THREE ESSAYS ON LARGE PANEL DATA ECONOMETRICS

By

SUNG HOON CHOI

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Economics
Written under the direction of
Yuan Liao
And approved by

_________________________

_________________________

_________________________

_________________________

New Brunswick, New Jersey
May, 2021
My dissertation consists of three chapters that focus on the development of new tools for use with big data, machine learning, and forecasting. In particular, I employ regularization methods from machine learning literature to improve the estimation of standard errors, efficient estimation of coefficients, and big data forecasts. Chapter 1 considers large factor models and proposes a new principal component analysis method that increases estimation accuracy and efficiency for big data forecasts. Chapters 2 and 3 develop the standard error of the ordinary least squares (OLS) estimation and the generalized least squares (GLS) estimation that take into account general dependences on the idiosyncratic error terms for linear panel data models.

All three chapters highlight the importance of cross-sectional heteroskedasticity and correlations of the error terms when the cluster structure is unknown. For example, idiosyncratic error variances can vary remarkably among different companies and the errors can be correlated among companies. In addition, the knowledge of cluster structure may be unavailable in practice. Taking these aspects into account is essential for obtaining more precise results for predictions or causal inference.

In Chapter 1, I propose a feasible weighted projected principal component analysis (FPPC) for factor models in which observable characteristics partially explain the latent factors. This novel method provides more efficient and accurate estimators than existing methods. To increase efficiency, I take into account both cross-sectional dependence and heteroskedasticity by using a consistent estimator of the inverse error covariance matrix as the weight matrix. To improve accuracy, I employ a projection approach using the
additionally observed characteristics because the projection removes noise components in high-dimensional factor analysis. By using the FPPC method, estimators of the factors and loadings have faster rates of convergence than those of the conventional factor analysis. Moreover, I propose an FPPC-based diffusion index forecasting model. The limiting distribution of the parameter estimates and the rate of convergence for forecast errors are obtained. Using U.S. bond market and macroeconomic data, I demonstrate that the proposed model outperforms models based on conventional principal component estimators. I also show that the proposed model performs well among a large group of machine learning techniques in forecasting excess bond returns.

Chapter 2, a joint work with Jushan Bai and Yuan Liao, develops a new standard-error estimator for linear panel data models. The proposed estimator is robust to heteroskedasticity, serial correlation, and cross-sectional correlation of unknown forms. The serial correlation is controlled by the Newey-West method. To control for cross-sectional correlations, we propose to use the thresholding method, without assuming the clusters to be known. We establish the consistency of the proposed estimator. Monte Carlo simulations show the method works well. We illustrate our method in an application of U.S. divorce law reform effects and find that cross-sectional correlations are non-negligible.

In Chapter 3, a co-authored paper with Jushan Bai and Yuan Liao, we consider the GLS estimation for linear panel data models. By estimating the large error covariance matrix consistently, the proposed feasible GLS estimator is more efficient than the OLS estimator in the presence of heteroskedasticity and both serial and cross-sectional correlations. The covariance matrix used for the feasible GLS is estimated via the banding and thresholding method. We establish the limiting distribution of the proposed estimator. A Monte Carlo study is considered. The proposed method is applied to the U.S. divorce rate data. We find that our more efficient estimators identify the significant effects of divorce law reforms on the divorce rate and provide tighter confidence intervals than existing methods.

1 This chapter is published in the Journal of Econometrics as a paper titled “Standard errors for panel data models with unknown clusters” (Bai et al., 2020).

2 This chapter is published in Empirical Economics as a paper titled “Feasible generalized least squares for panel data with cross-sectional and serial correlations” (Bai et al., 2021).
Acknowledgements

I wish to express my deepest gratitude to my advisor, Professor Yuan Liao. Without his extraordinary mentoring, boundless support, and unending encouragement, I might not have been able to complete the years of my doctoral studies compiled in this thesis. I am also deeply grateful to Professor Jushan Bai for his collaboration and constant guidance. In addition, I would like to express my warm and sincere thanks to my committee members, Professor John Landon-Lane and Professor Norman Swanson, for providing valuable suggestions and significant support. I am also indebted to Professor Xu Cheng, my outside committee member, for her insightful advice and suggestions.

I feel very fortunate to have the opportunity to have learned from many faculty members in the economic department at Rutgers. In particular, I appreciate Professors Roger Klein, Jennifer Hunt, Bruce Mizrach, and Xiye Yang for their rigorous training and enthusiastic education. I would also like to thank Hyun Hak Kim for his helpful advice and comments on my research and career development. In addition, I wish to show my gratitude to administrative staff members, especially Linda Zullinger, for their friendly assistance and support.

I have greatly benefited from and enjoyed my interactions with all my fellow colleagues during my studies. I am especially deeply indebted to Jungjun Choi for his comments and suggestions on this work. I will greatly miss the days and nights we discussed our research. I would also like to thank a number of previous and current graduate students at Rutgers for their valuable comments and suggestions: Yongbo Sim, Humberto Martinez Beltran, Kihwan Bae, Yixiao Jiang, Benda Yin, and Ryuichiro Izumi. My research in New Jersey Hall was greatly enhanced by thoughtful discussions and joyful chats with Sungmin Kang, Jun Liu, Hyukjun Kwon, Julián Fernández Mejía, Monica Mogollon Plazas, Ze Song, Andrew Garib, Xuesong Huang, Hyunji Ahn, Mark Avery, Min Kim, and many other young economists.
Finally, I would love to dedicate this dissertation to my family. Words cannot express my gratitude and appreciation to my parents in Seoul for their unconditional love and support. I also thank my parents-in-law for their best wishes to me. The last word of acknowledgment I have saved for my lovely wife, Yeseul Jung, who has given me endless support and unbounded love.
Dedication

To my wife and son,

Yeseul and Yoonsung,

and to my parents
# Contents

Abstract ii

Acknowledgements iv

Dedication vi

1 Feasible Weighted Projected Principal Component Analysis for Factor Models with an Application to Bond Risk Premia 1

1.1 Introduction 2

1.2 Feasible Weighted Projected Principal Component Analysis 6

1.2.1 Semiparametric factor model 6

1.2.2 Infeasible estimation 9

1.2.3 Implementation of FPPC 10

1.3 Asymptotic Analysis 12

1.3.1 Sparsity condition on $\Sigma_u$ 12

1.3.2 FPPC in conventional factor models 13

1.3.3 FPPC in semiparametric factor models 16

1.3.4 Diffusion index forecasting models 19

1.4 Monte Carlo Simulations 22

1.4.1 In-sample estimation 23

1.4.2 Out-of-sample forecast 24

1.5 Empirical Analysis: US Bond Risk Premia 26

1.5.1 Data 27

1.5.2 Experiment setup and forecast evaluation 28

1.6 Empirical Findings 31

1.6.1 In-sample analysis and economic interpretation of the factors 31
A.1.3 Individual factors ........................................ 99
A.2 Proof of Theorem 1.3.2 .................................... 102
  A.2.1 Convergence of loadings ............................... 103
  A.2.2 Convergence of factors ............................... 111
  A.2.3 Individual factors ..................................... 114
A.3 Proofs for Section 1.3.4 ................................. 115
  A.3.1 Proof of Theorem 1.3.3 ............................... 115
  A.3.2 Proof of Theorem 1.3.4 ............................... 117
A.4 Forecasting methods in Details ......................... 118
  A.4.1 Univariate autoregression model .................... 118
  A.4.2 Diffusion index models ............................ 118
  A.4.3 Modified diffusion index models .................. 119
  A.4.4 Penalized linear models ........................... 122
  A.4.5 Regression trees .................................. 123
  A.4.6 Neural networks .................................. 126

B Appendix for Chapter 2 ................................. 130
  B.1 Proofs ................................................. 130
    B.1.1 Technical lemmas .................................. 130
    B.1.2 Proof of Theorem 2.2.1. ......................... 133

C Appendix for Chapter 3 ................................. 136
  C.1 Proofs ................................................. 136
    C.1.1 Technical lemmas .................................. 136
    C.1.2 Proof of Theorem 3.2.1. ......................... 140
    C.1.3 Proof of Theorem 3.2.2. ......................... 144

Bibliography ................................................ 153
List of Tables

1.1 Three different principal component analysis methods. . . . . . . . . . . . 11
1.2 Canonical correlations of estimated loading or factor matrices: the larger
the better. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
1.3 Mean squared error of estimated common components \( \mathbf{AF}' \): the smaller
the better. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
1.4 Out-of-sample relative mean squared forecast error: the smaller the better. 26
1.5 Components of \( \mathbf{X}_t \) for U.S. bonds excess return forecasting. . . . . . 28
1.6 List of all forecasting models . . . . . . . . . . . . . . . . . . . . . . . . 29
1.7 In-sample adjusted \( R^2 \) of predictive regressions for excess bond returns:
the larger the better. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
1.8 Relative mean squared forecast errors of U.S. bonds excess return forecast-
ing: the smaller the better. Forecasting sample period: 1984:1-2016:4. 34
1.9 Out-of-sample \( R^2 \) (%) of U.S. bonds excess return forecasting: the larger
the better. Forecasting sample period: 1984:1-2016:4. . . . . . . . . . . . . 35
1.10 Out-of-sample \( R^2 \) (%) of U.S. bonds excess return forecasting in recessions
and expansions: the larger the better. Rolling window. . . . . . . . . . . . . 36
1.12 Relative mean squared forecast errors of U.S. bonds excess return forecast-
ing. Forecasting sample period: 1984:1-2016:4. . . . . . . . . . . . . . . 41
1.13 Out-of-sample \( R^2 \) (%) of U.S. bonds excess return forecasting: the larger
the better. Forecasting sample period: 1984:1-2016:4. . . . . . . . . . . . . 42
1.14 Pairwise model comparison using Giacomini-White tests for 2-year maturity. 43
1.15 Pairwise model comparison using Giacomini-White tests for 3-year maturity. 44
1.16 Pairwise model comparison using Giacomini-White tests for 4-year maturity. 45
1.17 Pairwise model comparison using Giacomini-White tests for 5-year maturity. 46
1.18 Relative mean squared forecast errors of U.S. bonds excess return forecasting for different prediction periods. Rolling window. 47

2.1 Null rejection probabilities, 5% level. Two-tailed test of $H_0 : \beta = 1$. Case 1: No cross-sectional correlation ($\gamma = 0$). 66

2.2 Null rejection probabilities, 5% level. Two-tailed test of $H_0 : \beta = 1$. Case 1: Cross-sectional correlation ($\gamma = 1$). 67

2.3 Null rejection probabilities, 5% level. Two-tailed test of $H_0 : \beta = 1$. Case 1: Both strong serial and cross-sectional correlations ($\rho = 0.9, \gamma = 1$). 68

2.4 Null rejection probabilities, 5% level. Two-tailed test of $H_0 : \beta = 1$. Case 2: Errors with Spatial AR(1) structure ($\psi = 0.5$). 68

2.5 Null rejection probabilities, 5% level. Two-tailed test of $H_0 : \beta = 1$. Case 3: Errors with Factor structure ($\rho_F = 0.9, \rho_\lambda = 0.3$). 69

2.6 Empirical application: effects of divorce law reform with state and year fixed effects: US state level data annual from 1956 to 1988, dependent variable is divorce rate per 1000 persons per year. OLS estimates and standard errors (using state population weights). 70

3.1 Performance of estimated $\beta_0$; true $\beta_0 = 1$; $i$-dependent serial correlation and weak cross-sectional correlation ($\gamma = 0.3$). 90

3.2 Performance of estimated $\beta_0$; true $\beta_0 = 1$; $i$-dependent serial correlation and strong cross-sectional correlation ($\gamma = 0.7$). 90

3.3 Empirical application: effects of divorce law reform with state and year fixed effects: US state level data annual from 1956 to 1988, dependent variable is divorce rate per 1000 persons per year. OLS and FGLS estimates and standard errors (using state population weights). 91
## List of Figures

1.1 Cross-sectional heteroskedasticity and correlations . . . . . . . . . . . . 3
1.2 Marginal $R^2$ for extracted factors . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
Chapter 1

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

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

First, it is essential to consider a large error covariance matrix estimator for efficient estimation. In linear regression models, for example, it is well known that the generalized least squares (GLS) estimators are more efficient than the ordinary least squares (OLS) estimators in the presence of cross-sectional heteroskedasticity. Similarly, factor models often require the idiosyncratic error components to be cross-sectionally heteroskedastic and correlated. Intuitively, the large error covariance matrix, $\Sigma_u = \text{cov}(u_t)$, is a non-diagonal matrix, and the diagonal entries may vary widely (e.g., Bai and Liao, 2016). Figure 1.1 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 analysis (PC) method does not require estimating $\Sigma_u$, and it essentially treats $u_{it}$ to be homoskedastic and uncorrelated over $i$. Hence, it is inefficient. In this paper, I consider consistently estimating the high-dimensional error covariance matrix and its inverse. Using the estimator for $\Sigma_u^{-1}$ 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 analysis method that incorporates observed covariates into factor models. However, this method assumes that the error covariance matrix is known and does not consider the large error covariance matrix estimator.

---

1 Choi (2012) studied efficient estimations using weighted least squares in the conventional factor model, but assumed $\Sigma_u$ to be known.
Figure 1.1: Cross-sectional heteroskedasticity and correlations

Note: The first figure shows the estimated error variance, \( \text{var}(u_{it}) \), 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 analysis (PPC), 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 estimate the factors more accurately than the conventional PC method.

I introduce a Feasible Weighted Projected Principal Component Analysis (FPPC) which takes into account the importance of accuracy and efficiency. 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 addition, 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 Oka (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 investigates
whether and how the FPPC method can improve predictive accuracy in the DI model.

I then apply my proposed FPPC-based DI model to the U.S. Treasury bond market. In this literature, the determinants of bond risk premia are thought to be crucial for both policymakers and investors (e.g., Campbell and Shiller, 1991; Cochrane and Piazzesi, 2005; Fama and Bliss, 1987). Such bond risk premia could be closely linked to macroeconomic factors. Among others, Ludvigson and Ng (2009) investigated critical linkages between bond returns and macroeconomic factors. The latent macroeconomic factors are estimated using the conventional PC method from a monthly balanced panel of macroeconomic time series. Moreover, they forecast excess returns of U.S. bonds using the conventional DI forecasting model. However, by using the FPPC method instead of the PC method, we can gain predictive accuracy on excess bond returns. In a recent paper, Bianchi et al. (2021) studied how machine learning methods (such as regression trees and neural networks) provide strong statistical evidence in predicting 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.

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 (see Giacomini and White, 2006).

This paper contributes to the literature in three ways. First, I develop a new methodology to estimate the parameters of semiparametric factor models. By using the consistent estimator of the 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 removes noise components gives more accurate estimators. To the best of my knowledge, this paper is the first to accurately and efficiently estimate factors 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. 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 are inefficient. Second, this paper includes a study of the DI forecasting model based on the FPPC method. The gain of forecasts using my model is substantial because it takes into account (i) the characteristics of the latent factors and (ii) cross-sectional correlation and heteroskedasticity. 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 (e.g., 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 rest of the chapter is organized as follows. Section 1.2 formally proposes the FPPC method. Section 1.3 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 1.4 provides simulation studies. In Section 1.5, the econometric framework for the empirical study is demonstrated. I then introduce the data and the experimental setup, including details of the statistics used to analyze results. Section 1.6 summarizes key empirical findings. Finally, Section 1.7 concludes. All proofs and descriptions of all machine learning forecasting methods are given in the Appendix.

Throughout this chapter, I use $\lambda_{\text{min}}(A)$ and $\lambda_{\text{max}}(A)$ to denote the minimum and maximum eigenvalues of a square matrix $A$. I also let $\|A\|_F = \sqrt{\text{tr}(A^TA)}$, $\|A\|_2 = \sqrt{\lambda_{\text{max}}(A^TA)}$, 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 $a$.

1.2 Feasible Weighted Projected Principal Component Analysis

1.2.1 Semiparametric factor model

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

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + u_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T, \quad (1.2.1)$$

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

$$f_{tk} = g_k(X_t) + \gamma_{tk}, \quad t = 1, \ldots, T, k = 1, \ldots, K, \quad (1.2.2)$$

where $X_t$ is a $d \times 1$ vector of observable covariates and $g$ is a unknown nonparametric function. For example, $X_t$ can be the Fama-French factors or aggregated macroeconomic variables. Here, $\gamma_{tk}$ is the component of common factors that cannot be explained by the covariates $X_t$. Recently, Fan et al. (2020) studied a similar model and proposed

\footnote{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.}
a robust estimator for heavy-tailed errors. Define $\gamma_t = (\gamma_{t1}, \cdots, \gamma_{tK})'$. I assume that $\{\gamma_t\}_{t \leq T}$ have mean zero, and are independent of $\{X_t\}_{t \leq T}$ and $\{u_{it}\}_{i \leq N, t \leq T}$. Then the model (2.1) and (1.2.2) can be represented using the following factor structure:

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} (g_k(X_t) + \gamma_{tk}) + u_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T.$$  \hfill (1.2.3)

The model (1.2.3) can be stacked and written in a full matrix notation as

$$Y = \Lambda (G(X) + \Gamma)' + U,$$  \hfill (1.2.4)

where $Y$ is the $N \times T$ matrix of $y_{it}$, $\Lambda$ is the $N \times K$ matrix of $\lambda_{ik}$, $G(X)$ is the $T \times K$ matrix of $g_k(X_t)$, $\Gamma$ is the $T \times K$ matrix of $\gamma_{tk}$ and $U$ is $N \times T$ matrix of $u_{it}$. Note that the common factor matrix can be decomposed by $F = G(X) + \Gamma$ from the model (1.2.2). Also $E(\Gamma | X) = 0$, where $G(X)$ and $\Gamma$ are orthogonal factor components so that $E[G(X)\Gamma'] = 0$. This paper assumes $K = \dim(F_t)$ and $d = \dim(X_t)$ to be constant. In addition, the number of factors $K$ is assumed to be unknown. 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 $g_k(X_t)$ does not depend on $i$, which means the common factors represent the time heterogeneity only. To estimate $g_k(X_t)$, $g_k(\cdot)$ is assumed to be additive for multivariate covariates $X_t$. Define, for each $k \leq K$ and for each $t \leq T$,

$$g_k(X_t) = \phi'(X_t)b_k + \sum_{l=1}^{d} R_{kl}(X_{tl}),$$  \hfill (1.2.5)

where

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

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

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 $l$th additive component of $g_k(X_t)$ for the $k$th 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 $\Phi(X) = (\phi(X_1), \ldots, \phi(X_T))'$ be a $T \times (Jd)$ matrix of basis functions, $B = (b_1, \ldots, b_K)'$ be a $K \times (Jd)$ matrix of sieve coefficients, and $R(X)$ be $T \times K$ matrix with the $(t, k)$th element $\sum_{l=1}^{d} R_{kl}(X_{tl})$. Then (1.2.5) can be written in the matrix form:

$$G(X) = \Phi(X)B' + R(X).$$

(1.2.6)

Then the model (1.2.4) can be rewritten as

$$Y = \Lambda(\Phi(X)B' + \Gamma)' + \Lambda R(X)' + U.$$  

(1.2.7)

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

$$YP = \Lambda B\Phi(X)' + \tilde{E},$$

(1.2.8)

where $\tilde{E} = \Lambda \Gamma'P + \Lambda R(X)'P + UP \approx 0$ because $\Gamma$ and $U$ are orthogonal to the function space spanned by $X$, and $\Lambda R(X)'$ 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. However, the idiosyncratic components are often cross-sectionally heteroskedastic and correlated in factor models. Since the PPC method does not require estimating the $N \times N$ covariance matrix, $\Sigma_u = \text{cov}(u_t)$, 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

3 Fan et al. (2016) considered the similar semi-parametric factor model, but the factor loading has the semiparametric structure, such as: $\lambda_{ik} = g_k(X_i) + \gamma_{ik}, i = 1, \ldots, N$. 
semiparametric factor models:

\[
\min_{\Lambda, B} \sum_{t=1}^{T} (y_t - \Lambda B \phi(X_t))'W(y_t - \Lambda B \phi(X_t))
\]

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 \( W = \Sigma_u^{-1} \). However, the optimal weight is usually infeasible. Hence, the proposed FPPC method requires a consistent estimator \( \hat{\Sigma}_u^{-1} \) as the feasible weight matrix. It is commonly used in the generalized method of moments literature. To implement the proposed method, we first project \( Y \) onto the sieve space spanned by \( \{X_t\}_{t \leq T} \), 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 \( \hat{\Sigma}_u^{-1} \) for \( \Sigma_u^{-1} \) under the conditional sparsity assumption. More specific estimation procedure is discussed in the following sections.

### 1.2.2 Infeasible estimation

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

\[
P = \Phi(X)(\Phi(X)'\Phi(X))^{-1}\Phi(X)',
\]

which is chosen as the projection matrix onto \( \mathcal{X} \). Let \( \Sigma_u \) 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 (1.2.9) with \( W = \Sigma_u^{-1} \) as the optimal weight matrix. Concentrating out \( B \) and using the normalization that \( \frac{1}{N} \Lambda'\Sigma_u^{-1}\Lambda = I_K \), the optimization problem is identical to maximizing \( \text{tr}(\Lambda'\Sigma_u^{-1}YPY'\Sigma_u^{-1}\Lambda) \). Let \( \Lambda^* = \Sigma_u^{-\frac{1}{2}}\Lambda \) and \( Y^* = \Sigma_u^{-\frac{1}{2}}Y \). The estimated (infeasible) weighted loading matrix, denoted by \( \hat{\Lambda}^* \), is \( \sqrt{N} \) times the eigenvectors corresponding to the \( K \) largest eigenvalues of the \( N \times N \) matrix \( Y^*PY'^* = \Sigma_u^{-\frac{1}{2}}YPY'\Sigma_u^{-\frac{1}{2}} \). Note that the infeasible estimator of \( \Lambda \) is \( \bar{\Lambda} = \Sigma_u^{\frac{1}{2}}\hat{\Lambda}^* \). Then given \( \hat{\Lambda}^* \),

\[
\hat{G}(X) = \frac{1}{N}PY'^*\hat{\Lambda}^*\n\]

is the estimator of \( G(X) \).

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

$$\hat{F} = \frac{1}{N} Y^{**'} \hat{\Lambda}^*.$$  

In addition, by (1.2.4), an estimator of $\Gamma$ is

$$\hat{\Gamma} = \hat{F} - \bar{G}(X) = \frac{1}{N} (I - P) Y^{**'} \hat{\Lambda}^*.$$  

1.2.3 Implementation of FPPC

The estimators are feasible only when a consistent estimator $\hat{\Sigma}_{u}^{-1}$ for $\Sigma_{u}^{-1}$ is obtained. Therefore, this paper considers $W = \hat{\Sigma}_u^{-1}$ as the asymptotically optimal weight matrix, which takes into account both heteroskedasticity and cross-sectional correlations simultaneously.

The estimator of $\Sigma_u$ and FPPC

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

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

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

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

where $(x)_+ = x$ if $x \geq 0$, and zero otherwise. Here $\text{sgn}(\cdot)$ denotes the sign function.

Note that other thresholding functions are possible such as hard thresholding. For the threshold value, I specify

$$\tau_{ij} = M \omega_{N,T} \sqrt{\tilde{R}_{ii} \tilde{R}_{jj}},$$

where $\omega_{N,T} = \left( \sqrt{\log N} \right) / \sqrt{T} + \frac{1}{\sqrt{N}}$. 


Table 1.1: Three different principal component analysis methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective Function</th>
<th>Eigenvectors of</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>$\sum_{t=1}^{T}(y_t - A F_t)'(y_t - A F_t)$</td>
<td>$Y Y'$</td>
</tr>
<tr>
<td>PPC</td>
<td>$\sum_{t=1}^{T}(y_t - \Lambda B \phi(X_t))' (y_t - \Lambda B \phi(X_t))$</td>
<td>$Y P Y'$</td>
</tr>
<tr>
<td>FPPC</td>
<td>$\sum_{t=1}^{T}(y_t - \Lambda B \phi(X_t))' \hat{\Sigma}_u^{-1}(y_t - \Lambda B \phi(X_t))$</td>
<td>$\hat{\Sigma}_u^{-1/2} Y P Y' \hat{\Sigma}_u^{-1/2}$</td>
</tr>
</tbody>
</table>

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 1.2.3. Intuitively, $\hat{\Sigma}_u$ thresholds off the small entries of the sample covariance matrix $\frac{1}{T} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$, where residuals are obtained from the PPC estimate.

Now, I introduce the FPPC estimators using $\hat{\Sigma}_u^{-1}$ as the feasible weight matrix. Let $\tilde{Y} = \hat{\Sigma}_u^{-\frac{1}{2}} Y$, $\tilde{A} = \hat{\Sigma}_u^{-\frac{1}{2}} \Lambda$, and $\tilde{U} = \hat{\Sigma}_u^{-\frac{1}{2}} U$. Then the estimated feasible weighted loading matrix for $\Sigma_u^{-\frac{1}{2}} A$, denoted by $\hat{\Lambda}$, is $\sqrt{N}$ times the eigenvectors corresponding to the $K$ largest eigenvalues of the $N \times N$ matrix $\tilde{Y} P \tilde{Y}' = \hat{\Sigma}_u^{-\frac{1}{2}} Y P Y' \hat{\Sigma}_u^{-\frac{1}{2}}$. Note that the estimator of $\Lambda$ is $\hat{\Lambda} = \hat{\Sigma}_u^{\frac{1}{2}} \Lambda$. With the estimated weighted factor loadings $\hat{\Lambda}$, the least-squares estimator of common factor matrix is

$$\hat{F} = \frac{1}{N} \tilde{Y}' \hat{\Lambda} = \frac{1}{N} Y' \hat{\Sigma}_u^{-1} \hat{\Lambda}. \quad (1.2.11)$$

Moreover, given $\hat{\Lambda}$,

$$\hat{G}(X) = \frac{1}{N} P \tilde{Y}' \hat{\Lambda}, \quad \hat{\Gamma} = \frac{1}{N} (I - P) \tilde{Y}' \hat{\Lambda}. \quad (1.2.12)$$

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

In Section 1.3, 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, $\hat{\Lambda} / \sqrt{N}$, are estimated from three different matrices. Table 1.1 shows the main differences of the estimators.

**Choice of threshold**

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

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

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

$$c_{\text{min}} = \inf \{C > 0 : \lambda_{\text{min}}(\hat{\Sigma}_u(M)) > 0, \forall M > C \}.$$

Then the resulting estimator is $\hat{\Sigma}_u(M^*)$.

### 1.3 Asymptotic Analysis

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

#### 1.3.1 Sparsity condition on $\Sigma_u$

In the literature, one of the commonly used assumptions to estimate a high-dimensional covariance matrix is the sparsity. This paper assumes $\Sigma_u$ 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 (2008a) and Fan et al. (2013). Consider the notion of generalized sparsity: let $\Sigma_u = (\Sigma_{u,ij})_{N \times N}$. For some $q \in [0, 1)$, define

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

(1.3.1)
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 \( \Sigma_u \).

**Assumption 1.3.1.** (i) There is \( q \in [0, 1) \) such that

\[
m_N \omega_{N,T}^{1-q} = o(1), \quad \text{where} \quad \omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}.
\]

(ii) There are constant \( c_1, c_2 > 0 \) such that \( \lambda_{\min}(\Sigma_u) > c_1 \) 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 \( \Sigma_u \) and its inverse. Condition (ii) requires that \( \Sigma_u \) 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 1.3.1.** Similar to Fan et al. (2013), for \( m_N \) and \( q \) defined in (1.3.1), we have

\[
\| \hat{\Sigma}_u^{-1} - \Sigma_u^{-1} \|_1 = O_P(m_N \omega_{N,T}^{1-q}),
\]

if (1.3.2) holds. When \( m_N \) grows slowly with \( N \), \( \hat{\Sigma}_u^{-1} \) 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

\[
\| \frac{1}{\sqrt{N}} \Lambda'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})u_t \| = o_P(1).
\]

Specifically, the above “absolute convergence” for the estimator would be too restrictive to be applicable when \( N > T \) (see Bai and Liao, 2017).

### 1.3.2 FPPC in conventional factor models

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

\[
Y = \Lambda F' + U.
\]
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.⁴ Recall that the projection matrix is defined as

\[ P = \Phi(X)(\Phi(X)'\Phi(X))^{-1}\Phi(X)'. \]

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

**Assumption 1.3.2.** (Genuine projection). There are positive constants \( c_1 \) and \( c_2 \) such that, with probability approaching one as \( T \to \infty \),

\[ c_1 < \lambda_{\min}(T^{-1}F'PF) < \lambda_{\max}(T^{-1}F'PF) < c_2. \]

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 \( \Phi(X) \) and \( F \) 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 nonsingular matrix \( M \), \( AF' = AM^{-1}MF' \), it has been well known that \( A \) and \( F \) are not separately identifiable without further restrictions (see Bai and Ng, 2013). Similar to Stock and Watson (2002a) and Bai (2003), the FPPC estimator estimates transformed factors and loadings.

**Assumption 1.3.3.** (Basis functions). (i) There are \( d_1, d_2 > 0 \) so that with probability approaching one as \( T \to \infty \),

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

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

Since \( T^{-1}\Phi(X)'\Phi(X) = T^{-1}\sum_{t=1}^{T} \phi(X_t)\phi(X_t)' \) and \( \phi(X_t) \) is a \( Jd \times 1 \) vector, where \( Jd \ll T \), the strong law of large numbers implies condition (i). This condition can

---

⁴ 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: \( \lambda_{ik} = g_k(X_i) + \gamma_{ik}, \) for \( i = 1, ..., N, k = 1, ..., K \).
be satisfied over normalizations of commonly used basis functions, e.g., Fourier basis, B-splines, polynomial basis. In addition, we may allow serial dependence and nonstationarity on \( \{X_t\}_{t \leq T} \) in this paper.

**Assumption 1.3.4. (Data generating process).** (i) \( \{F_t, u_t\}_{t \leq T} \) is strictly stationary; \( \{u_t\}_{t \leq T} \) is serially independent, that is, \( E(u_t u_s') = 0 \) if \( t \neq s \), for \( t, s \leq T \). \( Eu_t = 0 \) for all \( i \leq N, t \leq T \); \( \{u_t\}_{t \leq T} \) is independent of \( \{X_t, F_t\}_{t \leq T} \).

(ii) Strong mixing: There exist \( r_1, C > 0 \) such that for all \( T > 0 \),

\[
\sup_{A \in F_{-\infty}^0, B \in F_T^\infty} |P(A)P(B) - P(AB)| < \exp(-CT^{r_1}),
\]

where \( F_{-\infty}^0 \) and \( F_T^\infty \) denote the \( \sigma \)-algebras generated by \( \{(F_t, u_t) : -\infty \leq t \leq 0\} \) and \( \{(F_t, u_t) : T \leq t \leq \infty\} \), 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

\[
\max_{t \leq T} \sum_{s=1}^{T} |Eu_{it}u_{is}| < M, \\
\frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} |Eu_{it}u_{js}| < M, \\
\max_{t \leq T} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{s=1}^{T} \sum_{q=1}^{T} |\text{cov}(u_{it}u_{is}, u_{jqt})| < M.
\]

Condition (ii) allows factors and idiosyncratic components to be weakly serial dependent by requiring 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, \( \{u_t\}_{i \leq N, t \leq T} \). 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 1.2.3.

**Theorem 1.3.1.** (Conventional factor model). Suppose that Assumptions 1.3.1(ii)-1.3.4 hold and $m_N\delta_{N,T}^{1-q} = o(1)$, for $\delta_{N,T} = \sqrt{\log N/T} + \sqrt{J/T} + \frac{1}{\sqrt{N}}$. For an invertible matrix $M$, 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

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

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

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

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

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 $\hat{\Sigma}_u^{-1}$, it requires large-$N$ and large-$T$ for the factor and loading estimates. I discuss additional details further below.

1.3.3 FPPC in semiparametric factor models

In the semiparametric factor model, it is assumed that $f_{tk} = g_k(X_t) + \gamma_{tk}$. Here $g_k(X_t)$ is a nonparametric smooth function for the observed covariates, and $\gamma_{tk}$ is the unobserved random factor component, which is independent of $X_t$. In the matrix form, the model can be written as:

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

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

\[
\tilde{Y}P = \tilde{A}B\Phi(X)' + E, \quad (1.3.5)
\]

where \( E = \tilde{A}R(X)'P + \tilde{A}\Gamma'P + \tilde{U}P \) is approximately “small”, because \( R(X) \) is the sieve approximation error, and \( \Gamma \) and \( \tilde{U} \) are orthogonal to the function space spanned by \( X \). The sieve coefficient matrix \( B \) can be estimated by least squares from the above model (1.3.5) as:

\[
\hat{B} = (\hat{b}_1, \cdots, \hat{b}_K)' = \frac{1}{N} \tilde{A}'\tilde{Y}\Phi(X)[\Phi(X)\Phi(X)]^{-1}. \quad (1.3.6)
\]

Then the estimator for \( g_k(\cdot) \) is

\[
\hat{g}_k(x) = \phi(x)'\hat{b}_k \quad \forall x \in \mathcal{X}, k = 1, \cdots, K, \quad (1.3.7)
\]

where \( \mathcal{X} \) denotes the support of \( X_t \).

The estimators \( \hat{A}, \hat{G}(X) \) and \( \hat{F} \) are the FPPC estimators as defined in Section 1.2.3. Since \( F = G(X) + \Gamma, G(X) \) can be regarded as the projection of \( F \) onto the sieve space spanned by \( X \). Therefore, the following assumption is a sufficient condition for Assumption 1.3.2 in the semiparametric factor model.

**Assumption 1.3.5.** There are two positive constants \( c_1 \) and \( c_2 \) so that with probability approaching one as \( T \to \infty \),

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

Serial weak dependence for \( \{\gamma_t\}_{t \leq T} \) is imposed as following.

**Assumption 1.3.6.** (i) \( E\gamma_{tk} = 0 \) and \( \{X_t\}_{t \leq T} \) is independent of \( \{\gamma_{tk}\}_{t \leq T} \).

(ii) Define \( \gamma_t = (\gamma_{t1}, \cdots, \gamma_{tK})' \), and

\[
\nu_T = \max_{k \leq K} \frac{1}{T} \sum_{t \leq T} \text{var}(\gamma_{tk}).
\]
Then $\max_{k \leq K, t \leq T} E g_{k}(X_t)^2 \leq \infty, \nu_T < \infty$ and

$$
\max_{k \leq K, s \leq T} \sum_{t \leq T} |E \gamma_k \gamma_s| = O(\nu_T).
$$

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

**Assumption 1.3.7. (Accuracy of sieve approximation).** For all $l \leq d, k \leq K$,

(i) the factor component $g_{kl}(\cdot)$ belongs to a Hölder class $\mathcal{G}$ defined by

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

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 $l$th element of $X_t$, 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 1.3.2. (Semiparametric factor model).** Suppose $J = o(\sqrt{T})$ and Assumptions 1.3.1, 1.3.3-1.3.7 hold. There is an invertible matrix $H$, as $N, T, J \to \infty$, we have, for $\omega_{N,T} = \sqrt{\log N \over T} + {1 \over \sqrt{N}}$,

$$
1 \over N \| \widehat{\Lambda} - \Lambda H \|^2_F = O_P \left( 1 \over T \right),
$$

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

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

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

$$
\| \widehat{F}_t - H^{-1} F_t \| = O_P \left( m_N \omega_{N,T}^{1-q} \right).
$$
Note that $\hat{\mathbf{F}} = \hat{\mathbf{G}}(\mathbf{X}) + \hat{\mathbf{F}}$, 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 1.3.2.** Denote $\hat{\mathbf{L}} = (\hat{\lambda}_1, ..., \hat{\lambda}_N)'$, $\hat{\mathbf{G}}(\mathbf{X}) = (\hat{g}(\mathbf{X}_1), ..., \hat{g}(\mathbf{X}_T))'$, and $\hat{\mathbf{F}} = (\hat{\gamma}_1, ..., \hat{\gamma}_T)'$. For the factor loading, we have

$$\frac{1}{N} \sum_{i=1}^{N} \|\hat{\lambda}_i - \mathbf{H}'\lambda_i\|^2 = \text{OP}(\frac{1}{T}).$$

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

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

Moreover, when $N = O(1)$ and $\kappa$ 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 $(\hat{\lambda}_i, \hat{\mathbf{F}}_t)$, such as Stock and Watson (2002a) and Bai (2003): for some rotation matrix $\hat{\mathbf{H}}$,

$$\frac{1}{N} \sum_{i=1}^{N} \|\hat{\lambda}_i - \tilde{\mathbf{H}}'\lambda_i\|^2 = \text{OP}\left(\frac{1}{T} + \frac{1}{N}\right), \quad \frac{1}{T} \sum_{t=1}^{T} \|\hat{\mathbf{F}}_t - \tilde{\mathbf{H}}^{-1}\mathbf{F}_t\|^2 = \text{OP}\left(\frac{1}{T} + \frac{1}{N}\right).$$

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

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

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

### 1.3.4 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:

\[ z_{t+h} = \alpha' F_t + \beta' W_t + \epsilon_{t+h}, \]  

(1.3.8)

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

\[ y_t = \Lambda F_t + u_t. \]  

(1.3.9)

Note that, when \( z_t \) is a scalar, equations (1.3.8) and (1.3.9) constitute the DI model. In addition, the equation (1.3.8) is the FAVAR model of Bernanke et al. (2005), when \( h = 1 \) and \( z_{t+1} = (F'_{t+1}, W'_{t+1})' \). Intuitively, the common factor, \( F_t \), is known for the common shocks that generate comovements in economic time series.

Suppose the interest object is the conditional mean of (1.3.8), which is

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

where \( L_t = (F'_t, W'_t)' \). For example, if \( z_t \) employment rate, the estimated conditional mean can be interpreted as an estimate of the expected employment rate. Let \( \hat{\alpha} \) and \( \hat{\beta} \) be the least squares estimates from regression \( z_{t+h} \) on \( \hat{L}_t = (\hat{F}'_t, W'_t)' \), for \( t = 1, ..., T-h \), where \( \hat{F}_t \) is the estimated factors using FPPC. Then the feasible prediction would be

\[ \hat{z}_{T+h|T} = \hat{\alpha}' \hat{F}_T + \hat{\beta}' W_T = \hat{\delta}' \hat{L}_T. \]

Stock and Watson (2002a) proved that \( \hat{\delta} \) is consistent for \( \delta \) and \( \hat{z}_{T+h|T} \) 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 1.2.3. To do so, the following assumption is required.

**Assumption 1.3.8.** Let \( L_t = (F'_t, W'_t)' \). \( E\|L_t\|^4 \) is bounded for every \( t \).
(i) $E(\epsilon_{t+h}|z_t, L_t, z_{t-1}, L_{t-1}, \ldots) = 0$ for any $h > 0$, and $L_t$ and $\epsilon_t$ are independent of the idiosyncratic errors $u_{ts}$ for all $i$ and $s$.

(ii) $\frac{1}{T} \sum_{t=1}^{T} L_t L_t' \rightarrow^p \Sigma_L$, which is a positive definite matrix.

(iii) $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} L \epsilon_{t+h} \xrightarrow{d} N(0, \Sigma_{L, \epsilon})$, where $\Sigma_{L, \epsilon} = \plim \frac{1}{T} \sum_{t=1}^{T} \epsilon_{t+h}^2 L_t L_t'$.

Assumption 1.3.8 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 1.3.3 for the equation (1.3.9). 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 1.3.3.** *(Estimation)* Let $\hat{\delta} = (\hat{\alpha}', \hat{\beta}')'$ and $\delta = (\alpha' H, \beta')'$. Suppose the assumptions of Theorems 1.3.1-1.3.2 and Assumption 1.3.8 hold. For $q, m_N, \omega$, defined in (1.3.2), if $\sqrt{T} m_N^2 \omega_{N,T}^{2-2q} = o(1)$,

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

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

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

**Remark 1.3.3.** 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 $\Sigma_u$ 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 $\sqrt{T}/N \rightarrow 0$ is needed for the asymptotic normality of $\hat{\delta}$ 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

$$\hat{z}_{T+h|T} - z_{T+h|T} = (\hat{\alpha}' \hat{H} \hat{F}_T - \alpha' (H_T - H^{-1} F_T),$$
which contains two components, estimating $\delta$ and $F_t$.

**Theorem 1.3.4.** Let $\hat{z}_{T+h|T} = \hat{\delta} L_T$. Suppose that the assumptions of Theorem 1.3.3 hold. Then, for $\omega_{N,T} = \sqrt{\log N} + \frac{1}{\sqrt{N}}$,

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

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.

### 1.4 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,

$$y_t = A F_t + u_t,$$

and

$$F_t = \sigma_g g(X_t) + \sigma_\gamma \gamma_t,$$

for $t = 1, \cdots, T$,

where $A$ is drawn from i.i.d. Uniform$(0,1)$, and there are three factors ($K = 3$). I set the number of characteristics as $\text{dim}(X_t) = 3$. I introduce serial dependences on $X_t$ and $\gamma_t$ as follows:

$$X_t = \Psi X_{t-1} + \xi_t,$$

and

$$\gamma_t = \Psi \gamma_{t-1} + \nu_t,$$

for $t = 1, \cdots, T$,

with $X_0 = 0$, $\gamma_0 = 0$ and a $3 \times 3$ diagonal matrix $\Psi$. Each diagonal element of $\Psi$ is generated from Uniform$(0.3, 0.7)$. In addition, $\xi_t$ and $\nu_t$ are drawn from i.i.d. $N(0, I)$.

To address different correlations between $F_t$ and $g(X_t)$, 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 $g(\cdot)$ has the following model: $g(X_t) = (g_1(X_t), \cdots, g_K(X_t))^\prime$, where $g_k(X_t) = \sum_{l=1}^3 g_{kl}(X_{tl})$. 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 $g_k(X_t)$ and $\gamma_{k,t}$ 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 $\Sigma_u$ as follows: let $\{\varepsilon_{it}\}_{i \leq N, t \leq T}$ be i.i.d. $N(0, 1)$. Let

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

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

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 $\eta_t = (\eta_{1t}, \cdots, \eta_{N_t})'$ by $R_\eta$, 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 $\Sigma_u = D \Sigma_\eta D$, 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 1.2.3.

