Robust Factor Models with Covariates

Jianqing Fan* Yuan Ke† Yuan Liao‡

September 24, 2016

Abstract

We study factor models when the latent factors can be explained partially by several observed covariates. In financial factor models for instance, the unknown factors can be reasonably well predicted by a few covariates, such as the Fama-French factors. To incorporate the explanatory power of these covariates, we propose a two-step estimation procedure: (i) regress the data onto the observables, and (ii) take the principal components of the fitted data to estimate the loadings and factors. With those covariates, the factors can be estimated accurately even if the cross-sectional dimension is mild. The proposed estimator is robust to possibly heavy-tailed distributions, which are encountered in many applications. Empirically, we apply the model to forecasting US bond risk premia, and find that the observed economic covariates contain strong explanatory powers of the factors. The gain of forecast is more substantial when these covariates are incorporated to estimate the common factors than directly used for forecasts.

Keywords: Heavy tails, Forecasts, Large dimensions, Principal components

*Department of Operations Research and Financial Engineering, Bendheim Center for Finance, Princeton University

†Department of Operations Research and Financial Engineering, Princeton University

‡Department of Economics, Rutgers University

1 Introduction

This paper provides a general method for factor models when the factors depend on several observed explanatory variables. Consider the following static factor model:

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

where the latent factors $\mathbf{f}_t$ can be partially explained by a vector of observables $\mathbf{x}_t$:

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

for some nonparametric function $\mathbf{g} = E(\mathbf{f}_t|\mathbf{x}_t)$. Here $\mathbf{y}_t = (y_{1t}, ..., y_{NT})'$ is the outcome; $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N)'$ is an $N \times K$ matrix of unknown loadings; $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})'$ denotes the idiosyncratic vector. In (1.2), $\boldsymbol{\gamma}_t$ is interpreted as the factors' components that cannot be explained by the observables, whose covariance $\text{cov}(\boldsymbol{\gamma}_t)$ may or may not be close to zero. When $\text{cov}(\boldsymbol{\gamma}_t)$ is close to zero, the true factors are mostly explained by the observables $\mathbf{x}_t$; the latter is then interpreted as a good proxy of the true factors.

To incorporate the explanatory power of $\mathbf{x}_t$, we propose a *robust proxy-regressed* method to estimate the factors and loadings. The method consists of two major steps:

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

(ii) take the principal components of $(\widehat{\mathbf{y}}_1, ..., \widehat{\mathbf{y}}_T)$ to estimate the loadings and factors.

Since $\mathbf{x}_t$ is uncorrelated with $\mathbf{u}_t$, the regression step effectively removes the effects of idiosyncratic components. As a result, both the loadings and the $\mathbf{g}$ function can be identified (up to a rotation) under any given N. In addition, when $\boldsymbol{\gamma}_t$ is near zero ($\mathbf{x}_t$ nearly fully explains $\mathbf{f}_t$, which is a testable statement to be studied in Section 4), the estimated $\mathbf{g}(\mathbf{x}_t)$ can be directly used as factor estimators, whose rate of convergence is nearly $O_P(T^{-1} + (NT)^{-1/2})$,

and is faster than the usual estimates when $N = o(T^2)$. This shows that it is possible to estimate the factors well when the dimension is not very large relative to the sample size.

The proposed estimation procedure is robust to possibly heavy-tailed errors. Heavy-tailed and skewed data are commonly seen in many applications. Indeed, by examining the kurtosis of the dataset commonly used for diffusion index forecast (Stock and Watson, 2002; Ludvigson and Ng, 2009), we find that most of these variables have heavier tails than the t-distribution with degrees of freedom five. Most of the existing methods (PCA, MLE, etc.), however, are known to be sensitive to the tail distribution, whose scope of applications are therefore limited. We employ Huber (1964)'s robust M-estimation with a diverging regularity parameter in step (i) of our estimation. This demonstrates another advantage of our estimation procedure: the regression step 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). To our best knowledge, factor models with this type of distributions have not been studied previously.

We also consider testing

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

Under the null hypothesis, $\mathbf{x}_t$ fully explains the true factors. In empirical applications with "observed factors", what have been often used are in fact proxies of the true factors. Hence our proposed test can be applied to empirically validate the explanatory power of these "observed factors".

As a general methodology, our proposed method for factor models has wide range of applications. For instance, in factor-based forecasts, consider forecasting an outcome variable Z_t using a multi-index model:

$$Z_t = h(\boldsymbol{\psi}_1'\mathbf{X}_t, ..., \boldsymbol{\psi}_L'\mathbf{X}_t) + \varepsilon_t, \quad \mathbf{X}_t = (\mathbf{f}_t', \mathbf{x}_t')', \quad t = 1, ..., T, \tag{1.3}$$

where $(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_L)$ denote a set of regression indices. The model depends on a few unobserved factors $\mathbf{f}_t$, which can be estimated from a large number of variables $\mathbf{y}_t$ through model (1.1). For instance, in treatment effect studies, Z_t represents the outcomes of interest, and $\mathbf{y}_t$ denotes a potentially very large number of observed demographic variables for the individual t, and $\mathbf{x}_t$ contains the treatments. In the empirical study, we forecast the risk premia of U.S. government bonds. We find that the observed covariates contain strong explanatory powers of the factors. Incorporating them in the estimation of factors leads to a substantially improved out-of-sample forecast compared to the usual procedures that directly use them for forecasts.

Another application arises from *supervised singular value decomposition model*, (Li et al., 2016), and can be used to understand underlying patterns of genetic variations among tumors. For instance, in breast cancer data, a large number of gene expression measurements can be driven by a few common factors. In the Cancer Genome Atlas (TCGA) project (Network et al., 2012), we have additional information of cancer subtype for each sample, which is available as $\mathbf{x}_t$. 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 (SNP) data on the same group of subjects. It is known that SNPs drive underlying structure in the gene expression data, which can potentially lead to a better understanding in the studies.

