The Pennsylvania State University
The Graduate School

NONPARAMETRIC ESTIMATION OF SUFFICIENT FORECASTING WITH A DIVERGING NUMBER OF FACTORS

A Thesis in
Statistics
by
Xiufan Yu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2019

The thesis of Xiufan Yu was reviewed and approved* by the following:

Runze Li
Eberly Family Chair Professor of Statistics
Thesis Co-Advisor

Lingzhou Xue
Assistant Professor of Statistics
Thesis Co-Advisor

Bing Li
Professor of Statistics

Ephraim Hanks
Assistant Professor of Statistics
Chair of Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

The sufficient forecasting (Fan, Xue and Yao, 2017) provides an effective forecasting procedure to estimate sufficient indices from high-dimensional predictors in the presence of a possible nonlinear forecast function. In this paper, we first revisit the sufficient forecasting and explore its underlying connections to Fama-Macbeth regression and partial least squares. Then, we develop an inferential theory of sufficient forecasting within the high-dimensional framework with large cross sections, a large time dimension and a diverging number of factors. We derive the rate of convergence of the estimated factors and loadings and characterize the asymptotic behavior of the estimated sufficient forecasting directions without requiring the restricted linearity condition. The predictive inference of the estimated nonparametric forecasting function is obtained with nonparametrically estimated sufficient indices. We further demonstrate the power of the sufficient forecasting in an empirical study of financial markets.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere gratitude to my advisors Dr. Runze Li and Dr. Lingzhou Xue, for inspiring me with their enthusiasm and professionalism, for guiding me with their broad knowledge and remarkable insights, for supporting me with constant encouragements, and for everything.

I would also like to thank Dr. Bing Li and Dr. Ephraim Hanks for their precious time in reading my thesis and valuable advices on improving the contents of this thesis. I am also grateful to my coauthor Dr. Jiawei Yao, who helped me a lot in this project. Last but not least, I want to thank my parents, for their unconditional love, support and understanding.

This work was supported by the National Science Foundation, DMS 1505256, DMS 1811552 and National Institute on Drug Abuse (NIDA) grant P50DA039838.

Chapter 1

Introduction

Forecasting using high-dimensional predictors has received considerable attention in statistics, epidemiology, finance, macroeconomics, and many other fields. In a data-rich environment, it is not uncommon to assume that a few underlying common factors simultaneously drive the forecasting target and the high-dimensional predictors. Such a link between the predictors and the target opens a door for the efficient usage of large-scale predictive information. Not only does it reduce the dimension in predictive models, but more importantly, the use of principal components can typically characterize many economic predictors.

In most cases, the relationship between forecasting target and latent factors is nonlinear, which poses a significant challenge to extract information relevant to the target. Fan, Xue and Yao (2017) proposed a novel sufficient forecasting method to obtain predictive indices, in the presence of an unknown nonlinear forecasting function. However, it is of interest to understand the relation between the sufficient forecasting and other methods, and under what conditions we can use the estimated factors to forecast in a nonlinear environment. In this paper, we shall enrich the sufficient forecasting method by providing answers to these related questions.

By assuming linearity in the forecasting function, an abundant literature endeavors to use common factors for forecasting. In their seminal papers, Stock and Watson (1989, 2002*a*,*b*) first demonstrate the validity of using estimated principal components for forecasting. Bai and Ng (2006) conduct inferences on factor-augmented regressions to enable forecast. A further line of efforts to refine forecast

is by filtering out information unrelated to the forecasting target, either via the predictors or the estimated factors. For example, Bair, Hastie, Paul and Tibshirani (2006) applied correlation screening to obtain relevant predictors. Bai and Ng (2008) established thresholding criteria to rule out predictors not informative for the target. Kelly and Pruitt (2015) proposed a Three-Pass Regression Filter (3PRF) method that selectively identifies the subset of factors influencing the target while discarding those factors that are irrelevant. All of the works above consider the case of a linear forecasting function. The main difference between our paper and the existing work is that we develop inferential theories under a more general nonlinear forecasting equation.

This paper makes three main contributions. The first is to point out the close connections of the sufficient forecasting with existing panel data analysis, namely, the Fama-Macbeth regression (Fama and MacBeth 1973) and the three-pass regression filter, which involve time-series regressions as the first pass. We show that the characterization of sufficient forecasting is similar in spirit after transforming the underlying inverse forecasting indices curve using loading matrix $\mathbf{B}$. We also show that the sufficient forecasting can be derived as the solution to a constrained optimization problem more general than partial least squares method. The second contribution is to complement the asymptotic theories of the sufficient forecasting in the case of a diverging number of latent factors. We share the similar spirit of Ludvigson and Ng (2007) and Li, Li and Shi (2017) to allow the number of factors to increase with sample size, which avoids possible model misspecification and accommodates structural changes. The diverging number of factors also relaxes the restricted linearity condition that might lead to the time reversibility (Xia, Tong, Li and Zhu, 2002). Also, we study the decomposition of the estimated predictive directions to illustrate the source of estimation errors.

In our third contribution, we consider nonparametric estimation of the forecasting function with the estimated predictive indices, which serves as a final step of our methodology. An enduring interest of factor analysis lies in when and where we can treat estimated factors as known, and how well we can use them in any downstream inferences. We are not the first to investigate the nonparametric estimation with generated covariates. In a somewhat different setting, Mammen, Rothe and Schienle (2012) provided a two-stage analysis in which an estimated covariate is

generated in the first stage, and a nonparametric regression based on the estimates is obtained in the second stage. Inspired by their work, we extend the theories to panel data and factors obtained by sufficient forecasting method. We estimate the forecasting function using local linear regression, and show that the presence of nonparametrically estimated predictive indices affects the limiting behavior of the forecast only through a smoothed version of first-stage estimation error. Finally, in an empirical example, we apply the sufficient forecasting methodology to a broad set of equity stocks to forecast market returns. Our method improves upon existing linear methods regarding prediction accuracy.

The rest of this thesis is organized as follows.In Chapter 2, we provide a detailed literature review of existing factor-model based forecasting methods. In Chapter 3, we revisit the sufficient forecasting by offering an alternative interpretation of the methodology and connect it with existing approaches. In addition, we extend the sufficient forecasting method by allowing the number of factors to diverge, and establishes the extended asymptotic properties. Moreover, we study nonparametric estimation of the forecasting function. In the end, we apply the method to a real-data problem of asset return forecast. Some conclusion remarks and future work are discussed in Chapter 4.

Chapter 2

Literature Review

Recent advances in technology make it possible to collect and store massive predictive information at the same time. Under such a data-rich environment, it is of great importance to study the problem of forecasting using a large number of predictors. To be specific, let $\{y_t, t = 1, \cdots, T\}$ be the targeted time series variable to be forecast, and suppose the predictive information is collected on a large number of predictors $\{x_{it}, 1 \leq i \leq p,\ 1 \leq t \leq T\}$, where the number of predictors p is much larger than the number of observations T. The problem of interest is how to effectively utilize the vast predictive information to make accurate forecasts on the targets.

A common approach in statistics and ecnomics literature is to assume that the forecast target and high-dimensional predictors are driven by a few underlying common factors. The key is to extract information from large cross-sectional predictors and condense them into several highly informative factors. In this way, we could effectively reduce the dimensions in the predictive models, in the mean time, turning the high dimensionality into a blessing.

In most cases, the relationship between the forecasting target and latent factors is unclear. A large body of literature focuses on a linear model and its refinement. In this chapter, we shall first review three well-known methods which are well designed under the assumption of linearity, and then proceed to the recently proposed sufficient forecasting method that is applicable for nonlinear effects.

2.1 Forecasting Using Principal Components

Stock and Watson (2002*a*,*b*) propose a two-step process to forecast one time series using a large number of predictor series. They first simplify the high-dimensional problem by modeling the covariability of the predictors using a relatively small number of unobserved latent factors, which are estimated from principal components, and then make forecasting on targets using a linear regression. More precisely, let $\mathbf{x}_t = (x_{1t}, \cdots, x_{pt})'$ be a $p \times 1$ cross-sectional vector of p predictors at time t. They assume that $(\mathbf{X}_t, y_{t+h})$ admit an approximate factor model representation with K common latent factors $\mathbf{f}_t = (f_{1t}, \cdots, f_{Kt})'$,

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t \tag{2.1.1}$$

and

$$y_{t+h} = \boldsymbol{\beta}_f' \mathbf{f}_t + \boldsymbol{\beta}_w' \mathbf{w}_t + \epsilon_{t+h}, \tag{2.1.2}$$

where $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)'$ is a $p \times K$ loading matrix, $\mathbf{F} = (\mathbf{f}_1, \cdots, \mathbf{f}_T)'$ is $T \times K$ factor matrix , $\mathbf{u}_t = (u_{1t}, \cdots, u_{pt})'$ is a $p \times 1$ vector of idiosyncratic errors, $\mathbf{w}_t$ is a $m \times 1$ vector of additional observed variables (for example, lags of y_t) and h is the forecast horizon. $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_w$ are regression coefficients corresponding to the factors and additional observed variables, and ϵ_{t+h} is some stochastic error in the linear model. In this model, data for $\{y_t, \mathbf{X}_t, \mathbf{w}_t\}$ is observable, and the goal is to forecast y_{t+h}.

The first step is to estimate the latent factors $\{\mathbf{f}_t\}_{t=1}^T$ from the predictors $\{x_{it}\}_{1 \le i \le p, 1 \le t \le T}$. Consider $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{B}}$ to be the minimizers of least squares objective function

$$V(\mathbf{F}, \mathbf{B}) = (pT)^{-1} \sum_{i=1}^{p} \sum_{t=1}^{T} (x_{it} - \mathbf{b}_i' \mathbf{f}_t)^2 . \tag{2.1.3}$$

After concentrating out $\widehat{\mathbf{F}}$, minimizing (2.1.3) is equivalent to maximizing $\mathrm{tr}\,(\mathbf{B}'\mathbf{X}\mathbf{X}'\mathbf{B})$ subject to $\mathbf{B}'\mathbf{B}/p = \mathbf{I}_K$, where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ is a $p \times T$ data matrix. Hence, the minimization problem becomes a classical principal components problem, which can be solve by setting $\widehat{\mathbf{B}}$ equal to the eigenvectors of $\mathbf{X}\mathbf{X}'$ corresponding to its K largest eigenvalues. As a result, the principal components

estimator of $\mathbf{F}$ is

$$\widehat{\mathbf{F}} = \mathbf{X}'\widehat{\mathbf{B}}/p.$$

In this way, we form principal components of $\{\mathbf{x}_t\}_{t=1}^T$ to sever as estimates of the factors. And These estimated factors, together with $\mathbf{w}_t$'s, are then used in (2.1.4) to estimate the regression coefficients $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_w$. The forecast on y_{t+h} is finally constructed as

$$\widehat{y}_{t+h} = \widehat{\boldsymbol{\beta}}_f'\widehat{\mathbf{f}}_t + \widehat{\boldsymbol{\beta}}_w'\widehat{\mathbf{w}}_t. \tag{2.1.4}$$

In terms of theoretical analysis, they prove that under certain regularity conditions, the principal components of $\{\mathbf{x}_t\}_{t=1}^T$ are consistent estimators of true latent factors in the sense of subjecting to normalization. Moreover,a s p, T goes to infinity, the forecast $\widehat{y}_{t+h}$ constructed from the estimated factors and estimated coefficients converges to the infeasible forecast y_{t+h} which would be obtained if the factors and coefficients were known.

2.2 Targeted Diffusion Index Forecast

It has come to people's attention that in high-dimensional settings, it is likely that not all predictors are informative to the target variable. Stock and Watson (2002*a*,*b*) estimate the latent factors via principal components, leading to the results that all estimated factors are linear combinations of $\mathbf{x}_t$. So that the target y_{t+h} is forecast using all p predictors. In order to improve this problem, Bai and Ng (2008) propose a thresholding rule to select a subset of important predictors before estimating the factors. They consider a scenario that the series to be forecast is highly predictable by a subset of the p predictive series, and the subset may be different for various y. This subset of predictors is called targeted predictors.

The variable selection process is implemented by the harding thresholding rule. The detailed procedures are summarized as follows.

Step 1: Regress y_t on $\mathbf{w}_{t-h}$ and $x_{i,t-h}$ for each $i = 1, \cdots, p$ and let t_i denote the t-statistic associated with $x_{i,t-h}$.

Step 2: Sort $|t_1|, \cdots, |t_p|$ in descending order. This represents a ranking of the marginal predictive power of the i-th predictor.

Step 3: Let $k_\alpha^\star$ be the number of predictors whose absolute value of t-statistic, i.e. $|t_i|$, exceeds a threshold significance level α.

Step 4: Let $\mathbf{x}_t(\alpha) = \left(x_{[1t]}, \cdots, x_{[k_\alpha^\star, t]}\right)'$ be the corresponding set of $k_\alpha^\star$ targeted predictors. Calculate the estimated latent factors $\widehat{\mathbf{f}}_t$ from $\mathbf{x}_t(\alpha)$ by the method of principal components, proceed to estimate the linear forecasting equation, in addition, make the forecast using estimated forecasting model.

In a short summary, Bai and Ng (2008) pay attention to the existence of uninformative predictors. Under such circumstance, the principal components estimated from a large group of variables are in fact dominated by principal components estimated from a smaller set of predictors, which they call "targeted predictors". They show that instead of using all the predictors, applying thresholding rules to determine targeted predictors from which the latent factors are to be extracted would successfully reduces forecast errors.

2.3 The Three-Pass Regression Filter (3PRF)

Kelly and Pruitt (2015) consider another scenario in which only a subset of relevant factors drive the forecast target, regardless of the total number of common factors driving the cross section of predictors. The principal component (PC) methods condense the cross-sectional information based on covariance within the predictors. In addition, in order to achieve consistency, the PC-based method has to identify all the factors driving the panel of predictors, including those which may be irrelevant to the target variable. To overcome this problem, they develop a new method named the three-pass regression filter (3PRF), which only needs to estimate the relevant factors. The 3PRF condenses the cross-sectional information of predictors according to covariance with the forecast target.

The same as previous PC-based methods, the predictors are assumed to be described by an approximate factor model. And the target is a linear function of a subset of the latent factors plus some unforecastable noise. What's different is that they introduce a new variable called proxy, which is assume to be driven by target-relevant factors. Let $\mathbf{Z} = (\mathbf{z}_1', \cdots, \mathbf{z}_T')'$ be a $T \times L$ matrix of proxies. The 3PRT regression-based construction can be summarized as follows.

