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

January 6, 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 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 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.

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

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.

In empirical applications, researchers frequently encounter additional observable covariates that help explain the latent factors. In financial time series forecasts, researchers often collect additional variables that characterize financial markets, which are known to have explanatory powers of the factors. The Fama-French factors are well-known to be related to the factors that drive financial returns (Fama and French, 1992). 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. Also, some genetic studies collect both gene expression and single-nucleotide polymorphism data on the same group of subjects. These covariates are directly related to the latent factors. Another typical example of these covariates arises in situations when we observe some of the common factors.

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 $\|\mathbf{A}\|_F = \text{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ as the Frobenius norm, spectral norm, ℓ_1-norm, and element-wise norm of a matrix $\mathbf{A}$. Note that when $\mathbf{A}$ is a vector, both $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|$ are equal to the Euclidean norm. Finally, 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 $b_T = O(a_T)$.

1.1 Identifications of Usual Factor Models without Covariates

Q: this section is too long. Probably reduce to a few sentence. We can also shorten section 1.2 somewhat.

To appreciate the new features of model (1.2), we first briefly review the identification and estimation of regular factor models without covariates. Suppose $\{\mathbf{f}_t, \mathbf{u}_t\}_{t \le T}$ is a stationary process and that $\mathbf{f}_t$ and $\mathbf{u}_t$ are uncorrelated. Then 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 $\mathbf{f}_t$. Apparently, $\mathbf{\Lambda}\, \text{cov}(\mathbf{f}_t)\mathbf{\Lambda}'$ is a *low-rank* matrix so long as $K < N$. Since $\{\mathbf{f}_t, \mathbf{u}_t\}_{t \le T}$ is unobservable, what makes the decomposition (1.3) identifiable is the following four typical conditions (they are not necessary conditions):

(i) $N \to \infty$.

(ii) $\{\nu_{y,1} > ... > \nu_{y,K}\}$ should be distinct, representing the top K eigenvalues of $\text{cov}(\mathbf{y}_t)$.

(iii) All the eigenvalues of the $N \times N$ matrix $\text{cov}(\mathbf{u}_t)$ are bounded from above by a constant $c_u > 0$.

(iv) All the eigenvalues of the $K \times K$ matrices $\mathbf{\Lambda}'\mathbf{\Lambda}/N$ and $\text{cov}(\mathbf{f}_t)$ are bounded from below by a constant $c_\Lambda > 0$.

To see how this is helpful to identify $\mathbf{\Lambda}$, for simplicity of our explanations, we temporarily assume $\mathbf{\Lambda}'\mathbf{\Lambda}/N = \mathbf{I}_K$, and that $\text{cov}(\mathbf{f}_t)$ is a diagonal matrix. Then, it is easy to see that the diagonal entries of $\text{cov}(\mathbf{f}_t)$ are the first K eigenvalues of $\frac{1}{N}(\text{cov}(\mathbf{y}_t) - \text{cov}(\mathbf{u}_t))$, with the columns of $\frac{1}{\sqrt{N}}\mathbf{\Lambda}$ the corresponding eigenvectors. Regarding $\text{cov}(\mathbf{u}_t)$ as the perturbation,

let $\mathbf{\Xi}_y$ be an $N \times K$ matrix whose columns are the eigenvectors corresponding to the first K eigenvalues of $\text{cov}(\mathbf{y}_t)$. We apply the Davis-Kahan $\sin\theta$-theorem to reach:

$$\|\frac{1}{\sqrt{N}}\mathbf{\Lambda} - \mathbf{\Xi}_y\| = O(N^{-1}).$$

Thus, $\frac{1}{\sqrt{N}}\mathbf{\Lambda}$ is identified as the first K eigenvectors of $\text{cov}(\mathbf{y}_t)$ *asymptotically* 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)).

In the general setting that allows $\mathbf{\Lambda}'\mathbf{\Lambda}/N \neq \mathbf{I}_K$, and non-diagonal $\text{cov}(\mathbf{f}_t)$, the above arguments still hold, but $\mathbf{\Lambda}$ is identified up to a rotation transformation. We summarize the result as follows.

Proposition 1.1. *Suppose (i)-(iv) hold. Consider a general setting in which* $\mathbf{\Lambda}'\mathbf{\Lambda}/N$ *is not necessarily* $\mathbf{I}_K$ *and* $\text{cov}(\mathbf{f}_t)$ *is not necessarily diagonal. Then there is a* $K \times K$ *matrix* $\mathbf{H}_y$ *so that*

$$\|\frac{1}{\sqrt{N}}\mathbf{\Lambda}\mathbf{H}_y - \mathbf{\Xi}_y\| = O(N^{-1}).$$

Remark 1.1. Both (iii) (iv) can be weakened to allow the eigenvalues of $\mathbf{\Lambda}'\mathbf{\Lambda}/N$ to decay slowly, and that the eigenvalues of $\text{cov}(\mathbf{u}_t)$ to grow slowly. But this would slow down the identification rate in Proposition 1.1.

