Published in final edited form as:
Ann Stat. 2016 February ; 44(1): 219–254. doi:10.1214/15-AOS1364.

PROJECTED PRINCIPAL COMPONENT ANALYSIS IN FACTOR MODELS

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

[*]Princeton University [†]University of Maryland

Abstract

This paper introduces a Projected Principal Component Analysis (Projected-PCA), which employees principal component analysis to the projected (smoothed) data matrix onto a given linear space spanned by covariates. When it applies to high-dimensional factor analysis, the projection removes noise 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, when the factor loading matrices are related to the projected linear space. When the dimensionality is large, the factors can be estimated accurately even when the sample size is finite. 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' effects 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. The proposed method is illustrated by both simulated data and the returns of the components of the S&P 500 index.

Keywords and phrases

semi-parametric factor models; high dimensionality; loading matrix modeling; rates of covergence; sieve approximation

[‡]This project was supported by National Science Foundation grants DMS-1206464 and 2R01-GM072611-9, and the Research and Scholarship award from University of Maryland.
Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA. jqfan@princeton.edu, weichenw@princeton.edu
Address: Department of Mathematics, University of Maryland. College Park, MD 20742, USA. yuanliao@umd.edu

SUPPLEMENTARY MATERIAL
Supplement: Technical proofs Fan et al. (2015b)
(). This supplementary material contains all the remaining proofs.

1. Introduction

Factor analysis is one of the most useful tools for modeling common dependence among multivariate outputs. Suppose that we observe data $\{y_{it}\}_{i \le p, t \le T}$ that can be decomposed as

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + u_{it}, \quad i = 1, \cdots, p, \quad t = 1, \cdots, T \quad (1.1)$$

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 component that can not be explained by the static common component. Here p and T respectively denote the dimension and sample size of the data.

Model (1.1) has broad applications in the statistics literature. For instance, $\mathbf{y}_t = (y_{1t}, \cdots, y_{pt})'$ can be expression profiles or blood oxygenation level dependent (BOLD) measurements for the t^{th} microarray, proteomic or fMRI-image, whereas i represents a gene or protein or a voxel. See, for example, Desai and Storey (2012); Efron (2010); Fan et al. (2012); Friguet et al. (2009); Leek and Storey (2008). The separations between the common factors and idiosyncratic components are carried out by the low-rank plus sparsity decomposition. See, for example, Cai et al. (2013); Candès and Recht (2009); Fan et al. (2013); Koltchinskii et al. (2011); Ma (2013); Negahban and Wainwright (2011).

The factor model (1.1) has also been extensively studied in the econometric literature, in which $\mathbf{y}_t$ is the vector of economic outputs at time t or excessive returns for individual assets on day t. The unknown factors and loadings are typically estimated by the principal component analysis (PCA) and the separations between the common factors and idiosyncratic components are characterized via static pervasiveness assumptions. See, for instance, Bai (2003); Bai and Ng (2002); Breitung and Tenhofen (2011); Lam and Yao (2012); Stock and Watson (2002) among others. In this paper, we consider static factor model, which differs from the dynamic factor model (Forni et al., 2000, 2015; Forni and Lippi, 2001). The dynamic model allows more general infinite dimensional representations. For this type of model, the frequency domain PCA (Brillinger, 1981) was applied on the spectral density. The so-called *dynamic pervasiveness* condition also plays a crucial role in achieving consistent estimation of the spectral density.

Accurately estimating the loadings and unobserved factors are very important in statistical 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) (Desai and Storey, 2012; Efron, 2010; Fan et al., 2012; Friguet et al., 2009; Leek and Storey, 2008). 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 direct applications of (1.1) with large T infeasible. For instance, in financial applications, to pertain the stationarity in model (1.1) with time-invariant loading coefficients, a relatively short time series is often used. To make observed data less serially correlated, monthly returns are frequently used to reduce the serial correlations, yet a monthly data over three consecutive years contain merely 36 observations.

1.1. This paper

To overcome the aforementioned problems, and when relevant covariates are available, it may be helpful to incorporate them into the model. Let $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})'$ be a vector of d-dimensional covariates associated with the i^{th} variables. In the seminal papers by Connor and Linton (2007) and Connor et al. (2012), the authors studied the following semi-parametric factor model:

$$y_{it}=\sum_{k=1}^{K} g_k(\mathbf{X}_i) f_{tk}+u_{it}, \quad i=1,\cdots,p, t=1,\cdots,T, \tag{1.2}$$

where loading coefficients in (1.1) are modeled as $\lambda_{ik} = g_k(\mathbf{X}_i)$ for some functions $g_k(\cdot)$. For instance, in health studies, $\mathbf{X}_i$ can be individual characteristics (e.g. age, weight, clinical and genetic information); in financial applications $\mathbf{X}_i$ 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

$$\lambda_{ik}=g_k(\mathbf{X}_i)+\gamma_{ik}, \quad i=1,\cdots,p, k=1,\cdots,K, \tag{1.3}$$

where γ_{ik} is the component of loading coefficient that can not be explained by the covariates $\mathbf{X}_i$. Let $\gamma_i = (\gamma_{i1}, \cdots, \gamma_{iK})'$. We assume that $\{\gamma_i\}_{i\le p}$ have mean zero, and are independent of $\{\mathbf{X}_i\}_{i\le p}$ and $\{u_{it}\}_{i\le p,t\le T}$. In other words, we impose the following factor structure

$$y_{it}=\sum_{k=1}^{K}\{g_k(\mathbf{X}_i)+\gamma_{ik}\} f_{tk}+u_{it}, \quad i=1,\cdots,p, t=1,\cdots,T, \tag{1.4}$$

which reduces to model (1.2) when $\gamma_{ik} = 0$ and model (1.1) when $g_k(\cdot) = 0$. When $\mathbf{X}_i$ genuinely explains a part of loading coefficients λ_{ik}, the variability of γ_{ik} is smaller than that of λ_{ik}. Hence, the coefficient γ_{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 $\mathbf{Y}$ be the $p\times T$ matrix of y_{it}, $\mathbf{F}$ be the $T \times K$ matrix of f_{tk}, $\mathbf{G}(\mathbf{X})$ be the $p \times K$ matrix of $g_k(\mathbf{X}_i)$, Γ be the $p \times K$ matrix of γ_{ik}, and $\mathbf{U}$ be $p \times T$ matrix of u_{it}. Then model (1.4) can be written in a more compact matrix form:

$$\boldsymbol{Y}=\{\mathbf{G}(\mathbf{X})+\mathbf{\Gamma}\}\mathbf{F}^{'}+\mathbf{U}. \quad (1.5)$$

We treat the loadings **G**(**X**) and Γ as realizations of random matrices throughout the paper. This model is also closely related to the *supervised singular value decomposition* model, recently studied by Li et al. (2015). The authors showed that the model is useful in studying the gene expression and single-nucleotide polymorphism (SNP) data, and proposed an EM algorithm for parameter estimation.

We propose a projected-PCA estimator for both the loading functions and factors. Our estimator is constructed by first projecting **Y** onto the sieve space spanned by $\{\mathbf{X}_i\}_{i\le p}$, then applying PCA to the projected data or fitted values. Due to the approximate orthogonality condition of **X**, **U** and Γ, the projection of **Y** is approximately **G**(**X**)**F**′, as the smoothing projection suppresses the noise terms Γ 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 for factor analysis (e.g., Stock and Watson (2002), Bai and Ng (2002)) is no longer suitable in the current context. Moreover, the idea of projected-PCA is also potentially applicable to dynamic factor models of Forni et al. (2000), by first projecting the data onto the covariate space.

The asymptotic properties of the proposed 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 (HDLSS) situations (e.g., Jung and Marron (2009); Shen et al. (2013a,b)). 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 $O_P(p^{-1})$. Throughout the paper, K = dim($\mathbf{f}_t$) and d = dim($\mathbf{X}_i$) are assumed to be constant and do not grow.

Let Λ be a p×K matrix of $(\lambda_{ik})_{T\times K}$. Model (1.3) implies a decomposition of the loading matrix:

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

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

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

The first problem is about testing 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 explanatory power on 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 $\mathbf{G}(\mathbf{X}) \neq 0$, 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-PCA 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 similar to those in Ahn and Horenstein (2013) and Lam and Yao (2012). Different from these authors, our method applies to the projected data, and we demonstrate numerically that this can significantly improve the estimation accuracy.

We focus on the case when the observed covariates are time-invariant. When T is small, these covariates are approximately locally constant, so this assumption is reasonable in practice. On the other hand, there may exist individual characteristics that are time-variant (e.g., see Park et al. (2009)). We expect the conclusions in the current paper to still hold if some smoothness assumptions are added for the time varying components of the covariates. Due to the space limit, we provide heuristic discussions on this case in the supplementary material of this paper Fan et al. (2015b). In addition, note that in the usual factor model, $\mathbf{\Lambda}$ was assumed to be deterministic. In this paper, however, $\mathbf{\Lambda}$ is mainly treated to be stochastic, and potentially depend on a set of covariates. But we would like to emphasize that the results presented in Section 3 under the framework of more general factor models hold regardless of whether $\mathbf{\Lambda}$ is stochastic or deterministic. Finally, while some financial applications are presented in this paper, the projected-PCA is expected to be useful in broad areas of statistical applications (e.g., see Li et al. (2015) for applications in gene expression data analysis).

1.2. Notation and organization

Throughout this paper, for a matrix $\mathbf{A}$, let $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$ and $\|\mathbf{A}\|_2 = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_{\max} = \max_{ij} |A_{ij}|$ denote its Frobenius, spectral and max-norms. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a square matrix. For a vector $\mathbf{v}$, let $\|\mathbf{v}\|$ denote its Euclidean norm.

The rest of the paper is organized as follows. Section 2 introduces the new projected-PCA 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. Overview

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

$$\boldsymbol{Y}=\boldsymbol{\Lambda}\mathbf{F}'+\mathbf{U}. \quad (2.1)$$

Suppose we additionally observe a set of covariates $\{\mathbf{X}_i\}_{i\le p}$. The basic idea of the projected PCA is to smooth the observations $\{Y_{it}\}_{i\le p}$ for each given day t against its associated covariates. More specifically, let $\{\hat{Y}_{it}\}_{i\le p}$ be the fitted value after regressing $\{Y_{it}\}_{i\le p}$ on $\{\mathbf{X}_i\}_{i\le p}$ for each given t. This results in a smooth or projected observation matrix $\hat{\mathbf{Y}}$, which will also be denoted by $\mathbf{PY}$. The projected PCA then estimates the factors and loadings by running the PCA based on the projected data $\hat{\mathbf{Y}}$.

Here we heuristically describe the idea of projected PCA; rigorous analysis will be carried out afterwards. Let $\mathscr{X}$ be a space spanned by $\mathbf{X} = \{\mathbf{X}_i\}_{i\le p}$, which is orthogonal to the error matrix $\mathbf{U}$. Let $\mathbf{P}$ denote the projection matrix onto $\mathscr{X}$ (whose formal definition will be given in (2.6) below. At the population level, $\mathbf{P}$ approximates the conditional expectation operator $E(\cdot|\mathbf{X})$, which satisfies $E(\mathbf{U}|\mathbf{X}) = 0$), then $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{PU} \approx 0$. Hence, the projected data $\hat{\mathbf{Y}}$ is an approximately noiseless problem and its sample covariance has the following approximation:

$$\frac{1}{T}\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}=\frac{1}{T}\boldsymbol{Y}'\mathbf{PY} \approx \frac{1}{T}\mathbf{F}\boldsymbol{\Lambda}'\mathbf{P}\boldsymbol{\Lambda}\mathbf{F}' \quad (2.2)$$

Hence, $\mathbf{F}$ and $\mathbf{P}\Lambda$ can be recovered from the projected data $\hat{\mathbf{Y}}$ under some suitable normalization condition.

The normalization conditions we imposed are

$$\frac{1}{T}\mathbf{F}'\mathbf{F}=\mathbf{I}_K, \qquad \boldsymbol{\Lambda}'\mathbf{P}\boldsymbol{\Lambda}\ \text{is a diagonal matrix with distinct entries.} \quad (2.3)$$

Under this normalization, using (2.2), we conclude that the columns of $\mathbf{F}$ are approximately $\sqrt{T}$ times the first K eigenvectors of the $T \times T$ matrix $\frac{1}{T}\boldsymbol{Y}'\mathbf{PY}$. Therefore, the Projected-PCA naturally defines a factor estimator $\hat{\mathbf{F}}$ using the first K principal components of $\frac{1}{T}\boldsymbol{Y}'\mathbf{PY}$.

The projected loading matrix $\mathbf{P}\Lambda$ can also be recovered from the projected data $\mathbf{PY}$ in two (equivalent) ways. Given $\mathbf{F}$, from $\frac{1}{T}\mathbf{PYF}=\mathbf{P}\boldsymbol{\Lambda}+\frac{1}{T}\mathbf{PUF}$, we see $\mathbf{P}\boldsymbol{\Lambda} \approx \frac{1}{T}\mathbf{PYF}$. Alternatively, consider the $p \times p$ projected sample covariance:

$$\frac{1}{T}\mathbf{PYY}'\mathbf{P}=\mathbf{P}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{P}+\tilde{\boldsymbol{\Delta}},$$

where $\tilde{\mathbf{\Delta}}$ is a remaining term depending on **PU**. Right multiplying **P**Λ and ignoring terms depending on **PU**, we obtain $(\frac{1}{T}\mathbf{PYY}'\mathbf{P})\mathbf{P\Lambda} \approx \mathbf{P\Lambda}(\mathbf{\Lambda}'\mathbf{P\Lambda})$. Hence the (normalized) columns of **P**Λ approximate the first K eigenvectors of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$, the p×p sample covariance matrix based on the projected data. Therefore, we can either estimate **P**Λ by $\frac{1}{T}\mathbf{PY}\hat{\mathbf{F}}$ given $\hat{\mathbf{F}}$, or by the leading eigenvectors of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$. In fact, we shall see later that these two estimators are equivalent. If in addition, Λ = **P**Λ, that is, the loading matrix belongs to the space $\mathscr{X}$, then Λ can also be recovered from the projected data.

