

Scaled PCA: A New Approach to Dimension Reduction

Dashan Huang,^a Fuwei Jiang,^b Kunpeng Li,^c Guoshi Tong,^{d,*} Guofu Zhou^e

^aLee Kong Chian School of Business, Singapore Management University, 178899, Singapore; ^bSchool of Finance, Central University of Finance and Economics, 102206 China; ^cInternational School of Economics and Management, Capital University of Economics and Business, 100070 China; ^dFanhai International School of Finance, Fudan University, 200001 China; ^eOlin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

*Corresponding author

Contact: dashanhuang@smu.edu.sg (DH); jfuwei@gmail.com,  <https://orcid.org/0000-0003-1939-1075> (FJ); kunpenglithu@126.com (KL); gstong_academic@126.com (GT); zhou@wustl.edu,  <https://orcid.org/0000-0002-4104-403X> (GZ)

Received: April 1, 2020

Revised: November 12, 2020; January 22, 2021

Accepted: February 9, 2021

Published Online in Articles in Advance: June 1, 2021

<https://doi.org/10.1287/mnsc.2021.4020>

Copyright: © 2021 INFORMS

Abstract. This paper proposes a novel supervised learning technique for forecasting: scaled principal component analysis (sPCA). The sPCA improves the traditional principal component analysis (PCA) by scaling each predictor with its predictive slope on the target to be forecasted. Unlike the PCA that maximizes the common variation of the predictors, the sPCA assigns more weight to those predictors with stronger forecasting power. In a general factor framework, we show that, under some appropriate conditions on data, the sPCA forecast beats the PCA forecast, and when these conditions break down, extensive simulations indicate that the sPCA still has a large chance to outperform the PCA. A real data example on macroeconomic forecasting shows that the sPCA has better performance in general.

History: Accepted by Kay Giesecke, *Management Science* Special Section on Data-Driven Prescriptive Analytics.

Funding: F. Jiang acknowledges financial support from the National Natural Science Foundation of China [Grants 71872195, 72072193, and 71602198]. K. Li gratefully acknowledges financial support from the National Natural Science Foundation of China [Grants 71722011, 71571122, 71991470, and 71991471].

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/mnsc.2021.4020>.

Keywords: forecasting • PCA • big data • dimension reduction • machine learning

1. Introduction

Principal component analysis (PCA) is the oldest, dating back to Pearson (1901), and most widely used dimension-reduction method (Hastie et al. 2009). It transforms a large number of variables into orthogonal components so that these variables can be represented by a few principal components. It has wide applications in all areas of science, including, in particular, management, finance, economics, etc. Connor and Korajczyk (1986), Kelly et al. (2020), Kim et al. (2021), Giglio and Xiu (2021), and Lettau and Pelger (2020a, b) are a few examples of extending and applying PCA to finance. Today, in the age of big data, it is important to deal with the “curse of dimensionality”. Without dimension reduction, forecasting with conventional multivariate regressions suffers from in-sample overfitting and out-of-sample poor performance. Although PCA is useful in reducing a large number of predictors to just a few combinations of them, one recognized weakness is that it ignores the target information completely.

In this paper, we propose a novel dimension-reduction technique: scaled principal component analysis (sPCA), which scales each predictor with its predictive slope on

the target to be forecasted. By design, sPCA puts more weight on those predictors with stronger forecasting power. In contrast, PCA puts equal weight on all the predictors. Although PCA summarizes the common variation of the predictors, it ignores the target and is an unsupervised learning technique. Hence, if some predictors are noisier than others, they inevitably affect the forecast disproportionately. In the extreme case, the presence of irrelevant predictors would only add noise to the forecast, and it is possible to make the forecasting useless. The sPCA corrects exactly this deficiency by screening out such noisier predictors and assigning them shrinking weights. In this sense, sPCA is designed to let the target guide dimension reduction and is, in spirit, similar to the partial least squares (PLS) of Kelly and Pruitt (2013, 2015) for time series forecasting and to the risk-premium PCA of Lettau and Pelger (2020a, b) for cross-sectional asset pricing.

We extract sPCA factors in two steps. First, we run a predictive regression of the target on each predictor and scale the predictor with the regression slope. Second, we apply PCA to the scaled predictors to obtain principal components as the sPCA factors. In this way, sPCA tends to down-weight those predictors with weak forecasting power and overweight those with strong forecasting power. As a result, the sPCA

factors are more likely to outperform the PCA factors for forecasting purposes.

Theoretically, we consider sPCA in a partially relevant latent factor framework, in which each predictor loads on two groups of factors: one group of relevant factors that are truly associated with the target and the other group of irrelevant factors that are not useful for forecasting the target. In our model, the factors are allowed to be either strong or weak. Our analyses indicate that, when the factors are weak, if the number of predictors (N) and the number of periods (T) satisfy a mild condition, the sPCA forecast dominates the PCA forecast. If the factors are strong or the required condition does not hold, our extensive simulations show that the sPCA forecast still has a large chance to outperform the PCA forecast.

Empirically, we apply sPCA to forecast U.S. inflation, industrial production, unemployment, and the S&P 500 index volatility with a basket of 123 macro variables (Ludvigson and Ng 2007). We find that sPCA consistently generates more accurate forecasts than PCA both in- and out-of-sample, over a wide range of model specifications. The sPCA factor loadings are more tilted toward a smaller subset of the macro variables that display stronger forecasting power, and the PCA factor loadings are more disperse. In addition, sPCA produces comparable or better forecasting performance compared with other commonly used supervised learning techniques, such as LASSO, ridge regression, elastic-net, target PCA in Bai and Ng (2008), etc.

Overall, this paper complements existing approaches for dimension reduction in forecasting with a large number of predictors. Recent examples on the expanding literature include Connor et al. (2012), Kelly and Pruitt (2015), Huang et al. (2015), Light et al. (2017), Freyberger et al. (2020), Pelger (2020), Kelly et al. (2019), Gu et al. (2021, 2020), Lettau and Pelger (2020a, b), Chen et al. (2020), and Rapach and Zhou (2021), among others. Because sPCA generally improves the PCA forecasting performance, it can potentially be applied in many such areas to yield improved results.

The rest of the paper is organized as follows. Section 2 introduces the sPCA method and develops some asymptotic properties. Section 3 explores real data applications, which is followed by Section 4 with a brief conclusion.

2. Methodology

2.1. sPCA Method

Suppose there are N predictors, denoted by $X_t = (X_{1,t}, \dots, X_{N,t})'$ for $i = 1, \dots, N$ and $t = 1, \dots, T$, where T is the number of observations. We are interested in using these predictors to forecast a target variable y_{t+h} with a forecast horizon of h . Each $X_{i,t}$ is a relevant but imperfect predictor of the target. Hence, relying on a few of them is unlikely to capture the dynamics of the target well. However, including all the predictors in a

conventional multivariate regression suffers from the curse of dimensionality, which often leads to in-sample overfitting and out-of-sample poor performance. To address this issue, a common approach is to impose a factor structure on the predictors and extract the latent factors with a reduced dimension.

Specifically, we consider in this paper a partially relevant latent factor model on the joint dynamics of the N predictors X_t and the target y_{t+h} :

$$X_{i,t} = \mu_i + \lambda_i' f_t + e_{i,t} = \mu_i + \phi_i' g_t + \psi_i' h_t + e_{i,t}, \quad (1)$$

$$y_{t+h} = \alpha + \beta' g_t + \varepsilon_{t+h}, \quad (2)$$

where $f_t = (g_t', h_t')'$ are r -dimensional unobserved factors, of which g_t are r_1 -dimensional relevant factors that are associated with the target y_{t+h} and h_t are $(r - r_1)$ -dimensional irrelevant factors. For each predictor $i = 1, \dots, N$, $\lambda_i = (\phi_i', \psi_i')'$ denote the loadings on f_t .

The factor-augmented regression model of Bai and Ng (2006) is a special case of Equation (2). In their setup, any factor in f_t can forecast y_{t+h} . In our framework, in contrast, only the factors g_t , a subset of f_t , are relevant to the target, which seems more plausible in real data applications (see, e.g., Kelly and Pruitt 2015). The case of excluding the irrelevant factors h_t in Equation (1) for forecasting is explicitly analyzed in this paper because most of the theoretical results are based on all the factors.

Given the factor structure, a natural method to estimate the latent factors f_t is PCA. Specifically, according to Bai (2003), PCA estimates $f_t = (g_t', h_t')'$ as $\sqrt{T}$ times the eigenvectors associated with the r largest eigenvalues of M_{xx} where $M_{xx} = \frac{1}{N} \sum_{i=1}^N \dot{X}_i \dot{X}_i'$ denotes the $T \times T$ dimensional sample covariance matrix with $\dot{X}_i = (\dot{X}_{i,1}, \dot{X}_{i,2}, \dots, \dot{X}_{i,T})'$ and $\dot{X}_{i,t} = X_{i,t} - \frac{1}{T} \sum_{s=1}^T X_{i,s}$. Let $\hat{F} = (\hat{G}, \hat{H})$ with $\hat{G} = (\hat{g}_1, \hat{g}_2, \dots, \hat{g}_T)'$ and $\hat{H} = (\hat{h}_1, \hat{h}_2, \dots, \hat{h}_T)'$. The bulk of variation among X_t is summarized by the factors $\hat{F}$. One can conduct forecasting based on either all the factors (g_t and h_t) if one prefers a low bias or partial factors (g_t) if one prefers a parsimonious model and believes that the estimated partial factors are good enough. In either case, dimension reduction can be achieved because of $r \ll N$.

However, when the model is specified by Equations (1) and (2), PCA has a drawback of ignoring the target information. In particular, when factors are strong, PCA fails to differentiate between the target-relevant and -irrelevant latent factors, and there is no guarantee that the first r_1 principal components can best predict the target. When the factors are weak, PCA could fail to extract the signals from the large amount of noise, leading to biased forecasts even using all the factors. To overcome these deficiencies, sPCA is designed to modify PCA by incorporating the target information in the factor-extracting procedure. In so doing, we predict the target with sPCA in two steps:

1. Form a panel of scaled predictors, $(\hat{\gamma}_1 X_{1,t}, \dots, \hat{\gamma}_N X_{N,t})$, where the scaled coefficient $\hat{\gamma}_i$ is the estimated

slope from regressing the target on the i th (standardized) predictor:

$$y_{t+h} = v_i + \gamma_i X_{i,t} + u_{i,t+h}, \quad i = 1, \dots, N. \quad (3)$$

2. Apply PCA to $(\hat{\gamma}_1 X_{1,t}, \dots, \hat{\gamma}_N X_{N,t})$ to extract r factors and use them to predict the target. Specifically, calculate a $T \times T$ matrix $M_{XX}^\circ = \frac{1}{N} \sum_{i=1}^N \hat{\gamma}_i \dot{X}_i (\hat{\gamma}_i \dot{X}_i)'$, where $\dot{X}_i$ denotes the demeaned vector of predictor i . The sPCA factors, $\hat{F}^{\text{sPCA}}$, are equal to $\sqrt{T}$ times of the eigenvectors of the matrix M_{XX}° , which correspond to the r largest eigenvalues of M_{XX}° arranged in a descending order. Let $\hat{G}^{\text{sPCA}}$ be the first r_1 columns of $\hat{F}^{\text{sPCA}}$ and $\hat{g}_t^{\text{sPCA}}$ be the transpose of the t th row of $\hat{G}^{\text{sPCA}}$. With these estimates, one can conduct a full factor-based forecast by regressing y_{t+h} on a constant term and the estimated factors $\hat{f}_t^{\text{sPCA}}$ or conduct a partial factor-based forecast by regressing y_{t+h} on a constant term and the estimated $\hat{g}_t^{\text{sPCA}}$. Then the sPCA forecast for y_{t+h} is

$$\begin{aligned} \hat{y}_{t+h}^{\text{sPCA}} &= \hat{\alpha}^{\text{sPCA}} + (\hat{\pi}^{\text{sPCA}})' \hat{f}_t^{\text{sPCA}} \quad \text{or} \\ \hat{y}_{t+h}^{\text{sPCA}} &= \tilde{\alpha}^{\text{sPCA}} + (\tilde{\beta}^{\text{sPCA}})' \hat{g}_t^{\text{sPCA}}, \end{aligned} \quad (4)$$

where $(\hat{\alpha}^{\text{sPCA}}, (\hat{\pi}^{\text{sPCA}})')$ and $(\tilde{\alpha}^{\text{sPCA}}, (\tilde{\beta}^{\text{sPCA}})')$ are the respective slope estimates of the two predictive regressions.