It is important to note that this identification is only asymptotic (requires $N \to \infty$), in order to remove the effect of $\text{cov}(\mathbf{u}_t)$. In contrast, we show that with the additional common covariates and a *rank condition*, conditions (i)- (iii) can be removed, and condition (iv) is modified. The "exact identification" can be achieved, for any finite $N > K$. Neither do we need $N \to \infty$ to estimate $(\mathbf{\Lambda}, \mathbf{g}(\mathbf{x}_t))$. As we shall show in this paper, $N \to \infty$ is only required to improve the estimation accuracy for $\boldsymbol{\gamma}_t$.

1.2 Exact Identification with Covariates

This paper considers models with additional covariates, where the identification is done 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 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. 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 identifiable up to an orthogonal transformation.

The above discussion and (1.4) prompt 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 assume that $E(\mathbf{u}_t|\mathbf{x}_t) = 0$ so the regression step removes the error term $\mathbf{u}_t$ and $\{\widehat{\mathbf{y}}_t\}$ is approximately noiseless. 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] This demonstrates another advantage of our estimation procedure: step (i) projects the original data to the space of $\mathbf{x}_t$, whose distribution is no longer heavy-tailed, and is suitable for the PCA step (ii). In addition, the number of principal components taken in step (ii) equals K, the number of factors, and 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).

It turns out that there are two important objects that determine the rates of convergence for the estimators, which are essentially the signal and noise of the covariate model (1.2):

[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.

The signal: The $K \times K$ matrix $\mathbf{\Sigma}_{f|x} = E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$. It characterizes the strength of the explanatory power of the covariates on the unknown factors. We shall assume that its rank equals K, but allow its minimum eigenvalue to decay slowly to zero, hence allows the case of "weak signals".

The noise: The $K \times K$ covariance matrix of $\boldsymbol{\gamma}_t$: $\text{cov}(\boldsymbol{\gamma}_t) = E\boldsymbol{\gamma}_t\boldsymbol{\gamma}_t'$. We only require the eigenvalues of $\text{cov}(\boldsymbol{\gamma}_t)$ be bounded, and do not make further assumptions on the magnitudes of these eigenvalue sequences. In the special case that $\text{cov}(\boldsymbol{\gamma}_t) = 0$, $\mathbf{x}_t$ fully explains $\mathbf{f}_t$ in the sense that $\mathbf{f}_t = \mathbf{g}(\mathbf{x}_t)$ almost surely.

We allow N to be either finite or growing. Both loadings and $\mathbf{g}(\mathbf{x}_t)$ can be estimated consistently even if N is finite. In addition, when N grows, we allow $N/T \in [0, \infty]$. Our rates of convergence of the estimators are presented using $\text{cov}(\boldsymbol{\gamma}_t)$ and $\mathbf{\Sigma}_{f|x}$. We can achieve improved rates of convergence on both the estimated factors and the loadings, so long as the covariates can (at least partially) explain the latent factors. The degree of improvements depends on the strength of the signals.

1.3 Testing Proxy Factors for Financial Returns

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 explain 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 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. (2013)). While the efficiency test aims to test whether the alphas of excess returns are simultaneously zero for the specified factors, 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. 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.

1.4 Further Literature

Most existing works simply treat $\mathbf{x}_t$ as a set of additional regressors in (1.1). 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 $\mathbf{x}_t$. In addition, the time-varying-coefficient models are also related to our work, in which the coefficients can be explained by some influential variables. (e.g., Wang et al. (2009); Song et al. (2014); Lee et al. (2012)). Moreover, our expansion 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. Finally Section 7 concludes. We present the extensive simulation studies in Appendix **??** in the supplement. The supplement also contains all the technical proofs.

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 is: (i) associated with the latent factors $\mathbf{f}_t$, and (ii) mean-independent of the idiosyncratic term. Taking the conditional mean 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 $\{(\mathbf{y}_t, \mathbf{x}_t)\}_{t\leq T}$, but $\mathbf{\Sigma}_{f|x}$ is not because $\mathbf{f}_t$ is not observable. 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 $\mathbf{\Sigma}_{f|x}$. 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 to facilitate the above arguments. In this paper, they are not imposed. Then the same argument shows 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*, and is possible in the current context, 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 $\mathbf{H} = \mathbf{I}_K$.

Remark 2.1. Another key difference from the classical asymptotic identification is that Theorem 2.1 does not require any conditions on the $N \times N$ covariance matrix of $\mathbf{u}_t$. This is because Theorem 2.1 and the classical result (Proposition 1.1) rely on very different strategies to remove the effect of noises $\mathbf{u}_t$. Thanks to the condition $E(\mathbf{u}_t|\mathbf{x}_t) = 0$, we remove the effect of $\mathbf{u}_t$ by directly projecting onto the space of $\mathbf{x}_t$. In sharp contrast, under the classical setting without covariates, the effect of $\mathbf{u}_t$ was removed by requiring $\|\operatorname{cov}(\mathbf{u}_t)\|/N \to 0$ as $N \to \infty$.

