Augmented Factor Models with Applications to Validating Market Risk Factors and Forecasting Bond Risk Premia

Jianqing Fan[+], Yuan Ke[†] and Yuan Liao[‡]
[+]Princeton University, [†]University of Georgia, [‡] Rutgers University

April 1, 2019

Abstract

We study factor models augmented by observed covariates that have explanatory powers on the unknown factors. In financial factor models, the unknown factors can be reasonably well explained by a few observable proxies, such as the Fama-French factors. In diffusion index forecasts, identified factors are strongly related to several directly measurable economic variables such as consumption-wealth variable, financial ratios, and term spread. With those covariates, both the factors and loadings are identifiable up to a rotation matrix even only with a finite dimension. To incorporate the explanatory power of these covariates, we propose a smoothed or projected principal component analysis (PCA): (i) regress the data onto the observed covariates, and (ii) take the principal components of the fitted data to estimate the loadings and factors. This allows us to more accurately estimate the percentage of both explained and unexplained components in factors and thus to assess the explanatory power of covariates. We show that both the estimated factors and loadings can be estimated with improved rates of convergence compared to the benchmark method. The degree of improvement depends on the strength of the signals, representing the explanatory power of the covariates on the factors. The proposed estimator is robust to possibly heavy-tailed distributions. We apply the model to forecast US bond risk premia, and find that the observed macroeconomic characteristics contain strong explanatory powers of the factors. The gain of forecast is more substantial when the characteristics are incorporated to estimate the common factors than directly used for forecasts.

Keywords: Heavy tails, Forecasts; Principal components; identification.
JEL classification: C58, C38

1 Introduction

In this paper, we study the identification and estimations of factor models augmented by a set of additional covariates that are common to all individuals. Consider the following factor model:

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, \cdots, T. \tag{1.1}$$

Here $\mathbf{y}_t = (y_{1t}, ..., y_{Nt})'$ is the multivariate outcome for the t^{th} observation in the sample; $\mathbf{f}_t$ is the K-dimensional vector of latent factors; $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N)'$ is an $N \times K$ matrix of nonrandom factor loadings; $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})'$ denotes the vector of idiosyncratic errors. In addition to $\{\mathbf{y}_t\}_{t=1}^T$, we also observe variables, denoted by $\mathbf{x}_t$, that have some explanatory power on the unknown factors and hence impact on observed vector $\mathbf{y}_t$. We model $\mathbf{f}_t$ by using the model

$$\mathbf{f}_t = \mathbf{g}(\mathbf{x}_t) + \boldsymbol{\gamma}_t, \tag{1.2}$$

for some (nonparametric) function $\mathbf{g} = E(\mathbf{f}_t|\mathbf{x}_t)$. Here $\mathbf{g}(\mathbf{x}_t)$ is interpreted as the component of the factors that can be explained by the covariates, and $\boldsymbol{\gamma}_t$ is the components that cannot be explained by the covariates. We aim to provide an improved estimation procedure when the factors can be partially explained by several observed variables $\mathbf{x}_t$. In addition, by accurately estimating $\boldsymbol{\gamma}_t$, we can estimate the percentage of both explained and unexplained components in the factors, which describes the proxy/explanatory power of covariates.

Note that model (1.1) implies:

$$\text{cov}(\mathbf{y}_t) = \mathbf{\Lambda}\,\text{cov}(\mathbf{f}_t)\mathbf{\Lambda}' + \text{cov}(\mathbf{u}_t), \tag{1.3}$$

where $\text{cov}(\mathbf{y}_t)$ and $\text{cov}(\mathbf{u}_t)$ respectively denote the $N \times N$ variance-covariance matrices of $\mathbf{y}_t$ and $\mathbf{u}_t$; $\text{cov}(\mathbf{f}_t)$ denotes the $K \times K$ variance-covariance matrix of $\mathbf{f}_t$. Under usual factor models without covariates, $\frac{1}{\sqrt{N}}\mathbf{\Lambda}$ is identified *asymptotically* as the first K eigenvectors of $\text{cov}(\mathbf{y}_t)$ as $N \to \infty$ and can be estimated using the first K eigenvectors of the sample covariance matrix of $\mathbf{y}_t$ (e.g,, Stock and Watson (2002); Bai (2003)).

With additional covariates, on the other hand, *exact* identification can be achieved through covariance of the "smoothed data". By (1.1), assuming exogeneity of $\mathbf{x}_t$, we have $E(\mathbf{y}_t|\mathbf{x}_t) = \mathbf{\Lambda}E(\mathbf{f}_t|\mathbf{x}_t)$ so that it becomes a "noiseless" factor model with smoothed data $E(\mathbf{y}_t|\mathbf{x}_t)$ as the input and $E(\mathbf{f}_t|\mathbf{x}_t)$ as latent factors. The factor loadings and latent factors can be extracted from

$$\mathbf{\Sigma}_{y|x} = E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}. \tag{1.4}$$

It is easy to see from the model that

$$\mathbf{\Sigma}_{y|x} = \mathbf{\Lambda}\mathbf{\Sigma}_{f|x}\mathbf{\Lambda}', \tag{1.5}$$

where $\mathbf{\Sigma}_{f|x} = E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ is a $K \times K$ low-dimensional positive definite matrix. This decomposition is to be compared with (1.3), but the noise covariance $\text{cov}(\mathbf{u}_t)$ is removed. Therefore, as long as $\mathbf{\Sigma}_{f|x}$ is of full rank, $\mathbf{\Lambda}$ falls in the eigenspace generated by $\mathbf{\Sigma}_{y|x}$. In other words, $\mathbf{\Lambda}$ is exactly identifiable up to an orthogonal transformation. Because of such exact identification, we allow N to be finite as a special case. The number of factors is assumed to be known throughout the paper. In practice, K can be consistently estimated by many methods such as AIC, BIC-based criteria, or eigenvalue-ratio methods studied in Lam and Yao (2012); Ahn and Horenstein (2013).

The above discussion prompts us the following new method to estimate the factor loadings $\mathbf{\Lambda}$ that incorporates the explanatory power of $\mathbf{x}_t$: (See Section 3 for details of estimators)

(i) (robustly) regress $\{\mathbf{y}_t\}$ on $\{\mathbf{x}_t\}$ and obtain fitted value $\{\widehat{\mathbf{y}}_t\}$;

(ii) conduct the principal components analysis (PCA) on the fitted data $(\widehat{\mathbf{y}}_1, ..., \widehat{\mathbf{y}}_T)$ to estimate the factor loadings.

We employ a regression based on Huber (1964)'s robust M-estimation in step (i). The procedure involves a diverging truncation parameter, called adaptive Huber loss, to reduce the bias when the error distribution is asymmetric (Fan et al., 2017). This allows our procedure to be applicable to data with heavy tails.[1] There are two important quantities that determine the rates of convergence for the estimators: the "signal" $\mathbf{\Sigma}_{f|x} = E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ and the "noise" $\text{cov}(\boldsymbol{\gamma}_t)$. The rates of convergence are presented using these two quantities.

Under model (1.2), we can test $\boldsymbol{\gamma}_t = 0$ almost surely in the entire sampling period, under which the observed $\mathbf{x}_t$ fully explains the true factors. This is the same as testing

$$H_0 : \text{cov}(\boldsymbol{\gamma}_t) = 0.$$

While it is well known that the commonly used Fama-French factors have explanatory power for most of the variations of stock returns, it is questionable whether they fully explain the true (yet unknown) factors. These observed proxies are nevertheless used as the factors empirically, and the remaining components ($\boldsymbol{\gamma}_t$ and $\mathbf{u}_t$) have all been mistakenly regarded as the idiosyncratic components. The proposed test provides a diagnostic tool

[1] In this paper, by "heavy-tail" we mean tail distributions of $(\mathbf{u}_t, \mathbf{y}_t)$ that are heavier than the usual requirements on the high-dimensional factor model (which are either exponentially-tailed or have eighth or higher moments). But we do not allow large outliers on the covariates.

for the specification of common factors in empirical studies, and is different from the "efficiency test" in the financial econometric literature (e.g., Gibbons et al. (1989); Pesaran and Yamagata (2012); Gungor and Luger (2013); Fan et al. (2015)). While the efficiency test aims to test the asset pricing model through whether the alphas are zero for the specified factors, a rejection could be due to either mispecified factors or the existence of outperforming (underperforming) assets. In contrast, here we directly test whether the factor proxies are correctly specified. We test the specification of Fama French factors for the returns of S&P 500 constituents using rolling windows. We find that the null hypothesis is more often to be rejected using the daily data compared to the monthly data, due to a larger volatility of the unexplained factor components. The estimated overall volatility of factors varies over time and drops significantly during the acceptance period.

1.1 Further Literature

In empirical applications, researchers frequently encounter additional observable covariates that help explain the latent factors. In genomic studies, in the study of breast cancer data such as the Cancer Genome Atlas (TCGA) project (Network, 2012), there are additional information of cancer subtype for each sample. These cancer subtypes can be regarded as a partial driver of the factors for gene expression data. In financial time series forecasts, researchers often collect additional variables that characterize financial markets. The Fama-French factors are well-known to be related to the factors that drive financial returns (Fama and French, 1992).

Most existing works simply treat $\mathbf{x}_t$ as a set of additional regressors in (1.1), or additional outcomes combined with $\mathbf{y}_t$. This approach does not take advantage of the difference of observed variables (e.g. aggregated versus disaggregated macroeconomic variables; gene expressions versus clinical information) and the explanatory power of the covariates on the common factors, and hence does not lead to improved rates of convergence even if the signal is strong. The most related work is Li et al. (2016), who specified $\mathbf{f}_t$ as a linear function of $\mathbf{x}_t$. Also, Huang and Lee (2010) proposed to use the estimated $\mathbf{g}(\mathbf{x}_t)$ to forecast. Moreover, our expansion for $\mathbf{\Sigma}_{y|x}$ is also connected to the literature on asymptotic Bahadur-type representations for robust M-estimators, see, for example, Portnoy (1985), Mammen (1989), among others.

The "asymptotic identification" was described perhaps first by Chamberlain and Rothschild (1983). In addition, there has been a large literature on both the static and dynamic factor models, and we refer to Lawley and Maxwell (1971); Forni et al. (2005); Stock and Watson (2002); Bai and Ng (2002); Bai (2003); Doz et al. (2012); Onatski (2012a); Fan

et al. (2013), among many others.

The rest of the paper is organized as follows. Section 2 establishes the new identification of factor models. Section 3 formally defines our estimators and discusses possible alternatives. Section 4 presents the rates of convergence. Section 5 discusses the problem of testing the explanatory power. Section 6 applies the model to forecasting the excess return of US government bonds. We present the extensive simulation studies in Section 7 Finally Section 8 concludes. The main body of the proofs are given in the appendix, while the technical lemmas are referred to the supplementary material.

Throughout the paper, we use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote the minimum and maximum eigenvalues of a matrix $\mathbf{A}$. We define $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$. For two sequences, we write $a_T \gg b_T$ or $b_T \ll a_T$ if $b_T = o(a_T)$ and $a_T \asymp b_T$ if $a_T = O(b_T)$ and $b_T = O(a_T)$.

2 Identification of the covariate-based factor models

2.1 Identification

Suppose that there is a fixed d-dimensional observable vector $\mathbf{x}_t$ that satisfies $E(\mathbf{f}_t|\mathbf{x}_t) \neq 0$ (associated with the latent factors) and $E(\mathbf{u}_t|\mathbf{x}_t) = 0$ (idiosyncratic term impredicable by $\mathbf{x}_t$). Taking the conditional expectation on both sides of (1.1), we have

$$E(\mathbf{y}_t|\mathbf{x}_t) = \mathbf{\Lambda} E(\mathbf{f}_t|\mathbf{x}_t), \tag{2.1}$$

This implies

$$\mathbf{\Sigma}_{y|x} = \mathbf{\Lambda}\mathbf{\Sigma}_{f|x}\mathbf{\Lambda}', \tag{2.2}$$

where

$$\mathbf{\Sigma}_{y|x} := E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}, \qquad \mathbf{\Sigma}_{f|x} := E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}.$$

Note that $E(\mathbf{y}_t|\mathbf{x}_t)$ is identified by the data generating process with observables $\{(\mathbf{y}_t, \mathbf{x}_t)\}_{t\leq T}$. Since $N > K$, (2.2) implies that $\mathbf{\Sigma}_{y|x}$ is a low-rank matrix, whose rank is at most K. Furthermore, we assume $\mathbf{\Sigma}_{f|x}$ is also full rank, so $\mathbf{\Sigma}_{y|x}$ has exactly K nonzero eigenvalues.

To see how the equality (2.2) helps achieve the identification of $\mathbf{\Lambda}$ and $\mathbf{g}(\mathbf{x}_t)$, for the moment, suppose the following normalization holds:

$$\frac{1}{N}\mathbf{\Lambda}'\mathbf{\Lambda} = \mathbf{I}_K, \quad \mathbf{\Sigma}_{f|x} \text{ is a diagonal matrix.} \tag{2.3}$$

Then right multiplying (2.2) by $\mathbf{\Lambda}/N$, by the normalization condition,

$$\frac{1}{N}\mathbf{\Sigma}_{y|x}\mathbf{\Lambda} = \mathbf{\Lambda}\mathbf{\Sigma}_{f|x}.$$

We see that the (K) columns of $\frac{1}{\sqrt{N}}\mathbf{\Lambda}$ are the eigenvectors of $\mathbf{\Sigma}_{y|x}$, corresponding to its K nonzero eigenvalues, which also equal to the diagonal entries of $N\mathbf{\Sigma}_{f|x}$. Furthermore, left multiplying $\mathbf{\Lambda}'/N$ on both sides of (2.1), one can see that even if $\mathbf{f}_t$ is not observable, $E(\mathbf{f}_t|\mathbf{x}_t)$ is also identified as:

$$\mathbf{g}(\mathbf{x}_t) := E(\mathbf{f}_t|\mathbf{x}_t) = \frac{1}{N}\mathbf{\Lambda}'E(\mathbf{y}_t|\mathbf{x}_t).$$

The normalization (2.3) above is useful but is not required in this paper. Without them we show that $\mathbf{\Lambda}$ and $\mathbf{g}(\mathbf{x}_t)$ can be identified up to a rotation matrix transformation.

Let

$$\mathbf{\Sigma}_{\Lambda,N} \;:=\; \mathbf{\Lambda}'\mathbf{\Lambda}/N, \qquad\qquad \chi_N := \lambda_{\min}(E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}).$$

Assumption 2.1. *Suppose $\{\mathbf{f}_t, \mathbf{x}_t, \mathbf{u}_t\}_{t\leq T}$ are identically distributed. Assume:*

(i) Rank condition: $\chi_N > 0$.

(ii) There are positive constants $\underline{c}_\Lambda, \bar{c}_\Lambda > 0$, so that all the eigenvalues of the $K \times K$ matrix $\mathbf{\Sigma}_{\Lambda,N}$ are confined in $[\underline{c}_\Lambda, \bar{c}_\Lambda]$, regardless of whether $N \to \infty$ or not.

Condition (i) is the key condition on the explanatory power of $\mathbf{x}_t$ on factors, where χ_N represents the "signal strength" of the model. We postpone the discussion of this condition after Theorem 2.1. Condition (ii) in Assumption 2.1 can be weakened to allow the eigenvalues of $\mathbf{\Sigma}_{\Lambda,N}$ to slowly decay to zero. While doing so allows some of the factors to be *weak*, it does not provide any new statistical insights, but would bring unnecessary complications to our results and conditions. Therefore, we maintain the strong version as condition (ii).

Generally, we have the following theorem for identifying $(\mathbf{\Lambda}, \mathbf{g}(\mathbf{x}_t))$ (up to a rotation transformation).

Theorem 2.1. *Suppose $E(\mathbf{u}_t|\mathbf{x}_t) = 0$, Assumption 2.1 holds and $N > K$. Then there is an invertible $K \times K$ matrix $\mathbf{H}$ so that:*

(i) The columns of $\mathbf{\Lambda H}$ are the eigenvectors of $\mathbf{\Sigma}_{y|x}$ corresponding to the nonzero distinct eigenvalues.

(ii) Given $\mathbf{\Lambda H}$, $\mathbf{g}(\mathbf{x}_t) := E(\mathbf{f}_t|\mathbf{x}_t)$ *satisfies:*

$$\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t) = [(\mathbf{\Lambda H})'\mathbf{\Lambda H}]^{-1}\mathbf{\Lambda H}'E(\mathbf{y}_t|\mathbf{x}_t).$$

(iii) Let $\lambda_K(\mathbf{\Sigma}_{y|x})$ *denote the* K*th largest eigenvalue of* $\mathbf{\Sigma}_{y|x}$*, we have*

$$\lambda_K(\mathbf{\Sigma}_{y|x}) \geq N\chi_N \underline{c}_\Lambda.$$

where χ_N *and* $\underline{c}_\Lambda$ *are defined in Assumption 2.1. In addition, under the normalization conditions that* $E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ *is a diagonal matrix and that* $\mathbf{\Sigma}_{\Lambda,N} = \mathbf{I}_K$*, we have* $\mathbf{H} = \mathbf{I}_K$.