Pass 1: Run time series regression of x_{it} on $\mathbf{Z}$ for $i = 1, \cdots, p$

$$x_{it} = \phi_{0,i} + \mathbf{z}'\boldsymbol{\phi}_i + \epsilon_{it}.$$

Retain slope estimate $\widehat{\boldsymbol{\phi}}_i$, $i = 1, \cdots, p$.

Pass 2: Run cross section regression of x_{it} on $\widehat{\boldsymbol{\phi}}_i$ for $t = 1, \cdots, T$,

$$x_{it} = \phi_{0,t} + \widehat{\boldsymbol{\phi}}'\mathbf{F}_t + \varepsilon_{it}.$$

Retain slope estimate $\widehat{\mathbf{F}}_t$, $t = 1, \cdots, T$.

Pass 3: Run time series regression of y_{t+1} on predictive factors $\widehat{\mathbf{F}}_t$,

$$y_{t+1} = \beta_0 + \widehat{\mathbf{F}}'\boldsymbol{\beta} + \eta_{t+1}.$$

Deliver forecast $\hat{y}_{t+1}$.

The first pass runs p separate time series regressions, one for each predictor. The predictors x_{it} serve as the dependent variable and the proxies serve as the regressors. The estimated coefficients in this pass describe the sensitivity of the predictor to factors represented by the proxies. The second pass runs T cross section regressions with the predictors x_{it} being the dependent variable while the first-pass coefficients $\widehat{\boldsymbol{\phi}}_i$ as regressors. Fluctuations in the latent factors cause the cross section of predictors to fan out and compress over time. The first-stage coefficient estimates map the cross-sectional distribution of predictors to the latent factors, then the second-stage cross section regression use this map t back out estimates of the factors at each point in time. On top of the estimated predictive factors, the third-pass directly delivers forecasts on y_{t+1} based on ordinary least squares regression. They also derive the asymptotic distribution of the 3PRF forecasts and theoretical analysis proves that the 3PRF forecasts are consistent.

2.4 Sufficient Forecasting

Fan, Xue and Yao (2017) propose sufficient forecasting to forecast a single time series when there is large number of predictors and a possible nonlinear effect.

Suppose the information is collected on a large number of predictors x_{it} ($1 \leq i \leq p,\ 1 \leq t \leq T$), which are driven by some unobservable factors $\mathbf{f}_t$. The factor representation of the data is

$$x_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \quad 1 \leq i \leq p,\ 1 \leq t \leq T. \tag{2.4.1}$$

where $\mathbf{f}_t = (f_{1t}, \cdots, f_{Kt})'$ is the $K \times 1$ vector of common factors, $\mathbf{b}_i$ is the corresponding vector of factor loadings, and u_{it} is an idiosyncratic error. In matrix notation, the factor model is

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t,$$

where $\mathbf{x}_t = (x_{1t}, \cdots, x_{pt})'$ is the cross section of $p \times 1$ predictors, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)'$ is the $p \times K$ loading matrix and $\mathbf{u}_t = (u_{1t}, \cdots, u_{pt})'$ is the $p \times 1$ error term.

Then suppose the target variable y_{t+1} depends on the underlying factors only through L predictive indices $\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t$, that is,

$$y_{t+1} = g(\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t) + \epsilon_{t+1}. \tag{2.4.2}$$

where $\boldsymbol{\phi}_i$'s are unknown forecasting directions, $g(\cdot)$ is an arbitrary link function, and ϵ_{t+1} is some stochastic error independent of $\mathbf{f}_t$ and u_{it}.

Before proceeding to estimating parameters, they impose a few additional identification conditions on the model to settle the potential identifiability issue. Note that $\mathbf{B}\mathbf{f}_t = (\mathbf{B}\mathbf{A})(\mathbf{A}^{-1}\mathbf{f}_t)$ holds for any nonsingular matrix $\mathbf{A}$, they assume that the factor model (2.4.1) has the following canonical normalization

$$\text{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.4.3}$$

where $\mathbf{I}_K$ is a $K \times K$ identity matrix. Without loss of generality, they assume $E(\mathbf{f}_t) = \mathbf{0}$ and hence $E(\mathbf{x}_t) = \mathbf{0}$. Moreover, let $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L)$ denote a $K \times L$ matrix consisting of true forecasting directions. The directions $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ are not identifiable since there is no structural condition on $g(\cdot)$. Luckily, the good news is the subspace spanned by $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ can be identified (Cook, 2009), which is called the central subspace and denoted by $S_{y|\mathbf{f}}$. Therefore, they refer any orthonormal basis $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ as sufficient dimension reduction (SDR)direction, and their corresponding predictive indices $\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t$ as sufficient predictive

indices.

The idea of sufficient forecasting can be represented as a four-layer deep learning architecture as shown in Figure 2.1. The dimensionality is first reduced from p to K via a high-dimensional factor model implemented by the principal component analysis. Based on the inferred factors and estimated factor loadings, the sufficient dimension reduction directions are extracted following the idea of Sliced Inverse Regression (SIR) proposed by Li (1991). This will lead to consistent estimates of L sufficient predictive indices. Given these estimated low-dimensional indices, nonparametric regression techniques can be applied to estimate the link function $g(\cdot)$ and make forecasting on the target.

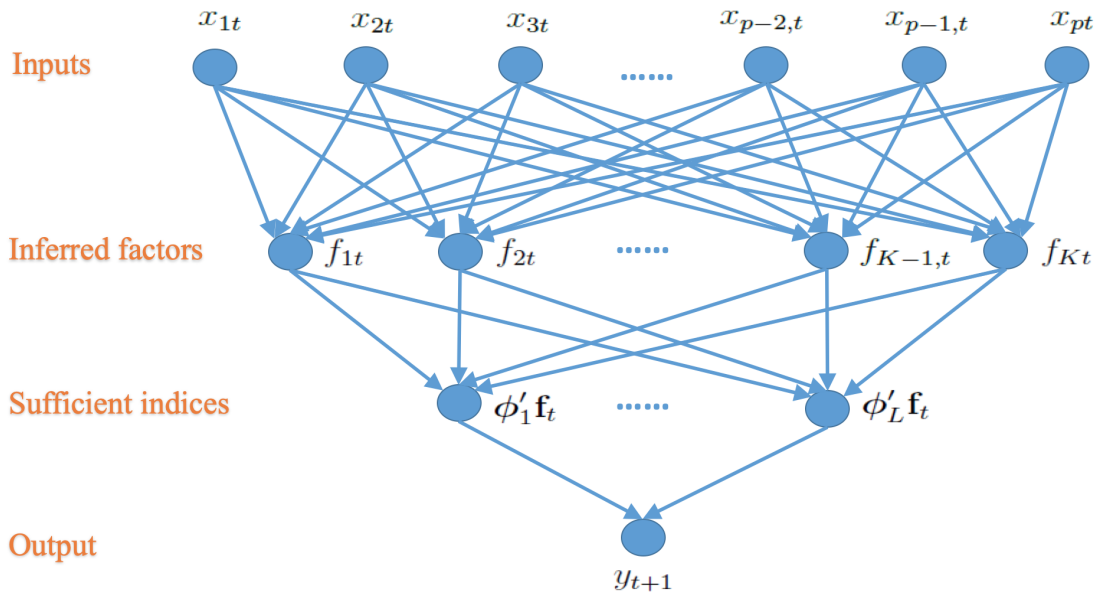

Figure 2.1. Sufficient Forecasting Represented by A Deep Learning Architecture. *Figure Courtesy of Fan, Xue and Yao (2017).*

To be more precisely, the sufficient forecasting method first extracts the latent factors $\widehat{\mathbf{F}}' = \left(\widehat{\mathbf{f}}_1, \cdots, \widehat{\mathbf{f}}_T\right)$ from the vast observable predictors. Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, $\mathbf{F}' = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$, and $\|\cdot\|_F$ denote the Frobenius norm. Consider the constrained least squares problems:

$$(\widehat{\mathbf{B}}, \widehat{\mathbf{F}}) = \text{argmin}_{(\mathbf{B},\mathbf{F})} \|\mathbf{X} - \mathbf{B}\mathbf{F}'\|_f^2$$

subject to

$$T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K,\ \mathbf{B}'\mathbf{B} \text{ is diagonal}$$

This is a classical principal components problem, and it has been widely used to extract underlying common factors in Statistics and Finance literature. The constrains correspond to the normalization condition, and $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{B}}$ are the estimated factor and loadings.

To extract information from the panel data, sufficient forecasting considers the covariance of conditional expectation of factors, $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$. To obtain forecasting directions $\boldsymbol{\phi}_i$'s, they propose to use the estimated factors $\widehat{\mathbf{f}}_t$ to form an estimator of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$. A slicing version of the covariance is

$$\boldsymbol{\Sigma}_{f|y} = \frac{1}{H}\sum_{h=1}^{H} E(\mathbf{f}_t|y_{t+1} \in I_h)E(\mathbf{f}_t'|y_{t+1} \in I_h), \tag{2.4.4}$$

where $H \geq L$ is fixed and the range of y_{t+1} is divided into H slices $I_1, \ldots, I_H$ such that $P(y_{t+1} \in I_h) = 1/H$. The sliced covariance is more appealing as we don't require $H \to \infty$. As discussed in Li (1991) and Fan et al. (2017), any direction orthogonal to $E(\mathbf{f}_t|y_{t+1})$ is also orthogonal to $\boldsymbol{\Sigma}_{f|y}$. Suppose we have the following *coverage condition*:

$$\langle \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_L \rangle = \langle \boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_L \rangle, \tag{2.4.5}$$

where $\boldsymbol{\psi}_i's$ are the eigenvectors of $\boldsymbol{\Sigma}_{f|y}$ corresponding to its L leading positive eigenvalues. One would recover the column space $\boldsymbol{\Phi}$ by exploiting the eigenvectors of $\boldsymbol{\Sigma}_{f|y}$. We remind that $\boldsymbol{\phi}_j$ in (2.4.2) is only identifiable up to a transformation, but the column space formed by $\boldsymbol{\Phi}$ could be identified. Since the sufficient forecasting pursues $\boldsymbol{\Sigma}_{f|y}$, we assume for simplicity that $\boldsymbol{\Phi}$ actually consists of the eigenvectors of $\boldsymbol{\Sigma}_{f|y}$. We also remind that condition (2.4.5) imposes an implicit condition on the link function $g(\cdot)$ in (2.4.2), e.g., the condition fails if $g(\cdot)$ is symmetric. Relaxing this condition is beyond the scope of this paper.

Sufficient forecasting sets out with estimated factors $\widehat{\mathbf{f}}_t$. Given T pairs $(y_{t+1}, \widehat{\mathbf{f}}_t)$, we divide them into H slices according to the order statistics of y_{t+1} and denote them by $(y_{(t+1)}, \widehat{\mathbf{f}}_{(t)})$. For ease of argument, we assume $T = cH$ for some integer c and write the sorted data as $(y_{(h,j)}, \widehat{\mathbf{f}}_{(h,j)})$, where in the double script $_{(h,j)}$, h refers to the slice number and j refers to the order number of an observation in the given

slice. Or formally,

$$y_{(h,j)} = y_{(c(h-1)+j+1)},\ \widehat{\mathbf{f}}_{(h,j)} = \widehat{\mathbf{f}}_{(c(h-1)+j)},$$

for $h = 1, \ldots, H$ and $j = 1, \ldots, c$. The sufficient forecasting estimator for (2.4.4) is

$$\widehat{\boldsymbol{\Sigma}}_{f|y} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\boldsymbol{\xi}}_h \widehat{\boldsymbol{\xi}}_h', \tag{2.4.6}$$

where $\widehat{\boldsymbol{\xi}}_h = c^{-1} \sum_{l=1}^{c} \widehat{\mathbf{f}}_{(h,l)}$ approximates the sliced inverse curve $\boldsymbol{\xi}_h = E(\mathbf{f}_t | y_{t+1} \in I_h)$. By eigenvalue decomposition of $\widehat{\boldsymbol{\Sigma}}_{f|y}$, we obtain the estimated forecasting directions $\widehat{\boldsymbol{\Phi}}$.

Chapter 3

Nonparametric Estimation of Sufficient Forecasting

3.1 Revisiting Sufficient Forecasting

In Chapter 2, we have a thorough review on the sufficient forecasting method proposed by Fan, Xue and Yao (2017). In this chapter, we shall take another look at this method and give an alternative interpretation by pointing out its close connections to many existing popular methods, such as Fama-Macbeth regression and partial least squares.

3.1.1 Connection to Fama-Macbeth Regression

The key idea in the sufficient forecasting is to consider the inverse regression curve $E(\mathbf{x}_t|y_{t+1})$. Suppose that the linearity condition on the underlying factors holds, that is, for any direction $\mathbf{b}'$ in $\mathbb{R}^K$,

$$E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t) = \sum_i c_i \boldsymbol{\phi}_i'\mathbf{f}_t \tag{3.1.1}$$

for some constants $c_i, i = 1, ..., K$. The following proposition shows that the curve $E(\mathbf{x}_t|y_{t+1})$ contains information of the forecasting directions $\boldsymbol{\Phi}$.

Proposition 3.1.1. *Under* (2.4.1)-(3.1.1)*, we have*

$$E(\mathbf{x}_t|y_{t+1}) = \mathbf{B}\mathbf{\Phi}\boldsymbol{\gamma}(y_{t+1}), \tag{3.1.2}$$

where the $L \times 1$ *vector* $\boldsymbol{\gamma}(y)$ *consists of the inverse regression curves of the forecasting indices,* $\boldsymbol{\gamma}(y_{t+1}) = E(\mathbf{\Phi}'\mathbf{f}_t|y_{t+1}) = E((\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t)'|y_{t+1})$.

