JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 1, 342–354
https://doi.org/10.1080/07350015.2020.1813589

Nonparametric Estimation and Conformal Inference of the Sufficient Forecasting With a Diverging Number of Factors

Xiufan Yu, Jiawei Yao, and Lingzhou Xue

Department of Statistics, Pennsylvania State University, University Park, PA

ABSTRACT

The sufficient forecasting (SF) provides a nonparametric procedure to estimate forecasting indices from high-dimensional predictors to forecast a single time series, allowing for the possibly nonlinear forecasting function. This article studies the asymptotic theory of the SF with a diverging number of factors and develops its predictive inference. First, we revisit the SF and explore its connections to Fama–MacBeth regression and partial least squares. Second, with 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 SF directions. Third, we use the local linear regression to estimate the possibly nonlinear forecasting function and obtain the rate of convergence. Fourth, we construct the distribution-free conformal prediction set for the SF that accounts for the serial dependence. Moreover, we demonstrate the finite-sample performance of the proposed nonparametric estimation and conformal inference in simulation studies and a real application to forecast financial time series.

ARTICLE HISTORY
Received October 2019
Accepted August 2020

KEYWORDS
Conformal inference; Factor models; Forecasting; High-dimensional predictors; Nonparametric estimation; Principal components

1. Introduction

Forecasting a single time series using high-dimensional predictors has received a lot of interests in macroeconomics, finance, business, epidemiology, and many other research fields. In a data-rich environment, it is usually reasonable to assume that a few underlying common factors simultaneously drive the forecasting target and the high-dimensional predictors. The use of principal components effectively reduces the dimensionality and more importantly provides a useful characterization of economic predictors.

By assuming the linear forecasting function, Stock and Watson (1989, 2002a, 2002b) demonstrated the validity of the estimated principal components in forecasting. Bai and Ng (2006) conducted inferences on factor-augmented regressions to enable the forecast. Further efforts refine the forecast by filtering out information unrelated to the target based on the predictors or estimated factors. Bair et al. (2006) applied the correlation screening to obtain relevant predictors, and Bai and Ng (2008) established the thresholding criteria to rule out predictors not informative for the target. Kelly and Pruitt (2015) proposed a three-pass regression filter method that follows the philosophy of partial least squares (PLS) and selectively identifies the subset of factors influencing the target while discarding factors that are irrelevant.

However, all of the aforementioned works may not perform well when the target and the latent factors have possibly nonlinear relationship. The possibly nonlinear and nonseparable forecasting function poses a significant challenge when extracting the information relevant to the target. Fan, Xue, and Yao (2017) proposed the sufficient forecasting (SF) procedure to obtain the sufficient predictive indices with provable theoretical guarantees, allowing for an unknown nonlinear forecasting function. In this article, we shall enrich the methodology, theory, and applicability of the SF by understanding the relation between the SF and other methods, exploring the nonparametric estimation of the forecasting function, and conducting the predictive inference for the SF.

First, we revisit the SF and point out its close connections to existing methods in panel data analysis such as the Fama–MacBeth (FM) regression (Fama and MacBeth 1973) and the three-pass regression filter (Kelly and Pruitt 2015). Both methods involve time-series regressions in their first pass. Proposition 2.1 will show that the characterization of the SF is similar in spirit to the FM regression after transforming the inverse regression curves of the forecasting indices using the loading matrix. Also, Proposition 2.2 will show that the SF can be derived as the solution to a constrained problem that includes the PLS as a special example.

Second, we characterize the asymptotic behavior of the estimated predictive directions for the SF with a diverging number of latent factors that increases with sample size. The diverging number of factors avoids possible model misspecification and accommodates potential structural changes (Ludvigson and Ng 2007; Li, Li, and Shi 2017; Luo, Xue, and Yao 2017). In addition, by using the known result that low dimensional projections from high-dimensional predictors is almost linear (Hall and Li 1993), the diverging number of factors provides the necessary guarantee that the linearity condition (Li 1991) approximately holds. Thus, we also relax the restricted linearity condition in the SF that might lead to the undesired time reversibility (Xia

CONTACT Lingzhou Xue *lzxue@psu.edu* Department of Statistics, Pennsylvania State University, University Park, PA 16802.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

et al. 2002). As will be shown in Theorem 3.2, the asymptotic decomposition of the estimated predictive directions illustrates the source of estimation errors in the SF.

Third, built on the asymptotic properties of estimated directions, we study the estimation of the possibly nonlinear and nonseparable forecasting function using the local linear regression (LLR). In the SF, it is important but challenging to provide the theoretical guarantee for the nonparametric estimation of the forecasting function with nonparametrically generated forecasting indices. To this end, we extend the theoretical analysis of Mammen, Rothe, and Schienle (2012) based on independent observations to the panel data setting for the SF. After a careful analysis, we derive the explicit rate of convergence for the nonparametric estimation of the forecasting function. This result fills an important gap in the methodology and theory of the SF (Fan, Xue, and Yao 2017).

Fourth, we introduce the permutation-based conformal inference to construct a prediction band for the SF with provable guarantees. It is very important for real-world applications to conduct the predictive inference without requiring stringent assumptions on the data generation process. Following the conformal inference (Vovk, Gammerman, and Shafer 2005; Lei et al. 2018), we conduct hypothesis tests on a grid of hypothesized values of the target and then calculate valid p-values based on the empirical quantiles of augmented samples. Thus, the prediction set is constructed according to the testing results. However, most of existing methods in the conformal inference are based on independent observations and require the exchangeability condition to guarantee the desired coverage. In this work, we follow the recent work by Chernozhukov, Wuthrich, and Zhu (2018) to conduct the conformal inference for the SF by incorporating blocks in the permutation scheme, which effectively preserves the temporal dependence structure in panel data.

The rest of this article is organized as follows. Section 2 revisits the SF and shows its connections to existing methods. Section 3 studies the asymptotic properties of the forecasting directions and also nonparametric estimation of the forecasting function, and Section 4 presents the conformal inference. Section 5 gives simulation studies and an empirical study to forecast financial time series. Section 6 includes a few concluding remarks. Technical details and additional numerical results are given in the supplementary materials.

2. Revisiting Sufficient Forecasting

We first present an alternative interpretation of the SF in Section 2.1 and then point out its connections to FM regression and PLS in Sections 2.2 and 2.3.

2.1. Sufficient Forecasting

Consider the factor representation of a large number of predictors in the form that

$$x_{it} = \mathbf{b}_i' \mathbf{f}_t + u_{it}, \quad 1 \le i \le p,\ 1 \le t \le T, \tag{2.1}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})'$ are the common factors, $\mathbf{b}_i$ are the corresponding 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 $\mathbf{x}_t = (x_{1t}, \ldots, x_{pt})'$ is the cross-section of $p \times 1$ predictors, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ is the $p \times K$ loading matrix and $\mathbf{u}_t = (u_{1t}, \ldots, u_{pt})'$ is the $p \times 1$ error term. The canonical normalization

$$\operatorname{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.2}$$

where $\mathbf{I}_K$ is a $K \times K$ identity matrix, serves as an identification condition of the loadings and factors. For simplicity, we also assume that $E(\mathbf{f}_t) = \mathbf{0}$ and hence $E(\mathbf{x}_t) = \mathbf{0}$.

The target y_{t+1} depends on the factors only through projected variables $\boldsymbol{\phi}_1' \mathbf{f}_t, \ldots, \boldsymbol{\phi}_L' \mathbf{f}_t$,

$$y_{t+1} = g(\boldsymbol{\phi}_1' \mathbf{f}_t, \ldots, \boldsymbol{\phi}_L' \mathbf{f}_t) + \epsilon_{t+1}, \tag{2.3}$$

where $\boldsymbol{\phi}_i$'s are unknown vectors, $g(\cdot)$ is an unknown link function, and ϵ_{t+1} is some stochastic error independent of $\mathbf{f}_t$ and u_{it}. Denote by $\Phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L)$ a $K \times L$ matrix. We may impose the identification condition $\Phi'\Phi = \mathbf{I}_L$ such that those directions form an orthonormal basis of the central space. Fan, Xue, and Yao (2017) first considered such models and proposed the SF procedure to estimate the directions $\boldsymbol{\phi}_i$'s. If $g(\cdot)$ is linear with $L = 1$, (2.1) and (2.3) constitute the diffusion index forecasting model of Stock and Watson (2002b).

The target y_{t+1} can be either an asset return or a general macroeconomic indicator. While our model is cast as a forecasting problem, the same theory applies when y_{t+1} is replaced by y_t, in which case y_t moves contemporaneously with the factors. When the target is a carefully constructed portfolio and the link function $g(\cdot)$ is linear with $L = 1$, there exists a large literature on constructing the meaningful portfolios, for example, Fama and French (1993), Jagannathan and Wang (1996), Lettau and Ludvigson (2001), and Li, Vassalou, and Xing (2006), to name a few. These proposed portfolios (or factors) can be subsequently used to determine individual asset's beta, a different goal than ours here. When the target is a macroeconomic factor, however, there is typically no guarantee that it will be linearly related to the latent factors.