### 1.4.1 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 1.2 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 $\left(\frac{1}{NT} \sum_{t,i} (\hat{\lambda}'_i f_t - \lambda'_i f_t)^2\right)^{1/2}$. The results are reported in Table 1.3.

According to Tables 1.2 and 1.3, 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 $X_t$ is not as informative, and hence the performance of PPC and FPPC deteriorates. For common
24

Table 1.2: Canonical correlations of estimated loading or factor matrices: the larger the better.

<table>
<thead>
<tr>
<th></th>
<th>Strong((w = 10))</th>
<th>Mild((w = 1))</th>
<th>Weak((w = 0.1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>T</td>
<td>PC</td>
<td>PPC</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.246</td>
<td>0.715</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.296</td>
<td>0.873</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.170</td>
<td>0.734</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.147</td>
<td>0.952</td>
</tr>
<tr>
<td>300</td>
<td>100</td>
<td>0.439</td>
<td>0.745</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.800</td>
<td>0.941</td>
</tr>
</tbody>
</table>

Loadings

<table>
<thead>
<tr>
<th></th>
<th>Strong((w = 10))</th>
<th>Mild((w = 1))</th>
<th>Weak((w = 0.1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>T</td>
<td>PC</td>
<td>PPC</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.180</td>
<td>0.618</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.185</td>
<td>0.695</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.189</td>
<td>0.750</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.131</td>
<td>0.836</td>
</tr>
<tr>
<td>300</td>
<td>100</td>
<td>0.527</td>
<td>0.883</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.780</td>
<td>0.922</td>
</tr>
</tbody>
</table>

Factors

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.

1.4.2 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:

\[
z_{t+1} = \alpha' F_t + \beta' W_t + \epsilon_{t+1},
\]

where \( W_t = 1 \), \( \beta = 1 \), and \( \epsilon_{t+1} \) is drawn from i.i.d. \( N(0, 1) \). To cover a variety of model settings, unknown coefficients, \( \alpha \), are generated from Uniform\((0, 1)\) for each simulation.

The unknown factor, \( F_t \), can be learned from a factor model: \( y_t = \Lambda F_t + u_t \). Here the same data-generating process is used as Section 1.4.1.

I conducted one-step ahead out-of-sample forecasting 50 times using rolling data windows. The moving window size is fixed as \( T \), and it is also the sample size for estimations. In each simulation, the total \( T + 50 \) observations are generated. To forecast \( z_{T+m+1} \) for \( m = 0, \cdots, 49 \), the observations from \( m + 1 \) to \( m + T \) are used. Specif-
Table 1.3: Mean squared error of estimated common components $\mathbf{A} \mathbf{F}'$: the smaller the better.

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>PC</th>
<th>PPC</th>
<th>FPPC</th>
<th>PC</th>
<th>PPC</th>
<th>FPPC</th>
<th>PC</th>
<th>PPC</th>
<th>FPPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>0.674</td>
<td>0.448</td>
<td><strong>0.281</strong></td>
<td>0.677</td>
<td>0.522</td>
<td><strong>0.363</strong></td>
<td>0.676</td>
<td>0.661</td>
<td><strong>0.561</strong></td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>0.638</td>
<td>0.376</td>
<td><strong>0.213</strong></td>
<td>0.644</td>
<td>0.430</td>
<td><strong>0.271</strong></td>
<td>0.640</td>
<td>0.598</td>
<td><strong>0.479</strong></td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.540</td>
<td>0.366</td>
<td><strong>0.260</strong></td>
<td>0.544</td>
<td>0.435</td>
<td><strong>0.336</strong></td>
<td>0.533</td>
<td>0.584</td>
<td><strong>0.512</strong></td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.457</td>
<td>0.251</td>
<td><strong>0.136</strong></td>
<td>0.450</td>
<td>0.279</td>
<td><strong>0.175</strong></td>
<td>0.396</td>
<td>0.418</td>
<td><strong>0.334</strong></td>
</tr>
<tr>
<td>300</td>
<td>100</td>
<td>0.375</td>
<td>0.298</td>
<td><strong>0.255</strong></td>
<td>0.348</td>
<td>0.360</td>
<td><strong>0.330</strong></td>
<td>0.331</td>
<td>0.496</td>
<td>0.475</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.231</td>
<td>0.179</td>
<td><strong>0.120</strong></td>
<td>0.205</td>
<td>0.210</td>
<td><strong>0.163</strong></td>
<td>0.198</td>
<td>0.341</td>
<td>0.327</td>
</tr>
</tbody>
</table>

Common components

The factors are estimated by PC, PPC, and FPPC methods and are denoted by $\{\hat{F}_{m+1}, \ldots, \hat{F}_{m+T}\}$. Then, $\hat{\alpha}$ and $\hat{\beta}$ are obtained by regressing $\{z_{m+2}, \ldots, z_{m+T}\}$ on $\{(\hat{F}_{m+1}', W_{m+1}), \ldots, (\hat{F}_{m+T-1}', W_{m+T-1})\}$. Finally, forecasts are $\hat{z}_{T+m+1|T+m} = \hat{\alpha}' \hat{F}_{m+T} + \hat{\beta} W_{m+T}$. This procedure continues for $m = 0, \ldots, 49$.

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):

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

where $\hat{z}_{T+m+1|T+m}$ 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 1.4. Overall, FPPC has smaller MSFEs than PPC and PC in this forecasting model. When the correlation between $X_t$ and $F_t$ 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 1.4.1.
1.5 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. (2021) 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

\[
r_{x_{t+1}}^{(n)} = r_{t+1}^{(n)} - y_t^{(1)}, \text{ for } t = 1, \ldots, T,
\]

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

### 1.5.1 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).\(^5\) 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 Oka (2014), among many others.

Finally, the observed characteristics \( X_t \) 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.

---

\(^5\) 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).
Table 1.5: Components of $X_t$ for U.S. bonds excess return forecasting.

<table>
<thead>
<tr>
<th>Series</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{1,t}$</td>
<td>Linear combination of five forward rates (CP)</td>
</tr>
<tr>
<td>$X_{2,t}$</td>
<td>Real gross domestic product (GDP)</td>
</tr>
<tr>
<td>$X_{3,t}$</td>
<td>Consumption price index (CPI) - Inflation</td>
</tr>
<tr>
<td>$X_{4,t}$</td>
<td>Non-agriculture employment</td>
</tr>
</tbody>
</table>

as studied by Stock and Watson (2014) and NBER (2008). A detailed description of these series is listed in Table 1.5. In addition, these data are also transformed and standardized.⁶

1.5.2 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 1.6 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 A.4.

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:

---

⁶ Note that I interpolate gross domestic product, which is reported quarterly, to a monthly frequency following Chow and Lin (1971).
Table 1.6: List of all forecasting models

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(SIC)</td>
<td>Autoregressive model with lags selected by the SIC</td>
</tr>
<tr>
<td>PCR</td>
<td>Principal components regression</td>
</tr>
<tr>
<td>FAAR</td>
<td>Factor augmented autoregressive model</td>
</tr>
<tr>
<td>DI</td>
<td>Diffusion index regression model with CP and factors</td>
</tr>
<tr>
<td>DI2</td>
<td>Diffusion index regression model with CP, lags, and factors</td>
</tr>
<tr>
<td>Bagging</td>
<td>Bagging with shrinkage, $c = 1.96$</td>
</tr>
<tr>
<td>Boosting</td>
<td>Component boosting, $M = 50$</td>
</tr>
<tr>
<td>Fac-Lasso</td>
<td>Factor-Lasso regression</td>
</tr>
<tr>
<td>Lasso</td>
<td>Lasso regression</td>
</tr>
<tr>
<td>Ridge</td>
<td>Ridge regression</td>
</tr>
<tr>
<td>EN</td>
<td>Elastic net regression</td>
</tr>
<tr>
<td>DT</td>
<td>Decision tree regression</td>
</tr>
<tr>
<td>G-Bst</td>
<td>Gradient boosting regression</td>
</tr>
<tr>
<td>RanForest</td>
<td>Random forest regression</td>
</tr>
<tr>
<td>NN1</td>
<td>Neural network with one hidden layer</td>
</tr>
<tr>
<td>NN2</td>
<td>Neural network with two hidden layers</td>
</tr>
<tr>
<td>NN3</td>
<td>Neural network with three hidden layers</td>
</tr>
<tr>
<td>H-NN1</td>
<td>Hybrid neural network with one hidden layer</td>
</tr>
<tr>
<td>H-NN2</td>
<td>Hybrid neural network with two hidden layers</td>
</tr>
<tr>
<td>H-NN3</td>
<td>Hybrid neural network with three hidden layers</td>
</tr>
<tr>
<td>FANN1</td>
<td>Factor augmented neural network with three hidden units</td>
</tr>
<tr>
<td>FANN2</td>
<td>Factor augmented neural network with five hidden units</td>
</tr>
<tr>
<td>FANN3</td>
<td>Factor augmented neural network with seven hidden units</td>
</tr>
</tbody>
</table>
1. mean square forecast error (MSFE), defined as

\[
\text{MSFE} = \frac{1}{P} \sum_{t=Q}^{T-1} (r_{x_{t+1}}^{(n)} - \hat{r}_{x_{t+1}}^{(n)})^2,
\]

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

\[
\text{Out-of-sample } R^2 = 1 - \frac{\sum_{t=Q}^{T-1} (r_{x_{t+1}}^{(n)} - \hat{r}_{x_{t+1}}^{(n)})^2}{\sum_{t=Q}^{T-1} (r_{x_{t+1}}^{(n)} - \bar{r}_{x_{t+1}}^{(n)})^2},
\]

where, for each maturity \( n \), \( \hat{r}_{x_{t+1}}^{(n)} \) is the forecast of bond excess returns using each model and \( \bar{r}_{x_{t+1}}^{(n)} \) is the historical average of bond excess return.\(^7\)

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: \( S_{DM} = \frac{\bar{d}}{\hat{\sigma}_d} \), where \( \bar{d} = \frac{1}{P} \sum_{t=1}^{P} (\hat{\sigma}_{1,t+1|t} - \hat{\sigma}_{2,t+1|t}) \), \( \hat{\sigma}_d \) is a heteroskedasticity and autocorrelation robust estimator of the standard deviation of \( \bar{d} \). Here \( \hat{\epsilon}_{1,t+1|t} \) and \( \hat{\epsilon}_{2,t+1|t} \) 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

\[\text{The historical average of bond excess return starts from 1964:1, and values are the same for both rolling and recursive window methods.}\]
Table 1.7: In-sample adjusted $R^2$ of predictive regressions for excess bond returns: the larger the better.

<table>
<thead>
<tr>
<th>Maturity</th>
<th>PC$_{130}$</th>
<th>PC$_{134}$</th>
<th>PPC</th>
<th>FPPC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample period: 1964:1-2016:4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 year</td>
<td>15.5</td>
<td>16.3</td>
<td>17.4</td>
<td>18.0</td>
</tr>
<tr>
<td>3 year</td>
<td>15.9</td>
<td>16.7</td>
<td>16.8</td>
<td>17.3</td>
</tr>
<tr>
<td>4 year</td>
<td>16.1</td>
<td>16.8</td>
<td>17.0</td>
<td>17.6</td>
</tr>
<tr>
<td>5 year</td>
<td>17.3</td>
<td>17.8</td>
<td>17.7</td>
<td>18.4</td>
</tr>
</tbody>
</table>

| 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 |

comparisons of all models in Section 1.6.2.

1.6 Empirical Findings

1.6.1 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:

$$ r_{x_{t+1}}^{(n)} = \alpha + \beta' \widehat{F}_t + \epsilon_{t+1}. \quad (1.6.1) $$

I investigate the unconditional predictive power of macro factors for future returns using different principal component methods. Note that the equation (1.6.1) is a restricted version 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 1.7 reports the adjusted $R^2$ statistics from in-sample regressions of the equation (1.6.1), for 2- to 5-year log excess bond returns. PC$_{130}$, which is a benchmark, denotes
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).

the conventional PC method with 130 macro variables and PC$_{134}$ with four additional characteristics $\mathbf{X}_t$ introduced in Table 1.5. 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 1.2.3. Two sample periods are considered: 1964:1-2016:4 and 1964:1-2003:12. From the table, I find that the estimated factors using FPPC outperform the factors using other methods. In the first panel of Table 1.7, 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 1.2 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, $\hat{F}_{1t}$ and $\hat{F}_{2t}$, are very similar to 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, $\hat{F}_{3t}$, is correlated with housing. The fourth factor, $\hat{F}_{4t}$, 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, $\hat{F}_{5t}$, is correlated with industrial production and housing, while the sixth factor, $\hat{F}_{6t}$, 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, $\hat{F}_{8t}$ is highly correlated with consumption orders and inventories. I interpret $\hat{F}_{3t}$ as a housing factor, $\hat{F}_{4t}$ as a stock market factor, both $\hat{F}_{4t}$ and $\hat{F}_{7t}$ as inflation factors, and both $\hat{F}_{5t}$ and $\hat{F}_{8t}$ as real factors.

### 1.6.2 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
Table 1.8: Relative mean squared forecast errors of U.S. bonds excess return forecasting: the smaller the better. Forecasting sample period: 1984:1-2016:4.

<table>
<thead>
<tr>
<th>Maturity</th>
<th>PCR ($L_t = F_t$)</th>
<th>DI ($L_t = (F'_t, CP_t)'$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC$<em>{130}$ PC$</em>{134}$ PPC FPPC</td>
<td>PC$<em>{130}$ PC$</em>{134}$ PPC FPPC</td>
</tr>
<tr>
<td></td>
<td>Rolling estimation window</td>
<td>Recursive estimation window</td>
</tr>
<tr>
<td>2 year</td>
<td>2.421 0.978 0.849 0.789</td>
<td>1.542 1.004 0.961 0.964</td>
</tr>
<tr>
<td>3 year</td>
<td>8.310 0.981 0.882 0.825</td>
<td>5.700 1.000 0.945 0.942</td>
</tr>
<tr>
<td>4 year</td>
<td>17.011 0.974 0.871 0.808</td>
<td>11.589 0.996 0.923 0.909</td>
</tr>
<tr>
<td>5 year</td>
<td>25.668 0.976 0.879 0.819</td>
<td>18.417 0.997 0.939 0.923</td>
</tr>
</tbody>
</table>

models are compared. Then, I explore the forecasting performance of all models outlined in Table 1.6. In addition, the replication of experiments using different forecasting periods was considered.

**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 1.8 and 1.9 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 $\text{MSFE}(M)/\text{MSFE}(\text{PC}_{130})$ 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.

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
Table 1.9: Out-of-sample $R^2$ (%) of U.S. bonds excess return forecasting: the larger the better. Forecasting sample period: 1984:1-2016:4.

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

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 1.10 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 1.11. 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 1.8 and 1.9. 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, $X_t$, has explanatory power on the latent factors. In addition, the results can be interpreted by the following...
Table 1.10: Out-of-sample $R^2(\%)$ of U.S. bonds excess return forecasting in recessions and expansions: the larger the better. Rolling window.

<table>
<thead>
<tr>
<th>Maturity</th>
<th>PC$_{130}$</th>
<th>PC$_{134}$</th>
<th>PPC</th>
<th>FPPC</th>
<th>PC$_{130}$</th>
<th>PC$_{134}$</th>
<th>PPC</th>
<th>FPPC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recessions</td>
<td>Expansions</td>
<td></td>
<td></td>
<td>Recessions</td>
<td>Expansions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 year</td>
<td>6.6</td>
<td>11.8</td>
<td>11.7</td>
<td>41.1</td>
<td>40.7</td>
<td>43.2</td>
<td>42.9</td>
<td></td>
</tr>
<tr>
<td>3 year</td>
<td>-4.3</td>
<td>8.2</td>
<td>11.6</td>
<td>40.4</td>
<td>40.2</td>
<td>42.8</td>
<td>42.8</td>
<td></td>
</tr>
<tr>
<td>4 year</td>
<td>-10.1</td>
<td>6.0</td>
<td>11.8</td>
<td>41.8</td>
<td>41.8</td>
<td>45.4</td>
<td>46.0</td>
<td></td>
</tr>
<tr>
<td>5 year</td>
<td>-9.8</td>
<td>5.7</td>
<td>14.2</td>
<td>39.7</td>
<td>39.7</td>
<td>42.5</td>
<td>43.1</td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Model</th>
<th>Maturity</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>5 year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCR ($L_t = F_t$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FPPC versus PC$_{130}$</td>
<td>-2.542**</td>
<td>-2.295**</td>
<td>-2.533**</td>
<td>-2.597***</td>
<td></td>
</tr>
<tr>
<td>FPPC versus PC$_{134}$</td>
<td>-2.521**</td>
<td>-2.254**</td>
<td>-2.391**</td>
<td>-2.511**</td>
<td></td>
</tr>
<tr>
<td>FPPC versus PPC</td>
<td>-2.442**</td>
<td>-2.405**</td>
<td>-2.640***</td>
<td>-2.581***</td>
<td></td>
</tr>
<tr>
<td>DI ($L_t = (F'_t, CP_t)'$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FPPC versus PC$_{130}$</td>
<td>-0.604</td>
<td>-1.020</td>
<td>-1.619</td>
<td>-1.396</td>
<td></td>
</tr>
<tr>
<td>FPPC versus PC$_{134}$</td>
<td>-0.698</td>
<td>-1.061</td>
<td>-1.620</td>
<td>-1.404</td>
<td></td>
</tr>
<tr>
<td>FPPC versus PPC</td>
<td>0.130</td>
<td>-0.157</td>
<td>-0.610</td>
<td>-0.698</td>
<td></td>
</tr>
</tbody>
</table>

|                |          |        |        |        |        |
| 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.
outlines: (i) 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.

**Forecasting performance**

Here, I investigate the one-year-ahead forecasting performance of linear and nonlinear machine learning models listed in Table 1.6. 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 1.12-1.17. Table 1.12 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 1.13. Tables 1.14-1.17 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 1.12-1.13 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 1.14-1.17 confirm that DI models based on the pro-
posed 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 1.6.2.

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 1.12 and 1.13. However, in the rolling window scheme, DI models perform better than the RanForest model for some maturities. For example, an inspection of Tables 1.12 and 1.13 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 1.14-1.17 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 1.12-1.17. 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.
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 1.18. 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 1.18. 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.

1.7 Conclusion

This paper examines a high-dimensional factor model that latent factors depend on a few observed covariate variables. This model is motivated by the fact that observed variables can partially explain the latent factors. The FPPC method estimates the unknown
factors and loadings efficiently by taking into account cross-sectional heteroskedasticity and correlations. 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>
<thead>
<tr>
<th>Method</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>5 year</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>5 year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR(SIC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR</td>
<td>0.898</td>
<td>0.858∗</td>
<td>0.824∗</td>
<td>0.800∗</td>
<td>1.042</td>
<td>0.978</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>FAAR</td>
<td>0.920</td>
<td>0.859</td>
<td>0.809∗</td>
<td>0.777∗</td>
<td>1.044</td>
<td>0.948</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>0.699∗</td>
<td>0.673∗</td>
<td>0.633∗</td>
<td>0.649∗</td>
<td>0.723∗</td>
<td>0.683∗</td>
<td>0.640∗</td>
</tr>
<tr>
<td></td>
<td>DI2</td>
<td>0.696∗</td>
<td>0.642∗</td>
<td>0.601∗</td>
<td>0.622∗</td>
<td>0.734∗</td>
<td>0.676∗</td>
<td>0.631∗</td>
</tr>
<tr>
<td></td>
<td>Bagging</td>
<td>0.685∗</td>
<td>0.649∗</td>
<td>0.609∗</td>
<td>0.658∗</td>
<td>0.752∗</td>
<td>0.704∗</td>
<td>0.656∗</td>
</tr>
<tr>
<td></td>
<td>Boosting</td>
<td>0.653∗</td>
<td>0.640∗</td>
<td>0.610∗</td>
<td>0.655∗</td>
<td>0.730∗</td>
<td>0.708∗</td>
<td>0.676∗</td>
</tr>
<tr>
<td></td>
<td>Fac-Lasso</td>
<td>0.660∗</td>
<td>0.635∗</td>
<td>0.598∗</td>
<td>0.618∗</td>
<td>0.739∗</td>
<td>0.687∗</td>
<td>0.636∗</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.986</td>
<td>0.894∗</td>
<td>0.851∗</td>
<td>0.830∗</td>
<td>0.982</td>
<td>0.881∗</td>
<td>0.834∗</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>0.975</td>
<td>0.955</td>
<td>0.883</td>
<td>0.884</td>
<td>1.136</td>
<td>1.054</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>EN</td>
<td>0.991</td>
<td>0.922∗</td>
<td>0.871∗</td>
<td>0.847∗</td>
<td>1.001</td>
<td>0.932∗</td>
<td>0.873∗</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.945</td>
<td>1.059</td>
<td>0.969</td>
<td>1.057</td>
<td>1.141</td>
<td>1.075</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>G-Bst</td>
<td>0.894∗</td>
<td>0.890∗</td>
<td>0.884∗</td>
<td>0.870∗</td>
<td>0.922∗</td>
<td>0.889∗</td>
<td>0.874∗</td>
</tr>
<tr>
<td></td>
<td>RanForest</td>
<td>0.647∗</td>
<td>0.600∗</td>
<td>0.663∗</td>
<td>0.644∗</td>
<td>0.652∗</td>
<td>0.647∗</td>
<td>0.629∗</td>
</tr>
<tr>
<td></td>
<td>NN1</td>
<td>0.942</td>
<td>0.905∗</td>
<td>0.853∗</td>
<td>0.846∗</td>
<td>0.897∗</td>
<td>0.860∗</td>
<td>0.822∗</td>
</tr>
<tr>
<td></td>
<td>NN2</td>
<td>1.092</td>
<td>1.017</td>
<td>1.021</td>
<td>1.019</td>
<td>1.131</td>
<td>1.071</td>
<td>1.057</td>
</tr>
<tr>
<td></td>
<td>NN3</td>
<td>1.084</td>
<td>1.020</td>
<td>1.023</td>
<td>1.000</td>
<td>1.118</td>
<td>1.084</td>
<td>1.063</td>
</tr>
<tr>
<td></td>
<td>H-NN1</td>
<td>0.749∗</td>
<td>0.706∗</td>
<td>0.696∗</td>
<td>0.749∗</td>
<td>0.804∗</td>
<td>0.799∗</td>
<td>0.772∗</td>
</tr>
<tr>
<td></td>
<td>H-NN2</td>
<td>0.810∗</td>
<td>0.792∗</td>
<td>0.759∗</td>
<td>0.766∗</td>
<td>1.003</td>
<td>0.989</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>H-NN3</td>
<td>0.797∗</td>
<td>0.795∗</td>
<td>0.774∗</td>
<td>0.765∗</td>
<td>0.987</td>
<td>0.989</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>FANN</td>
<td>0.691∗</td>
<td>0.673∗</td>
<td>0.646∗</td>
<td>0.677∗</td>
<td>0.767∗</td>
<td>0.728∗</td>
<td>0.691∗</td>
</tr>
<tr>
<td></td>
<td>FANN2</td>
<td>0.690∗</td>
<td>0.658∗</td>
<td>0.643∗</td>
<td>0.653∗</td>
<td>0.767∗</td>
<td>0.728∗</td>
<td>0.682∗</td>
</tr>
<tr>
<td></td>
<td>FANN3</td>
<td>0.699∗</td>
<td>0.670∗</td>
<td>0.650∗</td>
<td>0.650∗</td>
<td>0.773∗</td>
<td>0.730∗</td>
<td>0.681∗</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Method</th>
<th>Rolling 2 year</th>
<th>Rolling 3 year</th>
<th>Rolling 4 year</th>
<th>Rolling 5 year</th>
<th>Recursive 2 year</th>
<th>Recursive 3 year</th>
<th>Recursive 4 year</th>
<th>Recursive 5 year</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(SIC)</td>
<td>13.1</td>
<td>9.4</td>
<td>9.1</td>
<td>8.0</td>
<td>8.9</td>
<td>4.6</td>
<td>2.9</td>
<td>1.3</td>
</tr>
<tr>
<td>PCR</td>
<td>22.0</td>
<td>22.3</td>
<td>25.1</td>
<td>26.3</td>
<td>5.0</td>
<td>6.7</td>
<td>9.3</td>
<td>11.5</td>
</tr>
<tr>
<td>FAAR</td>
<td>20.0</td>
<td>22.2</td>
<td>26.5</td>
<td>28.5</td>
<td>4.9</td>
<td>9.5</td>
<td>13.9</td>
<td>17.0</td>
</tr>
<tr>
<td>DI</td>
<td>39.2</td>
<td>39.0</td>
<td>42.5</td>
<td>40.3</td>
<td>34.1</td>
<td>34.9</td>
<td>37.8</td>
<td>36.0</td>
</tr>
<tr>
<td>DI2</td>
<td>39.5</td>
<td>41.8</td>
<td>45.4</td>
<td>42.7</td>
<td>33.1</td>
<td>35.5</td>
<td>38.7</td>
<td>37.0</td>
</tr>
<tr>
<td>Bagging</td>
<td>40.5</td>
<td>41.2</td>
<td>44.7</td>
<td>39.4</td>
<td>31.5</td>
<td>32.8</td>
<td>36.3</td>
<td>33.9</td>
</tr>
<tr>
<td>Boosting</td>
<td>43.2</td>
<td>42.0</td>
<td>44.5</td>
<td>39.7</td>
<td>33.5</td>
<td>32.5</td>
<td>34.3</td>
<td>31.5</td>
</tr>
<tr>
<td>Fac-Lasso</td>
<td>42.6</td>
<td>42.5</td>
<td>45.6</td>
<td>43.1</td>
<td>32.7</td>
<td>34.5</td>
<td>38.2</td>
<td>37.0</td>
</tr>
<tr>
<td>Lasso</td>
<td>14.3</td>
<td>19.0</td>
<td>22.7</td>
<td>23.6</td>
<td>10.5</td>
<td>16.0</td>
<td>19.0</td>
<td>20.0</td>
</tr>
<tr>
<td>Ridge</td>
<td>15.3</td>
<td>13.5</td>
<td>19.8</td>
<td>18.6</td>
<td>-3.5</td>
<td>-0.6</td>
<td>7.7</td>
<td>6.7</td>
</tr>
<tr>
<td>EN</td>
<td>13.9</td>
<td>16.4</td>
<td>20.8</td>
<td>22.0</td>
<td>8.8</td>
<td>11.1</td>
<td>15.2</td>
<td>16.8</td>
</tr>
<tr>
<td>DT</td>
<td>17.9</td>
<td>4.0</td>
<td>11.9</td>
<td>2.7</td>
<td>-4.0</td>
<td>-2.6</td>
<td>16.7</td>
<td>13.0</td>
</tr>
<tr>
<td>G-Bst</td>
<td>22.3</td>
<td>19.3</td>
<td>19.7</td>
<td>19.9</td>
<td>16.0</td>
<td>15.2</td>
<td>15.1</td>
<td>16.5</td>
</tr>
<tr>
<td>RanForest</td>
<td>43.8</td>
<td>40.2</td>
<td>39.7</td>
<td>40.7</td>
<td>40.6</td>
<td>38.3</td>
<td>38.9</td>
<td>39.2</td>
</tr>
<tr>
<td>NN1</td>
<td>18.2</td>
<td>18.0</td>
<td>22.5</td>
<td>22.1</td>
<td>18.2</td>
<td>17.9</td>
<td>20.2</td>
<td>21.9</td>
</tr>
<tr>
<td>NN2</td>
<td>5.1</td>
<td>7.8</td>
<td>7.2</td>
<td>6.2</td>
<td>-3.0</td>
<td>-2.2</td>
<td>-2.7</td>
<td>-1.3</td>
</tr>
<tr>
<td>NN3</td>
<td>5.8</td>
<td>7.6</td>
<td>7.0</td>
<td>7.9</td>
<td>-1.9</td>
<td>-3.4</td>
<td>-3.3</td>
<td>-2.6</td>
</tr>
<tr>
<td>H-NN1</td>
<td>34.9</td>
<td>36.0</td>
<td>36.8</td>
<td>31.1</td>
<td>26.7</td>
<td>23.7</td>
<td>25.0</td>
<td>23.4</td>
</tr>
<tr>
<td>H-NN2</td>
<td>29.6</td>
<td>28.2</td>
<td>31.1</td>
<td>29.5</td>
<td>8.6</td>
<td>5.7</td>
<td>5.9</td>
<td>3.6</td>
</tr>
<tr>
<td>H-NN3</td>
<td>30.8</td>
<td>27.9</td>
<td>29.7</td>
<td>29.6</td>
<td>10.1</td>
<td>5.6</td>
<td>5.5</td>
<td>3.5</td>
</tr>
<tr>
<td>FANN</td>
<td>40.0</td>
<td>39.0</td>
<td>41.2</td>
<td>37.7</td>
<td>30.1</td>
<td>30.6</td>
<td>32.9</td>
<td>29.7</td>
</tr>
<tr>
<td>FANN2</td>
<td>40.0</td>
<td>40.4</td>
<td>41.5</td>
<td>39.9</td>
<td>30.1</td>
<td>30.6</td>
<td>33.8</td>
<td>29.4</td>
</tr>
<tr>
<td>FANN3</td>
<td>39.3</td>
<td>39.3</td>
<td>40.9</td>
<td>40.1</td>
<td>29.5</td>
<td>30.4</td>
<td>33.8</td>
<td>31.1</td>
</tr>
</tbody>
</table>
Table 1.14: Pairwise model comparison using Giacomini-White tests for 2-year maturity.

<table>
<thead>
<tr>
<th></th>
<th>PCR</th>
<th>FAAR</th>
<th>DI</th>
<th>DI2</th>
<th>Bagging</th>
<th>Boosting</th>
<th>Fac-Lasso</th>
<th>Lasso</th>
<th>Ridge</th>
<th>DT</th>
<th>G-Bst</th>
<th>RanForest</th>
<th>NN</th>
<th>H-NN</th>
<th>FANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(SIC)</td>
<td>0.93</td>
<td>0.68</td>
<td>4.43</td>
<td>5.66</td>
<td>5.23</td>
<td>6.62</td>
<td>5.40</td>
<td>-0.25</td>
<td>2.12</td>
<td>1.71</td>
<td>4.80</td>
<td>9.56</td>
<td>1.61</td>
<td>2.90</td>
<td>3.22</td>
</tr>
<tr>
<td>PCR</td>
<td>0.50</td>
<td>5.69</td>
<td>9.77</td>
<td>7.44</td>
<td>4.24</td>
<td>5.15</td>
<td>-1.76</td>
<td>-0.74</td>
<td>-1.50</td>
<td>0.22</td>
<td>-0.94</td>
<td>6.14</td>
<td>-1.79</td>
<td>3.02</td>
<td>3.06</td>
</tr>
<tr>
<td>FAAR</td>
<td>3.94</td>
<td>7.45</td>
<td>5.03</td>
<td>4.40</td>
<td>4.30</td>
<td>-0.95</td>
<td>-1.04</td>
<td>-0.88</td>
<td>0.22</td>
<td>-0.51</td>
<td>5.00</td>
<td>-1.80</td>
<td>2.59</td>
<td>2.62</td>
<td></td>
</tr>
<tr>
<td>DI</td>
<td>2.84</td>
<td>6.33</td>
<td>0.75</td>
<td>0.62</td>
<td>-4.88</td>
<td>-13.00</td>
<td>-4.47</td>
<td>-2.96</td>
<td>-2.71</td>
<td>0.81</td>
<td>-3.51</td>
<td>-0.78</td>
<td>-0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DI2</td>
<td>-0.12</td>
<td>1.66</td>
<td>0.14</td>
<td>-5.81</td>
<td>-13.14</td>
<td>-5.39</td>
<td>-2.69</td>
<td>-3.51</td>
<td>0.31</td>
<td>-5.04</td>
<td>-1.27</td>
<td>-1.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td>1.36</td>
<td>0.27</td>
<td></td>
<td>-5.64</td>
<td>-17.72</td>
<td>-5.16</td>
<td>-2.52</td>
<td>-3.46</td>
<td>0.53</td>
<td>-4.53</td>
<td>-1.34</td>
<td>-0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boosting</td>
<td>-0.78</td>
<td></td>
<td></td>
<td>-7.06</td>
<td>-14.07</td>
<td>-6.49</td>
<td>-5.51</td>
<td>-4.59</td>
<td>0.25</td>
<td>-4.89</td>
<td>-7.79</td>
<td>-6.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fac-Lasso</td>
<td>-6.00</td>
<td></td>
<td></td>
<td>-15.80</td>
<td>-5.54</td>
<td>-5.91</td>
<td>-3.96</td>
<td>0.52</td>
<td>-4.69</td>
<td>-1.56</td>
<td>-0.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-Bst</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RanForest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-NN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FANN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

See notes to Table 1.14. For further details, refer to Section 1.5.2.
Table 1.16: Pairwise model comparison using Giacomini-White tests for 4-year maturity.

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

See notes to Table 1.14. For further details, refer to Section 1.5.2.
Table 1.17: Pairwise model comparison using Giacomini-White tests for 5-year maturity.

<table>
<thead>
<tr>
<th></th>
<th>PCR</th>
<th>FAAR</th>
<th>DJ</th>
<th>DJ2</th>
<th>Bagging</th>
<th>Boosting</th>
<th>Fac-Lasso</th>
<th>Lasso</th>
<th>Ridge</th>
<th>EN</th>
<th>DT</th>
<th>G-Bst</th>
<th>RanForest</th>
<th>NN</th>
<th>H-NN</th>
<th>FANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(SIC)</td>
<td>7.19</td>
<td>7.18</td>
<td>7.85</td>
<td>9.79</td>
<td>8.70</td>
<td>5.89</td>
<td>8.90</td>
<td>0.72</td>
<td>15.22</td>
<td>0.48</td>
<td>5.44</td>
<td>9.41</td>
<td>14.26</td>
<td>1.80</td>
<td>4.69</td>
<td></td>
</tr>
<tr>
<td>PCR</td>
<td>0.38</td>
<td>2.88</td>
<td>3.94</td>
<td>2.05</td>
<td>0.95</td>
<td>2.45</td>
<td>-4.41</td>
<td>-0.49</td>
<td>-2.98</td>
<td>-1.31</td>
<td>-2.87</td>
<td>2.43</td>
<td>-0.61</td>
<td>-0.98</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>FAAR</td>
<td>0.98</td>
<td>2.85</td>
<td>1.12</td>
<td>0.27</td>
<td>1.06</td>
<td>-4.43</td>
<td>-1.10</td>
<td>-3.39</td>
<td>-1.81</td>
<td>-2.46</td>
<td>1.31</td>
<td>0.66</td>
<td>-0.62</td>
<td>0.22</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>DJ</td>
<td>0.22</td>
<td>-4.55</td>
<td>-2.21</td>
<td>0.29</td>
<td>-4.66</td>
<td>-10.87</td>
<td>-4.49</td>
<td>-8.76</td>
<td>-4.21</td>
<td>0.04</td>
<td>-2.22</td>
<td>-3.93</td>
<td>-1.66</td>
<td>-5.30</td>
<td>-2.42</td>
<td></td>
</tr>
<tr>
<td>DJ2</td>
<td>-2.57</td>
<td>-2.35</td>
<td>-0.07</td>
<td>-5.26</td>
<td>-12.33</td>
<td>-5.25</td>
<td>-10.81</td>
<td>-4.84</td>
<td>-0.12</td>
<td>-3.18</td>
<td>-5.30</td>
<td>-2.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td>-2.08</td>
<td>0.83</td>
<td>-4.37</td>
<td>-6.93</td>
<td>-4.23</td>
<td>-7.42</td>
<td>-3.97</td>
<td>0.23</td>
<td>-2.23</td>
<td>-4.05</td>
<td>-6.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boosting</td>
<td>1.18</td>
<td>-1.90</td>
<td>-5.34</td>
<td>-2.10</td>
<td>-5.41</td>
<td>-10.40</td>
<td>-4.27</td>
<td>-2.52</td>
<td>-2.47</td>
<td>-3.95</td>
<td>-2.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fac-Lasso</td>
<td>-4.88</td>
<td>-12.31</td>
<td>-4.64</td>
<td>-10.40</td>
<td>-4.27</td>
<td>-2.52</td>
<td>-2.47</td>
<td>-3.95</td>
<td>-2.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>-0.07</td>
<td>-1.67</td>
<td>-0.27</td>
<td>-0.79</td>
<td>5.90</td>
<td>1.74</td>
<td>1.01</td>
<td>1.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge</td>
<td>0.04</td>
<td>-0.15</td>
<td>0.01</td>
<td>3.60</td>
<td>0.32</td>
<td>2.47</td>
<td>7.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN</td>
<td>-0.14</td>
<td>-0.65</td>
<td>5.88</td>
<td>2.50</td>
<td>0.66</td>
<td>1.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>0.04</td>
<td>8.97</td>
<td>0.69</td>
<td>0.89</td>
<td>5.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-Bst</td>
<td>8.05</td>
<td>4.09</td>
<td>1.22</td>
<td>1.93</td>
<td>-3.57</td>
<td>-2.73</td>
<td>-0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RanForest</td>
<td>-1.58</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>2.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-NN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See notes to Table 1.14. For further details, refer to Section 1.5.2.
Table 1.18: Relative mean squared forecast errors of U.S. bonds excess return forecasting for different prediction periods. Rolling window.

<table>
<thead>
<tr>
<th>Method</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>5 year</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>5 year</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>5 year</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>0.835*</td>
<td>0.790**</td>
<td>0.766***</td>
<td>0.757***</td>
<td>0.965</td>
<td>0.956</td>
<td>0.945</td>
<td>0.946</td>
<td>1.058</td>
<td>0.951</td>
<td>0.848</td>
<td>0.719</td>
</tr>
<tr>
<td>FAAR</td>
<td>0.851</td>
<td>0.787**</td>
<td>0.746***</td>
<td>0.731***</td>
<td>0.863</td>
<td>0.764</td>
<td>0.715*</td>
<td>0.709*</td>
<td>1.404</td>
<td>1.441</td>
<td>1.281</td>
<td>1.063</td>
</tr>
<tr>
<td>DI</td>
<td>0.578***</td>
<td>0.525***</td>
<td>0.488***</td>
<td>0.513***</td>
<td>0.882</td>
<td>0.880</td>
<td>0.870</td>
<td>0.891</td>
<td>0.879</td>
<td>0.897</td>
<td>0.807</td>
<td>0.747</td>
</tr>
<tr>
<td>DI2</td>
<td>0.597***</td>
<td>0.532***</td>
<td>0.490***</td>
<td>0.517***</td>
<td>0.753</td>
<td>0.680*</td>
<td>0.653**</td>
<td>0.679*</td>
<td>1.055</td>
<td>1.096</td>
<td>0.999</td>
<td>0.920</td>
</tr>
<tr>
<td>Bagging</td>
<td>0.570***</td>
<td>0.531***</td>
<td>0.499***</td>
<td>0.549***</td>
<td>0.802</td>
<td>0.727</td>
<td>0.651**</td>
<td>0.734</td>
<td>0.985</td>
<td>1.043</td>
<td>1.019</td>
<td>0.940</td>
</tr>
<tr>
<td>Boosting</td>
<td>0.529***</td>
<td>0.513***</td>
<td>0.496***</td>
<td>0.525***</td>
<td>0.769</td>
<td>0.735*</td>
<td>0.677**</td>
<td>0.741*</td>
<td>1.005</td>
<td>1.036</td>
<td>0.995</td>
<td>0.996</td>
</tr>
<tr>
<td>Fac-Lasso</td>
<td>0.575***</td>
<td>0.524***</td>
<td>0.483***</td>
<td>0.514***</td>
<td>0.767</td>
<td>0.766</td>
<td>0.773</td>
<td>0.776</td>
<td>0.834</td>
<td>0.855</td>
<td>0.779</td>
<td>0.739</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.942</td>
<td>0.843***</td>
<td>0.810***</td>
<td>0.796***</td>
<td>1.046</td>
<td>1.007</td>
<td>0.962</td>
<td>0.933</td>
<td>1.064</td>
<td>0.869**</td>
<td>0.817***</td>
<td>0.785***</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.746</td>
<td>0.721</td>
<td>0.646*</td>
<td>0.672*</td>
<td>1.076</td>
<td>1.079</td>
<td>1.066</td>
<td>1.129</td>
<td>1.889</td>
<td>1.799</td>
<td>1.532</td>
<td>1.263</td>
</tr>
<tr>
<td>EN</td>
<td>0.973</td>
<td>0.907***</td>
<td>0.860***</td>
<td>0.835***</td>
<td>0.994</td>
<td>0.959</td>
<td>0.919</td>
<td>0.904</td>
<td>1.071</td>
<td>0.908*</td>
<td>0.828***</td>
<td>0.797***</td>
</tr>
<tr>
<td>DT</td>
<td>0.874</td>
<td>0.854</td>
<td>0.865</td>
<td>1.022</td>
<td>0.752</td>
<td>1.260</td>
<td>1.027</td>
<td>1.063</td>
<td>1.761</td>
<td>1.580</td>
<td>1.328</td>
<td>1.175</td>
</tr>
<tr>
<td>G-Bst</td>
<td>0.960</td>
<td>0.952</td>
<td>0.930</td>
<td>0.920</td>
<td>0.798*</td>
<td>0.827*</td>
<td>0.864</td>
<td>0.881</td>
<td>0.784**</td>
<td>0.740**</td>
<td>0.715***</td>
<td>0.665***</td>
</tr>
<tr>
<td>RanForest</td>
<td>0.688**</td>
<td>0.681**</td>
<td>0.693**</td>
<td>0.678**</td>
<td>0.612**</td>
<td>0.672*</td>
<td>0.655**</td>
<td>0.652**</td>
<td>0.522***</td>
<td>0.529***</td>
<td>0.544***</td>
<td>0.506***</td>
</tr>
<tr>
<td>NN1</td>
<td>1.002</td>
<td>0.915</td>
<td>0.872*</td>
<td>0.887</td>
<td>0.805**</td>
<td>0.914</td>
<td>0.824</td>
<td>0.833</td>
<td>0.963</td>
<td>0.835</td>
<td>0.822</td>
<td>0.714**</td>
</tr>
<tr>
<td>NN2</td>
<td>1.158</td>
<td>1.072</td>
<td>1.065</td>
<td>1.051</td>
<td>0.954</td>
<td>0.939</td>
<td>0.983</td>
<td>1.016</td>
<td>1.085</td>
<td>0.938</td>
<td>0.899*</td>
<td>0.902*</td>
</tr>
<tr>
<td>NN3</td>
<td>1.137</td>
<td>1.080</td>
<td>1.061</td>
<td>1.037</td>
<td>0.960</td>
<td>0.932</td>
<td>0.958</td>
<td>0.957</td>
<td>1.108</td>
<td>0.937</td>
<td>0.981</td>
<td>0.935</td>
</tr>
<tr>
<td>H-NN1</td>
<td>0.605***</td>
<td>0.549***</td>
<td>0.529***</td>
<td>0.552***</td>
<td>0.883</td>
<td>0.832</td>
<td>0.828</td>
<td>0.868</td>
<td>1.158</td>
<td>1.175</td>
<td>1.224</td>
<td>1.281</td>
</tr>
<tr>
<td>H-NN2</td>
<td>0.630***</td>
<td>0.578***</td>
<td>0.544***</td>
<td>0.552***</td>
<td>0.896</td>
<td>0.898</td>
<td>0.867</td>
<td>0.844</td>
<td>1.515</td>
<td>1.589</td>
<td>1.516</td>
<td>1.434</td>
</tr>
<tr>
<td>H-NN3</td>
<td>0.615***</td>
<td>0.597***</td>
<td>0.551***</td>
<td>0.542***</td>
<td>0.873</td>
<td>0.881</td>
<td>0.889</td>
<td>0.854</td>
<td>1.531</td>
<td>1.564</td>
<td>1.560</td>
<td>1.447</td>
</tr>
<tr>
<td>FANN</td>
<td>0.503***</td>
<td>0.522***</td>
<td>0.500***</td>
<td>0.533***</td>
<td>0.881</td>
<td>0.860</td>
<td>0.827</td>
<td>0.842</td>
<td>1.186</td>
<td>0.958</td>
<td>0.953</td>
<td>0.935</td>
</tr>
<tr>
<td>FANN2</td>
<td>0.520***</td>
<td>0.510***</td>
<td>0.492***</td>
<td>0.501***</td>
<td>0.869</td>
<td>0.867</td>
<td>0.847</td>
<td>0.86</td>
<td>1.122</td>
<td>0.876</td>
<td>0.926</td>
<td>0.874</td>
</tr>
<tr>
<td>FANN3</td>
<td>0.523***</td>
<td>0.520***</td>
<td>0.497***</td>
<td>0.497***</td>
<td>0.870</td>
<td>0.878</td>
<td>0.879</td>
<td>0.871</td>
<td>1.179</td>
<td>0.899</td>
<td>0.892</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Note: ***, **, and * denote the significance at 1, 5, and 10% levels based on the predictive accuracy test of Diebold and Mariano (1995), respectively.
Chapter 2

Standard Errors for Panel Data Models with Unknown Clusters

Note: This chapter is co-authored with Jushan Bai and Yuan Liao and is published in the Journal of Econometrics (Bai et al., 2020).
2.1 Introduction

Consider a linear panel regression with fixed-effects:

\[ y_{it} = x_{it}' \beta + \alpha_i + \mu_t + u_{it}, \]

where \( \alpha_i \) and \( \mu_t \) are individual fixed-effects and time fixed effects; \( x_{it} \) is a \( k \times 1 \) vector of explanatory variables; \( u_{it} \) is an unobservable error component. The outcome variable \( y_{it} \) and fixed effects are scalars, and \( \beta \) is a \( k \times 1 \) vector.

