Sufficient Forecasting Using Factor Models[*]

Jianqing Fan[†], Lingzhou Xue[‡] and Jiawei Yao[†]

[†]Princeton University and [‡]Pennsylvania State University

Abstract

We consider forecasting a single time series when there is a large number of predictors and a possible nonlinear effect. The dimensionality was first reduced via a high-dimensional factor model implemented by the principal component analysis. Using the extracted factors, we develop a link-free forecasting method, called the sufficient forecasting, which provides several sufficient predictive indices, inferred from high-dimensional predictors, to deliver additional predictive power. Our method is also applicable to cross-sectional sufficient regression using extracted factors. The connection between the sufficient forecasting and the deep learning architecture is explicitly stated. The sufficient forecasting correctly estimates projection indices of the underlying factors even in the presence of a nonparametric forecasting function. The proposed method extends the sufficient dimension reduction to high-dimensional regimes by condensing the cross-sectional information through factor models. We derive asymptotic properties for the estimate of the central subspace spanned by these projection directions as well as the estimates of the sufficient predictive indices. We also show that the natural method of running multiple regression of target on estimated factors yields a linear estimate that actually falls into this central subspace. Our method and theory allow the number of predictors to be larger than the number of observations. We finally demonstrate that the sufficient forecasting improves upon the linear forecasting in both simulation studies and an empirical study of forecasting macroeconomic variables.

Key Words: Regression; forecast; deep learning; approximate factor model; principal components; learning indices; sliced inverse regression; dimension reduction.

1 Introduction

Regression and forecasting in a data-rich environment have been an important research topic in statistics, economics and finance. Typical examples include forecasts of a macroeconomic output using a large number of employment and production variables (Stock and Watson, 1989; Bernanke et al., 2005), forecasts of the values of market prices and dividends using cross-sectional asset returns (Sharpe, 1964; Lintner, 1965), and studies of association between clinical outcomes and high throughput genomics and genetic information such as microarray gene expressions. The predominant framework to harness the vast predictive information is via the *factor model*, which proves effective in simultaneously modeling the commonality and cross-sectional dependence of the observed data. It is natural to think that the latent factors drives simultaneously the vast covariate information as well as the outcome. This achieves the dimensionality reduction in the regression and predictive models.

Turning the curse of dimensionality into blessing, the latent factors can be accurately extracted from the vast predictive variables and hence they can be reliably used to build models for response variables. For the same reason, factor models have been widely employed in many applications, such as portfolio management (Fama and French, 1992; Carhart, 1997; Fama and French, 2015), large-scale multiple testing (Leek and Storey, 2008; Fan et al., 2012), high-dimensional covariance matrix estimation (Fan et al., 2008, 2013), and forecasting using many predictors (Stock and Watson, 2002a,b). Leveraging on the recent developments on the factor models, in this paper, we are particularly interested in the problem of forecasting. Our techniques can also be applied to regression problems, resulting in sufficient regression.

With little knowledge of the relationship between the forecast target and the latent factors, most research focuses on a linear model and its refinement. Motivated by the classic principal component regression (Kendall, 1957; Hotelling, 1957), Stock and Watson (2002a,b) employed a similar idea to forecast a single time series from a large number of predictors: first use the principal component analysis (PCA) to estimate the underlying common factors, followed by a linear regression of the target on the estimated factors. The key insight here is to condense information from many cross-sectional predictors into several predictive indices. Boivin and Ng (2006) pointed out that the relevance of factors is important to the forecasting power, and may lead to improved forecast. As an improvement to this procedure, Bair et al. (2006) used correlation screening to remove irrelevant predictors before performing the PCA. In a similar fashion, Bai and Ng (2008) applied boosting and employed thresholding rules to select "targeted predictors", and Stock and Watson (2012) used shrinkage methods to downweight unrelated principal components. Kelly and Pruitt (2015) took into account the covariance with the forecast target in linear forecasting, and they proposed a three-pass

2

regression filter that generalizes partial least squares to forecast a single time series.

As mentioned, most of these approaches are fundamentally limited to linear forecasting. This yields only one index of extracted factors for forecasting. It does not provide sufficient predictive power when the predicting function is nonlinear and depends on multiple indices of extracted factors. Yet, as pointed out by Tjøstheim (1994), nonlinear models usually outperform linear models in time series analysis, especially in multi-step prediction. When the link function between the target and the factors is arbitrary and unknown, a thorough exploration of the factor space often leads to additional gains. To the best of our knowledge, there are only a few papers in forecasting a nonlinear time series using factor models. Bai and Ng (2008) discussed the use of squared factors (i.e., volatility of the factors) in augmenting forecasting equation. Ludvigson and Ng (2007) found that the square of the first factor estimated from a set of financial factors is significant in the regression model for mean excess returns. This naturally leads to the question of which effective factor, or more precisely, which effective direction of factor space to include for higher moments. Cunha et al. (2010) used a class of nonlinear factor models with shape constraints to study cognitive and noncognitive skill formation. The nonparametric forecasting function would pose a significant challenge in estimating effective predictive indices.

In this work, we shall address this issue from a completely different perspective to existing forecasting methods. We introduce a favorable alternative method called the *sufficient forecasting*. Our proposed forecasting method springs from the idea of the sliced inverse regression, which was first introduced in the seminal work by Li (1991). In the presence of some arbitrary and unknown (possibly nonlinear) forecasting function, we are interested in estimating a set of *sufficient predictive indices*, given which the forecast target is independent of unobserved common factors. To put it another way, the forecast target relates to the unobserved common factors only through these sufficient predictive indices. Such a goal is closely related to the estimation of the *central subspace* in the dimension reduction literature (Cook, 2009). In a linear forecasting model, such a central subspace consists of only one dimension. In contrast, when a nonlinear forecasting function is present, the central subspace can go beyond one dimension, and our proposed method can effectively reveal multiple sufficient predictive indices to enhance the prediction power. This procedure therefore greatly enlarges the scope of forecasting using factor models. As demonstrated in numerical studies, the sufficient forecasting has improved performance over benchmark methods, especially under a nonlinear forecasting equation.

In summary, the contribution of this work is at least twofold. On the one hand, our proposed sufficient forecasting advances existing forecasting methods, and fills the important gap between incorporating target information and dealing with nonlinear forecasting.

3

Our work identifies effective factors that influence the forecast target without knowing the nonlinear dependence. On the other hand, we present a promising dimension reduction technique through factor models. It is well-known that existing dimension reduction methods are limited to either a fixed dimension or a diverging dimension that is smaller than the sample size (Zhu et al., 2006). With the aid of factor models, our work alleviates what plagues sufficient dimension reduction in high-dimensional regimes by condensing the cross-sectional information, whose dimension can be much higher than the sample size.

The rest of this paper is organized as follows. Section 2 presents the complete methodological details of the sufficient forecasting using factor models. In Section 3, we establish the asymptotic properties for the sufficient forecasting, and we also prove that the simple linear estimate actually falls into the central subspace. Section 4 demonstrates the numerical performance of the sufficient forecasting in simulation studies and an empirical application. Section 5 includes a few concluding remarks. Technical proofs are rendered in the Appendix.

2   Sufficient forecasting

This section first presents a unified framework for forecasting using factor models. We then propose the sufficient forecasting procedure with unobserved factors, which estimates predictive indices without requiring estimating the unknown forecasting function.

2.1   Factor models and forecasting

Consider the following factor model with a target variable y_{t+1} which we wish to forecast:

$$y_{t+1} = h(\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t, \epsilon_{t+1}), \tag{2.1}$$

$$x_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \quad 1 \le i \le p, \ 1 \le t \le T, \tag{2.2}$$

where x_{it} is the i-th predictor observed at time t, $\mathbf{b}_i$ is a $K \times 1$ vector of factor loadings, $\mathbf{f}_t = (f_{1t}, \cdots, f_{Kt})'$ is a $K \times 1$ vector of common factors driving both the predictors and the response, and u_{it} is the error term, or the idiosyncratic component. For ease of notation, we write $\mathbf{x}_t = (x_{1t}, \cdots, x_{pt})'$, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)'$ and $\mathbf{u}_t = (u_{1t}, \cdots, u_{pt})'$. In (2.1), $h(\cdot)$ is an unknown link function, and ϵ_{t+1} is some stochastic error independent of $\mathbf{f}_t$ and u_{it}. The linear combinations $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ are K-dimensional orthonormal vectors. Clearly, the model is also applicable to the cross-sectional regression such as those mentioned at the introduction of the paper. Note that the sufficient forecasting can be represented in a deep learning architecture (Bengio, 2009; Bengio et al., 2013), consisting of four layers of nonlinear processes

for dimension reduction and forecasting. The connection between sufficient forecasting and deep learning is illustrated in Figure 1. An advantage of our deep learning model is that it admits scalable and explicit computational algorithm.
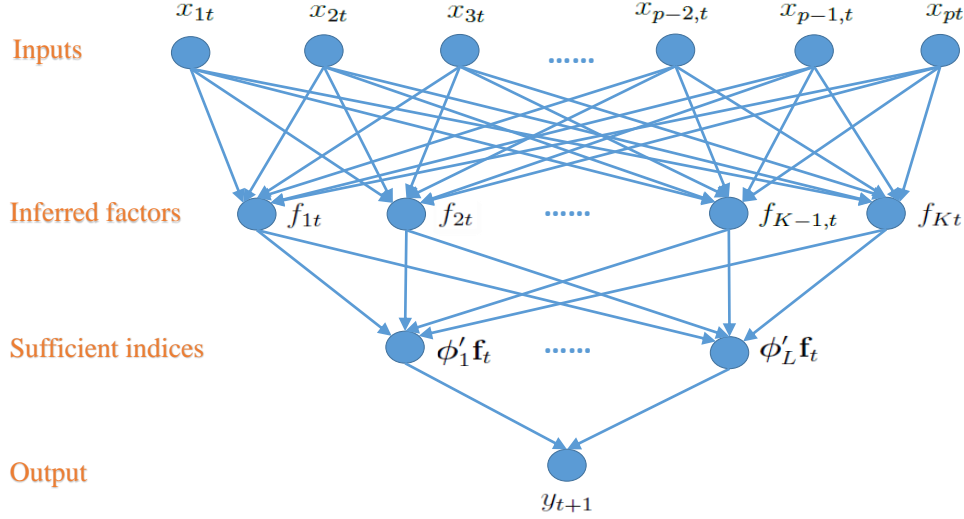

Figure 1: The sufficient forecasting is represented as a four-layer deep learning architecture. The factors $f_{1t}, \cdots, f_{Kt}$ are obtained from the inputs via principal component analysis.

Model (2.1) is a semiparametric multi-index model with latent covariates and unknown nonparametric function $h(\cdot)$. The target y_{t+1} depends on the factors $\mathbf{f}_t$ only through L predictive indices $\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t$, where L is no greater than K. In other words, these predictive indices are *sufficient* in forecasting y_{t+1}. Here, $\boldsymbol{\phi}_i$'s are also called the sufficient dimension reduction (SDR) directions, and their linear span forms the *central subspace* (Cook, 2009) denoted by $S_{y|\mathbf{f}}$. Note that the individual directions $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ are not identifiable without imposing any structural condition on $h(\cdot)$. However, the subspace $S_{y|\mathbf{f}}$ spanned by $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ can be identified. Therefore, throughout this paper, we refer any orthonormal basis $\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_L$ of the central subspace $S_{y|\mathbf{f}}$ as sufficient dimension reduction directions, and their corresponding predictive indices $\boldsymbol{\psi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\psi}_L'\mathbf{f}_t$ as sufficient predictive indices.

Suppose that the factor model (2.2) has the following canonical normalization

$$\text{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\mathbf{B} \text{ is diagonal,} \tag{2.3}$$

where $\mathbf{I}_K$ is a $K \times K$ identity matrix. This normalization serves as an identification condition of $\mathbf{B}$ and $\mathbf{f}_t$ because $\mathbf{B}\mathbf{f}_t = (\mathbf{B}\boldsymbol{\Omega})(\boldsymbol{\Omega}^{-1}\mathbf{f}_t)$ holds for any nonsingular matrix $\boldsymbol{\Omega}$. We assume for simplicity that the means of x_{it}'s and $\mathbf{f}_t$'s in (2.2) have already been removed, namely,

$E(x_{it}) = 0$ and $E(\mathbf{f}_t) = 0$.

