Feasible Weighted Projected Principal Component Analysis for Factor Models with an Application to Bond Risk Premia[*]

Sung Hoon Choi[†]
University of Connecticut

July, 2021

Abstract

I develop a feasible weighted projected principal component (FPPC) analysis for factor models in which observable characteristics partially explain the latent factors. This novel method provides more efficient and accurate estimators than existing methods. To increase estimation efficiency, I take into account both cross-sectional dependence and heteroskedasticity by using a consistent estimator of the inverse error covariance matrix as the weight matrix. To improve accuracy, I employ a projection approach using characteristics because it removes noise components in high-dimensional factor analysis. By using the FPPC method, estimators of the factors and loadings have faster rates of convergence than those of the conventional factor analysis. Moreover, I propose an FPPC-based diffusion index forecasting model. The limiting distribution of the parameter estimates and the rate of convergence for forecast errors are obtained. Using U.S. bond market and macroeconomic data, I demonstrate that the proposed model outperforms models based on conventional principal component estimators. I also show that the proposed model performs well among a large group of machine learning techniques in forecasting excess bond returns.

Keywords: Semiparametric factor models, High-dimensionality, Unknown factors, Conditional sparsity, Thresholding, Cross-sectional correlation, Heteroskedasticity, Diffusion index, Forecasting.

[*] I am grateful to my advisor, Yuan Liao, for his advice and encouragement. I also thank Jushan Bai, Norman Swanson, John Landon-Lane, Xu Cheng, Jungjun Choi, and seminar participants at University of Connecticut, Queen Mary University of London, HKUST, National Taiwan University, Bank of Korea, and Korea Development Institute for useful comments and suggestions. All errors are my own.

[†] Address: 365 Fairfield Way, U-1063, Storrs, CT 06269, USA. E-mail: `sung_hoon.choi@uconn.edu`.

1 Introduction

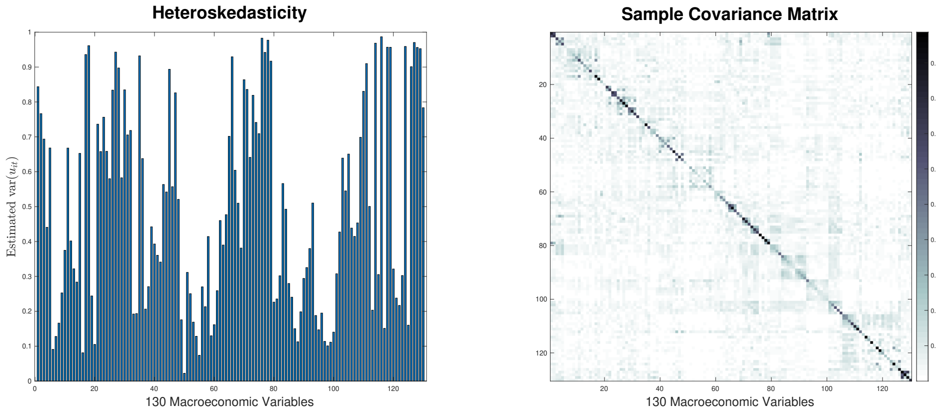
It is crucial to accurately and efficiently estimate latent factors in many economic and financial applications. For example, one would like to understand precisely how each individual stock depends on latent factors to examine its relative performance and risks. In addition, extracting more accurate factors improves forecasting performance with large datasets. This paper develops new statistical theories and methodologies for learning factors and big data forecasting, and demonstrates improved forecasting accuracy using excess returns on U.S. government bonds. In this paper, I demonstrate a novel theoretical econometric framework that incorporates the following two key aspects.

First, it is essential to consider a large error covariance matrix estimator for efficient estimation. In linear regression models, for example, it is well known that the generalized least squares estimators are more efficient than the ordinary least squares estimators in the presence of cross-sectional heteroskedasticity. Similarly, factor models often require the idiosyncratic error components to be cross-sectionally heteroskedastic and correlated. Intuitively, the large error covariance matrix is a non-diagonal matrix, and the diagonal entries may vary widely (e.g., Bai and Liao, 2016). Figure 1 shows that cross-sectional heteroskedasticity and correlations exist in real data, such as macroeconomic variables. However, the conventional principal component (PC) analysis method (e.g., Stock and Watson, 2002a; Bai, 2003) does not require estimating the error covariance matrix. This implies that the PC method essentially treats the error term to be homoskedastic and uncorrelated across the cross-sectional individuals. Hence, it is inefficient under cross-sectional heteroskedasticity with unknown dependence structures.[1] In this paper, I consider consistently estimating the high-dimensional error covariance matrix and its inverse. Using the inverse error covariance matrix estimator as the optimal weight matrix, a more efficient estimation than other existing methods can be obtained under cross-sectional heteroskedasticity and correlations.

Second, I consider factor models augmented by observed time series characteristics. The conventional PC method does not use extra information to improve estimation accuracy. However, we can improve it with more information and propose a "supervised" version of the PC method. Here, the eigenvectors (i.e., the principal components) are supervised by the additionally observed characteristics. This is very suitable in the context of asset pricing and economic forecasts. For example, a few observed covariates, such as aggregated macroeconomic variables (e.g., GDP, inflation, employment) or Fama-French factors, have explanatory powers for the latent factors. Fan et al. (2016) proposed a projected principal component (PPC) analysis, which employs the PC method to the projected data matrix onto a given linear space spanned by characteristics. Because the projection using characteristics removes

[1] Choi (2012) studied efficient estimations using weighted least squares in the approximate factor model by assuming the error covariance matrix to be known.

Figure 1: Cross-sectional heteroskedasticity and correlations in macroeconomic variables



Note: The first panel shows the estimated error variance for each cross-sectional individual using the estimated residuals by the regular PC method from 130 macroeconomic variables. The second panel displays an image of the sample error covariance matrix with scaled colors using the same estimated residuals.

noise components, it helps to estimate the factors more accurately than the conventional PC method.

This paper introduces a feasible weighted projected principal component (FPPC) analysis, which substantially improves both estimation efficiency and accuracy. The proposed estimator is constructed by first consistently estimating the error covariance matrix using the estimated residuals from the PPC method, then applying the PC method to the projected data combined with the inverse covariance estimator. Theoretically, I show that the rates of convergence of the FPPC estimators are faster than those of the regular PC estimators when both a cross-sectional dimension N and a time dimension T grow simultaneously.

Next, I suggest the FPPC-based diffusion index model. In the literature, the most popular application of factor models is factor-augmented regressions. For example, Stock and Watson (2002a) suggested the so-called diffusion index (DI) forecasting model, which uses factors estimated by the regular PC method to augment an autoregressive (AR) model. There has been a large literature on prediction with factor-based forecasting models, such as Stock and Watson (2002b), Bernanke et al. (2005), Bai and Ng (2006), Ludvigson and Ng (2009), and Kim and Swanson (2014), among many others. Importantly, more accurate and efficient estimations of the factors can substantially improve out-of-sample forecasts (see Bai and Liao, 2016). Therefore, this paper investigates whether and how the FPPC method can improve predictive accuracy in the DI model.

I apply the proposed FPPC-based DI model to the U.S. Treasury bond market. In the finance literature, the determinants of bond risk premia are crucial for both policymakers and investors (e.g., Fama and Bliss, 1987; Campbell and Shiller, 1991; Cochrane and Piazzesi, 2005). Such bond risk premia could be closely linked to macroeconomic factors. Among others, Ludvigson and Ng (2009) investigated critical linkages between bond returns and macroeconomic factors. The latent macroeconomic factors are estimated using the conventional PC method from a monthly balanced panel of macroeconomic time series. Moreover, they forecast excess returns of U.S. bonds using the conventional DI forecasting model. However, by using the proposed factor estimation method, the predictive accuracy of excess bond returns can be improved. In a recent paper, Bianchi et al. (2021) studied how machine learning methods (such as regression trees and neural networks) provide strong statistical evidence in predicting bond excess returns using both macroeconomic and yield data. Nevertheless, they did not consider and compare simple linear models such as AR and DI. Indeed, these linear models may perform well compared to nonlinear machine learning models in terms of out-of-sample forecasting.

In this empirical study, I compare the proposed FPPC method and the conventional PC and PPC methods in forecasting regression models using four characteristics including inflation, employment, forward factor, and GDP. I also compare and evaluate the forecasting performance of the proposed linear forecasting model and various machine learning models including penalized regressions (e.g., lasso, ridge, and elastic net), regression trees (e.g., gradient boosting and random forests), and neural networks. The experimental findings are based on the construction of one-year-ahead predictions using the sample period from January 1964 to April 2016. To evaluate the forecasting performances, I utilize out-of-sample R^2 and mean square forecast error (MSFE) criteria. Also, predictive accuracy tests of Diebold and Mariano (1995) and Giacomini and White (2006) are considered.

The empirical analysis points to several interesting findings. First, FPPC outperforms regular PC and PPC based on both in-sample and out-of-sample forecasting experiments. The forecasting gains associated with the FPPC method in the DI model range from approximately 6% to 30% compared to the PC method (in terms of out-of-sample R^2). These findings are robust to various forecasting periods and different factor-based models. Second, the FPPC-based DI models perform very well among a large group of machine learning techniques in the out-of-sample forecasts. These results confirm that nonlinear machine learning models may not be the best forecasting tools. Finally, based on MSFE criteria and point out-of-sample R^2, rolling window forecasts outperform recursive window forecasts for a majority of models considered in this paper. This indicates limited memory estimators are appropriate in out-of-sample forecasting because old data may no longer be informative (see Giacomini and White, 2006).

The rest of the paper is organized as follows. Section 2 formally proposes the FPPC method. Section 3 presents the assumptions and asymptotic analysis of the proposed estimators in both conventional and semiparametric factor models. Moreover, I study the FPPC-based DI model. Section 4 provides simulation studies. In Section 5, the econometric framework for the empirical study is demonstrated. I then introduce the data, the experimental setup, and key empirical findings. Finally, Section 6 concludes. All proofs, descriptions of all machine learning forecasting methods, and additional empirical results are given in the online supplement.

Throughout the paper, I use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of a square matrix A. I also let $\|A\|_F = \sqrt{\mathrm{tr}(A'A)}$, $\|A\|_2 = \sqrt{\lambda_{\max}(A'A)}$, and $\|A\|_1 = \max_i \sum_j |A_{ij}|$ denote the Frobenius norm, the spectral norm (also called the operator norm) and the L_1-norm of a matrix A, respectively. Note that if A is a vector, both $\|A\|$ and $\|A\|_F$ are equal to the Euclidean norm. In addition, $|a|$ is the absolute-value norm of a scalar a.

2 Feasible Weighted Projected Principal Component Analysis

2.1 Semiparametric factor model

This paper considers a semiparametric factor model defined by

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + u_{it}, \qquad i \leq N, t \leq T, \tag{2.1}$$

and

$$f_{tk} = g_k(\mathbf{X}_t) + \gamma_{tk}, \qquad t \leq T, k \leq K, \tag{2.2}$$

where $\{f_{t1}, \cdots, f_{tK}\}$ are latent common factors, $\{\lambda_{i1}, \cdots, \lambda_{iK}\}$ are corresponding factor loadings, and u_{it} is the idiosyncratic component. In addition, $\mathbf{X}_t$ is a $d \times 1$ vector of observable covariates, which partially explain the latent factors, and $g_k(\cdot)$ is a unknown nonparametric function.[2] For example, $\mathbf{X}_t$ can be the Fama-French factors or aggregated macroeconomic variables. Here, γ_{tk} is the component of common factors that cannot be explained by the covariates $\mathbf{X}_t$.[3] Define $\boldsymbol{\gamma}_t = (\gamma_{t1}, \cdots, \gamma_{tK})'$. I assume that $\{\boldsymbol{\gamma}_t\}_{t \leq T}$ have mean zero, and are independent of $\{\mathbf{X}_t\}_{t \leq T}$ and $\{u_{it}\}_{i \leq N, t \leq T}$. Then the model (2.1) and (2.2) can be represented using the following semiparametric factor structure:

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} \{g_k(\mathbf{X}_t) + \gamma_{tk}\} + u_{it}, \qquad i \leq N, t \leq T. \tag{2.3}$$

[2] Note that Connor and Linton (2007) studied the case of $\gamma_{tk} = 0$, which requires that the covariates fully explain the factor, and it is restrictive in many cases.

