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

Sung Hoon Choi[*]

September 14, 2020

⟨Job Market Paper - Link to the latest version⟩

Abstract

This paper considers factor models in which observed characteristics partially explain the latent factors. I propose a feasible weighted projected principal component (FPPC) analysis, which relies on a high-dimensional weight matrix. By using a consistent estimator of the inverse error covariance matrix as the weight matrix, we can take into account both cross-sectional dependence and heteroskedasticity. The rates of convergence for the FPPC estimators are much faster than those from the conventional principal component analysis. Moreover, I suggest 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 simulations and an empirical study with U.S. bond market data, I demonstrate that the proposed model outperforms benchmark models based on other principal component estimators. A substantial gain in predictive accuracy is achieved by (i) incorporating the characteristics and (ii) considering cross-sectional dependence and heteroskedasticity. Specifically, I forecast excess bond returns and find the proposed model performs well among a large group of machine learning techniques such as lasso, neural networks, and random forests.

Keywords: High-dimensionality, Unknown factors, Conditional sparsity, Thresholding, Cross-sectional correlation, Heteroskedasticity, Optimal weight matrix, Diffusion index, Forecasting, Machine learning.

[*] Department of Economics, Rutgers University, 75 Hamilton St,, New Brunswick, NJ 08901, USA. E-mail: shchoi@economics.rutgers.edu. Website: https://sites.google.com/view/sung-hoon-choi. I am grateful to my advisor, Yuan Liao, for his advice and encouragement. I also thank Jushan Bai, Norman Swanson, John Landon-Lane, Hyunhak Kim, and Jungjun Choi for their valuable comments and suggestions.

Contents

1 Introduction

Accurately and efficiently estimating latent factors is crucial 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 with large datasets. This paper provides a new factor estimation methodology for big data forecasting, which accomplishes two tasks: (i) it develops new statistical theories and methodologies for learning factors and for a factor-augmented linear regression model, and (ii) it demonstrates improved forecasting accuracy using excess returns on U.S. government bonds. In this paper, I demonstrate a novel theoretical econometric framework as follows.

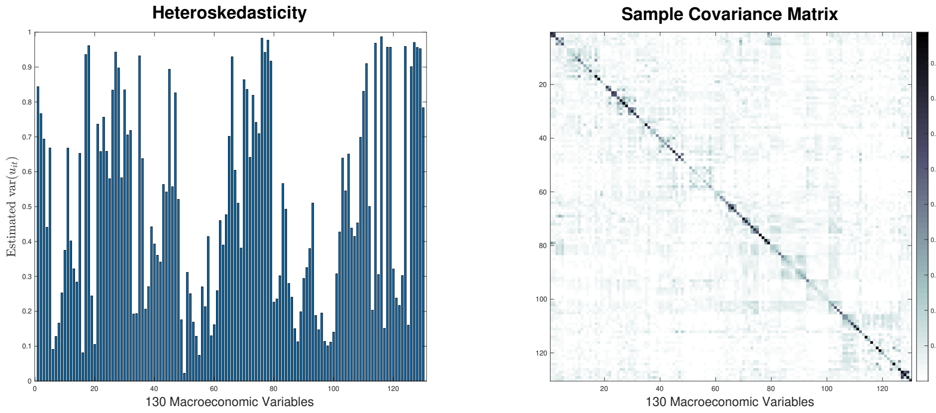
First, it is essential to consider a large error covariance matrix estimator for efficient estimation. In linear regression models, for example, it is well known that the generalized least squares (GLS) estimators are more efficient than the ordinary least squares (OLS) estimators in the presence of cross-sectional heteroskedasticity. Factor models often require the idiosyncratic error components to be cross-sectionally heteroskedastic and correlated. Intuitively, the large error covariance matrix, $\mathbf{\Sigma}_u = \text{cov}(u_t)$, is a non-diagonal matrix, and the diagonal entries may vary widely (e.g., Bai and Liao, 2016). For example, individual companies' idiosyncratic error variance can be remarkably different, and the errors can be correlated among companies. In addition, Figure 1 shows that cross-sectional heteroskedasticity and correlations exist in practice and implies that it is critical to consider estimating the error covariance matrix to take them into account. However, the conventional principal component (PC) method does not require estimating $\mathbf{\Sigma}_u$, and it essentially treats u_{it} to be homoskedastic and uncorrelated over i. Hence, it is inefficient.[1] In this paper, I consider consistently estimating the high-dimensional error covariance matrix and its inverse. Therefore, by using the estimator for $\mathbf{\Sigma}_u^{-1}$ 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 characteristics. In macroeconomic or financial applications, 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 noise components, it helps us estimate the factors more accurately than the conventional PC method even during the financial crisis.

To incorporate these two key aspects, I introduce a feasible weighted projected principal

[1] Choi (2012) studied efficient estimations using weighted least squares in the conventional factor model, but assumed $\mathbf{\Sigma}_u$ to be known.

Figure 1: Cross-sectional heteroskedasticity and correlations



Note: The first figure shows the estimated error variance, var(u_{it}), for each i using the estimated residuals by the regular PC method. The second figure displays an image of the sample covariance matrix with scaled colors.

component (FPPC) analysis. 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. This procedure substantially improves estimation accuracy and efficiency. In this paper, I carefully study the asymptotic properties of the proposed estimators. In particular, when both a cross-sectional dimension N and a time dimension T grow simultaneously, the rates of convergence of the FPPC estimators are faster than those of the regular PC estimators.

Next, I suggest the FPPC-based diffusion index model. In the literature, the most popular application of the factor model is factor-augmented regression. 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. Note there is extensive 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. Conversely, more accurate and efficient estimations of the factors can substantially improve out-of-sample forecasts (see Bai and Liao, 2016). Therefore, this paper addresses how the FPPC method can improve predictive accuracy in the DI model.

Turning to the empirical financial-economic literature, the determinants of bond risk premia are crucial for both policymakers and investors (e.g., Campbell and Shiller, 1991; Cochrane and Piazzesi, 2005; Fama and Bliss, 1987). 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 FPPC method instead of the PC method, we can gain predictive accuracy on excess bond returns. In a recent paper, Bianchi et al. (2019) studied how machine learning methods (such as regression trees and neural networks) provide strong statistical evidence in predicting excess bond 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 forecasting. In the financial economics literature using machine learning, however, some practitioners do not take them into account. These important issues motivate the following empirical study.

I investigate the macroeconomic factors used in estimating bond risk premia to answer the following key questions. Is there a substantial gain in the accuracy of predictions of excess bond returns using the FPPC method? Do other (nonlinear) machine learning methods outperform linear models? To answer these questions, in this paper I compare the proposed FPPC method and the conventional PC and PPC methods in forecasting regression models using inflation, employment, forward factor, and GDP as characteristics. I also compare and evaluate the forecasting performance of the proposed FPPC-based diffusion index model and various machine learning models including penalized regressions (e.g., lasso, ridge, elastic net), regression trees (e.g., gradient boosting, random forests), and neural networks. The experimental findings are based on the construction of one-year-ahead predictions for the full 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 forecast. These results are robust to different forecasting periods except for the period including the global financial crisis. 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.

This paper contributes to the literature in the following three ways. First, I develop a new methodology to estimate the parameters of semiparametric factor models using a consistent estimator of the error covariance matrix. By using the consistent error covariance matrix, the FPPC is more efficient than most of the existing methods under cross-sectional heteroskedasticity and correlations. Moreover, projection using characteristics that remove noise components gives accurate estimators. To the best of my knowledge, this paper is the first to integrate them in the literature. A large literature addresses the use of PC method to investigate the (static) factor models, such as Chamberlain and Rothschild (1983), Stock and Watson (2002a), Bai (2003), and Lam and Yao (2012), among others.[2] In addition, there are researchers who studied the semiparametric factor models (e.g., Connor and Linton, 2007; Fan et al., 2016). However, they did not consider estimating the error covariance matrix, so their estimators of factors and loadings are inefficient. Second, this paper includes a study of the DI model based on the FPPC method. I provide asymptotic distributions of coefficient estimators for the forecasting model. The convergence rate of forecast error is also obtained. Simulation studies show the proposed model outperforms other PC-based DI models, given the advantages of the FPPC method over the regular PC and PPC methods. Last, this paper contributes to the literature on bond return forecastability by showing that the proposed method provides better predictive power for excess bond returns than other PC methods (Ludvigson and Ng, 2009). I also compare the out-of-sample forecasting performance of the proposed model with other commonly used machine learning methods and find the FPPC-based DI performs very well among all models. The gain of forecasts using my method is substantial because it takes into account (i) the characteristics of the latent factors and (ii) cross-sectional correlation and heteroskedasticity.

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, $\|A\| = \|A\|_F$ is equal to the Euclidean norm. In addition, $|a|$ is the absolute-value norm of a scalar a.

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 and the experimental setup, including a description of all forecasting models and details of the statistics used to analyze results. Section 6 summarizes key empirical findings. Finally, Section 7 concludes.

[2] A static factor model differs from a dynamic factor model, which allows more general infinite dimensional representations. See, for example, Forni et al. (2000), Forni and Lippi (2001), and Forni et al. (2015).

2 Feasible Weighted Projected Principal Component Analysis

2.1 Semiparametric factor model

In this section, I introduce the semiparametric factor model. Consider a factor model defined by

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

where f_{tk} are common factors, λ_{ik} are corresponding factor loadings, and u_{it} is the idiosyncratic component of y_{it}. This paper considers the following semiparametric model:

$$f_{tk} = g_k(\mathbf{X}_t) + \gamma_{tk}, \qquad t = 1, ..., T, k = 1, ..., K, \tag{2.2}$$

where $\mathbf{X}_t$ is a $d \times 1$ vector of observable covariates and g is a unknown nonparametric function.[3] 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 $\mathbf{X}_t$. Recently, Fan et al. (2020) studied a similar model and proposed a robust estimator for heavy-tailed errors. 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 factor structure:

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

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

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

where $\mathbf{Y}$ is the $N \times T$ matrix of y_{it}, $\boldsymbol{\Lambda}$ is the $N \times K$ matrix of λ_{ik}, $\mathbf{G}(\mathbf{X})$ is the $T \times K$ matrix of $g_k(\mathbf{X}_t)$, $\boldsymbol{\Gamma}$ 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 $\mathbf{F} = \mathbf{G}(\mathbf{X}) + \boldsymbol{\Gamma}$ from the model (2.2). Also $E(\boldsymbol{\Gamma}|\mathbf{X}) = 0$, where $\mathbf{G}(\mathbf{X})$ and $\boldsymbol{\Gamma}$ are orthogonal factor components so that $E[\mathbf{G}(\mathbf{X})\boldsymbol{\Gamma}'] = 0$. This paper assumes $K = \dim(\mathbf{F}_t)$ and $d = \dim(\mathbf{X}_t)$ to be constant. In addition, the number of factors K is assumed to be known. 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., Ahn and Horenstein, 2013; Lam and Yao, 2012).

I assume that $g_k(\mathbf{X}_t)$ does not depend on i, which means the common factors represent the time heterogeneity only. To estimate $g_k(\mathbf{X}_t)$, $g_k(\cdot)$ is assumed to be additive for multivariate

[3] 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.

covariates $\mathbf{X}_t$. Define, for each $k \leq K$ and for each $t \leq T$,

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

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. Then each additive component $g_{kl}(X_{tl})$ is estimated using the sieve method. J denotes the number of sieve terms and it grows slowly as $T \to \infty$.

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

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

Here, I describe the main idea of the projection. Let $\mathcal{X}$ be a space spanned by $\mathbf{X}$, which is orthogonal to the error matrix $\mathbf{U}$. 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:

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

where $\widetilde{\mathbf{E}} = \mathbf{\Lambda}\mathbf{\Gamma}'\mathbf{P} + \mathbf{\Lambda}\mathbf{R}(\mathbf{X})'\mathbf{P} + \mathbf{U}\mathbf{P} \approx 0$ because $\mathbf{\Gamma}$ and $\mathbf{U}$ are orthogonal to the function space spanned by $\mathbf{X}$, and $\mathbf{\Lambda}\mathbf{R}(\mathbf{X})'$ 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 the accurate estimators.

Fan et al. (2016) proposed the projected principal component (PPC) method for the semi-parametric factor model.[4] However, the idiosyncratic components are often cross-sectionally

[4] Fan et al. (2016) considered the similar semi-parametric factor model, but the factor loading has the semiparametric structure, such as: $\lambda_{ik} = g_k(\mathbf{X}_i) + \gamma_{ik}, i = 1, ..., N$.

heteroskedastic and correlated in factor models. Since the PPC method does not require estimating the $N \times N$ covariance matrix, $\boldsymbol{\Sigma}_u = \mathrm{cov}(u_t)$, 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:

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

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 $W = \boldsymbol{\Sigma}_u^{-1}$. However, the optimal weight is usually infeasible. Hence, the proposed FPPC method requires a consistent estimator $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ as the feasible weight matrix. It is commonly used in the generalized method of moments literature. 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, we obtain the consistent estimator $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ for $\boldsymbol{\Sigma}_u^{-1}$ under the conditional sparsity assumption. More specific estimation procedure is discussed in the following sections.

2.2 Infeasible estimation

Let $\mathcal{X}$ be the sieve space spanned by the basis functions of $\mathbf{X}$. Define the $T \times T$ projection matrix

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

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

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

is the estimator of $\mathbf{G}(\mathbf{X})$.

The common factor component $\boldsymbol{\Gamma}$ that cannot be explained by the covariates can be estimated as follows. With the estimated weighted factor loadings $\widehat{\boldsymbol{\Lambda}^*}$, the least-squares estimator of common factor matrix is

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

In addition, by (2.4), an estimator of $\mathbf{\Gamma}$ 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 $\mathbf{\Sigma}_u^{-1}$ is obtained. Therefore, this paper considers $W = \widehat{\mathbf{\Sigma}}_u^{-1}$ as the asymptotically optimal weight matrix, which takes into account both heteroskedasticity and cross-sectional correlations simultaneously.

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

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

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

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

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

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	$\sum_{t=1}^{T}(y_t - \mathbf{\Lambda}\mathbf{B}\phi(\mathbf{X}_t))'\widehat{\mathbf{\Sigma}}_u^{-1}(y_t - \mathbf{\Lambda}\mathbf{B}\phi(\mathbf{X}_t))$	$\widehat{\mathbf{\Sigma}}_u^{-1/2}\mathbf{Y}\mathbf{P}\mathbf{Y}'\widehat{\mathbf{\Sigma}}_u^{-1/2}$

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 $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{\Sigma}}_u^{\frac{1}{2}}\widehat{\widetilde{\mathbf{\Lambda}}}$. With the estimated weighted factor loadings $\widehat{\widetilde{\mathbf{\Lambda}}}$, the least-squares estimator of common factor matrix 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 factor loadings, $\widehat{\mathbf{\Lambda}}/\sqrt{N}$, are estimated from three different matrices. Table 1 shows 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 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. This paper assumes $\mathbf{\Sigma}_u$ to be a sparse matrix, namely most of the off-diagonal entries are 0 or nearly so, to apply such a weight estimator by following similar conditions as those in Bickel and Levina (2008) and Fan et al. (2013). Consider the notion of generalized sparsity: let $\mathbf{\Sigma}_u = (\Sigma_{u,ij})_{N\times N}$. For some $q \in [0, 1)$, define

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

and it does not grow too fast as $N \to \infty$. In particular, when $q = 0$ (i.e., the exact sparsity case), $m_N = \max_{i\leq N} \sum_{j=1}^{N} 1\{\Sigma_{u,ij} \neq 0\}$, 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*

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

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

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

$$\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1 = O_P(m_N \omega_{N,T}^{1-q}), \tag{3.3}$$

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

Consider the asymptotic performance of the FPPC in the conventional factor model:

$$\mathbf{Y} = \mathbf{\Lambda}\mathbf{F}' + \mathbf{U}. \tag{3.4}$$

In financial application, 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.[5] 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.

[5] 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 $i = 1, ..., N, k = 1, ..., K$.

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. In addition, we may allow serial dependence and nonstationarity on $\{\mathbf{X}_t\}_{t \leq T}$ in this paper.

The following introduces regularity conditions about weak dependence and stationarity on $\{(\mathbf{F}_t, \mathbf{u}_t)\}$.

Assumption 3.4. *(Data generating process). (i) $\{\mathbf{F}_t, \mathbf{u}_t\}_{t \leq T}$ is strictly stationary; $Eu_{it} = 0$ for all $i \leq N, t \leq T$; $\{\mathbf{u}_t\}_{t \leq T}$ 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.
(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). The high-dimensional factor analysis requires both serially and cross-sectionally weak dependence on the error term, $\{u_{it}\}_{i\leq N, t\leq T}$. It is satisfied when the error covariance matrix is sufficiently sparse under the strong mixing condition.

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(ii)-3.4 hold and $m_N\delta_{N,T}^{1-q} = o(1)$, for $\delta_{N,T} = \sqrt{\frac{\log N}{T}} + \sqrt{\frac{J}{T}} + \frac{1}{\sqrt{N}}$. 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*

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

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

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. 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. I discuss additional details further below.

3.3 FPPC in semiparametric factor models

In the semiparametric factor model, it is assumed that $f_{tk} = g_k(\mathbf{X}_t) + \gamma_{tk}$. 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:

$$\widetilde{\mathbf{Y}}\mathbf{P} = \widetilde{\boldsymbol{\Lambda}}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})' + \mathbf{E}, \tag{3.5}$$

where $\mathbf{E} = \widetilde{\boldsymbol{\Lambda}}\mathbf{R}(\mathbf{X})'\mathbf{P} + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}'\mathbf{P} + \widetilde{\mathbf{U}}\mathbf{P}$ is approximately "small", because $\mathbf{R}(\mathbf{X})$ is the sieve approximation error, and $\boldsymbol{\Gamma}$ and $\widetilde{\mathbf{U}}$ are orthogonal to the function space spanned by $\mathbf{X}$. The sieve coefficient matrix $\mathbf{B}$ can be estimated by least squres from the above model (3.5) as:

$$\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_K)' = \frac{1}{N}\widehat{\widetilde{\boldsymbol{\Lambda}}}'\widetilde{\mathbf{Y}}\boldsymbol{\Phi}(\mathbf{X})[\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X})]^{-1}. \tag{3.6}$$

Then the estimator for $g_k(.)$ is

$$\widehat{g}_k(\mathbf{x}) = \phi(\mathbf{x})'\widehat{\mathbf{b}}_k \qquad \forall \mathbf{x} \in \mathcal{X}, k = 1, \cdots, K, \tag{3.7}$$

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

The estimators $\widehat{\boldsymbol{\Lambda}}$, $\widehat{\mathbf{G}}(\mathbf{X})$ and $\widehat{\mathbf{F}}$ are the FPPC estimators as defined in Section 2.3. Since $\mathbf{F} = \mathbf{G}(\mathbf{X}) + \boldsymbol{\Gamma}$, $\mathbf{G}(\mathbf{X})$ 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 $T \to \infty$,*

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

Serial weak dependence for $\{\boldsymbol{\gamma}_t\}_{t\le T}$ is imposed as following.

Assumption 3.6. *(i) $E\gamma_{tk} = 0$ and $\{\mathbf{X}_t\}_{t\le T}$ is independent of $\{\gamma_{tk}\}_{t\le T}$.*
(ii) Define $\boldsymbol{\gamma}_t = (\gamma_{t1}, \cdots, \gamma_{tK})'$, and

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

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

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

In addition, the assumption for the accuracy of the sieve approximation is considered.

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

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

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).

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 $\omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,*

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

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 $\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\gamma}}_1, ..., \widehat{\boldsymbol{\gamma}}_T)'$ satisfies

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

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

3.4 Diffusion index forecasting models

In this subsection, I study the forecasting regression model using the estimated factors, the so-called diffusion index (DI) forecasting model, which is originated from Stock and Watson (2002a). In the forecasting literature, this model has been used widely for prediction.

Consider the following forecasting equation:

$$z_{t+h} = \alpha'\mathbf{F}_t + \beta' W_t + \epsilon_{t+h}, \tag{3.8}$$

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 obtain $\widehat{\mathbf{F}}_t$ using principal components methods from the factor model:

$$y_t = \boldsymbol{\Lambda}\mathbf{F}_t + \mathbf{u}_t. \tag{3.9}$$

Note that, when z_t is a scalar, equations (3.8) and (3.9) constitute the DI model. In addition, the equation (3.8) is the FAVAR model of Bernanke et al. (2005), when $h = 1$ and $z_{t+1} = (\mathbf{F}'_{t+1}, W'_{t+1})'$. Intuitively, the common factor, $\mathbf{F}_t$, is known for the common shocks that generate comovements in economic time series.

Suppose the interest object is the conditional mean of (3.8), which is

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

where $L_t = (\mathbf{F}'_t, W'_t)'$. For example, if z_t employment rate, the estimated conditional mean can be interpreted as an estimate of the expected employment rate. 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 estimated factors using FPPC. Then the feasible prediction would be

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

Stock and Watson (2002a) proved that $\widehat{\delta}$ is consistent for δ and $\widehat{z}_{T+h|T}$ is consistent 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. These papers used the regular PC estimation method under the static factor model. On the other hand, this paper obtains the asymptotic distribution of the least squares estimates and the rate of convergence of the conditional mean based on the FPPC estimation method as discussed in Section 2.3. To do so, the following assumption is required.

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 $\Sigma_{L,\epsilon} = \operatorname{plim}\frac{1}{T}\sum_{t=1}^T \epsilon_{t+h}^2 L_t L_t'$.

Assumption 3.8 is standard for forecasting regression analysis. 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 standard assumptions in regressions and ensures that the parameters of the forecasting model can be identified.

In this section, I assume the semiparametric factor model as in Section 3.3 for the equation (3.9). All the theorems and proofs of 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. *(Estimation) Let $\widehat{\delta} = (\widehat{\alpha}', \widehat{\beta}')'$ and $\delta = (\alpha'\mathbf{H}, \beta')'$. Suppose the assumptions of Theorems 3.1-3.2 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 $\Pi = \operatorname{diag}(\mathbf{H}', \mathbf{I})$. 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 consider the convergence rate of the conditional mean, $z_{T+h|T}$. Define the forecast error as

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

which contains two components, 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 obtain the rate of convergence for FPPC, 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$(0,1)$, and there are three factors ($K = 3$). I set the number of characteristics as $\dim(\mathbf{X}_t) = 3$. 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. $N(\mathbf{0}, \mathbf{I})$. 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 I vary $w = \{10, 1, 0.1\}$, and 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 $l \leq d$. Note that, for each $k \leq K$, 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 $\{\varepsilon_{it}\}_{i \leq N, t \leq T}$ be i.i.d. $N(0,1)$. 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 i.i.d. $N(0, \sqrt{5})$. 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 $D = \text{diag}(d_i)$, where $\{d_i\}_{i \leq N}$ is drawn from i.i.d. Uniform$(0, \sqrt{5})$. Finally, define $\Sigma_u = D R_\eta D$, and generate $\{u_t\}_{t \leq T}$ as i.i.d. $N(0, \Sigma_u)$. Note that this generating procedure of the error term is similar to Bai and Liao (2017).

