Biometrika (2022), **109**, 2, *pp.* 473–487 doi: 10.1093/biomet/asab037
Advance Access publication 9 November 2021

Inverse moment methods for sufficient forecasting using high-dimensional predictors

BY WEI LUO

Center for Data Science, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China
weiluo@zju.edu.cn

LINGZHOU XUE

Department of Statistics, Pennsylvania State University,
318 Thomas Building, University Park, Pennsylvania 16802, U.S.A.
lzxue@stat.psu.edu

JIAWEI YAO

Department of Operations Research and Financial Engineering, Princeton University,
Sherrerd Hall, Princeton, New Jersey 08544, U.S.A.
jiaweiy@alumni.princeton.edu

AND XIUFAN YU

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 102G Crowley Hall, Notre Dame, Indiana 46556, U.S.A.
xyu24@nd.edu

SUMMARY

We consider forecasting a single time series using a large number of predictors in the presence of a possible nonlinear forecast function. Assuming that the predictors affect the response through the latent factors, we propose to first conduct factor analysis and then apply sufficient dimension reduction on the estimated factors to derive the reduced data for subsequent forecasting. Using directional regression and the inverse third-moment method in the stage of sufficient dimension reduction, the proposed methods can capture the nonmonotone effect of factors on the response. We also allow a diverging number of factors and only impose general regularity conditions on the distribution of factors, avoiding the undesired time reversibility of the factors by the latter. These make the proposed methods fundamentally more applicable than the sufficient forecasting method of Fan et al. (2017). The proposed methods are demonstrated both in simulation studies and an empirical study of forecasting monthly macroeconomic data from 1959 to 2016. Also, our theory contributes to the literature of sufficient dimension reduction, as it includes an invariance result, a path to perform sufficient dimension reduction under the high-dimensional setting without assuming sparsity, and the corresponding order-determination procedure.

Some key words: Factor model; Forecasting; High-dimensional asymptotics; Invariance property; Principal component; Sufficient dimension reduction.

1. INTRODUCTION

Forecasting using high-dimensional predictors is an increasingly important research topic in statistics, biostatistics, macroeconomics and finance. A large body of literature has contributed to forecasting in a data-rich environment, with various applications such as the forecasts of market prices, dividends and bond risks (Sharpe, 1964; Lintner, 1965; Ludvigson & Ng, 2009), macroeconomic outputs (Stock & Watson, 1989; Bernanke et al., 2005), macroeconomic uncertainty and fluctuations (Ludvigson & Ng, 2007; Jurado et al., 2015) and clinical outcomes based on massive genetic, genomic and imaging measurements. Motivated by principal component regression, the pioneering papers by Stock & Watson (2002a, b) systematically introduced the forecasting procedure using factor models, which has played an important role in macroeconomic analysis. Recently, Fan et al. (2017) extended Stock & Watson (2002a, b) to allow for a nonlinear forecast function and multiple nonadditive forecasting indices. Following Fan et al. (2017), we consider the following factor model with a target variable y_{t+1} that we aim to forecast:

$$y_{t+1} = g(\phi_1' f_t, \ldots, \phi_L' f_t, \epsilon_{t+1}), \tag{1}$$

$$x_{it} = b_i' f_t + u_{it}, \tag{2}$$

where $1 \leqslant i \leqslant p$, $1 \leqslant t \leqslant T$, x_{it} is the ith high-dimensional predictor observed at time t, b_i is a $K \times 1$ vector of factor loadings, f_t is a $K \times 1$ vector of common factors driving both predictor and response, $g(\cdot)$ is an unknown forecast function that is possibly nonadditive and nonseperable, u_{it} is an idiosyncratic error, and ϵ_{t+1} is an independent stochastic error. Here, $\phi_1, \ldots, \phi_L, b_1, \ldots, b_p$ and $f_1, \ldots, f_T$ are unobserved vectors. Model (1) equivalently assumes

$$y_{t+1} \perp\!\!\!\perp f_t \mid (\phi_1, \ldots, \phi_L)' f_t. \tag{3}$$

The linear space spanned by $\phi_1, \ldots, \phi_L$, denoted by $\mathcal{S}_{y|f}$, is the parameter of interest that is identifiable and known as the central subspace (Cook, 1998). Fan et al. (2017) introduced the sufficient forecasting scheme that uses factor analysis to estimate f_t and then applies sliced inverse regression (Li, 1991) to Model (1) with the estimated factors as the predictor. Such a combination provides a promising forecasting technique that not only extracts the underlying commonality of the high-dimensional predictor, but also models the complex dependence between the predictor and the forecast target. It allows the dimension of the predictor to diverge and even become much larger than the number of observations.

The consistency result of Fan et al. (2017) is nontrivial. If we replace the true factors f_t with a consistent estimate $\hat{f}_t$ in (3) and define the central subspace $\mathcal{S}_{y|\hat{f}}$ similarly, then $\mathcal{S}_{y|\hat{f}}$ may differ from $\mathcal{S}_{y|f}$ substantially. Thus, the naive method of applying existing dimension reduction methods to the estimated factors $\hat{f}_t$ may not necessarily lead to consistent estimation of $\mathcal{S}_{y|f}$, even if it consistently estimates $\mathcal{S}_{y|\hat{f}}$. Fan et al. (2017) effectively addressed this issue by developing an important invariance result between $E(f_t \mid y_{t+1})$ and $E(\hat{f}_t \mid y_{t+1})$, see Proposition 2.1 and Equation (2.9) of Fan et al. (2017). This invariance result provides an essential foundation for using the sliced inverse regression under Models (1) and (2).

Nonetheless, the applicability of Fan et al. (2017) is restricted by the requirements that the number of factors K must be fixed as p and T grow, and, for each set of factors, a linearity condition, see Condition 1 below, must hold. In particular, as $\mathcal{S}_{y|f}$ is unknown, the linearity condition is commonly strengthened to equivalently require an elliptically distributed f_t, which causes the undesired property of time reversibility (Xia et al., 2002). In addition, the consistency result of Fan et al. (2017) and Yu et al. (2021) hinges on an exhaustive estimation of $\mathcal{S}_{y|f}$, i.e., detecting all

the directions, for which $\phi_1' \Sigma_{f|y}\phi_1, \ldots, \phi_L' \Sigma_{f|y}\phi_L$ must be positive; see their Assumption (A2). This condition is violated, i.e., $\phi' \Sigma_{f|y}\phi$ being zero for some $\phi \in \mathcal{S}_{y|f}$, if $\phi' f_t \mid y_{t+1}$ has a symmetric distribution, which occurs when the forecast target was investigated using squared factors (Ludvigson & Ng, 2007; Bai & Ng, 2008). These limitations motivate us to construct more powerful forecasting methods based on the work of Fan et al. (2017).

In this paper we propose to use factor analysis and sufficient dimension reduction sequentially for sufficient forecasting, with second- or higher-order inverse moment methods being the working sufficient dimension reduction method. In the main text we focus on a commonly used second-order inverse moment method called directional regression (Li & Wang, 2007), and defer the development with the third-order inverse moment method to the Supplementary Material. Based on Models (1) and (2), the proposed method includes the following steps:

Step 1. Estimate the factor loadings B and the factors f_t in Model (2).

Step 2. Use the estimates $\hat{B}$ and $\hat{f}_t$ in directional regression to estimate $\mathcal{S}_{y|f}$.

Step 3. Use the nonparametric methods (Fan & Gijbels, 1996; Matzkin, 2002; Yu et al., 2021) to estimate $g(\cdot)$ in Model (1) and forecast y_{t+1}, based on the estimate of $(\phi_1' f_t, \ldots, \phi_L' f_t)$.