[3] A recent paper by Fan et al. (2020) considered a similar model and proposed a robust estimator for heavy-tailed distributions of the error term.

The model (2.3) can be stacked and written in a full matrix notation as

$$\mathbf{Y} = \mathbf{\Lambda}\{\mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}\}' + \mathbf{U}, \tag{2.4}$$

where $\mathbf{Y}$ is the $N \times T$ matrix of y_{it}, $\mathbf{\Lambda}$ is the $N \times K$ matrix of λ_{ik}, $\mathbf{G}(\mathbf{X})$ is the $T \times K$ matrix of $g_k(\mathbf{X}_t)$, $\mathbf{\Gamma}$ is the $T \times K$ matrix of γ_{tk} and $\mathbf{U}$ is $N \times T$ matrix of u_{it}. Note that the common factor matrix can be decomposed by $\mathbf{F} = \mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}$ from the model (2.2). Also $E(\mathbf{\Gamma}|\mathbf{X}) = 0$, where $\mathbf{G}(\mathbf{X})$ and $\mathbf{\Gamma}$ are orthogonal factor components so that $E[\mathbf{G}(\mathbf{X})\mathbf{\Gamma}'] = 0$. This paper assumes $K = \dim(\mathbf{F}_t)$ and $d = \dim(\mathbf{X}_t)$ to be constant. The number of factors, K, is assumed to be unknown. Because the considered model is a special case of the conventional latent factor model, in practice, the number of factors can be consistently estimated by existing methods such as AIC, BIC criteria (e.g., Bai and Ng, 2002), or eigenvalue ratio test methods (e.g., Lam and Yao, 2012; Ahn and Horenstein, 2013).

When the covariates $\mathbf{X}_t$ are multivariate, $g_k(\cdot)$ is assumed to be additive to estimate it nonparametrically as follows.[4] Define, for each $k \leq K$ and for each $t \leq T$,

$$g_k(\mathbf{X}_t) = \sum_{l=1}^{d} g_{kl}(X_{tl}) = \phi(\mathbf{X}_t)'\mathbf{b}_k + \sum_{l=1}^{d} R_{kl}(X_{tl}), \tag{2.5}$$

where

$$\mathbf{b}_k' = (b_{1,k1}, \cdots, b_{J,k1}, \cdots, b_{1,kd}, \cdots, b_{J,kd}) \in \mathbb{R}^{Jd},$$

$$\phi(\mathbf{X}_t)' = (\phi_1(X_{t1}), \cdots, \phi_J(X_{t1}), \cdots, \phi_1(X_{td}), \cdots, \phi_J(X_{td})) \in \mathbb{R}^{Jd}.$$

Here $\{\phi_1(x), \phi_2(x), \cdots\}$ is a set of basis functions, which spans a dense linear space of the functional space for $\{g_{kl}\}$; $\{b_{j,kl}\}_{j \leq J}$ are the sieve coefficients of the lth additive component of $g_k(\mathbf{X}_t)$ for the kth common factor; R_{kl} is an approximation error term; J denotes the number of sieve terms and it grows slowly as $T \to \infty$. Then each additive component $g_{kl}(\cdot)$, which is a nonparametric smooth fuction, can be estimated by the sieve method.

Let $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_K)'$ be a $K \times (Jd)$ matrix of sieve coefficients, $\mathbf{\Phi}(\mathbf{X}) = (\phi(\mathbf{X}_1), \cdots, \phi(\mathbf{X}_T))'$ be a $T \times (Jd)$ matrix of basis functions, and $\mathbf{R}(\mathbf{X})$ be $T \times K$ matrix with the (t, k)th element $\sum_{l=1}^{d} R_{kl}(X_{tl})$. Then (2.5) can be written in the matrix form:

$$\mathbf{G}(\mathbf{X}) = \mathbf{\Phi}(\mathbf{X})\mathbf{B}' + \mathbf{R}(\mathbf{X}). \tag{2.6}$$

Then the model (2.4) can be rewritten as

$$\mathbf{Y} = \mathbf{\Lambda}\{\mathbf{\Phi}(\mathbf{X})\mathbf{B}' + \mathbf{\Gamma}\}' + \mathbf{\Lambda}\mathbf{R}(\mathbf{X})' + \mathbf{U}. \tag{2.7}$$

[4] Note that $g_k(\mathbf{X}_t)$ does not depend on i and this implies that the common factors represent the heterogeneity of time series only.

Here, I describe the main idea of the projection. Let $\mathcal{X}$ be a sieve space spanned by the basis functions of $\mathbf{X}$. Let $\mathbf{P}$ denote the projection matrix onto $\mathcal{X}$. The projected data by operating $\mathbf{P}$ on both sides has the following sieve approximated representation:

$$\mathbf{YP} = \mathbf{\Lambda B \Phi}(\mathbf{X})' + \widetilde{\mathbf{E}}, \tag{2.8}$$

where $\widetilde{\mathbf{E}} = \mathbf{\Lambda\Gamma}'\mathbf{P} + \mathbf{\Lambda R}(\mathbf{X})'\mathbf{P} + \mathbf{UP} \approx 0$ because $\mathbf{\Gamma}$ and $\mathbf{U}$ are orthogonal to the function space spanned by the basis functions of $\mathbf{X}$, and $\mathbf{\Lambda R}(\mathbf{X})'$ is the sieve approximation error. In high-dimensional factor analysis, the projection removes those noise components, but the regular PC methods cannot remove them. Therefore, analyzing the projected data is an approximately noiseless problem and helps to obtain more accurate estimators.

Fan et al. (2016) proposed the projected principal component (PPC) method for the semi-parametric factor model.[5] Note that the idiosyncratic components are often cross-sectionally heteroskedastic and correlated in the approximate factor structure (e.g., Chamberlain and Rothschild, 1983). However, the PPC method does not require estimating the $N \times N$ covariance matrix, $\mathbf{\Sigma}_u = \mathrm{cov}(u_t)$, hence it essentially treats u_{it} to be homoskedastic and uncorrelated over i. As a result, it is inefficient under cross-sectional heteroskedasticity and correlations. Therefore, this paper considers the following a weighted least squares problem to efficiently estimate the approximate semiparametric factor models:

$$\min_{\mathbf{\Lambda},\mathbf{B}} \sum_{t=1}^{T} (y_t - \mathbf{\Lambda B}\phi(\mathbf{X}_t))' W (y_t - \mathbf{\Lambda B}\phi(\mathbf{X}_t)) \tag{2.9}$$

subject to certain normalization constraints. Here, W is an $N \times N$ positive definite weighted matrix. The first-order asymptotic optimal weight matrix is taken as $W = \mathbf{\Sigma}_u^{-1}$, which is not feasible in practice. Hence, the proposed FPPC method requires a consistent estimator $\widehat{\mathbf{\Sigma}}_u^{-1}$ for $\mathbf{\Sigma}_u^{-1}$ as the feasible weight matrix. To impliment the proposed method, we first project $\mathbf{Y}$ onto the sieve space spanned by $\{\mathbf{X}_t\}_{t\leq T}$, then employ the regular PC method to the projected data (i.e., the PPC method). Next, using the estimated residuals from the first step, the consistent estimator $\widehat{\mathbf{\Sigma}}_u^{-1}$ can be obtained by thresholding approach under the conditional sparsity assumption. More specific estimation procedure is discussed in the following sections.

[5] Fan et al. (2016) considered the similar semi-parametric factor model, where the factor loading has the semiparametric structure instead of the common factor, i.e., $\lambda_{ik} = g_k(\mathbf{X}_i) + \gamma_{ik}$, for each $i \leq N$, $k \leq K$.

2.2 Infeasible estimation

Recall that $\mathcal{X}$ is the sieve space spanned by the basis functions of $\mathbf{X}$. Define the $T \times T$ projection matrix onto $\mathcal{X}$:

$$\mathbf{P} = \mathbf{\Phi}(\mathbf{X})(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1}\mathbf{\Phi}(\mathbf{X})'. \tag{2.10}$$

Let $\mathbf{\Sigma}_u$ be the $N \times N$ covariance matrix of u_t, and assume that it is known for now. The common factors and loadings can be estimated by solving (2.9) with $W = \mathbf{\Sigma}_u^{-1}$ as the optimal weight matrix. Concentrating out $\mathbf{B}$ and using the normalization that $\frac{1}{N}\mathbf{\Lambda}'\mathbf{\Sigma}_u^{-1}\mathbf{\Lambda} = \mathbf{I}_K$, the optimization problem is identical to maximizing $\operatorname{tr}(\mathbf{\Lambda}'\mathbf{\Sigma}_u^{-1}\mathbf{Y}\mathbf{P}\mathbf{Y}'\mathbf{\Sigma}_u^{-1}\mathbf{\Lambda})$. Define $\mathbf{\Lambda}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{\Lambda}$ and $\mathbf{Y}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{Y}$. The estimated (infeasible) weighted loading matrix, denoted by $\widehat{\mathbf{\Lambda}^*}$, is $\sqrt{N}$ times the eigenvectors corresponding to the K largest eigenvalues of the $N \times N$ matrix $\mathbf{Y}^*\mathbf{P}\mathbf{Y}^{*\prime} = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{Y}\mathbf{P}\mathbf{Y}'\mathbf{\Sigma}_u^{-\frac{1}{2}}$. Note that the infeasible estimator of $\mathbf{\Lambda}$ is $\ddot{\mathbf{\Lambda}} = \mathbf{\Sigma}_u^{\frac{1}{2}}\widehat{\mathbf{\Lambda}^*}$. With the estimated weighted factor loadings $\widehat{\mathbf{\Lambda}^*}$, the least-squares estimator of common factor matrix is

$$\ddot{\mathbf{F}} = \frac{1}{N}\mathbf{Y}^{*\prime}\widehat{\mathbf{\Lambda}^*}.$$

In addition, given $\widehat{\mathbf{\Lambda}^*}$,

$$\ddot{\mathbf{G}}(\mathbf{X}) = \frac{1}{N}\mathbf{P}\mathbf{Y}^{*\prime}\widehat{\mathbf{\Lambda}^*}$$

is the estimator of $\mathbf{G}(\mathbf{X})$. By using (2.4), an estimator of $\mathbf{\Gamma}$, which cannot be explained by the covariates, is

$$\ddot{\mathbf{\Gamma}} = \ddot{\mathbf{F}} - \ddot{\mathbf{G}}(\mathbf{X}) = \frac{1}{N}(\mathbf{I} - \mathbf{P})\mathbf{Y}^{*\prime}\widehat{\mathbf{\Lambda}^*}.$$

2.3 Implementation of FPPC

The estimators are feasible only when a consistent estimator $\widehat{\mathbf{\Sigma}}_u^{-1}$ for the optimal weight $\mathbf{\Sigma}_u^{-1}$ is obtained. Therefore, this paper considers $W = \widehat{\mathbf{\Sigma}}_u^{-1}$, which takes into account both cross-sectional heteroskedasticity and correlations simultaneously, under the conditional sparsity assumption.

