Robust Factor Models with Explanatory Proxies

Jianqing Fan* Yuan Ke† Yuan Liao‡

March 21, 2016

Abstract

We provide an econometric analysis for the factor models when the latent factors can be explained partially by several observed explanatory proxies. In financial factor models for instance, the unknown factors can be reasonably well predicted 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. To incorporate the explanatory power of these observed characteristics, we propose a new 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 are encountered by many macroeconomic and financial time series. With those proxies, the factors can be estimated accurately even if the cross-sectional dimension is mild. Empirically, 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 these characteristics are incorporated to estimate the common factors than directly used for forecasts.

JEL Classification: C38, C53, C58

Key words: Huber loss, Heavy tails, Forecasts, Fama-French factors, Large dimensions

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

†Department of Operations Research and Financial Engineering, Princeton University

‡Department of Mathematics, University of Maryland

1 Introduction

This paper provides an econometric analysis for the factor models when the factors depend on several observed explanatory variables. Consider the following factor model:

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

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

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

for some function $\mathbf{g} = E(\mathbf{f}_t|\mathbf{w}_t)$, a nonparametric function. Here $\mathbf{x}_t = (x_{1t}, ..., x_{NT})'$ 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 $\mathbf{w}_t$; the latter is then interpreted as the good proxy of the true factors.

Factor models are found to be extremely useful for summarizing the information of a large number of economic variables. In economic applications, it is often the case that common factors are associated with some time-dependent observables. In financial factor models for instance, the factors are explained by a few observable proxies, such as the Fama-French factors (Fama and French, 1992, 2015). In diffusion index forecasts, Stock and Watson (2002a,b) and Ludvigson and Ng (2009) identified seven factors that represent production outcomes, the housing variables, stock markets, etc. The identified factors are strongly associated with several directly measurable economic variables such as consumption-wealth variable, financial ratios (ratios of price to dividends or earnings), and term spread.

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

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

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

Since $\mathbf{w}_t$ is uncorrelated with $\mathbf{u}_t$, the regression step effectively removes the effects of idiosyncratic components. As a result, the loadings and $\mathbf{g}$ function can be identified (up to a rotation) under any given N, as apposed to being identified asymptotically in traditional studies (as $N \to \infty$, e.g., Chamberlain and Rothschild (1983)). In addition, when $\boldsymbol{\gamma}_t$ is near zero ($\mathbf{w}_t$ nearly fully explains $\mathbf{f}_t$, which is a testable statement), the estimated $\mathbf{g}(\mathbf{w}_t)$ 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 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. It is well known that returns of many macroeconomic and financial time series are heavy-tailed and skewed. Indeed, by examining the kurtosis of the macroeconomic dataset commonly used for diffusion index forecast (Stock and Watson, 2002a; 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. In particular, when the number of cross-sectional units is large, these estimation methods require the tail distribution of the errors to exponentially decay in order to achieve good statistical properties (Fan et al., 2013). The sensitivity to the tail distributions, therefore, limits the application scopes of these estimators. We employ Huber (1964)'s robust M-estimation with a diverging regularity parameters 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 $\mathbf{w}_t$, whose distribution is no longer heavy-tailed, and is suitable for the PCA step (ii).

We consider two specific applications of the model with explanatory variables $\mathbf{w}_t$ for the common factors.

Testing Proxy Factors for Financial Returns

In model (1.2), we test

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

where $\mathbf{w}_t$ represents a set of observable proxies to the true factors, (e.g., Fama-French factors). The null hypothesis is equivalent to $\boldsymbol{\gamma}_t = 0$ almost surely in the entire sampling period, under which the observed proxies fully explain the true factors. 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 complements 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. (2015a)). 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 volatility of unexplained components varies over time and drops significantly during the acceptance period.

Multi-Index Regression

With more accurately estimated latent factors, we also consider a multi-index regression model using both latent factors and observed attributes:

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

where Y_t represents an observed scalar outcome and $(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_L)$ denote a set of regression indices. Here $\mathbf{f}_t$ and $\mathbf{w}_t$ are allowed to have overlapped components. For instance, in treatment effect studies, Y_t represents the outcome, and $\mathbf{w}_t$ denotes the treatments and a set of observed demographic variables for the individual t. In macroeconomic forecasts, $Y_t := y_{t+1}$ represents a scalar macroeconomic variable to forecast; and $\mathbf{w}_t$ denotes the observed characteristics at time t. The regression depends on a nonparametric link function h. In particular, it admits a factor-augmented linear model as a special case.

We estimate the common factors using the proposed *robust proxy-regressed* method from a large panel of variables. For prediction and forecast purposes, the multi-index model considered here does not require the identification of the index coefficients or complicated semi-parametric methods to individually estimate them. Using the "dimension reduction" techniques in the statistical literature (Li, 1991; Cook and Lee, 1999), our method only requires estimating the space spanned by the index coefficients.

In the empirical study, we apply the multi-index regression to forecast the risk premia of U.S. government bonds. We find that the observed macroeconomic characteristics contain strong explanatory powers of the factors. Incorporating these characteristics 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.

1.1 Further related literature, Organization and Notation

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 (2002a)), 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.

In the financial econometric literature, observed characteristics are often imposed to explain the loading matrix. For instance, Connor et al. (2012) considered a model where the loadings depend on a set of firm-specific characteristics (market capitalization, price-earning ratio, etc). They proposed a kernel-based method to iteratively estimate the loading matrix and factors. Their characteristics are observed cross-sectionally (firm-specific). Improved estimation of factors, on the other hand, is particularly important for predictions and forecasts. As is recently demonstrated by Bai and Liao (2016), more accurate estimations of the factors can substantially improve the out-of-sample forecasts. There is also an extensive literature on prediction/forecast based factor models. In addition to those discussed above, the literature includes, Stock and Watson (2002a); Bernanke et al. (2005); Bai and Ng (2008); Ludvigson and Ng (2010); Kim and Swanson (2014); Cheng and Hansen (2015), among many others.

Finally, the robust estimation is not rare in the econometric literature. For instance, it has been extensively studied in the time series literature (e.g., Andrews et al. (2007); Hill (2015)). The quantile regression is another type of robust estimation. However, quantile regression does not estimate conditional mean functions when the data are asymmetrically distributed.

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 new test on the specification of Fama-French factors, which tests whether the factor proxies fully explain the true factors. Section 5 applies the proposed method to multi-index nonlinear regression. Section 6 provides simulations and Section 7 applies the methods to an empirical application on bond risk premia. Finally Section 8 concludes. The appendix contains an empirical study of testing Fama-French factors, as well as 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 denote by $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|$, $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_{\max}$ the Frobenius norm, spectral norm (also called operator norm), ℓ_1-norm, and element-wise norm of a matrix $\mathbf{A}$, defined respectively by $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$. 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 Identification and Estimations

2.1 Identification

Suppose that there is a d-dimensional observable vector $\mathbf{w}_t$ that is: (i) associated with the latent factors $\mathbf{f}_t$, and (ii) mean-independent of the idiosyncratic term. We focus on the effect of observing $\mathbf{w}_t$ on the identification and estimation of the factors and loadings. Taking the conditional mean on both sides of (1.1), we have

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

and

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

Suppose the following normalization conditions hold: $\frac{1}{N}\mathbf{\Lambda}'\mathbf{\Lambda} = \mathbf{I}_K$, and that $E\{E(\mathbf{f}_t|\mathbf{w}_t)E(\mathbf{f}_t|\mathbf{w}_t)'\}$ 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:

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

Since $E(\mathbf{x}_t|\mathbf{w}_t)$ is identified by the data generating process with observables $\{(\mathbf{x}_t, \mathbf{w}_t)\}_{t\le T}$, 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 $E\{E(\mathbf{x}_t|\mathbf{w}_t)E(\mathbf{x}_t|\mathbf{w}_t)'\}$, which is assumed to have rank K. Let

$$\mathbf{\Sigma} := E\{E(\mathbf{x}_t|\mathbf{w}_t)E(\mathbf{x}_t|\mathbf{w}_t)'\}.$$

From (2.2), we have $\mathbf{\Sigma} = \mathbf{\Lambda} E\{E(\mathbf{f}_t|\mathbf{w}_t)E(\mathbf{f}_t|\mathbf{w}_t)'\}\mathbf{\Lambda}'$. Note that $E\{E(\mathbf{f}_t|\mathbf{w}_t)E(\mathbf{f}_t|\mathbf{w}_t)'\}$ is a $K \times K$ matrix, hence $\mathbf{\Sigma}$ has at most K nonzero eigenvalues. As a result, from (2.3), $\mathbf{\Lambda}$ corresponds to these K nonzero eigenvalues. Left multiplying $\mathbf{\Lambda}'/N$ on both sides of (2.1), one can see that $E(\mathbf{f}_t|\mathbf{w}_t)$ is also identified as:

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

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 $\mathbf{g}(\mathbf{w}_t)$ 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 $E(\cdot|\mathbf{w}_t)$ removes the effects of idiosyncratic components in the equality (2.1).

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

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

This assumption requires that the observed characteristics $\mathbf{w}_t$ should have an explanatory power for $\mathbf{f}_t$, which is essential whenever $\mathbf{w}_t$ is incorporated in the estimation procedure. For instance, when $\mathbf{w}_t$ represents the Fama-French factors (Fama and French, 1992), they are believed to be strongly associated with the "true" factors. In the ideal case that $\mathbf{w}_t$ fully explains $\mathbf{f}_t$, we have $E(\mathbf{f}_t|\mathbf{w}_t) = \mathbf{f}_t$ almost surely. Then this assumption is naturally satisfied so long as $E(\mathbf{f}_t\mathbf{f}_t')$ is well conditioned.