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 highly nonlinear, it is still possible to satisfy the full rank condition even if $d < K$, and we illustrate this in the simulation section. [2]

[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 $J \to \infty$.

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 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

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

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

$$\max_{i\le N} \|\mathbf{b}_{i,\alpha} - \mathbf{b}_i\| = O(\alpha_T^{-(\zeta_2+1)+\epsilon})$$

for an arbitrarily small $\epsilon > 0$, where ζ_2 is defined in Assumption 4.1. Hence the bias decreases as α_T grows. On the other hand, our theory requires the uniform convergence (in $i = 1, ..., N$) of (for $e_{it} = y_{it} - E(y_{it}|\mathbf{x}_t)$)

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

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]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, and thank an anonymous for this suggestion.

Q: We do not need this detail of CV. Suggest to remove.

As for the multi-fold cross validations, we randomly divide the sample into M folds with subsample sizes $\{T_1, \cdots, T_M\}$. Then we can define the *in-sample cross-validation* criterion for a given pair of C_α and J as

$$CV_{\text{in}}(C_\alpha, J) = \frac{1}{T}\sum_{m=1}^{M}\sum_{t=1}^{T_m}\sum_{i=1}^{N}\left(y_{it} - \Phi(\mathbf{x}_t)'\widetilde{\mathbf{b}}_i^{(-m)}(C_\alpha, J)\right)^2,$$

where $\widetilde{\mathbf{b}}_i^{(-m)}(C_\alpha, J)$ is the fitted parameter without using data in T_m:

$$\widetilde{\mathbf{b}}_i^{(-m)}(C_\alpha, J) = \arg\min_{\mathbf{b}\in\mathbb{R}^J}\frac{1}{T - T_m}\sum_{m'=[M]\setminus m}\sum_{t=1}^{T_{m'}}\rho\left(\frac{y_{it} - \Phi(\mathbf{x}_t)'\mathbf{b}}{C_\alpha\sqrt{\frac{T}{\log(NJ)}}}\right).$$

In practice, let $\mathcal{A}$ and $\mathcal{J}$ be two sets of gird points for C_α and J respectively. We use the following CV-based $(\widehat{\alpha_T}, \widehat{J})$ for (α_T, J):

$$(\widehat{C_\alpha}, \widehat{J}) = \arg\min_{(C_\alpha, J)\in\mathcal{A}\times\mathcal{J}} CV_{\text{in}}(C_\alpha, J), \quad \widehat{\alpha_T} = \widehat{C_\alpha}\sqrt{\frac{T}{\log(N\widehat{J})}}.$$

Alternatively, in out-of-sample forecast applications using estimated factors (see our empirical application). We aim to conduct an h-step ahead forecast of a scalar variable z_{T+h} using the data $\{z_t, \mathbf{x}_t, \mathbf{y}_t\}_{t\le T}$. Let $\widehat{z}_{t+h|t}(C_\alpha, J)$ be the predicted value of z_{t+h} using the data up to time t, which depends on (C_α, J) through the estimated factors. Then we can choose the tuning parameters by minimizing the *out-of-sample cross-validation* criterion (e.g. Hart, 1994):

$$CV_{\text{out}}(C_\alpha, J) = \sum_{t=1}^{T-h}\left(z_{t+h} - \widehat{z}_{t+h|t}(C_\alpha, J)\right)^2.$$

3.3 Some related alternative estimators

3.3.1 An alternative estimator of $\mathbf{\Sigma}_{y|x}$

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 of $E(\mathbf{Y}|\mathbf{x}_1, ..., \mathbf{x}_T)$ is simply $\mathbf{YP}$ and the covariance 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 suitable when the distribution of $\mathbf{u}_t$ has heavier tails. As expected, our numerical studies in Section **??** demonstrate that it performs well in light-tailed scenarios, but is less robust to heavy-tailed distributions.

3.3.2 Least squares on interactive effects

Plugging $\mathbf{f}_t = \mathbf{g}(\mathbf{w}_t) + \boldsymbol{\gamma}_t$ 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 closely 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 $\mathbf{h}(\mathbf{w}_t)$. 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.

3.3.3 Combination of forecasts

Huang and Lee (2010) proposed a related method called "combination of forecasts". They considered the following forecasting problem:

$$z_{t+1} = \boldsymbol{\beta}'\mathbf{f}_t + \epsilon_{t+1}, \quad \text{where}$$

[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.