2.3.1 The estimator of $\mathbf{\Sigma}_u$ and FPPC

A thresholding method is applied to estimate $\mathbf{\Sigma}_u^{-1}$, as suggested by Fan et al. (2013). Let $\widetilde{R}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}$, where $\widehat{u}_{it}$ is the estimated residuals using the PPC method introduced by Fan et al. (2016). Define $\widehat{\mathbf{\Sigma}}_u = (\widehat{\Sigma}_{u,ij})_{N\times N}$, where

$$\widehat{\Sigma}_{u,ij} = \begin{cases} \widetilde{R}_{ii}, & i = j \\ s_{ij}(\widetilde{R}_{ij}), & i \neq j \end{cases},$$

Table 1: Three different principal component methods.

	Objective function	Eigenvectors of
PC	$\sum_{t=1}^{T}(y_t - \mathbf{\Lambda}\mathbf{F}_t)'(y_t - \mathbf{\Lambda}\mathbf{F}_t)$	$\mathbf{Y}\mathbf{Y}'$
PPC	$\sum_{t=1}^{T}(y_t - \mathbf{\Lambda}\mathbf{B}\phi(\mathbf{X}_t))'(y_t - \mathbf{\Lambda}\mathbf{B}\phi(\mathbf{X}_t))$	$\mathbf{Y}\mathbf{P}\mathbf{Y}'$
FPPC	$\sum_{t=1}^{T}(y_t - \mathbf{\Lambda}\mathbf{B}\phi(\mathbf{X}_t))'\widehat{\mathbf{\Sigma}}_u^{-1}(y_t - \mathbf{\Lambda}\mathbf{B}\phi(\mathbf{X}_t))$	$\widehat{\mathbf{\Sigma}}_u^{-1/2}\mathbf{Y}\mathbf{P}\mathbf{Y}'\widehat{\mathbf{\Sigma}}_u^{-1/2}$

where $s_{ij}(\cdot) : \mathbb{R} \to \mathbb{R}$ is a "soft-thresholding function" with an entry dependent threshold τ_{ij} such that:

$$s_{ij}(z) = \mathrm{sgn}(z)(|z| - \tau_{ij})_+,$$

where $(x)_+ = x$ if $x \geq 0$, and zero otherwise. Here $\mathrm{sgn}(\cdot)$ denotes the sign function. Note that other thresholding functions such as hard-thresholding are possible. For the threshold value, I specify

$$\tau_{ij} = M\omega_{N,T}\sqrt{\widetilde{R}_{ii}\widetilde{R}_{jj}}, \text{ where } \omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$$

for some pre-determined threshold constant $M > 0$. In practice, the tuning parameter M can be chosen by multifold cross-validation, which is discussed in Section 2.3.2. Intuitively, $\widehat{\mathbf{\Sigma}}_u$ thresholds off the small entries of the sample covariance matrix $\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_t'$, where residuals are obtained from the PPC estimate.

Now, I introduce the FPPC estimators using $\widehat{\mathbf{\Sigma}}_u^{-1}$ as the feasible weight matrix. Let $\widetilde{\mathbf{Y}} = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{Y}$, $\widetilde{\mathbf{\Lambda}} = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{\Lambda}$, and $\widetilde{\mathbf{U}} = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{U}$. Then the estimated feasible weighted loading matrix for $\mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{\Lambda}$, denoted by $\widehat{\widetilde{\mathbf{\Lambda}}}$, is $\sqrt{N}$ times the eigenvectors corresponding to the K largest eigenvalues of the $N \times N$ matrix $\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}' = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{Y}\mathbf{P}\mathbf{Y}'\widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}$. Note that the estimator of $\mathbf{\Lambda}$ is $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{\Sigma}}_u^{\frac{1}{2}}\widehat{\widetilde{\mathbf{\Lambda}}}$. Similar to infeasible estimators in Section 2.2, given $\widehat{\widetilde{\mathbf{\Lambda}}}$, the estimator of common factor is

$$\widehat{\mathbf{F}} = \frac{1}{N}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}} = \frac{1}{N}\mathbf{Y}'\widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\mathbf{\Lambda}}. \tag{2.11}$$

Moreover, given $\widehat{\widetilde{\mathbf{\Lambda}}}$,

$$\widehat{\mathbf{G}}(\mathbf{X}) = \frac{1}{N}\mathbf{P}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}}, \quad \widehat{\mathbf{\Gamma}} = \frac{1}{N}(\mathbf{I} - \mathbf{P})\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}} \tag{2.12}$$

are estimators of $\mathbf{G}(\mathbf{X})$ and $\mathbf{\Gamma}$, respectively.

In Section 3, I present asymptotic theory for the proposed FPPC estimators in both conventional and semiparametric factor models. Note that regular PC, PPC and FPPC minimize different objective functions, depending on the model specification and the weight matrix. Thus the loading estimators, $\widehat{\mathbf{\Lambda}}/\sqrt{N}$, are obtained from three different matrices. Table 1 presents the main differences of the estimators.

2.3.2 Choice of threshold

The suggested covariance matrix estimator, $\widehat{\mathbf{\Sigma}}_u$, requires the choice of tuning parameters M, which is the threshold constant. Define $\widehat{\mathbf{\Sigma}}_u(M) = \widehat{\mathbf{\Sigma}}_u$, where the covariance estimator depends on M.

The thresholding constant, M, can be chosen through multifold cross-validation (e.g., Bickel and Levina, 2008; Fan et al., 2013). First we obtain the estimated $N \times 1$ vector residuals $\widehat{\mathbf{u}}_t$ by PPC, then divide the data into $P = \log(T)$ blocks $J_1, ..., J_P$ with block length $T/\log(T)$. Here we take one of the P blocks as the validation set. At the pth split, let $\widehat{\mathbf{\Sigma}}_u^p$ be the sample covariance matrix based on the validation set, defined by $\widehat{\mathbf{\Sigma}}_u^p = J_p^{-1} \sum_{t \in J_p} \widehat{\mathbf{u}}_t \widehat{\mathbf{u}}_t'$. Let $\widehat{\mathbf{\Sigma}}_u^{S,p}(M)$ be the soft-thresholding estimator with threshold constant M using the training data set $\{\widehat{\mathbf{u}}\}_{t \notin J_p}$. Then we choose the constant M^* by minimizing a cross-validation objective function

$$M^* = \arg \min_{c_{\min} < M < c_{\max}} \frac{1}{P} \sum_{j=1}^{P} \|\widehat{\mathbf{\Sigma}}_u^{S,p}(M) - \widehat{\mathbf{\Sigma}}_u^p\|_F^2,$$

where $c_{\max}$ is a large constant such that $\widehat{\mathbf{\Sigma}}_u(c_{\max})$ is a diagonal matrix, and $c_{\min}$ is the minimum constant that $\widehat{\mathbf{\Sigma}}_u(M)$ is positive definite for $M > c_{\min}$:

$$c_{\min} = \inf[C > 0 : \lambda_{\min}\{\widehat{\mathbf{\Sigma}}_u(M)\} > 0, \forall M > C].$$

Then, the resulting estimator of $\mathbf{\Sigma}_u$ is $\widehat{\mathbf{\Sigma}}_u(M^*)$.

3 Asymptotic Analysis

In this section, I provide assumptions and asymptotic performances of the proposed estimators in both conventional and semiparametric factor models.

3.1 Sparsity condition on $\mathbf{\Sigma}_u$

In the literature, one of the commonly used assumptions to estimate a high-dimensional covariance matrix is the sparsity, for example, Bickel and Levina (2008), Rothman et al. (2008), and Fan et al. (2013), among others. This paper assumes $\mathbf{\Sigma}_u = (\Sigma_{u,ij})_{N \times N}$ to be a sparse matrix, namely most of the off-diagonal entries are zero or nearly so. This special structure is known as the "conditional sparsity" given the common factors in an approximate factor model (see Fan et al., 2011). For some $q \in [0, 1)$, define

$$m_N = \max_{i \le N} \sum_{j=1}^{N} |\Sigma_{u,ij}|^q, \tag{3.1}$$

and it does not grow too fast as $N \to \infty$. In particular, when $q = 0$ (i.e., the exact sparsity case), $m_N = \max_{i \le N} \sum_{j=1}^{N} 1\{\Sigma_{u,ij} \neq 0\}$, which implies the maximum number of non-zero elements in each row.

The following assumption defines the conditional sparsity on $\mathbf{\Sigma}_u$.

Assumption 3.1. *(i) There is $q \in [0, 1)$ such that, for semiparametric factor models,*

$$m_N \omega_{N,T}^{1-q} = o(1), \text{ where } \omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}. \tag{3.2}$$

Or, for conventional factor models,

$$m_N \delta_{N,T}^{1-q} = o(1), \text{ where } \delta_{N,T} = \sqrt{\frac{\log N}{T}} + \sqrt{\frac{J}{T}} + \frac{1}{\sqrt{N}}. \tag{3.3}$$

(ii) There are constants $c_1, c_2 > 0$ such that $\lambda_{\min}(\mathbf{\Sigma}_u) > c_1$ and $\max_{i \le N} \sum_{j=1}^{N} |\Sigma_{u,ij}| < c_2$.

Condition (i) is needed for the $\| \cdot \|_1$-convergence of estimating $\mathbf{\Sigma}_u$ and its inverse. Condition (ii) requires that $\mathbf{\Sigma}_u$ be well conditioned. This is a standard assumption of idiosyncratic term in the approximate factor model literature, such as Bai (2003) and Bai and Ng (2008b).

Remark 3.1. Similar to Fan et al. (2013), for m_N and q defined in (3.1), we have

$$\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1 = O_P(m_N \omega_{N,T}^{1-q}), \tag{3.4}$$

if (3.2) holds. When m_N grows slowly with N, $\widehat{\mathbf{\Sigma}}_u^{-1}$ is consistent estimator with a nice convergence rate. In addition, when $m_N = O(1)$, $q = 0$ and $N > T$, the rate would be $O_P(\sqrt{\frac{\log N}{T}})$, which is minimax optimal rate as proved by Cai and Zhou (2012). On the other hand, for statistical inference purposes (e.g., deriving limiting distributions of estimated factors), we need to further strengthen the sparse condition to obtain $\|\frac{1}{\sqrt{N}} \mathbf{\Lambda}'(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}) u_t\| = o_P(1)$. Specifically, the above "absolute convergence" for the estimator would be too restrictive to be applicable when $N > T$ (see Bai and Liao, 2017).

3.2 FPPC in conventional factor models

This subsection studies the asymptotic performance of the FPPC in the conventional factor model:

$$\mathbf{Y} = \mathbf{\Lambda} \mathbf{F}' + \mathbf{U}. \tag{3.5}$$

Note that in financial applications, the latent factors are often treated to be weakly dependent time series, which satisfy strong mixing conditions. On the other hand, in many statistical application, the factors are assumed to be serially independent.

I introduce the conditions and asymptotic properties of the FPPC analysis.[6] Recall that the projection matrix is defined as

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

The following assumption is the most essential condition in this context.

Assumption 3.2. *(Genuine projection). There are positive constants c_1 and c_2 such that, with probability approaching one as $T \to \infty$,*

$$c_1 < \lambda_{\min}(T^{-1}\mathbf{F}'\mathbf{P}\mathbf{F}) < \lambda_{\max}(T^{-1}\mathbf{F}'\mathbf{P}\mathbf{F}) < c_2.$$

This assumption is a special type of "pervasive" condition on the factors. It requires that the observed characteristics have an explanatory power for the latent factors. Note that the dimensions of $\mathbf{\Phi}(\mathbf{X})$ and $\mathbf{F}$ are $T \times Jd$ and $T \times K$, respectively. Since the number of factors is assumed to be fixed in this paper, this assumption requires $Jd \geq K$. For any nonsigular matrix $\mathbf{M}$, $\mathbf{\Lambda}\mathbf{F}' = \mathbf{\Lambda}\mathbf{M}^{-1}\mathbf{M}\mathbf{F}'$, it has been well known that $\mathbf{\Lambda}$ and $\mathbf{F}$ are not separately identifiable without further restrictions (see Bai and Ng, 2013). Similar to Stock and Watson (2002a) and Bai (2003), the FPPC estimator estimates transformed factors and loadings.

