)^2 B, \quad C_2 = \frac{1}{p^2} R(X) \Phi(X) \Phi(X)' \Phi(X) B$$

$$C_3 = C_2', \quad C_4 = \frac{1}{p^2} R(X)' \Phi(X) \Phi(X)' R(X)$$

31
Lemma ?? implies $\|C_2\|_F = o_P(1) = \|C_4\|_F$. Hence

$$
\lambda_{\min}(\frac{1}{p}G(X)'PG(X)) = \lambda_{\min}(\frac{1}{p}G(X)'\Phi(X)(\frac{1}{p}\Phi(X)'\Phi(X))^{-1}\frac{1}{p}\Phi(X)'G(X))
$$

$$
\geq \lambda_{\min}(\frac{1}{p}\Phi(X)'\Phi(X))^{-1})\lambda_{\min}(\frac{1}{p}G(X)'\Phi(X)\frac{1}{p}\Phi(X)'G(X))
$$

$$
\geq d_{\max}^{-1}\lambda_{\min}(C_1) - o_P(1) \geq d_{\max}^{-1}\lambda_{\min}(\frac{\Phi(X)'\Phi(X)}{p})^2\lambda_{\min}(B'B) - o_P(1)
$$

$$
> d_{\max}^{-1}\lambda_{\min}(B'B) - o_P(1).
$$

This implies for some constant $C > 0$, $\lambda_{\min}(\frac{1}{p}G(X)'PG(X)) \geq C$ with probability approaching one.

### A.2 Convergence of $\hat{F}$

By the definition of eigenvalues, we have

$$
\frac{1}{T_p}(Y'PY)\hat{F} = \hat{F}V
$$

Let

$$
H = \frac{1}{T_p}B'\Phi(X)'\Phi(X)BF\hat{F}V^{-1}. \quad (A.1)
$$

Substituting $Y = \Phi(X)BF' + TF' + R(X)F' + U$, we have, for $F$ being the $T \times K$ matrix of

$$
\hat{F} - F = \left(\sum_{i=1}^{15}A_i\right)V^{-1} \quad (A.2)
$$

where

$$
A_1 = \frac{1}{T_p}FB'\Phi(X)'U\hat{F}, \quad A_2 = \frac{1}{T_p}U'\Phi(X)BF\hat{F}
$$

$$
A_3 = \frac{1}{T_p}U'PU\hat{F}, \quad A_4 = \frac{1}{T_p}FB'\Phi(X)'R(X)F'\hat{F}
$$

$$
A_5 = \frac{1}{T_p}FR(X)'\Phi(X)BF\hat{F}, \quad A_6 = \frac{1}{T_p}FR(X)'PR(X)F'\hat{F}
$$

$$
A_7 = \frac{1}{T_p}FR(X)'PU\hat{F}, \quad A_8 = \frac{1}{T_p}U'PR(X)F'\hat{F},
$$

$$
A_9 = \frac{1}{T_p}FB'\Phi(X)'TF'\hat{F}, \quad A_{10} = \frac{1}{T_p}FG'\Phi(X)BF\hat{F}
$$

$$
A_{11} = \frac{1}{T_p}FT'PU\hat{F}, \quad A_{12} = \frac{1}{T_p}U'PTF'\hat{F}, \quad A_{13} = \frac{1}{T_p}FR(X)'PTF'\hat{F},
$$

$$
A_{14} = \frac{1}{T_p}FT'PR(X)F'\hat{F}, \quad A_{15} = \frac{1}{T_p}GT'PTF'\hat{F}.
$$

The proof is divided into two steps. First, we bound $\frac{1}{p}\|\hat{F} - FH\|_F^2$, and then we prove the convergence for $\frac{1}{p}\|FH - F\|_F^2$. Note that $\frac{1}{p}\|FH - F\|_F^2 \leq \|H - I\|_F^2$, hence in the second step we
bound \(\|H - I\|_F^2\).

To bound \(\frac{1}{T} \|\hat{F} - FH\|_F^2\), note that there is a constant \(C > 0\), so that

\[
\frac{1}{T} \|\hat{F} - FH\|_F^2 \leq C \|V^{-1}\|_2^2 \sum_{i=1}^{15} \frac{1}{T} \|A_i\|_F^2.
\]

Hence we need to bound \(\frac{1}{T} \|A_i\|_F^2\) for \(i = 1, \ldots, 15\). The following lemma gives the stochastic bounds for individual terms.

**Lemma A.3.** (i) \(\frac{1}{T} \|A_1\|_F^2 = O_P(p^{-1})\), \(\frac{1}{T} \|A_2\|_F^2 = O_P(p^{-1})\),

(ii) \(\frac{1}{T} \|A_3\|_F^2 = O_P(J^2/p^2)\), \(\frac{1}{T} \|A_4\|_F^2 = O_P(J^{-\kappa})\), \(\frac{1}{T} \|A_5\|_F^2 = O_P(J^{-\kappa})\).

(iii) \(\frac{1}{T} \|A_6\|_F^2 = O_P(J^{-2\kappa})\), \(\frac{1}{T} \|A_7\|_F^2 = O_P(p^{-1}J^{-(\kappa-1)})\), \(\frac{1}{T} \|A_8\|_F^2 = O_P(p^{-1}J^{-(\kappa-1)})\).

(iv) \(\frac{1}{T} \|A_9\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\), \(\frac{1}{T} \|A_{10}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\).

(v) \(\frac{1}{T} \|A_{12}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\), \(\frac{1}{T} \|A_{13}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\).

\(\frac{1}{T} \|A_{14}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\), \(\frac{1}{T} \|A_{15}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\).

**Proof.** (i) Because \(\|F\|_F^2 = O_P(T)\), \(\|\hat{F}\|_F^2 = O_P(T)\). By Lemma ?? in the supplementary material, \(\|U\Phi(X)B\|_F^2 = O_P(pT)\). Hence \(\frac{1}{T} \|A_1\|_F^2 = O_P(p^{-1})\). The rate of convergence for \(\|A_2\|\) is obtained in the same way.

(ii) We have \(A_3 = \frac{1}{T^p} U'\Phi(X)'(\Phi(X)'\Phi(X)^{-1}\Phi(X)'U)\). By Lemma ??, \(\|U\Phi(X)\|_F = O_P(\sqrt{pT})\). By Assumption [3.2] \(\|(\Phi(X)'\Phi(X)^{-1}\Phi(X)'\Phi(X)')^{-1}\|_2 = O_P(p^{-1})\). So \(\frac{1}{T} \|A_3\|_F^2 = O_P(J^2/p^2)\).

Note that \(\|\Phi(X)B\|_2 \leq \|G(X)\|_2 + \|R(X)\|_2 = O_P(\sqrt{p})\), and \(\|R(X)\|_2 = O_P(pJ^{-\kappa})\) Thus \(\frac{1}{T} \|A_4\|_F^2 \leq \frac{1}{T^3p^2} \|F\|_F^2 \|A\|_F^2 \|A\|_F^2 \|R(X)\|_2^2 = O_P(J^{-\kappa})\).

Similarly, \(\frac{1}{T} \|A_6\|_F^2 = O_P(J^{-\kappa})\).

(iii) Note that \(\|P\|_2 = \|\Phi(X)'\Phi(X)^{-1/2}\Phi(X)'\Phi(X)'\Phi(X)'\Phi(X)^{-1/2}\|_2 = 1\). Hence \(\frac{1}{T} \|A_6\|_F^2 = O_P(J^{-2\kappa})\). In addition,

\(\|A_7\|_F \leq \frac{1}{T^p} \|F\|_F \|\hat{F}\|_F \|R(X)\|_F \|\Phi(X)\|_2 \|\Phi(X)'\Phi(X)^{-1}\|_2 \|\Phi(X)'\Phi(X)^{-1}\|_2 = O_P\left(\frac{TJ}{\sqrt{pJ^\kappa}}\right)\).