The above arguments are the fundament of the projected-PCA, and provide the rationale of our estimators to be defined in Section 2.3. We shall make the above arguments rigorous by showing that the projected error **PU** is asymptotically negligible, and therefore the idiosyncratic error term **U** can be completely removed by the projection step.

2.2. Semiparametric Factor Model

As one of the useful examples of forming the space $\mathscr{X}$ and the projection operator, this paper considers model (1.4), where $\mathbf{X}_i$'s and y_{it}'s are the only observable data, and $\{g_k(\cdot)\}_{k\le K}$ 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 $\mathbf{X}_i$'s being the observed “market characteristic variables”. We assume K to be known for now. In Section 6, we will propose a projected-eigenvalue-ratio method to consistently estimate K when it is unknown.

We assume that $g_k(\mathbf{X}_i)$ does not depend on t, which means the loadings represent the cross-sectional heterogeneity only. Such a model specification is reasonable since in many applications using factor models: to pertain the stationarity of the time series, the analysis can be conducted within each fixed time window with either a fixed or slowly-growing T. Through localization in time, it is not stringent to require the loadings be time-invariant. This also shows one of the attractive features of our asymptotic results: under mild conditions, our factor estimates are consistent even if T is finite.

To non-parametrically estimate $g_k(\mathbf{X}_i)$ without the curse of dimensionality when $\mathbf{X}_i$ is multivariate, we assume $g_k(\cdot)$ to be additive: for each $k \le K$, $i \le p$, there are $(g_{k1}, \cdots, g_{kd})$ nonparametric functions such that

$$g_k(\mathbf{X}_i)=\sum_{l=1}^{d} g_{kl}(X_{il}), \quad d=\dim(\mathbf{X}_i). \quad (2.4)$$

Each additive component of g_k is estimated by the sieve method. Define $\{\varphi_1(x), \varphi_2(x), \cdots\}$ to 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 $l \le d$,

$$g_{kl}(X_{il})=\sum_{j=1}^{J}b_{j,kl}\phi_j(X_{il})+R_{kl}(X_{il}),\quad k\le K, i\le p, l\le d. \tag{2.5}$$

Here $\{b_{j,kl}\}_{j\le J}$ are the sieve coefficients of the lth additive component of $g_k(\mathbf{X}_i)$, corresponding to the kth factor loading; R_{kl} is a "remaining function" representing the approximation error; 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 take the same basis functions in (2.5) purely for simplicity of notation.

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

$$\mathbf{b}_k'=(b_{1,k1},\cdots,b_{J,k1},\cdots,b_{1,kd},\cdots,b_{J,kd})\in\mathbb{R}^{Jd},$$
$$\phi(\mathbf{X}_i)'=(\phi_1(X_{i1}),\cdots,\phi_J(X_{i1}),\cdots,\phi_1(X_{id}),\cdots,\phi_J(X_{id}))\in\mathbb{R}^{Jd}.$$

Then, we can write

$$g_k(\mathbf{X}_i)=\phi(\mathbf{X}_i)'\mathbf{b}_k+\sum_{l=1}^{d}R_{kl}(X_{il}).$$

Let $\mathbf{B}=(\mathbf{b}_1,\cdots,\mathbf{b}_K)$ be a $(Jd)\times K$ matrix of sieve coefficients, $\Phi(\mathbf{X})=(\varphi(\mathbf{X}_1),\cdots,\varphi(\mathbf{X}_p))'$ be a $p\times(Jd)$ matrix of basis functions, and $\mathbf{R}(\mathbf{X})$ be $p\times K$ matrix with the (i,k)th element $\sum_{l=1}^{d}R_{kl}(X_{il})$. Then the matrix form of (2.4) and (2.5) is

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

Substituting this into (1.5), we write

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

We see that the residual term consists of two parts: the sieve approximation error $\mathbf{R}(\mathbf{X})\mathbf{F}'$ and the idiosyncratic $\mathbf{U}$. Furthermore, the random effect assumption on the coefficients Γ makes it also behave like noise and hence negligible when the projection operator $\mathbf{P}$ is applied.

2.3. The estimator

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

$$\mathbf{P}=\Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})'. \quad (2.6)$$

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

$$\hat{\mathbf{G}}(\mathbf{X})=\frac{1}{T}\mathbf{PY}\hat{\mathbf{F}}. \quad (2.7)$$

is the estimator of $\mathbf{G}(\mathbf{X})$.

The intuition can be readily seen from the discussions in Section 2.1, which also provides an alternative formulation of $\hat{\mathbf{G}}(\mathbf{X})$ as follows: let $\hat{\mathbf{D}}$ be a K×K diagonal matrix consisting of the largest K eigenvalues of the p×p matrix $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$. Let $\hat{\Xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_K)$ be a p×K matrix whose columns are the corresponding eigenvectors. According to the relation $(\frac{1}{T}\mathbf{PYY}'\mathbf{P})\mathbf{P\Lambda} \approx \mathbf{P\Lambda}(\mathbf{\Lambda}'\mathbf{P\Lambda})$ described in Section 2.1, we can also estimate $\mathbf{G}(\mathbf{X})$ or $\mathbf{P}\Lambda$ by

$$\hat{\mathbf{G}}(\mathbf{X})=\hat{\mathbf{\Xi}}\hat{\mathbf{D}}^{1/2}.$$

We shall show in Lemma A.1 that this is equivalent to (2.7). The weight matrix **P** projects the original data matrix onto the sieve space spanned by **X**. Therefore, unlike the traditional PC method for usual factor models (e.g., Bai (2003), Stock and Watson (2002)), the projected-PCA takes the principal components of the projected data **PY**. The estimator is thus invariant to the rotation-transformations of the sieve bases.

The estimation of the loading component Γ that can not be explained by the covariates can be estimated as follows. With the estimated factors $\hat{\mathbf{F}}$, the least-squares estimator of loading matrix is $\hat{\Lambda} = \mathbf{Y}\hat{\mathbf{F}}/T$, by using (2.1) and (2.3). Therefore, by (1.5), a natural estimator of Γ is

$$\hat{\mathbf{\Gamma}}=\hat{\mathbf{\Lambda}}-\hat{\mathbf{G}}(\mathbf{X})=\frac{1}{T}(\mathbf{I}-\mathbf{P})\boldsymbol{Y}\hat{\mathbf{F}}. \quad (2.8)$$

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

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

$$y_{it}=\mathbf{X}_i'\boldsymbol{\beta}_t+\mu_t+u_{it}, \quad i \le p, t \le T, \quad (2.9)$$

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

The semi-parametric factor model admits (2.9) as a special case. Note that (2.9) can be rewritten as $y_{it} = g(\mathbf{X}_i)'\mathbf{f}_t + u_{it}$ with $K = d + 1$ unobservable “factors” $\mathbf{f}_t = (\mu_t, \boldsymbol{\beta}_t')'$ and “loading” $g(\mathbf{X}_i) = (1, \mathbf{X}_i')'$. 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:

$$\boldsymbol{Y} = \boldsymbol{\Lambda}\mathbf{F}' + \mathbf{U}. \quad (3.1)$$

In the usual statistical applications for factor analysis, the latent factors are assumed to be serially independent, while in financial applications, the factors are often treated to be weakly dependent time series satisfying strong mixing conditions.

We now demonstrate by a simple example 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 β (otherwise, we can use a local constant approximation), where $\beta > 0$. Then, the model reduces to

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

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

$$\overline{y}_{\cdot t} = \beta f_t + \overline{\gamma}_{\cdot} f_t + \overline{u}_{\cdot t},$$

where $\bar{y}_{\cdot t}$, $\bar{\gamma}_{\cdot}$ and $\bar{u}_{\cdot t}$ denote the averages of their corresponding quantities over i. For the 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

$$\hat{\beta}=\left(\frac{1}{T}\sum_{t=1}^{T}\bar{y}_{\cdot t}^{2}\right)^{1/2}, \quad \text{and} \quad \hat{f}_t=\bar{y}_{\cdot t}/\hat{\beta}. \quad (3.2)$$

These estimators are special cases of the projected-PCA estimators. To see this, define $\bar{\mathbf{y}} = (\bar{y}_{\cdot 1}, \ldots, \bar{y}_{\cdot T})'$, and let $\mathbf{1}_p$ be a p-dimensional column vector of ones. Take a naive basis $\Phi(\mathbf{X}) = \mathbf{1}_p$; then the projected data matrix is in fact $\mathbf{PY} = \mathbf{1}_p\bar{\mathbf{y}}'$. Consider the $T \times T$ matrix $\mathbf{Y}'\mathbf{PY} = (\mathbf{1}_p\bar{\mathbf{y}}')'\mathbf{1}_p\bar{\mathbf{y}}' = p\bar{\mathbf{y}}\bar{\mathbf{y}}'$, whose largest eigenvalue is $p\|\bar{\mathbf{y}}\|^2$. From

$$\boldsymbol{Y}'\mathbf{PY}\frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|}=p\|\bar{\mathbf{y}}\|^{2}\frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|},$$

we have the first eigenvector of $\mathbf{Y}'\mathbf{PY}$ equals $\bar{\mathbf{y}}/\|\bar{\mathbf{y}}\|$. Hence the projected-PCA estimator of factors is $\hat{\mathbf{F}}=\sqrt{T}\bar{\mathbf{y}}/\|\bar{\mathbf{y}}\|$. In addition, the projected PCA estimator of the loading vector $\beta\mathbf{1}_p$ is

$$\frac{1}{T}\mathbf{1}_p\bar{\mathbf{y}}'\hat{\mathbf{F}}=\frac{1}{\sqrt{T}}\mathbf{1}_p\|\bar{\mathbf{y}}\|.$$

Hence the projected PCA-estimator of β equals $\|\bar{\mathbf{y}}\|/\sqrt{T}$. These estimators match with (3.2). Moreover, since the ignored two terms $\bar{\gamma}\cdot$ and $\bar{u}_{\cdot t}$ are of order $O_p(p^{-1/2})$, $\hat{\beta}$ and $\hat{f}_t$ converge whether or not T is large. Note that this simple example satisfies all the assumptions to be stated below, and $\hat{\beta}$ and $\hat{f}_t$ achieve the same rate of convergence as that of Theorem 4.1. We shall present more details about this example in Appendix G in the supplementary material.

3.1. Asymptotic Properties of Projected-PCA

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

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

The following assumption is the key condition of the projected-PCA.

Assumption 3.1 (Genuine projection)—*There are positive constants $c_{\min}$ and $c_{\max}$ such that, with probability approaching one (as $p \to \infty$),*

$$c_{\min}<\lambda_{\min}(p^{-1}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda})<\lambda_{\max}(p^{-1}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda})<c_{\max}.$$

Since the dimensions of $\Phi(\mathbf{X})$ and Λ are respectively $p \times Jd$ and $p \times K$, Assumption 3.1 requires $Jd \geq K$, which is reasonable since we assume K, the number of factors, to be fixed throughout the paper.

Assumption 3.1 is similar to the *pervasive* condition on the factor loadings (Stock and Watson (2002)). In our context, this condition requires the covariates $\mathbf{X}$ have non-vanishing explaining power on the loading matrix, so that the projection matrix $\Lambda'\mathbf{P}\Lambda$ has spiked eigenvalues. Note that it rules out the case when $\mathbf{X}$ is completely unassociated with the loading matrix Λ (e.g., when $\mathbf{X}$ is pure noise). One of the typical examples that satisfies this assumption is the semi-parametric factor model (model (1.4)). We shall study this specific type of factor model in Section 4, and prove Assumption 3.1 in the supplementary material Fan et al. (2015b).

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

Assumption 3.2 (Identification)—*Almost surely,* $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$ *and* $\Lambda'\mathbf{P}\Lambda$ *is a* $K \times K$ *diagonal matrix with distinct entries.*

This condition corresponds to the PC1 condition of Bai and Ng (2013), which separately identifies the factors and loadings from their product $\Lambda\mathbf{F}'$. It is often used in factor analysis for identification, and means that the columns of factors and loadings can be orthogonalized (also see Bai and Li (2012)).

Assumption 3.3 (Basis functions)

i. *(i) There are* $d_{\min}$ *and* $d_{\max} > 0$ *so that with probability approaching one (as* $p \to \infty$*),*

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

ii. $\max_{j \leq J, i \leq p, l \leq d} E\varphi_j(X_{il})^2 < \infty$.

Note that $p^{-1}\Phi(\mathbf{X})'\Phi(\mathbf{X}) = p^{-1}\sum_{i=1}^{p}\phi(\mathbf{X}_i)'\phi(\mathbf{X}_i)$ and $\varphi(\mathbf{X}_i)$ is a vector of dimensionality $Jd \ll p$. Thus, condition (i) can follow from the strong law of large numbers. For instance, $\{\mathbf{X}_i\}_{i \leq p}$ are weakly correlated and in the population level $E\varphi(\mathbf{X}_i)'\varphi(\mathbf{X}_i)$ is well-conditioned. In addition, this condition can be satisfied through proper normalizations of commonly used basis functions such as B-splines, wavelets, Fourier basis, etc. In the general setup of this paper, we allow $\{\mathbf{X}_i\}_{i \leq p}$'s to be cross-sectionally dependent and non-stationary. Regularity conditions about weak dependence and stationarity are imposed only on $\{(\mathbf{f}_t, \mathbf{u}_t)\}$ as follows.

We impose the strong mixing condition. Let $\mathscr{F}_{-\infty}^{0}$ and $\mathscr{F}_{T}^{\infty}$ denote the σ-algebras generated by $\{(\mathbf{f}_t, \mathbf{u}_t) : t \leq 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t) : t \geq T\}$ respectively. Define the mixing coefficient

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

Assumption 3.4 (Data generating process)

i. $\{\mathbf{u}_t, \mathbf{f}_t\}_{t \le T}$ *is strictly stationary. In addition,* $Eu_{it} = 0$ *for all* $i \le p$, $j \le K$; $\{\mathbf{u}_t\}_{t \le T}$ *is independent of* $\{\mathbf{X}_i, \mathbf{f}_t\}_{i \le p, t \le T}$.