I have simulated the data and reported for $N = \{50, 100, 300\}$ 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 cross-validation, as discussed in Section 2.3.2.

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, I define the averaged mean squared error (MSE) of estimated common components as $(\frac{1}{NT}\sum_{i,t}(\widehat{\lambda}_i'\widehat{f}_t - \lambda_i' f_t)^2)^{1/2}$. The results are reported in Table 3.

According to Tables 2 and 3, FPPC outperforms PPC and PC. Overall, the estimation becomes more accurate as the dimensionality increases. For loadings, FPPC performs better than PPC and PC except for the mild and weak explanatory power cases with larger dimensionality. When $w = 0.1$, on the other hand, the observed $\mathbf{X}_t$ is not as informative, and hence the performance of PPC and FPPC deteriorates. For common factors, however, FPPC always outperforms PPC and PC. In addition, for common components, FPPC gives the smallest mean squared errors for most cases, except when $w = 0.1$ has a larger dimensionality.

Table 2: Canonical correlations of estimated loading or factor matrices: the larger 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
300	100	0.439	0.745	**0.766**	**0.686**	0.621	0.675	**0.786**	0.344	0.461
	500	0.800	0.941	**0.942**	**0.928**	0.902	0.904	**0.952**	0.667	0.705
		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**
300	100	0.527	0.883	**0.990**	0.785	0.784	**0.978**	0.878	0.518	**0.879**
	500	0.780	0.922	**0.993**	0.910	0.927	**0.993**	0.940	0.841	**0.976**

Table 3: Mean squared error of estimated common components $\mathbf{\Lambda F}'$: 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
		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**
300	100	0.375	0.298	**0.255**	0.348	0.360	**0.330**	**0.331**	0.496	0.475
	500	0.231	0.179	**0.120**	0.205	0.210	**0.163**	**0.198**	0.341	0.327

4.2 Out-of-sample forecast

This subsection evaluates the performance of the proposed factor estimators on out-of-sample forecasts. Consider a diffusion index 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. $N(0,1)$. To cover a variety of model settings, unknown coefficients, α, are generated from Uniform$(0.5, 1)$ for each simulation. The unknown factor, $\mathbf{F}_t$, can be learned from a factor model: $\mathbf{y}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{u}_t$. Here the same data-generating process is used as Section 6.1.

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 average RMSFE are calculated as measurements of the forecasting performance based on 1000 replications.

The results are presented in Table 4. Overall, FPPC has smaller MSFEs than PPC and PC in this forecasting model. When the correlation between $\mathbf{X}_t$ and $\mathbf{F}_t$ is strong, PPC yields better forecasts than the regular PC method as expected, while FPPC outperforms PPC and PC. On the other hand, as the explanatory power gets weaker, the forecasting performances of PPC and FPPC decrease compared to PC. This phenomenon corresponds to the results of Section 4.1.

Table 4: 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**
300	100	1.000	0.981	**0.947**	1.000	1.001	**0.953**	1.000	1.070	**0.980**
	500	1.000	0.990	**0.958**	1.000	0.997	**0.963**	1.000	1.015	**0.963**

5 Empirical Analysis: US Bond Risk Premia

As an empirical study, I investigate the excess return of U.S. government bonds using the proposed FPPC-based diffusion index (DI) model. Fama and Bliss (1987) show that n-year excess bond returns are predictable by the spread between the n-year forward rate and the one-year yield. Cochrane and Piazzesi (2005) find that a so-called CP factor from five forward spreads explains a significant variation in one year ahead in excess bond returns with 2-5 year maturities. In the financial economic literature, a large body of research shows that risk premiums are forecastable by macroeconomic variables. Particularly, Ludvigson and Ng (2009) claim that common factors, extracted from a large number of economic time series, also have an important forecasting power besides the predictive information of the factor in Cochrane and Piazzesi (2005). Recently, Bianchi et al. (2019) asserted how machine learning methods (such as regression trees and neural networks) provide strong statistical evidence in predicting excess bond returns using both macroeconomic and yield information.

To obtain the common factors from large datasets, the conventional principal component (PC) method is popular, as Ludvigson and Ng (2009) implemented. In this paper, I shall explore how the newly proposed method, FPPC, performs in forecasting the excess bond returns. In addition, I empirically assess the predictive accuracy of a large group of models that are linear models (e.g., DI model) 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 $r_{t+1}^{(n)} = p_{t+1}^{(n-1)} - p_t^{(n)}$ as the log holding period return from buying an n-year bond at time t and selling it as an $n - 1$ year bond at time $t + 1$. Then the excess return with maturity of n-years is

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

where $y_t^{(1)}$ is the log yield on the one-year bond. Intuitively, one buys an n year maturity bond, sells it as an $n-1$ year bond in the next year and excess the one-year bond, which is the risk-free rate.

5.1 Data

I analyze monthly bond return data spanning from 1964:1 to 2016:4 ($T = 628$), which is the updated version of Ludvigson and Ng (2009) and Cochrane and Piazzesi (2005). 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. A specific description and transformation code of panel data is provided in McCracken and Ng (2016).[6] 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. 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 5. In addition, these data are also transformed and standardized.[7]

[6] The macroeconomic dataset is the FRED-MD monthly database. As of 2016:05, FRED-MD removed some variables (e.g., NAPMPI, NAPMEI, NAPM, etc.). Hence, I obtained the dataset up to 2016:04 to use the same variables as in Ludvigson and Ng (2009).

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

Table 5: 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

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 6 lists all forecasting models in the experiments. To avoid bogging down the reader with details of all methods, all models and specific implementation choices are described in Appendix D.

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 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 P as the number of ex-ante forecasts, and Q is the length of the rolling window or the initial length of the recursive window, hence $T = P + Q$ is the total length of the sample.

To evaluate the forecasting performance of various models, I utilize two statistics as follows:

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

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

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

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

where, for each maturity n, $\widehat{rx}_{t+1}^{(n)}$ is the forecast of bond excess returns using each

Table 6: List of all forecasting models

Method	Description
AR(SIC)	Autoregressive model with lags selected by the SIC
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
Bagging	Bagging with shrinkage, $c = 1.96$
Boosting	Component boosting, $M = 50$
Fac-Lasso	Factor-Lasso regression
Lasso	Lasso regression
Ridge	Ridge regression
EN	Elastic net regression
DT	Decision tree regression
G-Bst	Gradient boosting regression
RanForest	Random forest regression
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

model and $\overline{rx}_{t+1}^{(n)}$ is the historical average of bond excess return.[8]

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

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

where $\epsilon_{i,t+1|t}$ 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 $T, P, Q \to \infty$ 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 $\bar{d} = \frac{1}{P}\sum_{t=1}^{P}(\widehat{\epsilon}^2_{1,t+1|t} - \widehat{\epsilon}^2_{2,t+1|t})$, $\widehat{\sigma}_{\bar{d}}$ is a heteroskedasticity and autocorrelation robust estimator of the standard deviation of $\bar{d}$. Here $\widehat{\epsilon}_{1,t+1|t}$ and $\widehat{\epsilon}_{2,t+1|t}$ 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 in an out-of-sample forecast. 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 in Section 6.2.2.

6 Empirical Findings

6.1 In-sample analysis and economic interpretation of the factors

In this section, an in-sample regression analysis is conducted. I also economically interpret the extracted factors through the FPPC method using 130 macroeconomic variables with additional characteristics.

First, consider a principal component regression model as follows:

$$rx_{t+1}^{(n)} = \alpha + \beta' \widehat{\mathbf{F}}_t + \epsilon_{t+1}. \tag{6.1}$$

I investigate the unconditional predictive power of macro factors for future returns using different principal component methods. Note that the equation (6.1) is a restricted version

[8] The historical average of bond excess return starts from 1964:1, and values are the same for both rolling and recursive window methods.

Table 7: In-sample adjusted R^2 of U.S. bonds excess return forecasting: the larger the better.

Maturity	PC_{130}	PC_{134}	PPC	FPPC
	Sample period: 1964:1-2016:4			
2 year	15.5	16.3	17.4	18.0
3 year	15.9	16.7	16.8	17.3
4 year	16.1	16.8	17.0	17.6
5 year	17.3	17.8	17.7	18.4
	Sample period: 1964:1-2003:12			
2 year	23.5	26.1	28.0	30.4
3 year	22.4	25.5	25.5	27.4
4 year	22.2	25.5	25.2	27.2
5 year	21.6	24.8	24.6	26.5

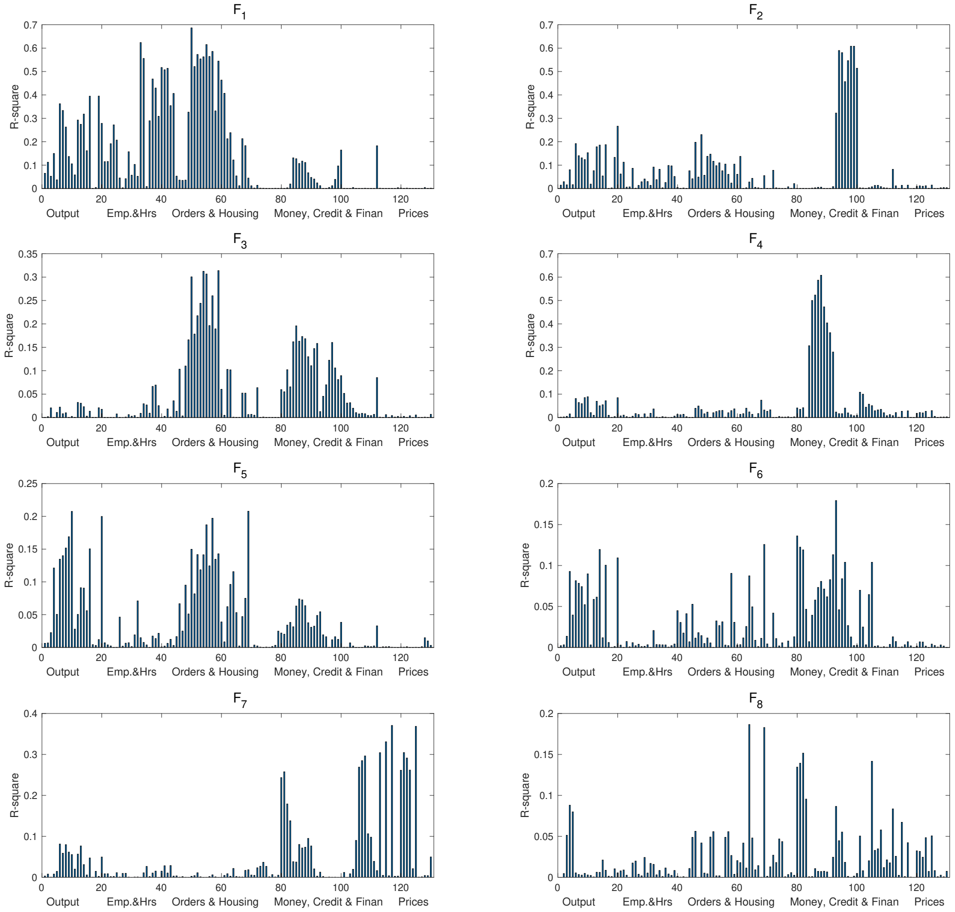
of the regular diffusion index model (see Ludvigson and Ng, 2009). This simple model is employed to simply compare the predictive power of different principal component methods, such as PC, PPC, and FPPC, for excess bond returns. I set the number of factors $K = 8$ for all methods, which is determined by the information criteria suggested in Bai and Ng (2002).

Table 7 reports the adjusted R^2 statistics from in-sample regressions of the equation (6.1), for 2- to 5-year log excess bond returns. PC_{130}, which is a benchmark, denotes the conventional PC method with 130 macro variables and PC_{134} with four additional characteristics $\mathbf{X}_t$ introduced in Table 5. In addition, the sieve basis for PPC and FPPC is chosen as the additive polynomial basis with $J = 5$. For FPPC, the threshold constant is selected by cross-validation as introduced in Section 2.3.2. Two sample periods are considered: 1964:1-2016:4 and 1964:1-2003:12.[9] From the table, I find that the estimated factors using FPPC outperform the factors using other methods. In the first panel of Table 7, the factors estimated by FPPC explain 18.0% of the variation one year ahead in the 2-year return, while the factors estimated by PC_{130} only explain about 15.5%. Interestingly, even though PC_{134} performs better than PC_{130} by adding additional covariates, it underperforms compared to the FPPC.

Next, I economically interpreted the extracted factors using my method and compared them with those using the regular PC method. By following Ludvigson and Ng (2009, 2016), I calculated the marginal R^2, which is the R^2 statistic from regressions of each of the 130 series onto each of the estimated factors. Each panel of Figure 2 displays the R^2 statistics as bar charts for each factor. I also group the 130 macroeconomic variables into five groups: (i) output and income, (ii) employment and hrs (i.e., labor market), (iii) orders (i.e., consumption; orders and inventories) and housing, (iv) money, credit, and finance, and (v) prices. Plots of the first two factors using my method, $\widehat{F}_{1t}$ and $\widehat{F}_{2t}$, are very similar to

[9] Ludvigson and Ng (2009) studied sample period 1964:1-2003:12.

Figure 2: Marginal R^2 for extracted factors



Note: Each panel of the figure shows the R^2 from regressing the 130 macroeconomic series onto each of the extracted factors. The factors are estimated using data from 1964:1 to 2016:4. A detailed description of the numbered series is presented in Appendix A of Ludvigson and Ng (2009).

that of PC, as shown in Ludvigson and Ng (2016). They interpret the first factor as the "real factor," which loads heavily on measures of employment, production, capacity utilization, and new manufacturing orders. The second factor loads heavily on variables in group (iv), especially several interest rate spreads.

Interestingly, other factors show different aspects. The third factor, $\widehat{F}_{3t}$, is correlated with housing. The fourth factor, $\widehat{F}_{4t}$, loads heavily on the nominal interest rates, such as the five-year government bond yield in group (iv). The rest of the extracted factors load much less heavily. The fifth factor, $\widehat{F}_{5t}$, is correlated with industrial production and housing, while the sixth factor, $\widehat{F}_{6t}$, loads heavily on measures of the aggregate stock market. The seventh factor loads heavily mostly on measures of inflation and price pressure, but explains little relation to the stock market. Lastly, $\widehat{F}_{8t}$ is highly correlated with consumption orders and inventories. I interpret $\widehat{F}_{3t}$ as a housing factor, $\widehat{F}_{6t}$ as a stock market factor, both $\widehat{F}_{4t}$ and $\widehat{F}_{7t}$ as inflation factors, and both $\widehat{F}_{5t}$ and $\widehat{F}_{8t}$ as real factors.

6.2 Forecasting performance

In this subsection, 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 ($P = 388$). Here the rolling forecast method uses information from the past 240 months, while the recursive forecast method uses information from the past $240 + s$ months for $s = 1, \cdots, 388$. Note that 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 6. In addition, the replication of experiments using different forecasting periods was considered.

6.2.1 Forecasting power of FPPC in diffusion index models

I first focus on linear forecasting models with obtained factors from different PC methods. 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.

Tables 8 and 9 present results of out-of-sample forecasting in linear models using a rolling and a recursive window. The first half columns in tables 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., $L_t = (F_t', CP_t)'$). The forecast performance is evaluated by the out-of-sample R^2 and the relative mean squared forecast error, defined as MSFE(M)/MSFE(PC$_{130}$) for each method M. Note that the column of PC$_{130}$ reports its MSFE. Here, PC$_{134}$ denotes the regular PC method using four additional characteristics in addition to 130 macro variables.

Table 8: Relative mean squared forecast errors of U.S. bonds excess return forecasting: the smaller 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	2.421	0.978	0.849	0.789	1.542	1.004	0.961	0.964
3 year	8.310	0.981	0.882	0.825	5.700	1.000	0.945	0.942
4 year	17.011	0.974	0.871	0.808	11.589	0.996	0.923	0.909
5 year	25.668	0.976	0.879	0.819	18.417	0.997	0.939	0.923
	Recursive estimation window							
2 year	2.494	0.962	0.954	0.931	1.782	0.997	0.896	0.902
3 year	8.459	0.963	0.990	0.969	6.421	0.990	0.892	0.889
4 year	17.081	0.962	0.995	0.971	12.876	0.985	0.892	0.880
5 year	25.710	0.965	1.004	0.977	20.079	0.984	0.919	0.903

First, FPPC results in notable improvements in out-of-sample predictive accuracy, when comparing MSFE values and out-of-sample R^2. For instance, in the DI model using the rolling window scheme, I find that FPPC generates an approximately 3.7-9.2% decrease in MSFE and 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. In particular, in the PCR model using the rolling window scheme, FPPC greatly improves the forecasting power compared to other methods. Moreover, Table 10 shows the out-of-sample R^2 of the DI model using the rolling window separately for the recession and expansion sub-samples as defined by the NBER recession index. Based on results, FPPC and PPC remarkably outperform PC in recessions.

Second, the DM test results are provided in Table 11. I compare FPPC to other 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. In the PCR model using the rolling window scheme, 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, but signs of DM test statistics are mostly negative. In the PCR model using the recursive window scheme, FPPC does not outperform other methods, and some statistics are positive, which corresponds to the results in Tables 8 and 9. Interestingly, for the DI model, FPPC mostly outperforms PCs, while FPPC does not yield significantly better results compared to PPC.

Overall, I confirmed that the information of characteristics, $\mathbf{X}_t$, has explanatory power on the latent factors. In addition, the results can be interpreted by the following outlines: (i)

Table 9: 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 10: 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

Table 11: Diebold-Mariano test statistics. Forecast sample period: 1984:1-2016:4.

Model	Maturity			
	2 year	3 year	4 year	5 year
	Rolling estimation window			
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
	Recursive estimation window			
PCR ($L_t = F_t$)				
FPPC versus PC$_{130}$	-1.219	-0.588	-0.551	-0.442
FPPC versus PC$_{134}$	-0.646	0.120	0.181	0.254
FPPC versus PPC	-0.925	-0.895	-1.053	-1.148
DI ($L_t = (F_t', CP_t)'$)				
FPPC versus PC$_{130}$	-2.139**	-2.459**	-2.890***	-2.464**
FPPC versus PC$_{134}$	-2.105**	-2.271**	-2.579**	-2.144**
FPPC versus PPC	0.268	-0.181	-0.747	-1.000

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

The estimated factors using characteristics yield significant improvement to forecast the US bond excess returns, and (ii) FPPC outperforms other principal component analysis methods, including PPC and PC in linear models.

6.2.2 Forecasting performance

Here, I investigate the one-year-ahead forecasting performance of linear and nonlinear machine learning models listed in Table 6. For factor-augmented models such as PCR, FAAR, DI, Bagging, Boosting, and Fac-Lasso, all PC, PPC, and FPPC methods are conducted. Because FPPC outperforms PC and PPC in most cases, the presented results of these models are all based on the FPPC method.

Several clear-cut findings are obtained through the inspection of the results contained in Tables 12-17. Table 12 reports relative MSFEs of all forecasting models, using rolling and recursive estimation window strategies. The AR(SIC) is used as a benchmark to generate relative MSFEs for all other models. In addition, out-of-sample R^2 of all forecasting models are tabulated in Table 13. Tables 14-17 report the pairwise test statistics of conditional predictive ability (GW) proposed by Giacomini and White (2006) for each maturity based on the rolling window scheme. The forecasting period of all tables is from 1984:1-2016:4 with a total of 388 months. I summarize the main empirical findings below.

First, rolling window forecasts (i.e., fixed window size, $Q = 240$) outperform recursive window forecasts for most of the models based on out-of-sample R^2 and MSFE 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. For example, PCR using the rolling window has 22% out-of-sample R^2, while PCR using the recursive window has only 5% for 2-year maturity.

Second, the FPPC-based diffusion index (DI) models "win" over most machine learning models, or are comparable with the best nonlinear machine learning models. For instance, in Tables 12-13 we see that DI2 using the rolling window scheme generates an approximately 30-40% decrease in MSFE when compared to the benchmark AR(SIC) model, and it has about 40-45% out-of-sample R^2 for each maturity. Additionally, GW test statistics reported in Tables 14-17 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 (such as 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. Note that the proposed FPPC method improves predictive accuracy compared to PC and PPC methods as discussed in

Section 6.2.1.

Third, the RanForest model outperforms all others among the models except for FPPC-based DI models. Evidently, RanForest has the smallest RMSFEs and largest out-of-sample R^2, especially in the recursive window scheme of Tables 12 and 13. However, in the rolling window scheme, DI models perform better than the RanForest model for some maturities. For example, an inspection of Tables 12 and 13 indicates that both RanForest and DI models have similar MSFE and out-of-sample R^2 values. Also, pairwise GW test statistics reported in Tables 14-17 support this claim. Among these models, there are no statistically significant GW test statistics for all maturities.

Fourth, the performance of FANN, which uses the extracted factors as regressors, stands out in various neural network architectures. Also, it is comparable with RanForest and DI models based on inspections of Tables 12-17. Among several types of neural networks, FANN outperforms NN and H-NN based on MSFE, out-of-sample R^2, and DM test statistics. Because a large number of predictors yield overfitting problems in general, the conventional NN and H-NN underperform FANN, which only use a small number of predictors. Interestingly, in addition, adding more hidden layers in NN and H-NN does not improve the predictive performance. Moreover, having five hidden units in FANN is the optimal choice in this experiment.

In summary, FPPC-based DI (including the modified models such as Bagging, Boosting, and Fac-Lasso), 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.

6.2.3 Robustness checks: Different forecasting sub-periods

Finally, I reexamined the experiments using different forecasting periods. Specifically, the full forecasting period from 1984:1-2016:4 is divided into three equal subperiods, including P1: 1984:1-1994:10, P2: 1994:11-2005:8, and P3: 2005:9-2016:4. On average, there are 130 months over each period. The relative MSFE results based on the rolling window scheme are presented in Table 18. The results in this table show that for P1, the MSFEs for all models are much larger than those in the other two subperiods. This is not surprising because the data in this period are more volatile than in the rest periods. Based on MSFEs and DM tests, DI models (including the modified models), H-NN, and FANN outperform other models. For instance, the DI model for 4-year maturity generates an approximately 51.2% decrease in MSFE when compared to the benchmark model. Interestingly, during P2, MSFEs for most models decline sharply to around 40-50% of the levels in P1. In other words, the MSFEs during this period for all models vary within a relatively narrow range. Hence, it may be challenging for a particular model to significantly outperform other models in this period. However, according to the relative MSFEs and DM test results, DI2 and RanForest outperform other models, and this corresponds to the previous results. Finally, during P3, the MSFEs of all models fall again to 40-60% of the levels seen during P2. Indeed, the bond excess return values during this period vary within a relatively narrow range compared to the values of the previous periods. In addition, this period includes the financial crisis of 2007-2008. Only the RanForest model stands out to be the best model for all maturities in this period, based on the results in Table 18. Note that RanForest has the most stable forecasting performance of all forecasting periods (i.e., not many outliers). Unfortunately, the (modified) diffusion index models do not seem to perform well in this period. Especially in the period immediately after the financial crisis, these linear models have poor predictive performance.