It is worth mentioning that the scaled predictors $(\gamma_1 X_{1,t}, \dots, \gamma_N X_{N,t})$ follow a latent factor structure as $\gamma_i X_{i,t} = \gamma_i \mu_i + \gamma_i \lambda_i f_t + \gamma_i e_{i,t}$, where γ_i is the probability limit of $\hat{\gamma}_i$ for each i . So the scaled predictors actually share the same factors f_t with the original predictors. Because the forecasted target y_{t+h} is related to the factors instead of the loadings, this naturally begs the question why the sPCA forecast can beat the PCA forecast, especially when all the factors are used to forecast y_{t+h} . The answer to this question is that sPCA screens out the irrelevant predictors by assigning them shrinking weights. This procedure is particularly important because, compared with strong factors, signals of weak factors usually do not dominate the noises sharply. Without a signal-strengthening procedure, traditional PCA could fail to extract the signals from the large amount of noise. We see this point clearly in Section 2.3.

2.2. Assumptions

We make the following five assumptions for the subsequent theoretical results. Hereafter, C denotes a generic constant that is large enough and could be different at each appearance.

Assumption 1. The factors satisfy $\sup_t E\|f_t\|^4 \leq C$ and $\frac{1}{T} \sum_{t=1}^T f_t f_t' \xrightarrow{p} \Sigma_F$ for some $r \times r$ positive definite matrix Σ_F .

Assumption 2. The factor loadings satisfy $\sup_i E\|\lambda_i\|^4 \leq C$. Let $\mathcal{I}_\lambda$ denote the set of units whose λ is not equal to

zero. We assume $\text{Card}(\mathcal{I}_\lambda) \asymp N^v$ for some $v \in (0, 1]$ and $\frac{1}{N^v} \sum_{i=1}^N \lambda_i \lambda_i' \xrightarrow{p} \Sigma_\lambda$ for some $r \times r$ positive definite matrix Σ_λ , where $\text{Card}(\cdot)$ denotes the cardinality of the input, that is, the number of elements of the input, and $a \asymp b$ means that there exist two constants c and C such that $cb \leq a \leq Cb$.¹

Assumption 3. Let $e_{it} = \sigma_i e_{it}^*$, where e_{it}^* is independent and identically distributed over i and t with the eighth moment bounded. In addition, $C^{-1} \leq \sigma_i \leq C$.

Assumption 4. Suppose that $\{\lambda_i\}$, $\{f_t\}$, and $\{e_{it}\}$ are mutually independent.

Assumption 5. Let $\mathcal{F}_t$ be the σ -field generated by $g_t, \varepsilon_t, g_{t-1}, \varepsilon_{t-1}, \dots$. Then, $E(\varepsilon_{t+h} | \mathcal{F}_t) = 0$ for any integer $h > 0$ and $\sup_t E(\varepsilon_t^4) \leq C$. In addition, $\{\varepsilon_{t+h}\}$ is independent with the three groups of variables in Assumption 4.

Assumptions 1 and 2 specify the strength of the factors. If $v \in (0, 1)$, our model is a weak factor model because it violates the strong factor assumption $N^{-1} \Lambda' \Lambda \xrightarrow{p} \Sigma_\lambda$; see Bai (2003) and Bai and Li (2012), among others, for a detailed illustration on strong factor models. The weak factor specification is of practical relevance. Recently, for example, Daniele et al. (2019) show that the eigenvalues of the return covariance matrix of the S&P 500 constituents can be well characterized by a mixed factor model with one strong factor and three weak factors. We do not consider the mixed-factor model in this paper for simplicity. Given that strong factor models are well understood in the literature, an extension of allowing partial factors to be strong can be made with some changes to the current arguments without theoretical challenges. Moreover, the assumption of a weak factor form is more sensible in our setting, in which the predictors' signal-to-noise ratios are usually low.

Assumption 3 assumes independence among the errors $\{e_{it}\}$. Our asymptotic analysis relies on the theoretical results about the largest eigenvalue of a random matrix, and so we can allow limited correlations in errors according to Onatski (2012). Assumptions 4 and 5 are standard for factor analysis with heterogenous loadings. As pointed out by Bai and Ng (2006), these assumptions are rather general.

Hereafter, to avoid unnecessary mathematical complexity, we confine the analysis to a simple two-factor model, one relevant factor g_t and one irrelevant factor h_t . To evaluate the forecasting performance, we follow the literature and use the asymptotic mean square forecast error (MSFE) as the evaluation criterion:

$$\text{MSFE} = \lim_{N, T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T (y_{t+h} - \hat{y}_{t+h})^2. \quad (5)$$

Given that the scaling coefficient from the regression of y_{t+h} on (standardized) predictor X_{it} in Equation (3) is

$$\hat{\gamma}_i = \frac{\frac{1}{T} \sum_{t=1}^T (X_{it} - \bar{X}_i)(y_{t+h} - \bar{y}_{t+h})}{\frac{1}{T} \sum_{t=1}^T (X_{it} - \bar{X}_i)^2} = \beta \phi_i + o_p(1), \quad (6)$$

where we use the standardized condition $\frac{1}{T} \sum_{t=1}^T (X_{it} - \bar{X}_i)^2 = 1$ for each i the scaled predictor $\hat{\gamma}_i X_{it}$ is very close to $\gamma_i X_{it}$, which is equal to

$$\begin{aligned} \gamma_i X_{it} &= \beta \phi_i [\mu_i + \phi_i g_t + \psi_i h_t + e_{it}] \\ &= \underbrace{\beta \phi_i \mu_i}_{\mu_i^o} + \underbrace{\beta \phi_i^2 g_t}_{\phi_i^o} + \underbrace{\beta \phi_i \psi_i h_t}_{\psi_i^o} + \underbrace{\beta \phi_i e_{it}}_{e_{it}^o}. \end{aligned} \quad (7)$$

In (7), the scaled series $\gamma_i X_{it}$ reflects the i th predictor's forecasting power on the target. A predictor with stronger forecasting power (i.e., higher absolute value of γ_i), on average, receives a larger weight, whereas a predictor with weaker forecasting power receives a smaller weight. In short, sPCA applies PCA to the scaled predictors $(\hat{\gamma}_1 X_{1,t}, \dots, \hat{\gamma}_N X_{N,t})$ rather than to the raw predictors $(X_{1,t}, \dots, X_{N,t})$ to better capture the predictive information contained in $X_{i,t}$.

2.3. Asymptotic Forecasting Performance

This section presents the asymptotic forecasting performance of sPCA and PCA. The results are used for comparison purposes. We first provide the asymptotic results on the estimated sPCA and PCA factors, respectively.

Proposition 1. Under Assumptions 1–5,

1. If $N^{1-\nu}/T^2 \rightarrow 0$, the estimated sPCA factors, denoted by $\hat{F}^{\text{sPCA}}$, are consistent and admit

$$\frac{1}{\sqrt{T}} \|\hat{F}^{\text{sPCA}} - \dot{F} R'_{\text{sPCA}}\| \asymp_p N^{-\nu/2} + T^{-1} + \frac{N^{1-\nu}}{T^2},$$

where $\dot{F} = (\dot{f}_1, \dot{f}_2, \dots, \dot{f}_T)'$ with $\dot{f}_t = f_t - \frac{1}{T} \sum_{s=1}^T f_s$, R_{sPCA} is some invertible rotational matrix, and $A \asymp_p B$ means that there exist two constants c and C such that $cB \leq A \leq CB$ with probability approaching one.

2a. Let $\hat{F}^{\text{PCA}}$ denote the estimated PCA factors. If $N^{1-\nu}/T \geq c$ for some $c > 0$, then $\hat{F}^{\text{PCA}}$ is not a consistent estimator of $\dot{F}$, in the sense that, for any invertible matrix R_{PCA} , the following holds:

$$\frac{1}{\sqrt{T}} \|\hat{F}^{\text{PCA}} - \dot{F} R'_{\text{PCA}}\| \geq c^*,$$

with a strictly positive probability for some constant c^* .

b. If $N^{1-\nu}/T \rightarrow 0$, $\hat{F}^{\text{PCA}}$ is consistent and admits

$$\frac{1}{\sqrt{T}} \|\hat{F}^{\text{PCA}} - \dot{F} R'_{\text{PCA}}\| \asymp_p N^{-\nu/2} + \frac{N^{1-\nu}}{T}.$$

We have three remarks on Proposition 1. First, the consistency condition for sPCA is obviously weaker than that for PCA. As such, if the values of N and T satisfy $N^{1-\nu}/T^2 \rightarrow 0$ and $N^{1-\nu}/T \geq c$ for some $c > 0$, the sPCA forecast dominates the PCA forecast because the former is consistent and the latter is not. Second, even $N^{1-\nu}/T \rightarrow 0$, the sPCA forecast still has a chance to

outperform the PCA forecast because it is likely that sPCA estimates the factors more precisely. Intuitively, if the forecast is consistent, the first-order MSFE can be decomposed into two components. One component arises from the regression errors in the predictive equation, which is independent of the estimation methods of the factors, and the other component is a result of the estimation errors of the factors. Apparently, a precise estimation of the factors leads to a smaller magnitude of MSFE. Third, when $\nu = 1$, the model reduces to a strong factor model. Our analysis indicates that

$$\frac{1}{\sqrt{T}} \|\hat{F}^{\text{PCA}} - \dot{F} R'_{\text{PCA}}\| \asymp_p N^{-1/2} + T^{-1}.$$

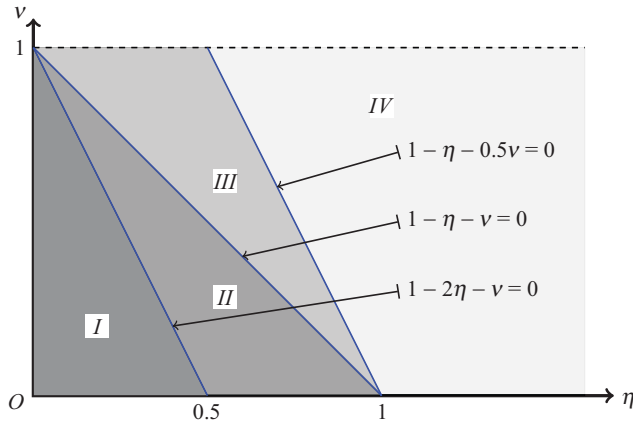
This result is consistent with the literature; see Bai (2003).²

2.3.1. Sufficient Conditions. With the results in Proposition 1, we can identify the sufficient conditions for sPCA to outperform PCA. If $N^{-\nu/2} + T^{-1}$ is dominated by $N^{1-\nu}/T^2$, it is certainly dominated by $N^{1-\nu}/T$. In this case, sPCA dominates PCA in estimating the factors. If $N^{-\nu/2} + T^{-1}$ dominates $N^{1-\nu}/T^2$, because T^{-1} and $N^{1-\nu}/T^2$ are both dominated by $N^{1-\nu}/T$, it suffices to compare $N^{-\nu/2}$ and $N^{1-\nu}/T$. If $N^{1-\nu}/T$ dominates $N^{-\nu/2}$, we still have that sPCA dominates PCA when estimating the factors. Given this discussion, we have the following proposition.