Hence \(\frac{1}{T} \|A_7\|_F^2 = O_P(p^{-1}J^{-(\kappa-1)})\). The rate of convergence for \(A_8\) can be bounded in the same way.

(iv) It follows from Lemma ?? that \(\|\Phi(X)'T\|_F^2 = O_P(J \sum_{i=1}^{p} \text{var}(\gamma_{ik}))\). Hence \(\frac{1}{T} \|A_9\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\). The rate for \(A_{10}\) follows similarly. \(\frac{1}{T} \|A_{11}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\).

(v) Similarly, \(\frac{1}{T} \|A_{12}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\).

\(\frac{1}{T} \|A_{13}\|_F^2 = O_P\left(\frac{1}{T^2p^2} \sum_{i=1}^{p} \text{var}(\gamma_{ik})\right)\). The rate for \(A_{14}\) is obtained
in the same way. Finally,
\[
\frac{1}{T} \|A_1\|_F^2 = O_P\left(\frac{1}{p^2} \|\Phi(X)\|_F^2\right) = O_P\left(\frac{1}{p^2} \sum_{i=1}^p \text{var}(\gamma_{ik})\right).
\]

It then follows from Lemmas A.2, A.3 and (A.2) that, \((J = O(\sqrt{p}))\)
\[
\frac{1}{T} \|\hat{F} - FH\|_F^2 = O_P\left(\frac{1}{p^2} + \frac{J^2}{p^2} + \frac{1}{J^c}\right).
\]

But the above convergence rate is not the final result. Given this rate, the rates for \(\frac{1}{T} \|A_1\|_F^2\), \(\frac{1}{T} \|A_3\|_F^2\), \(\frac{1}{T} \|A_7\|_F^2\) and \(\frac{1}{T} \|A_{11}\|_F^2\) can be further improved, as stated in the following lemma.

**Lemma A.4.** (Improved convergence rate)

(i) \(\frac{1}{T} \|A_1\|_F^2 = O_P(p^{-2} + p^{-1}J^{-\kappa} + p^{-1}T^{-1})\),

(ii) \(\frac{1}{T} \|A_3\|_F^2 = O_P(J^2/p^3 + J^4/p^4 + J^{2-\kappa}/p^2 + J^2 p^{-2} T^{-1})\),

(iii) \(\frac{1}{T} \|A_7\|_F^2 = O_P(J^{-1-\kappa}/p^2 + J^{-3-\kappa}/p^3 + J^{1-2\kappa}/p + J^{-\kappa}p^{-1}T^{-1})\),

(iv) \(\frac{1}{T} \|A_{11}\|_F^2 = O_P(J^2 \nu_p^{-2}(p^{-1} + T^{-1} + J^{-\kappa}))\), \((\text{recall } \nu_p = \max_{K \leq K} \frac{1}{p} \sum_{i=1}^p \text{var}(\gamma_{ik}))\).

**Proof.** (i) Note that \(\frac{1}{T} \|A_1\|_F^2 \leq \frac{1}{T^2 p^2} \|\Phi(X)\|_F^2 \|B'\Phi(X)/\hat{U}\|_F^2\).

\[
\|B'\Phi(X)/\hat{U}\|_F^2 \leq 2 \|B'\Phi(X)/\hat{U}(\hat{F} - FH)\|_F^2 + 2 \|B'\Phi(X)/\hat{U}FH\|_F^2,
\]

\[
\leq 2 \|B'\Phi(X)/\hat{U}\|_F^2 \|\hat{F} - FH\|_F^2 + 2 \|B'\Phi(X)/\hat{U}\|_F^2 \|H\|_2^2,
\]

\[
\leq O_P(pT) O_P(T/p + T/J^2/p^2 + T/J^\kappa) + O_P(pT) O_P(1)
\]

\[
= O_P(T^2 + T^2 J^2/p^2 + T^2/p + pT),
\]

where the third inequality follows from Lemmas ??, A.5 and (A.3). This implies \(\frac{1}{T} \|A_1\|_F^2 = O_P(p^{-2} + J^2/p^3 + p^{-1}J^{-\kappa} + p^{-1}T^{-1})\).

(ii) \(\frac{1}{T} \|A_3\|_F^2 \leq \frac{1}{T^2 p^2} \|U'\Phi(X)(\Phi(X)\Phi(X))^{-1}\|_2 \|\Phi(X)/\hat{U}\|_F^2 = O_P(T^2 p^2) \|\Phi(X)/\hat{U}\|_F^2\), where by (A.3) and Lemma ??,

\[
\|\Phi(X)/\hat{U}\|_F^2 \leq 2 \|\Phi(X)/\hat{U}\|_F^2 \|\hat{F} - FH\|_F^2 + 2 \|\Phi(X)/\hat{U}FH\|_F^2,
\]

\[
\leq O_P(pT) O_P(T/p + T/J^\kappa) + O_P(pTJ).
\]

This yields the result. Also note \(\|PU\hat{F}\|_F^2 = O_P(\frac{1}{p} \|\Phi(X)/\hat{U}\|_F^2\).

(iii) We have,
\[
\frac{1}{T} \|A_7\|_F^2 \leq \frac{1}{T^2 p^2} \|FR(X)\|_F^2 \|\Phi(X)(\Phi(X)\Phi(X))^{-1}\|_2 \|\Phi(X)/\hat{U}\|_F^2
\]

\[
= O_P(\frac{1}{T^2 p^2 J^\kappa}) \|\Phi(X)/\hat{U}\|_F^2
\]

\[
= O_P(J^{-1-\kappa}/p^2 + J^{-3-\kappa}/p^3 + J^{1-2\kappa}/p + J^{-\kappa}p^{-1}T^{-1})
\]

34
(iv) Note that \(\|F\Gamma'PU\|_F \leq \|F\Gamma'\Phi(X)(\Phi(X)'\Phi(X))^{-1}\|_2 \|\Phi(X)'U\hat{F}\|_F\). Hence the result follows. \(\square\)

**Lemma A.5.** \(\|H\|_2 = O_P(1)\).

**Proof.** We have, \(H = \frac{1}{T_p} (G(X) - R(X))' (G(X) - R(X)) F' \hat{F} V^{-1}\). The result then follows from Lemma A.2 and \(\|G(X)\|_2 = O_P(\sqrt{p})\).

The final rate of convergence for \(\frac{1}{T} \|\hat{F} - FH\|_F^2\) is summarized as follows.

**Proposition A.1.** As \(J = O(\sqrt{p})\) and \(\kappa \geq 1\),

\[
\frac{1}{T} \|\hat{F} - FH\|_F^2 = O_P \left( \frac{1}{p} + \frac{1}{J^\alpha} + \frac{J^2}{p^2 T} \right).
\]

**Proof.** The result follows from Lemmas A.3, A.4.

The convergence of \(\|H - I\|_F\) will be proved in Proposition A.3 below.