2.2 Discussions of Condition (i) of Assumption 2.1

In the model

$$\mathbf{f}_t = \mathbf{g}(\mathbf{x}_t) + \boldsymbol{\gamma}_t, \quad \mathbf{g}(\mathbf{x}_t) = E(\mathbf{f}_t|\mathbf{x}_t),$$

$\chi_N = \lambda_{\min}(\mathbf{\Sigma}_{f|x})$ represents the "signal" of the covariate model. We require $\chi_N > 0$ so that the rank of $\mathbf{\Sigma}_{y|x}$ is K. Only if this condition holds are we able to identify all the K factor loadings using the eigenvectors corresponding to the nonzero eigenvalues. From the estimation point of view, we are using the PCAs of the estimated $\mathbf{\Sigma}_{y|x}$, and can only consistently estimate its rank($\mathbf{\Sigma}_{y|x}$)-number of leading eigenvectors. So this condition is also essential to achieve the consistent estimation of the factor loadings.

Note that requiring $\mathbf{\Sigma}_{f|x}$ be of full rank might be restrictive in some cases. For instance, consider the linear case: $E(\mathbf{f}_t|\mathbf{x}_t) = \boldsymbol{\beta}\mathbf{x}_t$ for a $K \times d$ coefficient matrix $\boldsymbol{\beta}$, also suppose $E\mathbf{x}_t\mathbf{x}_t'$ is of full rank. Then $\mathbf{\Sigma}_{f|x} = \boldsymbol{\beta}E\mathbf{x}_t\mathbf{x}_t'\boldsymbol{\beta}'$, and is full-rank only if $d \geq K$. Thus we implicitly require, for linear models, the number of covariates should be at least as many as the number of latent factors. Note that if $E(\mathbf{f}_t|\mathbf{x}_t)$ is nonlinear, it is still possible to satisfy the full rank condition even if $d < K$, and we illustrate this in the simulation section. [2]

3 Definition of the estimators

The above identification strategy motivates us to estimate $\mathbf{\Lambda}$ and $\mathbf{g}(\mathbf{x}_t)$ respectively by $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{g}}(\mathbf{x}_t)$ as follows. Let $\widehat{\mathbf{\Sigma}}$ and $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ be some estimator of $\mathbf{\Sigma}_{y|x}$ and $E(\mathbf{y}_t|\mathbf{x}_t)$, whose

[2]Suppose $E(\mathbf{f}_t|\mathbf{x}_t)$ is nonlinear and can be well approximated by a series of orthogonal basis functions $\Phi(\mathbf{x}_t) = (\phi_1(\mathbf{x}_t), ..., \phi_J(\mathbf{x}_t))'$, where $E\phi_i(\mathbf{x}_t)\phi_j(\mathbf{x}_t) = 1\{i = j\}$, then for some $K \times J$ coefficient $\boldsymbol{\alpha}$, we have $E(\mathbf{f}_t|\mathbf{x}_t) \approx \boldsymbol{\alpha}'\Phi(\mathbf{x}_t)$ so $E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\} \approx \boldsymbol{\alpha}\boldsymbol{\alpha}'$. For nonlinear functions, it is not stringent to require $\boldsymbol{\alpha}\boldsymbol{\alpha}'$ be full rank since $K < J$ as $J \to \infty$.

definitions will be clear below. Then the columns of $\frac{1}{\sqrt{N}}\widehat{\mathbf{\Lambda}}$ are defined as the eigenvectors corresponding to the first K eigenvalues of $\widehat{\mathbf{\Sigma}}$, and

$$\widehat{\mathbf{g}}(\mathbf{x}_t) := \frac{1}{N}\widehat{\mathbf{\Lambda}}'\widehat{E}(\mathbf{y}_t|\mathbf{x}_t).$$

Recall that $\mathbf{f}_t = \mathbf{g}(\mathbf{x}_t) + \boldsymbol{\gamma}_t$. We estimate $\mathbf{f}_t$ using least squares:

$$\widehat{\mathbf{f}}_t := (\widehat{\mathbf{\Lambda}}'\widehat{\mathbf{\Lambda}})^{-1}\widehat{\mathbf{\Lambda}}'\mathbf{y}_t = \frac{1}{N}\widehat{\mathbf{\Lambda}}'\mathbf{y}_t.$$

Finally, we estimate $\boldsymbol{\gamma}_t$ by: $\widehat{\boldsymbol{\gamma}}_t = \widehat{\mathbf{f}}_t - \widehat{\mathbf{g}}(\mathbf{x}_t) = \frac{1}{N}\widehat{\mathbf{\Lambda}}'(\mathbf{y}_t - \widehat{E}(\mathbf{y}_t|\mathbf{x}_t))$. Estimating $\mathbf{g}(\mathbf{x}_t)$ and $\boldsymbol{\gamma}_t$ separately allows us to estimate and distinguish the percentage of explained and unexplained components in factors, as well as to quantify the explanatory power of covariates.

Below we introduce the estimators $\widehat{\mathbf{\Sigma}}$ and $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ to be used in this paper.

3.1 Robust estimation for $\widehat{\mathbf{\Sigma}}$

Recall that $\mathbf{\Sigma}_{y|x} = E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$, and let us first construct an estimator for $E(\mathbf{y}_t|\mathbf{x}_t)$ as follows. While many standard nonparametric regressions would work, here we choose an estimator that is robust to the tail-distributions of $\mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t)$.

Let $\Phi(\mathbf{x}_t) = (\phi_1(\mathbf{x}_t), ..., \phi_J(\mathbf{x}_t))'$ be a $J \times 1$ dimensional vector of sieve basis. Suppose $E(\mathbf{y}_t|\mathbf{x}_t)$ can be approximated by a sieve representation: $E(\mathbf{y}_t|\mathbf{x}_t) \approx \mathbf{B}\Phi(\mathbf{x}_t)$, where $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_N)'$ is an $N \times J$ matrix of sieve coefficients. To adapt to different heaviness of the tails of idiosyncratic components, we use the Huber loss function (Huber (1964)) to estimate the sieve coefficients. Define

$$\rho(z) = \begin{cases} z^2, & |z| < 1 \\ 2|z| - 1, & |z| \geq 1. \end{cases}$$

For some deterministic sequence $\alpha_T \to \infty$ (adaptive Huber loss), we estimate the sieve coefficients $\mathbf{B}$ by the following convex optimization:

$$\widehat{\mathbf{b}}_i = \arg\min_{b \in \mathbb{R}^J} \frac{1}{T}\sum_{t=1}^{T} \rho\left(\frac{y_{it} - \Phi(\mathbf{x}_t)'\mathbf{b}}{\alpha_T}\right), \quad \widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, ..., \widehat{\mathbf{b}}_N)'.$$

We then estimate $\mathbf{\Sigma}_{y|x}$ by

$$\widehat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{E}(\mathbf{y}_t|\mathbf{x}_t)\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)', \quad \text{where } \widehat{E}(\mathbf{y}_t|\mathbf{x}_t) = \widehat{\mathbf{B}}\Phi(\mathbf{x}_t).$$

An alternative method to the robust estimation of $\mathbf{\Sigma}_{y|x}$ is based on the sieve-least squares, corresponding to the case where $\alpha_T = \infty$. Let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$, which is $(N \times T)$, and

$$\mathbf{P} = \Phi'(\Phi\Phi')^{-1}\Phi, (T \times T), \quad \Phi = (\Phi(\mathbf{x}_1), ..., \Phi(\mathbf{x}_T)), (J \times T).$$

Then, the sieve least-squares estimator for $\mathbf{\Sigma}_{y|x}$ is $\widetilde{\mathbf{\Sigma}} = \frac{1}{T}\mathbf{YPY}'$. While this estimator is attractive due to its closed form, it is not as good as $\widehat{\mathbf{\Sigma}}$ when the distribution of $\mathbf{u}_t$ has heavier tails. As expected, our numerical studies in Section 7 demonstrate that it performs well in light-tailed scenarios, but is less robust to heavy-tailed distributions. Our theories are presented for $\widehat{\mathbf{\Sigma}}$, but most of the theoretical findings should carry over to $\widetilde{\mathbf{\Sigma}}$.

3.2 Choosing α_T and J

The selection of the sieve dimension J has been widely studied in the literature, e.g., Li (1987); Andrews (1991); Hurvich et al. (1998), among others. Another tunning parameter is α_T, which diverges in order to reduce the biases of estimating the conditional mean when the distribution of $\mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t)$ is asymmetric. Throughout the paper, we shall set

$$\alpha_T = C_\alpha \sqrt{\frac{T}{\log(NJ)}} \tag{3.1}$$

for some constant $C_\alpha > 0$, and choose (J, C_α) simultaneously using the multi-fold cross-validation[3]. The specified rate in (3.1) is due to a theoretical consideration, which leads to the "least biased robust estimation", as we now explain. The Huber-estimator is biased for estimating the mean coefficient in $E(y_{it}|\mathbf{x}_t)$, whose population counterpart is

$$\mathbf{b}_{i,\alpha} := \arg\min_{\mathbf{b}\in\mathbb{R}^J} E\rho\left(\frac{y_{it} - \Phi(\mathbf{x}_t)'\mathbf{b}}{\alpha_T}\right),$$

As α_T increases, it approaches the limit $\mathbf{b}_i := \arg\min_{\mathbf{b}\in\mathbb{R}^J} E[y_{it} - \mathbf{b}'\Phi(\mathbf{x}_t)]^2$ with the speed

$$\max_{i\le N} \|\mathbf{b}_{i,\alpha} - \mathbf{b}_i\| = O(\alpha_T^{-c_0})$$

for some constant $c_0 > 0$ that depends on the thickness of the tail distribution of $y_{it} - E(y_{it}|\mathbf{x}_t)$. Hence the bias decreases as α_T grows. On the other hand, our theory requires

[3]One can also allow α_T to depend on $\text{var}(y_{it}|\mathbf{x}_t)$ to allow for different scales across individuals. We describe this choice in the simulation section. In addition, the cross-validation can be based on either in-sample fit for $E(y_{it}|\mathbf{x}_t)$ or out-of-sample forecast, depending on the specific applications. In time series forecasts, one may also consider the time series cross validation (e.g. Hart, 1994) where the training and testing sets are defined through a moving window forecast.

the uniform convergence (in $i = 1, ..., N$) of (for $e_{it} = y_{it} - E(y_{it}|\mathbf{x}_t)$)

$$\max_{i \leq N} \| \frac{1}{T} \sum_{t=1}^{T} \dot{\rho}(\alpha_T^{-1} e_{it}) \Phi(\mathbf{x}_t) \|, \tag{3.2}$$

where $\dot{\rho}(\cdot)$ denotes the derivative of $\rho(\cdot)$. It turns out that α_T cannot grow faster than $O(\sqrt{\frac{T}{\log(NJ)}})$ in order to guard for robustness and to have a sharp uniform convergence for (3.2). Hence the choice (3.1) leads to the asymptotically least-biased robust estimation.

3.3 Alternative estimators

Plugging $\mathbf{f}_t = \mathbf{g}(\mathbf{x}_t) + \boldsymbol{\gamma}_t$ into (1.1), we obtain

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t) + \boldsymbol{\Lambda}\boldsymbol{\gamma}_t + \mathbf{u}_t, \quad \text{where } \mathbf{h}(\mathbf{x}_t) = \boldsymbol{\Lambda}\mathbf{g}(\mathbf{x}_t). \tag{3.3}$$

A related model is:

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t) + \boldsymbol{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \tag{3.4}$$

for a nonparametric function $\mathbf{h}(\cdot)$, or simply a linear form $\mathbf{h}(\mathbf{x}_t) = \boldsymbol{\beta}\mathbf{x}_t$. Models (3.3) and (3.4) were studied in the literature (Ahn et al., 2001; Bai, 2009; Moon and Weidner, 2015), where parameters are often estimated using least squares. For instance, we can estimate model (3.3) by

$$\min_{\mathbf{h}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}_t} \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{h}(\mathbf{x}_t) - \boldsymbol{\Lambda}\boldsymbol{\gamma}_t\|^2. \tag{3.5}$$

But this approach is not appropriate in the current context when $\mathbf{x}_t$ almost fully explains $\mathbf{f}_t$ for all $t = 1, ..., T$. In this case, $\boldsymbol{\gamma}_t \approx 0$, and least squares (3.5) would be inconsistent. [4] In addition, $\boldsymbol{\Lambda}$ in (3.4) would be very close to zero because the effects of $\mathbf{f}_t$ would be fully explained by $\mathbf{h}(\mathbf{x}_t)$. As a result, the factors in (3.4) cannot be consistently estimated (Onatski, 2012b) either. We conduct numerical comparisons with this method in the simulation section. In all simulated scenarios, the interactive effect approach gives the worst estimation performance.

Another simpler alternative is to combine $(\mathbf{x}_t, \mathbf{y}_t)$, and apply the classical methods on this enlarged dataset. One potential drawback is that the rates of convergence would not be improved, even if $\mathbf{x}_t$ has strong explanatory power on the factors. Another drawback, as mentioned before, is that it does not distinguish the disaggregated variables $\mathbf{x}_t$ and aggregated variables $\mathbf{y}_t$, which can provide very different information (e.g. Fama-French

[4] The inconsistency is due to the fact that $a\boldsymbol{\Lambda}\boldsymbol{\gamma}_t \approx \boldsymbol{\Lambda}\boldsymbol{\gamma}_t$ for *any* scalar a in the case $\boldsymbol{\gamma}_t \approx 0$. Thus $\boldsymbol{\Lambda}$ is not identifiable in the least squares problem.

factors versus returns of individual stocks).

4 Rates of Convergence

4.1 Assumptions

Let $e_{it} := y_{it} - E(y_{it}|\mathbf{x}_t)$. Suppose the conditional distribution of e_{it} given $\mathbf{x}_t = \mathbf{x}$ is absolutely continuous for almost all $\mathbf{x}$, with a conditional density $g_{e,i}(\cdot|\mathbf{x})$.

Assumption 4.1 (Tail distributions)**.** *(i) There are $\zeta_1, \zeta_2 > 2$, $C > 0$ and $M > 0$, so that for all $x > M$,*

$$\sup_{\mathbf{x}} \max_{i \le N} g_{e,i}(x|\mathbf{x}) \le Cx^{-\zeta_1}, \quad \sup_{\mathbf{x}} \max_{i \le N} E(e_{it}^2 1\{|e_{it}| > x\}|\mathbf{x}_t = \mathbf{x}) \le Cx^{-\zeta_2}. \tag{4.1}$$

(ii) $\Phi(\mathbf{x}_t)$ is a sub-Gaussian vector, that is, there is $L > 0$, for any $\boldsymbol{\nu} \in \mathbb{R}^J$ so that $\|\boldsymbol{\nu}\| = 1$,

$$P(|\boldsymbol{\nu}'\Phi(\mathbf{x}_t)| > x) \le \exp(1 - x^2/L), \quad \forall x \ge 0.$$

Assumption 4.2 (Sieve approximations)**.** *(i) For $k = 1, ..., K$, let $\mathbf{v}_k = \arg\min_{\mathbf{v}} E(f_{kt} - \mathbf{v}'\Phi(\mathbf{x}_t))^2$. Then there is $\eta \ge 2$, as $J \to \infty$,*

$$\max_{k \le K} \sup_{\mathbf{x}} |E(f_{tk}|\mathbf{x}_t = \mathbf{x}) - \mathbf{v}_k'\Phi(\mathbf{x})| = O(J^{-\eta}).$$

(ii) There are $c_1, c_2 > 0$ so that

$$c_1 \le \lambda_{\min}(E\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)') \le \lambda_{\max}(E\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)') \le c_2.$$

Recall $\boldsymbol{\gamma}_t = \mathbf{f}_t - E(\mathbf{f}_t|\mathbf{x}_t)$. Let γ_{kt} be its k th component.