Proposition 2. Under Assumptions 1–5, if $N^{1-\nu}/T^2 \rightarrow 0$ and $N^{1-\nu/2}/T \rightarrow \infty$, the sPCA forecast outperforms the PCA forecast.

As pointed out earlier, sPCA assigns different weights to different predictors, and in the extreme case when a predictor contains no relevant information, it assigns a weight of order $O_p(1/\sqrt{T})$ to the predictor to screen out noise, which sharpens the advantage of sPCA. Hence, the weaker the factors, that is, the smaller the ν , the more chance that sPCA outperforms PCA. This can be seen from Proposition 2. The condition $N^{1-\nu/2}/T \rightarrow \infty$ is more likely to hold if ν is small. As regards the condition $N^{1-\nu}/T^2 \rightarrow 0$, we note that it is imposed for the purpose of consistency of sPCA. When this condition breaks down, we conjecture that sPCA still has a superior performance because it delivers a relatively better estimation of factors although both the sPCA and PCA factors are not consistent in this case. However, we do not analyze further this case because of the lack of tools in the literature for inconsistent estimating. There are a few exceptions. For example, Onatski (2012) and Lettau and Pelger (2020a, b) consider inconsistency comparison in the very weak case ($\nu = 0$). However, their analysis depends critically on the tools from the random matrix theory, which seem not applicable in our context because of the shrinking weights in sPCA.

To further understand the implications of Proposition 2, we depict Figure 1 under the condition $T \asymp N^\eta$

Figure 1. (Color online) Partitioned Regions for Different Results

Notes. The graph is depicted under the assumption of $T \asymp N^\eta$ with $\eta \in (0, \infty)$. Region I: $1 - 2\eta - v > 0$. The sPCA and PCA forecasts are both inconsistent. Region II: $1 - \eta - v > 0$ but $1 - 2\eta - v < 0$. The sPCA forecast is consistent, but PCA is not. Region III: $1 - \eta - 0.5v > 0$ but $1 - \eta - v < 0$. The sPCA and PCA forecasts are both consistent, but the order magnitude of the MSFE for the sPCA is smaller. Region IV: $1 - \eta - 0.5v < 0$. The sPCA and PCA forecasts are both consistent, the MSFEs of two methods are in the same order of magnitude. Which method is better depends on the values of ϕ_i, ψ_i and the variance values of e_{it} .

with $\eta \in (0, \infty)$. In this figure, the rectangle area is partitioned into four regions, which are labeled I, II, III, and IV, respectively. According to Proposition 2, if (η, v) falls in regions II and III, the sPCA forecast outperforms the PCA forecast. More specifically, in region II, the sPCA forecast is consistent, but PCA is not; in region III, the sPCA and PCA forecasts are both consistent, but the magnitude of the sPCA's MSFE is smaller. In region I, we conjecture that sPCA could still outperform PCA although both methods deliver inconsistent forecasts. The case of region IV is discussed in detail in the following section. From Figure 1, one may have an impression that, when $N \leq T$, it suffices to consider the case of region IV. However, we point out that this is not true. Consider the case of $T = CN^\eta$ with $C = 10$. When $T = 100$, we have $N \leq T$ for all $\eta \geq 0.5$. So, even for $N \leq T$, it is still possible that $\eta < 1$. In other words, one should not use the observed values of N and T to roughly determine the region. Instead, given the fact that the increase of the number of predictors is speedier than the increase of the time periods, it is more plausible to assume $\eta < 1$. In a nutshell, regions II and III tend to represent the cases of practical relevance.

2.3.2. Other Cases. If N, T , and v satisfy $N^{1-v/2}/T \rightarrow 0$, or if (η, v) falls in region IV in Figure 1, Proposition 1 implies

$$\frac{1}{\sqrt{T}} \|\hat{F}^{\text{sPCA}} - \hat{F} R'_{\text{sPCA}}\| = O_p(N^{-v/2}), \quad \text{and} \\ \frac{1}{\sqrt{T}} \|\hat{F}^{\text{PCA}} - \hat{F} R'_{\text{PCA}}\| = O_p(N^{-v/2}).$$

In this case, the estimation errors of sPCA and PCA are in the same order of magnitude. Thus, if one expects to figure out which method is better, she has to calculate the concrete formulas of the MSFE. The following proposition gives the expressions of the first-order MSFE for both the sPCA and PCA forecasts, which is based on $v = 1$ for the strong factor case. However, the results for other points in region IV are almost the same with the mere change that N is replaced by N^v .

Proposition 3. Under Assumptions 1–5, as $N \rightarrow \infty$, $T \rightarrow \infty$, and $\sqrt{N}/T \rightarrow 0$,³ the first-order MSFEs of the PCA and sPCA forecasts have the following expressions:

$$\text{MSFE}_{\text{PCA}} = \frac{1}{T} 3\sigma_\varepsilon^2 + \frac{1}{T} \sum_{t=1}^T \beta^{*'} (\Lambda' \Lambda)^{-1} \Gamma_t^{\text{PCA}} (\Lambda' \Lambda)^{-1} \beta^*,$$

$$\text{MSFE}_{\text{sPCA}} = \frac{1}{T} 3\sigma_\varepsilon^2 + \frac{1}{T} \sum_{t=1}^T \beta^{*'} (\Lambda' W \Lambda)^{-1} \Lambda' \Gamma_t^{\text{sPCA}} \Lambda (\Lambda' W \Lambda)^{-1} \beta^*,$$

where $\beta^* = (\beta, 0)'$, $\Gamma_t^{\text{PCA}} = \frac{1}{N} \sum_{i=1}^N E(\lambda_i \lambda_i' e_{it}^2)$, $\Gamma_t^{\text{sPCA}} = \frac{1}{N} \sum_{i=1}^N E(\phi_i^4 \lambda_i \lambda_i' e_{it}^2)$, and $W = \text{diag}(\phi_1^2, \dots, \phi_N^2)$.

Proposition 3 indicates that, under current assumptions, it is difficult to determine which method dominates the other. On the one hand, if e_{it} is homoskedastic across i , the MSFE of PCA can be smaller than that of sPCA. On the other hand, if ϕ_i^2 happens to be the inverse of $\text{var}(e_{it})$, the MSFE of sPCA can be smaller than that of PCA. For this reason, in Section 2.5, we resort to simulations to investigate with what probability sPCA outperforms PCA. Our simulations indicate that the probability is about 0.7. This result looks plausible. To see the intuition, under the normalization $\frac{1}{T} \hat{F}' \hat{F} = I$, we have

$$\phi_i^2 + \psi_i^2 + \sigma_i^2 \approx 1, \quad (8)$$

where $\sigma_i^2 = \text{var}(e_{it})$. Then, a large absolute value of ϕ_i would generally lead to a small σ_i^2 . For sPCA, we assign a weight $\hat{\gamma}_i = \beta \phi_i + \mathbf{u}_i$ to the i th predictor. So the variance of the scaled data are approximately $\phi_i^2 \sigma_i^2$ for a generic i . Intuitively, PCA can be linked to the ordinary least square regressions (see, e.g., Bai 2003), and sPCA can be linked to the generalized least square regressions with a specific weighting matrix $W = \text{diag}(\phi_1^2, \phi_2^2, \dots, \phi_N^2)$. Theoretically, the most efficient estimator would use the inverse residual covariance matrix as a weight W . Hence, the closer the weighting matrix W gets to the inverse residual covariance matrix, the higher the efficiency. We refer readers to Pelger and Xiong (2021) for detailed discussions on this point. Given that ϕ_i^2 is generally negatively related to σ_i^2 , we have that, with a large probability, the weighting matrix W resembles the inverse residual covariance matrix. This means that, with a large probability, the sPCA factors are estimated more

precisely because the errors are more likely to be homoskedastic.

2.4. Single Factor Forecast

In this section, we briefly discuss the single-factor case of using the first sPCA factor to conduct forecasting, which has some theoretical interest as the factor represents an aggregate index of all the predictors. Moreover, it is worth mentioning that the identification in a partially relevant factor forecasting model is different from that in a fully relevant factor forecasting model, which has not attracted much attention in the literature. A detailed discussion of the identification issue in a general partially relevant factor model can be found in Huang et al. (2021).⁴

We provide a set of sufficient conditions under which the sPCA forecast outperforms the PCA forecast in the single-factor forecast case.

Proposition 4. Suppose $E(e_{it}^2) = \sigma_e^2$ for all i and t . We have

Case 1: $0 < \xi^\circ \leq \xi$ if $|\phi_i| > |\psi_i|$ and $\phi_i \psi_i \geq 0$ for all i .

Case 2: $0 > \xi^\circ \geq \xi$ if $|\phi_i| > |\psi_i|$ and $\phi_i \psi_i < 0$ for all i .

Case 3: $0 > \xi \geq \xi^\circ$ if $|\phi_i| < |\psi_i|$ and $\phi_i \psi_i \geq 0$ for all i .

Case 4: $0 < \xi \leq \xi^\circ$ if $|\phi_i| < |\psi_i|$ and $\phi_i \psi_i < 0$ for all i .

Here,

$$\xi^\circ = \frac{\sum_{i=1}^N \phi_i^\circ \psi_i^\circ}{\sum_{i=1}^N (\phi_i^{\circ 2} - \psi_i^{\circ 2})} = \frac{\sum_{i=1}^N \phi_i^3 \psi_i}{\sum_{i=1}^N \phi_i^2 (\phi_i^2 - \psi_i^2)}, \quad \text{and}$$

$$\xi = \frac{\sum_{i=1}^N \phi_i \psi_i}{\sum_{i=1}^N (\phi_i^2 - \psi_i^2)}.$$

The sPCA forecast outperforms the PCA forecast in the case of using the first factor to conduct forecasting, that is, $\text{MSFE}_{\text{sPCA}} \leq \text{MSFE}_{\text{PCA}}$ if the factor loadings ϕ_i and ψ_i in Equation (1) belong to one of these four cases.

Proposition 4 involves four cases on the signs of $\phi_i \psi_i$ and $|\phi_i| - |\psi_i|$ to simplify our discussions about the signs on the numerator and denominator of $\xi = \sum_{i=1}^N \phi_i \psi_i / \sum_{i=1}^N (\phi_i^2 - \psi_i^2)$. The intuition can be elaborated by the first case, in which the numerator and denominator of ξ are both positive. Let $z_i = \psi_i / \phi_i$ measure the noise-to-signal ratio of predictor i . Define $\omega_i = \phi_i \psi_i / (\phi_i^2 - \psi_i^2) = z_i / (1 - z_i^2)$. Because ω_i is an increasing function of z_i over the region of $z_i < 1$, a predictor with stronger forecasting power, that is, lower noise-to-signal ratio, has a lower value of ω_i . Denote ω_i^u and ω_i^d to be the numerator and denominator of ω_i , and so we have $\xi = \frac{\omega_1^u + \omega_2^u + \dots + \omega_N^u}{\omega_1^d + \omega_2^d + \dots + \omega_N^d}$. Then, the weighting scheme of sPCA puts a larger weight on predictor i if its $\omega_i = \omega_i^u / \omega_i^d$ has a smaller value, which can, in turn, reduce ξ because of the general inequality that $a/b \leq$

$\Sigma_i a_i / \Sigma_i b_i$ if $a/b \leq a_i/b_i$ for each i . The same argument carries through for the other three cases. Hence, the sPCA forecast outperforms the PCA forecast in terms of MSFE.