2.2 Definition of the estimators

2.2.1 The estimators

The identification strategy motivates us to estimate $\mathbf{\Lambda}$ and $E(\mathbf{f}_t|\mathbf{w}_t)$ respectively by $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{g}}(\mathbf{w}_t)$ as follows: let $\widehat{\mathbf{\Sigma}}$ be some covariance estimator of $\mathbf{\Sigma}$, whose definition will be clear below, and $\widehat{E}(\mathbf{x}_t|\mathbf{w}_t)$ be an estimator of $E(\mathbf{x}_t|\mathbf{w}_t)$. 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

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

Recall that $\mathbf{f}_t = \mathbf{g}(\mathbf{w}_t) + \boldsymbol{\gamma}_t$. We assume that $\text{cov}(\boldsymbol{\gamma}_t|\mathbf{w}_t) = \text{cov}(\boldsymbol{\gamma}_t)$ is independent of $\mathbf{w}_t$. When $\text{cov}(\boldsymbol{\gamma}_t)$ is small, $\widehat{\mathbf{g}}(\mathbf{w}_t)$ also serves as an estimator for the unknown factor $\mathbf{f}_t$. Above all, $\widehat{\mathbf{g}}(\mathbf{w}_t)$ is consistent for $\mathbf{f}_t$ so long as $\text{cov}(\boldsymbol{\gamma}_t) = o(1)$. In general, $\text{cov}(\boldsymbol{\gamma}_t) > 0$ might not

vanish and we estimate $\mathbf{f}_t$ directly using OLS:

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

Finally, we estimate $\boldsymbol{\gamma}_t$ by:

$$\widehat{\boldsymbol{\gamma}}_t = \widehat{\mathbf{f}}_t - \widehat{\mathbf{g}}(\mathbf{w}_t) = \frac{1}{N}\widehat{\mathbf{\Lambda}}'(\mathbf{x}_t - \widehat{E}(\mathbf{x}_t|\mathbf{w}_t)).$$

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

Several covariance estimators $\widehat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$ are available. Note that $\mathbf{\Sigma}$ is a high-dimensional matrix when N is large, hence it is difficult to estimate it consistently under usual matrix norms (e.g., Fronenius norm or spectral norm). Fortunately, as we show in Theorem 2.1 below, consistency for $\mathbf{\Sigma}$ is not a requirement for the consistency of $\widehat{\mathbf{\Lambda}}$, $\widehat{\mathbf{g}}(\mathbf{w}_t)$ or $\widehat{\mathbf{f}}_t$. The proposed estimators work so long as a "not-too-bad" estimator $\widehat{\mathbf{\Sigma}}$ is used. The required condition is mild.

Throughout the paper, we assume both N and T grow to infinity, while $K = \dim(\mathbf{f}_t)$ and $d = \dim(\mathbf{w}_t)$ 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 $\mathbf{\Sigma}_{\Lambda,N}^{1/2} E\{E(\mathbf{f}_t|\mathbf{w}_t)E(\mathbf{f}_t|\mathbf{w}_t)'\}\mathbf{\Sigma}_{\Lambda,N}^{1/2}$ are distinct.

Remark 2.1. We focus on strong factors throughout the paper, and condition (i) is the usual "pervasive condition" for approximate factor models. It requires that the common factors should impact on a non-negligible portion of the components of $\mathbf{x}_t$. As we take the principal components of $\widehat{\mathbf{\Sigma}}$ in the second step, we still require the first K eigenvalues of $\mathbf{\Sigma}$ to be large in order to signal the corresponding eigenvectors. This gives rise to the pervasive condition in the current context.

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* $E\{E(\mathbf{f}_t|\mathbf{w}_t)E(\mathbf{f}_t|\mathbf{w}_t)'\}$ *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 $\mathbf{\Sigma} = E\{E(\mathbf{x}_t|\mathbf{w}_t)E(\mathbf{x}_t|\mathbf{w}_t)'\}$. Hence we construct an estimator $\widehat{E}(\mathbf{x}_t|\mathbf{w}_t)$ first as follows.

Let $\Phi(\mathbf{w}_t) = (\phi_1(\mathbf{w}_t), ..., \phi_J(\mathbf{w}_t))'$ be a $J \times 1$ dimensional vector of sieve basis. Suppose $E(\mathbf{x}_t|\mathbf{w}_t)$ can be approximated by a sieve representation: $E(\mathbf{x}_t|\mathbf{w}_t) \approx \mathbf{B}\Phi(\mathbf{w}_t)$, 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 for different heaviness of the tails of idiosyncratic components to 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:

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

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

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

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

Throughout the paper, we shall set

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

for some constant $C > 0$. We recommend choose the constant C through a multifold cross-validation. We explain this choice in Section 3.2. Our method is particularly suitable for applications of financial and macroeconomic time series that often exhibit heavy tails (Balke and Fomby, 1994; Sakata and White, 1998; Atkinson et al., 1997). To our best knowledge, factor models with this type of distributions have not been studied previously.

2.3 Alternative estimation methods

We discuss some alternative estimation strategies.

2.3.1 Sieve-LS covariance estimator

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

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

Then, the fitted values of the least-squares estimate for $E(\mathbf{X}|\mathbf{w}_1, ..., \mathbf{w}_T)$ is simply $\mathbf{XP}$ and the sieve-LS covariance estimator for $\mathbf{\Sigma}$ is $\widetilde{\mathbf{\Sigma}} = \frac{1}{T}\mathbf{XPX}'$. 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 $\mathbf{w}_t$ still carries over. As expected, our numerical studies in Section 6 demonstrate that sieve-LS performs well in light-tailed scenarios, but is less robust to heavy-tailed distributions.

2.3.2 Panel data with interactive effects

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

$$\mathbf{x}_t = \mathbf{h}(\mathbf{w}_t) + \mathbf{\Lambda}\boldsymbol{\gamma}_t + \mathbf{u}_t, \quad \text{where } \mathbf{h}(\mathbf{w}_t) = \mathbf{\Lambda}\mathbf{g}(\mathbf{w}_t). \tag{2.5}$$

Alternatively, one may also consider the following model:

$$\mathbf{x}_t = \mathbf{h}(\mathbf{w}_t) + \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \tag{2.6}$$

for a nonparametric function $\mathbf{h}(\cdot)$, or simply a linear form $\mathbf{h}(\mathbf{w}_t) = \boldsymbol{\beta}\mathbf{w}_t$. Models (2.5) and (2.6) are known as the panel data models with interactive effects 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 (2.5) by

$$\min_{\mathbf{h},\mathbf{\Lambda},\boldsymbol{\gamma}_t} \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{h}(\mathbf{w}_t) - \mathbf{\Lambda}\boldsymbol{\gamma}_t\|^2. \tag{2.7}$$

But this approach is not appropriate in the current context when $\mathbf{w}_t$ almost fully explains $\mathbf{f}_t$ for all $t = 1, ..., T$. In this case, $\boldsymbol{\gamma}_t \approx 0$, and least squares (2.7) would be inconsistent. * In addition, $\mathbf{\Lambda}$ in (2.6) 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 (2.6) cannot be consistently estimated (Onatski, 2012) 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 Asymptotic Theory: The General Case

3.1 Assumptions

Let

$$e_{it} := x_{it} - E(x_{it}|\mathbf{w}_t).$$

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

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

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

*The inconsistency is due to the fact that $a\mathbf{\Lambda}\boldsymbol{\gamma}_t \approx \mathbf{\Lambda}\boldsymbol{\gamma}_t$ for *any* scalar a, since the true $\boldsymbol{\gamma}_t \approx 0$.

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

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

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

We allow e_{it} to have a tail distribution that is heavier than the exponential-type tails. Note that $e_{it} = u_{it} + \boldsymbol{\lambda}_i'\boldsymbol{\gamma}_t$. Therefore if the distribution of factors has a light tail (e.g., decays either exponentially or polynomially faster than that of u_{it}), then the required tail conditions on e_{it} directly carry over to u_{it}.

The following condition is regarding the sieve approximation.

Assumption 3.2. *For* $k = 1, ..., K$*, let* $\mathbf{v}_k = \arg\min_{\mathbf{v}} E(f_{kt} - \mathbf{v}'\Phi(\mathbf{w}_t))^2$*. Then there is* $\eta \geq 1$*, as* $J \to \infty$*,*

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

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

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

then this condition is satisfied by common basis such as the polynomials and B-splines with $\eta = 2(r + \alpha)/\dim(\mathbf{w}_t)$. If $E(f_{kt}|\mathbf{w}_t = \cdot)$ admits an additive structure and each component is in the Hölder class, then we can take $\eta = 2(r+\alpha)$. Furthermore, as co-movements of the cross-sectional units are driven by the common factors, this assumption ensures that $E(x_{it}|\mathbf{w}_t = \cdot)$ can be approximated by the sieve representation uniformly well across $i = 1, ..., N$.

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

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

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

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

(iii) (weak cross-sectional dependence)

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

Note that we allow for conditional heteroskedasticity and cross-sectional correlations in $\mathbf{u}_t$. A limitation of our theory is that serial independence is required, as required in Assumption 3.4 (ii). The serial independence is solely a technical condition for robust estimations. Unlike the usual principal components methods, the robust estimation based on Huber's loss does not have a closed form solution. In order to achieve sharp rates of convergence and limiting distributions, the asymptotic analysis relies on the uniform Bahadur representation (Bahadur (1966)) of the robust M-estimator:

$$\widehat{E}(x_{it}|\mathbf{w}_t) = E(x_{it}|\mathbf{w}_t) + \frac{1}{T}\sum_{s=1}^{T} \alpha_T \dot{\rho}(\alpha_T^{-1} e_{is}) \Phi(\mathbf{w}_s)' \mathbf{A} \Phi(\mathbf{w}_t) + z_{it} + \Delta_{it},$$

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

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

Here $\text{sgn}(z)$ denotes the sign function; $\mathbf{A}$ denotes the Hessian matrix of the expected Huber's loss function; z_{it} is the sieve approximation error of $E(x_{it}|\mathbf{w}_t)$; Δ_{it} is the remainder of the representation. The key technical argument is to bound the remainder uniformly over the cross-sectional units: $\max_{i \leq N} |\Delta_{it}|$. We appeal to the empirical process theories of van der Vaart and Wellner (1996) for this task, which requires the data be independently distributed.

Remark 3.1. When the data are not heavy-tailed, the sieve-LS estimator $\widetilde{\mathbf{\Sigma}}$ can be employed instead. In that case, the serial independence assumption can be replaced with strong mixing conditions to allow for serial correlations.

3.2 Choice of the tuning parameter

As mentioned earlier, throughout this paper, we take

$$\alpha_T = C\sqrt{\frac{T}{\log(NJ)}} \tag{3.2}$$

for a constant C chosen by the cross-validation. The Huber-estimator is biased for estimating the mean coefficient, whose population counterpart is

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

As α_T increases, the Huber loss behaves like a quadratic loss. In fact, we show in the appendix (Proposition D.1) that for $\mathbf{b}_i := \arg\min_{\mathbf{b}\in\mathbb{R}^J} E[x_{it} - \mathbf{b}'\Phi(\mathbf{w}_t)]^2$,

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

for an arbitrarily smal $\epsilon > 0$, where ζ_2 is defined in Assumption 3.1. Hence the bias decreases as α_T grows as expected. On the other hand, we shall investigate the uniform convergence (in $i = 1, ..., N$) of

$$\max_{i\leq N} \|\frac{1}{T}\sum_{t=1}^{T} \dot{\rho}(\alpha_T^{-1}e_{it})\Phi(\mathbf{w}_t)\|, \tag{3.3}$$

which is the leading term in the Bahadur expansion of the Huber-estimator. 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, where J is the number of sieve basis. Hence the choice (3.2) leads to the asymptotically least-biased robust estimation.