Assumption 4.3. *(i) (serial independence) $\{\mathbf{f}_t, \mathbf{u}_t, \mathbf{x}_t\}_{t \le T}$ is independent and identically distributed;*

(ii) (weak cross-sectional dependence) For some $C > 0$,

$$\sup_{\mathbf{x},\mathbf{f}} \max_{i \le N} \sum_{j=1}^{N} |E(u_{it}u_{jt}|\mathbf{x}_t = \mathbf{x}, \mathbf{f}_t = \mathbf{f})| < C.$$

(iii) $E(\mathbf{u}_t|\mathbf{f}_t, \mathbf{x}_t) = 0$, $\max_{i \le N} \|\boldsymbol{\lambda}_i\| < C$, and $\mathrm{cov}(\boldsymbol{\gamma}_t|\mathbf{x}_t) = \mathrm{cov}(\boldsymbol{\gamma}_t)$ almost surely, where $\mathrm{cov}(\boldsymbol{\gamma}_t|\mathbf{x}_t)$ denotes the conditional covariance matrix of $\boldsymbol{\gamma}_t$ given $\mathbf{x}_t$, assumed to exist.

Recall that

$$\mathbf{\Sigma}_{f|x} := E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}, \quad \chi_N := \lambda_{\min}(\mathbf{\Sigma}_{f|x}).$$

Assumption 4.4 (Signal-noise)**.** *(i) There is $C > 0$,*

$$\frac{\lambda_{\max}(\mathbf{\Sigma}_{f|x})}{\lambda_{\min}(\mathbf{\Sigma}_{f|x})} < C, \quad \frac{\lambda_{\max}(E\{\Phi(\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'E(\mathbf{f}_t|\mathbf{x}_t)\Phi(\mathbf{x}_t)'\})}{\lambda_{\min}(\mathbf{\Sigma}_{f|x})} < C.$$

(ii) There is $v > 1$, so that $\max_{k\le K} E[E(\gamma_{kt}^4|\mathbf{x}_t)]^v < \infty$.
(iii) We have $J^3 \log^2 N = O(T)$ and

$$J^2/T + J^{-\eta} + \sqrt{(\log N)/T} \ll \chi_N.$$

Assumption 4.1 allows distributions with relatively heavy tails on $y_{it} - E(y_{it}|\mathbf{x}_t)$. We still require sub-Gaussian tails for the sieve basis functions. Assumption 4.2 is regarding the accuracy of sieve approximations for nonparametric functions. Assumption 4.4 strengthens Assumption 2.1. We respectively regard $\lambda_{\min}(\mathbf{\Sigma}_{f|x})$ and $\mathrm{cov}(\boldsymbol{\gamma}_t)$ as the "signal" and "noise" when using $\mathbf{x}_t$ to explain common factors. The explanatory power is measured by these two quantities.

Assumption 4.3 (i) requires serial independence, and we admit that it can be restrictive in applications. Allowing for serial dependence is technically difficult due to the non-smooth Huber's loss. We derive a Bahadur expansion of the estimated eigenvectors for the spiked matrices, in the following form: for some rotation matrix $\mathbf{H}$,

$$\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H} = \frac{1}{NT}\sum_{t=1}^{T} \mathbf{\Lambda}\mathbf{g}(\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\sum_{i=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\Phi(\mathbf{x}_s)'\dot{\rho}(\alpha_T^{-1}e_{is})\alpha_T\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1} + \widetilde{\mathbf{\Delta}} \quad (4.2)$$

where $\dot{\rho}$ denotes the derivative of the Huber's loss function:

$$\dot{\rho}(z) = \begin{cases} 2z, & |z| < 1 \\ 2\,\mathrm{sgn}(z), & |z| \ge 1. \end{cases}$$

Here $\mathrm{sgn}(z)$ denotes the sign function; $\mathbf{e}_t := \mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t) = (e_{1t}, ..., e_{Nt})'$; $\mathbf{A}$ is the Hessian matrix of the expected Huber's loss function; $\widetilde{\mathbf{V}}$ is a K-dimensional diagonal matrix of the eigenvalues of $\widehat{\mathbf{\Sigma}}/N$. The second term $\widetilde{\mathbf{\Delta}}$ is a higher order random term. Such an expansion allows us to derive a much sharper bound for low-dimensional functionals of $\mathbf{\Delta}_\Lambda$ such as the estimated factors. To obtain the Bahadur representation of the estimated eigenvectors, we rely on the symmetrization and contraction theorems (e.g., van der Vaart and Wellner (1996)), which requires the data be independently distributed. Nevertheless, the idea of

using covariates would still be applicable for serial dependent data. For instance, it is not difficult to allow for weak serial correlations when the data are not heavy-tailed, by using the sieve least squares estimator $\widetilde{\mathbf{\Sigma}}$ (introduced in Section 3.1) in place of the Huber's estimator $\widehat{\mathbf{\Sigma}}$. We conduct numerical studies when the data are serially correlated in the simulations, and find that the proposed methods continue to perform well in the presence of serial correlations.

4.2 Rates of convergence

We present the rates of convergence in the following theorems, and discuss the statistical insights in the next subsection. Recall $\widehat{\mathbf{\Lambda}} = (\widehat{\boldsymbol{\lambda}}_i : i \leq N)$.

Theorem 4.1 (Loadings)**.** *Under Assumptions 2.1–4.4, there is an invertible matrix* $\mathbf{H}$*, as* $T, J \to \infty$*, and* N *either grows or stays constant,*

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 = O_P\left(\frac{J}{T} + \frac{1}{J^{2\eta-1}}\right)\chi_N^{-1}, \tag{4.3}$$

$$\max_{i\leq N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\| = O_P\left(\sqrt{\frac{J\log N}{T}} + \frac{1}{J^{\eta-1/2}}\right)\chi_N^{-1/2}. \tag{4.4}$$

The optimal rate for J in (4.3) is $J \asymp T^{1/(2\eta)}$, which results in

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 = O_P(T^{-(1-\frac{1}{2\eta})}\chi_N^{-1}). \tag{4.5}$$

Here η represents the smoothness of $E(\mathbf{f}_t|\mathbf{x}_t = \cdot)$, as defined in Assumption 4.2.

Define

$$J^* = \min\left\{(TN)^{1/(2\eta)}, (\frac{T}{\log N})^{1/(1+\eta)}\right\}.$$

Theorem 4.2 (Factors)**.** *Let* $J \asymp J^*$*. Suppose* $(J^*)^3 \log^2 N = O(T)$*, and Assumptions 2.1–4.4 hold. For* $\mathbf{H}$ *in Theorem 4.1, as* $T \to \infty$*, and* N *either grows or stays constant, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2 = O_P\left(r_{T,N}^* + (\frac{\log N}{T})^{2-\frac{3}{1+\eta}}\right),$$

where $r^*_{T,N} = \frac{J^{*2}}{T^2}\chi_N^{-1} + \frac{J^*\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|}{T} + (\frac{1}{TN})^{1-\frac{1}{2\eta}}$ *and*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 &= O_P\left(r^*_{T,N} + (\frac{\log N}{T})^{2-\frac{4}{1+\eta}}\right)\chi_N^{-1} \\ &\quad + O_P\left(\frac{1}{N}\right). \end{aligned} \tag{4.6}$$

These two convergences imply the rate of convergence of the estimated factors due to $\widehat{\mathbf{f}}_t = \widehat{\mathbf{g}}(\mathbf{x}_t) + \widehat{\boldsymbol{\gamma}}_t$.

Remark 4.1. For a general J, the rates of convergence of the two factor components are

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2 = O_P\left(r_{T,N} + \frac{J^3\log^2 N}{T^2}\right), \tag{4.7}$$

where $r_{T,N} = \frac{J^2}{T^2}\chi_N^{-1} + \frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + J^{1-2\eta} + \frac{J}{TN}$ and

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P\left(r_{T,N} + \frac{J^4\log^2 N}{T^2}\right)\chi_N^{-1} + O_P\left(\frac{1}{N}\right). \tag{4.8}$$

In fact $J \asymp J^*$ is the optimal choice in (4.7) ignoring the terms involving $\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|$ and χ_N. The convergence rates presented in Theorem 4.2 are obtained from (4.7) and (4.8) with this choice of J.

The presented rates connect well with the literature on both standard nonparametric sieve estimations and the high-dimensional factor models. To illustrate this, we discuss in more detail about the rate of convergence in (4.7). This rate is given by:

$$O_P\left(\underbrace{\frac{J^2}{T^2}\chi_N^{-1}}_{\substack{\text{effect of}\\ \text{estimating } \boldsymbol{\Lambda}}} + \underbrace{\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + \frac{J}{TN} + J^{1-2\eta}}_{\substack{\text{nonparametric sieve}\\ \text{estimation error}}} + \underbrace{\frac{J^3\log^2 N}{T^2}}_{\substack{\text{higher order from}\\ \text{Huber's M-estimation}}}\right).$$

More specifically, we have, for $\mathbf{e}_t = \boldsymbol{\Lambda}\boldsymbol{\gamma}_t + \mathbf{u}_t$,

$$\mathbf{y}_t = \boldsymbol{\Lambda}\mathbf{g}(\mathbf{x}_t) + \mathbf{e}_t, \quad E(\mathbf{e}_t|\mathbf{x}_t) = 0. \tag{4.9}$$

If $\boldsymbol{\Lambda}$ were known, we would estimate $\mathbf{g}(\cdot)$ by regressing the estimated $E(\mathbf{y}_t|\mathbf{x}_t)$ on $\boldsymbol{\Lambda}$. Then standard nonparametric results show that the rate of convergence in this "oracle sieve

regression" (knowing $\mathbf{\Lambda}$) would be

$$\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + \frac{J}{TN} + J^{1-2\eta}.$$

As we do not observe $\mathbf{\Lambda}$, we are running the regression (4.9) with $\widehat{\mathbf{\Lambda}}$ in place of $\mathbf{\Lambda}$. This leads to an additional term $\frac{J^2}{T^2}\chi_N^{-1}$ representing the effect of estimating $\mathbf{\Lambda}$, which also depends on the strength of the signal χ_N. Finally, Huber's M-estimation to estimate $E(\mathbf{y}_t|\mathbf{x}_t)$ gives rise to a higher order term $\frac{J^3\log^2 N}{T^2}$, and is often negligible.

4.3 The signal-noise regimes

We see that the rates depend on $\operatorname{cov}(\boldsymbol{\gamma}_t)$ and χ_N. Because $E\mathbf{f}_t\mathbf{f}_t' = \mathbf{\Sigma}_{f|x} + \operatorname{cov}(\boldsymbol{\gamma}_t)$, they are related through

$$c \leq \chi_N + \|\operatorname{cov}(\boldsymbol{\gamma}_t)\| \leq C_1 \tag{4.10}$$

for some $c, C_1 > 0$, assuming that there is $c > 0$ so that $\|E\mathbf{f}_t\mathbf{f}_t'\| > c$. For comparison, we state the rates of convergence of the benchmark PCA estimators: (e.g., Stock and Watson (2002); Bai (2003)) there is a rotation matrix $\tilde{\mathbf{H}}$, so that the PCA estimators $(\widetilde{\boldsymbol{\lambda}}_i, \widetilde{\mathbf{f}}_t)$ satisfy:

$$\frac{1}{N}\sum_{i=1}^{N}\|\widetilde{\boldsymbol{\lambda}}_i - \tilde{\mathbf{H}}'\boldsymbol{\lambda}_i\|^2 = O_P(\frac{1}{T}+\frac{1}{N}), \qquad \frac{1}{T}\sum_{t=1}^{T}\|\widetilde{\mathbf{f}}_t - \tilde{\mathbf{H}}^{-1}\mathbf{f}_t\|^2 = O_P(\frac{1}{T}+\frac{1}{N}). \tag{4.11}$$

The first interesting phenomena we observe is that both the estimated loadings and $\mathbf{g}(\mathbf{x}_t)$ are consistent even if N is finite, due to the "exact identification". In contrast, the PCA estimators requires a growing N. For more detailed comparisons, we consider three regimes based on the explanatory power of the factors using $\mathbf{x}_t$. To simplify our discussions, we consider the rate-optimal choices of J, and ignore the sieve approximation errors, so η is treated sufficiently large.

Regime I: strong explanatory power: $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\| \to 0$. Because of (4.10), χ_N is bounded away from zero. In this case, (4.5)-(4.6) approximately imply (for sufficiently large η):

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 &= O_P\left(\frac{1}{T}\right), \\
\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2 &= O_P\left(\frac{\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|}{T} + \frac{1}{TN} + (\frac{\log N}{T})^2\right), \\
\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}^{-1}\mathbf{f}_t\|^2 &= O_P\left(\frac{\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|}{T} + \frac{1}{N} + (\frac{\log N}{T})^2\right).
\end{aligned}$$

Compared to the rates of the usual PCA estimators in (4.11), either the new estimated loadings (when $N = o(T)$) or the new estimated factors (when $T = o(N)$) have a faster rate of convergence. Moreover, if $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\| = o((TN)^{-1} + T^{-2}\log^2 N)$, then $\widehat{\mathbf{g}}(\mathbf{x}_t)$ directly estimates the latent factor at a very fast rate of convergence:

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{f}_t\|^2 = O_P\left(\frac{1}{TN} + (\frac{\log N}{T})^2\right).$$

The improved rates are reasonable due to the strong explanatory powers from the covariates.

Regime II: mild explanatory power: $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|$ *is bounded away from zero;* χ_N *is either bounded away from zero or decays slower than* $\frac{N}{T}$ *in the case* $N = o(T)$. In this regime, $\mathbf{x}_t$ partially explains the factors, yet the unexplainable components are not negligible. (4.5)-(4.6) approximately become:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 &= O_P\left(\frac{1}{T}\chi_N^{-1}\right)\\ \frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}^{-1}\mathbf{f}_t\|^2 &= O_P\left(\frac{1}{T}\chi_N^{-1} + \frac{1}{N}\right).\end{aligned} \tag{4.12}$$

We see that the rate for the estimated loadings is still faster than the PCA when N is relatively small compared to T, while the rates for the estimated factors are the same. This is because

$$\underbrace{\frac{1}{T}\chi_N^{-1}}_{\text{new rate for loadings}} \ll \underbrace{\frac{1}{N}}_{\text{PCA rate for loadings}}$$

$$\underbrace{\frac{1}{T}\chi_N^{-1} + \frac{1}{N}}_{\text{new rate for factors}} \asymp \underbrace{\frac{1}{N}}_{\text{PCA rate for factors}}.$$

On one hand, due to the explanatory power from the covariates, the loadings can be estimated well without having to consistently estimate the factors. On the other hand, as the covariates only partially explain the factors, we cannot improve rates of convergence in estimating the unexplainable components in the latent factors. However, since $\boldsymbol{\gamma}_t$ has smaller variability than $\mathbf{f}_t$, it can still be better estimated in terms of a smaller constant factor.

Regime III: weak explanatory power: $\chi_N \to 0$ *and decays faster than* $\frac{N}{T}$ *when* $N \ll T$.

In this case, we have

$$\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 = O_P(\frac{1}{T}\chi_N^{-1}) = \frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}^{-1}\mathbf{f}_t\|^2.$$

While the new estimators are still consistent, they perform worse than PCA. This finding is still reasonable because the signal is so weak that the conditional expectation $E(\mathbf{y}_t|\mathbf{x}_t)$ loses useful information of the factors/loadings. Consequently, estimation efficiency is lost when running PCA on the estimated covariance $E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$.

In summary, improved rates of convergence can be achieved so long as the covariates can (partially) explain the latent factors, this corresponds to either the mild or the strong explanatory power case. The degree of improvements depend on the strength of the signals. In particular, the consistent estimation for factor loadings can also be achieved even under finite N. On the other hand, when the explanatory power is too weak, the rates of convergence would be slower than those of the benchmark estimator.

5 Testing the Explanatory Power of Covariates

We aim to test: (recall that $\boldsymbol{\gamma}_t = \mathbf{f}_t - E(\mathbf{f}_t|\mathbf{x}_t)$)

$$H_0 : \text{cov}(\boldsymbol{\gamma}_t) = 0. \tag{5.1}$$

Under H_0, $\mathbf{f}_t = E(\mathbf{f}_t|\mathbf{x}_t)$ over the entire sampling period $t = 1, ..., T$, implying that observed covariates $\mathbf{x}_t$ fully explain the true factors $\mathbf{f}_t$. In empirical applications with "observed factors", what have been often used are in fact $\mathbf{x}_t$. Hence our proposed test can be applied to empirically validate the explanatory power of these "observed factors".

The Fama-French three-factor model (Fama and French, 1992) is one of the most celebrated ones in empirical asset pricing. They modeled the excess return r_{it} on security or portfolio i for period t as

$$r_{it} = \alpha_i + b_i r_{Mt} + s_i \text{SMB}_t + h_i \text{HML}_t + u_{it},$$

where r_{Mt}, SMB_t and HML_t respectively represent the the excess returns of the market, the difference of returns between stocks with small and big market capitalizations ("small minus big"), and the difference of returns between stocks with high book to equity ratios and those with low book to equity ratios ("high minus low"). Ever since its proposal, there is much evidence that the three-factor model can leave the cross-section of expected

stock returns unexplained. Different factor definitions have been explored, e.g., Carhart (1997) and Novy-Marx (2013). Fama and French (2015) added profitability and investment factors to the three-factor model. They conducted GRS tests (Gibbons et al., 1989) on the five-factor models and its different variations. Their tests "reject all models as a complete description of expected returns".