### A.3 Convergence of loadings

Define \(\tilde{B} = \frac{1}{T} [\Phi(X)'\Phi(X)]^{-1}\Phi(X)'Y\hat{F}\). Then

\[
\tilde{G}(X) = \frac{1}{T} PY\hat{F} = \Phi(X)\tilde{B}.
\]

Substituting \(Y = \Phi(X)BF' + \Gamma F' + R(X)F' + U\), and using \(\frac{1}{T} F' F = I\),

\[
\tilde{B} = BH + \sum_{i=1}^{C_i} C_i,
\]

where

\[
\begin{align*}
C_1 &= \frac{1}{T} [\Phi(X)'\Phi(X)]^{-1}\Phi(X)'R(X)F'\hat{F}, & C_2 &= \frac{1}{T} [\Phi(X)'\Phi(X)]^{-1}\Phi(X)'U\hat{F} V^{-1} \\
C_3 &= \frac{1}{T} [\Phi(X)'\Phi(X)]^{-1}\Phi(X)'U(\hat{F} - FH), & C_4 &= \frac{1}{T} BF'(\hat{F} - FH), \\
C_5 &= \frac{1}{T} [\Phi(X)'\Phi(X)]^{-1}\Phi(X)'\Gamma F'\hat{F}.
\end{align*}
\]

Again, the proof is divided into two steps: bounding \(\|\tilde{B} - BH\|_F^2\) and bounding \(\|H - I\|_F^2\). In particular, \(\|\tilde{B} - BH\|_F^2 \leq C \sum_{i=1}^{C_i} \|C_i\|_F^2\) for some \(C > 0\). Each individual term is bounded as below:

**Proposition A.2.** (i) \(\|\tilde{B} - BH\|_F^2 = O_P(\frac{J}{p^2} + \frac{J^2}{p^4} + \frac{Jv}{p})\).

(ii) \(\frac{1}{p} \|\tilde{G}(X) - G(X)H\|_F^2 = O_P(\frac{J}{p^2} + \frac{J^2}{p^4} + \frac{Jv}{p})\).

(iii) \(\frac{1}{p} \|\tilde{F} - FH\|_F^2 = O_P(\frac{J}{p^2} + \frac{Jv}{p} + \frac{1}{p^2} + \frac{1}{p})\).
Proof. (i) By Lemmas ??, ?? and ?? in the supplementary material, straightforward calculation yields
\[
\|C_1\|_F^2 = O_P\left(\frac{1}{J^c}\right), \quad \|C_2\|_F^2 = O_P\left(\frac{J}{T^p}\right), \quad \|C_3\|_F^2 = O_P\left(\frac{J}{p^2} + \frac{J}{pJ^c}\right), \\
\|C_4\|_F^2 = O_P\left(\frac{J}{p^2} + \frac{J}{pI} + \frac{J}{J^c} + \frac{J^p}{p}\right), \quad \|C_5\|_F^2 = O_P\left(\frac{J^p}{p}\right).
\]
Therefore, \(\|\hat{B} - BH\|_F^2 \leq O(1) \sum_{i=1}^5 \|C_i\|_F^2 = O_P\left(\frac{J}{p^2} + \frac{J}{pI} + \frac{J}{J^c} + \frac{J^p}{p}\right).

(ii) Because \(G(X)H = \Phi(X)BH + R(X)H\), the result follows from
\[
\frac{1}{p} \|\hat{G}(X) - G(X)H\|_F^2 \leq \frac{2}{p} \|\Phi(X)(\hat{B} - BH)\|_F^2 + \frac{2}{p} \|R(X)H\|_F^2.
\]

(iii) Substituting \(Y = \Phi(X)BF' + GF' + R(X)F' + U\) into \(\hat{\Gamma} = \frac{1}{T}(I - P)Y\hat{F}\),
\[
\hat{\Gamma} - \Gamma H = \sum_{i=1}^6 D_i
\]
where
\[
D_1 = \frac{1}{T}(I - P)\Gamma F' cleared\, FH, \quad D_2 = \frac{1}{T}U(\hat{F} - FH), \\
D_3 = -P\Gamma H, \quad D_4 = (I - P)R(X)(H + \frac{1}{T}F'(\hat{F} - FH)), \quad D_5 = -\frac{1}{T}PU(\hat{F} - FH), \quad D_6 = \frac{1}{T}(I - P)UFH.
\]
The result then follows from Lemma ??.

### A.4 Convergence of \(H\)

Proposition A.3.
\[
\|H - I_K\|_F = O_P\left(\frac{1}{p} + \frac{1}{\sqrt{pT}} + \frac{1}{J^c/2} + \sqrt{\frac{p^2}{J}}\right).
\]

Proof. Let \(\delta_T = \frac{1}{p} + \frac{1}{\sqrt{pT}} + \frac{1}{J^c/2} + \sqrt{\frac{p^2}{J}}\). Then by Lemma ?? in the supplementary material, \(\frac{1}{T}\hat{F}'F = \frac{1}{T}(\hat{F} - FH)'F + H' = H' + O_P(\delta_T)\). Therefore
\[
H' = V^{-1}\frac{1}{T}\hat{F}'F\frac{1}{p}B'\Phi(X)'\Phi(X)B = V^{-1}H'H\frac{1}{p}B'\Phi(X)'\Phi(X)B + O_P(\delta_T).
\]

So we have \((\frac{1}{p}B'\Phi(X)'\Phi(X)B)H = HV + O_P(\delta_T)\). Moreover,
\[
\|\frac{1}{p}G'G - \frac{1}{p}B'\Phi(X)'\Phi(X)B\|_F = O_P\left(\frac{1}{J^c/2}\right). \quad \text{Hence we get}
\]
\[
\left(\frac{1}{p}G'G\right)H = HV + O_P(\delta_T).
\]
Also by Lemma ?? in the supplementary material, \( \|HH - I_K\|_F = O_P(\delta_T) \).

Let the elements of \( H \) be \( h_{ij} \), diagonal elements of \( \frac{1}{p}G'G \) be \( g_k \), diagonal elements of \( V \) be \( v_k \). So in fact we have obtained

\[
g_i h_{ij} = v_j h_{ij} + O_P(\delta_T) \text{ and } (\sum_{k=1}^{K} h_{ki}^2 - 1)^2 = O_P(\delta_T^2).
\]

\( \|HH - I_K\|_F = o_P(1) \) implies that \( h_{ii} \) is bounded away from zero with probability approaching one. Hence \( g_j h_{jj} = v_j h_{jj} + O_P(\delta_T) \) yields \( v_j = g_j + O_P(\delta_T) \). Hence \( (g_i - g_j) h_{ij} = O_P(\delta_T) \). We assume \( |g_i - g_j| \) is bounded away from zero almost surely when \( i \neq j \), thus \( h_{ij} = O_P(\delta_T) \) when \( i \neq j \).