To extract information from the panel data, SF considers the covariance of conditional expectation of factors, $\operatorname{cov}(E(\mathbf{f}_t|y_{t+1}))$. To obtain forecasting directions $\boldsymbol{\phi}_i$'s, Fan, Xue, and Yao (2017) extracted the estimated factors $\widehat{\mathbf{f}}_t$ from the observable predictors to form an estimator of $\operatorname{cov}(E(\mathbf{f}_t|y_{t+1}))$. A slicing version of the covariance is

$$\Sigma_{f|y} = \frac{1}{H} \sum_{h=1}^{H} E(\mathbf{f}_t | y_{t+1} \in I_h) E(\mathbf{f}_t' | y_{t+1} \in I_h), \tag{2.4}$$

where $H \ge L$ 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 do not require $H \to \infty$. As discussed in Li (1991) and Fan, Xue, and Yao (2017), any direction orthogonal to $E(\mathbf{f}_t|y_{t+1})$ is also orthogonal to $\Sigma_{f|y}$. Hence, the central curve $E(\mathbf{f}_t|y_{t+1})$ is confined in the eigenspace of $\Sigma_{f|y}$. Suppose we take one step further and have the following *coverage condition*:

$$\langle \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L \rangle = \langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L \rangle, \tag{2.5}$$

where $\boldsymbol{\psi}_i$'s are the eigenvectors of $\Sigma_{f|y}$ corresponding to its L leading positive eigenvalues. The notation $\langle \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L \rangle$

denotes the subspace spanned by vectors $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L\}$, that is, $\langle \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L \rangle = \left\{ \sum_{i=1}^{L} c_i \boldsymbol{\phi}_i, c_i \in \mathbb{R} \right\}$. The coverage condition ensures that the subspace spanned by the inverse conditional mean $E(\mathbf{f}_t|y_{t+1})$ coincides with the central space spanned by $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L\}$. This condition is common in the sufficient dimension reduction literature (e.g., Chiaromonte, Cook, and Li 2002; Cook 2004; Zhu, Miao, and Peng 2006; Fan, Xue, and Yao 2017; Lin, Zhao, and Liu 2018). Consequently, one would recover the column space Φ by exploiting the eigenvectors of $\Sigma_{f|y}$. Recall that $\boldsymbol{\phi}_j$ in (2.3) is only identifiable up to a transformation, but the column space formed by Φ could be identified. Since the SF pursues $\Sigma_{f|y}$, we assume for simplicity that Φ actually consists of the eigenvectors of $\Sigma_{f|y}$. Note that condition (2.5) imposes an implicit condition on the link function $g(\cdot)$ in (2.3), for example, the condition fails if $g(\cdot)$ is symmetric. We leave the relaxation of this condition for future study.

The SF 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, $y_{(h,j)} = y_{(c(h-1)+j+1)}$ and $\widehat{\mathbf{f}}_{(h,j)} = \widehat{\mathbf{f}}_{(c(h-1)+j)}$ for $h = 1, \ldots, H$ and $j = 1, \ldots, c$. The estimator for (2.4) is

$$\widehat{\Sigma}_{f|y} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\boldsymbol{\xi}}_h \widehat{\boldsymbol{\xi}}_h', \tag{2.6}$$

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 $\widehat{\Sigma}_{f|y}$, we obtain the estimated forecasting directions $\widehat{\Phi}$.

2.2. Connection to FM Regression

The SF studies 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 $\mathbb{R}^K$,

$$E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t) = \sum_{i=1}^{L} c_i \boldsymbol{\phi}_i'\mathbf{f}_t \tag{2.7}$$

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

Proposition 2.1. Under (2.1)–(2.7), we have $E(\mathbf{x}_t|y_{t+1}) = \mathbf{B}\Phi\boldsymbol{\gamma}(y_{t+1})$, where the $L \times 1$ vector $\boldsymbol{\gamma}(y)$ consists of the inverse regression curves of the forecasting indices, $\boldsymbol{\gamma}(y_{t+1}) = E(\Phi'\mathbf{f}_t|y_{t+1}) = E((\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)'|y_{t+1})$.

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 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; 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 $E(\mathbf{x}_t y_{t+1}) = E(E(\mathbf{x}_t|y_{t+1})y_{t+1}) = \mathbf{B}\Phi E(\boldsymbol{\gamma}(y_{t+1})y_{t+1})$, as $\mathbf{x}_t$ has been demeaned. Their results hence could only recover an average $\bar{\boldsymbol{\phi}}$ of the true directions, where $\bar{\boldsymbol{\phi}} = \Phi E(\boldsymbol{\gamma}(y_{t+1})y_{t+1}) = \sum_{i=1}^{L} E((\boldsymbol{\phi}_i'\mathbf{f}_t)y_{t+1})\boldsymbol{\phi}_i$. Fan, Xue, and Yao (2017) observed the same fact in the comparison between SF and principal component regression (PCR). Here, the benefit of using $E(\mathbf{x}_t|y_{t+1})$ is more clear.

2.3. Connection to Partial Least Squares

The ultimate goal of many forecasting problems translates into finding some predictive coefficient $\boldsymbol{\zeta}$ on individual predictors. On population level, SF 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 $\boldsymbol{\zeta}_i$ on the original predictors $\mathbf{x}_t$ via $\boldsymbol{\zeta}_i = \Lambda_b'\boldsymbol{\phi}_i$, where $\Lambda_b = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ only involves the loading matrix. One may observe that such $\boldsymbol{\zeta}_i$'s reside in the column space of the loading matrix $\mathbf{B}$. Had we obtained a direction $\widetilde{\boldsymbol{\zeta}}$ orthogonal to the column space of $\mathbf{B}$, the predictive index $\widetilde{\boldsymbol{\zeta}}'\mathbf{x}_t = \widetilde{\boldsymbol{\zeta}}'(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t) = \widetilde{\boldsymbol{\zeta}}'\mathbf{u}_t$ would be completely irrelevant to the target. This can also be understood as mitigating the impact of irrelevant factors if the idiosyncratic term 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 SF can be derived as the solution to the following constrained optimization.

Proposition 2.2. On population level, the ith SF predictive coefficient on the observed predictor $\mathbf{x}_t$ solves

$$\max_{\boldsymbol{\zeta}} \max_{\mathcal{T}(\cdot)} \text{corr}^2(\mathcal{T}(y_{t+1}), \boldsymbol{\zeta}'\mathbf{x}_t) \tag{2.8}$$

$$\text{subject to} \quad (\mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}')\boldsymbol{\zeta} = \mathbf{0} \tag{2.9}$$

$$\text{and} \quad \boldsymbol{\zeta}'\mathbf{B}\mathbf{B}'\boldsymbol{\zeta} = 1,\ \boldsymbol{\zeta}'\mathbf{B}\mathbf{B}'\boldsymbol{\zeta}_l = 0,\ l = 1, \ldots, i-1,$$

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

Remark 2.1. Proposition 2.2 shows that the direction $\boldsymbol{\zeta}$ 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) had a similar interpretation of their 3PRF model, but they resorted to proxies to determine relevant factor space. Our approach is more general, as it involves a general transformation $\mathcal{T}(\cdot)$ of the forecast target.

The characterization above also allows us to reveal its close connections to the PLS method. Note that the ith PLS direction solves

$$\max_{\boldsymbol{\zeta}} \text{corr}^2(y_{t+1}, \boldsymbol{\zeta}'\mathbf{x}_t)\text{cov}(\boldsymbol{\zeta}'\mathbf{x}_t) \tag{2.10}$$

$$\text{s.t.} \quad \boldsymbol{\zeta}'\boldsymbol{\zeta} = 1,\ \boldsymbol{\zeta}'\mathbf{S}\boldsymbol{\zeta}_l = 0,\ l = 1, \ldots, i-1,$$

where $\mathbf{S} = \text{cov}(\mathbf{x}_t)$; see Frank and Friedman (1993). Barbarino and Bura (2015, 2017) presented a connection of the sliced inverse regression (SIR) to PLS in a somewhat different paradigm. Rather than resorting to the latent factors $E(\mathbf{f}_t|y_{t+1})$

as an intermediate step for dimension reduction, they introduce an alternative approach by directly performing the SIR on top of the conditional mean of predictors $E(\mathbf{x}_t|y_{t+1})$. Their obtained predictive coefficients $\boldsymbol{\zeta}_i$ is shown to be the solution to the following maximization problem.

$$\max_{\boldsymbol{\zeta}} \text{corr}^2(E(\boldsymbol{\zeta}'\mathbf{x}_t|y_{t+1}), \boldsymbol{\zeta}'\mathbf{x}_t) \quad (2.11)$$
$$\text{s.t.} \quad \boldsymbol{\zeta}'\mathbf{S}\boldsymbol{\zeta} = 1,\ \boldsymbol{\zeta}'\mathbf{S}\boldsymbol{\zeta}_l = 0,\ l = 1, \ldots, i-1.$$

Since $\mathbf{S} = \mathbf{B}\mathbf{B}' + \Sigma_u$ in our setup, where $\Sigma_u = \text{cov}(\mathbf{u}_t)$, additional forecasting directions obtained from PLS or SIR on $E(\mathbf{x}_t|y_{t+1})$ would be mixed with undesired noises. The constraint $\boldsymbol{\zeta}'\mathbf{B}\mathbf{B}'\boldsymbol{\zeta} = 1$ in the constraint (2.9) pertains to the normalization of the forecasting directions $\boldsymbol{\phi}_i$'s on the latent factor, and is therefore inconsequential.