By studying both $E(f_t \mid y_{t+1})$ and $E(f_t f_t' \mid y_{t+1})$ in Step 2, we explore the full power of the factor space. To this end, we first provide an important invariance result, Lemma 1, for directional regression. With the help of this invariance result, we do not require the coincidence or closeness of two central subspaces $\mathcal{S}_{y|f}$ and $\mathcal{S}_{y|\hat{f}}$, so the proposed method can be applied to more general data, such as nonnormally distributed factors.

Our work extends the method, theory and applicability of the forecasting using factor models. Compared with Fan et al. (2017), we relax the linearity condition to the general moment conditions on f_t. From the discussion above, the proposed method does not require time reversibility of the factors, so it can be applied to the generalized forecasting model

$$y_{t+1} = g(\phi_1' f_t + \psi_1' \omega_t, \ldots, \phi_L' f_t + \psi_L' \omega_t, \epsilon_{t+1}), \tag{4}$$

where ω_t is an $m \times 1$ vector of the observed variables, e.g., lags of y_{t+1}. In addition, by using the higher-order inverse moments, the proposed method requires a weaker condition than Fan et al. (2017) and Yu et al. (2021) for exhaustive estimation of $\mathcal{S}_{y|f}$. In particular, it can detect nonmonotone effects of the factors on the response. Furthermore, we allow the number of underlying factors K to diverge as $p, T \to \infty$. By Lam & Yao (2012), Jurado et al. (2015) and Li et al. (2017), our method will deliver a more powerful forecast than Stock & Watson (2002a, b) and Fan et al. (2017).

Using directional regression as an illustration, the proposed method also provides a novel framework of performing sufficient dimension reduction with large-panel data under the high-dimensional setting, without the commonly adopted sparsity assumption, but with the assumption that the predictor affects the response only through the latent factors. The original direction regression (Li & Wang, 2007) can only deal with independently and identically distributed data under the low-dimensional setting. This enhances the applicability of model-free dimension reduction for high-dimensional data, when the sparsity assumption is not suitable.

The consistency of the proposed method hinges on the consistency of both factor analysis and directional regression based on the estimated factors, which we study next. For ease of presentation, we assume that both the number of factors K and the dimension L of $\mathcal{S}_{y|f}$ are known a priori. This does not affect the asymptotic development of the resulting estimator, as long as K and L can be consistently estimated; see the Supplementary Material for details. The consistent estimation of K and L is deferred to § 5. Throughout the article, we assume L to be fixed as K diverges.

2. CONSISTENCY OF FACTOR ANALYSIS

To make a forecast, we need to estimate the factor loadings B and the error covariance matrix Σ_u. Consider the following constrained least squares problem:

$$(\hat{B}_K, \hat{F}_K) = \arg\min_{(B,F)} \|X - BF'\|_{\mathrm{F}}^2, \quad \text{subject to } T^{-1}F'F = I_K,\ B'B \text{ diagonal}, \tag{5}$$

where $X = (x_1, \ldots, x_T)$, $F' = (f_1, \ldots, f_T)$ and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. The constraints $T^{-1}F'F = I_K$ and that $B'B$ is diagonal address the issue of identifiability during the minimization. As these conditions can always be satisfied for any BF' after appropriate matrix operations on B and F, they impose no additional restrictions on the factor model (2). It is known that the minimizers $\hat{F}_K$ and $\hat{B}_K$ of (5) are such that the columns of $\hat{F}_K/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $X'X$ and $\hat{B}_K = T^{-1}X\hat{F}_K$. To simplify the notation, let $\hat{B} = \hat{B}_K$ and $\hat{F} = \hat{F}_K$.

As both the dimension p of the predictor x_t and the number of factors K are diverging, it is necessary to regulate the magnitude of the factor loadings B and the idiosyncratic error u_t, so that the latter is negligible with respect to the former. We should also regulate the stationarity of the time series. In this paper we adopt the following assumptions. For simplicity in notation, we let $U = (u_{it})_{p\times T}$, $B = (b_1, \ldots, b_p)'$ and $\|B\|_{\max}$ be the maximum of the absolute values of all the entries in B. Let $\mathcal{F}_{\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras generated by $\{(f_t, u_t, \epsilon_{t+1}) : t \leqslant 0\}$ and $\{(f_t, u_t, \epsilon_{t+1}) : t \geqslant T\}$, respectively. Let $\alpha(T) = \sup_{A\in\mathcal{F}_{\infty}^{0}, B\in\mathcal{F}_{T}^{\infty}} |P(A)P(B) - P(AB)|$.

Assumption 1 (*Factors and loadings*).

(i) There exists $b > 0$ such that $\|b_i\| \leqslant b$ for $i = 1, \ldots, p$, and there exist two positive constants c_1 and c_2 such that $c_1 < p^{-1}\lambda_{\min}(B'B) < p^{-1}\lambda_{\max}(B'B) < c_2$.

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

Assumption 2 (*Data-generating process*). There are three independent groups, $\{f_t\}_{t\geqslant 1}$, $\{u_t\}_{t\geqslant 1}$ and $\{\epsilon_{t+1}\}_{t\geqslant 1}$, and they are strictly stationary, $\{K^{-2}E\|f_t\|^4 : K \in \mathbb{N}\}$ and $\{K^{-1}E(\|f_t\|^2 \mid y_{t+1}) : K \in \mathbb{N}\}$ are bounded sequences, and $\alpha(T) < c\rho^{\mathrm{T}}$ for $T \in \mathbb{Z}^+$ and some $\rho \in (0, 1)$.

Assumption 3 (*Residuals and dependence*). There is a constant $M > 0$ such that

(i) $E|u_{it}|^8 \leqslant M$;

(ii) $\|\Sigma_u\|_1 \leqslant M$;

(iii) for every (t, s), $E|p^{-1/2}\{u_s'u_t - E(u_s'u_t)\}|^4 \leqslant M$;

(iv) $U = LER$, where $L \in \mathbb{R}^{p\times p}$ and $R \in \mathbb{R}^{T\times T}$ are nonrandom positive definite matrices and $E = (e_{it})_{p\times T}$ includes independent elements with $E(e_{ti}) = 0$ and $E|e_{it}|^7 \leqslant M$.

Assumptions 1 and 3 ensure that signals dominate errors in the population level as p grows. Assumption 1 regulates the signal strength of factors contained in the predictor through the convergence rate of estimated factor loadings, and Assumption 3 regulates the idiosyncratic errors. Assumption 3(iv) regulates weak autocorrelation and cross-sectional correlation as in Li et al. (2017). Assumption 2 imposes independence between factors and idiosyncratic errors as in Lam & Yao (2012). Assumption 2 implies that the observations are only weakly dependent, so that the estimation accuracy grows with T. Assumptions 2 and 3(ii) imply that for every $i, j, t, s > 0$, $\max_{t\leqslant T} p^{-1}\sum_{i,j}|E(u_{it}u_{jt})| = O(1)$ and $(pT)^{-1}\sum_{i,j,t,s}|E(u_{it}u_{js})| = O(1)$; see Lemma 6 of Fan et al. (2013).

Under these assumptions, we have the following consistency result for estimating the factor loadings. Instead of the Frobenius norm used in (5), we use the spectral norm to measure the

magnitude of a matrix, defined as $\|A\| = \lambda_{\max}^{1/2}(A'A)$, the square root of the largest eigenvalue of the symmetric matrix $A'A$, for any matrix A.