$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t$, and $\mathbf{f}_t$ is also related to some observable $\mathbf{x}_t$. They estimated the common factors, denoted by $\widetilde{\mathbf{f}}_t$, by extracting the principal components from the estimated $\mathbf{\Sigma}_{y|x}$, and forecast z_{t+1} using the model $z_{t+1} = \boldsymbol{\beta}'\widetilde{\mathbf{f}}_t + \epsilon_{t+1}$. While this method is closely related to our proposal, we note three major differences. First, fundamentally, as we have explained, the extracted principal components from $\mathbf{\Sigma}_{y|x}$, are estimating $E(\mathbf{f}_t|\mathbf{x}_t)$, rather than $\mathbf{f}_t$. So $\widetilde{\mathbf{f}}_t$ is inconsistent for $\mathbf{f}_t$ unless $\boldsymbol{\gamma}_t = o_P(1)$. And using $\widetilde{\mathbf{f}}_t$ to forecast z_{t+1} would lose the predicting power from $\boldsymbol{\gamma}_t$, as long as the latter has a strong signal. In contrast, as we have emphasized, our method is valid regardless of the signal strength of $\boldsymbol{\gamma}_t$. Secondly, they used $\widetilde{\mathbf{\Sigma}}$ (as defined in Section 3.3.1) to estimate $\mathbf{\Sigma}_{y|x}$, which is not robust to heavy tails. Finally, Huang and Lee (2010) did not provide theoretical analysis for their proposed estimators. Instead, we formally study the rates of convergence for the estimated factors and loadings.

Another simpler alternative is to combine $(\mathbf{x}_t, \mathbf{y}_t)$, and apply the classical methods on this enlarged dataset. The potential drawback of this simpler approach is that the rates of convergence would not be improved, even if $\mathbf{x}_t$ has strong explanatory power on the factors. In sharp contrast, as we shall show in the next section, substantial improvements on the rates of convergence can be gained using the proposed method. Another drawback, as mentioned before, is that $\mathbf{x}_t$ and $\mathbf{y}_t$ can provide very different information (e.g. Fama-French factors versus returns of individual stocks).

4 Rates of Convergence

One of the popular technical tools in the literature on "spiked covariances" is the *perturbation theory* (e.g. Paul (2007); Birnbaum et al. (2013); Wang and Fan (2009)), which provides inequalities that directly upper bound the distance between the eigenvectors and their estimators. But this technique would not provide sharp rates for estimating the low-dimensional functionals of the eigenvectors. For instance, suppose we construct an estimator as a linear transformation of $\widehat{\mathbf{\Lambda}}$, a typical example is $\widehat{\mathbf{f}}_t = \frac{1}{N}\widehat{\mathbf{\Lambda}}'\mathbf{y}_t$, then it can be proved that for some rotation matrix $\mathbf{H}$, and $\mathbf{\Delta}_\Lambda = \widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H}$,

$$\widehat{\mathbf{f}}_t - \mathbf{H}^{-1}\mathbf{f}_t = \underbrace{\frac{1}{N}\mathbf{H}'\mathbf{\Lambda}'\mathbf{u}_t}_{\text{oracle term}} + \underbrace{\frac{1}{N}\mathbf{\Delta}_\Lambda'\mathbf{u}_t - \frac{1}{N}\widehat{\mathbf{\Lambda}}'\mathbf{\Delta}_\Lambda\mathbf{H}^{-1}\mathbf{f}_t}_{\text{effect of estimating } \mathbf{\Lambda}}$$

where the oracle term is the statistical error if $\mathbf{\Lambda}$ were known. The perturbation theory however, would only bound $\frac{1}{N}\mathbf{\Delta}_\Lambda'\mathbf{u}_t$ by $\|\frac{1}{N}\mathbf{\Delta}_\Lambda\|\|\mathbf{u}_t\|$ which is not sufficiently sharp, and may even obtain a rate dominating the oracle term. In fact, almost the entire literature only focuses on the eigenvectors $\mathbf{\Lambda}$ themselves, rather than their low-dimensional functionals.

Instead, we use a different approach by deriving a Bahadur expansion (Bahadur, 1966) of the estimated eigenvectors for the spiked matrices, in the following form:

$$\begin{aligned}\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H} &= \frac{1}{NT}\sum_{t=1}^{T}\mathbf{\Lambda}\mathbf{g}(\mathbf{x}_t)\Phi(\mathbf{x}_t)'\mathbf{A}\sum_{i=1}^{N}\frac{1}{T}\sum_{s=1}^{T}\Phi(\mathbf{x}_s)'\dot{\rho}(\alpha_T^{-1}e_{is})\alpha_T\widehat{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{-1} \\ &\quad +\widetilde{\mathbf{\Delta}}(\{\mathbf{x}_t,\mathbf{e}_t\}_{t\le T}),\end{aligned} \tag{4.1}$$

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; $\mathbf{e}_t := \mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t) = (e_{1t},...,e_{Nt})'$; $\mathbf{A} = (2E\Phi(\mathbf{x}_t)\Phi(\mathbf{x}_t)')^{-1}$ 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 $\widetilde{\mathbf{\Delta}}(\{\mathbf{x}_t,\mathbf{e}_t\}_{t\le T})$ is a higher order random term that depends on both $\{\mathbf{x}_t\}$ and $\{\mathbf{e}_t\}$. Such an expansion allows us to derive a much sharper bound for low-dimensional functionals of $\mathbf{\Delta}_\Lambda$ such as $\frac{1}{N}\mathbf{\Delta}_\Lambda'\mathbf{u}_t$. It is also potentially useful for deriving the limiting distributions of the estimators, which cannot be achieved by the perturbation theory. [5] The techniques we developed for deriving the Bahadur expansion for estimated eigenvectors might be of independent interest.