This paper is about the standard error of the fixed-effect ordinary least squares (OLS). One of the commonly used standard errors for OLS in empirical research is the White (1980) heteroskedasticity robust standard error in the cross-sectional setting. In the presence of serial and cross-sectional correlations, the conventional panel standard errors may be biased. Newey and West (1987) introduced heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimator for time series, which allows serial correlations (also see Andrews, 1991; Newey and West, 1994). The cluster standard errors suggested by Arellano (1987) are often reported in studies of the panel model. This estimator is robust to heteroskedasticity in the cross-section and also arbitrary serial correlation, but it focuses on the large-\( N \) small-\( T \) scenario. The case of large-\( N \) large-\( T \) is then studied by Ahn and Moon (2014), Hansen (2007a), among many others, while either cross-sectional or serial independence is required. Hansen (2007a) examined the covariance estimator when the time series dependence is left unrestricted. In addition, Vogelsang (2012) studied the asymptotic theory that is robust to heteroskedasticity, autocorrelation, and spatial correlation, which extended and generalized the asymptotic results of Hansen (2007a) for the conventional cluster standard errors including time fixed effects. Stock and Watson (2008) suggested a bias-adjusted heteroskedasticity-robust variance matrix estimator that handles serial correlations under any sequences of \( N \) or \( T \). Also, see Petersen (2009) who used a simulation study to examine different types of standard errors, including the clustered, Fama-MacBeth, and the modified version of Newey-West standard errors for panel data. In general, on the other hand, the conventional cluster standard errors assume that individuals across clusters are independent. Also, the cluster structure should be known such as schools, villages, industries, or states. See Arellano (2003), Cameron and Miller (2015), and
Greene (2003). However, the knowledge of clusters is not available in many applications.

In a recent interesting paper, Abadie et al. (2017) argue that clustering is an issue more of sampling design or experimental design. Clustered standard errors are not always necessary and researchers should be more thoughtful when applying them. One reason is that clustering may result in an unnecessarily wider confidence interval. Clustered standard errors are derived from the modeling perspective (model implied variance matrix) and are widely practiced, see, for example, Angrist and Pischke (2008), Cameron and Trivedi (2005), and Wooldridge (2003, 2010). In this paper, we continue to take the modeling perspective. Because of our use of thresholding method, the resulting confidence interval is not necessarily much wider, even if all cross-sectional units are allowed to be correlated. Furthermore, the proposed approach is also applicable when the knowledge of clustering is not available.

We provide a robust standard error that allows both serial and cross-sectional correlations. We do not impose parametric structures on the serial or cross-sectional correlations. We assume these correlations are weak and apply nonparametric methods to estimate the standard errors. To control for the autocorrelation in time series, we employ the Newey-West truncation. To control for the cross-sectional correlation, we assume sparsity for cross-section \((i, j)\) pairs, potentially resulting from the presence of cross-sectional clusters, but the knowledge on clustering (the number of clusters and the size of each cluster) is not assumed. We then estimate them by applying the thresholding approach of Bickel and Levina (2008a). We also show how to make use of information on clustering when available. In passing we point out that instead of robust standard errors, in a separate study, Bai et al. (2021) proposed a feasible GLS (FGLS) method to take into account heteroskedasticity and both serial and cross-sectional correlations. The FGLS is more efficient than OLS.

The regularization methods such as banding and thresholding employed in this paper have been used extensively in the recent machine learning literature for estimating high-dimensional parameters. Nonparametric machine learning techniques have been proved to be useful tools in econometric studies.

The rest of the chapter is organized as follows. In Section 2.2, we describe the models and standard errors as well as the asymptotic results of OLS. Monte Carlo studies
evaluating the finite sample performance of the estimators are presented in Section 2.3. Section 2.4 illustrates our methods in an application of US divorce law reform effects. Conclusions are provided in Section 2.5 and all proofs are given in the appendix.

Throughout this chapter, \( \nu_{\min}(A) \) and \( \nu_{\max}(A) \) denote the minimum and maximum eigenvalues of matrix \( A \). We use \( \|A\| = \sqrt{\nu_{\max}(A'A)} \), \( \|A\|_1 = \max_i \sum_j |A_{ij}| \) and \( \|A\|_F = \sqrt{\text{tr}(A'A)} \) as the operator norm, the \( \ell_1 \)-norm and the Frobenius norm of a matrix \( A \), respectively. Note that if \( A \) is a vector, \( \|A\| \) is the Euclidean norm, and \( |a| \) is the absolute-value norm of a scalar \( a \).

### 2.2 OLS and Standard Error Estimation

We consider the following model:

\[
y_{it} = x_{it}'\beta + u_{it},
\]

(2.2.1)

where \( \beta \) is a \( k \times 1 \) vector of unknown coefficients, \( x_{it} \) is a \( k \times 1 \) vector of regressors, and \( u_{it} \) represents the error term, often known as the idiosyncratic component. This formulation incorporates the standard fixed effects models as in Hansen (2007a). For example, \( x_{it}, y_{it} \) and \( u_{it} \) can be interpreted as variables resulting from removing the nuisance parameters from the equation, such as first-differencing to remove the fixed effects. Indeed, it is straightforward to allow additive fixed effects by using the usual demean procedure.

For a fixed \( t \), model (2.2.1) can be written as:

\[
y_t = x_t'\beta + u_t,
\]

(2.2.2)

where \( y_t = (y_{1t}, ..., y_{Nt})' \) \( (N \times 1) \), \( x_t = (x_{1t}, ..., x_{Nt})' \) \( (N \times k) \), and \( u_t = (u_{1t}, ..., u_{Nt})' \) \( (N \times 1) \). To economize notation, we define \( y_i = (y_{i1}, ..., y_{iT})' \) \( (T \times 1) \), \( x_i = (x_{i1}, ..., x_{iT})' \) \( (T \times k) \), and \( u_i = (u_{i1}, ..., u_{iT})' \) \( (T \times 1) \). So when the vector \( y \) is indexed by \( t \), it refers to an \( N \times 1 \) vector, and when \( y \) is indexed by \( i \) it refers to a \( T \times 1 \) vector. Similar meaning is applied to \( x \) and \( u \). There is no confusion when context is clear.

The (pooled) ordinary least square (OLS) estimator of \( \beta \) from equations (2.2.1) and
(2.2.2) may be defined as 

\[ \hat{\beta} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}x_{it}' \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}y_{it} = \left( \sum_{t=1}^{T} x_{t}x_{t}' \right)^{-1} \sum_{t=1}^{T} x_{t}'y_{t}. \]  

(2.2.3)

The variance of \( \hat{\beta} \) depends on both \( V_X \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}x_{it}' \), and particularly, 

\[
V \equiv Var\left( \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}u_{it} \right)
= \frac{1}{NT} \sum_{t=1}^{T} E x_{t}u_{t}u_{t}'x_{t} + \frac{1}{NT} \sum_{h=1}^{T-1} \sum_{t=h+1}^{T} [E x_{t}u_{t}u_{t-h}x_{t-h} + E x_{t}u_{t-h}u_{t-h}'x_{t-h}],
\]

(2.2.4)

The goal of this paper is to consistently estimate \( V \) in the presence of both serial and cross-sectional correlations in \( \{u_{it}\} \).

There are two types of clustered standard errors suggested by Arellano (1987). The original individual clustered version is 

\[
\hat{V}_{CX} = \frac{1}{NT} \sum_{i=1}^{N} x_{i}' \hat{u}_{i} \hat{u}_{i}' x_{i},
\]

with \( \hat{u}_{i} = y_{i} - x_{i}\hat{\beta} \) are the OLS residuals, and this estimator allows for arbitrary serial dependence and heteroskedasticity within individuals. In addition, \( \hat{V}_{CX} \) assumes no cross-section correlation.

The time-clustered version, which allows for heteroskedasticity and arbitrary cross-sectional correlation, is 

\[
\hat{V}_{CT} = \frac{1}{NT} \sum_{t=1}^{T} x_{t}' \hat{u}_{t} \hat{u}_{t}' x_{t},
\]

with \( \hat{u}_{t} = y_{t} - x_{t}\hat{\beta} \). Here \( \hat{V}_{CT} \) assumes no serial correlation.

The above clustered standard errors are robust to either arbitrary serial correlation or arbitrary cross-sectional correlation, respectively. In practice, however, since the dependence assumption is unknown, an over-rejection problem may occur. Specifically, if there exist both serial and cross-sectional correlations, these estimators are not robust anymore, as our numerical evidence shows in Section 2.3 (e.g., Tables 2.3 and 2.5).

To control for the serial correlation, a simple modification of \( \hat{V}_{CT} \) using Newey and
West (1987) is

\[
\hat{V}_{DK} = \frac{1}{NT} \sum_{t=1}^{T} x_t' \hat{u}_t' x_t + \frac{1}{NT} \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} [x_t' \hat{u}_t \hat{u}'_{t-h} x_{t-h} + x_{t-h} \hat{u}_{t-h} \hat{u}'_t x_t],
\]

(2.2.5)

where \( \omega(\cdot) \) is the kernel function and \( L \) is the bandwidth. This estimator is suggested by Driscoll and Kraay (1998). When \( N \) is large, however, (2.2.5) accumulates a large number of cross-sectional estimation noises.

More generally, let

\[
V_{ij} = \frac{1}{T} \sum_{t=1}^{T} E x_{it} u_{it} u_{jt} x_{jt} + \frac{1}{T} \sum_{h=1}^{T-1} \sum_{t=h+1}^{T} [E x_{it} u_{it} u_{jt} x_{jt} + E x_{i,t-h} u_{i,t-h} u_{jt} x_{jt}].
\]

Then equation (2.2.4) can be written as

\[
V = \frac{1}{N} \sum_{ij} V_{ij}.
\]

Unlike time series observations, cross-sectional observations have no natural ordering. They can be arranged in different orders. That is why cross-sectional correlation is more difficult to control. The usual cluster standard error makes the following assumption: let \( C_1, \ldots, C_G \) be disjoint subsets of \( \{1, \ldots, N\} \), so that they are known clusters and that \( V_{ij} = 0 \) when \( i \) and \( j \) belong to different clusters. So \( V \) can be expressed as

\[
V = \frac{1}{N} \sum_{g=1}^{G} \sum_{(i,j) \in C_g} V_{ij}.
\]

See Liang and Zeger (1986). Suppose the cardinality of each \( C_g \) is small (this would be the case if the number of clusters \( G \) is large) or grows slowly with \( N \), then we only need to estimate \( \sum_{g=1}^{G} \sum_{(i,j) \in C_g} 1 \) number of \( V_{ij} \)'s, greatly reducing the number of pairwise covariances. But as commented in the literature, this requires the knowledge of \( C_1, \ldots, C_G \), which in some applications, is not naturally available.

### 2.2.1 The estimator of \( V \) with unknown clusters

The key assumption we make is that conditionally on \( x_t, \{u_{it}\} \) is weakly correlated across both \( t \) and \( i \). Essentially, this means \( V_{ij} \) is zero or nearly so for most pairs of \( (i, j) \). There
is a partition \( \{(i, j) : i, j \leq N\} = S_s \cup S_l \) so that

\[
S_s = \{(i, j) : \|Ex_iu_tu_jx_{j,t+h}^t\| = 0 \forall h\}, \quad S_l = \{(i, j) : \|Ex_iu_tu_jx_{j,t+h}^t\| \neq 0 \exists h\},
\]

where the subscript “s” indicates “small”, and “l” indicates “large”. We assume that \((i, i) \in S_l\) for all \(i \leq N\), and importantly, most pairs \((i, j)\) belong to \(S_s\). Yet, we do not need to know which elements belong to \(S_s\) or \(S_l\). Then

\[
V = \frac{1}{N} \sum_{(i,j) \in S_l} V_{ij}.
\]

Furthermore, let \(\omega(h, L) = 1 - h/(L + 1)\) be the Bartlett kernel. Also see Andrews (1991) for other kernel functions. As suggested by Newey and West (1987), \(V_{ij}\) can be approximated by

\[
V_{u,ij} \equiv \frac{1}{T} \sum_{t=1}^{T} Ex_iu_tu_jx_{j,t}^t + \frac{1}{L} \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} [Ex_iu_tu_jx_{j,t}^t + Ex_{i,t-h}u_{i,t-h}u_{j,t}x_{j,t}^t].
\]

Then approximately,

\[
V \approx \frac{1}{N} \sum_{(i,j) \in S_l} V_{u,ij}.
\]

The above approximation plays the fundamental role of our standard error estimator. We estimate \(V_{ij}\) using Newey and West (1987), and estimate \(S_l\) using the cross-sectional thresholding.

To apply Newey and West (1987), we estimate \(V_{u,ij}\) by

\[
S_{u,ij} \equiv \frac{1}{T} \sum_{t=1}^{T} x_iu_tu_jx_{j,t}^t + \frac{1}{L} \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} [x_iu_tu_jx_{j,t}^t + x_{i,t-h}u_{i,t-h}u_{j,t}x_{j,t}^t],
\]

where \(\hat{u}_{it} = y_{it} - x_{it}'\hat{\beta}\). For a predetermined threshold value \(\lambda_{ij}\), we approximate \(S_l\) by

\[
\hat{S}_l = \{(i, j) : \|S_{u,ij}\| > \lambda_{ij}\}.
\]
Hence, a “matrix hard-thresholding” estimator of $V$ is

$$\hat{V}_{\text{Hard}} \equiv \frac{1}{N} \sum_{(i,j) \in \hat{S} \cup \{i=j\}} S_{u,ij}.$$ 

As for the threshold value, we specify

$$\lambda_{ij} = M \omega_{NT} \sqrt{\|S_{u,ii}\| \|S_{u,jj}\|}, \text{ where } \omega_{NT} = L \sqrt{\frac{\log(LN)}{T}}$$

for a constant $M > 0$. The converging sequence $\omega_{NT} \to 0$ is chosen to satisfy:

$$\max_{i,j \leq N} \|S_{u,ij} - V_{u,ij}\| = O_P(\omega_{NT}).$$

In practice, the thresholding constant, $M$, can be chosen through multifold cross-validation, which is discussed in the next subsection. In addition, we can obtain $\hat{V}_{DK}$ from $\hat{V}_{\text{Hard}}$ by setting $M = 0$. 

We also recommend a “matrix soft-thresholding” estimator as follows:

$$\hat{V}_{\text{Soft}} = \frac{1}{N} \sum_{i,j} \hat{S}_{u,ij},$$

where $\hat{S}_{u,ij}$ is

$$\hat{S}_{u,ij} = \begin{cases} 
S_{u,ij}, & \text{if } i = j, \\
A_{u,ij}, & \text{if } \|S_{u,ij}\| > \lambda_{ij}, \text{ and } i \neq j, \\
0, & \text{if } \|S_{u,ij}\| < \lambda_{ij}, \text{ and } i \neq j,
\end{cases}$$

where the $(k,k')$’s element of $A_{u,ij}$ is $(\text{sgn}(x)$ denotes the sign function)

$$A_{u,ij,kk'} = \begin{cases} 
\text{sgn}(S_{u,ij,kk'})[S_{u,ij,kk'} - \eta_{ij,kk'}], & \text{if } |S_{u,ij,kk'}| > \eta_{ij,kk'}, \\
0, & \text{if } |S_{u,ij,kk'}| < \eta_{ij,kk'},
\end{cases}$$

for the threshold value

$$\eta_{ij,kk'} = M \omega_{NT} \sqrt{|S_{u,ii,kk'}| \|S_{u,jj,kk'}|}, \text{ where } \omega_{NT} = L \sqrt{\frac{\log(LN)}{T}}$$

for some constant $M > 0$. 
Remark 2.2.1. The thresholding estimators for $V$ do not assume known cluster information (the number of clusters and the membership of clusters). The method can also be modified to take into account the clustering information when available, and is particularly suitable when the number of clusters is small, and the size of each cluster is large. The modification is to apply the thresholding method within each cluster. The conventional clustered standard errors lose a lot of degrees of freedom when the size of cluster is too large (because each cluster is effectively treated as a “single observation”), resulting in conservative confidence intervals. See Cameron and Miller (2015). The thresholding avoids this problem, while allowing correlations of unknown form within each cluster.

2.2.2 Choice of tuning parameters

Our suggested estimators, $\hat{V}_{\text{Hard}}$ and $\hat{V}_{\text{Soft}}$, require the choice of tuning parameters $L$ and $M$, which are the bandwidth and the threshold constant respectively. To choose the bandwidth $L$, we recommend using $L = 4(T/100)^{2/9}$ as Newey and West (1994) suggested.

In practice, $M$ can be chosen through multifold cross-validation. After obtaining the estimated residuals $\hat{u}_{it}$ by OLS, we split the data into two subsets, denoted by $\{\hat{u}_{it}\}_{t \in J_1}$ and $\{\hat{u}_{it}\}_{t \in J_2}$; let $T(J_1)$ and $T(J_2)$ be the sizes of $J_1$ and $J_2$, which are $T(J_1) + T(J_2) = T$ and $T(J_1) \approx T$. As suggested by Bickel and Levina (2008a), we can set $T(J_1) = T(1 - \log(T)^{-1})$ and $T(J_2) = T/\log(T)$; $J_1$ represents the training data set, and $J_2$ represents the validation data set.

The procedure requires splitting the data multiple times, say $P$ times. At the $p$th split, we denote by $\hat{V}^p$ the sample covariance matrix based on the validation set, defined by

$$\hat{V}^p = \frac{1}{N} \sum_{ij} S_{u,ij}^p,$$

with $S_{u,ij}^p$ defined similarly to $S_{u,ij}$ using data on $J_2$. Let $\hat{V}_n(M)$ be the thresholding estimator with threshold constant $M$ using the entire sample. Then we choose the constant $M^*$ by minimizing a cross-validation objective function.
\[ M^* = \arg \min_{0 < M < M_0} \frac{1}{P} \sum_{p=1}^{P} \| \hat{V}_s(M) - \hat{V}_p \|_F^2, \quad s \in \{ \text{Hard}, \text{Soft} \} \]

and the resulting estimator is \( \hat{V}_s(M^*) \). We use \( L = 4(T/100)^{2/9} \) for both \( \hat{V}_s(M) \) and \( \hat{V}_p \) and find that setting \( M_0 = 1 \) works well. So the minimization is taken over \( M \in (0, 1) \) through a grid search.

The above procedure modifies that of Bickel and Levina (2008a) in two aspects. One is to use the entire sample when computing \( \hat{V}_s \) instead of \( J_1 \). Since \( T(J_1) \) is close to \( T \), this modification does not change the result much, but simplifies the computation. The second modification is to use a consecutive block for the validation set because of time series, so that the serial correlation is not perturbed. Hence in view of the time series nature, we first divide the data into \( P = \log(T) \) blocks with block length \( T/\log(T) \). Each \( J_2 \) is taken as one of the \( P \) blocks when computing \( \hat{V}_p \), similar to the K-fold cross-validation. We have conducted simulations of the cross-validation in the presence of both correlations, and the results show that this procedure performs well. For instance, the cross-validation tends to choose smaller \( M \) as the cross-sectional correlation becomes stronger. Due to the page limit, however, those are not reported in this paper.

2.2.3 Consistency

Below we present assumptions under which \( \hat{V} \) (either \( \hat{V}_{\text{Hard}} \) or \( \hat{V}_{\text{Soft}} \)) consistently estimates \( V \). We define

\[ \alpha_{NT}(h) \equiv \sup_{X} \max_{t \leq T} [\| E(u_{t}u_{t-h} | X) \| + \| E(u_{t-h}u_{t} | X) \|] \]

and

\[ \rho_{ij,h} \equiv \sup_{X} \max_{t \leq T} [\| E(u_{i,t}u_{j,t-h} | X) \| + \| E(u_{i,t-h}u_{j,t} | X) \|], \]

where \( X = \{ x_{it} \}_{i \leq N, t \leq T} \). These coefficients give measures of autocovariances and cross-section covariances.

**Assumption 2.2.1.** (i) \( E(u_{t} | x_{t}) = 0 \).

(ii) Let \( \nu_1 \leq \ldots \leq \nu_k \) be the eigenvalues of \( (\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E x_{it} x_{it}') \). Then there exist constants \( c_1, c_2 > 0 \) such that \( c_1 < \nu_1 \leq \cdots \leq \nu_k < c_2 \).
Assumption 2.2.2. (weak serial and cross-sectional dependence).

(i) $\sum_{h=0}^{\infty} \alpha_{NT}(h) \leq C$ for some $C > 0$. In addition, there exist $\kappa \in (0, 1)$, $C > 0$ such that for all $T > 0$,

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

where $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the $\sigma$-algebras generated by $\{(x_t, u_t) : t \leq 0\}$ and $\{(x_t, u_t) : t \geq T\}$ respectively.

(ii) For some $q \in [0, 1)$, $\omega_{NT}^{1-q} \max_{i \leq N} \sum_{j=1}^{N} (\sum_{h=0}^{L} \rho_{ij,h})^q = o(1)$, where $\omega_{NT} = \sqrt{\log(LN)}$. Assumption 2.2.2 (i) is the standard alpha-mixing condition, adapted to the large-$N$ panel. Condition (ii) is new here. It requires weak cross-sectional correlations. It is similar to the “approximate sparse assumption” in Bickel and Levina (2008a). Note that we actually allow the presence of many “small” but nonzero $\|Ex_t u_t u'_j t + h^t x'_j t + h\|$. Clusters that have “large” $\|Ex_t u_t u'_j t + h^t x'_j t + h\|$ are unknown to us. Hence the appealing feature of our method is that we allow for unknown clusters.

Essentially the assumption $\omega_{NT}^{1-q} \max_{i \leq N} \sum_{j \leq N} (\sum_{h=0}^{L} \rho_{ij,h})^q = o(1)$ controls the order of elements in $S_l$. The following example presents a case of cross sectional weak correlations that satisfies condition (ii).

Example 2.2.1. Suppose uniformly for all $h = 0, \ldots, L$, $E(u_t u'_t | X)$ is an $N \times N$ block-diagonal matrix, where the size of each block is at most $S_{NT}$, which practically means that each cluster contains no more than $S_{NT}$ individuals, assuming clusters are mutually uncorrelated. Then $\rho_{ij,h} = 0$ for $(i, j)$ belong to different blocks. Within the same block, almost surely in $X$,

$$|E(u_t u'_t X) + |E(u_t u'_t u_j t + h X) \leq \alpha_{NT}(h), \sum_{h=0}^{\infty} \alpha_{NT}(h) < \infty$$

Then let $B(i)$ denote the block that $i$ belongs to, whose size is at most $S_{NT}$.

$$\omega_{NT}^{1-q} \max_{i \leq N} \sum_{j=1}^{N} (\sum_{h=0}^{L} \rho_{ij,h})^q = \omega_{NT}^{1-q} \max_{i \leq N} \sum_{j \in B(i)} (\sum_{h=0}^{L} \rho_{ij,h})^q \leq C\omega_{NT}^{1-q} S_{NT} \sum_{h=0}^{\infty} \alpha_{NT}(h)^{q/c}$$
for constants \( c, C > 0 \). The last term converges to zero so long as \( \omega_{NT}^{-1} S_{NT} \to 0 \). This then requires either fixed or slowly growing cluster size \( S_{NT} \).

**Assumption 2.2.3.** (i) For each fixed \( h, \omega(h, L) \to 1 \) as \( L \to \infty \) and \( \max_{h \leq L} |\omega(h, L)| \leq C \) for some \( C > 0 \).

(ii) Exponential tail: There exist \( r_1, r_2 > 0 \) and \( b_1, b_2 > 0 \), such that \( r_1^{-1} + r_2^{-1} + \kappa^{-1} > 1 \), and for any \( s > 0, i \leq N \),

\[
P(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1}), \quad P(|x_{it}| > s) \leq \exp(-(s/b_2)^{r_2}).
\]

(iii) There is \( c_1 > 0 \), for all \( i, \lambda \), \( \min(\text{var}(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{it} u_{it})) > c_1 \). Additionally, the eigenvalues of \( V \) and \( V_X \) are bounded away from both zero and infinity.

Condition (i) is well satisfied by various kernels for the HAC-type estimator. Condition (ii) ensures the Bernstein-type inequality for weakly dependent data. Note that it requires the underlying distributions to be thin-tailed. Allowing for heavy-tailed distributions is also an important issue. However, it would require a very different estimation method, and is out of the scope of this paper.

We have the following main theorem and all proofs are contained in the appendix.

**Theorem 2.2.1.** Under Assumption 2.2.1-2.2.3, as \( N, T \to \infty \),

\[
\sqrt{NT}[V_X^{-1}\tilde{V}V_X^{-1}]^{-1/2}(\hat{\beta} - \beta) \overset{d}{\to} \mathcal{N}(0, I).
\]

Theorem 2.2.1 allows us to construct a \((1 - \tau)\)% confidence interval for \( c'\beta \) for any given \( c \in \mathbb{R}^k \). The standard error of \( c'\hat{\beta}_{OLS} \) is

\[
\left( \frac{1}{NT} c' (V_X^{-1}\tilde{V}V_X^{-1}) c \right)^{1/2}
\]

and the confidence interval for \( c'\beta \) is \([c'\hat{\beta} \pm Z_{\tau} \hat{\sigma} / \sqrt{NT}]\) where \( Z_{\tau} \) is the \((1 - \tau)\)% quantile of standard normal distribution and \( \hat{\sigma} = (c'(V_X^{-1}\tilde{V}V_X^{-1}) c)^{1/2} \).
2.3 Monte Carlo Experiments

2.3.1 DGP and methods

In this section we examine the finite sample performance of the robust standard errors using simulation study. The data generating process (DGP) used for the simulation is produced by the fixed effect linear regression model

\[ y_{it} = \alpha_i + \mu_t + \beta_0 x_{it} + u_{it}, \]

where the true \(\beta_0 = 1\). The DGP allows for serial and cross-sectional correlations in \(x_{it}\) as follows:

\[ x_{it} = a_i \nu_{i+1,t} + \nu_{i,t} + b_i \nu_{i-1,t}, \quad \nu_{it} = \rho X \nu_{it-1} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, 1), \quad \nu_{i0} = 0, \]

\[ \alpha_i \sim N(0, 0.5), \quad \mu_t \sim N(0, 0.5), \]

where the constants \(\{a_i, b_i\}_{i=1}^N\) are i.i.d. Uniform(0, \(\gamma_X\)), which introduce cross-sectional correlation. In addition, \(\nu_{it}\) is modeled as AR(1) process with the autoregressive parameter \(\rho_X\). Throughout this simulation study, we set \(\rho_X = 0.3\) and \(\gamma_X = 1\).

We generate the error terms, \(u_{it}\), in three different cases as follows:

Case 1: \(u_{it} = c_i m_{i+1,t} + m_{i,t} + d_i m_{i-1,t}, \quad m_{it} = \rho m_{i,t-1} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, 1), \quad m_{i0} = 0, \)

Case 2: \(u_{it} = \psi \sum_{j=1}^{N} w_{ij} u_{jt} + \eta_{it}, \quad \eta_{it} \sim N(0, 1), \quad u_{i0} = 0, \)

Case 3: \(u_{it} = \sum_{k=1}^{r} \lambda_{ij} F_{tk} + \epsilon_{it}, \quad F_{tk} = \rho F_{t-1,k} + \xi_{tk}, \quad \lambda_{tk} = \rho \lambda_{t-1,k} + \zeta_{tk}, \quad \epsilon_{it} \sim N(0, 1), \quad \xi_{it} \sim N(0, 1), \quad \zeta_{it} \sim N(0, 1). \)

The regressor is uncorrelated with the error term \(u_{it}\) each other. In Case 1, we generate the error term similar to \(x_{it}\). The constants \(\{c_i, d_i\}_{i=1}^N\) are i.i.d. Uniform(0, \(\gamma\)), which introduce cross-sectional correlation, and heteroskedasticity when \(\gamma > 0\). \(m_{it}\) is modeled as AR(1) process with the autoregressive parameter \(\rho\). Varying \(\gamma > 0\) allows us to control for the strength of the cross-sectional correlation. Data are generated with four different structures of regressors and error terms: (a) no correlations (\(\rho = 0, \gamma = 0\)); (b) only
serial correlation ($\rho = 0.5, \gamma = 0$); (c) only cross-sectional correlation ($\rho = 0, \gamma = 1$); and (d) both serial and cross-sectional correlations ($\rho = \{0.3, 0.9\}, \gamma = 1$). In Case 2, the error terms are modeled as a spatial autoregressive (SAR(1)) process. The matrix $W = (w_{ij})_{N \times N}$ is a rook type weight matrix whose diagonal elements are zero. Note that the rows of $W$ are standardized, hence they sum to one. $\psi$ is the scalar spatial autoregressive coefficient with $|\psi| < 1$. In this paper, we report the case of $\psi = 0.5$.

Importantly, SAR(1) model does not produce the serial correlation on the error term. In Case 3, we consider an error factor structure. Both factors and factor loadings follow AR(1) processes, which introduce both serial and cross-sectional correlations. We set $r = 2$, and consider the cases of $\rho_\lambda = 0.3$ and $\rho_F = 0.9$.

In this simulation study, we examined $t$-statistics for testing the null hypothesis $H_0 : \beta_0 = 1$ against the alternative $H_1 : \beta_0 \neq 1$. In each simulation we compare the proposed estimator with that of other common five types of standard errors for $\hat{\beta}$: the standard White estimator given by $\hat{V}_{\text{White}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{u}_{it} \tilde{u}_{it}' \hat{u}_{it}^2$, where $\tilde{x}_{it}$ is demeaned version of regressor. Two types of clustered standard errors, $\hat{V}_{\text{CX}}$ and $\hat{V}_{\text{CT}}$, as defined in Section 2.2. In addition, we use two types of Newey and West HAC estimators for the panel version as follows:

$$\hat{V}_{\text{DK}} = \frac{1}{NT} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{u}_{it} \tilde{u}_{it}' \hat{x}_{it} + \frac{1}{NT} \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} \frac{1}{\tilde{x}_{it}' \tilde{u}_{it} \tilde{u}_{it}' \hat{x}_{it} + \tilde{x}_{t-h} \tilde{u}_{t-h} \tilde{u}_{t-h}' \hat{x}_{t-h}}$$

and

$$\hat{V}_{\text{HAC}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{x}_{it} \tilde{u}_{it}^2$$

$$+ \frac{1}{NT} \sum_{i=1}^{N} \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} \frac{1}{\tilde{x}_{it}' \tilde{u}_{it} \tilde{u}_{it}' \hat{x}_{it} + \tilde{x}_{t-h} \tilde{u}_{t-h} \tilde{u}_{t-h}' \hat{x}_{t-h}}$$.

Note that $\hat{V}_{\text{HAC}}$ assumes cross-sectional independence, while $\hat{V}_{\text{DK}}$ allows arbitrary cross-sectional dependence. In addition, $\hat{V}_{\text{DK}}$ and $\hat{V}_{\text{HAC}}$ can be obtained from our proposed estimator with $M = 0$ and a large constant $M$, respectively.

Results are given for sample sizes $N = 50, 200$ and $T = 100, 200$. For each $\{N, T\}$ combination, we set $L = 3, 7, 11$ as the bandwidth for $\hat{V}_{\text{HAC}}, \hat{V}_{\text{DK}}$, and the proposed estimator, $\hat{V}_{\text{Hard}}$. We also use Bartlett kernel for these three estimators. For the thresholding constant parameters of $\hat{V}_{\text{Hard}}$, we set $M = 0.10, 0.15, 0.20, 0.25$ in all cases. The
simulation is replicated for one thousand times for each case and the nominal significance level is 0.05. Simulation results are reported in Tables 2.1 - 2.5.

2.3.2 Results

Tables 2.1 - 2.5 present the simulation results, where each table corresponds to different cases. Each table reports null rejection probabilities for 5% level tests based on six different standard errors. As expected, a common feature in all tables is that when both $N$ and $T$ are small, all six estimators have rejection probabilities greater than 0.05. This might happen even when the errors are drawn from i.i.d. standard normal, and this problem becomes more noticeable in the presence of serial, cross-sectional, or both correlations. A number of interesting findings based on tables are summarized below.

Tables 2.1-2.3 show the results of Case 1. In Table 2.1, Panel A indicates that all the estimators perform well due to no correlation. Especially, White standard error estimators give rejection probabilities close to 0.05. In Panel B, when the serial correlation is introduced, the performances of $\hat{V}_{CX}$ and $\hat{V}_{HAC}$ are markedly better than others except for small sample size. In addition, our proposed estimator, $\hat{V}_{Hard}$, also performs well when we use both larger threshold constant $M$ and bandwidth $L$. Since there is only a serial correlation in the error term, these estimators take this correlation into account and perform well. As the size of bandwidth increases, the standard error estimated by $\hat{V}_{HAC}$ increases to a level similar to the results of $\hat{V}_{CX}$ and the tendency to over-reject diminishes. Since the Newey-West technique gives the weight, which is less than one, the estimated standard error may be underestimated. Hence, the traditional cluster standard error, $\hat{V}_{CX}$, dominates the standard error of Newey-West panel version, $\hat{V}_{HAC}$. Note that the unreported rejection probabilities of $\hat{V}_{DK}$ exponentially increases as the bandwidth $L$ increases.

Table 2.2 considers the case of cross-sectionally correlated errors and regressors. In Panel A, except the case of small sample size, $\hat{V}_{CT}$ and $\hat{V}_{DK}$ with small bandwidth $L$ have rejection probabilities close to 0.05 in the first panel. Also, $\hat{V}_{Hard}$ with small $L$ and $M$ performs well. Importantly, notice that the rejection rate of $\hat{V}_{DK}$ and $\hat{V}_{Hard}$ tend to over-reject substantially as the lag length $L$ increases. In addition, as the cross-section size $N$ increases, the over-rejection problem becomes worse, as we mentioned in Section 2.2. This tendency is easy to explain. Since $\hat{V}_{DK}$ is an estimator based on a single time
series and it is zero when full weight is given to the sample autocovariance, the bias in $\hat{V}_{DK}$ initially falls but then increases as the lag length increases, while the variance of $\hat{V}_{DK}$ is initially increasing but eventually becomes decreasing. Hence, $\hat{V}_{DK}$ is biased downward substantially, and its t-statistics tends to over-reject when a large bandwidth is used. On the other hand, in the case of the small size of $L$ and $M$, $\hat{V}_{\text{Hard}}$ gives less bias on the estimated standard error.

Panel B of Table 2.2 allows the serial correlation as well as the cross-sectional correlation. Not surprisingly, all estimators except $\hat{V}_{\text{Hard}}$ and $\hat{V}_{DK}$ tend to over-reject substantially. In the small sample, these two estimators get worse than the case of the first panel. In the large sample, however, rejection probabilities of $\hat{V}_{\text{Hard}}$ and $\hat{V}_{DK}$ are close to 0.05. Importantly, $\hat{V}_{\text{Hard}}$ outperforms $\hat{V}_{DK}$ by choosing $M$ properly. Unreported results of $\hat{V}_{DK}$ with larger bandwidth, $L$, show much larger rejection probabilities than that of $\hat{V}_{\text{Hard}}$. This indicates that we can obtain unbiased standard error estimator and appropriate rejection rates using our proposed estimators, $\hat{V}_{\text{Hard}}$. Table 2.3 is the result of strong serial correlation with the cross-sectional dependence. When the serial correlation gets stronger, such as $\rho = 0.9$, all estimators tend to over-reject exponentially in small samples. However, $\hat{V}_{\text{Hard}}$ and $\hat{V}_{DK}$ outperform other estimators as the dimensionality increases.

Table 2.4 considers the error with SAR(1) structure, which does not require the serial correlation on the error term. Similar to the results reported in the first panel of Table 2.2, $\hat{V}_{CT}$ gives rejection probabilities close to 0.05. $\hat{V}_{DK}$ and $\hat{V}_{\text{Hard}}$ with small bandwidth $L$ also perform well. Moreover, $\hat{V}_{\text{Hard}}$ with proper thresholding constant $M$ gives less bias than $\hat{V}_{DK}$ on the estimated standard error.

Finally, Table 2.5 presents the results of the error factor structure. Similar to the results of Table 2.3, all estimators except $\hat{V}_{\text{Hard}}$ and $\hat{V}_{DK}$ tend to over-reject. Rejection probabilities of $\hat{V}_{\text{Hard}}$ and $\hat{V}_{DK}$ are relatively close to 0.05 when the sample size is large.

### 2.4 Empirical study: Effects of divorce law reforms

In this section, we re-examine the empirical work of the association between divorce law reforms and divorce rates using our proposed OLS standard error. There are many empirical studies on the effects of divorce law reforms on divorce rates. Friedberg (1998)
found that state law reforms significantly increased divorce rates with controls for state and year fixed effects. Wolfers (2006) investigated the question of whether law reform continues to have an impact on the divorce rate by including dummy variables for the first two years after the reforms, 3-4 years, 5-6 years, and so on. Specifically, he studied the following fixed effect panel data model

$$y_{it} = \alpha_i + \mu_t + \sum_{k=1}^{8} \beta_k X_{it,k} + \delta_i t + u_{it},$$

(2.4.1)

where $y_{it}$ is the divorce rate for state $i$ and year $t$; $\alpha_i$ and $\mu_t$ are the state and year fixed effects; $X_{it,k}$ is a binary regressor that representing the treatment effect $2k$ years after the reform; $\delta_i t$ a linear time trend. Wolfers (2006) suggested that there might be two sides of the same treatment yield this phenomenon: a number of divorces gradually shifted after the earlier dissolution of bad matches, after the reform.

Both Friedberg (1998) and Wolfers (2006) estimated OLS regressions using state population weight for each year. In addition, they estimated standard errors under the assumption that errors are homoskedastic, serially and cross-sectionally uncorrelated. However, ignoring these correlations might lead to bias in the standard error estimators. We re-estimated the model of Wolfers (2006) using proposed OLS standard error estimators.