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*

$$\max_{j \le p} \sum_{i=1}^{p} |Eu_{it}u_{jt}| < C_2, \quad \frac{1}{pT} \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{t=1}^{T} \sum_{s=1}^{T} |Eu_{it}u_{js}| < C_2,$$
$$\max_{i \le p} \frac{1}{pT} \sum_{k=1}^{p} \sum_{m=1}^{p} \sum_{t=1}^{T} \sum_{s=1}^{T} |\mathrm{cov}(u_{it}u_{kt}, u_{is}u_{ms})| < C_2.$$

iv. *Exponential tail: there exist* $r_2 \ge 0$, $r_3 > 0$ *satisfying* $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$ *and* $b_1, b_2 > 0$, *such that for any* $s > 0$, $i \le p$ *and* $j \le K$,

$$P(|u_{it}| > s) \le \exp(-(s/b_1)^{r_2}), \quad P(|f_{jt}| > s) \le \exp(-(s/b_2)^{r_3}).$$

Assumption 3.4 is standard, especially condition (iii) is commonly imposed for high-dimensional factor analysis (e.g., Bai (2003); Stock and Watson (2002)), which requires $\{u_{it}\}_{i \le p, t \le T}$ be weakly dependent both serially and cross-sectionally. It is often satisfied when the covariance matrix $E\mathbf{u}_t\mathbf{u}_t'$ is sufficiently sparse under the strong mixing condition. We provide primitive conditions of condition (iii) in the supplementary material Fan et al. (2015b).

Formally, we have the following theorem:

Theorem 3.1—*Consider the conventional factor model (*3.1*) with Assumptions 3.1–3.4. The projected-PCA estimators* $\widehat{\mathbf{F}}$ *and* $\hat{\mathbf{G}}(\mathbf{X})$ *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 with* $Jd \ge K$*),*

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

To compare with the regular PC method, the convergence rate for the estimated factors is improved for small *T*. In particular, the projected-PCA does not require $T \to \infty$, and also has a good rate of convergence for the loading matrix up to a projection transformation. Hence we have achieved a finite-*T* consistency, which is particularly interesting in the "high-dimensional-low-sample-size" (HDLSS) context, considered by Jung and Marron (2009). In contrast, the conventional PC method achieves a rate of convergence of $O_P(1/p + 1/T^2)$ for estimating factors, and $O_P(1/T + 1/p)$ for estimating loadings. See Remarks 4.1, 4.2 below for additional details.

3.2. Projected-PCA consistency in the HDLSS context

In recent years, substantial work has been done on the PCA consistency on the *spiked covariance model* (e.g., Johnstone (2001) and Paul (2007)), and is extended to the HDLSS context by Ahn et al. (2007), Jung and Marron (2009) and Shen et al. (2013a). In a high-dimensional factor model $\mathbf{y}_t = \Lambda\mathbf{f}_t + \mathbf{u}_t$, let $\Sigma = \text{cov}(\mathbf{y}_t)$ be the $p \times p$ covariance matrix of $\mathbf{y}_t$. Let $\Xi = (\xi_1, \ldots, \xi_K)$ be the leading eigenvectors of Σ. Under the *pervasiveness condition*, the first K eigenvalues of the $p \times p$ covariance matrix $\Sigma = \text{cov}(\mathbf{y}_t)$ grow at rate $O(p)$. Due to the presence of these very spiked eigenvalues, Fan et al. (2013) showed that the leading eigenvectors of Σ can be consistently estimated by those of the $p \times p$ sample covariance matrix $\frac{1}{T}\mathbf{YY}'$ as both $p, T \to \infty$. However, either the consistency fails to hold with a finite T or the rate of convergence is slow when T grows slowly as in the HDLSS context.

With a genuine projection $\mathbf{P}$ that satisfies Assumption 3.1, the projected-PCA estimates Ξ using the leading eigenvectors of the sample covariance matrix based on the projected data $\mathbf{PY}$. Specifically, recall that $\hat{\Xi}$ is a $p\times K$ matrix whose columns are the eigenvectors corresponding to the first largest K eigenvalues of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$, and $\hat{\mathbf{D}}$ is a diagonal matrix consisting of the largest K eigenvalues of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$. The consistency of projected-PCA can be achieved up to a projection error $\frac{1}{\sqrt{p}}\|\mathbf{\Lambda}-\mathbf{P\Lambda}\|_F$ even if T is finite, and the rate of convergence is faster when T also grows.

Let $\tilde{\mathbf{V}}$ be an orthogonal matrix whose columns are the eigenvectors of $\Lambda'\Lambda$, corresponding to the eigenvalues in a decreasing order. Let Σ_u be the $p\times p$ covariance matrix of $\mathbf{u}_t$. We have the following result on the convergence of eigenspace spanned by the spiked eigenvalues.

Theorem 3.2—*Under the conditions of Theorem 3.1, we have, as p* $\to \infty$ *(J, T may either grow simultaneously with p satisfying* $J=o(\sqrt{p})$ *or stay constant with Jd ≥ K), for* $\mathbf{V} = \tilde{\mathbf{V}}'(\Lambda'\Lambda)^{1/2}\hat{\mathbf{D}}^{-1/2}$,

$$\|\hat{\mathbf{\Xi}}-\mathbf{\Xi V}\|_F=O_P\left(\frac{1}{p}\|\sum\nolimits_u\|_2+\sqrt{\frac{J}{pT}+\frac{J}{p}+\frac{1}{\sqrt{p}}}\|\mathbf{P\Lambda}-\mathbf{\Lambda}\|_F\right).$$

Q: Not clear. What is the order of V? How does it related to Theorem 3.1? Why readers need to know all details below?

Briefly speaking, $\hat{\Xi}$ approximates the space spanned by the columns of Ξ, which consists of leading eigenvectors of Σ. In addition, in the high-dimensional factor model, we shall prove in the appendix that, for $\bar{\Lambda} = \Lambda\tilde{\mathbf{V}}$,

$$\|\mathbf{\Xi}-\mathbf{\Lambda}\tilde{\mathbf{V}}(\overline{\mathbf{\Lambda}}'\overline{\mathbf{\Lambda}})^{-1/2}\|_F=O\left(\frac{1}{p}\|\sum\nolimits_u\|_2\right). \quad (3.3)$$

As a result, Theorem 3.2 follows from (3.3). There are three sources of estimation errors: (i) the error from (3.3), (ii) the error from approximating $\mathbf{P}\Lambda$ by $\hat{\Xi}\hat{\mathbf{D}}^{1/2}$, which depends on the projected noise **PU**, and (iii) the projection error $\frac{1}{\sqrt{p}}\|\mathbf{\Lambda}-\mathbf{P}\mathbf{\Lambda}\|_2$. Note that the errors from (i) and (ii) are both asymptotically negligible as $p \to \infty$, and does not require a diverging *T*. The error of the third type depends on the nature of the loading matrix. In the special case when Λ belongs to the space spanned by **X**, corresponding to **G**(**X**) = Λ, this term is also asymptotically negligible as $p \to \infty$.

4. Projected-PCA in Semi-parametric Factor Models

4.1. Sieve approximations

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

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

In the matrix form,

$$\boldsymbol{Y}=\{\mathbf{G}(\mathbf{X})+\mathbf{\Gamma}\}\mathbf{F}'+\mathbf{U},$$

and **G**(**X**) does not vanish (pervasive condition, see Assumption 4.2 below).

The estimators $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$ (**X**) are the projected-PCA estimators as defined in Section 2.3. We now define the estimator of the nonparametric function $g_k(\cdot)$, *k* = 1, …, *K*. In the matrix form, the projected data has the following sieve approximated representation:

$$\mathbf{PY}=\Phi(\mathbf{X})\mathbf{BF}'+\tilde{\mathbf{E}}, \quad (4.1)$$

where $\tilde{\mathbf{E}}$ = **P**Γ**F**′+**PR**(**X**)**F**′+**PU** is "small" because Γ 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**$_1$, …, **b**$_K$) can be estimated by least squares from the projected model (4.1): Ignore $\tilde{\mathbf{E}}$, replace **F** with $\hat{\mathbf{F}}$, and solve (4.1) to obtain

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

We then estimate $g_k(\cdot)$ by

$$\hat{g}_k(\mathbf{x})=\phi(\mathbf{x})'\hat{\mathbf{b}}_k, \quad \forall \mathbf{x}\in\mathscr{X}, \quad k=1,\ldots,K,$$

where $\mathscr{X}$ denotes the support of **X**$_i$.

4.2. Asymptotic analysis

When $\Lambda = \mathbf{G}(\mathbf{X}) + \Gamma$, $\mathbf{G}(\mathbf{X})$ can be understood as the projection of Λ onto the sieve space spanned by $\mathbf{X}$. Hence the following assumption is a specific version of Assumptions 3.1 and 3.2 in the current context.

Assumption 4.1

i. *Almost surly, $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$ and $\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})$ is a $K \times K$ diagonal matrix with distinct entries.*

ii. *There are two positive constants $c_{\min}$ and $c_{\max}$ so that with probability approaching one (as $p \to \infty$),*

$$c_{\min}<\lambda_{\min}(p^{-1}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X}))<\lambda_{\max}(p^{-1}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X}))<c_{\max}.$$

In this section, we do not need to assume $\{\gamma_i\}_{i\leq p}$ to be i.i.d. for the estimation purpose. Cross-sectional weak dependence as in Condition (ii) would be sufficient. The i.i.d. assumption will be only needed when we consider specification tests in Section 5. Define $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iK})'$, and

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

Assumption 4.2

i. *$E\gamma_{ik} = 0$ and $\{\mathbf{X}_i\}_{i\leq p}$ is independent of $\{\gamma_{ik}\}_{i\leq p}$.*

ii. $\max_{k\leq K,i\leq p} Eg_k(\mathbf{X}_i)^2 < \infty$, $\nu_p < \infty$ *and*

$$\max_{k\leq K,j\leq p}\sum_{i\leq p}|E\gamma_{ik}\gamma_{jk}|=O(\nu_p).$$

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

Assumption 4.3 (Accuracy of sieve approximation)—$\forall l \leq d$, $k \leq K$,

i. *the loading component $g_{kl}(\cdot)$ belongs to a Hölder class $\mathscr{G}$ defined by*

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

for some $L > 0$;

ii. *The sieve coefficients $\{b_{k,jl}\}_{j\leq J}$* satisfy for $\kappa = 2(r + \alpha) \geq 4$, *as $J \to \infty$,*

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

where $\mathscr{X}_l$ is the support of the lth element of $\mathbf{X}_i$, and J is the sieve dimension.

iii. $\max_{k,j,l} b_{k,jl}^2 < \infty$.

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

Theorem 4.1—*Suppose $J = o(\sqrt{p})$. Under Assumptions 3.3, 3.4, 4.1–4.3, as p, J → ∞, T can be either divergent or bounded, we have that*

$$\frac{1}{T}\left\|\hat{\mathbf{F}}-\mathbf{F}\right\|_F^2 = O_P\left(\frac{1}{p}+\frac{1}{J^\kappa}\right),$$
$$\frac{1}{p}\left\|\hat{\mathbf{G}}(\mathbf{X})-\mathbf{G}(\mathbf{X})\right\|_F^2 = O_P\left(\frac{J}{p^2}+\frac{J}{pT}+\frac{J}{J^\kappa}+\frac{J\nu_p}{p}\right),$$
$$\max_{k\le K}\sup_{\mathbf{x}\in\mathscr{X}}\left|\hat{g}_k(\mathbf{x})-g_k(\mathbf{x})\right| = O_P\left(\frac{J}{p}+\frac{J}{\sqrt{pT}}+\frac{J}{J^{\kappa/2}}+J\sqrt{\frac{\nu_p}{p}}\right)\max_{j\le J}\sup_x\left|\phi_j(x)\right|.$$

In addition, if T → ∞ simultaneously with p and J, then

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

The optimal $J^* = (p\min\{T, p, \nu_p^{-1}\})^{1/\kappa}$ 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 $\kappa \ge 4$. With $J = J^*$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\mathbf{f}}_t-\mathbf{f}_t\right\|^2 = O_P\left(\frac{1}{p}\right),$$
$$\frac{1}{p}\sum_{i=1}^{p}\left|\hat{g}_k(\mathbf{X}_i)-g_k(\mathbf{X}_i)\right|^2 = O_P\left(\frac{1}{(p\min\{T,p,\upsilon_p^{-1}\})^{1-1/\kappa}}\right), \quad \forall k$$
$$\max_{k\le K}\sup_{\mathbf{x}\in\mathscr{X}}\left|\hat{g}_k(\mathbf{x})-g_k(\mathbf{x})\right| = O_P\left(\frac{\max_{j\le J}\sup_x|\phi_j(x)|}{(p\min\{T,p,\nu_p^{-1}\})^{1/2-1/\kappa}}\right),$$

and $\hat{\Gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$ satisfies:

$$\frac{1}{p}\sum_{i=1}^{p}\left\|\hat{\boldsymbol{\gamma}}_i-\boldsymbol{\gamma}_i\right\|^2 = O_P\left(\frac{1}{(p\min\{T,p,\upsilon_p^{-1}\})^{1-1/\kappa}}+\frac{1}{T}\right).$$

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

[*] One can first conduct the analysis conditioning on the event $\{\hat{K} = K\}$, then argue that the results still hold unconditionally as $P(\hat{K} = K) \to 1$

Remark 4.1—The rates of convergence for factors and nonparametric functions do not require $T \to \infty$. When $T = O(1)$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\hat{\mathbf{f}}_t-\mathbf{f}_t\|^2=O_P\left(\frac{1}{p}\right),\quad \frac{1}{p}\sum_{i=1}^{p}|\hat{g}_k(\mathbf{X}_i)-g_k(\mathbf{X}_i)|^2=O_P\left(\frac{1}{p^{1-1/\kappa}}\right).$$

The rates still converge fast when p is large, demonstrating the blessing of dimensionality. This is an attractive feature of the projected-PCA in the HDLSS context, as in many applications, the stationarity of a time series and the time-invariance assumption on the loadings hold 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 Fan et al. (2015a) (Lemma C.1), the regular PCA method has the following convergence rate

$$\frac{1}{T}\sum_{t=1}^{T}\|\tilde{\mathbf{f}}_t-\mathbf{f}_t\|^2=O_P\left(\frac{1}{p}+\frac{1}{T^2}\right),$$

which is inconsistent when T is bounded.