Assumption 3.3. *(Basis functions). (i) There are d_1, $d_2 > 0$ so that with probability approaching one as $T \to \infty$,*

$$d_1 < \lambda_{\min}(T^{-1}\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X})) < \lambda_{\max}(T^{-1}\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X})) < d_2.$$

(ii) $\max_{j\leq J, t\leq T, l\leq d} E\phi_j(X_{tl})^2 \leq \infty$.

Since $T^{-1}\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}) = T^{-1}\sum_{t=1}^{T}\phi(\mathbf{X}_t)\phi(\mathbf{X}_t)'$ and $\phi(\mathbf{X}_t)$ is a $Jd \times 1$ vector, where $Jd \ll T$, the strong law of large numbers implies condition (i). This condition can be satisfied over normalizations of commonly used basis functions, e.g., Fourier basis, B-splines, polinomial basis.

Assumption 3.4. *(Data generating process). (i) $\{\mathbf{F}_t, \mathbf{u}_t\}_{t\leq T}$ is strictly stationary. $E\mathbf{u}_t = 0$ for all $t \leq T$; $\{\mathbf{u}_t\}_{t\leq T}$ is independent of $\{\mathbf{X}_t, \mathbf{F}_t\}_{t\leq T}$.*
(ii) Strong mixing: There exist r_1, $C > 0$ such that for all $T > 0$,

$$\sup_{A\in\mathcal{F}_{-\infty}^{0}, B\in\mathcal{F}_{T}^{\infty}} |P(A)P(B) - P(AB)| < \exp(-CT^{r_1}),$$

where $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras generated by $\{(\mathbf{F}_t, \mathbf{u}_t) : -\infty \leq t \leq 0\}$ and $\{(\mathbf{F}_t, \mathbf{u}_t) : T \leq t \leq \infty\}$, respectively.

[6] The conditions are symmetric to that of Fan et al. (2016), because they considered the case of the loading matrix is explained by characteristics covariates: $\lambda_{ik} = g_k(\mathbf{X}_i) + \gamma_{ik}$, for $i = 1, ..., N, k = 1, ..., K$.

(iii) Exponential tail: There exist $r_2, r_3 > 0$ satisfying $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$ and $b_1, b_2 > 0$, such that for any $s > 0, i \leq N$ and $k \leq K$,

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

(iv) Weak dependence: There exists a positive constant $M < \infty$ so that

$$\max_{s \leq T} \sum_{t=1}^{T} |Eu_{it}u_{is}| < M,$$

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} |Eu_{it}u_{js}| < M,$$

$$\max_{t \leq T} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{s=1}^{T} \sum_{q=1}^{T} |\mathrm{cov}(u_{it}u_{is}, u_{jt}u_{jq})| < M.$$

Condition (ii) allows factors and idiosyncratic components to be weakly serial dependent by requring the strong-mixing. Condition (iii) ensures the Bernstein-type inequality for weakly dependent data. Note that the underlying distributions are assumed to be thin-tailed. Allowing for heavy-tailed distributions is also an important issue, but it would require a very different estimation method (see Fan et al., 2020). Condition (iv) is commonly imposed in high-dimensional factor analysis, such as Stock and Watson (2002a) and Bai (2003).

Formally, the following theorem presents the rates of convergence for the FPPC estimators defined in Section 2.3.

Theorem 3.1. *(Conventional factor model). Suppose that Assumptions 3.1-3.4 hold. For an invertible matrix $\mathbf{M}$, as $N, T \to \infty$, and J can be either divergent with T satisfying $J = o(\sqrt{T})$ or bounded with $Jd \geq K$, we have*

$$\frac{1}{N} \|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda M}\|_F^2 = O_P\left(\frac{J}{T}\right),$$

$$\frac{1}{T} \|\widehat{\mathbf{G}}(\mathbf{X}) - \mathbf{PFM}\|_F^2 = O_P\left(\frac{J^2}{T^2} + \frac{J}{T} m_N^2 \delta_{N,T}^{2-2q}\right).$$

In addition, for any $t \leq T$,

$$\|\widehat{\mathbf{F}}_t - \mathbf{M}^{-1}\mathbf{F}_t\| = O_P\left(m_N \delta_{N,T}^{1-q}\right).$$

The convergence rate for the estimated loadings can be faster than that of the conventional PC method. In addition, the FPPC has a nice convergence rate, which is much faster than the regular PC, for the factor matrix up to a projection transformation (see Remark 3.2). Note that the PPC estimates, which do not employ the error covariance matrix estimator, are consistent even if N is finite. However, since FPPC method exploits the consistent estimator $\widehat{\mathbf{\Sigma}}_u^{-1}$, it requires large-$N$ and large-T for the factor and loading estimates.

3.3 FPPC in semiparametric factor models

In this subsection, I consider the semiparametric factor structure: $f_{tk} = g_k(\mathbf{X}_t) + \gamma_{tk}$. Here $g_k(\mathbf{X}_t)$ is a nonparametric smooth function for the observed covariates, and γ_{tk} is the unobserved random factor component, which is independent of $\mathbf{X}_t$. In the matrix form, the model can be written as:

$$\mathbf{Y} = \mathbf{\Lambda}\{\mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}\}' + \mathbf{U}.$$

Recall that $\widetilde{\mathbf{Y}} = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{Y}$ and $\widetilde{\mathbf{U}} = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{U}$. Then the projected data has the following sieve approximated representation:

$$\widetilde{\mathbf{Y}}\mathbf{P} = \widetilde{\mathbf{\Lambda}}\mathbf{B}\mathbf{\Phi}(\mathbf{X})' + \mathbf{E}, \tag{3.6}$$

where $\mathbf{E} = \widetilde{\mathbf{\Lambda}}\mathbf{R}(\mathbf{X})'\mathbf{P} + \widetilde{\mathbf{\Lambda}}\mathbf{\Gamma}'\mathbf{P} + \widetilde{\mathbf{U}}\mathbf{P}$ is approximately "small", because $\mathbf{R}(\mathbf{X})$ is the sieve approximation error, and $\mathbf{\Gamma}$ and $\widetilde{\mathbf{U}}$ are orthogonal to the function space spanned by $\mathbf{X}$. Note that, similar to Fan et al. (2016), the sieve coefficient matrix $\mathbf{B}$ can be estimated by least squres from the above model (3.6) as:

$$\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_K)' = \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widetilde{\mathbf{Y}}\mathbf{\Phi}(\mathbf{X})[\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X})]^{-1}. \tag{3.7}$$

Then the estimator for $g_k(.)$ is

$$\widehat{g}_k(\mathbf{x}) = \phi(\mathbf{x})'\widehat{\mathbf{b}}_k \quad \forall \mathbf{x} \in \mathcal{X}, k = 1, \cdots, K, \tag{3.8}$$

where $\mathcal{X}$ denotes the support of $\mathbf{X}_t$.

The estimators $\widehat{\mathbf{\Lambda}}$, $\widehat{\mathbf{G}}(\mathbf{X})$ and $\widehat{\mathbf{F}}$ are the FPPC estimators as defined in Section 2.3. Since $\mathbf{F} = \mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}$, $\mathbf{G}(\mathbf{X})$ can be regarded as the projection of $\mathbf{F}$ onto the sieve space spanned by $\mathbf{X}$. Therefore, the following assumption is a sufficient condition for Assumption 3.2 in the semiparametric factor model.

Assumption 3.5. *There are two positive constants c_1 and c_2 so that with probability approaching one as $T \to \infty$,*

$$c_1 < \lambda_{\min}(T^{-1}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) < \lambda_{\max}(T^{-1}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) < c_2.$$

Define $\boldsymbol{\gamma}_t = (\gamma_{t1}, \cdots, \gamma_{tK})'$. In this paper, serial weak dependence for $\{\boldsymbol{\gamma}_t\}_{t \le T}$ is imposed as follows.

Assumption 3.6. *(i) $E\gamma_{tk} = 0$ and $\{\mathbf{X}_t\}_{t \le T}$ is independent of $\{\gamma_{tk}\}_{t \le T}$.*

(ii) $\max_{k\leq K, t\leq T} Eg_k(\mathbf{X}_t)^2 \leq \infty, \nu_T < \infty$ *and*

$$\max_{k\leq K, s\leq T} \sum_{t\leq T} |E\gamma_{tk}\gamma_{sk}| = O(\nu_T),$$

where

$$\nu_T = \max_{k\leq K} \frac{1}{T}\sum_{t\leq T} \text{var}(\gamma_{tk}).$$

The following assumption is needed for the accuracy of the sieve approximation.

Assumption 3.7. *(Accuracy of sieve approximation). For all $l \leq d, k \leq K$,*
(i) the factor component $g_{kl}(\cdot)$ belongs to a Hölder class $\mathcal{G}$ defined by

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

for some $L > 0$.
(ii) the sieve coefficients $\{b_{j,kl}\}_{j\leq J}$ satisfy for $\kappa = 2(r+\alpha) \geq 4$, as $J \to \infty$,

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

where $\mathcal{X}_l$ is the support of the lth element of $\mathbf{X}_t$, and J is the sieve dimension.
(iii) $\max_{k,j,l} b_{j,kl}^2 < \infty$.

Note that condition (ii) is satisfied by common basis. For example, when $\{\phi_j\}$ is B-splines or polynomial basis, condition (i) implies condition (ii), as discussed in Chen (2007) and Fan et al. (2016).

Theorem 3.2. *(Semiparametric factor model). Suppose $J = o(\sqrt{T})$ and Assumptions 3.1, 3.3-3.7 hold. There is an invertible matrix $\mathbf{H}$, as $N, T, J \to \infty$, we have, for $\omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,*

$$\begin{aligned}
\frac{1}{N}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda H}\|_F^2 &= O_P\left(\frac{1}{T}\right), \\
\frac{1}{T}\|\widehat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X})\mathbf{H}\|_F^2 &= O_P\left(\frac{1}{J^{\kappa}} + \frac{J\nu_T}{T} + \frac{J}{T}m_N^2\omega_{N,T}^{2-2q}\right), \\
\frac{1}{T}\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma H}\|_F^2 &= O_P\left(\frac{1}{N} + \frac{1}{J^{\kappa}} + \frac{J}{T^2} + \frac{J\nu_T}{T} + \frac{1}{T}m_N^2\omega_{N,T}^{2-2q}\right).
\end{aligned}$$

In addition, for any $t \leq T$,

$$\|\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t\| = O_P\left(m_N\omega_{N,T}^{1-q}\right).$$

Note that $\widehat{\mathbf{F}} = \widehat{\mathbf{G}}(\mathbf{X}) + \widehat{\mathbf{\Gamma}}$, hence the convergence rate for the estimated common factor can be obtained by two convergences. We have the following remark about the rates of convergence above compared with those using the conventional PC method.