Various methods have been developed in the literature to estimate the common factors, including principal components analysis (PCA, e.g., Connor and Korajczyk (1986); Stock and Watson (2002)), maximum likelihood estimate (MLE, Doz et al. (2012); Bai and Li (2012)), Kalman filtering (Doz et al., 2011), among others. We study the *static factor model*, which is different from the *dynamic factor model*. The dynamic model allows more general infinite dimensional representations using the frequency domain PCA (Brillinger,

1981). We refer to Forni et al. (2000, 2005); Hallin and Liška (2007) for the literature, among others. There is also an extensive literature on prediction/forecast based factor models. Improved estimation of factors is particularly important for predictions and forecasts. The literature includes, Stock and Watson (2002); Bernanke et al. (2005); Bai and Ng (2008); Ludvigson and Ng (2010); Kim and Swanson (2014); Cheng and Hansen (2015), among many others.

The rest of the paper is organized as follows. Section 2 overviews the method and defines the estimators. Section 3 presents the general asymptotic theory of the estimators. Section 4 proposes a test on the explanatory power of the covariates on the true factors. Section 5 provides simulations and Section 6 applies the methods to an empirical application on bond risk premia. Finally Section 7 concludes. The appendix contains additional numerical studies and all the technical proofs.

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 also 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 (also called operator norm), ℓ_1-norm, and elementwise 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 $a_T \gg b_T$ 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 Covariate-Based Estimator

2.1 Identification

Suppose that there is a 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

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

and thus

$$E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)' = \mathbf{\Lambda} E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\mathbf{\Lambda}'. \tag{2.2}$$

Suppose the following normalization conditions hold: $\frac{1}{N}\mathbf{\Lambda}'\mathbf{\Lambda} = \mathbf{I}_K$, and that $E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ is a diagonal matrix, with distinct diagonal entries. Then taking expectation on both sides of (2.2), and right multiplying by $\mathbf{\Lambda}/N$, by the normalization condition, we reach:

$$\frac{1}{N}E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}\mathbf{\Lambda} = \mathbf{\Lambda} E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}. \tag{2.3}$$

Since $E(\mathbf{y}_t|\mathbf{x}_t)$ is identified by the data generating process with observables $\{(\mathbf{y}_t, \mathbf{x}_t)\}_{t\leq T}$, we see that the columns of $\frac{1}{\sqrt{N}}\mathbf{\Lambda}$ (up to a sign change) are identified as the eigenvectors corresponding to the first $K = \dim(\mathbf{f}_t)$ eigenvalues of $E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$. Left multiplying $\mathbf{\Lambda}'/N$ on both sides of (2.1), one can see that $E(\mathbf{f}_t|\mathbf{x}_t)$ is also identified as:

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

We impose the normalization conditions above to facilitate our heuristic arguments. In this paper, these normalization conditions are not imposed. Then the same argument shows that $\mathbf{\Lambda}$ and $\mathbf{g}(\mathbf{x}_t)$ can be identified up to a matrix transformation. It is important to note that here the identification of $\mathbf{\Lambda}$ (or its transformation) is "exact" in the sense that it can be written as leading eigenvectors of identified covariance matrices for any given N. This is in contrast to the "asymptotic identification" (as $N \to \infty$) as in Chamberlain and Rothschild (1983) where the loading matrix (or its transformation) is identified only when

there are sufficiently large number of cross-sectional units. Here the exact identification is achieved due to the fact that the conditional expectation operation $E(\cdot|\mathbf{x}_t)$ removes the effects of idiosyncratic components in the equality (2.1).

The key assumption to be made about the role of $\mathbf{x}_t$ is as follows:

Assumption 2.1. *There are $\underline{c}, \bar{c} > 0$ so that all the eigenvalues of $E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ are confined in $[\underline{c}, \bar{c}]$.*

This assumption requires that the observed characteristics $\mathbf{x}_t$ should have an explanatory power for $\mathbf{f}_t$, which is essential whenever $\mathbf{x}_t$ is incorporated in the estimation procedure.

2.2 Definition of the estimators

The identification strategy motivates us to estimate $\mathbf{\Lambda}$ and $E(\mathbf{f}_t|\mathbf{x}_t)$ respectively by $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{g}}(\mathbf{x}_t)$ as follows. Let $\mathbf{\Sigma} := E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$, and let $\widehat{\mathbf{\Sigma}}$, $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ be some estimator of $\mathbf{\Sigma}$ 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 assume that $\text{cov}(\boldsymbol{\gamma}_t|\mathbf{x}_t) = \text{cov}(\boldsymbol{\gamma}_t)$ is independent of $\mathbf{x}_t$. In general, $\text{cov}(\boldsymbol{\gamma}_t) > 0$ might not vanish and we estimate $\mathbf{f}_t$ directly using OLS:

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

Finally, we estimate $\boldsymbol{\gamma}_t$ by: $\widehat{\boldsymbol{\gamma}}_t = \widehat{\mathbf{f}}_t - \widehat{\mathbf{g}}(\mathbf{x}_t) = \frac{1}{N}\widehat{\mathbf{\Lambda}}'(\mathbf{y}_t - \widehat{E}(\mathbf{y}_t|\mathbf{x}_t))$.

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

Several covariance estimators for $\mathbf{\Sigma}$ are available. As we show in Theorem 2.1 below, consistency for $\mathbf{\Sigma}$ is not a requirement for the consistency of $\widehat{\mathbf{\Lambda}}$, $\widehat{\mathbf{g}}(\mathbf{x}_t)$ or $\widehat{\mathbf{f}}_t$. The proposed estimators work so long as a "not-too-bad" estimator $\widehat{\mathbf{\Sigma}}$ is used.