From the second equation, \( \sum_{k=1}^{K} h_{ki}^2 = 1 + O_P(\delta_T) \). So \( h_{ii} = \pm 1 + O_P(\sqrt{\delta_T}) \). From here, we can assume \( h_{ii} = 1 + O_P(\sqrt{\delta_T}) \), otherwise we can always multiply the corresponding columns of \( \hat{F} \) and \( \hat{G}(X) \) by \(-1\). Note that \( h_{ii} - 1 = \frac{1}{2}(h_{ii}^2 - 1 - (h_{ii} - 1)^2) = O_P(\delta_T) \). Finally,

\[
\|H - I_K\|_F^2 = \sum_{i \neq j} h_{ij}^2 + \sum_{i=1}^{K} (h_{ii} - 1)^2 = O_P(\delta_T^2).
\]

### A.5 Proof of Theorem 4.1

\[
\|H - I_K\|_F^2 = O_P\left(\frac{1}{p^2} + \frac{1}{pT} + \frac{1}{J^\kappa} + \frac{\nu_P}{p}\right).
\]

It follows from Propositions A.1 A.2 and A.3 and \( J^2 = O(p) \) that

\[
\frac{1}{T} \|\hat{F} - F\|_F^2 = O_P\left(\frac{1}{p} + \frac{1}{J^\kappa} + \frac{J^2}{p^2 T}\right),
\]

\[
\frac{1}{p} \|\hat{G}(X) - \hat{G}(X)\|_F^2 = O_P\left(\frac{J}{p^2} + \frac{J}{pT} + \frac{J\nu_p}{p}\right),
\]

\[
\|\hat{B} - B\|_F^2 = O_P\left(\frac{J}{p^2} + \frac{J}{pT} + \frac{J\nu_p}{p}\right),
\]

\[
\frac{1}{p} \|\hat{\Gamma} - \Gamma\|_F^2 = O_P\left(\frac{1}{J^\kappa} + \frac{J\nu_p}{p} + \frac{1}{T}\right).
\]

This also implies, \( \max_k \|b_k - \hat{b}_k\| = O_P\left(\frac{J^{1/2}}{p} + \frac{J^{1/2}}{\sqrt{pT}} + \frac{J^{1/2}}{J^{\kappa/2}} + \frac{J^{1/2} v_p^{1/2}}{p^{1/2}}\right) \),

\[
\max_{k \leq K} \|b_k - \hat{b}_k\| \leq \sup_{x \in \mathcal{X}} |\phi(x)| \max_k \|b_k - \hat{b}_k\| + \sup_{x \in \mathcal{X}} \sum_{l=1}^{d} R_{kl}(x_l) \\
= O_P\left(\frac{J^{1/2}}{p} + \frac{J^{1/2}}{\sqrt{pT}} + \frac{J^{1/2}}{J^{\kappa/2}} + \frac{J^{1/2} v_p^{1/2}}{p^{1/2}}\right) \sup_{x \in \mathcal{X}} \|\phi(x)\|.
\]
B Proofs for Section 5

We first prove the limiting distribution of the standardized $S_T$ under $H_0^2$. By Proposition [B.1] below, one of the key steps is to show that estimating $\Sigma_u^{-1}$ by $\hat{\Sigma}_u^{-1}$ does not affect the asymptotic behavior of the leading term, that is,

$$\text{tr}((\frac{1}{T}UF)'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\frac{1}{T}UF)) = o_P(\frac{\sqrt{p}}{T}).$$

(B.1)

However, since $p$ can be either larger or comparable with $T$, the above result cannot be simply implied from the crude bounds like

$$\|\frac{1}{T}UF\|_{\text{max}} \|\frac{1}{T}UF\|_1 \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1, \quad \|\frac{1}{T}UF\|_p^2 \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_2,$$

even when $\Sigma_u^{-1}, \hat{\Sigma}_u^{-1}$ are diagonal matrices. As we shall see in lemma [B.1] below, proving (B.1) is highly technically involved. We need to require $p = o(T^2)$, but still allow $p/T \to \infty$.

**Proposition B.1.** If $p, T, J$ satisfy $\max\{T^{2/3}, JT^{1/2}, J^2\} = o(p)$ and $\max\{Tp^{1/2}, p\} = o(J^\kappa)$, we have

$$\text{tr}(\hat{\Lambda}'(I - P)\hat{\Sigma}_u^{-1}(I - P)\hat{\Lambda}) = \frac{1}{T^2} \text{tr}(F'U\hat{\Sigma}_u^{-1}UF) + o_P(\frac{\sqrt{p}}{T}).$$

**Proof.** We have,

$$\text{tr}(\hat{\Lambda}'(I - P)\hat{\Sigma}_u^{-1}(I - P)\hat{\Lambda})$$

$$= \text{tr}((\sum_{i=1}^6 D_i + \frac{1}{T}UF)'\hat{\Sigma}_u^{-1}(\sum_{i=1}^6 D_i + \frac{1}{T}UF))$$

$$= \text{tr}((\frac{1}{T}UF)'\hat{\Sigma}_u^{-1}\frac{1}{T}UF)) + \text{tr}((\sum_{i=1}^6 D_i)'\hat{\Sigma}_u^{-1}(\sum_{i=1}^6 D_i)) + 2\text{tr}((\sum_{i=1}^6 D_i)'\hat{\Sigma}_u^{-1}\frac{1}{T}UF)$$

$$= S_1 + S_2 + 2S_3,$$

where we define

$$D_1 = \frac{1}{T}GF'(\hat{F} - F), D_2 = \frac{1}{T}U(\hat{F} - F),$$

$$D_3 = -\frac{1}{T}PGF'(\hat{F} - F), D_4 = -\frac{1}{T}PU(\hat{F} - F),$$

$$D_5 = (I - P)R, D_6 = -\frac{1}{T}PUF.$$

Therefore, it suffices to show $S_2, S_3 = o_P(\sqrt{p}/T)$. 

38
From the lemmas given in the above proofs, we have the following results:

\[
\|D_1 + D_3\|_F \leq \left\| \frac{1}{T} (I - P) GF' (\hat{F} - FH) \right\|_F + \left\| \frac{1}{T} (I - P) GF' (H - I) \right\|_F
\]
\[
= O_P\left( \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}} + \sqrt{\frac{p}{J^\kappa}} \right),
\]

where \( J = o(\sqrt{p}) \) and \( \|G\| = O_P(\sqrt{p}) \) by assumption and \( \|I - P\| = O_P(1) \); 

\[
\|D_2 + D_4\|_F \leq \left\| \frac{1}{T} (I - P) U (\hat{F} - FH) \right\|_F + \left\| \frac{1}{T} (I - P) UF (H - I) \right\|_F
\]
\[
= \frac{1}{\sqrt{T}} \times O_P\left( \|U\| \left( \|\hat{F} - FH\|_F + \|F\|_F \|H - I\|_F \right) \right)
\]
\[
= \frac{1}{\sqrt{T}} \times O_P\left( \sqrt{T} + \sqrt{p} \left( O_P\left( \frac{1}{\sqrt{p}} + \frac{1}{J^{\kappa/2}} \right) + O_P\left( \frac{1}{p} + \frac{1}{\sqrt{pT}} + \frac{1}{J^{\kappa/2}} \right) \right) \right)
\]
\[
= O_P\left( \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}} + \sqrt{\frac{1}{J^\kappa}} + \sqrt{\frac{p}{TJ^\kappa}} \right),
\]

where again we used \( J = o(\sqrt{p}) \), \( \|I - P\| = O_P(1) \) and the random matrix fact \( \|U\| = O_P(\sqrt{T} + \sqrt{p}) \);

\[
\|D_5\|_F \leq \|I - P\| \|R\|_F = O_P\left( \sqrt{\frac{p}{J^\kappa}} \right);
\]

\[
\|D_6\|_F \leq O_P\left( \|\frac{1}{T\sqrt{p}} \Phi(X)' UF\|_F \right) = O_P\left( \sqrt{\frac{J}{T}} \right).
\]

Hence,

\[
S_2 = O_P\left( \sum_{i=1}^{6} \|D_i\|^2_F \right) = O_P\left( \frac{J}{T} + \frac{1}{p} + \frac{p}{J^\kappa} \right).
\]

So \( S_2 = o_P(\sqrt{p}/T) \) if \( J = o(\sqrt{p}) \), \( T^2 = o(p^2) \) and \( T\sqrt{p} = o(J^\kappa) \). The three conditions are satisfied by our assumption of the dimension regime. Then let us work on \( S_3 \).

\[
S_3 = \sum_{i=1}^{6} \text{tr}\left( \frac{1}{T} D_i' \hat{\Sigma}_u^{-1} UF \right).
\]