Throughout the paper, we assume T grows to infinity, while $K = \dim(\mathbf{f}_t)$ and $d = \dim(\mathbf{x}_t)$ are constant. In addition, N may either grow or stay constant.

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 $g_{e,i}(\cdot|\mathbf{x})$.

[5] While deriving the limiting distributions for the estimated factors and loadings is out of the scope of this paper, we expect that they can be derived based on the Bahadur expansion, and shall leave it for the future research.

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

$$\sup_{\mathbf{x}} \max_{i \le N} g_{e,i}(x|\mathbf{x}) \le C x^{-\zeta_1}, \quad \sup_{\mathbf{x}} \max_{i \le N} E(e_{it}^2 1\{|e_{it}| > x\}|\mathbf{x}_t = \mathbf{x}) \le C x^{-\zeta_2}. \tag{4.2}$$

(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 (Weak dependences). *(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

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

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

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

(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

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

4.2 Discussions of the assumptions

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. We refer to many advances in the robust regression literature, see for instance, Rousseeuw and Leroy (2005) and Maronna et al. (2006). Assumption 4.2 is regarding the accuracy of sieve approximations for non-parametric functions. It also implies a similar condition on the sieve approximation to $E(y_{it}|\mathbf{x}_t = \cdot)$. Suppose $E(f_{kt}|\mathbf{x}_t = \cdot)$ belongs to a Hölder class: for some $r, \alpha > 0$,

$$\mathcal{G} = \{h : |h^{(r)}(\mathbf{x}) - h^{(r)}(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|^{\alpha}\},$$

then this condition is satisfied by common basis such as the polynomials and B-splines with $\eta = 2(r + \alpha)/\dim(\mathbf{x}_t)$.

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. To obtan 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 3.3.1) 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.

Finally, Assumption 4.4 strengthens Assumption 2.1. We respectively regard $\lambda_{\min}(\mathbf{\Sigma}_{f|x})$ and $\text{cov}(\boldsymbol{\gamma}_t)$ as the "signal" and "noise" when using $\mathbf{x}_t$ to explain common factors. The explanatory power is measured by these two quantities. Condition (i) is not necessary, but since both $\lambda_{\max}(\mathbf{\Sigma}_{f|x})$ and $\lambda_{\max}(E\{\Phi(\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'E(\mathbf{f}_t|\mathbf{x}_t)\Phi(\mathbf{x}_t)'\})$ appear in the rates of convergence, it facilitates our presentation when the factor is multivariate. Also, requiring them be of the same order as χ_N is not stringent since the dimension K is small. Furthermore, the condition $J^2/T + J^{-\eta} + \sqrt{\log N/T} \ll \chi_N$ ensures that the signal is stronger than the statistical error for estimating $\mathbf{\Sigma}_{y|x}$.

4.3 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,*

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

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

Remark 4.1. 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. Because $\mathbf{\Lambda}$ is exactly identified (up to a rotation) as the leading eigenvectors of $E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$, there are two main sources of the estimation errors reflected in the achieved rate: (i) nonparametric estimating $E(\mathbf{y}_t|\mathbf{x}_t)$, and (ii) given $E(\mathbf{y}_t|\mathbf{x}_t)$, estimating the leading eigenvectors. As for (i), the presented rate connects well with the standard nonparametric literature (e.g., Shen and Wong (1994); Birgé and Massart (1998)) where $\frac{J}{T}$ and $\frac{1}{J^{2\eta-1}}$ respectively represent the "variance" and "bias" of the sieve estimation, and $J \asymp T^{1/(2\eta)}$ balances these two. As for (ii), the accuracy of estimating the eigenvectors depends on the signal strength of the corresponding eigenvalues. This is measured by χ_N.

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 4.2. 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.

Remark 4.3. The presented rates also 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 results show that the rate of convergence in this "oracle sieve regression" (knowing $\boldsymbol{\Lambda}$) would be

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

Here $\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T}+\frac{J}{TN}$ and $J^{1-2\eta}$ respectively represent the usual "variance" and "bias" terms of the sieve nonparametric regression, and the "variance" arises from two components in $\mathbf{e}_t$. As we do not observe $\boldsymbol{\Lambda}$, we are running the regression (4.9) with $\widehat{\boldsymbol{\Lambda}}$ in place of $\boldsymbol{\Lambda}$. This leads to an additional term $\frac{J^2}{T^2}\chi_N^{-1}$ representing the effect of estimating $\boldsymbol{\Lambda}$, 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.