Throughout the paper, we assume both N and T grow to infinity, while $K = \dim(\mathbf{f}_t)$ and $d = \dim(\mathbf{x}_t)$ are constant. Write $\mathbf{\Sigma}_{\Lambda,N} := \frac{1}{N}\mathbf{\Lambda}'\mathbf{\Lambda}$.

Assumption 2.2. *(i) All the eigenvalues of the $K \times K$ matrix $\mathbf{\Sigma}_{\Lambda,N}$ are bounded away from both zero and infinity;*
(ii) The eigenvalues of $\mathbf{\Sigma}_{\Lambda,N}^{1/2} E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}\mathbf{\Sigma}_{\Lambda,N}^{1/2}$ are distinct.

Theorem 2.1. *Suppose Assumptions 2.1 and 2.2 hold. Let $\widehat{\mathbf{\Sigma}}$ be such that*

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = o_P(N) \tag{2.4}$$

Then there exists an invertible $K \times K$ matrix $\mathbf{H}$ (whose dependence on N and T is suppressed for notational simplicity) such that, as $N, T \to \infty$,

$$\frac{1}{N}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H}\|_F^2 = o_P(1).$$

In addition, if the normalization conditions hold: $\mathbf{\Sigma}_{\Lambda,N} = \mathbf{I}_K$, and $E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ is a diagonal matrix, then $\mathbf{H} = \mathbf{I}_K$.

Recall that (2.4) uses the spectral norm for matrices. A useful sufficient condition for (2.4) is the element-wise convergence: $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = o_P(1)$, which is a very weak convergence requirement. Recall that $\mathbf{\Sigma} = E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$. Hence we construct an estimator $\widehat{E}(\mathbf{y}_t|\mathbf{x}_t)$ first as follows.

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. This setup includes structured nonparametric models such as the additive model and the parametric model (e.g., linear models). To adapt to different heaviness of the tails of idiosyncratic components and robustify the estimation, 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 $\alpha_T \to \infty$, we estimate the sieve coefficients $\mathbf{B}$ by the following convex optimization:

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

We then estimate $\mathbf{\Sigma}$ by

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

We regard α_T as a tuning parameter, which diverges in order to reduce the biases of estimating the conditional mean $E(\mathbf{y}_t|\mathbf{x}_t)$ when the distribution of $\mathbf{y}_t - E(\mathbf{y}_t|\mathbf{x}_t)$ is asymmetric. Throughout the paper, we shall set

$$\alpha_T = C\sqrt{\frac{T}{\log(NJ)}} \tag{2.5}$$

for some constant $C > 0$. We recommend choose the constant C through a multifold cross-

validation. This choice leads to the "least biased robust estimation", and we explain it in Section D in the appendix.

2.2.2 An alternative estimation method

Below we briefly discuss an alternative method, called "Sieve-LS covariance estimator".

Recall that $\Phi(\mathbf{x}_t) = (\phi_1(\mathbf{x}_t), ..., \phi_J(\mathbf{x}_t))'$ is a $J \times 1$ dimensional vector of sieve basis. Let

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

Then, the least-squares estimate of $E(\mathbf{Y}|\mathbf{x}_1, ..., \mathbf{x}_T)$ is simply $\mathbf{YP}$ and the sieve-LS covariance estimator for $\mathbf{\Sigma}$ is $\widetilde{\mathbf{\Sigma}} = \frac{1}{T}\mathbf{YPY}'$. While the sieve-LS is attractive due to its closed form, it is not suitable when the idiosyncratic distribution has heavier tails.

Nevertheless, in complicated real data analysis, applied researchers might like to use simple but possibly not robust estimators, for sake of simplicity. In that case, $\widetilde{\mathbf{\Sigma}}$ can still serve as an alternative estimator for $\mathbf{\Sigma}$. Our major estimation procedure for incorporating the information from $\mathbf{x}_t$ still carries over. As expected, our numerical studies in Section 5 demonstrate that sieve-LS performs well in light-tailed scenarios, but is less robust to heavy-tailed distributions.

3 Asymptotic Theory

3.1 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 3.1 (Tail distributions)**.** *(i) There are $\zeta_1, \zeta_2 > 2$, $C > 0$ and $M > 0$, so that for all $x > M$,*

$$\sup_{\mathbf{x}} \max_{i \le N} g_{e,i}(x|\mathbf{x}) \le C x^{-\zeta_1}, \quad \sup_{\mathbf{x}} \max_{i \le N} E(e_{it}^2 1\{|e_{it}| > x\}|\mathbf{x}_t = \mathbf{x}) \le C x^{-\zeta_2}. \tag{3.1}$$

(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 $\|\boldsymbol{\nu}\| = 1$,

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

(iii) For $\gamma_{kt} := f_{kt} - E(f_{kt}|\mathbf{x}_t)$, there is $v > 1$, so that $\max_{k \le K} E[E(\gamma_{kt}^4|\mathbf{x}_t)]^v < \infty$.

We allow e_{it} to have a tail distribution that is heavier than the exponential-type tails.

Assumption 3.2. *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 $\eta \ge 1$, as $J \to \infty$,*

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

Suppose $E(f_{kt}|\mathbf{x}_t = \cdot)$ belongs to a Hölder class: for some $r, \alpha > 0$,

$$\mathcal{G} = \{h : |h^{(r)}(x_1) - h^{(r)}(x_2)| \le L|x_1 - x_2|^\alpha\},$$

then this condition is satisfied by common basis such as the polynomials and B-splines with $\eta = 2(r + \alpha)/\dim(\mathbf{x}_t)$. If $E(f_{kt}|\mathbf{x}_t = \cdot)$ admits an additive structure and each component is in the Hölder class, then we can take $\eta = 2(r + \alpha)$.

Assumption 3.3. *There are $c_1, c_2 > 0$ so that*

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

Assumption 3.4. *(i)* $E(\mathbf{u}_t|\mathbf{f}_t, \mathbf{x}_t) = 0$, *and* $\max_{i\leq N} \|\boldsymbol{\lambda}_i\| < \infty$.