We note that these analyses are based on the assumption of stationarity. However, when the predictors are highly persistent, say local-to-unit process or fractional integrated, we caution the readers that our sPCA method could have a chance to break down, depending on the sources of the persistence. For example, if the persistence of predictors comes from idiosyncratic errors or irrelevant factors or both, our method may not work because the useless information dominates the useful one. However, we note that, when this happens, PCA also breaks down for the same reason (see, e.g., Bai 2004, Bai and Ng 2004). So it is not an issue specific to our method although the estimation accuracy of scaling values in sPCA depends on persistence. Developing a method that is immune to the persistence of predictors is an interesting topic but beyond the scope of this paper, and we leave it for future research.

2.5. Simulation Evidence

In this section, we conduct Monte Carlo experiments to compare the *finite sample* forecasting accuracies of sPCA with PCA. Our simulation design is based on a simple two-latent-factor model with one being target relevant. To be more specific, we begin by simulating the time series of two factors that are independently and normally distributed with zero mean and unit variance, that is, $g_t \sim N(0, 1)$ and $h_t \sim N(0, 1)$. The target is generated by $y_{t+h} = g_t + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \sim N(0, 1)$ so that the infeasible best forecast has the lowest MSFE of one. There are N observable predictors $X_{i,t}$ ($i = 1, \dots, N$), which load on both the relevant factor g_t and the irrelevant factor h_t . Idiosyncratic noises are also generated from the normal population with zero mean and standard deviation of σ_i and are independent across predictors and over time.

2.5.1. Weak Factor Case. We first examine the performance in the weak factor setting, in which we set all ϕ_i and ψ_i to be zero except n predictors with $n \ll N$. The n predictors are randomly drawn from the independent uniform distribution with support $[0, 1]$ denoted by $U[0, 1]$. The standard deviations of idiosyncratic noises σ_i ($i = 1, \dots, N$) are drawn independently from $U[0, 1]$. To allow for heterogeneity over time, we consider multiplying each σ_i with a time-dependent random variable σ_t ($t = 1, \dots, T$), which is drawn from a uniform distribution with support $[0.5, 1.5]$.

Table 1 presents the simulated forecasting performance of sPCA and PCA with different degrees of

Table 1. The MSFEs of the sPCA and PCA Forecasts with Weak Factors

n	sPCA			PCA		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
Panel A: Heterogenous idiosyncratic errors (cross-sectionally)						
50	1.262	1.048	1.056	1.507	1.109	1.111
40	1.278	1.033	1.038	1.532	1.140	1.135
30	1.269	1.066	1.070	1.509	1.292	1.270
20	1.274	1.107	1.107	1.499	1.429	1.421
10	1.292	1.194	1.194	1.565	1.571	1.569
Panel B: Heterogenous idiosyncratic errors (cross-sectionally and time series)						
50	1.268	1.027	1.032	1.513	1.225	1.220
40	1.246	1.030	1.039	1.492	1.337	1.311
30	1.311	1.131	1.135	1.548	1.504	1.489
20	1.291	1.110	1.109	1.557	1.558	1.557
10	1.362	1.272	1.280	1.720	1.719	1.716

Notes. This table reports the out-of-sample MSFEs of the sPCA and PCA forecasts in the case of weak factors and heterogenous noises. Simulations are based on a two-latent-factor model with one relevant factor. The sample size is $(T, N) = (250, 500)$. Both factors are independent and identically normally distributed and are weak in the sense that only $n = 50, 40, 30, 20$, or 10 out of $N = 500$ predictors are loading on the two latent factors. Nonzero loadings on the relevant factor ϕ_i and on the irrelevant factor ψ_i are drawn independently from a uniform distribution with support $[0, 1]$. Panel A considers the case in which the idiosyncratic errors are heterogenous cross-sectionally, and the standard deviations of the idiosyncratic errors σ_i follow a uniform distribution with support $[0, 1]$. Panel B adds the time series heterogeneity, and the standard deviations of the idiosyncratic errors are further multiplied by a random variable σ_t , which is drawn from a uniform distribution with support $[0.5, 1.5]$. We use the first 200 observations for parameter training and the rest 50 for out-of-sample evaluation. Reported are the median MSFE of one, two, and three sPCA and PCA factors on the basis of 100 simulations.

weakness. We focus on their MSFEs in an out-of-sample environment and consider forecasting with one, two, and three estimated sPCA and PCA factors, respectively. These three cases correspond to underestimation, correct estimation, and overestimation of the number of factors. The results are obtained with 100 repetitions with a sample size of $n = 500$ and $T = 250$. We split the data into two parts and use the first 200 time periods for sample training and the remaining 50 periods for out-of-sample evaluation. We report the median out-of-sample MSFEs for the sPCA and PCA forecasts. Panel A considers the cross-sectional heteroskedasticity alone, and Panel B considers both the cross-sectional and the time-series heteroskedasticities.

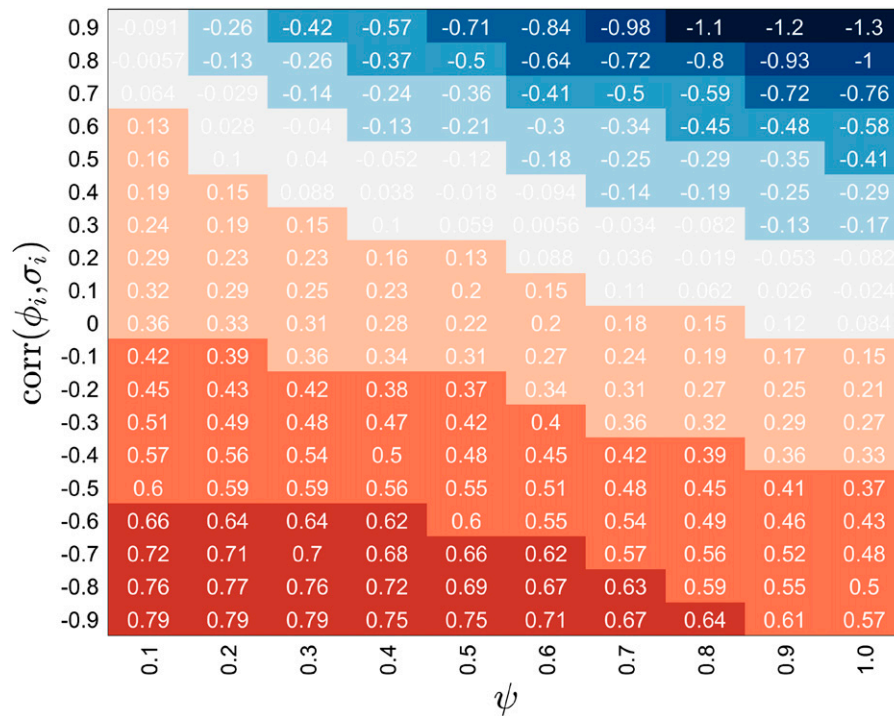
Panel A of Table 1 indicates that the first sPCA factor has an MSFE less than 1.30 across the values of n/N . This result contrasts with the first PCA factor that produces an MSFE above 1.50 in general. In the case of two factors, the MSFE of the sPCA forecast ranges from 1.03 to 1.19, and the MSFE of the PCA ranges from 1.11 to 1.57 as n decreases from 50 to 10, which implies that the sPCA forecast dominates the PCA forecast unanimously. The MSFE of the PCA is far from the best value of one, and therefore, none of these forecasts is consistent, confirming our theoretical conclusion. As expected, including the overestimated three factors leads to little change in the MSFE of PCA and sPCA. Panel B shows that the conclusion remains

true when we consider the cross-sectional and time-series heterogenous noises simultaneously. In summary, consistent with our theoretical analyses, the simulated results in Table 1 show that sPCA dominates PCA in the weak factor setting, and the dominance becomes more pronounced when the degree of weakness grows.

2.5.2. Strong Factor Case. We next compare the forecasting performance of sPCA with PCA in the strong factor setting, in which the number of non-zero loadings is $n = N$. Because the asymptotic MSFEs of the sPCA and the PCA depend on the joint distribution of factor loadings ϕ_i and ψ_i and the standard deviation of idiosyncratic noise σ_i , we consider different parameter specifications and compute the asymptotic MSFEs through simulations. We draw ϕ_i and σ_i randomly from a correlated uniform distribution with support $[0, 1]$ and draw ψ_i independently from the uniform distribution with support $[0, \psi]$, where ψ governs the average relative importance of the systematic noise. This design allows us to highlight the role of $\rho = \text{corr}(\phi_i, \sigma_i)$ in shaping the asymptotic efficiency of sPCA relative to PCA.

Figure 2 displays a heat map of the differences in the asymptotic MSFE between PCA and sPCA in which a positive (negative) value indicates the outperformance

Figure 2. (Color online) Asymptotic Forecasting Performance Difference Between sPCA and PCA



Notes. This figure plots the asymptotic forecasting performance difference between sPCA and PCA in a two-latent-factor model with heterogeneous idiosyncratic errors. We calculate this asymptotic performance difference as the asymptotic MSFE of PCA minus that of sPCA divided by the number of predictors $n = 500$. A brighter positive value indicates that the sPCA forecast is more accurate than the PCA forecast. Loadings on the relevant factor, ϕ_i , and the standard deviation of idiosyncratic error, σ_i , are drawn randomly from correlated uniform distributions with support $[0, 1]$ and loadings on the irrelevant factor are drawn independently from a uniform distribution with support $[0, \psi]$.

(underperformance) of sPCA. We find that sPCA tends to deliver a more accurate forecast (i.e., a positive difference) when ϕ_i and σ_i are negatively correlated or when ψ is small. In these cases, a higher signal value of ϕ_i tends to be associated with a lower idiosyncratic noise σ_i and a lower systematic noise ψ . Similarly, a lower signal value of ϕ_i tends to be associated with a higher idiosyncratic noise σ_i and systematic noise ψ , which suggests that the signal-to-noise ratios are more dispersed in the cross-section. By incorporating the target information, sPCA overweights the predictors with a high signal-to-noise ratio and down-weights the predictors with a low signal-to-noise ratio. In this way, sPCA effectively enhances the efficiency in extracting the factors relative to PCA.

Table 2 presents the out-of-sample MSFEs with one, two, and three sPCA and PCA factors by assuming $(N, T) = (50, 250)$ and $(N, T) = (100, 250)$, respectively. The results show that the first sPCA factor unanimously generates a lower MSFE than the first PCA factor across all the parameter specifications. The performance improvement tends to be larger as the value of ψ increases. This is intuitive because a higher ψ means a larger

component of the irrelevant factor attributing to the predictors, and so it is more likely to contaminate the first PCA factor. The first sPCA factor is less affected by the irrelevant factor as it is target driven and can filter out the systematic noise. When we employ two or three factors, the outperformance of sPCA relative to PCA becomes weaker because both the sPCA and PCA forecasts are consistent in this case. Overall, sPCA tends to outperform PCA in general, especially when the correlation between ϕ_i and σ_i is negative, which is consistent with the heat map of Figure 2.