As a special case, when we have *a priori* knowledge that the link function $h(\cdot)$ in (2.1) is in fact linear. In this case, no matter what L is, there is only a single index of the factors $\mathbf{f}_t$ in the prediction equation, i.e. $L = 1$. Now, (2.1) reduces to

$$y_{t+1} = \boldsymbol{\phi}_1' \mathbf{f}_t + \epsilon_{t+1}.$$

Such linear forecasting problems using many predictors have been studied extensively in the econometrics literature, for example, Stock and Watson (2002*a*,*b*), Bai and Ng (2008), and Kelly and Pruitt (2015), among others. Existing methods are mainly based on the principal component regression (PCR), which is a multiple linear regression of the forecast target on the estimated principal components. The linear forecasting framework does not thoroughly explore the predictive power embedded in the underlying common factors. It can only create a single index of the unobserved factors to forecast the target. The presence of a nonlinear forecasting function with multiple forecasting indices would only deteriorate the situation. While extensive efforts have been made such as Boivin and Ng (2006), Bai and Ng (2009), and Stock and Watson (2012), these approaches are more tailored to handle issues of using too many factors, and they are fundamentally limited to the linear forecasting.

2.2 Sufficient forecasting

Traditional analysis of factor models typically focuses on the covariance with the forecast target $\text{cov}(\mathbf{x}_t, y_{t+1})$ and the covariance within the predictors $\mathbf{x}_t$, denoted by a $p \times p$ matrix

$$\boldsymbol{\Sigma}_x = \mathbf{B}\text{cov}(\mathbf{f}_t)\mathbf{B}' + \boldsymbol{\Sigma}_u, \tag{2.4}$$

where $\boldsymbol{\Sigma}_u$ is the error covariance matrix of $\mathbf{u}_t$. Usually $\text{cov}(\mathbf{x}_t, y_{t+1})$ and $\boldsymbol{\Sigma}_x$ are not enough to construct an optimal forecast, especially in the presence of a nonlinear forecasting function. To fully utilize the target information, our method considers the covariance matrix of the inverse regression curve, $E(\mathbf{x}_t|y_{t+1})$. By conditioning on the target y_{t+1} in model (2.2), we obtain

$$\text{cov}(E(\mathbf{x}_t|y_{t+1})) = \mathbf{B}\text{cov}(E(\mathbf{f}_t|y_{t+1}))\mathbf{B}',$$

where we used the assumption that $E(\mathbf{u}_t|y_{t+1}) = 0$. This salient feature does not impose any structure on the covariance matrix of $\mathbf{u}_t$. Under model (2.1), Li (1991) showed that $E(\mathbf{f}_t|y_{t+1})$ is contained in the central subspace $S_{y|\mathbf{f}}$ spanned by $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L$ provided 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 $\mathbf{b} \in \mathbb{R}^K$. This important result implies that $S_{y|\mathbf{f}}$ contains the linear span of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$. Thus, it is promising to

estimate sufficient directions by investigating the top L eigenvectors of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$.

However, the factors $\mathbf{f}_t$ are *unobserved* in practice. It is difficult to directly estimate the covariance of the inverse regression curve $E(\mathbf{f}_t|y_{t+1})$ and obtain sufficient predictive indices. Instead of using the conventional nonparametric regression method, we follow Li (1991) to explore the conditional information of these underlying factors. To this end, we introduce the sliced covariance estimate

$$\boldsymbol{\Sigma}_{f|y} = \frac{1}{H} \sum_{h=1}^{H} E(\mathbf{f}_t|y_{t+1} \in I_h) E(\mathbf{f}'_t|y_{t+1} \in I_h), \tag{2.5}$$

where the range of y_{t+1} is divided into H slices $I_1, \ldots, I_H$ such that $P(y_{t+1} \in I_h) = 1/H$. Remark that $H \geq \max\{L, 2\}$ is required in $\boldsymbol{\Sigma}_{f|y}$ to estimate the sufficient forecasting directions. Note that

$$E(\mathbf{f}_t|y_{t+1} \in I_h) = E(E(\mathbf{f}_t|y_{t+1})|y_{t+1} \in I_h),$$

and that $E(\mathbf{f}_t|y_{t+1})$ is confined in the central subspace $S_{y|\mathbf{f}}$ spanned by directions $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L$. One observes that for any vector $\boldsymbol{\phi}_\perp$ orthogonal to $S_{y|\mathbf{f}}$, $\boldsymbol{\phi}'_\perp E(\mathbf{f}_t|y_{t+1}) = 0$ holds and hence $\boldsymbol{\phi}'_\perp \boldsymbol{\Sigma}_{f|y} = 0$ also holds. As a result, the eigenvectors of $\boldsymbol{\Sigma}_{f|y}$ corresponding to its L largest positive eigenvalues would form an orthogonal basis for $S_{y|\mathbf{f}}$, resulting in sufficient directions.

To estimate $\boldsymbol{\Sigma}_{f|y}$ in (2.5) with unobserved factors, a natural solution is to first consistently estimate factors $\mathbf{f}_t$ from the factor model (2.2) and then use the estimated factors $\widehat{\mathbf{f}}_t$ and the observed target y_{t+1} to approximate the sliced estimate $\boldsymbol{\Sigma}_{f|y}$. Alternatively, we can start with the observed predictors $\mathbf{x}_t$. Since $E(\mathbf{u}_t|y_{t+1}) = 0$, we then have

$$E(\mathbf{x}_t|y_{t+1} \in I_h) = \mathbf{B} E(\mathbf{f}_t|y_{t+1} \in I_h),$$

and

$$E(\mathbf{f}_t|y_{t+1} \in I_h) = \boldsymbol{\Lambda}_b E(\mathbf{x}_t|y_{t+1} \in I_h),$$

where $\boldsymbol{\Lambda}_b = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. Hence,

$$\boldsymbol{\Sigma}_{f|y} = \frac{1}{H} \sum_{h=1}^{H} \boldsymbol{\Lambda}_b E(\mathbf{x}_t|y_{t+1} \in I_h) E(\mathbf{x}'_t|y_{t+1} \in I_h) \boldsymbol{\Lambda}'_b, \tag{2.6}$$

Notice that $\boldsymbol{\Lambda}_b$ is also *unknown* and needs to be estimated. With a consistent estimator $\widehat{\boldsymbol{\Lambda}}_b$, we can use $\mathbf{x}_t$ and y_{t+1} to approximate the sliced estimate $\boldsymbol{\Sigma}_{f|y}$.

In the sequel, we present two estimators, $\widehat{\boldsymbol{\Sigma}}^1_{f|y}$ in (2.9) and $\widehat{\boldsymbol{\Sigma}}^2_{f|y}$ in (2.10), estimated by using respectively (2.5) and (2.6). Interestingly, we show that these two estimators are

exactly equivalent when factors and factor loadings are estimated from the constrained least squares. We denote two equivalent estimators $\widehat{\boldsymbol{\Sigma}}_{f|y}^1$ and $\widehat{\boldsymbol{\Sigma}}_{f|y}^2$ by $\widehat{\boldsymbol{\Sigma}}_{f|y}$. As in Remark 2.1, these two are in general not equivalent.

In Section 3, we shall show that under mild conditions, $\boldsymbol{\Sigma}_{f|y}$ is consistently estimated by $\widehat{\boldsymbol{\Sigma}}_{f|y}$ as $p, T \to \infty$. Furthermore, the eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{f|y}$ corresponding to the L largest eigenvalues, denoted as $\widehat{\boldsymbol{\psi}}_j$ ($j = 1, \cdots, L$), will converge to the corresponding eigenvectors of $\boldsymbol{\Sigma}_{f|y}$, which actually span the aforementioned central subspace $S_{y|\mathbf{f}}$. This will yield consistent estimates of sufficient predictive indices $\widehat{\boldsymbol{\psi}}_1'\widehat{\mathbf{f}}_t, \cdots, \widehat{\boldsymbol{\psi}}_L'\widehat{\mathbf{f}}_t$. We summarize the sufficient forecasting procedure with unobserved factors in Algorithm 1.

Algorithm 1 The sufficient forecasting

- Step 1: Obtain the estimated factors $\{\widehat{\mathbf{f}}_t\}_{t=1,\ldots,T}$ from (2.7) and (2.8);

- Step 2: Construct $\widehat{\boldsymbol{\Sigma}}_{f|y}$ as in (2.9) or (2.10);

- Step 3 Obtain $\widehat{\boldsymbol{\psi}}_1, \cdots, \widehat{\boldsymbol{\psi}}_L$ from the L largest eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{f|y}$;

- Step 4: Construct the predictive indices $\widehat{\boldsymbol{\psi}}_1'\widehat{\mathbf{f}}_t, \cdots, \widehat{\boldsymbol{\psi}}_L'\widehat{\mathbf{f}}_t$ and forecast y_{t+1}.

To make the forecasting procedure concrete, we elucidate how factors and factor loadings are estimated. We temporarily assume that the number of underlying factors K is known to us. Consider the following constrained least squares problem:

$$(\widehat{\mathbf{B}}_K, \widehat{\mathbf{F}}_K) \quad = \arg\min_{(\mathbf{B}, \mathbf{F})} \quad \|\mathbf{X} - \mathbf{B}\mathbf{F}'\|_F^2, \tag{2.7}$$

$$\text{subject to} \quad T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.8}$$

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. This is a classical principal components problem, and it has been widely used to extract underlying common factors (Stock and Watson, 2002a; Bai and Ng, 2002; Fan et al., 2013; Bai and Ng, 2013). The constraints in (2.8) correspond to the normalization (2.3). The minimizers $\widehat{\mathbf{F}}_K$ and $\widehat{\mathbf{B}}_K$ are such that the columns of $\widehat{\mathbf{F}}_K/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $\mathbf{X}'\mathbf{X}$ and

$$\widehat{\mathbf{B}}_K = T^{-1}\mathbf{X}\widehat{\mathbf{F}}_K.$$

To simplify notation, we use $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}_K$, $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}_K$ and $\widehat{\mathbf{F}}' = (\widehat{\mathbf{f}}_1, \cdots, \widehat{\mathbf{f}}_T)$ throughout this paper.

To effectively estimate $\boldsymbol{\Sigma}_{f|y}$ in (2.5) and (2.6), we shall replace the conditional expecta-

tions $E(\mathbf{f}_t|y_{t+1} \in I_h)$ and $E(\mathbf{x}_t|y_{t+1} \in I_h)$ by their sample counterparts. Denote the ordered statistics of $\{(y_{t+1}, \widehat{\mathbf{f}}_t)\}_{t=1,\dots,T-1}$ by $\{(y_{(t+1)}, \widehat{\mathbf{f}}_{(t)})\}_{t=1,\dots,T-1}$ according to the values of y, where $y_{(2)} \le \cdots \le y_{(T)}$ and we only use information up to time T. We divide the range of y into H slices, where H is typically fixed. Each of the first $H-1$ slices contains an even number of observations $c > 0$, and the last slice may have less than c observations, which exerts little influence asymptotically. For ease of presentation, we introduce a double script $_{(h,j)}$ in which $h = 1, \dots, H$ refers to the slice number and $j = 1, \dots, c$ is the index of an observation in a given slice. Thus, we can write $\{(y_{(t+1)}, \widehat{\mathbf{f}}_{(t)})\}_{t=1,\dots,T-1}$ as

$$\{(y_{(h,j)}, \widehat{\mathbf{f}}_{(h,j)}) : \ y_{(h,j)} = y_{(c(h-1)+j+1)}, \ \widehat{\mathbf{f}}_{(h,j)} = \widehat{\mathbf{f}}_{(c(h-1)+j)}\}_{h=1,\dots,H; \ j=1,\dots,c}.$$

Based on the estimated factors $\widehat{\mathbf{f}}_t$, we have the estimate $\widehat{\boldsymbol{\Sigma}}^1_{f|y}$ in the form of

$$\widehat{\boldsymbol{\Sigma}}^1_{f|y} = \frac{1}{H} \sum_{h=1}^{H} [\frac{1}{c} \sum_{l=1}^{c} \widehat{\mathbf{f}}_{(h,l)}][\frac{1}{c} \sum_{l=1}^{c} \widehat{\mathbf{f}}_{(h,l)}]'. \tag{2.9}$$

Analogously, by (2.6), we have an alternative estimator:

$$\widehat{\boldsymbol{\Sigma}}^2_{f|y} = \widehat{\boldsymbol{\Lambda}}_b (\frac{1}{H} \sum_{h=1}^{H} [\frac{1}{c} \sum_{l=1}^{c} \mathbf{x}_{(h,l)}][\frac{1}{c} \sum_{l=1}^{c} \mathbf{x}_{(h,l)}]') \widehat{\boldsymbol{\Lambda}}'_b. \tag{2.10}$$

where $\widehat{\boldsymbol{\Lambda}}_b = (\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}$ is the plug-in estimate of $\boldsymbol{\Lambda}_b$. The following proposition shows that the estimates of $\boldsymbol{\Sigma}_{f|y}$ based on either $\widehat{\boldsymbol{\Sigma}}^1_{f|y}$ or $\widehat{\boldsymbol{\Sigma}}^2_{f|y}$ are exactly equivalent.