(ii) (serial independence) $\{\mathbf{f}_t, \mathbf{u}_t, \mathbf{x}_t\}_{t\leq T}$ *is independent and identically distributed;*

(iii) (weak cross-sectional dependence)

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

Assumption 3.4 (ii) requires serial independence, which can be restrictive. It is not difficult to allow for serial correlations when the data are not heavy-tailed, by using the sieve-LS estimator $\widetilde{\boldsymbol{\Sigma}}$ in place of the Huber's estimator $\widehat{\boldsymbol{\Sigma}}$. The analytical form of $\widetilde{\boldsymbol{\Sigma}}$ greatly facilitates the technical analysis so that the serial independence assumption can be weakened to strong mixing conditions. When the data are heavy-tailed, however, allowing for serial dependence is technically difficult due to the non-smooth Huber's loss. Nevertheless, our idea of using covariates would still be applicable. We conduct numerical studies when the data are serially correlated in the Appendix, and find that the proposed methods continue to perform well in the presence of serial correlations.

3.2 Asymptotic properties

We have the following result. Recall that $\mathbf{g}(\mathbf{x}_t) = E(\mathbf{f}_t|\mathbf{x}_t)$, and $\boldsymbol{\gamma}_t = \mathbf{f}_t - \mathbf{g}(\mathbf{x}_t)$.

Theorem 3.1 (Loadings)**.** *Suppose* $J^2 \log^3 N = O(T)$ *and* $J = O(N)$. *Under Assumptions 2.1–3.4, there is an invertible matrix* $\mathbf{H}$, *as* $N, T, J \to \infty$, *we have*

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 = O_P\left(\frac{J}{T} + \frac{1}{J^{2\eta-1}}\right), \tag{3.2}$$

$$\max_{i\leq N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\| = O_P\left(\sqrt{\frac{J\log N}{T}} + \frac{1}{J^{\eta-1/2}}\right). \tag{3.3}$$

Remark 3.1. The optimal rate for J in (3.2) is $J \asymp T^{1/(2\eta)}$, which results in

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

Here η represents the smoothness of $E(\mathbf{f}_t|\mathbf{x}_t = \cdot)$, as defined in Assumption 3.2. When η is sufficiently large, the rate is close to $O_P(T^{-1})$, which is faster than the rate of the usual principal components (PC) estimator when N is relatively small compared to T. In fact, the PC estimator $\widetilde{\boldsymbol{\lambda}}_i$ (e.g., Bai (2003)) satisfies, for some $\widetilde{\mathbf{H}}$,

$$\frac{1}{N}\sum_{i=1}^{N}\|\widetilde{\boldsymbol{\lambda}}_i - \widetilde{\mathbf{H}}'\boldsymbol{\lambda}_i\|^2 = O_P(T^{-1} + N^{-1}).$$

The estimation improvement is essentially due to a better estimation of the factors when N is relatively small. In the contrary, the usual PC estimator cannot estimate the factors well when N is small.

Define

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

Theorem 3.2 (Factors)**.** *Let $J \asymp J^*$. Suppose $(J^*)^2\log^3 N = O(T)$, $J^* = O(N)$, and Assumptions 2.1–3.4 hold. For $\mathbf{H}$ in Theorem 3.1, as $N, T \to \infty$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2 = O_P\left(\frac{J^*\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|}{T} + (\frac{1}{TN})^{1-1/(2\eta)} + (\frac{\log N}{T})^{2-3/(1+\eta)}\right), \tag{3.4}$$

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P\left(\frac{J^*\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|}{T} + \frac{1}{N} + (\frac{1}{TN})^{1-1/\eta} + (\frac{\log N}{T})^{2-4/(1+\eta)}\right), \tag{3.5}$$

where $\operatorname{cov}(\boldsymbol{\gamma}_t)$ denotes the covariance matrix of $\boldsymbol{\gamma}_t$.

Remark 3.2. The term $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|$ reflects the impact of the components in the factors that cannot be explained by $\mathbf{x}_t$. In the special case when $\operatorname{cov}(\boldsymbol{\gamma}_t) = 0$, we have $\mathbf{f}_t = \mathbf{g}(\mathbf{x}_t)$, and (3.4) implies

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

When η is large, this rate is faster than the usual PC estimator $\widetilde{\mathbf{f}}_t$, since the latter has the following rate of convergence (e.g., Stock and Watson (2002); Bai (2003)):

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

On the other hand, when $\operatorname{cov}(\boldsymbol{\gamma}_t)$ is bounded away from zero and η is large, the rate of convergence for $\widehat{\boldsymbol{\gamma}}_t$ is approximately

$$\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P(\frac{1}{T} + \frac{1}{N})$$

which is the same as that of the PC estimator for $\mathbf{f}_t$.

Remark 3.3. For a general J, the rates of convergence of the two factor components are:

$$\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\mathbf{g}}(\mathbf{x}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}_t)\|^2 = O_P\left(\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|}{T} + \frac{J}{TN} + \frac{J^3 \log N \log J}{T^2} + \frac{1}{J^{2\eta-1}} \right), \quad (3.6)$$

$$\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P\left(\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_s)\|}{T} + \frac{1}{N} + \frac{J^4 \log N \log J}{T^2} + \frac{1}{J^{2\eta-1}} \right). \quad (3.7)$$

In fact $J \asymp J^*$ is the optimal choice in (3.6) ignoring the term involving $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|$.

4 Testing the Explanatory Power

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

$$H_0 : \text{cov}(\boldsymbol{\gamma}_t) = 0. \tag{4.1}$$

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 $\mathbf{f}_t$.