Remark 3.2. Denote $\widehat{\mathbf{\Lambda}} = (\widehat{\boldsymbol{\lambda}}_1, ..., \widehat{\boldsymbol{\lambda}}_N)'$, $\widehat{\mathbf{G}}(\mathbf{X}) = (\widehat{\mathbf{g}}(\mathbf{X}_1), ..., \widehat{\mathbf{g}}(\mathbf{X}_T))'$, and $\widehat{\mathbf{\Gamma}} = (\widehat{\gamma}_1, ..., \widehat{\gamma}_T)'$. For the factor loading, we have

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

For the factor components, consider $m_N = O(1)$ and $q = 0$ as a simple case. Define the optimal $J^* = (T\min\{N, T/\log N, \nu_T^{-1}\})^{1/(\kappa+1)}$. With $J = J^*$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{g}}(\mathbf{X}_t) - \mathbf{H}^{-1}\mathbf{g}(\mathbf{X}_t)\|^2 = O_P\left(\frac{1}{(T\min\{N, T/\log N, \nu_T^{-1}\})^{1-1/(\kappa+1)}}\right).$$

Moreover, when $N = O(1)$ and κ is sufficiently large, the rate is close to $O_P(T^{-1})$. This implies that, when $\mathbf{F}_t = \mathbf{g}(\mathbf{X}_t)$, the rates of factors and loadings are faster than the rates of the regular PC method estimators $(\widetilde{\boldsymbol{\lambda}}_i, \widetilde{\mathbf{F}}_t)$, such as Stock and Watson (2002a) and Bai (2003): for some rotation matrix $\widetilde{\mathbf{H}}$,

$$\frac{1}{N}\sum_{i=1}^{N}\|\widetilde{\boldsymbol{\lambda}}_i - \widetilde{\mathbf{H}}'\boldsymbol{\lambda}_i\|^2 = O_P\left(\frac{1}{T} + \frac{1}{N}\right), \quad \frac{1}{T}\sum_{t=1}^{T}\|\widetilde{\mathbf{F}}_t - \widetilde{\mathbf{H}}^{-1}\mathbf{F}_t\|^2 = O_P\left(\frac{1}{T} + \frac{1}{N}\right).$$

On the other hand, when the common factor cannot be fully explained by the covariates, we have $\widehat{\mathbf{\Gamma}} = (\widehat{\gamma}_1, ..., \widehat{\gamma}_T)'$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\gamma}_t - \mathbf{H}^{-1}\boldsymbol{\gamma}_t\|^2 = O_P\left(\frac{1}{N} + \frac{1}{(T\min\{N, T/\log N, \nu_T^{-1}\})^{1-1/(\kappa+1)}}\right),$$

which requires $N \to \infty$ to be consistent.

3.4 Diffusion index forecasting models

Next, I study the forecasting regression model using the estimated factors, the so-called diffusion index (DI) forecasting model, originally proposed by Stock and Watson (2002a). In the forecasting literature, this model has been used widely for prediction. In this paper, I investigate the limiting distribution of the parameter estimates and the rate of convergence for forecast errors based on factor estimates via the FPPC method.

Consider the following forecasting equation:

$$z_{t+h} = \alpha'\mathbf{F}_t + \beta' W_t + \epsilon_{t+h}, \tag{3.9}$$

where h is a forecasting horizon, $\mathbf{F}_t$ is unobservable factors and W_t are observable variables (e.g., lags of z_t). Because $\mathbf{F}_t$ is latent, we can obtain $\widehat{\mathbf{F}}_t$ using principal components methods from the factor model:

$$y_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{u}_t. \tag{3.10}$$

Note that, when z_t is a scalar, equations (3.9) and (3.10) constitute the DI model. In addition, when $h = 1$ and $z_{t+1} = (\mathbf{F}'_{t+1}, W'_{t+1})'$, the equation (3.9) is the FAVAR model suggested by Bernanke et al. (2005). Intuitively, the common factor, $\mathbf{F}_t$, can be summarized from the large economic time series y_t, and it is known for the common shocks that generate comovements.

Define $L_t = (\mathbf{F}'_t, W'_t)'$. For now, assume that $\mathbf{F}_t$ is observable and $E(\epsilon_{T+h}|L_T, L_{T-1}, ...) = 0$. Suppose the conditional mean of (3.9) is the target of interest:

$$z_{T+h|T} = E(z_{T+h}|L_T, L_{T-1}, ...) = \alpha'\mathbf{F}_t + \beta' W_t \equiv \delta' L_T.$$

However, it is not feasible to predict because of unobservable parameters, α, β, and $\mathbf{F}_t$, in practice. Let $\widehat{\alpha}$ and $\widehat{\beta}$ be the least squares estimates from regression z_{t+h} on $\widehat{L}_t = (\widehat{\mathbf{F}}'_t, W'_t)'$, for $t = 1, ..., T-h$, where $\widehat{\mathbf{F}}_t$ is the FPPC factor estimates based on the semiparametric factor model. Note that the true parameter α cannot be identified, because $\widehat{\mathbf{F}}_t$ and $\widehat{\alpha}$ estimate rotations of $\mathbf{F}_t$ and α, respectively. The feasible prediction can be obtained by

$$\widehat{z}_{T+h|T} = \widehat{\alpha}'\widehat{\mathbf{F}}_T + \widehat{\beta}' W_T = \widehat{\delta}'\widehat{L}_T.$$

For example, if z_t employment rate, the estimated conditional mean can be represented by an estimate of the expected employment rate. Stock and Watson (2002a) showed consistency of $\widehat{z}_{T+h|T}$ for $z_{T+h|T}$. Bai and Ng (2006) established the limiting distributions of the least squares estimates and forecast errors so that inference can be conducted. Note that these papers used the regular PC estimation method under the static factor model. On the other hand, I consider the FPPC-based DI forecasting model and its properties. The following assumption is standard for forecasting regression analysis.

Assumption 3.8. *Let $L_t = (\mathbf{F}'_t, W'_t)'$. $E\|L_t\|^4$ is bounded for every t.*
(i) $E(\epsilon_{t+h}|z_t, L_t, z_{t-1}, L_{t-1}, ...) = 0$ for any $h > 0$, and L_t and ϵ_t are independent of the idiosyncratic errors u_{is} for all i and s.
(ii) $\frac{1}{T}\sum_{t=1}^{T} L_t L'_t \xrightarrow{p} \Sigma_L$, which is a positive definite matrix.
(iii) $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} L_t\epsilon_{t+h} \xrightarrow{d} N(0, \Sigma_{L,\epsilon})$, where $\Sigma_{L,\epsilon} = \text{plim}\,\frac{1}{T}\sum_{t=1}^{T}\epsilon^2_{t+h} L_t L'_t$.

Condition (i) implies that the idiosyncratic errors from the factor model and all the random variables in the forecasting model are independent. Conditions (ii)-(iii) are needed for regression analysis.

In this section, I assume the semiparametric factor model studied in Section 3.3 instead of the equation (3.10). All the theorems and proofs based on the conventional factor model can be obtained similarly. The limiting distribution for OLS estimators of the DI model is discussed in the following theorem.

Theorem 3.3. *Let $\widehat{\delta} = (\widehat{\alpha}', \widehat{\beta}')'$ and $\delta = (\alpha'\mathbf{H}, \beta')'$. Suppose the assumptions of Theorem 3.2 and Assumption 3.8 hold. For q, m_N, and $\omega_{N,T}$ defined in (3.2), if $\sqrt{T} m_N^2 \omega_{N,T}^{2-2q} = o(1)$,*

$$\sqrt{T}(\widehat{\delta} - \delta) \xrightarrow{d} N(0, \Sigma_\delta),$$

where $\Sigma_\delta = \Pi'^{-1}\Sigma_L^{-1}\Sigma_{L,\epsilon}\Sigma_L^{-1}\Pi'$ with $\Pi = \mathrm{diag}(\mathbf{H}', \mathbf{I})$. A heteroskedasticity consistent estimator for Σ_δ is

$$\widehat{\Sigma}_\delta = \left(\frac{1}{T}\sum_{t=1}^{T-h} \widehat{L}_t \widehat{L}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T-h} \widehat{\epsilon}_{t+h}^2 \widehat{L}_t \widehat{L}_t'\right) \left(\frac{1}{T}\sum_{t=1}^{T-h} \widehat{L}_t \widehat{L}_t'\right)^{-1}.$$

Remark 3.3. Consider a special case where $m_N = O(1)$ and $q = 0$ (i.e., a strictly sparse case), which means the number of nonzero elements in each row of $\mathbf{\Sigma}_u$ is bounded. Then the condition $\sqrt{T} m_N^2 \omega_{N,T}^{2-2q} = o(1)$ becomes $\frac{\log N}{\sqrt{T}} + \frac{\sqrt{T}}{N} = o(1)$, which holds if $\sqrt{T} = o(N)$. Implicitly, requiring $\sqrt{T}/N \to 0$ is needed for the asymptotic normality of $\widehat{\delta}$ as Bai and Ng (2006) imposed.

I now study the convergence rate for the estimated conditional mean. Consider the following equation:

$$\widehat{z}_{T+h|T} - z_{T+h|T} = (\widehat{\delta} - \delta)'\widehat{L}_T + \alpha'\mathbf{H}(\widehat{\mathbf{F}}_T - \mathbf{H}^{-1}\mathbf{F}_T),$$

that contains two components which are caused by estimating δ and $\mathbf{F}_t$.

Theorem 3.4. *Let $\widehat{z}_{T+h|T} = \widehat{\delta}'\widehat{L}_T$. Suppose that the assumptions of Theorem 3.3 hold. Then, for $\omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,*

$$\widehat{z}_{T+h|T} - z_{T+h|T} = O_P(m_N \omega_{N,T}^{1-q}).$$

The overall rate of convergence is similar to Bai and Ng (2006), which is $\min[\sqrt{T}, \sqrt{N}]$. Note that obtaining the asymptotic properties of the DI forecasts requires the limiting distributions of the estimated factors (e.g., Bai, 2003). However, because this paper only establishes the rate of convergence for FPPC factor estimates, formal theoretical studies on this issue are left to future research.

4 Monte Carlo Simulations

In this section, I conduct numerical experiments to compare the proposed FPPC method with other existing methods. Consider the following semiparametric factor model,

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{u}_t, \text{ and } \mathbf{F}_t = \sigma_g \mathbf{g}(\mathbf{X}_t) + \sigma_\gamma \boldsymbol{\gamma}_t, \text{ for } t = 1, \cdots T,$$

where $\mathbf{\Lambda}$ is drawn from i.i.d. Uniform$(0, 1)$. I set $\dim(\mathbf{X}_t) = 3$ and three factors (i.e., $K = 3$). I introduce serial dependences on $\mathbf{X}_t$ and $\boldsymbol{\gamma}_t$ as follows:

$$\mathbf{X}_t = \Psi\mathbf{X}_{t-1} + \boldsymbol{\xi}_t, \text{ and } \boldsymbol{\gamma}_t = \Psi\boldsymbol{\gamma}_{t-1} + \boldsymbol{\nu}_t, \text{ for } t = 1, \cdots T,$$

with $\mathbf{X}_0 = \mathbf{0}$, $\boldsymbol{\gamma}_0 = \mathbf{0}$ and a 3×3 diagonal matrix Ψ. Each diagonal element of Ψ is generated from Uniform$(0.3, 0.7)$. In addition, $\boldsymbol{\xi}_t$ and $\boldsymbol{\nu}_t$ are drawn from i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I})$. To address different correlations between $\mathbf{F}_t$ and $\mathbf{g}(\mathbf{X}_t)$, define $\sigma_g^2 = \frac{w}{1+w}$ and $\sigma_\gamma^2 = \frac{1}{1+w}$. Here, three different values, 10, 1, and 0.1, are used for w. Note that the larger w represents the stronger explanatory power.

The unknown function $\mathbf{g}(\cdot)$ has the following model: $\mathbf{g}(\mathbf{X}_t) = (g_1(\mathbf{X}_t), \cdots, g_K(\mathbf{X}_t))'$, where $g_k(\mathbf{X}_t) = \sum_{l=1}^3 g_{kl}(X_{tl})$. The three characteristic functions are $g_{1l} = x, g_{2l} = x^2 - 1$, and $g_{3l} = x^3 - 2x$, for all $l \le d$. For each $k \le K$, I standardize the $g_k(\mathbf{X}_t)$ and $\boldsymbol{\gamma}_{k,t}$ such that they have mean of zero and standard deviation of one.