We need to bound \( S_3 \) term by term, and lemmas ??-?? in the supplementary material are used. Note that under the assumption \( J = o(\sqrt{p}) \), rate results of in those lemmas could be written in a much simpler form. Lemmas ?? and ?? imply the following results.

\[
\text{tr}\left( \frac{1}{T} D_1' \hat{\Sigma}_u^{-1} UF \right) = \frac{1}{T^2} \text{tr}\left( (\hat{F} - F)' G' \hat{S}_u^{-1} UF \right) \leq \frac{1}{T^2} \| (\hat{F} - F)' F \|_F \| G' \hat{S}_u^{-1} UF \|_F
\]
\[
= O_P\left( \frac{1}{T} + \frac{1}{\sqrt{pT}} + \sqrt{\frac{p}{TJ^\kappa}} \right).
\]
\[
\text{tr}(\frac{1}{T} \hat{D}_3 \hat{\Sigma}_u^{-1} \mathbf{UF}) = \frac{1}{T^2} \text{tr}((\hat{F} - F)'F \hat{G}' \hat{P} \hat{\Sigma}_u^{-1} \mathbf{UF}) \leq \frac{1}{T^2} \| (\hat{F} - F)'F \|_F \| \hat{G}' \hat{P} \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F
\]
\[
\leq \frac{1}{T^2} \| (\hat{F} - F)'F \|_F (\| \hat{B}' \hat{\Phi}(X)' \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F + \| \hat{R}' \|_F \| \hat{P} \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F)
\]
\[
= O_P(\frac{1}{T} + \frac{1}{\sqrt{pT}} + \sqrt{\frac{p}{TJ^\kappa}}).
\]

\[
\text{tr}(\frac{1}{T} \hat{D}_4 \hat{\Sigma}_u^{-1} \mathbf{UF}) = \frac{1}{T^2} \text{tr}((\hat{F} - F)'U \hat{P} \hat{\Sigma}_u^{-1} \mathbf{UF})
\]
\[
\leq \frac{1}{T^2} \| (\hat{F} - F)'U \hat{\Phi}(X) \|_F \| [\frac{1}{p} \hat{\Phi}(X)' \hat{\Phi}(X)]^{-1} \| \| \hat{\Phi}(X)' \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F
\]
\[
= \frac{1}{T^2} O_P(\sqrt{JT} + \sqrt{\frac{pT}{J^\kappa}}) = O_P(\frac{J}{\sqrt{pT}} + \frac{J}{\sqrt{TJ^\kappa}}).
\]

\[
\text{tr}(\frac{1}{T} \hat{D}_5 \hat{\Sigma}_u^{-1} \mathbf{UF}) = \frac{1}{T} \text{tr}(\hat{R}'(I - P) \hat{\Sigma}_u^{-1} \mathbf{UF}) \leq \frac{\sqrt{K}}{T} (\| \hat{R}' \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F + \| \hat{R}' \hat{P} \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F)
\]
\[
\leq \frac{\sqrt{K}}{T} (\| \hat{R}' \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F
\]
\[
+ \| \hat{R}' \|_F \| [\frac{1}{p} \hat{\Phi}(X) \| [\frac{1}{p} \hat{\Phi}(X)' \hat{\Phi}(X)]^{-1} \| \| \hat{\Phi}(X)' \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F
\]
\[
= \frac{1}{T} O_P(\sqrt{pTJ^{-\kappa}} + \sqrt{pJ^{-\kappa}} \sqrt{JT}) = O_P(\sqrt{\frac{pJ}{TJ^\kappa}}).
\]

\[
\text{tr}(\frac{1}{T} \hat{D}_6 \hat{\Sigma}_u^{-1} \mathbf{UF}) = \frac{1}{T^2} \text{tr}(\hat{F}'U' \hat{P} \hat{\Sigma}_u^{-1} \mathbf{UF}) \leq \frac{1}{T^2} \| [\frac{1}{p} \hat{\Phi}(X)' \hat{\Phi}(X)]^{-1} \| \| \hat{\Phi}(X)' \mathbf{UF} \|_F \| \hat{\Phi}(X)' \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F
\]
\[
= \frac{1}{T^2} O_P(PTJ) = O_P(\frac{J}{T}).
\]

Lemmas (iii) and (iv) in the supplementary material imply the following results.

\[
\text{tr}(\frac{1}{T} \hat{D}_3 \hat{\Sigma}_u^{-1} \mathbf{UF}) = \frac{1}{T^2} \text{tr}((\hat{F} - F)'F \hat{\Sigma}_u^{-1} \mathbf{UF}) \leq \frac{1}{T^2} \sqrt{K} \| (\hat{F} - F)'F \hat{\Sigma}_u^{-1} \mathbf{UF} \|_F
\]
\[
= O_P(\frac{1}{T} + \frac{1}{\sqrt{pT}} + \sqrt{\frac{p^2}{T^2J^\kappa}} + o_P(\sqrt{p})).
\]

Combining the above 6 terms together, we derive

\[
S_3 = O_P(\frac{J}{T} + \frac{J}{\sqrt{pT}} + \sqrt{\frac{p}{T^3}} + \sqrt{\frac{p^2}{T^2J^\kappa}} + o_P(\sqrt{p})) = o_P(\sqrt{\frac{p}{T}}),
\]

if $J = o(\sqrt{p})$, $TJ^2 = O(p^2)$ and $\max\{JT, p\} = o(J^\kappa)$. These three conditions are again satisfied by our assumption of the dimension regime.

\[\square\]
Lemma B.1. Suppose $p \log^4 p = o(T^2)$, $J \log^3 p = o(T)$ and $p \log^2 p = o(J^{2\kappa})$. Then

$$\text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\frac{1}{T} \mathbf{U} \mathbf{F})) = o_P(\frac{$\sqrt{p}$}{T}).$$

Proof. We have, $\text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\frac{1}{T} \mathbf{U} \mathbf{F}))$ equals

$$\text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\frac{1}{T} \mathbf{U} \mathbf{F}))
\leq \text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\Sigma_u - \hat{\Sigma}_u)(\frac{1}{T} \mathbf{U} \mathbf{F}))
+ \text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \hat{\Sigma}_u)(\frac{1}{T} \mathbf{U} \mathbf{F}))
\leq ||\mathbf{U} \mathbf{F}||_F^2 ||\Sigma_u^{-1}||_2 ||\Sigma_u - \hat{\Sigma}_u||_2 ||\Sigma_u - \hat{\Sigma}_u||_2
+ \text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \hat{\Sigma}_u)(\frac{1}{T} \mathbf{U} \mathbf{F})).$$

Because $||\mathbf{U} \mathbf{F}||_F^2 = O_P(pT)$, $||\Sigma_u^{-1}||_2 = O_P(1)$, $||\Sigma_u - \hat{\Sigma}_u||_2 = O_P(\sqrt{\frac{\log p}{T} + J^{-\kappa/2}})$, the first term on the right hand side is $O_P(\frac{\log p}{T^2} + \frac{p}{J^2T}) = o_P(\frac{$\sqrt{p}$}{T})$, given that $p \log p)^2 = o(T^2)$ and $p = o(J^{2\kappa})$.