7 Conclusions

This paper examines a high-dimensional factor model in which the factors depend on a few observed covariate variables. This model is motivated by the fact that observed variables can partially explain the latent factors. I propose the feasible weighted projected principal component (FPPC) analysis, which takes into account cross-sectional heteroskedasticity and correlations, to estimate the unknown factors and loadings. In addition, I study the FPPC-based diffusion index model. 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 forecast bond risk premia. Moreover, I find that the proposed linear forecasting model performs well among other nonlinear machine learning models in terms of out-of-sample forecasting.

Table 12: 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
FANN	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 13: Out-of-sample R^2 (%) of U.S. bonds excess return forecasting: the larger 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)	13.1	9.4	9.1	8.0	8.9	4.6	2.9	1.3
PCR	22.0	22.3	25.1	26.3	5.0	6.7	9.3	11.5
FAAR	20.0	22.2	26.5	28.5	4.9	9.5	13.9	17.0
DI	39.2	39.0	42.5	40.3	34.1	34.9	37.8	36.0
DI2	39.5	41.8	45.4	42.7	33.1	35.5	38.7	37.0
Bagging	40.5	41.2	44.7	39.4	31.5	32.8	36.3	33.9
Boosting	43.2	42.0	44.5	39.7	33.5	32.5	34.3	31.5
Fac-Lasso	42.6	42.5	45.6	43.1	32.7	34.5	38.2	37.0
Lasso	14.3	19.0	22.7	23.6	10.5	16.0	19.0	20.0
Ridge	15.3	13.5	19.8	18.6	-3.5	-0.6	7.7	6.7
EN	13.9	16.4	20.8	22.0	8.8	11.1	15.2	16.8
DT	17.9	4.0	11.9	2.7	-4.0	-2.6	16.7	13.0
G-Bst	22.3	19.3	19.7	19.9	16.0	15.2	15.1	16.5
RanForest	43.8	40.2	39.7	40.7	40.6	38.3	38.9	39.2
NN1	18.2	18.0	22.5	22.1	18.2	17.9	20.2	21.9
NN2	5.1	7.8	7.2	6.2	-3.0	-2.2	-2.7	-1.3
NN3	5.8	7.6	7.0	7.9	-1.9	-3.4	-3.3	-2.6
H-NN1	34.9	36.0	36.8	31.1	26.7	23.7	25.0	23.4
H-NN2	29.6	28.2	31.1	29.5	8.6	5.7	5.9	3.6
H-NN3	30.8	27.9	29.7	29.6	10.1	5.6	5.5	3.5
FANN	40.0	39.0	41.2	37.7	30.1	30.6	32.9	29.7
FANN2	40.0	40.4	41.5	39.9	30.1	30.6	33.8	29.4
FANN3	39.3	39.3	40.9	40.1	29.5	30.4	33.8	31.1

Table 14: 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 utilized 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.

Table 15: Pairwise model comparison using Giacomini-White tests for 3-year maturity.

	PCR	FAAR	DI	DI2	Bagging	Boosting	Fac-Lasso	Lasso	Ridge	EN	DT	G-Bst	RanForest	NN	H-NN	FANN
AR(SIC)	1.27	2.21	**5.32**	**6.79**	**6.33**	**5.80**	**5.70**	2.89	-0.08	**6.09**	0.69	3.32	**6.64**	**4.62**	2.93	3.36
PCR		1.35	4.47	**8.22**	**6.57**	4.48	**4.67**	-0.38	-0.57	-0.54	-1.61	-0.31	3.34	-0.19	**5.40**	3.89
FAAR			2.64	**6.11**	3.56	4.53	3.16	-0.68	-0.82	-0.89	-1.12	-0.45	2.65	-0.28	3.48	2.13
DI				4.19	**7.29**	0.04	0.67	**-4.80**	**-17.88**	-4.42	**-10.76**	-3.56	0.03	**-4.79**	-2.99	-1.85
DI2					-2.29	-2.77	-0.10	**-5.84**	**-20.83**	**-5.68**	**-14.36**	**-5.10**	-0.22	**-6.67**	-2.47	-2.06
Bagging						3.82	0.04	**-5.56**	**-18.32**	**-5.27**	**-15.43**	**-4.88**	-0.07	**-6.24**	-2.31	-1.51
Bsting							0.50	**-4.93**	**-11.95**	**-4.90**	**-11.52**	-4.53	0.00	-3.76	-3.20	**-6.91**
Fac-Lasso								**-5.61**	**-20.24**	**-5.21**	**-13.37**	-4.32	-0.54	**-5.19**	-2.12	-1.26
Lasso									-0.15	-0.80	-1.11	0.11	**6.47**	0.59	2.36	2.50
Ridge										0.14	0.65	0.18	**4.98**	0.83	**4.93**	**8.86**
EN											-0.92	0.20	6.17	1.03	2.38	2.57
DT												1.05	**13.14**	0.88	**6.53**	**7.54**
G-Bst													**8.90**	1.62	2.21	2.37
RanForest														**-6.67**	-2.69	-1.02
NN															1.64	1.69
H-NN																0.38

See notes to Table 14. For further details, refer to Section 5.2.

Table 16: Pairwise model comparison using Giacomini-White tests for 4-year maturity.

	PCR	FAAR	DI	DI2	Bagging	Boosting	Fac-Lasso	Lasso	Ridge	EN	DT	G-Bst	RanForest	NN	H-NN	FANN
AR(SIC)	2.98	4.24	**6.32**	**8.87**	**8.70**	**6.24**	**6.75**	4.26	0.26	**6.80**	2.44	3.12	**7.21**	**6.25**	2.17	3.33
PCR		0.58	4.45	**7.73**	**6.06**	3.09	4.32	-1.04	-0.30	-0.98	-0.95	-1.33	2.53	0.05	1.89	1.92
FAAR			2.02	**5.04**	3.64	2.27	2.16	-2.09	-0.73	-1.99	-2.65	-1.68	0.93	-1.52	0.24	0.36
DI				1.25	1.94	-0.32	0.58	**-5.08**	**-11.84**	**-4.79**	**-8.33**	**-5.01**	-0.25	-2.99	-2.20	-1.63
DI2					-1.08	-1.51	-0.10	**-6.47**	**-15.54**	**-6.47**	**-14.38**	**-6.80**	-0.76	**-5.58**	-4.07	-2.79
Bagging						-3.06	-0.01	**-6.53**	**-12.93**	**-6.42**	**-15.07**	**-6.93**	-0.72	**-5.37**	**-5.32**	-3.34
Bsting							0.40	-4.04	**-6.76**	-4.17	**-9.28**	**-4.81**	-0.18	-2.67	**-5.41**	**-5.23**
Fac-Lasso								**-6.01**	**-14.22**	**-5.58**	**-14.00**	**-5.63**	-2.75	-3.58	-2.94	-3.45
Lasso									-0.04	-1.04	-2.98	-0.70	**5.79**	1.10	1.31	1.73
Ridge										0.04	-0.16	0.05	3.06	0.44	1.92	**6.02**
EN											-3.17	-0.53	**5.75**	1.97	1.17	1.84
DT												4.16	**10.48**	1.00	3.52	**4.61**
G-Bst													**8.95**	2.84	1.45	2.48
RanForest														-3.76	-2.06	-0.68
NN															1.67	0.84
H-NN																0.88

See notes to Table 14. For further details, refer to Section 5.2.

Table 17: Pairwise model comparison using Giacomini-White tests for 5-year maturity.

	PCR	FAAR	DI	DI2	Bagging	Boosting	Fac-Lasso	Lasso	Ridge	EN	DT	G-Bst	RanForest	NN	H-NN	FANN
AR(SIC)	**7.19**	**7.18**	**7.85**	**9.79**	**8.70**	**5.89**	**7.93**	**8.90**	0.72	**15.22**	0.48	**5.44**	**9.41**	**14.26**	1.80	**4.69**
PCR		0.38	2.88	3.94	2.05	0.95	2.45	-4.41	-0.49	-2.98	-1.31	-2.87	2.43	-0.61	-0.98	0.42
FAAR			0.98	2.85	1.12	0.27	1.06	-4.43	-1.10	-3.39	-1.81	-2.46	1.31	-0.66	-0.22	0.76
DI				0.22	-4.55	-2.21	0.29	**-4.66**	**-10.87**	-4.49	**-8.76**	-4.21	0.04	-2.22	-3.93	-1.66
DI2					-2.57	-2.35	-0.07	**-5.26**	**-12.33**	**-5.25**	**-10.81**	**-4.84**	-0.12	-3.18	**-5.30**	-2.42
Bagging						-2.08	0.83	-4.37	**-6.93**	-4.23	**-7.42**	-3.97	0.23	-2.23	-4.05	**-6.79**
Bsting							1.18	-1.90	**-5.34**	-2.10	**-5.41**	-2.47	0.48	-1.49	-3.29	2.94
Fac-Lasso								**-4.88**	**-12.31**	**-4.64**	**-10.40**	-4.27	-2.52	-2.47	-3.95	-2.33
Lasso									-0.07	-1.67	-0.27	-0.79	**5.90**	1.74	1.01	1.57
Ridge										0.04	-0.15	0.01	3.60	0.32	2.47	**7.72**
EN											-0.14	-0.65	**5.88**	2.50	0.66	1.74
DT												0.04	**8.97**	0.69	0.89	**5.32**
G-Bst													**8.05**	4.09	1.22	1.93
RanForest														-3.57	-2.73	-0.48
NN															-1.58	0.72
H-NN																2.51

See notes to Table 14. For further details, refer to Section 5.2.

Table 18: Relative mean squared forecast errors of U.S. bonds excess return forecasting for different prediction periods. Rolling window.

	P1 (1984:1-1994:10)				P2 (1994:11-2005:8)				P3 (2005:9-2016:4)			
Method	2 year	3 year	4 year	5 year	2 year	3 year	4 year	5 year	2 year	3 year	4 year	5 year
AR(SIC)	3.793	13.994	29.938	45.453	1.789	6.943	13.065	20.44	0.765	2.890	6.711	12.392
PCR	0.835*	0.790**	0.766***	0.757***	0.965	0.956	0.945	0.946	1.058	0.951	0.848	0.719
FAAR	0.851	0.787**	0.746***	0.731***	0.863	0.764	0.715*	0.709*	1.404	1.441	1.281	1.063
DI	0.578***	0.525***	0.488***	0.513***	0.882	0.880	0.879	0.891	0.879	0.897	0.807	0.747
DI2	0.597***	0.532***	0.490***	0.517***	0.753	0.680*	0.653**	0.679*	1.055	1.096	0.999	0.920
Bagging	0.570***	0.531***	0.499***	0.549***	0.802	0.727	0.651**	0.734	0.985	1.043	1.019	0.940
Boosting	0.529***	0.513***	0.496***	0.525***	0.769	0.735*	0.677**	0.741*	1.005	1.036	0.995	0.996
Fac-Lasso	0.575***	0.524***	0.483***	0.514***	0.767	0.766	0.773	0.776	0.834	0.855	0.779	0.739
Lasso	0.942	0.843***	0.810***	0.796***	1.046	1.007	0.962	0.933	1.064	0.869**	0.817***	0.785***
Ridge	0.746	0.721	0.646*	0.672*	1.076	1.079	1.096	1.129	1.889	1.799	1.532	1.263
EN	0.973	0.907***	0.860***	0.835***	0.994	0.959	0.919	0.904	1.071	0.908*	0.828***	0.797***
DT	0.874	0.854	0.865	1.022	0.752	1.260	1.027	1.063	1.761	1.580	1.328	1.175
G-Bst	0.960	0.952	0.930	0.920	0.798**	0.827*	0.864	0.881	0.784**	0.740**	0.715***	0.665***
RanForest	0.688**	0.681**	0.693**	0.678**	0.612**	0.672*	0.655**	0.652**	0.522***	0.529***	0.544***	0.506***
NN1	1.002	0.915	0.872*	0.887	0.805**	0.914	0.824	0.833	0.963	0.835	0.822	0.714**
NN2	1.158	1.072	1.065	1.051	0.954	0.939	0.983	1.016	1.085	0.938	0.899*	0.902*
NN3	1.137	1.080	1.061	1.037	0.960	0.932	0.958	0.957	1.108	0.937	0.981	0.935
H-NN1	0.605***	0.549***	0.522***	0.552***	0.883	0.832	0.828	0.868	1.158	1.175	1.224	1.281
H-NN2	0.630***	0.578***	0.544***	0.552***	0.896	0.898	0.867	0.844	1.515	1.589	1.516	1.434
H-NN3	0.615***	0.597***	0.551***	0.542***	0.873	0.881	0.889	0.854	1.531	1.564	1.560	1.447
FANN	0.503***	0.522***	0.500***	0.533***	0.881	0.860	0.827	0.842	1.186	0.958	0.953	0.935
FANN2	0.520***	0.510***	0.492***	0.501***	0.869	0.867	0.847	0.86	1.122	0.876	0.926	0.874
FANN3	0.523***	0.520***	0.497***	0.497***	0.870	0.878	0.879	0.871	1.179	0.899	0.892	0.854

Note: ***, **, and * denote the significance at 1, 5, and 10% levels based on the predictive accuracy test of Diebold and Mariano (1995), respectively.

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.

——— (2009): "Boosting diffusion indices," *Journal of Applied Econometrics*, 24, 607–629.

——— (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 (2019): "Bond risk premia with machine learning," *USC-INET Research Paper*.

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

Breiman, L. (1996): "Bagging predictors," *Machine learning*, 24, 123–140.

——— (2001): "Random forests," *Machine learning*, 45, 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984): *Classification and regression trees*, CRC press.

Bühlmann, P. and B. Yu (2003): "Boosting with the L 2 loss: regression and classification," *Journal of the American Statistical Association*, 98, 324–339.

Bühlmann, P., B. Yu, et al. (2002): "Analyzing bagging," *The Annals of Statistics*, 30, 927–961.

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 (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.

FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2000): "The generalized dynamic-factor model: Identification and estimation," *Review of Economics and statistics*, 82, 540–554.

FORNI, M., M. HALLIN, M. LIPPI, AND P. ZAFFARONI (2015): "Dynamic factor models with infinite-dimensional factor spaces: One-sided representations," *Journal of econometrics*, 185, 359–371.

FORNI, M. AND M. LIPPI (2001): "The generalized dynamic factor model: representation theory," *Econometric theory*, 1113–1141.

FREUND, Y. AND R. E. SCHAPIRE (1995): "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, Springer, 23–37.

FRIEDMAN, J. H. (2001): "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, 1189–1232.

GIACOMINI, R. AND H. WHITE (2006): "Tests of conditional predictive ability," *Econometrica*, 74, 1545–1578.

GU, S., B. KELLY, AND D. XIU (2020): "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 33, 2223–2273.

HANSEN, C. AND Y. LIAO (2019): "The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications," *Econometric Theory*, 35, 465–509.

HE, K., X. ZHANG, S. REN, AND J. SUN (2016): "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

HOERL, A. E. AND R. W. KENNARD (1970): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

HORNIK, K., M. STINCHCOMBE, H. WHITE, ET AL. (1989): "Multilayer feedforward networks are universal approximators." *Neural networks*, 2, 359–366.

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.

——— (2016): "A Factor Analysis of Bond Risk Premia," *Handbook of Empirical Economics and Finance*, 313.

Masters, T. (1993): *Practical neural network recipes in C++*, Morgan Kaufmann.

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," .

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.

——— (2012): "Generalized shrinkage methods for forecasting using many predictors," *Journal of Business & Economic Statistics*, 30, 481–493.

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

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67, 301–320.

Appendices

Appendix A Proof of Theorem 3.1

Throughout the proofs, $T \to \infty$ and N may either grow simultaneously with T or stay constant. For two fixed dimensions matrices A and B, and a sequence a_T, we can say $\|A - B\|_F = o_P(a_T)$ when $A = B + o_P(a_T)$.

Theorem A.1. *Consider the conventional factor model. Under the Assumption 3.1, 3.2-3.4, when* $\|\mathbf{\Sigma}_u^{-1}\|_1 = O(1)$,

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

for $\delta_{N,T} = \sqrt{\frac{\log N}{T}} + \sqrt{\frac{J}{T}} + \frac{1}{\sqrt{N}}$.

Proof. Let $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{\Lambda}}\widehat{\mathbf{F}}'$, which is the residual from the Projected-PC (PPC) method. Note that, with the similar proofs as those of Fan et al. (2016) and Fan et al. (2013), we have $\max_{i \le N} \|\widehat{\lambda}_i - \mathbf{M}'\lambda_i\| = O_P(\sqrt{\frac{J}{T}})$. Then $\max_{i \le N} \frac{1}{T}\sum_{t=1}^T (\widehat{u}_{it} - u_{it})^2 = O_P(\frac{1}{N} + \frac{J}{T})$, which also implies

$$\max_{i \le N, j \le N} |\frac{1}{T}\sum_{t=1}^T (\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})| = O_P(\frac{1}{\sqrt{N}} + \sqrt{\frac{J}{T}}).$$

In addition, $\max_{\le N, j \le N} |\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - \Sigma_{u,ij}| = O_P(\sqrt{\frac{\log N}{T}})$. Then we have, for $\delta_{N,T} = \sqrt{\frac{\log N}{T}} + \sqrt{\frac{J}{T}} + \frac{1}{\sqrt{N}}$,

$$\max_{i \le N, j \le N} |\frac{1}{T}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt} - \Sigma_{u,ij}| = O_P(\delta_{N,T}).$$

By Theorem 5 of Fan et al. (2013) using sparsity property, we then have $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_1 = O_P(m_N \delta_{N,T}^{1-q})$. For the second statement, we have

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1 &\le \|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{\Sigma}_u^{-1}\|_1 + \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{\Sigma}_u^{-1}\|_1 \\
&\le \|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\|_1 \|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_1 \|\mathbf{\Sigma}_u^{-1}\|_1 + \|\mathbf{\Sigma}_u^{-1}\|_1^2 \|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_1 \\
&= O_P(m_N \delta_{N,T}^{1-q}) \|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\|_1 + O_P(m_N \delta_{N,T}^{1-q}).
\end{aligned}$$

Therefore, we have $(1 + o_P(1))\|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\|_1 = O_P(m_N \delta_{N,T}^{1-q})$, and it implies the results. □

A.1 Convergence of loadings

Let $\widetilde{\mathbf{Y}} = \widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\mathbf{Y}$, $\widetilde{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\boldsymbol{\Lambda}$, and $\widetilde{\mathbf{U}} = \widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\mathbf{U}$. Then the regular factor model (3.4) can be written as

$$\widetilde{\mathbf{Y}} = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}' + \widetilde{\mathbf{U}}.$$

Let $\mathbf{K}$ denote a $K \times K$ diagonal matrix of the first K eigenvalues of $\frac{1}{NT}\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}'$. Then by definition of eigenvalues, we have

$$\frac{1}{NT}\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\boldsymbol{\Lambda}}} = \widehat{\widetilde{\boldsymbol{\Lambda}}}\mathbf{K}.$$

Let $\mathbf{M} = \frac{1}{NT}\mathbf{F}'\mathbf{P}\mathbf{F}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}\mathbf{K}^{-1}$. Then

$$\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M} = \sum_{i=1}^{3}\mathbf{A}_i\mathbf{K}^{-1}, \tag{A.1}$$

where

$$\mathbf{A}_1 = \frac{1}{NT}\widetilde{\boldsymbol{\Lambda}}\mathbf{F}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}, \quad \mathbf{A}_2 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}, \quad \mathbf{A}_3 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\mathbf{F}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}.$$

Lemma A.1. $\|\mathbf{K}\|_2 = O_P(1) = \|\mathbf{K}^{-1}\|_2$, *and* $\|\mathbf{M}\|_2 = O_P(1)$.

Proof. Note that $\mathbf{K}$ is the diagonal matrix of the first K eigenvalue of

$$\frac{1}{NT}\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}' = \frac{1}{NT}\widetilde{\mathbf{Y}}\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\widetilde{\mathbf{Y}}'.$$

Then the eigenvalues of $\mathbf{K}$ are the same as those of

$$\mathbf{W} = \frac{1}{NT}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}\boldsymbol{\Phi}(\mathbf{X})'\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}$$

By substituting $\widetilde{\mathbf{Y}} = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}' + \widetilde{\mathbf{U}}$, and $\frac{1}{N}\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}} = \mathbf{I}_K$, we have $\mathbf{W} = \sum_{i=1}^{4}\mathbf{W}_i$, where

$$\begin{aligned}
\mathbf{W}_1 &= \frac{1}{T}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}\boldsymbol{\Phi}(\mathbf{X})'\mathbf{F}\mathbf{F}'\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2},\\
\mathbf{W}_2 &= \frac{1}{T}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}\boldsymbol{\Phi}(\mathbf{X})'\frac{\mathbf{F}\widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{U}}}{N}\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2},\\
\mathbf{W}_3 &= \frac{1}{T}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}\boldsymbol{\Phi}(\mathbf{X})'\frac{\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\mathbf{F}'}{N}\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2} = \mathbf{W}_2',\\
\mathbf{W}_4 &= \frac{1}{T}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}\boldsymbol{\Phi}(\mathbf{X})'\frac{\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}}{N}\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}.
\end{aligned}$$

Note that, Assumption 3.2 and 3.3 implies that $\|\boldsymbol{\Phi}(\mathbf{X})\|_2 = \lambda_{\max}^{1/2}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X})) = O_P(\sqrt{T})$, $\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1/2}\|_2 = \lambda_{\max}^{1/2}((\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}) = O_P(1/\sqrt{T})$. $\|\mathbf{P}\mathbf{F}\|_2 = \lambda_{\max}^{1/2}(\frac{1}{T}\mathbf{F}'\mathbf{P}\mathbf{F})\sqrt{T} =$

$O_P(\sqrt{T})$. By Lemma A.3,

$$\begin{aligned}
\|\mathbf{W}_2\|_2 &\leq \frac{1}{NT}\|(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}\|_2^2\|\mathbf{\Phi}(\mathbf{X})\|_2\|\mathbf{F}\|_F\|\widetilde{\mathbf{\Lambda}}'\widetilde{\mathbf{U}}\mathbf{\Phi}(\mathbf{X})\|_F \\
&\leq O_P(\frac{1}{NT})(\|\mathbf{\Lambda}'(\widehat{\mathbf{\Sigma}}_u^{-1}-\mathbf{\Sigma}_u^{-1})\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F+\|\mathbf{\Lambda}'\mathbf{\Sigma}_u^{-1}\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F) \\
&= O_P\left(\sqrt{\frac{J}{NT}}+\frac{\sqrt{J\log N}}{T}\right), \\
\|\mathbf{W}_4\|_2 &\leq \frac{1}{NT}\|(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}\|_2^2\|\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\|_F^2\|\widehat{\mathbf{\Sigma}}_u^{-1/2}\|_2^2 \\
&= O_P\left(\frac{J}{T}\right).
\end{aligned}$$

For the kth eigenvalue, we have $|\lambda_k(\mathbf{W})-\lambda_k(\mathbf{W}_1)\leq\|\mathbf{W}-\mathbf{W}_1\|_2=o_P(1)$. Hence it suffices to prove that the first K eigenvalues of $\mathbf{W}_1$ are bounded away from zero and infinity, which are also the first K eigenvalues of $\frac{1}{T}\mathbf{F}'\mathbf{P}\mathbf{F}$. This is assumed in Assumption 3.2. Therefore, $\|\mathbf{K}^{-1}\|_2=O_P(1)=\|\mathbf{K}\|_2$, and this also implies $\|\mathbf{M}\|_2=O_P(1)$. □

Lemma A.2. *(i)* $\|\mathbf{A}_1\|_F^2=O_P(JN/T)$, *(ii)* $\|\mathbf{A}_2\|_F^2=O_P(J^2N/T^2)$, *(iii)* $\|\mathbf{A}_3\|_F^2=O_P(JN/T)$.