Proof. It suffices to show that $E(\mathbf{f}_t|y_{t+1}) = \mathbf{\Phi}\boldsymbol{\gamma}(y_{t+1})$. The linearity condition (2.4.3) implies that $E(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t) = \mathbf{C}\mathbf{\Phi}'\mathbf{f}_t$ for some $K \times K$ matrix $\mathbf{C}$. Right multiplying both sides by $\mathbf{f}_t'\mathbf{\Phi}$ and taking expectation, we have

$$\mathbf{C} = \mathbf{C}\mathbf{\Phi}'E(\mathbf{f}_t\mathbf{f}_t')\mathbf{\Phi} = E(E(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)\mathbf{f}_t'\mathbf{\Phi}) = E(\mathbf{f}_t\mathbf{f}_t'\mathbf{\Phi}) = \mathbf{\Phi}.$$

Hence $E(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t) = \mathbf{\Phi}\mathbf{\Phi}'\mathbf{f}_t$. It follows that

$$\begin{aligned} E(\mathbf{f}_t|y_{t+1}) &= E(E(\mathbf{f}_t|\Phi'\mathbf{f}_t, y_{t+1})|y_{t+1}) \\ &= E(E(\mathbf{f}_t|\Phi'\mathbf{f}_t)|y_{t+1}) \\ &= E(\mathbf{\Phi}\mathbf{\Phi}'\mathbf{f}_t|y_{t+1}) \\ &= \mathbf{\Phi}\boldsymbol{\gamma}(y_{t+1}), \end{aligned}$$

where $\boldsymbol{\gamma}(y_{t+1}) = E(\mathbf{\Phi}'\mathbf{f}_t|y_{t+1})$, and we have used the fact that $E(\mathbf{f}_t\mathbf{f}_t') = \mathbf{I}$ as in (2.4.3) and $\mathbf{\Phi}'\mathbf{\Phi} = \mathbf{I}$. □

Note that the loading matrix $\mathbf{B}$ transforms the underlying curve $E(\mathbf{f}_t|y_{t+1})$ to $E(\mathbf{x}_t|y_{t+1})$. Since $\mathbf{x}_t$ is readily observable, the time series regression on the target unveils their loadings on the forecasting indices. The characterization is similar in spirit to the first pass of the Fama-Macbeth (FM) procedure or the recently proposed three-pass regression filter (3PRF), where they run time-series regressions for each predictor (asset) to obtain exposure to market factor or economic proxies, see, e.g., Fama and MacBeth (1973), Cochrane (2001), and Kelly and Pruitt (2015). An important distinction, however, is that their consideration is based on $\text{cov}(\mathbf{x}_t, y_{t+1})$. In our setup, this is equivalent to $E(\mathbf{x}_t y_{t+1}) = E(E(\mathbf{x}_t|y_{t+1})y_{t+1}) = \mathbf{B}\mathbf{\Phi}E(\boldsymbol{\gamma}(y_{t+1})y_{t+1})$, as $\mathbf{x}_t$ has been demeaned. Their results hence could only recover an average $\bar{\boldsymbol{\phi}}$ of the true directions, where $\bar{\boldsymbol{\phi}} = \Phi E(\boldsymbol{\gamma}(y)y_{t+1}) = \sum_{i=1}^{L} E((\boldsymbol{\phi}_i'\mathbf{f}_t)y_{t+1})\boldsymbol{\phi}_i$. Fan et al. (2017) observed the same fact in

the comparison between sufficient forecasting and principal component regression. Here, the benefit of using $E(\mathbf{x}_t|y_{t+1})$ is more clear.

3.1.2 Connection to Partial Least Squares

The ultimate goal of many forecasting problems translates into finding some predictive coefficient $\boldsymbol{\xi}$ on individual predictors. On population level, sufficient forecasting first recovers latent factor directions $\boldsymbol{\phi}_i$ of the target from $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$ and then obtains forecasting direction $\boldsymbol{\xi}_i$ on the original predictors $\mathbf{x}_t$ via $\boldsymbol{\xi}_i = \boldsymbol{\Lambda}_b'\boldsymbol{\phi}_i$, where $\boldsymbol{\Lambda}_b = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ only involves the loading matrix. One may observe that such $\boldsymbol{\xi}_i$'s reside in the column space of the loading matrix $\mathbf{B}$. Had we obtained a direction $\tilde{\boldsymbol{\xi}}$ orthogonal to the column space of $\mathbf{B}$, the predictive index $\tilde{\boldsymbol{\xi}}'\mathbf{x}_t = \tilde{\boldsymbol{\xi}}'(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t) = \tilde{\boldsymbol{\xi}}'\mathbf{u}_t$ would be completely irrelevant to the target. This can also be understood as mitigating the impact of irrelevant factors if the idiosyncratic admits further factor structure, in which case the target is only driven by a strict subset of factors that explain the cross section of the predictors. In fact, the sufficient forecasting can be derived as the solution to the following constrained optimization.

Theorem 3.1.1. *On population level, the i-th sufficient forecasting predictive coefficient on the observed predictor* $\mathbf{x}_t$ *solves*

$$\max_{\boldsymbol{\xi}} \max_{T} \text{Cov}(T(y_{t+1}), \boldsymbol{\xi}'\mathbf{x}_t) \tag{3.1.3}$$

$$\textit{subject to} \quad (\mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}')\boldsymbol{\xi} = \mathbf{0} \tag{3.1.4}$$

$$\textit{and} \quad \boldsymbol{\xi}'\mathbf{B}\mathbf{B}'\boldsymbol{\xi} = 1,\ \boldsymbol{\xi}'\mathbf{B}\mathbf{B}'\boldsymbol{\xi}_l = 0,\ l = 1, ..., i-1,$$

where maximum is taken over all bounded transform $T(\cdot)$ *and vectors* $\boldsymbol{\xi} \in \mathbb{R}^p$.

Proof. Since $\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ is a projection matrix that maps $\mathbb{R}^K$ into the the column space of $\mathbf{B}$, $\mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ is an orthogonal projection. We may write $\boldsymbol{\alpha} = \mathbf{B}\boldsymbol{\psi}$ for some $K \times 1$ vector $\boldsymbol{\psi}$, as the constraint (3.1.4) requires that $\boldsymbol{\alpha}$ stay in the column space of $\mathbf{B}$. By (2.4.3), $\mathbf{B}'\mathbf{B}$ is diagonal, so we write $\boldsymbol{\alpha} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\boldsymbol{\phi} = \boldsymbol{\Lambda}_b'\boldsymbol{\phi}$ for some $\boldsymbol{\phi}$ equivalently, where $\boldsymbol{\Lambda}_b = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. One observes that this representation connects $\boldsymbol{\alpha}$ to the direction $\boldsymbol{\phi}$ in the latent factor space. As a

result, $\boldsymbol{\alpha}'\mathbf{x}_t = \boldsymbol{\phi}'\boldsymbol{\Lambda}_b(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t) = \boldsymbol{\phi}'\mathbf{f}_t + \boldsymbol{\phi}'\boldsymbol{\Lambda}_b\mathbf{u}_t$. Since $\mathbf{u}_t$ and y_{t+1} are independent, the optimization hence translates into

$$\begin{aligned} & \max_{\boldsymbol{\phi}} \max_{T} \operatorname{Cov}(T(y_{t+1}), \boldsymbol{\phi}'\mathbf{f}_t) \\ \text{subject to} \quad & \|\boldsymbol{\phi}\| = 1,\ \boldsymbol{\phi}'\boldsymbol{\phi}_l = 0,\ l = 1, ..., i-1, \end{aligned}$$

for the i-th predictive coefficient. It can be shown that the optimal transformation for any direction $\boldsymbol{\phi}$ is $T_{op}(y_{t+1}) = E(\boldsymbol{\phi}'\mathbf{f}_t|y_{t+1}) = \boldsymbol{\phi}'E(\mathbf{f}_t|y_{t+1})$. By conditional expectation, we have

$$\begin{aligned} \operatorname{Cov}(T_{op}(y_{t+1}), \boldsymbol{\phi}'\mathbf{f}_t) &= E(T_{op}(y_{t+1})\mathbf{f}_t'\boldsymbol{\phi}) = E[T_{op}(y_{t+1})E(\mathbf{f}_t'\boldsymbol{\phi}|y_{t+1})] \\ &= \boldsymbol{\phi}'E(E(\mathbf{f}_t|y_{t+1})E(\mathbf{f}_t|y_{t+1})')\boldsymbol{\phi} \\ &= \boldsymbol{\phi}'\operatorname{Cov}(E(\mathbf{f}_t|y_{t+1}))\boldsymbol{\phi}. \end{aligned}$$

Therefore, the eigenvalue decomposition of $\operatorname{Cov}(E(\mathbf{f}_t|y_{t+1}))$ solves the maximization problem (3.1.3), which completes the proof. □

Remark 3.1.1. Theorem 3.1.1 shows that the direction $\boldsymbol{\xi}$ we search for lives in the kernel of the projection matrix $\mathbf{I} - \mathbf{M}_b = \mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. As discussed earlier, it ensures that noises irrelevant to y_{t+1} drop out of the forecast. Kelly and Pruitt (2015) has a similar interpretation of their 3PRF model, but they resort to proxies to determine relevant factor space. Our approach is more general, as it involves a general transformation $T(\cdot)$ of the forecast target.

Remark 3.1.2. The characterization above also allows us to reveal its close connections to the partial least squares (PLS) method. The i-th PLS direction solves

$$\begin{aligned} & \max_{\boldsymbol{\xi}} \operatorname{Cov}(y_{t+1}, \boldsymbol{\xi}'\mathbf{x}_t) \\ \text{s.t.} \quad & \|\boldsymbol{\xi}\| = 1,\ \boldsymbol{\xi}'\mathbf{S}\boldsymbol{\xi}_l = 0,\ l = 1, ..., i-1, \end{aligned}$$

where $\mathbf{S} = \operatorname{cov}(\mathbf{x}_t)$; see Frank and Friedman (1993). Since $\mathbf{S} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}_u$ in our setup, where $\boldsymbol{\Sigma}_u = \operatorname{cov}(\mathbf{u}_t)$, additional forecasting directions obtained from PLS would be mixed with undesired noises. The constraint $\boldsymbol{\xi}'\mathbf{B}\mathbf{B}'\boldsymbol{\xi} = 1$ in the constraint (3.1.4) pertains to the normalization of the forecasting directions $\boldsymbol{\phi}_i$'s on the latent

factor, and is therefore inconsequential. Again, our approach is more flexible and gracefully handles non-linearity with the presence of $T(\cdot)$ in the optimization.

3.2 Asymptotics with Diverging K

With the growing popularity of various dimension reduction techniques, the use of factors in statistical models to summarize information becomes more and more popular (Fan, Fan and Lv, 2008; Fan, Liao and Mincheva, 2013; Fan, Liao and Wang, 2015). However, it is often non-trivial to conduct inference for such models, as the factors are estimated. In this Chapter, we extend the sufficient forecasting method of Fan, Xue and Yao (2017) by allowing the number of factors K to increase as $p, T \to \infty$ and conduct corresponding theoretica analysis. We also provide with a set of sufficient conditions for consistent estimation of sufficient forecasting directions.

3.2.1 Assumptions

Assumption 3.2.1 (Factors and Loadings)**.**

(i) There exists $b > 0$ such that $\sup_{p\in\mathbb{N}} \|\mathbf{B}\|_{\max} \leq b$, and there exist two positive constants c_1 and c_2 such that

$$c_1 < p^{-1}\lambda_{\min}(\mathbf{B}'\mathbf{B}) < p^{-1}\lambda_{\max}(\mathbf{B}'\mathbf{B}) < c_2.$$

(ii) Identification: $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$, and $\mathbf{B}'\mathbf{B}$ is a diagonal matrix with distinct entries.

Next, we impose the strong mixing condition on the data generating process. Denote by $\mathcal{F}_{\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ the σ-algebras generated by $\{(\mathbf{f}_t, \mathbf{u}_t, \epsilon_{t+1}) : t \leq 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t, \epsilon_{t+1}) : t \geq T\}$ respectively. Define the mixing coefficient as

$$\alpha(T) = \sup_{A\in\mathcal{F}_{\infty}^{0}, B\in\mathcal{F}_{T}^{\infty}} |P(A)P(B) - P(AB)|.$$

Assumption 3.2.2 (Data Generating Process)**.** *$\{\mathbf{f}_t\}_{t\geq 1}$, $\{\mathbf{u}_t\}_{t\geq 1}$ and $\{\epsilon_{t+1}\}_{t\geq 1}$ are three independent groups, and all of them are strictly stationary.*

(i) Both $\{K^{-2}E\|\mathbf{f}_t\|^4 : p \in \mathbb{N}\}$ *and* $\{K^{-1}E(\|\mathbf{f}_t\|^2|y_{t+1}) : p \in \mathbb{N}\}$ *are bounded sequences.*

(ii) There exist some constants $C > 0$ *and* $l > 0$ *that* $E(\exp(l|\epsilon_{t+1}|)) \leq C$ *for any* $t \geq 1$.

(iii) The mixing coefficient $\alpha(T) < c\rho^T$ *for* $T \in \mathbb{Z}^+$, *some* $c > 0$ *and some* $\rho \in (0,1)$. *In addition,* $\alpha(T) \leq exp(-cT^{\gamma_1})$ *for all* $T \in \mathbb{Z}^+$ *and some positive constants* γ_1 *and* c.

Assumption 3.2.3 (Residuals and Dependence)**.** *There exists a positive constant* $M < \infty$ *that does not depend on* p *or* T, *such that*

(i) $E(\mathbf{u}_t) = \mathbf{0}$, *and* $E|u_{it}|^8 \leq M$.

(ii) $\|\mathbf{\Sigma}_u\|_1 \leq M$, *and for every* $i, j, t, s > 0$, $(pT)^{-1}\sum_{i,j,t,s}|E(u_{it}u_{js})| \leq M$

(iii) For every (t, s), $E|p^{-1/2}(\mathbf{u}_s'\mathbf{u}_t - E(\mathbf{u}_s'\mathbf{u}_t))|^4 \leq M$.

3.2.2 Asymptotic Properties of Estimated Factors and Loadings

In this section, we lay out the asymptotic properties pertaining to the estimated factors and loadings, which serve as our cornerstone for forecasting. We extend the method of Fan, Xue and Yao (2017) by allowing the number of factors K to increase as $p, T \to \infty$, which not only avoids the possible model misspecification (Li, Li and Shi, 2017) but also accommodate the potential structural changes (Ludvigson and Ng, 2007). Specifically, the estimation of factors is based on the asymptotic principal components as follows:

$$\begin{aligned}(\widehat{\mathbf{B}}_K, \widehat{\mathbf{F}}_K) &= \arg\min_{(\mathbf{B},\mathbf{F})} \|\mathbf{X} - \mathbf{B}\mathbf{F}'\|_F^2, \\ &\text{subject to } T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal},\end{aligned} \tag{3.2.1}$$

where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T), \mathbf{F}' = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$, and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Since $\mathbf{B}$ and $\mathbf{F}$ are not separately identified, the normalization $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$ and that $\mathbf{B}'\mathbf{B}$ is diagonal are necessary and correspond to (2.4.3).