A leading application example of this test is the validation of the Fama-French three-factor model in finance (Fama and French, 1992), which is one of the most celebrated ones in the empirical asset pricing. There has been much evidence that the three-factor model can leave a large percentage of the cross-sectional variations of stock returns unexplained. Different factor definitions have been explored and new factors have also be found. On the other hand, the Fama-French (FF) factors are good proxies for the true factors. Consequently, they form a natural choice for $\mathbf{x}_t$ and have explanatory power on the latent factors. Our general results in Section 3 immediately apply to the estimation of the loadings and true factors, incorporating the extra information from observing $\mathbf{x}_t$. Testing H_0 validates whether FF factors fully explain the true factors.

4.1 The test statistic

Our test is based on a 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.

4.2 Limiting distribution under H_0

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

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

Thus the limiting distribution is determined by that of $\bar{S} := \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t'\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\mathbf{u}_t$. Note that a cross-sectional central limit theorem implies, as $N \to \infty$,

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

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

We now state the null distribution in the following theorem.

Theorem 4.1. *Suppose* $\{u_{it}\}_{i\leq N}$ *is cross-sectionally independent, and Assumption 4.1 and assumptions of Theorem 3.2 hold. Then, when* $NJ^4 \log N \log J = o(T^{3/2})$, $T = o(N^2)$, $N\sqrt{T} = o(J^{2\eta-1})$, *as* $T, N \to \infty$,

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

Remark 4.1. 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 $T/N \to \infty$. Furthermore, the condition $N\sqrt{T} = o(J^{2\eta-1})$ requires the function $E(\mathbf{f}_t|\mathbf{x}_t = \cdot)$ be sufficiently smooth so that the sieve approximation error is negligible.

5 Simulation Studies

5.1 Model settings

In this section, we use simulated examples to demonstrate the finite sample performance of the proposed *robust proxy-regressed* (RPR) method and compare it with Sieve-LS (Section 2.2.2) and the regular PCA. We respectively specify five factors ($K = 5$) and five characteristics, and the sample size is $N = 50,\ T = 100$. The sieve basis is chosen as the additive

Fourier basis with $J = 5$. As shown in (2.5), the tuning parameter α_T in the Huber loss is of form $\alpha_T = C\sqrt{\frac{T}{\log(NJ)}}$. We selected the constant C by the 5 fold cross validation.

Consider the following model,

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \quad \text{and} \quad \mathbf{f}_t = \mathbf{g}(\mathbf{x}_t) + \boldsymbol{\gamma}_t, \quad t = 1, \cdots, T, \tag{5.1}$$

where $\boldsymbol{\gamma}_t$ are drawn from i.i.d. $N(\mathbf{0}, \sigma\mathbf{I})$ and σ is set to be 0.01, 0.3 and 1. The smaller the σ is, the more strongly $\mathbf{f}_t$ and $\mathbf{g}(\mathbf{x}_t)$ are correlated. In this study, $\mathbf{\Lambda}$ and $\mathbf{x}_t$ are drawn from i.i.d. standard normal distribution. The unknown function $\mathbf{g}(\cdot)$ is set to be one of the following 3 models:

(I) $\mathbf{g}(\mathbf{x}_t) = \mathbf{D}\mathbf{x}_t$, where $\mathbf{D}$ is a $K \times K$ matrix with each entry drawn from $U[1, 2]$;

(II) $\mathbf{g}(\mathbf{x}_t) = \sin(0.5\pi\mathbf{x}_t)$;

(III) $\mathbf{g}(\mathbf{x}_t) = \mathbf{0}$, which implies that $\mathbf{x}_t$ is irrelevant to $\mathbf{f}_t$.

For Model (III), we focus on $\sigma = 1$, as the observables $\mathbf{x}_t$ are independent of latent factors and other choices would yield similar results. In addition, $\mathbf{u}_t$ is drawn from either the Normal distribution $N(0, 8)$ or the Log-Normal distribution e^{1+2Z}, where Z is standard normal, which is asymmetric and heavy-tailed. The generated $\mathbf{u}_t$ has been centralized to have zero mean.

We also studied the numerical performances when the data are serially dependent. Due to the space limit, the serial dependence case is reported in the appendix.

5.2 In-sample Estimation

First, we compare the in-sample model fitting among RPR, Sieve-LS and PCA under different scenarios. For each scenario, we conduct 200 simulations. 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 1. Sieve-LS and RPR are comparable for light-tail distributions. This implies that we pay little price for robustness. However, when the error distributions have heavy tails, RPR yields much better estimation than other methods as expected. Sieve-LS out-performs PCA when $\mathbf{x}_t$ and $\mathbf{f}_t$ are well correlated. In general, PCA yields the worst estimation performance as it does not exploit the information in $\mathbf{x}_t$. When $\sigma = 1$, the observed $\mathbf{x}_t$ is not as informative and hence the performance of RPR and Sieve-LS deteriorates.

Table 1: Median of 5 canonical correlations of the estimated loadings/factors and the true ones when $N = 50, T = 100$: the larger the better

			Model (I)			Model (II)		
	$\mathbf{u}_t$	σ	RPR	Sieve-LS	PCA	RPR	Sieve-LS	PCA
Loadings	$N(0,\ 8)$	0.01	0.93	0.93	0.85	0.91	0.91	0.85
		0.3	0.91	0.91	0.90	0.87	0.87	0.87
		1.0	0.90	0.90	0.97	0.86	0.86	0.95
	LogN	0.01	0.68	0.33	0.26	0.67	0.34	0.25
		0.3	0.66	0.31	0.26	0.65	0.33	0.26
		1.0	0.63	0.30	0.28	0.62	0.29	0.27
Factors	$N(0,\ 8)$	0.01	0.94	0.94	0.77	0.95	0.95	0.85
		0.3	0.86	0.86	0.83	0.89	0.89	0.87
		1.0	0.85	0.85	0.95	0.88	0.88	0.95
	LogN	0.01	0.66	0.43	0.30	0.65	0.38	0.27
		0.3	0.64	0.41	0.33	0.60	0.36	0.29
		1.0	0.60	0.37	0.36	0.57	0.34	0.31