Next, the idiosyncratic errors are generated using a $N \times N$ banded covariance matrix Σ_u as follows: let $\{\varepsilon_{it}\}_{i \le N, t \le T}$ be i.i.d. $\mathcal{N}(0, 1)$. Let

$$\eta_{1t} = \varepsilon_{1t}, \eta_{2t} = \varepsilon_{2t} + a_1\varepsilon_{1t}, \eta_{3t} = \varepsilon_{3t} + a_2\varepsilon_{2t} + b_1\varepsilon_{1t},$$

$$\eta_{i+1,t} = \varepsilon_{i+1,t} + a_i\varepsilon_{it} + b_{i-1}\varepsilon_{i-1,t} + c_{i-2}\varepsilon_{i-2,t},$$

where the constants $\{a_i, b_i, c_i\}_{i=1}^N$ are generated from i.i.d. $\mathcal{N}(0, \sqrt{5})$. Here I denote the correlation matrix of $\eta_t = (\eta_{1t}, \cdots, \eta_{Nt})'$ by R_η, which is a banded matrix. Then the cross-sectional heteroskedasticity is introduced as follows: let $D = \mathrm{diag}(d_i)$, where $\{d_i\}_{i \le N}$ is drawn from i.i.d. Uniform$(0, \sqrt{5})$. Finally, define $\Sigma_u = DR_\eta D$, and generate $\{u_t\}_{t \le T}$ as i.i.d. $\mathcal{N}(0, \Sigma_u)$. Note that this generating procedure of the error term is similar to that of Bai and Liao (2017).

For sample sizes, I consider $N = 50, 100$ and $T = 100, 200, 500$. The additive polynomial basis with $J = 5$ is used for the sieve basis. The threshold constant M for FPPC is chosen by the 5-fold cross-validation, as discussed in Section 2.3.2.

Table 2: Canonical correlations of estimated loading or factor matrices (the larger the better) and mean squared error of $\mathbf{\Lambda F}'$ (the smaller the better).

		Strong($w=10$)			Mild($w=1$)			Weak($w=0.1$)		
N	T	PC	PPC	FPPC	PC	PPC	FPPC	PC	PPC	FPPC
					Loadings					
50	100	0.246	0.715	0.790	0.296	0.515	0.665	0.326	0.239	0.340
	200	0.296	0.873	0.887	0.358	0.762	0.814	0.418	0.358	0.477
100	100	0.170	0.734	0.787	0.303	0.583	0.695	0.442	0.276	0.432
	500	0.147	0.952	0.954	0.421	0.913	0.918	0.755	0.628	0.715
					Factors					
50	100	0.180	0.618	0.901	0.228	0.487	0.859	0.269	0.261	0.572
	200	0.185	0.695	0.919	0.248	0.652	0.922	0.312	0.357	0.700
100	100	0.189	0.750	0.963	0.322	0.639	0.953	0.467	0.361	0.807
	500	0.131	0.836	0.971	0.378	0.849	0.977	0.695	0.687	0.946
					Common components					
50	100	0.674	0.448	0.281	0.677	0.522	0.363	0.676	0.661	0.561
	200	0.638	0.376	0.213	0.644	0.430	0.271	0.640	0.598	0.479
100	100	0.540	0.366	0.260	0.544	0.435	0.336	0.533	0.584	0.512
	500	0.457	0.251	0.136	0.450	0.279	0.175	0.396	0.418	0.334

4.1 In-sample estimation

In this section, I first show in-sample numerical experiment results to compare the proposed FPPC with the conventional PC and PPC methods. The factor loadings and common factors using each method are estimated. For each estimator, the canonical correlation between the estimators and parameters can be regarded as a measurement of the estimation accuracy because the factors and loading may be estimated up to a rotation matrix (e.g., Bai and Liao, 2016). The simulation is replicated 1000 times for each scenario. Table 2 shows the sample mean of the smallest canonical correlations for several competing methods. In addition, the averaged mean squared errors (MSE) of estimated common components, $(\frac{1}{NT}\sum_{i,t}(\widehat{\lambda}_i'\widehat{f}_t - \lambda_i' f_t)^2)^{1/2}$, are also compared.

The results in Table 2 are summarized as follows. The estimation accuracy increases when the dimensionality is increased. For loadings, FPPC performs better than PPC and PC except for the weak explanatory power cases with larger dimensionality. When $w = 0.1$, PPC and FPPC do not perform well because the observed $\mathbf{X}_t$ is not as informative. On the other hand, FPPC always outperforms PPC and PC in terms of common factors. In addition, FPPC provides the smallest MSE of the common components.

Table 3: Out-of-sample relative mean squared forecast error (with PC as the benchmark): the smaller the better.

		Strong($w = 10$)			Mild($w = 1$)			Weak($w = 0.1$)		
N	T	PC	PPC	FPPC	PC	PPC	FPPC	PC	PPC	FPPC
50	100	1.000	0.952	0.861	1.000	0.963	0.863	1.000	1.021	0.896
	200	1.000	0.947	0.866	1.000	0.948	0.853	1.000	1.001	0.874
100	100	1.000	0.964	0.892	1.000	0.977	0.884	1.000	1.035	0.903
	500	1.000	0.961	0.891	1.000	0.966	0.898	1.000	1.002	0.909

4.2 Out-of-sample forecast

This section illustrates the improvement of out-of-sample forecasts based on the proposed factor estimators. Consider a linear forecasting model as follows:

$$z_{t+1} = \alpha' \mathbf{F}_t + \beta' W_t + \epsilon_{t+1},$$

where $W_t = 1$, $\beta = 1$, and ϵ_{t+1} is drawn from i.i.d. $\mathcal{N}(0, 1)$. To cover a variety of model settings, unknown coefficients, α, are generated from Uniform$(0.5, 1)$ for each simulation. Here the unknown factor can be learned from a factor model: $\mathbf{y}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{u}_t$. The data generating process for the factor model is the same as Section 4.1.

I conducted one-step ahead out-of-sample forecasting 50 times using rolling data windows. The moving window size is fixed as T, and it is also the sample size for estimations. In each simulation, the total $T+50$ observations are generated. To forecast z_{T+m+1} for $m = 0, \cdots, 49$, the observations from $m + 1$ to $m + T$ are used. Specifically, the factors are estimated by PC, PPC, and FPPC methods and are denoted by $\{\widehat{\mathbf{F}}_{m+1}, \cdots, \widehat{\mathbf{F}}_{m+T}\}$. Then, $\widehat{\alpha}$ and $\widehat{\beta}$ are obtained by regressing $\{z_{m+2}, \cdots, z_{m+T}\}$ on $\{(\widehat{\mathbf{F}}'_{m+1}, W_{m+1})', \cdots, (\widehat{\mathbf{F}}'_{m+T-1}, W_{m+T-1})'\}$. Finally, forecasts are $\widehat{z}_{T+m+1|T+m} = \widehat{\alpha}'\widehat{\mathbf{F}}_{m+T} + \widehat{\beta} W_{m+T}$. This procedure continues for $m = 0, \cdots, 49$.

The mean squared forecasting errors (MSFE) are compared based on the PC, PPC, and FPPC estimates of the factor space. I use PC as a benchmark and report the relative mean squared forecasting errors (RMSFE):

$$\text{RMSFE} = \frac{\sum_{m=0}^{49} (z_{T+m+1} - \widehat{z}_{T+m+1|T+m})^2}{\sum_{m=0}^{49} (z_{T+m+1} - \widehat{z}^{PC}_{T+m+1|T+m})^2},$$

where $\widehat{z}_{T+m+1|T+m}$ is the forecast z_{T+m+1} based on FPPC or PPC. For each case, the averaged RMSFE are calculated as measurements of the forecasting performance based on 1000 replications.

The results are presented in Table 3. Because both cross-sectional correlations and observed characteristics are taken into account, FPPC performs significantly better than PPC and PC. This implies that more accurate and efficient estimations of the factors also result in better forecasting performances. Note that, when $\mathbf{X}_t$ has weak explanatory power of the common factors, PPC and FPPC are slightly worse compare to the cases of strong or mild explanatory powers. This phenomenon corresponds to the results of Section 4.1.

5 Application: US Bond Risk Premia

As an empirical study, I forecast the excess return of U.S. government bonds using the proposed FPPC-based DI model. Ludvigson and Ng (2009) emphasized that common factors, extracted from a large number of economic time series using the conventional PC method, have an important forecasting power in addition to the predictive information in the yield curve studied by Cochrane and Piazzesi (2005). In this section, I shall explore how the newly proposed FPPC method performs in forecasting the excess bond returns. In addition, I empirically assess the predictive accuracy of a large group of models including linear models and other (nonlinear) machine learning models.

As in Ludvigson and Ng (2009) and Cochrane and Piazzesi (2005), the following definitions and notations are used. The bond excess return is defined as the one-year bond return in excess of the risk-free rate. Specifically, let $p_t^{(n)}$ denote the log price of n-year discount bond at time t, and then the log yield is $y_t^{(n)} = -(1/n)p_t^{(n)}$. The log forward rates are defined as $f_t^{(n)} = p_t^{(n-1)} - p_t^{(n)}$. I define $r_{t+12}^{(n)} = p_{t+12}^{(n-1)} - p_t^{(n)}$ as the log 1 year holding period return from buying an n-year bond at time t and selling it as an $n-1$ year bond at time $t+12$. Then the excess return with maturity of n-years is defined by

$$rx_{t+12}^{(n)} = r_{t+12}^{(n)} - y_t^{(1)}, \text{ for } t = 1, ..., T,$$

where $y_t^{(1)}$ is the log yield on the one-year bond.

5.1 Data

I analyze monthly bond return data spanning from the period 1964:1–2016:4 ($T = 628$), which is an updated version of the Ludvigson and Ng (2009) dataset. The bond return data are obtained from the Fama-Bliss dataset from the Center for Research in Securities Prices (CRSP), which contains observations from one-year to five-year zero-coupon bond prices. These are used to calculate excess bond returns, yields, and forward rates, as discussed above.

The factors are estimated by using several principal component methods (i.e., PC, PPC, and FPPC) from a monthly balanced panel of disaggregated 130 macroeconomic time series.

Table 4: Components of $\mathbf{X}_t$ for U.S. bonds excess return forecasting.

	Series
$X_{1,t}$	Linear combination of five forward rates (CP)
$X_{2,t}$	Real gross domestic product (GDP)
$X_{3,t}$	Consumption price index (CPI) - Inflation
$X_{4,t}$	Non-agriculture employment

A specific description and transformation code of panel data is provided in McCracken and Ng (2016).[7] The series are sorted by broad categories of macroeconomic series: real output and income, employment and hours, real retail, manufacturing and trade sales, consumer spending, housing starts, inventories and inventory sales ratios, orders and unfilled orders, compensation and labor costs, capacity utilization measures, price indexes, bond and stock market indexes, and foreign exchange measures. This set of variables has been widely used in the literature such as Stock and Watson (2002a), Bai and Ng (2008a), and Kim and Swanson (2014), among many others.

Finally, the observed characteristics $\mathbf{X}_t$ are required to employ the PPC or FPPC methods. As for the characteristics, I choose a single forward factor (CP) suggested by Cochrane and Piazzesi (2005) and three aggregated macroeconomic series.[8] These aggregate series are widely used to describe the co-movement of the macroeconomic activities, as studied by Stock and Watson (2014) and NBER (2008). A detailed description of these series is listed in Table 4. In addition, these data are also transformed and standardized.