3.3 Asymptotic properties

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

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*

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\|^2 = O_P\left(\frac{J}{T} + \frac{1}{J^{2\eta-1}}\right), \tag{3.4}$$

$$\max_{i\leq N} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}'\boldsymbol{\lambda}_i\| = O_P\left(\sqrt{\frac{J\log N}{T}} + \frac{1}{J^{\eta-1/2}}\right). \tag{3.5}$$

Remark 3.2. The optimal rate for J in (3.4) 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 $E(x_{it}|\mathbf{w}_t = \cdot)$. 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*

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

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

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

Remark 3.3. The term $\|\operatorname{cov}(\boldsymbol{\gamma}_t)\|$ reflects the impact of the components in the factors that cannot be explained by $\mathbf{w}_t$. In the special case when $\operatorname{cov}(\boldsymbol{\gamma}_t) = 0$, we have $\mathbf{f}_t = \mathbf{g}(\mathbf{w}_t)$, and (3.6) implies

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

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 (2002a); 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 $\text{cov}(\boldsymbol{\gamma}_t)$ 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 3.4. For a general J, the rates of convergence of the two factor components are:

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{g}}(\mathbf{w}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{w}_t)\|^2 = O_P\left(\frac{J\|\,\text{cov}(\boldsymbol{\gamma}_t)\|}{T} + \frac{J}{TN} + \frac{J^3\log N\log J}{T^2} + \frac{1}{J^{2\eta-1}}\right), \quad (3.8)$$

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P\left(\frac{J\|\,\text{cov}(\boldsymbol{\gamma}_s)\|}{T} + \frac{1}{N} + \frac{J^4\log N\log J}{T^2} + \frac{1}{J^{2\eta-1}}\right). \quad (3.9)$$

In fact $J \asymp J^*$ is the optimal choice in (3.8) ignoring the term involving $\|\,\text{cov}(\boldsymbol{\gamma}_t)\|$.

We now present the asymptotic distribution for $\widehat{\boldsymbol{\gamma}}_t$, which can be used to derive the confidence interval for (rotated) $\boldsymbol{\gamma}_t$ for each fixed t. We introduce an additional assumption and some notation. Let $\boldsymbol{\Sigma}_u$ denote the covariance matrix of $\mathbf{u}_t$. Assumption 3.5 below is the cross-sectional central limit theorem, and is commonly imposed for the limiting distribution of estimated factors (e.g., Bai (2003)).

Assumption 3.5. *Suppose* $\lim_{N\to\infty}\frac{1}{N}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u\boldsymbol{\Lambda} = \mathbf{Q}$*, and for each fixed* t*,*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\lambda}_i u_{it} \to^d \mathcal{N}(0,\mathbf{Q}).$$

Define $\boldsymbol{\Sigma}_F = E\{E(\mathbf{f}_t|\mathbf{w}_t)E(\mathbf{f}_t|\mathbf{w}_t)'\}$, $\boldsymbol{\Sigma}_\Lambda = \lim_{N\to\infty}\frac{1}{N}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$. Let $\mathbf{V}$ be a $K\times K$ diagonal matrix, whose diagonal elements are the eigenvalues of $\boldsymbol{\Sigma}_\Lambda^{1/2}\boldsymbol{\Sigma}_F\boldsymbol{\Sigma}_\Lambda^{1/2}$, and $\boldsymbol{\Gamma}$ be the $K\times K$ matrix whose columns are the corresponding eigenvectors. Let $\mathbf{J} := \mathbf{V}^{-1/2}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_F^{1/2}$, $\mathbf{G} := E\mathbf{f}_t\Phi(\mathbf{w}_t)'$, and $\mathbf{S} := (E\Phi(\mathbf{w}_t)\Phi(\mathbf{w}_t)')^{-1}$. Finally, let

$$\mathbf{M}_t = \text{cov}(\boldsymbol{\gamma}_t)\boldsymbol{\alpha}_t'\mathbf{S}\boldsymbol{\alpha}_t - \text{cov}(\boldsymbol{\gamma}_t)\boldsymbol{\beta}_t\boldsymbol{\alpha}_t'\mathbf{S}\mathbf{G}' - (\text{cov}(\boldsymbol{\gamma}_t)\boldsymbol{\beta}_t\boldsymbol{\alpha}_t'\mathbf{S}\mathbf{G}')' + \mathbf{G}\mathbf{S}\mathbf{G}'\boldsymbol{\beta}_t'\,\text{cov}(\boldsymbol{\gamma}_t)\boldsymbol{\beta}_t,$$

where $\boldsymbol{\alpha}_t = \Phi(\mathbf{w}_t) - \mathbf{G}'\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\gamma}_t$, $\boldsymbol{\beta}_t = \boldsymbol{\Sigma}_F^{-1}\boldsymbol{\gamma}_t$. It can be shown that $\mathbf{M}_t$ is positive definite.

Theorem 3.3 (Limiting distribution for $\widehat{\boldsymbol{\gamma}}_t$)**.** *Suppose* $J^4 \log^2 N = o(T)$, $J^{1-\eta} = o(1/\sqrt{N} + 1/\sqrt{T})$. *Then under Assumptions 2.1–3.5, for each fixed* $t = 1, ..., T$,

$$(T^{-1}\mathbf{J}\boldsymbol{\Sigma}_\Lambda \mathbf{M}_t \boldsymbol{\Sigma}_\Lambda \mathbf{J}' + N^{-1}\mathbf{J}\mathbf{Q}\mathbf{J}')^{-1/2}(\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t) \to^d \mathcal{N}(0, \mathbf{I}).$$

All terms in the asymptotic variance can be estimated using their sample counterparts: respectively define $\widehat{\boldsymbol{\Sigma}}_F = \frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{g}}(\mathbf{w}_t)\widehat{\mathbf{g}}(\mathbf{w}_t)'$, $\widehat{\boldsymbol{\Sigma}}_\Lambda = \frac{1}{N}\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Lambda}}$, $\widehat{\mathbf{G}} = \frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{f}}_t \Phi(\mathbf{w}_t)'$, $\widehat{\mathbf{S}} = (\frac{1}{T}\sum_{t=1}^T \Phi(\mathbf{w}_t)\Phi(\mathbf{w}_t)')^{-1}$, $\widehat{\mathrm{cov}}(\boldsymbol{\gamma}_t) = \frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{\gamma}}_t\widehat{\boldsymbol{\gamma}}_t'$, and $\widehat{\mathbf{Q}} = \widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\widehat{\boldsymbol{\Lambda}}$, where the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_u$ for $\boldsymbol{\Sigma}_u$ can be easily constructed based on the residuals in the absence of cross-sectional correlations (see section 4 below, where we also present a covariance estimator allowing for cross-sectional correlations). Finally, $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\mathbf{M}_t$ can be estimated similarly:

$$\widehat{\mathbf{M}}_t = \widehat{\mathrm{cov}}(\boldsymbol{\gamma}_t)\widehat{\boldsymbol{\alpha}}_t'\widehat{\mathbf{S}}\widehat{\boldsymbol{\alpha}}_t - \widehat{\mathrm{cov}}(\boldsymbol{\gamma}_t)\widehat{\boldsymbol{\beta}}_t\widehat{\boldsymbol{\alpha}}_t'\widehat{\mathbf{S}}\widehat{\mathbf{G}}' - (\widehat{\mathrm{cov}}(\boldsymbol{\gamma}_t)\widehat{\boldsymbol{\beta}}_t\widehat{\boldsymbol{\alpha}}_t'\widehat{\mathbf{S}}\widehat{\mathbf{G}}')' + \widehat{\mathbf{G}}\widehat{\mathbf{S}}\widehat{\mathbf{G}}'\widehat{\boldsymbol{\beta}}_t'\widehat{\mathrm{cov}}(\boldsymbol{\gamma}_t)\widehat{\boldsymbol{\beta}}_t,$$

where $\widehat{\boldsymbol{\alpha}}_t = \Phi(\mathbf{w}_t) - \widehat{\mathbf{G}}'\widehat{\boldsymbol{\Sigma}}_F^{-1}\widehat{\boldsymbol{\gamma}}_t$, and $\widehat{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\Sigma}}_F^{-1}\widehat{\boldsymbol{\gamma}}_t$.

The following result provides the limiting distribution of $\widehat{\boldsymbol{\gamma}}_t$ with consistent covariance estimators.

Corollary 3.1. *Assume the assumptions of Theorem 3.3 and cross-sectional independence. As* $T, N \to \infty$,

$$(T^{-1}\widehat{\boldsymbol{\Sigma}}_\Lambda \widehat{\mathbf{M}}_t \widehat{\boldsymbol{\Sigma}}_\Lambda + N^{-1}\widehat{\mathbf{Q}})^{-1/2}(\widehat{\boldsymbol{\gamma}}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t) \to^d \mathcal{N}(0, \mathbf{I}).$$

4 Application: Testing Fama-French Factors

The Fama-French three-factor model (Fama and French, 1992) is one of the most celebrated ones in the empirical asset pricing. It takes into account the size and value effects, in addition to the market risk. Ever since its proposal, there is much evidence that the three-factor model can leave the cross-section of expected stock returns unexplained. To isolate exposures to the different dimensions of returns, 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 modeled the excess return r_{it} on security or portfolio i for period t as

$$r_{it} = \alpha_i + b_i r_{Mt} + s_i \mathrm{SMB}_t + h_i \mathrm{HML}_t + c_i \mathrm{RMW}_t + d_i \mathrm{CMA}_t + u_{it},$$

where r_{Mt}, SMB_t and HML_t are the three factors of (Fama and French, 1992), respectively representing 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"). Two additional factors were included: RMW (profitability) is the difference between the returns on diversified portfolios of stocks with robust and weak profitability, and CMA (investment) is the difference between the returns on diversified portfolios of low and high investment stocks. In addition, Fama and French (2015) 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 $\mathbf{w}_t$. 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. Our general results in Section 3 immediately apply to the estimation of the loadings and true factors, incorporating the extra information from observing $\mathbf{w}_t$.