5.3 Out-of-sample Forecast

Consider $y_{t+1} = \boldsymbol{\beta}'\mathbf{f}_t + \epsilon_t$, 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 $U[0.5,\ 1.5]$ 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 by RPR, Sieve-LS or PCA. In each simulation, we generate $T + 50$ observations in total. For $s = 1, \cdots, 50$, we use the T observations $(y_s, ..., y_{T+s-1})$ to forecast y_{T+s}. We use PCA as the benchmark and define the relative mean squared error (RMSE) as:

$$\text{RMSE} = \frac{\sum_{s=1}^{50}(\widehat{y}_{T+s|T+s-1} - y_{T+s})^2}{\sum_{s=1}^{50}(\widetilde{y}^{PCA}_{T+s|T+s-1} - y_{T+s})^2},$$

where $\widehat{y}_{T+s|T+s-1}$ is the forecast y_{T+s} based on either RPR or Sieve-LS while $\widetilde{y}^{PCA}_{T+s|T+s-1}$ is the forecast based on PCA. For each scenario, we conduct 200 simulations and calculate the averaged RMSE as a measurement of the one-step-ahead out-of-sample forecast.

The results are presented in Table 2. Again, when the tails of error distributions are light, RPR and Sieve-LS perform comparably. But RPR outperforms Sieve-LS when the errors have heavy tails. On the other hand, Sieve-LS outperforms PCA when the correlation between $\mathbf{x}_t$ and $\mathbf{f}_t$ is strong. In general, the RPR method performs best under heavy-tailed cases. This suggests that in light-tailed scenarios, the sieve-LS is a good choice for the proposed proxy-regressed method when $\mathbf{x}_t$ has explanatory powers about the factors. In more general scenarios, the RPR is more robust to the tail distribution and does not pay much price even for the light tailed distributions.

Table 2: Mean RMSE of forecast when $N = 50, T = 100$: the smaller the better (with PCA as the benchmark)

		Model (I)		Model (II)		Model (III)	
$\mathbf{u}_t$	σ	RPR	Sieve-LS	RPR	Sieve-LS	RPR	Sieve-LS
	0.01	0.89	0.89	0.91	0.90		
$N(0,\ 8)$	0.3	0.94	0.93	0.93	0.93		
	1.0	1.04	1.03	1.04	1.03	1.45	1.44
	0.01	0.51	0.57	0.58	0.70		
LogN	0.3	0.50	0.57	0.60	0.72		
	1.0	0.48	0.59	0.62	0.75	1.19	1.18

6 Empirical Study of US Bond Risk Premia

6.1 Motivations

We apply our method to forecasting the risk premia of U.S. government bonds. The bond risk premia is defined through the one year excess bond return with n year maturity, which means 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. 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 $y_{T+1}^{(n)}$, the excess return with maturity of n years in period $T+1$, where

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

For a long time, the literature has found a significant predictive variation of the excess returns of U.S. government bonds. Recently, Ludvigson and Ng (2009, 2010) predicted the bond risk premia with observable variables based on a factor model. They achieved the out-

of-sample R^2 to be about 21% when forecasting one year excess bond return with maturity of two years. Using the proposed 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 find that the observed covariates have a strong explanatory power of the factors. Incorporating them in the factor estimation leads to a significantly better forecast rather than using them in forecast directly. In addition, the factors are robustly estimated, as many series in the dataset are heavy-tailed. Our method achieve $R^2 \approx 38.1\%$ using linear forecast model, and 44.8% using the nonlinear multi-index forecast model.

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 series (Ludvigson and Ng, 2010). The covariates $\mathbf{x}_t$ are listed in Table 3. Throughout this study, the sieve basis of $\mathbf{x}_t$ 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 been selected by the 5 fold cross validation.

6.2 Heavy-tailed data and robust estimations

We first study the excess kurtosis for the time series to assess the tail distributions. The left panel of Figure 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 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. This result illustrates the benefit of using our proposed robust method compared to the non-robust ones.

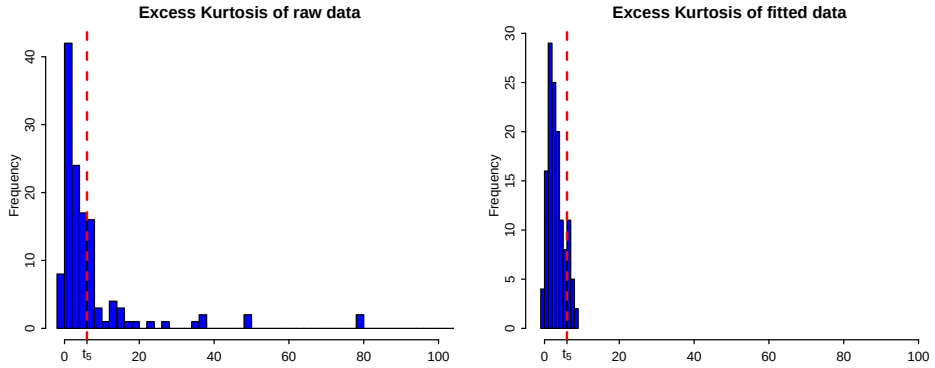

Figure 1: Excess kurtosis of the macroeconomic panel data

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.