5.2 Experiment setup and forecast evaluation

This paper considers a variety of estimation techniques including simple linear models (AR and DI) as well as various (linear and nonlinear) machine learning methods such as penalized regression (e.g., lasso, ridge, elastic net), regression trees (e.g., decision tree, gradient boosting, random forest), neural networks (e.g., hybrid neural network, factor augmented neural network). The modified diffusion index models using statistical learning algorithms (e.g., bagging, boosting, factor-lasso) are also considered. Table 5 lists all forecasting models in the experiments. To avoid bogging down the reader with details of all methods, models and specific implementation choices are described in the supplementary material.

All forecasting models are estimated using either rolling or recursive estimation windows, and all models and parameters are reestimated at each point in time, prior to the construction

[7] The macroeconomic dataset is the FRED-MD monthly database. As of May 2016, FRED-MD removed some variables (e.g., NAPMPI, NAPMEI, NAPM, etc.). Hence, I obtained the dataset up to April 2016 to use the same variables as the Ludvigson and Ng (2009) dataset.

[8] Note that I interpolate gross domestic product, which is reported quarterly, to a monthly frequency following Chow and Lin (1971).

Table 5: List of all forecasting models

	Method	Description
Benchmark	AR(SIC)	Autoregressive model with lags selected by the SIC
Diffusion Index Models	PCR	Principal components regression
	FAAR	Factor augmented autoregressive model
	DI	Diffusion index regression model with CP and factors
	DI2	Diffusion index regression model with CP, lags, and factors
Modified DI Models	Bagging	Bagging with shrinkage, $c = 1.96$
	Boosting	Component boosting, $M = 50$
	Fac-Lasso	Factor-Lasso regression
Penalized Linear Regressions	Lasso	Lasso regression
	Ridge	Ridge regression
	EN	Elastic net regression
Regression Trees	DT	Decision tree regression
	G-Bst	Gradient boosting regression
	RanForest	Random forest regression
Neural Networks	NN1	Neural network with one hidden layer
	NN2	Neural network with two hidden layers
	NN3	Neural network with three hidden layers
	H-NN1	Hybrid neural network with one hidden layer
	H-NN2	Hybrid neural network with two hidden layers
	H-NN3	Hybrid neural network with three hidden layers
	FANN1	Factor augmented neural network with three hidden units
	FANN2	Factor augmented neural network with five hidden units
	FANN3	Factor augmented neural network with seven hidden units

of each new forecast. In the rolling estimation scheme, three different window sizes are examined (i.e., 180, 240, and 300 months). The recursive estimation scheme begins with the same in-sample period, but a new observation is added to the sample in each period. I denote B as the number of ex-ante forecasts, and Q is the length of the rolling window or the initial length of the recursive window.

To evaluate the forecasting performance of various models, I utilize two statistics including

1. mean square forecast error (MSFE), defined as

$$\text{MSFE} = \frac{1}{B}\sum_{t=Q}^{T-12}(rx_{t+12}^{(n)} - \widehat{rx}_{t+12|t}^{(n)})^2,$$

2. and out-of-sample R^2 suggested by Campbell and Thompson (2007), defined as

$$\text{Out-of-sample } R^2 = 1 - \frac{\sum_{t=Q}^{T-12}(rx_{t+12}^{(n)} - \widehat{rx}_{t+12|t}^{(n)})^2}{\sum_{t=Q}^{T-12}(rx_{t+12}^{(n)} - \overline{rx}_{t+12}^{(n)})^2},$$

   where, for each maturity n, $\widehat{rx}_{t+12|t}^{(n)}$ is the forecast of bond excess returns using each model and $\overline{rx}_{t+12}^{(n)}$ is the historical average of bond excess return.

Note that the out-of-sample R^2 values can be negative, indicating that the forecasting performance of the particular model is even worse than the historical averages. However, squared error loss measures such as MSFE may yield misleading decision-making by forecasts in terms of profit measure. Therefore, I use the predictive accuracy test of Diebold and Mariano (1995), called the DM test, for forecast performance evaluations. The DM test has a null hypothesis that the two models being compared have equal predictive accuracy, and its statistic has asymptotic $N(0,1)$ limiting distribution. The null hypothesis of equal predictive accuracy of two forecasting models is

$$H_0 : E[l(\epsilon_{1,t+12|t})] - E[l(\epsilon_{2,t+12|t})] = 0,$$

where $\epsilon_{i,t+12|t}$ is the prediction error of i-th model for $i = 1, 2$ and $l(\cdot)$ is the quadratic loss function. Here, we assume that parameter estimation error vanishes as $T, B, Q \to \infty$ and that each pair of two models is nonnested. The actual DM test statistic is followed by: $S_{DM} = \bar{d}/\widehat{\sigma}_{\bar{d}}$, where $\bar{d} = \frac{1}{B}\sum_{t=1}^{B}(\widehat{\epsilon}_{1,t+12|t}^2 - \widehat{\epsilon}_{2,t+12|t}^2)$, $\widehat{\sigma}_{\bar{d}}$ is a heteroskedasticity and autocorrelation robust estimator of the standard deviation of $\bar{d}$. Here $\widehat{\epsilon}_{1,t+12|t}$ and $\widehat{\epsilon}_{2,t+12|t}$ denote the forecast error estimates using Model 1 and Model 2, respectively. Thus, a negative and significant value of S_{DM} indicates that Model 1 outperforms Model 2. However, the DM testing framework cannot be used for the comparisons between nested models. Therefore, I also constructed conditional predictive ability (GW) test statistics suggested by Giacomini and White (2006) for pairwise model comparisons of all models.

5.3 Empirical findings

I conduct a one-year-ahead out-of-sample forecasting investigation using various forecasting techniques. Forecasts are constructed based on both rolling and recursive estimation windows for out-of-sample forecast periods from January 1984 to April 2016. Here I set the length of the window size as $Q = 240$, and other out-of-sample results of different window sizes ($Q = 180$ and 300) are available upon request from the author. First, the forecasting performance of several principal component methods based on diffusion index models are compared. Then, I explore the forecasting performance of all models outlined in Table 5. In addition, the in-sample analysis and the replication of experiments using different forecasting periods are considered in the supplementary material.

5.3.1 Forecasting power of FPPC in diffusion index models

I first focus on linear forecasting models: $rx_{t+12}^{(n)} = \alpha + \delta' L_t + \epsilon_{t+12}$, where L_t contains the unobserved factors. This model is employed to compare the out-of-sample forecasting performance of different principal component methods, such as PC, PPC, and FPPC, for excess bond returns. I set the number of factors $K = 8$ for all methods, which is determined by the information criteria suggested in Bai and Ng (2002). In addition, the additive polynomial basis with $J = 5$ is used for the sieve basis of PPC and FPPC. The threshold constant of FPPC is chosen by cross-validation as introduced in Section 2.3.2 for each estimation period.

For each fixed and recursive data window ended at time t, the factors are estimated from the panel data of 130 macroeconomic series and four characteristics via different PC methods. Table 6 presents results of the out-of-sample R^2 in both rolling and recursive window cases. The first half columns in the table are results of the principal component regression (PCR) model (i.e., $L_t = F_t$), while the second half columns are results of the DI model (i.e., $L_t = (F_t', CP_t)'$). Here, PC$_{134}$ denotes the regular PC method using four additional characteristics in addition to 130 macro variables. Note that the results of relative MSFE are referred to the supplementary material.

First, FPPC results in notable improvements in out-of-sample predictive accuracy based on Table 6. For instance, in the DI model using the rolling window scheme, I find that FPPC generates an approximately 6.2-15.8% increase in out-of-sample R^2, when compared to the benchmark method (i.e., PC$_{130}$). Additionally, rolling window forecasts outperform recursive window forecasts in both models. Interestingly, in the PCR model using the rolling window scheme, FPPC greatly improves the forecasting power compared to other methods. Moreover, Table 7 shows the out-of-sample R^2 of the DI model using rolling window separately for the

Table 6: Out-of-sample R^2 (%) of U.S. bonds excess return forecasting: the larger the better. Forecasting sample period: 1984:1-2016:4.

	PCR ($L_t = F_t$)				DI ($L_t = (F_t', CP_t)'$)			
Maturity	PC_{130}	PC_{134}	PPC	FPPC	PC_{130}	PC_{134}	PPC	FPPC
	Rolling estimation window							
2 year	0.9	3.0	15.9	22.0	36.9	36.6	39.4	39.2
3 year	5.5	7.3	16.6	22.3	35.2	35.2	38.7	39.0
4 year	7.0	9.4	19.0	25.1	36.6	36.9	41.5	42.5
5 year	9.7	11.9	20.6	26.3	35.2	35.4	39.2	40.3
	Recursive estimation window							
2 year	-2.1	1.8	2.6	5.0	27.0	27.3	34.6	34.1
3 year	3.8	7.3	4.7	6.7	27.0	27.7	34.8	34.9
4 year	6.6	10.1	7.1	9.3	29.6	30.7	37.2	37.8
5 year	9.6	12.7	9.2	11.5	29.4	30.5	35.1	36.0

Table 7: Out-of-sample R^2(%) of U.S. bonds excess return forecasting in recessions and expansions: the larger the better. Rolling window.

	DI ($\widehat{L}_t = (\widehat{\mathbf{F}}_t', CP_t)'$)							
Maturity	PC_{130}	PC_{134}	PPC	FPPC	PC_{130}	PC_{134}	PPC	FPPC
	Recessions				Expansions			
2 year	6.6	7.7	11.8	11.7	41.1	40.7	43.2	42.9
3 year	-4.3	-2.3	8.2	11.6	40.4	40.2	42.8	42.8
4 year	-10.1	-7.4	6.0	11.8	41.8	41.8	45.4	46.0
5 year	-9.8	-7.5	5.7	14.2	39.7	39.7	42.5	43.1

recession and expansion sub-samples as defined by the NBER recession index. By using characteristics, FPPC and PPC remarkably outperform PC in recessions.

Second, the DM test results are provided in Table 8. I compare FPPC to other PC methods, assuming that each pair of models being compared is nonnested. A negative and significant DM statistic indicates that FPPC outperforms the other method in out-of-sample forecasts. For the PCR model, FPPC provides significantly better forecasts at 1% and 5% levels compared to PCs and PPC. For the DI model, FPPC is not statistically significant compared to other methods. This implies that the added forward factor of Cochrane and Piazzesi (2005) also explains a significant variation in excess bond returns. Nevertheless, signs of DM test statistics are mostly negative.

Overall, FPPC outperforms other PC methods based on linear models in terms of out-of-sample forecasting, because (i) the information of characteristics has explanatory power on the

Table 8: Diebold-Mariano test statistics. Forecast sample period: 1984:1-2016:4. Rolling window.

Model	Maturity			
	2 year	3 year	4 year	5 year
PCR ($L_t = F_t$)				
FPPC versus PC_{130}	-2.542**	-2.295**	-2.533**	-2.597***
FPPC versus PC_{134}	-2.521**	-2.254**	-2.391**	-2.511**
FPPC versus PPC	-2.442**	-2.405**	-2.640***	-2.581***
DI ($L_t = (F_t', CP_t)'$)				
FPPC versus PC_{130}	-0.604	-1.020	-1.619	-1.396
FPPC versus PC_{134}	-0.698	-1.061	-1.620	-1.404
FPPC versus PPC	0.130	-0.157	-0.610	-0.698

Note: ***, **, and * denote significance at 1, 5, and 10% levels respectively.

latent factors, and (ii) the cross-sectional heteroskedasticity and dependence are considered.

5.3.2 Forecasting performance

I now investigate the one-year-ahead forecasting performances of linear and nonlinear machine learning models listed in Table 5. For factor-augmented models (i.e., PCR, FAAR, DI, Bagging, Boosting, and Fac-Lasso), all PC, PPC, and FPPC methods are conducted. Because FPPC outperforms others in most cases, I present the results of factor-augmented models based on the FPPC method.