On the other hand, the Fama-French factors, though imperfect, are good proxies for the true unknown factors. Consequently, they form a natural choice for $\mathbf{x}_t$. These observables are actually diversified portfolios, which have explanatory power on the latent factors $\mathbf{f}_t$, as supported by financial economic theories as well as empirical studies. The test proposed in this section validates the specification of these common covariates as "factors".

5.1 The Test Statistic

Our test is based on a Wald-type weighted quadratic statistic

$$S(\mathbf{W}) := \frac{N}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\gamma}}_t'\mathbf{W}\widehat{\boldsymbol{\gamma}}_t = \frac{1}{TN}\sum_{t=1}^{T}(\mathbf{y}_t - \widehat{E}(\mathbf{y}_t|\mathbf{x}_t))'\widehat{\boldsymbol{\Lambda}}\mathbf{W}\widehat{\boldsymbol{\Lambda}}'(\mathbf{y}_t - \widehat{E}(\mathbf{y}_t|\mathbf{x}_t)).$$

The weight matrix normalizes the test statistic, taken as $\mathbf{W} = \text{AVar}(\sqrt{N}\widehat{\boldsymbol{\gamma}}_t)^{-1}$, where $\text{AVar}(\widehat{\boldsymbol{\gamma}}_t)$ represents the asymptotic covariance matrix of $\widehat{\boldsymbol{\gamma}}_t$ under the null, and is given by

$$\text{AVar}(\sqrt{N}\widehat{\boldsymbol{\gamma}}_t) = \frac{1}{N}\mathbf{H}'\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u\boldsymbol{\Lambda}\mathbf{H}.$$

As $\boldsymbol{\Sigma}_u$ is a high-dimensional covariance matrix, to simplify the technical arguments, in this section we assume $\{u_{it}\}$ to be cross-sectionally uncorrelated, and estimate $\boldsymbol{\Sigma}_u$ by:

$$\widehat{\boldsymbol{\Sigma}}_u = \text{diag}\{\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}^2, i = 1, ..., N\}, \quad \widehat{u}_{it} = y_{it} - \widehat{\boldsymbol{\lambda}}_i'\widehat{\mathbf{f}}_t.$$

The feasible test statistic is defined as

$$S := S(\widehat{\mathbf{W}}), \quad \widehat{\mathbf{W}} := (\frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Sigma}}_u\widehat{\boldsymbol{\Lambda}})^{-1}.$$

We reject the null hypothesis for large values of S. It is straightforward to allow $\boldsymbol{\Sigma}_u$ to be a non-diagonal but a sparse covariance, and proceed as in Bickel and Levina (2008). We expect the asymptotic analysis to be quite involved, and do not pursue it in this paper.

We show that the test statistic has the following asymptotic expansion:

$$S = \bar{S} + o_P(\frac{1}{\sqrt{T}}),$$

where

$$\bar{S} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t'\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Sigma}_u\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{u}_t.$$

Thus the limiting distribution is determined by that of $\bar{S}$. Note that a cross-sectional central limit theorem implies, as $N \to \infty$,

$$(\frac{1}{N}\mathbf{\Lambda}'\mathbf{\Sigma}_u\mathbf{\Lambda})^{-1/2}\frac{1}{\sqrt{N}}\mathbf{u}_t'\mathbf{\Lambda} \to^d \mathcal{N}(0, \mathbf{I}_K).$$

Hence each component of $\bar{S}$ can be roughly understood as χ^2-distributed with degrees of freedom K being the number of common factors, whose variance is $2K$. This motivates the following assumption.

Assumption 5.1. *Suppose* $\frac{1}{T}\sum_{t=1}^{T} \mathrm{var}(\mathbf{u}_t'\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Sigma}_u\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{u}_t) \to 2K$ *as* $T, N \to \infty$.

We now state the null distribution in the following theorem.

Theorem 5.1. *Suppose Assumption 5.1 and assumptions of Theorem 4.2 hold. In addition, we further assume that* $\{u_{it}\}_{i\le N}$ *is cross-sectionally independent. Then, when* $J^4 N \log N = o(T^{3/2})$, $T = o(N^2)$, $N\sqrt{T} = o(J^{2\eta-1})$, *as* $T, N \to \infty$,

$$\sqrt{\frac{T}{2K}}(S - K) \to^d \mathcal{N}(0, 1).$$

5.2 Testing market risk factors for S&P 500 returns

We test the explanatory power of the observable proxies for the true factors using S&P 500 returns. We calculate the excess returns for the stocks in S&P 500 index that are collected from the Center for Research in Securities Prices (CRSP). We consider three groups of proxy factors ($\mathbf{x}_t$) with increasing information: (1) Fama-French 3 factors (FF3); (2) Fama-French 5 factors (FF5); and (3) Fama-French 5 factors plus 9 sector SPDR ETF's (FF5+ETF9). Here the sector SPDR ETF's, which are intended to track the 9 largest S&P sectors. The detailed descriptions of sector SPDR ETF's are listed in Table 5.1. For each given group of observable proxies, we set the number of common factors K equals the number of observable proxies.

Table 5.1: Sector SPDR ETF's (data available from Yahoo finance)

Code	Sector	Code	Sector	Code	Sector
XLE	Energy	XLB	Materials	XLI	Industrials
XLY	Consumer discretionary	XLP	Consumer staples	XLV	Health care
XLF	Financial	XLK	Information technology	XLU	Utilities

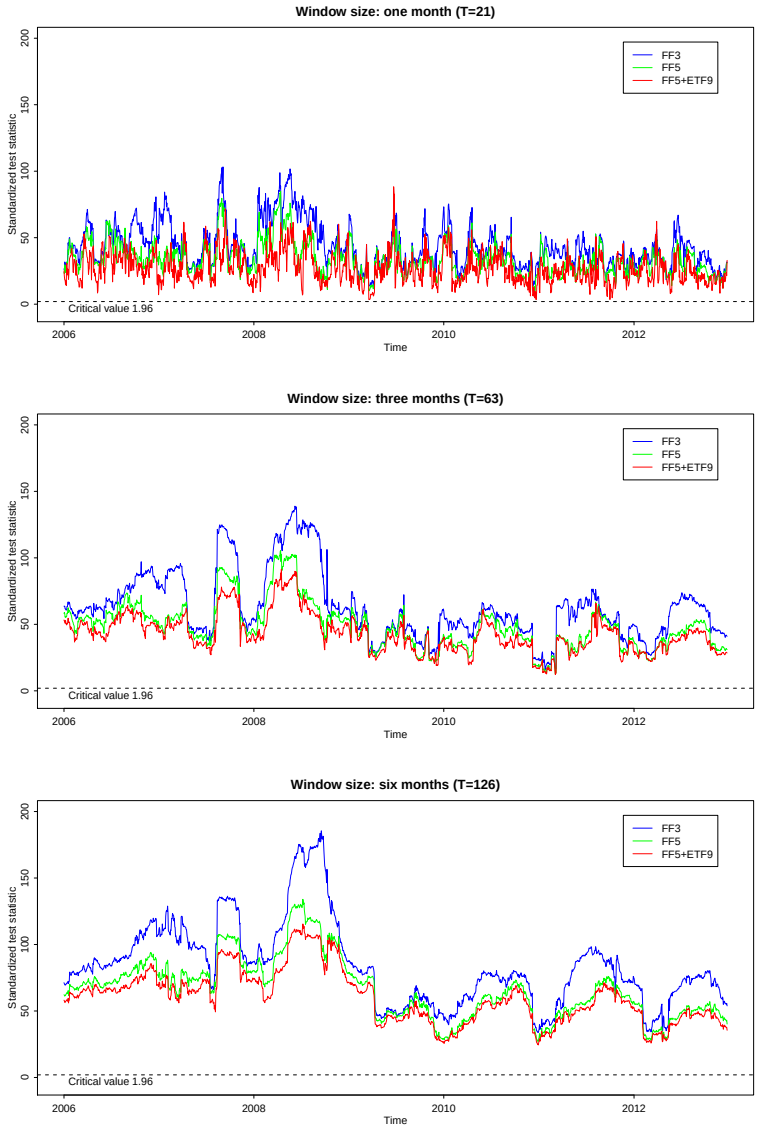

Figure 5.1: S&P 500 daily returns: plots for standardized test statistic S for various window sizes. The dotted line is critical value 1.96.

We consider tests using both daily and monthly data. For the daily data, we collect 393 stocks that have complete daily closing prices from January 2005 to December 2013,

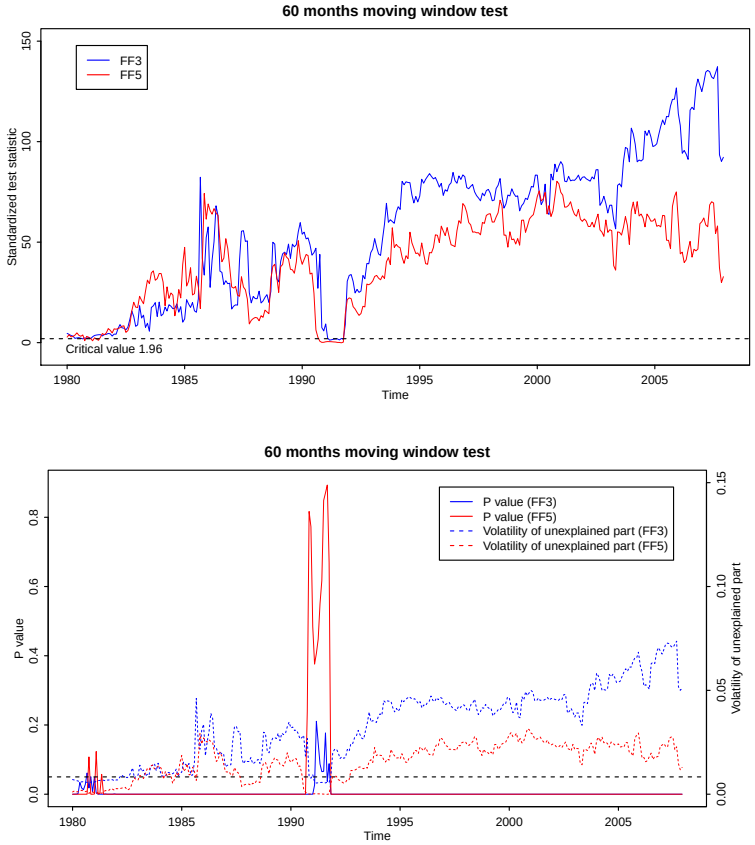

Figure 5.2: S&P 500 monthly returns: plots for standardized test statistic S, P-value and the volatility of the part of factors that can not be explained by the proxy factors.

with a time span of 2265 trading days. We apply moving window tests with the window size (T) equals one month, three months or six months. The testing window moves one trading day forward per test. Within each testing window, we calculate the standardized test statistic S for three groups of proxy factors.

As for the monthly excess returns, we use stocks that have complete record from January 1980 to December 2012, which contains 202 stocks with a time span of 396 months. Here we only consider the first two groups of proxy factors as sector SPDR ETF's are introduced since 1998. The window size equals sixty months and moves one month forward per test. Within each testing window, besides standardized test statistic and p-value, we also estimate the volatility of $\boldsymbol{\gamma}_t$, the part of factors that can not be explained by $\mathbf{x}_t$ as:

$$\widehat{\mathrm{Vol}}(\boldsymbol{\gamma}_t) = \frac{1}{21T}\sum_{t=1}^{T}\widehat{\boldsymbol{\gamma}}_t'\widehat{\boldsymbol{\gamma}}_t,$$

where there are 21 trading days per month. The sieve basis is chosen as the additive Fourier

basis with $J = 5$. We set the tuning parameter $\alpha_T = C\sqrt{\frac{T}{\log(NJ)}}$ with constant C selected by the 5-fold cross validation.

For the daily data, the plots of S under various scenarios are reported in Figure 5.1. Under all scenarios, the null hypothesis ($H_0 : \text{cov}(\boldsymbol{\gamma}_t) = 0$) is rejected as S is always larger than the critical value 1.96. This suggests a strong evidence that the proxy factors can not fully explain the estimated common factors. Under all window sizes, a larger group of proxy factors tends to yield smaller statistics, demonstrating stronger explanatory power for estimated common factors. Also, we find the test statistics increase while the window size increases.

The results for the monthly data are reported in Figure 5.2. For both Fama-French 3 factors and 5 factors, the null hypothesis is rejected most of the time except in early 1980s and 1990s. When the null hypothesis is accepted, Fama-French 5 factors tend to yield larger p-values. The estimated volatility of unexplained part are close to zero over these two periods. For the rest of the time, the standardized test statistics are much larger than the critical value 1.96 and hence the p-values are close to zero. Also the estimated volatilities are not close to zero. This indicates the proxy factors can not fully explain the estimated common factors during these testing periods.

6 Forecast the excess return of US government bonds

We apply our method to forecast the excess return of U.S. government bonds. The bond excess return is the one-year bond return in excess of the risk-free rate. To be more specific, we buy an n year bond, sell it as an $n-1$ year bond in the next year and excess the one-year bond yield as the risk-free rate. Let $p_t^{(n)}$ be the log price of an n-year discount bond at time t. Denote $\zeta_t^{(n)} \equiv -\frac{1}{n}p_t^{(n)}$ as the log yield with n year maturity, and $r_{t+1}^{(n)} \equiv p_{t+1}^{(n-1)} - p_t^{(n)}$ as the log holding period return. The goal of one-step-ahead forecast is to forecast $z_{T+1}^{(n)}$, the excess return with maturity of n years in period $T+1$, where

$$z_{t+1}^{(n)} = r_{t+1}^{(n)} - \zeta_t^{(1)}, \quad t = 1, \cdots, T.$$

For a long time, the literature has found a significant predictive power of the excess returns of U.S. government bonds. For instance, Ludvigson and Ng (2009, 2010) predicted the bond excess returns with observable variables based on a factor model using 131 (disaggregated) macroeconomics variables. They achieved the out-of-sample $R^2 \approx 21\%$ when forecasting bond excess return with two year maturity. Using the proposed method, this section develops a new way of incorporating the explanatory power of the observed char-

acteristics, and investigates the robustness of the conclusions in existing literature.

We analyze monthly data spanned from January 1964 to December 2003, which is available from the Center for Research in Securities Prices (CRSP). The factors are estimated from a macroeconomic dataset consisting of 131 disaggregated macroeconomic time series (Ludvigson and Ng, 2010). The covariates $\mathbf{x}_t$ are 8 aggregated macro-economic time series, listed in Table 6.1.

Table 6.1: Components of $\mathbf{x}_t$

$x_{1,t}$	Linear combination of five forward rates
$x_{2,t}$	Real gross domestic product (GDP)
$x_{3,t}$	Real category development index (CDI)
$x_{4,t}$	Non-agriculture employment
$x_{5,t}$	Real industrial production
$x_{6,t}$	Real manufacturing and trade sales
$x_{7,t}$	Real personal income less transfer
$x_{8,t}$	Consumer price index (CPI)

6.1 Heavy-tailed data and robust estimations

We first examine the excess kurtosis for the time series to assess the tail distributions. The left panel of Figure 6.1 shows 43 among the 131 series have excess kurtosis greater than 6. This indicates the tails of their distributions are fatter than the t-distribution with degrees of freedom 5. On the other hand, the right panel of Figure 6.1 reports the histograms of excess kurtosis of the "fitted data" $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ (the robust estimator of $E(\mathbf{y}_t|\mathbf{x}_t)$ using Huber loss), which demonstrates that most series in the fitted data are no longer severely heavy-tailed.

The tuning parameter in the Huber loss is of order $\alpha_T = C_\alpha\sqrt{\frac{T}{\log(NT)}}$. In this study, the constant C_α and the degree of sieve approximation J are selected by the out-of-sample 5-fold cross validation as described in Section 3.2.

6.2 Forecast results

We compare the rolling window forecast performance between our proposed *smoothed PCA* (SPCA) method and two competitors. The first competitor, denoted as SPCA-LS, is similar to SPCA except $\mathbf{\Sigma}_{y|x}$ is estimated by the sieve least-squares estimator $\widetilde{\mathbf{\Sigma}}$ rather than the robust estimator $\widehat{\mathbf{\Sigma}}$. We refer to Section 3.1 for the detailed definitions of $\widetilde{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}$.