The same data as in Wolfers (2006) are used, but we exclude Indiana, New Mexico, and Louisiana due to missing observations around divorce law reforms. As a result, we obtain a balanced panel data contain the divorce rates, state-level reform years, and binary regressors from 1956 to 1988 over 48 states. We fit models both with and without linear time trend, and also calculate our standard errors, as well as OLS, White, cluster, and HAC standard errors. We set lag choices $L = 3$ for HAC and our standard errors as suggested by Newey and West (1994) ($L = 4(T/100)^{2/9}$). The threshold values $M$ chosen by the cross-validation method is $M = 0.2$ for the model without state-specific linear trends, and $M = 0.1$ with state-specific linear trends. These $M$ values are relatively small, implying the existence of cross-sectional correlations. The estimated $\beta_1, \cdots, \beta_8$ with and without linear time trend and their different types of standard errors are presented respectively in Table 2.6. Note that robust standard errors are not necessarily larger than the usual OLS standard errors, as shown in columns corresponding to $se_{CT}$,
In Table 2.6, OLS estimates with and without linear time trend are similar to each other. These estimates are also closely comparable to the results obtained in Wolfers (2006). The OLS estimates indicate that divorce rates rose soon after the law reform. However, within a decade, divorce rates had fallen over time. Most of the coefficient estimates are statistically significant at the 5% level using usual OLS standard errors. According to the cluster standard errors, however, the only significant estimates are 11-15+ after the reform in the model without linear time trend. We use our method of correcting standard error estimates for heteroskedasticity, serial correlation, and also cross-sectional correlation. In the model without linear trend, the estimates for 3-4 and 7-15+ are significant. On the other hand, the estimates for 1-4 are significant when linear trend is added. Our estimated standard errors are close to those of \( se_{CT} \) and \( se_{DK} \), which allow arbitrary cross-section correlations. The result indicates non-negligible cross-sectional correlations. The result is also consistent with Kim and Oka (2014), who used the interactive fixed effects approach. The latter approach is suitable for models with strong cross-sectional correlations.

2.5 Conclusion

This paper studies the standard error problem for the OLS estimator in linear panel models, and proposes a new standard-error estimator that is robust to heteroskedasticity, serial and cross-sectional correlations when clusters are unknown. Simulated experiments demonstrate the robustness of the new standard-error estimator to various correlation structures.
Table 2.1: Null rejection probabilities, 5% level. Two-tailed test of $H_0: \beta = 1$. Case 1: No cross-sectional correlation ($\gamma = 0$).

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>L \ M</th>
<th>$\hat{V}_{\text{Hard}}$</th>
<th>$\hat{V}_{\text{HAC}}$</th>
<th>$\hat{V}_{\text{DK}}$</th>
<th>$\hat{V}_{\text{CX}}$</th>
<th>$\hat{V}_{\text{CT}}$</th>
<th>$\hat{V}_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.15</td>
<td>0.20</td>
<td>0.25</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>3</td>
<td>0.067</td>
<td>0.065</td>
<td>0.065</td>
<td>0.067</td>
<td>0.057</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>0.082</td>
<td>0.071</td>
<td>0.056</td>
<td>0.055</td>
<td>0.057</td>
<td>0.088</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>3</td>
<td>0.054</td>
<td>0.053</td>
<td>0.053</td>
<td>0.051</td>
<td>0.046</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>0.056</td>
<td>0.054</td>
<td>0.056</td>
<td>0.053</td>
<td>0.047</td>
<td>0.056</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>3</td>
<td>0.062</td>
<td>0.065</td>
<td>0.059</td>
<td>0.060</td>
<td>0.047</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>0.079</td>
<td>0.065</td>
<td>0.057</td>
<td>0.047</td>
<td>0.047</td>
<td>0.091</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>3</td>
<td>0.051</td>
<td>0.051</td>
<td>0.053</td>
<td>0.052</td>
<td>0.048</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>0.058</td>
<td>0.052</td>
<td>0.049</td>
<td>0.046</td>
<td>0.047</td>
<td>0.060</td>
</tr>
</tbody>
</table>

A. No serial correlation: $\rho = 0$

B. Serial correlation: $\rho = 0.5$
Table 2.2: Null rejection probabilities, 5% level. Two-tailed test of $H_0 : \beta = 1$. Case 1: Cross-sectional correlation ($\gamma = 1$).

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>L</th>
<th>M</th>
<th>$\hat{V}_{\text{Hard}}$</th>
<th>$\hat{V}_{HAC}$</th>
<th>$\hat{V}_{DK}$</th>
<th>$\hat{V}_{CX}$</th>
<th>$\hat{V}_{CT}$</th>
<th>$\hat{V}_{W}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.15</td>
<td>0.20</td>
<td>0.25</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>3</td>
<td></td>
<td>.054</td>
<td>.054</td>
<td>.055</td>
<td>.055</td>
<td>.142</td>
<td>.055</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.066</td>
<td>.064</td>
<td>.063</td>
<td>.069</td>
<td>.141</td>
<td>.068</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.078</td>
<td>.077</td>
<td>.082</td>
<td>.109</td>
<td>.145</td>
<td>.079</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>3</td>
<td></td>
<td>.046</td>
<td>.049</td>
<td>.046</td>
<td>.047</td>
<td>.133</td>
<td>.049</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.054</td>
<td>.055</td>
<td>.060</td>
<td>.060</td>
<td>.134</td>
<td>.052</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.059</td>
<td>.062</td>
<td>.063</td>
<td>.073</td>
<td>.135</td>
<td>.058</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>3</td>
<td></td>
<td>.060</td>
<td>.060</td>
<td>.060</td>
<td>.064</td>
<td>.148</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.069</td>
<td>.073</td>
<td>.075</td>
<td>.080</td>
<td>.149</td>
<td>.067</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.086</td>
<td>.085</td>
<td>.096</td>
<td>.126</td>
<td>.151</td>
<td>.084</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>3</td>
<td></td>
<td>.050</td>
<td>.051</td>
<td>.051</td>
<td>.050</td>
<td>.121</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.057</td>
<td>.058</td>
<td>.057</td>
<td>.057</td>
<td>.122</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.063</td>
<td>.062</td>
<td>.064</td>
<td>.079</td>
<td>.123</td>
<td>.062</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A. No serial correlation: $\rho = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>3</td>
<td></td>
<td>.070</td>
<td>.069</td>
<td>.069</td>
<td>.067</td>
<td>.150</td>
<td>.069</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.074</td>
<td>.075</td>
<td>.073</td>
<td>.078</td>
<td>.150</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.083</td>
<td>.079</td>
<td>.093</td>
<td>.108</td>
<td>.150</td>
<td>.085</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>3</td>
<td></td>
<td>.058</td>
<td>.058</td>
<td>.058</td>
<td>.058</td>
<td>.150</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.055</td>
<td>.060</td>
<td>.062</td>
<td>.061</td>
<td>.142</td>
<td>.056</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.059</td>
<td>.063</td>
<td>.071</td>
<td>.082</td>
<td>.142</td>
<td>.060</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>3</td>
<td></td>
<td>.078</td>
<td>.076</td>
<td>.077</td>
<td>.072</td>
<td>.162</td>
<td>.080</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.083</td>
<td>.087</td>
<td>.083</td>
<td>.084</td>
<td>.160</td>
<td>.086</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.097</td>
<td>.089</td>
<td>.103</td>
<td>.133</td>
<td>.159</td>
<td>.101</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>3</td>
<td></td>
<td>.055</td>
<td>.055</td>
<td>.054</td>
<td>.056</td>
<td>.132</td>
<td>.056</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>.053</td>
<td>.051</td>
<td>.056</td>
<td>.057</td>
<td>.130</td>
<td>.057</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>.057</td>
<td>.059</td>
<td>.065</td>
<td>.078</td>
<td>.130</td>
<td>.061</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B. Serial correlation: $\rho = 0.3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3: Null rejection probabilities, 5% level. Two-tailed test of $H_0: \beta = 1$. Case 1:
Both strong serial and cross-sectional correlations ($\rho = 0.9, \gamma = 1$).

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>L \ M</th>
<th>$\hat{V}_{Hard}$</th>
<th>$\hat{V}_{HAC}$</th>
<th>$\hat{V}_{DK}$</th>
<th>$\hat{V}_{CX}$</th>
<th>$\hat{V}_{CT}$</th>
<th>$\hat{V}_{W}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>3</td>
<td>.096 .097 .098 .098</td>
<td>.180 .098 .168 .145</td>
<td>.271</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.101 .101 .100 .101</td>
<td>.173 .102 .168 .145</td>
<td>.271</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.088 .087 .090 .093</td>
<td>.185 .087 .180 .157</td>
<td>.286</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>3</td>
<td>.089 .087 .087 .086</td>
<td>.193 .089 .146 .144</td>
<td>.256</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.092 .089 .095 .097</td>
<td>.178 .097 .146 .144</td>
<td>.256</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.100 .096 .109 .128</td>
<td>.173 .109 .146 .144</td>
<td>.256</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>3</td>
<td>.069 .069 .069 .067</td>
<td>.146 .068 .125 .121</td>
<td>.226</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.066 .068 .069 .067</td>
<td>.136 .069 .125 .121</td>
<td>.226</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.072 .071 .076 .087</td>
<td>.133 .072 .125 .121</td>
<td>.226</td>
<td>.291</td>
<td>.291</td>
<td>.271</td>
</tr>
</tbody>
</table>

Table 2.4: Null rejection probabilities, 5% level. Two-tailed test of $H_0: \beta = 1$. Case 2:
Errors with Spatial AR(1) structure ($\psi = 0.5$).

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>L \ M</th>
<th>$\hat{V}_{Hard}$</th>
<th>$\hat{V}_{HAC}$</th>
<th>$\hat{V}_{DK}$</th>
<th>$\hat{V}_{CX}$</th>
<th>$\hat{V}_{CT}$</th>
<th>$\hat{V}_{W}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>3</td>
<td>.061 .060 .062 .057</td>
<td>.124 .059 .143 .053</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.068 .068 .073 .077</td>
<td>.125 .067 .143 .053</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.086 .083 .089 .110</td>
<td>.128 .085 .143 .053</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>3</td>
<td>.046 .046 .047 .047</td>
<td>.113 .046 .130 .043</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.052 .053 .053 .059</td>
<td>.114 .048 .130 .043</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.052 .059 .065 .083</td>
<td>.112 .061 .130 .043</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
<td>.111</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>3</td>
<td>.061 .057 .058 .057</td>
<td>.123 .062 .120 .051</td>
<td>.122</td>
<td>.122</td>
<td>.122</td>
<td>.122</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.070 .071 .070 .081</td>
<td>.122 .068 .120 .051</td>
<td>.122</td>
<td>.122</td>
<td>.122</td>
<td>.122</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.082 .080 .101 .117</td>
<td>.121 .088 .120 .051</td>
<td>.122</td>
<td>.122</td>
<td>.122</td>
<td>.122</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>3</td>
<td>.055 .055 .053 .050</td>
<td>.124 .056 .125 .049</td>
<td>.123</td>
<td>.123</td>
<td>.123</td>
<td>.123</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.064 .061 .062 .064</td>
<td>.123 .064 .125 .049</td>
<td>.123</td>
<td>.123</td>
<td>.123</td>
<td>.123</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.069 .070 .083 .099</td>
<td>.123 .068 .125 .049</td>
<td>.123</td>
<td>.123</td>
<td>.123</td>
<td>.123</td>
</tr>
</tbody>
</table>
Table 2.5: Null rejection probabilities, 5% level. Two-tailed test of $H_0: \beta = 1$. Case 3: Errors with Factor structure ($\rho_F = 0.9, \rho_\lambda = 0.3$).

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th>$L \setminus M$</th>
<th>$\hat{V}_{\text{Hard}}$</th>
<th>$\hat{V}_{\text{HAC}}$</th>
<th>$\hat{V}_{\text{DK}}$</th>
<th>$\hat{V}_{\text{CX}}$</th>
<th>$\hat{V}_{\text{CT}}$</th>
<th>$\hat{V}_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>3</td>
<td>.081</td>
<td>.081</td>
<td>.078</td>
<td>.081</td>
<td>.129</td>
<td>.078</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.091</td>
<td>.084</td>
<td>.079</td>
<td>.082</td>
<td>.115</td>
<td>.093</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.103</td>
<td>.086</td>
<td>.094</td>
<td>.086</td>
<td>.115</td>
<td>.108</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>3</td>
<td>.083</td>
<td>.082</td>
<td>.081</td>
<td>.081</td>
<td>.117</td>
<td>.080</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.066</td>
<td>.066</td>
<td>.065</td>
<td>.073</td>
<td>.107</td>
<td>.065</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.072</td>
<td>.076</td>
<td>.080</td>
<td>.084</td>
<td>.104</td>
<td>.072</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>3</td>
<td>.076</td>
<td>.074</td>
<td>.076</td>
<td>.074</td>
<td>.109</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.077</td>
<td>.080</td>
<td>.077</td>
<td>.074</td>
<td>.105</td>
<td>.083</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.081</td>
<td>.076</td>
<td>.084</td>
<td>.094</td>
<td>.103</td>
<td>.096</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>3</td>
<td>.072</td>
<td>.072</td>
<td>.073</td>
<td>.069</td>
<td>.115</td>
<td>.071</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td>.070</td>
<td>.067</td>
<td>.071</td>
<td>.074</td>
<td>.106</td>
<td>.068</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>.074</td>
<td>.073</td>
<td>.073</td>
<td>.075</td>
<td>.104</td>
<td>.072</td>
</tr>
</tbody>
</table>
Table 2.6: Empirical application: effects of divorce law reform with state and year fixed effects: US state level data annual from 1956 to 1988, dependent variable is divorce rate per 1000 persons per year. OLS estimates and standard errors (using state population weights).

<table>
<thead>
<tr>
<th>Effects</th>
<th>$\hat{\beta}_{OLS}$</th>
<th>$se_{OLS}$</th>
<th>$se_W$</th>
<th>$se_{CX}$</th>
<th>$se_{CT}$</th>
<th>$se_{HAC}$</th>
<th>$se_{DK}$</th>
<th>$se_{Hard}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Without state-specific linear time trends</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–2 years</td>
<td>.256</td>
<td>.086*</td>
<td>.140</td>
<td>.189</td>
<td>.139</td>
<td>.172</td>
<td>.155</td>
<td>.148</td>
</tr>
<tr>
<td>3–4 years</td>
<td>.209</td>
<td>.086*</td>
<td>.081*</td>
<td>.159</td>
<td>.075*</td>
<td>.114</td>
<td>.104*</td>
<td>.089*</td>
</tr>
<tr>
<td>5–6 years</td>
<td>.126</td>
<td>.086</td>
<td>.073</td>
<td>.168</td>
<td>.064*</td>
<td>.105</td>
<td>.088</td>
<td>.069</td>
</tr>
<tr>
<td>7–8 years</td>
<td>.105</td>
<td>.086</td>
<td>.070</td>
<td>.165</td>
<td>.059</td>
<td>.100</td>
<td>.065</td>
<td>.040*</td>
</tr>
<tr>
<td>9–10 years</td>
<td>-.122</td>
<td>.085</td>
<td>.060*</td>
<td>.161</td>
<td>.041*</td>
<td>.088</td>
<td>.058*</td>
<td>.054*</td>
</tr>
<tr>
<td>11–12 years</td>
<td>-.344</td>
<td>.085*</td>
<td>.071*</td>
<td>.173*</td>
<td>.043*</td>
<td>.101*</td>
<td>.056*</td>
<td>.075*</td>
</tr>
<tr>
<td>13–14 years</td>
<td>-.496</td>
<td>.085*</td>
<td>.074*</td>
<td>.188*</td>
<td>.050*</td>
<td>.110*</td>
<td>.054*</td>
<td>.062*</td>
</tr>
<tr>
<td>15+ years</td>
<td>-.508</td>
<td>.081*</td>
<td>.089*</td>
<td>.223*</td>
<td>.048*</td>
<td>.139*</td>
<td>.061*</td>
<td>.077*</td>
</tr>
<tr>
<td>Panel B: With state-specific linear time trends</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–2 years</td>
<td>.286</td>
<td>.064*</td>
<td>.152</td>
<td>.206</td>
<td>.143*</td>
<td>.185</td>
<td>.145*</td>
<td>.140*</td>
</tr>
<tr>
<td>3–4 years</td>
<td>.254</td>
<td>.071*</td>
<td>.099*</td>
<td>.171</td>
<td>.102*</td>
<td>.140</td>
<td>.134</td>
<td>.126*</td>
</tr>
<tr>
<td>5–6 years</td>
<td>.186</td>
<td>.079*</td>
<td>.102</td>
<td>.206</td>
<td>.110</td>
<td>.145</td>
<td>.148</td>
<td>.143</td>
</tr>
<tr>
<td>7–8 years</td>
<td>.177</td>
<td>.086*</td>
<td>.109</td>
<td>.230</td>
<td>.120</td>
<td>.153</td>
<td>.155</td>
<td>.146</td>
</tr>
<tr>
<td>9–10 years</td>
<td>-.037</td>
<td>.093</td>
<td>.111</td>
<td>.241</td>
<td>.120</td>
<td>.156</td>
<td>.164</td>
<td>.154</td>
</tr>
<tr>
<td>11–12 years</td>
<td>-.247</td>
<td>.100*</td>
<td>.128</td>
<td>.268</td>
<td>.141</td>
<td>.179</td>
<td>.196</td>
<td>.183</td>
</tr>
<tr>
<td>13–14 years</td>
<td>-.386</td>
<td>.108*</td>
<td>.137*</td>
<td>.296</td>
<td>.164*</td>
<td>.193*</td>
<td>.218</td>
<td>.209</td>
</tr>
<tr>
<td>15+ years</td>
<td>-.414</td>
<td>.120*</td>
<td>.158*</td>
<td>.337</td>
<td>.186*</td>
<td>.221</td>
<td>.251</td>
<td>.243</td>
</tr>
</tbody>
</table>

**Note:** Standard errors with asterisks indicate significance at 5% level using $N(0,1)$ critical values; $se_{OLS}$ and $se_W$ refer to OLS and White standard errors respectively; $se_{CX}$ and $se_{CT}$ are clustered standard errors suggested by Arellano (1987); $se_{HAC}$ and $se_{DK}$ are two types of Newey-West HAC estimator as explained in the text; $se_{Hard}$ is our standard error. Bartlett kernel with lag length $L = 3$ is used for $se_{HAC}$, $se_{DK}$ and $se_{Hard}$. The threshold value for $se_{Hard}$ by the cross-validation is $M = 0.2$ (for the first panel) and $M = 0.1$ (for the second panel).
Chapter 3

Feasible Generalized Least Squares for Panel Data with Cross-Sectional and Serial Correlations

Note: This chapter is co-authored with Jushan Bai and Yuan Liao and is published in Empirical Economics (Bai et al., 2021).
3.1 Introduction

Heteroskedasticity, cross-sectional and serial correlations are important problems in the error terms of panel regression models. There are two approaches to deal with these problems. The first approach is to use the ordinary least squares (OLS) estimator but with a robust standard error that is robust to heteroskedasticity and correlations, for example, White (1980); Newey and West (1987); Liang and Zeger (1986); Arellano (1987); Driscoll and Kraay (1998); Hansen (2007a); Vogelsang (2012), among others. A widely used class of robust standard errors are clustered standard errors, for example, Petersen (2009), Wooldridge (2010) and Cameron and Miller (2015). Bai et al. (2020) proposed a robust standard error with unknown clusters. In an interesting paper by Abadie et al. (2017), they argued for caution in the application of clustered standard errors since they may give rise to conservative confidence intervals. The second approach is to use the generalized least squares estimator (GLS) that directly takes into account heteroskedasticity, and cross-sectional and serial correlations in the estimation. It is well known that GLS is more efficient than OLS.

This paper focuses on the second approach. For panel models, the underlying covariance matrix involves a large number of parameters. It is important to make GLS operational. We thus consider feasible generalized least squares (FGLS). Hansen (2007b) studied FGLS estimation that takes into account serial correlation and clustering problems in fixed effects panel and multilevel models. His approach requires the cluster structure to be known. This gives motivation to our paper. We assume the unknown cluster structure, and control heteroskedasticity, both serial and cross-sectional correlations by estimating the large error covariance matrix consistently. In cross-sectional setting, Romano and Wolf (2017) obtained asymptotically valid inference of the FGLS estimator, combined with heteroskedasticity-consistent standard errors without knowledge of the conditional heteroskedasticity functional form. Moreover, Miller and Startz (2018) adapted machine learning methods (i.e., support vector regression) to take into account the misspecified form of heteroskedasticity.

In this paper, we consider (i) balanced panel data, (ii) the case of large-$N$ large-$T$, and (iii) both serial and cross-sectional correlations, but unknown structure of clusters. We introduce a modified FGLS estimator that eliminates the cross-sectional and se-
rial correlation bias by proposing a high-dimensional error covariance matrix estimator. In addition, our proposed method is applicable when the knowledge of clusters is not available. Following an idea suggested in Bai and Liao (2017), in this paper, the FGLS involves estimating an \( NT \times NT \) dimensional inverse covariance matrix \( \Omega^{-1} \), where

\[
\Omega = (Eu_t'u_s')
\]

where each block \( Eu_t'u_s' \) is an \( N \times N \) autocovariance matrix. Here parametric structures on the serial or cross-sectional correlations are not imposed. By assuming weak dependencies, we apply nonparametric methods to estimate the covariance matrix. To control the autocorrelation in time series, we employ the idea of Newey-West truncation. This method, in the FGLS setting, is equivalent to “banding”, previously proposed by Bickel and Levina (2008b) for estimating large covariance matrices. We apply it to banding out off-diagonal \( N \times N \) blocks that are far from the diagonal block. In addition, to control for the cross-sectional correlation, we assume that each of the \( N \times N \) block matrices are sparse, potentially resulting from the presence of cross-sectional correlations within clusters. We then estimate them by applying the thresholding approach of Bickel and Levina (2008a). We apply thresholding separately to the \( N \times N \) blocks, which are formed by time lags \( Eu_{t-h}u'_{t-h} : h = 0, 1, 2, ... \). This allows the cluster-membership to be potentially changing over time. A contribution of this paper is the theoretical justification for estimating the large error covariance matrix.

For the FGLS, it is crucial for the asymptotic analysis to prove that the effect of estimating \( \Omega \) is first-order negligible. In the usual low-dimensional settings that involve estimating optimal weight matrix, such as the optimal GMM estimations, it has been well known that consistency for the inverse covariance matrix estimator is sufficient for the first-order asymptotic theory, e.g., Hansen (1982), Newey (1990), Newey and McFadden (1994). However, it turns out that when the covariance matrix is of high-dimensions, not even the optimal convergence rate for estimating \( \Omega^{-1} \) is sufficient. In fact, proving the first-order equivalence between the FGLS and the infeasible GLS (that uses the true \( \Omega^{-1} \)) is a very challenging problem under the large \( N \), large \( T \) setting. We provide a new theoretical argument to achieve this goal.

The banding and thresholding methods, which we employ in this paper, are two
of the useful regularization methods. In the recent machine learning literature, these
methods have been extensively exploited for estimating high-dimensional parameters.
Moreover, in the econometric literature, nonparametric machine learning techniques
have been verified to be powerful tools: Bai and Ng (2017); Chernozhukov et al. (2016,
2017); Wager and Athey (2018), etc.

The rest of the chapter is organized as follows. In Section 3.2, we describe the model
and the large error covariance matrix estimator. Also we introduce the implementation
of FGLS estimator and its limiting distribution. Section 3.3 presents Monte Carlo
studies evaluating the finite sample performance of the estimators. In Section 3.4, we
apply our methods to study the US divorce rate problem. Conclusions are provided in
Section 3.5. All proofs are given in Appendix C.1.

Throughout this chapter, let \( \nu_{\min}(A) \) and \( \nu_{\max}(A) \) denote the minimum and maxi-
num eigenvalues of matrix \( A \) respectively. Also we use \( \|A\| = \sqrt{\nu_{\max}(A'A)} \), \( \|A\|_1 = \max_i \sum_j |A_{ij}| \) and \( \|A\|_F = \sqrt{\text{tr}(A'A)} \) as the operator norm, \( \ell_1 \)-norm and the Frobenius
norm of a matrix \( A \), respectively. Note that if \( A \) is a vector, \( \|A\| = \|A\|_F \) is equal to
the Euclidean norm.

### 3.2 Feasible Generalized Least Squares

We consider a linear model \(^1\)

\[
y_{it} = x_{it}' \beta + u_{it}.
\]

(3.2.1)

The model (3.2.1) can be stacked and represented in full matrix notation as

\[
Y = X \beta + U,
\]

(3.2.2)

where \( Y = (y_1', \ldots, y_T')' \) is the \( NT \times 1 \) vector of \( y_{it} \) with each \( y_t \) being an \( N \times 1 \)
vector; \( X = (x_1', \ldots, x_T')' \) is the \( NT \times d \) matrix of \( x_{it} \) with each \( x_t \) being an \( N \times d \);
\( U = (u_1', \ldots, u_T')' \) is the \( NT \times 1 \) vector of \( u_{it} \) with each \( u_t \) being an \( N \times 1 \) vector.

Let \( \Omega = (Eu_{it}u_{is}') \) be an \( NT \times NT \) matrix, consisting of many “blocks” matricies.
The \((t,s)\)th block is an \( N \times N \) covariance matrix \( Eu_{it}u_{is}' \). We consider the following

\(^1\) For technical simplicity we focus on a simple model where there are no fixed effects. It is straightforward to allow additive fixed effects \( \alpha_i + \mu_t \) by applying the de-meaning first. The theories would be slightly more sophisticated, though such extensions are straightforward.
(infeasible) GLS estimator of $\beta$:

$$
\tilde{\beta}^{\text{inf}}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.
$$

(3.2.3)

Note that $\Omega$ is a high-dimensional conditional covariance matrix, which is very difficult to estimate. We aim to achieve the following: (i) obtain a “good” estimator of $\Omega^{-1}$, allowing an arbitrary form of weak dependence in $u_{it}$, and (ii) show that the effect of replacing $\Omega^{-1}$ by $\hat{\Omega}^{-1}$ is asymptotically negligible.

We start with a population approximation for $\Omega$ in order to gain the intuitions. Then, we suggest the estimator for $\Omega$ that takes into account both correlations problem.

### 3.2.1 Population approximation

We start with a “banding” approximation to control serial correlations. Recall that $\Omega = (Eu_tu'_s)$, where the $(t, s)$ block is $Eu_tu'_s$. By assuming serial stationarity and strong mixing condition, $Eu_tu'_s$ depends on $(t, s)$ only through $h = t - s$. Specifically, with slight abuse of notation, we can write $\Omega_{t,s} = \Omega_{h} = Eu_{t}u'_{t-h}$. Note for $i \neq j$, it is possible that $Eu_{it}u'_{jt-h} \neq Eu_{i,t-h}u_{j,t}$, so $\Omega_{h}$ is possibly non-symmetric for $h > 0$. On the other hand, $\Omega$ is symmetric due to $\Omega_{s,t} = \Omega'_{t,s}$. The diagonal blocks are the same, and all equal $\Omega_0 = Eu_tu'_t$, while magnitudes of the elements of the off-diagonal blocks $\Omega_h = Eu_tu'_{t-h}$ decay to zero as $|h| \to \infty$ under the weak serial dependence assumption.

In the Newey-West spirit, $\Omega$ can be approximated by $\Omega^{NW} = (\Omega^{NW}_{t,s})$, where each block can be written as $\Omega^{NW}_{t,s} = \Omega^{NW}_h$ for $h = t - s$. Here $\Omega^{NW}_h$ is an $N \times N$ block matrix, defined as:

$$
\Omega^{NW}_h = \begin{cases} 
Eu_tu'_{t-h}, & \text{if } |h| \leq L \\
0, & \text{if } |h| > L,
\end{cases}
$$

for some pre-determined $L \to \infty$. For instance, as suggested by Newey and West (1994), we can set $L$ equal to $4(T/100)^{(2/9)}$. Note that $\Omega^{NW}_h = \Omega^{NW}_{-h}$. We regard $\Omega^{NW} = (\Omega^{NW}_h)$ as the “population banding approximation”.

Next, we focus on the $N \times N$ block matrix $\Omega_h = Eu_tu'_{t-h}$ to control cross-sectional correlations. Under the intuition that $u_{it}$ is cross-sectional weakly dependent, we assume $\Omega_h$ is a sparse matrix, that is, $\Omega_{h,ij} = Eu_{it}u_{jt-h}$ is “small” for “many” pairs $(i, j)$. Then $\Omega_h$ can be approximated by a sparse matrix $\Omega^{BL}_h = (\Omega^{BL}_{h,ij})_{N \times N}$ (Bickel and Levina,
where

\[ \Omega^{BL}_{h,ij} = \begin{cases} 
E u_{it} u_{jt,-h}, & \text{if } |E u_{it} u_{jt,-h}| > \tau_{ij} \\
0, & \text{if } |E u_{it} u_{jt,-h}| \leq \tau_{ij},
\end{cases} \]

for some pre-determined threshold \( \tau_{ij} \to 0 \). We regard \( \Omega^{BL}_h \) as the “population sparse approximation”.

In summary, we approximate \( \Omega \) by an \( NT \times NT \) matrix \( (\tilde{\Omega}^{NT}_{t,s}) \), where each block \( \tilde{\Omega}^{NT}_{t,s} \) is an \( N \times N \) matrix, defined as: for \( h = t - s \),

\[
\tilde{\Omega}^{NT}_{t,s} := \begin{cases} 
\Omega^{BL}_h, & \text{if } |h| \leq L \\
0, & \text{if } |h| > L.
\end{cases}
\]

Therefore, we use “banding” to control the serial correlation, and “sparsity” to control the cross-sectional correlation. Note that an advantage of the method proposed in this paper is that it does not assume known cluster information (i.e., the number of clusters and the membership of clusters). Moreover, this method could also be modified to take into account the clustering information when available.

### 3.2.2 Implementation of Feasible GLS

**The estimator of \( \Omega \) and FGLS**

Given the intuition of the population approximation, we construct the large covariance estimator as follows. First, we denote the OLS estimator of \( \beta \) by \( \hat{\beta}_{OLS} \) and the corresponding residuals by \( \hat{u}_{it} = y_{it} - x_{it}' \hat{\beta}_{OLS} \).

Now we estimate the \( N \times N \) block matrix \( \Omega_h = E u_{it} u_{jt,-h}' \). To do so, let

\[
\tilde{R}_{h,ij} = \begin{cases} 
\frac{1}{T} \sum_{t=h+1}^{T} \hat{u}_{it} \hat{u}_{jt,-h}, & \text{if } h \geq 0 \\
\frac{1}{T} \sum_{t=1}^{T-h} \hat{u}_{it} \hat{u}_{jt,-h}, & \text{if } h < 0
\end{cases}, \quad \text{and} \quad \tilde{\sigma}_{h,ij} = \begin{cases} 
\tilde{R}_{h,ii}, & \text{if } i = j \\
s_{ij}(\tilde{R}_{h,ij}), & \text{if } i \neq j
\end{cases},
\]

where \( s_{ij}() : \mathbb{R} \to \mathbb{R} \) is a “soft-thresholding function” with an entry dependent threshold \( \tau_{ij} \) such that

\[ s_{ij}(z) = \text{sgn}(z)(|z| - \tau_{ij})_+ , \]

where \((x)_+ = x \) if \( x \geq 0 \), and zero otherwise. Here \( \text{sgn}(\cdot) \) denotes the sign function, and
other thresholding functions, e.g., hard thresholding, are possible. For the threshold value, we specify

$$\tau_{ij} = M \gamma_T \sqrt{|\tilde{R}_{0,ii}| |\tilde{R}_{0,jj}|},$$

for some pre-determined value $M > 0$, where $\gamma_T = \sqrt{\frac{\log(LN)}{T}}$ is such that

$$\max_{h \leq L} \max_{i,j \leq N} |\tilde{R}_{h,ij} - E_u u_{i,t-h}| = O_P(\gamma_T).$$

Note that the constant thresholding parameter could be allowed as Bickel and Levina (2008a). In practice, however, it is more desirable to have entry dependent threshold, $\tau_{ij}$. $M$ can be chosen by multifold cross-validation. Then define

$$\tilde{\Omega}_h = (\tilde{\sigma}_{h,ij})_{N \times N}. \quad (3.2.4)$$

Next, we define the $(t,s)$th block $\hat{\Omega}_{t,s}$ as an $N \times N$ matrix: for $h = t - s$,

$$\hat{\Omega}_{t,s} = \begin{cases} 
\omega(|h|, L) \hat{\Omega}_h, & \text{if } |h| \leq L \\
0, & \text{if } |h| > L.
\end{cases}$$

Here $\omega(h, L)$ is the kernel function (see Andrews, 1991 and Newey and West, 1994). We let $\omega(h, L) = 1 - h/(L + 1)$ be the Bartlett kernel function, where $L$ is the bandwidth. Our final estimator of $\Omega$ is an $NT \times NT$ matrix:

$$\hat{\Omega} = (\hat{\Omega}_{t,s}).$$

Here $\hat{\Omega}$ is a nonparametric estimator, which does not require an assumed parametric structure on $\Omega$. Note that, for the large sample size, the proposed estimator may require a huge computational cost due to use of an $NT \times NT$ matrix.

Finally, given $\hat{\Omega}$, we propose the feasible GLS (FGLS) estimator of $\beta$ as

$$\hat{\beta}_{FGLS} = [X'\hat{\Omega}^{-1}X]^{-1}X'\hat{\Omega}^{-1}Y.$$ 

**Remark 3.2.1** (Universal thresholding). We apply thresholding separately to the $N \times N$ blocks, $(\tilde{\sigma}_{h,ij})_{N \times N}$, which are estimated lagged blocks for $E_u u_{i,t-h} : h = 0,1,2,\ldots$. 
This allows the cluster-membership to be potentially changing over-time, that is, the identities of zeros and nonzero elements of $E u_{t} u_{t-h}$ can change over $h$. If it is known that the cluster-membership (i.e., identities of nonzero elements) is time-invariant, then one would set $\tilde{\sigma}_{h,ij} = 0$ if $\max_{h \leq L} |\tilde{R}_{h,ij}| \leq \tau_{ij}$ for $i \neq j$. This potentially would increase the finite sample accuracy of identifying the cluster-membership.

**Choice of tuning parameters**

Our suggested covariance matrix estimator, $\hat{\Omega}$, requires the choice of tuning parameters $L$ and $M$, which are the bandwidth and the threshold constant respectively. We write $\hat{\Omega}(M, L) = \hat{\Omega}$, where the covariance estimator depends on $M$ and $L$. First, to choose the bandwidth $L$, we suggest using $L^* = 4(T/100)^{2/9}$, which is proposed by Newey and West (1994). For a small size of $T$, we also recommend $L \leq 3$.

The thresholding constant, $M$, can be chosen through multifold cross-validation. We randomly split the data $P$ times. We divide the data into $P = \log(T)$ blocks $J_1, \ldots, J_P$ with block length $T/\log(T)$ and take one of the $P$ blocks as the validation set. At the $p$th split, we denote by $\hat{\Omega}_0^p$ the sample covariance matrix based on the validation set, defined by $\hat{\Omega}_0^p = \frac{1}{|J_p| - 1} \sum_{t \in J_p} \tilde{u}_t \tilde{u}_t'$. Let $\hat{\Omega}_0^{sp}(M)$ be the thresholding estimator with threshold constant $M$ using the training data set $\{\tilde{u}_t\}_{t \notin J_p}$. Finally, we choose the constant $M^*$ by minimizing the cross-validation objective function

$$M^* = \arg \min_{c < M < C} \frac{1}{P} \sum_{j=1}^{P} \|\tilde{\hat{\Omega}}_0^{sp}(M) - \hat{\Omega}_0^p\|_F^2,$$

where $C$ is a large constant such that $\tilde{\hat{\Omega}}_0^p(C)$ is a diagonal matrix, and $c$ is a constant that guarantees the positive definiteness of $\hat{\Omega}(M, L)$ for $M > c$: for each fixed $L$,

$$c = \inf\{M > 0 : \lambda_{\min}\{\hat{\Omega}(C, L)\} > 0, \forall C > M\}.$$

Here $\tilde{\hat{\Omega}}_0^{sp}(M)$ is the soft-thresholded estimator as defined in the equation (3.2.4). Then the resulting estimator of $\Omega$ is $\hat{\Omega}(M^*, L^*)$. 
3.2.3 The effect of $\hat{\Omega}^{-1} - \Omega^{-1}$

A key step of proving the asymptotic property for $\hat{\beta}_{FGLS}$ is to show that it is asymptotically equivalent to $\hat{\beta}_{GLS}^{inf}$, that is:

$$
\frac{1}{\sqrt{NT}}X' (\hat{\Omega}^{-1} - \Omega^{-1})U = o_P(1). \quad (3.2.5)
$$

In the usual low-dimensional settings that involve estimating optimal weight matrix, such as the optimal GMM estimations, it has been well known that consistency for the inverse covariance matrix estimator is sufficient for the first-order asymptotic theory, e.g., Hansen (1982), Newey (1990), Newey and McFadden (1994). It turns out, when the covariance matrix is of high-dimensions, not even the optimal convergence rate of $\|\hat{\Omega} - \Omega\|$ is sufficient. In fact, proving equation (3.2.5) is a very challenging problem. In the general case when both cross-sectional and serial correlations are present, our strategy is to use a careful expansion for $\frac{1}{\sqrt{NT}}X' (\hat{\Omega}^{-1} - \Omega^{-1})U$. We shall proceed in two steps:

Step 1: Show that $\frac{1}{\sqrt{NT}}X' (\hat{\Omega}^{-1} - \Omega^{-1})U = \frac{1}{\sqrt{NT}}W'(\hat{\Omega} - \Omega)\varepsilon + o_P(1)$, where $W = \Omega^{-1}X$, and $\varepsilon = \Omega^{-1}U$.

Step 2: Show that $\frac{1}{\sqrt{NT}}W'(\hat{\Omega} - \Omega)\varepsilon = o_P(1)$.

Now we suppose $\omega(h, L) = 1, \Omega \approx \Omega^{NW}$ and let $A_{bh} = \{(i, j) : |Eu_{it}u_{j,t-h}| \neq 0\}, A_{sh} = \{(i, j) : |Eu_{it}u_{j,t-h}| = 0\}$. As for Step 2, we shall show,

$$
\frac{1}{\sqrt{NT}}W'(\hat{\Omega} - \Omega)\varepsilon \approx \frac{1}{\sqrt{NT}} \sum_{|h| \leq L} \sum_{i, j \in A_{bh}} \sum_{t = h+1}^{T} w_{it}\varepsilon_{j,t-h} \frac{1}{T} \sum_{s = h+1}^{T} (u_{is}u_{j,s-h} - Eu_{it}u_{j,t-h}). \quad (3.2.6)
$$

Here $w_{it}$ is defined such that, we can write $W = (w_{1}', \ldots, w_{T}')'$ with $w_{t}$ being an $N \times d$ matrix of $w_{it}$; $\varepsilon_{it}$ is defined similarly. We then further argue that the right hand side of (3.2.6) is $o_P(1)$ by applying a high-level Assumption 3.2.4, which essentially saying the right hand side of (3.2.6) is $o_P(1)$.

To appreciate the need of this high-level condition, let us consider a simple example as follows.

**A simple example.** To illustrate the key technical issue, consider a simple and ideal case where $u_{it}$ is known, and independent across both $i$ and $t$, but with cross-
sectional heteroskedasticity. In this case, the covariance matrix of the $NT \times 1$ vector $U$ is a diagonal matrix, with diagonal elements $\sigma_i^2 = Eu_{it}^2$.

$$
\Omega = \begin{pmatrix} D & & \\ & D & \\ & & \ddots \\ & & & D \end{pmatrix}, \quad \text{where} \quad D = \begin{pmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_N^2 \end{pmatrix}.
$$

Then a natural estimator for $\Omega$ is

$$
\hat{\Omega} = \begin{pmatrix} \hat{D} & & \\ & \hat{D} & \\ & & \ddots \\ & & & \hat{D} \end{pmatrix}, \quad \text{where} \quad \hat{D} = \begin{pmatrix} \hat{\sigma}_1^2 & & \\ & \hat{\sigma}_2^2 & \\ & & \ddots \\ & & & \hat{\sigma}_N^2 \end{pmatrix},
$$

and $\hat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^{T} u_{it}^2$, because $u_{it}$ is known. Then the GLS becomes:

$$
\left( \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}x_{it}' \hat{\sigma}_i^{-2} \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}y_{it} \hat{\sigma}_i^{-2}.
$$

A key step is to prove that the effect of estimating $D$ is asymptotically negligible:

$$
\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}u_{it}(\hat{\sigma}_i^{-2} - \sigma_i^{-2}) = o_P(1).
$$

It can be shown that the problem reduces to proving:

$$
A \equiv \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}u_{it} \sigma_i^{-2} \left( \frac{1}{T} \sum_{s=1}^{T} (u_{is}^2 - Eu_{is}^2) \right) \sigma_i^{-2} = o_P(1).
$$

In fact, straightforward calculations yield

$$
EA = \frac{\sqrt{NT}}{T} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E(x_{it}E(u_{it}^3|x_{it})) \sigma_i^{-4}.
$$

Generally, if $u_{it}|x_{it}$ is non-Gaussian and asymmetric, $E(u_{it}^3|x_{it}) \neq 0$. Hence we require $N/T \to 0$ to have $EA \to 0$. Hence, to allow for non-Gaussian and asymmetric conditional distributions, in the GLS setting it turns out $N = o(T)$ is required.

We shall not explicitly impose $N = o(T)$ in this paper as a formal assumption,
but instead impose Assumption 3.2.4. On one hand, when the distribution of \( u_{it} \) is symmetric, we do not require \( N = o(T) \) because as is shown in the above example, \( E(u_{it}^3|x_{it}) = 0 \) is sufficient and holds for symmetric distributions. On the other hand, when \( u_{it} \) is non-symmetric, Assumption 3.2.4 then implicitly requires \( N = o(T) \). Note that \( N = o(T) \) is a strong assumption in many microeconomic applications for panel data models. But as illustrated in the above simple example, if \( u_{it}|x_{it} \) is not symmetric, it is required for feasible GLS even if \( \Omega \) is diagonal. One possible approach to weakening this assumption is to remove the higher order bias from \( \hat{\Omega} \). Higher order debiasing is a complicated procedure in the presence of general weak dependences. This is left for future research.

3.2.4 Asymptotic results of FGLS

We impose the following conditions, regulating the sparsity and serial weak dependence.

**Assumption 3.2.1.**

(i) \( \{u_t, x_t\}_{t \geq 1} \) is strictly stationary. In addition, each \( u_t \) has zero mean vector, and \( \{u_t\}_{t \geq 1} \) and \( \{x_t\}_{t \geq 1} \) are independent.

(ii) There are constants \( c_1, c_2 > 0 \) such that \( \lambda_{\min}(\Omega_h) > c_1 \) and \( \|\Omega_h\|_1 < c_2 \) for each fixed \( h \).