The comparison of (2.8) with (2.10) and (2.11) not only suggests the inherent connections of SF with PLS and SIR, but also implies the benefits brought by employing $E(\mathbf{f}_t|y_{t+1})$. Taking advantages of the latent factor structure, our approach precludes the undesirable effect of noises $\mathbf{u}_t$ on forecasting directions $\boldsymbol{\zeta}$. In addition, our approach is more flexible and gracefully handles nonlinearity with the presence of $\mathcal{T}(\cdot)$ in the optimization.

3. Nonparametric Estimation

After presenting the assumptions in Section 3.1, we present the asymptotic properties of the forecasting directions in Sections 3.2 and 3.3 and then study the nonparametric estimation of the forecasting function in Section 3.4.

3.1. Assumptions

To begin with, we provide a set of technical conditions for the consistent estimation of SF directions.

Assumption 3.1 (Factors and loadings).

(i) There exists $b > 0$ such that $\sup_{p\in\mathbb{N}} ||\mathbf{B}||_{\max} \le b$, 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, where $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$.

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 $\{(\mathbf{f}_t, \mathbf{u}_t, \epsilon_{t+1}) : t \le 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t, \epsilon_{t+1}) : t \ge T\}$, 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 (Data-generating process). $\{\mathbf{f}_t\}_{t\ge1}$, $\{\mathbf{u}_t\}_{t\ge1}$ and $\{\epsilon_{t+1}\}_{t\ge1}$ 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 $E(\exp(l|\epsilon_{t+1}|)) \le C$ for any $t \ge 1$.

(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, $\alpha(T) \le exp(-cT^{\gamma_1})$ for all $T \in \mathbb{Z}^+$ and some positive constants γ_1 and c.

Assumption 3.3 (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 $E|u_{it}|^8 \le M$.

(ii) $||\Sigma_u||_1 \le M$, and for every $i, j, t, s > 0$, $(pT)^{-1}\sum_{i,j,t,s} |E(u_{it}u_{js})| \le M$.

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

3.2. Asymptotics with Diverging K

In this subsection, 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 accommodates the potential structural changes (Ludvigson and Ng 2007). Specifically, the estimation of factors is based on the asymptotic principal components as follows:

$$(\widehat{\mathbf{B}}_K, \widehat{\mathbf{F}}_K) = \arg\min_{(\mathbf{B},\mathbf{F})} ||\mathbf{X} - \mathbf{B}\mathbf{F}'||_F^2, \quad (3.1)$$
$$\text{subject to } T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T), \mathbf{F}' = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$, 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 (2.2). 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 presented in the following proposition.

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

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

Next, suppose $K = o(\min\{p^{1/3}, T^{1/2}\})$, 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 (2.6) with the following accuracy

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

Remark 3.1. Part (2) of our Proposition 3.1 is in a similar vein to Proposition 3.2 of Li, Li, and Shi (2017). Note that our approach allows $K = o(\min\{p^{1/3}, T\})$ to achieve the same error rate for estimating the factors as Li, Li, and Shi (2017), which requires $K = o(\min\{p^{1/17}, T^{1/16}\})$. Also, we prove the asymptotic properties of the estimated factor loadings $\widehat{\mathbf{B}}$ and

estimated conditional expectation $\widehat{\boldsymbol{\xi}}_h = \widehat{E}(\mathbf{f}_t|y_{t+1} \in I_h)$, which were not studied by Li, Li, and Shi (2017).

Proposition 3.1 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.

3.3. 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 $\widehat{\Phi}$, without requiring knowledge of the procedure of factor estimation. Let $\Xi = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H)$ be the collection of the sliced regression curves and $\widehat{\Xi} = (\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_H)$ be the corresponding estimate. It is straightforward to translate (2.4) and (2.6) into matrix notation: $\Sigma_{f|y} = H^{-1}\Xi\Xi'$ and $\widehat{\Sigma}_{f|y} = H^{-1}\widehat{\Xi}\widehat{\Xi}'$. Denote by $\Delta = \widehat{\Xi} - \Xi$ the difference between the estimated and true regression curves. We make the following high-level assumption for Δ.

Assumption 3.4. There is a positive sequence $\omega_{p,T,K} = o(1)$ such that $||\Delta|| = O_p(\omega_{p,T,K})$.

The estimation accuracy of Δ 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 (3.1).

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

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

each row of which depicts the projections of SIR functions on different forecasting directions.

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

Theorem 3.1. Suppose Γ is of full rank, the coverage condition (2.5) holds, and the L largest eigenvalues of $\Sigma_{f|y}$ are positive and distinct. Under Assumptions 3.1–3.4 and the linearity condition that $E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)$ is a linear function of $\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t$ for any $\mathbf{b} \in \mathbb{R}^p$, the SF direction estimates $\widehat{\Phi}$ have the following approximation,

$$\widehat{\Phi} = \Phi + (\mathbf{I} - \Phi\Phi')\Delta\Gamma(\Gamma'\Gamma)^{-1} + o_p(\omega_{p,T,K}). \tag{3.2}$$

The proof given in the supplementary materials sheds light on how such decomposition is possible. A direct implication of Theorem 3.1 is that the estimated forecasting directions are consistent, that is, $\widehat{\Phi} = \Phi + O_p(\omega_{p,T,K})$. This result generalizes the findings in Theorem 3.1 of Fan, Xue, and Yao (2017), and provides finer details of the estimation quality.

Theorem 3.1 depends on the connection between the column space of $\Sigma_{f|y}$ and the true forecasting directions given in Φ, 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 et al. (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 SF direction estimates $\widehat{\Phi}$ 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 Υ be an orthonormal basis of the orthogonal complement of the central subspace. For $0 < c < 1$, let $B(c) = \{\mathbf{f}_t : ||\mathbf{f}_t||^2 \le K(1-c)\}$ 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 3.5. The factors $\mathbf{f}_t$ satisfy that

(i) The conditional covariance $\text{cov}(\mathbf{f}_t|\Phi'\mathbf{f}_t)$ is degenerate.
(ii) $P(|K^{-1}||\mathbf{f}_t||^2 - 1| \ge c) = o(K^{-1})$ and $E\{K||\mathbf{f}_t||^{-2}I(B(c))\} = o(K^{-1})$ for any $0 < c < 1$.
(iii) $E\{\sup_{|\sin(\mathbf{f}_t,\mathbf{e})|\le c}\gamma(\mathbf{e})\} = o(K^{-1/2}c^{-K})$ for some $0 < c \le 1$.
(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 $||E(\Upsilon'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)||^4 \le u(\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)$ almost surely.

It is worth pointing out that Assumption 3.5 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 3.5, the following theorem provides an important consistency result for the SF without requiring the linearity condition.

Theorem 3.2. Suppose Γ is of full rank, the coverage condition (2.5) holds, and the L largest eigenvalues of $\Sigma_{f|y}$ are positive and distinct. Under Assumptions 3.1–3.5, the SF direction estimates $\widehat{\Phi}$ have the following approximation,

$$\widehat{\Phi} = \Phi + (\mathbf{I} - \Phi\Phi')\Delta\Gamma(\Gamma'\Gamma)^{-1} + o_p(1). \tag{3.3}$$

3.4. Estimation of the Forecasting Function

The nonparametric regression model (2.3) can be written as

$$y_{t+1} = g(\boldsymbol{r}(\mathbf{x}_t)) + \epsilon_{t+1}, \quad \text{with } \mathbb{E}(\epsilon_{t+1}) = 0 \tag{3.4}$$

where $\boldsymbol{r}(\mathbf{x}_t) = (\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t)'$ denotes L-dimensional predictive indices extracted from p-dimensional predictors $\mathbf{x}_t$. Our goal now is to provide a nonparametric estimation of the unknown forecasting function $g(\mathbf{z}) = \mathbb{E}(y_{t+1}|\boldsymbol{r}(\mathbf{x}_t) = \mathbf{z})$ given the independence of $\{\mathbf{f}_t\}$ and $\{\epsilon_{t+1}\}$. Note that $\boldsymbol{r}(\mathbf{x}_t)$ can be consistently estimated by $\widehat{\boldsymbol{r}}(\mathbf{x}_t) = (\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\phi}}_L'\widehat{\mathbf{f}}_t)$ using the SF

procedure, as established in previous section. For ease of notation, we define $\boldsymbol{r}_t = \boldsymbol{r}(\mathbf{x}_t)$ and $\widehat{\boldsymbol{r}}_t = \widehat{\boldsymbol{r}}(\mathbf{x}_t)$. Note that covariates $\boldsymbol{r}_t$ are not observed but have to be estimated nonparametrically from data. As shown in Mammen, Rothe, and Schienle (2012), many economic applications require the nonparametric estimation of the regression function with nonparametrically generated covariates $\widehat{\boldsymbol{r}}(\mathbf{x}_t)$, such as the simultaneous nonparametric equation models (Newey, Powell, and Vella 1999; Imbens and Newey 2009) and the structural equations for treatment effects (Heckman and Vytlacil 2005).