Proposition 2.1. *Let $\widehat{\mathbf{f}}_t$ and $\widehat{\mathbf{B}}$ be estimated by solving the constrained least square problem in (2.7) and (2.8). Then, the two estimators (2.9) and (2.10) of $\boldsymbol{\Sigma}_{f|y}$ are equivalent, i.e.*

$$\widehat{\boldsymbol{\Sigma}}^1_{f|y} = \widehat{\boldsymbol{\Sigma}}^2_{f|y}.$$

Remark 2.1. There are alternative ways for estimating factors $\mathbf{F}$ and factor loadings $\mathbf{B}$. For example, Forni et al. (2000) studied factor estimation based on projection, Liu et al. (2012) and Xue and Zou (2012) employed rank-based methods to robust estimation of the covariance matrices, and Connor et al. (2012) applied a weighted additive nonparametric estimation procedure to estimate characteristic-based factor models. These methods do not necessarily lead to the equivalence of $\widehat{\boldsymbol{\Sigma}}^1_{f|y}$ and $\widehat{\boldsymbol{\Sigma}}^2_{f|y}$ as in Proposition 2.1.

2.3 Choices of tuning parameters

Sufficient forecasting includes determination of the number of slices H, the number of predictive indices L, and the number of factors K. In practice, H has little influence on the estimated directions, which was first pointed out in Li (1991). The choice of H differs from that of a smoothing parameter in nonparametric regression, which may lead to a slower rate of convergence. As shown in Section 3, we always have the same rate of convergence for $\widehat{\boldsymbol{\phi}}_i$ no matter how H is specified in the sufficient forecasting. The reason is that (2.5) gives the same eigenspace as long as $H \geq \max\{L, 2\}$.

In terms of the choice of L, the first L eigenvalues of $\boldsymbol{\Sigma}_{f|y}$ must be significantly different from zero compared to the estimation error. Several methods such as Li (1991) and Schott (1994) have been proposed to determine L. For instance, the average of the smallest $K - L$ eigenvalues of $\boldsymbol{\Sigma}_{f|y}$ would follow a χ^2-distribution if the underlying factors are normally distributed, where K is the number of underlying factors. Should that average be large, there would be at least $L + 1$ predictive indices in the sufficient forecasting model.

Pertaining to many factor analysis, the number of latent factors K might be unknown to us. There are many existing approaches to determining K in the literature, e.g., Bai and Ng (2002), Hallin and Liška (2007), Alessi et al. (2010). Recently, Lam et al. (2012) and Ahn and Horenstein (2013) proposed a ratio-based estimator by maximizing the ratio of two adjacent eigenvalues of $\mathbf{X}'\mathbf{X}$ arranged in descending order, i.e.

$$\hat{K} = \underset{1 \leq i \leq kmax}{\arg\max} \hat{\lambda}_i / \hat{\lambda}_{i+1},$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_T$ are the eigenvalues. The estimator enjoys good finite-sample performances and was motivated by the following observation: the K largest eigenvalues of $\mathbf{X}'\mathbf{X}$ grow unboundedly as p increases, while the others remain bounded. We note here that once a consistent estimator of K is found, the asymptotic results in this paper hold true for the unknown K case by a conditioning argument. Unless otherwise specified, we shall assume a known K in the sequel.

3 Asymptotic properties

We define some necessary notation here. For a vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes its Euclidean norm. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_1$ represent its spectral norm and ℓ_1 norm respectively, defined as the largest singular value of $\mathbf{M}$ and its maximum absolute column sum. $\|\mathbf{M}\|_\infty$ is its matrix ℓ_∞ norm defined as the maximum absolute row sum. For a symmetric matrix

$\mathbf{M}$, $\|\mathbf{M}\|_1 = \|\mathbf{M}\|_\infty$. Denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest eigenvalue of a symmetric matrix respectively.

3.1 Assumptions

We first detail the assumptions on the forecasting model (2.1) and the associated factor model (2.2), in which common factors $\{\mathbf{f}_t\}_{t=1,\ldots,T}$ and the forecasting function $h(\cdot)$ are unobserved and only $\{(\mathbf{x}_t, y_t)\}_{t=1,\ldots,T}$ are observable.

Assumption 3.1 (Factors and Loadings). *For some $M > 0$,*
(1) The loadings $\mathbf{b}_i$ satisfy that $\|\mathbf{b}_i\| \leq M$ for $i = 1, \ldots, p$. As $p \to \infty$, there exist two positive constants c_1 and c_2 such that

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

(2) Identification: $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$, and $\mathbf{B}'\mathbf{B}$ is a diagonal matrix with distinct entries.
(3) Linearity: $E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t)$ is a linear function of $\boldsymbol{\phi}_1'\mathbf{f}_t, \cdots, \boldsymbol{\phi}_L'\mathbf{f}_t$ for any $\mathbf{b} \in \mathbb{R}^p$, where $\boldsymbol{\phi}_i$'s come from model (2.1).

Condition (1) is often known as the *pervasive* condition (Bai and Ng, 2002; Fan et al., 2013) such that the factors impact a non-vanishing portion of the predictors. Condition (2) corresponds to the PC1 condition in Bai and Ng (2013), which eliminates rotational inderterminacy in the individual columns of $\mathbf{F}$ and $\mathbf{B}$. Condition (3) ensures that the (centered) inverse regression curve $E(\mathbf{f}_t|y_{t+1})$ is contained in the central subspace, and it is standard in the dimension reduction literature. The linearity condition is satisfied when the distribution of $\mathbf{f}_t$ is elliptically symmetric (Eaton, 1983), and it is also asymptotically justified when the dimension of $\mathbf{f}_t$ is large (Hall and Li, 1993).

We impose the strong mixing condition on the data generating process. Let $\mathcal{F}_\infty^0$ and $\mathcal{F}_T^\infty$ denote the $\sigma-$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

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

Assumption 3.2 (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 processes, and they are strictly stationary. The factor process satisfies that $E\|\mathbf{f}_t\|^4 < \infty$ and $E(\|\mathbf{f}_t\|^2|y_{t+1}) < \infty$. In addition, $\alpha(T) < c\rho^T$ for all $T \in \mathbb{Z}^+$ and some $\rho \in (0, 1)$.*

Moreover, we make the following assumption on the residuals and dependence of the

factor model (2.2). Such conditions are similar to those in Bai (2003), which are needed to consistently estimate the common factors as well as the factor loadings.

Assumption 3.3 (Residuals and Dependence)**.** *There exists a positive constant $M < \infty$ that does not depend on p and T, such that*
(1) $E(\mathbf{u}_t) = \mathbf{0}$, and $E|u_{it}|^8 \leq M$.
(2) $\|\mathbf{\Sigma}_u\|_1 \leq M$, and for every $i, j, t, s > 0$, $(pT)^{-1} \sum_{i,j,t,s} |E(u_{it}u_{js})| \leq M$
(3) 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   Rate of convergence

The following theorem establishes the convergence rate of the sliced covariance estimate of inverse regression curve (i.e. $\widehat{\mathbf{\Sigma}}_{f|y}$ as in (2.9) or (2.10)) under the spectral norm. It further implies the same convergence rate of the estimated SDR directions associated with sufficient predictive indices. For simplicity, we assume that K and H are fixed, though they can be extended to allow K and H to depend on T. This assumption enables us to obtain faster rate of convergence. Note that the regression indices $\{\boldsymbol{\phi}_i\}_{i=1}^L$ and the PCA directions $\{\boldsymbol{\psi}_i\}_{i=1}^L$ are typically different, but they can expand the same linear space.

Theorem 3.1. *Suppose that Assumptions 3.1-3.3 hold and let $\omega_{p,T} = p^{-1/2} + T^{-1/2}$. Then under the forecasting model (2.1) and the associated factor model (2.2), we have*

$$\|\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{\Sigma}_{f|y}\| = O_p(\omega_{p,T}). \tag{3.1}$$

If the L largest eigenvalues of $\mathbf{\Sigma}_{f|y}$ are positive and distinct, the eigenvectors $\widehat{\boldsymbol{\psi}}_1, \cdots, \widehat{\boldsymbol{\psi}}_L$ associated with the L largest eigenvalues of $\widehat{\mathbf{\Sigma}}_{f|y}$ give the consistent estimate of directions $\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_L$ respectively, with rates

$$\|\widehat{\boldsymbol{\psi}}_j - \boldsymbol{\psi}_j\| = O_p(\omega_{p,T}), \tag{3.2}$$

for $j = 1, \cdots, L$, where $\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_L$ form another orthonormal basis for the central subspace $S_{y|\mathbf{f}}$ spanned by $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$.

When $L = 1$ and the link function $h(\cdot)$ in model (2.1) is linear, Stock and Watson (2002a) showed that having estimated factors as regressors in the linear forecasting does not affect consistency of the parameter estimates in a forecasting equation. Our Theorem 3.1 extends the theory and method of Stock and Watson (2002a) to show that it is also true for any unknown link function $h(\cdot)$ and multiple predictive indices by using the sufficient forecasting.

12

The two estimates coincide in the linear forecasting case, but our procedure can also deal with the nonlinear forecasting cases.

As a consequence of Theorem 3.1, the sufficient predictive indices can be consistently estimated as well. We summarize this result in the following corollary.

Corollary 3.1. *Under the same conditions of Theorem 3.1, for any $j = 1, \ldots, L$, we have*

$$\widehat{\boldsymbol{\psi}}_j' \widehat{\mathbf{f}}_t \to_p \boldsymbol{\psi}_j \mathbf{f}_t. \tag{3.3}$$

Translating into the original predictors $\mathbf{x}_t$ and letting $\widehat{\boldsymbol{\xi}}_j = \widehat{\boldsymbol{\Lambda}}_b' \widehat{\boldsymbol{\psi}}_j$, we then have

$$\widetilde{\boldsymbol{\xi}}_j' \mathbf{x}_t \to_p \boldsymbol{\psi}_j \mathbf{f}_t. \tag{3.4}$$

Remark 3.1. In view of (3.3) and (3.4) in Corollary 3.1, the sufficient forecasting finds not only the sufficient dimension reduction directions $\widehat{\boldsymbol{\psi}}_1, \cdots, \widehat{\boldsymbol{\psi}}_L$ for estimated factors $\widehat{\mathbf{f}}_t$ but also the sufficient dimension reduction directions $\widehat{\boldsymbol{\xi}}_1, \cdots, \widehat{\boldsymbol{\xi}}_L$ for predictor $\mathbf{x}_t$. The sufficient forecasting effectively obtains the projection of the p-dimensional predictor $\mathbf{x}_t$ onto the L-dimensional subspace $\widehat{\boldsymbol{\xi}}_1' \mathbf{x}_t, \cdots, \widehat{\boldsymbol{\xi}}_L' \mathbf{x}_t$ by using the factor model, where the number of predictors p can be larger than the number of observations T. It is well-known that the traditional sliced inverse regression and many other dimension reduction methods are limited to either a fixed dimension or a diverging dimension that is smaller than the sample size (Zhu et al., 2006). By condensing the cross-sectional information into indices, our work alleviates what plagues the sufficient dimension reduction in high-dimensional regimes, and the factor structure in (2.1) and (2.2) effectively reduces the dimension of predictors and extends the methodology and applicability of the sufficient dimension reduction.