4.1 Testing the explanatory power of the factor proxies

We shall use $\mathbf{w}_t$ as the "observed proxy" of the true factors, such as the Fama-French factors. We are interested in testing: (recall that $\boldsymbol{\gamma}_t = \mathbf{f}_t - E(\mathbf{f}_t|\mathbf{w}_t)$.)

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

Under H_0, $\mathbf{f}_t = E(\mathbf{f}_t|\mathbf{w}_t)$ over the entire sampling period $t = 1, ..., T$, implying that observed Fama-French factors $\mathbf{w}_t$ fully explain the true factors $\mathbf{f}_t$. The GRS test and related tests, in contrast, are designed to test the "zero-alpha" hypothesis ($H_0 : \alpha_i = 0$ for all $i = 1, ..., N$) using $\mathbf{w}_t$ as the factors in the empirical asset pricing model. The "zero-alpha" test is used to assess the proxy of the true factors only when the market is mean-variance efficient (Gungor and Luger, 2013). In contrast, our proposed test aims directly at the question whether the observed proxy $\mathbf{w}_t$ is adequate or not, without assuming the efficient market hypothesis.

Our method can also be used to test whether there are any missing factors in financial studies. To illustrate this, consider a simple example where there are four true factors, which

are characterized by four variables: $w_{1t}, w_{2t}, w_{3t}, w_{4t}$ as follows:

$$\begin{pmatrix} f_{1t} \\ f_{2t} \\ f_{3t} \\ f_{4t} \end{pmatrix} = \begin{pmatrix} F_1(w_{1t}, w_{2t}, w_{3t}, w_{4t}) \\ F_2(w_{1t}, w_{2t}, w_{3t}, w_{4t}) \\ F_3(w_{1t}, w_{2t}, w_{3t}, w_{4t}) \\ F_4(w_{1t}, w_{2t}, w_{3t}, w_{4t}) \end{pmatrix}, \quad t = 1, ..., T,$$

where $F_i(w_{1t}, w_{2t}, w_{3t}, w_{4t})$ are unknown functions (e.g., $F_i(w_{1t}, w_{2t}, w_{3t}, w_{4t}) = w_{it}$, $i = 1, ..., 4$). Suppose however, only w_{1t}, w_{2t}, w_{3t} are identified and measured (e.g., Fama-French three factors), but w_{4t} is missing. Then in our notation, $\mathbf{w}_t = (w_{1t}, w_{2t}, w_{3t})$, and we write the factors as

$$\begin{pmatrix} f_{1t} \\ f_{2t} \\ f_{3t} \\ f_{4t} \end{pmatrix} = \begin{pmatrix} g_1(w_{1t}, w_{2t}, w_{3t}) + \gamma_{1t} \\ g_2(w_{1t}, w_{2t}, w_{3t}) + \gamma_{2t} \\ g_3(w_{1t}, w_{2t}, w_{3t}) + \gamma_{3t} \\ g_4(w_{1t}, w_{2t}, w_{3t}) + \gamma_{4t} \end{pmatrix}, \quad \mathbf{g} = (g_1, ..., g_4)', \quad \boldsymbol{\gamma}_t = (\gamma_{1t}, ..., \gamma_{4t})'.$$

If at least one of the true factors depends on w_{4t}, then at least one of the γ_{it}'s must be nonzero. This can be detected by testing (4.1) as the test is equivalent to testing $\boldsymbol{\gamma}_t = 0$ for $t = 1, ..., T$ almost surely.

In our empirical study (Appendix A) on the S&P 500 constituents, we find that the null hypothesis is more often to be rejected using the daily data compared to the monthly data, possibly because daily data tend to demonstrate larger volatilities on $\boldsymbol{\gamma}_t$. In addition, the estimated overall volatilities of factors are significantly smaller during the acceptance period.

4.2 Test statistic

Our test statistic is based on a weighted quadratic statistic

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

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 variance-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 facilitate the technical arguments, in this section we assume $\{u_{it}\}$ to be cross-sectionally uncorrelated, and estimate $\mathbf{\Sigma}_u$ by:

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

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.

4.3 Sparse idiosyncratic covariance with heavy-tailed data

It is reasonable to allow cross-sectional dependence by assuming $\mathbf{\Sigma}_u$ to be a sparse covariance, in the sense that most off-diagonal entries of $\mathbf{\Sigma}_u$ are either zero or nearly so. The sparsity condition is a natural extension of the standard setup of approximate factor models (Chamberlain and Rothschild, 1983), and has been recently used in the financial econometrics literature by, e.g., Fan et al. (2015a); Gagliardini et al. (2016). Following the construction of Fan et al. (2015a), we can estimate $\mathbf{\Sigma}_u$ by, for some estimator $\widehat{\sigma}_{ij}$ for $E(u_{it}u_{jt})$,

$$(\widehat{\mathbf{\Sigma}}_u)_{ij} = \begin{cases} \widehat{\sigma}_{ij}, & \text{if } i = j, \\ h_{ij}(\widehat{\sigma}_{ij}), & \text{if } i \neq j. \end{cases}$$

Here $h_{ij}(x) = \text{sgn}(x)(|x| - \tau_{ij})_+$ is taken as a the soft-thresholding function, commonly used in the statistical literature, where $(x)_+ = \max\{x, 0\}$. The threshold value τ_{ij} is chosen to guarantee that

$$\max_{ij} |\widehat{\sigma}_{ij} - E(u_{it}u_{jt})| = O_P(\tau_{ij}). \tag{4.2}$$

The resulting estimator is a sparse covariance matrix, which thresholds off most off-diagonal entries ($(\widehat{\mathbf{\Sigma}}_u)_{ij}$ is zero so long as $|\widehat{\sigma}_{ij}| < \tau_{ij}$). When u_{it} is possibly heavy-tailed, the uniform convergence (4.2) requires a robust variance estimator $\widehat{\sigma}_{ij}$, which can be defined as

$$\widehat{\sigma}_{ij} := \arg\min_{\sigma} \frac{1}{T}\sum_{t=1}^{T}\rho\left(\frac{\widehat{u}_{it}\widehat{u}_{jt} - \sigma}{\alpha_T}\right).$$

While these extensions are straightforward, we expect that the asymptotic analysis might be quite involved, and do not pursue it in this paper.

4.4 Limiting distribution under H_0

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

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

Thus the limiting distribution is determined by that of $\bar{S} := \frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t'\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Sigma}_u\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{u}_t$. Note that the cross-sectional central limit theorem (Assumption 3.5) 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 4.1. *Suppose as* $T, N \to \infty$, $\frac{1}{T}\sum_{t=1}^{T} var(\mathbf{u}_t'\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Sigma}_u\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{u}_t) \to 2K$.

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. There are three technical conditions in this theorem that characterize the relationship among (N, T, J): The first condition (i) $NJ^4 \log N \log J = o(T^{3/2})$ requires that T should be large relative to N. High-dimensional estimation errors accumulate in the test statistic as N increases. Hence this condition controls the error accumulations under a large panel. Condition (ii) $T = o(N^2)$, on the other hand, is commonly required to guarantee the asymptotic accuracy of estimating the unknown factors. Importantly, we allow either $N/T \to \infty$ or $T/N \to \infty$. Finally, condition (iii) $N\sqrt{T} = o(J^{2\eta-1})$ requires the function

$E(\mathbf{f}_t|\mathbf{w}_t = \cdot)$ be sufficiently smooth so that the sieve approximation error is negligible, and does not play a role in the limiting distribution.

5 Application: Multi-index regression

5.1 The model

Consider a multi-index regression model:

$$Y_t = h(\boldsymbol{\psi}_1'\mathbf{z}_t, ..., \boldsymbol{\psi}_L'\mathbf{z}_t) + \varepsilon_t, \quad \mathbf{z}_t = (\mathbf{f}_t', \mathbf{w}_t')', \quad t = 1, ..., T, \tag{5.1}$$

where Y_t represents an observed scalar outcome; $\mathbf{f}_t$ is a set of latent factors that have a predicting power about Y_t; $\mathbf{w}_t$ is a set of observed conditioning variables that might be associated with $\mathbf{f}_t$ and Y_t. For instance, in treatment effect studies, Y_t represents the outcome, and $\mathbf{w}_t$ denotes the treatments and a set of observed demographic variables for the individual t. In macroeconomic forecasts, $Y_t := y_{t+1}$ represents a scalar macroeconomic variable to be forecast. The common factors are often inferred by using a large panel data:

$$\mathbf{x}_t = \boldsymbol{\Lambda}\mathbf{f}_t + \mathbf{u}_t. \tag{5.2}$$

Our goal is to predict/forecast Y_T using the data $\{Y_t, \mathbf{w}_t, \mathbf{x}_t\}_{t=1}^{T-1}$ and $\{\mathbf{w}_T, \mathbf{x}_T\}$. Forecasts based on the estimated factors of large datasets have been extensively studied, where y_{t+1} represents industrial production outputs, excess returns of stocks or U.S. government bonds (Stock and Watson, 2002a,b; Ludvigson and Ng, 2009, 2010), among many others.

Here the non-parametric link function h depends on L indices $\{\boldsymbol{\psi}_l'\mathbf{z}_t\}_{l\leq L}$ with unknown coefficients $\boldsymbol{\psi}_l$'s, and $L < \dim(\mathbf{z}_t)$ is imposed for the dimension reduction. While numerous regression models focus on linear models, some empirical evidence also suggests the possibility of nonlinear dynamics. For instance, nonlinearity is an important part of the theories that attempt to explain the Great Recession with a financial accelerator mechanism (Bemanke et al., 1996). Omitted nonlinearity can lead to biases in predictions. On the other hand, a full nonparametric model may suffer from the curse of dimensionality of $\mathbf{z}_t$. Our goal is to enhance the prediction using improved estimated factors, incorporating the explanatory power of the observed conditioning variables.

5.2 The intuitions of the method

Since the function h is unknown, neither the index parameters nor the unknown factors are separately identifiable. However, we can find a proper transformation

$$\widetilde{\mathbf{z}}_t = \mathbf{M}^{-1}\mathbf{z}_t, \quad \widetilde{\boldsymbol{\psi}}_i = \mathbf{M}'\boldsymbol{\psi}_i, \quad i = 1, ..., L.$$

such that $E\widetilde{\mathbf{z}}_t\widetilde{\mathbf{z}}_t' = \mathbf{I}$ and $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$ is identified, where $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$ denotes the space spanned by the index coefficients. As a result, we consider the following observationally equivalent model:

$$Y_t = h(\widetilde{\boldsymbol{\psi}}_1'\widetilde{\mathbf{z}}_t, ..., \widetilde{\boldsymbol{\psi}}_L'\widetilde{\mathbf{z}}_t) + \varepsilon_t, \tag{5.3}$$

For the ease of reading and to avoid complicated notations in this section, we give the definition of $\mathbf{M}$ in the appendix.