Such conditions describe but do not impose any structure on the data, nor would a diverging K place any restrictions on the data $\mathbf{X}$. The solution for $\mathbf{F}$, denoted by $\widehat{\mathbf{F}}_K$, is $\sqrt{T}$ times the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $\mathbf{X}'\mathbf{X}$. The solution for $\mathbf{B}$, denoted by $\widehat{\mathbf{B}}_K$, is $T^{-1}\mathbf{X}\widehat{\mathbf{F}}_K$. To simplify notation, we let $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}_K$ and $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}_K$.

The asymptotic properties of the estimated factors and loadings are summarized in the following proposition.

Proposition 3.2.1. *Under Assumptions 3.1-3.3, suppose* $K = o(\min\{p^{1/3}, T\})$, *then*

1) $\frac{1}{p}\|\widehat{\mathbf{B}} - \mathbf{B}\|^2 = O_p(\frac{K^3}{p} + \frac{K}{T})$,

2) $\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{F}\|^2 = O_p(\frac{K^3}{p} + \frac{K}{T})$.

Further suppose that $K = o(\min\{p^{1/4}, T^{1/2}\})$, *then the conditional expectation* $\boldsymbol{\xi}_h = E(\mathbf{f}_t | y_{t+1} \in I_h)$ *is approximated by* $\widehat{\boldsymbol{\xi}}_h = c^{-1}\sum_{l=1}^{c}\widehat{\mathbf{f}}_{(h,l)}$ *as in* (2.4.6) *with the following accuracy*

3) $\|\widehat{\boldsymbol{\xi}}_h - \boldsymbol{\xi}_h\| = O_p(\frac{K^{3/2}}{p^{1/2}} + \frac{K}{T^{1/2}})$.

Proposition 3.2.1 suggests that the cross-sectional average of estimation errors in loadings and the time-series average of estimation errors in factors, as measured in the spectral norm, all vanish when $p, T \to \infty$. The convergence rate depends both on the panel structure p, T and on the factor structure K. Also, the sliced inverse curve $\boldsymbol{\xi}_h$ can be consistently estimated by its sample counterpart $\widehat{\boldsymbol{\xi}}_h$. We note that while there are alternative methods for estimating factors and loadings, such as the quasi-maximum likelihood by Bai and Li (2012), principal component estimation remains a simple and popular choice.

3.2.3 Asymptotic Properties of Estimated Directions

Suppose we have at our disposal a reasonably well estimated factor $\widehat{\mathbf{f}}_t$ for forecasting. We now show how they affect the accuracy of forecasting directions $\widehat{\boldsymbol{\Phi}}$, without requiring knowledge of the procedure of factor estimation. Let

$\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_H)$ be the collection of the sliced regression curves and $\widehat{\boldsymbol{\Xi}} = (\widehat{\boldsymbol{\xi}}_1, ..., \widehat{\boldsymbol{\xi}}_H)$ be the corresponding estimate. It is straightforward to translate (2.4.4) and (2.4.6) into matrix notation: $\boldsymbol{\Sigma}_{f|y} = H^{-1}\boldsymbol{\Xi}\boldsymbol{\Xi}'$ and $\widehat{\boldsymbol{\Sigma}}_{f|y} = H^{-1}\widehat{\boldsymbol{\Xi}}\widehat{\boldsymbol{\Xi}}'$. Denote by $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}$ the difference between the estimated and true regression curves. We make the following high-level assumption for $\boldsymbol{\Delta}$.

Assumption 3.2.4. *There exists a positive sequence* $\omega_{p,T,K} = o(1)$ *such that* $||\boldsymbol{\Delta}|| = O_p(\omega_{p,T,K})$

The estimation accuracy of $\boldsymbol{\Delta}$ hinges on the quality of the estimated factors, which involves cross-sectional and time-series dimensions as well as the factor structure. Evidently, $\omega_{p,T,K} = K^{3/2}/p^{1/2} + K/T^{1/2}$ should we follow the principal component estimation (3.2.1).

Define an $H \times L$ matrix $\boldsymbol{\Gamma} = \boldsymbol{\Xi}'\boldsymbol{\Phi}$, or,

$$\boldsymbol{\Gamma} = \begin{pmatrix} E(\mathbf{f}_t'\boldsymbol{\phi}_1|y_{t+1} \in I_1) & \cdots & E(\mathbf{f}_t'\boldsymbol{\phi}_L|y_{t+1} \in I_1) \\ \vdots & \ddots & \vdots \\ E(\mathbf{f}_t'\boldsymbol{\phi}_1|y_{t+1} \in I_H) & \cdots & E(\mathbf{f}_t'\boldsymbol{\phi}_L|y_{t+1} \in I_H) \end{pmatrix},$$

each row of which depicts the projections of sliced inverse regression functions on different forecasting directions.

The following theorem characterizes the behavior of estimated forecasting directions given the commonly used linearity assumption.

Theorem 3.2.1. *Suppose* $\boldsymbol{\Gamma}$ *is of full rank, the coverage condition* (2.4.5) *holds, and the* L *largest eigenvalues of* $\boldsymbol{\Sigma}_{f|y}$ *are positive and distinct. Under Assumptions 3.1–3.4 and the linearity condition that* $E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t)$ *is a linear function of* $\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t$ *for any* $\mathbf{b} \in \mathbb{R}^p$*, the sufficient forecasting direction estimates* $\widehat{\boldsymbol{\Phi}}$ *have the following approximation,*

$$\widehat{\boldsymbol{\Phi}} = \boldsymbol{\Phi} + (\mathbf{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}')\boldsymbol{\Delta}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1} + o_p(\omega_{p,T,K}). \tag{3.2.2}$$

The proof given in the appendix sheds light on how such decomposition is possible. A direct implication of Theorem 3.2.1 is that the estimated forecasting directions are consistent, i.e., $\widehat{\boldsymbol{\Phi}} = \boldsymbol{\Phi} + O_p(\omega_{p,T,K})$. This result generalizes the

findings in Theorem 3.1 of Fan, Xue and Yao (2017), and provides finer details of the estimation quality.

Theorem 3.2.1 depends on the connection between the column space of $\mathbf{\Sigma}_{f|y}$ and the true forecasting directions given in $\mathbf{\Phi}$, where the linearity condition plays an important role (Li, 1991). One important family of distributions that satisfies the linearity condition is the well-known elliptically symmetric distributions. However, as pointed out in Xia, Tong, Li and Zhu (2002), when lags are included in the forecasting, the elliptical symmetry implies an undesirable time reversibility. Thus, it is essential to relax the linearity condition. In the sequel, we establish the consistency of the sufficient forecasting direction estimates $\widehat{\mathbf{\Phi}}$ with a diverging K without requiring the restricted linearity condition.

Let $\sin(\cdot,\cdot)$ be the sine of the angle between two real vectors of equal dimension under the usual Euclidean inner product, $\gamma(\cdot)$ be the density function of $\|\mathbf{f}_t\|^{-1}\mathbf{f}_t$ with respect to the uniform distribution on the unit hypersphere in $\mathbb{R}^K$, and $\mathbf{\Upsilon}$ be an orthonormal basis of the orthogonal complement of the central subspace. For $0 < c < 1$, let $B(c) = \{\mathbf{f}_t : \|\mathbf{f}_t\|^2 \leq K(1-c)\}$ and $I(B(c))$ be the indicator function of $\mathbf{f}_t$ for $B(c)$. We introduce the following assumption to relax the linearity condition when there is a diverging number of factors.

Assumption 3.2.5. *The factors* $\mathbf{f}_t$ *satisfy that*

(i) The conditional covariance $\mathrm{cov}(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)$ *is degenerate.*

(ii) $P(|K^{-1}\|\mathbf{f}_t\|^2 - 1| \geq c) = o(K^{-1})$ *and* $E\{K\|\mathbf{f}_t\|^{-2}I(B(c))\} = o(K^{-1})$ *for any* $0 < c < 1$.

(iii) $E\{\sup_{|\sin(\mathbf{f}_t,\mathbf{e})|\leq c}\gamma(\mathbf{e})\} = o(K^{-1/2}c^{-K})$ *for some* $0 < c \leq 1$.

(iv) There exists a function $u(\cdot) : \mathbb{R}^L \to \mathbb{R}$ *such that* $E\{u(\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)\}$ *exists and* $\|E(\mathbf{\Upsilon}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)\|^4 \leq u(\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)$ *almost surely.*

It is worth pointing out that Assumption 3.2.5 is similar to regularity conditions in Hall and Li (1993). Specifically, (i)–(iii) are introduced to show that the linearity condition (Li, 1991) approximately holds given a diverging number of factors, and (iv) helps control the remainder term of this approximation.

Given Assumption 3.2.5, the following theorem provides an important consistency result for the sufficient forecasting without requiring the linearity condition.

Theorem 3.2.2. *Suppose* $\mathbf{\Gamma}$ *is of full rank, the coverage condition* (2.4.5) *holds, and the* L *largest eigenvalues of* $\mathbf{\Sigma}_{f|y}$ *are positive and distinct. Under Assumptions 3.1–3.5, the sufficient forecasting direction estimates* $\widehat{\mathbf{\Phi}}$ *have the following approximation,*

$$\widehat{\mathbf{\Phi}} = \mathbf{\Phi} + (\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}')\mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1} + o_p(1). \tag{3.2.3}$$

3.3 Nonparametric Estimation

3.3.1 Local Linear Regression

The nonparametric regression model (2.4.2) can be written as

$$y_{t+1} = g(\boldsymbol{r}(\mathbf{f}_t)) + \epsilon_{t+1}, \quad \text{with } \mathbb{E}(\epsilon_{t+1}) = 0 \tag{3.3.1}$$

where $\boldsymbol{r}(\mathbf{f}_t) = (\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t)'$ denotes the L-dimensional predictive indices constructed from K-dimensional factors. Our goal now is to provide a nonparametric estimation of the unknown forecast function $g(\mathbf{z}) = \mathbb{E}(y_{t+1}|\boldsymbol{r}(\mathbf{f}_t) = \mathbf{z})$ given the independence of $\{\mathbf{f}_t\}$ and $\{\epsilon_{t+1}\}$. Note that $\boldsymbol{r}(\mathbf{f}_t)$ can be consistently estimated by $\hat{\boldsymbol{r}}(\hat{\mathbf{f}}_t) = (\hat{\boldsymbol{\phi}}_1'\hat{\mathbf{f}}_t, \cdots, \hat{\boldsymbol{\phi}}_L'\hat{\mathbf{f}}_t)$ using the sufficient forecasting procedure, as established in previous sections. For ease of notation, we define $\boldsymbol{r}_t = \boldsymbol{r}(\mathbf{f}_t)$ and $\hat{\boldsymbol{r}}_t = \hat{\boldsymbol{r}}(\hat{\mathbf{f}}_t)$. Note that covariates $\boldsymbol{r}_t$ are not observed but have to be estimated nonparametrically from data. As shown in Mammen et al. (2012), many economic applications require the nonparametric estimation of the regression function with nonparametrically generated covariates $\hat{\boldsymbol{r}}(\hat{\mathbf{f}}_t)$, such as the simultaneous nonparametric equation models (Newey et al., 1999; Imbens and Newey, 2009) and the structural equations for treatment effects (Heckman and Vytlacil, 2005).

In what follows, we estimate $\hat{g}(\mathbf{z})$ through a nonparametric regression of y_{t+1} on $\hat{\boldsymbol{r}}(\mathbf{f}_t)$ by using local linear smoothing technique (Fan and Gijbels, 1996), i.e., $\hat{g}(\mathbf{z}) = \hat{\alpha}$ is obtained by

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \underset{\alpha, \boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=0}^{T-1} \left(y_{t+1} - \alpha - \boldsymbol{\beta}'(\hat{\boldsymbol{r}}_t - \mathbf{z})\right)^2 K_h(\hat{\boldsymbol{r}}_t - \mathbf{z}) \tag{3.3.2}$$

where $K_h(\mathbf{u}) = \frac{1}{h^L}\prod_{j=1}^{L}\mathcal{K}(\frac{u_j}{h})$ is a product kernel with univariate kernel function

K, and $h > 0$ is smoothing bandwidth.

Specifically, let $\mathbf{e}_1 = (1, 0, \cdots, 0)' \in \mathbb{R}^{L+1}$, $Y = (y_1, \cdots, y_T)'$, $\mathbf{v}_t(\boldsymbol{r}, \mathbf{z}) = (1, (\boldsymbol{r}(\mathbf{f}_t) - \mathbf{z})')'$, $\hat{Z} = (\mathbf{v}_0(\hat{\boldsymbol{r}}, \mathbf{z}), \cdots, \mathbf{v}_{T-1}(\hat{\boldsymbol{r}}, \mathbf{z}))'$. $W_{\hat{\boldsymbol{r}}} = \text{diag}(K_h(\hat{\boldsymbol{r}}_0 - \mathbf{z}), \cdots, K_h(\hat{\boldsymbol{r}}_{T-1} - \mathbf{z}))$ is a diagonal weighting matrix. Then the solution to the local linear regression (3.3.2) is

$$\hat{g}(\mathbf{z}) = e_1' \left(\hat{Z}' \hat{W} \hat{Z} \right)^{-1} \left(\hat{Z}' \hat{W} Y \right) \tag{3.3.3}$$

3.3.2 Prediction Consistency

Before proceeding, we provide a set of assumptions that will be needed to achieve consistent prediction.

Assumption 3.3.1 (Regularity)**.**

(i) The density function $f_{\mathbf{z}}(\mathbf{z})$ of random vector $\mathbf{z} = \boldsymbol{r}(\mathbf{f})$ is twice continuously differentiable and bounded away from 0 on a compact support $I_{\mathbf{z}}$.

(ii) The regression function $g(\mathbf{z})$ is twice continuously differentiable on $I_{\mathbf{z}}$

(iii) The kernel $\mathcal{K}(\cdot)$ is a symmetric, twice continuously differentiable, compactly supported density function.