2.6. Number of sPCA Factors

Determining the number of factors is an important topic in factor analysis, which has received much attention over the last two decades. Most of the existing studies focus on determining the number of strong factors (see, e.g., Bai and Ng 2002, Ahn and Horenstein 2013). Because the current paper also allows the presence of weak factors, the method to determine the number of factors must be consistent with this assumption. Fortunately, the literature has designed methods for weak factors (e.g., Onatski

Table 2. The MSFEs of the sPCA and PCA Forecasts with Strong Factors

Corr(ϕ_i, σ_i)	Ψ	sPCA			PCA		
		One factor	Two factors	Three factors	One factor	Two factors	Three factors
Panel A: $T = 200, N = 50$							
−0.75	0.5	1.115	1.024	1.028	1.179	1.043	1.048
−0.75	1	1.208	0.998	1.003	1.403	1.007	1.013
−0.50	0.5	1.119	1.047	1.055	1.189	1.061	1.065
−0.50	1	1.238	1.026	1.032	1.457	1.035	1.039
0	0.5	1.138	1.033	1.041	1.227	1.031	1.034
0	1	1.273	1.053	1.057	1.529	1.050	1.054
0.50	0.5	1.156	1.042	1.048	1.267	1.048	1.054
0.50	1	1.310	1.062	1.068	1.584	1.049	1.056
0.75	0.5	1.188	1.060	1.066	1.323	1.047	1.048
0.75	1	1.330	1.070	1.077	1.568	1.055	1.058
Panel B: $T = 200, N = 100$							
−0.75	0.5	1.130	1.048	1.056	1.188	1.056	1.061
−0.75	1	1.234	1.031	1.040	1.433	1.036	1.044
−0.50	0.5	1.118	1.034	1.042	1.186	1.041	1.042
−0.50	1	1.230	1.000	1.006	1.450	1.007	1.009
0	0.5	1.134	1.023	1.030	1.223	1.023	1.025
0	1	1.255	1.001	1.009	1.499	1.002	1.009
0.50	0.5	1.159	1.025	1.030	1.280	1.020	1.026
0.50	1	1.351	1.074	1.084	1.609	1.072	1.075
0.75	0.5	1.143	0.991	0.997	1.284	0.988	0.990
0.75	1	1.342	1.070	1.078	1.602	1.063	1.066

Notes. This table reports the out-of-sample MSFEs of the sPCA and PCA forecasts in the case of strong factors and heterogenous noises. Simulations are based on a two-latent-factor model with one relevant factor. The sample size $(T, N) = (250, 50)$ in Panel A and $(T, N) = (250, 100)$ in Panel B. Loadings on the relevant factor ϕ_i and standard deviation of idiosyncratic error σ_i are drawn from correlated uniform distributions with support $[0, 1]$, whereas loadings on the irrelevant factor are drawn independently from a uniform distribution with support $[0, \psi]$. We use the first 200 observations for parameter training and the rest 50 for out-of-sample evaluation. Reported are the median MSFE of one, two, and three sPCA and PCA factors on the basis of 100 simulations.

2012). In addition, some methods, such as Ahn and Horenstein (2013), seem to continue to work in the weak factor setting. Overall, one may rely on methods in the existing literature to determine the number of weak factors in real applications.

3. Empirical Evidence

In this section, we apply sPCA to macroeconomic forecasting with real data. We compare the in- and out-of-sample forecasting performance of sPCA with that of PCA. We also compare sPCA with alternative big-data forecasting methods, such as target PCA (tPCA), PLS, LASSO, and ridge regression.

3.1. Data

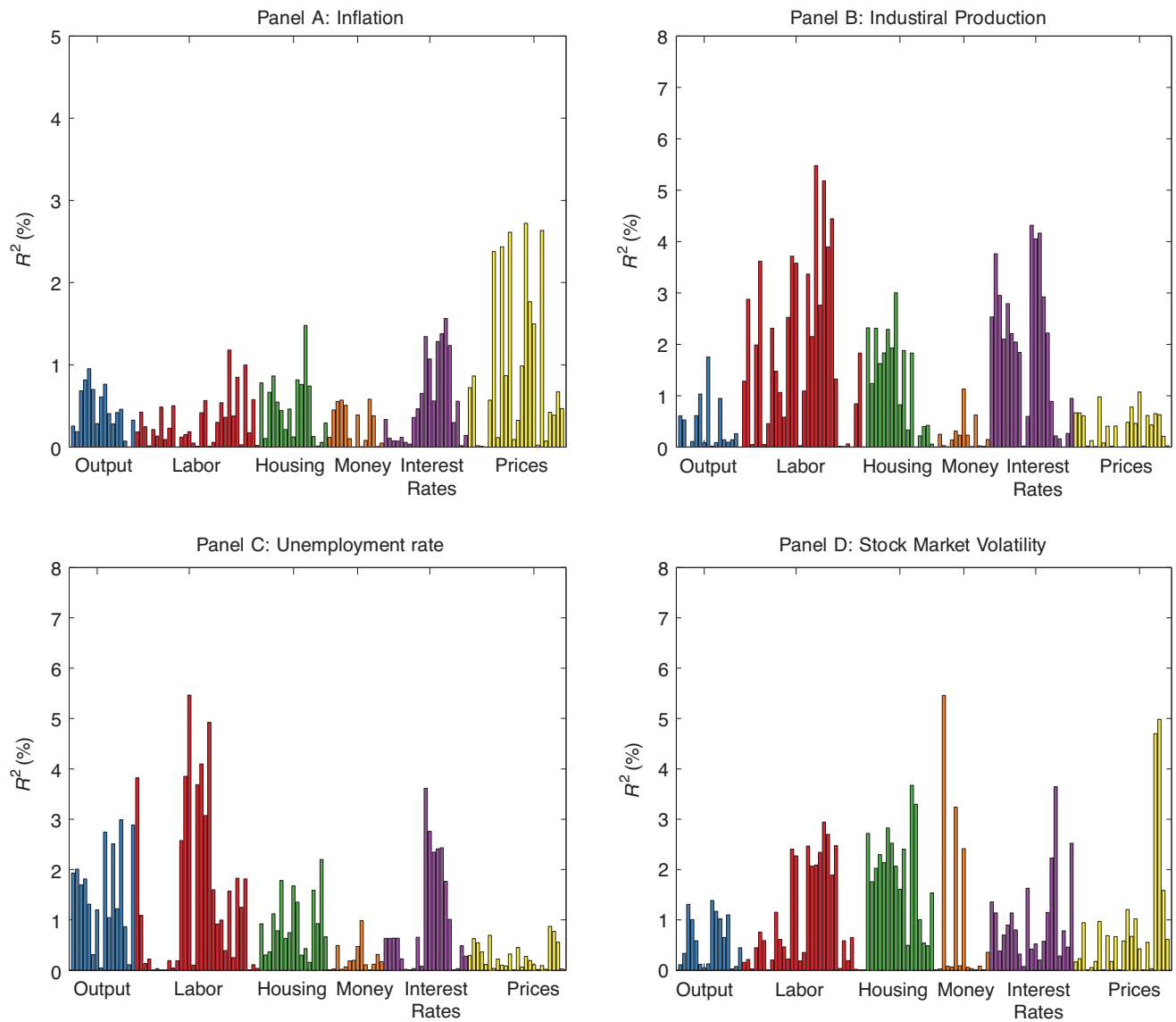
We consider 123 macro variables from the FRED-MD database spanning from January 1960 to December 2019, which are maintained by St. Louis Fed.⁵ As described in McCracken and Ng (2016), the FRED-MD database represents a recent effort by the authors and St. Louis Fed staff to compile a standard macroeconomic database to facilitate big data macro research. It extends the widely used Stock and Watson (2006)

data set and covers broad economic categories, such as output and income, labor force and unemployment, consumption expenditure and housing indicators, money stock and credit, and price indices. The detailed variables and transformation codes to ensure stationarity of each macro variable are provided in the online appendix.

We apply sPCA to these 123 macro variables to forecast the one-month-ahead U.S. inflation, industrial production (IP) growth, change in unemployment rate, and the S&P 500 index volatility. Among these four applications, inflation, IP growth, and unemployment rate predictions are widely examined in the macroeconomic forecasting literature, and the S&P 500 index volatility prediction with macro variables is carefully explored by Ludvigson and Ng (2007).

Figure 3 plots the R^2 of predicting one-month-ahead inflation, IP growth, change in unemployment rate, and the S&P 500 index volatility with one of the 123 macro variables, respectively. To highlight the incremental forecasting power, we control for lagged values of the target with the number of lags selected by the Bayesian information criterion (BIC).

Figure 3. (Color online) In-Sample Forecasting of Macro Variables



Notes. This figure plots in-sample R^2 (in percentage) of predicting one-month-ahead inflation (Panel A), industrial production growth (Panel B), unemployment rate (Panel C), and S&P 500 index volatility (Panel D) using each of the 123 macro variables from the FRED-MD data set of McCracken and Ng (2016), consisting of output and income (No. 1–16), labor market (No. 17–47), consumption and housing (No. 48–64), money and credit (No. 65–78), interest and exchange rate (No. 79–99), and prices (No. 100–123). To highlight the incremental predictive power of each variable, we control for lags of the target with the number of lags selected by BIC. Macro variables are collected at a monthly frequency and the sample period is January 1960–December 2019.

Panel A of Figure 3 shows that, among different categories, interest rate- and price-related variables have the highest predictive power for future inflation, followed by housing-, labor-, and output-related variables. Panels B and C show that labor market conditions and interest rate-related variables display the highest predictive power for future IP growth and unemployment rate. Panel D shows that several money- and price-related variables and housing- and interest rate-related variables have the highest predictive

ability for the S&P 500 index volatility. Overall, Figure 3 demonstrates that one should not treat individual predictors equally in extracting factors for prediction because their forecasting abilities vary.

3.2. In-Sample Results

This section examines the in-sample forecasting performance.

First, Table 3 reports, in a descending order, the first 15 eigenvalues of the covariance matrix of the raw

Table 3. Eigenvalues of the Covariance Matrixes of the Raw and Scaled Macro Variables

	PCA	sPCA			
		Inflation	IP	Unemploy	Volatility
1st	0.15	0.19	0.33	0.37	0.22
2nd	0.07	0.14	0.10	0.11	0.10
3rd	0.07	0.13	0.07	0.07	0.08
4th	0.05	0.08	0.07	0.06	0.08
5th	0.04	0.06	0.06	0.04	0.06
6th	0.03	0.03	0.03	0.03	0.04
7th	0.03	0.03	0.03	0.03	0.03
8th	0.02	0.03	0.02	0.02	0.03
9th	0.02	0.02	0.02	0.02	0.03
10th	0.02	0.02	0.02	0.02	0.02
11th	0.02	0.02	0.02	0.02	0.02
12th	0.02	0.02	0.02	0.02	0.02
13th	0.02	0.01	0.01	0.01	0.02
14th	0.02	0.01	0.01	0.01	0.02
15th	0.02	0.01	0.01	0.01	0.02

Notes. This table reports the 1st to 15th eigenvalues in a descending order for the covariance matrixes of the raw and scaled macro variables, in which the scaling parameter is the predictive slope of the variable on the forecasted target. The eigenvalues are normalized to have a sum of one.

and scaled macro variables, for which the scaling parameter is the predictive slope of the variable on the forecasted target. The eigenvalues are normalized to sum up to one and reflect the dominance of each factor in explaining the total variations of the macro variables. We observe that the distributions of the eigenvalues are more concentrated among the first few factors for sPCA than PCA. Specifically, the first PCA factor explains about 15% of the total variation, and the first sPCA factor explains 20%–37% of the total variation depending on the target to be forecasted. The eigenvalues corresponding to the second and third sPCA factors are also much larger than that corresponding to the PCA factors. This result suggests that sPCA, which scales each predictive variable by its predictive power, reorders the latent factors according to their predictive power.

Second, we examine the compositions of the PCA and sPCA factors and plot their loadings on each macro variable. Figure 4 shows that the first PCA factor is related to the real economic condition, which loads mainly on output-, labor-, and housing-related variables. The second and third PCA factors are nominal factors, which load mainly on interest rate- and price-related variables. The fourth and fifth PCA factors load heavily on housing- and interest rate-related variables, and the seventh PCA factor has larger loadings on price-related variables.

In contrast, Figure 5 shows that the loadings of the sPCA factors are much more concentrated within only

a few key variables. Specifically, to forecast inflation, the first sPCA factor loads on interest rate-related variables and the second sPCA factor on price-related variables. To forecast IP growth, the first sPCA factor loads mainly on output- and labor-related variables akin to the first PCA factor, and the third sPCA factor loads solely on interest rate-related variables. To forecast the unemployment rate, the first sPCA factor loads mainly on output- and labor-related variables. Finally, to forecast S&P 500 index volatility, the first sPCA factor loads on labor- and housing-related variables, the second factor on interest rate- and price-related variables, and the fourth factor has larger loadings on interest rate- and money-related variables, respectively.