(iii) Exponential tail: There exist \( r_1, r_2 > 0 \) and \( b_1, b_2 > 0 \), and for any \( s > 0 \), \( i \leq N \) and \( l \leq d \),

\[
P(|u_{it}| > s) \leq \exp(-s/b_1^{r_1}), \quad P(|x_{it,l}| > s) \leq \exp(-s/b_2^{r_2}).
\]

(iv) Strong mixing: There exist \( \kappa \in (0, 1) \) such that \( r_1^{-1} + r_2^{-1} + \kappa^{-1} > 1 \), and \( C > 0 \) such that for all \( T > 0 \),

\[
\sup_{A \in \mathcal{F}_0^0, B \in \mathcal{F}_T^\infty} |P(A)P(B) - P(AB)| < \exp(-CT^\kappa),
\]

where \( \mathcal{F}_0^0 \) and \( \mathcal{F}_T^\infty \) denote the \( \sigma \)-algebras generated by \( \{(x_t, u_t) : t \leq 0\} \) and \( \{(x_t, u_t) : t \geq T\} \) respectively.

Condition (ii) requires that \( \Omega_h \) be well conditioned. Condition (iii) ensures the Bernstein-type inequality for weakly dependent data, which requires the underlying distributions to be thin-tailed. Condition (iv) is the standard \( \alpha \)-mixing condition, adapted
Assumption 3.2.2. (i) There exists a constant $C > 0$ such that for all $i \leq N$ and $t \leq T$, $E\|x_{it}\|^4 < C$ and $Eu_{it}^4 < C$.

(ii) Define $\xi_T(L) = \max_{t \leq T} \sum_{|h| > L} \|Eu_{t-h}u_t\|$. Then $\xi_T(L) \to 0$.

(iii) Define $f_T(L) = \max_{t \leq T} \sum_{|h| \leq L} \|Eu_{t-h}(1 - \omega(|h|, L))\|$. Then $f_T(L) \to 0$.

Assumption 3.2.2 allows us to prove the convergence rate of the covariance matrix estimator. Condition (ii) is an extension of the standard weak serial dependence condition to the high-dimensional case in panel data literature. It allows us to employ banding or Newey-West truncation procedure. Condition (iii) is well satisfied by various kernel functions for the HAC-type estimator. For the Bartlett kernel, for example,

$$\max_{t \leq T} \sum_{|h| \leq L} \|Eu_{t-h}(1 - \omega(|h|, L))\| \leq \frac{1}{L} \max_{|h| = 0} \sum_{i \leq T} \|Eu_{t-h}\| |h|$$

converges to zero as $L \to \infty$ as long as $\max_{t \leq T} \sum_{|h| = 0} \|Eu_{t-h}\| |h| < \infty$.

In this paper, we assume $\Omega_h$ to be a sparse matrix for each $h$ and impose similar conditions as those in Bickel and Levina (2008a) and Fan et al. (2013): write $\Omega_h = (\Omega_{h,ij})_{N \times N}$, where $\Omega_{h,ij} = Eu_{it}u_{j,t-h}$. For some $q \in [0, 1)$, we define

$$m_N = \max_{|h| \leq L} \max_{i \leq N} \sum_{j=1}^{N} |\Omega_{h,ij}|^q,$$

as a measurement of the sparsity. We would require that $m_N$ should be either fixed or grow slowly as $N \to \infty$. In particular, when $q = 0$, $m_N = \max_{|h| \leq L} \max_{i \leq N} \sum_{j=1}^{N} 1(\Omega_{h,ij} \neq 0)$, which corresponds to the exact sparsity case.

Let

$$\gamma_T = \sqrt{\log(LN)/T}.$$

The following theorem shows the convergence rate of the estimated large covariance matrix. For technical simplicity, we assume that there is no fixed effects so that we do not take the de-meaning procedure. Extending to the more complete estimators with de-meaning is straightforward, but should require more technical arguments to show that the effect from added dependences due to the de-meaning is negligible.

Theorem 3.2.1. Under the Assumptions 3.2.1-3.2.2, when $\|\Omega^{-1}\|_1 = O(1)$, for $q \in$
\[0,1\) such that \(Lm_N\gamma_T^{1-q} = o(1)\), and for \(\xi_T\) defined in Assumption 3.2.2,
\[
\|\hat{\Omega} - \Omega\| = O_P(Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L)) = \|\hat{\Omega}^{-1} - \Omega^{-1}\|.
\]

The following conditions are required to prove Step 1 in the previous section.

**Assumption 3.2.3.** For any \(NT \times NT\) matrix \(M\), we denote \((M)_{ts,ij}\) as the \((i,j)\)th element of the \((t,s)\)th block of the matrix \(M\).

(i) \(\sum_{|h|>L} \|\Omega_h\|_1 = O(L^{-\alpha})\), for a constant \(\alpha > 0\).

(ii) \(\max_{i \leq N, t \leq T} \sum_{s=1}^{T} \sum_{j=1}^{N} |(\Omega^{-1})_{ts,ij}| = O(1)\).

(iii) There is \(q \in [0,1)\) such that \(Lm_N\gamma_T^{1-q} = o(1)\) holds. In addition,
\[
\sqrt{NT}L^2m_N^2\gamma_T^{3-2q} = o(1), \quad L^{-\alpha}T\sqrt{NT}m_N\gamma_T^{1-q} = o(1).
\]

(iv) Define \(\gamma^* = Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L)\). Then \(\sqrt{NT}\gamma^{*3} = o(1)\).

Conditions (i)-(ii) require the weak cross-sectional correlations. Conditions (iii)-(iv) are the sparsity assumptions. In addition, the sparsity assumptions assume that \(m_N\) should not be too large.

**Remark 3.2.2.** To understand Assumption 3.2.3, consider a simple case where \(Eu_{it}u_{j,t-h}\) is nonzero for only finitely many pairs \(i \neq j\). This corresponds to \(q = 0\) and \(m_N = O(1)\). Then conditions (iii) and (iv) require
\[
\sqrt{N}L^3\log^{3/2}(NT) = o(T)
\]
\[
L^{-\alpha}T^{3/2}N^{1/2}\gamma_T = o(1), \quad \sqrt{NT}(\xi_T(L) + f_T(L))^3 = o(1)
\]

**Proposition 3.2.1.** Under the Assumption 3.2.1-3.2.2, for \(q \in [0,1)\) and \(\alpha > 0\) such that Assumption 3.2.3 holds,
\[
\sqrt{NT}(\hat{\beta}_{FGLS} - \beta) = \Gamma^{-1} \left( \frac{1}{\sqrt{NT}} X'\Omega^{-1}U \right) + \Gamma^{-1} \left( \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\hat{\Omega} - \Omega)\Omega^{-1}U \right) + o_P(1),
\]
where \(\Gamma = E(X'\Omega^{-1}X/NT)\).

In addition, we impose the following assumption, which allows us to prove that the
second term on the right hand side in the above equation is \( o_P(1) \).

**Assumption 3.2.4.** Let \( A_{bh} = \{(i, j) : |Eu_{it}u_{j,t-h}| \neq 0 \} \). Then

\[
\frac{1}{\sqrt{NT}} \sum_{h=0}^{L} \sum_{i,j \in A_{bh}} G_{1,ij}(h)G_{2,ij}(h) = o_P(1),
\]

where \( G_{1,ij}(h) = \frac{1}{\sqrt{T}} \sum_{t=h+1}^{T}(u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h}) \) and \( G_{2,ij}(h) = \frac{1}{\sqrt{T}} \sum_{t=h+1}^{T} w_{it}e_{j,t-h} \).

We have the following limiting distribution by using the result of Theorem 3.2.1.

**Theorem 3.2.2.** Suppose \( \text{var}(U|X) = \text{var}(U) = \Omega \). Under the Assumptions 3.2.1-3.2.4, for \( q \in [0, 1) \) and \( \alpha > 0 \) such that Assumption 3.2.3 (i) holds, as \( N,T \to \infty \),

\[
\sqrt{NT}(\hat{\beta}_{FGLS} - \beta) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1}),
\]

where \( \Gamma = E(X'\Omega^{-1}X/NT) \). The consistent estimator of \( \Gamma \) is \( \hat{\Gamma} = X'\hat{\Omega}^{-1}X/NT \).

The asymptotic variance of the FGLS estimator is \( \text{Avar}(\hat{\beta}_{FGLS}) = \Gamma^{-1}/NT \), and an estimator of it is \( (X'\hat{\Omega}^{-1}X)^{-1} \). Asymptotic standard errors can be obtained in the usual fashion from the asymptotic variance estimates.

### 3.3 Monte Carlo evidence

#### 3.3.1 DGP and methods

In this section we compare the proposed FGLS estimator with OLS estimator. We consider the fixed effect linear regression model, although this paper focuses on the simple linear model for technical simplicity. Hence the de-meaning procedure is applied first. The data generating process (DGP) used for the simulations is given by

\[
y_{it} = \alpha_i + \mu_t + \beta_0 x_{it} + u_{it},
\]

where the true \( \beta_0 = 1 \) and fixed effects \( \alpha_i, \mu_t \) are generated from \( \mathcal{N}(0, 0.5) \). The DGP allows for serial and cross-sectional correlation in both \( x_{it} \) and \( u_{it} \), which are generated by \( (NT) \times (NT) \) covariance matrices, \( \Omega_X \) and \( \Omega_U \), as follows: let \( R_\eta = (R_{\eta,ij}) \) denote an \( N \times N \) block diagonal correlation matrix. We fix the number of clusters as \( G = 25 \).
Hence, each diagonal block is a $N/G \times N/G$ matrix with the off-diagonal entries $(i, j)$ in the same cluster, $R_{n_{ij}}$ for $i \neq j$, which are generated from i.i.d. Uniform$(0, \gamma)$. In this study, we set the level of cross-sectional correlation in each cluster as $\gamma = 0.3$, or 0.7. For the cross-sectional heteroskedasticity, let $D = \text{diag}\{d_i\}$, where $\{d_i\}_{1 \leq N}$ are i.i.d. Uniform$(1, m)$. Finally, we define the $N \times N$ covariance matrix of $u_t$ as $\Sigma_u = DR\eta D$. In this case, we report results when $m = \sqrt{5}$. For the covariance matrix of the regressor, we simply set $\Sigma_x = R\eta$, which does not have heteroskedasticity.

Now we introduce $i$-dependent serial correlation for the regressor and the error as follows: first let $\sigma_{ii} = \rho_i$ if $i = j$ and $\sigma_{ij} = \rho_i \rho_j$ if $i \neq j$. Then we define the $(NT) \times (NT)$ covariance matrix, $\Omega_U = (\Omega_{t,s})$. The $(t, s)$th block is an $N \times N$ covariance matrix, given by $\Omega_{t,s} = (\Omega_{t,s}(i, j))$, where $\Omega_{t,s}(i, j) = \Sigma_{u_{ij}} \sigma_{ij}^{[|t-s|]}$. The large covariance matrix of the regressor, $\Omega_X$, is generated similarly. The level of $i$-dependent $\rho_i$ of the regressor and the error is generated from i.i.d. Uniform$(0, 0.6)$, seperately.

Note that the $(t, s)$th block covariance decays exponentially as $|t - s|$ increases.

Finally we generate the $NT \times 1$ vectors $(u'_1, \ldots, u'_T)' = \Omega_U^{1/2} \zeta$, where $\zeta$ is an $N \times 1$ vector, whose entries are generated from i.i.d. $N(0, 5)$. Similarly, the regressor is generated by $(x'_1, \ldots, x'_T)' = \Omega_X^{1/2} \xi$, where $\xi$ is an $NT \times 1$ vector, whose entries are generated from i.i.d. $N(0, 1)$. Note that $x_{it}$ is uncorrelated with $u_{it}$.

In this numerical study, we use sample sizes $N = 50, 100$ and $T = 50, 100, 150$, and the simulation is replicated for one thousand times in all cases. For each $\{N, T\}$ combination, we set the bandwidth $L = 3$ in all cases. The threshold constant, $M$, is obtained by the cross-validation method as suggested in Section 3.2.2. For instance, when $T = 100$, the number of folds to split is $\log(100) \approx 5$. In general, the cross-validation chooses $M$ between 1.4 and 1.8. Interestingly, as the level of cross-section correlation increases, the cross-validation tends to choose smaller $M$, so that the number of non-thresholded elements increases. Hence it takes into account the strength of cross-sectional correlation. We use the Bartlett kernel for our FGLS estimator. Results are summarized in Tables 3.1-3.2.

\[2\] The procedure of proposed estimators require use of an $NT \times NT$ matrix as discussed in Section 3.2.2. Indeed, when $NT$ is large, the procedure appears to be computationally demanding. Hence, we focus on the small sample size in this study.
3.3.2 Results

Tables 3.1-3.2 present the simulation results, where each table corresponds to a different level of cross-sectional correlation, \( \gamma = \{0.3, 0.7\} \). In each table, the mean and standard deviation of the estimators are reported. FGLS(Diag) refers to the FGLS estimator using the diagonal covariance matrix, which only takes into account heteroskedasticity. RMSE is the ratio of the mean squared error of FGLS to that of OLS. The mean and standard deviation of the estimated standard errors for OLS and FGLS are also reported. The robust unknown clustered standard error, suggested by Bai et al. (2020), is used for OLS. For FGLS, we report the results of the standard error as introduced in Theorem 3.2.2. The difference between the standard deviation of the estimators and the mean of standard errors can be explained as the bias of estimated standard errors. In addition, we present null rejection probabilities for the 5% level tests using the traditional \( N(0, 1) \) critical value based on each standard errors.

According to Tables 3.1-3.2, we see that both methods are almost unbiased, while our proposed FGLS has indeed smaller standard deviation of \( \hat{\beta} \) than that of OLS and FGLS(Diag). In all cases, the RMSE of our proposed FGLS is significantly smaller than one. Hence the results confirm that the FGLS estimator is more efficient than the OLS and the FGLS(Diag) estimators in presence of heteroskedasticity, serial and cross-sectional correlations. Regarding the \( t \)-test, in Table 3.1, the rejection probabilities of FGLS and OLS are close to 0.05 when \( T \) is large, while those of FGLS(Diag) tend to over-reject. Since the FGLS(Diag) estimator does not take into account the serial and the cross-sectional correlations, its standard errors are underestimated. On the other hand, in Table 3.2, we find that the standard errors of all estimators are underestimated and the \( t \)-test rejection probabilities are much larger than 0.05, especially when \( T \) is relatively smaller than \( N \) (e.g., \( N = 100 \) and \( T = 50 \)). This is due to the strong cross-sectional correlation within clusters. However, the rejection probabilities of FGLS and OLS are much smaller than those of FGLS(Diag). In summary, FGLS does improve efficiency in terms of mean squared error; also we obtain unbiased standard error estimator and appropriate rejection rate as \( T \) increases.
3.4 Empirical study: Effects of divorce law reforms on divorce rates

In the literature, the cause of the sharp increase in the U.S. divorce rate in the 1960-1970s is an important research question. During 1970s, more than half of states in the U.S. liberalized the divorce system, and the effects of reforms on divorce rates have been investigated by many such as Allen (1992) and Peters (1986). With controls for state and year fixed effects, Friedberg (1998) suggested that state law reforms significantly increased divorce rates. Also, she assumed that unilateral divorce laws affected divorce rates permanently. However, divorce rates from 1975 have been subsequently decreasing according to empirical evidence. Therefore the question of whether law reforms also affect the divorce rate decrease has arisen. Wolfers (2006) revisited this question by using a treatment effect panel data model, and identified only temporal effects of reforms on divorce rates. In particular, he used dummy variables for the first two years after the reforms, 3-4 years, 5-6 years, and so on. More specifically, the following fixed effect panel data model was considered:

$$y_{it} = \alpha_i + \mu_t + \sum_{k=1}^{8} \beta_k X_{it,k} + \delta_i t + u_{it}, \quad (3.4.1)$$

where $y_{it}$ is the divorce rate for state $i$ and year $t$, $\alpha_i$ a state fixed effect, $\mu_t$ a time fixed effects, and $\delta_i t$ a linear time trend with unknown coefficient $\delta_i$. $X_{it}$ is a binary regressor which denotes the treatment effect $2k$ years after the reform. Wolfers (2006) suggested that “the divorce rate rose sharply following the adoption of unilateral divorce laws, but this rise was reversed within about a decade”. He also concluded that “15 years after reform the divorce rate is lower as a result of the adoption of unilateral divorce, although it is hard to draw any strong conclusions about long-run effects”.

Both Friedberg (1998) and Wolfers (2006) used a weighted model by multiplying all variables by the square root of state population. In addition, they used ordinary OLS standard error, which does not take into account heteroskedasticity, serial and cross-sectional correlations. However, standard errors might be biased when one disregards these correlations. Therefore, we re-estimated the model of Wolfers (2006) using the proposed FGLS method and OLS with the heteroskedastic standard errors of White.
(1980), the clustered standard error of Arellano (1987), and the robust standard error of Bai et al. (2020).

The same dataset as in Wolfers (2006) is used, which includes the divorce rate, state-level reform years, binary regressors, and state population. Due to missing observations around divorce law reforms, we exclude Indiana, New Mexico and Louisiana. As a result, we obtain balanced panel data from 1956 to 1988 for 48 states. We fit the models both with and without linear time trend, and use OLS and FGLS in each model to estimate $\beta$. In the FGLS estimation, we set bandwidth $L = 3$ as proposed by Newey and West (1994). The thresholding values are chosen by the cross-validation method as discussed in Section 3.2.2, more specifically, $M = 1.8$ and $M = 1.9$ for the model with and without linear time trends, respectively. The Bartlett kernel is used in the OLS robust standard error and FGLS estimation. The estimated $\beta_1, \cdots, \beta_8$ with and without linear time trend and standard errors are summarized in Table 3.3 below.

The OLS and FGLS estimates in both models are similar to each other. The results show that divorce rates rose soon after the law reform. However, within a decade, divorce rates had fallen over time. Interestingly, FGLS confirms the negative effects of the law reforms on the divorce rates, specifically, 11-15+ years after the reform in the model with state-specific linear time trends, and 9-15+ years after the reform in the model without state-specific linear time trends. In addition, the FGLS estimates for 1-6 and 1-4 years are positive and statistically significant in the models with and without linear time trends, respectively. For OLS, the coefficient estimates for 3-4 and 7-15+ are significant in the model without linear time trends based on $se_{BCL}$. In contrast, the OLS estimates are statistically significant only for 1-4 years when a linear time trend is added. According to the clustered standard error, $se_{CX}$, note that only 11-15+ are statistically significant in the model without trends.

According to OLS and FGLS estimation results with and without a linear time trend, we make the following conlcusion: in the first 8 years, the overall trend of divorce rate is increasing, but the law reform reduces the divorce rate after 3-4 years. However, 8 years after the reform, we observe that the law reform has a negative effect on divorce rate. Note that Wolfers (2006) de-emphasized the negative coefficient at the end of the periods, as these are not robust to inclusion of state-specific quadratic trends, which we did not employ in this paper. Overall, the results of FGLS estimates are consistent with
3.5 Conclusion

In this paper, we propose a large covariance matrix estimator and a modified version of FGLS that takes into account both serial and cross-sectional correlations in linear panel models that are robust to heteroskedasticity, serial and cross-sectional correlations. The covariance matrix estimator is asymptotically unbiased with an improved convergence rate. It is shown to be more efficient than other existing methods in panel data literature. From simulated experiments, we confirmed that our FGLS estimates are more efficient than OLS estimates.
Table 3.1: Performance of estimated $\beta_0$; true $\beta_0 = 1$; $i$-dependent serial correlation and weak cross-sectional correlation ($\gamma = 0.3$).

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>OLS Diag</th>
<th>OLS Our</th>
<th>FGLS Diag</th>
<th>FGLS Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>1.001</td>
<td>0.080</td>
<td>1.002</td>
<td>0.075</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>1.000</td>
<td>0.061</td>
<td>1.000</td>
<td>0.055</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>1.000</td>
<td>0.045</td>
<td>1.000</td>
<td>0.041</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>1.000</td>
<td>0.058</td>
<td>1.000</td>
<td>0.053</td>
</tr>
<tr>
<td>100</td>
<td>150</td>
<td>1.000</td>
<td>0.041</td>
<td>1.000</td>
<td>0.037</td>
</tr>
<tr>
<td>150</td>
<td>150</td>
<td>1.000</td>
<td>0.034</td>
<td>1.000</td>
<td>0.029</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>mean(s.e.)</th>
<th>std(s.e.)</th>
<th>t-test rejection prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 0.079</td>
<td>0.004</td>
<td>0.054</td>
</tr>
<tr>
<td>100 0.058</td>
<td>0.002</td>
<td>0.052</td>
</tr>
<tr>
<td>150 0.047</td>
<td>0.001</td>
<td>0.047</td>
</tr>
<tr>
<td>100 0.058</td>
<td>0.002</td>
<td>0.058</td>
</tr>
<tr>
<td>100 0.040</td>
<td>0.001</td>
<td>0.053</td>
</tr>
<tr>
<td>150 0.032</td>
<td>0.001</td>
<td>0.069</td>
</tr>
</tbody>
</table>

**Note:** OLS and FGLS comparison. RMSE is the ratio of the mean squared error of FGLS to that of OLS. The t-test rejection prob. is t-test rejection rates for 5% level tests. Robust standard error suggested by Bai et al. (2020) is used for OLS. Reported results are based on 1000 replications. The threshold value, $M$, is chosen through the cross-validation method as discussed in Section 3.2.2. For the bandwidth, we set $L = 3$.

Table 3.2: Performance of estimated $\beta_0$; true $\beta_0 = 1$; $i$-dependent serial correlation and strong cross-sectional correlation ($\gamma = 0.7$).

<table>
<thead>
<tr>
<th>N</th>
<th>T</th>
<th>OLS Diag</th>
<th>OLS Our</th>
<th>FGLS Diag</th>
<th>FGLS Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.084</td>
<td>1.001</td>
<td>0.079</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>1.000</td>
<td>0.063</td>
<td>1.000</td>
<td>0.057</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>1.000</td>
<td>0.051</td>
<td>1.000</td>
<td>0.047</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>1.000</td>
<td>0.070</td>
<td>1.000</td>
<td>0.063</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>1.000</td>
<td>0.048</td>
<td>0.999</td>
<td>0.044</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>0.079</td>
<td>0.004</td>
<td>0.066</td>
<td>0.002</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.059</td>
<td>0.002</td>
<td>0.048</td>
<td>0.001</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>0.047</td>
<td>0.001</td>
<td>0.039</td>
<td>0.001</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.059</td>
<td>0.002</td>
<td>0.048</td>
<td>0.001</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>0.040</td>
<td>0.000</td>
<td>0.034</td>
<td>0.000</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>0.032</td>
<td>0.000</td>
<td>0.027</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Note:** See notes to Table 3.1.
Table 3.3: Empirical application: effects of divorce law reform with state and year fixed effects: US state level data annual from 1956 to 1988, dependent variable is divorce rate per 1000 persons per year. OLS and FGLS estimates and standard errors (using state population weights).

<table>
<thead>
<tr>
<th>Effects:</th>
<th>(\hat{\beta}_{OLS})</th>
<th>(se_{W})</th>
<th>(se_{CX})</th>
<th>(se_{BCL})</th>
<th>(\hat{\beta}_{FGLS})</th>
<th>(se_{FGLS})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Without state-specific linear time trends</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–2 years</td>
<td>0.256</td>
<td>0.140</td>
<td>0.189</td>
<td>0.148</td>
<td>0.133</td>
<td>0.046*</td>
</tr>
<tr>
<td>3–4 years</td>
<td>0.209</td>
<td>0.081*</td>
<td>0.159</td>
<td>0.089*</td>
<td>0.165</td>
<td>0.056*</td>
</tr>
<tr>
<td>5–6 years</td>
<td>0.126</td>
<td>0.073</td>
<td>0.168</td>
<td>0.069</td>
<td>0.100</td>
<td>0.059</td>
</tr>
<tr>
<td>7–8 years</td>
<td>0.105</td>
<td>0.070</td>
<td>0.165</td>
<td>0.040*</td>
<td>0.026</td>
<td>0.061</td>
</tr>
<tr>
<td>9–10 years</td>
<td>-0.122</td>
<td>0.060*</td>
<td>0.161</td>
<td>0.054*</td>
<td>-0.129</td>
<td>0.061*</td>
</tr>
<tr>
<td>11–12 years</td>
<td>-0.344</td>
<td>0.071*</td>
<td>0.173*</td>
<td>0.075*</td>
<td>-0.253</td>
<td>0.062*</td>
</tr>
<tr>
<td>13–14 years</td>
<td>-0.496</td>
<td>0.074*</td>
<td>0.188*</td>
<td>0.062*</td>
<td>-0.324</td>
<td>0.063*</td>
</tr>
<tr>
<td>15+ years</td>
<td>-0.508</td>
<td>0.089*</td>
<td>0.223*</td>
<td>0.077*</td>
<td>-0.325</td>
<td>0.067*</td>
</tr>
<tr>
<td>Panel B: With state-specific linear time trends</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–2 years</td>
<td>0.286</td>
<td>0.152</td>
<td>0.206</td>
<td>0.140*</td>
<td>0.171</td>
<td>0.044*</td>
</tr>
<tr>
<td>3–4 years</td>
<td>0.254</td>
<td>0.099*</td>
<td>0.171</td>
<td>0.126*</td>
<td>0.220</td>
<td>0.058*</td>
</tr>
<tr>
<td>5–6 years</td>
<td>0.186</td>
<td>0.102</td>
<td>0.206</td>
<td>0.143</td>
<td>0.175</td>
<td>0.067*</td>
</tr>
<tr>
<td>7–8 years</td>
<td>0.177</td>
<td>0.109</td>
<td>0.230</td>
<td>0.146</td>
<td>0.097</td>
<td>0.075</td>
</tr>
<tr>
<td>9–10 years</td>
<td>-0.037</td>
<td>0.111</td>
<td>0.241</td>
<td>0.154</td>
<td>-0.073</td>
<td>0.082</td>
</tr>
<tr>
<td>11–12 years</td>
<td>-0.247</td>
<td>0.128</td>
<td>0.268</td>
<td>0.183</td>
<td>-0.240</td>
<td>0.089*</td>
</tr>
<tr>
<td>13–14 years</td>
<td>-0.386</td>
<td>0.137*</td>
<td>0.295</td>
<td>0.209</td>
<td>-0.329</td>
<td>0.098*</td>
</tr>
<tr>
<td>15+ years</td>
<td>-0.414</td>
<td>0.158*</td>
<td>0.337</td>
<td>0.243</td>
<td>-0.382</td>
<td>0.108*</td>
</tr>
</tbody>
</table>

**Note:** Standard errors with asterisks indicate significance at 5% level using \(N(0,1)\) critical values. For OLS standard errors, \(se_{W}\) and \(se_{CX}\) refer to the heteroskedastic standard errors by White (1980) and the clustered standard errors by Arellano (1987), respectively; \(se_{BCL}\) is the robust standard error suggested by Bai et al. (2020). The threshold values for FGLS by the cross-validation are \(M = 1.9\) and \(M = 1.8\) for Panel A and B, respectively.
Appendix A

Appendix for Chapter 1

A.1 Proof of Theorem 1.3.1

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.1.1.** Consider the conventional factor model. Under the Assumption 1.3.1, 1.3.2-1.3.4, when $\|\Sigma - \Sigma^{-1}u\|_1 = O(1)$,

$$\|\hat{\Sigma}_u - \Sigma_u\|_1 = O_P(m_N\delta_{N,T}^{\frac{1}{2}}) = \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1,$$

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

**Proof.** Let $\hat{U} = Y - \hat{\Lambda}\hat{F}'$, 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 $\max_{i \leq N} \|\hat{\lambda}_i - M'\lambda_i\| = O_P(\sqrt{\frac{J}{T}})$. Then $\max_{i \leq N} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2 = O_P(\frac{1}{N} + \frac{T}{J})$, which also implies

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

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

$$\max_{i \leq N, j \leq N} \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt} - \Sigma_{u,ij} = O_P(\delta_{N,T}).$$
By Theorem 5 of Fan et al. (2013) using sparsity property, we then have \( \|\hat{\Sigma}_u - \Sigma_u\|_1 = O_P(m_N \delta_{N,T}^{1-q}). \) For the second statement, we have
\[
\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 \leq \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 \|\hat{\Sigma}_u - \Sigma_u\|_1 + \|\Sigma_u^{-1}(\hat{\Sigma}_u - \Sigma_u)\Sigma_u^{-1}\|_1
\]
\[
= O_P(m_N \delta_{N,T}^{1-q}) \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 + O_P(m_N \delta_{N,T}^{1-q}).
\]
Therefore, we have \((1 + o_P(1))\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 = O_P(m_N \delta_{N,T}^{1-q}), \) and it implies the results.

A.1.1 Convergence of loadings

Let \( \tilde{Y} = \hat{\Sigma}_u^{-\frac{1}{2}} Y, \tilde{\Lambda} = \hat{\Sigma}_u^{-\frac{1}{2}} \Lambda, \) and \( \tilde{U} = \hat{\Sigma}_u^{-\frac{1}{2}} U. \) Then the regular factor model (1.3.4) can be written as
\[
\tilde{Y} = \tilde{\Lambda} F' + \tilde{U}.
\]

Let \( K \) denote a \( K \times K \) diagonal matrix of the first \( K \) eigenvalues of \( \frac{1}{NT} \tilde{Y} P \tilde{Y}'. \) Then by definition of eigenvalues, we have
\[
\frac{1}{NT} \tilde{Y} P \tilde{Y}' \tilde{\Lambda} = \tilde{\Lambda} K.
\]

Let \( M = \frac{1}{NT} F' P \tilde{\Lambda} \tilde{\Lambda}^{-1} = \tilde{U} F' \tilde{\Lambda} \tilde{\Lambda}^{-1}. \) Then
\[
\hat{\Lambda} - \tilde{\Lambda} M = \sum_{i=1}^{3} A_i K^{-1}, \quad (A.1.1)
\]
where
\[
A_1 = \frac{1}{NT} \tilde{F}' P \tilde{U}' \hat{\Lambda}, \quad A_2 = \frac{1}{NT} \tilde{U} P \tilde{U}' \hat{\Lambda}, \quad A_3 = \frac{1}{NT} \tilde{U} P \tilde{F}' \hat{\Lambda}.
\]

Lemma A.1.1. \( \|K\|_2 = O_P(1) = \|K^{-1}\|_2, \) and \( \|M\|_2 = O_P(1). \)

Proof. Note that \( K \) is the diagonal matrix of the first \( K \) eigenvalue of
\[
\frac{1}{NT} \tilde{Y} P \tilde{Y}' = \frac{1}{NT} \tilde{Y} \Phi(X)(\Phi(X)' \Phi(X))^{-1} \Phi(X)' \tilde{Y}'.
\]
Then the eigenvalues of $K$ are the same as those of

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

By substituting $\tilde{Y} = \tilde{A}F' + \tilde{U}$, and $\frac{1}{N} \tilde{A}' \tilde{A} = I_K$, we have $W = \sum_{i=1}^{4} W_i$, where

$$W_1 = \frac{1}{T} (\Phi(X)'\Phi(X))^{-1/2} \Phi(X)' F F' (\Phi(X)'\Phi(X))^{-1/2},$$
$$W_2 = \frac{1}{T} (\Phi(X)'\Phi(X))^{-1/2} \tilde{F} \tilde{\Lambda}' \tilde{U} (\Phi(X)'\Phi(X))^{-1/2},$$
$$W_3 = \frac{1}{T} (\Phi(X)'\Phi(X))^{-1/2} \tilde{U}' \tilde{\Lambda} F' (\Phi(X)'\Phi(X))^{-1/2} = W_2',$n
$$W_4 = \frac{1}{T} (\Phi(X)'\Phi(X))^{-1/2} \tilde{U}' \tilde{U} \tilde{\Lambda}' F (\Phi(X)'\Phi(X))^{-1/2}.$$

Note that, Assumption 1.3.2 and 1.3.3 implies that $\|\Phi(X)\|_2 = \lambda^1/2_{\text{max}}(\Phi(X)'\Phi(X)) = O_P(\sqrt{T})$, $\|\Phi(X)'\Phi(X))^{-1/2}\|_2 = \lambda^1/2_{\text{max}}(\Phi(X)'\Phi(X))^{-1} = O_P(1/\sqrt{T})$. $\|PF\|_2 = \lambda^1/2_{\text{max}}(\frac{1}{T}F'PF)\sqrt{T} = O_P(\sqrt{T})$. By Lemma A.1.3,

$$\|W_2\|_2 \leq \frac{1}{NT} \|(\Phi(X)'\Phi(X))^{-1/2}\|_2^2 \|\Phi(X)\|_2 \|F\|_F \|\tilde{F}' \tilde{U} \Phi(X)\|_F 
\leq O_P(\frac{1}{NT})(\|\tilde{\Lambda}'(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})U\Phi(X)\|_F + \|\tilde{\Lambda}'\Sigma_u^{-1}U\Phi(X)\|_F) 
= O_P \left( \sqrt{\frac{J}{NT}} + \frac{\sqrt{J} \log N}{T} \right),$$
$$\|W_4\|_2 \leq \frac{1}{NT} \|(\Phi(X)'\Phi(X))^{-1/2}\|_2^3 \|\Phi(X)'U\|_F^2 \|\Sigma_u^{-1/2}\|_2^2 
= O_P \left( \frac{J}{T} \right).$$

For the $k$th eigenvalue, we have $|\lambda_k(W) - \lambda_k(W_1)| \leq \|W - W_1\|_2 = o_P(1)$. Hence it suffices to prove that the first $K$ eigenvalues of $W_1$ are bounded away from zero and infinity, which are also the first $K$ eigenvalues of $\frac{1}{T}F'PF$. This is assumed in Assumption 1.3.2. Therefore, $\|K^{-1}\|_2 = O_P(1) = \|K\|_2$, and this also implies $\|M\|_2 = O_P(1)$. \hfill $\Box$

**Lemma A.1.2.** (i) $\|A_1\|_F^2 = O_P(JN/T)$, (ii) $\|A_2\|_F^2 = O_P(J^2N/T^2)$, (iii) $\|A_3\|_F^2 = O_P(JN/T)$.

**Proof.** Note that $\|\hat{\Sigma}_u^{-1/2}\|_2 = O_P(1)$, $\|\tilde{\Lambda}\|_F^2 = \|\hat{\Sigma}_u^{-1/2}\tilde{\Lambda}\|_F^2 = O_P(N) = \|\tilde{\Lambda}\|_F^2$, and As-.
Lemma A.1.3. (i) $\max_{i \leq N} \max_{t \leq T} \sum_{s=1}^{T} |Eu_{it}u_{is}| = O(1)$, 
$\max_{k \leq K, t \leq T} \max_{i \leq N} \sum_{j=1}^{N} \text{cov}(\lambda_{ik}u_{it}, \lambda_{jk}u_{jt}) = O(1)$, 
$\max_{t \leq T, s \leq T} \max_{i \leq N} \sum_{j=1}^{N} \text{cov}(u_{it}u_{is}, u_{jt}u_{js}) = O(1)$.
(ii) $\|U'A\|_F = O(\sqrt{NT})$.
(iii) $\|UP(X)\|_F^2 = O(P(JNT))$, $\|\Phi(X)'U'A\|_F^2 = O(P(JNT))$, $\|PU'\|_F^2 = O(P(JN))$.
(iv) $\|\Phi(X)'U'S^{-1}_a\|_F^2 = O(P(JNT))$, $\|U'S^{-1}_a\|_F^2 = O(P(NT))$.

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

$$E\|U'A\|_F^2 = E \sum_{k=1}^{K} \sum_{t=1}^{T} (\sum_{i=1}^{N} u_{it}\lambda_{ik})^2 = \sum_{k=1}^{K} \sum_{t=1}^{T} \text{var}(\sum_{i=1}^{N} u_{it}\lambda_{ik})$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{T} \text{var}(u_{it}\lambda_{ik}) + \sum_{i \leq N, j \leq N} \text{cov}(u_{it}\lambda_{ik}, u_{jt}\lambda_{jk}) = O(NT).$$

(iii) From part (i) and Assumptions 1.3.3 (ii) and 1.3.4 (iv),

$$E\|U\Phi(X)\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{d=1}^{d} \sum_{s=1}^{T} \sum_{t=1}^{T} E\phi_j(X_{it})\phi_j(X_{it})Eu_{it}u_{is}$$

$$\leq Jd T \max_{j \leq J, t \leq T} |E\phi_j(X_{it})\phi_j(X_{it})| \sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{s \leq T} E|u_{it}u_{is}|$$

$$\leq Jd T \max_{j \leq J, t \leq T} E\phi_j(X_{it})^2 \sum_{i=1}^{N} \sum_{s \leq T} \sum_{t=1}^{T} E|u_{it}u_{is}| = O(JNT).$$
where the Cauchy Schwarz inequality implies the second inequality. On the other hand,

\[ E\|\Phi(X)U'\Lambda\|^2_F = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=1}^{d} E\left(\sum_{t=1}^{T} \sum_{i=1}^{N} \phi_j(X_{it})u_{it}\lambda_{ik}\right)^2 \]

\[ = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=1}^{d} \text{var}\left(\sum_{t=1}^{T} \sum_{i=1}^{N} \phi_j(X_{it})u_{it}\lambda_{ik}\right) \]

\[ = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=1}^{d} \sum_{t=1}^{T} \sum_{i=1}^{N} \text{var}(\phi_j(X_{it})u_{it}\lambda_{ik}) \]

\[ + \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=1}^{d} \sum_{t=1}^{T} \sum_{i \neq m; i, m \leq N} (E\phi_j(X_{it})^2\lambda_{ik}\lambda_{mk}) E\|u_{im}\|^2 \]

\[ + \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=1}^{d} \sum_{t \neq s; t, s \leq T} \sum_{i, m \leq N} (E\phi_j(X_{it})\phi_j(X_{is})\lambda_{ik}\lambda_{mk}) E\|u_{im}\|E\|u_{is}\| \]

\[ \leq O(JNT) + O(JNT) \max_{t \leq T, i \leq N} \sum_{m=1}^{N} |E\|u_{im}\| + O(JNT) \frac{1}{NT} \sum_{i, m \leq N} \sum_{t, s \leq T} |E\|u_{im}\|E\|u_{is}\| \]

\[ = O(JNT). \]

Finally, since \(\|\Phi(X)\Phi(X)^{-1}\|_2 = \lambda_{\min}^{-1}(T^{-1}\Phi(X)\Phi(X))T^{-1} = O(T^{-1}),\)

\[ \|PU'\|_F \leq \|\Phi(X)\|_2 \|\Phi(X)\|^{-1}_2 \|\Phi(X)'U'\|_F = O(\sqrt{JN}). \]

(iv) Note that we have derived the same rates for the case \(\Sigma_u^{-1} = I\) in parts (ii) and (iii). Here we only need to define \(\Lambda^* = \Sigma_u^{-\frac{1}{2}}\Lambda\) and \(U^* = \Sigma_u^{-\frac{1}{2}}U\) and prove \(\Lambda^*\) and \(U^*\) possess the same properties as \(\Lambda\) and \(U\). The results follow using the same argument as in parts (ii) and (iii) that \(\Sigma_u^{-\frac{1}{2}}\Lambda\) has bounded row sums. \(\square\)

Lemma A.1.4. In the conventional factor model,

(i) \(\|\tilde{\Lambda}'A_1\|_F = O_P\left(\frac{JN}{T} + N\sqrt{\frac{J}{T}m_N\delta_{N,T}^{1-q}}\right),\)

\(\|\tilde{\Lambda}'A_2\|_F = O_P\left(\frac{JN}{T}m_N^2\delta_{N,T}^{2-2q} + \frac{JN}{T}\sqrt{m_N\delta_{N,T}^{1-q}}\right),\)

\(\|\tilde{\Lambda}'A_3\|_F = O_P(N\sqrt{\frac{J}{T}m_N\delta_{N,T}^{1-q}}).\)

(ii) \(\|\tilde{\Lambda}'(\tilde{\Lambda} - \tilde{\Lambda}M)\|_F = O_P\left(\frac{JN}{T} + \sqrt{\frac{J}{T}m_N\delta_{N,T}^{1-q}}\right),\) and \(\|\tilde{\Lambda}'(\tilde{\Lambda} - \tilde{\Lambda}M)\|_F = O_P\left(\frac{JN}{T} + \sqrt{\frac{J}{T}m_N\delta_{N,T}^{1-q}}\right).\)

Proof. Note that \(\|\Phi'U\Sigma_u^{-1}\Lambda\|_F = O_P(\sqrt{JNT})\) and

\[ \|PU'\tilde{\Lambda}\|_F \leq \|\Phi(\Phi')^{-1}\|_2\|\Phi'U(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})\Lambda\|_F + \|\Phi'U\Sigma_u^{-1}\Lambda\|_F \leq O_P(N\sqrt{Jm_N\delta_{N,T}^{1-q}}). \]
Hence

\[ \|\tilde{\Lambda}' \tilde{A}\|_F \leq \frac{1}{N^2T} \|\tilde{\Lambda}' \tilde{A} \tilde{F}' \tilde{P}\|_F(\|\tilde{P} \tilde{U}' \tilde{\Lambda} \tilde{M}\|_F + \|\tilde{P} \tilde{U}' (\tilde{\Lambda} - \tilde{\Lambda} M)\|_F) \]

\[ = O_P \left( \frac{JN}{T} + N \sqrt{\frac{J}{T} m_N \delta_{N,T}^{1-q}} \right), \]

\[ \|\tilde{\Lambda}' \tilde{A}_2\|_F \leq \frac{1}{N^2T} \|\tilde{\Lambda}' \tilde{U} \tilde{P}\|_F(\|\tilde{P} \tilde{U}' \tilde{\Lambda} \tilde{M}\|_F + \|\tilde{P} \tilde{U}' (\tilde{\Lambda} - \tilde{\Lambda} M)\|_F) \]

\[ = O_P \left( \frac{JN}{T} m_N^2 \delta_{N,T}^{2-2q} + \frac{JN}{T} \sqrt{\frac{J}{T} m_N \delta_{N,T}^{1-q}} \right), \]

\[ \|\tilde{\Lambda}' \tilde{A}_3\|_F \leq \frac{1}{N^2T} \|\tilde{\Lambda}' \tilde{A} \tilde{F}' \tilde{P}\|_F \|\tilde{P} \tilde{F}\|_2 \|\tilde{A}\|_F = O_P \left( N \sqrt{\frac{J}{T} m_N \delta_{N,T}^{1-q}} \right). \]

For part (ii),

\[ \frac{1}{N} \|\tilde{\Lambda}' (\tilde{\Lambda} - \tilde{\Lambda} M)\|_F \leq \frac{1}{N} \|K^{-1}\|_2 \sum_{i=1}^{3} \|\tilde{A}' A_i\|_F = O_P(J/T + \sqrt{J/T m_N \delta_{N,T}^{1-q}}). \]

In addition, the result follows from the following inequality:

\[ \frac{1}{N} \|\tilde{\Lambda}' (\tilde{\Lambda} - \tilde{\Lambda} M)\|_F \leq \frac{1}{N} \|\tilde{\Lambda} - \tilde{\Lambda} M\|_F^2 + \frac{1}{N} \|M' \tilde{\Lambda}' (\tilde{\Lambda} - \tilde{\Lambda} M)\|_F = O_P(J/T + \sqrt{J/T m_N \delta_{N,T}^{1-q}}) \]

Lemma A.1.5. \( \|M' M - I_K\|_F = O_P(\frac{J}{T} + \sqrt{J/T m_N \delta_{N,T}^{1-q}}). \) Therefore, \( \|M^{-1}\|_2 = O_P(1). \)

Proof. Note that \( N^{-1} A' A = N^{-1} A' \tilde{A} = N^{-1} \tilde{A}' \tilde{A}, \) by the identification condition. Then

\[ M' M = \frac{1}{N} (\tilde{\Lambda} M)' \tilde{\Lambda} M = \frac{1}{N} (\tilde{\Lambda} M - \tilde{\Lambda})' \tilde{\Lambda} M + \frac{1}{N} \tilde{\Lambda}' (\tilde{\Lambda} M - \tilde{\Lambda}) + I_K. \]

This implies the following convergence rate,

\[ \|M' M - I_K\|_F \leq \frac{1}{N} \|\tilde{\Lambda} M - \tilde{\Lambda}\|_F \|M\|_2 + \frac{1}{N} \|\tilde{\Lambda}' (\tilde{\Lambda} M - \tilde{\Lambda})\|_F = O_P(J/T + \sqrt{J/T m_N \delta_{N,T}^{1-q}}). \]

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

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

□
Lemma A.1.6. (i) \( \| \frac{1}{N} \tilde{U}' A_1 \|_F^2 = O_P(J/N + (J/T)m_N^2 \delta_{N,T}^{2-2q}) \), 
\( \| \frac{1}{N} \tilde{U}' A_2 \|_F^2 = O_P(J^2/NT + J/T^2) \), 
\( \| \frac{1}{N} \tilde{U}' A_3 \|_F^2 = O_P(J/N + 1/T) \).