(iv) The bandwidth h satisfies $h \sim T^{-\eta}$, and $\eta < \frac{1}{L}$

Assumption 3.3.2 (Accuracy)**.** *For some $\delta > \eta$, the estimation of $\hat{r}(\mathbf{x})$ satisfies*

$$\sup ||\hat{\boldsymbol{r}} - \boldsymbol{r}||_\infty = o_P(T^{-\delta})$$

Assumption 3.3.3 (Complexity)**.** *There exists sequences of sets $\mathcal{M}_{T,j}$ such that*

(i) $Pr(r_j \in \mathcal{M}_{T,j}) \to 1$ as $n \to \infty$ for all $j = 1, \cdots, L$

(ii) For a constant $C_M > 0$ and a function $r_{n,j}$ with $||r_{T,j} - r_{0,j}||_\infty = o(T^{-\delta})$, the set $\bar{\mathcal{M}}_{T,j} = \mathcal{M}_{T,j} \cap \{r_j : ||r_j - r_{T,j}||_\infty \leq T^{-\delta}\}$ can be covered by at most $C_M exp(\lambda^{-\alpha_j} T^{\xi_j})$ balls with $||\cdot||_\infty$-radius λ for all $\lambda \leq T^{-\delta}$, where $0 < \alpha_j \leq 2$, $\xi_j \in \mathbb{R}$.

We have the following theorems on the estimation consistency of the forecasting function.

Theorem 3.3.1. *Suppose Assumptions 3.2.2 and 3.3.1-3.3.3 hold and* $\kappa = min\{\kappa_1, \kappa_2, \kappa_3\}$*, then*

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\hat{g}(\mathbf{z}) - \tilde{g}(\mathbf{z}) + \nabla' g(\mathbf{z})\hat{\Delta}(\mathbf{z})| = O_p(T^{-\kappa}) \tag{3.3.4}$$

with

$$\kappa_1 < \delta + 1 - (L+1)\eta - (\frac{1}{\gamma_1} + 1)\max(\delta\alpha_j + \xi_j), \quad \kappa_2 < \delta + \eta, \quad \kappa_3 < 2\delta - \eta,$$

where $\tilde{g}(\mathbf{z})$ *is an infeasible estimator, expected to be fitted by the true value* $\boldsymbol{r}$ *instead of the estimated* $\hat{\boldsymbol{r}}$*, i.e.,* $\tilde{g}(\mathbf{z}) = \tilde{\alpha}$ *is obtained by*

$$(\tilde{\alpha}, \tilde{\boldsymbol{\beta}}) = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=0}^{T-1} \left(y_{t+1} - \alpha - \boldsymbol{\beta}'(\boldsymbol{r}_t - \mathbf{z})\right)^2 K_h(\boldsymbol{r}_t - \mathbf{z}). \tag{3.3.5}$$

$\hat{\Delta}(\mathbf{z}) = \bar{\boldsymbol{\alpha}}$ *is obtained by*

$$(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=0}^{T-1} \left((\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) - \boldsymbol{\alpha} - \boldsymbol{\beta}'(\boldsymbol{r}_t - \mathbf{z})\right)^2 K_h(\boldsymbol{r}_t - \mathbf{z}). \tag{3.3.6}$$

Theorem 3.3.2. *Under the same assumption as theorem 3.3.1, the consistency of nonparametric estimator with generated covariates* $\hat{g}(\mathbf{z})$ *is given by*

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\hat{g}(\mathbf{z}) - g(\mathbf{z})| = O_p(\sqrt{\log(T)T^{-1+L\eta}} + T^{-2\eta} + T^{-\delta} + T^{-\kappa}). \tag{3.3.7}$$

3.4 Empirical Analysis

In this section, we apply the sufficient forecasting method to financial data, and assess the predictability of the daily market return using the cross-section of stock returns. Our dataset is drawn from the Center for Research in Security Prices (CRSP) database. The market return is proxied by S&P 500 index return, whereas the cross-section of equities consists of 310 large-cap stock returns from

2007 to 2016 without missing data. Most of the existing literature on market return predictability (Fama and French (1993), Kelly and Pruitt (2015)) focus on monthly frequency and rely on portfolios information (e.g., from Kenneth French's website). By contrast, we examine the issue using individual stock returns at the daily frequency. Not only is such data readily available for a long time from various sources, but it also provides enough sample to conduct accurate estimation.

The focus of our study is a rolling out-of-sample forecast implementation. At date t, our forecast target y_{t+1} is the market return that is realized over the next day $t+1$, while our estimate is based on time t information. Specifically, our factors $\mathbf{f}_s$ are constructed through (3.2.1) using daily stock returns $\mathbf{x}_s (s = t-755, .., t)$ of the past three years of trading days. We then collect $\{(\mathbf{f}_s, y_{s+1}) : s = t-755, ..., t-1\}$ (or simply the raw data $\{(\mathbf{x}_s, y_{s+1})\}$) to estimate our predictive model, whose parameters are t-measurable. Finally, we use the latest information $\mathbf{f}_t$ (or $\mathbf{x}_t$) alone with the estimated model to make forecast. The evaluation of different models is based on the closeness of their estimated market returns and true market returns from 2010 to 2016.

Our models consist of sufficient forecasting (SF), principal component regression (PCR) and partial least squares (PLS). For the first two models, we use seven estimated factors extracted from the return panel, which on average account for around 60% of the variation in the cross-section of stock returns. We denote by SF(i) the sufficient forecasting with $i = 1, 2$ predictive indexes. To reveal its potential, we consider both linear (L-SF) and nonlinear sufficient (NL-SF) forecasting in building the predictive regression, where the NL-SF uses local linear regression.

Table 3.1. Correlation matrix of predicted market returns via different models from 2010 to 2016

	L-SF(1)	L-SF(2)	NL-SF(1)	NL-SF(2)	PCR	PLS
L-SF(1)	1.00	0.90	0.67	0.55	0.75	0.54
L-SF(2)	-	1.00	0.60	0.67	0.83	0.62
NL-SF(1)	-	-	1.00	0.66	0.49	0.39
NL-SF(2)	-	-	-	1.00	0.56	0.44
PCR	-	-	-	-	1.00	0.77
PLS	-	-	-	-	-	1.00

Table 3.1 reports the time-series correlation of the predicted market returns via different models. We first observe that those predictions are positively correlated, indicating that these models are making similar bets on the next-day market returns. Second, predictions from sufficient forecasting are more correlated with PCR than PLS, as both SF and PCR depend on the estimated factors from in the first step. In comparison, PLS starts directly from the return panel $\mathbf{x}_t$ and ignores the factor structure, resulting in quite different forecasting directions on the original predictors. Third, nonlinear forecasts can be very different from linear forecasts. Figure 3.1 plots the histograms of predicted market returns by each method. It is noticeable that the forecasts made by SF are more concentrated around zero, while PLS's forecast has much wider distributions.

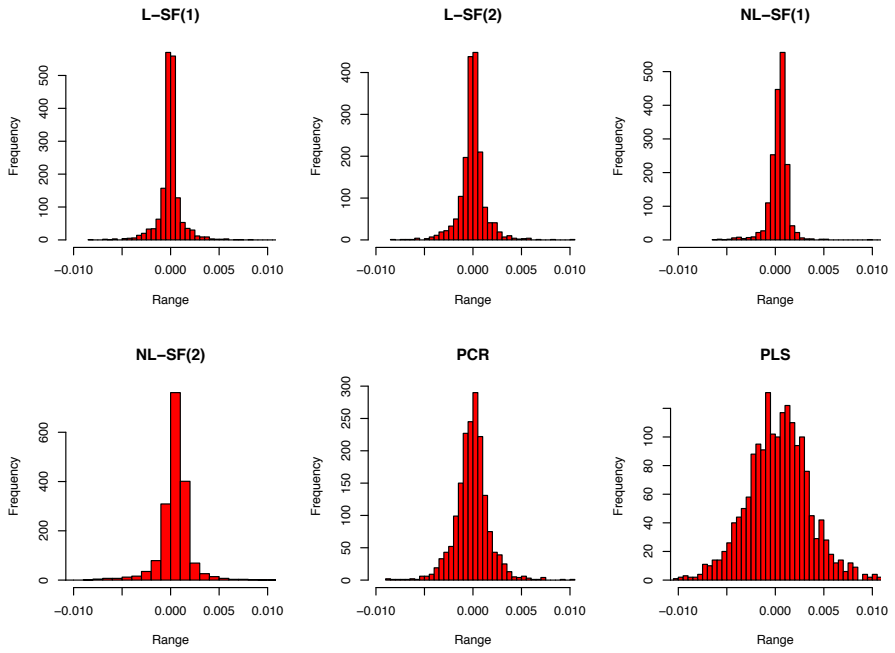

Figure 3.1. Histograms of Predicted Market Returns

Table 3.2. Prediction summary

	L-SF(1)	L-SF(2)	NL-SF(1)	NL-SF(2)	PCR	PLS
cor	7.73%	7.78%	7.50%	8.25%	6.72%	6.08%
RMSE	0.998	0.999	0.995	1.002	1.010	1.068

Notes: Predictability measures by different models. The first row reports the correlation between realized and predicted market returns. The second row lays out the relative mean square error (RMSE) to the mean market return in the evaluation period.

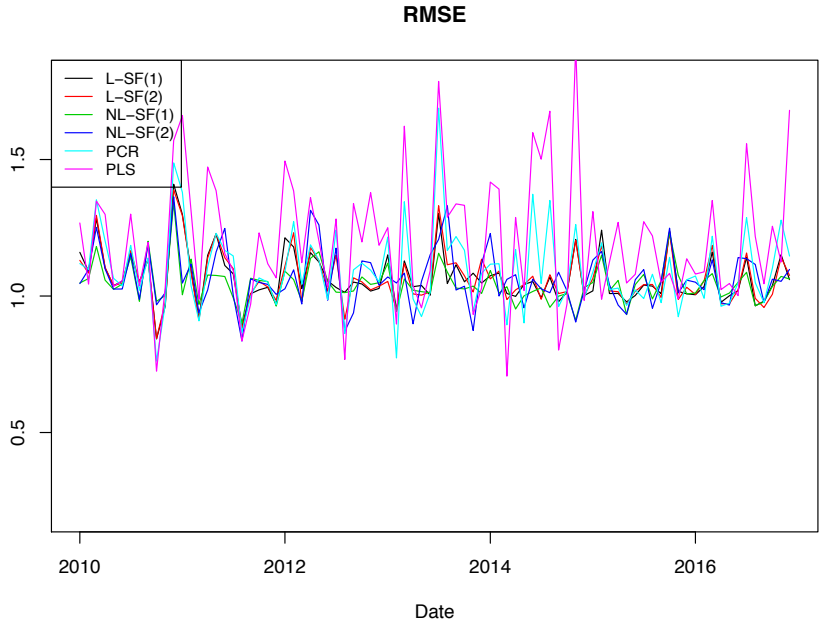

Figure 3.2. Monthly RMSE over the Evaluation Period

Table 3.2 shows the predictive power of the four methods. For each method M, we consider its correlation with the target, and its mean squared error relative to the out-of-sample mean,

$$\text{RMSE}(M) = \frac{\sum_{t \in S}(y_t - \hat{y}_t)^2}{\sum_{t \in S}(y_t - \bar{y}_t)^2},$$

where S is the evaluation sample and $\bar{y}_t$ is the mean of y_t in S. As we can see, sufficient forecasting methods yield slightly better performance than the other methods, mostly because it is a more concise model and is less prone to overfit. By exploring the nonlinear nature of the market returns, NL-SF(2) delivers additional predictive power in terms of correlation metrics. Figure 3.2 further shows that the RMSEs for L-SF, NL-SF and PCR are relatively consistent through different months, and the fact that they are close to 1 indicates that these methods' predictability of market returns can get as near as market average daily returns. On the other hand, PLS does not exploit the factor structure in the cross-section of stock returns and requires further refinement to improve its predictive power.

3.5 Proofs

3.5.1 Proof of Proposition 3.2.1

Proof. We leverage on the existing results concerning the asymptotic properties of estimated loadings and factors, and sketch the proof as follows. First, we let $\mathbf{V}$ denote the $K \times K$ diagonal matrix consisting of the K largest eigenvalues of the sample covariance matrix $(pT)^{-1}\mathbf{X}'\mathbf{X}$ in descending order. Define a $K \times K$ matrix

$$\mathbf{H} = \mathbf{V}^{-1}\frac{\widehat{\mathbf{F}}'\mathbf{F}}{T}\frac{\mathbf{B}'\mathbf{B}}{p}, \tag{3.5.1}$$

where $\mathbf{F}' = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$ and $\widehat{\mathbf{F}}$ is estimated from principal component analysis. Under Assumptions 3.1-3.4, the following facts can be sequentially developed, which are detailed in Luo et al. (2017). First, for $K = o(\min\{p^{1/3}, T\})$, the factors are estimated consistently up to the rotation matrix $\mathbf{H}$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 = O_p(\frac{K^3}{p} + \frac{K}{T}).$$

In addition, in matrix notation, we have the following facts,

$$\frac{1}{T}\|(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')'\mathbf{F}\| \quad = \quad O_p(\frac{K^2}{p} + \frac{K}{T}),$$

$$\frac{1}{T}\|(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')'\widehat{\mathbf{F}}\| \quad = \quad O_p(\frac{K^3}{p} + \frac{K}{T}).$$

The above results, together with Assumptions 3.1-3.4 and in particular the identification assumption in Assumption 3.1, imply that the rotation matrix is asymptotically identity,

$$\|\mathbf{H} - \mathbf{I}_K\| = O_p(\frac{K^3}{p} + \frac{K}{T}).$$

It then follows that $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_p(p^{1/2}(K^{3/2}p^{-1/2} + K^{1/2}T^{-1/2}))$ and $\|\widehat{\mathbf{F}} - \mathbf{F}\| = O_p(T^{1/2}(K^{3/2}p^{-1/2} + K^{1/2}T^{-1/2}))$. Let $\mathbf{\Lambda}_b = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ and $\widehat{\mathbf{\Lambda}}_b = (\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}'$. We have

$$\|\widehat{\mathbf{\Lambda}}_b - \mathbf{\Lambda}_b\| = O_p(p^{-1/2}(K^{3/2}p^{-1/2} + K^{1/2}T^{-1/2})).$$

Next, letting $\widehat{\mathbf{m}}_h = \frac{1}{c}\sum_{l=1}^{c}\mathbf{x}_{(h,l)}$ and $\mathbf{m}_h = E(\mathbf{x}_t|y_{t+1} \in I_h)$ its population version, we have

$$\begin{aligned}\|\widehat{\mathbf{m}}_h - \mathbf{m}_h\| &= \|\frac{1}{c}\sum_{l=1}^{c}(\mathbf{B}\mathbf{f}_{(h,l)} + \mathbf{u}_{(h,l)}) - \mathbf{B}E(\mathbf{f}_t|y_{t+1} \in I_h)\| \\ &\leq \|\mathbf{B}\| \cdot \|\frac{1}{c}\sum_{l=1}^{c}\mathbf{f}_{(h,l)} - E(\mathbf{f}_t|y_{t+1} \in I_h)\| + \|\frac{1}{c}\sum_{l=1}^{c}\mathbf{u}_{(h,l)}\|.\end{aligned}$$