Proof. Note that $\|\widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\|_2=O_P(1)$, $\|\widetilde{\mathbf{\Lambda}}\|_F^2=\|\widehat{\mathbf{\Sigma}}_u^{-1/2}\mathbf{\Lambda}\|_F^2=O_P(N)=\|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F^2$ and Assumption 3.2 implies $\|\mathbf{P}\mathbf{F}\|_2^2=O_P(T)$. By Lemma A.3,

$$\begin{aligned}
\|\mathbf{A}_1\|_F^2 &= \left\|\frac{1}{NT}\widetilde{\mathbf{\Lambda}}\mathbf{F}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}\right\|_F^2 \leq \frac{1}{N^2T^2}\|\widetilde{\mathbf{\Lambda}}\|_F^2\|\mathbf{F}'\mathbf{P}\|_2^2\|\mathbf{P}\mathbf{U}'\|_F^2\|\widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\|_2^2\|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F^2=O_P(JN/T), \\
\|\mathbf{A}_2\|_F^2 &= \left\|\frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}\right\|_F^2 \leq \frac{1}{N^2T^2}\|\mathbf{P}\mathbf{U}'\|_F^4\|\widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\|_2^4\|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F^2=O_P(J^2N/T^2), \\
\|\mathbf{A}_3\|_F^2 &= \left\|\frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\mathbf{F}\widetilde{\mathbf{\Lambda}}'\widehat{\widetilde{\mathbf{\Lambda}}}\right\|_F^2 \leq \frac{1}{N^2T^2}\|\widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\|_2^2\|\mathbf{U}\mathbf{P}\|_F^2\|\mathbf{P}\mathbf{F}\|_2^2\|\widetilde{\mathbf{\Lambda}}\|_F^2\|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F^2=O_P(JN/T).
\end{aligned}$$

□

Therefore, since $\|\widehat{\mathbf{\Sigma}}_u\|_2<\infty$, it follows from Lemmas A.1 and A.2 that

$$\frac{1}{N}\|\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}\mathbf{M}\|_F^2\leq O_P\left(\frac{1}{N}\right)\|\widehat{\widetilde{\mathbf{\Lambda}}}-\widetilde{\mathbf{\Lambda}}\mathbf{M}\|_F^2\leq O_P\left(\frac{1}{N}\|\mathbf{K}^{-1}\|_2\right)(\|\mathbf{A}_1\|_F^2+\|\mathbf{A}_2\|_F^2+\|\mathbf{A}_3\|_F^2)=O_P(J/T).$$

Lemma A.3. *(i)* $\max_{i\leq N}\max_{t\leq T}\sum_{s=1}^T|Eu_{it}u_{is}|=O(1)$,
$\max_{k\leq K,t\leq T}\max_{i\leq N}\sum_{j=1}^N|\mathrm{cov}(\lambda_{ik}u_{it},\lambda_{jk}u_{jt})|=O(1)$,
$\max_{t\leq T,s\leq T}\max_{i\leq N}\sum_{j=1}^N|\mathrm{cov}(u_{it}u_{is},u_{jt}u_{js})|=O(1)$,
(ii) $\|\mathbf{U}'\mathbf{\Lambda}\|_F=O_P(\sqrt{NT})$.
(iii) $\|\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F^2=O_P(JNT)$, $\|\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda}\|_F^2=O_P(JNT)$, $\|\mathbf{P}\mathbf{U}'\|_F^2=O_P(JN)$.
(iv) $\|\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\mathbf{\Sigma}_u^{-1}\mathbf{\Lambda}\|_F^2=O_P(JNT)$, $\|\mathbf{U}'\mathbf{\Sigma}_u^{-1}\mathbf{\Lambda}\|_F^2=O_P(NT)$.

Proof. (i) The results follow from Davydov's inequality, which are similar to Lemma B.1 in the supplementary material of Fan et al. (2016).
(ii) From part (i), we have

$$
\begin{aligned}
E\|\mathbf{U}'\mathbf{\Lambda}\|_F^2 &= E\sum_{k=1}^K\sum_{t=1}^T(\sum_{i=1}^N u_{it}\lambda_{ik})^2 = \sum_{k=1}^K\sum_{t=1}^T \mathrm{var}(\sum_{i=1}^N u_{it}\lambda_{ik}) \\
&= \sum_{k=1}^K\sum_{t=1}^T[\sum_{i=1}^N \mathrm{var}(u_{it}\lambda_{ik}) + \sum_{i\le N, j\le N}\mathrm{cov}(u_{it}\lambda_{ik}, u_{jt}\lambda_{jk})] = O(NT).
\end{aligned}
$$

(iii) From part (i) and Assumptions 3.3 (ii) and 3.4 (iv),

$$
\begin{aligned}
E\|\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F^2 &= \sum_{i=1}^N\sum_{j=1}^J\sum_{l=1}^d E(\sum_{t=1}^T \phi_j(X_{tl})u_{it})^2 = \sum_{i=1}^N\sum_{j=1}^J\sum_{l=1}^d\sum_{t=1}^T\sum_{s=1}^T E\phi_j(X_{tl})\phi_j(X_{sl})Eu_{it}u_{is} \\
&\le JdT\max_{j\le J, l\le d, t,s\le T}|E\phi_j(X_{tl})\phi_j(X_{sl})|\sum_{i=1}^N\max_{s\le T}\sum_{t=1}^T|Eu_{it}u_{is}| \\
&\le JdT\max_{j\le J, l\le d, t,s\le T}E\phi_j(X_{tl})^2\sum_{i=1}^N\max_{s\le T}\sum_{t=1}^T|Eu_{it}u_{is}| = O(JNT),
\end{aligned}
$$

where the Cauchy Schwarz inequality implies the second inequality. On the other hand,

$$
\begin{aligned}
&E\|\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda}\|_F^2 = \sum_{k=1}^K\sum_{j=1}^J\sum_{l=1}^d E(\sum_{t=1}^T\sum_{i=1}^N \phi_j(X_{tl})u_{it}\lambda_{ik})^2 \\
&= \sum_{k=1}^K\sum_{j=1}^J\sum_{l=1}^d \mathrm{var}(\sum_{t=1}^T\sum_{i=1}^N \phi_j(X_{tl})u_{it}\lambda_{ik}) \\
&= \sum_{k=1}^K\sum_{j=1}^J\sum_{l=1}^d\sum_{t=1}^T\sum_{i=1}^N \mathrm{var}(\phi_j(X_{tl})u_{it}\lambda_{ik}) + \sum_{k=1}^K\sum_{j=1}^J\sum_{l=1}^d\sum_{t=1}^T\sum_{i\ne m; i,m\le N}(E\phi_j(X_{tl})^2\lambda_{ik}\lambda_{mk})Eu_{it}u_{mt} \\
&\quad + \sum_{k=1}^K\sum_{j=1}^J\sum_{l=1}^d\sum_{t\ne s; t,s\le T}\sum_{i,m\le N}(E\phi_j(X_{tl})\phi_j(X_{sl})\lambda_{ik}\lambda_{mk})Eu_{it}u_{ms} \\
&\le O(JNT) + O(JNT)\max_{t\le T, i\le N}\sum_{m=1}^N|Eu_{it}u_{mt}| + O(JNT)\frac{1}{NT}\sum_{i,m\le N}\sum_{t,s\le T}|Eu_{it}u_{ms}| = O(JNT).
\end{aligned}
$$

Finally, since $\|(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1}\|_2 = \lambda_{\min}^{-1}(T^{-1}\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))T^{-1} = O_P(T^{-1})$,

$$
\|\mathbf{P}\mathbf{U}'\|_F \le \|\mathbf{\Phi}(\mathbf{X})\|_2\|(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1}\|_2\|\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\|_F = O_P(\sqrt{JN}).
$$

(iv) Note that we have derived the same rates for the case $\mathbf{\Sigma}_u^{-1} = \mathbf{I}$ in parts (ii) and (iii).

Here we only need to define $\boldsymbol{\Lambda}^* = \boldsymbol{\Sigma}_u^{-\frac{1}{2}}\boldsymbol{\Lambda}$ and $\mathbf{U}^* = \boldsymbol{\Sigma}_u^{-\frac{1}{2}}\mathbf{U}$ and prove $\boldsymbol{\Lambda}^*$ and $\mathbf{U}^*$ possess the same properties as $\boldsymbol{\Lambda}$ and $\mathbf{U}$. The results follow using the same argument as in parts (ii) and (iii) that $\boldsymbol{\Sigma}_u^{-\frac{1}{2}}\boldsymbol{\Lambda}$ has bounded row sums. □

Lemma A.4. *In the conventional factor model,*
(i) $\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_1\|_F = O_P(\frac{JN}{T} + N\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$,
$\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_2\|_F = O_P(\frac{JN}{T}m_N^2\delta_{N,T}^{2-2q} + \frac{JN}{T}\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$,
$\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_3\|_F = O_P(N\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$.
(ii) $\frac{1}{N}\|\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F = O_P(\frac{J}{T}+\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$, *and* $\frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F = O_P(\frac{J}{T}+\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$.

Proof. Note that $\|\boldsymbol{\Phi}'\mathbf{U}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Lambda}\|_F = O_P(\sqrt{JNT})$ and

$$\|\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\|_F \leq \|\boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\|_2(\|\boldsymbol{\Phi}'\mathbf{U}(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\boldsymbol{\Lambda})\|_F + \|\boldsymbol{\Phi}'\mathbf{U}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Lambda}\|_F) \leq O_P(N\sqrt{J}m_N\delta_{N,T}^{1-q}).$$

Hence

$$\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_1\|_F \leq \frac{1}{NT}\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}}\mathbf{F}'\mathbf{P}\|_F(\|\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\mathbf{M}\|_F + \|\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \boldsymbol{\Lambda}\mathbf{M})\|_F) = O_P\left(\frac{JN}{T} + N\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q}\right),$$

$$\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_2\|_F \leq \frac{1}{NT}\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{U}}\mathbf{P}\|_F(\|\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\mathbf{M}\|_F + \|\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F) = O_P\left(\frac{JN}{T}m_N^2\delta_{N,T}^{2-2q} + \frac{JN}{T}\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q}\right),$$

$$\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_3\|_F \leq \frac{1}{NT}\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}}\mathbf{F}'\mathbf{P}\|_F\|\mathbf{P}\mathbf{F}\|_2\|\widetilde{\boldsymbol{\Lambda}}\|_F\|\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F = O_P\left(N\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q}\right).$$

For part (ii),

$$\frac{1}{N}\|\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F \leq \frac{1}{N}\|\mathbf{K}^{-1}\|_2\sum_{i=1}^{3}\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_i\|_F = O_P(J/T + \sqrt{J/T}m_N\delta_{N,T}^{1-q}).$$

In addition, the result follows from the following inequality:

$$\frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F \leq \frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M}\|_F^2 + \frac{1}{N}\|\mathbf{M}'\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F = O_P(\frac{J}{T} + \sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$$

□

Lemma A.5. $\|\mathbf{M}'\mathbf{M} - \mathbf{I}_K\|_F = O_P(\frac{J}{T} + \sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$. *Therefore,* $\|\mathbf{M}^{-1}\|_2 = O_P(1)$.

Proof. Note that $N^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = N^{-1}\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}} = N^{-1}\widehat{\widetilde{\boldsymbol{\Lambda}}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}$, by the identification condition. Then

$$\mathbf{M}'\mathbf{M} = \frac{1}{N}(\widetilde{\boldsymbol{\Lambda}}\mathbf{M})'\widetilde{\boldsymbol{\Lambda}}\mathbf{M} = \frac{1}{N}(\widetilde{\boldsymbol{\Lambda}}\mathbf{M} - \widehat{\widetilde{\boldsymbol{\Lambda}}})'\widetilde{\boldsymbol{\Lambda}}\mathbf{M} + \frac{1}{N}\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widetilde{\boldsymbol{\Lambda}}\mathbf{M} - \widehat{\widetilde{\boldsymbol{\Lambda}}}) + \mathbf{I}_K.$$

This implies the following convergence rate,

$$\|\mathbf{M}'\mathbf{M} - \mathbf{I}_K\|_F \leq \frac{1}{N}\|(\widetilde{\boldsymbol{\Lambda}}\mathbf{M} - \widehat{\widetilde{\boldsymbol{\Lambda}}})'\widetilde{\boldsymbol{\Lambda}}\|_F\|\mathbf{M}\|_2 + \frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widetilde{\boldsymbol{\Lambda}}\mathbf{M} - \widehat{\widetilde{\boldsymbol{\Lambda}}})\|_F = O_P(\frac{J}{T} + \sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q}).$$

In addition, it implies $\lambda_{\min}(\mathbf{M}'\mathbf{M}) \geq 1 - o_P(1)$. Therefore,

$$\|\mathbf{M}^{-1}\|_2^2 = \lambda_{\max}(\mathbf{M}^{-1}(\mathbf{M}^{-1})') = \lambda_{\max}((\mathbf{M}'\mathbf{M})^{-1}) = \lambda_{\min}(\mathbf{M}'\mathbf{M}) = O_P(1).$$

□

Lemma A.6. *(i)* $\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_1\|_F^2 = O_P(J/N + (J/T)m_N^2\delta_{N,T}^{2-2q})$,
$\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_2\|_F^2 = O_P(J^2/NT + J/T^2)$, $\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_3\|_F^2 = O_P(J/N + 1/T)$.
(ii) $\|\frac{1}{N}\widetilde{\mathbf{U}}(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F^2 = O_P(J/N + 1/T)$.

Proof. (i) Note that, by Lemmas A.3 and B.9,

$$\|\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\|_F^2 \leq \|\mathbf{U}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\boldsymbol{\Lambda}\|_F^2 + \|\mathbf{U}'\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Lambda}\|_F^2 = O_P(NT + (N^2 + NT)m_N^2\delta_{N,T}^{2-2q}),$$

$$\|\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 \leq \|\mathbf{U}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 + \|\mathbf{U}'\boldsymbol{\Sigma}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 = O_P(JNT^2 + N^2T).$$

Hence

$$\begin{aligned}
&\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_1\|_F^2 \leq \frac{1}{N^4T^2}\|\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\|_F^2\|\mathbf{F}'\mathbf{P}\|_2^2\|\mathbf{P}\mathbf{U}'\|_F^2\|\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\|_2^2\|\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 = O_P(J/N + (J/T)m_N^2\delta_{N,T}^{2-2q}),\\
&\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_2\|_F^2 \leq \frac{1}{N^4T^2}\|\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2^2\|\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\|_2^2\|\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 = O_P(J^2/NT + J/T^2),\\
&\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_3\|_F^2 \leq \frac{1}{N^4T^2}\|\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\|_2^2\|\mathbf{P}\mathbf{F}\|_2^2\|\widetilde{\boldsymbol{\Lambda}}\|_F^2\|\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 = O_P(J/N + 1/T).
\end{aligned}$$

Part (ii) follows from part (i). □

A.2 Convergence of factors

For the estimated $\widehat{\mathbf{G}}(\mathbf{X})$, note that

$$\widehat{\mathbf{G}}(\mathbf{X}) = \frac{1}{N}\mathbf{P}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\boldsymbol{\Lambda}}} = \frac{1}{N}\mathbf{P}\mathbf{F}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}} + \frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}} = \mathbf{P}\mathbf{F}\mathbf{M} + \mathbf{E},$$

where $\mathbf{E} = \frac{1}{N}\mathbf{P}\mathbf{F}\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M}) + \frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M}) + \frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}\mathbf{M}$.
By Lemma A.4,

$$\left\|\frac{1}{N}\mathbf{P}\mathbf{F}\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\right\|_F \leq O_P\left(\frac{\sqrt{T}}{N}\right)\|\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{M})\|_F = O_P\left(\frac{J}{\sqrt{T}} + \sqrt{J}m_N\delta_{N,T}^{1-q}\right).$$

By Lemma A.3, $\|\frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M})\|_F \leq \frac{1}{N}\|\mathbf{P}\mathbf{U}'\|_F\|\widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\|\|\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M}\|_F = O_P(J/\sqrt{T})$, and $\|\frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{M}\|_F = O_P(\sqrt{J}m_N\delta_{N,T}^{1-q})$. Hence,

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

As for the estimated factor matrix $\widehat{\mathbf{F}}$, note that $\widehat{\mathbf{F}} = \frac{1}{N}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}}$. Substituting $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{\Lambda}}\mathbf{F}' + \widetilde{\mathbf{U}}$,

$$\widehat{\mathbf{F}} = \mathbf{FM} + \sum_{i=1}^{3}\mathbf{B}_i,$$

where

$$\mathbf{B}_1 = \frac{1}{N}\mathbf{F}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M}),\ \ \mathbf{B}_2 = \frac{1}{N}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M}),\ \ \mathbf{B}_3 = \frac{1}{N}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{M}.$$

Note that $\|\mathbf{U}\|^2 = O_P(N+T)$. By Lemmas A.3, A.4, and A.6,

$$\|\mathbf{B}_1\|_F^2 = O_P\left(\frac{J^2}{T} + Jm_N^2\delta_{N,T}^{2-2q}\right), \|\mathbf{B}_2\|_F^2 = O_P\left(\frac{J}{N} + \frac{1}{T}\right), \|\mathbf{B}_3\|_F^2 = O_P\left(\frac{T}{N} + \frac{T^2}{N^2}m_N^2\delta_{N,T}^{2-2q}\right).$$

Therefore,

$$\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{FM}\|_F^2 \leq O(\frac{1}{T})\sum_{i=1}^{3}\|\mathbf{B}_i\|_F^2 = O_P\left(\frac{1}{N} + \frac{J^2}{T^2} + \left(\frac{J}{T} + \frac{T}{N^2}\right)m_N^2\delta_{N,T}^{2-2q}\right).$$

A.3 Individual factors

Since $\widehat{\mathbf{F}} = \frac{1}{N}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}}$, $\widehat{\mathbf{F}}_t = \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widetilde{\mathbf{\Lambda}}\mathbf{F}_t + \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widetilde{\mathbf{u}}_t$. Using $\widetilde{\mathbf{\Lambda}} = \widetilde{\mathbf{\Lambda}} - \widehat{\widetilde{\mathbf{\Lambda}}}\mathbf{M}^{-1} + \widehat{\widetilde{\mathbf{\Lambda}}}\mathbf{M}^{-1}$ and $\frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widehat{\widetilde{\mathbf{\Lambda}}} = \mathbf{I}_K$, we have

$$\widehat{\mathbf{F}}_t - \mathbf{M}^{-1}\mathbf{F}_t = \sum_{i=1}^{3}\mathbf{D}_i,$$

where

$$\mathbf{D}_1 = \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'(\widetilde{\mathbf{\Lambda}}\mathbf{M} - \widehat{\widetilde{\mathbf{\Lambda}}})\mathbf{M}^{-1}\mathbf{F}_t, \quad \mathbf{D}_2 = \frac{1}{N}(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M})'\widetilde{\mathbf{u}}_t, \quad \mathbf{D}_3 = \frac{1}{N}\mathbf{M}'\widetilde{\mathbf{\Lambda}}'\widetilde{\mathbf{u}}_t.$$

Then, by Lemmas A.4 and A.5,

$$\|\mathbf{D}_1\| \leq \|\frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F\|\mathbf{M}^{-1}\|_2\|\mathbf{F}_t\| = O_P\left(\frac{J}{T} + \sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q}\right).$$

Note that $\|N^{-1}\mathbf{\Lambda}'\mathbf{u}_t\| = O_P(1/\sqrt{N})$ by Lemma A.8. Let $\mathbf{\Lambda}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{\Lambda}$ and $\mathbf{u}_t^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{u}_t$. Then $\|N^{-1}\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\| = O_P(1/\sqrt{N})$.

$$\begin{aligned}
\|\mathbf{D}_3\| &\le \frac{1}{N}\|\mathbf{M}\|_2(\|\mathbf{\Lambda}'(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\mathbf{u}_t\| + \|\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\|) \\
&\le O_P(\frac{1}{N})(\|\mathbf{\Lambda}\|_F\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1\|\mathbf{u}_t\| + \|\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\|) \\
&\le O_P(m_N\delta_{N,T}^{1-q} + \frac{1}{\sqrt{N}}) = O_P(m_N\delta_{N,T}^{1-q}).
\end{aligned}$$

Note that $\mathbf{D}_2 = \|\frac{1}{N}\widetilde{\mathbf{u}}_t'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M})\| \le \|\mathbf{K}^{-1}\|_2\sum_{i=1}^3\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_i\|$, where

$$\mathbf{A}_1 = \frac{1}{NT}\widetilde{\mathbf{\Lambda}}\mathbf{F}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}, \quad \mathbf{A}_2 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}, \quad \mathbf{A}_3 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\mathbf{F}\widetilde{\mathbf{\Lambda}}'\widehat{\widetilde{\mathbf{\Lambda}}}.$$

Then, by Lemma A.7, $\|\mathbf{D}_2\| = O_P(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$.

Therefore, for each $t \le T$,

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

Lemma A.7. *(i)* $\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_1\| = O_P(m_N\delta_{N,T}^{1-q}(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q} + \frac{J}{T}))$,
(ii) $\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_2\| = O_P(m_N\delta_{N,T}^{1-q}(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q} + \frac{J}{T}\sqrt{\frac{J}{T}}))$,
(iii) $\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_3\| = O_P(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q})$.