(ii) \( \| \frac{1}{N} \tilde{U}' (\tilde{A} - \tilde{A}M) \|_F^2 = O_P(J/N + 1/T) \).

Proof. (i) Note that, by Lemmas A.1.3 and A.2.9,

\[ \| \tilde{U}' \tilde{A} \|_F^2 \leq \| U'(\tilde{\Sigma}_u^{-1} - \Sigma_u^{-1})A \|_F^2 + \| U'\Sigma_u^{-1}A \|_F^2 = O_P(NT + (N^2 + NT)m_N^2 \delta_{N,T}^{2-2q}) \],

\[ \| \tilde{U}' \tilde{\Phi}(X) \|_F^2 \leq \| U'(\tilde{\Sigma}_u^{-1} - \Sigma_u^{-1})U\Phi(X) \|_F^2 + \| U'\Sigma_u^{-1}U\Phi(X) \|_F^2 = O_P(JNT^2 + N^2T) \].

Hence

\[ \| \frac{1}{N} \tilde{U}' A_1 \|_F^2 \leq \frac{1}{N^2T^2} \| \tilde{U}' \tilde{A} \|_F^2 \| F'P \|_2^2 \| P \tilde{U}' \|_F^2 \| \tilde{\Sigma}_u^{-1} \|_2^2 \| \tilde{\Lambda} \|_F^2 = O_P(J/N + (J/T)m_N^2 \delta_{N,T}^{2-2q}) \],

\[ \| \frac{1}{N} \tilde{U}' A_2 \|_F^2 \leq \frac{1}{N^2T^2} \| \tilde{U}' \tilde{\Phi}(X) \|_F^2 \| (\Phi(X)'\Phi(X))^{-1} \|_2^2 \| \tilde{\Phi}(X) \|_F^2 \| \tilde{\Sigma}_u^{-1} \|_2^2 \| \tilde{\Lambda} \|_F^2 = O_P(J^2/NT + J/T^2) \],

\[ \| \frac{1}{N} \tilde{U}' A_3 \|_F^2 \leq \frac{1}{N^2T^2} \| \tilde{U}' \tilde{\Phi}(X) \|_F^2 \| (\Phi(X)'\Phi(X))^{-1} \|_2^2 \| PF \|_2^2 \| \tilde{\Phi} \|_F^2 \| \tilde{\Lambda} \|_F^2 = O_P(J/N + 1/T) \].

Part (ii) follows from part (i). \( \Box \)

A.1.2 Convergence of factors

For the estimated \( \tilde{G}(X) \), note that

\[ \tilde{G}(X) = \frac{1}{N} P \tilde{Y}' \tilde{\Lambda} = \frac{1}{N} P \tilde{F}' \tilde{A} + \frac{1}{N} P \tilde{U}' \tilde{\Lambda} = PF + E, \]

where \( E = \frac{1}{N} P \tilde{F}' (\tilde{A} - \tilde{A}M) + \frac{1}{N} P \tilde{U}' (\tilde{A} - \tilde{A}M) + \frac{1}{N} P \tilde{U}' \tilde{A}M \).

By Lemma A.1.4,

\[ \left\| \frac{1}{N} P \tilde{F}' (\tilde{A} - \tilde{A}M) \right\|_F \leq O_P \left( \frac{\sqrt{T}}{N} \right) \left\| \tilde{A}' (\tilde{A} - \tilde{A}M) \right\|_F = O_P \left( \frac{J}{\sqrt{T}} + \sqrt{J}m_N \delta_{N,T}^{1-q} \right). \]
By Lemma A.1.3, $\| \frac{1}{N} \mathbf{P} \hat{\mathbf{U}}'(\hat{\mathbf{A}} - \hat{\mathbf{A}} \mathbf{M}) \|_F \leq \frac{1}{N} \| \mathbf{P} \mathbf{U}' \|_F \| \hat{\Sigma}_{u}^{-2} \| \| \hat{\mathbf{A}} - \hat{\mathbf{A}} \mathbf{M} \|_F = O_P(J/\sqrt{T})$, and $\| \frac{1}{N} \mathbf{P} \hat{\mathbf{U}}' \hat{\mathbf{A}} \mathbf{M} \|_F = O_P(\sqrt{J} N \delta_{N,T}^{1-q})$. Hence,

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

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

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

where

$$\mathbf{B}_1 = \frac{1}{N} \mathbf{F} \hat{\mathbf{A}}' (\hat{\mathbf{A}} - \hat{\mathbf{A}} \mathbf{M}), \quad \mathbf{B}_2 = \frac{1}{N} \hat{\mathbf{U}}' (\hat{\mathbf{A}} - \hat{\mathbf{A}} \mathbf{M}), \quad \mathbf{B}_3 = \frac{1}{N} \hat{\mathbf{U}}' \hat{\mathbf{A}} \mathbf{M}.$$

Note that $\| \mathbf{U} \|_2^2 = O_P(N + T)$. By Lemmas A.1.3, A.1.4, and A.1.6,

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

$$\| \mathbf{B}_3 \|_F^2 = O_P \left( \frac{T}{N} + \frac{T}{N^2} m N \delta_{N,T}^{2-2q} \right).$$

Therefore,

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

A.1.3 Individual factors

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

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

where

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

Then, by Lemmas A.1.4 and A.1.5,

$$\| \hat{\mathbf{D}}_1 \| \leq \frac{1}{N} \| \hat{\mathbf{A}} (\hat{\mathbf{A}} - \hat{\mathbf{A}} \mathbf{H}) \|_F \| \mathbf{M}^{-1} \|_2 \| \mathbf{F}_t \| = O_P \left( \frac{J}{T} + \sqrt{\frac{J}{T} m N \delta_{N,T}^{1-q}} \right).$$
Note that $\|N^{-1}\Lambda'\mathbf{u}_t\| = O_P(1/\sqrt{N})$ by Lemma A.1.8. Let $\Lambda^* = \Sigma_u^{-\frac{1}{2}} \Lambda$ and $\mathbf{u}_t^* = \Sigma_u^{-\frac{1}{2}} \mathbf{u}_t$. Then $\|N^{-1}\Lambda^*\mathbf{u}_t^*\| = O_P(1/\sqrt{N})$.

$$\|D_3\| \leq \frac{1}{N}\|M\|_2(\|\Lambda' (\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})\mathbf{u}_t\| + \|\Lambda' \mathbf{u}_t^*\|)$$

$$\leq O_P\left(\frac{1}{N}\|\Lambda\|_F\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1\|\mathbf{u}_t\| + \|\Lambda^* \mathbf{u}_t^*\|\right)$$

$$\leq O_P(m_N\delta_{N,T}^{1-\frac{q}{2}} + \frac{1}{\sqrt{N}}) = O_P(m_N\delta_{N,T}^{1-\frac{q}{2}}).$$

Note that $D_2 = \frac{1}{NT} \hat{\mathbf{u}}_t (\hat{\Lambda} - \tilde{\Lambda} M) \| \leq \|K^{-1}\|_2 \sum_{i=1}^3 \frac{1}{N} \hat{\mathbf{u}}_t A_i$, where

$$A_1 = \frac{1}{NT} \hat{\Lambda} F' P \hat{U}' \hat{\Lambda}, \quad A_2 = \frac{1}{NT} \hat{U} P \hat{U}' \hat{\Lambda}, \quad A_3 = \frac{1}{NT} \hat{U} P F \hat{\Lambda}'$$

Then, by Lemma A.1.7, $\|D_2\| = O_P(\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}})$.

Therefore, for each $t \leq T$,

$$\|\hat{F}_t - M^{-1} F_t\| = O_P\left(\frac{J}{T} + \sqrt{\frac{1}{T} m_N \delta_{N,T}^{1-\frac{q}{2}}} + O_P\left(\frac{\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}}} + O_P(\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2} + J \frac{1}{\sqrt{T}}} + O_P(\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}}).$$

Lemma A.1.7. (i) $\frac{1}{N} \hat{\mathbf{u}}_t A_1 = O_P(m_N \delta_{N,T}^{1-\frac{q}{2}} (\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}} + \frac{1}{T}))$,

(ii) $\frac{1}{N} \hat{\mathbf{u}}_t A_2 = O_P(m_N \delta_{N,T}^{1-\frac{q}{2}} (\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}} + \frac{1}{T}))$,

(iii) $\frac{1}{N} \hat{\mathbf{u}}_t A_3 = O_P(\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}})$.

Proof. (i) Note that $\|\Lambda^* \mathbf{u}_t^*\| = O_P(\sqrt{N})$ by Lemma A.1.8, and $\|F'P\|_2 = O_P(\sqrt{T})$.

$$\frac{1}{N} \hat{\mathbf{u}}_t A_1 = \frac{1}{N^2 T} \|\hat{\mathbf{u}}_t \hat{\Lambda} F' P \hat{U}' \hat{\Lambda}\|$$

$$\leq \frac{1}{N^2 T} \|\hat{\mathbf{u}}_t (\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}) \Lambda F' P \hat{U}' \hat{\Lambda}\| + \frac{1}{N^2 T} \|\mathbf{u}_t^* \Lambda^* F' P \hat{U}' \hat{\Lambda}\|$$

$$\leq \frac{1}{N^2 T} \|\hat{\mathbf{u}}_t\| \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 \|\Lambda\|_F \|F'P\|_2(\|P \hat{U}' \hat{\Lambda} M\|_F + \|P \hat{U}' (\hat{\Lambda} - \tilde{\Lambda} M)\|_F$$

$$+ \frac{1}{N^2 T} \|\mathbf{u}_t^*\| \|\Lambda^*\|_F \|F'P\|_2(\|P \hat{U}' \tilde{\Lambda} M\|_F + \|P \hat{U}' (\hat{\Lambda} - \tilde{\Lambda} M)\|_F)$$

$$= O_P((\frac{1}{N^2 T} m_N \delta_{N,T}^{1-\frac{q}{2}} + \frac{1}{N^2 T}) (N \sqrt{J} m_N \delta_{N,T}^{1-\frac{q}{2}} + \frac{J N}{\sqrt{T}})$

$$= O_P(m_N \delta_{N,T}^{1-\frac{q}{2}} (\sqrt{\frac{J}{T}} m_N \delta_{N,T}^{1-\frac{q}{2}} + \frac{1}{T})$$. 

(ii) Note that, by Lemma A.1.8,

$$\|\tilde{\mathbf{u}}_t U P\| \leq \|\mathbf{u}_t (\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}) UP\| + \|\mathbf{u}_t^* U^* P\| = O_P(N \sqrt{J} m_N \delta_{N,T}^{1-\frac{q}{2}}).$$
Then,
\[
\|\frac{1}{N} \tilde{u}_t A_2 \| = \frac{1}{N^2 T} \| \tilde{u}_t \bar{U} \bar{P} \tilde{A} \|
\geq \frac{1}{N^2 T} \| \tilde{u}_t \bar{U} \bar{P} (\| P \bar{U} \hat{\Lambda} \|_F + \| P \bar{U} (\hat{\Lambda} - \bar{\Lambda}) \|_F) = O_P(m_N \delta_{N,T}^{1-q}(\sqrt{\frac{1}{T} m_N \delta_{N,T}^{1-q}} + J \sqrt{\frac{1}{T}})).
\]

(iii) Note that \(\| F' \|_2 = O_P(\sqrt{T}) \) and \(\| \tilde{u}_t \bar{U} \bar{P} \| = O_P(N \sqrt{J m_N \delta_{N,T}^{1-q}}). \) Then,
\[
\| \frac{1}{N} \tilde{u}_t A_3 \| = \frac{1}{N^2 T} \| \tilde{u}_t \bar{U} \bar{P} \bar{A} \tilde{A} \| \leq O_P(\frac{1}{N^T}) \| \tilde{u}_t \bar{U} \bar{P} \| \| \bar{P} \|_2 = O_P(\sqrt{T} m_N \delta_{N,T}^{1-q}).
\]

**Lemma A.1.8.** (i) \(\| A'u_t \|^2 = O_P(N) = \| A' \Sigma^{-1} u_t \|^2.\)

(ii) \(\| u_t' \bar{U} \Phi(X) \| = O_P(N \sqrt{J} + \sqrt{J N T}) = \| u_t' \Sigma^{-1} \bar{U} \Phi(X) \|.\)

**Proof.** (i) Note that \(\| \lambda_k' u_t \|^2 = O_P(1)E(\| \lambda_k' u_t \|^2 = O_P(1) \text{var}(\lambda_k' u_t) = O_P(1) \lambda_k \text{var}(u_t) \lambda_k \leq O_P(1) \| \lambda_k \|^2 \| \text{var}(u_t) \| = O_P(N). \) Then \(\| A'u_t \|^2 = \sum_{k=1}^{K} \| \lambda_k' u_t \|^2 = O_P(N). \)

(ii) Note that \(\| u_t' \bar{U} \Phi(X) \| = \| \sum_{i=1}^{N} u_{it} u_t' \Phi(X) \|, \) where \(u_i = (u_{i1}, ..., u_{iT})'. \) Also
\[
E \| \sum_{i=1}^{N} u_{it} u_t' \Phi(X) \|^2 = \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \sum_{s} E(\sum_{i} \sum_{s} u_{is} u_{it} \phi_k(X_{sl}))^2
\]
\[
\leq 2 \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \sum_{s} \text{var}(u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl})) + \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \sum_{s} (u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl}))^2
\]
\[
\leq O(N^2 J) + 2 \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \text{var}(u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl}))
\]
\[
+ 2 \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \text{var}(u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl}))
\]
\[
= O(N^2 J) + 2 \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \text{var}(u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl}))
\]
\[
+ 2 \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \text{cov}(u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl})) (u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl}))
\]
\[
= O(N^2 J + J N T) + 4 \sum_{k=1}^{J} \sum_{l=1}^{d} \sum_{i} \text{cov}(u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl})) (u_{is} u_{it} - E u_{is} u_{it}) \phi_k(X_{sl}))
\]
\[
= O(N^2 J + J N T).
\]
A.2 Proof of Theorem 1.3.2

Theorem A.2.1. Consider the semiparametric factor model. Under the Assumption 1.3.1, 1.3.3-1.3.7, when \( \|\Sigma_u^{-1}\|_1 = O(1) \),

\[
\|\hat{\Sigma}_u - \Sigma_u\|_1 = O_P(m_N\omega_{N,T}^{1-q}) = \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1,
\]

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

Proof. Let \( \hat{U} = Y - \hat{\Lambda}\hat{F} \), 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 \( \max_{i \leq N} \|\hat{\lambda}_i - H'\lambda_i\| = O_P(\frac{1}{\sqrt{T}}) \). Then \( \max_{i \leq N} \frac{1}{T} \sum_{t=1}^T (\hat{u}_{it} - u_{it})^2 = O_P(\frac{1}{N} + \frac{1}{T}) \), which also implies

\[
\max_{i \leq N,j \leq N} \frac{1}{T} \sum_{t=1}^T (\hat{u}_{it}\hat{u}_{jt} - u_{it}u_{jt}) = O_P(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{T}}).
\]

In addition, \( \max_{i \leq N,j \leq N} \frac{1}{T} \sum_{t=1}^T u_{it}u_{jt} = O_P(\frac{\log N}{T}) \). Then we have, for \( \omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}} \),

\[
\max_{i \leq N,j \leq N} \frac{1}{T} \sum_{t=1}^T \hat{u}_{it}\hat{u}_{jt} - \Sigma_{u,ij} = O_P(\omega_{N,T}).
\]

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

\[
\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 \leq \|(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\hat{\Sigma}_u - \Sigma_u)\Sigma_u^{-1}\|_1 + \|\Sigma_u^{-1}(\hat{\Sigma}_u - \Sigma_u)\Sigma_u^{-1}\|_1
\]

\[
\leq \|(\hat{\Sigma}_u^{-1} - \Sigma_u^{-1})\|_1 \|\hat{\Sigma}_u - \Sigma_u\|_1 \|\Sigma_u^{-1}\|_1 + \|\Sigma_u^{-1}\|_1 \|\hat{\Sigma}_u - \Sigma_u\|_1
\]

\[
= O_P(m_N\omega_{N,T}^{1-q})\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 + O_P(m_N\omega_{N,T}^{1-q}).
\]

Therefore, we have \( (1 + o_P(1))\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 = O_P(m_N\omega_{N,T}^{1-q}) \), and it implies the results. \( \square \)
A.2.1 Convergence of loadings

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

$$\frac{1}{NT}(\tilde{Y}P\tilde{Y}')\tilde{\Lambda} = \tilde{\Lambda}K.$$ 

Let

$$H = \frac{1}{NT}(QQ'\tilde{\Lambda}' + Q\tilde{U}')\tilde{\Lambda}^{-1},$$

where $Q = B\Phi(X)' + \Gamma'P + R(X)'P$. We shall show that $\|H\|_2 = O_P(1)$ in Lemma A.2.3. Note that $\tilde{\Lambda} = \hat{\Sigma}^{-\frac{1}{2}}\Lambda$, and $\tilde{U} = \hat{\Sigma}^{-\frac{1}{2}}U$. Substituting $\tilde{Y} = \hat{\Lambda}B\Phi(X)' + \hat{\Lambda}R(X)' + \hat{\Lambda}\Gamma' + \tilde{U}$, we have

$$\hat{\Lambda} - \tilde{\Lambda}H = \left(\sum_{i=1}^{4} A_i\right)K^{-1},$$

where

$$A_1 = \frac{1}{NT} \tilde{U} \Phi(X)B'\tilde{\Lambda}'\tilde{\Lambda}, A_2 = \frac{1}{NT} \tilde{U} R(X)\tilde{\Lambda}'\tilde{\Lambda}, A_3 = \frac{1}{NT} \tilde{U} \Gamma'\tilde{\Lambda}'\tilde{\Lambda}, A_4 = \frac{1}{NT} \tilde{U} \tilde{U}'\tilde{\Lambda}'\tilde{\Lambda}.$$

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

$$\frac{1}{N} \|\hat{\Lambda} - \tilde{\Lambda}H\|_F^2 \leq C \|K^{-1}\|_2^2 \sum_{i=1}^{4} \frac{1}{N} \|A_i\|_F^2.$$

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

**Lemma A.2.1.** (i) $\frac{1}{N} \|A_1\|_F^2 = O_P(T^{-1})$,

(ii) $\frac{1}{N} \|A_2\|_F^2 = O_P(T^{-1}J^{-1-\kappa})$,

(iii) $\frac{1}{N} \|A_3\|_F^2 = O_P(J^2v_T/T^2)$,

(iv) $\frac{1}{N} \|A_4\|_F^2 = O_P(J^2/T^2)$.

**Proof.** (i) Because $\|\tilde{\Lambda}\|_F^2 = O_P(N) = \|\Lambda\|_F^2$, $\|\tilde{\Lambda}\|_F^2 = O_P(N)$. Note that $\|\tilde{\Sigma}^{-\frac{1}{2}}\|_2 = O_P(1)$. By Lemma A.2.2, $\|U\Phi(X)B\|_F^2 = O_P(NT)$. Then $\frac{1}{N} \|A_1\|_F^2 = O_P(T^{-1})$.

(ii) Note that $\|R(X)\|_F^2 = O_P(TJ^{-\kappa})$. By Lemma A.1.3, $\|U\Phi(X)\|_F = O_P(\sqrt{JNT})$. 

By Assumption 1.3.3, \(|(\Phi(X)'\Phi(X))^{-1}|_2 = O_P(T^{-1}) \). Then,

\[
\|A_2\|_F \leq \frac{1}{NT} \|\tilde{\Sigma}_u^{-1}\|_2 \|U\Phi(X)\|_F \|((\Phi(X)'\Phi(X))^{-1}|_2 \|\Phi(X)\|_2 \|R(X)\|_F \|\tilde{A}\|_F \|\tilde{A}\|_F
\]

\[
= O_P \left( \sqrt{\frac{JN}{TJr}} \right).
\]

Therefore, \(\frac{1}{N} \|A_2\|_F^2 = O_P(T^{-1}J^{1-k})\).

(iii) It follows from Lemma A.2.2 that \(\|\Phi(X)'\|_F^2 = O_P(JT\nu_T)\). Then

\[
\frac{1}{N} \|A_3\|_F^2 = O_P(\frac{1}{N\nu_T} \|U\Phi(X)\|_F^2 \|\Phi(X)'\|_F^2) = O_P(J^2\nu_T/T^2).
\]

(iv) By Lemma A.1.3 and \(\|\tilde{\Sigma}_u^{-\frac{1}{2}}\|_2 = O_P(1), \frac{1}{N} \|A_4\|_F^2 = O_P(J^2/T^2). \)

By Lemma A.2.3, \(\|K^{-1}\|_2 = O_P(1)\). Note that \(\|\tilde{\Sigma}_u\|_2 < \infty\). Therefore, as \(J = o(\sqrt{T}) \) and \(\kappa \geq 1\),

\[
\frac{1}{N} \|\hat{A} - \Lambda H\|_F^2 \leq O_P \left( \frac{1}{N} \right) \|\hat{A} - \hat{\Lambda} H\|_F^2 \leq O_P \left( \frac{1}{N} \|K^{-1}\|_2^2 \right) \sum_{i=1}^4 \|A_i\|_F^2 = O_P(1/T).
\]

**Lemma A.2.2.** (i) \(\|U\Phi(X)B\|_F^2 = O_P(NT)\) and \(\|A'U\Phi(X)B\|_F^2 = O_P(NT) = \|A'\Sigma_u^{-1}U\Phi(X)B\|_F^2\).

(ii) \(\|\Phi(X)'T\|_F^2 = O_P(JT\nu_T), \|B\Phi(X)'T\|_F^2 = O_P(T\nu_T)\).

**Proof.** (i) Note that \(B\Phi(X)'U' = G(X)'U' - R(X)'U'\) and \(X_t\) and \(u_i\) are independent. Then

\[
E\|G(X)'U'\|_F^2 = \sum_{k=1}^K \sum_{t=1}^T \left( \sum_{i=1}^N g_k(X_t)u_{it} \right)^2
\]

\[
= \sum_{k=1}^K \sum_{t=1}^T \sum_{s=1}^T E_{g_k(X_t)g_k(X_s)}E_{u_{it}u_{is}}
\]

\[
\leq TK \max_{k \leq K} E_{g_k(X_t)}^2 \sum_{i=1}^N \max_{s \leq T} \sum_{t=1}^T |E_{u_{it}u_{is}}| = O(NT).
\]

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

\[
E\|R(X)'U'\|_F^2 = \sum_{k=1}^K \sum_{t=1}^T \left( \sum_{i=1}^N R_{tk}u_{it} \right)^2
\]

\[
= \sum_{k=1}^K \sum_{t=1}^T \sum_{s=1}^T \sum_{l=1}^d ER_{tk}R_{sk}E_{u_{it}u_{is}}
\]
where \( \max_{k \leq K} E R_{ik}^2 = O(J^{-\kappa}) \). Therefore, \( \| \mathbf{B} \Phi(\mathbf{X})' \mathbf{U}' \|^2_F = O_P(NT) \).

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

\[
E\| \mathbf{G}(\mathbf{X})' \mathbf{U}' \Lambda \|^2_F = \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} g_t(\mathbf{X}_t) u_{it} \lambda_{ik}^2 = \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \text{cov}(\sum_{i=1}^{N} g_t(\mathbf{X}_t) u_{it} \lambda_{ik}, \sum_{i=1}^{N} g_t(\mathbf{X}_t) u_{it} \lambda_{jk}) \]

\[
= D_1 + D_2.
\]

Here, note that \( \text{var}(g_t(\mathbf{X}_t) u_{it} \lambda_{ik}) \) and \( |E g_t(\mathbf{X}_t) u_{it} \lambda_{ik}| \) are bounded uniformly in \( k \leq K, l \leq K, i \leq N, \) and \( j \leq N \). Then, by Assumption 1.3.4,

\[
D_1 = \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \text{var}(g_t(\mathbf{X}_t) u_{it} \lambda_{ik}) + \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \text{cov}(g_t(\mathbf{X}_t) u_{it} \lambda_{ik}, g_t(\mathbf{X}_t) u_{jt} \lambda_{jk}) \]

\[
= O(NT) + \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} g_t(\mathbf{X}_t) u_{it} \lambda_{ik} g_t(\mathbf{X}_t) u_{jt} \lambda_{jk} \]

\[
\leq O(NT) + NTK^2 \sum_{k=1}^{K} \sum_{l=1}^{K} \max_{i \leq K, j \leq N} \text{var}(g_t(\mathbf{X}_t) u_{it} \lambda_{ik}) \max_{i \leq N} \sum_{j=1}^{N} |E u_{it} u_{jt}| = O(NT).
\]

\[
D_2 = \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{i \leq K, j \leq N} \sum_{t \leq T, s \leq T} \text{cov}(g_t(\mathbf{X}_t) u_{it} \lambda_{ik}, g_t(\mathbf{X}_t) u_{it} \lambda_{jk}) |E u_{it} u_{js}| \]

\[
\leq k^2 \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t \leq T, s \leq T} \sum_{i=1}^{N} \sum_{j=1}^{N} |E g_t(\mathbf{X}_t) u_{it} \lambda_{ik} g_t(\mathbf{X}_t) u_{it} \lambda_{jk}| \sum_{t \leq T, s \leq T} \sum_{i=1}^{N} \sum_{j=1}^{N} |E u_{it} u_{js}| = O(NT).
\]

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

(ii) Because \( \mathbf{X}_t \) and \( \gamma_{tk} \) are independent and \( E \gamma_{tk} = 0 \) by Assumption 1.3.6,

\[
E\| \mathbf{G} \|^2_F = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=1}^{L} E(\sum_{t=1}^{T} \phi_j(\mathbf{X}_t) \gamma_{tk})^2
\]
Lemma A.2.3.

\[ \| \| \Phi(X)^{\gamma_{tk}} \| \|^2 = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{d=1}^{d} \text{var} \left( \sum_{t=1}^{T} \phi_j(X_t)^{\gamma_{tk}} \right) \]

\[ = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{d=1}^{d} \text{var} \left( \phi_j(X_t)^{\gamma_{tk}} \right) + \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{d=1}^{d} \sum_{t=1}^{T} \text{cov} \left( \phi_j(X_t)^{\gamma_{tk}}, \phi_j(X_{\tau})^{\gamma_{sk}} \right) \]

\[ = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{d=1}^{d} \sum_{t=1}^{T} E\phi_j(X_t)^{2\gamma_{tk}} + \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{d=1}^{d} \sum_{t=1}^{T} E\phi_j(X_t)^{\phi_j(X_{\tau})E\gamma_{tk}} \gamma_{sk} \]

\[ \leq JdT \max_{j \leq J, l \leq T, l \leq d} E\phi_j(X_t)^{2\gamma_{tk}} \sum_{k=1}^{K} \max_{k \leq K} \frac{1}{T} \sum_{t=1}^{T} \text{Var} \left( \gamma_{tk} \right) \]

\[ + JdT \max_{j \leq J, l \leq d, t, t \leq s \leq T} E\phi_j(X_t)^{2} \sum_{k=1}^{K} \max_{k \leq K} \sum_{t=1}^{T} \left| \text{Var} \left( \gamma_{tk} \gamma_{sk} \right) \right| \]

\[ = O(JT\nu_T). \]

Also, \( B\Phi(X)^{\gamma_{tk}} = G(X)^{\gamma_{tk}} - R(X)^{\gamma_{tk}} \). Because

\[ E\| G(X)^{\gamma_{tk}} \|^{2} = \sum_{k=1}^{K} \sum_{t=1}^{T} E(\sum_{t=1}^{T} \text{var} \left( \phi_j(X_t)^{\gamma_{tk}} \right) = O(T\nu_T), \]

and \( E\| R(X)^{\gamma_{tk}} \|^{2} = O(J^{-\alpha}T\nu_T) \). Hence \( B\Phi(X)^{\gamma_{tk}} \|^{2} = O(T\nu_T). \)

\[ \square \]

Lemma A.2.3. \( \| K \|_{2} = O(1), \quad \| K^{-1} \|_{2} = O(1), \quad \text{and} \quad \| H \|_{2} = O(1). \)

**Proof.** For the general factor model, under Assumption 1.3.2, Lemma A.1.1 showed that \( K, K^{-1} \) and \( M \) all have bounded spectral norms. Now we consider the semiparametric factor model with \( F = F(X) + \Gamma \), and it implies \( \| K^{-1} \|_{2} = O(1) \) as well under Assumption 1.3.2. Note that \( \| Y \|_{2}^{2} = O(P(NT)) = \| \tilde{Y} \|_{2}^{2} \), and \( \| P \|_{2} = O(P(1)). \) Then it follows immediately that \( \| bK \|_{2} = O(1). \) Now we consider to prove \( \| K^{-1} \|_{2} = O(1). \)

Note that the eigenvalues of \( K \) are the same as those of

\[ W = \frac{1}{NT} (\Phi(X)^{\gamma_{tk}})^{-1/2} \Phi(X)^{\gamma_{tk}} \tilde{Y} \tilde{Y} \tilde{Y} \tilde{Y} \tilde{Y} \Phi(X)(\Phi(X)^{\gamma_{tk}})^{-1/2}. \]

Then, using \( \tilde{Y} = \tilde{F} + \tilde{U} \), and \( \frac{1}{N} \tilde{A}^T \tilde{A} = I_K \), we have \( W = \sum_{i=1}^{4} W_i \), where

\[ W_1 = \frac{1}{T} (\Phi(X)^{\gamma_{tk}})^{-1/2} \Phi(X)^{\gamma_{tk}} \tilde{F} \tilde{F} \Phi(X)(\Phi(X)^{\gamma_{tk}})^{-1/2}, \]

\[ W_2 = \frac{1}{T} (\Phi(X)^{\gamma_{tk}})^{-1/2} \Phi(X)^{\gamma_{tk}} \frac{\tilde{F} \tilde{A} \tilde{U}}{N} \Phi(X)(\Phi(X)^{\gamma_{tk}})^{-1/2}, \]

\[ W_3 = W_2', \]
By Lemma A.1.3, we have \( H \). where \( O \).

(i) Proof.

By (A.2.1), Lemma A.2.4. Under the assumptions of Theorem 1.3.2, where \( Q = B\Phi(X)' + \Gamma'P + R(X)'P \). It can be rewritten as

\[
H = \left( \sum_{i=1}^{12} H_i \right) K^{-1},
\]

where

\[
H_1 = \frac{1}{NT} B\Phi(X)' \Phi(X) B' \tilde{\Lambda}, \\
H_2 = \frac{1}{NT} B\Phi(X)' R(X) \tilde{\Lambda}, \\
H_3 = \frac{1}{NT} B\Phi(X)' T \tilde{\Lambda}, \\
H_4 = \frac{1}{NT} B\Phi(X)' \tilde{U} \tilde{\Lambda}, \\
H_5 = \frac{1}{NT} R(X)' (B \tilde{\Lambda})^\top, \\
H_6 = \frac{1}{NT} R(X)' (\tilde{\Lambda})^\top, \\
H_7 = \frac{1}{NT} R(X)' PR \tilde{\Lambda}, \\
H_8 = \frac{1}{NT} R(X)' P \tilde{U} \tilde{\Lambda}, \\
H_9 = \frac{1}{NT} \Gamma' \Phi(X) B' \tilde{\Lambda}, \\
H_{10} = \frac{1}{NT} \Gamma' PR \tilde{\Lambda}, \\
H_{11} = \frac{1}{NT} \Gamma' P \tilde{U} \tilde{\Lambda}, \\
H_{12} = \frac{1}{NT} \Gamma' P \tilde{U} \tilde{\Lambda}.
\]

By \( \|G(X)\|_2 = O_P(\sqrt{T}) \), \( \|H_1\|_F = \frac{1}{NT} \| (G(X) - R(X))' (G(X) - R(X)) \tilde{\Lambda} \|_F = O_P(1) \), which is the dominating term. The result then follows from \( \|K^{-1}\|_2 = O_P(1) \) and Lemma A.2.6.

Lemma A.2.4. Under the assumptions of Theorem 1.3.2,

(i) \( \frac{1}{N} \| \tilde{\Lambda}' (\tilde{\Lambda} - \tilde{\Lambda}H) \|_F = O_P\left( \frac{mN \log^q T}{\sqrt{T}} \right) \).

(ii) \( \frac{1}{N} \| \tilde{\Lambda}' (\tilde{\Lambda} - \tilde{\Lambda}H) \|_F = O_P\left( \frac{mN \log^q T}{\sqrt{T}} \right) \).

Proof. By (A.2.1), \( \tilde{\Lambda}' (\tilde{\Lambda} - \tilde{\Lambda}H) = \sum_{i=1}^{4} \tilde{\Lambda}' A_i K^{-1} \). We evaluate each term in the sum and those can be bounded more tightly. Specifically, by Lemma A.2.2 and A.1.3,

\[
\frac{1}{N} \| \tilde{\Lambda}' A_1 \|_F^2 = \frac{1}{N} \| \tilde{\Lambda}' \tilde{U} \Phi(X) B' \tilde{\Lambda} \|_F^2
\]
\[
\begin{aligned}
&\leq O_P\left(\frac{1}{N^2T^2}\right)\|\Lambda'(\tilde{\Sigma}_u^{-1} - \Sigma_u^{-1})U\Phi(\mathbf{X})B\|_F^2 + \|\Lambda'\Sigma_u^{-1}U\Phi(\mathbf{X})B\|_F^2 \\
&= O_P\left(m_N^2\omega_{N,T}^{2-\eta}/T\right),
\end{aligned}
\]

\[
\frac{1}{N^2}\|\tilde{\Lambda}'A_2\|_F^2 \leq O_P\left(\frac{1}{N^2T^2}\right)\|\Lambda'\tilde{\Sigma}_u^{-1}U\Phi(\mathbf{X})\|_F^2 + \|\Phi(\mathbf{X})\Phi(\mathbf{X})^{-1}\|_F^2\|\Phi(\mathbf{X})\|_F^2 \|\tilde{\Lambda}'A_2\|_F^2 \\
&\leq O_P\left(\frac{J\nu T}{N^2T^2}\right)\|\Lambda'(\tilde{\Sigma}_u^{-1} - \Sigma_u^{-1})U\Phi(\mathbf{X})\|_F^2 + \|\Lambda'\Sigma_u^{-1}U\Phi(\mathbf{X})\|_F^2 \\
&= O_P\left(J\nu T m_N^2\omega_{N,T}^{2-\eta}\right),
\]

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

\[
\begin{aligned}
&\frac{1}{N^2}\|\tilde{\Lambda}'A_3\|_F^2 = \frac{1}{N^4T^2}\|\tilde{\Lambda}'\tilde{U}\tilde{P}\tilde{G}\tilde{\Lambda}\|_F^2 \\
&\leq O_P\left(\frac{1}{N^2T^2}\right)\|\Lambda'\tilde{\Sigma}_u^{-1}U\Phi(\mathbf{X})\|_F^2 \|\Phi(\mathbf{X})\Phi(\mathbf{X})^{-1}\|_F^2 \|\Phi(\mathbf{X})\|_F^2 \|\tilde{\Lambda}'A_3\|_F^2 \\
&\leq O_P\left(\frac{J\nu T}{N^2T^2}\right)\|\Lambda'(\tilde{\Sigma}_u^{-1} - \Sigma_u^{-1})U\Phi(\mathbf{X})\|_F^2 + \|\Lambda'\Sigma_u^{-1}U\Phi(\mathbf{X})\|_F^2 \\
&= O_P\left(J\nu T m_N^2\omega_{N,T}^{2-\eta}\right),
\end{aligned}
\]

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

\[
\frac{1}{N^2}\|\tilde{\Lambda}'(\tilde{\Lambda} - \tilde{\Lambda}H)\|_F = O_P(1)\sum_{i=1}^4 \frac{1}{N^2}\|\tilde{\Lambda}'A_i\|_F^2 = O_P\left(\frac{m_N^2\omega_{N,T}^{2-\eta}}{T}\right).
\]

(ii) The result follows from the following inequality:

\[
\frac{1}{N}\|\tilde{\Lambda}'(\tilde{\Lambda} - \tilde{\Lambda}H)\|_F \leq \frac{1}{N}\|\tilde{\Lambda} - \tilde{\Lambda}H\|_F + \frac{1}{N}\|\tilde{H}'\tilde{\Lambda}'(\tilde{\Lambda} - \tilde{\Lambda}H)\|_F.
\]

\[
\square
\]

Lemma A.2.5. \(\|H'H - I_K\|_F = O_P\left(\frac{m_N\omega_{N,T}^{1-\eta}}{\sqrt{T}}\right)\). Therefore, \(\|H^{-1}\|_2 = O_P(1)\).
Proof. Note that \(N^{-1} \tilde{A}' \tilde{A} = N^{-1} \tilde{A}' \tilde{A} = I_K\), by the identification condition. Then

\[
H'H = \frac{1}{N}(\tilde{A}H')'\tilde{A} = \frac{1}{N}(\tilde{A}H - \hat{A})'\hat{A}H + \frac{1}{N} \hat{A}'(\tilde{A}H - \hat{A}) + I_K.
\]

This implies the following convergence rate,

\[
\|H'H - I_K\|_F \leq \frac{1}{N}\| (\tilde{A}H - \hat{A})'\hat{A}\|_F \|H\|_2 + \frac{1}{N} \| \hat{A}'(\tilde{A}H - \hat{A})\|_F = O_P\left(\frac{mN^{1-q}}{\sqrt{T}}\right).
\]

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

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

Lemma A.2.6. \((i)\) \(\|H_2\|_F = O_P(\sqrt{J^{-\kappa}}) = \|H_3\|_F\), \(\|H_3\|_F = O_P(\sqrt{\nu_T/T}) = \|H_9\|_F\), \(\|H_4\|_F = O_P(1/\sqrt{T})\), \(\|H_12\|_F = O_P(J^\kappa/\sqrt{T})\), \(\|H_6\|_F = O_P(J^{-\kappa})\), \(\|H_10\|_F\), \(\|H_11\|_F = O_P(J^\kappa/\sqrt{T})\).

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

\[
\|H_2\|_F \leq \frac{1}{\sqrt{T}}\|B\Phi(X)'\|_2\|R(X)\|_F\|\tilde{A}\|_F = O_P(\sqrt{J^{-\kappa}}).
\]

Similarly, \(\|H_5\|_F = O_P(\sqrt{J^{-\kappa}})\).

\((i)\) The results follows from Lemma A.2.2. \(\|H_3\|_F = O_P(\sqrt{\nu_T/T}) = O_P(\sqrt{\nu_T/T})\).