The method we introduce below does not require the identification of the individual coefficients. The estimation procedure is summarized as follows.

Step 1 Estimate $\widetilde{\mathbf{z}}_t$ by $\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t$, where

$$\widehat{\mathbf{z}}_t = (\widehat{\mathbf{f}}_t', \mathbf{w}_t')', \quad \widehat{\mathbf{M}} = (\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{z}}_t\widehat{\mathbf{z}}_t')^{1/2}. \tag{5.4}$$

Here $\widehat{\mathbf{f}}_t$ is the proposed *robust proxy-regressed* factor estimator, with $\mathbf{w}_t$ incorporated in the estimation procedure.

Step 2 Find $\{\widehat{\boldsymbol{\psi}}_1, ..., \widehat{\boldsymbol{\psi}}_L\}$ so that $\text{span}\{\widehat{\boldsymbol{\psi}}_1, ..., \widehat{\boldsymbol{\psi}}_L\}$ consistently estimates $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$. Then our estimated indices are written as $\widehat{\boldsymbol{\psi}}_i'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t$, $i = 1, ..., L$.

Step 3 Finally, obtain a nonparametric estimator $\widehat{h}$ using any smoothing technique by regressing Y_t onto the estimated indices $\widehat{\boldsymbol{\psi}}_i'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t$, for $t = 1, ..., T-1$. We then predict Y_T by

$$\widehat{Y}_T := \widehat{h}(\widehat{\boldsymbol{\psi}}_1'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_T, ..., \widehat{\boldsymbol{\psi}}_L'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_T).$$

In particular, when $Y_t = y_{t+1}$ in the forecast context, we forecast y_{T+1} by $\widehat{y}_{T+1|T} := \widehat{Y}_T$.

A standard procedure of estimating the common factors is to apply the principal components (PC) estimator on (5.2). In contrast, we employ the proposed factor estimators. Our estimator potentially have two advantages compared to the PC estimator: (i) macroeconomic variables and financial excess returns are heavy-tailed. We shall also demonstrate it

in our empirical study. (ii) The conditioning variables $\mathbf{w}_t$ can have strong explanatory powers on the factors. Indeed, numerous empirical studies of the macroeconomic dataset used in Stock and Watson (2002a) and Ludvigson and Ng (2009) have identified several factors related to housing variables and stock markets. These factors are believed to be strongly associated with, e.g., the consumption-wealth variable, financial ratios (ratios of price to dividends or earnings), and term spread (e.g., Lettau and Ludvigson (2010); Campbell and Shiller (1988); Fama and French (1988); Campbell (1991)). By incorporating these variables as $\mathbf{w}_t$, the estimation of $\mathbf{f}_t$ can be improved, which can potentially lead to significantly better predictions.

In step 2 we shall directly estimate $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$. In the statistical literature on dimension reductions (Li, 1991; Cook and Lee, 1999), this space is also called *dimension-reduction subspace*. We now explain the rationale of using this space for the multi-index regression. We assume that the indices are sufficient for Y_t in the sense that the conditional distribution of $Y_t|\widetilde{\mathbf{z}}_t$ satisfies:

$$Y_t|\widetilde{\mathbf{z}}_t =^d Y_t|(\widetilde{\boldsymbol{\psi}}_1'\widetilde{\mathbf{z}}_t, ..., \widetilde{\boldsymbol{\psi}}_L'\widetilde{\mathbf{z}}_t) \tag{5.5}$$

for all values of $\widetilde{\mathbf{z}}_t$ in its marginal sample space. As is shown by Cook and Lee (1999), (5.5) still holds when $(\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L)$ is replaced with any $(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_L)$ such that

$$\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\} = \text{span}\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_L\}.$$

In other words, all the information of $Y_t|\widetilde{\mathbf{z}}_t$ is preserved by using $(\boldsymbol{\xi}_1'\widetilde{\mathbf{z}}_t, ..., \boldsymbol{\xi}_L'\widetilde{\mathbf{z}}_t)$. This then implies,

$$h(\widetilde{\boldsymbol{\psi}}_1'\widetilde{\mathbf{z}}_t, ..., \widetilde{\boldsymbol{\psi}}_L'\widetilde{\mathbf{z}}_t) = E(Y_t|\widetilde{\mathbf{z}}_t) = E(Y_t|\boldsymbol{\xi}_1'\widetilde{\mathbf{z}}_t, ..., \boldsymbol{\xi}_L'\widetilde{\mathbf{z}}_t). \tag{5.6}$$

We shall propose $(\widehat{\boldsymbol{\psi}}_1, ..., \widehat{\boldsymbol{\psi}}_L)$ so that $\widehat{\boldsymbol{\psi}}_i'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t \to^P \boldsymbol{\xi}_i'\widetilde{\mathbf{z}}_t$ for a particular set of $\{\boldsymbol{\xi}_i\}$. Hence

$$\widehat{h}(\widehat{\boldsymbol{\psi}}_1'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t, ..., \widehat{\boldsymbol{\psi}}_L'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t) \to^P h(\widetilde{\boldsymbol{\psi}}_1'\widetilde{\mathbf{z}}_t, ..., \widetilde{\boldsymbol{\psi}}_L'\widetilde{\mathbf{z}}_t).$$

5.3 Estimating the space spanned by index coefficients

We now describe the identification and estimation of $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$, which is based on the *sliced inverse regression* method of Li (1991). Fix $H \geq \max\{L, 2\}$, and divide the range of data $\{Y_1, ..., Y_{T-1}\}$ into H disjoint "slices" $I_1, ..., I_H$ such that $P(Y_t \in I_h) = 1/H$.

Specifically, let F_y denote the cumulative distribution function (CDF) of Y_t, and let

$$I_h = [F_y^{-1}((h-1)/H), F_y^{-1}(h/H)].$$

In practice, we replace F_y with the sample CDF to construct I_h. Now define a "sliced covariance matrix"

$$\mathbf{\Sigma}_{z|y} := \frac{1}{H}\sum_{h=1}^{H} E(\widetilde{\mathbf{z}}_t|Y_t \in I_h)E(\widetilde{\mathbf{z}}_t|Y_t \in I_h)'.$$

It follows from Li (1991) that $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$ can be identified as the space spanned by the eigenvectors $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_L$ of $\mathbf{\Sigma}_{z|y}$, corresponding to the first L eigenvalues [†]:

$$\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\} = \text{span}\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_L\}.$$

We shall present this proposition and its proof in Appendix F.

Importantly, $\mathbf{\Sigma}_{z|y}$ is easy to estimate using its sample counterpart, whose leading eigenvectors span a subspace that consistently estimates $\text{span}\{\widetilde{\boldsymbol{\psi}}_1, ..., \widetilde{\boldsymbol{\psi}}_L\}$. Let $\widehat{E}(\widetilde{\mathbf{z}}_t|Y_t \in I_h)$ be the sample analogue of $E(\widetilde{\mathbf{z}}_t|Y_t \in I_h)$, which is the sample average of $\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t$ for all $Y_t \in I_h$, for $t = 1, ..., T-1$:

$$\widehat{E}(\widetilde{\mathbf{z}}_t|Y_t \in I_h) := \frac{\widehat{\mathbf{M}}^{-1}\sum_{t=1}^{T-1}\widehat{\mathbf{z}}_t 1\{Y_t \in I_h\}}{\sum_{t=1}^{T-1} 1\{Y_t \in I_h\}},$$

where $\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t$ is as defined in (5.4). We then let $\{\widehat{\boldsymbol{\psi}}_1, ..., \widehat{\boldsymbol{\psi}}_L\}$ be the eigenvectors of the first L eigenvalues of

$$\widehat{\mathbf{\Sigma}}_{z|y} = \frac{1}{H}\sum_{h=1}^{H} \widehat{E}(\widetilde{\mathbf{z}}_t|Y_t \in I_h)\widehat{E}(\widetilde{\mathbf{z}}_t|Y_t \in I_h)'.$$

It is shown by Li (1991) and many authors in the statistical literature that the eigenvectors of $\mathbf{\Sigma}_{z|y}$ is non-sensitive to H. Indeed, the choice of H has very little impact on the estimated eigenvector space. This was elucidated by Fan et al. (2015). The following result presents the estimation consistency of the index space. Similar results of this type was recently obtained by Fan et al. (2015). Theorem 5.1 in contrast, provides a robust estimator for the factors that takes advantages of the information contained in the observables $\mathbf{w}_t$. In addition, our procedure is suitable for possibly heavy-tailed economic data. As recently noted by Bai and Liao (2016), more accurate estimation of the common factors can lead to better out-of-sample forecast performances.

[†]The formal statement of this proposition is given in Appendix F.1.

Define

$$b_{NT} = \sqrt{\frac{J\|\operatorname{cov}(\boldsymbol{\gamma}_t)\| + \log N}{T}} + \frac{1}{\sqrt{N}} + \frac{1}{J^{\eta-1/2}}$$

Theorem 5.1. *For any given $H \geq L$, if the largest L eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{z}|y}$ are distinct, then*

$$\max_{i\leq L} \|\widehat{\boldsymbol{\psi}}_i - \boldsymbol{\xi}_i\| = O_P(b_{NT}).$$

In addition, for each fixed t, $\widehat{\mathbf{M}}^{-1}\mathbf{z}_t - \widetilde{\mathbf{z}}_t = O_P(b_{NT})$, and

$$\frac{1}{T}\sum_t \|\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t - \widetilde{\mathbf{z}}_t\|^2 = O_P(b_{NT}^2).$$

Theorem 5.1 immediately implies $\widehat{\boldsymbol{\psi}}_i'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t$ consistently estimates $\boldsymbol{\xi}_i'\widetilde{\mathbf{z}}_t$ for each fixed t. Then by regressing Y_t onto $\{\widehat{\boldsymbol{\psi}}_i'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_t\}_{i\leq L}$, the function $\widehat{h}(\widehat{\boldsymbol{\psi}}_1'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_T, ..., \widehat{\boldsymbol{\psi}}_L'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{z}}_T)$ consistently estimates $E(Y_T|\widetilde{\boldsymbol{\psi}}_1'\widetilde{\mathbf{z}}_T, ..., \widetilde{\boldsymbol{\psi}}_L'\widetilde{\mathbf{z}}_T) = E(Y_T|\mathbf{f}_T, \mathbf{w}_T)$, whenever $\widehat{h}$ is consistent. The latter is a conditional mean predictor for Y_T.