Proof. (i) Note that $\|\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\| = O_P(\sqrt{N})$ by Lemma A.8, and $\|\mathbf{F}'\mathbf{P}\|_2 = O_P(\sqrt{T})$.

$$\begin{aligned}
\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_1\| &= \frac{1}{N^2T}\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{\Lambda}}\mathbf{F}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}\| \\
&\le \frac{1}{N^2T}\|\mathbf{u}_t'(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\mathbf{\Lambda}\mathbf{F}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}\| + \frac{1}{N^2T}\|\mathbf{u}_t^{*'}\mathbf{\Lambda}^*\mathbf{F}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}\| \\
&\le \frac{1}{N^2T}\|\mathbf{u}_t\|\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1\|\mathbf{\Lambda}\|_F\|\mathbf{F}'\mathbf{P}\|_2(\|\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{M}\|_F + \|\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M})\|_F) \\
&\quad + \frac{1}{N^2T}\|\mathbf{u}_t^{*'}\mathbf{\Lambda}^*\|\|\mathbf{F}'\mathbf{P}\|_2(\|\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{M}\|_F + \|\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M})\|_F) \\
&= O_P((\frac{1}{N\sqrt{T}}m_N\delta_{N,T}^{1-q} + \frac{1}{N\sqrt{NT}})(N\sqrt{J}m_N\delta_{N,T}^{1-q} + \frac{JN}{\sqrt{T}}) \\
&= O_P(m_N\delta_{N,T}^{1-q}(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q} + \frac{J}{T})).
\end{aligned}$$

(ii) Note that, by Lemma A.8,

$$\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{U}}\mathbf{P}\| \le \|\mathbf{u}_t'(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\mathbf{U}\mathbf{P}\| + \|\mathbf{u}_t^{*'}\mathbf{U}^*\mathbf{P}\| = O_P(N\sqrt{J}m_N\delta_{N,T}^{1-q}).$$

Then,

$$\begin{aligned}\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_2\| &= \frac{1}{N^2T}\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\mathbf{\Lambda}}}\| \\ &\leq \frac{1}{N^2T}\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{U}}\mathbf{P}\|(\|\mathbf{P}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{M}\|_F + \|\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{M})\|_F) \\ &= O_P(m_N\delta_{N,T}^{1-q}(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q} + \frac{J}{T}\sqrt{\frac{J}{T}})).\end{aligned}$$

(iii) Note that $\|\mathbf{F}'\mathbf{P}\|_2 = O_P(\sqrt{T})$ and $\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{U}}\mathbf{P}\| = O_P(N\sqrt{J}m_N\delta_{N,T}^{1-q})$. Then,

$$\|\frac{1}{N}\widetilde{\mathbf{u}}_t'\mathbf{A}_3\| = \frac{1}{N^2T}\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{U}}\mathbf{P}\mathbf{F}\widetilde{\mathbf{\Lambda}}'\widehat{\widetilde{\mathbf{\Lambda}}}\| \leq O_P(\frac{1}{NT})\|\widetilde{\mathbf{u}}_t'\widetilde{\mathbf{U}}\mathbf{P}\|\|\mathbf{P}\mathbf{F}\|_2 = O_P(\sqrt{\frac{J}{T}}m_N\delta_{N,T}^{1-q}).$$

□

Lemma A.8. *(i)* $\|\mathbf{\Lambda}'\mathbf{u}_t\|^2 = O_P(N) = \|\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{u}_t\|^2$.
(ii) $\|\mathbf{u}_t'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\| = O_P(N\sqrt{J} + \sqrt{JNT}) = \|\mathbf{u}_t'\mathbf{\Sigma}^{-1}\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|$.

Proof. (i) Note that $|\lambda_k'\mathbf{u}_t|^2 = O_P(1)E|\lambda_k'\mathbf{u}_t|^2 = O_P(1)\text{var}(\lambda_k'\mathbf{u}_t) = O_P(1)\lambda_k'\text{var}(\mathbf{u}_t)\lambda_k \leq O_P(1)\|\lambda_k\|^2\|\text{var}(\mathbf{u}_t)\| = O_P(N)$. Then $\|\mathbf{\Lambda}'\mathbf{u}_t\|^2 = \sum_{k=1}^K(\lambda_k'\mathbf{u}_t)^2 = O_P(N)$.
(ii) Note that $\|\mathbf{u}_t'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\| = \|\sum_{i=1}^N u_{it}\mathbf{u}_i'\mathbf{\Phi}(\mathbf{X})\|$, where $\mathbf{u}_i = (u_{i1}, ..., u_{iT})'$. Also

$$\begin{aligned}&E\|\sum_{i=1}^N u_{it}\mathbf{u}_i'\mathbf{\Phi}(\mathbf{X})\|^2 = \sum_{k=1}^J\sum_{l=1}^d E(\sum_i\sum_s u_{is}u_{it}\phi_k(X_{sl}))^2 \\ &\leq 2\sum_{k=1}^J\sum_{l=1}^d E(\sum_i\sum_s(u_{is}u_{it} - Eu_{is}u_{it})\phi_k(X_{sl}))^2 + 2\sum_{k=1}^J\sum_{l=1}^d E(\sum_i\sum_s(Eu_{is}u_{it})\phi_k(X_{sl}))^2 \\ &\leq 2\sum_{k=1}^J\sum_{l=1}^d \text{var}(\sum_i\sum_s(u_{is}u_{it} - Eu_{is}u_{it})\phi_k(X_{sl})) + 2\sum_{k=1}^J\sum_{l=1}^d N^2E(u_{is}^2)^2E\phi_k(X_{sl}))^2 \\ &\leq O(N^2J) + 2\sum_{k=1}^J\sum_{l=1}^d\sum_{i=1}^N \text{var}(\sum_s(u_{is}u_{it} - Eu_{is}u_{it})\phi_k(X_{sl})) \\ &\quad + 2\sum_{k=1}^J\sum_{l=1}^d\sum_{i\neq j}\sum_{q=1}^T\sum_{s=1}^T E((u_{is}u_{it} - Eu_{is}u_{it})(u_{iq}u_{it} - Eu_{iq}u_{it}))E(\phi_k(X_{sl})\phi_k(X_{ql})) \\ &= O(N^2J) + 2\sum_{k=1}^J\sum_{l=1}^d\sum_{i=1}^N\sum_{s=1}^T \text{var}((u_{is}u_{it} - Eu_{is}u_{it})\phi_k(X_{sl})) \\ &\quad + 2\sum_{k=1}^J\sum_{l=1}^d\sum_{i=1}^N\sum_{q\neq s} \text{cov}((u_{is}u_{it} - Eu_{is}u_{it})\phi_k(X_{sl}), (u_{iq}u_{it} - Eu_{iq}u_{it})\phi_k(X_{ql})) \\ &= O(N^2J + JNT) + 4\sum_{k=1}^J\sum_{l=1}^d\sum_{i=1}^N\sum_{t\neq s} \text{cov}(u_{is}u_{it}, u_{it}u_{it})E(\phi_k(X_{sl})\phi_k(X_{tl})) = O(N^2J + JNT).\end{aligned}$$

□

Appendix B Proofs of Theorem 3.2

Theorem B.1. *Consider the semiparametric factor model. Under the Assumption 3.1, 3.3-3.7, when* $\|\mathbf{\Sigma}_u^{-1}\|_1 = O(1)$,

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

for $\omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$.

Proof. Let $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{\Lambda}}\widehat{\mathbf{F}}'$, which is the residual from the Projected-PC (PPC) method. Note that, with the similar proofs as those of Fan et al. (2016) and Fan et al. (2013), we have $\max_{i \le N} \|\widehat{\lambda}_i - \mathbf{H}'\lambda_i\| = O_P(\frac{1}{\sqrt{T}})$. Then $\max_{i\le N} \frac{1}{T}\sum_{t=1}^T (\widehat{u}_{it} - u_{it})^2 = O_P(\frac{1}{N} + \frac{1}{T})$, which also implies

$$\max_{i\le N, j\le N} |\frac{1}{T}\sum_{t=1}^T (\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt})| = O_P(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{T}}).$$

In addition, $\max_{\le N, j \le N} |\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - \Sigma_{u,ij}| = O_P(\sqrt{\frac{\log N}{T}})$. Then we have, for $\omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,

$$\max_{i\le N, j\le N} |\frac{1}{T}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt} - \Sigma_{u,ij}| = O_P(\omega_{N,T}).$$

By Theorem 5 of Fan et al. (2013) using sparsity property, we then have $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_1 = O_P(m_N\omega_{N,T}^{1-q})$. For the second statement, we have

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1 &\le \|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{\Sigma}_u^{-1}\|_1 + \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\mathbf{\Sigma}_u^{-1}\|_1 \\
&\le \|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\|_1\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_1\|\mathbf{\Sigma}_u^{-1}\|_1 + \|\mathbf{\Sigma}_u^{-1}\|_1^2\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_1 \\
&= O_P(m_N\omega_{N,T}^{1-q})\|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\|_1 + O_P(m_N\omega_{N,T}^{1-q}).
\end{aligned}$$

Therefore, we have $(1+o_P(1))\|(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\|_1 = O_P(m_N\omega_{N,T}^{1-q})$, and it implies the results. □

B.1 Convergence of loadings.

Recall that $\mathbf{K}$ denote the $K \times K$ diagonal matrix consisting the first K largest eigenvalues of $(NT)^{-1}\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}'$, where $\widetilde{\mathbf{Y}} = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{Y}$, in descending order. By the definition of eigenvalues, we have

$$\frac{1}{NT}(\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}')\widehat{\widetilde{\boldsymbol{\Lambda}}} = \widehat{\widetilde{\boldsymbol{\Lambda}}}\mathbf{K}.$$

Let

$$\mathbf{H} = \frac{1}{NT}(\mathbf{Q}\mathbf{Q}'\widetilde{\boldsymbol{\Lambda}}' + \mathbf{Q}\widetilde{\mathbf{U}}')\widehat{\widetilde{\boldsymbol{\Lambda}}}\mathbf{K}^{-1},$$

where $\mathbf{Q} = \mathbf{B}\boldsymbol{\Phi}(\mathbf{X})' + \boldsymbol{\Gamma}'\mathbf{P} + \mathbf{R}(\mathbf{X})'\mathbf{P}$. We shall show that $\|\mathbf{H}\|_2 = O_P(1)$ in Lemma B.3. Note that $\widetilde{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\boldsymbol{\Lambda}$, and $\widetilde{\mathbf{U}} = \widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\mathbf{U}$. Substituting $\widetilde{\mathbf{Y}} = \widetilde{\boldsymbol{\Lambda}}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})' + \widetilde{\boldsymbol{\Lambda}}\mathbf{R}(\mathbf{X})' + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}' + \widetilde{\mathbf{U}}$, we have

$$\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{H} = \left(\sum_{i=1}^{4}\mathbf{A}_i\right)\mathbf{K}^{-1},$$

where

$$\mathbf{A}_1 = \frac{1}{NT}\widetilde{\mathbf{U}}\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\;\; \mathbf{A}_2 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\mathbf{R}(\mathbf{X})\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\;\; \mathbf{A}_3 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\;\; \mathbf{A}_4 = \frac{1}{NT}\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}.$$

To show the convergence of $\widehat{\boldsymbol{\Lambda}}$, note that there is a constant $C > 0$, so that

$$\frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{H}\|_F^2 \le C\|\mathbf{K}^{-1}\|_2^2\sum_{i=1}^{4}\frac{1}{N}\|\mathbf{A}_i\|_F^2.$$

Hence we need to bound $\frac{1}{N}\|\mathbf{D}_i\|_F^2$ for $i = 1, ..., 4$. The following lemma gives the stochastic bounds for individual terms.

Lemma B.1. *(i)* $\frac{1}{N}\|\mathbf{A}_1\|_F^2 = O_P(T^{-1})$,
(ii) $\frac{1}{N}\|\mathbf{A}_2\|_F^2 = O_P(T^{-1}J^{1-\kappa})$,
(iii) $\frac{1}{N}\|\mathbf{A}_3\|_F^2 = O_P(J^2 v_T/T^2)$,
(iv) $\frac{1}{N}\|\mathbf{A}_4\|_F^2 = O_P(J^2/T^2)$,

Proof. (i) Because $\|\widetilde{\boldsymbol{\Lambda}}\|_F^2 = O_P(N) = \|\boldsymbol{\Lambda}\|_F^2$, $\|\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 = O_P(N)$. Note that $\|\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\|_2 = O_P(1)$. By Lemma B.2, $\|\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT)$. Then $\frac{1}{N}\|\mathbf{A}_1\|_F^2 = O_P(T^{-1})$.
(ii) Note that $\|\mathbf{R}(\mathbf{X})\|_F^2 = O_P(TJ^{-\kappa})$. By Lemma A.3, $\|\mathbf{U}'\boldsymbol{\Phi}(\mathbf{X})\|_F = O_P(\sqrt{JNT})$. By Assumption 3.3, $\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2 = O_P(T^{-1})$. Then,

$$\|\mathbf{A}_2\|_F \le \frac{1}{NT}\|\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\|_2\|\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2\|\boldsymbol{\Phi}(\mathbf{X})\|_2\|\mathbf{R}(\mathbf{X})\|_F\|\widetilde{\boldsymbol{\Lambda}}\|_F\|\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F = O_P\left(\sqrt{\frac{JN}{TJ^{\kappa}}}\right).$$

Therefore, $\frac{1}{N}\|\mathbf{A}_2\|_F^2 = O_P(T^{-1}J^{1-k})$.
(iii) It follows from Lemma B.2 that $\|\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Gamma}\|_F^2 = O_P(JTv_T)$. Then
$\frac{1}{N}\|\mathbf{A}_3\|_F^2 = O_P(\frac{1}{NT^4}\|\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Gamma}\|_F^2) = O_P(J^2 v_T/T^2)$.
(iv) By Lemma A.3 and $\|\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}\|_2 = O_P(1)$, $\frac{1}{N}\|\mathbf{A}_4\|_F^2 = O_P(J^2/T^2)$. □

By Lemma B.3, $\|\mathbf{K}^{-1}\|_2 = O_P(1)$. Note that $\|\widehat{\mathbf{\Sigma}}_u\|_2 < \infty$. Therefore, as $J = o(\sqrt{T})$ and $\kappa \geq 1$,

$$\frac{1}{N}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\mathbf{H}\|_F^2 \leq O_P\left(\frac{1}{N}\right)\|\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}\|_F^2 \leq O_P\left(\frac{1}{N}\|\mathbf{K}^{-1}\|_2^2\right)\sum_{i=1}^4 \|\mathbf{A}_i\|_F^2 = O_P(1/T).$$

Lemma B.2. *(i)* $\|\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT)$, $\|\mathbf{\Lambda}'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT) = \|\mathbf{\Lambda}'\mathbf{\Sigma}_u^{-1}\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2$. *(ii)* $\|\mathbf{\Phi}(\mathbf{X})'\mathbf{\Gamma}\|_F^2 = O_P(JTv_T)$, $\|\mathbf{B}\mathbf{\Phi}(\mathbf{X})'\mathbf{\Gamma}\|_F^2 = O_P(T\nu_T)$.

Proof. (i) Note that $\mathbf{B}\mathbf{\Phi}(\mathbf{X})'\mathbf{U}' = \mathbf{G}(\mathbf{X})'\mathbf{U}' - \mathbf{R}(\mathbf{X})'\mathbf{U}'$ and $\mathbf{X}_t$ and $\mathbf{u}_i$ are independent. Then

$$\begin{aligned}
E\|\mathbf{G}(\mathbf{X})'\mathbf{U}'\|_F^2 &= \sum_{k=1}^K\sum_{i=1}^N E(\sum_{t=1}^T g_k(\mathbf{X}_t)u_{it})^2 \\
&= \sum_{k=1}^K\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T Eg_k(\mathbf{X}_t)g_k(\mathbf{X}_s)Eu_{it}u_{is} \\
&\leq TK\max_{k\leq K} Eg_k(\mathbf{X}_t)^2\sum_{i=1}^N\max_{s\leq T}\sum_{t=1}^T|Eu_{it}u_{is}| = O(NT).
\end{aligned}$$

Note that $R_{tk} \equiv \sum_{l=1}^d R_{kl}(X_{tl})$ is (t,k)th element of $\mathbf{R}(\mathbf{X})$. Then

$$\begin{aligned}
E\|\mathbf{R}(\mathbf{X})'\mathbf{U}'\|_F^2 &= \sum_{k=1}^K\sum_{i=1}^N E(\sum_{t=1}^T R_{tk}u_{it})^2 \\
&= \sum_{k=1}^K\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T ER_{tk}R_{sk}Eu_{it}u_{is} \\
&\leq TK\max_{k\leq K} ER_{tk}^2\sum_{i=1}^N\max_{s\leq T}\sum_{t=1}^T|Eu_{it}u_{is}| = O(NTJ^{-\kappa}),
\end{aligned}$$

where $\max_{k\leq K} ER_{tk}^2 = O(J^{-\kappa})$. Therefore, $\|\mathbf{B}\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\|_F^2 = O_P(NT)$.

Note that $\mathbf{B}\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda} = \mathbf{G}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda} - \mathbf{R}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda}$, and $\mathbf{X}_t$ and $\mathbf{u}_i$ are independent. Here,

$$\begin{aligned}
&E\|\mathbf{G}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda}\|_F^2 = \sum_{k=1}^K\sum_{l=1}^K E(\sum_{t=1}^T\sum_{i=1}^N g_l(\mathbf{X}_t)u_{it}\lambda_{ik})^2 = \sum_{k=1}^K\sum_{l=1}^K \mathrm{var}(\sum_{t=1}^T\sum_{i=1}^N g_l(\mathbf{X}_t)u_{it}\lambda_{ik}) \\
&= \sum_{k=1}^K\sum_{l=1}^K\sum_{t=1}^T \mathrm{var}(\sum_{i=1}^N g_l(\mathbf{X}_t)u_{it}\lambda_{ik}) + \sum_{k=1}^K\sum_{l=1}^K\sum_{t\neq s;t,s\leq T}\mathrm{cov}(\sum_{i=1}^N g_l(\mathbf{X}_t)u_{it}\lambda_{ik}, \sum_{i=1}^N g_l(\mathbf{X}_s)u_{is}\lambda_{ik}) \\
&\equiv D_1 + D_2.
\end{aligned}$$

Here, note that $\mathrm{var}(g_l(\mathbf{X}_t)u_{it}\lambda_{ik})$ and $|Eg_l(\mathbf{X}_t)^2\lambda_{ik}\lambda_{jk}|$ are bounded uniformly in $k \leq K$,

$l \leq K$, $i \leq N$, and $j \leq N$. Then, by Assumption 3.4,

$$\begin{aligned} D_1 &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathrm{var}(g_l(\mathbf{X}_t)u_{it}\lambda_{ik}) + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{t=1}^{T}\sum_{i\neq j;i,j\leq N}\mathrm{cov}(g_l(\mathbf{X}_t)u_{it}\lambda_{ik}, g_l(\mathbf{X}_t)u_{jt}\lambda_{jk}) \\ &= O(NT) + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{t=1}^{T}\sum_{i\neq j;i,j\leq N} E(g_l(\mathbf{X}_t)^2\lambda_{ik}\lambda_{jk})Eu_{it}u_{jt}) \\ &\leq O(NT) + NTK^2 \max_{k,l\leq K,i,j\leq N}|g_l(\mathbf{X}_t)^2\lambda_{ik}\lambda_{jk}| \max_{i\leq N}\max_{t\leq T}\sum_{j=1}^{N}|Eu_{it}u_{jt}| = O(NT). \\ D_2 &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{t\neq s;t,s\leq T}\sum_{i=1}^{N}\sum_{j=1}^{N}E(g_l(\mathbf{X}_t)\lambda_{ik}g_l(\mathbf{X}_s)\lambda_{jk})Eu_{it}u_{js} \\ &\leq k^2 \max_{k\leq K;t,s\leq T;i,j\leq N}|E(g_l(\mathbf{X}_t)\lambda_{ik}g_l(\mathbf{X}_s)\lambda_{jk})| \sum_{t,s\leq T}\sum_{i,j\leq N}|Eu_{it}u_{js}| = O(NT). \end{aligned}$$

Hence we have $\|\mathbf{G}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda}\|_F^2 = O_P(NT)$. $\|\mathbf{R}(\mathbf{X})'\mathbf{U}'\mathbf{\Lambda}\|_F^2$ can be bounded in the same way. Therefore, $\|\mathbf{\Lambda}'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT)$. In addition, we need to define $\mathbf{\Lambda}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{\Lambda}$ and $\mathbf{U}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{U}$ and prove $\mathbf{\Lambda}^*$ and $\mathbf{U}^*$ possess the same properties as $\mathbf{\Lambda}$ and $\mathbf{U}$. Then we have $\|\mathbf{\Lambda}'\mathbf{\Sigma}_u^{-1}\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT)$.

(ii) Because X_{tl} and γ_{tk} are independent and $E\gamma_{tk} = 0$ by Assumption 3.6,

$$\begin{aligned} E\|\mathbf{\Gamma}'\mathbf{\Phi}\|_F^2 &= \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{d}E(\sum_{t=1}^{T}\phi_j(X_{tl})\gamma_{tk})^2 \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{d}\mathrm{var}(\sum_{t=1}^{T}\phi_j(X_{tl})\gamma_{tk}) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{d}\sum_{t=1}^{T}\mathrm{var}(\phi_j(X_{tl})\gamma_{tk}) + \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{d}\sum_{t\neq s,t,s\leq T}\mathrm{cov}(\phi_j(X_{tl})\gamma_{tk}, \phi_j(X_{sl})\gamma_{sk}) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{d}\sum_{t=1}^{T}E\phi_j(X_{tl})^2E\gamma_{tk}^2 + \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{l=1}^{d}\sum_{t\neq s,t,s\leq T}E\phi_j(X_{tl})\phi_j(X_{sl})E\gamma_{tk}\gamma_{sk} \\ &\leq JdT \max_{j\leq J,t\leq T,l\leq d}E\phi_j(X_{tl})^2\sum_{k=1}^{K}\max_{k\leq K}\frac{1}{T}\sum_{t=1}^{T}Var(\gamma_{tk}) \\ &\quad + JdT \max_{j\leq J,l\leq d,t,s\leq T}E\phi_j(X_{tl})^2\sum_{k=1}^{K}\max_{k\leq K,s\leq T}\sum_{t=1}^{T}|E\gamma_{tk}\gamma_{sk}| \\ &= O(JT\nu_T). \end{aligned}$$

Also, $\mathbf{B\Phi}(\mathbf{X})'\mathbf{\Gamma} = \mathbf{G}(\mathbf{X})'\mathbf{\Gamma} - \mathbf{R}(\mathbf{X})'\mathbf{\Gamma}$. Because

$$E\|\mathbf{G}(\mathbf{X})'\mathbf{\Gamma}\|_F^2 = \sum_{k=1}^{K}\sum_{l=1}^{K} E(\sum_{t=1}^{T} g_l(\mathbf{X}_t)\gamma_{tk})^2 = O(T\nu_T),$$

and $E\|\mathbf{R}(\mathbf{X})'\mathbf{\Gamma}\|_F^2 = O(J^{-\kappa}T\nu_T)$. Hence $\|\mathbf{B\Phi}(\mathbf{X})'\mathbf{\Gamma}\|_F^2 = O_P(T\nu_T)$. □

Lemma B.3. $\|\mathbf{K}\|_2 = O_P(1)$, $\|\mathbf{K}^{-1}\|_2 = O_P(1)$, *and* $\|\mathbf{H}\|_2 = O_P(1)$.