The difficulty arises in bounding the second term on the right hand side. We aim to show, $\text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \hat{\Sigma}_u)(\frac{1}{T} \mathbf{U} \mathbf{F})) = o_P(T \sqrt{\frac{p}{T}})$. As we argued before, simple inequalities like Holder’s or Cauchy-Schwarz do not work. Note that $\Sigma_u$ is a diagonal matrix. Write $w_i = \frac{1}{\mathbf{E}u_{it}^2}$,

$$\text{tr}((\frac{1}{T} \mathbf{U} \mathbf{F})'(\hat{\Sigma}_u^{-1} - \hat{\Sigma}_u)(\frac{1}{T} \mathbf{U} \mathbf{F})) = \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} \hat{u}_{is}^2 - u_{is}^2 \sum_{t=1}^{T} w_i u_{it} f_{tk}^2 + \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} u_{is}^2 - E u_{is}^2 \sum_{t=1}^{T} w_i u_{it} f_{tk}^2$$

Denote by $\lambda_i'$ as the ith row of $\mathbf{A}$. For part (1), note that $\frac{1}{T} \sum_t \hat{u}_{it}^2 - u_{it}^2 = \frac{1}{T} \sum_t (\hat{u}_{it} - u_{it})^2 + \frac{2}{T} \sum_t (\hat{u}_{it} - u_{it}) u_{it}$, and

$$\hat{u}_{it} - u_{it} = (\hat{\alpha}_i - \alpha_i)'(\hat{f}_t - f_t) + (\hat{\alpha}_i - \alpha_i)'f_t + \lambda_i'(\hat{f}_t - f_t).$$

On the other hand, $\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2 = O_P(\frac{\log p}{T} + \frac{J^2}{p^2T} + \frac{1}{J})$, $\max_{i,k} (\sum_t w_i u_{it} f_{tk})^2 = O_P(T \log p)$, hence,

(1) \quad \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} (\hat{u}_{is} - u_{is})^2 (\sum_{t=1}^{T} w_i u_{it} f_{tk})^2 + 2 \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} (\hat{u}_{is} - u_{is}) u_{is} (\sum_{t=1}^{T} w_i u_{it} f_{tk})^2

\quad = \quad \frac{p}{T} O_P(\frac{\log p}{T} + \frac{J^2}{p^2T} + \frac{1}{J}) O_P(T \log p)

\quad + 2 \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} (\hat{\alpha}_i - \alpha_i)' f_{s} u_{is} (\sum_{t=1}^{T} w_i u_{it} f_{tk})^2
+2 \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} (\hat{\lambda}_i - \lambda_i)(\hat{f}_s - f_s)u_{is}(\sum_{t=1}^{T} w_{it}f_{tk})^2 \\
+2 \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} \lambda_i(\hat{f}_s - f_s)u_{is}(\sum_{t=1}^{T} w_{it}f_{tk})^2 \\
= OP(p \log^2 p + \frac{Tp(\log p)}{J^{1/2}}) \\
+ \max_i ||\hat{\lambda}_i - \lambda_i||_2 \max_i \left\{ \frac{1}{T} \sum_s f_{si} \right\} \max_i \left( \sum_{t=1}^{T} w_{it}f_{tk} \right)^2 \\
+ \max_i ||\hat{\lambda}_i - \lambda_i||_2 \max_i \left\{ \frac{1}{T} \sum_s (\hat{f}_s - f_s)u_{is} \right\} \max_i \left( \sum_{t=1}^{T} w_{it}f_{tk} \right)^2 \\
+2 \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} \lambda_i^2(\hat{f}_s - f_s)u_{is}(\sum_{t=1}^{T} w_{it}f_{tk})^2 \\
= OP(p \log^2 p + \frac{Tp(\log p)}{J^{1/2}}) + \sqrt{JpT(\log p)^{3/2} + p^{3/4}T^{1/4}J^{1/4}(\log p)^{3/2}} \\
+2 \sum_{i=1}^{p} \sum_{k=1}^{K} \frac{1}{T} \sum_{s=1}^{T} \lambda_i^2(\hat{f}_s - f_s)u_{is}(\sum_{t=1}^{T} w_{it}f_{tk})^2 \\
= OP(p \log^2 p + \frac{Tp(\log p)}{J^{1/2}}) + \sqrt{JpT(\log p)^{3/2}} \\
+p^{3/4}T^{1/4}J^{1/4}(\log p)^{3/2} + T\sqrt{J} \log p + \frac{TJ \log p}{J^{1/2}}).

The last equality follows from Lemma ?? in the supplementary material. Therefore, part (1) is $o_P(T\sqrt{p})$. It also follows from Lemma ?? that part (2) is $o_P(T\sqrt{p})$. This finishes the proof.

**Proof of Theorem 5.1**

By Proposition B.1 and Lemma B.1 we have $S = \frac{1}{T}\text{tr}(F'\Sigma_u^{-1}UF) + o_P(\sqrt{p})$. So it suffices to argue the normality for

$$
\frac{1}{T^2}\text{tr}(F'\Sigma_u^{-1}UF) = [\text{vec}(\frac{1}{T}UF)]'(\Sigma_u^{-1} \otimes I_K)[\text{vec}(\frac{1}{T}UF)]
$$

Let

$$
M_t = \text{vec}(UF) = (f_{u1t}', \ldots, f_{updt}')', \quad (pK) \times 1.
$$

Then it is easy to see $W = \Sigma_u^{-1} \otimes I_K = \left[ \frac{1}{T} \sum_{t=1}^{T}\text{cov}(M_t|F) \right]^{-1}$, because $\frac{1}{T} \sum_{t=1}^{T} f_{it}'f_{it}' = I_K$. Denote $x = T^{-1/2}W^{1/2}M_t$, so $TS = x'x + o_P(\sqrt{p})$ and $x|F \sim N(0, I_{pK})$. Therefore $x'x$, conditioning on $F$, is chi-square distributed with degree of freedom $pK$, and

$$
x'x - \frac{pK}{\sqrt{2pK}} \mid F \xrightarrow{d} N(0, 1),
$$

almost surely in $F$. We now need to prove the convergence unconditional on $F$.

The conditional normality means for any $t$, let $z_t$ be the $t$th quantile of standard normal, then almost surely in $F$,

$$
P \left( \frac{x'x - \frac{pK}{\sqrt{2pK}}}{\mid F} < t \mid F \right) \rightarrow z_t
$$
Since the probability is always bounded by one, by dominated convergence theorem,

\[ EP\left(\frac{x'x - pK}{\sqrt{2pK}} < t|F\right) = P\left(\frac{x'x - pK}{\sqrt{pK}} < t\right) \to z_t, \]

which holds for any fixed \( t \). Hence we have shown \( (x'x - pK)/\sqrt{2pK} \to^d N(0, 1) \). Hence

\[ TS - pK \sqrt{2pK} = x'x - pK \sqrt{pK} + o_P(\sqrt{p/2pK}) \to^d N(0, 1). \]

The proof of the Limiting distribution of \( S_G \) is given in the supplementary material, Appendix ??.

### C Proofs for Section 6

Let \( \lambda_k(A) \) denote the \( k \)th largest eigenvalue of matrix \( A \) and \( \sigma_k(A) \) the \( k \)th singular value of \( A \). In the supplementary material, we cite several lemmas which are useful for proving the behavior of eigenvalues \( \lambda_k(Y'PY/(pT)) \).

The following proofs use the technical lemmas ??-?? in the supplementary material.

#### C.1 Behavior of \( \lambda_k(Y'PY/(pT)) \) for \( k \leq K \)

We write \( G = G(X) \). Recall that \( \lambda_k(A) \) denotes the \( k \)th largest eigenvalue of \( A \).