Remark 4.2—When both p and T are large, the projected-PCA estimates factors as well as the regular PCA does, and achieves a faster rate of convergence for the estimated loadings when γ_{ik} vanishes. In this case, $\lambda_{ik} = g_k(\mathbf{X}_i)$, the loading matrix is estimated by $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{G}}(\mathbf{X})$, and

$$\frac{1}{p}\sum_{i=1}^{p}|\hat{\lambda}_{ik}-\lambda_{ik}|^2=\frac{1}{p}\sum_{i=1}^{p}|\hat{g}_k(\mathbf{X}_i)-g_k(\mathbf{X}_i)|^2=O_P\left(\frac{1}{(pT)^{1-1/\kappa}}+\frac{1}{p^{2-2/\kappa}}\right).$$

In contrast, the regular PCA method as in Stock and Watson (2002) yields

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

Comparing these rates, we see that when $g_k(\cdot)$'s are sufficiently smooth (larger κ), the rate of convergence for the estimated loadings is also improved.

5. Semiparametric Specification Test

The loading matrix always has the following orthogonal decomposition:

$$\mathbf{\Lambda}=\mathbf{G}(\mathbf{X})+\mathbf{\Gamma},$$

where Γ is interpreted as the loading component that cannot be explained by **X**. We consider two types of specification tests: testing $H_0^1{:}\mathbf{G}(\mathbf{X})=0$, and $H_0^2{:}\mathbf{\Gamma}=0$. 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 tests 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:

$$H_0^1{:}\mathbf{G}(\mathbf{X})=0 \ a.s.$$

Due to the approximate orthogonality of **X** and Γ, we have $\mathbf{P}\Lambda \approx \mathbf{G}(\mathbf{X})$. Hence, the null hypothesis is approximately equivalent to

$$H_0{:}\mathbf{P\Lambda}=0 \ a.s.$$

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

$$S_G=\frac{1}{p}\mathrm{tr}(\mathbf{W}_1\tilde{\mathbf{\Lambda}}'\mathbf{P}\tilde{\mathbf{\Lambda}}), \quad \mathbf{W}_1=\left(\frac{1}{p}\tilde{\mathbf{\Lambda}}'\tilde{\mathbf{\Lambda}}\right)^{-1},$$

where the $K \times K$ matrix $\mathbf{W}_1$ is the weight matrix. The null hypothesis is rejected when S_G is large.

The projected PCA estimator is inappropriate under the null hypothesis as the projection is not genuine. We therefore use the least squares estimator $\tilde{\Lambda} = \mathbf{Y}\tilde{\mathbf{F}}/T$, leading to the test statistic

$$S_G=\frac{1}{T^2p}\mathrm{tr}(\mathbf{W}_1\tilde{\mathbf{F}}'\mathbf{Y}'\mathbf{P}\mathbf{Y}\tilde{\mathbf{F}}).$$

Here, we take $\tilde{\mathbf{F}}$ as the regular PC estimator: the columns of $\tilde{\mathbf{F}}/\sqrt{T}$ are the first K eigenvectors of the $T \times T$ data matrix $\mathbf{Y}'\mathbf{Y}$.

5.2. Testing Γ = 0

Connor et al. (2012) applied the semi-parametric factor model to analyzing financial returns, who assumed that $\mathbf{\Gamma} = 0$, that is, the loading matrix can be fully explained by the observed covariates. It is therefore natural to test the following null hypothesis of specification:

$$H_0^2{:}\boldsymbol{\Gamma}=0 \ a.s.$$

Recall that $\mathbf{G}(\mathbf{X}) \approx \mathbf{P}\Lambda$ so that $\Lambda \approx \mathbf{P}\Lambda + \Gamma$. Therefore essentially the specification testing problem is equivalent to testing:

$$H_0{:}\mathbf{P}\boldsymbol{\Lambda}=\boldsymbol{\Lambda} \ a.s.$$

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

$$\mathrm{tr}(\hat{\boldsymbol{\Gamma}}'\mathbf{W}_2\hat{\boldsymbol{\Gamma}})=\mathrm{tr}(\hat{\boldsymbol{\Lambda}}'(\mathbf{I}-\mathbf{P})'\mathbf{W}_2(\mathbf{I}-\mathbf{P})\hat{\boldsymbol{\Lambda}}),$$

for some $p \times p$ positive definite weight matrix$\mathbf{W}_2$, where $\widehat{\mathbf{F}}$ is the projected-PCA estimator for factors and $\widehat{\Lambda} = \mathbf{Y}\widehat{\mathbf{F}}/T$. To control the size of the test, we take $\mathbf{W}_2=\sum_u^{-1}$, where Σ_u is a diagonal covariance matrix of $\mathbf{u}_t$ under H_0, assuming that $(u_{1t}, \cdots, u_{pt})$ are uncorrelated.

We replace $\sum_u^{-1}$ with its consistent estimator: let $\hat{\mathbf{U}} = \mathbf{Y} - \widehat{\Lambda}\widehat{\mathbf{F}}'$. Define

$$\widehat{\sum}_u=T^{-1}\mathrm{diag}\{\hat{\mathbf{U}}\hat{\mathbf{U}}'\}=T^{-1}\mathrm{diag}\{\boldsymbol{Y}(\mathbf{I}-T^{-1}\hat{\mathbf{F}}\hat{\mathbf{F}}')\boldsymbol{Y}'\}.$$

Then the operational test statistic is defined to be

$$S_\Gamma=\mathrm{tr}(\hat{\boldsymbol{\Lambda}}'(\mathbf{I}-\mathbf{P})'\widehat{\sum}_u^{-1}(\mathbf{I}-\mathbf{P})\hat{\boldsymbol{\Lambda}}).$$

The null hypothesis is rejected for large values of S_Γ.

5.3. Asymptotic null distributions

For the testing purpose we assume $\{\mathbf{X}_i, \gamma_i\}$ to be i.i.d., and let $T, p, J \to \infty$ simultaneously. The following assumption regulates the relation between T and p.

Assumption 5.1—Suppose

- **i.** $\{\mathbf{X}_i, \gamma_i\}_{i \le p}$ *are independent and identically distributed;*
- **ii.** $T^{2/3} = o(p)$, *and* $p(\log p)^4 = o(T^2)$.
- **iii.** *J and* κ *satisfy:* $J=o(\min\{\sqrt{p}, \sqrt{T}\})$, *and* $\max\{T\sqrt{p}, p\}=o(J^\kappa)$.

Condition (ii) 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 $\sum_u^{-1}$ 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 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. (2013b)). The iid assumption of covariates $\mathbf{X}_i$ in Condition (i) can be relaxed with further distributional assumptions on γ_i (e.g., assuming γ_i to be Gaussian). The conditions on J in Condition (iii) is consistent with those of the previous sections.

We focus on the case when $\mathbf{u}_t$ is Gaussian, and show that under H_0^1,

$$S_G = (1+o_P(1))\frac{1}{p}\mathrm{tr}(\mathbf{W}_1\mathbf{\Gamma}'\mathbf{P}\mathbf{\Gamma}),$$

and under H_0^2

$$S_\Gamma = (1+o_P(1))\frac{1}{T^2}\mathrm{tr}(\mathbf{F}'\mathbf{U}'\sum\nolimits_u^{-1}\mathbf{U}\mathbf{F}),$$

whose conditional distributions (given $\mathbf{F}$) under the null are χ^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 3.3, 3.4, 4.2, 5.1 hold. Then under* H_0^1,

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

where $K = \dim(\mathbf{f}_t)$ *and* $d = \dim(\mathbf{X}_i)$. *In addition, suppose Assumptions 4.1 and 4.3 further hold,* $\{\mathbf{u}_t\}_{t\leq T}$ *is i.i.d.* $N(0, \Sigma_u)$ *with a diagonal covariance matrix* Σ_u *whose elements are bounded away from zero and infinity. Then under* H_0^2,

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

In practice, when a relatively small sieve dimension J is used, one can instead use the upper α-quantile of the χ^2_{JdK} distribution for pS_G.

Remark 5.1—We require u_{it} be independent across t, which ensures that the covariance matrix of the leading term $\mathrm{vec}(\frac{1}{\sqrt{T}}\mathbf{U}\mathbf{F}')$ to have a simple form $\sum_u^{-1} \otimes \mathbf{I}_K$. This assumption can be relaxed to allow for weakly dependent $\{\mathbf{u}_t\}_{t \le T}$, but many autocovariance terms will be involved in the covariance matrix. One may regularize standard autocovariance matrix estimators such as Newey and West (1987) and Andrews (1991) to account for the high dimensionality. Moreover, we assume Σ_u be diagonal to facilitate estimating $\sum_u^{-1}$, which can also be weakened to allow for a non-diagonal but sparse Σ_u. Regularization methods such as thresholding (Bickel and Levina (2008)) can then be employed, though they are expected to be more technically involved.

6. Estimating the Number of Factors from Projected Data

We now address the problem of estimating $K = \dim(\mathbf{f}_t)$ when it is unknown. Once consistent estimation of K is obtained, all the results 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 to select the largest ratio of the adjacent eigenvalues of $\mathbf{Y}'\mathbf{Y}$, based on the fact that the K largest eigenvalues of the sample covariance matrix grow fast as p increases, while the remaining eigenvalues either remain bounded or grow slowly.

We extend Ahn and Horenstein (2013)'s theory in two ways. First, when the loadings depend on the observable characteristics, it is more desirable to work on the projected data $\mathbf{PY}$. Due to the orthogonality condition of $\mathbf{U}$ and $\mathbf{X}$, the projected data matrix is approximately equal to $\mathbf{G}(\mathbf{X})\mathbf{F}'$. The projected matrix $\mathbf{PY}(\mathbf{PY})'$ thus allows us to study the eigenvalues of the principal matrix component $\mathbf{G}(\mathbf{X})\mathbf{G}(\mathbf{X})'$, which directly connects with the strengths of those factors. Since the non-vanishing eigenvalues of $\mathbf{PY}(\mathbf{PY})'$ and $(\mathbf{PY})'\mathbf{PY} = \mathbf{Y}'\mathbf{PY}$ are the same, we can work directly with the eigenvalues of the matrix $\mathbf{Y}'\mathbf{PY}$. Secondly, we allow $p/T \to \infty$.

Let $\lambda_k(\mathbf{Y}'\mathbf{PY})$ denote the kth largest eigenvalue of the projected data matrix $\mathbf{Y}'\mathbf{PY}$. We assume $0 < K < Jd/2$, which naturally holds if the sieve dimension J slowly grows. The estimator is defined as:

$$\hat{K} = \arg \max_{0<k<Jd/2} \frac{\lambda_k(\mathbf{Y}'\mathbf{PY})}{\lambda_{k+1}(\mathbf{Y}'\mathbf{PY})}.$$

The following assumption is similar to that of Ahn and Horenstein (2013). Recall that $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_T)$ is a $p \times T$ matrix of the idiosyncratic components, and $\sum_u = E\mathbf{u}_t\mathbf{u}_t'$ denote the $p \times p$ covariance matrix of $\mathbf{u}_t$.

Assumption 6.1—*The error matrix* $\mathbf{U}$ *can be decomposed as*

$$\mathbf{U}=\sum\nolimits_u^{1/2}\mathbf{E}\mathbf{M}^{1/2}, \quad (6.1)$$

where,

i. *the eigenvalues of* Σ_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* × *T stochastic matrix, where* e_{it} *is independent in both i and t, and* $\mathbf{e}_t = (e_{1t}, \ldots, e_{pt})'$ *are i.i.d. isotropic sub-Gaussian vectors, that is, there is* $C >$ 0, *for all* $s > 0$,

$$\sup_{\|\mathbf{v}\|=1} \sup_{m\ge 1} P(|\mathbf{v}'\mathbf{e}_t| > s) \le \exp(1-s^2/C).$$

iv. *There are* $d_{\min}, d_{\max} > 0$, *almost surely,*

$$d_{\min} \le \lambda_{\min}(\Phi(\mathbf{X})'\Phi(\mathbf{X})/p) \le \lambda_{\max}(\Phi(\mathbf{X})'\Phi(\mathbf{X})/p) \le d_{\max}.$$

This assumption allows the matrix **U** to be both cross-sectionally and serially dependent. The *T* × *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**. In addition, we require $\mathbf{u}_t$ to be sub-Gaussian to apply random matrix theories of Vershynin (2010). For instance, when $\mathbf{u}_t$ is $N(\mathbf{0}, \Sigma_u)$, for any $\|\mathbf{v}\| = 1$, $\mathbf{v}'\mathbf{e}_t \sim N(0, 1)$. Thus condition (iii) is satisfied. Finally, the *almost surely* condition of (iv) seems somewhat strong, but is still satisfied by bounded basis functions (e.g., Fourier basis) and follows from the strong law of large numbers given that $E\varphi(\mathbf{X}_i)\varphi(\mathbf{X}_i)'$ is well conditioned.

We show in the supplementary material that when Σ_u is diagonal (u_{it} is cross-sectionally independent), both the sub-Gaussian assumption and condition (iv) can be relaxed.

The following theorem is the main result of this section.