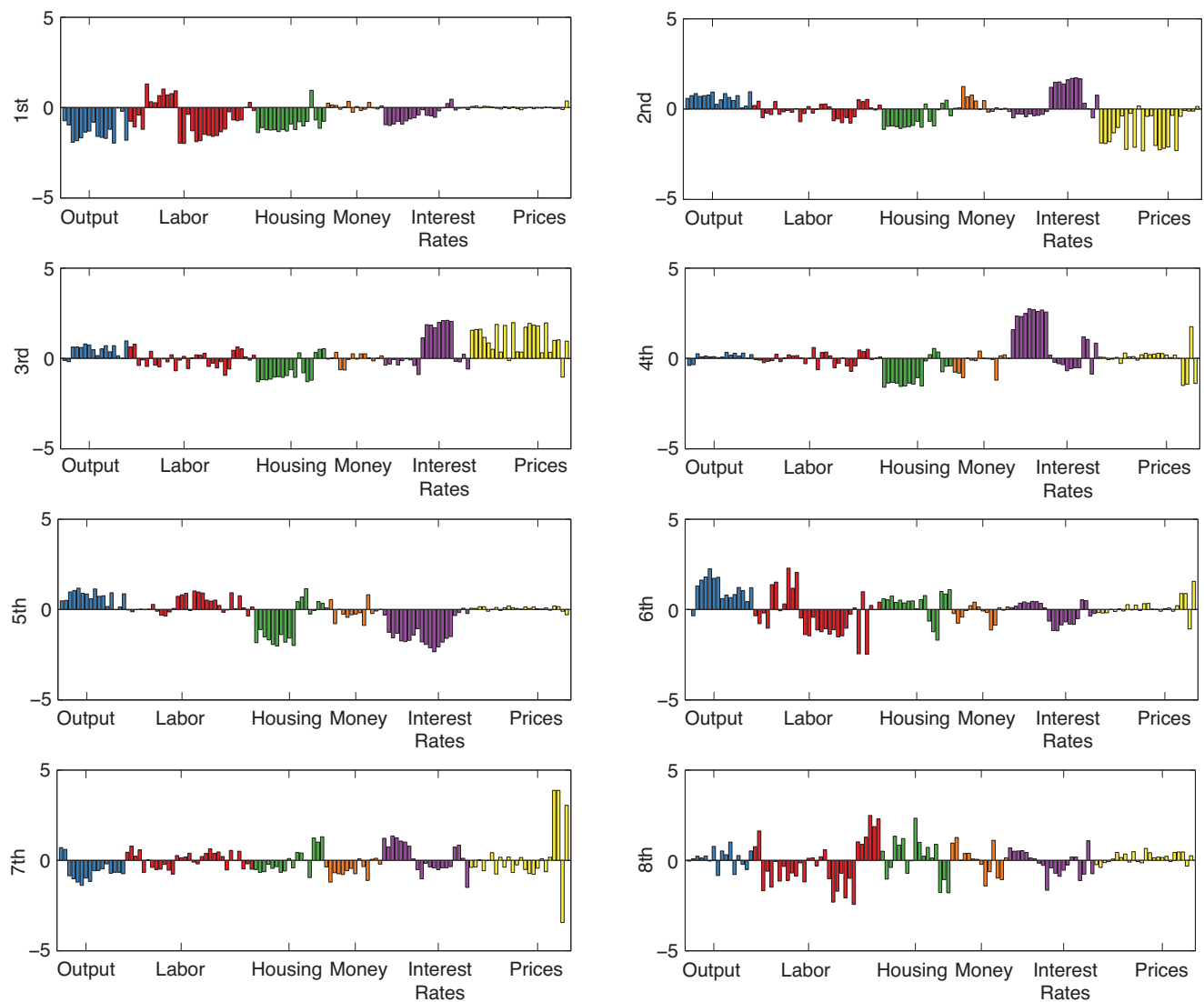
Third, we compare the in-sample forecasting performance between sPCA and PCA. Figure 6 depicts the adjusted R^2 (in percentage) of predicting the one-month-ahead inflation (Panel A), IP growth (Panel B), change in unemployment rate (Panel C), and S&P 500 index volatility (Panel D) by the PCA and sPCA factors, respectively. Within each panel, we consider the forecasting performance with number of factors ranging from 1 to 15.

Panel A of Figure 6 presents the inflation prediction results. It shows that the first PCA factor generates an R^2 less than 1%, which increases to 8% with seven PCA factors and to 9% with 15 PCA factors. Hence, the first PCA factor, which mainly captures the real economic condition, has little forecasting power on inflation. The inflation-relevant factors, which heavily load on interest rate and price variables, have lower ranks as the second and seventh PCA factors. In contrast, by using the sPCA to forecast inflation, the R^2 is about 5.5% with the first factor and about 15% with the first 15 factors.

It is evident in Panel A that sPCA effectively pushes up the eigenvalues corresponding to the inflation-relevant factors so that the first couple of sPCA factors contain the main predictive information, which naturally outperforms the same number of PCA factors. Although including more PCA factors improves the forecasting performance, sPCA consistently beats PCA even with 15 factors (15% versus 9%), lending empirical support to our theoretical result in Section 2.3.1. We observe a similar pattern in Panel D on the S&P 500 index volatility prediction.

Panel B of Figure 6 presents the IP growth prediction results. It shows that the first sPCA and PCA factors have similar predictive power because they both load on the real economic condition-related variables, such as output and labor. With five factors, the R^2 of sPCA becomes larger and reaches about 16%, and the PCA underperforms with an R^2 about 10%. However, with 15 factors, PCA eventually catches up, and it performs similarly to sPCA with an R^2 about 17%. We obtain

Figure 4. (Color online) Loadings of the PCA Factors on Macro Variables



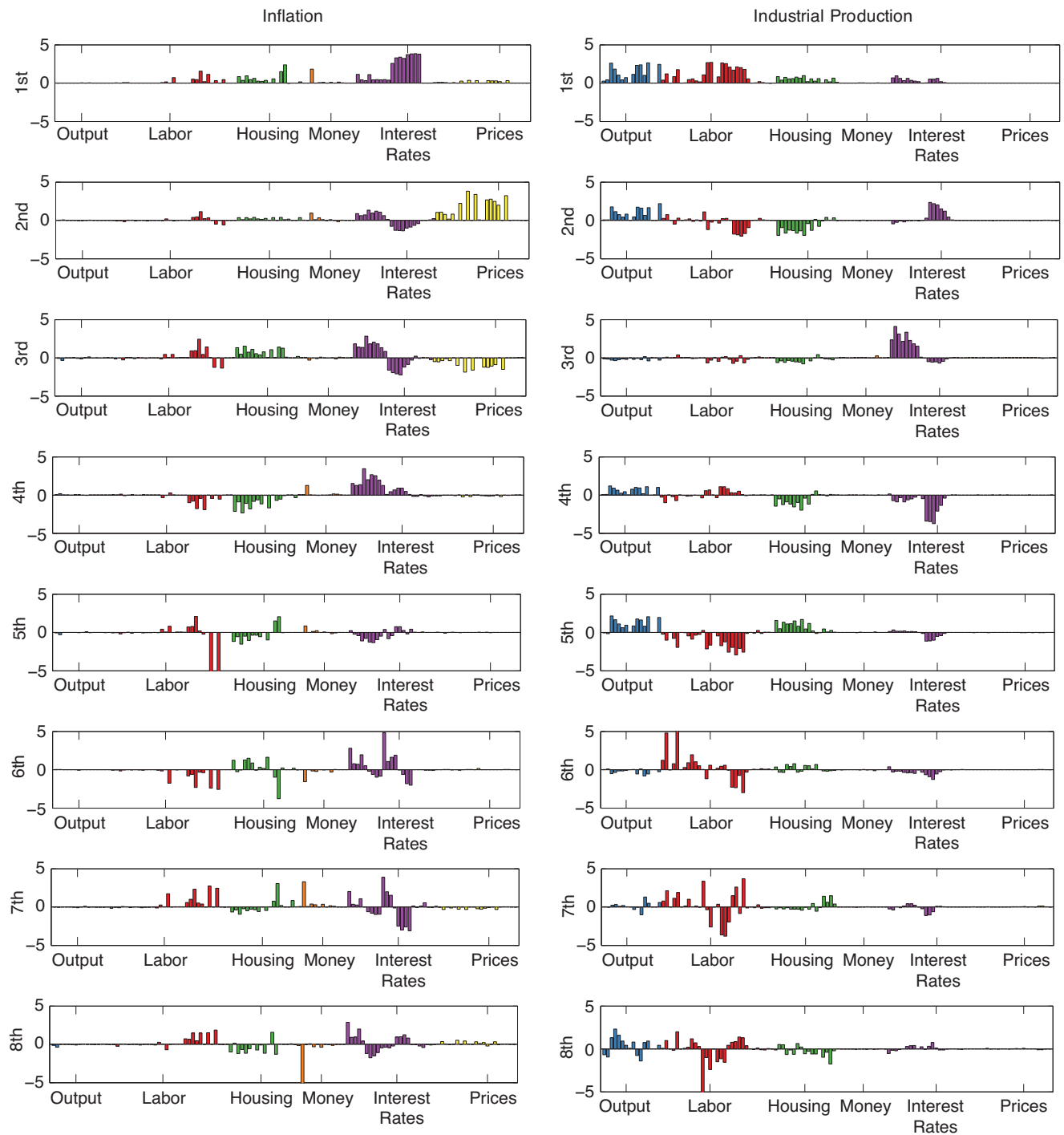
Notes. This figure plots the loadings of the first to eighth PCA factors on the 123 macro variables. The macro variables are collected at a monthly frequency from the FRED-MD data set of McCracken and Ng (2016), consisting of output and income (No. 1–16), labor market (No. 17–47), consumption and housing (No. 48–64), money and credit (No. 65–78), interest and exchange rate (No. 79–99), and prices (No. 100–123). The sample period is January 1960–December 2019.

similar results in Panel C on the unemployment rate prediction. Overall, the results in Panels B and C indicate that, when just using the first few factors, the sPCA method is more likely to dominate the PCA method. When including more factors, the performance of PCA could eventually catch up and converge to that of sPCA, consistent with our theoretical results in Section 2.3.2.

3.3. Out-of-Sample Results

In this section, we explore the out-of-sample forecast performance.