THEOREM 1. *Let* $\Lambda_b = (B'B)^{-1}B'$ *and* $\hat{\Lambda}_b = (\hat{B}'\hat{B})^{-1}\hat{B}'$. *Given* $K = o(\min\{p^{1/3}, T\})$ *and Assumptions 1, 2 and 3(i)–(iii), we have*

(a) $\|\hat{B} - B\| = O_p\{p^{1/2}(K^{3/2}p^{-1/2} + K^{1/2}T^{-1/2})\}$,
(b) $\|\hat{\Lambda}_b - \Lambda_b\| = O_p\{p^{-1/2}(K^{3/2}p^{-1/2} + K^{1/2}T^{-1/2})\}$.

Theorem 1 extends the existing consistency result for estimating the factor loadings (Lam et al., 2011; Fan et al., 2013, 2017) by pinpointing the effect of diverging K. Because the dimension p of factor loadings B is diverging, the estimation error $\hat{B} - B$ accumulates as p grows. For a p-dimensional vector whose entries are constantly 1, its spectral norm is $p^{1/2}$, which diverges to infinity. Thus, we should treat $p^{1/2}$ as the unit magnitude of the spectral norm of matrices with p rows, in which sense statement (a) of Theorem 1 justifies the estimation consistency of the factor loadings B. As the error term u_t shrinks as p grows under Assumption 3, the convergence rate of the factor loading estimation largely depends on p; a higher-dimensional predictor means a more accurate estimation. The convergence rate in this theorem can be further improved if we impose stronger assumptions on the negligibility of the error terms in the factor model (2).

Given $\hat{B}$, it is easy to see that $\hat{f}_t = \hat{\Lambda}_b B f_t + \hat{\Lambda}_b u_t$. Thus, together with the negligibility of the error term u_t, the consistency of $\hat{B}$ and $\hat{\Lambda}_b$ indicates the closeness between the true factors f_t and the estimated factors $\hat{f}_t$, of which the latter will be used in the subsequent sufficient dimension reduction. The error covariance matrix Σ_u can be estimated by thresholding the sample covariance matrix of the estimated residual $x_t - \hat{B}\hat{f}_t$, denoted by $\hat{\Sigma}_u = (\hat{\sigma}_{ij}^u)_{p\times p}$, as in Cai & Liu (2011), Xue et al. (2012) and Fan et al. (2013, 2016).

3. DIRECTIONAL REGRESSION BASED ON AN INVARIANCE RESULT

3.1. *An invariance result*

Had the true factors f_t been observed, directional regression would estimate the central subspace $\mathcal{S}_{y|f}$ as the column space of

$$M_{\mathrm{dr}} = E\{2\mathrm{var}(f_t) - E[(f_t - g_s)(f_t - g_s)' \mid y_{t+1}, \eta_{s+1}]\}^2, \tag{6}$$

where (g_s, η_{s+1}) is a hypothetical independent copy of (f_t, y_{t+1}). The term $\mathrm{var}(f_t)$ can be replaced with the identity matrix as in Li & Wang (2007), but we keep it in this form for convenience in the theoretical work developed later. For the resulting directions being included in $\mathcal{S}_{y|f}$, f_t needs to satisfy the following conditions:

Condition 1 (*Linearity*). $E(b'f_t \mid \phi_1'f_t, \ldots, \phi_L'f_t)$ is a linear function of $(\phi_1'f_t, \ldots, \phi_L'f_t)$ for any $b \in \mathbb{R}^K$;

Condition 2 (*Constant variance*). $\mathrm{var}(f_t \mid \phi_1'f_t, \ldots, \phi_L'f_t)$ is degenerate.

Since $\mathcal{S}_{y|f}$ is unknown, Conditions 1 and 2 are commonly strengthened such that they are satisfied for basis matrices of any L-dimensional subspace of $\mathbb{R}^K$. The strengthened conditions equivalently require the factors to be jointly normally distributed. To assess these conditions, one can treat f_t as the response and $(\phi_1'f_t, \ldots, \phi_L'f_t)$ as the predictor in regression; then, Condition 1 is

the linearity assumption on the regression function and Condition 2 is the homoscedasticity assumption on the error term. In this sense, we follow the convention in the literature of regression to treat Condition 2 as less worrisome than Condition 1 in practice. We tentatively assume Condition 1 and relax it in § 4.

Under general conditions, the column space of M_{dr} is L-dimensional, which, together with Conditions 1 and 2, means the exhaustive recovery of $\mathcal{S}_{y|f}$. These conditions are proposed in Li & Wang (2007) and reviewed in the Supplementary Material. They are weaker than those required for the exhaustiveness of sliced inverse regression, as more information about $f_t \mid y_{t+1}$, i.e., the second moment, is used. We assume these conditions throughout the paper, including § 4 where Condition 1 is violated.

To pinpoint the effect of using the estimated factors in directional regression, we next propose an invariance result for M_{dr}. As mentioned in § 1, a similar invariance result for sliced inverse regression can be found in Fan et al. (2017) where only the inverse first moment is involved; see their equation (2.6). To simplify the discussion, in the rest of the subsection we assume an oracle scenario where B is known a priori, which gives

$$\hat{f}_t = f_t + u_t^*, \tag{7}$$

where $u_t^* = \Lambda_b u_t$ is independent of f_t. Let u_s^* be an independent copy of u_t^* in (7) and let $\hat{g}_s = g_s + u_s^*$. Since B is known, $\hat{g}_s$ is an independent copy of $\hat{f}_t$.

LEMMA 1 (THE INVARIANCE RESULT). *Under Model (2), M_{dr} defined in (6) is invariant if the true factors f_t and g_s are replaced with the estimated factors $\hat{f}_t$ and $\hat{g}_s$.*

Using the estimated factors, one would naturally treat $\mathcal{S}_{y|\hat{f}}$ as the working parameter in the stage of sufficient dimension reduction. However, as no distributional assumptions are imposed on u_t^*, both Conditions 1 and 2 can be violated for $\hat{f}_t$, which causes inconsistency of directional regression for recovering $\mathcal{S}_{y|\hat{f}}$. In addition, $\mathcal{S}_{y|\hat{f}}$ itself may deviate from the parameter of interest $\mathcal{S}_{y|f}$, as the identity between the two essentially requires the normality of both f_t and u_t^* (Li & Yin, 2007). The invariance result provides the key to address these issues; that is, we can bypass $\mathcal{S}_{y|\hat{f}}$ and directly estimate $\mathcal{S}_{y|f}$ using the estimated factors, as if the true factors were used. As $\mathrm{var}(\hat{f}_t)$ is no longer the identity matrix, M_{dr} adopted here modifies its original form in Li & Wang (2007). This modification is crucial as it averages out the effect of the estimation error u_t^*. It also means that the column space of the working M_{dr} does differ from $\mathcal{S}_{y|\hat{f}}$.

3.2. *Consistency of directional regression*

In reality, the hypothetical independent copies (g_s, η_{s+1}) and (f_t, y_{t+1}) do not exist in the observed data, so we expand (6) and estimate an equivalent form of M_{dr},

$$\begin{aligned} M_{\mathrm{dr}} = {} & 2E[\{\mathrm{var}(f_t) - E(f_t f_t' \mid y_{t+1})\}^2] + 2E^2\{E(f_t \mid y_{t+1})E'(f_t \mid y_{t+1})\} \\ & + 2E\{E'(f_t \mid y_{t+1})E(f_t \mid y_{t+1})\} \cdot E\{E(f_t \mid y_{t+1})E'(f_t \mid y_{t+1})\}. \end{aligned} \tag{8}$$

By Lemma 1, we can replace f_t with $\hat{f}_t$, in which B is replaced with $\hat{B}$. For ease of estimation, in the sufficient dimension reduction literature it has been a common practice to employ the slicing technique: we partition the sample of y_{t+1} into H slices with equal sample proportion. In the population level, it corresponds to partitioning the support of y_{t+1} into H slices with equal probability, and using the corresponding indicator, denoted by y_{t+1}^D, as the new working response variable.