Theorem 6.1—*Under assumptions of Theorem 4.1 and Assumption 6.1, as p, T* $\to \infty$, *if J satisfies* $J=o(\min\{\sqrt{p}, T\})$ *and K* < *Jd*/2 *(J may either grow or stay constant), we have*

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

7. Numerical Studies

This section presents numerical results to demonstrate the performance of projected-PCA 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 through 2013, and their corresponding market capitalization and book value from Compustat. There are 337 stocks in our data set, whose daily excess returns were calculated. We considered four characteristics $\mathbf{X}$ as in Connor et al. (2012) for each stock: size, value, momentum and volatility, which were calculated using the data before a certain data analyzing window so that characteristics are treated known. See Connor et al. (2012) for detailed descriptions of these characteristics. All four characteristics are standardized to have mean zero and unit variance. Note that the construction 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 $\{y_{it}\}_{i\le 337, t\le 63}$ and characteristics $\mathbf{X}_i$ as the input data and setting $K = 3$, we fit loading functions $g_k(\mathbf{X}_i)=\alpha_{ik}+\sum_{l=1}^{4} g_{kl}(X_{il})$ for $k = 1, 2, 3$ using the projected-PCA 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 3. 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 $\{\mathbf{u}_t\}_{t\le T}$ from $N(\mathbf{0}, \mathbf{D}\Sigma_0\mathbf{D})$ where $\mathbf{D}$ is diagonal and Σ_0 sparse. Generate the diagonal elements of $\mathbf{D}$ from Gamma(α, β) with $\alpha = 7.06$, $\beta = 536.93$ (calibrated from the real data), and generate the off-diagonal elements of Σ_0 from $N(\mu_u, \sigma_u^2)$ with $\mu_u = -0.0019$, $\sigma_u = 0.1499$. Then truncate Σ_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 the i.i.d. Gaussian distribution with mean 0 and standard deviation 0.0027, calibrated with real data.
4. Generate $\mathbf{f}_t$ from a stationary VAR model $\mathbf{f}_t = \mathbf{A}\mathbf{f}_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(\mathbf{0}, \Sigma_\varepsilon)$. The model parameters are calibrated with the market data and listed in Table 1.
5. Finally, generate $\mathbf{X}_i \sim N(\mathbf{0}, \Sigma_X)$. Here Σ_X is a 4×4 correlation matrix estimated from the real data.

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 3 to demonstrate the performance of projected-PCA. Note that the "true" loading curves in the simulation are taken from the estimates calibrated using the real data. The estimates based on simulated data capture the shape of the true curve, though we also notice slight biases at boundaries. But in general, projected-PCA fits the model well.

We also compare our method with the regular PC method (e.g., Stock and Watson (2002)). The mean values of $\|\widehat{\Lambda} - \Lambda\|_{\max}$, $\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_F/\sqrt{p}$, $\|\widehat{\mathbf{F}} - \mathbf{F}_0\|_{\max}$ and $\|\widehat{\mathbf{F}} - \mathbf{F}_0\|_F/\sqrt{T}$ are plotted in Figures 1 and 2. where $\Lambda = \mathbf{G}_0(\mathbf{X}) + \Gamma$ (see section 7.3 for definitions of $\mathbf{G}_0(\mathbf{X})$ and $\mathbf{F}_0$). The breakdown error or $\mathbf{G}_0(\mathbf{X})$ and Γ are also depicted in Figure 1. In comparison, projected-PCA outperforms PC in estimating both factors and loadings including the nonparametric curves $\mathbf{G}(\mathbf{X})$ and random noise Γ. The estimation errors for $\mathbf{G}(\mathbf{X})$ of projected-PCA decrease as the dimension increases, which is consistent with our asymptotic theory.

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 $\{\mathbf{f}_t\}_{t \leq T}$ from the stationary VAR(1) model, that is $\mathbf{f}_t = \mathbf{A}\mathbf{f}_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(0, \mathbf{I})$. We consider $\Gamma = 0$.

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 satisfy the identifiability conditions, we calculate a transformation matrix $\mathbf{H}$ such that $\frac{1}{T}\mathbf{HF}'\mathbf{FH} = \mathbf{I}_K$, $\mathbf{H}^{-1}\mathbf{G}'\mathbf{GH}'^{-1}$ is diagonal. Let the final true factors and loadings be $\mathbf{F}_0 = \mathbf{FH}$, $\mathbf{G}_0 = \mathbf{GH}'^{-1}$. For each p, we run the simulation for 500 times.

We estimate the loadings and factors using both projected-PCA and PC. For projected-PCA, 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 $\mathbf{PYF}_0/T$, where $\mathbf{F}_0$ is the true factor matrix of simulated data.

The estimation error measured in max and standardized Frobenius norms for both loadings and factors are reported in Figures 4 and 5. The plots demonstrate the good performance of projected-PCA in estimating both loadings and factors. In particular, it works well when we encounter small T but a large p. In this design, $\Gamma = 0$, so the accuracy of estimating $\Lambda = \mathbf{G}_0$ is significantly improved by using the projected-PCA. 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. It is also clearly seen that when p is fixed, the improvement on estimating factors is not significant as T grows. This matches with our convergence results for the factor estimators.

It is also interesting to compare projected-PCA 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-PCA is not as good as SLS. But the two methods behave similarly as p increases. This further confirms the theory and intuition that as the dimension becomes larger, the effects of estimating the unknown factors are negligible.

7.4. Estimating number of factors

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

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-PCA 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-PCA still obtains a closer estimated number of factors.

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 and S_Γ, 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 $H_0^1{:}\mathbf{G}(\mathbf{X})=0$), we have very strong evidence of the existence of non-vanishing covariate 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 $H_0^2{:}\mathbf{\Gamma}=0$), we see for a majority of period, the null hypothesis is rejected. In other words, the characteristics of assets cannot fully explain 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 $\mathbf{\Gamma} = 0$ 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.

8. Conclusions

This paper proposes and studies a high-dimensional 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 PCA to estimate the unknown factors, loadings, and number of factors. After projecting the response variable onto the sieve space spanned by the covariates, the projected-PCA yields a significant improvement on the rates of convergence than the regular methods. In particular, consistency can be achieved without a diverging sample size, as long as the dimensionality grows. This demonstrates that the proposed method is useful in the typical HDLSS situations. In addition, we propose new specification tests for the orthogonal decomposition of the loadings, which fill 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 provide theoretical basis for employing projected-PCA methods. On the other hand, our empirical study also shows that the firm characteristics can not fully explain the factor loadings so that the proposed generalized factor model is more appropriate.

References

Ahn J, Marron J, Muller KM, Chi YY. The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika. 2007; 94:760–766.

Ahn S, Horenstein A. Eigenvalue ratio test for the number of factors. Econometrica. 2013; 81:1203–1227.

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

Andrews D. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica. 1991; 59:817–858.

Bai J. Inferential theory for factor models of large dimensions. Econometrica. 2003; 71:135–171.

Bai J, Li Y. Statistical analysis of factor models of high dimension. Annals of Statistics. 2012; 40:436–465.

Bai J, Ng S. Determining the number of factors in approximate factor models. Econometrica. 2002; 70:191–221.

Bai J, Ng S. Principal components estimation and identification of the factors. Journal of Econometrics. 2013; 176:18–29.

Bickel P, Levina E. Covariance regularization by thresholding. Annals of Statistics. 2008; 36:2577–2604.

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

Breitung, Tenhofen. Gls estimation of dynamic factor models. Journal of the American Statistical Association. 2011; 106:1150–1166.

Brillinger, DR. Time series: data analysis and theory. Vol. 36. Siam; 1981.

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

Candès EJ, Recht B. Exact matrix completion via convex optimization. Foundations of Computational Mathematics. 2009; 9:717–772.

Chen, X. Handbook of Econometrics, North Holland. 2007. Large sample sieve estimation of semi-nonparametric models; p. 76

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

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

Desai KH, Storey JD. Cross-dimensional inference of dependent high-dimensional data. Journal of the American Statistical Association. 2012; 107:135–151.

Efron B. Correlated z-values and the accuracy of large-scale statistical estimates. Journal of the American Statistical Association. 2010; 105:1042–1055. [PubMed: 21052523]

Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary covariance dependence (with discussion). Journal of the American Statistical Association. 2012; 107:1019–1035. [PubMed: 24729644]

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

Fan J, Liao Y, Shi X. Risks of large portfolios. Journal of Econometrics. 2015a; 186:367–387. [PubMed: 26195851]

Fan J, Liao Y, Wang W. Supplementary appendix to the paper "projected principal component analysis in factor models". 2015b

Forni M, Hallin M, Lippi M, Reichlin L. The generalized dynamic-factor model: Identification and estimation. Review of Economics and statistics. 2000; 82:540–554.

Forni M, Hallin M, Lippi M, Zaffaroni P. Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. Journal of econometrics. 2015; 185:359–371.

Forni M, Lippi M. The generalized dynamic factor model: representation theory. Econometric theory. 2001; 17:1113–1141.

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

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

Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Annals of statistics. 2001:295–327.

Jung S, Marron J. Pca consistency in high dimension, low sample size context. Annals of Statistics. 2009; 37:3715–4312.

Koltchinskii V, Lounici K, Tsybakov AB. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics. 2011; 39:2302–2329.

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

Leek JT, Storey JD. A general framework for multiple testing dependence. Proceedings of the National Academy of Sciences. 2008; 105:18718–18723.

Li G, Yang D, Nobel AB, Shen H. Supervised singular value decomposition and its asymptotic properties. Journal of Multivariate Analysis. 2015

Lorentz, G. Approximation of functions. 2. American Mathematical Society; Rhode Island: 1986.

Ma Z. Sparse principal component analysis and iterative thresholding. The Annals of Statistics. 2013; 41:772–801.

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

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

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

Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica. 2007; 17:1617.

Shen D, Shen H, Marron J. Consistency of sparse pca in high dimension, low sample size contexts. Journal of Multivariate Analysis. 2013a; 115:317–333.

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

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

Vershynin R. Introduction to the non-asymptotic analysis of random matrices. 2010 arXiv preprint arXiv:1011.3027.

APPENDIX A: PROOFS FOR SECTION 3

Throughout the proofs, $p \to \infty$, and T may either grow simultaneously with p or stay constant. For two matrices **A**, **B** with fixed dimensions, and a sequence a_T, by writing **A** = **B** + $o_P(a_T)$, we mean $\|\mathbf{A} - \mathbf{B}\|_F = o_P(a_T)$.

In the regular factor model $\mathbf{Y} = \mathbf{\Lambda}\mathbf{F}' + \mathbf{U}$, let **K** denote a $K\times K$ diagonal matrix of the first K eigenvalues of $\frac{1}{T_p}\mathbf{Y}'\mathbf{P}\mathbf{Y}$. Then by definition, $\frac{1}{T_p}\mathbf{Y}'\mathbf{P}\mathbf{Y}\hat{\mathbf{F}} = \hat{\mathbf{F}}\mathbf{K}$. Let $\mathbf{M} = \frac{1}{T_p}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}\mathbf{F}'\hat{\mathbf{F}}\mathbf{K}^{-1}$. Then

$$\hat{\mathbf{F}} - \mathbf{F}\mathbf{M} = \sum_{i=1}^{3}\mathbf{D}_i\mathbf{K}^{-1}, \quad \text{(A.1)}$$

where

$$\mathbf{D}_1 = \frac{1}{T_p}\mathbf{F}\mathbf{\Lambda}'\mathbf{P}\mathbf{U}\hat{\mathbf{F}}, \quad \mathbf{D}_2 = \frac{1}{T_p}\mathbf{U}'\mathbf{P}\mathbf{U}\hat{\mathbf{F}}, \quad \mathbf{D}_3 = \frac{1}{T_p}\mathbf{U}'\mathbf{P}\mathbf{\Lambda}\mathbf{F}'\hat{\mathbf{F}}.$$

We now describe the structure of the proofs for

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

Note that $\hat{\mathbf{F}} - \mathbf{F} = \hat{\mathbf{F}} - \mathbf{F}\mathbf{M} + \mathbf{F}(\mathbf{M} - \mathbf{I})$. Hence we need to bound $\frac{1}{T}\|\hat{\mathbf{F}} - \mathbf{F}\mathbf{M}\|_F^2$ and $\frac{1}{T}\|\mathbf{F}(\mathbf{M} - \mathbf{I})\|_F^2$ respectively.

Step 1: prove that $\frac{1}{T}\|\hat{\mathbf{F}} - \mathbf{F}\mathbf{M}\|_F^2 = O_P(J/p)$.

Due to the equality (A.1), it suffices to bound $\|\mathbf{K}^{-1}\|_2$ as well as the $\frac{1}{T}\|.\|_F^2$ norm of $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{D}_3$ respectively. These are obtained in Lemmas A.2, A.3 below.

Step 2: prove that $\frac{1}{T}\|\mathbf{F}'(\hat{\mathbf{F}} - \mathbf{F}\mathbf{M})\|_F = O_P(\sqrt{J/(pT)} + J/p)$.

Still by the equality (A.1), $\frac{1}{T}\|\mathbf{F}'(\hat{\mathbf{F}} - \mathbf{F}\mathbf{M})\|_F \le \frac{1}{T}\|\mathbf{K}^{-1}\|_2\sum_{i=1}^{3}\|\mathbf{F}'\mathbf{D}_i\|_F$. Hence this step is achieved by bounding $\|\mathbf{F}'\mathbf{D}_i\|_F$ for i = 1, 2, 3. Note that in this step, we shall not apply a simple inequality $\|\mathbf{F}'\mathbf{D}_i\|_F \le \|\mathbf{F}\|_F \|\mathbf{D}_i\|_F$, which is too crude. Instead, with the help of the result $\frac{1}{T}\|\hat{\mathbf{F}} - \mathbf{F}\mathbf{M}\|_F^2 = O_p(J/p)$ achieved in Step 1, sharper upper bounds for $\|\mathbf{F}'\mathbf{D}_i\|_F$ can be achieved. We do so in Lemma **??** in the supplementary material.