Figure 7 plots the out-of-sample R^2_{OS} (in percentage) of predicting one-month-ahead inflation (Panel A), IP growth (Panel B), unemployment rate (Panel C), and S&P 500 index volatility (Panel D) by up to eight PCA and sPCA factors, respectively. The R^2_{OS} is computed against an autoregressive model with lagged target as a benchmark, and a higher R^2_{OS} value indicates better forecasting performance. Although the in-sample R^2 is generally nondecreasing as we include more factors, the out-of-sample R^2_{OS} is not necessarily monotonic because the increased estimation errors could

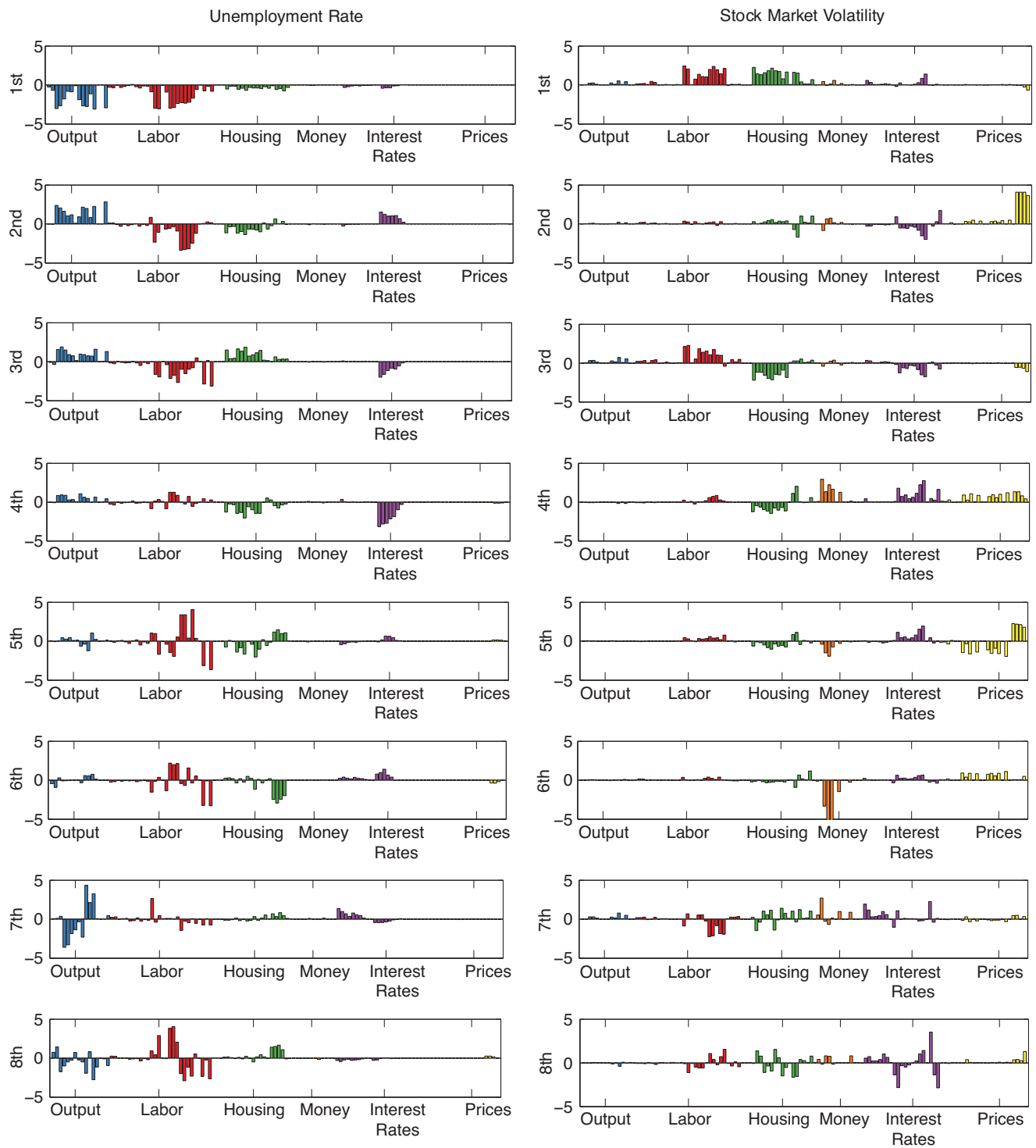
Figure 5. (Color online) Loadings of the sPCA Factors on Macro Variables

outweigh the marginal benefit of including an additional factor, which translates into deteriorated out-of-sample forecasting performance. All factors and predictive regressions are recursively estimated with an expanding window scheme. The initial

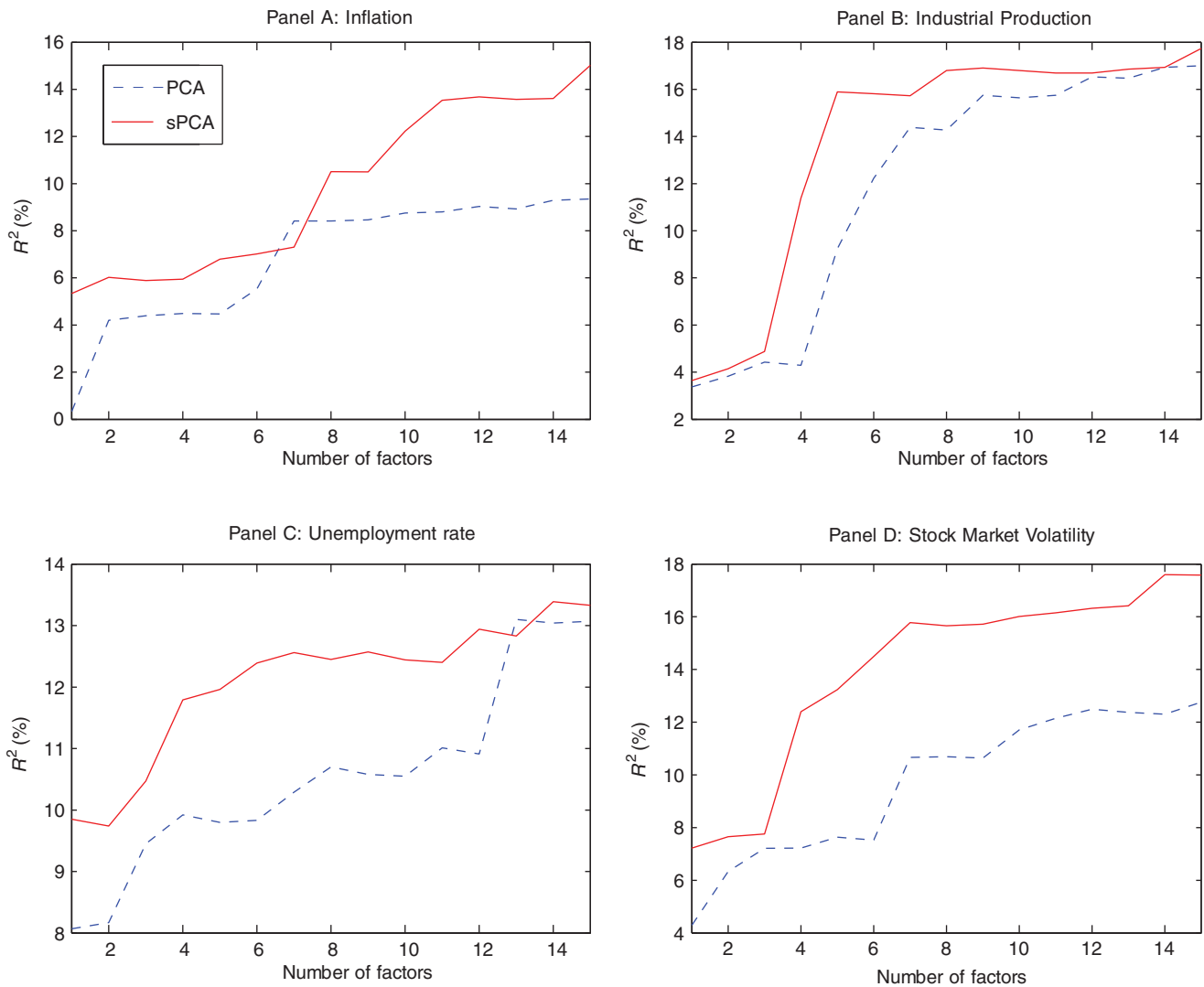
estimation window ranges from January 1960 to December 1984, and the out-of-sample evaluation period is from January 1985 to December 2019.

Panel A of Figure 7 shows that, to forecast inflation, the first sPCA factor generates an R_{OS}^2 of 7%, which

Figure 5. (Continued)



Notes. This figure plots the loadings of the first to eighth sPCA factors on 123 macro variables when predicting the one-month-ahead inflation, industrial production growth, unemployment rate, and S&P 500 index volatility, respectively. The macro variables are collected at a monthly frequency from the FRED-MD data set of McCracken and Ng (2016), consisting of output and income (No. 1–16), labor market (No. 17–47), consumption and housing (No. 48–64), money and credit (No. 65–78), interest and exchange rate (No. 79–99), and prices (No. 100–123). The sample period is January 1960–December 2019.

Figure 6. (Color online) In-Sample Forecasting Performance of the PCA and sPCA Factors

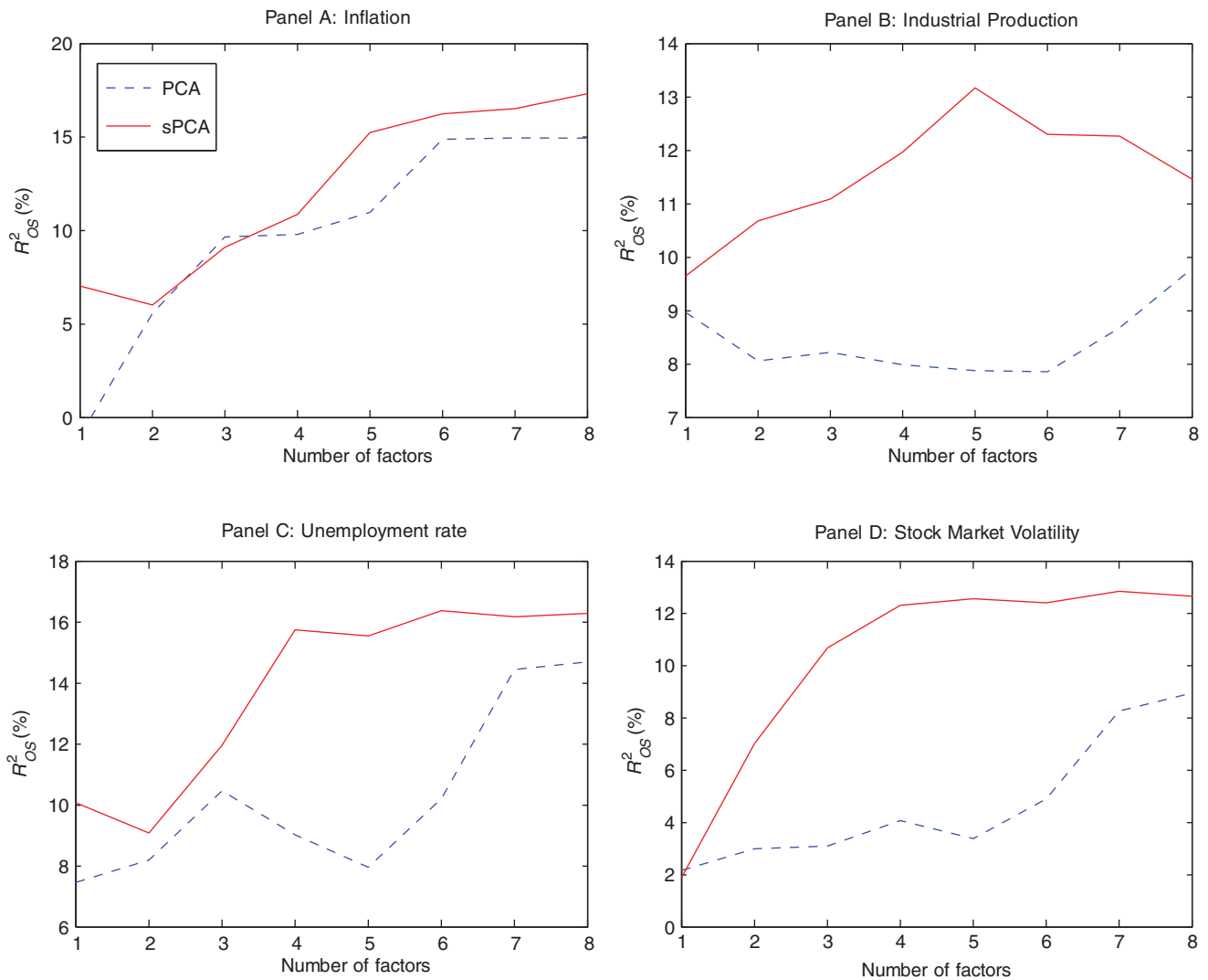
Notes. This figure plots the in-sample R^2 (in percentage) of predicting the one-month-ahead inflation (Panel A), industrial production growth (Panel B), unemployment rate (Panel C), and S&P 500 index volatility (Panel D) by using the PCA and sPCA factors extracted from 123 macro variables, respectively. We control for the lags of the target the number of lags selected by BIC. The macro variables are collected at a monthly frequency from the FRED-MD data set of McCracken and Ng (2016), consisting of output and income (No. 1–16), labor market (No. 17–47), consumption and housing (No. 48–64), money and credit (No. 65–78), interest and exchange rate (No. 79–99), and prices (No. 100–123). The sample period is January 1960–December 2019.

increases to about 18% when including eight sPCA factors. In sharp contrast, the first PCA factor has no predictability. Although an eight-factor PCA model catches up and generates an R^2_{OS} of 15%, it still underperforms sPCA. Panels B–D exhibit similar patterns when forecasting IP growth, unemployment, and S&P 500 index volatility. In summary, Figure 7 suggests that sPCA beats PCA in out-of-sample forecasting at least up to the first eight factors, which is consistent with the earlier in-sample results.