Similarly, \(\|H_9\|_F\) attains the same rate. \(\|H_12\|_F = O_P(J^\kappa/\sqrt{T})\). In addition, \(\|H_4\|_F = O_P(\sqrt{T})\), \(\|H_10\|_F\), \(\|H_11\|_F = O_P(J^\kappa/\sqrt{T})\). In addition, \(\|H_6\|_F = O_P(J^{-\kappa})\). In addition,

\[
\|H_8\|_F \leq \frac{1}{\sqrt{T}}\|R(X)\|_F\|P\tilde{U}'\|_F = O_P(1/\sqrt{T})\).
\]

\((i)\) \(\|H_7\|_F = O_P(\sqrt{T})\), \(\|H_9\|_F\) attains the same rate. \(\|H_12\|_F = O_P(\sqrt{\nu_T/T})\). In addition, \(\|H_4\|_F = O_P(\sqrt{T})\), \(\|H_10\|_F\), \(\|H_11\|_F = O_P(J^\kappa/\sqrt{T})\). In addition,

\[
\|H_7\|_F = O_P(\sqrt{T})\|P\|_F = O_P(\sqrt{\nu_T/T})\).
\]

The convergence rate for
\( H_{10} \) can be bounded in the same way. Finally,

\[
\|H_{11}\|_F \leq O_P(\frac{1}{T}\|\Gamma'\Phi(X)\|^2_F\|\Phi(X)'\Phi(X)^{-1}\|_2) = O_P(J\nu_T/T).
\]

\[\square\]

**Lemma A.2.7.** Consider the semiparametric factor model. Under Assumption 1.3.2, there are constants \( c_1, c_2 > 0 \) such that with probability approaching one,

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

**Proof.** Note that \( G(X) = \Phi(X)B' + R(X), \) and

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

We show that all the terms of right hand sides are negligible. By Lemma A.2.2,

\[
\frac{1}{T}G(X)'\Phi(X)' = \frac{1}{T}B\Phi(X)'T + \frac{1}{T}R(X)'\Phi(X)' = O_P(\sqrt{\nu_T/T}).
\]

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

\[
P(\lambda_{\min}(\frac{1}{T}F'PF) > c_1/3) \geq P(\lambda_{\min}(\frac{1}{T}G(X)'G(X)) - \|\frac{1}{T}F'PF - \frac{1}{T}G(X)'G(X)\|_F > c_1/3)
\]

\[
\geq P(\lambda_{\min}(\frac{1}{T}G(X)'G(X)) > \frac{5c_1}{6}, A)
\]

\[
\geq 1 - P(A^c) - P(\lambda_{\min}(\frac{1}{T}G(X)'G(X)) \leq \frac{5c_1}{6}) = 1 - o(1).
\]

In addition,

\[
P(\lambda_{\max}(\frac{1}{T}F'PF) < 2c_2) \geq P(\lambda_{\min}(\frac{1}{T}G(X)'G(X)) + \|\frac{1}{T}F'PF - \frac{1}{T}G(X)'G(X)\|_F > 2c_2)
\]

\[
\geq P(\lambda_{\max}(\frac{1}{T}G(X)'G(X)) < \frac{3c_2}{2}, A)
\]

\[
\geq 1 - P(A^c) - P(\lambda_{\max}(\frac{1}{T}G(X)'G(X)) \leq \frac{3c_1}{2}) = 1 - o(1).
\]
A.2.2 Convergence of factors

Define $\hat{B} = \frac{1}{N} \hat{\Lambda} \hat{\Phi}(X)(\Phi(X)'\Phi(X))^{-1}$. Then

$$\hat{G}(X) = \frac{1}{N} P\hat{Y}'\hat{\Lambda} = \Phi(X)\hat{B}'$$

Substituting $\hat{Y} = \hat{\Lambda}B\Phi(X)' + \hat{\Lambda}R(X)' + \hat{\Lambda}\Gamma' + \hat{U}$, and using $\hat{\Lambda}'\hat{\Lambda}/N = I_K$,

$$\hat{B}' - B'H = \sum_{i=1}^{5} C_i,$$

where

$$C_1 = \frac{1}{N}(\Phi(X)'\Phi(X))^{-1}\Phi(X)'R(X)\hat{n}'\hat{\Lambda}, \quad C_2 = \frac{1}{N}(\Phi(X)'\Phi(X))^{-1}\Phi(X)'\hat{U}'\hat{\Lambda}H,$$

$$C_3 = \frac{1}{N}(\Phi(X)'\Phi(X))^{-1}\Phi(X)'\hat{U}'(\hat{\Lambda} - \hat{\Lambda})H, \quad C_4 = \frac{1}{N}B'\hat{\Lambda}'(\hat{\Lambda} - \hat{\Lambda}H)$$

$$C_5 = \frac{1}{N}(\Phi(X)'\Phi(X))^{-1}\Phi(X')T\hat{\Lambda}'\hat{\Lambda}.$$ 

Lemma A.2.8. (i) $\|C_1\|_F^2 = O_P(\frac{1}{T})$, $\|C_2\|_F^2 = O_P(\frac{1}{T}m^2 N^2 \omega^2/T^2)$,

$\|C_3\|_F^2 = O_P(\frac{1}{T}), \quad \|C_4\|_F^2 = O_P(\frac{1}{T}m^2 N^2 \omega^2/T^2), \quad \|C_5\|_F^2 = O_P(\frac{1}{T}m^2 N^2 \omega^2/T^2)$.

(ii) $\|\hat{B}' - B'H\|_F^2 = O_P(\frac{1}{T} + \frac{J\nu}{T} + \frac{1}{T}m^2 N^2 \omega^2/T^2)$.

Proof. (i) By Lemmas A.1.3, A.2.1, A.2.2 and A.2.4,

$$\|C_1\|_F^2 \leq O_P(\|(\Phi(X)'\Phi(X))^{-1}\frac{1}{2}\|\Phi(X)'\|_2^2\|\Phi(X)'\Phi(X)'\|_2) = O_P(J^{-\kappa})$$

$$\|C_2\|_F^2 \leq O_P(\frac{1}{N^2}\|(\Phi(X)'\Phi(X))^{-1}\frac{1}{2}\|\Phi(X)'\Phi(X)'\hat{\Sigma}^{-1}\|_2^2)$$

$$\leq O_P\left(\frac{1}{N^2T^2}\|(\Phi(X)'\Phi(X))^{-1}\frac{1}{2}\|\Phi(X)'\Phi(X)'\hat{\Sigma}^{-1}\|_2^2\right) = O_P(Jm^2 \omega^2/T^2),$$

$$\|C_3\|_F^2 \leq O_P\left(\frac{1}{N^2T^2}\|\Phi(X)'\Phi(X)'\|_2^2\|\hat{\Lambda} - \hat{\Lambda}H\|_F^2\right) = O_P(J/T^2),$$

$$\|C_4\|_F^2 \leq \|B\|_F^2 \left(\frac{1}{N}\hat{n}'(\hat{\Lambda} - \hat{\Lambda}H)\right)^2 = O_P(Jm^2 \omega^2/T^2),$$

$$\|C_5\|_F^2 \leq O_P(\|(\Phi(X)'\Phi(X))^{-1}\frac{1}{2}\|\Phi(X)'\Phi(X)'\|_2^2) = O_P(Jm^2 \omega^2/T^2).$$
Lemma A.2.9.

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

\[
\|\hat{B}' - B'H\|_F^2 \leq O(1) \sum_{i=1}^5 \|C_i\|_F^2 = O_P \left( \frac{1}{J^\kappa} + \frac{J\nu_I}{T} + \frac{J}{T} \right) m_{N,\omega_{N,T}}^{2-2q}.
\]

Because \(G(X)H = \Phi(X)B'H + R(X)H\), by Lemma A.2.8

\[
\frac{1}{T}\|\hat{G}(X) - G(X)H\|_F^2 \leq \frac{2}{T}\|\Phi(X)(\hat{B}' - B'H)\|_F^2 + \frac{2}{T}\|R(X)H\|_F^2
\]

\[= O_P(\|\hat{B}' - B'H\|_F^2 + J^{-\kappa}) \leq O_P \left( \frac{1}{J^\kappa} + \frac{J\nu_I}{T} + \frac{J}{T} m_{N,\omega_{N,T}}^{2-2q} \right).
\]

Substituting \(\bar{Y} = \bar{A}B\Phi(X)' + \bar{A}R(X)' + \bar{A}\Gamma' + \bar{U} + \bar{U}'\) into \(\hat{\Gamma} = \frac{1}{N}(I - P)\hat{Y}\hat{A},\)

\[\hat{\Gamma} - \Gamma H = \sum_{i=1}^6 D_i,
\]

where

\[
D_1 = \frac{1}{N}(I - P)\Gamma' (\hat{\Lambda} - \tilde{\Lambda}H), \quad D_2 = \frac{1}{N} \tilde{U}' (\hat{\Lambda} - \tilde{\Lambda}H),
\]

\[
D_3 = -\Gamma' H, \quad D_4 = (I - P)R(X)(H + \frac{1}{N}\hat{\Lambda}' (\hat{\Lambda} - \tilde{\Lambda}H)),
\]

\[
D_5 = -\frac{1}{N}P\tilde{U}' (\hat{\Lambda} - \tilde{\Lambda}H), \quad D_6 = \frac{1}{N}(I - P)\tilde{U}' \hat{\Lambda}H.
\]

Then, by Lemma A.2.11,

\[
\frac{1}{T}\|\hat{\Gamma} - \Gamma H\|_F^2 \leq O_P \left( \frac{1}{T} \sum_{i=1}^6 \|D_i\|_F^2 \right) = O_P \left( \frac{1}{N} + \frac{J}{T^2} + \frac{1}{J^\kappa} + \frac{J\nu_I}{T} + \frac{J}{T} m_{N,\omega_{N,T}}^{2-2q} \right).
\]

Lemma A.2.9. (i) \(\|U'U\Phi(X)\|_F^2 = O_P(JNT^2 + N^2T)\), and \(\|U'U\Phi(X)B'\|_F^2 = O_P(NT^2 + N^2T)\).

(ii) \(\|U'\Sigma_u^{-1}U\Phi(X)\|_F^2 = O_P(JNT^2 + N^2T)\), and \(\|U'\Sigma_u^{-1}U\Phi(X)B'\|_F^2 = O_P(NT^2 + N^2T)\).

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

Then

\[
E\|U'U - EU'U\Phi(X)\|_F^2 = \sum_{i=1}^T \sum_{l=1}^d \sum_{j=1}^J \text{var} \left( \sum_{s=1}^T \phi_j(X_{st})s_{is} \right)
\]
where

\[
\sum_{t=1}^{T} \sum_{d=1}^{D} \sum_{j=1}^{J} \sum_{t=1}^{T} E\phi_j(X_{sl})^2E s_t^2 + \sum_{t=1}^{T} \sum_{d=1}^{D} \sum_{j=1}^{J} \sum_{s\neq q,s,q\neq T} E\phi_j(X_{sl})\phi_j(X_{ql})\text{cov}(s_{ts}, s_{tq})
\]

by Assumption 1.3.4. Example (iv) follows from Lemma A.2.10. \(\|\cdot\|\) follows from Lemma A.2.11.

Proof. Given Lemma A.2.9, the proof is straightforward calculations.

\[\text{Lemma A.2.10. (i) } \|\frac{1}{N}\hat{U}'A_1\|_F^2 = O_P\left(\frac{1}{N} + \frac{1}{T}\right), \|\frac{1}{N}\hat{U}'A_2\|_F^2 = O_P\left(\frac{1}{N\sqrt{J}} + \frac{1}{\sqrt{T}}\right),
\]

\[\text{(ii) } \|\frac{1}{N}\hat{U}'A_3\|_F^2 = O_P\left(\frac{J_n\sqrt{m}}{N} + \frac{J}{T}\right), \|\frac{1}{N}\hat{U}'A_4\|_F^2 = O_P\left(\frac{J^2}{NT} + \frac{J}{T^2}\right).\]

Proof. Given Lemma A.2.9, the proof is straightforward calculations.

\[\text{Lemma A.2.11. (i) } \|\frac{1}{N}(I - P)\Gamma \hat{A}'\hat{A} - \hat{A}\|_F^2 = O_P(\nu_T m^2 N^{2 -2q}),
\]

\[\text{(ii) } \|\frac{1}{N}\hat{U}'(\hat{\Lambda} - \hat{A})\|_F^2 = O_P\left(\frac{1}{N} + \frac{1}{T}\right),
\]

\[\text{(iii) } \|\hat{P}\hat{\Gamma}\hat{H}\|_F^2 = O_P(J_T),
\]

\[\text{(iv) } \|\hat{U}'(\hat{\Lambda} - \hat{A})\|_F^2 = O_P\left(\frac{T}{N}\right),
\]

\[\text{(v) } \|\frac{1}{N}P\hat{U}'(\hat{\Lambda} - \hat{A})\|_F^2 = O_P\left(\frac{T}{N}\right),
\]

\[\text{(vi) } \|\frac{1}{N}(I - P)\hat{U}'\hat{A}\|_F^2 = O_P\left(\frac{T}{N} + m^2 N^{2 - 2q}\right).\]

Proof. Note that \(\|\hat{\Gamma}\|_F^2 = O_P(T\nu_T)\) and \(\|\hat{R}(\hat{X})\|_F^2 = O_P(TJ^{-\kappa})\). (i), (iii)-(v) follow from Lemmas A.1.3, A.2.2 and A.2.4. (ii) follows from Lemma A.2.10. (vi) follows from Cauchy-Schwarz inequality.

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

\[
\hat{F} = \hat{F} + \hat{F} = \sum_{i=1}^{5} E_i,
\]
Lemma A.2.12. (i) \( \| \frac{1}{N} \Phi(X) B \tilde{A}'(\tilde{\Lambda} - \tilde{\Lambda}H) \|_F^2 = O_P(\frac{1}{N} \nu T^2 m_2^2 \omega_{N,T}^2) \),
(ii) \( \| \frac{1}{N} R(X) \tilde{A}'(\tilde{\Lambda} - \tilde{\Lambda}H) \|_F^2 = O_P(\frac{1}{N} \nu T^2 m_2^2 \omega_{N,T}^2) \),
(iii) \( \| \frac{1}{N} \Gamma \tilde{A}'(\tilde{\Lambda} - \tilde{\Lambda}H) \|_F^2 = O_P(\nu T^2 m_2^2 \omega_{N,T}^2) \),
(iv) \( \| \frac{1}{N} \tilde{U}'(\tilde{\Lambda} - \tilde{\Lambda}H) \|_F^2 = O_P(\frac{1}{N} \nu T^2 m_2^2 \omega_{N,T}^2) \),
(v) \( \| \frac{1}{N} \tilde{U}' \tilde{\Lambda}H \|_F^2 = O_P(\frac{N^2}{N} T^2 m_2^2 \omega_{N,T}^2) \).

Proof. Note that \( \| \Phi(X) B \|_2 \leq \| G(X) \|_2 + \| R(X) \|_2 = O_P(\sqrt{T}) \), and \( \| R(X) \|_F^2 = O_P(TJ^{-\nu}) \). In addition, \( \| N^{-1} U^t \Lambda^t H \|_2^2 = O_P(N^{-2} T^2 \sum_{t=1}^T E(\| \Lambda^t u_t \|^2) = O_P(T/N) \).

(i)-(iii), (v) follow from Lemmas A.1.3, A.2.2 and A.2.4. (iv) follows from Lemma A.2.10.

A.2.3 Individual factors

Since \( \tilde{F} = \frac{1}{N} \tilde{Y}' \tilde{A} \), \( \tilde{F}_t = \frac{1}{N} \tilde{\Lambda}^- \Lambda^t F_t + \frac{1}{N} \tilde{\Lambda}^- \Lambda^t u_t \). Using \( \tilde{\Lambda} = \tilde{\Lambda} - \tilde{\Lambda}H^{-1} + \tilde{\Lambda}H^{-1} \) and \( \frac{1}{N} \tilde{\Lambda} \Lambda = I_K \), we have

\[ \tilde{F}_t - H^{-1} F_t = \sum_{i=1}^3 W_i, \]

where

\[ W_1 = \frac{1}{N} \tilde{\Lambda}^- (\tilde{\Lambda}H - \tilde{\Lambda}) H^{-1} F_t, \quad W_2 = \frac{1}{N} (\tilde{\Lambda} - \tilde{\Lambda}H) u_t, \quad W_3 = \frac{1}{N} H^t \tilde{\Lambda} u_t. \]

Then, by Lemmas A.2.4 and A.2.5,

\[ \| W_1 \| \leq \frac{1}{N} \tilde{\Lambda}^- (\tilde{\Lambda} - \tilde{\Lambda}H) \| F \| H^{-1} \| F_t \| = O_P(\frac{1}{\sqrt{T}} m_2^{1-\nu}) \].
Note that \( \|N^{-1}\Lambda'\mathbf{u}_t\| = O_P(1/\sqrt{N}) \) by Lemma A.1.8. Let \( \Lambda^* = \Sigma_u^{-\frac{1}{2}}\Lambda \) and \( \mathbf{u}_t^* = \Sigma_u^{-\frac{1}{2}}\mathbf{u}_t \). Then \( \|N^{-1}\Lambda^*\mathbf{u}_t^*\| = O_P(1/\sqrt{N}) \). By Lemma A.2.3,

\[
\|W_3\| \leq \frac{1}{N}\|H\|_2(\|A'(\Sigma_u^{-1} - \Sigma_u^{-1})\mathbf{u}_t\| + \|A'\mathbf{u}_t\|) \\
\leq O_P\left(\frac{1}{N}\|A\|_F\|\Sigma_u^{-1} - \Sigma_u^{-1}\|_1\|\mathbf{u}_t\| + \|A'\mathbf{u}_t\|\right) \\
\leq O_P(m_N\omega_{N,T}^{1/2} + \frac{1}{\sqrt{N}}) = O_P(m_N\omega_{N,T}^{1/2}).
\]

For each fixed \( t \), it follows from \( \frac{1}{N}\|\tilde{A} - \tilde{A}H\|_2^2 = O_P(T^{-1}) \) and \( \frac{1}{N}\sum_{i=1}^N u_{it}^2 = O_P(1) \) that \( \|W_2\| = O_P(T^{-1/2}) \). Therefore, for each \( t \leq T \),

\[
\|\hat{F}_t - H^{-1}F_t\| = O_P(m_N\omega_{N,T}^{1/2}).
\]

### A.3 Proofs for Section 1.3.4

#### A.3.1 Proof of Theorem 1.3.3

**Lemma A.3.1.** Let \( L_t = (F_t', W_t')' \) and \( \hat{L}_t = (\hat{F}_t', W_t')' \). Under Assumptions 1.3.1, 1.3.3-1.3.8, we have, for \( \omega_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}} \).

1. \( \frac{1}{T}\sum_{t=1}^T (\hat{F}_t - H^{-1}F_t)L_t' = O_P(m_N^2\omega_{N,T}^{2-2q}) \),
2. \( \frac{1}{T}\sum_{t=1}^T (\hat{F}_t - H^{-1}F_t)\hat{L}_t' = O_P(m_N^2\omega_{N,T}^{2-2q}) \),
3. \( \frac{1}{T}\sum_{t=1}^T (\hat{F}_t - H^{-1}F_t)\epsilon_{t+h} = O_P(m_N^2\omega_{N,T}^{2-2q}) \).

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

\[
\frac{1}{T}\sum_{t=1}^T (\hat{F}_t - H^{-1}F_t)L_t' = \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N \tilde{\lambda}_i(H'\tilde{\lambda}_i - \tilde{\lambda}_i)'H^{-1}F_tL_t' \\
+ \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N \tilde{u}_{it}(\tilde{\lambda}_i - H'\tilde{\lambda}_i)L_t' + \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N H'\tilde{\lambda}_i\tilde{u}_{it}L_t' \\
= (I) + (II) + (III).
\]

For (I), we have

\[
I = \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N (\hat{\lambda}_i - H'\tilde{\lambda}_i)(\hat{\lambda}_i - H'\tilde{\lambda}_i)'H^{-1}F_tL_t' + \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N H'\tilde{\lambda}_i(\hat{\lambda}_i - H'\tilde{\lambda}_i)'H^{-1}F_tL_t'.
\]
Note that, by Assumption 1.3.8, \( T^{-1} \sum_{t=1}^{T} (E\|\mathbf{F}_t\|^2)^{1/2} (E\|L_t\|^2)^{1/2} = O(1) \). Also, by Theorem 1.3.2, \( N^{-1} \sum_{i=1}^{N} \|\hat{\lambda}_i - \mathbf{H}'\hat{\lambda}_i\|^2 = O_P(T^{-1}) \). Then the first term is bounded by
\[
\frac{1}{N} \sum_{i=1}^{N} \|\hat{\lambda}_i - \mathbf{H}'\hat{\lambda}_i\|^2 H^{-1} \frac{1}{T} \sum_{t=1}^{T} \mathbf{F}_t L_t' = O_P(T^{-1}).
\]
by the result of Lemma A.2.5. Similarly, the second term is bounded by \( O_P(T^{-1/2} m_N \omega_{N,T}^{1/2}) \) from the result in Lemma A.2.4. Thus, \((I) = O_P(T^{-1/2} m_N \omega_{N,T}^{1/2})\).

For \((II)\), note that we have \( \frac{1}{N} \sum_{i=1}^{N} |\hat{\lambda}_i - \mathbf{H}'\hat{\lambda}_i|_{\mathbb{F}_t} |u_{it}| = O_P(m_N^2 \omega_{N,T}^{2-2q}) \) 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 \( \mathbf{H}' \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{\lambda}_i L_t' \hat{u}_{it} = O_P(1) O_P(\frac{1}{\sqrt{NT}}) \). Therefore, we have
\[
I + II + III = O_P(T^{-1/2} m_N \omega_{N,T}^{1/2}) + O_P(m_N^2 \omega_{N,T}^{2-2q}) + O_P(\frac{1}{\sqrt{NT}}) = O_P(m_N^2 \omega_{N,T}^{2-2q}).
\]
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 \( \epsilon_t \) instead of \( L_t \).

The forecasting model (1.3.8) can be written as
\[
z_{t+h} = \alpha' \mathbf{F}_t + \beta' \mathbf{W}_t + \epsilon_{t+h}
= \alpha' \mathbf{H}^{-1} \hat{\mathbf{F}}_t + \beta' \mathbf{W}_t + \epsilon_{t+h} + \alpha' \mathbf{H}^{-1} (\mathbf{H}' \mathbf{F}_t - \hat{\mathbf{F}}_t)
= \hat{\mathbf{L}}' \delta + \epsilon_{t+h} + \alpha' \mathbf{H}^{-1} (\mathbf{H}' \mathbf{F}_t - \hat{\mathbf{F}}_t).
\]
In matrix notation, the model can be rewritten as
\[
Z = \hat{\mathbf{L}} \delta + \epsilon + (\mathbf{F} \mathbf{H} - \hat{\mathbf{F}}) \mathbf{H}^{-1} \alpha,
\]
where \( Z = (z_{h+1}, ..., z_T)' \), \( \epsilon = (\epsilon_{h+1}, ..., \epsilon_T)' \), and \( \hat{\mathbf{L}} = (\hat{\mathbf{L}}_1, ..., \hat{\mathbf{L}}_{T-h})' \). Then, the ordinary least squares estimator of \( \delta \) is \( \hat{\delta} = (\hat{\mathbf{L}}' \hat{\mathbf{L}})^{-1} \hat{\mathbf{L}}' Z \). Note that by Lemma A.3.1, \( T^{-1/2} \hat{\mathbf{L}}' (\mathbf{F} \mathbf{H} - \hat{\mathbf{F}}) = O_P(T^{1/2} m_N^2 \omega_{N,T}^{2-2q}) \) and \( T^{-1/2} (\mathbf{F} - \mathbf{H})' \epsilon = O_P(T^{1/2} m_N^2 \omega_{N,T}^{2-2q}) \). Then, if \( T^{1/2} m_N^2 \omega_{N,T}^{2-2q} \rightarrow 0 \),
\[
\sqrt{T} (\hat{\delta} - \delta) = (T^{-1/2} \hat{\mathbf{L}}' \hat{\mathbf{L}})^{-1} T^{-1/2} \hat{\mathbf{L}}' \epsilon + (T^{-1/2} \hat{\mathbf{L}}' \hat{\mathbf{L}})^{-1} T^{-1/2} \hat{\mathbf{L}}' (\mathbf{F} - \hat{\mathbf{F}}) \mathbf{H}^{-1} \alpha
\]
\[
\begin{align*}
= (T^{-1}\hat{L}'\hat{L})^{-1}T^{-1/2}\hat{L}'\epsilon + o_P(1),
\end{align*}
\]

and
\[
\frac{\hat{L}'\epsilon}{\sqrt{T}} = \begin{bmatrix}
\frac{\hat{F} - FH'\epsilon}{\sqrt{T}} + \frac{HF'\epsilon}{\sqrt{T}} \\
\frac{W'\epsilon}{\sqrt{T}}
\end{bmatrix} = \begin{bmatrix}
H'F'\epsilon \\
W'\epsilon
\end{bmatrix} + o_P(1).
\]

Then, for a block diagonal matrix $\Pi = \text{diag}(H', I)$,
\[
\sqrt{T}(\hat{\delta} - \delta) = (T^{-1}\hat{L}'\hat{L})^{-1}\Pi(T^{-1/2}L'\epsilon) + o_P(1).
\]

Since $T^{-1/2}L'\epsilon \overset{d}{\rightarrow} N(0, \Sigma_{L,\epsilon})$ by Assumption 1.3.8, we obtain the result in Theorem 1.3.3.

### A.3.2 Proof of Theorem 1.3.4

From Theorem 1.3.2, note that $\|\hat{F}_t - H^{-1}F_t]\| = O_P(m_N\omega_{N,T}^{1-\eta})$ and $\sqrt{T}(\hat{\delta} - \delta)$ is asymptotically normal. Therefore,
\[
\hat{z}_{T+h|T} - z_{T+h|T} = \alpha'\hat{F}_T + \beta'W_T - \alpha'F_T - \beta'W_T
\]
\[
= (\hat{\alpha} - H'\alpha)'\hat{F}_T + \alpha'H(\hat{F}_T - H^{-1}F_t) + (\hat{\beta} - \beta)'W_T
\]
\[
= \frac{1}{\sqrt{T}}\hat{L}'_T\sqrt{T}(\hat{\delta} - \delta) + \alpha'H(\hat{F}_T - H^{-1}F_t)
\]
\[
= O_P(m_N\omega_{N,T}^{1-\eta}).
\]
A.4 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. (2021) and Gu et al. (2020).

A.4.1 Univariate autoregression model

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

\[ r_{x_{t+1}}^{(n)} = \alpha + \phi(L)r_{x_t}^{(n)} + \epsilon_{t+1}, \]

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.

A.4.2 Diffusion index models

This paper considers Diffusion Index models of the following form:

\[ r_{x_{t+1}}^{(n)} = \alpha + \beta_W' W_t + \beta_F' F_t + \epsilon_{t+1}, \]  \hspace{1cm} (A.4.1)

where \( W_t \) is observable variables, \( \epsilon_t \) is a disturbance term, and \( \alpha \) and \( \beta \) are parameters estimated using least squares. Since \( F_t \) 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 Oka, 2014; Stock and Watson, 2014, etc). Varying the components of \( W_t \) yields different type of models as follows.
Principal components regression

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

$$\hat{r}_{x_{t+1}}^{(n)} = \hat{\alpha} + \hat{\beta}_F \hat{F}_t,$$

where $\hat{F}_t$ is estimated using principal component methods using $\{Y_i\}_{i=1}^T$, which is 130 macroeconomic variables. Note that PCR is a special case of DI model by replacing $W_t$ with constant term in equation (A.4.1).

Factor augmented autoregression

Based on equation (A.4.1), forecasts are computed as

$$\hat{r}_{x_{t+1}}^{(n)} = \hat{\alpha} + \hat{\beta}_W(L) \hat{r}_{x_{t}}^{(n)} + \hat{\beta}_F' \hat{F}_t$$

where $\hat{F}_t$ 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.

Diffusion index

Diffusion index forecasts are computed as

$$\hat{r}_{x_{t+1}}^{(n)} = \hat{\alpha} + \hat{\beta}_W W_t + \hat{\beta}_F' \hat{F}_t$$

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

A.4.3 Modified diffusion index models

Instead of the conventional diffusion index model, I also conduct some statistical learning methods for estimating $\hat{\beta}_F$. For example, Kim and Oka (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.

**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 $\hat{Y}_b^* = \hat{\beta}_b^* X_b^*$ be a bootstrap sample based predictor for $b = 1, \ldots, B$ denotes the $b$-th bootstrap sample. Then the bagging predictor is $\hat{Y}_{\text{bagging}} = \frac{1}{B} \sum_{b=1}^{B} \hat{Y}_b^*$.

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:

$$\hat{r}_t^{(n)} = \hat{\alpha} + \hat{\beta}_W W_t + \sum_{j=1}^{r} \psi(t_j) \hat{\beta}_{Fj} \hat{F}_{t,j},$$

where $\hat{\alpha}, \hat{\beta}$ are the least squares estimator from a regression of $r_x_{t+1}^{(n)}$ on $W_t$, $\hat{\beta}_{Fj}$ is a least squares estimator from a regression of residuals, $Z_t = r_x_{t+1}^{(n)} - \hat{\alpha} - \hat{\beta}_W W_t$ on $\hat{F}_{t,j}$, and $t_j$ is the $t$-statistic associated with $\hat{\beta}_{Fj}$. Specifically, $t_j = \sqrt{T} \hat{\beta}_{Fj}/s_e$, where $s_e$ is a Newey-West standard error. In addition, I follow Stock and Watson (2012) and Kim and Oka (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, $\Phi$ is the standard normal CDF, and $\phi$ is the standard normal density with $c = 1.96$. Here I set $W_t$ be the CP factor with lags of $r_x_{t}^{(n)}$, and the number of lags $p$ is determined using the SIC.

**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

$$\hat{Y}^M = \sum_{m=1}^{M} \gamma_m f(X; \beta_m),$$
where the $\gamma_m, m = 1, \ldots, M$ 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, $\hat{F}_t$, and denote $\hat{\mu}^M(\hat{F}_t)$ as the output of boosting algorithm. Finally, the boosting estimator is

$$\hat{r}_x^{(n)}_{t+1} = \hat{\alpha} + \hat{\beta}'_W W_t + \hat{\mu}^M(\hat{F}_t),$$

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

**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 following model:

$$r_x^{(n)}_{t+1} = \alpha + \beta'_WW_t + \beta'_F F_t + \theta'u_t + \epsilon_{t+1},$$

$$y_t = \Lambda F_t + u_t, \quad t = 1, \ldots, T,$$

where $\theta$ 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 $r_x^{(n)}_{T+1|T}$.

Predictions are constructed using the following steps:

1. Obtain $\{\hat{F}_t, \hat{u}_t\}_{t \leq T}$ by principal component methods from the factor model.
2. Estimate the diffusion index model, and obtain
\[ \hat{z}_{t+1} = rx_{t+1}^{(n)} - \hat{\alpha} + \hat{\beta}'_W W_t + \hat{\beta}'_F \hat{F}_t. \]

3. Estimate \( \theta \) using lasso on
\[ \hat{z}_{t+1} = \theta \hat{u}_t + \epsilon_{t+1}. \]

4. Then forecasts are computed as
\[ \hat{r}x_{t+1}^{(n)} = \hat{\alpha} + \hat{\beta}'_W W_t + \hat{\beta}'_F \hat{F}_t + \hat{\theta}' \hat{u}_t. \]

**A.4.4 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.

**Ridge regression**

Ridge regression, which is introduced by Hoerl and Kennard (1970), solves the following problem:
\[
\min L(\lambda, \theta) = \sum_{t=1}^{T-1} [rx_{t+1}^{(n)} - \alpha - \beta' y_t]^2 + \lambda \sum_{j=1}^{p} \beta_j^2,
\]
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 coefficient and shrinks them toward zero.

**Lasso regression**

Lasso regression, which is introduced by Tibshirani (1996), solves the following problem:
\[
\min L(\lambda, \theta) = \sum_{t=1}^{T-1} [rx_{t+1}^{(n)} - \alpha - \beta' y_t]^2 + \lambda \sum_{j=1}^{p} |\beta_j|,
\]
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 coefficient and exactly set to zero.

**Elastic net**

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

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

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, \( \lambda \), \( \lambda_1 \), and \( \lambda_2 \), 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.

**A.4.5 Regression trees**

Regression trees are popular in the machine learning literature, because 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 \( \sum (x_{r_{t+1}}^{(n)} - f(y_t))^2 \), we can estimate \( \hat{c}_m \) by averaging the excess bond returns \( x_{r_{t+1}} \) in region \( R_m \) as following:

\[
\hat{c}_m = E[x_{r_{t+1}}^{(n)} | y_t \in R_m].
\]
Since finding the optimal binary partition via minimum sum of squares is infeasible, I proceed the following tree-based method.

**Decision tree**

I first introduce a popular method 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 regression 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 $\bar{R}$ in $\{R_l(d), l = 1, ..., 2^{d-1}\}$:
     
     Given splitting variable $j$ and each threshold level $\tau$, define a split regions $R_1(j, \tau) = \{Z | Z_j \leq \tau, Z_j \cap \bar{R}\}$ and $R_2(j, \tau) = \{Z | Z_j > \tau, Z_j \cap \bar{R}\}$.

     In the splitting regions set
     
     $$c_1(j, \tau) \leftarrow \frac{1}{|R_1(j, \tau)|} \sum_{y_t \in R_1(j, \tau)} r_{x_{t+1}}^{(n)}(y_t)$$ and $c_2(j, \tau) \leftarrow \frac{1}{|R_2(j, \tau)|} \sum_{y_t \in R_2(j, \tau)} r_{x_{t+1}}^{(n)}(y_t)$.

     Select the optimal split:

     $$(j^*, \tau^*) = \arg\min_{j, \tau} \left[ \sum_{y_t \in R_1(j, \tau)} (r_{x_{t+1}}^{(n)} - c_1(j, \tau))^2 + \sum_{y_t \in R_2(j, \tau)} (r_{x_{t+1}}^{(n)} - c_2(j, \tau))^2 \right].$$

     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:

$$\hat{f}(y_t) = \sum_{k=1}^{2^L} \text{avg}(r_{x_{t+1}}^{(n)} | y_t \in R_k(L))I(y_t \in R_k(L)).$$
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 $f_0(y_t) = \arg \min_\theta \sum_{i=1}^{N} L(rx_{t+1}^{(n)}, \theta)$. Let $L(\cdot, \cdot)$ be the loss function.\(^1\)

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:
      $$
      \hat{\theta}_{jm} = \arg \min_\theta \sum_{y_t \in S_{jm}} L(rx_{t+1}^{(n)}, f_{m-1}(y_t) + \theta).
      $$

   (c) Update the learner $f_m(y_t) = f_{m-1}(y_t) + \nu \sum_{j=1}^{J_m} \hat{\theta}_{jm} I(y_t \in S_{jm})$, where $\nu \in (0, 1]$ is a hyperparameter.

3. The gradient boosted regression tree output is
   $$
   \hat{f}(y_t) = f_M(y_t).
   $$

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 $\nu$ 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 de-correlated trees which are estimated separately and then averaged out.

**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
   
   $$\hat{f}(y_t) = \frac{1}{B} \sum_{b=1}^{B} T_b(y_t).$$

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 optimized by cross-validation. To reduce the computa-
tional costs, the hyperparameters are determined using the whole sample first before
we conduct the out-of-sample forecast.

**A.4.6 Neural networks**

Neural network models, which emulate the neural architecture of brain and its pro-
cesses, 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

\[ r_{x_{t+1}}^{(n)} = b_0 + \sum_{j=1}^{q} b_j G(\gamma_j y_t + \gamma_{0j}), \]

where \( G(\cdot) \) is a nonlinear “activation function”, \( q \) is the fixed number of hidden units, and \( y_t \) is the \( p \) dimensional input.\(^2\) 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 surprisingsly, 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

\[
\begin{align*}
Z^{(1)} &= G^{(1)}(b^{(0)} + W^{(0)} Y), \\
Z^{(2)} &= G^{(2)}(b^{(1)} + W^{(1)} Z^{(1)}), \\
&\vdots \\
Z^{(L)} &= G^{(L)}(b^{(L-1)} + W^{(L-1)} Z^{(L-1)}), \\
r_{x_{t+1}}^{(n)} &= b^{(L)} + W^{(L)} Z^{(L)};
\end{align*}
\]

where \( G^{(l)} \) is activation function, \( W^{(l)} \) is weight matrices, and \( b^{(l)} \) is activation levels for \( l = 1, \ldots, L \). 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.

---

\(^2\) 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.
choose sigmoid function as activation function at all nodes, that is,

\[ G(z) = \frac{1}{1 + \exp(-z)}. \]

Simple network models with a few layers and nodes in each layer for small dataset. In general, however, selecting a best network architecture using cross-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 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. (2021).

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

**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.

**Hybrid Neural Network**

A hybrid class of models that combines the least squares with neural networks is considered. 130 macroeconomic variables with a linear combination 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:

\[ r x_{t+1}^{(n)} = \alpha + \beta CP_t + u_{t+1}. \]

In the second step, the residual, \( \hat{u}_{t+1} \), 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 A.4.6.

**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 procedure 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 \( \{\hat{F}_t\}_{t=1,...,T} \) using the PC methods (e.g., Stock and Watson, 2002a).

- **Step 2:** Denote \( \hat{L}_t = (\hat{F}_t', W_t')' \). We estimate \( g(\cdot) \) using the smooth sigmoid neural network sieve estimator. Specifically, \( \hat{g}(\hat{L}_t) = \arg\max L_n(\theta) \), where \( L_n(\theta) = -T^{-1} \sum_{t=1}^{T} \frac{1}{2} [z_{t+h} - g(\hat{L}_t)]^2 \).

- **Step 3:** Forecast \( z_{T+h} \) using \( \hat{L}_T \).
Appendix B

Appendix for Chapter 2

B.1 Proofs

Throughout the proof, \( \max_i, \max_t, \max_h, \max_{ij}, \max_{it}, \sum_i, \sum_t, \) and \( \sum_{ij} \) denote \( \max_{i \leq N}, \max_{t \leq T}, \max_{h \leq L}, \max_{i,j}, \max_{i,t}, \sum_{i=1}^N, \sum_{t=1}^T, \) and \( \sum_{i=1}^N \sum_{j=1}^N \) respectively.

B.1.1 Technical lemmas

First let

\[
V_L = \frac{1}{NT} \sum_t E x'_t u_t u'_t x_t + \frac{1}{NT} \sum_{h=1}^L \omega(h, L) \sum_{t=h+1}^T \left[ E x'_t u_t u'_{t-h} x_{t-h} + E x'_{t-h} u_{t-h} u'_t x_t \right].
\]

We need following lemmas to prove the main results.

**Lemma B.1.1.** (i) \( \|V - V_L\| \leq C \sum_{h=L}^{T-1} \alpha_{NT}(h) + C \sum_{h=1}^L (1 - \omega(h, L)) \alpha_{NT}(h) \).

(ii) \( \max_i |V_{u,ii} - \text{var} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T x_{it} u_{it} \right) | = o(1) \).

(iii) \( \min_i \lambda_{\min}(V_{u,ii}) > c \).

**Proof.** (i) First note that

\[
\|E x'_t u_t u'_{t-h} x_{t-h} + E x'_{t-h} u_{t-h} u'_t x_t \| \leq E \|x_t\| E(u_t u'_{t-h}|X) \|x_{t-h}\| + E \|x_{t-h}\| E(u_{t-h} u'_t|X) \|x_t\|
\]

\[
\leq \alpha_{NT}(h) E \|x_t\| \|x_{t-h}\| \leq CN \alpha_{NT}(h).
\]
Hence for some $C, c > 0,$

$$\|V - V_L\| \leq \left\| \frac{1}{NT} \sum_{h=1}^{T} \sum_{t=0}^{T} \left[ Ex_{t} u_{t-h} x_{t-h} + Ex_{t} u_{t-h} u_{t} x_{t} \right] \right\| + \left\| \frac{1}{NT} \sum_{h=1}^{L} (1 - \omega(h, L)) \sum_{t=0}^{T} \left[ Ex_{t} u_{t-h} x_{t-h} + Ex_{t} u_{t-h} u_{t} x_{t} \right] \right\| \leq C \frac{1}{T} \sum_{h=1}^{L} \alpha_{NT}(h^c) + C \frac{1}{T} \sum_{h=1}^{L} (1 - \omega(h, L)) \sum_{t=0}^{T} \alpha_{NT}(h^c) \leq C \sum_{h \geq L} \alpha_{NT}(h^c) + C \sum_{h=1}^{L} |1 - \omega(h, L)| \alpha_{NT}(h^c) = o(1).$$

The second term of the last equation goes to zero due to Assumption 2.2.2(iii) and the dominated convergence theorem, noting that $|1 - \omega(h, L)| \alpha_{NT}(h^c) \leq C \alpha_{NT}(h^c)$ and $\alpha_{NT}(h^c)$ is summable over $h$.

(ii) The proof for $\max_i |V_{n, i} - \var(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{it} u_{it})| = o(1)$ follows from the same argument.

(iii) The result follows from (ii) and the assumption that $\min_i \lambda_{\min}(\var(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{it} u_{it})) > c$.

**Lemma B.1.2.** Suppose $\log N = o(T)$. For $f(t, h, L) = \omega(h, L) 1_{\{t > h\}},$

$$\max_{h} \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} u_{it} \varepsilon_{jt, t-h} x'_{jt, t-h} f(t, h, L) - Ex_{it} u_{it} \varepsilon_{jt, t-h} x'_{jt, t-h} f(t, h, L) \right\| = O_P\left(\sqrt{\log(N) \frac{\log(LN)}{T}}\right).$$

**Proof.** The left hand side can be written as

$$\max_{h} \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} Z_{h, i,j, t} \right\|,$$

where $Z_{h, i,j, t} = f(t, h, L)(x_{it} \varepsilon_{jt, t-h} x'_{jt, t-h} - Ex_{it} \varepsilon_{jt, t-h} x'_{jt, t-h})$.