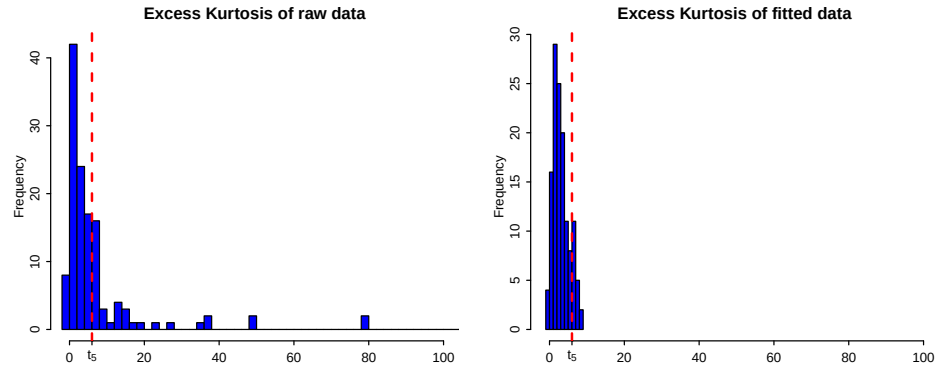

Figure 6.1: Excess kurtosis of the macroeconomic panel data. Left panel shows 43 among 131 series in the raw data are heavy tailed. Right panel shows the robustly fitted data $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ are no longer severely heavy-tailed.

The SPCA-LS method can be considered as a "non-robust" version of SPCA. The second competitor is the benchmark PCA.

We conduct one-month-ahead out-of-sample forecast of the bond risk premia. The forecast uses the information in the past 240 months, starting from January 1984 and rolling forward to December 2003. Within each window ended at time t, we fit the following factor model with SPCA, SPCA-LS and PCA respectively,

$$\mathbf{y}_s = \mathbf{\Lambda}\mathbf{f}_s + \mathbf{u}_s, \quad s = t - 239, \ldots, t,$$

where $\mathbf{y}_s$ is the panel data of 131 macroeconomics variables. For all three methods, we set the number of factors $K = 8$. For SPCA and SPCA-LS, the sieve basis is chosen as the additive polynomial basis whose dimension J is chosen by 5-fold cross-validation.

Then we consider two models to predict the bond risk premia at time $s = t-238, \ldots, t$:

$$\text{Linear model:} \qquad z_s = \alpha + \boldsymbol{\beta}'\mathbf{W}_{s-1} + \epsilon_s, \tag{6.1}$$

$$\text{Multi-index model:} \qquad z_s = h(\boldsymbol{\psi}_1'\mathbf{W}_{s-1}, \ \cdots, \ \boldsymbol{\psi}_L'\mathbf{W}_{s-1}) + \epsilon_s, \tag{6.2}$$

where α is the intercept and h is a nonparametric function. The covariate $\mathbf{W}_{s-1}$ is either $\mathbf{f}_{s-1}$ or an augmented vector $(\mathbf{f}_{s-1}', \mathbf{x}_{s-1}')'$. Here, the latent factors $\mathbf{f}_{s-1}$ is estimated by either SPCA, SPCA-LS or PCA as described above. The multi-index model allows more general nonlinear forecasts and is estimated by the sliced inverse regression (Li, 1991). The number

of indices L is estimated by the ratio-based method suggested in Lam and Yao (2012) and is usually 2 or 3. We approximate h using an additive model $h(\boldsymbol{\psi}_1'\mathbf{W}_{s-1},\ \cdots,\ \boldsymbol{\psi}_L'\mathbf{W}_{s-1}) = \sum_{l=1}^{L} g_l(\boldsymbol{\psi}_l'\mathbf{W}_{s-1})$, which is the projection pursuit model (Friedman and Stuetzle, 1981). Each individual nonparametric function $g_l(\cdot)$ is smoothed by the local linear approximation.

After that, we predict z_{t+1} as:

$$\text{Linear predictor:} \qquad \widehat{z}_{t+1|t} = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}'\mathbf{W}_t, \tag{6.3}$$

$$\text{Multi-index predictor:} \qquad \widehat{z}_{t+1|t} = \sum_{l=1}^{\widehat{L}} \widehat{g}_l(\widehat{\boldsymbol{\psi}}_l'\mathbf{W}_t), \tag{6.4}$$

where $\widehat{\alpha}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\psi}}$, $\widehat{L}$ and $\widehat{g}_l(\cdot)$ are estimated from (6.1) and (6.2).

The forecast performance is assessed by the out-of-sample R^2, defined as

$$R^2 = 1 - \frac{\sum_{t=240}^{479} (z_{t+1} - \widehat{z}_{t+1|t})^2}{\sum_{t=240}^{479} (z_{t+1} - \bar{z}_t)^2},$$

where $\bar{z}_t$ is the sample mean of z_t over the sample period $[t-239, t]$. The R^2 of various methods are reported in Table 6.2. We notice that factors estimated by SPCA and SPCA-LS can explain more variations in bond excess returns with all maturities than the ones estimated by PCA. SPCA yields a 59.3% out-of-sample R^2 for forecasting the bond excess returns with two year maturity, which is much higher than the best out-of-sample predictor found in Ludvigson and Ng (2009). It is also observed that the forecast based on either SPCA or SPCA-LS cannot be improved by adding any covariate in $\mathbf{x}_t$. We argue that, in this application, the information of $\mathbf{x}_t$ should be mainly used as the explanatory power for the factors.

We summarize the observed results in the following aspects:

1. The factors estimated using additional covariates lead to significantly improved out-of-sample forecast on the US bond excess returns compared to the ones estimated by PCA.

2. As many series in the panel data are heavy-tailed, the robust-version of our method (SPCA) can result in improved out-of-sample forecasts.

3. The multi-index models yield significantly larger out-of-sample R^2's than those of the linear forecast models.

4. The observed covariates $\mathbf{x}_t$ (e.g. forward rates, employment and inflation) contain strong explanatory powers for the latent factors. The gain of forecasting bond excess returns is more substantial when these covariates are incorporated to estimate the common factors (using the proposed procedure) than directly used for forecasts.

Table 6.2: Forecast out-of-sample R^2 (%): the larger the better.

$\mathbf{W}_t$	SPCA				SPCA-LS				PCA			
	Maturity(Year)				Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5	2	3	4	5
	linear model											
$\mathbf{f}_t$	55.4	51.2	45.7	41.8	49.6	45.3	41.5	39.8	45.4	41.5	36.6	33.1
$(\mathbf{f}_t', \mathbf{x}_t')'$	54.7	50.5	45.4	41.4	49.3	44.9	40.8	38.7	46.2	42.4	37.1	33.7
	multi-index model											
$\mathbf{f}_t$	59.3	55.6	50.5	46.1	53.8	51.3	46.5	44.4	47.6	43.7	39.9	36.4
$(\mathbf{f}_t', \mathbf{x}_t')'$	58.9	54.8	50.1	45.5	53.1	50.9	45.6	42.2	48.9	44.1	40.2	37.0

7 Simulation Studies

7.1 Model settings

We use simulated examples to demonstrate the finite sample performance of the proposed method, which is denoted by SPCA (*smoothed PCA*), and compare it with SPCA-LS (which uses $\widetilde{\mathbf{\Sigma}}$, the least-squares based smoothed PCA, described in Section 3.1) and the benchmark PCA.

Consider the following data generating process,

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \quad \text{and} \quad \mathbf{f}_t = \tilde{\sigma}(g)\mathbf{g}^0(\mathbf{x}_t) + \tilde{\sigma}(\gamma)\boldsymbol{\gamma}_t^0, \quad t = 1, \cdots, T, \tag{7.1}$$

where $\mathbf{\Lambda}$ is drawn from i.i.d. standard Normal distribution and $\mathbf{u}_t$ is drawn from either the i.i.d standard Normal distribution or i.i.d. re-scaled Log-Normal distribution $c_1\{\exp(1 + 1.2\zeta) - c_2\}$, where $\zeta \sim \mathcal{N}(0, 1)$ and $c_1, c_2 > 0$ are chosen such that u_{it} has mean zero and variance 1. We set $K = 5$.

Here $\tilde{\sigma}(g)$ and $\tilde{\sigma}(\gamma)$ respectively represent the signal and noise levels. Set $\tilde{\sigma}(g)^2 + \tilde{\sigma}(\gamma)^2 = 1$ and $\tilde{\sigma}(g)^2/\tilde{\sigma}(\gamma)^2 = \omega$, where ω controls the ratio between the explained and unexplained parts in the latent factors. To address different signal-noise regimes, we set $\omega = 10, 1$ and 0.1 to represent strong, mild and weak explanatory powers respectively. The baseline $\boldsymbol{\gamma}_t^0$ is drawn from i.i.d. standard Normal distribution and the baseline function $\mathbf{g}^0(\cdot)$ is set to be one of the following two models:

(I) LINEAR MODEL: We set $d = K$ and $\mathbf{x}_t$ is drawn from i.i.d. standard Normal distribution. Let $\mathbf{g}^0(\mathbf{x}_t) = \mathbf{D}\mathbf{x}_t$, where $\mathbf{D}$ is a $K \times K$ matrix with each entry drawn from $U[1,\ 2]$;

(II) NONLINEAR MODEL: We set $d = 1$ and x_t is drawn from i.i.d. uniform distribution $[0, 1]$. Let $\mathbf{g}^0(x_t) = \{g_1^0(x_t), \cdots, g_K^0(x_t)\}'$ with $g_k^0(x_t) = a_k \cos(2\pi k x_t) + b_k \sin(2\pi k x_t)$ for $k = 1, \cdots K$. The coefficients a_k and b_k are calibrated from a nonlinear test function $\theta(x) = \sin(x) + 2\exp(-30x^2)$ with $x \in [0, 1]$ so that $\mathbf{g}^0$ forms its leading Fourier bases. To save the space, we refer to the example 2 of Dimatteo et al. (2001) for the plot of $\theta(x)$.

For each $k \leq K$, we normalize $g_k^0(\mathbf{x}_t)$ and $\gamma_{t,k}^0$ such that they have means zero, and standard deviations one.

Throughout this section, the number of factors is estimated by the eigen-ratio method. In the following simulated examples, the eigen-ratio method can correctly select $K = 5$ in most replications. The sieve basis is chosen as the additive polynomial basis whose dimension J is chosen by 5-fold cross-validation. To account the scale of the noise variance, we also consider the tuning parameter in the Huber loss to admit $\alpha_{T,i} = C_\alpha \tilde{\sigma}_i \sqrt{\frac{T}{\log(NT)}}$, where $\tilde{\sigma}_i = \sqrt{\frac{1}{T}\sum_t (y_{it} - \tilde{E}(y_{it}|\mathbf{x}_t))^2}$ and $\tilde{E}(y_{it}|\mathbf{x}_t)$ is smoothed by sieve least squares using additive polynomial basis of order 5. In Subsection 7.2, the tuning parameters C_α and J are selected by the in-sample 5-fold cross validation, while in subsection 7.3, they are chosen using the out-of-sample 5-fold cross validation.

7.2 In-sample Estimation

First, we compare the in-sample model fitting among SPCA, SPCA-LS and PCA under different scenarios. For each scenario, we conduct 200 replications. As the factors and loading may be estimated up to a rotation matrix, the canonical correlations between the parameter and its estimator can be used to measure the estimation accuracy (Bai, 2003). For Model (I) and (II) we report the sample mean of the median of 5 canonical correlations between the true loading and factors and the estimated ones.

The results are presented in Table 7.2. SPCA-LS and SPCA are comparable for light-tail distributions, and are both slightly better than PCA. This implies that we pay little price for the robustness and that the proposed estimators are potentially better than PCA when N is relatively small, due to the merit of the "finite-N" asymptotics of the proposed estimators. However, when the error distributions have heavy tails, SPCA yields much better estimation than other methods as expected. SPCA-LS out-performs PCA when $\mathbf{x}_t$

has strong or mild explanatory powers of $\mathbf{f}_t$ which is in line with the discussion in Section 4.3. When $\omega = 0.1$, the observed $\mathbf{x}_t$ is not as informative and hence the performance of SPCA and SPCA-LS are close to regular PCA.

7.3 Out-of-sample Forecast

We now consider using latent factors in a linear forecast model $z_{t+1} = \boldsymbol{\beta}'\mathbf{f}_t + \epsilon_{t+1}$, where ϵ_t is drawn from i.i.d. standard normal distribution. For each simulation, the unknown coefficients in $\boldsymbol{\beta}$ are independently drawn from uniform distribution $[0.5,\ 1.5]$ to cover a variety of model settings.

We conduct one-step ahead rolling window forecast using the linear model by estimating $\boldsymbol{\beta}$ and $\mathbf{f}_t$. The factors are estimated from (7.1) by SPCA, SPCA-LS or PCA. In each replication, we generate $T + 50$ observations in total. For $s = 1, \cdots, 50$, we use the T observations $(z_s, ..., z_{T+s-1})$ to forecast z_{T+s}. We use PCA as the benchmark and define the relative mean squared error (RelMSE) as:

$$\text{RelMSE} = \frac{\sum_{s=1}^{50}(\widehat{z}_{T+s|T+s-1} - z_{T+s})^2}{\sum_{s=1}^{50}(\tilde{z}^{PCA}_{T+s|T+s-1} - z_{T+s})^2},$$

where $\widehat{z}_{T+s|T+s-1}$ is the forecast of z_{T+s} based on either SPCA or SPCA-LS while $\tilde{z}^{PCA}_{T+s|T+s-1}$ is the forecast based on PCA. For each scenario, we simulate 200 replications and calculate the averaged RelMSE as a measurement of the one-step-ahead out-of-sample forecast.

The results are presented in Table 7.1. Again, when the tails of error distributions are light, SPCA and SPCA-LS perform comparably. But SPCA outperforms SPCA-LS when the errors have heavy tails. On the other hand, both SPCA and SPCA-LS outperform PCA when $\mathbf{x}_t$ exhibits strong or mild explanatory powers of $\mathbf{f}_t$, but are slightly worse when ω is small. In general, the SPCA method performs the best under heavy-tailed cases.

7.4 Compare with the interactive effect approach

Here we consider three pairs of sample sizes: $N = 40,\ T = 150$; $N = 60,\ T = 100$ and $N = 60,\ T = 150$. We compare the proposed SPCA method with SPCA-LS (Section 3.1), regular PCA and pure least squares (LS), which models the covariates and estimates the

Table 7.1: **Out-of-sample Forecast**: Mean RelMSE of forecast when $N = 40, T = 100$: the smaller the better (with PCA as the benchmark)

		Model (I)		Model (II)	
$\mathbf{u}_t$	ω	SPCA	SPCA-LS	SPCA	SPCA-LS
Normal	10	0.86	0.85	0.88	0.87
	1	0.91	0.91	0.92	0.92
	0.1	1.01	1.01	1.02	1.01
LogN	10	0.45	0.60	0.49	0.64
	1	0.52	0.62	0.51	0.66
	0.1	0.55	0.65	0.56	0.70

parameters by simply using

$$\min_{\mathbf{\Lambda},\{\mathbf{f}_t\},\boldsymbol{\beta}} \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{\Lambda}\mathbf{f}_t - \mathbf{x}_t'\boldsymbol{\beta}\|^2.$$

In Tables 7.2–7.3, we report sample mean of the median of 5 canonical correlations between the true loading and factors and the estimated ones. Under various sample size combinations, the findings are similar as discussed in Section 7.2: (1) both SPCA and SPCA-LS outperform PCA under light-tail distributions when $\mathbf{x}_t$ has strong or mild explanatory powers of $\mathbf{f}_t$; (2) when the error distributions have heavy tails, SPCA outperforms other methods as expected; (3)when $\mathbf{x}_t$ has weak explanatory power, the performance of SPCA and SPCA-LS are close to regular PCA; (4) under all simulated scenarios, the LS approach gives the worst estimation performance.

7.5 Serial dependent case

In this subsection, we compare the in-sample model fitting among SPCA, SPCA-LS and PCA under serial dependences. The simulation settings are similar as in Section 7.1 except both $\mathbf{x}_t$ and $\boldsymbol{\gamma}_t$ are generated from a stationary VAR(1) model as follows

$$\mathbf{x}_t = \mathbf{\Pi}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\gamma}_t = \mathbf{\Pi}\boldsymbol{\gamma}_{t-1} + \eta_t, \quad t = 1, \cdots, T,$$

with $\mathbf{x}_0 = \mathbf{0}$ and $\boldsymbol{\gamma}_0 = \mathbf{0}$. The (i,j)th entry of $\mathbf{\Pi}$ is set to be 0.5 when $i = j$ and $0.4^{|i-j|}$ when $i \neq j$. In addition, $\boldsymbol{\varepsilon}_t$ and η_t are drawn form i.i.d. $N(\mathbf{0}, \mathbf{I})$.