Under Assumption 3.2 and due to central limit theorem on weak dependent data (Theorem 19.2 in Billingsley (1999)), the sample mean $\frac{1}{c}\sum_{l=1}^{c}\mathbf{f}_{(h,l)}$ converges to the population mean $E(\mathbf{f}_t|y_{t+1} \in I_h)$ at the rate of $O_p(K^{1/2}T^{-1/2})$. So does $\mathbf{u}_t|y_{t+1} \in I_h$. Therefore,

$$\|\widehat{\mathbf{m}}_h - \mathbf{m}_h\| \quad = \quad O_p(p^{1/2}) \cdot O_p(K^{1/2}T^{-1/2}) + O_p(K^{1/2}T^{-1/2}) = O_p(\sqrt{pK/T}).$$

Note that $\mathbf{x}_t$ and y_{t+1} are conditionally independent given $\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t$. Then by using this fact and Assumption (3.4), we know that $\|E(\mathbf{f}_t|y_{t+1} \in I_h)\| = O_p(1)$. Hence, we have

$$\|\mathbf{m}_h\| = \|E(\mathbf{x}_t|y_{t+1} \in I_h)\| \leq \|\mathbf{B}\| \cdot \|E(\mathbf{f}_t|y_{t+1} \in I_h)\| = O_p(p^{1/2}).$$

The above two inequalities immediately imply that

$$\|\widehat{\mathbf{m}}_h\| \leq \|\widehat{\mathbf{m}}_h - \mathbf{m}_h\| + \|\mathbf{m}_h\| = O_p(p^{1/2}).$$

Now, the triangle inequality implies that

$$\begin{aligned}
\|\widehat{\boldsymbol{\xi}}_h - \boldsymbol{\xi}_h\| &= \|\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h - \boldsymbol{\Lambda}_b\mathbf{m}_h\| \\
&\leq \|\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h - \boldsymbol{\Lambda}_b\widehat{\mathbf{m}}_h\| + \|\boldsymbol{\Lambda}_b\widehat{\mathbf{m}}_h - \boldsymbol{\Lambda}_b\mathbf{m}_h\| \\
&\leq \|\widehat{\boldsymbol{\Lambda}}_b - \boldsymbol{\Lambda}_b\| \cdot \|\widehat{\mathbf{m}}_h\| + \|\boldsymbol{\Lambda}_b\| \cdot \|\widehat{\mathbf{m}}_h - \mathbf{m}_h\| \\
&= O_p(p^{-1/2}(K^{3/2}p^{-1/2} + K^{1/2}T^{-1/2})) \cdot O_p(p^{1/2}) \\
&\quad + O_p(p^{-1/2}) \cdot O_p((pK/T)^{1/2}) \\
&= O_p(K^{3/2}/p^{1/2} + K/T^{1/2}). \quad \square
\end{aligned}$$

3.5.2 Proof of Theorem 3.2.1

Proof. We begin with the following lemma, which provides a compact representation of $\widehat{\boldsymbol{\Sigma}}_{f|y}$.

Lemma 3.5.1. *Suppose $\boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ is nonsingular and let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}$, as defined in Section 3.2. Then $\widehat{\boldsymbol{\Sigma}}_{f|y}$ can be represented as*

$$\widehat{\boldsymbol{\Sigma}}_{f|y} = \frac{1}{H}(\boldsymbol{\Phi}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{1/2} + \boldsymbol{\Delta}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1/2})(\boldsymbol{\Phi}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{1/2} + \boldsymbol{\Delta}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1/2})' + o_p(\omega_{p,T,K}).$$

Proof of Lemma 3.5.1. First note that by definition,

$$\begin{aligned}
H\widehat{\boldsymbol{\Sigma}}_{f|y} &= \widehat{\boldsymbol{\Xi}}\widehat{\boldsymbol{\Xi}}' = (\boldsymbol{\Xi} + \boldsymbol{\Delta})(\boldsymbol{\Xi} + \boldsymbol{\Delta})' \\
&= \boldsymbol{\Xi}\boldsymbol{\Xi}' + \boldsymbol{\Delta}\boldsymbol{\Xi}' + \boldsymbol{\Xi}\boldsymbol{\Delta}' + o_p(\omega_{p,T,K}).
\end{aligned}$$

By Proposition 3.1.1, we have $\boldsymbol{\xi}_h = E(\mathbf{f}_t|y_{t+1} \in I_h) = \boldsymbol{\Phi}E(\boldsymbol{\Phi}'\mathbf{f}_t|y_{t+1} \in I_h) = \boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\xi}_h$. It follows that $\boldsymbol{\Xi} = \boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\Xi} = \boldsymbol{\Phi}\boldsymbol{\Gamma}'$. Thus,

$$\begin{aligned}
H\widehat{\boldsymbol{\Sigma}}_{f|y} &= \boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\Phi}' + \boldsymbol{\Delta}\boldsymbol{\Gamma}\boldsymbol{\Phi}' + \boldsymbol{\Gamma}'\boldsymbol{\Delta}' + o_p(1/\sqrt{T}) \\
&= (\boldsymbol{\Phi}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{1/2} + \boldsymbol{\Delta}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1/2})(\boldsymbol{\Phi}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{1/2} + \boldsymbol{\Delta}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1/2})' + o_p(\omega_{p,T,K}),
\end{aligned}$$

which completes the proof. □

The next lemma is based on simple linear algebra results, but it is placed under our specific context.

Lemma 3.5.2. *Suppose that the leading L eigenvalues of $\widehat{\mathbf{\Sigma}}_{f|y}$ are positive and distinct. Let $\mathbf{\Psi}$ be a $K \times L$ matrix that satisfies the following two properties,*

(a) $\mathbf{\Psi}\mathbf{\Psi}'\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi}\mathbf{\Psi}' = \widehat{\mathbf{\Sigma}}_{f|y} + o_p(\omega_{p,T,K})$,

(b) $\mathbf{\Psi}'\mathbf{\Psi} = \mathbf{I} + o_p(\omega_{p,T,K})$.

Then $\mathbf{\Psi}$ consists of the eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$ corresponding to the L nonzero eigenvalues up to an order $o_p(\omega_{p,T,K})$.

Proof of Lemma 3.5.2. Part a) and b) imply that

$$\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi} = \mathbf{\Psi}\mathbf{\Psi}'\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi}\mathbf{\Psi}'\mathbf{\Psi} + o_p(\omega_{p,T,K}) = \mathbf{\Psi}\mathbf{\Psi}'\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi} + o_p(\omega_{p,T,K}).$$

Denoting $\mathbf{A} = \mathbf{\Psi}'\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi}$, which is a $L \times L$ full rank matrix, we have

$$\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi} = \mathbf{\Psi}\mathbf{A} + o_p(\omega_{p,T,K}).$$

We write the eigen-decomposition of $\mathbf{A}$ as $\mathbf{AU} = \mathbf{UD}$, where $\mathbf{D}$ is diagonal and $\mathbf{U}$ is orthonormal. Right multiplying the above equation by $\mathbf{U}$, we have $\widehat{\mathbf{\Sigma}}_{f|y}\mathbf{\Psi}\mathbf{U} = \mathbf{\Psi}\mathbf{U}\mathbf{D} + o_p(\omega_{p,T,K})$. As a result, $\mathbf{\Psi}\mathbf{U}$ is a matrix of L eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$ corresponding to the L nonzero eigenvalues. Since $\widehat{\mathbf{\Sigma}}_{f|y}$ is symmetric and its eigenvalues of are positive and distinct, the associated eigenvectors are independent. It follows that up to a sign change, $\mathbf{U}$ is a permutation matrix and the columns of $\mathbf{\Psi}$ are a rearrangement of the eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$. Without loss of generality. we assume that all nonzero elements in $\mathbf{U}$ are one. Therefore $\mathbf{\Psi}$ consists of the eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$ corresponding to the L nonzero eigenvalues up to an order $o_p(\omega_{p,T,K})$. □

We now establish the result of Theorem 3.2.1. We show that $\widehat{\mathbf{\Phi}}$ as defined in (3.2.2) satisfies the two properties in Lemma 3.5.2, and hence consists of the eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$. Lemma 3.5.1 b) is trivially satisfied by definition.

Note that

$$\begin{aligned}
&\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}' \\
= \;& [\mathbf{\Phi} + (\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}')\mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1} + o_p(\omega_{p,T,K})][\mathbf{\Phi} + (\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}')\mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1} + o_p(\omega_{p,T,K})]' \\
= \;& \mathbf{\Phi}\mathbf{\Phi}' + (\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}')\mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1}\mathbf{\Phi}' + \mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1}\mathbf{\Gamma}'\mathbf{\Delta}'(\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}') + o_p(\omega_{p,T,K}),
\end{aligned}$$

Using the fact that $\mathbf{\Phi}'\mathbf{\Phi} = \mathbf{I}$, we have

$$\begin{aligned}
&\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}'(\mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + \mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2}) \\
= \;& \mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + (\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}')\mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2} + \mathbf{\Phi}\mathbf{\Phi}'\mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2} + o_p(\omega_{p,T,K}) \\
= \;& \mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + \mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2} + o_p(\omega_{p,T,K}).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}'\widehat{\mathbf{\Sigma}}_{f|y}\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}' \\
= \;& \frac{1}{H}[\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}'(\mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + \mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2})][\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}'(\mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + \mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2})]' + o_p(\omega_{p,T,K}) \\
= \;& \frac{1}{H}(\mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + \mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2})(\mathbf{\Phi}(\mathbf{\Gamma}'\mathbf{\Gamma})^{1/2} + \mathbf{\Delta}\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1/2})' + o_p(\omega_{p,T,K}) \\
= \;& \widehat{\mathbf{\Sigma}}_{f|y} + o_p(\omega_{p,T,K}),
\end{aligned}$$

which verifies Lemma 3.5.2 a).

From Lemma 3.5.1 we know that $\|\mathbf{\Sigma}_{f|y} - \widehat{\mathbf{\Sigma}}_{f|y}\| = o_p(1)$. A direct application of Davis-Kahan $\sin(\theta)$ theorem implies that the first L eigenvalues of $\widehat{\mathbf{\Sigma}}_{f|y}$ are positive and distinct with probability approaching one. By Lemma 3.5.2, $\widehat{\mathbf{\Phi}}$ is consistent with the eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$ up to an order $o_p(\omega_{p,T,K})$. This completes the proof. □

3.5.3 Proof of Theorem 3.2.2

Proof. From Lemma 3.5.1 we know that $\|\mathbf{\Sigma}_{f|y} - \widehat{\mathbf{\Sigma}}_{f|y}\| = o_p(1)$. Hence, it remains to show that $\lambda_{L+1} = o_P(1)$ and $\|\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}' - \mathbf{\Phi}\mathbf{\Phi}'\|_F = o_P(1)$ without assuming the linearity condition.

Let $\mathbf{\Upsilon}$ be the linear subspace complement to $\mathbf{\Phi}$ such that $\mathbf{\Phi}\mathbf{\Phi}' + \mathbf{\Upsilon}\mathbf{\Upsilon}' = \mathbf{I}$.

From the law of total variance, we know that

$$\text{cov}(E(\mathbf{f}_t|y_{t+1})) = \text{cov}(\mathbf{f}_t) - E(\text{cov}(\mathbf{f}_t|y_{t+1})).$$

Next, we estimate $\text{cov}(\mathbf{f}_t) - E(\text{cov}(\mathbf{f}_t|y_{t+1}))$ to find the approximate central subspace, which is the equivalent expression of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$. Let $\mathbf{M}_t = \text{cov}(\mathbf{f}_t) - \text{cov}(\mathbf{f}_t|y_{t+1})$, and then $\text{cov}(E(\mathbf{f}_t|y_{t+1})) = E(\mathbf{M}_t)$. Note that

$$\begin{aligned}
\mathbf{M}_t &= \text{cov}(\mathbf{f}_t) - E(\mathbf{f}_t\mathbf{f}_t'|y_{t+1}) \\
&= [\text{cov}(\mathbf{f}_t) - E\{\text{cov}(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)|y_{t+1}\}] - E\{E(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)E'(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)|y_{t+1}\} \\
&= \mathbf{\Phi}\mathbf{\Phi}' - E\{E(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)E'(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)|y_{t+1}\}
\end{aligned}$$

where the constant conditional variance condition is used. To simplify notation, we define $\mathbf{\Delta}_t = -E\{E(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)E'(\mathbf{f}_t|\mathbf{\Phi}'\mathbf{f}_t)|y_{t+1}\}$. Next, by using the fact that $\mathbf{\Phi}\mathbf{\Phi}' + \mathbf{\Upsilon}\mathbf{\Upsilon}' = \mathbf{I}$, we have

$$\begin{aligned}
\mathbf{M}_t &= \mathbf{\Phi}\mathbf{\Phi}' + (\mathbf{\Phi}\mathbf{\Phi}' + \mathbf{\Upsilon}\mathbf{\Upsilon}')\mathbf{\Delta}_t \\
&= (\mathbf{\Phi}\mathbf{\Phi}' + \mathbf{\Phi}\mathbf{\Phi}'\mathbf{\Delta}_t\mathbf{\Phi}\mathbf{\Phi}') + (\mathbf{\Phi}\mathbf{\Phi}'\mathbf{\Delta}_t\mathbf{\Upsilon}\mathbf{\Upsilon}' + \mathbf{\Upsilon}\mathbf{\Upsilon}'\mathbf{\Delta}_t\mathbf{\Phi}\mathbf{\Phi}' + \mathbf{\Upsilon}\mathbf{\Upsilon}'\mathbf{\Delta}_t\mathbf{\Upsilon}\mathbf{\Upsilon}') \\
&\equiv \mathbf{N}_t + \mathbf{\Delta}_t,
\end{aligned}$$