6 Simulation Studies

6.1 Model settings

In this section, we use simulated examples to demonstrate the finite sample performance of the proposed *robust proxy-regressed* method and compare it with existing ones. Throughout this section, we respectively specify five factors ($K = 5$) and five characteristics, and the sample size is $N = 50,\ T = 100$. More results based on different sample sizes ($N = 100,\ T = 50$ and $N = 100,\ T = 100$) are presented in Appendix B, and the results are very similar.

The sieve basis is chosen as the additive Fourier basis with $J = 5$. As discussed in Section 3.2, 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,

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\Lambda}\mathbf{f}_t + \mathbf{u}_t, \\ \mathbf{f}_t &= \mathbf{g}(\mathbf{w}_t) + \boldsymbol{\gamma}_t, \quad t = 1, \cdots, T, \end{aligned} \tag{6.1}$$

where $\sigma = \sqrt{\operatorname{var}(\boldsymbol{\gamma}_t)}$ is set to be 0.01, 0.3 and 1. The smaller the σ is, the more $\mathbf{f}_t$ and

$\mathbf{g}(\mathbf{w}_t)$ are correlated. In this study, $\mathbf{\Lambda}$, $\mathbf{w}_t$ and $\boldsymbol{\gamma}_t$ 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) $\mathbf{g}(\mathbf{w}_t) = \mathbf{D}\mathbf{w}_t$, where $\mathbf{D}$ is a $K \times K$ matrix with each entry drawn from $U[1,\ 2]$;

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

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

In addition, $\mathbf{u}_t$ is drawn from one of the following four distributions:

(1) Normal distribution ($N(0, 8)$);

(2) Mixture Normal distribution (MixN) $0.5*N(-1, 4)+0.5*N(8, 1)$, which is asymmetric and light tailed;

(3) Two times the $t-$distribution with degrees of freedom 3 ($2t_3$), which is symmetric and heavy-tailed;

(4) Log-normal distribution (LogN) 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. For Model (III), we only consider $\sigma = 1$, as the observables $\mathbf{w}_t$ are independent of latent factors and other choices would have yield similar results.

In the presentation below, we shall abbreviate the proposed *robust proxy-regressed* method as RPR. For comparisons, we estimate the factors and loadings in (6.1) by the proposed RPR, Sieve-LS (Section 2.3.1), the regular PCA and INT under different scenarios. In particular, the INT method estimates the factors and loadings from the panel data with interactive effects (INT) model (Bai, 2009) as follows

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{g}(\mathbf{w}_t) + \mathbf{\Lambda}\boldsymbol{\gamma}_t + \mathbf{u}_t, \quad t = 1,\ \cdots,T,$$

by solving the least squares problem (2.5).

6.2 In-sample Estimation

First, we compare the in-sample model fitting performance among RPR, Sieve-LS, PCA and INT under different scenarios. Let $F = (f_1, \cdots, f_T)'$ be the $T \times K$ matrix of factors. We use PCA as a benchmark and define the relative estimation error as

$$\frac{\|\widehat{\Lambda}\widehat{F}' - \Lambda F'\|_F^2}{\|\tilde{\Lambda}\tilde{F}' - \Lambda F'\|_F^2},$$

where $\tilde{\Lambda}$ and $\tilde{F}$ are the estimators of Λ and F obtained by PCA, $\widehat{\Lambda}$ and $\widehat{F}$ are the estimators of Λ and F obtained by one of the methods to be compared. For each scenario, we conduct 200 simulations and calculate the average of relative estimation error. Results are presented in Table 1. Besides, one may be also interested in the estimation accuracy of factors or loadings alone rather than their products. 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 and Liao, 2016). For Model (I) and (II) we report the sample mean of the median of 5 canonical correlations between the true loading matrix and estimated one and the true factors and estimated ones(Bai (2003)). The results are presented in Table 2 and 3.

According to Tables 1– 3, Sieve-LS and RPR are comparable for light-tail distributions. This implies that we do not pay much the 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 $\mathbf{w}_t$ and $\mathbf{f}_t$ are well correlated. In general, PCA gives the worst estimation performance as it does not exploit the information in $\mathbf{w}_t$. When $\sigma = 1$, the observabed $\mathbf{w}_t$ is not as informative and hence the performance of RPR and Sieve-LS deteriorates. The interactive-effect based method (INT) has worse estimation performance then Sieve-LS under light tailed scenarios and performs similarly to PCA under heavy-tailed cases.

6.3 Out-of-sample Forecast

Consider y_{t+1} as a linear function of $\mathbf{f}_t$:

$$y_{t+1} = \boldsymbol{\beta}'\mathbf{f}_t + \epsilon_t,$$

Table 1: Mean relative estimation error of $\Lambda F'$ (%) 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	INT	RPR	Sieve-LS	INT	RPR	Sieve-LS	INT
	0.01	0.76	0.75	0.85	0.78	0.78	0.84			
$N(0,\ 8)$	0.3	0.83	0.82	0.91	0.85	0.85	0.90			
	1.0	1.59	1.58	1.92	1.37	1.36	1.77	2.29	2.28	2.44
	0.01	0.78	0.78	0.86	0.86	0.86	0.90			
MixN	0.3	0.91	0.90	0.95	0.92	0.92	0.94			
	1.0	1.72	1.70	2.13	1.44	1.43	1.64	2.67	2.64	2.75
	0.01	0.62	0.84	0.94	0.56	0.85	0.95			
$2t_3$	0.3	0.63	0.85	0.96	0.57	0.85	0.95			
	1.0	0.64	0.86	0.99	0.58	0.86	1.00	1.18	1.18	1.20
	0.01	0.66	0.81	0.93	0.64	0.83	0.94			
LogN	0.3	0.66	0.82	0.96	0.65	0.84	0.95			
	1.0	0.67	0.83	0.99	0.65	0.84	0.97	1.16	1.16	1.23

Table 2: Median of 5 canonical correlations of estimated loading matrix and true one when $N = 50, T = 100$: the larger the better

		Model (I)				Model (II)			
$\mathbf{u}_t$	σ	RPR	Sieve-LS	PCA	INT	RPR	Sieve-LS	PCA	INT
	0.01	0.93	0.93	0.85	0.90	0.91	0.91	0.85	0.90
$N(0,\ 8)$	0.3	0.91	0.91	0.90	0.88	0.87	0.87	0.87	0.83
	1.0	0.90	0.90	0.97	0.85	0.86	0.86	0.95	0.82
	0.01	0.96	0.96	0.92	0.94	0.94	0.94	0.91	0.92
MixN	0.3	0.94	0.94	0.93	0.93	0.91	0.91	0.91	0.90
	1.0	0.93	0.93	0.98	0.89	0.91	0.91	0.96	0.88
	0.01	0.57	0.37	0.29	0.34	0.58	0.36	0.27	0.33
$2t_3$	0.3	0.55	0.35	0.30	0.32	0.56	0.35	0.28	0.32
	1.0	0.54	0.34	0.32	0.32	0.53	0.33	0.31	0.31
	0.01	0.68	0.33	0.26	0.30	0.67	0.34	0.25	0.31
LogN	0.3	0.66	0.31	0.26	0.27	0.65	0.33	0.26	0.30
	1.0	0.63	0.30	0.28	0.28	0.62	0.29	0.27	0.28

Both RPR and Sieve-LS are the proposed characteristic-based methods. RPR uses robust estimator for $\mathbf{\Sigma}$ while Sieve-LS uses non-robust least squares covariance estimator.

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, PCA or INT. In each simulation, we

Table 3: Median of 5 canonical correlations between estimated factors and true ones when $N = 50, T = 100$: the larger the better

		Model (I)				Model (II)			
$\mathbf{u}_t$	σ	RPR	Sieve-LS	PCA	INT	RPR	Sieve-LS	PCA	INT
	0.01	0.94	0.94	0.77	0.84	0.95	0.95	0.85	0.92
$N(0,\ 8)$	0.3	0.86	0.86	0.83	0.82	0.89	0.89	0.87	0.88
	1.0	0.85	0.85	0.95	0.81	0.88	0.88	0.95	0.86
	0.01	0.96	0.96	0.86	0.92	0.97	0.97	0.97	0.97
MixN	0.3	0.91	0.91	0.89	0.90	0.93	0.93	0.92	0.92
	1.0	0.89	0.90	0.96	0.86	0.92	0.92	0.96	0.91
	0.01	0.68	0.43	0.27	0.37	0.64	0.40	0.27	0.36
$2t_3$	0.3	0.66	0.40	0.29	0.36	0.61	0.37	0.27	0.33
	1.0	0.63	0.37	0.33	0.34	0.58	0.34	0.29	0.30
	0.01	0.66	0.43	0.30	0.40	0.65	0.38	0.27	0.34
LogN	0.3	0.64	0.41	0.33	0.38	0.60	0.36	0.29	0.32
	1.0	0.60	0.37	0.36	0.36	0.57	0.34	0.31	0.31

generate $T + 50$ observations in total. For $s = 1, \cdots, 50$, we use the T observations right before time $T + s$ to forecast y_{T+s}. We use PCA as a 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 RPR, Sieve-LS or INT while $\widetilde{y}^{PCA}_{T+s|T+s-1}$ is the forecast based on PCA. For RPR and PCA, they are both based on model (5.3) except the factors there are estimated by two different method. For the INT method, the factors are estimated by using RPR. For each scenario, we conduct 200 simulations and calculate the averaged RMSE as a measurement of the one-step ahead out-of-sample forecast performance.

The results are presented in Table 4. 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 $\mathbf{w}_t$ and $\mathbf{f}_t$ is strong. The INT model has worse forecast performance than Sieve-LS. 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 $\mathbf{w}_t$ has explanatory powers about the factors. In more general scenarios, the RPR is more robust to the tail distribution and does not pay much the price even for the light tailed distributions.