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

$$(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=1}^{T-1} \left(y_{t+1} - \alpha - \boldsymbol{\beta}'(\widehat{\boldsymbol{r}}_t - \mathbf{z})\right)^2 K_h(\widehat{\boldsymbol{r}}_t - \mathbf{z}), \tag{3.5}$$

where $K_h(\mathbf{u}) = h^{-L}\prod_{j=1}^{L}\mathcal{K}(u_j/h)$ is a product kernel with univariate kernel function $\mathcal{K}$, and $h > 0$ is smoothing bandwidth. Specifically, let $\mathbf{e}_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^{L+1}$, $Y = (y_2, \ldots, y_T)'$, $\mathbf{v}_t(\boldsymbol{r}, \mathbf{z}) = \left(1, (\boldsymbol{r}_t - \mathbf{z})'\right)'$, $\widehat{Z} = (\mathbf{v}_1(\widehat{\boldsymbol{r}}, \mathbf{z}), \ldots, \mathbf{v}_{T-1}(\widehat{\boldsymbol{r}}, \mathbf{z}))'$. $W_{\widehat{\boldsymbol{r}}} = \operatorname{diag}(K_h(\widehat{\boldsymbol{r}}_1 - \mathbf{z}), \ldots, K_h(\widehat{\boldsymbol{r}}_{T-1} - \mathbf{z}))$ is a diagonal weighting matrix. Then the solution to the local linear regression (LLR) (3.5) is $\widehat{g}(\mathbf{z}) = \mathbf{e}_1'\left(\widehat{Z}'\widehat{W}\widehat{Z}\right)^{-1}\left(\widehat{Z}'\widehat{W}Y\right)$.

Before proceeding, we provide a set of assumptions for consistent predictive inference.

Assumption 3.6 (Regularity).

(i) The density function $f_{\mathbf{z}}(\mathbf{z})$ of random vector $\mathbf{z} = \boldsymbol{r}(\mathbf{x})$ 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 3.7 (Accuracy). For some $\delta > \eta$, the estimation of $\widehat{\boldsymbol{r}}(\mathbf{x})$ satisfies that

$$||\widehat{\boldsymbol{r}} - \boldsymbol{r}||_\infty = o_p(T^{-\delta}).$$

Remark 3.2. Fan, Xue, and Yao (2017) prove that under Assumptions 3.1–3.3, Assumption 3.7 holds for the predictive indices estimated by the SF approach.

Assumption 3.8 (Complexity). There exists sequences of sets $\mathcal{M}_{T,j}$ such that

(i) $Pr(\widehat{r}_j \in \mathcal{M}_{T,j}) \to 1$ as $T \to \infty$ for all $j = 1, \ldots, L$.
(ii) For a constant $C_M > 0$ and a function $r_{T,j}$ with $||r_{T,j} - r_{0,j}||_\infty = o(T^{-\delta})$, the set $\bar{\mathcal{M}}_{T,j} = \mathcal{M}_{T,j} \cap \{r_j : ||r_j - r_{T,j}||_\infty \le T^{-\delta}\}$ can be covered by at most $C_M exp(\lambda^{-\alpha_j} T^{\xi_j})$ balls with $||\cdot||_\infty$-radius λ for all $\lambda \le T^{-\delta}$, where $0 < \alpha_j \le 2$, $\xi_j \in \mathbb{R}$.

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

Theorem 3.3. Given Assumptions 3.1–3.3, 3.6, and 3.8, let $\kappa = min\{\kappa_1, \kappa_2, \kappa_3\}$, and then, we have

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\widehat{g}(\mathbf{z}) - \widetilde{g}(\mathbf{z}) + \nabla' g(\mathbf{z})\widehat{\Delta}(\mathbf{z})| = O_p(T^{-\kappa}) \tag{3.6}$$

with $\kappa_1 < \delta + 1 - (L+1)\eta - (\frac{1}{\gamma_1} + 1)\max(\delta\alpha_j + \xi_j)$, $\kappa_2 < \delta + \eta$, $\kappa_3 < 2\delta - \eta$, where $\widetilde{g}(\mathbf{z})$ is an infeasible estimator, expected to be fitted by the true value $\boldsymbol{r}_t$ instead of the estimated $\widehat{\boldsymbol{r}}_t$, that is, $\widetilde{g}(\mathbf{z}) = \widetilde{\alpha}$ is obtained by $(\widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}) = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=1}^{T-1}\left(y_{t+1} - \alpha - \boldsymbol{\beta}'(\boldsymbol{r}_t - \mathbf{z})\right)^2 K_h(\boldsymbol{r}_t - \mathbf{z})$. $\widehat{\Delta}(\mathbf{z}) = \bar{\boldsymbol{\alpha}}$ is obtained by $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=1}^{T-1}\left((\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t) - \boldsymbol{\alpha} - \boldsymbol{\beta}'(\boldsymbol{r}_t - \mathbf{z})\right)^2 K_h(\boldsymbol{r}_t - \mathbf{z})$.

Theorem 3.4. Under the same assumptions as in Theorem 3.3, the consistency of nonparametric estimator with generated covariates $\widehat{g}(\mathbf{z})$ is given by

$$\sup_{\mathbf{z}\in I_{\mathbf{z}}} |\widehat{g}(\mathbf{z}) - g(\mathbf{z})| = O_p(\sqrt{\log(T)T^{-1+L\eta}} + T^{-2\eta} + T^{-\delta} + T^{-\kappa}). \tag{3.7}$$

4. Conformal Inference

It is of interest to construct a prediction set based on a time series $\{(\mathbf{x}_t, y_{t+1})\}_{t=1}^{T-1}$ for a future response y_{T+1} given a new feature $\mathbf{x}_T$. Fan, Xue, and Zou (2016) used multi-task quantile regression to construct the provable prediction interval, and Lei et al. (2018) proposed a distributed-free prediction using conformal inference for regression problems. Standard conformal prediction relies on the assumption of exchangeability, and most existing works expose the iid assumption to fulfill the exchangeability condition. To extend to dependent cases such as time series data, Chernozhukov, Wuthrich, and Zhu (2018) introduced a randomization method that accounts for potential serial dependence. By including block structures which preserve the dependence structure in the permutation scheme, their proposed methodology established theoretical guarantees for conformal inference by permutations covering most common types of series models, including strongly mixing processes as a special case.

To conduct conformal prediction in conjunction with the SF, let y be a hypothesized value for y_{T+1}, and $\mu(\mathbf{x}) = \mathbb{E}(y_{t+1}|\mathbf{x}_t = \mathbf{x})$. Note that because ϵ_{t+1} is independent of $\mathbf{x}_t$ as shown in SF model (3.4), we have $\mu(\mathbf{x}) = g(\boldsymbol{r}(\mathbf{x}))$. Through the SF algorithm and the nonparametric estimation from Section 3, we obtain a valid estimator of $\mu(\cdot)$. Let $\widehat{\mu}_y(\cdot)$ denote the estimator constructed based on the augmented data $\{(\mathbf{x}_t, y_{t+1})\}_{t=1}^{T}$, and let $\widehat{\mu}_y^{\pi}(\cdot)$ denote the estimator based on the permuted data $\{(\mathbf{x}_{\pi(t)}, y_{\pi(t)+1})\}_{t=1}^{T}$, where π is a permutation of $\{1, \ldots, T\}$.

The underlying mechanism is testing candidate values for y_{T+1} and constructing prediction sets based on test inversion. Consider the residual-based conformity score

$$R_y := |y - \widehat{\mu}_y(\mathbf{x}_T)| \quad \text{and} \quad R_y^{\pi} := |y - \widehat{\mu}_y^{\pi}(\mathbf{x}_T)|,$$

and define the randomization p-value for testing $H_0 : y_{T+1} = y$ as

$$\widehat{p}(y) := \frac{1}{|\Pi|}\sum_{\pi\in\Pi} \mathbb{1}\{R_y^{\pi} \ge R_y\}, \tag{4.1}$$

Algorithm 1 Conformal inference for SF

Input: observed data $\{\mathbf{x}_t, y_{t+1}\}_{t=1}^{T-1}$, new predictor $\mathbf{x}_T$, $\mathcal{Y}$ (candidate values of y_{T+1}), $1-\alpha$ (significance level), K (number of factors), L (number of indices), H (number of slices).