The performance under 200 replications are presented in Table 7.4 below. Our numerical findings for the independent data continue to hold for serially dependent data: both SPCA and SPCA-LS outperform PCA when $\mathbf{x}_t$ and $\mathbf{f}_t$ are serially correlated. SPCA gives the

best performance when the error distributions are heavy-tailed.

7.6 Conditional week cross-sectional dependent case

In this subsection, we compare the in-sample model fitting among SPCA, SPCA-LS and PCA under conditional week cross-sectional dependency. The simulation settings are similar as in Section 7.1 except we allow $\mathbf{u}_t$ to be cross-sectional dependent. First we generate $\mathbf{v}_t$ from i.i.d. re-scaled Log-Normal distribution as introduced in Section 7.1. Then we generate $\mathbf{u}_t = \mathbf{\Omega}_u^{1/2}\mathbf{v}_t$, where $\mathbf{\Omega}_u$ is a correlation matrix whose (i,j)th entry equals $0.5^{|i-j|}$. We set $N = 60$ and $T = 150$. The performance under 200 replications are presented in Table 7.5 below. Similar are the results with cross-sectional independent errors, SPCA give the best performance among the three methods.

Table 7.2: **In-sample Estimation**: Median of 5 canonical correlations of the estimated loadings/factors and the true ones when $N = 40, T = 100$: the larger the better

			Model (I)			Model (II)		
	$\mathbf{u}_t$	ω	SPCA	SPCA-LS	PCA	SPCA	SPCA-LS	PCA
Loadings	Normal	10	0.91	0.91	0.82	0.90	0.90	0.75
		1	0.88	0.89	0.82	0.84	0.84	0.75
		0.1	0.83	0.83	0.82	0.77	0.79	0.75
	LogN	10	0.81	0.50	0.36	0.77	0.48	0.31
		1	0.77	0.45	0.36	0.73	0.42	0.31
		0.1	0.72	0.41	0.36	0.70	0.39	0.31
Factors	Normal	10	0.90	0.90	0.74	0.90	0.90	0.72
		1	0.82	0.83	0.74	0.81	0.82	0.72
		0.1	0.75	0.76	0.74	0.74	0.74	0.72
	LogN	10	0.83	0.54	0.31	0.81	0.57	0.26
		1	0.80	0.53	0.31	0.77	0.50	0.26
		0.1	0.75	0.48	0.31	0.74	0.46	0.26

7.7 Testing the explanatory power

We now study the performance of the proposed test statistic under various scenarios. Consider the following data generating process,

$$\mathbf{y}_t \;=\; \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \quad \text{and} \quad \mathbf{f}_t = \mathbf{D}\mathbf{x}_t + \delta\boldsymbol{\gamma}_t, \quad t = 1, \cdots, T, \tag{7.2}$$

where $\mathbf{\Lambda}$ and $\boldsymbol{\gamma}_t$ are drawn from i.i.d. standard Normal distribution, $\mathbf{D}$ is a $K \times K$ matrix with each entry drawn from $U[1,\ 2]$ and δ is a constant which can be set as either 0 or 1.

Table 7.3: **In-sample Estimation**: Median of 5 canonical correlations of the estimated loadings/factors and the true ones when $N = 60, T = 150$: the larger the better

			Model (I)				Model (II)			
	$\mathbf{u}_t$	ω	SPCA	SPCA-LS	PCA	LS	SPCA	SPCA-LS	PCA	LS
Loading	Normal	10	0.95	0.95	0.88	0.82	0.93	0.93	0.85	0.78
		1	0.92	0.92	0.88	0.83	0.88	0.88	0.85	0.79
		0.1	0.85	0.86	0.88	0.86	0.84	0.84	0.85	0.83
	LogN	10	0.86	0.59	0.44	0.38	0.84	0.55	0.41	0.34
		1	0.83	0.55	0.44	0.40	0.80	0.52	0.41	0.36
		0.1	0.79	0.48	0.44	0.43	0.75	0.44	0.41	0.39
Factors	Normal	10	0.94	0.94	0.83	0.75	0.91	0.91	0.81	0.74
		1	0.86	0.86	0.83	0.78	0.83	0.83	0.81	0.76
		0.1	0.81	0.82	0.83	0.82	0.79	0.79	0.81	0.79
	LogN	10	0.85	0.66	0.40	0.33	0.84	0.64	0.37	0.30
		1	0.81	0.60	0.40	0.35	0.80	0.61	0.37	0.32
		0.1	0.77	0.54	0.40	0.38	0.75	0.56	0.37	0.35

Table 7.4: **Dependent data**: Median of canonical correlations of the estimated loadings/factors and the true ones when $N = 40, T = 100$: the larger the better

			Model (I)			Model (II)		
	$\mathbf{u}_t$	ω	SPCA	SPCA-LS	PCA	SPCA	SPCA-LS	PCA
Loadings	Normal	10	0.89	0.90	0.78	0.87	0.87	0.73
		1	0.84	0.84	0.78	0.82	0.82	0.73
		0.1	0.80	0.81	0.78	0.76	0.77	0.73
	LogN	10	0.75	0.47	0.25	0.73	0.45	0.22
		1	0.69	0.41	0.25	0.69	0.39	0.22
		0.1	0.64	0.38	0.25	0.62	0.35	0.22
Factors	Normal	10	0.88	0.89	0.71	0.88	0.88	0.68
		1	0.81	0.82	0.71	0.80	0.81	0.68
		0.1	0.73	0.74	0.71	0.72	0.72	0.68
	LogN	10	0.80	0.59	0.24	0.78	0.55	0.19
		1	0.74	0.51	0.24	0.72	0.49	0.19
		0.1	0.70	0.45	0.24	0.69	0.40	0.19

Table 7.5: **Cross-sectional Dependent error**: Median of canonical correlations of the estimated loadings/factors and the true ones when $N = 60, T = 150$: the larger the better

		Model (I)			Model (II)		
	ω	SPCA	SPCA-LS	PCA	SPCA	SPCA-LS	PCA
Loadings	10	0.81	0.56	0.42	0.79	0.51	0.38
	1	0.76	0.53	0.42	0.74	0.49	0.38
	0.1	0.72	0.47	0.42	0.71	0.40	0.38
Factors	10	0.80	0.63	0.40	0.79	0.60	0.36
	1	0.75	0.57	0.40	0.73	0.57	0.36
	0.1	0.73	0.50	0.40	0.71	0.52	0.36

Further, $\mathbf{x}_t$ and $\mathbf{u}_t$ are jointly generated from one of the following three cases.

(I) INDEPENDENT CASE: $\mathbf{x}_t$ is drawn from i.i.d. standard Normal distribution. $\mathbf{u}_t$ is independent of $\mathbf{x}_t$ and drawn from i.i.d. re-scaled Log-Normal distribution $c_1\{\exp(1+1.2\zeta) - c_2\}$, where $\zeta \sim \mathcal{N}(0,1)$ and $c_1, c_2 > 0$ are chosen such that u_{it} has mean zero and variance 1.

(II) SERIAL DEPENDENT CASE: Both $\mathbf{x}_t$ and $\mathbf{u}_t$ are generated from stationary VAR(1) models as follows

$$\mathbf{x}_t = \mathbf{\Pi}^{(1)}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \mathbf{u}_t = \mathbf{\Pi}^{(2)}\mathbf{u}_{t-1} + \eta_t, \quad t = 1, \cdots, T,$$

with $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{u}_0 = \mathbf{0}$. In both $\mathbf{\Pi}^{(1)}$ and $\mathbf{\Pi}^{(2)}$, the (i,j)th entry is set to be 0.5 when $i = j$ and $0.4^{|i-j|}$ when $i \neq j$. In addition, $\boldsymbol{\varepsilon}_t$ is drawn form i.i.d. standard Normal distribution and η_t is drawn from i.i.d. re-scaled Log-Normal distribution as described in (I).

(III) CROSS-SECTIONAL DEPENDENT CASE: First we generate $\mathbf{x}_t^*$ and $\mathbf{u}_t^*$ similar as in (I). Then we generate

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{pmatrix} = \mathbf{\Omega}^{1/2} \begin{pmatrix} \mathbf{x}_t^* \\ \mathbf{u}_t^* \end{pmatrix},$$

where $\mathbf{\Omega}$ is a correlation matrix whose diagonal entries equal 1 and off-diagonal entries equal 0.5.

Let $N = 50$, $T = 200$ and $K = 3$. We set the significance level $\alpha = 0.05$ and repeat the testing on explanatory power of covariates over 1000 replications. For the first 500 replications, we set $\delta = 0$ and hence the null hypothesis $H_0 : \text{cov}(\boldsymbol{\gamma}_t) = 0$ is true. For the rest 500 replications, we set $\delta = 1$ and hence the null hypothesis is false. The number of factors K is estimated by the eigen-ratio method suggested in Lam and Yao (2012). The sieve basis is chosen as the additive polynomial basis whose dimension J is chosen by 5-fold cross-validation. The false positive rate and true negative rate are reported in Table 7.6.

We see observe size distortions as dependence is presence, and the power is also affected. As a result, modified covariance estimators should be used in place of $\widehat{\mathbf{\Sigma}}_u$, such as the sparse covariance estimator.

Table 7.6: Size and power of test

	Independent	Serial Dependent	Cross-sectional dependent
Size	0.048	0.066	0.082
Power	0.926	0.892	0.848

8 Conclusions

We study factor models when the factors depend on observed explanatory characteristics. The proposed method incorporates the explanatory power of these observed covariates, and is robust to possibly heavy-tailed distributions. We focus on the case $\dim(\mathbf{x}_t)$ is finite, and on the rates of convergence for the estimated factors and loadings. Under various signal-noise ratios, substantial improved rates of convergence can be gained.

Related to the above, the idea could be easily extended to the case that $\dim(\mathbf{x}_t)$ is slowly growing (with respect to (N, T)). On the other hand, allowing $\dim(\mathbf{x}_t)$ to be fast-growing would require some dimension-reduction treatment combined with covariate selections. In addition, selecting the covariates would be also useful as the quality of the signal is crucial. We shall leave these open questions for future studies.

9 Acknowledgement

Fan gratefully acknowledges the support of NSF grant DMS-1712591.

A Proof of Theorem 2.1

Proof. Let $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_N$ be the eigenvectors of $\boldsymbol{\Sigma}_{y|x}$, corresponding to the eigenvalues $\lambda_1(\boldsymbol{\Sigma}_{y|x}) \geq \lambda_2(\boldsymbol{\Sigma}_{y|x}) ... \geq \lambda_N(\boldsymbol{\Sigma}_{y|x})$. Due to $\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{f|x}\boldsymbol{\Lambda}'$, and by the assumption that $\lambda_{\min}(\boldsymbol{\Sigma}_{f|x}) > 0$, the rank of $\boldsymbol{\Sigma}_{y|x}$ equals K. Hence $\lambda_i(\boldsymbol{\Sigma}_{y|x}) = 0$ for all $i > K$.

(i) Let $\mathbf{L} = \boldsymbol{\Sigma}_{\Lambda,N}^{1/2}\boldsymbol{\Sigma}_{f|x}\boldsymbol{\Sigma}_{\Lambda,N}^{1/2}$. Let $\mathbf{M}$ be a $K \times K$ matrix, whose columns are the eigenvectors of $\mathbf{L}$. Then $\mathbf{D} := \mathbf{M}'\mathbf{L}\mathbf{M}$ is a diagonal matrix, with diagonal elements being the eigenvalues of $\mathbf{L}$. Let $\mathbf{H} = \boldsymbol{\Sigma}_{\Lambda,N}^{-1/2}\mathbf{M}$. Then

$$\frac{1}{N}\boldsymbol{\Sigma}_{y|x}\boldsymbol{\Lambda}\mathbf{H} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{f|x}\boldsymbol{\Sigma}_{\Lambda,N}\mathbf{H} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\Lambda,N}^{-1/2}\mathbf{L}\mathbf{M} \underbrace{=}_{\mathbf{M}\mathbf{M}'=\mathbf{I}} \boldsymbol{\Lambda}\mathbf{H}\mathbf{M}'\mathbf{L}\mathbf{M} = \boldsymbol{\Lambda}\mathbf{H}\mathbf{D}.$$

In addition, $(\boldsymbol{\Lambda}\mathbf{H})'(\boldsymbol{\Lambda}\mathbf{H}) = N\mathbf{M}'\mathbf{M} = N\mathbf{I}_K$, hence the columns of $\boldsymbol{\Lambda}\mathbf{H}/\sqrt{N}$ are the eigenvectors of $\boldsymbol{\Sigma}_{y|x}$, corresponding to the K nonzero eigenvalues.

(ii) From $E(\mathbf{y}_t|\mathbf{x}_t) = \boldsymbol{\Lambda}\mathbf{g}(\mathbf{x}_t)$, we have $(\boldsymbol{\Lambda}\mathbf{H})'E(\mathbf{y}_t|\mathbf{x}_t) = (\boldsymbol{\Lambda}\mathbf{H})'\boldsymbol{\Lambda}\mathbf{H}\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)$. This leads to the desired expression of $\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)$.

(iii) The nonzero eigenvalues of $\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{f|x}\boldsymbol{\Lambda}'$ equal those of

$$\boldsymbol{\Sigma}_{f|x}^{1/2}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{f|x}^{1/2} = N\boldsymbol{\Sigma}_{f|x}^{1/2}\boldsymbol{\Sigma}_{\Lambda,N}\boldsymbol{\Sigma}_{f|x}^{1/2},$$

which are also the same as those of $N\boldsymbol{\Sigma}_{\Lambda,N}^{1/2}\boldsymbol{\Sigma}_{f|x}\boldsymbol{\Sigma}_{\Lambda,N}^{1/2} = N\mathbf{L}$. Note that

$$\lambda_{\min}(N\mathbf{L}) \geq N\lambda_{\min}(\boldsymbol{\Sigma}_{f|x})\lambda_{\min}(\boldsymbol{\Sigma}_{\Lambda,N}) \geq N\chi_N\underline{c}_\Lambda.$$

□

B Proofs for Section 4

Here we present the main body of the proof, and refer to the supplementary material for additional technical lemmas.

B.1 A bird's-eye view of the major technical steps

We first provide a bird's-eye view of the major steps in the proof. The key intermediate result is to prove the following Bahadur representation of the estimated eigenvectors:

$$\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\mathbf{H} = \frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\Lambda}\mathbf{g}(\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\sum_{i=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\Phi(\mathbf{x}_s)'\dot{\rho}(\alpha_T^{-1}e_{is})\alpha_T\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1}$$

$$+\widetilde{\mathbf{\Delta}}(\{\mathbf{x}_t, \mathbf{e}_t\}_{t\leq T}), \tag{B.1}$$

for some invertible matrix $\mathbf{H}$. Here the first term on the right hand side is the leading term that results in the presented rate of convergence in Theorem 4.1, where $\dot{\rho}(\cdot)$ denotes the derivative of Huber's loss function; $\widetilde{\mathbf{V}}$ is a K-dimensional diagonal matrix of the eigenvalues of $\widehat{\mathbf{\Sigma}}/N$. The second term $\widetilde{\mathbf{\Delta}}(\{\mathbf{x}_t, \mathbf{e}_t\}_{t\leq T})$ is a higher order random term that depends on both $\{\mathbf{x}_t\}$ and $\{\mathbf{e}_t\}$, where $\mathbf{e}_t = \mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t) = (e_{1t}, ..., e_{Nt})$.

To have an general idea of how we prove (B.1), recall that $\widehat{\mathbf{\Sigma}}/N := \frac{1}{TN}\sum_{t=1}^{T} \widehat{E}(\mathbf{y}_t|\mathbf{x}_t)\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)'$, where each element of $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ is $\widehat{E}(y_{it}|\mathbf{x}_t) = \widehat{\mathbf{b}}_i'\Phi(\mathbf{x}_t)$ with $\widehat{\mathbf{b}}_i$ being the M-estimator of the sieve coefficients of $E(y_{it}|\mathbf{x}_t)$, obtained by minimizing the Huber's loss:

$$\widehat{\mathbf{b}}_i = \arg\min_{\mathbf{b}\in\mathbb{R}^J} Q_i(\mathbf{b}), \qquad Q_i(\mathbf{b}) = \frac{1}{T}\sum_{t=1}^{T} \alpha_T^2 \rho\left(\frac{y_{it} - \Phi(\mathbf{x}_t)'\mathbf{b}}{\alpha_T}\right).$$

Then by the definition of $\widehat{\mathbf{\Lambda}}$,

$$\frac{1}{N}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}. \tag{B.2}$$

The above is the key equality we shall use to derive (B.1). To use this equality, we need to obtain the Bahadur representations of $\widehat{\mathbf{b}}_i$ and $\widehat{E}(y_{it}|\mathbf{x}_t)$ in the following steps.