Step 3: prove that $\|\mathbf{M} - \mathbf{I}\|_F^2 = O_P(J/(pT) + (J/p)^2)$.

This step is achieved in Lemma A.4 below, which uses the result in Step 2.

Before proceeding to Step 1, we first show that the two alternative definitions for $\hat{\mathbf{G}}(\mathbf{X})$ described in Section 2.3 are equivalent.

Lemma A.1 $\frac{1}{T}\mathbf{PY}\hat{\mathbf{F}}=\hat{\mathbf{\Xi}}\hat{\mathbf{D}}^{1/2}$

Proof

Consider the singular value decomposition: $\frac{1}{\sqrt{T}}\mathbf{PY}=\mathbf{V}_1\mathbf{S}\mathbf{V}_2'$, where $\mathbf{V}_1$ is a p×p orthogonal matrix, whose columns are the eigenvectors of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$; $\mathbf{V}_2$ is a $T \times T$ matrix whose columns are the eigenvectors of $\frac{1}{T}\boldsymbol{Y}'\mathbf{PY}$; $\mathbf{S}$ is a $p \times T$ rectangular diagonal matrix, with diagonal entries as the square roots of the non-zero eigenvalues of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$. In addition, by definition, $\hat{\mathbf{D}}$ is a $K \times K$ diagonal matrix consisting of the largest K eigenvalues of $\frac{1}{T}\mathbf{PYY}'\mathbf{P}$; $\hat{\Xi}$ is a $p \times K$ matrix whose columns are the corresponding eigenvectors. The columns of $\hat{\mathbf{F}}/\sqrt{T}$ are the eigenvectors of $\frac{1}{T}\boldsymbol{Y}'\mathbf{PY}$, corresponding to the first K eigenvalues.

With these definitions, we can write $\mathbf{V}_1 = (\hat{\Xi}, \tilde{\mathbf{V}}_1)$, $\mathbf{V}_2=(\hat{\mathbf{F}}/\sqrt{T},\tilde{\mathbf{V}}_2)$, and

$$\mathbf{S}=\begin{pmatrix}\hat{\mathbf{D}}^{1/2} & \mathbf{0}\\ \mathbf{0} & \tilde{\mathbf{D}}\end{pmatrix},\quad \hat{\mathbf{F}}'\tilde{\mathbf{V}}_2=\mathbf{0},\hat{\mathbf{F}}'\hat{\mathbf{F}}/T=\mathbf{I}_K,$$

for some matrices $\tilde{\mathbf{V}}_1$, $\tilde{\mathbf{V}}_2$ and $\tilde{\mathbf{D}}$. It then follows that

$$\frac{1}{T}\mathbf{PY}\hat{\mathbf{F}}=\mathbf{V}_1\mathbf{S}\mathbf{V}_2'\frac{1}{\sqrt{T}}\hat{\mathbf{F}}=(\hat{\mathbf{\Xi}},\tilde{\mathbf{V}}_1)\begin{pmatrix}\hat{\mathbf{D}}^{1/2} & \mathbf{0}\\ \mathbf{0} & \tilde{\mathbf{D}}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{F}}'/\sqrt{T}\\ \tilde{\mathbf{V}}_2'\end{pmatrix}\frac{1}{\sqrt{T}}\hat{\mathbf{F}}=\hat{\mathbf{\Xi}}\hat{\mathbf{D}}^{1/2}.$$

Lemma A.2

$\|\mathbf{K}\|_2 = O_P(1)$, $\|\mathbf{K}^{-1}\|_2 = O_P(1)$, $\|\mathbf{M}\|_2 = O_P(1)$.

Proof

The eigenvalues of $\mathbf{K}$ are the same as those of

$$\mathbf{W}=\frac{1}{Tp}(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\Phi(\mathbf{X})'\mathbf{YY}'\Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2},$$

Substituting $\mathbf{Y} = \mathbf{\Lambda F}' + \mathbf{U}$, and $\mathbf{F}'\mathbf{F}/T = \mathbf{I}_K$, we have $\mathbf{W}=\sum_{i=1}^{4}\mathbf{W}_i$, where

$$\begin{array}{rcl}\mathbf{W}_1 & = & \frac{1}{p}(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\Phi(\mathbf{X})'\mathbf{\Lambda}\mathbf{\Lambda}'\Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}, \\ \mathbf{W}_2 & = & \frac{1}{p}(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\Phi(\mathbf{X})'(\frac{\mathbf{\Lambda}\mathbf{F}'\mathbf{U}'}{T})\Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}, \\ \mathbf{W}_3 & = & \mathbf{W}_2', \\ \mathbf{W}_4 & = & \frac{1}{p}(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\Phi(\mathbf{X})'\frac{\mathbf{U}\mathbf{U}'}{T}\Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}.\end{array}$$

By Assumption 3.3, $\|\Phi(\mathbf{X})\|_2=\lambda_{\max}^{1/2}(\Phi(\mathbf{X})'\Phi(\mathbf{X}))=O_P(\sqrt{p})$,

$$\begin{array}{rcl}\|(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\|_2 & = & \lambda_{\max}^{1/2}((\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1})=\lambda_{\min}^{-1/2}(\frac{1}{p}\Phi(\mathbf{X})'\Phi(\mathbf{X}))p^{-1/2}=O_P(p^{-1/2}) \\ \|\mathbf{P}\mathbf{\Lambda}\|_2 & = & \lambda_{\max}^{1/2}(\frac{1}{p}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda})p^{1/2}=O_P(p^{1/2}).\end{array}$$

Hence

$$\|\mathbf{W}_2\|_2 \le \frac{1}{p}\|(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\|_2^2\|\Phi(\mathbf{X})\|_2\|\mathbf{\Lambda}\|_F\|\frac{1}{T}\mathbf{F}'\mathbf{U}'\Phi(\mathbf{X})\|_F=O_P(\frac{1}{pT})\|\mathbf{F}'\mathbf{U}'\Phi(\mathbf{X})\|_F.$$

By Lemma **??** in the supplementary material, $\|W_2\|_2=O_P(\frac{\sqrt{J}}{\sqrt{pT}})$. Similarly,

$$\|\mathbf{W}_4\|_2 \le \frac{1}{pT}\|(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1/2}\|_2^2\|\Phi(\mathbf{X})'\mathbf{U}\|_F^2=O_P(\frac{1}{p^2T})\|\Phi(\mathbf{X})'\mathbf{U}\|_F^2=O_P(\frac{J}{p}).$$

Using the inequality that for the kth eigenvalue, $|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| \le \|\mathbf{W} - \mathbf{W}_1\|_2$, we have $|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| = O_P(T^{-1/2} + p^{-1})$, for $k = 1, \cdots, K$. Hence it suffices to prove that the first K eigenvalues of $\mathbf{W}_1$ are bounded away from both zero and infinity, which are also the first K eigenvalues of $\frac{1}{p}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}$. This holds under the theorem's assumption (Assumption 3.1). Thus $\|\mathbf{K}^{-1}\|_2 = O_P(1) = \|\mathbf{K}\|_2$, which also implies $\|\mathbf{M}\|_2 = O_P(1)$.

Lemma A.3

(i) $\|\mathbf{D}_1\|_F^2=O_P(TJ/p)$, *(ii)* $\|\mathbf{D}_2\|_F^2=O_P(J/p^2)$, *(iii)* $\|\mathbf{D}_3\|_F^2=O_P(TJ/p)$, *(iv)* $\frac{1}{T}\|\hat{\mathbf{F}}-\mathbf{F}\mathbf{M}\|_F^2=O_P(J/p)$.

Proof

It follows from Lemma **??** in the supplementary material that $\|\mathbf{P}\mathbf{U}\|_F=O_P(\sqrt{TJ})$. Also, $\|\mathbf{F}\|_F^2=O_P(T)=\|\hat{\mathbf{F}}\|_F^2$ and Assumption 3.1 implies $\|\mathbf{P}\mathbf{\Lambda}\|_2^2=O_P(p)$. So

$$\begin{aligned}\|\mathbf{D}_1\|_F^2 &= \|\tfrac{1}{Tp}\mathbf{F}\mathbf{\Lambda}'\mathbf{P}\mathbf{U}\hat{\mathbf{F}}\|_F^2 \le \tfrac{1}{T^2p^2}\|\mathbf{F}\|_F^2\|\hat{\mathbf{F}}\|_F^2\|\mathbf{P}\mathbf{\Lambda}\|_2^2\|\mathbf{P}\mathbf{U}\|_F^2 = O_P(TJ/p)\\ \|\mathbf{D}_2\|_F^2 &= \|\tfrac{1}{Tp}\mathbf{U}'\mathbf{P}\mathbf{U}\hat{\mathbf{F}}\|_F^2 \le \tfrac{1}{T^2p^2}\|\mathbf{P}\mathbf{U}\|_F^2\|\hat{\mathbf{F}}\|_F^2 = O_P(J/p^2)\\ \|\mathbf{D}_3\|_F^2 &= \|\tfrac{1}{Tp}\mathbf{U}'\mathbf{P}\mathbf{\Lambda}\mathbf{F}'\hat{\mathbf{F}}\|_F^2 \le \tfrac{1}{T^2p^2}\|\mathbf{P}\mathbf{U}\|_F^2\|\mathbf{P}\mathbf{\Lambda}\|_2^2\|\mathbf{F}\|_F^2\|\hat{\mathbf{F}}\|_F^2 = O_P(TJ/p).\end{aligned}$$

By Lemma A.2, $\|\mathbf{K}^{-1}\|_2 = O_P(1)$. Part (iv) then follows directly from

$$\frac{1}{T}\|\hat{\mathbf{F}}-\mathbf{F}\mathbf{M}\|_F^2 \le O_P(\frac{1}{T}\|\mathbf{K}^{-1}\|_2)(\|\mathbf{D}_1\|_F^2+\|\mathbf{D}_2\|_F^2+\|\mathbf{D}_3\|_F^2).$$

Lemma A.4

In the regular factor model $\|\mathbf{M}-\mathbf{I}\|_F = O_P(\sqrt{J/(pT)}+J/p)$.

Proof

By Lemma **??** in the supplementary material and the triangular inequality, $\|\frac{1}{T}(\hat{\mathbf{F}}-\mathbf{F}\mathbf{M})'\mathbf{F}\| = O_P(\sqrt{J/(pT)}+J/p)$. Hence

$$\hat{\mathbf{F}}'\mathbf{F}/T = \mathbf{M}' + \frac{1}{T}(\hat{\mathbf{F}}-\mathbf{F}\mathbf{M})'\mathbf{F} = \mathbf{M}' + O_P(\sqrt{J/(pT)}+J/p)$$

Right multiplying **M** to both sides $\hat{\mathbf{F}}'\mathbf{F}\mathbf{M}/T = \mathbf{M}'\mathbf{M} + O_P(\sqrt{J/(pT)}+J/p)$. In addition,

$$\|\hat{\mathbf{F}}'(\hat{\mathbf{F}}-\mathbf{F}\mathbf{M})/T\|_F \le \frac{1}{T}\|\hat{\mathbf{F}}-\mathbf{F}\mathbf{M}\|_F^2 + \|\mathbf{F}'(\hat{\mathbf{F}}-\mathbf{F}\mathbf{M})/T\|_F = O_P(\sqrt{J/(pT)}+J/p).$$

Hence

$$\mathbf{I} = \mathbf{M}'\mathbf{M} + O_P(\sqrt{J/(pT)}+J/p).$$

In addition, from $\mathbf{M} = \frac{1}{Tp}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}\mathbf{F}'\hat{\mathbf{F}}\mathbf{K}^{-1} = \frac{1}{p}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}\mathbf{M}\mathbf{K}^{-1} + O_P(\sqrt{J/(pT)}+J/p)$,

$$\mathbf{M}\mathbf{K} = \frac{1}{p}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}\mathbf{M} + O_P(\sqrt{J/(pT)}+J/p).$$

Because $\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}$ is diagonal, the same proofs of those of Proposition **??** lead to the desired result.