Output: $\widehat{C}_{1-\alpha}$ (a prediction set for y_{T+1} with coverage probability $1-\alpha$)

Procedures:

(1) **for** $y \in \mathcal{Y}$ **do**

(i) Perform the SF procedure (Fan, Xue, and Yao 2017) on $\{\mathbf{x}_t, y_{t+1}\}_{t=1}^{T-1}$ and $\{\mathbf{x}_T, y\}$ to construct the predictive indices $[\widehat{\boldsymbol{r}}_1', \ldots, \widehat{\boldsymbol{r}}_T']' = \widehat{\mathbf{F}}\widehat{\Phi}$.

(ii) Run the (local) linear regression on $\{\widehat{\boldsymbol{r}}_t, y_{t+1}\}_{t=1}^{T-1}$ and $\{\widehat{\boldsymbol{r}}_T, y\}$.

(iii) Obtain $R_{y_2}, \ldots, R_{y_T}, R_y$ as the absolute value of fitted residuals based on (ii).

(iv) Calculate the randomization p-value $\widehat{p}(y) = (\sum_{t=2}^{T} \mathbb{1}\{R_{y_t} \geq R_y\} + 1)/T$.

end for

(2) Construct the prediction set $\widehat{C}_{1-\alpha} = \{y \in \mathcal{Y} : \widehat{p}(y) \geq \alpha\}$.

where Π is a group of permutations.

The standard conformal prediction with exchangeability assumptions can be viewed as choosing the group to be the set of all permutations. Yet when the exchangeability fails, it is essential to choose a group of permutations that preserve the dependence structure in the data. Following Chernozhukov, Wuthrich, and Zhu (2018), we consider the permutation $\pi_j(t) = \text{mod}(t-j, T) + 1$, $t = 1, \ldots, T$, and the collection of all permutations is given by

$$\Pi = \{\pi_j : 1 \leq j \leq T\}. \tag{4.2}$$

The randomization p-value (4.1) becomes $\widehat{p}(y) := \frac{1}{T}\sum_{\pi\in\Pi} \mathbb{1}\left\{R_y^{\pi} \geq R_y\right\}$. For a given $\alpha \in (0,1)$, the $(1-\alpha)$ confidence set contains the set of y whose p-values are larger than α, that is,

$$C_{1-\alpha} = \left\{y \in \mathbb{R} : \widehat{p}(y) > \alpha\right\}. \tag{4.3}$$

In practice, as it is impractical to search over the entire real line $\mathbb{R}$, we consider a grid of candidate values of y_{T+1}, denoted by $\mathcal{Y}$. We summarize the conformal prediction interval construction for the SF procedure in Algorithm 1.

To study the theoretical properties of the conformal prediction interval, we first define $R_y^{\star} := |y - \mu(\mathbf{x}_T)|$ as the oracle conformity score function. We can view $\{R_y^{\pi\star}\}_{\pi\in\Pi}$ as a time series $\{u_t\}_{t=1}^{T}$ for the Π defined by (4.2). In addition, we can see that $u_t = \varepsilon_{T+2-t}$. Chernozhukov, Wuthrich, and Zhu (2018) proved that as long as the conformity score R_y^{π} satisfies certain accuracy and ergodicity conditions, the group of blocking permutations preserves the dependence structure in the data, contributing to an approximately valid conformal interval. In combination with the SF, we can see that the resulting conformal prediction interval retains its approximate validity.

Theorem 4.1. Under the same assumptions as in Theorem 3.4, and suppose $\{\epsilon_{t+1}\}_{t=0}^{T}$ is stationary and strongly mixing with $\sum_{T=1}^{\infty} \alpha_{mixing}(T) \leq M$ for a constant M, the conformal prediction set constructed from (4.3) has approximate coverage $1-\alpha$, that is, $\left|P\left(y_{T+1} \in C_{1-\alpha}\right) - (1-\alpha)\right| = o_p(1)$ as $T \to \infty$.

5. Numerical Studies

This section conducts both Monte Carlo experiments and an empirical study to evaluate the proposed methods. Section 5.1 demonstrates the additional predictive power of using nonparametrically estimated link function compared to a linear regression (LR) on the predictive indices. In addition, we examine the sensitivity of forecasting performances using SF with respect to the parameter K and its estimator. Those results are presented in the supplementary materials. Section 5.2 verifies the approximate validity of conformal prediction intervals. In Section 5.3, we apply our methods to financial data, and assess the predictability of the daily market return using the cross-section of stock returns.

5.1. Nonparametric Estimation

We first examine the potential benefits brought by nonparametrically estimated link functions. To be more specific, after obtaining the sufficient predictive indices $\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\phi}}_L'\widehat{\mathbf{f}}_t$ as regressors, we compare the performance of SF using a LR to the performance with link function estimated through LLR.

We set underlying factor model as $x_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}$. Following Li, Li, and Shi (2017), we set the number of factors K to increase with the cross-section p in the form of $K = [1.5\log(p)]$ so as to investigate the scenarios of a large K. Here, $[x]$ denotes the integer part of a real number x. To mimic the serial dependence, we generate f_{jt} and u_{it} following two AR(1) processes as $f_{jt} = \alpha_j f_{j,t-1} + e_{jt}$, $u_{it} = \rho_i u_{it-1} + \nu_{it}$, where α_j, ρ_i are drawn from $U[0.2, 0.8]$ and fixed during simulations. We simulate the noises e_{jt}, and ν_{it} independently from standard normal distribution, so do factor loadings $\mathbf{b}_i'$.

We consider the scenarios when the underlying link function is a linear function and a nonlinear function. Both types of models play an influential role and have been widely employed in economics and statistics (Rajan and Zingales 1988; Stock and Watson 2002b). Sharing the spirit of Fan, Xue, and Yao (2017), we consider the following linear forecasting model (Model 1) and the forecasting model with factor interaction (Model 2):

- Model 1: $y_{t+1} = \boldsymbol{\phi}'\mathbf{f}_t + \epsilon_{t+1}$ with $\boldsymbol{\phi} = (0.8, 0.5, 0.3, \mathbf{0}_{K-3}')'$;
- Model 2: $y_{t+1} = f_{1t}(f_{2t} + f_{3t} + 1) + \epsilon_{t+1}$;

where ϵ_{t+1} is drawn from standard normal distribution. For both models there are K common factors driving the predictors whereas only the first three are associated with the response y_{t+1}. For Model 1, the target is a linear function of the latent factors plus some noise. For Model 2, the interaction between factors is present. The true sufficient directions in the central space $S_{y|\mathbf{f}}$ can be represented by $\boldsymbol{\phi}_1 = (1, \mathbf{0}_{K-1}')'$ and $\boldsymbol{\phi}_2 = (0, 1, 1, \mathbf{0}_{K-3}')'/\sqrt{2}$. When implementing the SF algorithms and the PCR, we estimate K with $\widehat{K}$ using the $IC(k)$ criterion presented in Section 2 of the supplementary materials.

To assess the predictability of different approaches, we examine the in-sample R^2 and the out-of-sample R^2 suggested by Campbell and Thompson (2008). Let $p = 100, 200, 500, 1000$ and $T = 100, 200, 500$. For each setup, we generate 1000 independent replications, and report the average in-sample R^2 and out-of-sample R^2 in percentage.

Table 1. Forecasting performance using in-sample and out-of-sample R^2 (Model 1).

		In-sample R^2(%)				Out-of-sample R^2(%)			
p	T	SF1-LLR	SF1-LR	PCR	PC1	SF1-LLR	SF1-LR	PCR	PC1
100	100	50.9	49.6	50.8	20.1	36.6	38.3	41.3	16.4
100	200	50.1	49.4	49.7	21.0	43.9	44.6	45.1	19.3
100	500	49.1	48.9	49.0	20.5	46.8	47.1	47.2	19.9
200	100	51.4	49.9	51.3	17.4	34.8	36.4	40.0	13.4
200	200	50.4	49.7	50.1	17.8	43.5	44.2	44.8	16.2
200	500	49.6	49.3	49.4	17.7	47.1	47.4	47.5	17.1
500	100	52.0	50.5	52.5	13.5	32.4	34.0	38.3	10.2
500	200	50.9	50.2	50.7	13.7	42.4	43.1	43.9	11.7
500	500	49.8	49.6	49.7	14.0	46.7	46.9	47.1	13.4
1000	100	52.4	50.9	53.2	12.8	30.9	32.8	37.6	9.2
1000	200	51.1	50.4	51.0	12.8	42.2	42.9	43.7	10.7
1000	500	49.9	49.6	49.7	12.7	46.5	46.7	46.9	11.8

Table 2. Forecasting performance using in-sample and out-of-sample R^2 (Model 2).