Table 4: Mean relative mean squared error 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	INT	RPR	Sieve-LS	INT	RPR	Sieve-LS	INT
	0.01	0.89	0.89	0.95	0.91	0.90	0.96			
$N(0,\ 8)$	0.3	0.94	0.93	0.98	0.93	0.93	1.00			
	1.0	1.04	1.03	1.06	1.04	1.03	1.10	1.45	1.44	1.42
	0.01	0.74	0.74	0.90	0.79	0.78	0.92			
MixN	0.3	0.83	0.81	0.95	0.86	0.85	0.97			
	1.0	1.10	1.07	1.12	1.11	1.07	1.15	1.55	1.52	1.50
	0.01	0.18	0.31	0.53	0.44	0.63	0.75			
$2t_3$	0.3	0.18	0.32	0.60	0.44	0.64	0.82			
	1.0	0.20	0.37	0.74	0.46	0.65	0.89	1.28	1.26	1.29
	0.01	0.51	0.57	0.70	0.58	0.70	0.81			
LogN	0.3	0.50	0.57	0.75	0.60	0.72	0.86			
	1.0	0.48	0.59	0.79	0.62	0.75	0.91	1.19	1.18	1.20

7 Empirical Study of US Bond Risk Premia

7.1 Econometric Motivation

In this section, we study 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. Then, we denote the one year excess return with maturity of n years in period $t+1$ as

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

For a long time, the literature has found a significant predictive variation of the excess returns of U.S. government bonds. Theoretical and empirical researches show the linkage between the excess bond returns and a few macroeconomic variables, see Fama and Bliss

(1987); Campbell and Cochrane (1999); Brandt and Wang (2003); Wachter (2006); Campbell and Shiller (1991); Piazzesi and Swanson (2008), among others. Recently, Ludvigson and Ng (2009, 2010) use diffusion index model (Stock and Watson, 2002b) to predict the bond risk premia with observable variables plus a few common factors estimated from a large macroeconomic panel data. Their study showed a best subset of the estimated factors together with the single index in Cochrane and Piazzaesi (2005) can explain about 21% of one year excess bond return with maturity of two years, in terms of out-of-sample multiple R^2. By incorporating the observed characteristics in estimating the factors, our method achieve 38.1% out-of-sample R^2 using linear forecast model, and 44.8% using the nonlinear multi-index forecast model.

This empirical application develops a new way of incorporating the explanatory power of the observed characteristics, and investigates the robustness of the conclusions. First, we find that the observed characteristics have a strong explanatory power of the factors. Incorporating them in factor estimation leads to a significantly better forecast rather than using them in forecast directly. Second, the factors are robustly estimated by the proposed method as we find many series in the macroeconomic panel dataset are heavy-tailed. Finally, our forecast is based on the multi-index regression introduced in Section 5. This method not only allows a nonlinear modeling but also serves as a further dimension reduction tool.

We analyze monthly data spanned from January 1964 to December 2003. The excess bond returns are calculated based on the one- through five-year zero coupon U.S. Treasure bond prices (Cochrane and Piazzaesi, 2005), which is available from the Center for Research in Securities Prices (CRSP). The factors are estimated from a macroeconomic dataset consisting of 131 series. A detailed description and transformation code of this panel data can be found in the Appendix A of Ludvigson and Ng (2010). The observed characteristics $\mathbf{w}_t$ are chosen to be the single index in Cochrane and Piazzaesi (2005) and seven aggregate macroeconomic series that calculated from particular sub-panels, see Table 5 for detailed description. These aggregate series are widely used to describe the co-movement of the macroeconomic activities, e.g. (NBER, 2008; Stock and Watson, 2010).

Throughout this study, the sieve basis of $\mathbf{w}_t$ 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.

7.2 Heavy-tailed data and robust estimations

We studied the excess kurtosis for the time series to assess the tail distributions. Figure 1 shows 43 among the 131 series have an excess kurtosis greater than 6. This indicates the tails of their distributions are fatter than the t-distribution with degrees of freedom 5. Figure 2 plots the estimated idiosyncratic error $\widehat{\mathbf{u}}_t = \mathbf{x}_t - \widehat{E}(\mathbf{x}_t|\mathbf{w}_t)$, which preserves the heavy-tailed behavior. On the other hand, Figure 2 reports the histograms of excess kurtosis of the "fitted data" $\widehat{E}(\mathbf{x}_t|\mathbf{w}_t)$ (the robust estimator of $E(\mathbf{x}_t|\mathbf{w}_t)$ using Huber loss, which demonstrates that most series in the fitted data are no longer severely heavy-tailed. In comparison with the raw data (Figure 1), the excess kurtosis of the fitted variables dramatically decreases.

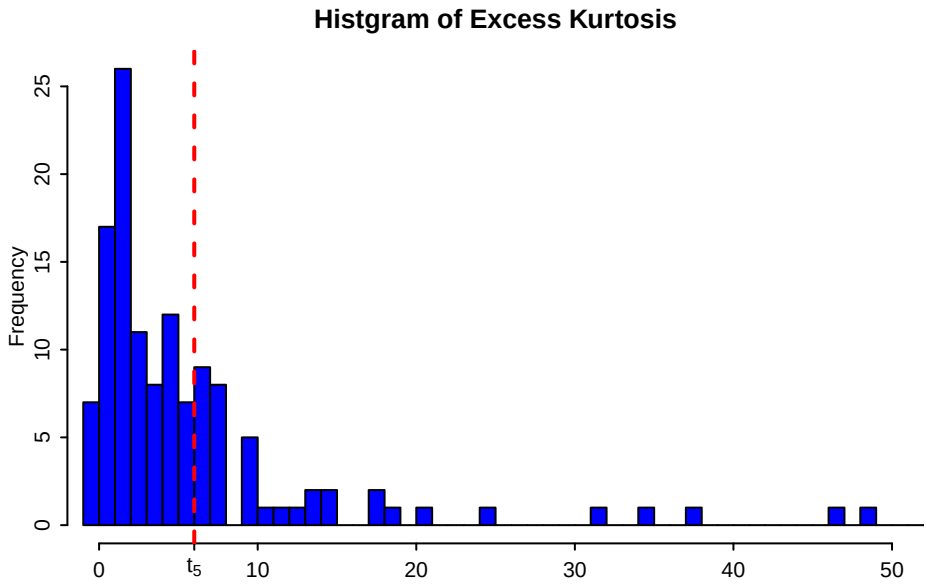

Figure 1: Excess kurtosis of the macroeconomic panel data

7.3 Forecast with factors

We first study the forecast power of the estimated factors. We apply the one-month ahead out-of-sample forecast of the bond risk premia with maturity of two to five years. The forecast uses the information in the past 240 months, starting from January 1984 and rolling forward to December 2003. The number of factors is determined by the information criteria developed in Bai and Ng (2002), which suggests eight common factors. We consider and

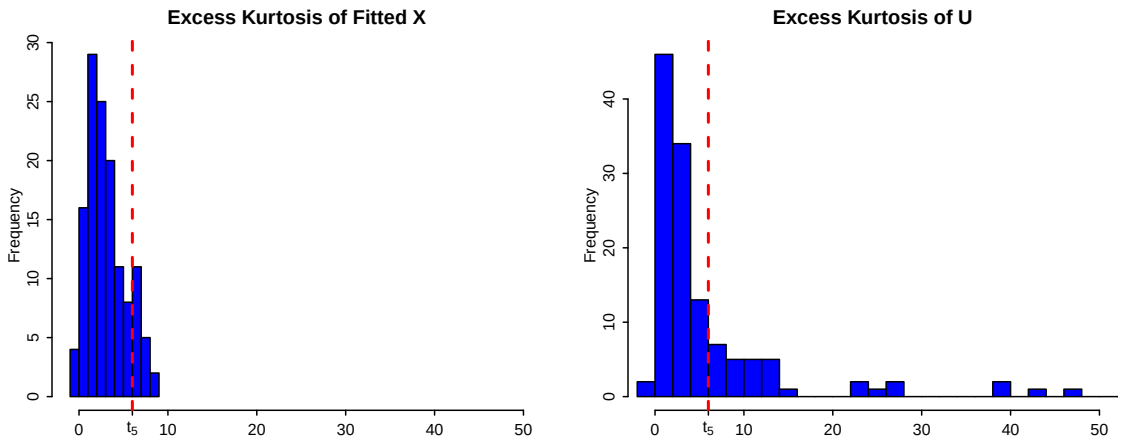

Figure 2: Excess kurtosis of the fitted data and idiosyncratic error

The fitted data $\widehat{\mathbf{x}}_t$ after regressing on $\mathbf{w}_t$ are no longer severely heavy-tailed. The estimated idiosyncratic errors preserve the heavy-tailed behavior with 36 series have tails fatter than t-distribution with degrees of freedom 5 and the largest excess kurtosis is greater than 160.

Table 5: Components of $\mathbf{w}_t$

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

compare four approaches to estimating the factors: RPR, Sieve Least Squares (Sieve-LS), PCA and PCA2 based on an enlarged panel data that includes both 131 series and $\mathbf{w}_t$ (PCA2). Also we consider two forecast models as follows:

$$\text{Linear model:} \qquad y_{t+1} = \alpha + \boldsymbol{\beta}'\mathbf{f}_t + \epsilon_t, \tag{7.1}$$

$$\text{Multi-index model:} \qquad y_{t+1} = \alpha + h(\boldsymbol{\psi}_1'\mathbf{f}_t, \ \cdots, \ \boldsymbol{\psi}_L'\mathbf{f}_t) + \epsilon_t, \tag{7.2}$$

where α is the intercept and h is a nonparametric function. The number of sufficient indices L is estimated by the ratio-based method used in Lam and Yao (2012) and Fan et al. (2015) and should be no larger than the number of factors. The estimated L is usually 2 or 3. Given the estimated indices, (7.2) is fitted by the "gam" package in R which approximates h by an additive function, resulting in the projection pursuit model $h(\boldsymbol{\psi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\psi}_L'\mathbf{f}_t) = \sum_{l=1}^{L} g_l(\boldsymbol{\psi}_l'\mathbf{f}_t)$. Each individual nonparametric function $g_l(\cdot)$ is smoothed by the local linear approximation. We use the default kernel function in the package and the bandwidth is selected by the cross validation. The parametric part (weights and intercept) is fitted by a backfitting method which iteratively minimizes the partial residuals.

Let $\widehat{y}_{T+t+1|T+t}$ be the forecast 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 all scenarios are reported in Table 6. Furthermore, we present the plots of forecast results of RPR and PCA based on the linear model and the multi-index model in Appendix B.