Step 1: bias of sieve coefficients. Define, for $i = 1, ..., N$,

$$\mathbf{b}_i := \arg\min_{\mathbf{b}\in\mathbb{R}^J} E[y_{it} - \mathbf{b}'\Phi(\mathbf{x}_t)]^2, \quad \mathbf{b}_{i,\alpha} = \arg\min_{\mathbf{b}\in\mathbb{R}^J} E\alpha_T^2 \rho\left(\frac{y_{it} - \Phi(\mathbf{x}_t)'\mathbf{b}}{\alpha_T}\right).$$

Note that the sieve expansion of $E(\mathbf{y}_t|\mathbf{x}_t)$ is $\mathbf{b}_i'\Phi(\mathbf{x}_t)$ (to be proved in Lemma C.1). But $\widehat{\mathbf{b}}_i$ is biased for estimating $\mathbf{b}_i$, and asymptotically converges to $\mathbf{b}_{i,\alpha}$. As $\alpha_T \to \infty$, $\mathbf{b}_{i,\alpha}$ is expected to converge to $\mathbf{b}_i$ uniformly in $i \leq N$. This is true given some moment conditions on $\mathbf{e}_t := \mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t)$.

Step 2: Expansion of $\widehat{\mathbf{b}}_i - \mathbf{b}_{i,\alpha}$.

The first order condition gives $\nabla Q_i(\widehat{\mathbf{b}}_i) = 0$. But we cannot directly expand this equation because ∇Q_i is not differentiable. As in many M-estimations, define $\bar{Q}_i(\mathbf{b}) = EQ_i(\mathbf{b})$, and $\boldsymbol{\mu}_i(\mathbf{b}) = \nabla Q_i(\mathbf{b}) - \nabla\bar{Q}_i(\mathbf{b})$. So we have

$$0 = \nabla\bar{Q}_i(\widehat{\mathbf{b}}_i) - \boldsymbol{\mu}_i(\widehat{\mathbf{b}}_i),$$

and $\nabla\bar{Q}_i$ is differentiable. We shall apply the standard empirical process theory for independent data (the symmetrization and contraction theorems, e.g., Bühlmann and van de Geer (2011)) to prove the stochastic equicontinuity of $\boldsymbol{\mu}_i(\mathbf{b})$ and thus the convergence of

$\max_i \|\boldsymbol{\mu}_i(\widehat{\mathbf{b}}_i) - \boldsymbol{\mu}_i(\mathbf{b}_{i,\alpha})\|$. This will eventually lead to an expansion of $\widehat{\mathbf{b}}_i - \mathbf{b}_{i,\alpha}$, to be given in Lemma C.3.

Step 3: Expansion of $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t) - E(\mathbf{y}_t|\mathbf{x}_t)$. Combining steps 1 and 2 will eventually lead to

$$\widehat{E}(y_{it}|\mathbf{x}_t) = E(y_{it}|\mathbf{x}_t) + \Phi(\mathbf{x}_t)'\mathbf{A}\frac{1}{T}\sum_{s=1}^{T}\alpha_T\dot{\rho}(\alpha_T^{-1}e_{is})\Phi(\mathbf{x}_s) + R_{it} \tag{B.3}$$

where R_{it} is a high-order remainder term that depends on $\mathbf{x}_t$, and $\mathbf{A} = (2E\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)')^{-1}$ is the Hessian matrix. We shall bound $\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}R_{it}$ in Proposition C.3.

Step 4: Expansion of $\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\mathbf{H}$. Substituting the expansion of $\widehat{E}(y_{it}|\mathbf{x}_t)$ to (B.3), with $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ replaced by its expansions, we will eventually obtain (B.1). Then (B.1) can be directly applied to obtain the rate of convergence for the estimated loadings. This will be done in Section C.2, where we show that the remainder term is of a smaller order than the leading term.

Importantly, both the signal strength $\chi_N = \lambda_{\min}(\boldsymbol{\Sigma}_{f|x})$ and the "noise" $\text{cov}(\boldsymbol{\gamma}_t)$ plays an essential role in (B.2), which are to be reflected in the rate of convergence.

B.2 Estimating the loadings

Throughout the proofs, as $T, J \to \infty$, N either grows or stays constant.

Write $\mathbf{M}_\alpha$ be an $N \times J$ matrix, whose ith row is given by

$$\mathbf{M}_{i,\alpha}' := \frac{1}{T}\sum_{s=1}^{T}\alpha_T\dot{\rho}(\alpha_T^{-1}e_{is})\Phi(\mathbf{x}_s)'.$$

Write $\mathbf{R}_t = (R_{1t}, ..., R_{Nt})'$, where R_{it} was defined in Proposition C.3. Then the Bahadur representation in Proposition C.3 can be written in the vector form: $\mathbf{A} = (2E\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)')^{-1}$,

$$\widehat{E}(\mathbf{y}_t|\mathbf{x}_t) = E(\mathbf{y}_t|\mathbf{x}_t) + \mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t) + \mathbf{R}_t = \boldsymbol{\Lambda}E(\mathbf{f}_t|\mathbf{x}_t) + \mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t) + \mathbf{R}_t. \tag{B.4}$$

Let $\widetilde{\mathbf{V}}$ be a $K \times K$ diagonal matrix, whose diagonal elements are the first K eigenvalues of $\widehat{\boldsymbol{\Sigma}}/N := \frac{1}{TN}\sum_{t=1}^{T}\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)'$. By the definition of $\widehat{\boldsymbol{\Lambda}}$, $\frac{1}{N}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}$. Plugging in (B.4), with $\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)'$ we have,

$$\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\mathbf{H} = \sum_{i=1}^{8}\mathbf{B}_i, \qquad \mathbf{H} = \frac{1}{TN}\sum_{t=1}^{T}E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1} \tag{B.5}$$

where for $\mathbf{A} = (2E\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)')^{-1}$,

$$\begin{aligned}
\mathbf{B}_1 &= \frac{1}{TN}\sum_{t=1}^{T}\mathbf{\Lambda}E(\mathbf{f}_t|\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \qquad \mathbf{B}_2 = \frac{1}{TN}\sum_{t=1}^{T}\mathbf{\Lambda}E(\mathbf{f}_t|\mathbf{x}_t)\mathbf{R}_t'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \\
\mathbf{B}_3 &= \frac{1}{TN}\sum_{t=1}^{T}\mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\mathbf{\Lambda}'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1} \qquad \mathbf{B}_4 = \frac{1}{TN}\sum_{t=1}^{T}\mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \\
\mathbf{B}_5 &= \frac{1}{TN}\sum_{t=1}^{T}\mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)\mathbf{R}_t'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \qquad \mathbf{B}_6 = \frac{1}{TN}\sum_{t=1}^{T}\mathbf{R}_tE(\mathbf{f}_t|\mathbf{x}_t)'\mathbf{\Lambda}'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \\
\mathbf{B}_7 &= \frac{1}{TN}\sum_{t=1}^{T}\mathbf{R}_t\Phi(\mathbf{x}_t)'\mathbf{A}\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \qquad \mathbf{B}_8 = \frac{1}{TN}\sum_{t=1}^{T}\mathbf{R}_t\mathbf{R}_t'\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1}.
\end{aligned}$$

We derive the rates of convergence by examining each term of (B.5).

B.2.1 Proof of Theorem 4.1: $\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2$

Proposition B.1. *Suppose $J^3\log^2 N = O(T)$, $\eta \geq 2$, and $J^2/T + J^{-\eta} \ll \chi_N$. Then*

$$\frac{1}{N}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H}\|_F^2 = O_P(\frac{J}{T} + J^{1-2\eta})\chi_N^{-1}.$$

Proof. From Lemma C.5 and Proposition C.3, we obtain

$$\begin{aligned}
\frac{1}{N}\|\mathbf{M}_\alpha\|^2 + \max_i \frac{1}{T}\sum_{t=1}^{T}R_{it}^2 &= O_P(\frac{J}{T} + J^{1-2\eta} + \alpha_T^{-2(\zeta_1-1)}\frac{J^3\log N}{T} + \frac{J^3\log N\log J}{T^2}) \\
&\leq O_P(\frac{J}{T} + J^{1-2\eta})
\end{aligned}$$

under the assumption $J^3\log^2 N = O(T)$, $\alpha_T = C\sqrt{T/\log(NJ)}$ and $\zeta_1 > 2$. Hence from Lemma C.5 and Proposition C.3,

$$\begin{aligned}
\frac{1}{N}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H}\|_F^2 &= O_P(\frac{1}{N}\sum_{i=1}^{8}\|\mathbf{B}_i\|_F^2) \\
&= O_P(\frac{1}{N}\|\mathbf{M}_\alpha\|^2 J\max_i\frac{1}{T}\sum_{t=1}^{T}R_{it}^2/\chi_N^2) \\
&\quad + O_P(\frac{1}{N}\|\mathbf{M}_\alpha\|^2\chi_N^{-1} + \frac{1}{N}\|\mathbf{M}_\alpha\|_F^4/(N\chi_N^2)) \\
&\quad + O_P(\max_i\frac{1}{T}\sum_{t=1}^{T}R_{it}^2\chi_N^{-1} + (\max_i\frac{1}{T}\sum_{t=1}^{T}R_{it}^2)^2/\chi_N^2) \\
&\leq O_P(\chi_N^{-2}J)(\frac{1}{N}\|\mathbf{M}_\alpha\|^2 + \max_i\frac{1}{T}\sum_{t=1}^{T}R_{it}^2)^2
\end{aligned}$$

$$
\begin{aligned}
&+O_P(\chi_N^{-1})(\frac{1}{N}\|\mathbf{M}_\alpha\|^2+\max_i\frac{1}{T}\sum_{t=1}^T R_{it}^2)\\
\leq\ & O_P(\frac{J}{T}+J^{1-2\eta})\chi_N^{-1}[1+(\frac{J}{T}+J^{1-2\eta})\chi_N^{-1}J]\\
\leq\ & O_P(\frac{J}{T}+J^{1-2\eta})\chi_N^{-1}.
\end{aligned}
$$

The last equality is due to $(\frac{J}{T}+J^{1-2\eta})\chi_N^{-1}J=O(1)$, granted by $\eta\geq 2$, and $J^2/T+J^{-\eta}\ll \chi_N$. Q.E.D.

B.2.2 Proof of Theorem 4.1: $\max_{i\leq N}\|\boldsymbol{\lambda}_i-\mathbf{H}'\boldsymbol{\lambda}_i\|$

$$
\begin{aligned}
\mathbf{B}_1 &= \frac{1}{TN}\sum_{t=1}^T \boldsymbol{\Lambda}E(\mathbf{f}_t|\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\mathbf{M}_\alpha'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \qquad \mathbf{B}_2=\frac{1}{TN}\sum_{t=1}^T \boldsymbol{\Lambda}E(\mathbf{f}_t|\mathbf{x}_t)\mathbf{R}_t'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1},\\
\mathbf{B}_3 &= \frac{1}{TN}\sum_{t=1}^T \mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1} \qquad \mathbf{B}_4=\frac{1}{TN}\sum_{t=1}^T \mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\mathbf{M}_\alpha'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1},\\
\mathbf{B}_5 &= \frac{1}{TN}\sum_{t=1}^T \mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)\mathbf{R}_t'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \qquad \mathbf{B}_6=\frac{1}{TN}\sum_{t=1}^T \mathbf{R}_tE(\mathbf{f}_t|\mathbf{x}_t)'\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1},\\
\mathbf{B}_7 &= \frac{1}{TN}\sum_{t=1}^T \mathbf{R}_t\Phi(\mathbf{x}_t)'\mathbf{A}\mathbf{M}_\alpha'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1}, \qquad \mathbf{B}_8=\frac{1}{TN}\sum_{t=1}^T \mathbf{R}_t\mathbf{R}_t'\widehat{\boldsymbol{\Lambda}}\widetilde{\mathbf{V}}^{-1}.
\end{aligned}
$$

Proof. By Lemma C.9 $\max_{i\leq N}\|\mathbf{M}_{i,\alpha}\| = O_P(J^{-\eta}\sqrt{J}+\sqrt{J(\log N)/T})$. Let $\mathbf{B}_{i1},...,\mathbf{B}_{i8}$ respectively denote the ith row of $\mathbf{B}_1,...,\mathbf{B}_8$. We have

$$
\begin{aligned}
\max_i\|\mathbf{B}_{i1}\| &\leq \chi_N^{-1/2}O_P(\|\mathbf{M}_\alpha\widehat{\boldsymbol{\Lambda}}\|/N)\leq O_P(\chi_N^{-1/2}\max_i\|\mathbf{M}_{i,\alpha}\|)\\
\max_i\|\mathbf{B}_{i2}\| &\leq \chi_N^{-1/2}O_P(\max_i\frac{1}{T}\sum_{t=1}^T R_{it}^2)^{1/2})\\
\max_i\|\mathbf{B}_{i3}\| &\leq \chi_N^{-1/2}O_P(\max_i\|\mathbf{M}_{i,\alpha}\|)=O_P(J^{-\eta}\sqrt{J}+\sqrt{J(\log N)/T})\chi_N^{-1/2}\\
\max_i\|\mathbf{B}_{i4}\| &\leq O_P(\max_i\|\mathbf{M}_{i,\alpha}\|)O_P(\|\mathbf{M}_\alpha'\widehat{\boldsymbol{\Lambda}}\|/N)\chi_N^{-1}\\
\max_i\|\mathbf{B}_{i5}\| &\leq O_P(\max_i\|\mathbf{M}_{i,\alpha}\|)O_P(\sqrt{J}\max_i\frac{1}{T}\sum_t R_{it}^2)^{1/2}\chi_N^{-1}\\
\max_i\|\mathbf{B}_{i6}\| &\leq O_P(\max_i\frac{1}{T}\sum_t R_{it}^2)^{1/2})\chi_N^{-1/2}\\
\max_i\|\mathbf{B}_{i7}\| &\leq O_P(\max_i\frac{1}{T}\sum_t R_{it}^2)^{1/2}\sqrt{J})O_P(\|\mathbf{M}_\alpha\widehat{\boldsymbol{\Lambda}}\|/N)\chi_N^{-1}\\
\max_i\|\mathbf{B}_{i8}\| &\leq O_P(\max_i\frac{1}{T}\sum_t R_{it}^2\chi_N^{-1}).
\end{aligned}
$$

Hence

$$\begin{aligned}
&\max_{i\leq N}\|\boldsymbol{\lambda}_i-\mathbf{H}'\boldsymbol{\lambda}_i\| \leq O_P(\max_i\|\mathbf{B}_{i2}\|+\max_i\|\mathbf{B}_{i3}\|)\\
=\;& O_P(J^{-\eta}\sqrt{J}+\sqrt{J(\log N)/T}+\alpha_T^{-(\zeta_1-1)}\sqrt{\frac{J^3\log N}{T}})\chi_N^{-1/2}\\
=\;& O_P(J^{-\eta}\sqrt{J}+\sqrt{J(\log N)/T})\chi_N^{-1/2},
\end{aligned}$$

where the last equality follows from

$$\alpha_T^{-(\zeta_1-1)}\sqrt{\frac{J^3\log N}{T}}=O(\sqrt{\frac{J\log N}{T}})$$

under assumptions $(\log N)^2J^3=O(T)$ and $\zeta_1>2$.

B.3 Proof of Theorem 4.2: factors

Recall that $\widehat{\mathbf{g}}(\mathbf{x}_t)=\frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$. By (B.4), $\widehat{\mathbf{g}}(\mathbf{x}_t)-\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)=\sum_{i=1}^4\mathbf{C}_{ti}$, where

$$\begin{aligned}
\mathbf{C}_{t1} &= \frac{1}{N}(\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\mathbf{H})'(\boldsymbol{\Lambda}\mathbf{H}-\widehat{\boldsymbol{\Lambda}})\mathbf{H}^{-1}E(\mathbf{f}_t|\mathbf{x}_t), \qquad \mathbf{C}_{t3}=\frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t),\\
\mathbf{C}_{t2} &= -\frac{1}{N}\mathbf{H}'\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\mathbf{H})\mathbf{H}^{-1}E(\mathbf{f}_t|\mathbf{x}_t), \qquad \mathbf{C}_{t4}=\frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\mathbf{R}_t.
\end{aligned}$$

The convergence of $\frac{1}{T}\sum_{t=1}^T\|\widehat{\mathbf{g}}(\mathbf{x}_t)-\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2$ in this theorem is proved in the following proposition.