Because the slice indicator y_{t+1}^D is a measurable function of the original response y_{t+1}, f_t must affect y_{t+1}^D through y_{t+1}. Thus, the working central subspace $\mathcal{S}_{y^D|f}$ is always a subspace of the central subspace of interest $\mathcal{S}_{y|f}$. The two spaces further coincide for large H. Because the dimension L of $\mathcal{S}_{y|f}$ is fixed as K grows, without loss of generality we fix H as K grows and assume the identity between $\mathcal{S}_{y^D|f}$ and $\mathcal{S}_{y|f}$. Such identity is confirmed by an omitted simulation study that shows the robustness of the proposed method to the choice of H, for a reasonable range of H, e.g., from three to ten. The same phenomenon has also been commonly observed in the literature (Li, 1991; Li & Wang, 2007).

Using y_{t+1}^D, the inverse moments $E(\hat{f}_t \mid y_{t+1})$ and $E(\hat{f}_t\hat{f}_t' \mid y_{t+1})$ in M_{dr} become the marginal moments of $\hat{f}_t$ within each slice, and can be estimated by the usual sample moments. Hence, the slicing technique simplifies the estimation. In detail, the implementation of Step 2 is as follows: Let $y_{(0)/H} = -\infty$, and, for $i = 1, \ldots, H$, let $y_{(i)/H}$ be the (i/H)th quantile of $\{y_1, \ldots, y_T\}$. Let $y_{t+1}^D = i$ if $y_{t+1} \in (y_{(i)/H}, y_{(i+1)/H}]$. Estimate $E(\hat{f}_t \mid y_{t+1}^D = i)$ by $\sum_{t=1}^T \hat{f}_t I(y_{t+1}^D = i)/(T/H)$ and $E(\hat{f}_t\hat{f}_t' \mid y_{t+1}^D = i)$ by $\sum_{t=1}^T \hat{f}_t\hat{f}_t' I(y_{t+1}^D = i)/(T/H)$. Estimate $\mathrm{var}(\hat{f}_t)$ by I_K. Plug these into (8) to derive $\hat{M}_{\mathrm{dr}}$. Estimate $\mathcal{S}_{y|f}$ by the space spanned by $(\hat{\phi}_1, \ldots, \hat{\phi}_L)$, the leading L eigenvectors of $\hat{M}_{\mathrm{dr}}$.

To estimate $\mathrm{var}(\hat{f}_t)$ in (8), one can alternatively use $I_K + \hat{\Sigma}_{u^*}$ by the restriction $\mathrm{var}(f_t) = I_K$, where $\hat{\Sigma}_{u^*}$ is the thresholding covariance estimator. An omitted simulation study shows that the resulting estimator of M_{dr} performs similarly.

THEOREM 2. *Suppose* $K = o\{\min(p^{1/3}, T^{1/2})\}$. *Under Assumptions* 1, 2 *and* 3*(i)–(iii), and Conditions* 1 *and* 2, $(\hat{\phi}_1, \ldots, \hat{\phi}_L)$ *span a consistent estimator of* $\mathcal{S}_{y|f}$ *in the sense that*

$$\|(\hat{\phi}_1, \ldots, \hat{\phi}_L)(\hat{\phi}_1, \ldots, \hat{\phi}_L)' - (\phi_1, \ldots, \phi_L)(\phi_1, \ldots, \phi_L)'\|_{\mathrm{F}} = O_P(K^{3/2}p^{-1/2} + KT^{-1/2}).$$

In connection with Theorem 1, this theorem justifies that the estimation error of $\mathcal{S}_{y|f}$ comes from two parts. The first part, which is of order $O_P(K^{3/2}p^{-1/2})$, is inherited from factor analysis. This part represents the price we pay for estimating the factor loadings B, and it depends on the dimension p of the original predictor. By contrast, the second part, which is of order $O_P(KT^{-1/2})$, does not depend on p and is newly generated in the sufficient dimension reduction stage. From the proof of Theorem 2, it represents the price we pay for estimating the unknown inverse second moment involved in the kernel matrix. Therefore, this part would persist even if no error were generated in factor analysis.

4. RELAXING THE LINEARITY CONDITION

As mentioned in § 3, Condition 1 can be regarded as a parametric assumption and can be violated in real applications. For example, this occurs when one incorporates the lag variables of y_{t+1} in forecasting and considers Model (4). In this section we address this issue in two ways: first, we justify the consistency of the proposed method without Condition 1 under the setting that the number of factors K must diverge; second, we weaken Condition 1 and generalize the proposed method accordingly following the spirit of Dong & Li (2010) under the setting that K is fixed.

When Condition 1 is violated, Theorem 2 still holds if we treat $(\phi_1, \ldots, \phi_L)$ as the L leading eigenvectors of M_{dr}. Thus, the consistency of the proposed methodology depends on the closeness between the column space of M_{dr} and the central subspace $\mathcal{S}_{y|f}$, which hinges on the approximation of Condition 1. Fortunately, the latter has been justified in Hall & Li (1993) for all large K.

THEOREM 3. *Suppose* $K \to \infty$ *and* $K = o\{\min(p^{1/3}, T^{1/2})\}$. *Under Assumptions* 1, 2 *and* 3*(i)–(iii), Condition* 2, *and other regularity conditions in the Supplementary Material,* $\hat{\phi}_1, \ldots, \hat{\phi}_L$ *span a consistent estimator of* $\mathcal{S}_{y|f}$ *in the sense that*

$$\|(\hat{\phi}_1, \ldots, \hat{\phi}_L)(\hat{\phi}_1, \ldots, \hat{\phi}_L)' - (\phi_1, \ldots, \phi_L)(\phi_1, \ldots, \phi_L)'\|_{\mathrm{F}} = o_P(1).$$

In the literature, the Hall & Li (1993) result on the approximation of Condition 1 was used heuristically to support the effectiveness of inverse moment methods when Condition 1 is violated; see, for example, Cook & Weisberg (1991) and Li & Wang (2007). As far as we are aware, this is the first attempt to rigorously build the consistency of inverse moment methods using the Hall & Li (1993) result.

When K is small and the factors clearly violate Condition 1, the approximation result in Hall & Li (1993) no longer applies. In this case, we treat K as fixed, and relax Condition 1 to:

Condition 1′ $E(f_t \mid \phi_1' f_t, \ldots, \phi_L' f_t)$ is a linear combination of $\{h_i(\phi_1' f_t, \ldots, \phi_L' f_t) : i = 1, \ldots, q\}$.

One can set the basis functions in Condition 1′ to be power functions, trigonometric functions, etc. In addition to Condition 1′ we require Condition 2, which, as mentioned in § 1, is quite mild. These conditions closely resemble those in Dong & Li (2010). We generalize directional regression from the eigendecomposition of M_{dr} to minimizing

$$\begin{aligned}\kappa(\psi_1, \ldots, \psi_L) = E(2I_p - E\{(f_t - g_s)^{\otimes 2} \mid y_{t+1}, \eta_{s+1}\} - 2E\{E^{\otimes 2}(f_t \mid \psi_1' f_t, \ldots, \psi_L' f_t)\} \\ + E[\{E(f_t \mid \psi_1' f_t, \ldots, \psi_L' f_t) - E(g_s \mid \psi_1' g_s, \ldots, \psi_L' g_s)\}^{\otimes 2} \mid y_{t+1}, \eta_{s+1}])^{\otimes 2}\end{aligned}$$

over all the semiorthogonal matrices $(\psi_1, \ldots, \psi_L)$, where $v^{\otimes 2}$ denotes vv' for any real vector v and $E(f_t \mid \psi_1' f_t, \ldots, \psi_L' f_t)$ is modelled parametrically as if Condition 1′ held for $(\psi_1, \ldots, \psi_L)$. Using the estimated factors $\hat{f}_t$ and $\hat{g}_s$ and the slicing strategy, we can similarly construct $\hat{\kappa}(\cdot)$.