Table 3: 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)

6.3 Forecast results

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. We compare three approaches to estimating the factors: RPR (the proposed method with robust estimation), Sieve Least Squares (Sieve-LS, the proposed method without robust estimation) and the usual PCA. Also we consider two forecast models as follows:

$$\text{Linear model:} \qquad y_{t+1} = \alpha + \boldsymbol{\beta}'\mathbf{X}_t + \epsilon_t, \tag{6.1}$$

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

where α is the intercept and h is a nonparametric function. The covariate $\mathbf{X}_t$ takes one of the following three forms: (i) $\mathbf{f}_t$; (ii) $(\mathbf{f}_t', \mathbf{x}_t')'$; (iii) $(\mathbf{f}_t', x_{i,t})'$, $i = 1, \cdots, 8$. The multi-index model allows more general nonlinear forecasts. The number of indices L is estimated by the ratio-based method suggested in Lam and Yao (2012) and should be no larger than the number of factors. The estimated L is usually 2 or 3. We approximate h using a weighted additive model $h(\boldsymbol{\psi}_1'\mathbf{f}_t,\ \cdots,\ \boldsymbol{\psi}_L'\mathbf{f}_t) = \sum_{l=1}^{L} \theta_l g_l(\boldsymbol{\psi}_l'\mathbf{f}_t)$. Each individual nonparametric function $g_l(\cdot)$ is smoothed by the local linear approximation.

Let $\widehat{y}_{T+t+1|T+t}$ be the forecast of y_{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

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

where $\bar{y}_t$ is the sample mean of y_t over the sample period $[1+t, T+t]$. The R^2 of various scenarios are reported in Tables 4 and 5 respectively. We notice that factors estimated by RPR and Sieve-LS can explain more variations in bond risk premia with all maturities than the ones estimated by PCA. The effect of estimating factors using PCA is negligible in forecasts only if $\sqrt{T}/N \to 0$ (Bai and Ng, 2008). However, the panel data studied here has a relatively small number of series ($N = 131$) compared with the length of the time span ($T = 240$). On the contrary, using the covariates, the proposed methods (RPR and Sieve-LS) can improve the estimation of factors even if N is relatively mild. RPR yields a 44.8% out-of-sample R^2 for forecasting the bond risk premia with two year maturity, which is much higher than the best out-of-sample predictor found in Ludvigson and Ng (2009). It is alsoobserved that the forecast based on either RPR or Sieve-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 RPR and Sieve-LS lead to significantly improved out-of-sample forecast on the US bond risk premia compared to the ones estimated by PCA.

2. As many series in the panel data are heavy-tailed, the RPR 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 risk premia is more substantial when these covariates are incorporated to estimate the common factors (using the proposed procedure) than directly used for forecasts.

Table 4: Forecast out-of-sample R^2 (%) for **linear model**: the larger the better. The bold figures represent larger R^2 than forecast with factors alone under the same scenario.

$\mathbf{X}_t$	RPR				Sieve-LS				PCA			
	Maturity(Year)				Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5	2	3	4	5
$\mathbf{f}_t$	38.1	32.9	25.7	23.0	37.4	33.4	25.4	22.6	32.6	28.2	23.3	19.7
$(\mathbf{f}_t', \mathbf{x}_t')'$	37.9	32.6	25.6	22.8	37.1	31.9	25.3	22.1	23.9	21.4	17.4	17.5
$(\mathbf{f}_t', x_{1,t})'$	38.0	32.8	25.5	22.9	37.1	29.2	24.3	21.9	**33.3**	27.9	22.9	19.4
$(\mathbf{f}_t', x_{4,t})'$	38.1	32.8	25.7	22.9	36.1	29.1	24.7	22.3	**33.6**	27.7	22.3	**20.0**
$(\mathbf{f}_t', x_{8,t})'$	38.1	32.9	25.6	22.9	37.3	32.4	25.3	22.5	**34.8**	**32.0**	**27.1**	**24.3**

Table 5: Forecast out-of-sample R^2 (%) for **multi-index model**: the larger the better. The bold figures represent larger R^2 than forecast with factors alone under the same scenario.

$\mathbf{X}_t$	RPR				Sieve-LS				PCA			
	Maturity(Year)				Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5	2	3	4	5
$\mathbf{f}_t$	44.8	43.2	38.9	37.6	41.2	39.1	35.2	34.1	34.5	32.1	27.3	23.7
$(\mathbf{f}_t', \mathbf{x}_t')'$	41.7	39.0	35.6	34.1	41.1	35.7	32.2	30.0	30.8	26.3	24.6	22.0
$(\mathbf{f}_t', x_{1,t})'$	43.4	38.2	34.5	30.9	39.5	37.3	32.2	28.8	**39.4**	**36.9**	**31.7**	**28.5**
$(\mathbf{f}_t', x_{4,t})'$	41.5	39.8	35.4	33.2	38.3	35.6	32.0	29.1	**36.2**	**34.4**	**30.7**	**28.2**
$(\mathbf{f}_t', x_{8,t})'$	41.1	38.9	34.6	30.2	39.0	36.3	31.6	26.8	**35.0**	**33.2**	**28.6**	**24.2**

7 Conclusions

We study factor models when the factors depend on observed explanatory characteristics. In financial factor pricing models for instance, the factors are approximated by a few observable proxies, such as the Fama-French factors. To incorporate the explanatory power

of these observed covariates, we propose a two-step estimation procedure: (i) regress the data onto the observables, and (ii) take the principal components of the fitted data to estimate the loadings and factors. The proposed estimator is robust to possibly heavy-tailed distributions, which is found to be the case for many time series in applications. The factors can be estimated accurately even if the cross-sectional dimension is mild. Empirically, we apply the model to forecasting US bond risk premia, and find that the observed economic covariates contain strong explanatory powers of the factors. The gain of forecast is more substantial when these covariates are incorporated to estimate the common factors than directly used for forecasts.

A Proof of Theorem 2.1

The proof of Theorem 2.1 is given below. The proofs of remaining theorems are given in the supplementary material. We denote by $\{\widehat{\lambda}_i, \widehat{\boldsymbol{\xi}}_i\}$ and $\{\lambda_i, \boldsymbol{\xi}_i\}$ respectively as the eigenvalues-vectors of $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma} = E\{E(\mathbf{y}_t|\mathbf{x}_t)E(\mathbf{y}_t|\mathbf{x}_t)'\}$, where the eigenvalues are in descending order.