3.3   Linear forecasting under link violation

Having extracted common factors from the vast predictors, the PCR is a simple and natural method to run a multiple regression and forecast the outcome. Such a linear estimate is easy to construct and provides a benchmark for our analysis. When the underlying relationship between the response and the latent factors is nonlinear, directly applying linear forecasts would violate the link function $h(\cdot)$ in (2.2). But what does this method do in the case of model misspecification? In what follows, we shall see that asymptotically the linear forecast falls into the central subspace $S_{y|\mathbf{f}}$, namely, it is a weighted average of the sufficient predictive indices.

With the underlying factors $\{\widehat{\mathbf{f}}_t\}$ estimated as in (2.7) and (2.8), the response y_{t+1} is then regressed on $\widehat{\mathbf{f}}_t$ via a multiple regression. In this case, the design matrix is orthogonal due to

13

the normalization (2.8). Therefore, the least-squares estimate of regression coefficients from regressing y_{t+1} on $\widehat{\mathbf{f}}_t$ is

$$\widehat{\boldsymbol{\phi}} = \frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1} \widehat{\mathbf{f}}_t. \tag{3.5}$$

To study the behavior of regression coefficient $\widehat{\boldsymbol{\phi}}$, we shall assume normality of the underlying factors $\mathbf{f}_t$. The following theorem shows that, regardless of the specification of the link function $h(\cdot)$, $\widehat{\boldsymbol{\phi}}$ falls into the central subspace $S_{y|\mathbf{f}}$ spanned by $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$ as $p, T \to \infty$. More specifically, it converges to the sufficient direction

$$\bar{\boldsymbol{\phi}} = \sum_{i=1}^{L} E((\boldsymbol{\phi}_i' \mathbf{f}_t) y_{t+1}) \boldsymbol{\phi}_i,$$

which is a weighted average of sufficient directions $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_L$. In particular, when the link function is linear and $L = 1$, the PCR yields an asymptotically consistent direction.

Theorem 3.2. *Consider model (2.1) and (2.2) under assumptions of Theorem 3.1. Suppose the factors $\mathbf{f}_t$ are normally distributed and that $E(y_t^2) < \infty$. Then, we have*

$$\|\widehat{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}}\| = O_p(\omega_{p,T}). \tag{3.6}$$

As a consequence of Theorem 3.2, the linear forecast $\widehat{\boldsymbol{\phi}}' \widehat{\mathbf{f}}_t$ is a weighted average of the sufficient predictive indices $\boldsymbol{\phi}_1' \mathbf{f}_t, \cdots, \boldsymbol{\phi}_L' \mathbf{f}_t$.

Corollary 3.2. *Under the same conditions of Theorem 3.2, we have*

$$\widehat{\boldsymbol{\phi}}' \widehat{\mathbf{f}}_t \to_p \bar{\boldsymbol{\phi}}' \mathbf{f}_t = \sum_{i=1}^{L} E((\boldsymbol{\phi}_i' \mathbf{f}_t) y_{t+1}) \boldsymbol{\phi}_i' \mathbf{f}_t. \tag{3.7}$$

It is interesting to see that when $L = 1$, $\bar{\boldsymbol{\phi}} = E((\boldsymbol{\phi}_1' \mathbf{f}_t) y_{t+1}) \boldsymbol{\phi}_1$. Thus, the least squares coefficient $\widehat{\boldsymbol{\phi}}$ in (3.5), using the estimated factors as covariates, delivers an asymptotically consistent estimate of the forecasting direction $\boldsymbol{\phi}_1$ up to a multiplicative scalar. Thus, we can now get a scatter plot for the data $\{(\widehat{\boldsymbol{\phi}}' \widehat{\mathbf{f}}_t, y_{t+1})\}_{t=1}^{T-1}$ to check if the linear fit suffices and employs a nonparametric smoothing to get a better fit when needed. This can correct the bias of PCR when needed.

PCR can only deliver one index of extracted factors for forecasting. When there exist multiple sufficient predictive indices (i.e., $L \geq 2$), the above robustness no longer holds. The estimated coefficient $\widehat{\boldsymbol{\phi}}$ belongs to the central subspace, and it only reveals one dimension of the central subspace $S_{y|\mathbf{f}}$, leading to limited predictive power. The central subspace $S_{y|\mathbf{f}}$,

in contrast, is entirely contained in $\mathbf{\Sigma}_{f|y}$ as in (2.5). By estimating $\mathbf{\Sigma}_{f|y}$, the sufficient forecasting tries to recover all the effective directions and would therefore capture more driving forces to forecast.

As demonstrated in Section 4, the proposed method is comparable to the linear forecasting when $L = 1$, but significantly outperforms the latter when $L \geq 2$ in both simulation studies and empirical applications.

4 Numerical studies

In this section, we conduct both Monte Carlo experiments and an empirical study to assess the numerical performance of the sufficient forecasting using factor models.

4.1 Linear predictive model

We first consider the case when the target is a linear function of the latent factors plus some noise. Prominent examples include asset return predictability, where we could use the cross section of book-to-market ratios to forecast aggregate market returns (Campbell and Shiller, 1988; Polk et al., 2006; Kelly and Pruitt, 2013). To this end, we specify our data generating process as

$$y_{t+1} = \boldsymbol{\phi}'\mathbf{f}_t + \sigma_y\epsilon_{t+1},$$
$$x_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it},$$

where we let $K = 5$ and $\boldsymbol{\phi} = (0.8, 0.5, 0.3, 0, 0)'$, namely there are 5 common factors that drives the predictors and a linear combination of those five factors predicts the outcome. Factor loadings $\mathbf{b}_i$'s are drawn from the standard normal distribution. To account for serial correlation, we set f_{jt} and u_{it} as two AR(1) processes respectively,

$$f_{jt} = \alpha_j f_{jt-1} + e_{jt}, \quad u_{it} = \rho_i u_{it-1} + \nu_{it}.$$

We draw α_j, ρ_i from $\sim U[0.2, 0.8]$ and fix them during simulations, while the shocks e_{jt}, ν_{it} and ϵ_{t+1} are standard normal respectively. σ_y is taken as the variance of the factors, so that the infeasible best forecast using $\boldsymbol{\phi}'\mathbf{f}_t$ has an R^2 of approximately 50%.

We examine both the in-sample and out-of-sample performance between the principal component regression and the proposed sufficient forecasting. Akin to the in-sample R^2 typically used in linear regression, the out-of-sample R^2, suggested by Campbell and Thompson

(2008), is defined here as

$$R_{\text{OS}}^2 = 1 - \frac{\sum_{t=[T/2]}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=[T/2]}^{T}(y_t - \bar{y}_t)^2},$$

(4.1)

where in each replication we use the second half of the time series as the testing sample. $\hat{y}_t$ is the fitted value from the predictive regression using all information before t and $\bar{y}_t$ is the mean from the test sample. Note that R_{OS}^2 can be negative.

Table 1: Performance of forecasts using in-sample and out-sample R^2

		In-sample R^2 (%)			Out-of-sample R^2 (%)		
p	T	SF(1)	PCR	PC1	SF(1)	PCR	PC1
50	100	46.9	47.7	7.8	35.1	39.5	2.4
50	200	46.3	46.5	6.6	42.3	41.7	4.4
100	100	49.3	50.1	8.9	37.6	40.3	3.0
100	500	47.8	47.8	5.5	43.6	43.5	1.1
500	100	48.5	48.8	7.9	40.0	43.1	4.7
500	500	48.2	48.3	7.2	48.0	47.9	6.0

Notes: In-sample and out-of-sample median R^2, recorded in percentage, over 1000 simulations. SF(1) denotes the sufficient forecast using one single predictive index, PCR denotes principal component regression, and PC1 uses only the first principal component.

Numerical results are summarized in Table 1, where SF(1) denotes the sufficient forecast using only one single predictive index, PCR denotes the principal component regression, and PC1 uses only the first principal component. Note that the first principal component is not necessarily the same as the sufficient regression direction ϕ. Hence, PC1 can perform poorly. As shown in Table 1, the sufficient forecasting SF(1), by finding one single projection of the latent factors, yields comparable results as PCR, which has a total of five regressors. The good performance of SF(1) is consistent with Theorem 3.1, whereas the good performance of PCR is due to the fact that the true predictive model is linear (see Stock and Watson (2002a) and Theorem 3.2). In contrast, using the first principal component alone has very poor performance in general, as it is not the index for the regression target.

4.2 Predictive Model with Factor Interaction

We next turn to the case when the interaction between factors is present. Interaction models have been employed by many researchers in both economics and statistics. For

example, in an influential work, Rajan and Zingales (1998) examined the interaction between financial dependence and economic growth. Consider the model

$$y_{t+1} = f_{1t}(f_{2t} + f_{3t} + 1) + \epsilon_{t+1},$$

where ϵ_{t+1} is taken to be standard normal. The data generating process for the predictors x_{it}'s is set to be the same as that in the previous simulation, but we let K=7, i.e., 7 factors drive the predictors with only the first 3 factors associated with the response (in a nonlinear fashion). The true sufficient directions are the vectors in the plane $S_{f|y}$ generated by $\phi_1 = (1, 0, 0, 0, 0, 0, 0)'$ and $\phi_2 = (0, 1, 1, 0, 0, 0, 0)'/\sqrt{2}$. Following the procedure in Algorithm 1, we can find the estimated factors $\widehat{\mathbf{f}}_t$ as well as the forecasting directions $\widehat{\phi}_1$ and $\widehat{\phi}_2$.

Note that the forecasting model only involves a strict subset of the underlying factors. Had we made a linear forecast, all the estimated factors would likely be selected, resulting in issues such as use of irrelevant factors (i.e., only the first three factors suffice, but the first 7 factors are used). This is evident through the correlations between the response and the estimated factors, as in Table 2 and results in both bias (fitting a wrong model) and variance issue.

Table 2: $|\text{Corr}(y_{t+1}, \widehat{\mathbf{f}}_{it})|$

p	T	$\widehat{\mathbf{f}}_{1t}$	$\widehat{\mathbf{f}}_{2t}$	$\widehat{\mathbf{f}}_{3t}$	$\widehat{\mathbf{f}}_{4t}$	$\widehat{\mathbf{f}}_{5t}$	$\widehat{\mathbf{f}}_{6t}$	$\widehat{\mathbf{f}}_{7t}$
500	500	0.1819	0.1877	0.1832	0.1700	0.1737	0.1712	0.1663

Notes: Median correlation between the forecast target and the estimated factors in absolute values, based on 1000 replications.

Next, we compare the numerical performance between the principal component regression and the proposed sufficient forecasting. As in Section 4.1, we again examine their in-sample and out-of-sample forecasting performances using the in-sample R^2 and the out-of-sample R^2 respectively. Furthermore, we will measure the distance between any estimated regression direction $\widehat{\phi}$ and the central subspace $S_{f|y}$ spanned by ϕ_1 and ϕ_2. We ensure that the true factors and loadings meet the identifiability conditions by calculating an invertible matrix $\mathbf{H}$ such that $\frac{1}{T}\mathbf{H}\mathbf{F}'\mathbf{F}\mathbf{H}' = \mathbf{I}_K$ and $\mathbf{H}^{-1}\mathbf{B}'\mathbf{B}\mathbf{H}'^{-1}$ is diagonal. The space we are looking for is then spanned by $\mathbf{H}^{-1}\phi_1$ and $\mathbf{H}^{-1}\phi_2$. For notation simplicity, we still represent the rotated forecasting space as $S_{f|y}$. We employ the squared multiple correlation coefficient $R^2(\widehat{\phi})$ to evaluate the effectiveness of $\widehat{\phi}$, where

$$R^2(\widehat{\phi}) = \max_{\phi \in S_{f|y}, \|\phi\|=1} (\widehat{\phi}'\phi)^2.$$

This measure is commonly used in the dimension reduction literature; see Li (1991). We calculate $R^2(\widehat{\phi})$ for the sufficient forecasting directions $\widehat{\phi}_1, \widehat{\phi}_2$ and the PCR direction $\widehat{\phi}_{pcr}$.