Under fairly general assumptions (Dong & Li, 2010), there exists the unique minimizer of $\kappa(\cdot)$ up to orthogonal column transformations, which spans the central subspace $\mathcal{S}_{y|f}$; we omit these assumptions here. Intuitively, a minimizer of $\hat{\kappa}(\cdot)$ spans a consistent estimator of $\mathcal{S}_{y|f}$.

THEOREM 4. *Let* $(\hat{\phi}_1, \ldots, \hat{\phi}_L)$ *denote any minimizer of* $\hat{\kappa}(\psi_1, \ldots, \psi_L)$. *Under Assumptions* 1–3 *and Conditions* 1′ *and* 2, *we have*

$$\|(\hat{\phi}_1, \ldots, \hat{\phi}_L)(\hat{\phi}_1, \ldots, \hat{\phi}_L)' - (\phi_1, \ldots, \phi_L)(\phi_1, \ldots, \phi_L)'\|_{\mathrm{F}} = O_P(p^{-1/2} + T^{-1/2}).$$

By Theorems 3 and 4, we can apply the proposed forecasting method or its generalization without concerning Condition 1, for both fixed and diverging K. For example, we now allow the predictor x_t, as well as the factors f_t, to contain discrete components.

5. DETERMINING K AND L

We now discuss how to determine the number of factors K and the dimension L of the central subspace $\mathcal{S}_{y|f}$. The problem is commonly called order determination in the literature of dimension reduction (Luo & Li, 2016).

In the literature, various order-determination methods have been proposed to estimate K, including Bai & Ng (2002, 2008), Ludvigson & Ng (2009), Onatski (2010),

Ahn & Horenstein (2013), and Jurado et al. (2015). Recently, Li et al. (2017) extended Bai & Ng's approach to the case of diverging K, and estimated K by

$$\hat{K} = \arg \min_{0 \leqslant k \leqslant K_{\max}} \log(p^{-1}T^{-1}\|X - T^{-1}X\hat{F}_k\hat{F}_k'\|_{\mathrm{F}}^2) + k \cdot q(p,T),$$

where $K_{\max}$ is a prescribed upper bound that possibly increases with p and T, and $\hat{F}_k$ denotes the solution to (5) with k being the working number of factors; $q(p,T)$ is a penalty function such that $q(p,T) = o(1)$ and $(K_{\max}^6/p + K_{\max}^4/T)^{-1}q(p,T) \to \infty$. We adopt the Li et al. (2017) approach, and follow their suggestion to take $q(p,T) = (p+T)(pT)^{-1}\log\{pT(p+T)^{-1}\}$.

To estimate the dimension L of the central subspace $\mathcal{S}_{y|f}$, multiple methods have been proposed, including sequential tests (Li, 1991; Li & Wang, 2007), the bootstrap procedure (Ye & Weiss, 2003), the cross-validation method (Xia et al., 2002; Wang & Xia, 2008), the BIC-type procedure (Zhu et al., 2006) and the ladle estimator (Luo & Li, 2016), among which we adopt the BIC-type procedure and extend it to the high-dimensional case. For a positive semidefinite matrix parameter M whose columns span $\mathcal{S}_{y|f}$ and its sample estimator $\hat{M}$, let $\{\lambda_1, \ldots, \lambda_K\}$ and $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_K\}$ be their respective eigenvalues in descending order. By definition, λ_L must be positive. We introduce a constant $c \in (0,1)$ and set K_c, the nearest integer to cK, as an upper bound of L. This is reasonable because L is fixed and usually small in practice. We modify the objective function in Zhu et al. (2006) to $G : \{1, \ldots, K_c\} \to \mathbb{R}$ with

$$G(l) = (T/2)\textstyle\sum_{i=1+\min(\tau,l)}^{K_c}\{\log(\hat{\lambda}_i + 1) - \hat{\lambda}_i\} - C_T l(2K - l + 1)/2, \tag{9}$$

where τ is the number of positive $\hat{\lambda}_i$. We then estimate L as the maximizer $\hat{L}$ of $G(\cdot)$. Due to the introduction of the nontrivial upper bound K_c, we do not need to impose additional constraints on K or $\|\hat{M} - M\|$ for the consistency of $\hat{L}$. This improves the result in Zhu et al. (2006).

THEOREM 5. *Suppose* $\|\hat{M} - M\| = o_P(1)$. *If* C_T *satisfies* $C_T K T^{-1} \to 0$ *and* $\|\hat{M} - M\|^2 = o_P(C_T K T^{-1})$, *then* $\hat{L}$ *converges to* L *in probability.*

A candidate for C_T is $K^{-1}T\|\hat{M} - M\|$. Referring to Theorem 2, if we apply the BIC-type procedure to directional regression, then we can choose C_T to be $K^{1/2}p^{-1/2}T + T^{1/2}$.

6. SIMULATION STUDIES

We now present a numerical example to illustrate the performance of the proposed forecasting method that uses directional regression in the sufficient dimension reduction stage. The data-generating process is specified as the following:

$$y_{t+1} = g(\phi_1'f_t, \phi_2'f_t) + \sigma\epsilon_{t+1}, \qquad x_{it} = b_i'f_t + u_{it}.$$

We fix $\phi_1 = (1,1,1,0_{K-3}')/\sqrt{3}$, $\phi_2 = (1,0_{K-3}',1,3)/\sqrt{11}$. Following Li et al. (2017), we set the number of factors K to increase with p in the form $K = [1.5\log(p)]$, where $[x]$ denotes the integer part of a real number x. The factor loadings b_i are independently sampled from $U[-1,2]$. We generate the latent factors $f_{j,t}$ and the error terms u_{it} from two AR(1) processes, $f_{j,t} = \alpha_j f_{j,t-1} + e_{jt}$ and $u_{it} = \rho_i u_{i,t-1} + v_{it}$, with α_j, ρ_i drawn from $U[0.2, 0.8]$ and fixed during the simulation; the noises e_{jt}, v_{it}, are $N(0,1)$. We set $\epsilon_{t+1} \sim N(0,1)$ and $\sigma = 0.2$.