Proof of Theorem 3.1

Proof

It follows from Lemmas A.3 (iv) and A.4 that

$$\frac{1}{T}\|\hat{\mathbf{F}}-\mathbf{F}\|_F^2 \le \frac{2}{T}\|\hat{\mathbf{F}}-\mathbf{F}\mathbf{M}\|_F^2+2\|\mathbf{M}-\mathbf{I}\|_F^2=O_p(\frac{J}{p}).$$

As for the estimated loading matrix, note that

$$\hat{\mathbf{G}}(\mathbf{X})=\frac{1}{T}\mathbf{P}\mathbf{Y}\hat{\mathbf{F}}=\frac{1}{T}\mathbf{P}\mathbf{\Lambda}\mathbf{F}'\hat{\mathbf{F}}+\frac{1}{T}\mathbf{P}\mathbf{U}\hat{\mathbf{F}}=\mathbf{P}\mathbf{\Lambda}+\mathbf{E},$$

where $\mathbf{E}=\frac{1}{T}\mathbf{P}\mathbf{\Lambda}\mathbf{F}'(\hat{\mathbf{F}}-\mathbf{F})+\frac{1}{T}\mathbf{P}\mathbf{U}(\hat{\mathbf{F}}-\mathbf{F})+\frac{1}{T}\mathbf{P}\mathbf{U}\mathbf{F}\cdot$

By Lemmas **??** and A.4,

$$\|\frac{1}{T}\mathbf{P}\mathbf{\Lambda}\mathbf{F}'(\hat{\mathbf{F}}-\mathbf{F})\|_F \le O_P\left(\frac{\sqrt{p}}{T}\right)\|\mathbf{F}'(\hat{\mathbf{F}}-\mathbf{F}\mathbf{M})\|_F+O_P(\sqrt{p})\|\mathbf{M}-\mathbf{I}\|_F=O_P\left(\sqrt{\frac{J}{T}}+\frac{J}{\sqrt{p}}\right).$$

By Lemma **??**, $\|\frac{1}{T}\mathbf{P}\mathbf{U}(\hat{\mathbf{F}}-\mathbf{F})\|_F \le \frac{1}{T}\|\mathbf{P}\mathbf{U}\|_2\|\hat{\mathbf{F}}-\mathbf{F}\|_F=O_P(\frac{J}{\sqrt{p}})$, and from Lemma **??** $\|\frac{1}{T}\mathbf{P}\mathbf{U}\mathbf{F}\|_F=O_P(\sqrt{\frac{J}{T}})$. Hence $\|\mathbf{E}\|_F=O_P(\sqrt{\frac{J}{T}}+\frac{J}{\sqrt{p}})$, which implies

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

Proof of Theorem 3.2

Proof

Since $\hat{\mathbf{G}}(\mathbf{X}) = \mathbf{\Xi}\hat{\mathbf{D}}^{\hat{1}/2}$, by Theorem 3.1, $\frac{1}{p}\|\hat{\mathbf{\Xi}}\hat{\mathbf{D}}^{1/2}-\mathbf{P}\mathbf{\Lambda}\|_F^2=O_P(\frac{J}{pT}+\frac{J^2}{p^2})$. Hence $\frac{1}{p}\|\hat{\mathbf{\Xi}}\hat{\mathbf{D}}^{1/2}-\mathbf{\Lambda}\|_F^2=O_P(\frac{J}{pT}+\frac{J^2}{p^2})+\frac{1}{p}\|\mathbf{P}\mathbf{\Lambda}-\mathbf{\Lambda}\|_F^2$. By lemma A.2, $\|(\widehat{\mathbf{D}/p})^{-1}\|_2 = \|\mathbf{K}^{-1}\|_2 = O_P(1)$, which implies

$$\|\hat{\mathbf{\Xi}}-\mathbf{\Lambda}\hat{\mathbf{D}}^{-1/2}\|_F^2=O_P\left(\frac{J}{pT}+\frac{J^2}{p^2}+\frac{1}{p}\|\mathbf{P}\mathbf{\Lambda}-\mathbf{\Lambda}\|_F^2\right).$$

On the other hand, define $\bar{\Lambda} = \Lambda\tilde{\mathbf{V}} = (\bar{\Lambda}_1, \ldots \bar{\Lambda}_K)$. Then $\bar{\Lambda}'\bar{\Lambda}$ is diagonal and $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\overline{\boldsymbol{\Lambda}}_j = \overline{\boldsymbol{\Lambda}}_j \|\overline{\boldsymbol{\Lambda}}_j\|_2^2$, $j = 1, \ldots, K$. This implies that the columns of $\Lambda(\bar{\Lambda}'\bar{\Lambda})^{-1/2}$ are the eigenvectors of $\Lambda\Lambda'$ corresponding to the largest K eigenvalues. In addition, in the factor model, we have the following matrix decomposition: for $\Sigma_u = \mathrm{cov}(\mathbf{u}_t)$, $\Sigma = \Lambda\Lambda' + \Sigma_u$. Hence by the same argument of the proof of Proposition 2.2 in Fan et al. (2013),

$$\|\boldsymbol{\Xi} - \boldsymbol{\Lambda}\tilde{\mathbf{V}}(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{-1/2}\|_F = O\left(\frac{1}{p}\|\sum\nolimits_u\|_2\right).$$

Using $\tilde{\mathbf{V}}\tilde{\mathbf{V}}' = \mathbf{I}$, we have

$$\begin{aligned}\|\hat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}\tilde{\mathbf{V}}'(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{1/2}\hat{\mathbf{D}}^{-1/2}\|_F \leq\ & \|\hat{\boldsymbol{\Xi}} - \boldsymbol{\Lambda}\hat{\mathbf{D}}^{-1/2}\|_F + \|\boldsymbol{\Lambda}\hat{\mathbf{D}}^{-1/2} - \boldsymbol{\Xi}(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{1/2}\tilde{\mathbf{V}}'\hat{\mathbf{D}}^{-1/2}\|_F + \|\boldsymbol{\Xi}(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{1/2}\tilde{\mathbf{V}}'\hat{\mathbf{D}}^{-1/2} - \boldsymbol{\Xi}\tilde{\mathbf{V}}'(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{1/2}\hat{\mathbf{D}}^{-1/2} \\ \leq\ & \|\hat{\boldsymbol{\Xi}} - \boldsymbol{\Lambda}\hat{\mathbf{D}}^{-1/2}\|_F + \|(\boldsymbol{\Lambda}\tilde{\mathbf{V}}(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{-1/2} - \boldsymbol{\Xi})(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{1/2}\tilde{\mathbf{V}}'\hat{\mathbf{D}}^{-1/2}\|_F + \|\boldsymbol{\Xi}((\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{1/2}\tilde{\mathbf{V}}' - \tilde{\mathbf{V}}'(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{1/2})\hat{\mathbf{D}}^{-1/2}\|_F.\end{aligned}$$

On the right hand side, the first term is $O_P\left(\frac{J}{pT} + \frac{J^2}{p^2} + \frac{1}{p}\|\mathbf{P}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\|_F^2\right)$, as is proved above. Still by $\|(p^{-1}\hat{\mathbf{D}})^{-1/2}\|_2 = O_P(1)$, the second term is bounded by

$$\|\boldsymbol{\Lambda}\tilde{\mathbf{V}}(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}})^{-1/2} - \boldsymbol{\Xi}\|_F \|(\overline{\boldsymbol{\Lambda}}'\overline{\boldsymbol{\Lambda}}/p)^{1/2}\|_F \|\tilde{\mathbf{V}}\|_2 \|(\hat{\mathbf{D}}/p)^{-1/2}\|_2 = O(\|\sum\nolimits_u\|_2/p).$$

Finally, since $(\bar{\Lambda}'\bar{\Lambda})^{1/2} = \tilde{\mathbf{V}}'(\Lambda'\Lambda)^{1/2}\tilde{\mathbf{V}}$, so $(\bar{\Lambda}'\bar{\Lambda})^{1/2}\tilde{\mathbf{V}}' - \tilde{\mathbf{V}}'(\Lambda'\Lambda)^{1/2} = 0$, which implies the third term is zero. Hence

$$\|\hat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}\tilde{\mathbf{V}}'(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{1/2}\hat{\mathbf{D}}^{-1/2}\|_F \leq O_P\left(\frac{J}{pT} + \frac{J^2}{p^2} + \frac{1}{p}\|\mathbf{P}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\|_F^2\right) + O\left(\frac{1}{p}\|\sum\nolimits_u\|_2\right).$$

All the remaining proofs are given in the supplementary material.

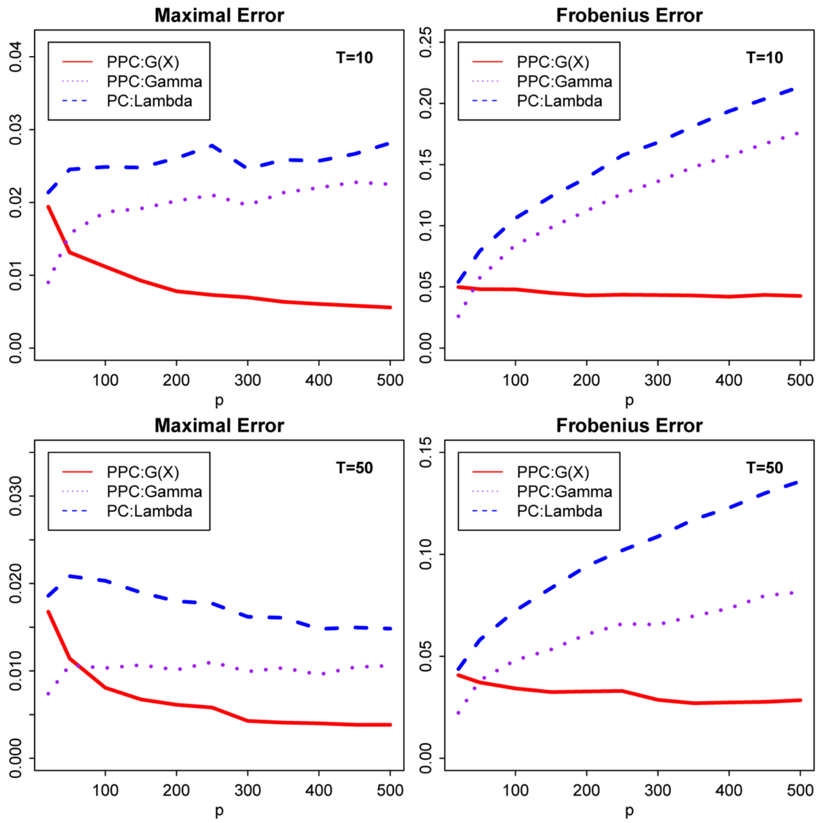


Fig. 1.

Averaged $\|\widehat{\Lambda}-\Lambda\|$ by projected-PCA (PPCA, red solid) and regular PC (dashed blue) and $\|\hat{\mathbf{G}}-\mathbf{G}_0\|$, $\|\widehat{\Gamma}-\Gamma\|$ by PPCA over 500 repetitions. Left panel:$\|.\|_{\max}$, right panel: $\|\cdot\|_F/\sqrt{p}$.

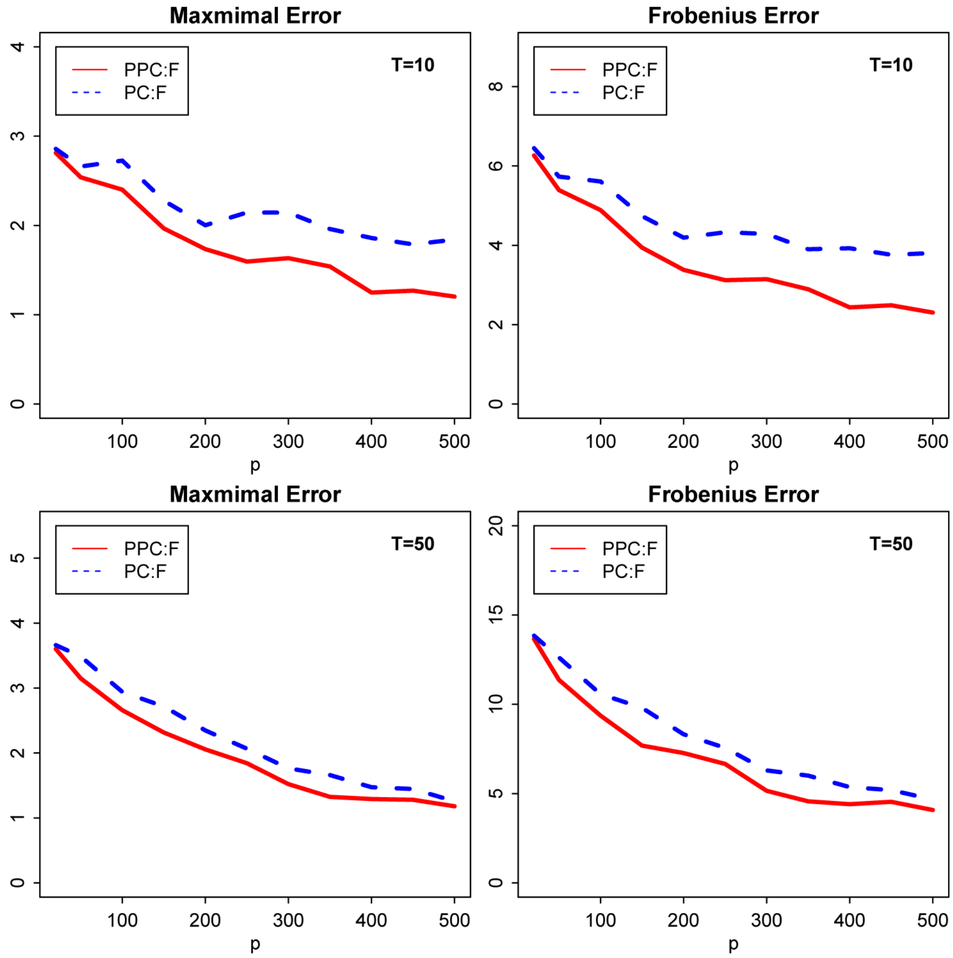


Fig. 2.

Averaged $\|\hat{\mathbf{F}} - \mathbf{F}_0\|_{\max}$ and $\|\hat{\mathbf{F}} - \mathbf{F}_0\|_F / \sqrt{T}$ over 500 repetitions, by projected-PCA (PPCA, solid red) and regular PC (dashed blue).

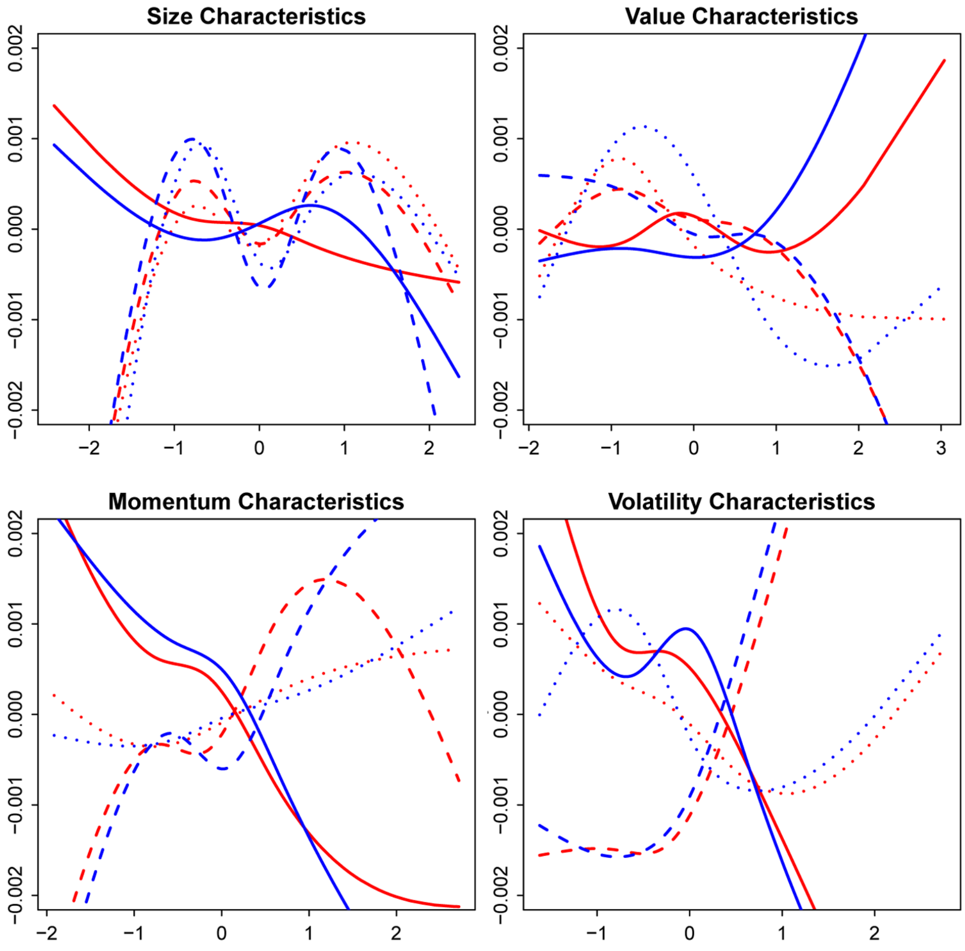


Fig. 3.