Table 3: Performance of estimated index $\widehat{\phi}$ using $R^2(\widehat{\phi})$ (%)

		Sufficient forecasting		PCR
p	T	$R^2(\widehat{\phi}_1)$	$R^2(\widehat{\phi}_2)$	$R^2(\widehat{\phi}_{pcr})$
100	100	84.5 (16.3)	64.4 (25.0)	91.4 (9.1)
100	200	92.9 (7.8)	83.0 (17.5)	95.7 (4.6)
100	500	96.9 (2.2)	94.1 (5.0)	96.7 (1.6)
500	100	85.4 (15.9)	57.9 (24.8)	91.9 (8.8)
500	200	92.9 (6.7)	82.9 (16.1)	95.9 (4.1)
500	500	97.0 (2.0)	94.5 (4.5)	98.2 (1.5)

Notes: Median squared multiple correlation coefficients in percentage over 1000 replications. The values in parentheses are the associated standard deviations. PCR denotes principal component regression.

Table 4: Performance of forecasts using in-sample and out-sample R^2

		In-sample R^2 (%)			Out-of-sample R^2 (%)		
p	T	SFi	PCR	PCRi	SFi	PCR	PCRi
100	100	46.2	38.5	42.4	20.8	12.7	13.5
100	200	57.7	35.1	38.6	41.6	24.0	24.7
100	500	77.0	31.9	34.9	69.7	29.1	31.5
500	100	49.5	38.7	22.7	21.9	16.6	19.2
500	200	58.9	34.7	39.0	40.2	22.2	24.0
500	500	79.8	32.5	35.6	72.3	26.9	28.2

Notes: In-sample and out-of-sample median R^2 in percentage over 1000 replications. SFi uses first two predictive indices and includes their interaction effect; PCR uses all principal components; PCRi extends PCR by including an extra interaction term built on the first two principal components of the covariance matrix of predictors.

In the sequel, we draw two conclusions from the simulation results summarized in Tables 3 and 4 based on 1000 independent replications.

Firstly, we observe from Table 3 that the length of time-series T has an important effect on $R^2(\widehat{\phi}_1)$, $R^2(\widehat{\phi}_2)$ and $R^2(\widehat{\phi}_{pcr})$. As T increases from 100 to 500, all the squared multiple correlation coefficients increase too. In general, the PCR and the proposed sufficient forecasting perform very well in terms of the measure $R^2(\widehat{\phi})$. The good performance of $R^2(\widehat{\phi}_{pcr})$

is a testimony of the correctness of Theorem 3.2. However, the PCR is limited to the linear forecasting function, and it only picks up one effective sufficient direction. The sufficient forecasting successfully picks up both effective sufficient directions in the forecasting equation, which is consistent to Theorem 3.1.

Secondly, we evaluate the power of sufficient forecasting using the two predictive indices $\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_t$ and $\widehat{\boldsymbol{\phi}}_2'\widehat{\mathbf{f}}_t$. Similar to the previous section, we build a simple linear regression to forecast, but include three terms, $\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_1$, $\widehat{\boldsymbol{\phi}}_2'\widehat{\mathbf{f}}_2$ and $(\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_1)\cdot(\widehat{\boldsymbol{\phi}}_2'\widehat{\mathbf{f}}_2)$, as regressors to account for interaction effects. In-sample and out-of-sample R^2 are reported. For comparison purposes, we also report results from principal component regression (PCR), which uses the 7 estimated factors extracted from predictors. In addition, we add the interaction between the first two principal components on top of the PCR, and denote this method as PCRi. Note that the principal component directions differ from the sufficient forecast directions, and it does not take into account the target information. Therefore, PCRi model is very different from the sufficient regression with interaction terms. As shown in Table 4, the in-sample R^2's of PCR hover around 35%, and its out-of-sample R^2's are relatively low. Including interaction between the first two PCs of the covariance of the predictors does not help much, as the interaction term is incorrectly specified for this model. SFi picks up the correct form of interaction and exhibit better performance, especially when T gets reasonably large.

4.3 An empirical example

As an empirical investigation, we use sufficient forecasting to forecast several macroeconomic variables. Our dataset is taken from Stock and Watson (2012), which consists of quarterly observations on 108 U.S. low-level disaggregated macroeconomic time series from 1959:I to 2008:IV. Similar datasets have been extensively studied in the forecasting literature (Bai and Ng, 2008; Ludvigson and Ng, 2009). These time series are transformed by taking logarithms and/or differencing to make them approximately stationary. We treat each of them as a forecast target y_t, with all the others forming the predictor set $\mathbf{x}_t$. The out-of-sample performance of each time series is examined, measured by out-of-sample R^2 in (4.1) . The procedure involves fully recursive factor estimation and parameter estimation starting half-way of the sample, using data from the beginning of the sample through quarter t for forecasting in quarter $t + 1$.

We present the forecasting results for a few representatives in each macroeconomic category. The time series are chosen such that when doing sufficient forecasting, the second eigenvalue of $\widehat{\boldsymbol{\Sigma}}_{f|y}$ exceeds 50% of the first eigenvalue in the first half training sample. This implies that a single predictive index may not be enough to forecast the given target, so

we could consider the effect of second predictive index. Note that in some categories, the second eigenvalues $\widehat{\boldsymbol{\Sigma}}_{f|y}$ for all the time series are very small; in this case, we randomly pick a representative. As described in previous simulations, we use SF(1) to denote sufficient forecasting with one single predictive index, which corresponds to a linear forecasting equation with one regressor. SF(2) employs an additional predictive index on top of SF(1), and is fit by local linear regression. In all cases, we use 7 estimated factors extracted from 107 predictors, which are found to explain over 40 percent for the variation in the data; see Bai and Ng (2013).

Table 5: Out-of-sample Macroeconomic Forecasting

Category	Label	SF(1)	SF(2)	PCR	PC1
GDP components	GDP264	11.2	14.9	13.8	9.2
IP	IPS13	17.8	21.0	20.6	-2.9
Employment	CES048	24.1	26.4	23.4	25.7
Unemployment rate	LHU680	19.0	21.7	13.6	21.1
Housing	HSSOU	1.9	4.4	-0.8	7.4
Inventories	PMNO	19.8	20.2	18.6	6.3
Prices	GDP275_3	9.1	11.3	10.2	0.2
Wages	LBMNU	32.7	34.2	34.0	29.1
Interest rates	FYFF	3.6	2.4	-31.5	1.6
Money	CCINRV	4.8	16.0	1.3	-0.5
Exchange rates	EXRCAN	-4.7	4.5	-7.9	-1.3
Stock prices	FSDJ	-9.2	8.8	-14.5	-7.7
Consumer expectations	HHSNTN	-3.7	-5.0	-4.4	0.0

Notes: Out-of-sample R^2 for one–quarter ahead forecasts. SF(1) uses a single predictive index built on 7 estimated factors to make a linear forecast. SF(2) is fit by local linear regression using the first two predictive indices. PCR uses all 7 principal components as linear regressors. PC1 uses the first principal component.

Table 5 shows that SF(1) yields comparable performance as PCR. There are cases where SF(1) exhibits more predictability than PCR, e.g., Inventories and Interest rates. This is due to the fact that there are possible nonlinear effects from extracted factors to the target. In this case, the first predictive index obtained from our procedure accounts for such a nonlinear effect whereas the index created by PCR finds only the best linear approximation and can not find a second index. SF(2) improves predictability in many cases, where the target time series might not be linear in the latent factors. Taking CCINRV (consumer credit outstanding) for

example, Figure 2 plots the eigenvalues of its corresponding $\widehat{\Sigma}_{f|y}$, the estimated regression surface and the running out-of-sample R^2's by different sample split date. As can been seen from the plot, there is a non-linear effect of the two underlying macro factors on the target. By taking such effect into account, SF(2) consistently outperforms the other methods.
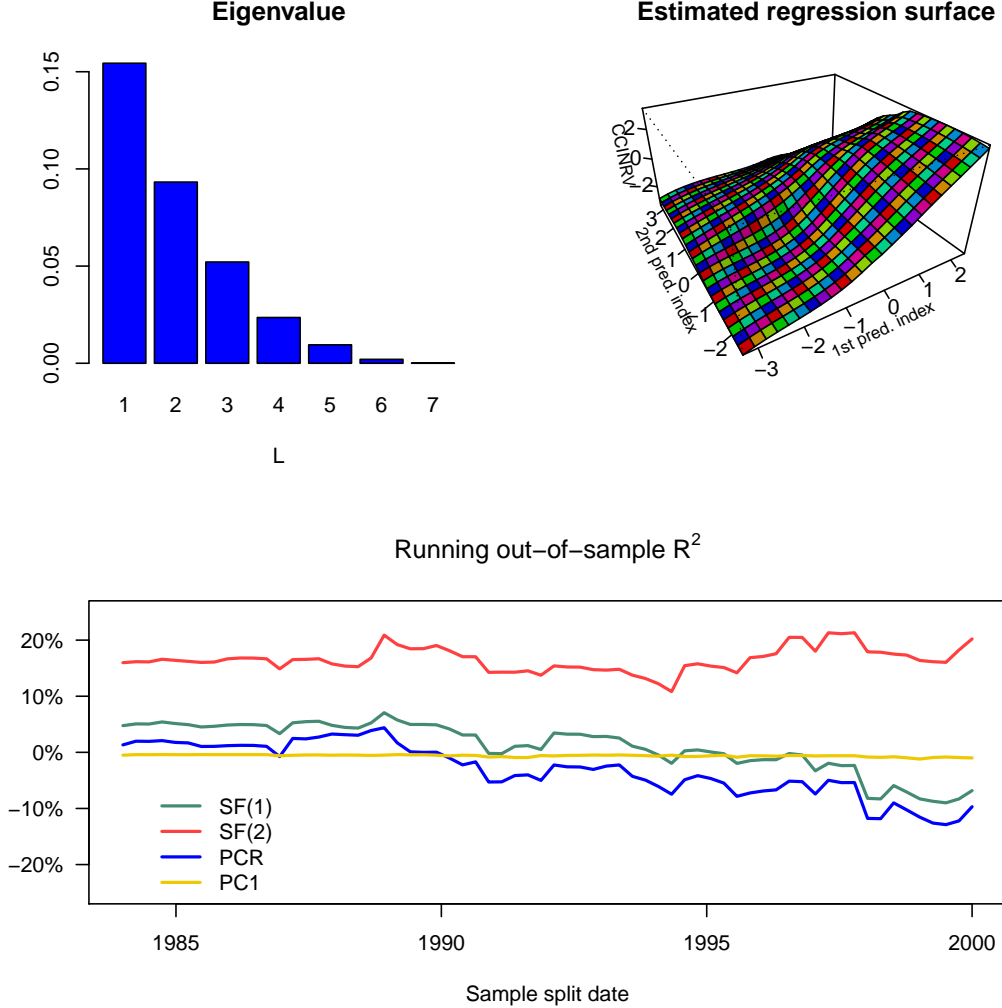

Figure 2: Forecasting results for CCINRV (consumer credit outstanding). The top left panel shows the eigenvalues of $\widehat{\Sigma}_{f|y}$. The top right panel gives a 3-d plot of the estimated regression surface. The lower panel displays the running out-of-sample R^2's for four methods described in Table 5.