Proposition B.2. *As $T\to\infty$ and N either grows or stays constant,*

$$\frac{1}{T}\sum_{t=1}^T\|\widehat{\mathbf{g}}(\mathbf{x}_t)-\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2=O_P(\frac{J^2}{T^2}\chi_N^{-1}+\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T}+J^{1-2\eta}+\frac{J}{TN}+\frac{J^3\log^2N}{T^2}).$$

Proof. Recall

$$a_T^2:=\frac{J}{T}+J^{1-2\eta},\quad b_{NT}^2:=\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T}+\frac{J}{TN}+\frac{J}{T}\alpha_T^{-\zeta_2}.$$

By Lemma C.8, $\|\mathbf{H}\|=O_P(1)=\|\mathbf{H}^{-1}\|$. Also, by Proposition B.1 and Lemmas C.6, C.7,

$$\begin{aligned}
&\frac{1}{N}\|\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\mathbf{H}\|_F^2=O_P(a_T^2\chi_N^{-1}).\\
&\frac{1}{N}\|\mathbf{M}_\alpha'\widehat{\boldsymbol{\Lambda}}\|_F=O_P(a_T^2\chi_N^{-1/2})+O_P(b_{NT})
\end{aligned}$$

$$\|\frac{1}{N}\mathbf{\Lambda}'(\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}\mathbf{H})\| \leq O_P(\chi_N^{-1/2})(\frac{1}{N}\|\mathbf{M}'_\alpha\widehat{\mathbf{\Lambda}}\|_F + (\max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2)^{1/2})$$

Therefore, as $\frac{1}{T}\sum_t \|E(\mathbf{f}_t|\mathbf{x}_t)\|^2 = O_P(\chi_N)$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{t1}\|^2 &\leq O_P(1)[\frac{1}{N}\|\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}\mathbf{H}\|^2]^2\chi_N \leq O_P(a_T^4\chi_N^{-1}) \\
\frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{t2}\|^2 &\leq O_P(1)[\frac{1}{N}\mathbf{\Lambda}'(\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}\mathbf{H})]^2\chi_N \\
&\leq O_P(a_T^4\chi_N^{-1} + b_{NT}^2 + \max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2) \\
\frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{t4}\|^2 &= O_P(\max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2).
\end{aligned}$$

Finally, let $\boldsymbol{\beta}_i$ denote the ith row of $\frac{1}{N}\widehat{\mathbf{\Lambda}}'\mathbf{M}_\alpha\mathbf{A}$, $i \leq K$. Then

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{t3}\|^2 = \frac{1}{T}\sum_{t=1}^T \|\frac{1}{N}\widehat{\mathbf{\Lambda}}'\mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t)\|^2 = \sum_{i=1}^K \frac{1}{T}\sum_{t=1}^T (\boldsymbol{\beta}_i'\Phi(\mathbf{x}_t))^2 \\
\leq\ & \sum_{i=1}^K \|\boldsymbol{\beta}_i\|^2 \|\frac{1}{T}\sum_{t=1}^T \Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)'\| \\
=\ & O_P(1)\|\frac{1}{N}\widehat{\mathbf{\Lambda}}'\mathbf{M}_\alpha\mathbf{A}\|_F^2 = O_P(\frac{1}{N^2}\|\widehat{\mathbf{\Lambda}}'\mathbf{M}_\alpha\|^2) \\
\leq\ & O_P(b_{NT}^2 + a_T^4\chi_N^{-1}).
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2 \leq O_P(1)\sum_{i=1}^4 \frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{ti}\|^2 \\
\leq\ & O_P(a_T^4\chi_N^{-1} + b_{NT}^2 + \max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2) \\
\leq\ & O_P(\frac{J^2}{T^2}\chi_N^{-1} + J^{2-4\eta}\chi_N^{-1} + J^{1-2\eta} + \frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + \frac{J}{TN} + \frac{J^3\log N\log J}{T^2} \\
& + \frac{J}{T}\alpha_T^{-\zeta_2} + \alpha_T^{-2(\zeta_1-1)}\frac{J^3\log N}{T}) \\
\leq^{(1)}\ & O_P(\frac{J^2}{T^2}\chi_N^{-1} + \frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + J^{1-2\eta} + \frac{J}{TN} + \frac{J^3\log^2 N}{T^2})
\end{aligned}$$

where (1) is due to $\zeta_1, \zeta_2 > 2$, and $J^3 \log^2 N = O(T)$,

$$\frac{J}{T}\alpha_T^{-\zeta_2} + \alpha_T^{-2(\zeta_1-1)}\frac{J^3 \log N}{T} + \frac{J^3 \log N \log J}{T^2} = O(\frac{J^3 \log^2 N}{T^2})$$

and $\chi_N \gg J^{-\eta}$ (so $J^{2-4\eta}\chi_N^{-1} = O(J^{1-2\eta})$). Q.E.D.

Proposition B.3.

$$\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P(\frac{1}{N}) + O_P(\chi_N^{-1})(\frac{J^4(\log N)^2}{T^2} + \frac{J^2}{T^2}\chi_N^{-1} + \frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + J^{1-2\eta} + \frac{J}{TN}).$$

Note that $\mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t) = \boldsymbol{\Lambda}\boldsymbol{\gamma}_t + \mathbf{u}_t$. and $\widehat{\boldsymbol{\gamma}}_t = \frac{1}{N}\widehat{\boldsymbol{\Lambda}}'(\mathbf{y}_t - \widehat{E}(\mathbf{y}_t|\mathbf{x}_t))$. Hence from (B.4)

$$\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t = \frac{1}{N}\mathbf{H}'\boldsymbol{\Lambda}'\mathbf{u}_t + \mathbf{D}_{t1} + \mathbf{D}_{t2} + \mathbf{C}_{t3} + \mathbf{C}_{t4} \tag{B.6}$$

where $\mathbf{C}_{t3}, \mathbf{C}_{t4}$ are as defined earlier, and

$$\begin{aligned}
\mathbf{D}_{t1} &= \frac{1}{N}\widehat{\boldsymbol{\Lambda}}'(\boldsymbol{\Lambda}\mathbf{H} - \widehat{\boldsymbol{\Lambda}})\mathbf{H}^{-1}\boldsymbol{\gamma}_t, \qquad \mathbf{D}_{t2} = \frac{1}{N}(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\mathbf{H})'\mathbf{u}_t \\
\mathbf{C}_{t3} &= \frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\mathbf{M}_\alpha\mathbf{A}\Phi(\mathbf{x}_t), \qquad \mathbf{C}_{t4} = \frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\mathbf{R}_t.
\end{aligned}$$

Hence for a constant $C > 0$, $\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 \le C(\sum_{i=1}^{2}\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{D}_{ti}\|^2 + \sum_{i=3}^{4}\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{C}_{ti}\|^2)$. We look at terms on the right hand side one by one. First of all,

$$\begin{aligned}
E\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\gamma}_t\boldsymbol{\gamma}_t' - \operatorname{cov}(\boldsymbol{\gamma}_t)\|_F^2 &= \sum_{i=1}^{K}\sum_{j=1}^{K}\operatorname{var}(\frac{1}{T}\sum_{t=1}^{T}\gamma_{it}\gamma_{jt}) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{1}{T}\operatorname{var}(\gamma_{it}\gamma_{jt}) \\
&= O(T^{-1})\max_{i,j\le K}\operatorname{var}(\gamma_{it}\gamma_{jt}).
\end{aligned}$$

This implies $\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\gamma}_t\boldsymbol{\gamma}_t'\| \le O_P(c_T)$ where

$$c_T := \|\operatorname{cov}(\boldsymbol{\gamma}_t)\| + (\frac{1}{T}\max_{i,j\le K}\operatorname{var}(\gamma_{it}\gamma_{jt}))^{1/2}.$$

As for $\mathbf{D}_{t1}$, let $\mathbf{G} = \frac{1}{N}\widehat{\boldsymbol{\Lambda}}'(\boldsymbol{\Lambda}\mathbf{H} - \widehat{\boldsymbol{\Lambda}})\mathbf{H}^{-1}$ and let $\mathbf{G}_i'$ denote its ith row, $i \le K$. By (C.5),

and $\|\mathbf{H}^{-1}\| = O_P(1)$,

$$\|\mathbf{G}\|^2 \leq O_P(\chi_N^{-1})(\frac{1}{N}\|\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\|_F + (\max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2)^{1/2})^2 + O_P(a_T^4\chi_N^{-2}).$$

Then

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \|\mathbf{D}_{t1}\|^2 &= \sum_{i=1}^K \frac{1}{T}\sum_{t=1}^T (\mathbf{G}_i'\boldsymbol{\gamma}_t)^2 = \sum_{i=1}^K \mathbf{G}_i' \frac{1}{T}\sum_{t=1}^T \boldsymbol{\gamma}_t\boldsymbol{\gamma}_t'\mathbf{G}_i \\ &\leq \|\frac{1}{T}\sum_{t=1}^T \boldsymbol{\gamma}_t\boldsymbol{\gamma}_t'\|\|\mathbf{G}\|_F^2 \\ &= \|\mathbf{G}\|_F^2 O_P(c_T) \\ &\leq c_T O_P(\chi_N^{-1})(\frac{1}{N}\|\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\|_F + (\max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2)^{1/2})^2 + O_P(c_T a_T^4\chi_N^{-2}).\end{aligned}$$

Terms $\mathbf{C}_{t3}$ and $\mathbf{C}_{t4}$ were bounded in the proof of Proposition B.2:

$$\frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{t3}\|^2 + \frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{t4}\|^2 \leq O_P(\max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2 + \frac{1}{N^2}\|\widehat{\mathbf{\Lambda}}'\mathbf{M}_\alpha\|^2).$$

Term $\mathbf{D}_{t2}$ is given in Lemma C.12:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T \|\mathbf{D}_{t2}\|^2 \\ =\ & O_P(\chi_N^{-1})(\frac{1}{N^3}\|\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\|^2 + \frac{1}{N}\max_i \frac{1}{T}\sum_{t=1}^T R_{it}^2 + \frac{1}{N^2T}\sum_{s=1}^T \|\mathbf{u}_s'\mathbf{M}_\alpha\|^2 + \frac{1}{N^2T^2}\sum_{s=1}^T\sum_{t=1}^T |\mathbf{u}_s'\mathbf{R}_t|^2).\end{aligned}$$

By Lemmas C.10, C.11,

$$\begin{aligned}\frac{1}{N^2T}\sum_{s=1}^T \|\mathbf{u}_s'\mathbf{M}_\alpha\|^2 &\leq O_P(\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{TN} + \frac{J}{T^2} + \frac{J}{N^2T} + \frac{J}{T}\alpha_T^{-\zeta_2}) \\ \frac{1}{N^2T^2}\sum_{s=1}^T\sum_{t=1}^T |\mathbf{u}_s'\mathbf{R}_t|^2 &\leq O_P(\frac{J^4\log N\log J}{T^2} + \frac{J^{2-2\eta}}{N} + \frac{J^4\log N}{T}\alpha_T^{-2(\zeta_1-1)}).\end{aligned}$$

So combined with Lemmas C.6, Proposition C.3,

$$\sum_{i=3}^4 \frac{1}{T}\sum_{t=1}^T \|\mathbf{C}_{ti}\|^2 + \sum_{i=1}^2 \frac{1}{T}\sum_{t=1}^T \|\mathbf{D}_{ti}\|^2 = O_P(c_T a_T^4\chi_N^{-2})$$

$$
\begin{aligned}
&+O_P(1+c_T\chi_N^{-1}+N^{-1}\chi_N^{-1})(\frac{1}{N^2}\|\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\|^2+\max_i\frac{1}{T}\sum_{t=1}^T R_{it}^2)\\
&+O_P(\chi_N^{-1})(\frac{1}{N^2T}\sum_{s=1}^T\|\mathbf{u}_s'\mathbf{M}_\alpha\|^2+\frac{1}{N^2T^2}\sum_{s=1}^T\sum_{t=1}^T|\mathbf{u}_s'\mathbf{R}_t|^2)\\
\leq^{(1)}\ & O_P(\chi_N^{-1})(\frac{1}{N^2}\|\mathbf{M}_\alpha'\widehat{\mathbf{\Lambda}}\|^2+\max_i\frac{1}{T}\sum_{t=1}^T R_{it}^2+\frac{1}{N^2T}\sum_{s=1}^T\|\mathbf{u}_s'\mathbf{M}_\alpha\|^2+\frac{1}{N^2T^2}\sum_{s=1}^T\sum_{t=1}^T|\mathbf{u}_s'\mathbf{R}_t|^2)\\
&+O_P(c_Ta_T^4\chi_N^{-2})\\
\leq^{(2)}\ & O_P(\chi_N^{-1})(\frac{J^4(\log N)^2}{T^2}+\frac{J^2}{T^2}\chi_N^{-1}+\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T}+J^{1-2\eta}+\frac{J}{TN}+a_T^4\chi_N^{-1})\\
\leq^{(3)}\ & O_P(\chi_N^{-1})(\frac{J^4(\log N)^2}{T^2}+\frac{J^2}{T^2}\chi_N^{-1}+\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T}+J^{1-2\eta}+\frac{J}{TN}).
\end{aligned}
$$

where (1) follows from that $1+c_T\chi_N^{-1}+\chi_N^{-1}N^{-1}=O(\chi_N^{-1})$; (2) is due to $\frac{J}{T}\alpha_T^{-\zeta_2}+\alpha_T^{-2(\zeta_1-1)}\frac{J^4\log N}{T}+\frac{J^4\log N\log J}{T^2}=O(\frac{J^4\log^2 N}{T^2})$ and that $c_T=O(1)$ due to Assumption 4.1; (3) is due to $J^{-\eta}\chi_N^{-1}=O(1)$.

Finally, $\frac{1}{T}\sum_{t=1}^T\|\frac{1}{N}\mathbf{H}'\mathbf{\Lambda}'\mathbf{u}_t\|^2=O_P(\frac{1}{TN^2}\sum_{t=1}^T E\|\mathbf{\Lambda}'\mathbf{u}_t\|^2)=O_P(\frac{1}{N})$. Hence

$$
\frac{1}{T}\sum_{t=1}^T\|\widehat{\boldsymbol{\gamma}}_t-\mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2=O_P(\frac{1}{N})+O_P(\chi_N^{-1})(\frac{J^4(\log N)^2}{T^2}+\frac{J^2}{T^2}\chi_N^{-1}+\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T}+J^{1-2\eta}+\frac{J}{TN}).
$$

References

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

Ahn, S., Lee, Y. and Schmidt, P. (2001). Gmm estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics* **101** 219–255.

Andrews, D. (1991). Asymptotic optimality of generalized c_l, cross-validation, and generalized crossvalidation in regression with heteroskedastic errors. *Journal of Econometrics* **47** 359-377.

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

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279.

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

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

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data, methods, theory and applications*. The first edition ed. Springer, New York.

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

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analyssi in large asset markets. *Econometrica* **51** 1305–1324.

Dimatteo, I., Genovese, C. and Kass, R. (2001). Bayesian curve fitting with free-knot splines. *Biometrika* **88** 1055–1071.

Doz, C., Giannone, D. and Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics* **94** 1014–1024.

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

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics* **116** 1–22.

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 247–265.

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

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

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

FriedmanJ.H. and StuetzleW. (1981). Projection pursuit regression. *Journal of the American statistical Association* **76**, 817-823.

Gibbons, M., Ross, S. and Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica* **57** 1121–1152.

Gungor, S. and Luger, R. (2013). Testing linear factor pricing models with large cross sections: A distribution-free approach. *Journal of Business & Economic Statistics* **31** 66–77.

Hart, J. D. H. (1994). Automated kernel smoothing of dependent data by using time series cross- validation. *Journal of the Royal Statistical Society, Series B* **56** 529–542.

Huang, H. and Lee, T.-H. (2010). To combine forecasts or to combine information? *Econometric Reviews* **29** 534–570.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101.

Hurvich, C., Simonoff, J. and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60** 271-293.

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

Lawley, D. and Maxwell, A. (1971). *Factor analysis as a statistical method*. The second edition ed. Butterworths, London.

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

Li, K. (1987). Asymptotic optimality for c_p, c_l cross-validation, and generalized cross-validation: Discrete index set. *Annals of Statistics* **15** 958-975.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327.

Ludvigson, S. and Ng, S. (2009). Macro factors in bond risk premia. *Review of Financial Studies* **22** 5027–5067.

Ludvigson, S. and Ng, S. (2010). A factor analysis of bond risk premia. *Handbook of Empirical Economics and Finance* 313–372.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics* 382–400.

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

Network, C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics* **108** 1–28.

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

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

Pesaran, H. and Yamagata, T. (2012). Testing capm with a large number of assets. Tech. rep., University of South California.

Portnoy, S. (1985). Asymptotic behavior of m estimators of p regression parameters when p2/n is large; ii. normal approximation. *The Annals of Statistics* 1403–1417.

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

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. The first edition ed. Springer.