**Lemma C.1.** There is \( C > 0 \), with probability approaching one, for all \( k = 1, \cdots, K \),

\[ \lambda_k\left(\frac{FG'PGF'}{p^\alpha T}\right) > C. \]

**Proof.** We have

\[
\lambda_k\left(\frac{FG'PGF'}{p^\alpha T}\right) = \lambda_k\left(\frac{G'PG}{p^\alpha}\right) \geq \lambda_{Jd}(\Phi^{-1}(X)'\Phi(bX))^{-1})\lambda_k\left(\frac{GG'\Phi(X)\Phi(X)'}{p^{1+\alpha}}\right)
\]

\[
\geq \frac{1}{C_2} \lambda_k\left(\frac{GG'}{p^\alpha}\right)\sigma_{Jd}(\frac{\Phi(X)}{\sqrt{p}})\sigma_{Jd}(\frac{\Phi(X)'}{\sqrt{p}}) \geq \frac{C_1}{C_2} \lambda_k\left(\frac{GG'}{p^\alpha}\right) = \frac{C_1}{C_2} \lambda_k\left(\frac{GG'}{p^\alpha}\right) \geq C_1. 
\]

**Lemma C.2.** Suppose \( J = o(p^\alpha) \), for all \( k = 1, \cdots, K \),

\[ \lambda_k\left(\frac{Y'PY}{pT}\right) = \lambda_k\left(\frac{FG'PGF'}{p^\alpha T}\right) + O_P\left(\frac{\sqrt{Tp^\alpha}}{p}\right). \]
Proof. Because $Y = GF' + \bar{U}$, where $\bar{U} = \Gamma F' + U$

$$\frac{Y'PY}{T} = \frac{FG'PGF'}{T} + I' + II$$

where

$$I = \frac{FG'P\bar{U}}{T}, \quad II = \frac{\bar{U}'P\bar{U}}{T}.$$ 

So

$$\|I\| \leq \sqrt{p}\|G'\|\frac{\Phi(X)}{\sqrt{p}}\|\|\frac{1}{p}\Phi(X)'\Phi(X)^{-1}\|\|\frac{1}{p\sqrt{T}}\Phi(X)'\bar{U}\|\frac{F}{\sqrt{T}} = O_P(\sqrt{Jp^\alpha}).$$

Then

$$\|II\| \leq p\|\frac{1}{p\sqrt{T}}\bar{U}'\Phi(X)\|\|\frac{1}{p}\Phi(X)'\Phi(X)^{-1}\|\|\frac{1}{p\sqrt{T}}\Phi(X)\bar{U}\| = O_P(J).$$

The above two inequalities use the fact that $\|\Phi(X)'U\|^2_F = O_P(pTJ)$ and $\|\Phi(X)'TF'\|^2_F = O_P(pTJ)$ from Lemma ??.

Hence, if we write

$$\Delta = \frac{Y'PY}{T} - \frac{FG'PGF'}{T}$$

then $\|\Delta/p\| = O_P(\sqrt{Jp^\alpha}/p)$. By Weyl’s theorem Lemma ??,

$$\frac{1}{p}|\lambda_k(\frac{Y'PY}{pT}) - \lambda_k(\frac{FG'PGF'}{pT})| \leq \frac{1}{p}\|\Delta\| = O_P(\sqrt{Jp^\alpha}/p).$$

Proposition C.1. Suppose $J = o(p^\alpha)$, for $k = 1, \ldots, K - 1$,

$$\frac{\lambda_k(Y'PY/(pT))}{\lambda_{k+1}(Y'PY/(pT))} = O_P(1).$$

Proof. Write

$$\delta_k = \lambda_k(Y'PY/(p^\alpha T)) - \lambda_k(FG'PGF'/(p^\alpha T)).$$

The previous lemma shows $|\delta_k| = O_P(\sqrt{T}/p^\alpha)$.

Then

$$\frac{\lambda_k(Y'PY/(pT))}{\lambda_{k+1}(Y'PY/(pT))} = \frac{\lambda_k(FG'PGF'/(p^\alpha T)) + \delta_k}{\lambda_{k+1}(FG'PGF'/(p^\alpha T))} \leq \frac{\lambda_k(FG'PGF'/(p^\alpha T))}{\lambda_{k+1}(FG'PGF'/(p^\alpha T))} + \frac{\delta_k}{\lambda_{k+1}(FG'PGF'/(p^\alpha T))}.$$ 

Note that (1)

$$\frac{|\delta_k|}{\lambda_{k+1}(FG'PGF'/(p^\alpha T))} \leq \frac{|\delta_k|}{\lambda_k(FG'PGF'/(p^\alpha T))}.$$

By Lemma [C.1] with probability approaching one, it is upper bounded by $C|\delta_k|$ for some $C > 0$.

(2)

$$\frac{\lambda_k(FG'PGF'/(p^\alpha T))}{\lambda_{k+1}(FG'PGF'/(p^\alpha T))} \leq \frac{\lambda_1(FG'PGF'/(p^\alpha T))}{\lambda_{k+1}(FG'PGF'/(p^\alpha T))}.$$ 

The top $\lambda_1(FG'PGF'/(p^\alpha T)) = \lambda_1(G'PG'/(p^\alpha)) < C'$ with probability approaching one. Hence this ratio is upper bounded by some constant $C$ with probability approaching one.
(3) With probability approaching one,
\[
\frac{\delta_{k+1}}{\lambda_{k+1}(FG'PGF'/(p^aT))} \leq \frac{\delta_{k+1}}{\lambda_K(FG'PGF'/(p^aT))} \leq C\delta_{k+1},
\]
Combining (1)-(3),
\[
\frac{\lambda_k(Y'PY/(pT))}{\lambda_{k+1}(Y'PY/(pT))} \leq (C + C\delta_k) \frac{1}{1 - C\delta_{k+1}}.
\]

C.2 Behavior of $\lambda_k(Y'PY/(pT))$ for $k > K$

Lemma C.3. Let $m = \min\{T, p\}, M = \max\{T, p\}$. Under assumptions of Theorem 6.1 for any $0 < b < 1$ such that $2K < \lfloor bm \rfloor < Jd$, there exist constants $c$ and $C$ such that for $j = 1, 2, \ldots, \lfloor bm \rfloor - 2K$,
\[
c + o_P(1) \leq m\lambda_{k+j}(Y'PY/(pT)) \leq C + o_P(1),
\]
where $\lfloor bm \rfloor$ denote the nearest integer of $bm; c, C$ and $o_P(1)$ are uniform in $j \leq \lfloor bm \rfloor - 2K$.

Proof. By the assumptions, there exist $c_1$ and $c_2$ such that uniformly in $T$ and $p$,
\[
\lambda_1(\Sigma_u) \leq c_1, \lambda_1(M) \leq c_1,
\]
\[
\lambda_p(\Sigma_u) \geq c_2, \lambda_T(M) \geq c_2.
\]
From Lemma ??, we obtain
\[
\lambda_1(U'PU/M) = \lambda_1(E'S_u^{1/2}PS_u^{1/2}EM/M) \leq \lambda_1(E'E/M)\lambda_1(\Sigma_u)\lambda_1(P)\lambda_1(M) \leq c_1^2\lambda_1(E'E/M),
\]
and from Lemma ??,
\[
\lambda_{\lfloor bm \rfloor}(U'PU/M) = \lambda_{\lfloor bm \rfloor}(E'S_u^{1/2}PS_u^{1/2}EM/M) \geq \lambda_{\lfloor bm \rfloor}(E'S_u^{1/2}PS_u^{1/2}EM/M)\lambda_T(M) \geq c_2\lambda_{\lfloor bm \rfloor}(\Phi(X)'\Phi(X))^{-1}\lambda_{\lfloor bm \rfloor}(E'S_u^{1/2}PS_u^{1/2}EM/M)\lambda_T(M) \geq \frac{c_2}{C_2}\sigma_{jd}(\Phi(X)'\Phi(X))^{-1}\lambda_{\lfloor bm \rfloor}(E'E/M)\lambda_p(\Sigma_u) \geq \frac{c_2C_1}{C_2}\lambda_{\lfloor bm \rfloor}(E'E/M).
\]
Hence,
\[
c_2^2C_1C_2^{-1}\lambda_{\lfloor bm \rfloor}(E'E/M) \leq \lambda_{\lfloor bm \rfloor}(U'PU/M) \leq \lambda_1(U'PU/M) \leq c_1^2\lambda_1(E'E/M).
\]