5   Discussion

We have introduced the sufficient forecasting in a high-dimensional environment to forecast a single time series, which enlarges the scope of traditional factor forecasting. The key

feature of the sufficient forecasting is its ability in extracting multiple predictive indices and providing additional predictive power beyond the simple linear forecasting, especially when the target is an unknown nonlinear function of underlying factors. We explicitly point out the interesting connection between the proposed sufficient forecasting and the deep learning (Bengio, 2009; Bengio et al., 2013). Furthermore, the proposed method extends the sufficient dimension reduction to high-dimensional regimes by condensing the cross-sectional information through factor models. We have demonstrated the efficacy through Monte Carlo experiments. Our empirical results on macroeconomic forecasting also suggest that such procedure can contribute to substantial improvement beyond conventional linear models.

6 Appendix

We first cite two lemmas from Fan et al. (2013), which are needed subsequently in the proofs.

Lemma 6.1. *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are two semi-positive definite matrices, and that* $\lambda_{\min}(\mathbf{A}) > c_{p,T}$ *for some sequence* $c_{p,T} > 0$. *If* $\|\mathbf{A} - \mathbf{B}\| = o_p(c_{p,T})$, *then*

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = O_p(c_{p,T}^{-2})\|\mathbf{A} - \mathbf{B}\|.$$

Lemma 6.2. *Let* $\{\lambda_i\}_{i=1}^p$ *be the eigenvalues of* $\mathbf{\Sigma}$ *in descending order and* $\{\boldsymbol{\xi}\}_{i=1}^p$ *be their associated eigenvectors. Correspondingly, let* $\{\widehat{\lambda}_i\}_{i=1}^p$ *be the eigenvalues of* $\widehat{\mathbf{\Sigma}}$ *in descending order and* $\{\widehat{\boldsymbol{\xi}}\}_{i=1}^p$ *be their associated eigenvectors. Then,*

(a) Weyl's Theorem:

$$|\widehat{\lambda}_i - \lambda_i| \leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|.$$

(b) $\sin(\theta)$ *Theorem (Davis and Kahan, 1970):*

$$\|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\| \leq \frac{\sqrt{2} \cdot \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{\min(|\widehat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \widehat{\lambda}_{i+1}|)}.$$

6.1 Proof of Proposition 2.1

Proof. It suffices to show that $\widehat{\mathbf{f}}_t = \widehat{\mathbf{\Lambda}}_b \mathbf{x}_t$, or equivalently, $\widehat{\mathbf{F}}' = \widehat{\mathbf{\Lambda}}_b \mathbf{X}$ in matrix form. First we let $\mathbf{M} = \mathrm{diag}(\lambda_1, \cdots, \lambda_K)$, where λ_i are the largest K eigenvalues of $\mathbf{X}'\mathbf{X}$. By construction, we have $(\mathbf{X}'\mathbf{X})\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{M}$. Thus,

$$\widehat{\mathbf{B}}'\widehat{\mathbf{B}} = T^{-2}\widehat{\mathbf{F}}'(\mathbf{X}'\mathbf{X})\widehat{\mathbf{F}} = T^{-2}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M} = T^{-1}\mathbf{M}.$$

Now, it is easy to see that

$$(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}'\mathbf{X} = T\mathbf{M}^{-1}(T^{-1}\widehat{\mathbf{F}}'\mathbf{X}')\mathbf{X} = \mathbf{M}^{-1}(\mathbf{X}'\mathbf{X}\widehat{\mathbf{F}})' = \widehat{\mathbf{F}}'.$$

This concludes the proof of Proposition 2.1.

$\square$

6.2 Proof of Theorem 3.1

Let $\mathbf{V}$ denote the $K \times K$ diagonal matrix consisting of the K largest eigenvalues of the sample covariance matrix $T^{-1}\mathbf{X}'\mathbf{X}$ in the descending order. Define a $K \times K$ matrix

$$\mathbf{H} = (1/T)\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}, \tag{6.1}$$

where $\mathbf{F}' = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$. In factor models, the rotation matrix $\mathbf{H}$ is typically used for identifiability. With identification in Assumptions 3.1 (2), Bai and Ng (2013) showed that $\mathbf{H}$ is asymptotically an identity matrix. The result continues to hold under our setting.

Lemma 6.3. *Under Assumptions 3.1 (1)–(2), 3.2 and 3.3, we have*

$$||\mathbf{H} - \mathbf{I}_K|| = O_p(\omega_{p,T})$$

Proof. In view of Assumption A and Appendix B of Bai and Ng (2013), we only need to verify that Assumption A (c.ii) in Bai and Ng (2013) is satisfied under our models and assumptions, since all other assumptions of theirs are met here. To verity Assumption A (c.ii), it is enough to show that $\sum_{t=1}^{T} |E(u_{it}u_{js})| \leq M_1$ and $\sum_{i=1}^{p} |E(u_{it}u_{js})| \leq M_2$ for some positive constants $M_1, M_2 < \infty$.

Consider $E(u_{it}u_{js})$. Since $E|u_{it}|^8 \leq M$ for some constant M and any t by Assumption 3.3(1), by using Davydov's inequality (Corollary 16.2.4 in Athreya and Lahiri (2006)), there exist constants C_1 and C_2 such that for all (i,j),

$$|E(u_{it}u_{js})| \leq C_1 \left(\alpha\left(|t-s|\right)\right)^{1-\frac{1}{8}-\frac{1}{8}} \leq C_2 \rho^{\frac{3}{4}|t-s|},$$

where $\alpha(\cdot)$ is the mixing coefficient, $\rho \in (0,1)$, and $\alpha\left(|t-s|\right) < c\rho^{|t-s|}$ by Assumption 3.2. Hence, with $M_1 = 2C_2/(1-\rho^{\frac{3}{4}})$, we have

$$\sum_{t=1}^{T} |E(u_{it}u_{js})| = \sum_{t=1}^{T} C_2 \rho^{\frac{3}{4}|t-s|} \leq 2C_2/(1-\rho^{\frac{3}{4}}) = M_1.$$

In addition, it is easy to verify that there is a constant C_3 such that

$$|E(u_{it}u_{js})| \leq C_3|E(u_{it}u_{jt})| = C_3|\boldsymbol{\Sigma}_u(i,j)|.$$

By assumption, we also have

$$\sum_{i=1}^{p} |E(u_{it}u_{js})| \leq C_3 \sum_{i=1}^{p} |\boldsymbol{\Sigma}_u(i,j)| \leq C_3 \|\boldsymbol{\Sigma}_u\|_1 \leq C_3 M.$$

Thus, Assumption A (c.ii) in Bai and Ng (2013) is satisfied with $M_2 = C_3 M$. Therefore, with identification in Assumption 3.1(2), we have

$$||\mathbf{H} - \mathbf{I}_K|| = O_p(\omega_{p,T}),$$

which completes the proof of Lemma 6.3.

$\square$

The next lemma shows that the normalization matrix $\boldsymbol{\Lambda}_b$ can be consistently estimated by the plug-in estimate $\widehat{\boldsymbol{\Lambda}}_b$ under the spectral norm, which is an important result to prove Theorem 3.1.

Lemma 6.4. *Under Assumptions 3.1 (1)–(2), 3.2 and 3.3, we have*

(a) $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_p(p^{1/2}\omega_{p,T})$,

(b) $\|\widehat{\mathbf{\Lambda}}_b - \mathbf{\Lambda}_b\| = O_p(p^{-1/2}\omega_{p,T})$.

Proof. (a) First of all,

$$
\begin{aligned}
\|\widehat{\mathbf{B}} - \mathbf{B}\| &= \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}' + \mathbf{B}(\mathbf{H}' - \mathbf{I}_K)\| \\
&\leq \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'\| + \|\mathbf{B}(\mathbf{H}' - \mathbf{I}_K)\|,
\end{aligned}
$$

where $\mathbf{H}$ is defined in (6.1). The second term is bounded by $\|\mathbf{B}\|\|\mathbf{H}' - \mathbf{I}_K)\| = O_p(p^{1/2}\omega_{p,T})$ by Lemma 6.3 and using the fact $\|\mathbf{B}\|^2 = \lambda_{\max}(\mathbf{B}'\mathbf{B}) < c_2 p$ from Assumption 3.1. It remains to bound the first term. Using the fact that $\|\cdot\| \leq \|\cdot\|_F$, we have

$$
\|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'\|^2 \leq \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'\|_F^2 = \sum_{i=1}^p \|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|^2 \tag{6.2}
$$

Next, note that both $\widehat{\mathbf{b}}_i = (1/T)\sum_{t=1}^T x_{it}\widehat{\mathbf{f}}_t$ and $(1/T)\sum_{t=1}^T \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t' = \mathbf{I}_K$ hold as a result of the constrained least-squares problem (2.7) and (2.8). Then, we have

$$
\begin{aligned}
\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i &= \frac{1}{T}\sum_{t=1}^T (\widehat{\mathbf{f}}_t x_{it} - \mathbf{H}\mathbf{f}_t x_{it}) + \frac{1}{T}\sum_{t=1}^T \mathbf{H}\mathbf{f}_t(\mathbf{f}_t'\mathbf{b}_i + u_{it}) - \mathbf{H}\mathbf{b}_i \\
&= \frac{1}{T}\sum_{t=1}^T x_{it}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + \frac{1}{T}\sum_{t=1}^T \mathbf{H}\mathbf{f}_t u_{it} \tag{6.3}
\end{aligned}
$$

where we have used the fact that $\frac{1}{T}\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K = 0$. The remaining two terms on the right-hand side can be separately bounded as follows.

Under Assumptions 3.1 (1)–(2), 3.2 and 3.3, we have the following rate of convergence of the estimated factors,

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

This result is proved in Theorem 1 of Bai and Ng (2002).

For the first term in (6.3), it is easy to see that $E(x_{it}^2) = O(1)$ and thus $T^{-1}\sum_{t=1}^T x_{it}^2 = O_p(1)$. Hence, we use the Cauchy–Schwarz inequality to yield

$$
\|\frac{1}{T}\sum_{t=1}^T x_{it}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)\| \leq \left(\frac{1}{T}\sum_{t=1}^T x_{it}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^T \|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2\right)^{1/2} = O_p(\omega_{p,T}).
$$

For the second term, by using the Cauchy–Schwarz inequality again, we have

$$
\|\frac{1}{T}\sum_{t=1}^T \mathbf{H}\mathbf{f}_t u_{it}\| \leq \|\mathbf{H}\| \cdot \|\frac{1}{T}\sum_{t=1}^T \mathbf{f}_t u_{it}\|.
$$

To bound $\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|^2$, we use the simple fact that $(a+b)^2 \leq 2(a^2+b^2)$. Therefore we combine

24

two upper bonds to obtain that

$$\|\widehat{\mathbf{B}} - \mathbf{BH}'\|^2 \leq 2\sum_{i=1}^{p}\left(\|\frac{1}{T}\sum_{t=1}^{T}x_{it}(\widehat{\mathbf{f}}_t - \mathbf{Hf}_t)\|^2 + \|\frac{1}{T}\sum_{t=1}^{T}\mathbf{Hf}_t u_{it}\|^2\right)$$

$$\leq 2\sum_{i=1}^{p}\left(O_p(\omega_{p,T}^2) + O_p(1)\cdot(T^{-1}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t u_{it}\|^2)\right).$$

To bound $\sum_{i=1}^{p}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t u_{it}\|^2$, as pointed out in the proof of Lemma 4 of Bai and Ng (2002), we only need to prove that $E(\frac{1}{p}\sum_{i=1}^{p}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t u_{it}\|^2) \leq M_1$ for some positive constant M_1. Next, we use the Cauchy–Schwarz inequality and the independence between $\{\mathbf{f}_t\}_{t\geq 1}$ and $\{\mathbf{u}_t\}_{t\geq 1}$ to obtain

$$E(\frac{1}{p}\sum_{i=1}^{p}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t u_{it}\|^2) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E(u_{is}u_{it})E(\mathbf{f}_s'\mathbf{f}_t)$$

$$\leq \frac{1}{pT}\sum_{i=1}^{p}\sum_{s=1}^{T}\sum_{t=1}^{T}|E(u_{is}u_{it})|\cdot\sqrt{E\|\mathbf{f}_s\|^2 E\|\mathbf{f}_t\|^2}$$

$$\leq M\cdot C \;(\equiv M_1)$$

where $E\|\mathbf{f}_s\|^2 \leq C$ holds due to Assumption 3.2, and $\frac{1}{pT}\sum_{i=1}^{p}\sum_{s=1}^{T}\sum_{t=1}^{T}|E(u_{is}u_{it})| \leq M$ holds due to Assumption 3.3(2). Then, we can follow Bai and Ng (2002) to show that

$$\sum_{i=1}^{p}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t u_{it}\|^2 = O_p(p).$$