Proof. For the general factor model, under Assumption 3.2, Lemma A.1 showed that $\mathbf{K}$, $\mathbf{K}^{-1}$ and $\mathbf{M}$ all have bounded spectral norms. Now we consider the semiparametric factor model with $\mathbf{F} = \mathbf{F}(\mathbf{X}) + \mathbf{\Gamma}$, and it implies $\|\mathbf{K}^{-1}\|_2 = O_P(1)$ as well under Assumption 3.2. Note that $\|\mathbf{Y}\|_F^2 = O_P(NT) = \|\widetilde{\mathbf{Y}}\|_F^2$, and $\|\mathbf{P}\|_2 = O_P(1)$. Then it follows immediately that $\|bK\|_2 = O_P(1)$. Now we consider to prove $\|\mathbf{K}^{-1}\|_2 = O_P(1)$. Note that the eigenvalues of $\mathbf{K}$ are the same as those of

$$\mathbf{W} = \frac{1}{NT}(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}\mathbf{\Phi}(\mathbf{X})'\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}\mathbf{\Phi}(\mathbf{X})(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}.$$

Then, using $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{\Lambda}}\mathbf{F}' + \widetilde{\mathbf{U}}$, and $\frac{1}{N}\widetilde{\mathbf{\Lambda}}'\widetilde{\mathbf{\Lambda}} = \mathbf{I}_K$, we have $\mathbf{W} = \sum_{i=1}^4 \mathbf{W}_i$, where

$$\begin{aligned}
\mathbf{W}_1 &= \frac{1}{T}(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}\mathbf{\Phi}(\mathbf{X})'\mathbf{F}\mathbf{F}'\mathbf{\Phi}(\mathbf{X})(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2},\\
\mathbf{W}_2 &= \frac{1}{T}(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}\mathbf{\Phi}(\mathbf{X})'\frac{\mathbf{F}\widetilde{\mathbf{\Lambda}}'\widetilde{\mathbf{U}}}{N}\mathbf{\Phi}(\mathbf{X})(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2},\\
\mathbf{W}_3 &= \mathbf{W}_2',\\
\mathbf{W}_4 &= \frac{1}{T}(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}\mathbf{\Phi}(\mathbf{X})'\frac{\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}}{N}\mathbf{\Phi}(\mathbf{X})(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1/2}.
\end{aligned}$$

By Lemma A.3, we have $\|\mathbf{W}_2\|_2 = O_P(\sqrt{J/NT} + \sqrt{J\log N}/T)$, and $\|\mathbf{W}_4\|_2 = O_P(J/T)$. Therefore, for $k = 1, ..., K$, $|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| \leq \|\mathbf{W} - \mathbf{W}_1\|_2 = o_P(1)$. Hence, it suffices to prove that the first K eigenvalues of $\mathbf{W}_1$ are bounded away from zero, which are also the first K eigenvalues of $\frac{1}{T}\mathbf{F}'\mathbf{P}\mathbf{F} = \frac{1}{T}(\mathbf{G}(\mathbf{X})+\mathbf{\Gamma})'\mathbf{P}(\mathbf{G}(\mathbf{X})+\mathbf{\Gamma})$. In the semiparametric factor model. this can be formally justified and is proved in Lemma B.7. Therefore, $\|\mathbf{K}^{-1}\|_2 = O_P(1)$.

Note that

$$\mathbf{H} = \frac{1}{NT}(\mathbf{Q}\mathbf{Q}'\widetilde{\mathbf{\Lambda}}' + \mathbf{Q}\widetilde{\mathbf{U}}')\widehat{\widetilde{\mathbf{\Lambda}}}\mathbf{K}^{-1},$$

where $\mathbf{Q} = \mathbf{B\Phi}(\mathbf{X})' + \mathbf{\Gamma}'\mathbf{P} + \mathbf{R}(\mathbf{X})'\mathbf{P}$. It can be rewritten as

$$\mathbf{H} = (\sum_{i=1}^{12}\mathbf{H}_i)\mathbf{K}^{-1},$$

where

$$\mathbf{H}_1 = \frac{1}{NT}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_2 = \frac{1}{NT}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})'\mathbf{R}(\mathbf{X})\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_3 = \frac{1}{NT}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},$$
$$\mathbf{H}_4 = \frac{1}{NT}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})'\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_5 = \frac{1}{NT}\mathbf{R}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_6 = \frac{1}{NT}\mathbf{R}(\mathbf{X})'\mathbf{P}\mathbf{R}(\mathbf{X})\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},$$
$$\mathbf{H}_7 = \frac{1}{NT}\mathbf{R}(\mathbf{X})'\mathbf{P}\boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_8 = \frac{1}{NT}\mathbf{R}(\mathbf{X})'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_9 = \frac{1}{NT}\boldsymbol{\Gamma}'\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}$$
$$\mathbf{H}_{10} = \frac{1}{NT}\boldsymbol{\Gamma}'\mathbf{P}\mathbf{R}(\mathbf{X})\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_{11} = \frac{1}{NT}\boldsymbol{\Gamma}'\mathbf{P}\boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}},\ \ \mathbf{H}_{12} = \frac{1}{NT}\boldsymbol{\Gamma}'\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}.$$

Note that, by $\|\mathbf{G}(\mathbf{X})\|_2 = O_P(\sqrt{T})$, $\|\mathbf{H}_1\|_F = \frac{1}{NT}\|(\mathbf{G}(\mathbf{X}) - \mathbf{R}(\mathbf{X}))'(\mathbf{G}(\mathbf{X}) - \mathbf{R}(\mathbf{X}))\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F = O_P(1)$, which is the dominating term. The result then follows from $\|\mathbf{K}^{-1}\|_2 = O_P(1)$ and Lemma B.6. □

Lemma B.4. *Under the assumptions of Theorem 3.2,*
(i) $\frac{1}{N}\|\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F = O_P(\frac{m_N\omega_{N,T}^{1-q}}{\sqrt{T}})$.
(ii) $\frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F = O_P(\frac{m_N\omega_{N,T}^{1-q}}{\sqrt{T}})$.

Proof. By (B.1), $\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{H}) = \sum_{i=1}^4 \widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_i\mathbf{K}^{-1}$. We evaluate each term in the sum and those can be bounded more tightly. Specifically, by Lemma B.2 and A.3,

$$\begin{aligned}
\frac{1}{N^2}\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_1\|_F^2 &= \frac{1}{N^4T^2}\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{U}}\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 \\
&\le O_P(\frac{1}{N^2T^2})(\|\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 + \|\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2) \\
&= O_P(m_N^2\omega_{N,T}^{2-2q}/T), \\
\frac{1}{N^2}\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_2\|_F^2 &= \frac{1}{N^4T^2}\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{U}}\mathbf{P}\mathbf{R}(\mathbf{X})\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 \\
&\le O_P(\frac{1}{N^2T^2})\|\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2^2\|\boldsymbol{\Phi}(\mathbf{X})\|_2^2\|\mathbf{R}(\mathbf{X})\|_F^2 \\
&\le O_P(\frac{J^{-\kappa}}{N^2T^2})(\|\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 + \|\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2) \\
&= O_P(\frac{1}{TJ^{\kappa-1}}m_N^2\omega_{N,T}^{2-2q}),
\end{aligned}$$

where we used $\|\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(N^2Tm_N^2\omega_{N,T}^{2-2q})$, and $\|\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 = O_P(JN^2Tm_N^2\omega_{N,T}^{2-2q})$. Similarly,

$$\begin{aligned}
\frac{1}{N^2}\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_3\|_F^2 = \frac{1}{N^4T^2}&\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{U}}\mathbf{P}\boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2 \\
&\le O_P(\frac{1}{N^2T^2})\|\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2^2\|\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Gamma}\|_F^2 \\
&\le O_P(\frac{J\nu_T}{N^2T^3})(\|\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 + \|\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2)
\end{aligned}$$

$$
\begin{aligned}
&= O_P(\frac{J^2\nu_T}{T^2} m_N^2 \omega_{N,T}^{2-2q}),\\
\frac{1}{N^2}\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_4\|_F^2 &= \frac{1}{N^4T^2}\|\widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}\|_F^2\\
&\leq O_P(\frac{1}{N^4T^4})\|\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2\|\boldsymbol{\Phi}(\mathbf{X})'\mathbf{U}'\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F^2 + O_P(\frac{1}{N^4T^4})\|\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^4\\
&= O_P(\frac{J^2}{T^2} m_N^4 \omega_{N,T}^{4-4q}),
\end{aligned}
$$

where we used $\|\boldsymbol{\Phi}(\mathbf{X})'\mathbf{U}'\widehat{\boldsymbol{\Sigma}}_u^{-\frac{1}{2}}(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F^2 = O_P(JN^2)$. Therefore, combining all these terms, we obtain

$$
\frac{1}{N^2}\|\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F^2 = O_P(1)\sum_{i=1}^{4}\frac{1}{N^2}\|\widetilde{\boldsymbol{\Lambda}}'\mathbf{A}_i\|_F^2 = O_P(\frac{m_N^2\omega_{N,T}^{2-2q}}{T}).
$$

(ii) The result follows from the following inequality:

$$
\frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F \leq \frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{H}\|_F^2 + \frac{1}{N}\|\mathbf{H}'\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}}-\widetilde{\boldsymbol{\Lambda}}\mathbf{H})\|_F.
$$

□

Lemma B.5. *$\|\mathbf{H}'\mathbf{H}-\mathbf{I}_K\|_F = O_P(\frac{m_N\omega_{N,T}^{1-q}}{\sqrt{T}})$. Therefore, $\|\mathbf{H}^{-1}\|_2 = O_P(1)$.*

Proof. Note that $N^{-1}\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}} = N^{-1}\widehat{\widetilde{\boldsymbol{\Lambda}}}'\widehat{\widetilde{\boldsymbol{\Lambda}}} = \mathbf{I}_K$, by the identification condition. Then

$$
\mathbf{H}'\mathbf{H} = \frac{1}{N}(\widetilde{\boldsymbol{\Lambda}}\mathbf{H})'\widetilde{\boldsymbol{\Lambda}}\mathbf{H} = \frac{1}{N}(\widetilde{\boldsymbol{\Lambda}}\mathbf{H}-\widehat{\widetilde{\boldsymbol{\Lambda}}})'\widetilde{\boldsymbol{\Lambda}}\mathbf{H} + \frac{1}{N}\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widetilde{\boldsymbol{\Lambda}}\mathbf{H}-\widehat{\widetilde{\boldsymbol{\Lambda}}}) + \mathbf{I}_K.
$$

This implies the following convergence rate,

$$
\|\mathbf{H}'\mathbf{H}-\mathbf{I}_K\|_F \leq \frac{1}{N}\|(\widetilde{\boldsymbol{\Lambda}}\mathbf{H}-\widehat{\widetilde{\boldsymbol{\Lambda}}})'\widetilde{\boldsymbol{\Lambda}}\|_F\|\mathbf{H}\|_2 + \frac{1}{N}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}'(\widetilde{\boldsymbol{\Lambda}}\mathbf{H}-\widehat{\widetilde{\boldsymbol{\Lambda}}})\|_F = O_P(\frac{m_N\omega_{N,T}^{1-q}}{\sqrt{T}}).
$$

In addition, it implies $\lambda_{\min}(\mathbf{H}'\mathbf{H}) \geq 1 - o_P(1)$. Therefore,

$$
\|\mathbf{H}^{-1}\|_2^2 = \lambda_{\max}(\mathbf{H}^{-1}(\mathbf{H}^{-1})') = \lambda_{\max}((\mathbf{H}'\mathbf{H})^{-1}) = \lambda_{\min}(\mathbf{H}'\mathbf{H}) = O_P(1).
$$

□

Lemma B.6. *(i) $\|\mathbf{H}_2\|_F = O_P(\sqrt{J^{-\kappa}}) = \|\mathbf{H}_5\|_F$,*
(ii) $\|\mathbf{H}_3\|_F = O_P(\sqrt{\nu_T/T}) = \|\mathbf{H}_9\|_F$, $\|\mathbf{H}_4\|_F = O_P(1/\sqrt{T})$, $\|\mathbf{H}_{12}\|_F = O_P(J\sqrt{\nu_T}/T)$,
(iii) $\|\mathbf{H}_6\|_F = O_P(J^{-\kappa})$, $\|\mathbf{H}_8\|_F = O_P(1/\sqrt{TJ^{\kappa-1}})$,
(iv) $\|\mathbf{H}_7\|_F = O_P(\sqrt{\nu_T/(TJ^{\kappa-1})} = \|\mathbf{H}_{10}\|_F$, $\|\mathbf{H}_{11}\|_F = O_P(J\nu_T/T)$.

Proof. (i) Note that $\|\widetilde{\mathbf{\Lambda}}\|_F = O_P(\sqrt{N}) = \|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F$. Also $\|\mathbf{B\Phi}(\mathbf{X})'\|_2 \leq \|\mathbf{G}(\mathbf{X})\|_2 + \|\mathbf{R}(\mathbf{X})\|_2 = O_P(\sqrt{T})$, and $\|\mathbf{R}(\mathbf{X})\|_F^2 = O_P(TJ^{-\kappa})$. Then

$$\|\mathbf{H}_2\|_F \leq \frac{1}{NT}\|\mathbf{B\Phi}(\mathbf{X})'\|_2\|\mathbf{R}(\mathbf{X})\|_F\|\widetilde{\mathbf{\Lambda}}\|_F\|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F = O_P(\sqrt{J^{-\kappa}}).$$

Similarlly, $\|\mathbf{H}_5\|_F = O_P(\sqrt{J^{-\kappa}})$.
(ii) The results follows from Lemma B.2. $\|\mathbf{H}_3\|_F = O_P(\frac{1}{T}\|\mathbf{B\Phi}(\mathbf{X})'\mathbf{\Gamma}\|_F) = O_P(\sqrt{\nu_T/T})$. Similarlly, $\|\mathbf{H}_9\|_F$ attains the same rate. $\|\mathbf{H}_{12}\|_F = O_P(\frac{1}{T^2\sqrt{N}}\|\mathbf{\Gamma}'\mathbf{\Phi}(\mathbf{X})\|_F\|\mathbf{\Phi}(\mathbf{X})'\widetilde{\mathbf{U}}\|_F) = O_P(J\sqrt{\nu_T}/T)$. In addition, $\|\mathbf{H}_4\|_F = O_P(\frac{1}{T\sqrt{N}}\|\mathbf{B\Phi}(\mathbf{X})'\widetilde{\mathbf{U}}\|_F) = O_P(1/\sqrt{T})$.
(iii) Note that $\|\mathbf{P}\|_2 = O_P(1)$. Hence $\|\mathbf{H}_6\|_F = O_P(J^{-\kappa})$. In addition,

$$\|\mathbf{H}_8\|_F \leq \frac{1}{NT}\|\mathbf{R}(\mathbf{X})\|_F\|\mathbf{P}\widetilde{\mathbf{U}}'\|_F\|\widehat{\widetilde{\mathbf{\Lambda}}}\|_F = O_P(1/\sqrt{TJ^{\kappa-1}}).$$

(iv) $\|\mathbf{H}_7\|_F = O_P(\frac{1}{T}\|\mathbf{R}(\mathbf{X})\|_F\|\mathbf{P\Gamma}\|_F) = O_P(\sqrt{\nu_T/(TJ^{\kappa-1}})$. The convergence rate for $\mathbf{H}_{10}$ can be bounded in the same way. Finally,

$$\|\mathbf{H}_{11}\|_F \leq O_P(\frac{1}{T}\|\mathbf{\Gamma}'\mathbf{\Phi}(\mathbf{X})\|_F^2\|(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1}\|_2) = O_P(J\nu_T/T).$$

□

Lemma B.7. *Consider the semiparametric factor model. Under Assumption 3.2, there are constants c_1, $c_2 > 0$ such that with probability approaching one,*

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

Proof. Note that $\mathbf{G}(\mathbf{X}) = \mathbf{\Phi}(\mathbf{X})\mathbf{B}' + \mathbf{R}(\mathbf{X})$, and

$$\frac{1}{T}\mathbf{F}'\mathbf{PF} - \frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X}) = \frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{P\Gamma} + \frac{1}{T}\mathbf{\Gamma}'\mathbf{PG}(\mathbf{X}) + \frac{1}{T}\mathbf{\Gamma}'\mathbf{P\Gamma} - \frac{1}{T}\mathbf{G}(\mathbf{X})'(\mathbf{I}-\mathbf{P})\mathbf{G}(\mathbf{X}).$$

We show that alll the terms of right hand sides are negligible. By Lemma B.2,

$$\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{P\Gamma} = \frac{1}{T}\mathbf{B\Phi}(\mathbf{X})'\mathbf{\Gamma} + \frac{1}{T}\mathbf{R}(\mathbf{X})'\mathbf{P\Gamma} = O_P(\sqrt{\nu_T/T}).$$

Similarly, $\frac{1}{T}\mathbf{\Gamma}'\mathbf{PG}(\mathbf{X})$ attains the same rate. $\frac{1}{T}\mathbf{\Gamma}'\mathbf{P\Gamma} = \frac{1}{T}\|\mathbf{P\Gamma}\|_F^2 = O_P(J\nu_T/T)$. Finally, $\frac{1}{T}\mathbf{G}(\mathbf{X})'(\mathbf{I}-\mathbf{P})\mathbf{G}(\mathbf{X}) = \frac{1}{T}\mathbf{R}(\mathbf{X})'(\mathbf{I}-\mathbf{P})\mathbf{R}(\mathbf{X}) = O_P(J^{-\kappa})$. Define the event $A = \{\|\frac{1}{T}\mathbf{F}'\mathbf{PF} - \frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})\|_F < c_1/2\}$. Then it implies that A occurs with probability approaching one,

$$P(\lambda_{\min}(\frac{1}{T}\mathbf{F}'\mathbf{PF}) > c_1/3) \geq P(\lambda_{\min}(\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) - \|\frac{1}{T}\mathbf{F}'\mathbf{PF} - \frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})\|_F > c_1/3)$$
$$\geq P(\lambda_{\min}(\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) > \frac{5c_1}{6}, A) \geq 1 - P(A^c) - P(\lambda_{\min}(\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) \leq \frac{5c_1}{6}) = 1 - o(1).$$

In addition,

$$\begin{aligned}
&P(\lambda_{\max}(\frac{1}{T}\mathbf{F}'\mathbf{P}\mathbf{F}) < 2c_2) \geq P(\lambda_{\min}(\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) + \|\frac{1}{T}\mathbf{F}'\mathbf{P}\mathbf{F} - \frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})\|_F > 2c_2) \\
&\geq P(\lambda_{\max}(\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) < \frac{3c_2}{2}, A) \geq 1 - P(A^c) - P(\lambda_{\max}(\frac{1}{T}\mathbf{G}(\mathbf{X})'\mathbf{G}(\mathbf{X})) \leq \frac{3c_1}{2}) = 1 - o(1).
\end{aligned}$$

□

B.2 Convergence of factors

Define $\widehat{\mathbf{B}} = \frac{1}{N}\widehat{\widetilde{\boldsymbol{\Lambda}}}'\widetilde{\mathbf{Y}}\boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}$. Then

$$\widehat{\mathbf{G}}(\mathbf{X}) = \frac{1}{N}\mathbf{P}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\boldsymbol{\Lambda}}} = \boldsymbol{\Phi}(\mathbf{X})\widehat{\mathbf{B}}'.$$

Substituting $\widetilde{\mathbf{Y}} = \widetilde{\boldsymbol{\Lambda}}\mathbf{B}\boldsymbol{\Phi}(\mathbf{X})' + \widetilde{\boldsymbol{\Lambda}}\mathbf{R}(\mathbf{X})' + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}' + \widetilde{\mathbf{U}}$, and using $\widetilde{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{\Lambda}}/N = \mathbf{I}_K$,

$$\widehat{\mathbf{B}}' - \mathbf{B}'\mathbf{H} = \sum_{i=1}^{5}\mathbf{C}_i,$$

where

$$\begin{aligned}
&\mathbf{C}_1 = \frac{1}{N}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\mathbf{R}(\mathbf{X})\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}, \quad \mathbf{C}_2 = \frac{1}{N}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\widetilde{\mathbf{U}}'\widetilde{\boldsymbol{\Lambda}}'\mathbf{H}, \\
&\mathbf{C}_3 = \frac{1}{N}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\widetilde{\mathbf{U}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}'\mathbf{H}), \quad \mathbf{C}_4 = \frac{1}{N}\mathbf{B}'\widetilde{\boldsymbol{\Lambda}}'(\widehat{\widetilde{\boldsymbol{\Lambda}}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{H}) \\
&\mathbf{C}_5 = \frac{1}{N}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}'\widehat{\widetilde{\boldsymbol{\Lambda}}}.
\end{aligned}$$

Lemma B.8. *(i)* $\|\mathbf{C}_1\|_F^2 = O_P(\frac{1}{J^\kappa})$, $\quad \|\mathbf{C}_2\|_F^2 = O_P(\frac{J}{T}m_N^2\omega_{N,T}^{2-2q})$, $\|\mathbf{C}_3\|_F^2 = O_P(\frac{J}{T^2})$, $\quad \|\mathbf{C}_4\|_F^2 = O_P(\frac{J}{T}m_N^2\omega_{N,T}^{2-2q})$, $\quad \|\mathbf{C}_5\|_F^2 = O_P(\frac{J\nu_T}{T})$.
(ii) $\|\widehat{\mathbf{B}}' - \mathbf{B}'\mathbf{H}\|_F^2 = O_P(\frac{1}{J^\kappa} + \frac{J\nu_T}{T} + \frac{J}{T}m_N^2\omega_{N,T}^{2-2q})$.