		In-sample R^2(%)					Out-of-sample R^2(%)				
p	T	SF2-LLR	SF2-LR	SFi	PCR	PCRi	SF2-LLR	SF2-LR	SFi	PCR	PCRi
100	100	46.4	27.8	36.5	30.1	34.3	18.0	9.3	14.1	10.8	12.5
100	200	51.7	26.7	43.2	27.5	31.1	35.0	16.7	28.2	17.9	20.3
100	500	59.2	25.5	57.0	25.6	29.1	51.5	21.2	48.4	21.5	24.4
200	100	45.7	28.3	36.4	31.2	34.7	15.5	8.5	12.3	9.7	10.3
200	200	49.2	26.5	40.9	27.5	30.4	31.0	15.8	25.4	17.0	18.8
200	500	58.9	25.8	56.7	26.0	28.6	50.2	21.2	47.3	21.6	23.8
500	100	43.2	28.3	34.6	32.5	35.2	10.8	6.8	9.4	7.7	7.6
500	200	47.2	27.5	39.6	28.8	31.1	27.4	14.9	22.7	16.4	17.4
500	500	57.6	26.4	55.2	26.7	28.4	47.5	21.2	44.5	21.6	22.9
1000	100	43.3	29.3	35.3	33.8	36.4	9.7	6.6	8.4	6.5	6.2
1000	200	47.1	28.3	39.7	29.8	31.8	26.2	15.0	21.8	16.5	17.2
1000	500	56.7	26.1	54.6	26.5	28.1	45.8	20.6	43.3	21.2	22.3

Table 1 presents the prediction performance of four forecasting approaches for linear forecasting model. SF1-LLR denotes the SF using only one predictive index, with link function estimated via LLR. SF1-LR is also the SF approach with $L = 1$, except that the response is fitted using LR on the predictive index. PCR stands for the the principal component regression, and PC1 uses only the first principal component. Roughly speaking, SF1-LLR, SF1-LR, and PCR yield comparable performance when the true predictive model is linear. However, PC1 performs poorly, indicating using only the first principal component is not enough for prediction. The good performance of PCR in linear forecasting has been well studied by Stock and Watson (2002b) and the satisfactory results of SF1-LR is guaranteed in Fan, Xue, and Yao (2017). Since the LLR has no bias when the true model is linear (Fan and Gijbels 1996), SF1-LLR exhibits similar behavior to SF1-LR in Model 1.

Table 2 displays the comparison among various forecasting approaches in the presence of interaction between factors and nonlinear link functions. SF2-LLR and SF2-LR are named in the same fashion as SF1-LLR and SF1-LR, except the difference in the number of predictive indices $L = 2$ rather than $L = 1$. SFi and PCRi were introduced in Fan, Xue, and Yao (2017) as an effective way to account for interactions in the model. SFi fits a multivariate LR on the first two predictive indices and includes their interaction effect, and PCi extends PCR by including an extra interaction term built on the first two principal components. As can be seen from the table, SF with linear link function and PCR no longer perform well because of the existence of nonlinear effects. SFi and PCRi can improve over SF2-LR and PCR in terms of in-sample performance, as including the interaction term takes part of interaction into account. However, they do not help much with the out-of-sample performance, since the interaction term is incorrectly specified. As a comparison, SF2-LLR can effectively model the nonlinearity and perform well in both in-sample and out-of-sample predictions.

5.2. Conformal Inference

In this subsection, we take a look into the finite-sample performance of the conformal prediction sets. We adopt the same data generating process as in Section 5.1. In addition, we let the mis-coverage rate α to be 0.1. For each simulation setting, we generate 1000 replications, and report the empirical coverage rates and the average length of the confidence intervals. The results are displayed in Tables 3 and 4 for Models 1 and 2, respectively.

As is seen, for all methods, the empirical coverage rates are approximately equal to $1 - \alpha = 0.90$, which provides numerical evidence on the approximate validity mentioned in Theorem 4.1. For linear forecasting model, SF1-LLR, SF1-LR, and PCR share similar performance, yet PC1 yields to a wider confidence set. For the nonlinear model, both SF2-LR and PCR have relatively wider prediction intervals. PCi improve a bit, and SFi and SF2-LLR can generate shorter intervals.

These outcomes resound with the conclusions from Section 5.1 in the sense that a more accurate forecasting method likely leads to a shorter conformal prediction interval (Lei et al. 2018). Intuitively, this happens because the conformal intervals are essentially constructed based on the quantile of residuals.

Table 3. Empirical coverage and average length of conformal prediction intervals (Model 1).

		Coverage (%)				Length			
p	T	SF1-LLR	SF1-LR	PCR	PC1	SF1-LLR	SF1-LR	PCR	PC1
100	100	90.8	91.9	91.4	91.1	3.481	3.446	3.400	4.135
100	200	89.6	90.6	90.2	90.4	3.361	3.348	3.335	4.104
100	500	91.1	91.2	90.9	89.9	3.289	3.284	3.282	4.067
200	100	90.6	90.9	90.6	91.4	3.518	3.487	3.425	4.211
200	200	90.4	90.4	90.8	89.7	3.354	3.339	3.327	4.161
200	500	90.7	90.8	90.4	90.3	3.288	3.282	3.281	4.143
500	100	88.6	89.5	89.6	89.1	3.553	3.524	3.446	4.289
500	200	90.7	89.9	90.2	90.2	3.375	3.361	3.344	4.269
500	500	90.7	90.9	90.9	91.0	3.284	3.279	3.273	4.243
1000	100	90.0	91.0	90.6	90.7	3.593	3.565	3.452	4.314
1000	200	90.0	89.9	89.9	90.7	3.383	3.368	3.346	4.294
1000	500	90.8	90.5	90.7	88.8	3.290	3.284	3.281	4.264

Table 4. Empirical coverage and average length of conformal prediction intervals (Model 2).

		Coverage (%)					Length				
p	T	SF2-LLR	SF2-LR	SFi	PCR	PCRi	SF2-LLR	SF2-LR	SFi	PCR	PCRi
100	100	89.3	89.2	89.1	89.1	88.8	5.137	5.712	5.417	5.617	5.506
100	200	90.9	90.6	90.9	90.6	90.4	4.643	5.538	4.899	5.485	5.399
100	500	89.8	88.8	89.6	88.9	89.6	4.099	5.419	4.207	5.406	5.306
200	100	91.2	91.0	90.9	89.6	90.9	5.215	5.717	5.468	5.621	5.545
200	200	90.2	90.3	90.9	90.3	90.1	4.736	5.539	5.011	5.500	5.419
200	500	90.2	90.1	90.0	90.2	90.1	4.124	5.429	4.237	5.411	5.339
500	100	89.8	90.0	89.4	90.3	90.3	5.486	5.866	5.700	5.701	5.653
500	200	89.2	88.7	88.4	89.0	89.0	4.876	5.581	5.130	5.518	5.469
500	500	90.8	91.7	89.6	91.8	91.6	4.214	5.430	4.326	5.408	5.364
1000	100	89.4	89.7	89.8	89.3	88.9	5.601	5.939	5.778	5.749	5.699
1000	200	88.8	90.4	89.6	90.3	90.2	4.951	5.632	5.186	5.555	5.509
1000	500	88.5	88.4	90.0	88.9	89.2	4.267	5.449	4.349	5.431	5.388

As we improve the accuracy of forecasting approaches, the fitted residuals become smaller, so that the resulting conformal intervals decrease in length.

5.3. An Empirical Study

Stock market returns are volatile and hard to predict, but we want to examine whether the cross-section of individual stock returns contains any predictive information of the market. 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) focuses on monthly frequency and relies on portfolios information. By contrast, we examine the issue using individual stock returns at 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. In the real-world practice, daily factor models provided by Barra Inc. and Axioma Inc. are widely used to explain the cross-section of stock returns.

We use a rolling out-of-sample forecast implementation. At date t, our 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. Factors $\mathbf{f}_s$ are constructed through (3.1) using daily stock returns $\mathbf{x}_s$ ($s = t-755, \ldots, t$) of the past three years of trading days. We then collect $\{(\mathbf{f}_s, y_{s+1}) : s = t-755, \ldots, t-1\}$ (or simply the raw data $\{(\mathbf{x}_s, y_{s+1})\}$). Finally, we use $\mathbf{f}_t$ (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.

Table 5. Correlation matrix of predicted market returns via different models.

	SF1-LR	SF2-LR	SF1-LLR	SF2-LLR	PCR	PLS
SF1-LR	1.00	0.90	0.68	0.55	0.76	0.54
SF2-LR	–	1.00	0.61	0.67	0.84	0.62
SF1-LLR	–	–	1.00	0.66	0.50	0.39
SF2-LLR	–	–	–	1.00	0.57	0.45
PCR	–	–	–	–	1.00	0.78
PLS	–	–	–	–	–	1.00

Our models consist of SF, PCR, and 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 SF1 and SF2 the SF with $L = 1, 2$ predictive indices. To reveal its potential, we consider both linear (SF-LR) and nonlinear (SF-LLR) sufficient forecasting in building the predictive regression, where the nonlinear link function is estimated through LLR.