We consider four different choices of the link function $g(\cdot)$:

Model I: $y_{t+1} = 0.4(\phi_1'f_t)^2 + 3\sin(\phi_2'f_t/4) + \sigma\epsilon_{t+1}$;

Table 1. *Performance of estimated $\hat{\phi}$ using median $R^2(\hat{\phi})$ (%) with standard deviations in parentheses over 1000 replications*

		SIR		DR		SEE	
p	T	$R^2(\hat{\phi}_1)$	$R^2(\hat{\phi}_2)$	$R^2(\hat{\phi}_1)$	$R^2(\hat{\phi}_2)$	$R^2(\hat{\phi}_1)$	$R^2(\hat{\phi}_2)$
Model I							
100	100	75.0 (21.3)	28.4 (27.4)	82.9 (14.8)	79.9 (21.9)	80.4 (26.8)	27.0 (23.5)
100	200	88.7 (10.4)	17.7 (27.6)	94.5 (5.4)	91.5 (8.5)	83.4 (26.6)	21.7 (20.7)
100	500	95.9 (3.6)	14.4 (28.2)	98.4 (1.4)	96.0 (3.4)	87.6 (26.9)	30.8 (20.7)
200	100	63.2 (24.5)	26.6 (24.8)	74.6 (20.3)	67.9 (24.4)	40.9 (23.4)	13.0 (18.5)
500	200	76.6 (16.1)	16.1 (23.2)	86.8 (20.1)	80.2 (22.1)	26.6 (15.4)	9.4 (15.8)
500	500	90.5 (5.5)	9.2 (22.4)	96.0 (29.9)	87.7 (26.0)	24.2 (13.4)	7.6 (13.5)
Model II							
100	100	95.8 (3.5)	21.0 (25.7)	95.8 (3.5)	26.4 (26.6)	89.7 (22.5)	33.0 (20.1)
100	200	97.8 (1.8)	32.4 (27.7)	97.9 (1.8)	43.4 (28.7)	90.4 (15.0)	30.2 (19.5)
100	500	99.1 (0.7)	63.8 (27.0)	99.1 (0.7)	74.8 (23.8)	91.9 (20.5)	48.7 (20.9)
200	100	94.6 (3.6)	17.6 (22.4)	94.2 (10.6)	21.4 (23.4)	81.6 (26.8)	21.2 (18.7)
500	200	95.9 (2.1)	18.2 (22.6)	95.5 (11.9)	24.7 (23.4)	37.8 (26.5)	13.9 (17.4)
500	500	98.4 (0.9)	41.1 (25.6)	97.9 (15.2)	48.3 (26.3)	30.7 (24.9)	13.1 (17.1)
Model III							
100	100	33.4 (26.7)	26.1 (23.4)	83.0 (19.7)	47.6 (28.2)	40.1 (30.7)	29.9 (18.4)
100	200	34.8 (27.3)	23.8 (22.7)	94.9 (4.1)	83.2 (22.9)	68.4 (35.1)	20.2 (18.1)
100	500	33.0 (28.1)	24.2 (23.4)	98.4 (1.4)	97.6 (2.1)	77.2 (34.6)	21.5 (16.7)
200	100	29.5 (25.9)	19.8 (20.4)	75.0 (23.3)	36.5 (25.7)	37.9 (26.8)	12.9 (17.9)
500	200	20.3 (23.7)	15.2 (10.1)	88.9 (22.2)	48.8 (27.8)	20.5 (16.1)	8.6 (14.5)
500	500	21.3 (23.1)	14.5 (18.1)	95.6 (29.6)	92.9 (28.0)	14.0 (13.5)	6.6 (13.7)
Model IV							
100	100	61.8 (29.1)	31.3 (26.0)	85.6 (14.2)	79.1 (23.5)	64.4 (27.8)	43.9 (18.4)
100	200	75.1 (26.4)	41.6 (27.9)	94.5 (4.9)	93.5 (5.2)	71.7 (34.1)	51.1 (19.4)
100	500	89.4 (15.0)	67.8 (27.4)	98.1 (1.7)	97.7 (1.9)	88.2 (37.7)	66.6 (17.0)
200	100	51.9 (28.6)	29.0 (24.8)	79.6 (19.7)	71.0 (24.3)	41.5 (25.9)	12.2 (18.2)
500	200	59.4 (27.9)	30.2 (24.4)	87.5 (21.7)	86.2 (20.2)	19.5 (15.4)	7.4 (13.5)
500	500	83.3 (17.8)	54.9 (26.9)	95.1 (28.3)	94.6 (26.4)	10.2 (13.3)	4.8 (13.1)

SIR, sliced inverse regression; DR, directional regression; SEE, semiparametric efficient estimator.

Model II: $y_{t+1} = 3\sin(\phi_1' f_t/4) + 3\sin(\phi_2' f_t/4) + \sigma\epsilon_{t+1}$;
Model III: $y_{t+1} = 0.4(\phi_1' f_t)^2 + |\phi_2' f_t|^{1/2} + \sigma\epsilon_{t+1}$;
Model IV: $y_{t+1} = (\phi_1' f_t)(\phi_2' f_t + 1) + \sigma\epsilon_{t+1}$.

The proposed forecasting by directional regression is compared with the forecasting by sliced inverse regression (Fan et al., 2017), the linear principal components estimator, and the semiparametric efficient estimator proposed by Ma & Zhu (2013). Models I and III include at least one symmetric component, which cannot be estimated well by sliced inverse regression. Model II is favourable to sliced inverse regression. Model IV contains the interaction component to examine the ability of each method to detect such nonlinear effects.

To gauge the quality of the estimated directions, we adopt the squared multiple correlation coefficient $R^2(\hat{\phi}) = \max_{\phi \in \mathcal{S}_{y|f}, \|\phi\|=1} (\phi' \hat{\phi})^2$, where $\mathcal{S}_{y|f}$ is spanned by ϕ_1 and ϕ_2. We ensure that the true factors and loadings meet the identifiability conditions by calculating H such that $T^{-1}HF'FH' = I_K$ and $H^{-1}B'BH^{-1}$ is diagonal. The rotated central subspace is then understood as $H^{-1}\mathcal{S}_{y|f}$, which is still denoted as $\mathcal{S}_{y|f}$, see Fan et al. (2017).

Table 2. *Comparison of out-of-sample median R^2 in percentage (%) over 1000 replications*

p	T	SIR	DR	PC	SEE	SIR	DR	PC	SEE
			Model I				Model II		
100	100	−11.7	28.8	−0.4	1.4	94.6	94.8	93.3	78.0
100	200	−3.9	72.1	18.0	9.9	95.7	95.8	94.6	79.4
100	500	0.4	92.2	27.4	11.4	96.1	96.2	94.9	79.3
200	100	−11.4	18.6	−6.9	−4.0	95.3	95.6	94.2	61.8
500	200	−5.3	57.5	−1.1	−0.7	96.2	96.5	94.8	45.9
500	500	−0.9	91.4	13.8	1.7	97.1	97.1	95.8	45.4
			Model III				Model IV		
100	100	−9.4	34.8	17.8	17.2	−0.2	23.6	21.2	18.4
100	200	1.0	77.1	30.8	22.7	13.5	53.7	35.8	28.4
100	500	5.2	90.5	38.0	25.5	29.6	57.3	43.2	30.8
200	100	−9.7	21.5	3.8	6.3	−2.3	16.9	6.8	7.6
500	200	−4.4	62.5	6.6	2.6	5.6	46.0	9.7	5.3
500	500	−1.3	89.5	19.1	5.2	22.4	58.3	21.6	48.5

SIR, sliced inverse regression; DR, directional regression; PC, principal components; SEE, semiparametric efficient estimator.

Table 1 compares the estimation of sliced inverse regression and directional regression in simulation studies, where the linear principal components estimator is omitted, as it produces only one directional estimate. It is evident that directional regression has substantial improvement over sliced inverse regression in Models I, III and IV, with higher $R^2(\hat{\phi})$ and lower variance. This is not surprising as directional regression explores higher conditional moments, and hence incorporates more information. The semiparametric efficient estimator is slightly better than sliced inverse regression in these cases, but it also fails to capture ϕ_2 accurately, partially due to its semiparametric nature, which typically requires lengthy steps to converge. In Model II, sliced inverse regression, directional regression and the semiparametric efficient estimator yield comparable results. We also observe that directional regression has outstanding performance in small samples, which makes it favourable in practice.