As for the convergence rate of $\boldsymbol{\gamma}_t$ in (4.6), the additional term $O_P(N^{-1})$ is similar to the leading term in the convergence rate for estimtated factors (e.g., Bai (2003)). In the current context, this term appears because $\boldsymbol{\gamma}_t$ captures the unknown components in the factors that cannot be explained by the covariates. Even though it has the same rate as estimating $\mathbf{f}_t$, it has a smaller variability than $\mathbf{f}_t$ does and hence is easier to be estimated. This also reflects the gain of using covariates.

4.4 The signal-noise regimes

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

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

for some $c, C_1 > 0$, assuming that there is $c > 0$ so that $\|E\mathbf{f}_t\mathbf{f}_t'\| > c$. 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", we can estimate the loadings well without consistently estimating the latent factors. In addition, if $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\| \to 0$, then $\widehat{\mathbf{g}}(\mathbf{x}_t)$ can also consistently estimate the latent factors even if N is finite. In contrast, the benchmark PCA estimators rely on a diverging N to compensate for the effect of estimating the unknown factors. 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:

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

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, when $N \ll T$, and $\chi_N \gg \frac{N}{T}$,

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

$$\underbrace{\frac{1}{T}\chi_N^{-1}+\frac{1}{N}}_{\text{new rate for factors}} \asymp \underbrace{\frac{1}{N}}_{\text{PCA rate for factors}}.^{6}$$

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

[6] In the case $N > T$, in this regime χ_N is bounded away from zero. Then the rate for the estimated factors is T^{-1}, which is still the same as the PCA estimators.

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 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 $\mathbf{\Sigma}_u$ 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 $\mathbf{\Sigma}_u$ by:

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

The feasible test statistic is defined as

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

We reject the null hypothesis for large values of S. It is straightforward to allow $\mathbf{\Sigma}_u$ 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.

5.2 The Limiting Distribution under H_0

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

We now state the null distribution in the following theorem.

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

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

Remark 5.1. The required rate conditions are determined by the nature of the high-dimensional Wald-test statistic. They respectively reflect three sources of statistical errors, (i) The high-dimensional Wald statistic requires estimating the $N \times N$ covariance matrix of $\mathbf{u}_t$, whose estimation accuracy prefers a larger T but a smaller (J, N), giving rise to the condition $J^4 N \log N = o(T^{3/2})$. This is essential for inferences using large-covariance matrices; (ii) Since the factors are estimated, the condition $T = o(N^2)$ is required to guarantee the asymptotic accuracy of estimating the unknown factors. Importantly, we allow either $N/T \to \infty$ or $T/N \to \infty$; (iii) Finally, the sieve approximation error for $\mathbf{g}(\mathbf{x}_t)$ should be negligible, giving rise to $N\sqrt{T} = o(J^{2\eta-1})$. In other words, the last two rate constraints can be relaxed if a simpler model is considered: a model with observed factors and parametric $\mathbf{g}(\mathbf{x}_t)$.

5.3 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. For each given group of observable proxies, we set the number of common factors K equals the number of observable proxies. We calculate the daily excess returns for the stocks in S&P 500 index that have complete daily closing prices from January 2005 to December 2013. The data, collected from CRSP, contains 393 stocks with a time span of 2265 trading days. Two stylized features of S&P 500 daily returns are asymmetry and heavy tails. The proxy factors $(\mathbf{x}_t)$ are chosen to be the Fama-French 3 or 5 factors and 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. In this study, we consider three groups of proxy factors 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).

Q: transpose the table to make it empty; requires probably 4 rows

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. The sieve basis is chosen as the additive Fourier basis with $J = 5$. We set the

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

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

tuning parameter $\alpha_T = C\sqrt{\frac{T}{\log(NJ)}}$ with constant C selected by the 5-fold cross validation. The plots of S under various scenarios are reported in Figure 5.1.

According to 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.

Moreover, we also used the monthly excess returns for the stocks in S&P 500 index 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:

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

where there are 21 trading days per month.