Proof. (i) By Lemmas A.3, B.1, B.2 and B.4,

$$\begin{aligned}
\|\mathbf{C}_1\|_F^2 &\leq O_P(\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2^2\|\boldsymbol{\Phi}(\mathbf{X})\|_2^2\|\mathbf{R}(\mathbf{X})\|_F^2) = O_P(J^{-\kappa}), \\
\|\mathbf{C}_2\|_F^2 &\leq O_P(\frac{1}{N^2})\|(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\|_2^2\|\boldsymbol{\Phi}(\mathbf{X})'\mathbf{U}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Lambda}\|_F^2 \\
&\leq O_P(\frac{1}{N^2T^2})(\|\boldsymbol{\Lambda}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2 + \|\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_u^{-1}\mathbf{U}\boldsymbol{\Phi}(\mathbf{X})\|_F^2) \\
&= O_P(\frac{J}{T}m_N^2\omega_{N,T}^{2-2q}),
\end{aligned}$$

$$
\begin{aligned}
&\|\mathbf{C}_3\|_F^2 \leq O_P(\frac{1}{N^2T^2})\|\mathbf{\Phi}(\mathbf{X})'\mathbf{U}'\|_F^2\|\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}\|_F^2 = O_P(J/T^2),\\
&\|\mathbf{C}_4\|_F^2 \leq \|\mathbf{B}\|_F^2\|\frac{1}{N}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\frac{J}{T}m_N^2\omega_{N,T}^{2-2q}),\\
&\|\mathbf{C}_5\|_F^2 \leq O_P(\|(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))^{-1}\|_2^2\|\mathbf{\Phi}(\mathbf{X})'\mathbf{\Gamma}\|_F^2) = O_P(J\nu_T/T).
\end{aligned}
$$

(ii) By the results in (i), we have

$$
\|\widehat{\mathbf{B}}' - \mathbf{B}'\mathbf{H}\|_F^2 \leq O(1)\sum_{i=1}^{5}\|\mathbf{C}_i\|_F^2 = O_P\left(\frac{1}{J^\kappa} + \frac{J\nu_T}{T} + \frac{J}{T}m_N^2\omega_{N,T}^{2-2q}\right).
$$

□

Because $\mathbf{G}(\mathbf{X})\mathbf{H} = \mathbf{\Phi}(\mathbf{X})\mathbf{B}'\mathbf{H} + \mathbf{R}(\mathbf{X})\mathbf{H}$, by Lemma B.8

$$
\begin{aligned}
\frac{1}{T}\|\widehat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X})\mathbf{H}\|_F^2 &\leq \frac{2}{T}\|\mathbf{\Phi}(\mathbf{X})(\widehat{\mathbf{B}}' - \mathbf{B}'\mathbf{H})\|_F^2 + \frac{2}{T}\|\mathbf{R}(\mathbf{X})\mathbf{H}\|_F^2\\
&= O_P(\|\widehat{\mathbf{B}}' - \mathbf{B}'\mathbf{H}\|_F^2 + J^{-\kappa}) = O_P\left(\frac{1}{J^\kappa} + \frac{J\nu_T}{T} + \frac{J}{T}m_N^2\omega_{N,T}^{2-2q}\right).
\end{aligned}
$$

Substituting $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{\Lambda}}\mathbf{B}\mathbf{\Phi}(\mathbf{X})' + \widetilde{\mathbf{\Lambda}}\mathbf{R}(\mathbf{X})' + \widetilde{\mathbf{\Lambda}}\mathbf{\Gamma}' + \widetilde{\mathbf{U}}$ into $\widehat{\mathbf{\Gamma}} = \frac{1}{N}(\mathbf{I} - \mathbf{P})\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}}$,

$$
\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\mathbf{H} = \sum_{i=1}^{6}\mathbf{D}_i,
$$

where

$$
\begin{aligned}
&\mathbf{D}_1 = \frac{1}{N}(\mathbf{I} - \mathbf{P})\mathbf{\Gamma}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}), \quad \mathbf{D}_2 = \frac{1}{N}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}),\\
&\mathbf{D}_3 = -\mathbf{P}\mathbf{\Gamma}\mathbf{H}, \quad \mathbf{D}_4 = (\mathbf{I} - \mathbf{P})\mathbf{R}(\mathbf{X})(\mathbf{H} + \frac{1}{N}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})),\\
&\mathbf{D}_5 = -\frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}), \quad \mathbf{D}_6 = \frac{1}{N}(\mathbf{I} - \mathbf{P})\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{H}.
\end{aligned}
$$

Then, by Lemma B.11,

$$
\frac{1}{T}\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\mathbf{H}\|_F^2 \leq O(\frac{1}{T})\sum_{i=1}^{6}\|\mathbf{D}_i\|_F^2 = O_P(\frac{1}{N} + \frac{J}{T^2} + \frac{1}{J^\kappa} + \frac{J\nu_T}{T} + \frac{1}{T}m_N^2\omega_{N,T}^{2-2q}).
$$

Lemma B.9. *(i)* $\|\mathbf{U}'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F^2 = O_P(JNT^2 + N^2T)$, *and* $\|\mathbf{U}'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT^2 + N^2T)$.
(ii) $\|\mathbf{U}'\mathbf{\Sigma}_u^{-1}\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F^2 = O_P(JNT^2 + N^2T)$, *and* $\|\mathbf{U}'\mathbf{\Sigma}_u^{-1}\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT^2 + N^2T)$.

Proof. (i) Note that $\|E\mathbf{U}'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\|_F^2 = O_P(N^2T)$. Let $s_{ts} = \sum_{i=1}^N (u_{it}u_{is} - Eu_{it}u_{is})$. Then

$$
\begin{aligned}
E\|(\mathbf{U}'\mathbf{U} - E\mathbf{U}'\mathbf{U})\mathbf{\Phi}(\mathbf{X})\|_F^2 &= \sum_{t=1}^T\sum_{l=1}^d\sum_{j=1}^J \mathrm{var}(\sum_{s=1}^T \phi_j(X_{sl})s_{ts}) \\
&= \sum_{t=1}^T\sum_{l=1}^d\sum_{j=1}^J\sum_{s=1}^T E\phi_j(X_{sl})^2Es_{ts}^2 + \sum_{t=1}^T\sum_{l=1}^d\sum_{j=1}^J\sum_{s\neq q;s,q\leq T} E\phi_j(X_{sl})\phi_j(X_{ql})\mathrm{cov}(s_{ts},s_{tq}) \\
&= \sum_{t=1}^T\sum_{l=1}^d\sum_{j=1}^J\sum_{s=1}^T E\phi_j(X_{sl})^2\mathrm{var}(\sum_{i=1}^N u_{it}u_{is}) \\
&\quad + \sum_{t=1}^T\sum_{l=1}^d\sum_{j=1}^J\sum_{s\neq q;s,q\leq T} E\phi_j(X_{sl})\phi_j(X_{ql})\sum_{i=1}^N\sum_{m=1}^N \mathrm{cov}(u_{it}u_{is},u_{mt}u_{mq}) \\
&= O(JNT^2),
\end{aligned}
$$

by Assumption 3.4.

Note that $\|\mathbf{U}'\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 \leq \|\mathbf{U}\|^2\|\mathbf{U}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_F^2 = O_P(NT^2 + N^2T)$ by Lemma B.2.

(ii) We only need to define $\mathbf{U}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{U}$ and prove $\mathbf{U}^*$ possess the same properties as $\mathbf{U}$. Then the results follow using the same argument as in (i). □

Lemma B.10. *(i)* $\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_1\|_F^2 = O_P(\frac{1}{N} + \frac{1}{T})$, $\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_2\|_F^2 = O_P(\frac{1}{NJ^{\kappa-1}} + \frac{1}{TJ^\kappa})$,
(ii) $\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_3\|_F^2 = O_P(\frac{J^2\nu_T}{NT} + \frac{J\nu_T}{T^2})$, $\|\frac{1}{N}\widetilde{\mathbf{U}}'\mathbf{A}_4\|_F^2 = O_P(\frac{J^2}{NT} + \frac{J}{T^2})$.

Proof. Given Lemma B.9, the proof is stratightforward calculations. □

Lemma B.11. *(i)* $\|\frac{1}{N}(\mathbf{I}-\mathbf{P})\mathbf{\Gamma}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\nu_T m_N^2\omega_{N,T}^{2-2q})$,
(ii) $\|\frac{1}{N}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\frac{1}{N} + \frac{1}{T})$,
(iii) $\|\mathbf{P}\mathbf{\Gamma}\mathbf{H}\|_F^2 = O_P(J\nu_T)$,
(iv) $\|(\mathbf{I}-\mathbf{P})\mathbf{R}(\mathbf{X})(\mathbf{H} + \frac{1}{N}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}))\|_F^2 = O_P(\frac{T}{J^\kappa})$,
(v) $\|\frac{1}{N}\mathbf{P}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\frac{J}{T})$,
(vi) $\|\frac{1}{N}(\mathbf{I}-\mathbf{P})\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{H}\|_F^2 = O_P(\frac{T}{N} + m_N^2\omega_{N,T}^{2-2q})$.

Proof. Note that $\|\mathbf{\Gamma}\|_F^2 = O_P(T\nu_T)$ and $\|\mathbf{R}(\mathbf{X})\|_F^2 = O_P(TJ^{-\kappa})$. (i), (iii)-(v) follow from Lemmas A.3, B.2 and B.4. (ii) follows from Lemma B.10. (vi) follows from Cauchy-Schwarz inequality. □

As for the estimated factor matrix $\widehat{\mathbf{F}}$, note that $\widehat{\mathbf{F}} = \frac{1}{N}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}}$. Substituting $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{\Lambda}}\mathbf{B}\mathbf{\Phi}(\mathbf{X})' + \widetilde{\mathbf{\Lambda}}\mathbf{R}(\mathbf{X})' + \widetilde{\mathbf{\Lambda}}\mathbf{\Gamma}' + \widetilde{\mathbf{U}}$, then

$$
\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H} = \sum_{i=1}^5 \mathbf{E}_i,
$$

where

$$\mathbf{E}_1 = \frac{1}{N}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}), \quad \mathbf{E}_2 = \frac{1}{N}\mathbf{R}(\mathbf{X})'\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}),$$
$$\mathbf{E}_3 = \frac{1}{N}\mathbf{\Gamma}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}), \quad \mathbf{E}_4 = \frac{1}{N}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}), \quad \mathbf{E}_5 = \frac{1}{N}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{H}.$$

Then, by Lemma B.12,

$$\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_F^2 \le O(\frac{1}{T})\sum_{i=1}^{5}\|\mathbf{E}_i\|_F^2 = O_P\left(\frac{1}{N} + \frac{1}{T}m_N^2\omega_{N,T}^{2-2q}\right).$$

Lemma B.12. *(i)* $\|\frac{1}{N}\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(m_N^2\omega_{N,T}^{2-2q})$,
(ii) $\|\frac{1}{N}\mathbf{R}(\mathbf{X})'\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\frac{1}{J^\kappa}m_N^2\omega_{N,T}^{2-2q})$,
(iii) $\|\frac{1}{N}\mathbf{\Gamma}\widetilde{\mathbf{\Lambda}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\nu_T m_N^2\omega_{N,T}^{2-2q})$,
(iv) $\|\frac{1}{N}\widetilde{\mathbf{U}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F^2 = O_P(\frac{1}{N} + \frac{1}{T})$,
(v) $\|\frac{1}{N}\widetilde{\mathbf{U}}'\widetilde{\mathbf{\Lambda}}\mathbf{H}\|_F^2 = O_P(\frac{T}{N} + m_N^2\omega_{N,T}^{2-2q})$,

Proof. Note that $\|\mathbf{\Phi}(\mathbf{X})\mathbf{B}'\|_2 \le \|\mathbf{G}(\mathbf{X})\|_2 + \|\mathbf{R}(\mathbf{X})\|_2 = O_P(\sqrt{T})$, and $\|\mathbf{R}(\mathbf{X})\|_F^2 = O_P(TJ^{-\kappa})$. In addition, $\|N^{-1}\mathbf{U}^{*'}\mathbf{\Lambda}^*\mathbf{H}\|_F^2 = O_P(N^{-2}\sum_{t=1}^T E\|\mathbf{\Lambda}'\mathbf{u}_t\|)^2) = O_P(T/N)$. (i)-(iii), (v) follow from Lemmas A.3, B.2 and B.4. (iv) follows from Lemma B.10. □

B.3 Individual factors

Since $\widehat{\mathbf{F}} = \frac{1}{N}\widetilde{\mathbf{Y}}'\widehat{\widetilde{\mathbf{\Lambda}}}$, $\widehat{\mathbf{F}}_t = \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widetilde{\mathbf{\Lambda}}\mathbf{F}_t + \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widetilde{\mathbf{u}}_t$. Using $\widetilde{\mathbf{\Lambda}} = \widetilde{\mathbf{\Lambda}} - \widehat{\widetilde{\mathbf{\Lambda}}}\mathbf{H}^{-1} + \widehat{\widetilde{\mathbf{\Lambda}}}\mathbf{H}^{-1}$ and $\frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'\widehat{\widetilde{\mathbf{\Lambda}}} = \mathbf{I}_K$, we have

$$\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t = \sum_{i=1}^{3}\mathbf{W}_i,$$

where

$$\mathbf{W}_1 = \frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'(\widetilde{\mathbf{\Lambda}}\mathbf{H} - \widehat{\widetilde{\mathbf{\Lambda}}})\mathbf{H}^{-1}\mathbf{F}_t, \quad \mathbf{W}_2 = \frac{1}{N}(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})'\widetilde{\mathbf{u}}_t, \quad \mathbf{W}_3 = \frac{1}{N}\mathbf{H}'\widetilde{\mathbf{\Lambda}}'\widetilde{\mathbf{u}}_t.$$

Then, by Lemmas B.4 and B.5,

$$\|\mathbf{W}_1\| \le \|\frac{1}{N}\widehat{\widetilde{\mathbf{\Lambda}}}'(\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H})\|_F\|\mathbf{H}^{-1}\|_2\|\mathbf{F}_t\| = O_P\left(\frac{1}{\sqrt{T}}m_N\omega_{N,T}^{1-q}\right).$$

Note that $\|N^{-1}\mathbf{\Lambda}'\mathbf{u}_t\| = O_P(1/\sqrt{N})$ by Lemma A.8. Let $\mathbf{\Lambda}^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{\Lambda}$ and $\mathbf{u}_t^* = \mathbf{\Sigma}_u^{-\frac{1}{2}}\mathbf{u}_t$. Then $\|N^{-1}\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\| = O_P(1/\sqrt{N})$. By Lemma B.3,

$$\begin{aligned}
\|\mathbf{W}_3\| &\le \frac{1}{N}\|\mathbf{H}\|_2(\|\mathbf{\Lambda}'(\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\mathbf{u}_t\| + \|\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\|) \\
&\le O_P(\frac{1}{N})(\|\mathbf{\Lambda}\|_F\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_1\|\mathbf{u}_t\| + \|\mathbf{\Lambda}^{*'}\mathbf{u}_t^*\|) \\
&\le O_P(m_N\omega_{N,T}^{1-q} + \frac{1}{\sqrt{N}}) = O_P(m_N\omega_{N,T}^{1-q}).
\end{aligned}$$

For each fixed t, it follows from $\frac{1}{N}\|\widehat{\widetilde{\mathbf{\Lambda}}} - \widetilde{\mathbf{\Lambda}}\mathbf{H}\|^2 = O_P(T^{-1})$ and $\frac{1}{N}\sum_{i=1}^N u_{it}^2 = O_P(1)$ that $\|\mathbf{W}_2\| = O_P(T^{-1/2})$. Therefore, for each $t \le T$,

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

Appendix C Proofs for Section 3.4

C.1 Proof of Theorem 3.3

Lemma C.1. *Let $L_t = (\mathbf{F}_t', W_t')'$, and $\widehat{L}_t = (\widehat{\mathbf{F}}_t', W_t')'$. Under Assumptions 3.1, 3.3-3.8, we have, for $\omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,*
(i) $\frac{1}{T}\sum_{t=1}^T(\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t)L_t' = O_P(m_N^2\omega_{N,T}^{2-2q})$,
(ii) $\frac{1}{T}\sum_{t=1}^T(\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t)\widehat{L}_t' = O_P(m_N^2\omega_{N,T}^{2-2q})$,
(iii) $\frac{1}{T}\sum_{t=1}^T(\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t)\epsilon_{t+h} = O_P(m_N^2\omega_{N,T}^{2-2q})$.

Proof. (i) Note that $\widehat{\widetilde{\mathbf{\Lambda}}} = (\widehat{\widetilde{\boldsymbol{\lambda}}}_1, ..., \widehat{\widetilde{\boldsymbol{\lambda}}}_N)'$ and $\widetilde{\mathbf{u}}_t = (\widetilde{u}_{1t}, ..., \widetilde{u}_{Nt})' = \widehat{\mathbf{\Sigma}}_u^{-\frac{1}{2}}\mathbf{u}_t$. Then, we can write

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T(\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t)L_t' &= \frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N\widehat{\widetilde{\boldsymbol{\lambda}}}_i(\mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i - \widehat{\widetilde{\boldsymbol{\lambda}}}_i)'\mathbf{H}^{-1}\mathbf{F}_tL_t' \\
&\quad + \frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N\widetilde{u}_{it}(\widehat{\widetilde{\boldsymbol{\lambda}}}_i - \mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i)L_t' + \frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N\mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i\widetilde{u}_{it}L_t' \\
&= (I) + (II) + (III).
\end{aligned}$$

For (I), we have

$$I = \frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N(\widehat{\widetilde{\boldsymbol{\lambda}}}_i - \mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i)(\widehat{\widetilde{\boldsymbol{\lambda}}}_i - \mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i)'\mathbf{H}^{-1}\mathbf{F}_tL_t' + \frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N\mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i(\widehat{\widetilde{\boldsymbol{\lambda}}}_i - \mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i)'\mathbf{H}^{-1}\mathbf{F}_tL_t'.$$

Note that $T^{-1}\sum_{t=1}^T(E\|\mathbf{F}_t\|^2)^{1/2}(E\|L_t\|^2)^{1/2} = O(1)$ by Assumption 3.8 and, $N^{-1}\sum_{i=1}\|\widehat{\widetilde{\boldsymbol{\lambda}}}_i -$

$\mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i\|^2 = O_P(T^{-1})$ by Theorem 3.2. Then the first term is bounded by

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

by the result of Lemma B.5. Similarly, the second term is bounded by $O_P(T^{-1/2}m_N\omega_{N,T}^{1-q})$ from the result in Lemma B.4. Thus, $(I) = O_P(T^{-1/2}m_N\omega_{N,T}^{1-q})$.

For (II), note that we have $\|\frac{1}{N}\sum_{i=1}^{N}(\widehat{\widetilde{\boldsymbol{\lambda}}}_i - \mathbf{H}'\widetilde{\boldsymbol{\lambda}}_i)\widetilde{u}_{it}\| = O_P(m_N^2\omega_{N,T}^{2-2q})$ using the same proofs as those of Lemma A.6 of working version of Bai and Liao (2017). Then Cauchy-Schwarz inequality implies that $(II) = O_P(m_N^2\omega_{N,T}^{2-2q})$.

For (III), we have $\mathbf{H}'\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\widetilde{\boldsymbol{\lambda}}_i L_t'\widetilde{u}_{it} = O_P(1)O_P(\frac{1}{\sqrt{NT}})$. Therefore, we have

$$I + II + III = O_P(T^{-1/2}m_N\omega_{N,T}^{1-q}) + O_P(m_N^2\omega_{N,T}^{2-2q}) + O_P(\frac{1}{\sqrt{NT}}) = O_P(m_N^2\omega_{N,T}^{2-2q}).$$

Next, the proof for (ii) can be easily obtained by adapting Lemma A.1 of Bai and Ng (2006). Finally, the proof for (iii) is similar to (i), with ϵ_t instead of L_t. □

The forecasting model (3.8) can be written as

$$\begin{aligned}
z_{t+h} &= \alpha'\mathbf{F}_t + \beta' W_t + \epsilon_{t+h} \\
&= \alpha'\mathbf{H}^{-1'}\widehat{\mathbf{F}}_t + \beta' W_t + \epsilon_{t+h} + \alpha'\mathbf{H}^{-1'}(\mathbf{H}'\mathbf{F}_t - \widehat{\mathbf{F}}_t) \\
&= \widehat{L}_t'\delta + \epsilon_{t+h} + \alpha'\mathbf{H}^{-1'}(\mathbf{H}'\mathbf{F}_t - \widehat{\mathbf{F}}_t).
\end{aligned}$$

In matrix notation, the model can be rewritten as

$$Z = \widehat{L}\delta + \epsilon + (\mathbf{F}\mathbf{H} - \widehat{\mathbf{F}})\mathbf{H}^{-1}\alpha,$$

where $Z = (z_{h+1}, ..., z_T)'$, $\epsilon = (\epsilon_{h+1}, ..., \epsilon_T)'$, and $\widehat{L} = (\widehat{L}_1, ..., \widehat{L}_{T-h})'$. Then, the ordinary least squares estimator of δ is $\widehat{\delta} = (\widehat{L}'\widehat{L})^{-1}\widehat{L}'Z$. Note that by Lemma C.1, $T^{-1/2}\widehat{L}'(\mathbf{F}\mathbf{H} - \widehat{\mathbf{F}}) = O_P(T^{1/2}m_N^2\omega_{N,T}^{2-2q})$ and $T^{-1/2}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})'\epsilon = O_P(T^{1/2}m_N^2\omega_{N,T}^{2-2q})$. Then, if $T^{1/2}m_N^2\omega_{N,T}^{2-2q} \to 0$,

$$\begin{aligned}
\sqrt{T}(\widehat{\delta} - \delta) &= (T^{-1}\widehat{L}'\widehat{L})^{-1}T^{-1/2}\widehat{L}'\epsilon + (T^{-1}\widehat{L}'\widehat{L})^{-1}T^{-1/2}\widehat{L}'(\mathbf{F}\mathbf{H} - \widehat{\mathbf{F}})\mathbf{H}^{-1}\alpha \\
&= (T^{-1}\widehat{L}'\widehat{L})^{-1}T^{-1/2}\widehat{L}'\epsilon + o_P(1),
\end{aligned}$$

and

$$\frac{\widehat{L}'\epsilon}{\sqrt{T}} = \begin{bmatrix} \frac{(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})'\epsilon}{\sqrt{T}} + \frac{\mathbf{H}'\mathbf{F}'\epsilon}{\sqrt{T}} \\ \frac{W'\epsilon}{\sqrt{T}} \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{H}'\mathbf{F}'\epsilon}{\sqrt{T}} \\ \frac{W'\epsilon}{\sqrt{T}} \end{bmatrix} + o_P(1).$$

Then, for a block diagonal matrix $\Pi = \mathrm{diag}(\mathbf{H}', I)$,

$$\sqrt{T}(\widehat{\delta} - \delta) = (T^{-1}\widehat{L}'\widehat{L})^{-1}\Pi(T^{-1/2}L'\epsilon) + o_P(1).$$

Since $T^{-1/2}L'\epsilon \xrightarrow{d} N(0, \Sigma_{L,\epsilon})$ by Assumption 3.8, we obtain the result in Theorem 3.3.

C.2 Proof of Theorem 3.4

From Theorem 3.2, note that $\|\widehat{\mathbf{F}}_t - \mathbf{H}^{-1}\mathbf{F}_t\| = O_P(m_N \omega_{N,T}^{1-q})$ and $\sqrt{T}(\widehat{\delta} - \delta)$ is asymptotically normal. Therefore,

$$\begin{aligned}
\widehat{z}_{T+h|T} - z_{T+h|T} &= \widehat{\alpha}'\widehat{\mathbf{F}}_T + \widehat{\beta}'W_T - \alpha'\mathbf{F}_T - \beta'W_T \\
&= (\widehat{\alpha} - \mathbf{H}'\alpha)'\widehat{\mathbf{F}}_T + \alpha'\mathbf{H}(\widehat{\mathbf{F}}_T - \mathbf{H}^{-1}\mathbf{F}_T) + (\widehat{\beta} - \beta)'W_T \\
&= \frac{1}{\sqrt{T}}\widehat{L}_T'\sqrt{T}(\widehat{\delta} - \delta) + \alpha'\mathbf{H}(\widehat{\mathbf{F}}_T - \mathbf{H}^{-1}\mathbf{F}_T) \\
&= O_P(m_N \omega_{N,T}^{1-q}).
\end{aligned}$$

Appendix D Forecasting methods in Details

This section introduces the collection of all methods including machine learning techniques that I use in this paper. For machine learning models, I split the sample data into validation, training and test set by following common practice. The validation dataset is used to estimate hyperparameters in the models and avoid overfitting problems. For more details on machine learning estimation strategy, see Bianchi et al. (2019) and Gu et al. (2020).