45
We decompose $Y'PY$ by

$$Y'PY = (FG' + U')P(\bar{G}F' + U) = \bar{F}G'P\bar{G}F' + UP[I - \bar{G}(G'PG)^{-1}\bar{G}]PU,$$

where $\bar{G} = G + \Gamma$ and $\bar{F} = F + U'PG(G'PG)^{-1}$. So from Lemma ??,

$$\lambda_{K+j}(Y'PY) \leq \lambda_{K+1}(\bar{F}G'P\bar{G}F') + \lambda_j(U'P[I - \bar{G}(G'PG)^{-1}\bar{G}]PU)$$

$$= \lambda_j(U'P[I - G(G'PG)^{-1}G]PU) \leq \lambda_j(U'PU),$$

where the last inequality is due to Lemma ???. From the decomposition we also notice that

$$\lambda_{K+j}(Y'PY) \geq \lambda_{K+j}(U'P[I - \bar{G}(G'PG)^{-1}\bar{G}]PU)$$

$$= \lambda_{K+j}(U'P[I - G(G'PG)^{-1}G]PU) + \lambda_{K+1}(U'PG(G'PG)^{-1}GPU)$$

$$\geq \lambda_{2K+j}(U'PU).$$

The last equality is because $U'PG(G'PG)^{-1}GPU$ has rank $K$ and the last inequality is again due to Lemma ??.

Hence, for $j \leq [bm] - 2K$,

$$c_2^2C_1C_2^{-1}\lambda_{[bm]}(E'E/M) \leq \lambda_{[bm]}(U'PU/M)$$

$$\leq m\lambda_{K+j}(YPY/(pT)) \leq \lambda_1(U'PU/M) \leq c_1^2\lambda_1(E'E/M).$$

From Lemma ??, since $m/M \rightarrow \gamma \in [0, 1]$, it is clear that $\lambda_1(E'E/M) \xrightarrow{p} (1 + \sqrt{\gamma})^2$. Therefore, $m\lambda_{K+j}(YPY/(pT)) \leq C + o_P(1)$. From Lemma ??, by induction we could conclude that

$$\lambda_{[bm]}(E'E/M) \geq \lambda_{[bm]}(E'[bm]E_{[bm]}/M) \xrightarrow{p} (1 - \sqrt{\beta})^2,$$

where $E_{[bm]}$ is the $M \times [bm]$ submatrix of $E$. So $m\lambda_{K+j}(YPY/(pT)) \geq c + o_P(1)$. It is clear from the proof that $c, C$ and convergence in probability are uniform in all $j \leq [bm] - 2K$. □

Define

$$\Theta_k = \frac{\lambda_k(YPY/(pT))}{\lambda_{k+1}(YPY/(pT))}.$$

**Proposition C.2.** Let $m = \min\{T, p\}$. Under the assumptions of Theorem 6.7,

$$\max_{K < k \leq [m/2]} \Theta_k = O_P(1), \quad \Theta_K \xrightarrow{p} \infty.$$

**Proof.** By Lemmas C.2 and C.3 choose $b \in (0, 1)$ such that $[m/2] < [bm] - K$, so uniformly for
\[ k = 1, 2, \cdots, [m/2], k \neq K, \]

\[ \Theta_k \leq \frac{C + o_p(1)}{c + o_p(1)} = O_p(1), \quad \text{and} \quad \Theta_K \geq \frac{\lambda_K \left( \frac{\text{FG}'\text{PGF'}}{p^2} \right)}{(C + o_p(1))p^{1-\alpha}m^{-1}}. \]

Because \( p^{1-\alpha}/T \to 0 \) and \( J/p^\alpha \to 0 \), \( \Theta_K \overset{p}{\to} \infty. \)

**Proof of Theorem 6.1**

Define

\[ A = \{k \leq \min\{T, p\}/2 : k \neq K\}. \]

By Proposition C.1 and C.2 for any \( \epsilon > 0 \), there is \( C > 0 \),

\[ P(\max_{k \in A} \Theta_k > C) < \epsilon. \]

Then

\[ P(\widehat{K} \neq K) \leq P(\max_{k \in A} \Theta_k \geq \Theta_K) \leq P(\Theta_K \leq \max_{k \in A} \Theta_k \leq C) + P(\max_{k \in A} \Theta_k > C) < \epsilon. \]

It implies \( P(\widehat{K} \neq K) \to 0 \) because \( \epsilon > 0 \) is arbitrary.

Further technical lemmas used in the proof are given in the supplementary material.

**References**

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

Alessi, L., Barigozzi, M. and Capassoc, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics and Probability Letters* 80 1806–1813.

Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59 817–858.

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

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* 70 191–221.
Bai, J. and Ng, S. (2013). Principal components estimation and identification of the factors. *Journal of Econometrics* **176** 18–29.

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

Breitung and Pigorsch (2009). A canonical correlation approach for selecting the number of dynamic factors. *Oxford Bulletin of Economics and Statistics* **75** 23–36.

Breitung and Tenhofen (2011). Gls estimation of dynamic factor models. *Journal of the American Statistical Association* **106** 1150–1166.

Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics* **41** 3074–3110.

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics* **40** 2014–2042.

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38** 2118–2144.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9** 717–772.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics, North Holland* **76**.

Choi, I. (2012). Efficient estimation of factor models. *Econometric Theory* **28** 274–308.

Chudik, A. and Pesaran, M. (2013). Large panel data models with cross-sectional dependence: a survey. Tech. rep., University of South California.

Connor, G. and Linton, O. (2007). Semiparametric estimation of a characteristic-based factor model of stock returns. *Journal of Empirical Finance* **14** 694–717.

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

Desai, K. H. and Storey, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association* **107** 135–151.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* **105** 1042–1055.
Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence (with discussion). *Journal of the American Statistical Association* **107** 1019–1035.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* **39** 3320–3356.

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 Shi, X. (2014). Risks of large portfolios. Tech. rep., Princeton University.

Friguet, C., Kloareg, M. and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* **104** 1406–1415.

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.

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

Jung, S. and Marron, J. (2009). Pca consistency in high dimension, low sample size context. *Annals of Statistics* **37** 3715–4312.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39** 2302–2329.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 42–54.

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

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105** 18718–18723.
Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.

Lorentz, G. (1986). *Approximation of functions*. The second edition ed. American Mathematical Society, Rhode Island.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics* **41** 772–801.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069–1097.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168** 244–258.

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

Rigollet, P. and Tsybakov, A. (2012). Estimation of covariance matrices under sparsity constraints. *arXiv preprint arXiv:1205.1210*.

Shen, D., Shen, H., Zhu, H. and Marron, J. (2013). Surprising asymptotic conical structure in critical sample eigen-directions. Tech. rep., University of North Carolina.

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.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional non-paranormal graphical models. *The Annals of Statistics* **40** 2541–2571.