Now, it immediately follows from (6.2) that

$$\|\widehat{\mathbf{B}} - \mathbf{BH}'\|^2 = O_p(p\omega_{p,T}^2) + O_p(p/T) = O_p(p\omega_{p,T}^2).$$

(b) Begin by noting that

$$\|\widehat{\mathbf{B}}\| \leq \|\widehat{\mathbf{B}} - \mathbf{B}\| + \|\mathbf{B}\| = O_p(p^{\frac{1}{2}}\omega_{p,T}) + O_p(p^{1/2}) = O_p(p^{1/2}).$$

We use the triangle inequality to bound

$$\|\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{B}'\mathbf{B}\| \leq \|\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{B}'\widehat{\mathbf{B}}\| + \|\mathbf{B}'\widehat{\mathbf{B}} - \mathbf{B}'\mathbf{B}\|$$

$$\leq \|\widehat{\mathbf{B}} - \mathbf{B}\|\cdot\|\widehat{\mathbf{B}}\| + \|\mathbf{B}'\|\cdot\|\widehat{\mathbf{B}} - \mathbf{B}\|$$

Note that $\|\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{B}'\mathbf{B}\| = O_p(p\omega_{p,T}) = o_p(p)$ and $\lambda_{\min}(\mathbf{B}'\mathbf{B}) > c_1 p$ hold by Assumption 3.1. Now, we use Lemma 6.1 to obtain

$$\|(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1} - (\mathbf{B}'\mathbf{B})^{-1}\| = O_p(p^{-2})\cdot\|\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{B}'\mathbf{B}\| = O_p(p^{-1}\omega_{p,T}).$$

It follows that

$$
\begin{aligned}
\|\widehat{\boldsymbol{\Lambda}}_b - \boldsymbol{\Lambda}_b\| & = \|(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}' - (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\| \\
& = \|(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}' - (\mathbf{B}'\mathbf{B})^{-1}\widehat{\mathbf{B}}'\| + \|(\mathbf{B}'\mathbf{B})^{-1}\widehat{\mathbf{B}}' - (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\| \\
& \leq \|(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1} - (\mathbf{B}'\mathbf{B})^{-1}\| \cdot \|\widehat{\mathbf{B}}\| + \|(\mathbf{B}'\mathbf{B})^{-1}\| \cdot \|\widehat{\mathbf{B}}' - \mathbf{B}'\| \\
& = O_p(p^{-1}\omega_{p,T}) \cdot O_p(p^{1/2}) + O_p(p^{-1}) \cdot O_p(p^{1/2}\omega_{p,T}) \\
& = O_p(p^{-1/2}\omega_{p,T}).
\end{aligned}
$$

which completes the proof of Lemma 6.4.

$\square$

The following lemma lays the foundation of inverse regression, which can be found in Li (1991).

Lemma 6.5. *Under model (2.1) and Assumption 3.1 (3), the centered inverse regression curve* $E(\mathbf{f}_t|y_{t+1}) - E(\mathbf{f}_t)$ *is contained in the linear subspace spanned by* $\boldsymbol{\phi}_k'\mathrm{cov}(\mathbf{f}_t), k = 1, \ldots, L$.

We are now ready to complete the proof of Theorem 3.1.

Proof of Theorem 3.1. In the first part, we derive the probability bound for $\|\widehat{\boldsymbol{\Sigma}}_{f|y} - \boldsymbol{\Sigma}_{f|y}\|$. Let $\widehat{\mathbf{m}}_h = \frac{1}{c}\sum_{l=1}^{c}\mathbf{x}_{(h,l)}$ denote the average of the predictors within the corresponding slice I_h, and $\mathbf{m}_h = E(\mathbf{x}_t|y_{t+1} \in I_h)$ represents its population version. By definition, $\boldsymbol{\Sigma}_{f|y} = H^{-1}\sum_{h=1}^{H}(\boldsymbol{\Lambda}_b\mathbf{m}_h)(\boldsymbol{\Lambda}_b\mathbf{m}_h)'$. For fixed H, note that

$$
\widehat{\boldsymbol{\Sigma}}_{f|y} - \boldsymbol{\Sigma}_{f|y} = H^{-1}\sum_{h=1}^{H}[(\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h)(\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h)' - (\boldsymbol{\Lambda}_b\mathbf{m}_h)(\boldsymbol{\Lambda}_b\mathbf{m}_h)'].
$$

Thus,

$$
\|\widehat{\boldsymbol{\Sigma}}_{f|y} - \boldsymbol{\Sigma}_{f|y}\| \leq H^{-1}\sum_{h=1}^{H}\left(\|\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h - \boldsymbol{\Lambda}_b\mathbf{m}_h\| \cdot \|\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h\| + \|\boldsymbol{\Lambda}_b\mathbf{m}_h\| \cdot \|\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h - \boldsymbol{\Lambda}_b\mathbf{m}_h\|\right) \tag{6.4}
$$

We now bound each of the above terms.

From the definition, we immediately 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, 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 $O_p(T^{-1/2})$. This holds true as the random variable $\mathbf{f}_t|y_{t+1} \in I_h$ is still stationary with finite second moments, and the sum of the α-mixing coefficients converges (Theorem 19.2 in Billingsley (1999)). This applies to $\mathbf{u}_t|y_{t+1} \in I_h$ as well. Therefore,

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

In addition to the inequality above, we also have

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

26

Thus, it follows that
$$\|\widehat{\mathbf{m}}_h\| \le \|\widehat{\mathbf{m}}_h - \mathbf{m}_h\| + \|\mathbf{m}_h\| = O_p(p^{1/2}).$$

Now, we use both the triangle inequality and the Cauchy–Schwarz inequality to obtain that

$$
\begin{aligned}
\|\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h - \mathbf{\Lambda}_b\mathbf{m}_h\| &\le \|\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h - \mathbf{\Lambda}_b\widehat{\mathbf{m}}_h\| + \|\mathbf{\Lambda}_b\widehat{\mathbf{m}}_h - \mathbf{\Lambda}_b\mathbf{m}_h\| \\
&\le \|\widehat{\mathbf{\Lambda}}_b - \mathbf{\Lambda}_b\| \cdot \|\widehat{\mathbf{m}}_h\| + \|\mathbf{\Lambda}_b\| \cdot \|\widehat{\mathbf{m}}_h - \mathbf{m}_h\| \\
&= O_p(p^{-1/2}\omega_{p,T}) \cdot O_p(p^{1/2}) + O_p(p^{-1/2}) \cdot O_p(\sqrt{p/T}) \\
&= O_p(\omega_{p,T}),
\end{aligned}
$$

where we use $\|\widehat{\mathbf{\Lambda}}_b - \mathbf{H}\mathbf{\Lambda}_b\| = O_p(p^{-1/2}\omega_{p,T})$ by Lemma 6.4 in the second inequality, and that $\|\mathbf{\Lambda}_b\| \le \|(\mathbf{B}'\mathbf{B})^{-1}\| \cdot \|\mathbf{B}\| = O_p(p^{-1/2})$ in the third equality.

Also we have that

$$\|\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h\| \le (\|\widehat{\mathbf{\Lambda}}_b - \mathbf{\Lambda}_b\| + \|\mathbf{\Lambda}_b\|) \cdot \|\widehat{\mathbf{m}}_h\| = O_p(p^{-1/2}) \cdot O_p(p^{1/2}) = O_p(1)$$

and

$$\|\mathbf{\Lambda}_b\mathbf{m}_h\| \le \|\mathbf{\Lambda}_b\| \cdot \|\mathbf{m}_h\| = O_p(1).$$

Hence, from (6.4), we conclude the desired result as follows:

$$\|\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{\Sigma}_{f|y}\| = H^{-1}\sum_{h=1}^{H}(O_p(\omega_{p,T}) \cdot O_p(1) + O_p(1) \cdot O_p(\omega_{p,T})) = O_p(\omega_{p,T}).$$

In the second part, we show that $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L$ are the desired SDR directions and further derive the probability bound for $\|\widehat{\boldsymbol{\psi}}_j - \boldsymbol{\psi}_j\|$ with $j = 1, \ldots, L$. Let $\{\lambda_i\}_{i=1}^{K}$ be the eigenvalues of $\mathbf{\Sigma}_{f|y}$ in the descending order and $\{\widehat{\lambda}_i\}_{i=1}^{K}$ be the eigenvalues of $\widehat{\mathbf{\Sigma}}_{f|y}$ in the descending order. Let $\widehat{\boldsymbol{\psi}}_j$ be the eigenvector associated with the eigenvalue $\widehat{\lambda}_j$, and $\boldsymbol{\psi}_j$ is the eigenvector associated with the eigenvalue λ_j.

In view of Lemma 6.5, $E(\mathbf{f}_t|y_{t+1})$ is contained in the central subspace $S_{y|\mathbf{f}}$ spanned by $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L$, since we have the normalization $\text{cov}(\mathbf{f}_t) = \mathbf{I}_K$ and $E(\mathbf{f}_t) = \mathbf{0}$. Recall the fact that the eigenvectors of $\mathbf{\Sigma}_{f|y}$ corresponding to its L largest eigenvalues would form another orthogonal basis for $S_{y|\mathbf{f}}$. Hence, the eigenvectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L$ of $\mathbf{\Sigma}_{f|y}$ constitute the SDR directions for model (2.1).

Note that $|\widehat{\lambda}_{j-1} - \lambda_{j-1}| \le \|\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{\Sigma}_{f|y}\|$ holds by Weyl's Theorem in Lemma 6.2 (a). Similarly, we have $|\lambda_j - \widehat{\lambda}_{j+1}| \ge |\lambda_j - \lambda_{j+1}| - O_p(\omega_{p,T})$. Then, we have

$$
\begin{aligned}
|\widehat{\lambda}_{j-1} - \lambda_j| &\ge |\lambda_{j-1} - \lambda_j| - |\widehat{\lambda}_{j-1} - \lambda_{j-1}| \\
&\ge |\lambda_{j-1} - \lambda_j| - O_p(\omega_{p,T}),
\end{aligned}
$$

where the fact that $|\widehat{\lambda}_{j-1} - \lambda_{j-1}| \le \|\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{\Sigma}_{f|y}\|$ is used.

Since $\mathbf{\Sigma}_{f|y}$ have distinct eigenvalues, both $|\widehat{\lambda}_{j-1} - \lambda_j|$ and $|\lambda_j - \widehat{\lambda}_{j+1}|$ are bounded away from zero with probability tending to one for $j = 1, \ldots, L$. Now, a direct application of $\sin(\theta)$ Theorem (Davis and Kahan, 1970) in Lemma 6.2 (b) shows that

$$\|\widehat{\boldsymbol{\psi}}_j - \boldsymbol{\psi}_j\| \le \frac{\sqrt{2} \cdot \|\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{\Sigma}_{f|y}\|}{\min(|\widehat{\lambda}_{j-1} - \lambda_j|, |\lambda_j - \widehat{\lambda}_{j+1}|)} = O_p(1) \cdot O_p(\omega_{p,T}) = O_p(\omega_{p,T}),$$

where $j = 1, \ldots, L$,

Therefore, we complete the proof of Theorem 3.1.