We next investigate the predictive power of directional regression through the out-of-sample R^2, i.e., $R^2 = 1 - \sum_{t=T+1}^{T+n_T}(y_t - \hat{y}_t)^2 / \sum_{t=T+1}^{T+n_T}(y_t - \bar{y}_t)^2$, where we use a fixed length $n_T = 50$ of testing samples to evaluate the out-of-sample performance, and $\hat{y}_t$ is the predicted value using all information prior to t. The fitting is done by building an additive model in Step 3 of the proposed estimator. In the case of the principal components estimator, $\hat{K}$ smooth functions are constructed for the estimated factors. In contrast, only $\hat{L}$ smooth functions are applied in the cases of sliced inverse regression, directional regression and the semiparametric efficient estimator. $\hat{K}$ and $\hat{L}$ are obtained using the procedures introduced in § 5. It is clear from Table 2 that directional regression enjoys great performance in almost all the cases. Similar to directional regression, the semiparametric efficient estimator is better than sliced inverse regression, as it explores the structural dimension more thoroughly with different forms of the target. However, the semiparametric efficient estimator is often limited to a large sample size to produce accurate estimation. The principal components estimator is more robust in the presence of symmetric components, but fails to capture the interaction effect in general. To investigate the accuracy of $\hat{K}$ and $\hat{L}$ used above, which are obtained from § 5, we carry out simulations to investigate the accuracy of the estimation procedures, and examine the sensitivity of forecasting performance with respect to $\hat{K}$ and $\hat{L}$. In addition, we conduct experiments to show the effectiveness of the proposed method when the linearity condition is violated for factors f_t. Due to space limitations, these numerical results are presented in the Supplementary Material.

Table 3. *Root mean squared error in out-of-sample forecast (median/max/min) relative to the linear diffusion index. In each group, the median, maximum and minimum of the root mean squared error is reported*

	SIR(1)	SIR(2)	DR(1)	DR(2)	NL-PC
Group ($h = 1$)					
Output and income	1.03/1.61/0.96	1.02/1.13/0.94	0.99/1.19/0.92	1.02/1.14/0.90	1.21/1.38/1.05
Consumption	1.00/2.10/0.80	0.95/1.05/0.74	0.92/1.02/0.86	1.00/1.05/0.81	1.16/1.44/1.04
Labour market	1.02/2.27/0.71	1.00/1.21/0.42	0.97/1.13/0.52	0.98/1.16/0.42	1.21/1.53/0.46
Housing	1.04/1.32/0.64	0.92/1.08/0.52	0.83/1.04/0.50	0.79/0.94/0.44	0.83/0.97/0.49
Money and credit	0.94/1.04/0.86	0.97/1.05/0.90	0.96/1.10/0.86	1.04/1.24/0.92	1.14/1.41/1.07
Stock market	0.99/1.39/0.90	1.02/1.12/0.83	0.92/1.08/0.88	1.04/1.07/0.91	1.36/1.39/1.14
Interest rates	1.04/1.79/0.79	0.93/1.17/0.61	0.90/1.04/0.59	0.92/1.15/0.62	1.12/1.32/0.73
Prices	0.97/1.42/0.80	0.99/1.05/0.83	0.95/1.12/0.81	0.97/1.12/0.88	1.12/1.47/0.92
Group ($h = 6$)					
Output and income	1.07/1.47/0.93	0.97/1.23/0.81	0.99/1.18/0.89	1.05/1.27/0.95	1.28/1.52/0.97
Consumption	1.16/1.73/0.90	0.90/1.12/0.67	0.94/1.16/0.71	1.03/1.14/0.73	1.28/1.66/0.77
Labour market	1.15/2.02/0.68	0.89/1.22/0.39	0.90/1.26/0.48	0.98/1.39/0.43	1.24/1.42/0.45
Housing	0.96/1.29/0.66	0.85/0.95/0.51	0.73/0.89/0.50	0.69/0.86/0.47	0.78/1.02/0.55
Money and credit	0.95/3.51/0.76	1.01/3.65/0.83	0.99/1.52/0.76	1.02/1.74/0.78	1.23/2.90/0.92
Stock market	0.91/1.20/0.83	0.94/1.05/0.89	0.89/1.08/0.84	1.00/1.03/0.94	1.23/1.27/0.83
Interest rates	1.01/1.61/0.75	0.90/1.12/0.64	0.84/1.13/0.50	0.88/1.18/0.58	1.11/1.46/0.70
Prices	1.16/1.37/0.51	1.03/1.12/0.82	1.11/1.37/0.94	1.14/1.36/0.95	1.17/1.35/1.11
Group ($h = 12$)					
Output and income	1.24/1.67/0.79	1.01/1.45/0.76	0.99/1.22/0.76	1.01/1.36/0.86	1.17/1.34/0.92
Consumption	1.27/1.60/0.83	1.08/1.44/0.62	1.09/1.32/0.65	1.06/1.38/0.66	1.16/1.38/0.87
Labour market	1.07/1.76/0.67	0.83/1.40/0.41	0.91/1.44/0.54	0.89/1.41/0.46	1.13/1.39/0.56
Housing	0.85/1.35/0.59	0.69/0.93/0.46	0.67/0.91/0.40	0.68/0.83/0.36	0.89/1.16/0.54
Money and credit	1.14/2.03/0.41	1.03/2.16/0.80	1.05/1.52/0.85	1.00/1.40/0.82	1.20/1.69/0.87
Stock market	1.09/1.20/0.89	1.01/1.13/0.84	0.96/1.17/0.94	1.08/1.16/0.75	1.06/1.14/0.89
Interest rates	1.00/1.31/0.75	0.82/1.22/0.59	0.80/1.27/0.53	0.85/1.18/0.51	1.07/1.62/0.70
Prices	1.18/1.40/0.53	1.21/1.40/0.66	1.19/1.31/0.71	1.21/1.33/0.77	1.25/1.52/0.94

SIR(i), sufficient forecasting using i indices; DR, sufficient directional forecasting; NL-PC, a nonlinear additive model on all the estimated factors.

7. MACRO INDEX FORECAST

We now analyse how the diffusion indices constructed by the proposed directional regression affect real-data forecasts. We use a monthly macro dataset consisting of 134 macroeconomic time series recently composed by McCracken & Ng (2016), which are classified into eight groups: (i) output and income, (ii) labour market, (iii) housing, (iv) consumption, orders and inventories, (v) money and credit, (vi) bond and exchange rates, (vii) prices, and (viii) stock market. The dataset spans from 1959 to 2016. For a given target time series, we model the multi-step-ahead variable as $y^h_{t+h} = g(\phi_1' f_t, \ldots, \phi_L' f_t) + \epsilon^h_{t+h}$, where $y^h_{t+h} = h^{-1} \sum_{i=1}^h y_{t+i}$ is the variable to forecast, as in Stock & Watson (2002a).