Table 5 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 SF are more correlated with PCR than PLS, as both SF and PCR depend on the estimated factors from 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,

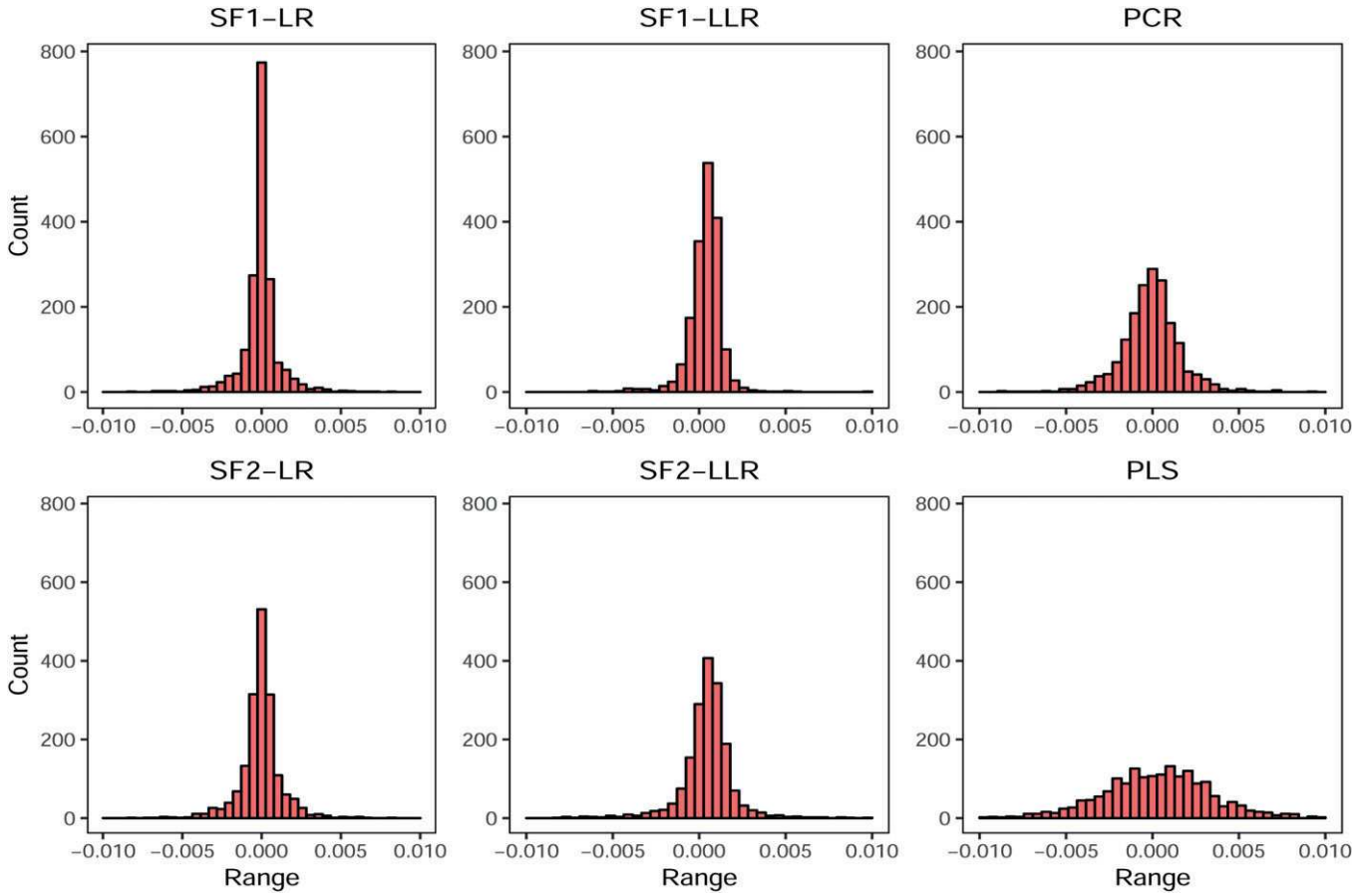

Figure 1. Histograms of predicted market returns.

Table 6. Prediction summary.

	SF1-LR	SF2-LR	SF1-LLR	SF2-LLR	PCR	PLS
corr	7.76%	7.89%	7.58%	8.15%	6.85%	6.21%
RMSE	0.998	0.999	0.992	0.998	1.010	1.071

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.

Table 7. Empirical coverage and average length of conformal prediction intervals.

		SF2-LM	SF2-LLR	PCR	PLS
90% CI	Coverage (%)	92.68	92.45	92.74	92.28
	Length ($\times 10^{-2}$)	3.820	3.815	3.836	3.849
95% CI	Coverage (%)	96.77	96.82	96.82	96.48
	Length ($\times 10^{-2}$)	5.091	5.097	5.110	5.075

nonlinear forecasts can be very different from linear forecasts. Figure 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.

Table 6 shows the predictive power of the six methods. For each method M, we consider its correlation with the target and its mean squared error relative to the out-of-sample mean,

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

where S is the evaluation sample and $\bar{y}_t$ is the mean of y_t in S. When RMSE is larger than 1, it indicates that the model does not beat the out-of-sample mean, which is not uncommon in the forecast of stock returns. Many previously studied predictors of stock returns in the literature typically perform well in sample but become insignificant out-of-sample, often performing worse than forecasts based on the historical mean return (Goyal and Welch 2008). As shown in Table 6, SF methods yield slightly better performance than the other methods, mostly because it is a more concise model and is less prone to over-fit. By exploring the nonlinear nature of market returns, SF2-LLR delivers additional predictive power in terms of correlation metrics. Figure 2 shows that the RMSEs for SF-LR, SF-LLR, 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.

In addition to the forecasting performance via different methods, we also examine the corresponding conformal prediction intervals associated with each forecasting approach. Figure 3 displays the time series of the SP500 daily returns over the evaluation periods from 2010 to 2016, together with the 90% and 95% conformal prediction bands constructed by SF-LR, SF-LLR, PCR, and PLS, respectively. Their corresponding empirical coverage rates and average length are reported in Table 7. The graphs show that both SF and PCR generate similar prediction

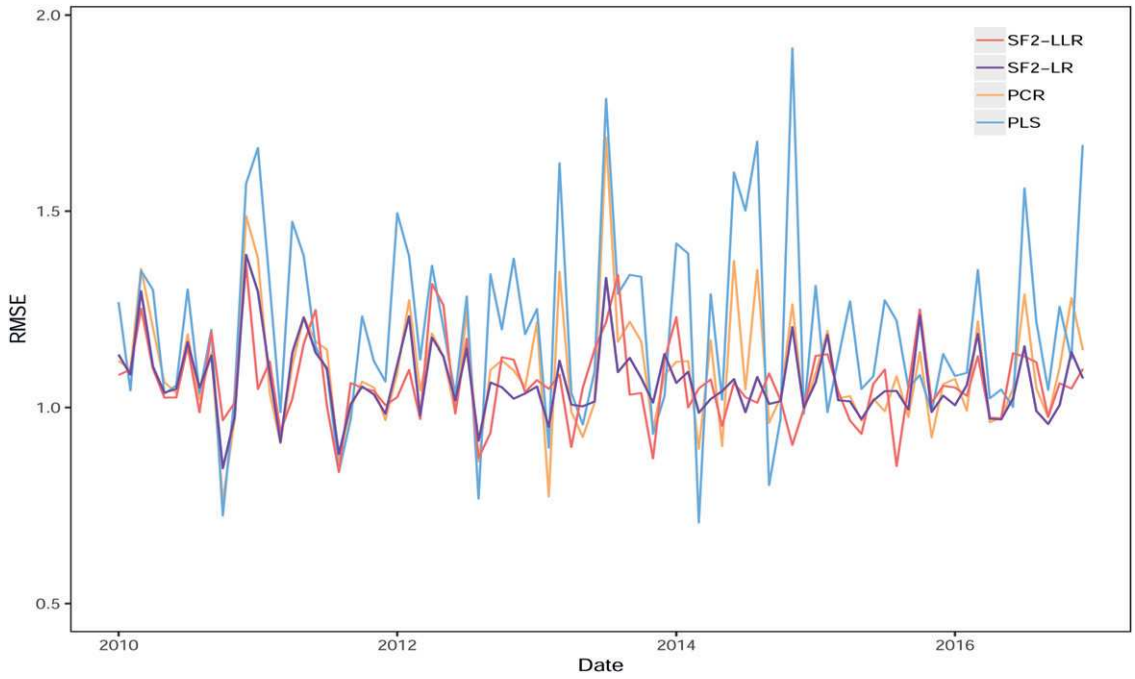

Figure 2. Monthly RMSE over the evaluation period.