Table 6: Forecast performance in out-of-sample R^2 (%): the larger the better

Maturity	Linear model				Multi-index model			
	RPR	Sieve-LS	PCA	PCA2	RPR	Sieve-LS	PCA	PCA2
2 Year	38.1	37.4	32.6	34.2	44.8	41.2	34.5	40.3
3 Year	32.9	32.4	28.2	28.5	43.2	39.1	32.1	37.9
4 Year	25.7	25.4	23.3	23.9	38.9	35.2	27.3	34.6
5 Year	23.0	22.6	19.7	19.8	37.6	34.1	23.7	31.9

Our results justify the conclusions in Ludvigson and Ng (2009) by showing the bond risk premia is forecastable by the common factors of macroeconomic data. The multi-index model with factors estimated by the Robust Projected PCA has an 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).

From Table 6, we notice the factors estimated by RPR and Sieve-LS can explain more variation in bond risk premia with all maturities than the ones estimated by PCA. We

interpret the observed results from this table in the following aspects:

1. Both RPR and Sieve-LS outperform PCA and PCA2.

   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 characteristics, the proposed methods (RPR and Sieve-LS) can improve the estimation of factors even if N is relatively mild.

2. RPR outperforms Sieve-LS.

   Many series in this panel data are heavy-tailed. The RPR method can robustly estimate the factors and result in better forecasts.

3. PCA2 slightly outperforms PCA.

   Furthermore, the factors estimated by PCA2 has better forecast performance than the ones estimated by PCA as the former method includes a larger panel.

4. Multi-index models outperform linear models

   According to Table 6, the proposed multi-index model always has higher R^2 than linear model with the same estimated factors.

7.4 Forecast using $\mathbf{w}_t$

Previously, the observed characteristics $\mathbf{w}_t$ were only used in the proposed estimators for the common factors, and were not used directly in the forecasting model (7.1) or (7.2). Now we study the effect of $\mathbf{w}_t$ in forecasting bond risk premia and consider the following models

$$\text{Factor-augmented linear model:} \qquad y_{t+1} = \alpha + \boldsymbol{\beta}'\mathbf{z}_t + \epsilon_t, \tag{7.3}$$

$$\text{Factor-augmented multi-index model:} \qquad y_{t+1} = \alpha + h(\boldsymbol{\psi}_1'\mathbf{z}_t,\ \cdots,\ \boldsymbol{\psi}_L'\mathbf{z}_t) + \epsilon_t, \tag{7.4}$$

where $\mathbf{z}_t$ is one of the following three forms: (i) $\mathbf{w}_t$; (ii) $(\mathbf{f}_t', \mathbf{w}_t')'$; (iii) $(\mathbf{f}_t', w_{i,t})'$, $i = 1, \cdots, 8$. The data-driven method for choosing L often picks $\widehat{L} = 5$ when $\mathbf{z}_t = (\mathbf{f}_t', \mathbf{w}_t')'$ and $\widehat{L} = 3$ when $\mathbf{z}_t$ takes the other two forms.

The selected forecast results are reported in Tables 7 and 8 respectively. The full results are deferred to Appendix B. The results show the forecast based on the factor-augmented

multi-index model performs better than the factor-augmented linear model. While comparing with model (7.1) or (7.2), the stories are similar.

Table 7: Forecast out-of-sample R^2 (%) for factor-augmented linear model: the larger the better. The bold figures represent larger R^2 than forecast with factors alone under the same scenario.

$\mathbf{z}_t$	RPR				Sieve-LS				PCA			
	Maturity(Year)				Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5	2	3	4	5
$(\mathbf{f}_t', \mathbf{w}_t')'$	37.9	32.6	25.6	22.8	37.1	31.9	25.3	22.1	23.9	21.4	17.4	17.5
$(\mathbf{f}_t', w_{1,t})'$	38.0	32.8	25.5	22.9	37.1	29.2	24.3	21.9	**33.3**	27.9	22.9	19.4
$(\mathbf{f}_t', w_{4,t})'$	38.1	32.8	25.7	22.9	36.1	29.1	24.7	22.3	**33.6**	27.7	22.3	**20.0**
$(\mathbf{f}_t', w_{8,t})'$	38.1	32.9	25.6	22.9	37.3	32.4	25.3	22.5	**34.8**	**32.0**	**27.1**	**24.3**
$\mathbf{w}_t$	6.1	5.5	4.7	4.5	NA				NA			

Table 8: Forecast out-of-sample R^2 (%) for factor-augmented multi-index model: the larger the better. The bold figures represent larger R^2 than forecast with factors alone under the same scenario.

$\mathbf{z}_t$	RPR				Sieve-LS				PCA			
	Maturity(Year)				Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5	2	3	4	5
$(\mathbf{f}_t', \mathbf{w}_t')'$	41.7	39.0	35.6	34.1	41.1	35.7	32.2	30.0	30.8	26.3	24.6	22.0
$(\mathbf{f}_t', w_{1,t})'$	43.4	38.2	34.5	30.9	39.5	37.3	32.2	28.8	**39.4**	**36.9**	**31.7**	**28.5**
$(\mathbf{f}_t', w_{4,t})'$	41.5	39.8	35.4	33.2	38.3	35.6	32.0	29.1	**36.2**	**34.4**	**30.7**	**28.2**
$(\mathbf{f}_t', w_{8,t})'$	41.1	38.9	34.6	30.2	39.0	36.3	31.6	26.8	**35.0**	**33.2**	**28.6**	**24.2**
$\mathbf{w}_t$	13.6	10.8	10.0	6.8	NA				NA			

For the factor-augmented multi-index model, $\mathbf{w}_t$ itself gives an out-of-sample R^2 of 13.6% in predicting the bond risk premia with two year maturity and its forecast power decreases as the maturity increases. The factors and $\mathbf{w}_t$ together ($\mathbf{z}_t = (\mathbf{f}_t', \mathbf{w}_t')'$) have slightly worse forecast performance than the factors alone.

Adding each covariate in $\mathbf{w}_t$ to augmenting the prediction leads to some interesting findings. The forecast performance based on PCA method gets sizable improvement when adding $w_{1,t}$ (Linear combination of five forward rates), $w_{4,t}$ (Non-agriculture employment) or $w_{8,t}$ (CPI). This coincides with the findings in existing literature that forward rates, employment and inflation have predictive power in bond risk premia (Cochrane and Piazzaesi,

2005; Campbell and Cochrane, 1999; Wachter, 2006; Brandt and Wang, 2003). However, the forecast performance based on either RPR or Sieve-LS cannot be improved by adding any covariate in $\mathbf{w}_t$. We argue that, in this application, the information of $\mathbf{w}_t$ should be mainly used as the explanatory power for the factors. And our proposed (RPR and Sieve-LS) have efficiently exploited this information.

Therefore, we conclude:

1. The observed macroeconomic characteristics $\mathbf{w}_t$ (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 characteristics are incorporated to estimate the common factors (using the proposed procedure) than directly used for forecasts.

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

3. The factors estimated by RPR lead to significantly improved out-of-sample forecast on the US bond risk premia compared to the ones estimated by PCA.

4. As many series in the panel data are heavy-tailed, the proposed method can robustly estimate the factors and result in improved out-of-sample forecasts.

8 Conclusions

We provide an econometric analysis for the factor models when the factors depend on several 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. 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. To incorporate the explanatory power of these observed characteristics, we propose a new 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 macroeconomic time series. 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 macroeconomic characteristics contain strong explanatory

powers of the factors. The gain of forecast is more substantial when these characteristics are incorporated to estimate the common factors than directly used for forecasts.

References

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.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843.

Andrews, B., Davis, R. A. and Breidt, F. J. (2007). Rank-based estimation for all-pass time series models. *Annals of Statistics* 844–869.

Atkinson, A. C., Koopman, S. J. and Shephard, N. (1997). Detecting shocks: outliers and breaks in time series. *Journal of Econometrics* **80** 387–422.

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

Bai, J. and Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics* **191** 1–18.

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

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

Balke, N. S. and Fomby, T. B. (1994). Large shocks, small shocks, and economic fluctuations: Outliers in macroeconomic time series. *Journal of Applied Econometrics* **9** 181–200.

Bemanke, B., Gertler, M. and Gilchrist, S. (1996). The financial accelerator and the flight to quality. *The Review of Economics and Statistics* **78** 1–15.

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.

Brandt, M. W. and Wang, K. Q. (2003). Time-varying risk aversion and unexpected inflation. *Journal of Monetary Economics* **50** 1457–1498.

Brillinger, D. (1981). *Time series: data analysis and theory.* vol 36, Siam.

Campbell, J. (1991). A variance decomposition for stock returns. *Economic Journal* **101** 157–79.

Campbell, J. and Shiller, R. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* **1** 195–228.

Campbell, J. Y. and Cochrane, J. H. (1999). Explaining the poor performance of consumption-based asset pricing models. Tech. rep., National bureau of economic research.

Campbell, J. Y. and Shiller, R. J. (1991). Yield spreads and interest rate movements: A bird's eye view. *The Review of Economic Studies* **58** 495–514.

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.

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

Cochrane, J. H. and Piazzaesi, M. (2005). Bond risk premia. *The American Economic Review* **95** 138–160.

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.

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

Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association* **94** 1187–1200.

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 Bliss, R. R. (1987). The information in long-maturity forward rates. *The American Economic Review* **77** 680–692.

Fama, E. F. and French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics* **22** 3–25.

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., 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.

Fan, J., Xue, L. and Yao, J. (2015). Sufficient forecasting using factor models. *arXiv preprint arXiv:1505.07414* .

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., 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.

Gagliardini, P., Ossola, E. and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity datasets. *Econometrica*. Forthcoming.

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.

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.

Hill, J. B. (2015). Robust estimation and inference for heavy-tailed garch. *Bernoulli* **21** 1629–1669.

Huber, P. (1964). Robust estimation of a location parameter. *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.

Lettau, M. and Ludvigson, S. (2010). Measuring and modeling variation in the risk-return trade-off. *Handbook of Financial Econometrics* **1** 617–690.

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.

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

NBER (2008). Determination of the december 2007 peak in economic activity. Tech. rep., National bureau of economic research: Business Cycle Dating Committee.

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

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

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

Piazzesi, M. and Swanson, E. T. (2008). Futures prices as risk-adjusted forecasts of monetary policy. *Journal of Monetary Economics* **55** 677–691.

Sakata, S. and White, H. (1998). High breakdown point conditional dispersion estimation with application to s & p 500 daily returns volatility. *Econometrica* **66** 529–567.

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

Stock, J. and Watson, M. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* **20** 147–162.

Stock, J. and Watson, M. (2010). Estimating turning points using large data sets. Tech. rep., National bureau of economic research.

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

Wachter, J. A. (2006). A consumption-based model of the term structure of interest rates. *Journal of Financial Economics* **79** 365–399.