We follow McCracken & Ng (2016) to pre-process the data. We also employ the Ljung–Box test with various lags to test for uncorrelatedness in residuals, which suggests the appropriateness of using our proposed methods. Forecasts of y^h_{t+h} are constructed based on a moving window with fixed length ($T = 120$) to account for timeliness. For each fixed window, the factors in the forecasting equation are estimated by the method of principal components using all time series except the target. As noted by McCracken & Ng (2016), eight factors have good explanatory power in various

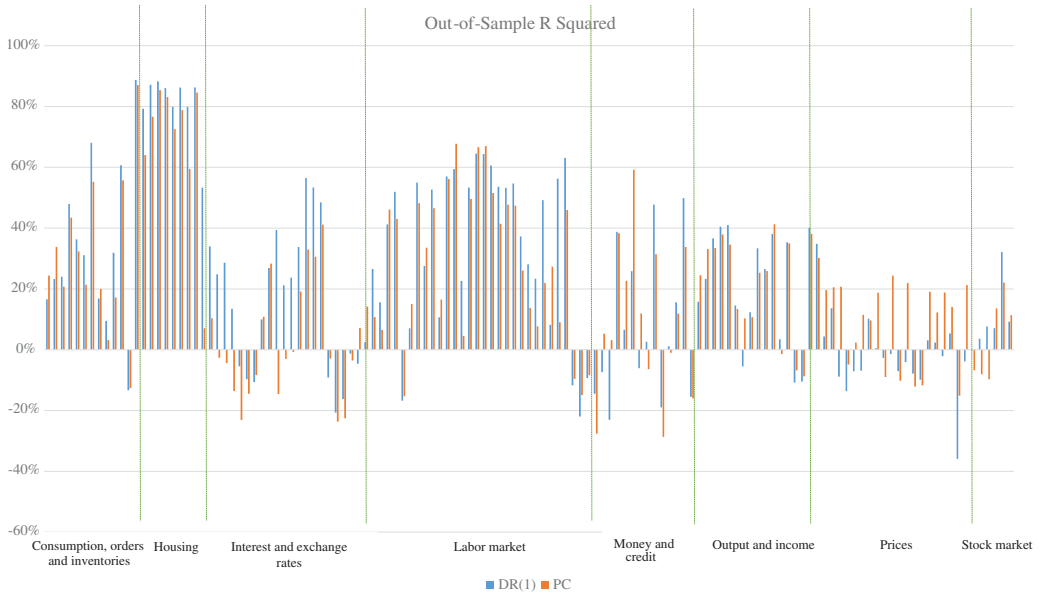

Fig. 1. Six-month-ahead forecasting out-of-sample R^2 for the 134 macroeconomic series organized into eight groups.

cases, so we set $K = 8$ throughout the exercise. For each method M, we compare out-of-sample forecasting performances using the relative mean squared error to the principal components estimator method, $\text{RMSE}(M) = \text{MSE}(M) \,/\, \text{MSE}(\text{PC})$, where $\text{MSE}(M) = m^{-1} \sum_{t=T+1}^{T+m} (y_t - \hat{y}_t)^2$, which we evaluate on the last $m = 240$ months (20 years). The methods we consider here include sliced inverse regression and directional regression with sufficient forecasting with $L = i$. Both methods use an additive model in specifying the forecasting equation. We also impose an additive model to the estimated factors, denoted by NL-PC, to see how much we can leverage on the nonlinearity without projecting principal components.

We report the results in Table 3 for $h = 1, 6, 12$ on the maximum, minimum and median of the root mean squared error in each broad sector. Several features are noteworthy. First, a nonlinear additive model built on estimated factors does not buy us more predictive power, except in the housing sector, where most of the nonlinear methods improve prediction accuracy. Second, the one-step-ahead out-of-sample forecast favours DR(1), as we observe the median root mean squared errors are uniformly less than 1 and some of the reductions in the root mean squared error are substantial. Moving from short horizon to long horizon changes the predictability of the targets, but DR(1) manages to improve the forecast over the principal components method in many instances. Finally, as an illustration, we plot the out-of-sample R^2 for the six-month-ahead forecast using DR(1) and principal components in Fig. 1. Notably, macro time series in the housing and labour market sectors have higher predictability than in the rates and stock market sectors.

Acknowledgement

The authors thank the editor, the associate editor and the referees for their helpful comments and constructive suggestions. Luo was supported in part by the National Science Foundation of China (12001484). Xue and Yu were supported in part by the National Science Foundation (DMS-1811552, DMS-1953189 and CCF-2007823). All authors contributed equally to this work.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes proofs of the theoretical results.

REFERENCES

AHN, S. C. & HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–27.
BAI, J. & NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.
BAI, J. & NG, S. (2008). Forecasting economic time series using targeted predictors. *J. Economet.* **146**, 304–17.
BERNANKE, B., BOIVIN, J. & ELIASZ, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Quart. J. Econ.* **120**, 387–422.
CAI, T. & LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Assoc.* **106**, 672–84.
COOK, R. D. (1998). *Regression Graphics*. New York: Wiley.
COOK, R. D. & WEISBERG, S. (1991). Comment on 'Sliced inverse regression for dimension reduction (with discussion)'. *J. Am. Statist. Assoc.* **86**, 328–32.
DONG, Y. & LI, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika*, **97**, 279–94.
FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modelling And Its Applications*. Boca Raton, FL: CRC Press.
FAN, J., LIAO, Y. & MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Statist. Soc.* B **75**, 603–80.
FAN, J., XUE, L. & YAO, J. (2017). Sufficient forecasting using factor models. *J. Economet.* **201**, 292–306.
FAN, J., XUE, L. & ZOU, H. (2016). Multitask quantile regression under the transnormal model. *J. Am. Statist. Assoc.* **111**, 1726–35.
HALL, P. & LI, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867–89.
JURADO, K., LUDVIGSON, S. C. & NG, S. (2015). Measuring uncertainty. *Am. Econ. Rev.* **105**, 1177–216.
LAM, C. & YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40**, 694–726.
LAM, C., YAO, Q. & BATHIA, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, 901–18.
LI, B. & WANG, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **102**, 997–1008.
LI, B. & YIN, X. (2007). On surrogate dimension reduction for measurement error regression: An invariance law. *Ann. Statist.* **35**, 2143–72.
LI, H., LI, Q. & SHI, Y. (2017). Determining the number of factors when the number of factors can increase with sample size. *J. Economet.* **197**, 76–86.
LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316–27.
LINTNER, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Rev. Econ. Statist.* **47**, 13–37.
LUDVIGSON, S. & NG, S. (2007). The empirical risk return relation: A factor analysis approach. *J. Finan. Econ.* **83**, 171–222.
LUDVIGSON, S. & NG, S. (2009). Macro factors in bond risk premia. *Rev. Finan. Studies* **22**, 5027–67.
LUO, W. & LI, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–87.
MA, Y. & ZHU, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41**, 250–68.
MATZKIN, R. L. (2002). Nonparametric estimation of nonadditive random functions. *Econometrica* **71**, 1339–75.
MCCRACKEN, M. W. & NG, S. (2016). FRED-MD: A monthly database for macroeconomic research. *J. Bus. Econ. Statist.* **34**, 574–89.
ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Statist.* **92**, 1004–16.
SHARPE, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finan.* **19**, 425–42.
STOCK, J. H. & WATSON, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroecon. Ann.* **4**, 351–409.
STOCK, J. H. & WATSON, M. W. (2002a). Forecasting using principal components from a large number of predictors. *J. Am. Statist. Assoc.* **97**, 1167–79.
STOCK, J. H. & WATSON, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econ. Statist.* **20**, 147–62.
WANG, H. & XIA, Y. (2008). Sliced regression for dimension reduction. *J. Am. Statist. Assoc.* **103**, 811–21.

Xia, Y., Tong, H., Li, W. & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space (with discussion). *J. R. Statist. Soc.* B **64**, 363–410.

Xue, L., Ma, S. & Zou, H. (2012). Positive definite ℓ_1 penalized estimation of large covariance matrices. *J. Am. Statist. Assoc.* **107**, 1480–91.

Ye, Z. & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–79.

Yu, X., Yao, J. & Xue, L. (2021). Nonparametric estimation and conformal inference of the sufficient forecasting with a diverging number of factors. *J. Bus. Econ. Statist.* **40**, 342–54.

Zhu, L., Miao, B. & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Am. Statist. Assoc.* **101**, 630–43.

[*Received on* 21 *April* 2020. *Editorial decision on* 28 *May* 2021]