Table 9 reports the relative MSFEs of the rolling window and recursive window forecasts. I set the AR(SIC) model as a benchmark to generate relative MSFEs for other models. Note that, out-of-sample R^2 of all forecasting models are tabulated in the online suppement. Table 10 reports the pairwise test statistics of conditional predictive ability (GW) proposed by Giacomini and White (2006) for 2-year maturity based on the rolling window scheme.[9] I describe the main empirical findings in the following aspects.

First, the FPPC-based diffusion index (DI) models "win" over most machine learning models, or are comparable with the best nonlinear machine learning models (e.g., RanForest and FANN). For instance, in Table 9, DI models using the rolling window scheme generates an approximately 35% decrease in MSFEs when compared to the benchmark AR(SIC) model for each maturity. Additionally, GW test results confirm that DI models based on the proposed FPPC method exhibit one of the best forecasting performances among all models considered in this experiment. Less importantly, the modified diffusion index models (e.g.,

[9] The GW test statistics of 3-5 year maturities are listed in the supplementary material. The results are similar to those of 2-year maturity case.

Bagging, Boosting, and Fac-Lasso) also perform well, but these are not statistically significant compared to DI models. Moreover, the FPPC-based PCR model outperforms some of the machine learning models, including the conventional neural network and penalized linear models.

Second, rolling window forecasts outperform recursive window forecasts for most of the models based on MSFE and out-of-sample R^2 values. This implies that the proper in-sample size yields better forecasting performance than the redundant sample size, especially for PCR, FAAR, and penalized regression models. In the forecasting literature, however, many practitioners only conduct the recursive window scheme, which may lead to misleading empirical findings.

Lastly, RanForest also performs very well among all employed models in this paper. However, the inspection indicates that both RanForest and FPPC-based DI models have similar MSFE and out-of-sample R^2 values. Based on GW test statistics, there is no statistically significant difference between these models for all maturities. In addition, among several types of neural networks, FANN, which only uses a few estimated factors as predictors, outperforms NN and H-NN. This implies that a large number of predictors may yield the general overfitting problem.

In summary, FPPC-based DI models, RanForest, and FANN are the best performing models based on MSFE, out-of-sample R^2, and GW test statistic values. Importantly, I find that there is no guarantee that nonlinear machine learning will yield superior forecasting performance compared to the DI linear models.

6 Conclusions

This paper examines a high-dimensional factor model that latent factors depend on a few observed covariate variables. This model is motivated by the fact that observed variables can partially explain the latent factors. The FPPC method estimates the unknown factors and loadings efficiently by taking into account cross-sectional heteroskedasticity and correlations of the error terms. Also, the projection approach using the characteristics can improve estimation accuracy. I apply the proposed estimation method to the diffusion index forecasts. The rates of convergence of factors, factor loadings, and forecast errors are considered. My empirical evidence shows that the proposed method using aggregated macroeconomic variables as characteristics yields a substantial gain of forecasting bond risk premia. Moreover, I find that the proposed linear forecasting model performs well compared to other nonlinear machine learning models in terms of out-of-sample forecasting.

Table 9: Relative mean squared forecast errors of U.S. bonds excess return forecasting: the smaller the better. Forecasting sample period: 1984:1-2016:4.

	Rolling				Recursive			
Method	2 year	3 year	4 year	5 year	2 year	3 year	4 year	5 year
AR(SIC)	2.122	7.968	16.622	26.166	2.225	8.390	17.767	28.047
PCR	0.898	0.858*	0.824**	0.800***	1.042	0.978	0.934	0.897
FAAR	0.920	0.859	0.809**	0.777***	1.044	0.948	0.887	0.841*
DI	0.699***	0.673***	0.633***	0.649***	**0.723*****	**0.683*****	0.640***	0.649***
DI2	0.696***	**0.642*****	**0.601*****	**0.622*****	0.734***	**0.676*****	**0.631*****	**0.639*****
Bagging	0.685***	0.649***	**0.609*****	0.658***	0.752***	0.704***	0.656***	0.670***
Boosting	**0.653*****	**0.640*****	0.610***	0.655***	**0.730*****	0.708***	0.676***	0.694***
Fac-Lasso	**0.660*****	**0.635*****	**0.598*****	**0.618*****	0.739***	0.687***	**0.636*****	**0.638*****
Lasso	0.986	0.894***	0.851***	0.830***	0.982	0.881***	0.834***	0.811***
Ridge	0.975	0.955	0.883	0.884	1.136	1.054	0.950	0.946
EN	0.991	0.922***	0.871***	0.847***	1.001	0.932***	0.873***	0.843***
DT	0.945	1.059	0.969	1.057	1.141	1.075	0.858	0.882
G-Bst	0.894**	0.890**	0.884***	0.870***	0.922*	0.889***	0.874***	0.847***
RanForest	**0.647*****	0.660***	0.663***	**0.644*****	**0.652*****	**0.647*****	**0.629*****	**0.617*****
NN1	0.942	0.905*	0.853***	0.846***	0.897**	0.860***	0.822***	0.792***
NN2	1.092	1.017	1.021	1.019	1.131	1.071	1.057	1.026
NN3	1.084	1.020	1.023	1.000	1.118	1.084	1.063	1.040
H-NN1	0.749***	0.706***	0.696***	0.749***	0.804**	0.799**	0.772**	0.776**
H-NN2	0.810*	0.792**	0.759**	0.766**	1.003	0.989	0.969	0.977
H-NN3	0.797**	0.795**	0.774**	0.765**	0.987	0.989	0.973	0.978
FANN1	0.691***	0.673***	0.646***	0.677***	0.767**	0.728***	0.691***	0.712***
FANN2	0.690***	0.658***	0.643***	0.653***	0.767**	0.728***	0.682***	0.716***
FANN3	0.699***	0.670***	0.650***	0.650***	0.773**	0.730***	0.681***	0.698***

Note: ***, **, and * denote significance at 1, 5, and 10% levels based on the predictive accuracy test of Diebold and Mariano (1995), respectively. Entries in bold denote point MSFE "best three" forecasting models for a given maturity.

Table 10: Pairwise model comparison using Giacomini-White tests for 2-year maturity.

	PCR	FAAR	DI	DI2	Bagging	Boosting	Fac-Lasso	Lasso	Ridge	EN	DT	G-Bst	RanForest	NN	H-NN	FANN
AR(SIC)	0.93	0.68	4.43	**5.66**	**5.23**	**6.62**	**5.40**	2.68	-0.25	2.12	1.71	**4.80**	**9.56**	1.61	2.90	3.22
PCR		0.50	**5.69**	**9.77**	**7.44**	4.24	**5.15**	-1.76	-0.74	-1.50	0.22	-0.94	**6.14**	-1.79	3.02	3.06
FAAR			3.94	**7.45**	**5.03**	4.40	4.30	-0.95	-1.04	-0.88	0.22	-0.51	**5.00**	-1.80	2.59	2.62
DI				2.84	**6.33**	0.75	0.62	**-4.88**	**-13.00**	-4.47	-2.96	-2.71	0.81	-3.51	-0.78	-0.65
DI2					-0.12	1.66	0.14	**-5.81**	**-13.14**	**-5.39**	-2.69	-3.51	0.31	**-5.04**	-1.27	-1.58
Bagging						1.36	0.27	**-5.64**	**-17.72**	**-5.16**	-2.52	-3.46	0.53	-4.53	-1.34	-0.66
Bsting							-0.78	**-7.06**	**-14.07**	**-6.49**	**-5.51**	-4.59	0.25	**-4.89**	**-7.79**	**-6.08**
Fac-Lasso								**-6.00**	**-15.80**	**-5.54**	**-5.91**	-3.96	0.52	**-4.69**	-1.56	-0.53
Lasso									-0.58	1.58	1.86	2.62	**9.87**	**5.02**	3.33	3.61
Ridge										0.55	3.47	0.65	**8.10**	0.80	3.40	**7.50**
EN											1.86	3.05	**9.17**	4.60	3.15	3.37
DT												-0.50	**12.62**	-0.37	1.09	2.06
G-Bst													**10.26**	-2.71	1.78	2.04
RanForest														**-10.01**	-4.44	-3.37
NN															1.32	1.73
H-NN																1.65

Note: This table reports pairwise Giacomini and White (2006) test statistics comparing the out-of-sample bond excess returns among seventeen models. I ustilized the absolute error loss function. Positive numbers indicate the column model outperforms the row model, while negative numbers indicate the row model outperforms the column model. Bold font indicates the difference is significant at the 10% level or better for individual tests. Note that NN, H-NN, and FANN denote that the best performing models among different numbers of hidden layers or hidden units.

References

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

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

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

——— (2017): “Inferences in panel data with interactive effects using large covariance matrices,” *Journal of econometrics*, 200, 59–78.

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

——— (2006): “Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions,” *Econometrica*, 74, 1133–1150.

——— (2008a): “Forecasting economic time series using targeted predictors,” *Journal of Econometrics*, 146, 304–317.

——— (2008b): “Large Dimensional Factor Analysis,” *Foundations and Trends in Econometrics*, 3, 89–163.

——— (2013): “Principal components estimation and identification of static factors,” *Journal of Econometrics*, 176, 18–29.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005): “Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach,” *The Quarterly journal of economics*, 120, 387–422.

Bianchi, D., M. Büchner, and A. Tamoni (2021): “Bond risk premiums with machine learning,” *The Review of Financial Studies*, 34, 1046–1089.

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

Cai, T. T. and H. H. Zhou (2012): “Optimal rates of convergence for sparse covariance matrix estimation,” *The Annals of Statistics*, 40, 2389–2420.

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

Campbell, J. Y. and S. B. Thompson (2007): "Predicting excess stock returns out of sample: Can anything beat the historical average?" *The Review of Financial Studies*, 21, 1509–1531.

Chamberlain, G. and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Choi, I. (2012): "Efficient estimation of factor models," *Econometric Theory*, 28, 274–308.

Chow, G. and A.-l. Lin (1971): "Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series," *The Review of Economics and Statistics*, 53, 372–75.

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

Connor, G. and O. Linton (2007): "Semiparametric estimation of a characteristic-based factor model of common stock returns," *Journal of Empirical Finance*, 14, 694–717.

Diebold, F. and R. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Fama, E. F. and R. R. Bliss (1987): "The information in long-maturity forward rates," *The American Economic Review*, 680–692.

Fan, J., Y. Ke, and Y. Liao (2020): "Augmented factor models with applications to validating market risk factors and forecasting bond risk premia," *Journal of Econometrics.*

Fan, J., Y. Liao, and M. Mincheva (2011): "High dimensional covariance matrix estimation in approximate factor models," *Annals of statistics*, 39, 3320.

——— (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 603–680.

Fan, J., Y. Liao, and W. Wang (2016): "Projected principal component analysis in factor models," *Annals of statistics*, 44, 219.

Giacomini, R. and H. White (2006): "Tests of conditional predictive ability," *Econometrica*, 74, 1545–1578.

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

LAM, C. AND Q. YAO (2012): "Factor modeling for high-dimensional time series: inference for the number of factors," *The Annals of Statistics*, 40, 694–726.

LUDVIGSON, S. C. AND S. NG (2009): "Macro factors in bond risk premia," *The Review of Financial Studies*, 22, 5027–5067.

MCCRACKEN, M. W. AND S. NG (2016): "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 34, 574–589.

NBER (2008): "Determination of the December 2007 peak in economic activity," .

ROTHMAN, A. J., P. J. BICKEL, E. LEVINA, AND J. ZHU (2008): "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, 2, 494–515.

STOCK, J. H. AND M. W. WATSON (2002a): "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, 97, 1167–1179.

——— (2002b): "Macroeconomic forecasting using diffusion indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

——— (2014): "Estimating turning points using large data sets," *Journal of Econometrics*, 178, 368–381.