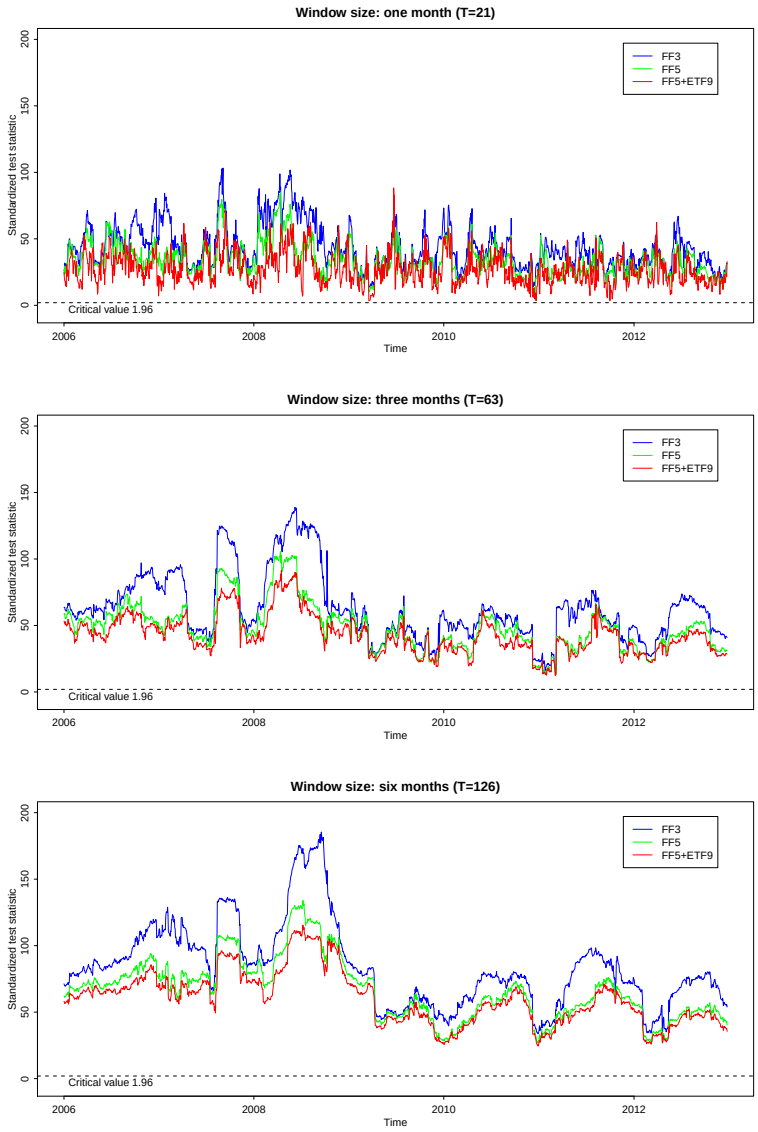

Figure 5.1: S&P 500 daily returns: plots for standardized test statistic S for various window sizes. The dotted line is critical value 1.96.

The results 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

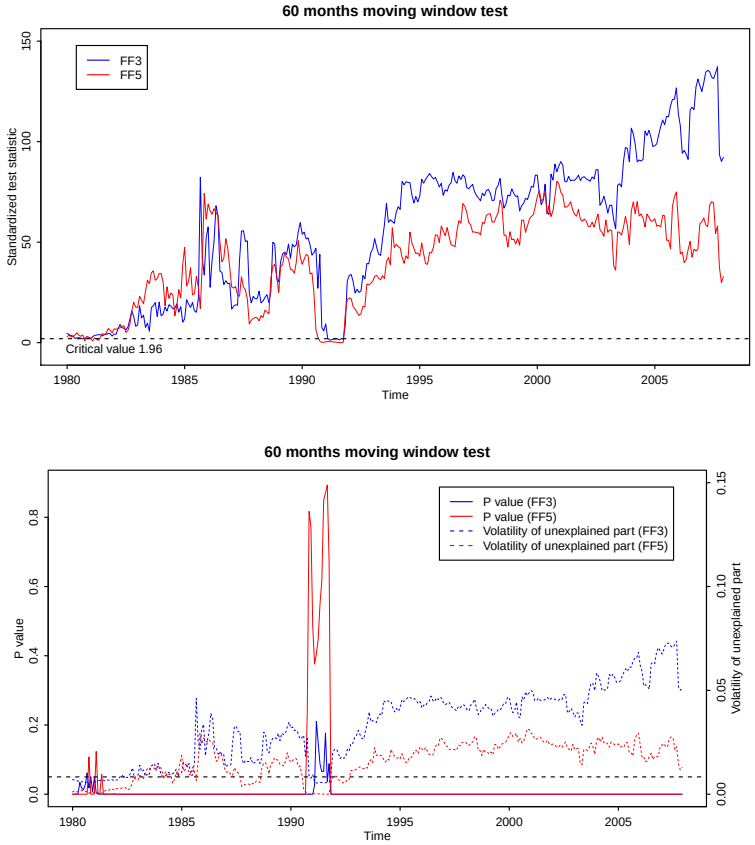

Figure 5.2: S&P 500 monthly returns: plots for standardized test statistic S, P-value and the volatility of the part of factors that can not be explained by the proxy factors.

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 one year excess bond return with maturity of two years. Using the proposed SPCA method, this section develops a new way of incorporating the explanatory power of the observed characteristics, 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.