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 $T = 50$, $p = 300$.

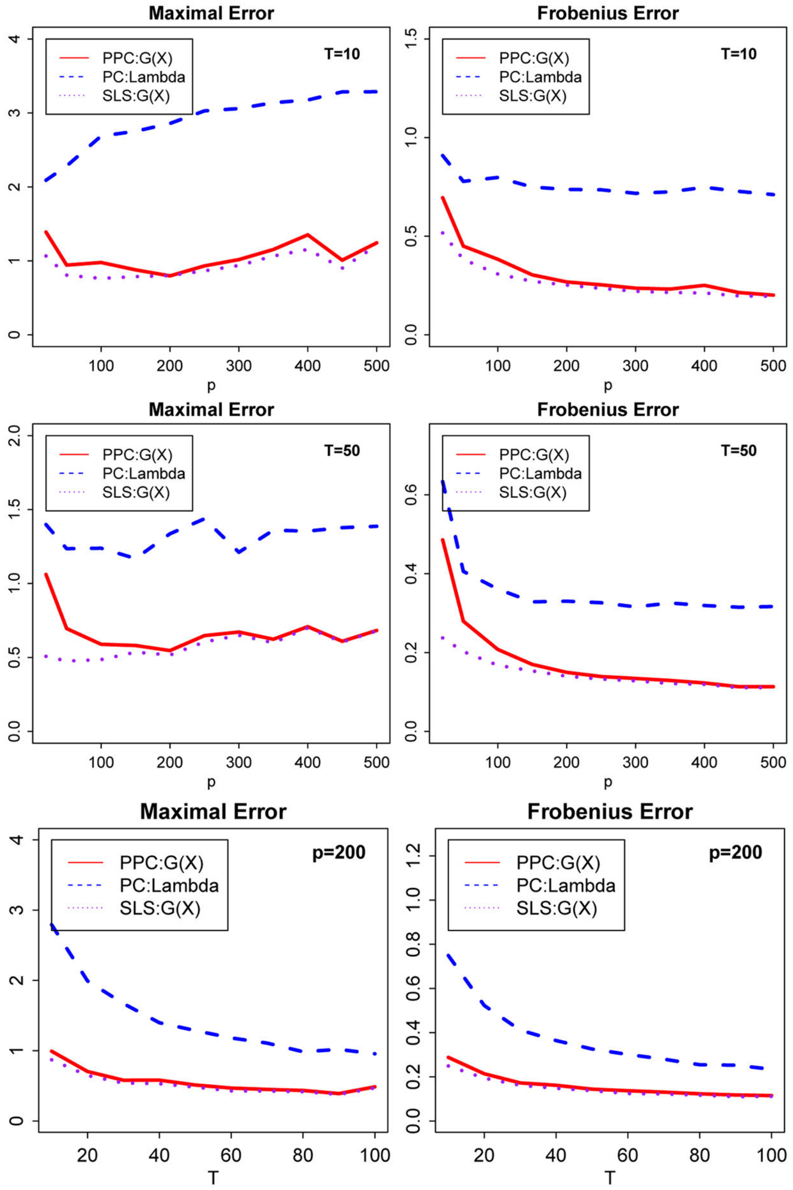


Fig. 4.

Averaged $\|\hat{\mathbf{G}} - \mathbf{G}_0\|_{\max}$ and $\|\hat{\mathbf{G}} - \mathbf{G}_0\|_F / \sqrt{p}$ over 500 repetitions. PPCA, PC and SLS respectively represent projected-PCA, regular PCA and sieve least squares with known factors: Design 2. Here $\boldsymbol{\Gamma} = 0$, so $\boldsymbol{\Lambda} = \mathbf{G}_0$. Upper two panels: p grows with fixed T; bottom panels: T grows with fixed p.

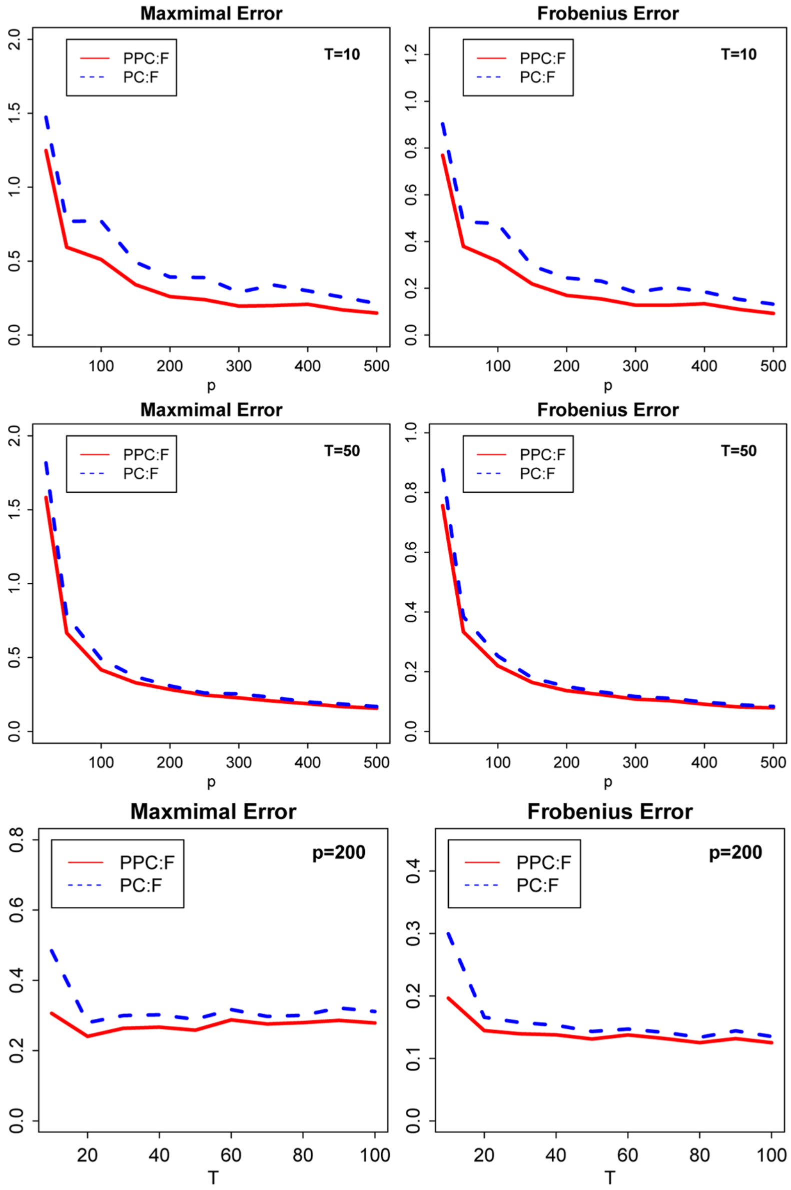


Fig. 5.

Average estimation error of factors over 500 repetitions, i.e. $\|\hat{\mathbf{F}} - \mathbf{F}_0\|_{\max}$ and $\|\hat{\mathbf{F}} - \mathbf{F}_0\|_F / \sqrt{T}$ by projected-PCA (solid red) and PC (dashed blue): Design 2. Upper two panels: *p* grows with fixed *T*; bottom panels: *T* grows with fixed *p*.

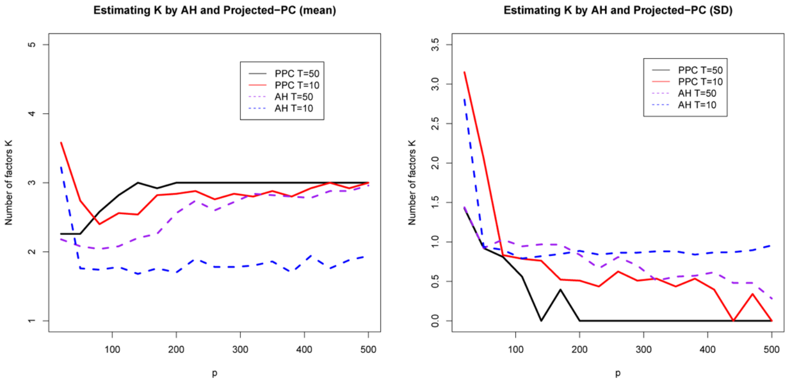
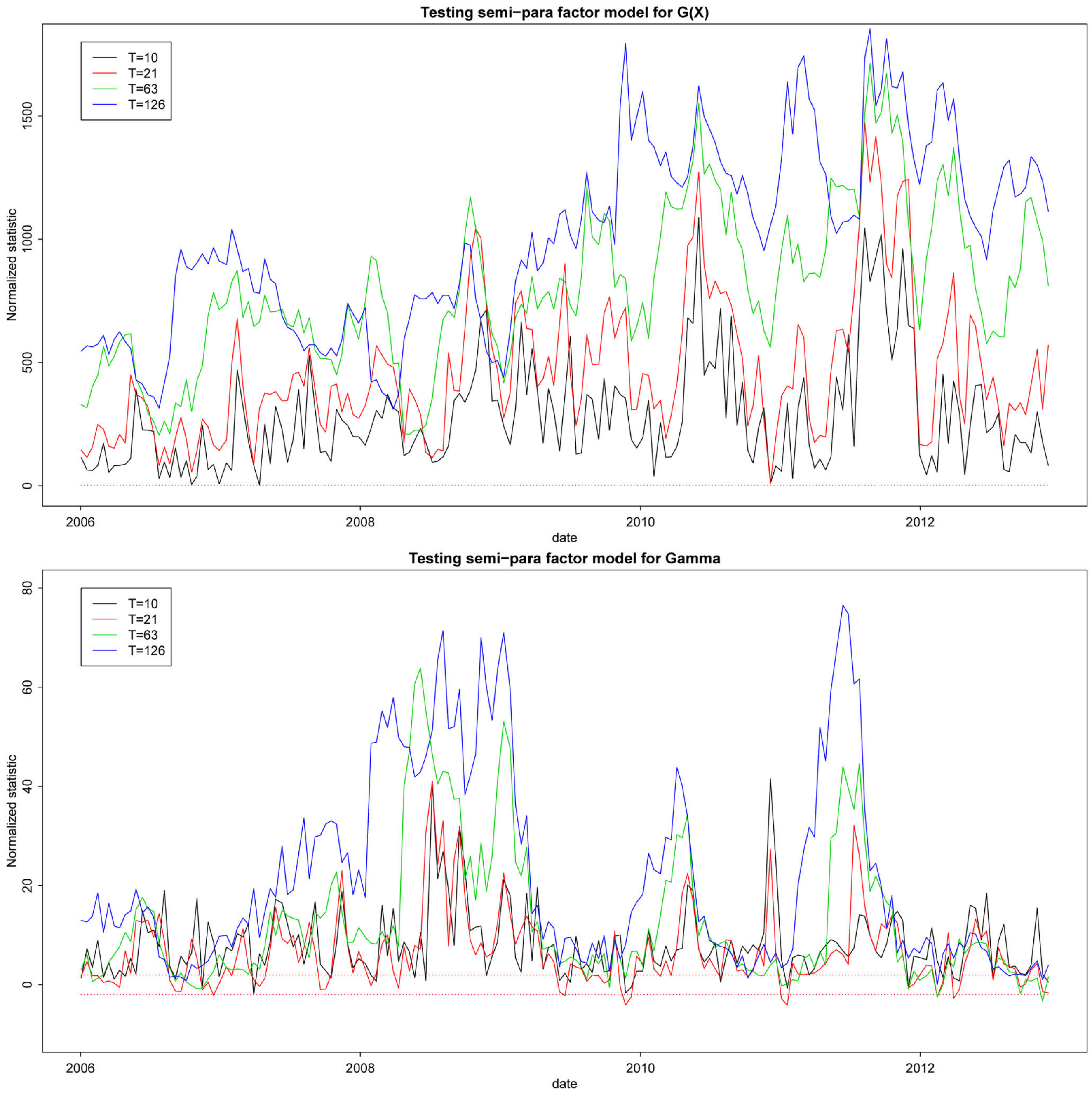


Fig 6.

Mean and standard deviation of the estimated number of factors over 50 repetitions. True *K* = 3. PPCA and AH respectively represent the methods of projected-PCA and Ahn and Horenstein (2013). Left panel: Mean; Right panel: standard deviation.


Testing semi−para factor model for G(X)

Testing semi−para factor model for Gamma

Fig 7.

Normalized S_G, S_γ from 2006/01/03 to 2012/11/30. The dotted lines are ± 1.96.

Table 1

Parameters used for the factor generating process, obtained by calibration to the real data.

Σ_ε			**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