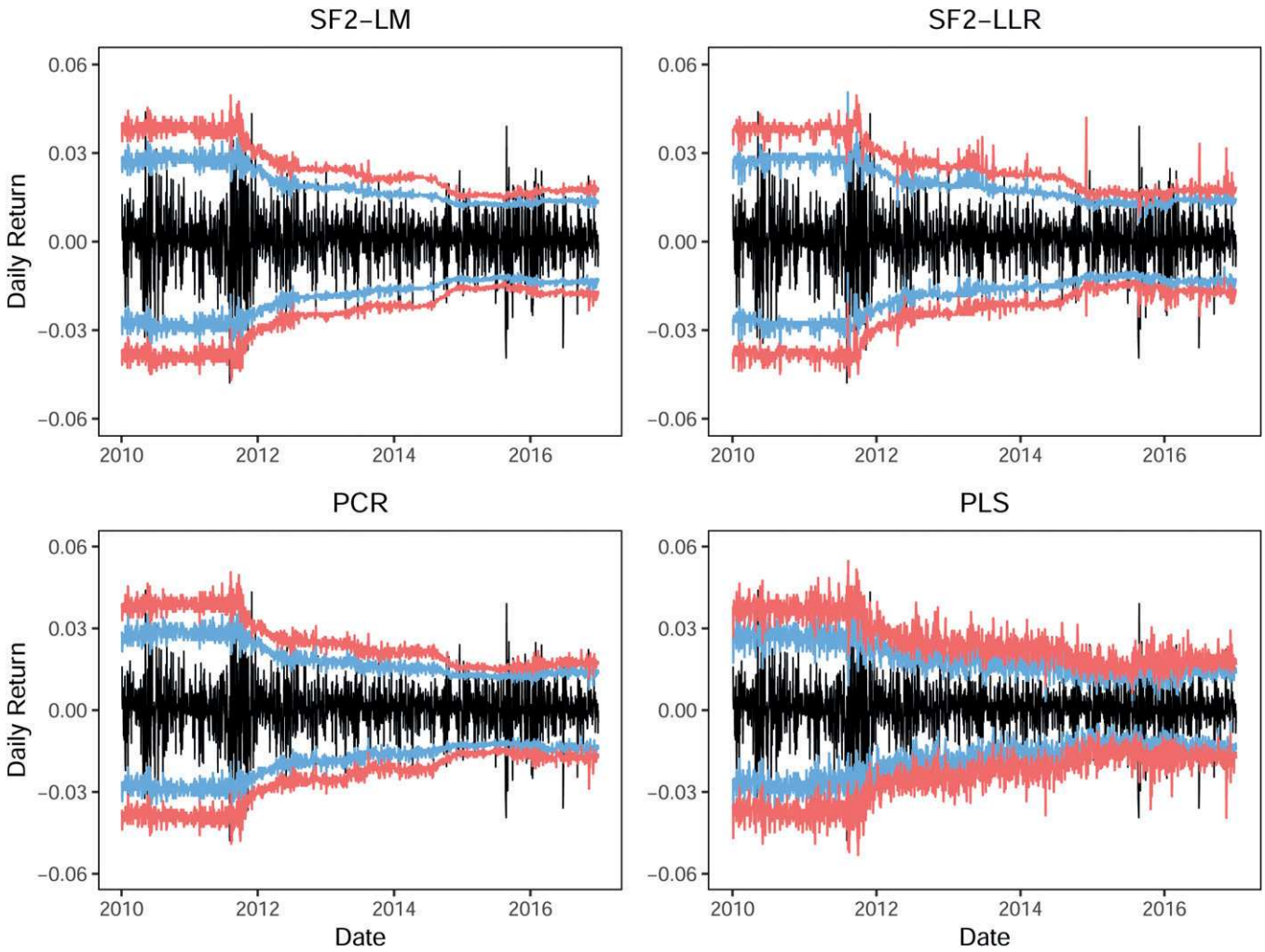

Figure 3. Conformal prediction intervals over the evaluation period. NOTES: The black lines denote the SP500 daily returns, the red lines denote the 95% prediction intervals, and the blue lines denote the 90% prediction intervals.

bands, while the PLS yields to a slightly different band which oscillates more often. All prediction intervals can achieve the desired coverage, more precisely, a slightly higher coverage rate than specified.

6. Conclusion

This article revisits the SF and studies its nonparametric estimation and predictive inference in a large panel data settings. We point out the close connection between the SF and existing methods (such as FM regression and PLS) in their way of using the target information. The estimated forecasting directions are shown to be consistent under a diverging number of latent factors, and the asymptotic behavior of the estimated SF directions is carefully characterized. We also show that the estimated predicted indices can be directly used to estimate the forecasting function, which is often a nontrivial issue in factor analysis. Moreover, we introduce the conformal inference to construct valid prediction sets for SF. In numerical studies, we demonstrate that allowing nonlinearity in forecasting functions can yield additional gains and illustrate the approximate validity of prediction sets constructed from the conformal inference for SF.

Supplementary Materials

The supplementary materials consist of three distinct sections, including the complete proofs of propositions and theorems, the discussion on the choice of turning parameters as well as additional numerical results about the sensitivity of forecasting performances of SF with regards to the estimated number of factors.

Acknowledgments

The authors would like to thank Joint Editor Christian Hansen, associate editor, and two referees for their helpful comments and suggestions.

Funding

This article was supported in part by NSF grants DMS-1811552, DMS-1953189, and CCF-2007823.

References

Bai, J., and Li, K. (2012), "Statistical Methods of Factor Models of High Dimension," *The Annals of Statistics*, 40, 436–465. [346]

Bai, J., and Ng, S. (2006), "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74(4), 1133–1150. [342]

——— (2008), "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304–317. [342]

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association*, 101, 119-137. [342]

Barbarino, A., and Bura, E. (2015), "Forecasting With Sufficient Dimension Reductions," Working Paper. [344]

——— (2017), "A Unified Framework for Dimension Reduction in Forecasting," Working Paper. [344]

Campbell, J. Y., and Thompson, S. B. (2008), "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *The Review of Financial Studies*, 21, 1798–1828. [348]

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018), "Exact and Robust Conformal Inference Methods for Predictive Machine Learning With Dependent Data," *Proceedings of the 31st Conference on Learning Theory, PMLR*, 75, 732–749. [343,347,348]

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimensions Reduction in Regressions With Categorical Predictors," *The Annals of Statistics*, 30, 475–497. [344]

Cochrane, J. H. (2001), *Asset Pricing*, Princeton, NJ: Princeton University Press. [344]

Cook, R. D. (2004), "Testing Predictor Contributions in Sufficient Dimension Reduction," *The Annals of Statistics*, 32, 1062–1092. [344]

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. [343,350]

Fama, E. F., and MacBeth, J. D. (1973), "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636. [342,344]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Boca Raton, FL: CRC Press. [347,349]

Fan, J., Xue, L., and Yao, J. (2017), "Sufficient Forecasting Using Factor Models," *Journal of Econometrics*, 201, 292–306. [342,343,344,345,346,347,348,349]

Fan, J., Xue, L., and Zou, H. (2016), "Multi-Task Quantile Regression Under the Transnormal Model," *Journal of the American Statistical Association*, 111, 1726–1735. [347]

Frank, L. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [344]

Goyal, A., and Welch, I. (1993), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21, 1455–1508. [351]

Hall, P., and Li, K.-C. (1993), "On Almost Linearity of Low Dimensional Projections From High Dimensional Data," *The Annals of Statistics*, 21, 867–889. [342,346]

Heckman, J. J., and Vytlacil, E. (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. [347]

Imbens, G. W., and Newey, W. K. (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. [347]

Jagannathan, R., and Wang, Z. (1996), "The Conditional CAPM and the Cross-Section of Expected Returns," *Journal of Finance*, 51, 3–53. [343]

Kelly, B., and Pruitt, S. (2015), "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors," *Journal of Econometrics*, 186, 294–316. [342,344,350]

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094–1111. [343,347,349]

Lettau, M., and Ludvigson, S. (2001), "Consumption, Aggregate Wealth, and Expected Stock Returns," *Journal of Finance*, 56, 815–849. [343]

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, 76–86. [342,345,346,348]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327. [342,343,346]

Li, Q., Vassalou, M., and Xing, Y. (2006), "Sector Investment Growth Rates and the Cross Section of Equity Returns," *The Journal of Business*, 79, 1637–1665. [343]

Lin, Q., Zhao, Z., and Liu, J. S. (2018), "On Consistency and Sparsity for Sliced Inverse Regression in High Dimensions," *The Annals of Statistics*, 46, 580–610. [344]

Ludvigson, S. C., and Ng, S. (2007), "The Empirical Risk–Return Relation: A Factor Analysis Approach," *Journal of Financial Economics*, 83, 171–222. [342,345]

Luo, W., Xue, L., and Yao, J. (2017), "Inverse Moment Methods for Sufficient Forecasting Using High-Dimensional Predictors," arXiv no. 1705.00395. [342]

Mammen, E., Rothe, C., and Schienle, M. (2012), "Nonparametric Regression With Nonparametrically Generated Covariates," *The Annals of Statistics*, 40, 1132–1170. [343,347]

Newey, W. K., Powell, J. L., and Vella, F. (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603. [347]

Rajan, R. G., and Zingales, L. (1988), "Financial Dependence and Growth," *American Economic Review*, 88, 559–586. [348]

Stock, J. H., and Watson, M. W. (1989), "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual*, 4, 351–409. [342]

——— (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [342]

——— (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162. [342,343,348,349]

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, New York: Springer. [343]

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, 363–410. [343,346]

Zhu, L., Miao, B., and Peng, H. (2006), "On Sliced Inverse Regression With High-Dimensional Covariates," *Journal of the American Statistical Association*, 101, 630–643. [344]