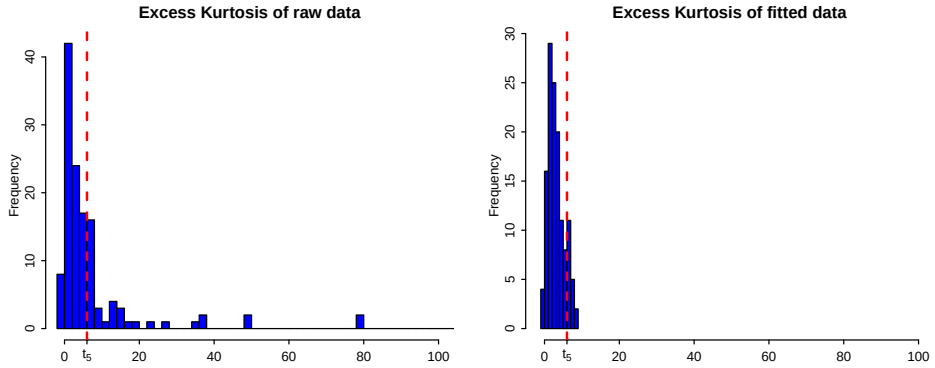

Figure 6.1: Excess kurtosis of the macroeconomic panel data. Left panel shows 43 among 131 series in the raw data are heavy tailed. Right panel shows the robustly fitted data $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ are no longer severely heavy-tailed.

6.2 Forecast results

We conduct one-month-ahead out-of-sample forecast of the bond excess returns. The forecast uses the information in the past 240 months, starting from January 1984 and rolling forward to December 2003. We compare three approaches to estimating the factors: SPCA, SPCA-LS, and the usual PCA. Also we consider two forecast models as follows:

$$\text{Linear model:} \qquad z_{t+1} = \alpha + \boldsymbol{\beta}'\mathbf{W}_t + \epsilon_{t+1}, \tag{6.1}$$

$$\text{Multi-index model:} \qquad z_{t+1} = h(\boldsymbol{\psi}_1'\mathbf{W}_t,\ \cdots,\ \boldsymbol{\psi}_L'\mathbf{W}_t) + \epsilon_{t+1}, \tag{6.2}$$

where α is the intercept and h is a nonparametric function. The covariate $\mathbf{W}_t$ is either $\mathbf{f}_t$ or an augmented vector $(\mathbf{f}_t', \mathbf{x}_t')'$. Here, the latent factors $\mathbf{f}_t$ are used by the three methods mentioned above in order to compare their effectiveness. The multi-index model allows more general nonlinear forecasts and are estimated by using 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 a weighted additive model $h(\boldsymbol{\psi}_1'\mathbf{W}_t, \ \cdots, \ \boldsymbol{\psi}_L'\mathbf{W}_t) = \sum_{l=1}^{L} g_l(\boldsymbol{\psi}_l'\mathbf{W}_t)$. Each individual nonparametric function $g_l(\cdot)$ is smoothed by the local linear approximation.

The performance of each method is assessed by the out-of-sample R^2. Let $\widehat{z}_{T+t+1|T+t}$ be the forecast of z_{T+t+1} using the data of the previous T months: $1+t, ..., T+t$ for $T = 240$ and $t = 0, ..., 239$. The forecast performance is assessed by the out-of-sample R^2, defined as

$$R^2 = 1 - \frac{\sum_{t=0}^{239}(z_{T+t+1} - \widehat{z}_{T+t+1|T+t})^2}{\sum_{t=0}^{239}(z_{T+t+1} - \bar{z}_t)^2},$$

where $\bar{z}_t$ is the sample mean of z_t over the sample period $[1+t, T+t]$. The R^2 of various methods are reported in Tables 6.2 and 6.3 respectively. 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 44.6% 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 by SPCA and SPCA-LS 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 SPCA method can robustly estimate the factors and 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 of 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 (%) for **linear model**: 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
$\mathbf{f}_t$	38.0	32.7	25.6	22.9	37.4	33.4	25.4	22.6	23.0	20.7	16.8	16.5
$(\mathbf{f}_t', \mathbf{x}_t')'$	37.7	32.4	25.4	22.7	37.1	31.9	25.3	22.1	23.9	21.4	17.4	17.5

Table 6.3: Forecast out-of-sample R^2 (%) for **multi-index model**: 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
$\mathbf{f}_t$	44.6	43.0	38.8	37.3	41.2	39.1	35.2	34.1	30.1	25.5	23.2	21.3
$(\mathbf{f}_t', \mathbf{x}_t')'$	41.5	38.7	35.2	33.8	41.1	35.7	32.2	30.0	30.8	26.3	24.6	22.0

7 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.

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** 359377.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics* **37** 577–580.

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.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013). Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics* **41** 1055.

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.

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** 271293.

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.

Lee, Y. K., Mammen, E. and Park, B. U. (2012). Projection-type estimation for varying coefficient regression models. *Bernoulli* **18** 177–205.

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** 958975.

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.

Maronna, R., Martin, R. D. and Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

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.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* 1617–1642.

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.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, vol. 589. John wiley & sons.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics* 580–615.

Song, S., Härdle, W. K. and Ritov, Y. (2014). Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series. *The Econometrics Journal* **17** S101–S131.

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.

Wang, L., Kai, B. and Li, R. (2009). Local rank inference for varying coefficient models. *Journal of the American Statistical Association* **104** 1631–1645.

Wang, W. and Fan, J. (2015). Asymptotics of empirical eigen-structure for high dimensional spiked covariance. *Annals of Statistics* **45** 1342-1374. 1631–1645.