$\square$

6.3 Proof of Theorem 3.2

Proof. First we write $\widehat{\phi} = \frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}\widehat{\mathbf{f}}_t$ in terms of the true factors $\mathbf{f}_t$, i.e.,

$$
\begin{aligned}
\widehat{\phi} &= \frac{\widehat{\mathbf{\Lambda}}_b}{T-1}\sum_{t=1}^{T-1} y_{t+1}\mathbf{x}_t \\
&= \frac{\widehat{\mathbf{\Lambda}}_b}{T-1}\sum_{t=1}^{T-1}(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t)y_{t+1},
\end{aligned}
$$

where we used the fact that $\widehat{\mathbf{f}}_t = \widehat{\mathbf{\Lambda}}_b\mathbf{x}_t$ as shown in Proposition 2.1.

Using the triangular inequality, we have

$$
\begin{aligned}
\|\widehat{\phi} - \bar{\phi}\| &\leq \|\widehat{\phi} - (\widehat{\mathbf{\Lambda}}_b\mathbf{B})\bar{\phi}\| + \|(\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I}_K)\bar{\phi}\| \\
&= \|\frac{(\widehat{\mathbf{\Lambda}}_b\mathbf{B})}{T-1}\sum_{t=1}^{T-1}\mathbf{f}_t y_{t+1} + \frac{\widehat{\mathbf{\Lambda}}_b}{T-1}\sum_{t=1}^{T-1}\mathbf{u}_t y_{t+1} - (\widehat{\mathbf{\Lambda}}_b\mathbf{B})\bar{\phi}\| + \|(\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I}_K)\bar{\phi}\| \\
&\leq \|\frac{(\widehat{\mathbf{\Lambda}}_b\mathbf{B})}{T-1}\sum_{t=1}^{T-1}(y_{t+1}\mathbf{f}_t - \bar{\phi})\| + \|\frac{1}{T-1}\widehat{\mathbf{\Lambda}}_b\sum_{t=1}^{T-1}\mathbf{u}_t y_{t+1}\| + \|(\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I}_K)\bar{\phi}\|.
\end{aligned}
$$

In the sequel, we will bound three terms on the right hand side of this inequality respectively.

By Lemmas 6.3 and 6.4, we have

$$
\|\widehat{\mathbf{\Lambda}}_b\mathbf{B}\| \leq (\|\widehat{\mathbf{\Lambda}}_b - \mathbf{\Lambda}_b\| + \|\mathbf{\Lambda}_b\|) \cdot \|\mathbf{B}\| = O_p(1)
$$

and

$$
\|\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I}_K\| = \|\widehat{\mathbf{\Lambda}}_b - \mathbf{\Lambda}_b\| \cdot \|\mathbf{B}\| = O_p(\omega_{p,T}).
$$

Since $\|\bar{\phi}\| = O_p(1)$, the third term on the right hand side of the inequality is $O_p(\omega_{p,T})$, i.e.,

$$
\|(\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I}_K)\bar{\phi}\| = O_p(\omega_{p,T}).
$$

For the second term, note that $\mathbf{u}_t$ is independent of y_{t+1}, hence $E(\mathbf{u}_t y_{t+1}) = 0$. By law of large numbers and $\|\widehat{\mathbf{\Lambda}}_b\| = O_p(p^{-1/2})$, we have

$$
\|\frac{1}{T-1}\widehat{\mathbf{\Lambda}}_b\sum_{t=1}^{T-1}\mathbf{u}_t y_{t+1}\| = O_p(p^{-1/2}) \cdot O_p(T^{-1/2}) = O_p(\omega_{p,T}).
$$

It remains to bound the first term $\|\frac{1}{T-1}\sum_{t=1}^{T-1}(y_{t+1}\mathbf{f}_t - \bar{\phi})\|$. To this end, we express $\mathbf{f}_t$ along the basis $\phi_1, \cdots, \phi_L$ and their orthogonal hyperplane, namely,

$$
\mathbf{f}_t = \sum_{j=1}^{L}\langle \mathbf{f}_t, \phi_j\rangle\phi_j + \mathbf{f}_t^{\perp}.
$$

By the orthogonal decomposition of normal distribution, $\langle \mathbf{f}_t, \phi_j\rangle$ and $\mathbf{f}_t^{\perp}$ are independent. In

28

addition, y_{t+1} depends on $\mathbf{f}_t$ only through $\phi_1'\mathbf{f}_t, \cdots, \phi_L'\mathbf{f}_t$, and then, it is conditionally independent of $\mathbf{f}_t^\perp$. It follows from contraction property that y_{t+1} and $\mathbf{f}_t^\perp$ are independent, unconditionally. Thus,

$$E(y_{t+1}\mathbf{f}_t^\perp) = E(y_{t+1})E(\mathbf{f}_t^\perp) = \mathbf{0}.$$

Recall that $\bar{\phi} = \sum_{j=1}^L E((\phi_j'\mathbf{f}_t)y_{t+1})\phi_j$. Now, we use the triangle inequality to obtain

$$
\begin{aligned}
&\|\frac{1}{T-1}\sum_{t=1}^{T-1}(y_{t+1}\mathbf{f}_t - \bar{\phi})\| \\
=\ &\|\frac{1}{T-1}\sum_{t=1}^{T-1}[\sum_{j=1}^L(\phi_j'\mathbf{f}_t)y_{t+1}\phi_j + y_{t+1}\mathbf{f}_t^\perp - \sum_{j=1}^L E((\phi_j'\mathbf{f}_t)y_{t+1})\phi_j]\| \\
\leq\ &\sum_{j=1}^L\|\frac{1}{T-1}\sum_{t=1}^{T-1}[(\phi_j'\mathbf{f}_t)y_{t+1} - E((\phi_j'\mathbf{f}_t)y_{t+1})]\phi_j\| + \|\frac{1}{T-1}\sum_{t=1}^{T-1}y_{t+1}\mathbf{f}_t^\perp\| \\
\leq\ &\sum_{j=1}^L|\frac{1}{T-1}\sum_{t=1}^{T-1}[(\phi_j'\mathbf{f}_t)y_{t+1} - E((\phi_j'\mathbf{f}_t)y_{t+1})]| + \|\frac{1}{T-1}\sum_{t=1}^{T-1}y_{t+1}\mathbf{f}_t^\perp\|,
\end{aligned}
$$

where we used the fact that $\|\phi_j\| = 1$ for any j. Note that each term here is $O_p(T^{-1/2})$ by Law of Large Numbers. Then, we have,

$$\|\frac{(\widehat{\boldsymbol{\Lambda}}_b\mathbf{B})}{T-1}(\sum_{t=1}^{T-1}y_{t+1}\mathbf{f}_t - \bar{\phi})\| \leq \|\widehat{\boldsymbol{\Lambda}}_b\mathbf{B}\| \cdot O_p(T^{-\frac{1}{2}}) = O_p(\omega_{p,T}).$$

This concludes the desired bound that $\|\widehat{\phi} - \bar{\phi}\| = O_p(\omega_{p,T})$.

Therefore, we complete the proof of Theorem 3.2. $\qquad\square$

References

Ahn, S. C. and Horenstein, A. R. (2013), 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**(3), 1203–1227.

Alessi, L., Barigozzi, M. and Capasso, M. (2010), 'Improved penalization for determining the number of factors in approximate factor models', *Statistics & Probability Letters* **80**(23), 1806–1813.

Athreya, K. B. and Lahiri, S. N. (2006), *Measure Theory and Probability Theory*, Springer, New York.

Bai, J. (2003), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**(1), 135–171.

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. (2009), 'Boosting diffusion indices', *Journal of Applied Econometrics* **24**(4), 607–629.

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**(473), 119–137.

Bengio, Y. (2009), 'Learning deep architectures for ai', *Foundations and Trends in Machine Learning* **2**(1), 1–127.

Bengio, Y., Courville, A. and Vincent, P. (2013), 'Representation learning: A review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1798–1828.

Bernanke, B. S., Boivin, J. and Eliasz, P. (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach', *The Quarterly Journal of Economics* **120**(1), 387–422.

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

Boivin, J. and Ng, S. (2006), 'Are more data always better for factor analysis?', *Journal of Econometrics* **132**(1), 169–194.

Campbell, J. Y. and Shiller, R. J. (1988), 'Stock prices, earnings, and expected dividends', *Journal of Finance* **43**(3), 661–676.

Campbell, J. Y. and Thompson, S. B. (2008), 'Predicting excess stock returns out of sample: Can anything beat the historical average?', *Review of Financial Studies* **21**(4), 1509–1531.

Carhart, M. M. (1997), 'On persistence in mutual fund performance', *Journal of Finance* **52**(1), 57–82.

Connor, G., Hagmann, M. and Linton, O. (2012), 'Efficient semiparametric estimation of the fama–french model and extensions', *Econometrica* **80**(2), 713–754.

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

Cunha, F., Heckman, J. J. and Schennach, S. M. (2010), 'Estimating the technology of cognitive and noncognitive skill formation', *Econometrica* **78**(3), 883–931.

Davis, C. and Kahan, W. M. (1970), 'The rotation of eigenvectors by a perturbation. iii', *SIAM Journal on Numerical Analysis* **7**(1), 1–46.

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

Fama, E. F. and French, K. R. (1992), 'The cross-section of expected stock returns', *Journal of Finance* **47**(2), 427–465.

Fama, E. F. and French, K. R. (2015), 'A five-factor asset pricing model', *Journal of Financial Economics* **116**(1), 1–22.

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., Han, X. and Gu, W. (2012), 'Estimating false discovery proportion under arbitrary covariance dependence (with discussion)', *Journal of the American Statistical Association* **107**(499), 1019–1035.

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.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), 'The generalized dynamic-factor model: identification and estimation', *The Review of Economics and Statistics* **82**(4), 540–554.

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.

Hallin, M. and Liška, R. (2007), 'Determining the number of factors in the general dynamic factor model', *Journal of the American Statistical Association* **102**(478), 603–617.

Hotelling, H. (1957), 'The relations of the newer multivariate statistical methods to factor analysis', *British Journal of Statistical Psychology* **10**(2), 69–79.

Kelly, B. and Pruitt, S. (2013), 'Market expectations in the cross-section of present values', *Journal of Finance* **68**(5), 1721–1756.

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.

Kendall, M. (1957), *A Course in Multioariate Analysis.*, London: Griffin.

Lam, C., Yao, Q. et al. (2012), 'Factor modeling for high-dimensional time series: inference for the number of factors', *The Annals of Statistics* **40**(2), 694–726.

Leek, J. T. and Storey, J. D. (2008), 'A general framework for multiple testing dependence', *Proceedings of the National Academy of Sciences* **105**(48), 18718–18723.

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

Lintner, J. (1965), 'The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets', *The Review of Economics and Statistics* **47**(1), 13–37.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L. et al. (2012), 'High-dimensional semiparametric gaussian copula graphical models', *The Annals of Statistics* **40**(4), 2293–2326.

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

Ludvigson, S. and Ng, S. (2009), 'Macro factors in bond risk premia', *Review of Financial Studies* **22**(12), 5027–5067.

Polk, C., Thompson, S. and Vuolteenaho, T. (2006), 'Cross-sectional forecasts of the equity premium', *Journal of Financial Economics* **81**(1), 101–141.

Rajan, R. G. and Zingales, L. (1998), 'Financial dependence and growth', *American Economic Review* **88**(3), 559–586.

Schott, J. R. (1994), 'Determining the dimensionality in sliced inverse regression', *Journal of the American Statistical Association* **89**(3), 141–148.

Sharpe, W. F. (1964), 'Capital asset prices: a theory of market equilibrium under conditions of risk', *Journal of Finance* **19**(3), 425–442.

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.

Stock, J. H. and Watson, M. W. (2012), 'Generalized shrinkage methods for forecasting using many predictors', *Journal of Business & Economic Statistics* **30**(4), 481–493.

Tjøstheim, D. (1994), 'Nonlinear time series, a selective review', *Scandinavian Journal of Statistics* **21**(2), 97–130.

Xue, L. and Zou, H. (2012), 'Regularized rank-based estimation of high-dimensional nonparanormal graphical models', *The Annals of Statistics* **40**(5), 2541–2571.

Zhu, L., Miao, B. and Peng, H. (2006), 'On sliced inverse regression with high-dimensional covariates', *Journal of the American Statistical Association* **101**(474), 630–643.