It is obvious that the column space of $\mathbf{N}_t$ is included in the central subspace. Assumption 3.2.5 and Diaconis and Freedman (1984) naturally imply that $\|E(\mathbf{N}_t^2)\| = O_P(1)$. Assumption 3.2.5 and the result in Hall and Li (1993) (such as Theorem 3.2, Lemma 4.1 and Section 5) further show that $\|\Delta_t\| = o_P(1)$. Thus, we have

$$\|\mathbf{M}_t - E(\mathbf{N}_t)\| = o_P(1).$$

Let $\lambda_L(E(\mathbf{N}_t))$ the L-th largest eigenvalue of $E(\mathbf{N}_t)$. Note that $\mathbf{\Upsilon}'E(\mathbf{N}_t) = \mathbf{0}$ and $\lambda_L(E(\mathbf{N}_t))$ is non-diminishing with probability tending to one. Therefore, we immediately use Davis-Kahan $\sin(\theta)$ theorem to obtain that $\lambda_{L+1} = o_P(1)$ and $\|\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}' - \mathbf{\Phi}\mathbf{\Phi}'\|_F = o_P(1)$. The proof of Theorem 3.2.2 is complete. □

3.5.4 Proof of Theorem 3.3.1

Proof. Define the smoothing operator $S(\mathbf{u};\mathbf{v})$ as follows:

$$S(\mathbf{u};\mathbf{v}) = \hat{\alpha}, \text{ where } (\hat{\alpha},\hat{\boldsymbol{\beta}}) = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{T} \left(v_i - \alpha - \boldsymbol{\beta}'(\mathbf{u}-\mathbf{z})\right)^2 K_h(\mathbf{u}-\mathbf{z})$$

Intuitively the smoothing operator is additive for $\mathbf{v}$. The additivity can be easily verified based on the form of (3.3.3). Substitute equations (3.5.3) and (3.5.2) for y_{t+1} in local linear regression formula (3.3.2) and (3.3.5) respectively.

$$y_{t+1} = g(\boldsymbol{r}_t) + \epsilon_{t+1} = g(\mathbf{z}) + \epsilon_{t+1} + g(\boldsymbol{r}_t) - g(\mathbf{z}) - \nabla' g(\mathbf{z})(\boldsymbol{r}_t - \mathbf{z}) + \nabla' g(\mathbf{z})(\boldsymbol{r}_t - \mathbf{z}) \tag{3.5.2}$$

$$y_{t+1} = g(\mathbf{z}) + \epsilon_{t+1} + g(\boldsymbol{r}_t) - g(\mathbf{z}) - \nabla' g(\mathbf{z})(\boldsymbol{r}_t - \mathbf{z}) - \nabla' g(\mathbf{z})(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t) + \nabla' g(\mathbf{z})(\hat{\boldsymbol{r}}_t - \mathbf{z}) \tag{3.5.3}$$

Then by additivity $\hat{g}(\mathbf{z}) = S(\hat{\boldsymbol{r}}; \mathbf{y} = (y_1, \cdots, y_T)')$ can be written as

$$\hat{g}(\mathbf{z}) = g(\mathbf{z}) + \hat{g}_A(\mathbf{z}) + \hat{g}_B(\mathbf{z}) + \hat{g}_C(\mathbf{z}) + \hat{g}_D(\mathbf{z}) \tag{3.5.4}$$

where

$$\begin{aligned}
\hat{g}_A(\mathbf{z}) &= S(\hat{\boldsymbol{r}}; \boldsymbol{\varepsilon} = (\epsilon_1, \cdots, \epsilon_T)') \\
\hat{g}_B(\mathbf{z}) &= S(\hat{\boldsymbol{r}}; ((g(\boldsymbol{r}_t) - g(\mathbf{z}) - \nabla' g(\mathbf{z})(\boldsymbol{r}_t - \mathbf{z}))_{t=0,\cdots,T-1})') \\
\hat{g}_C(\mathbf{z}) &= S(\hat{\boldsymbol{r}}; (-\nabla' g(\mathbf{z})(\hat{\boldsymbol{r}}_t - \boldsymbol{r}_t)_{t=0,\cdots,T-1})') \\
\hat{g}_D(\mathbf{z}) &= S(\hat{\boldsymbol{r}}; (\nabla' g(\mathbf{z})(\hat{\boldsymbol{r}}_t - \mathbf{z})_{t=0,\cdots,T-1})')
\end{aligned}$$

Similarly, $\tilde{g}(\mathbf{z}) = S(\boldsymbol{r}; \mathbf{y} = (y_1, \cdots, y_T)')$ can be written as

$$\tilde{g}(\mathbf{z}) = g(\mathbf{z}) + \tilde{g}_A(\mathbf{z}) + \tilde{g}_B(\mathbf{z}) + \tilde{g}_C(\mathbf{z}) + \tilde{g}_D(\mathbf{z}) \tag{3.5.5}$$

where

$$\begin{aligned}
\tilde{g}_A(\mathbf{z}) &= S(\boldsymbol{r}; \boldsymbol{\varepsilon} = (\epsilon_1, \cdots, \epsilon_T)') \\
\tilde{g}_B(\mathbf{z}) &= S(\boldsymbol{r}; ((g(\boldsymbol{r}_t) - g(\mathbf{z}) - \nabla' g(\mathbf{z})(\boldsymbol{r}_t - \mathbf{z}))_{t=0,\cdots,T-1})') \\
\tilde{g}_D(\mathbf{z}) &= S(\boldsymbol{r}; (\nabla' g(\mathbf{z})(\boldsymbol{r}_t - \mathbf{z})_{t=0,\cdots,T-1})')
\end{aligned}$$

We also set $\tilde{g}_C(\mathbf{z}) = \nabla' g(\mathbf{z})\hat{\Delta}(\mathbf{z})$.

In what follows, we provide several technical lemmas to prove Theorem 3.3.1.

Lemma 3.5.3. *Under the same condition as Theorem 3.3.1*

$$\sup_{\mathbf{z}\in I_{\mathbf{z}},\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)}\in\bar{\mathcal{M}}_T} |\frac{1}{T}\mathbf{e}_i'\left(M_{r^{(1)}}W_{r^{(1)}} - M_{r^{(2)}}W_{r^{(2)}}\right)\boldsymbol{\varepsilon}| = O_p(T^{-\kappa_1}) \tag{3.5.6}$$

for $i = 1, \cdots, L+1$, *where* $\mathbf{e}_i \in \mathbb{R}^{L+1}$ *is the unit vector with i-th element being 1 while others being 0,* $\boldsymbol{\varepsilon} = (\epsilon_1, \cdots, \epsilon_T)'$, $\mathbf{w}_t(\boldsymbol{r}, \mathbf{z}) = \left(1, \left(\frac{(\boldsymbol{r}(\mathbf{f}_t)-\mathbf{z})}{h}\right)'\right)'$, $M_{\boldsymbol{r}} = (\mathbf{w}_0(\boldsymbol{r}, \mathbf{z}), \cdots, \mathbf{w}_{T-1}(\boldsymbol{r}, \mathbf{z}))'$. $W_{\boldsymbol{r}}$ *is diagonal matrix with* $W_{\boldsymbol{r}} = \mathrm{diag}(K_h(\boldsymbol{r}_0 - \mathbf{z}), \cdots, K_h(\boldsymbol{r}_{T-1} - \mathbf{z}))$.

Proof of Lemma 3.5.3. We just proof the lemma when $i = 1$. For $i = 2, \cdots, L$, the lemma follows by similar arguments. If $\kappa_1 \leq (\delta - \eta)$, we have

$$\begin{aligned}
&|\frac{1}{T}\mathbf{e}_1'\left(M_{r^{(1)}}W_{r^{(1)}} - M_{r^{(2)}}W_{r^{(2)}}\right)\boldsymbol{\varepsilon}| \\
=&|\frac{1}{T}\sum_{t=0}^{T-1}\left(K_h(\boldsymbol{r}_t^{(1)} - \mathbf{z}) - K_h(\boldsymbol{r}_t^{(2)} - \mathbf{z})\right)\epsilon_{t+1}| \\
=&O_p(T^{-1+L\eta-(\delta-\eta)}).
\end{aligned}$$

Then, the lemma holds naturally. In the following statements, we assume $\kappa_1 > (\delta - \eta)$. For C_1, C_2 large enough, there exists C_ϵ such that

$$P(\max_t |\epsilon_t| > C_\epsilon \log(T)) \leq T^{-C_1} \quad \textit{and} \quad |\mathbb{E}\epsilon_t\mathbb{I} \leq C_\epsilon \log(T)| \leq T^{-C_2}$$

Define

$$\Delta_t(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}) = \left(K_h(\boldsymbol{r}_t^{(1)} - \mathbf{z}) - K_h(\boldsymbol{r}_t^{(2)} - \mathbf{z})\right)\epsilon_t^*$$

with

$$\epsilon_t^* = \epsilon_t\mathbb{I}\{|\epsilon_t| \leq C_{\epsilon_t}\log(T)\} - \mathbb{E}\left(\epsilon_t\mathbb{I}\{|\epsilon_t| \leq C_{\epsilon_t}\log(T)\}\right)$$

For $s \geq 0$, let $\bar{\mathcal{M}}_{s,T,j}^*$ be a set of functions chosen such that for any $\boldsymbol{r} \in \bar{\mathcal{M}}_T$, there exists $\boldsymbol{r}^* \in \bar{\mathcal{M}}_{s,T,j}^*$ such that $||\boldsymbol{r} - \boldsymbol{r}^*||_\infty \leq 2^{-s}T^{-\delta}$. By Assumption 3.3.3, $\bar{\mathcal{M}}_{s,T,j}^*$ can be chosen with $\#\bar{\mathcal{M}}_{s,T,j}^* \leq C\exp((2^{-s}n^{-\delta})^{-\alpha_j}T^{\xi_j})$. Denote $\bar{\mathcal{M}}_{T,s}^* = \bar{\mathcal{M}}_{s,T,1}^* \times \cdots \times \bar{\mathcal{M}}_{s,T,L}^*$. For any $\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)} \in \bar{\mathcal{M}}_T$, choose $\boldsymbol{r}^{(1,s)}, \boldsymbol{r}^{(2,s)} \in \bar{\mathcal{M}}_{s,T}^*$.
By Chain arguments,

$$\Delta_t(\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)}) = \quad \Delta_t(\boldsymbol{r}^{(1,0)},\boldsymbol{r}^{(2,0)}) - \sum_{s=1}^{G_T}\Delta_t(\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)}) + \sum_{s=1}^{G_T}\Delta_t(\boldsymbol{r}^{(2,s-1)},\boldsymbol{r}^{(2,s)})$$
$$- \Delta_t(\boldsymbol{r}^{(1,G_T)},\boldsymbol{r}^{(1)}) + \Delta_t(\boldsymbol{r}^{(2,G_T)},\boldsymbol{r}^{(2)})$$

where G_T satisfies $G_T > ((1+c_G)(\kappa_1-(\delta-\eta))+L\eta)\log(T)/\log(2)$ for a constant $c_G > 0$. Hence

$$T_1 = |\frac{1}{T}\sum_{t=0}^{T-1}\Delta_t(\boldsymbol{r}^{(l,G_T)},\boldsymbol{r}^{(l)})| \leq C\log(T)2^{-G_T}T^{-(\delta-\eta)+L\eta} \leq CT^{-\kappa_1}$$

Now for any $a > c_G$, let $c_a = (\sum_{s=1}^{\infty}2^{-as})^{-1}$

$$\begin{aligned}
&P\left(\frac{1}{T}\sup_{\boldsymbol{r}\in\bar{\mathcal{M}}_T}|\sum_{t=0}^{T-1}\sum_{s=1}^{G_T}\Delta_t(\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)})| > T^{-\kappa_1}\right)\\
&\leq \sum_{s=1}^{G_T}P\left(\frac{1}{T}\sup_{\boldsymbol{r}\in\bar{\mathcal{M}}_T}|\sum_{t=0}^{T-1}\Delta_t(\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)})| > c_a2^{-as}T^{-\kappa_1}\right)\\
&\leq \sum_{s=1}^{G_T}\#\bar{\mathcal{M}}^*_{s-1,T}\#\bar{\mathcal{M}}^*_{s,T}P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\boldsymbol{r}^{(1,s)}_*,\boldsymbol{r}^{(1,s)}_{**}) > c_a2^{-as}T^{-\kappa_1}\right)\\
&+\sum_{s=1}^{G_T}\#\bar{\mathcal{M}}^*_{s-1,T}\#\bar{\mathcal{M}}^*_{s,T}P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\tilde{\boldsymbol{r}}^{(1,s)}_*,\tilde{\boldsymbol{r}}^{(1,s)}_{**}) < c_a2^{-as}T^{-\kappa_1}\right)\\
&= T_2+T_3
\end{aligned}$$

where $\boldsymbol{r}^{(1,s)}_*, \tilde{\boldsymbol{r}}^{(1,s)}_* \in \bar{\mathcal{M}}^*_{s-1,T}$, $\boldsymbol{r}^{(1,s)}_{**}, \tilde{\boldsymbol{r}}^{(1,s)}_{**} \in \bar{\mathcal{M}}^*_{s,T}$ are chosen such that

$$\begin{aligned}
&P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\boldsymbol{r}^{(1,s)}_*,\boldsymbol{r}^{(1,s)}_{**}) > c_a2^{-as}T^{-\kappa_1}\right)\\
&= \max_{\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)})}P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)})) > c_a2^{-as}T^{-\kappa_1}\right)
\end{aligned}$$

and

$$P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\tilde{\boldsymbol{r}}^{(1,s)}_*,\tilde{\boldsymbol{r}}^{(1,s)}_{**}) < c_a2^{-as}T^{-\kappa_1}\right)$$

$$= \max_{\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)})} P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\boldsymbol{r}^{(1,s-1)},\boldsymbol{r}^{(1,s)})) > c_a 2^{-as}T^{-\kappa_1}\right).$$

Note that for some $m < (\delta-\eta)+1-L\eta$,

$$\begin{aligned}
&\sup_t P\left(\frac{1}{T}\Delta_t(\boldsymbol{r}_*^{1,s},\boldsymbol{r}_{**}^{1,s})T^m > x\right)\\
\leq&\frac{\mathbb{E}(\exp(\frac{1}{T}\Delta_t(\boldsymbol{r}_*^{1,s},\boldsymbol{r}_{**}^{1,s})T^m))}{exp(x)}\\
\leq&\exp(c\frac{1}{T}T^{L\eta}T^{-(\delta-\eta)\log(T)T^m} - x)\\
\leq&\exp(1-x)
\end{aligned}$$