Step 1. Show that $\lambda_i = (c_i + o(1))N$, $i = 1, ..., K$, for some constants $c_1 > c_2 > ... > c_K$ that are in $[\underline{c}, \bar{c}]$. In fact, from (2.2), we have $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}\boldsymbol{\Lambda}'$. Hence the first K eigenvalues are the same as those of

$$E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}^{1/2}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}^{1/2}$$

which are also the same as those of $N\boldsymbol{\Sigma}_{\Lambda,N}^{1/2}E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}\boldsymbol{\Sigma}_{\Lambda,N}^{1/2}$. By Assumptions 2.1 and 2.2, these eigenvalues are distinct, and can be written as $N(c_i + o(1))$ for some $c_i > 0$, $i = 1, ..., K$. Let them be ordered so that $c_1 > ... > c_K$. This proves the claim.

Step 2. Show that the first K eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ satisfy: with probability approaching

one, $(c_i + c_{i+1} + o(1))N/2 \leq \widehat{\lambda}_i \leq (c_i + c_{i-1} + o(1))N/2, \quad i = 1, ..., K$. The remaining eigenvalues are either bounded or growing uniformly at rate $o_P(N)$.

In fact, by Weyl's theorem (cited in Fan et al. (2013)), uniformly in $i \leq N$, $\max_{i \leq N} |\widehat{\lambda}_i - \lambda_i| = o_P(N)$. Therefore, with probability approaching one, by step 1, for $i \leq K-1$,

$$(c_i+c_{i+1}+o(1))N/2 = \lambda_i-(c_i-c_{i+1})N/2 \leq \widehat{\lambda}_i \leq \lambda_i+(c_{i-1}-c_i)N/2 = (c_i+c_{i-1}+o(1))N/2.$$

For $i = K$, we have $(c_K + o(1))N/2 \leq \widehat{\lambda}_i \leq (c_K + c_{K-1} + o(1))N/2$.

In addition, note that $\lambda_{K+1} = ... = \lambda_N = 0$. Hence $\widehat{\lambda}_N \leq ... \leq \widehat{\lambda}_{K+1} = o_P(N)$.

Step 3. Show that $\max_{i \leq K} \|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\| = o_P(1)$. We respectively lower bound $|\widehat{\lambda}_{i-1} - \lambda_i|$ and $|\lambda_i - \widehat{\lambda}_{i+1}|$ for $i \leq K$. As for the first term, for any $i \leq K$, by Step 2, $\widehat{\lambda}_{i-1} - \lambda_i \geq (c_{i-1} + c_i + o(1))N/2 - c_i N = (c_{i-1} - c_i + o(1))N/2$ with probability approaching one. As for the second term, when $i \leq K-1$, we have $\lambda_i - \widehat{\lambda}_{i+1} \geq (c_i + o(1))N - (c_{i+1} + c_i)N/2 = (c_i - c_{i-1} + o(1))N/2$. When $i = K$, we have $\lambda_i - \widehat{\lambda}_{i+1} \geq (c_i + o(1))N/2$ with probability approaching one. Hence By Davis-Kahan Theorem (cited in Fan et al. (2013)), $\max_{i \leq K} \|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\| \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| / \min_{i \leq K} \min\{|\widehat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \widehat{\lambda}_{i+1}|\} = o_P(N)/N = o_P(1)$.

Step 4. Complete the proof. Let

$$\mathbf{L} = \boldsymbol{\Sigma}_{\Lambda,N}^{1/2} E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}\boldsymbol{\Sigma}_{\Lambda,N}^{1/2}.$$

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 $\mathbf{L}$, which are distinct values bounded away from zero by Assumption 2.2. Let $\mathbf{H} = \boldsymbol{\Sigma}_{\Lambda,N}^{-1/2}\mathbf{M}$. Then

$$\frac{1}{N}\boldsymbol{\Sigma}\boldsymbol{\Lambda}\mathbf{H} = \boldsymbol{\Lambda}E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}\boldsymbol{\Sigma}_{\Lambda,N}\mathbf{H} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\Lambda,N}^{-1/2}\mathbf{L}\boldsymbol{\Sigma}_{\Lambda,N}^{-1/2}\boldsymbol{\Sigma}_{\Lambda,N}\mathbf{H} = \boldsymbol{\Lambda}\mathbf{H}\mathbf{M}'\mathbf{L}\mathbf{M} = \boldsymbol{\Lambda}\mathbf{H}\mathbf{D}.$$

In addition, note that $(\mathbf{\Lambda H})'(\mathbf{\Lambda H}) = N\mathbf{I}_K$, hence the columns of $\mathbf{\Lambda H}/\sqrt{N}$ are the eigenvectors of $\mathbf{\Sigma}$, corresponding to the K nonzero eigenvalues. By step 3, we have

$$\frac{1}{\sqrt{N}}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda H}\| = o_P(1).$$

In addition, when $\mathbf{\Sigma}_{\Lambda,N} = \mathbf{I}_K$ and $E\{E(\mathbf{f}_t|\mathbf{x}_t)E(\mathbf{f}_t|\mathbf{x}_t)'\}$ is diagonal, $\mathbf{M} = \mathbf{I}_K$. Thus $\mathbf{H} = \mathbf{I}_K$.

References

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

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

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics* **146** 304–317.

Bernanke, B., Boivin, J. and Eliasz, P. (2005). Measuring monetary policy: A factor augmented vector autoregressive (favar) approach. *Quarterly Journal of Economics* **120** 387–422.

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

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*, vol. 36. SIAM.

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

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics* **186** 280–293.

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

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

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.

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.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *The Review of Economics and Statistics* **82** 540–554.

Forni, M., Hallin, M., Marc annd Lippi and Reichlin, L. (2005). The generalized dynamic factor model. *JASA* **100** 830–840.

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

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

Kim, H. H. and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics* **178** 352–367.

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.

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.

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.

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

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.