For convenience, assume that $\dim(Z_{h, i,j, t}) = 1$ and there is no serial correlation. Set $\alpha_n = \sqrt{\log(LN) \frac{\log(LN^2)}{T}}$ and $c^2 = 2C$ for $c, C > 0$. Then, by using Bernstein inequality and exponential tail conditions (e.g., Merlevède et al., 2011), and that $f(t, h, L)$ is bounded,

$$P(\max_{h \leq L} \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} Z_{h, i,j, t} \right\| > c\alpha_n) \leq LN^2 \max_{h \leq L} \max_{i,j} P\left(\left\| \frac{1}{T} \sum_{t=1}^{T} Z_{h, i,j, t} \right\| > c\alpha_n\right) \to 0.$$  

Then $\max_{h} \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} Z_{h, i,j, t} \right\| = O_P\left(\sqrt{\log(LN^2) \max_{h \leq L} \frac{1}{T} \sum_{t=1}^{T} \var(Z_{h, i,j, t})} \right) = O_P\left(\sqrt{\log(LN)} \frac{\log(LN)}{T}\right).$  

\[\square\]
Lemma B.1.3. Suppose $\log N = o(T)$. For $f(t, h, L) = \omega(h, L)1\{t > h\}$,

$$\max_h \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} \hat{u}_{it} \hat{u}_{jt,t-h} x_{jt,t-h} f(t, h, L) - x_{it} u_{it} u_{jt,t-h} x_{jt,t-h} f(t, h, L) \right\| = O_P(\frac{1}{T} \sqrt{\frac{\log(LN)}{N}}).$$

Proof. The left hand side is bounded by $a_1 + a_2 + a_3$, where

$$a_1 = \max_h \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} \hat{u}_{it} - u_{it} \right\| \max_h \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} \hat{u}_{it} - u_{it} \right\| \max_h \left\| x_{it} - u_{it} \right\| \leq O_P(\frac{1}{NT}) \max_i \left\| x_{it} \right\| ^4 = O_P(\frac{1}{NT}).$$

For simplicity, let’s assume dim($x_{it}$) = 1. Then

$$a_1 \leq \| \hat{\beta} - \beta \|^2 \max_h \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} \hat{u}_{it} x_{jt,t-h} x_{jt,t-h} f(t, h, L) \right\| \leq O_P(\frac{1}{NT}) \max_i \left\| x_{it} \right\| ^4 = O_P(\frac{1}{NT}).$$

By using Bernstein’s inequality in Merlevède et al. (2011) for weakly dependent data and exponential tail conditions, and that $f(t, h, L)$ is bounded,

$$a_2 \leq \| \hat{\beta} - \beta \| \max_h \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} u_{it} x_{jt,t-h} x_{jt,t-h} f(t, h, L) \right\| \leq O_P(\frac{1}{\sqrt{NT}}) O_P(\sqrt{\frac{\log(LN)}{T}}) = O_P(\frac{1}{T} \sqrt{\frac{\log(LN)}{N}}).$$

$a_3$ is bounded using the same argument. Together,

$$\max_h \max_{i,j} \left\| \frac{1}{T} \sum_{t=1}^{T} x_{it} \hat{u}_{it} \hat{u}_{jt,t-h} x_{jt,t-h} f(t, h, L) - x_{it} u_{it} u_{jt,t-h} x_{jt,t-h} f(t, h, L) \right\| = O_P(\frac{1}{T} \sqrt{\frac{\log(LN)}{N}}).$$

$\square$
B.1.2 Proof of Theorem 2.2.1.

It suffice to prove \( \| \hat{V} - V \| = o_p(1) \). By Lemma B.1.1, we have

\[
\| \hat{V} - V \| \leq \| \hat{V} - V_L \| + C \sum_{h > L} \alpha_{NT}(h) + C \sum_{h=1}^{L} (1 - \omega(h, L)) \alpha_{NT}(h).
\]

The remaining proof is that of \( \| \hat{V} - V_L \| = o_p(1) \), given below.

Main proof of the convergence of \( \| \hat{V} - V_L \| \)

Note that \( V_L = \frac{1}{N} \sum_{ij} V_{u,ij}, \hat{V} = \frac{1}{N} \sum_{ij} \hat{S}_{u,ij} \). Hence

\[
\| \hat{V} - V_L \| \leq \frac{1}{N} \sum_{\hat{S}_{u,ij} = 0} \| V_{u,ij} - \hat{S}_{u,ij} \| + \frac{1}{N} \sum_{\hat{S}_{u,ij} \neq 0} \| V_{u,ij} - \hat{S}_{u,ij} \|.
\]

Note that \( \| S_{u,ij} - V_{u,ij} \| < \frac{1}{2} \lambda_{ij} \) for \( \forall (i, j) \) and \( C_1 > 0 \)

\[
\| S_{u,ii} \| \geq \| V_{u,ii} \| - \| S_{u,ii} - V_{u,ii} \| \\
\geq \| V_{u,ii} \| - \max_{ij} \| S_{u,ii} - V_{u,ii} \| \\
\geq \| V_{u,ii} \| - C \omega_{NT} > C_1.
\]

From Assumption 2.2.3, \( \| V_{u,ii} \| > c_1 > 0 \), then, \( \lambda_{ij} = M \omega_{NT} \sqrt{\| S_{u,ii} \| \| S_{u,jj} \|} > c_1 M \omega_{NT} > 2c_1 \omega_{NT} \). Then, \( \frac{\lambda_{ij}}{2} > c \omega_{NT} \geq \max_{ij} \| S_{u,ij} - V_{u,ij} \| \). Therefore, \( \| S_{u,ij} - V_{u,ij} \| < \frac{1}{2} \lambda_{ij} \) for \( \forall (i, j) \)

Recall \( \rho_{ij,h} = \sup_X \max_t |E(u_{it}u_{jt,h}|X)| + |E(u_{it,h}u_{jt}|X)| \). Then,

\[
\| V_{u,ij} \| \leq \frac{1}{T} \sum_{t=0}^{T} E_{x_{it}u_{it}u_{xt}x_{jt}} + \frac{1}{T} \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} [E_{x_{it}u_{it}u_{jt,h}x_{jt}} + E_{x_{it}u_{jt,h}u_{jt,h}x_{jt}}] \\
\leq C \rho_{ij,0}/2 + C \sum_{h=1}^{L} \omega(h, L) \sum_{t=h+1}^{T} \rho_{ij,h} \leq C \sum_{h=0}^{L} \rho_{ij,h}.
\]

Hence, on the event \( \max_{ij} \| S_{u,ij} - V_{u,ij} \| \leq C \omega_{NT} \),

\[
\frac{1}{N} \sum_{\hat{S}_{u,ij} = 0} \| V_{u,ij} - \hat{S}_{u,ij} \| \leq \frac{1}{N} \sum_{\hat{S}_{u,ij} = 0} \| V_{u,ij} \| \leq \frac{1}{N} \sum_{ij} \| V_{u,ij} \| 1\{|S_{u,ij}| < \lambda_{ij} \}
\]
\[
\begin{align*}
&= \frac{1}{N} \sum_{ij} \|V_{u,ij}\| 1\{\|V_{u,ij}\| < \|S_{u,ij}\| + \|S_{u,ij} - V_{u,ij}\|, \|S_{u,ij}\| < \lambda_{ij}\} \\
&\leq \frac{1}{N} \sum_{ij} \|V_{u,ij}\| \frac{(1.5\lambda_{ij})^{1-q}}{\|V_{u,ij}\|^{1-q}} 1\{\|V_{u,ij}\| < 1.5\lambda_{ij}\} \\
&\leq \frac{1}{N} \sum_{ij} \|V_{u,ij}\|^q (1.5\lambda_{ij})^{1-q} \leq C \omega_{NT}^{-q} \frac{1}{N} \sum_{ij} \|V_{u,ij}\|^q \\
&\leq C \omega_{NT}^{-q} \max_i \left( \sum_j (\sum_{h=0}^L \rho_{ij,h})^q \right).
\end{align*}
\]

On the other hand, on the event \( \max_{ij} \|S_{u,ij} - V_{u,ij}\| \leq C \omega_{NT} \),

\[
\begin{align*}
\frac{1}{N} \sum_{S_{u,ij} \neq 0} \|V_{u,ij} - \hat{S}_{u,ij}\| &\leq \frac{1}{N} \sum_{S_{u,ij} \neq 0} \|V_{u,ij} - S_{u,ij}\| + \frac{1}{N} \sum_{S_{u,ij} \neq 0} \|S_{u,ij} - \hat{S}_{u,ij}\| \\
&\leq \frac{1}{N} \sum_{S_{u,ij} \neq 0} 0.5\lambda_{ij} + \frac{1}{N} \sum_{S_{u,ij} \neq 0} \lambda_{ij} \leq \frac{1}{N} \sum_{ij} 1.5\lambda_{ij} 1\{\|S_{u,ij}\| > \lambda_{ij}\} \\
&= \frac{1}{N} \sum_{ij} 1.5\lambda_{ij} 1\{\|V_{u,ij}\| > \|S_{u,ij}\| - \|S_{u,ij} - V_{u,ij}\|, \|S_{u,ij}\| > \lambda_{ij}\} \\
&\leq \frac{1}{N} \sum_{ij} 1.5\lambda_{ij} \frac{\|V_{u,ij}\|^q}{(0.5\lambda_{ij})^q} 1\{\|V_{u,ij}\| > 0.5\lambda_{ij}\} \\
&\leq \frac{1}{N} \sum_{ij} C \lambda_{ij}^{1-q} \|V_{u,ij}\|^q \leq \frac{1}{N} \sum_{ij} \|V_{u,ij}\|^q C \omega_{NT}^{1-q} \\
&\leq C \omega_{NT}^{-q} \max_i \left( \sum_j (\sum_{h=0}^L \rho_{ij,h})^q \right).
\end{align*}
\]

Hence \( \|\hat{V} - V_L\| \leq C \omega_{NT}^{-q} \max_i \left( \sum_j (\sum_{h=0}^L \rho_{ij,h})^q \right) \). Therefore, we have

\[
\|\hat{V} - V\| \leq O_P(\omega_{NT}^{-q} \max_i \left( \sum_j (\sum_{h=0}^L \rho_{ij,h})^q \right) + \sum_{h=0}^{T-1} \alpha_{NT}(h) + \sum_{h=1}^{\sqrt{\log(LN)}} C \sum_{h=0}^{T-1} (1 - \omega(h, L)) \alpha_{NT}(h).
\]

**Remaining proofs:** \( \max_{ij} \|S_{u,ij} - V_{u,ij}\| = O_P(\omega_{NT}) \), where \( \omega_{NT} = L \sqrt{\frac{\log(LN)}{T}} \). Recall

\[
\begin{align*}
S_{u,ij} &= \frac{1}{T} \sum_t x_{it} \hat{u}_{it} \hat{u}_{jt} x_{jt}' + \frac{1}{T} \sum_{h=1}^L \omega(h, L) \sum_{t=h+1}^T \left[ x_{it} \hat{u}_{it} \hat{u}_{jt} x_{jt}' \right], \\
V_{u,ij} &= \frac{1}{T} \sum_t E \hat{u}_{it} \hat{u}_{jt} x_{jt}' \omega(h, L) \sum_{t=h+1}^T \left[ E \hat{u}_{it} \hat{u}_{jt} x_{jt}' \right].
\end{align*}
\]
Let
\[ M_{u,ij} \equiv \frac{1}{T} \sum_t x_{it}u_{it}u_{jt}x_{jt} + \frac{1}{T} \sum_{t=1}^{T} \sum_{h=1}^{L} \omega(h, L) [x_{it}u_{it}u_{jt-h}x_{jt-h} + x_{i,t-h}u_{i,t-h}u_{jt}x_{jt}]. \]

We first bound \( \max_{ij} \| M_{u,ij} - V_{u,ij} \| \), then bound \( \max_{ij} \| S_{u,ij} - M_{u,ij} \| \).

**Proof of** \( \max_{ij} \| M_{u,ij} - V_{u,ij} \| = O_P(L \sqrt{\frac{\log(LN)}{T}}) \)

Given Lemma B.1.2, we have

\[
\begin{align*}
\max_{ij} \| M_{u,ij} - V_{u,ij} \| & \leq O_P(\sqrt{\frac{\log(LN)}{T}}) \\
& \quad + 2 \max_{ij} \left\| \frac{1}{T} \sum_{h=1}^{L} \sum_{t=1}^{T} [x_{it}u_{it}u_{jt-h}x_{jt-h}f(t, h, L) - x_{it}u_{it}u_{jt-h}x_{jt-h}f(t, h, L)] \right\| \\
& \leq O_P(\sqrt{\frac{\log(LN)}{T}}) + \text{LO}_P(\sqrt{\frac{\log(LN)}{T}}) = O_P(L \sqrt{\frac{\log(LN)}{T}}).
\end{align*}
\]

**Proof of** \( \max_{ij} \| M_{u,ij} - S_{u,ij} \| = O_P(L \sqrt{\frac{\log(LN)}{N}}) \)

Given Lemma B.1.3, we have

\[
\begin{align*}
\max_{ij} \| M_{u,ij} - S_{u,ij} \| & \leq O_P(\sqrt{\frac{\log(LN)}{N}}) \\
& \quad + 2 \max_{ij} \left\| \frac{1}{T} \sum_{h=1}^{L} \sum_{t=1}^{T} [x_{it}u_{it}u_{jt-h}x_{jt-h}f(t, h, L) - x_{it}u_{it}u_{jt-h}x_{jt-h}f(t, h, L)] \right\| \\
& \leq O_P(\sqrt{\frac{\log(LN)}{N}}) + \text{LO}_P(\sqrt{\frac{\log(LN)}{N}}) = O_P(\frac{L}{T} \sqrt{\frac{\log(LN)}{N}}).
\end{align*}
\]

Together,

\[
\max_{ij} \| V_{u,ij} - S_{u,ij} \| = O_P(L \sqrt{\frac{\log(LN)}{T}}) + O_P(\frac{L}{T} \sqrt{\frac{\log(LN)}{N}}) = O_P(L \sqrt{\frac{\log(LN)}{T}}).
\]

\( \square \)
Appendix C

Appendix for Chapter 3

C.1 Proofs

Throughout the proof, max$_i$, max$_t$, max$_h$, max$_{ij}$, and max$_{it}$ denote max$_i \leq N$, max$_t \leq T$, max$_h \leq L$, max$_{i,j} \leq N$, and max$_{i,t} \leq T$ respectively. In addition, for technical simplicity, we assume that there is no fixed effects so that we do not take the de-meaning procedure. Extending to the more complete estimators with de-meaning is straightforward, but should require more technical arguments to show that the effect from added dependences due to the de-meaning is negligible.

C.1.1 Technical lemmas

**Lemma C.1.1.** Under the Assumptions 3.2.1-3.2.2, for $\gamma_T = \sqrt{\frac{\log(LN)}{T}}$, for $h \geq 0$,

$$\max_{h \leq L} \max_{i,j \leq N} \left\| \frac{1}{T} \sum_{t=h+1}^{T} x_{it} u_{jt-h} \right\| = O_P(\gamma_T).$$

*Proof.* Let $\gamma_{ij,t,h} = x_{it} u_{jt-h} I_{\{h+1 \leq t \leq T\}}$, where $I_A$ is the indicator function of the set $A$. To simplify notation, we assume $d = \dim(x_{it}) = 1$. By Lemma A.2 of Fan et al. (2011) and Assumption 3.2.1 (iii), $\gamma_{ij,t,h}$ satisfies the exponential tail condition. We set $\alpha_T = \sqrt{\frac{\log(LN)}{T}}$ and $c_1^2 = 3c_2$ for $c_1, c_2 > 0$. Using Bernstein inequality for weakly dependent data in Merlevède et al. (2011) and the Bonferroni method, we have

$$P \left( \max_{h \geq L} \max_{i,j \leq N} \left| \frac{1}{T} \sum_{t=1}^{T} \gamma_{ij,t,h} \right| > c_1 \alpha_T \right) \leq LN^2 \max_{h \leq L} \max_{i,j \leq N} P \left( \left| \frac{1}{T} \sum_{t=1}^{T} \gamma_{ij,t,h} \right| > c_1 \alpha_T \right) \to 0.$$
Then

$$\max_h \max_{ij} \left\| \frac{1}{T} \sum_{t=h+1}^{T} x_{it} u_{j,t-h} \right\| = O_P\left( \sqrt{\log(LN^2) \max_{ij,h} \frac{1}{T} \sum_{t=1}^{T} \text{var}(\gamma_{ij,t,h})} \right)$$

$$= O_P\left( \sqrt{\log(LN)} T \right).$$

\[\square\]

**Lemma C.1.2.** Under the Assumptions 3.2.1-3.2.2, for \( \gamma_T = \sqrt{\frac{\log(LN)}{T}} \),

(i) \( \max_{h \leq L} \max_{i,j \leq N} |\tilde{R}_{h,ij} - \Omega_{h,ij}| = O_P(\gamma_T) \), where \( \tilde{R}_{h,ij} = \frac{1}{T} \sum_{t=h+1}^{T} \hat{u}_{it} \hat{u}_{j,t-h} \) for \( h \geq 0 \).

(ii) \( \max_{|h| \leq L} \| \tilde{\Omega}_h - \Omega_h \|_1 = O_P(m_N^{-1-q}). \)

**Proof.** (i) First, we write

$$\max_{h \leq L} \max_{i,j \leq N} \left| \frac{1}{T} \sum_{t=h+1}^{T} \hat{u}_{it} \hat{u}_{j,t-h} - Eu_{it}u_{j,t-h} \right| \leq \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} (\hat{u}_{it} - u_{it})(\hat{u}_{j,t-h} - u_{j,t-h}) \right|$$

$$+ \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h} \right|$$

$$+ \max_h \max_{ij} \left| \frac{L}{T} |Eu_{it}u_{j,t-h}| \right|$$

$$\equiv a_1 + a_2 + a_3.$$

Then, \( a_1 \) is bounded by \( a_{11} + a_{12} + a_{13} \), where

$$a_{11} \equiv \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} (\hat{u}_{it} - u_{it})(\hat{u}_{j,t-h} - u_{j,t-h}) \right|$$

$$a_{12} \equiv \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} (\hat{u}_{it} - u_{it})u_{j,t-h} \right|$$

$$a_{13} \equiv \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} u_{it}(\hat{u}_{j,t-h} - u_{j,t-h}) \right|.$$

First, by the Cauchy-Schwarz inequality,

$$a_{11} = \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} x'_{it}(\hat{\beta} - \beta)x'_{j,t-h}(\hat{\beta} - \beta) \right|$$

$$\leq \|\hat{\beta} - \beta\|^2 \max_h \max_{ij} \left| \frac{1}{T} \sum_{t=h+1}^{T} \|x_{it}\| \|x_{j,t-h}\| \right|$$
\[ \leq O_P \left( \frac{1}{NT} \right) \max_i \frac{1}{T} \sum_{t=1}^{T} \|x_{it}\|^2 = O_P \left( \frac{1}{NT} \right). \]

Note that \( \max_i \frac{1}{T} \sum_{t=1}^{T} \|x_{it}\|^2 \) is bounded by the exponential tail condition and Bernstein’s inequality using the same argument of Lemma 3.1 of Fan et al. (2011). Next, by Lemma C.1.1,

\[ a_{12} \leq \|\beta - \hat{\beta}\| \max_{h} \max_{ij} \left\| \frac{1}{T} \sum_{t=h+1}^{T} x_{it}u_{j,t-h} \right\| \]

\[ \leq O_P(\frac{1}{\sqrt{NT}}) \max_{h} \max_{ij} \left\| \frac{1}{T} \sum_{t=h+1}^{T} x_{it}u_{j,t-h} \right\| = O_P \left( \frac{1}{T} \sqrt{\frac{\log(LN)}{N}} \right). \]

Similarly, \( a_{13} \) is bounded using the same argument. Then, we have \( a_1 = O_P(\frac{1}{T} \sqrt{\frac{\log(LN)}{N}}). \)

Next, we let \( Z_{h,i,j,t} = u_{it}u_{j,t-h} - Eu_{it}u_{j,t-h} \), which satisfies the exponential tail condition by Assumption 3.2.1 and Lemma A.2 of Fan et al. (2011). Then \( a_2 \) can be written as \( \max_h \max_{ij} \frac{1}{T} \sum_{t=1}^{T} Z_{h,i,j,t} \). Set \( \alpha_T = \sqrt{\frac{\log(LN)}{T}} \) and \( c_1^2 = 3c_2 \) for \( c_1, c_2 > 0 \). Then, using Bernstein’s inequality in Merlevède et al. (2011) and the same argument as in the proof of Lemma C.1.1,

\[ P \left( \max_{h \leq L} \max_{ij} \left\| \frac{1}{T} \sum_{t=1}^{T} Z_{h,i,j,t} \right\| > c_1 \alpha_T \right) \leq LN^2 \max_{h \leq L} \max_{ij} P \left( \left\| \frac{1}{T} \sum_{t=1}^{T} Z_{h,i,j,t} \right\| > c_1 \alpha_T \right) \to 0. \]

Hence, we have \( a_2 = O_P(\sqrt{\frac{\log(LN)}{T}}). \) In addition, \( a_3 = O_P(\frac{L}{T}) \), which can be proved easily. Together,

\[ \max_h \max_{ij} \left\| \frac{1}{T} \sum_{t=h+1}^{T} \tilde{u}_{it}\tilde{u}_{j,t-h} - Eu_{it}u_{j,t-h} \right\| = O_P \left( \sqrt{\frac{\log(LN)}{T}} \right). \]

(ii) Following Theorem 5 of Fan et al. (2013), we then have \( \max_{|h| \leq L} \|\tilde{\Omega}_h - \Omega_h\|_1 = O_P(m_N^{-\gamma} T^{-q}) \), where \( \tilde{\Omega}_h \) is defined in (3.2.4). \( \square \)

**Lemma C.1.3.** For \( h \leq L \) and \( v \leq L \), let \( Q_{imp}^{hv} = \sum_{t=1}^{T} w_{it} \sum_{j=1}^{N} (\Omega^{-1})_{t+h,m+v,jp}. \)
Then, under the Assumption 3.2.1-3.2.2,

\[
\max_{h,v \leq L} \max_{i,p \leq N} \left\| \frac{1}{NT} \sum_{q=1}^{N} \sum_{m=1}^{T} \varepsilon_{qm} Q_{hv}^{imp} \right\| = O_{P}\left( \sqrt{\frac{\log(LN)}{NT}} \right).
\]

**Proof.** First, we define \( W = \Omega^{-1}X = (w'_1, \cdots, w'_T)' \) \((NT \times d)\), and \( \varepsilon = \Omega^{-1}U = (\varepsilon'_1, \cdots, \varepsilon'_T)' \) \((NT \times 1)\). Let \( w_{it}' \) and \( \varepsilon_{it} \) denote the \( i \)th row of \( w_t \) and the \( i \)th element of \( \varepsilon_t \), respectively. For simplicity, we assume \( d = 1 \). Let \( \zeta_{ipq,mhv} = \varepsilon_{qm} Q_{hv}^{imp} \). Note that due to \( \|\Omega^{-1}\|_1 < \infty \), we know \( E(\zeta_{ipq,mhv}^2) < \infty \). Then \( \max_{h,v} \max_{i,p} \frac{1}{NT} \sum_{q=1}^{T} \sum_{m=1}^{T} \text{var}(\zeta_{ipq,mhv}) \) is bounded. Set \( \alpha_{NT} = \sqrt{\frac{\log(LN)}{NT}} \) and \( c_1^2 = 3c_2 \) for \( c_1, c_2 > 0 \). Then, using Bernstein’s inequality and the same argument as in the proof of Lemma C.1.1,

\[
P\left( \max_{h,v \leq L} \max_{i,p \leq N} \left\| \frac{1}{NT} \sum_{q=1}^{N} \sum_{m=1}^{T} \zeta_{ipq,mhv} \right\| > c_1 \alpha_{NT} \right) \\
\leq L^2 N^2 \max_{h,v \leq L} \max_{i,j} P\left( \left\| \frac{1}{NT} \sum_{q=1}^{N} \sum_{m=1}^{T} \zeta_{ipq,mhv} \right\| > c_1 \alpha_{NT} \right) \\
\rightarrow 0.
\]

Therefore, we have \( \max_{h,v \leq L} \max_{i,p \leq N} \left\| \frac{1}{NT} \sum_{q=1}^{N} \sum_{m=1}^{T} \varepsilon_{qm} Q_{hv}^{imp} \right\| = O_{P}(\sqrt{\frac{\log(LN)}{NT}}). \)

**Lemma C.1.4.** Consider a symmetric block matrix \( A = (A_{ij}) \in \mathbb{R}^{dn \times dn} \) where \( A_{ij} \in \mathbb{R}^{d \times d} \). Then

\[
\| A \| \leq \max_{i} \sum_{j=1}^{n} \| A_{ij} \|.
\]

**Proof.** Suppose \( \sigma(\cdot) \) is the spectrum of a matrix, which is the set of its eigenvalues. By Gershgorin’s Theorem for block matrices (see Salas (1999)), if we define

\[
G_i \equiv \sigma(A_{ii}) \cup T_i,
\]

where \( T_i = \{ \lambda \notin \sigma(A_{ii}) : \|(A_{ii} - \lambda I_d)^{-1}\|^{-1} \leq \sum_{j=1,j \neq i}^{n} \|A_{ij}\| \} \), then

\[
\sigma(A) \subset \bigcup_{i=1}^{n} G_i.
\]

Note that this theorem means the eigenvalue of \( A \) either equals \( \sigma(A_{ii}) \) or in that specific region.
Let $\lambda \in \bigcup_{i=1}^{n} G_i$. If $\lambda \in \sigma(A_{ii})$ for some $i$, then $|\lambda| \leq \|A_{ii}\| \leq \max_{i} \sum_{j=1}^{n} \|A_{ij}\|$. If $\lambda \notin \sigma(A_{ii})$ for all $i$, then we know $\lambda \in T_i$ for some $i$. Now we consider two cases: (i) $\|A_{ii}\| < |\lambda|$, and (ii) $\|A_{ii}\| \geq |\lambda|$, where $i$ such that $\lambda \in T_i$. For the case of (i), note that if a matrix $M$ is such that $\|M\| < 1$, then 

$$\frac{1}{1 + \|M\|} \leq \| (I - M)^{-1} \| \leq \frac{1}{1 - \|M\|}. \tag{C.1.1}$$

Then we have 

$$|\lambda| - \|A_{ii}\| \leq |\lambda| \left( 1 - \frac{\|A_{ii}\|}{|\lambda|} \right) \leq |\lambda| \left\| \left( I_d - \frac{A_{ii}}{|\lambda|} \right)^{-1} \right\|^{-1} = \|(|\lambda|I_d - A_{ii})^{-1}\|^{-1} \leq \sum_{j=1, j \neq i}^{n} \|A_{ij}\|.$$ 

Therefore, we have $|\lambda| \leq \sum_{j=1}^{n} \|A_{ij}\| \leq \max_{i} \sum_{j=1}^{n} \|A_{ij}\|$. Note that we have the second inequality since $\frac{\|A_{ii}\|}{|\lambda|} < 1$ with the inequality (C.1.1). For part (ii), if $\|A_{ii}\| \geq |\lambda|$, then $|\lambda| \leq \sum_{j=1}^{n} \|A_{ij}\| \leq \max_{i} \sum_{j=1}^{n} \|A_{ij}\|$. Therefore, $|\lambda| \leq \max_{i} \sum_{j=1}^{n} \|A_{ij}\|$ for all $\lambda \in \bigcup_{i=1}^{n} G_i$.

Finally, since $\sigma(A) \subset \bigcup_{i=1}^{n} G_i$, we know that for all $\lambda \in \sigma(A)$, $|\lambda| \leq \max_{i} \sum_{j=1}^{n} \|A_{ij}\|$. Therefore, we have $\|A\| \leq \max_{i} \sum_{j=1}^{n} \|A_{ij}\|$. \qed

### C.1.2 Proof of Theorem 3.2.1.

For any $NT \times NT$ blocked matrix $M = (m_{t,s})$ where the $m_{t,s}$ is the $(t, s)$th block $N \times N$ matrix. In addition, for any $0 \leq L < T$, we define $B_L(M) = [(m_{t,s})1(|t - s| \leq L)]$, which is an $NT \times NT$ matrix. Then we can write 

$$\|\hat{\Omega} - \Omega\| \leq \|B_L(\Omega) - \Omega\| + \|\hat{\Omega} - B_L(\Omega)\|.$$ 

First, we assume that $\xi_T(L) = \max_{t} \sum_{|h| > L} \|Eu_t u'_{t-h}\| = o(1)$ in Assumption 3.2.2(ii). This implies that off-diagonal $N \times N$ blocks that are far from the diagonal block are
negligible due to weak dependences. As for the first part, by Lemma C.1.4,

\[ \| B_L(\Omega) - \Omega \| \leq \max_t \sum_{s : |s - t| > L} \| E u_t u'_s \| = \xi_T(L) \to 0. \]

Next, note that \( f_T(L) = \max_t \sum_{|h| \leq L} \| E u_t'_{-h}(1 - \omega(|h|, L)) \| = o(1) \) (see Assumption 3.2.2(iii)). Then by Lemmas C.1.2 and C.1.4, for \( C < \infty \),

\[ \| \tilde{\Omega} - B_L(\Omega) \| \leq \max_t \sum_{s : |t - s| \leq L} \| \tilde{\Omega}_{t,s} - E u_t u'_s \|| \leq L \max_{|h| \leq L} \| (\tilde{\Omega}_h - \Omega_h) \omega(|h|, L) \| + \max_t \sum_{|h| \leq L} \| E u_t u'_{-h}(1 - \omega(|h|, L)) \| \]

\[ \leq CL \max_{|h| \leq L} \| \tilde{\Omega}_h - \Omega_h \| + f_T(L) \]

\[ = O_P(Lm_N\gamma_T^{1-q}) + f_T(L), \]

where \( \tilde{\Omega}_h \) is defined in (3.2.4). Therefore,

\[ \| \tilde{\Omega} - \Omega \| = O_P(Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L)). \]

We now show the second statement of Theorem 3.2.1. By the triangular inequality, we have

\[ \| \tilde{\Omega}^{-1} - \Omega^{-1} \| \leq \| (\tilde{\Omega}^{-1} - \Omega^{-1})(\tilde{\Omega} - \Omega) \Omega^{-1} \| + \| \Omega^{-1} (\tilde{\Omega} - \Omega) \Omega^{-1} \| \]

\[ \leq \| \tilde{\Omega}^{-1} - \Omega^{-1} \| \| \Omega^{-1} \| \| \tilde{\Omega} - \Omega \| + \| \Omega^{-1} \|^2 \| \tilde{\Omega} - \Omega \| \]

\[ = O_P(Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L))\| \tilde{\Omega}^{-1} - \Omega^{-1} \| + O_P(Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L)). \]

Hence we have \( (1 + o_P(1))\| \tilde{\Omega}^{-1} - \Omega^{-1} \| = O_P(Lm_N\gamma_T^{1-q} + \xi_T(L) + f_T(L)), \) that implies the result. \( \square \)

**Proof of Proposition 3.2.1.** First the left hand side of equation (3.2.5) can be extended as

\[ \frac{1}{\sqrt{NT}} X'(\tilde{\Omega}^{-1} - \Omega^{-1}) U = \frac{1}{\sqrt{NT}} X'\Omega^{-1}(\tilde{\Omega} - \Omega)\Omega^{-1} U \]
\[
\begin{align*}
&+ \frac{1}{\sqrt{NT}} X' \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U \\
&+ \frac{1}{\sqrt{NT}} X' (\hat{\Omega}^{-1} - \Omega^{-1}) (\hat{\Omega} - \Omega) \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U \\
&\equiv a + b + c.
\end{align*}
\]

Now we shall show that \( \frac{1}{\sqrt{NT}} X' (\hat{\Omega}^{-1} - \Omega^{-1}) U = \frac{1}{\sqrt{NT}} X' \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U + o_P(1) \).

First, we define \( W = \Omega^{-1} X \) and \( \varepsilon = \Omega^{-1} U \). Then, \( W = (w'_1, \ldots, w'_T)' \) with \( w_t \) being an \( N \times d \) matrix of \( w_{it} \), and \( \varepsilon_{it} \) is defined similarly. For any \( NT \times NT \) matrix \( M \), we denote \( (M)_{t,s} \) or \( (M)_{h} \) as the \( (t,s) \)th block matrix for \( h = t - s \). Moreover, we denote \( (M)_{t,s,ij} \) or \( (M)_{h,ij} \) as the \( (i,j) \)th element of the \( (t,s) \)th block matrix. Under Assumption 3.2.3, we show that \( b = o_P(1) \) as follows:

We write, for \( h = t - s \) and \( v = k - m \),

\[
\begin{align*}
b &= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{s=1}^{T} \sum_{k=1}^{T} \sum_{m=1}^{T} w'_t (\hat{\Omega} - \Omega)_{t,s} (\Omega^{-1})_{s,k} (\hat{\Omega} - \Omega)_{k,m} \varepsilon_m \\
&= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{|h| \leq L} \sum_{k=1}^{T} \sum_{m=1}^{T} w'_t (\hat{\Omega} - \Omega)_{h} (\Omega^{-1})_{t-h,k} (\hat{\Omega} - \Omega)_{k,m} \varepsilon_m \\
&- \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{|h| > L} \sum_{k=1}^{T} \sum_{m=1}^{T} w'_t \Omega_h (\Omega^{-1})_{t-h,k} (\hat{\Omega} - \Omega)_{k,m} \varepsilon_m \\
&= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{|h| \leq L} \sum_{|v| \leq L} \sum_{m=1}^{T} w'_t (\hat{\Omega} - \Omega)_{h} (\Omega^{-1})_{t-h,m+v} (\hat{\Omega} - \Omega)_{v} \varepsilon_m \\
&- \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{|h| \leq L} \sum_{|v| > L} \sum_{m=1}^{T} w'_t (\hat{\Omega} - \Omega)_{h} (\Omega^{-1})_{t-h,m+v} \Omega_v \varepsilon_m \\
&- \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{|h| > L} \sum_{k=1}^{T} \sum_{m=1}^{T} w'_t \Omega_h (\Omega^{-1})_{t-h,k} (\hat{\Omega} - \Omega)_{k,m} \varepsilon_m \\
&\equiv b_1 + b_2 + b_3.
\end{align*}
\]

First, define \( Q_{imp}^{hv} = \sum_{t=1}^{T} w_{it} \sum_{j=1}^{N} (\Omega^{-1})_{t-h,m+v,jp} \) as in Lemma C.1.3. We have, by
In addition, by Lemma C.1.2 and Bernstein inequality,

\[ \|b_1\| = \left\| \frac{1}{\sqrt{NT}} \sum_{|h| \leq L} \sum_{|v| \leq L} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{l=1}^{T} \sum_{m=1}^{T} (\hat{\Omega} - \Omega)_{h,ij} (\hat{\Omega} - \Omega)_{v,pq} \sum_{t=1}^{T} \sum_{m=1}^{T} w_{it} \varepsilon_{qm} (\Omega^{-1})_{t-h,m+v,pq} \right\| \]

\[ \leq \frac{1}{\sqrt{NT}} \max_{|h| \leq L} \left\| \frac{\Omega_h - \Omega_h}{1} \right\|_{2} \max_{i,p} \left\| \sum_{|v| \leq L} \sum_{q=1}^{V} \sum_{m=1}^{M} \sum_{j=1}^{T} \sum_{p=1}^{P} \varepsilon_{qm} Q_{\text{imp}}^{i,p,v} \right\| \]

\[ \leq O(L^2 \sqrt{NT}) \max_{|h| \leq L} \left\| \Omega_h - \Omega_h \right\|_{2} \max_{i,p,h,v} \left\| \frac{1}{NT} \sum_{q=1}^{V} \sum_{m=1}^{M} \sum_{j=1}^{T} \sum_{p=1}^{P} \varepsilon_{qm} Q_{\text{imp}}^{i,p,v} \right\| \]

\[ = O_P(\sqrt{TL^2m_N^2\gamma_T^{3-2q}}) = o_P(1). \]

In addition, by Lemma C.1.2 and Bernstein inequality,

\[ \|b_2\| = \left\| \frac{1}{\sqrt{NT}} \sum_{|h| \leq L} \sum_{|v| \leq L} \sum_{l=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{m=1}^{M} \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{t=1}^{T} \sum_{m=1}^{T} w_{it} (\hat{\Omega} - \Omega)_{h,ij} (\Omega^{-1})_{t-h,m+v,pq} \varepsilon_{qm} \right\| \]

\[ \leq \frac{L}{\sqrt{NT}} \max_{|h| \leq L} \left\| \frac{\Omega_h - \Omega_h}{1} \right\|_{2} \max_{i,p,h,v} \left\| \sum_{|v| \leq L} \sum_{q=1}^{Q} \sum_{m=1}^{M} \sum_{j=1}^{T} \sum_{p=1}^{P} (\Omega^{-1})_{t-h,m+v,pq} \right\| \]

\[ \leq O(L \sqrt{NT}) \max_{|h| \leq L} \left\| \Omega_h - \Omega_h \right\|_{1} \sum_{|v| \leq L} \| \Omega_v \|_1 \]

\[ = O_P(L^{1-a} \sqrt{NT} \gamma_T^{1-a}) = o_P(1) \]

and

\[ \|b_3\| = \left\| \frac{1}{\sqrt{NT}} \sum_{|h| \leq L} \sum_{|v| \leq L} \sum_{l=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{m=1}^{M} \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{t=1}^{T} \sum_{m=1}^{T} w_{it} \Omega_h,ij (\Omega^{-1})_{t-h,k,jp} (\hat{\Omega} - \Omega)_{km,pq} \varepsilon_{qm} \right\| \]

\[ \leq \frac{T}{\sqrt{NT}} \max_{|v| \leq L} \left\| \frac{\Omega_h - \Omega_h}{1} \right\|_{2} \max_{i,j} \left\| \sum_{|h| \leq L} \sum_{q=1}^{Q} (\hat{\Omega} - \Omega)_{v,pq} \right\| \]

\[ = O_P(T^{1-\alpha} \sqrt{NT} \gamma_T^{1-\alpha}) = o_P(1) \]
\[ \times \max_{i} \sum_{t=1}^{T} \| w_{it} \| \max_{m} \sum_{q=1}^{N} | \varepsilon_{qm} | \max_{\ell,h,v,p} \sum_{m=1}^{T} \sum_{j=1}^{N} | (\Omega^{-1})_{t-h,m+v,j,p} | \]
\[ \leq O(T^{\sqrt{NT}}) \max_{v} \| \hat{\Omega}_{v} - \Omega_{v} \|_{1} \sum_{|h| > L} \| \Omega_{h} \|_{1} \]
\[ = O_{p}(L^{-o} T^{\sqrt{NT}} (\gamma_{f}^{1-q}) = o_{p}(1). \]

Therefore, we have \( \| b \| = o_{p}(1) \). Next, we define \( \gamma^{*} = L \gamma_{f}^{1-q} + \xi_{T}(L) + f_{T}(L). \)

By Theorem 3.2.1, \( \| \hat{\Omega} - \Omega \| = O_{p}(\gamma^{*}) = \| \hat{\Omega}^{-1} - \Omega^{-1} \|. \) Then, under the Assumption 3.2.3(v), when \( \| \Omega^{-1} \|_{1} = O(1) \), we have

\[ \| c \| = \left\| \frac{1}{\sqrt{NT}} X^\prime (\hat{\Omega}^{-1} - \Omega^{-1}) (\hat{\Omega} - \Omega) \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U \right\| \]
\[ \leq \| \hat{\Omega}^{-1} - \Omega^{-1} \| \| \hat{\Omega} - \Omega \|^{2} \sqrt{NT} = O_{p}(\sqrt{NT} \gamma^{3}) = o_{p}(1). \]

Therefore, we have \( \frac{1}{\sqrt{NT}} X^\prime (\hat{\Omega}^{-1} - \Omega^{-1}) U \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U + o_{p}(1). \)

From Theorem 3.2.1, it is easy to that \( \frac{1}{\sqrt{NT}} X^\prime \hat{\Omega}^{-1} X = \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} X + o_{p}(1). \) Also, by the weak law of large numbers, \( (\frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} X)^{-1} = \Gamma^{-1} + o_{p}(1), \) where \( \Gamma = E(\frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} X). \)

Then

\[ \sqrt{NT}(\hat{\beta}_{FGLS} - \beta) = \left( \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} X \right)^{-1} \left( \frac{1}{\sqrt{NT}} X^\prime \hat{\Omega}^{-1} U \right) + o_{p}(1) \]
\[ = \left( \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} X \right)^{-1} \left( \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} U + \frac{1}{\sqrt{NT}} X^\prime (\hat{\Omega}^{-1} - \Omega^{-1}) U \right) + o_{p}(1) \]
\[ = \Gamma^{-1} \left( \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} U + \frac{1}{\sqrt{NT}} X^\prime (\hat{\Omega}^{-1} - \Omega^{-1}) U \right) + o_{p}(1) \]
\[ = \Gamma^{-1} \left( \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} U \right) + \Gamma^{-1} \left( \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U \right) + o_{p}(1). \]

\[ \boxed{\text{C.1.3 Proof of Theorem 3.2.2.}} \]

It suffices to prove \( \left\| \frac{1}{\sqrt{NT}} X^\prime \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U \right\| = o_{p}(1). \)

Let \( A_{bh} = \{ (i,j) : |E u_{it} u_{j,t-h} | \neq 0 \} \). Also, let \( W = \Omega^{-1} X, \) and \( \varepsilon = \Omega^{-1} U. \) In addition, \( w_{it} \) and \( \varepsilon_{it} \) are defined as in the proof of Proposition 3.2.1. We define \( G_{T,i,j}^{1}(h) = \frac{1}{\sqrt{T}} \sum_{t=h+1}^{T} (u_{it} u_{j,t-h} - E u_{it} u_{j,t-h}) \) and \( G_{T,i,j}^{2}(h) = \frac{1}{\sqrt{T}} \sum_{t=h+1}^{T} w_{it} \varepsilon_{j,t-h}. \) Then under the
Assumption 3.2.4, there is $C > 0$ so that

$$\left\| \frac{1}{\sqrt{NT}} X' \Omega^{-1} (\hat{\Omega} - \Omega) \Omega^{-1} U \right\| = \left\| \frac{1}{\sqrt{NT}} W' (\hat{\Omega} - \Omega) \varepsilon \right\|$$

$$\leq \left\| \frac{C}{\sqrt{NT}} \sum_{h=0}^{L} \sum_{i,j \in A_{b_h}} \sum_{t=h+1}^{T} w_{it} \varepsilon_{j,t-h} \frac{1}{T} \sum_{s=h+1}^{T} \left( u_{is} u_{j,s-h} - Eu_{it} u_{j,t-h} \right) \right\| + o_P(1)$$

$$= \left\| \frac{C}{\sqrt{NT}} \sum_{h=0}^{L} \sum_{i,j \in A_{b_h}} G_{T,ij}^{1}(h) G_{T,ij}^{2}(h) \right\| + o_P(1) = o_P(1).$$
Bibliography