D.1 Univariate autoregression model

I set a univariate AR(p) model as our main benchmark:

$$rx_{t+1}^{(n)} = \alpha + \phi(L) rx_t^{(n)} + \epsilon_{t+1},$$

with the number of lags, p, is selected using the SIC. In the experiments, coefficients are estimated using least squares under both rolling and recursive data window methods. The window sizes are $Q = 240$ months for the rolling method, and $Q = 240 + s$ months for the recursive method.

D.2 Diffusion index models

This paper considers Diffusion Index models of the following form:

$$rx_{t+1}^{(n)} = \alpha + \beta_W' W_t + \beta_F' \mathbf{F}_t + \epsilon_{t+1}, \tag{D.1}$$

where W_t is observable variables, ϵ_t is a disturbance term, and α and β are parameters estimated using least squares. Since $\mathbf{F}_t$ is unobservable variables, I estimate it using principal component methods. In this experiment, I perform PC, PPC, and FPPC as discussed in previous sections. This diffusion index forecasting model is widely studied in the literature (e.g., Bai and Ng, 2002, 2006; Kim and Swanson, 2014; Stock and Watson, 2014, etc). Varying the components of W_t yields different type of models as follows.

D.2.1 Principal components regression

Forecasts from a principal components regression (PCR) are computed as

$$\widehat{rx}_{t+1}^{(n)} = \widehat{\alpha} + \widehat{\beta}_F' \widehat{\mathbf{F}}_t,$$

where $\widehat{\mathbf{F}}_t$ is estimated using principal component methods using $\{Y_t\}_{t=1}^T$, which is 130 macroeconomic variables. Note that PCR is a special case of DI model by replacing W_t with constant term in equation (D.1).

D.2.2 Factor augmented autoregression

Based on equation (D.1), forecasts are computed as

$$\widehat{rx}_{t+1}^{(n)} = \widehat{\alpha} + \widehat{\beta}_W(L)rx_t^{(n)} + \widehat{\beta}_F'\widehat{\mathbf{F}}_t$$

where $\widehat{\mathbf{F}}_t$ is estimated using principal component methods as the first step. This model combines an AR(p) model with the above PCR model. The number of lags p is determined using the SIC.

D.2.3 Diffusion index

Diffusion index forecasts are computed as

$$\widehat{rx}_{t+1}^{(n)} = \widehat{\alpha} + \widehat{\beta}_W'W_t + \widehat{\beta}_F'\widehat{\mathbf{F}}_t$$

where W_t could be either the CP factor or the CP with lags of $rx_t^{(n)}$ in this experiment. Again, the latent factors are estimated using principal component methods, and the number of lags p is determined using the SIC.

D.3 Modified diffusion index models

Instead of the conventional diffusion index model, I also conduct some statistical learning methods for estimating $\widehat{\beta}_F$. For example, Kim and Swanson (2014) examined various "robust" estimation techniques including statistical learning algorithms and penalized regression methods (i.e., ridge regression, least angle regression, and the elastic net) to forecast macroeconomic variables. This paper considers the statistical learning techniques and a recent novel model called Factor-Lasso.

D.3.1 Bagging

Bootstrap aggregating (Bagging), which is introduced by Breiman (1996), first draw bootstrap samples from the original data, and averages the constructed prediction of bootstrap samples. Let $\widehat{Y}_b^* = \widehat{\beta}_b^* X_b^*$ be a bootstrap sample based predictor for $b = 1, ..., B$ denotes the b-th bootstrap sample. Then the bagging predictor is $\widehat{Y}_{bagging} = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_b^*$.

The bagging estimator can be represented in shrinkage form as Bühlmann et al. (2002) and Stock and Watson (2012). In this paper, I also perform the following bagging estimator:

$$\widehat{rx}_{t+1}^{(n)} = \widehat{\alpha} + \widehat{\beta}_W'W_t + \sum_{j=1}^{r}\psi(t_j)\widehat{\beta}_{Fj}'\widehat{\mathbf{F}}_{t,j},$$

where $\widehat{\alpha}, \widehat{\beta}$ are the least squares estimator from a regression of $rx_{t+1}^{(n)}$ on W_t, $\widehat{\beta}_{Fj}$ is a least squares estimator from a regression of residuals, $Z_t = rx_{t+1}^{(n)} - \widehat{\alpha} - \widehat{\beta}'_W W_t$ on $\widehat{\mathbf{F}}_{t,j}$, and t_j is the t-statistic associated with $\widehat{\beta}_{Fj}$. Specifically, $t_j = \sqrt{T}\widehat{\beta}_{Fj}/s_e$, where s_e is a Newey-West standard error. In addition, I follow Stock and Watson (2012) and Kim and Swanson (2014), and set $\psi(t) = 1 - \Phi(t+c) + \Phi(t-c) + t^{-1}[\phi(t+c) - \phi(t-c)]$, where c is the pretest critical value, Φ is the standard normal CDF, and ϕ is the standard normal density with $c = 1.96$. Here I set W_t be the CP factor with lags of $rx_t^{(n)}$, and the number of lags p is determined using the SIC.

D.3.2 Boosting

Boosting, which is introduced by Freund and Schapire (1995), constructs a user-determined set of functions such as least square estimators, and it is often called "learners". Then, boosting uses the set repeatedly on filtered data which are outputs from previous iterations of the learning algorithm. The output of a boosting takes the following form

$$\widehat{Y}^M = \sum_{m=1}^{M} \gamma_m f(X;\beta_m),$$

where the γ_m, $m = 1, ..., M$ are the weights, and $f(X;\beta_m)$ are functions of the dataset, X. Friedman (2001) proposed "L_2 Boosting " that employ the simple approach of refitting "base learners" to residuals from previous iterations. Bühlmann and Yu (2003) develop a boosting algorithm fitting "learners" using one predictor at a time for large numbers of predictors with i.i.d dataset. To deal with time-series, Bai and Ng (2009) modified the algorithm and this paper uses "Component-Wise L_2 Boosting" algorithm with least squares "learners".

In diffusion index contexts, I consider the boosting for both original W_t data and extracted factors, $\widehat{\mathbf{F}}_t$, and denote $\widehat{\mu}^M(\widehat{\mathbf{F}}_t)$ as the output of boosting algorithm. Finally, the boosting estimator is

$$\widehat{rx}_{t+1}^{(n)} = \widehat{\alpha} + \widehat{\beta}'_W W_t + \widehat{\mu}^M(\widehat{\mathbf{F}}_t),$$

where W_t is the CP factor with lags of $rx_t^{(n)}$, and the number of lags p is determined using the SIC.

D.3.3 Factor-Lasso

Factor-Lasso, which is recently introduced by Hansen and Liao (2019), is a nested model of large factor and variable selection. The main idea is that they take into account the variation in observables not captured by factors. Specifically, consider the folloing model:

$$rx_{t+1}^{(n)} = \alpha + \beta'_W W_t + \beta'_F \mathbf{F}_t + \theta' u_t + \epsilon_{t+1},$$

$$y_t = \mathbf{\Lambda F}_t + u_t, \quad t = 1, ..., T,$$

where θ is a $N \times 1$ vector and it assumed to be sparse. Including u_t is to capture idiosyncratic information in y_t, but only a few y have "useful remaining information" after factors are controlled. Note that lasso estimation of $\theta' u_t$ may affect confidence interval for $rx^{(n)}_{T+1|T}$.

Predictions are constructed using the following steps:

1. Obtain $\{\widehat{\mathbf{F}}_t, \widehat{u}_t\}_{t \leq T}$ by principal component methods from the factor model.

2. Estimate the diffusion index model, and obtain

$$\widehat{z}_{t+1} = rx^{(n)}_{t+1} - \widehat{\alpha} + \widehat{\beta}'_W W_t + \widehat{\beta}'_F \widehat{\mathbf{F}}_t.$$

3. Estimate θ using lasso on

$$\widehat{z}_{t+1} = \theta \widehat{u}_t + \epsilon_{t+1}.$$

4. Then forecasts are computed as

$$\widehat{rx}^{(n)}_{t+1} = \widehat{\alpha} + \widehat{\beta}'_W W_t + \widehat{\beta}'_F \widehat{\mathbf{F}}_t + \widehat{\theta}' \widehat{u}_t.$$

D.4 Penalized linear models

Instead of dimension reduction technique such as principal component analysis, imposing sparsity in the set of regressors through a penalty term is a common strategy to deal with a large set of predictors. By selecting a subset of variables, which have the strong predictive power among a large number of predictors, the penalized regression mitigate the overfitting problem. There are three popular methods with different types of penalty terms to the least squares loss function as follows.

D.4.1 Ridge regression

Ridge regression, which is introduced by Hoerl and Kennard (1970), solves the following problem:

$$\min L(\lambda, \theta) = \sum_{t=1}^{T-1} [xr^{(n)}_{t+1} - \alpha - \beta' y_t]^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\theta = (\alpha, \beta')$, $\beta = (\beta_1, ..., \beta_p)$, and $\lambda \geq 0$. Here y_t is both large macroeconomic and forward rates panel data. The ridge penalty term regularizes the regression cofficient and shrinks them toward zero.

D.4.2 Lasso regression

Lasso regression, which is introduced by Tibshirani (1996), solves the following problem:

$$\min L(\lambda, \theta) = \sum_{t=1}^{T-1} [xr_{t+1}^{(n)} - \alpha - \beta' y_t]^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\theta = (\alpha, \beta')$, $\beta = (\beta_1, ..., \beta_p)$, and $\lambda \geq 0$. Note that the L_2 ridge penalty term, $\sum_{j=1}^{p} \beta_j^2$, is replaced by the L_1 lasso penalty term, $\sum_{j=1}^{p} |\beta_j|$. The lasso penalty term regularizes the regression cofficient and exactly set to zero.

D.4.3 Elastic net

Elastic net regression, which is introduced by Zou and Hastie (2005), solves the following problem:

$$\min L(\lambda, \theta) = \sum_{t=1}^{T-1} [xr_{t+1}^{(n)} - \alpha - \beta' y_t]^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|,$$

where $\theta = (\alpha, \beta')$, $\beta = (\beta_1, ..., \beta_p)$, and $\lambda_1, \lambda_2 \geq 0$. Note that the L_2 ridge penalty term, $\sum_{j=1}^{p} \beta_j^2$, is replaced by the L_1 lasso penalty term, $\sum_{j=1}^{p} |\beta_j|$. Ridge and Lasso regressions are special cases of Elastic net by setting $\lambda_1 = \lambda$, $\lambda_2 = 0$ or $\lambda_2 = \lambda$, $\lambda_1 = 0$, respectively.

The tuning parameters, λ, λ_1, and λ_2, are predetermined by cross-validation using the in-sample data before performing forecasts. Penalized regressions introduced above still do consider the linear relations. To address the nonlinear relation, I also consider regression trees and neural networks.

D.5 Regression trees

Regression trees are popular in the machine learning literature, becuase it is simple, but powerful. Suppose first we have a partition of input space into M regions, $R_1, ..., R_M$. Then, we fit a simple linear model in each region of the vector of input y_t:

$$f(y_t) = \sum_{m=1}^{M} c_m I(y_t \in R_m).$$

Here y_t is both macroeconomic variables and the whole set of forward rates in this experiment. If we minimize the sum of squared residuals $\sum(xr_{t+1}^{(n)} - f(y_t))^2$, we can estimate $\widehat{c}_m$ by averaging the excess bond returns xr_{t+1} in region R_m as following:

$$\widehat{c}_m = E[xr_{t+1}^{(n)} | y_t \in R_m].$$

Since finding the optimal binary partition via minimum sum of squares is infeasible, I proceed the following tree-based method.

D.5.1 Decision tree

I first introduce a popular mehod called the Classification and Regression Tree (CART), introduced by Breiman et al. (1984). The CART is a universal underlying algorithm utilized for the estimation of regeression trees, and other tree-based methods such as Random Forests and Gradient Boosted Regression Trees are the modified versions of it. Algorithm 1 is a greedy algorithm to grow a complete binary regression tree.

Algorithm 1 Classification and Regression Trees (CART)

```
1. Initialize the stump. $R_1(0)$ denotes the range of all covariates, $R_l(d)$ denote the $l$-th node of depth $d$.

2. For $d = 1, ..., L$:

   • For $\tilde{R}$ in $\{R_l(d), l = 1, ..., 2^{d-1}\}$:

     Given splitting variable $j$ and each threshold level $\tau$, define a split regions

       $R_1(j,\tau) = \{Z|Z_j \leq \tau, Z_j \cap \tilde{R}\}$ and $R_2(j,\tau) = \{Z|Z_j > \tau, Z_j \cap \tilde{R}\}$.

     In the splitting regions set

       $c_1(j,\tau) \leftarrow \frac{1}{|R_1(j,\tau)|} \sum_{y_t \in R_1(j,\tau)} rx_{t+1}^{(n)}(y_t)$ and $c_2(j,\tau) \leftarrow \frac{1}{|R_2(j,\tau)|} \sum_{y_t \in R_2(j,\tau)} rx_{t+1}^{(n)}(y_t)$.

     Select the optimal split:

       $(j^*, \tau^*) = \underset{j,\tau}{\text{argmin}} \left[ \sum_{y_t \in R_1(j,\tau)} (rx_{t+1}^{(n)} - c_1(j,\tau))^2 + \sum_{y_t \in R_2(j,\tau)} (rx_{t+1}^{(n)} - c_2(j,\tau))^2 \right].$

     Update the nodes:

       $R_{2l-1}(d) \leftarrow R_1(j^*, \tau^*)$ and $R_{2l}(d) \leftarrow R_2(j^*, \tau^*)$

3. The output of a fully grown regression tree is given by:

       $\widehat{f}(y_t) = \sum_{k=1}^{2^L} \text{avg}(rx_{t+1}^{(n)} | y_t \in R_k(L)) I(y_t \in R_k(L)).$
```

D.5.2 Gradient boosting

A gradient boosting procedure, which is introduced by Friedman (2001), is a method for reducing the variance of the model estimates and increasing precision for regression and classification. Algorithm 2 summarizes the gradient boosted regression trees.

Algorithm 2 Gradient Boosted Regression Trees

```
1. Start $f_0(y_t) = \operatorname{argmin}_\theta \sum_{i=1}^{N} L(rx_{t+1}^{(n)}, \theta)$. Let $L(\cdot, \cdot)$ be the loss function.[10]
2. For $m = 1, ..., M$:
    (a) For $i = 1, ..., N$:
        compute negative gradient of loss function evaluated for current state of regressor $f = f_{m-1}$
            $\varepsilon_{im} = -\left[\frac{\partial L(rx_{t+1}^{(n)}, f(y_t))}{\partial f(y_t)}\right]_{f=f_{m-1}}.$
    (b) Train a regression tree with target $\varepsilon_{im}$ to get the terminal regions $S_{jm}$ for $j = 1, ..., J_m$. For $j = 1, ..., J_m$ compute:
            $\widehat{\theta}_{jm} = \underset{\theta}{\operatorname{argmin}} \sum_{y_t \in S_{jm}} L(rx_{t+1}^{(n)}, f_{m-1}(y_t) + \theta).$
    (c) Update the learner $f_m(y_t) = f_{m-1}(y_t) + \nu \sum_{j=1}^{J_m} \widehat{\theta}_{jm} I(y_t \in S_{jm})$, where $\nu \in (0, 1]$ is a hyperparameter.
3. The gradient boosted regression tree output is
            $\widehat{f}(y_t) = f_M(y_t).$
```

In practice, one need to determine the hyperparameters (i.e., learning rate, number of stage, minimum number of samples, maximum number of nodes, etc.) by cross-validation. Note that, the learning rate ν shrinks the contribution of each tree, while M captures the number of stages in the estimation.

Boosting procedure allows growing trees in an adaptive way to reduce the bias. Hence, trees are not identically distributed. The following alternative method builds a set of decorrelated trees which are estimated separately and then averaged out.

D.5.3 Random forests

Random forests, which is introduced by Breiman (2001), is a substantial modification of bootstrap aggregation (bagging) through averaging the outcome of independently drawn precesses to reduce the variance estimates. Bagging implies that the regression trees are identically distributed, and the variance of the average estimates depends on the variance of each tree times the correlation between trees. Random forests aim to reduce the correlation among trees in different bootstrap samples. Algorithm 3 summarizes the random forests procedure.

Algorithm 3 Random Forests

```
1. Obtain $B$ bootstrap samples from the original dataset.
2. For $b = 1, ..., B$, grow full trees following Algorithm 1 (i.e., CART algorithm introduced by Breiman et al. (1984)) with the following adjustments:
    (a) Select $p*$ variables from the original set of $p$ variables.
    (b) Choose the best variable/split-point, $(j, \tau)$, from $p*$ variables in Algorithm 1.
    (c) Split the node into two daughter nodes.
    (d) Recursively repeat the above procedures for each terminal node of the tree, until the minimum node size, $n_{min}$, is reached.
3. Denote the obtained tree by $T_b(y_t)$ for each $b$. The random forest output is
    $$\widehat{f}(y_t) = \frac{1}{B}\sum_{b=1}^{B} T_b(y_t).$$
```

Like other tree-based methods, the tuning parameters such as the depth of the trees, the number of bootstrap sample, the number of trees, and the size of the randomly selected sub-set of predictors are optivized by cross-validation. To reduce the computational costs, the hyperparameters are determined using the whole sample first before we conduct the out-of-sample forecast.

D.6 Neural networks

Neural network models, which emulate the neural architecture of brain and its processes, is the one of most powerful modeling method in machine learning. Hornik et al. (1989) prove that multilayer feed-forward networks are "universal approximators" for any smooth predictive association, when the complexity of the network (i.e., the number of hidden units) is allowed to grow with the sample size. In this paper, I implement the artificial neural network or multi-layer perceptrons with various neural networks structure.

Specifically, neural network structure can be described as follows. The models basically consist of an "input layer" of predictors, single or several "hidden layers", and an "output layer" which yields outcome prediction from aggregation of hidden layers. A simple one hidden layer model have the form

$$rx_{t+1}^{(n)} = b_0 + \sum_{j=1}^{q} b_j G(\gamma_j' y_t + \gamma_{0j}),$$

where $G(\cdot)$ is a nonlinear "activation function", q is the fixed number of hidden units, and y_t is the p dimensional input.[11] Then, there are a total of $(p+1) \times q + (q+1)$ parameters, where q parameters to reach each neuron and $(q+1)$ weights to aggregate the neurons into a single output. Note that the simplest neural network is a linear regression model when there is no hidden layer.

By adding hidden layers, one can construct build deeper network model, so-called "deep learning". Not surprixingly, the model with multiple hidden layers are often better approximator than single hidden layer model (see He et al., 2016). Let $Z^{(l)}$ denote the l-th hidden layer, containing $q^{(l)}$ number of hidden units, among L layers. Then the explicit structure of a deep prediction procedure is

$$\begin{aligned}
Z^{(1)} &= G^{(1)}(b^{(0)} + W^{(0)}Y), \\
Z^{(2)} &= G^{(2)}(b^{(1)} + W^{(1)}Z^{(1)}), \\
&\vdots \\
Z^{(L)} &= G^{(L)}(b^{(L-1)} + W^{(L-1)}Z^{(L-1)}), \\
rx_{t+1}^{(n)} &= b^{(L)} + W^{(L)}Z^{(L)},
\end{aligned}$$

where $G^{(l)}$ is activation function, $W^{(l)}$ is weight matrices, and $b^{(l)}$ is activation levels for $l = 1, ..., L$. When constructing a neural network, there are many choices such as the number of hidden layers, the number of neurons in each layer, and the choice of activation function for each of layers.

In practice, there are many potential choices for the nonlinear activation function such as sigmoid, softmax, hyperbolic, and rectified linear unit (ReLU) functions. I choose sigmoid function as activation function at all nodes, that is,

$$G(z) = 1/(1 + \exp(-z)).$$

Simple network models with a few layers and nodes in each layer for small dataset. In general, however, selecting a best network architecture using corss-validation is very demanding procedure, since it depends on the choice of activation functions, number of neurons, etc. However, training and regularizing neural networks may reduce the burden, and one only

[11] In this experiment, the inputs could be (i) macroeconomic variables with CP factor, (ii) both of forward rates and macroeconomic variable, or (iii) extracted factors with the CP factor.

need to determine the total number of layers and the maximum number of neurons in each layer. In addition, determining number of neurons in each layer follows the geometric pyramid rule as Masters (1993) suggested. To train a neural network, I use stochastic gradient decent (SGD) with tolerance for the optimization as $1e-3$ and maximum number of iterations as 3000. In addition, early stopping is considered to prevent over-fitting problem and improve the performance of trained model. For a more detailed training process, see Gu et al. (2020) and Bianchi et al. (2019).

In this paper, I consider three alternative specifications by varing the type of inputs as follows.

D.6.1 Neural Network

All 130 macroeconomic variables and the forward rates are inputs. This is a conventional neural network model. I consider the models having up to four hidden layers. NN1 denotes a neural network with a single hidden layer of 16 neurons; NN2 has two hidden layers with 16 and 8 neurons; and NN3 has three hidden layers with 16, 8, 4 neurons, respectively.

D.6.2 Hybrid Neural Network

A hybrid class of models that combines the least squares with neural networks is considered. 130 macroeconomic variables with a linear comibnation of forward rates (i.e., CP) are inputs. This hybrid model is based on a two step specification method. In the first step, the least square regression is conducted to predict the forecasting target using the model:

$$rx_{t+1}^{(n)} = \alpha + \beta CP_t + u_{t+1}.$$

In the second step, the residual, $\widehat{u}_{t+1}$, resulted from the first step is deployed as our forecasting target, and employ neural network with macroeconomic variables. I denote HNN as a hybrid neural network, and HNN1, HNN2, and HNN3, follow the same number of layers and neurons as the neural network specification in Section D.6.1.

D.6.3 Factor Augmented Neural Network

For a factor augmented neural network (FANN) model, extracted factors from 130 macroeconomic variables and CP factor are inputs. First, the estimated factors are obtained by the principal components (PC) methods. The number of factors are determined by Bai and Ng (2002). Then, the conventional neural network model is considered using the factors from macroeconomic variables and the CP factor as predictors. Due to the dimension reduction precedure from PC methods, there are only a few number of inputs in neural network model.

Hence, I only consider a single hidden layer model, and FANN1, FANN2, and FANN3 have 3, 5, 7 neurons, respectively. See Algorithm 4 for the detailed procedure.

Algorithm 4 Factor Augmented Neural Network (FANN) Forecasting

- Step 1: Obtain the estimated factors $\{\widehat{\mathbf{F}}_t\}_{t=1,\ldots,T}$ using the PC methods (e.g., Stock and Watson, 2002a).
- Step 2: Denote $\widehat{L}_t = (\widehat{\mathbf{F}}'_t, W'_t)'$. We estimate $g(\cdot)$ using the smooth sigmoid neural network sieve estimator. Specifically, $\widehat{g}(\widehat{L}_t) = \operatorname{argmax} L_n(\theta)$, where $L_n(\theta) = -T^{-1}\sum_{t=1}^{T-h}\frac{1}{2}[z_{t+h} - g(\widehat{L}_t)]^2$.
- Step 3: Forecast z_{T+h} using $\widehat{L}_T$.