Hence $\gamma = \frac{1}{\frac{1}{\gamma_1}+1} < 1$. Then by the Bernstein's inequality [Theorem 1 of Merlevède et al. (2009, 2011)], for any small enough $\alpha > 0$

$$\begin{aligned}
&P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\boldsymbol{r}_*^{(1,s)},\boldsymbol{r}_{**}^{(1,s)}) > c_a 2^{-as}T^{-\kappa_1}\right)\\
&= P\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(\boldsymbol{r}_*^{(1,s)},\boldsymbol{r}_{**}^{(1,s)})T^m > c_a 2^{-as}T^{-\kappa_1}T^m\right)\\
&= T\exp\left(-\frac{(c_a 2^{-as}T^{-\kappa_1}T^m)^\gamma}{C_1}\right) + \exp\left(-\frac{(c_a 2^{-as}T^{-\kappa_1}T^m)^2}{C_2(1+TV)}\right)\\
&+ \exp\left(-\frac{(c_a 2^{-as}T^{-\kappa_1}T^m)^2}{C_3 T}\exp\left(\frac{(c_a 2^{-as}T^{-\kappa_1}T^m)^{\gamma(1-\gamma)}}{C_4(\log(c_a 2^{-as}T^{-\kappa_1}T^m)^\gamma}\right)\right)\\
&\leq CT\exp(-CT^{-\gamma\kappa_1+\gamma m}) \leq C\exp(T^\alpha - CT^{-\gamma\kappa_1+\gamma m}).
\end{aligned}$$

Then

$$\begin{aligned}
T_2 &\leq C\sum_{s=1}^{G_T}\prod_j \exp\left((2^{-s}n^{-\delta})^{-\alpha_j}T^{\xi_j}\right) Pr\left(\sum_{t=0}^{T-1}\frac{1}{T}\Delta_t(r_1^{*,s},r_1^{**,s}) > c_a 2^{-as}T^{-\kappa_1}\right)\\
&\leq C\exp\left(CT^{\max(\delta\alpha_j+\xi_j)} + T^\alpha - CT^{-\gamma\kappa_1+\gamma m}\right)\\
&\leq \exp(-cT^c)
\end{aligned}$$

Similarly, $T_3 \leq \exp(-cT^c)$

Also,

$$T_4 = P\left(\sup_{\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)}\in\bar{\mathcal{M}}_T} |\frac{1}{T}\sum_{t=0}^{T-1}\Delta_t(\boldsymbol{r}^{(1,0)},\boldsymbol{r}^{(2,0)}| > T^{-\kappa_1}\right) \leq \exp(-cT^c)$$

Combining T_1, T_2, T_3 and T_4,

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} P\left(\sup_{\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)}\in\bar{\mathcal{M}}_T} |\frac{1}{T}\sum_{t=0}^{T-1}K_h(\boldsymbol{r}_t^{(1)}-\mathbf{z})\epsilon_{t+1}^* - \frac{1}{T}\sum_{t=0}^{T-1}K_h(\boldsymbol{r}_t^{(2)}-\mathbf{z})\epsilon_{t+1}^*| > CT^{-\kappa_1}\right)$$
$$\leq \exp(-cT^c).$$

Furthermore,

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}}\sup_{\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)}\in\bar{\mathcal{M}}_T} |\frac{1}{T}\sum_{t=0}^{T-1}K_h(\boldsymbol{r}_t^{(1)}-\mathbf{z})\epsilon_{t+1} - \frac{1}{T}\sum_{t=0}^{T-1}K_h(\boldsymbol{r}_t^{(2)}-\mathbf{z})\epsilon_{t+1}| < CT^{-\kappa_1}$$

The lemma then follows. □

Lemma 3.5.4. *Under the same condition as Theorem 3.3.1*

$$\sup_{\mathbf{z}\in I_{\mathbf{z}},\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)}\in\bar{\mathcal{M}}_T} |\frac{1}{T}\left(M_{\boldsymbol{r}^{(1)}}^{a'}W_{\boldsymbol{r}^{(1)}}M_{\boldsymbol{r}^{(1)}}^{b} - M_{\boldsymbol{r}^{(2)}}^{a'}W_{\boldsymbol{r}^{(2)}}M_{\boldsymbol{r}^{(2)}}^{b}\right)_{j,l}| = O_p(T^{-(\delta-\eta))}) \tag{3.5.7}$$

for $j,l = 1,\cdots,L, j\neq l$ *and* $0\leq a+b\leq 2$, $a,b\geq 0$, *with* $M_{\boldsymbol{r}}^a = \left(\left(\frac{\boldsymbol{r}_0-\mathbf{z}}{h}\right)^a,\cdots,\left(\frac{\boldsymbol{r}_{T-1}-\mathbf{z}}{h}\right)^a\right)'$.

Proof of Lemma 3.5.4. Note that

$$\sup_{\mathbf{z}}|K_h(\boldsymbol{r}_1-\mathbf{z}) - K_h(\boldsymbol{r}_2-\mathbf{z})| = O_p(T^{-(\delta-\eta)+L\eta})$$

$$\sup_{\mathbf{z}\in I_{\mathbf{z}},\boldsymbol{r}\in\bar{\mathcal{M}}}|\frac{1}{T}K_h(\boldsymbol{r}-\mathbf{z})| \leq CT^{-1+L\eta}\sup_{\mathbf{z}\in I_{\mathbf{z}}}\#\{i:|r_{0,j}-z_j|\leq CT^{-\eta}\} = O_p(1)$$

The lemma holds after a simple calculation. □

Lemma 3.5.5. *Under the same condition as Theorem 3.3.1, then*

$$\sup_{\mathbf{z}\in I_{\mathbf{z}},0\leq t\leq T-1}|\left(g(\boldsymbol{r}_t)-g(\mathbf{z})-\nabla' g(\mathbf{z})(\boldsymbol{r}_t-\mathbf{z})\right)\mathbb{I}\{||(\boldsymbol{r}_t-\mathbf{z})/h||_1\leq 1\}| = O_p(T^{-2\eta}) \tag{3.5.8}$$

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} || \frac{1}{T} \left(\hat{M}' W_{\hat{\boldsymbol{r}}} \hat{M} - M' W_{\boldsymbol{r}} M \right) || = O_p(T^{-(\delta-\eta)}) \tag{3.5.9}$$

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} || \frac{1}{T} M' W_{\boldsymbol{r}} M - f_{\mathbf{z}}(\mathbf{z}) B_{\mathcal{K}} || = O_p(T^{-\eta} + T^{-(1-L\eta)/2} \sqrt{\log(T)}) \tag{3.5.10}$$

where $B_{\mathcal{K}} = \text{diag}(1, \int u^2 \mathcal{K}(u) du, \cdots, \int u^2 \mathcal{K}(u) du)$ *is an* $(L+1)\times(L+1)$ *matrix.*

Proof of Lemma 3.5.5. (3.5.8) follows from a simple calculation. (3.5.9) is a special case of Lemma 3.5.4. (3.5.10) is a standard result from Kernel smoothing regression. □

Now we are ready to complete the proof of Theorem 3.3.1. By Lemma 3.5.3 and 3.5.4,

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\hat{g}_A(\mathbf{z}) - \tilde{g}_A(\mathbf{z})| = O_p(T^{-\kappa_1}) \tag{3.5.11}$$

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\hat{g}_B(\mathbf{z}) - \tilde{g}_B(\mathbf{z})| = O_p(T^{-\kappa_2}) \tag{3.5.12}$$

By Lemma 3.5.4 and 3.5.5,

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\hat{g}_C(\mathbf{z}) - \tilde{g}_C(\mathbf{z})| = O_p(T^{-\kappa_3}) \tag{3.5.13}$$

It is easy to conclude $\hat{g}_D(\mathbf{z}) = \tilde{g}_D(\mathbf{z}) = 0$ from the construction. Theorem 3.3.1 then follows. □

3.5.5 Proof of Theorem 3.3.2

Proof. By using Theorem 6 of Masry (1996), we show that under the assumptions 3.2.2 and 3.3.1-3.3.3, $\tilde{g}(\mathbf{z})$ has the property that

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\tilde{g}(\mathbf{z}) - g(\mathbf{z})| = O_p(\sqrt{\log(T) T^{-1} h^{-L}} + h^2) \tag{3.5.14}$$

Most conditions in this theorem are trivially satisfied. In addition, Assumption 3.3.1(iv) and Assumption 3.2.2(ii) will ensure the properties that the bandwidth goes to zero slowly enough and the strong mixing coefficient fulfills the summability condition. Now directly combine equation (3.5.14) with results from Theorem 3.3.1, Theorem 3.3.2 then follows. □

Chapter 4

Conclusion and Future Work

This paper revisits the sufficient forecasting and studies its nonparametric estimation and predictive inference in a large panel data settings. We point out the close connection between the sufficient forecasting and existing methods (such as Fama-Macbeth regression and partial least squares) in their way of utilizing target information. The estimated forecasting directions are shown to be consistent under a diverging number of latent factors, and a general decomposition of the estimates is obtained. We also show that the estimated predicted indices can be directly used to estimate the forecasting function, which is often a non-trivial issue in factor analysis. In the empirical example of predicting market returns, we demonstrate that allowing non-linearity in forecasting functions can yield additional gains.

There are a number of extensions that can be made on top of this work. For example, in many forecasting models, lagged variables are sometimes used as predictors in the forecasting model. Without linearity assumption, how to utilize lagged variables in sufficient forecasting would become an intriguing question.

Bibliography

Bai, J. and Li, K. (2012), 'Statistical methods of factor models of high dimension', *The Annals of Statistics* **40**(1), 436–465.

Bai, J. and Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bai, J. and Ng, S. (2008), 'Forecasting economic time series using targeted predictors', *Journal of Econometrics* **146**(2), 304–317.

Bai, J. and Ng, S. (2013), 'Principal components estimation and identification of the factors', *Journal of Econometrics* **176**, 18–29.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), 'Prediction by supervised principal components', *Journal of the American Statistical Association* **101**(483), 119–137.

Billingsley, P. (1999), *Convergence of Probability Measures*, 2nd edn, John Wiley & Sons.

Cochrane, J. H. (2001), *Asset Pricing*, Princeton University Press.

Cook, R. D. (2009), 'Regression Graphics: Ideas for Studying Regressions through Graphics', Vol.482 , John Wiley & Sons.

Diaconis, P. and Freedman, D. (1984), 'Asymptotics of graphical projection pursuit', *The Annals of Statistics* **12**(3), 793–815.

Eaton, M. L. (1983), *Multivariate Statistics: a Vector Space Approach*, Wiley, New York.

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* **33**, 3–56.

Fama, E. F. and MacBeth, J. D. (1973), 'Risk, return, and equilibrium: Empirical tests', *Journal of Political Economy* **81**(3), 607–636.

Fan, J., Fan, Y. and Lv, J. (2008), 'High dimensional covariance matrix estimation using a factor model', *Journal of Econometrics* **147**(1), 186–197.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling And Its Applications*, CRC Press.

Fan, J., Liao, Y. and Mincheva, M. (2013), 'Large covariance estimation by thresholding principal orthogonal complements (with discussion)', *Journal of the Royal Statistical Society: Series B* **75**(4), 603–680.

Fan, J., Liao, Y. and Wang, W. (2016), 'Projected principal component analysis in factor models', *The Annals of Statistics* **44**(1), 219254

Fan, J., Xue, L. and Yao, J. (2017), 'Sufficient forecasting using factor models', *Journal of Econometrics* **201**(2), 292–306.

Frank, L. E. and Friedman, J. H. (1993), 'A statistical view of some chemometrics regression tools', *Technometrics* **35**(2), 109–135.

Hall, P. and Li, K.-C. (1993), 'On almost linearity of low dimensional projections from high dimensional data', *The Annals of Statistics* **21**(2), 867–889.

Heckman, J. J. and Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**(3), 669–738.

Imbens, G. W. and Newey, W. K. (2009), 'Identification and estimation of triangular simultaneous equations models without additivity', *Econometrica* **77**(5), 1481–1512.

Jagannathan, R. and Wang, Z. (1996), 'The conditional capm and the cross-sectionof expected returns', *Journal of Finance* **51**(1), 3–53.

Kelly, B. and Pruitt, S. (2015), 'The three-pass regression filter: a new approach to forecasting using many predictors', *Journal of Econometrics* **186**(2), 294–316.

Lettau, M. and Ludvigson, S. (2001), 'Consumption, aggregate wealth, and expected stock returns', *Journal of Finance* **56**(3), 815–849.

Li, K.-C. (1991), 'Sliced inverse regression for dimension reduction (with discussion)', *Journal of the American Statistical Association* **86**(414), 316–327.

Li, H., Li, Q. and Shi, Y. (2017), 'Determining the number of factors when the number of factors can increase with sample size', *Journal of Econometrics* **197**(1), 76–86.

Li, Q., Vassalou, M. and Xing, Y. (2006), 'Sector investment growth rates and the cross section of equity returns', *The Journal of Business* **79**(3), 1637–1665.

Ludvigson, S. C. and Ng, S. (2007), 'The empirical risk–return relation: A factor analysis approach', *Journal of Financial Economics* **83**(1), 171–222.

Luo, W., Xue, L. and Yao, J. (2017), 'Inverse moment methods for sufficient forecasting using high-dimensional predictors', *arXiv preprint arXiv:1705.00395* .

Mammen, E., Rothe, C. and Schienle, M. (2012), 'Nonparametric regression with nonparametrically generated covariates', *The Annals of Statistics* **40**(2), 1132–1170.

Masry, E. (1996), 'Multivariate local polynomial regression for time series: uniform strong consistency and rates', *Journal of Time Series Analysis* **17**(6), 571–599.

Merlevède, F., Peligrad, M. and Rio, E. (2011), 'A bernstein type inequality and moderate deviations for weakly dependent sequences', *Probability Theory and Related Fields* **151**(3), 435–474.

Merlevède, F., Peligrad, M., Rio, E. et al. (2009), Bernstein inequality and moderate deviations under strong mixing conditions, *in* 'High Dimensional Probability V: the Luminy volume', Institute of Mathematical Statistics, pp. 273–292.

Newey, W. K., Powell, J. L. and Vella, F. (1999), 'Nonparametric estimation of triangular simultaneous equations models', *Econometrica* **67**(3), 565–603.

Stock, J. H. and Watson, M. W. (1989), 'New indexes of coincident and leading economic indicators', *NBER Macroeconomics Annual* **4**, 351–409.

Stock, J. H. and Watson, M. W. (2002*a*), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.

Stock, J. H. and Watson, M. W. (2002*b*), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business & Economic Statistics* **20**(2), 147–162.

Xia, Y., Tong, H., Li, W. and Zhu, L.-X. (2002), 'An adaptive estimation of dimension reduction space (with discussion)', *Journal of the Royal Statistical Society Series B* **64**(3), 363–410.