Table 4 compares the out-of-sample forecasting performance of the sPCA with alternative machine learning

methods that also apply to a large number of predictors. Specifically, we consider the tPCA, PLS, LASSO, elastic-net, and ridge regressions, respectively. Among these methods, tPCA relies on the hard threshold criteria in Bai and Ng (2008) to select targeted predictors, PLS extracts multiple factors iteratively as suggested in Kelly and Pruitt (2013), the penalized forecast regressions are estimated with the standardized values of macro variables, and the strength of penalties are determined recursively through threefold cross-validations. We also explore the LASSO regression on principal components of the macro variables, that is, eigenvector-rotated macro

Figure 7. (Color online) Out-of-Sample Forecasting Performance of the PCA and sPCA Factors



Notes. This figure plots the out-of-sample R^2_{OS} (in percentage) of predicting one-month-ahead inflation (Panel A), industrial production growth (Panel B), unemployment rate (Panel C), and S&P 500 index volatility (Panel D) by using the PCA and sPCA factors extracted from 123 macro variables, respectively. All factors and predictive regressions are recursively estimated with an expanding window scheme. The initial estimation window ranges from January 1960 to December 1984, and the out-of-sample evaluation period is from January 1985 to December 2019.

variables. Aside from the baseline sPCA that scales each predictor with its regression slope, we also consider a modified version, $sPCA^*$, that scales each predictor with the t -value of the regression slope. The empirical results in Table 4 show that, across all four target-driven methods considered, sPCA displays comparable, and in many cases better, performance with these alternative forecasting methods.

4. Conclusion

In this paper, we propose a novel supervised learning technique, sPCA, for forecasting with many predictors.

The sPCA improves the traditional PCA by scaling each predictor with its predictive slope on the target to be forecasted and assigns more (less) weights to those predictors that have stronger (weaker) forecasting power.

Theoretically, we motivate our sPCA with a general factor framework that allows the factors to be either weak or strong and show that sPCA can outperform PCA under a wide range of appropriate conditions on data. Extensive simulations also support the superior performance of sPCA in a finite sample setting. In real data applications, we show that sPCA works well in forecasting the U.S. inflation, industrial production growth, unemployment rate, and S&P 500 index

Table 4. Out-of-Sample R_{OS}^2 of Forecasting with Macro Variables

	Inflation	IP	Unemploy	Volatility
PCA	10.98***	7.88***	7.96***	3.39*
sPCA	15.24***	13.17***	15.55***	12.56**
sPCA ^o	17.52***	13.20***	15.76***	11.36**
tPCA	15.08***	9.10***	14.72***	10.28**
PLS	12.27***	0.10***	8.84***	7.97**
LASSO	16.66***	8.22***	12.60***	14.57***
ENet	16.94***	8.86***	11.91***	14.04***
Ridge	9.40***	9.02***	10.61***	12.63**
LASSO (PCs)	17.64***	4.20***	9.02***	6.10***

Notes. This table reports the out-of-sample R_{OS}^2 (in percentage) of predicting the one-month-ahead inflation, industrial production, unemployment rate, and S&P 500 index volatility with 123 macro variables. The methods include the PCA, sPCA, tPCA, PLS, LASSO, elastic-net, and ridge regression. sPCA^o refers to an alternative sPCA method that scales each predictor by its t -value in regression rather than the regression slope. We also consider LASSO regression on principal components of macro variables. For the PCA, sPCA, tPCA, PLS, and sPCA^o, we employ the first five factors. For tPCA, we use the hard threshold criteria in Bai and Ng (2008) to select the target predictors. For LASSO, elastic-net, and ridge regression, we estimate forecast via a threefold cross-validation. All factors and predictive regressions are recursively estimated with an expanding window scheme. The initial estimation window ranges from January 1960 to December 1984 and the out-of-sample evaluation period is from January 1985 to December 2019. Statistical significance for R_{OS}^2 is based on the Clark and West (2007) test.

*, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

volatility with a panel of 123 macro variables, and it outperforms PCA both in- and out-of-sample in general. In addition, sPCA performs similarly or better than several other supervised learning techniques commonly used in a big data environment.

Acknowledgments

The authors thank Kay Giesecke (the department editor), an anonymous associate editor, and three anonymous referees for very insightful and helpful comments that improved the paper significantly. They also thank Hui Chen, Zhanhui Chen, Anna Cieslak, Zhi Da, James D. Hamilton, Wenxin Huang, Ben Jacobsen, Toshio Kimura, Junye Li, Roger Loh, Sydney C. Ludvigson, Neil Pearson, Markus Pelger, David Rapach, Allan Timmermann, Jun Tu, Dacheng Xiu, and Hao Zhou as well as seminar and conference participants at the Central University of Finance and Economics, Peking University, Tsinghua University, Washington University in St. Louis, and 2018 Singapore Management University Finance Summer Camp for valuable comments and suggestions. Authors are listed alphabetically and contributed equally to this paper.

Endnotes

¹ We sincerely thank the associate editor for inspiring us to consider sPCA in a weak factor framework.

² Bai (2003) shows that $\frac{1}{\sqrt{T}}\|\hat{F}^{PCA} - \hat{F}R_{PCA}^*\| = O_p(N^{-1/2}) + O_p(T^{-1/2})$. However, we note that this result is not the sharpest. In the subsequent analysis on the limiting distribution, it is seen that this result can be improved by replacing $T^{-1/2}$ with T^{-1} .

³ Note that, when $\nu = 1$, the condition $N^{1-\nu/2}/T \rightarrow 0$ reduces to $\sqrt{N}/T \rightarrow 0$.

⁴ Huang et al. (2021) further compare the asymptotic forecast performance of sPCA with the PLS approach (Kelly and Pruitt 2013, 2015) on which Lin et al. (2018) provide a combination method interpretation.

⁵ Please see <http://research.stlouisfed.org/econ/mccracken/sel/>.

References

- Ahn SC, Horenstein AR (2013) Eigenvalue ratio test for the number of factors. *Econometrica* 81(3):1203–1227.
- Bai J (2003) Inferential theory for factor models of large dimensions. *Econometrica* 71(1):135–171.
- Bai J (2004) Estimating cross-section common stochastic trends in nonstationary panel data. *J. Econometrics* 122(1):137–183.
- Bai J, Li K (2012) Statistical analysis of factor models of high dimension. *Ann. Statist.* 40(1):436–465.
- Bai J, Ng S (2002) Determining the number of factors in approximate factor models. *Econometrica* 70(1):135–172.
- Bai J, Ng S (2004) A PANIC attack on unit roots and cointegration. *Econometrica* 72(4):1127–1177.
- Bai J, Ng S (2006) Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74(4):1133–1150.
- Bai J, Ng S (2008) Forecasting economic time series using targeted predictors. *J. Econometrics* 146(2):304–317.
- Chen L, Pelger M, Zhu J (2020) Deep learning in asset pricing. Working paper, Stanford University, Stanford, CA.
- Clark TE, West KD (2007) Approximately normal tests for equal predictive accuracy in nested models. *J. Econometrics* 138(1):291–311.
- Connor G, Korajczyk RA (1986) Performance measurement with the arbitrage pricing theory: A new framework for analysis. *J. Financial Econom.* 15(3):373–394.
- Connor G, Hagmann M, Linton O (2012) Efficient semiparametric estimation of the Fama–French model and extensions. *Econometrica* 80(2):713–754.
- Daniele M, Pohlmeier W, Zagidullina A (2019) Sparse approximate factor estimation for high-dimensional covariance matrices. Working paper, University of Konstanz, Germany.
- Freyberger J, Neuhierl A, Weber M (2020) Dissecting characteristics nonparametrically. *Rev. Financial Stud.* 33(5):2326–2377.
- Giglio S, Xiu D (2021) Asset pricing with omitted factors. *J. Political Econom.* Forthcoming.
- Gu S, Kelly BT, Xiu D (2021) Autoencoder asset pricing models. *J. Econom.* 222(1):429–450.
- Gu S, Kelly BT, Xiu D (2020) Empirical asset pricing via machine learning. *Rev. Financial Stud.* 33(5):2223–2273.
- Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, New York).
- Huang D, Jiang F, Tu J, Zhou G (2015) Investor sentiment aligned: A powerful predictor of stock returns. *Rev. Financial Stud.* 28(3):791–837.
- Huang D, Jiang F, Li K, Tong G, Zhou G (2021) Are bond returns predictable with real-time macro data? Working paper, Singapore Management University, Singapore.
- Kelly B, Pruitt S (2013) Market expectations in the cross-section of present values. *J. Finance* 68(5):1721–1756.
- Kelly B, Pruitt S (2015) The three-pass regression filter: A new approach to forecasting using many predictors. *J. Econometrics* 186(2):294–316.
- Kelly BT, Pruitt S, Su Y (2019) Characteristics are covariances: A unified model of risk and return. *J. Financial Econom.* 134(3):501–524.

- Kelly BT, Pruitt S, Su Y (2020) Instrumented principal component analysis. Working paper, Yale University, New Haven, CT.
- Kim S, Korajczyk RA, Neuhierl A (2021) Arbitrage portfolios. *Rev. Financial Stud.* Forthcoming.
- Lettau M, Pelger M (2020a) Estimating latent asset-pricing factors. *J. Econometrics* 218(1):1–31.
- Lettau M, Pelger M (2020b) Factors that fit the time series and cross-section of stock returns. *Rev. Financial Stud.* 33(5): 2274–2325.
- Light N, Maslov D, Rytchkov O (2017) Aggregation of information about the cross section of stock returns: A latent variable approach. *Rev. Financial Stud.* 30(4):1339–1381.
- Lin H, Wu C, Zhou G (2018) Forecasting corporate bond returns with a large set of predictors: An iterated combination approach. *Management Sci.* 64(9):4218–4238.
- Ludvigson SC, Ng S (2007) The empirical risk-return relation: A factor analysis approach. *J. Financial. Econom.* 83(1):171–222.
- McCracken MW, Ng S (2016) FRED-MD: A monthly database for macroeconomic research. *J. Bus. Econom. Statist.* 34(4):574–589.
- Onatski A (2012) Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* 168(2):244–258.
- Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos. Magazine J. Sci.* 2:559–572.
- Pelger M (2020) Understanding systematic risk: A high-frequency approach. *J. Finance* 75(4):2179–2220.
- Pelger M, Xiong R (2021) Interpretable sparse proximate factors for large dimensions. *J. Bus. Econom. Statist.* Forthcoming.
- Rapach D, Zhou G (2021) Sparse macro factors. Working paper, Saint Louis University, St. Louis, MO.
- Stock JH, Watson MW (2006) Forecasting with many predictors. Elliot G, Granger CW, Timmermann A, eds. *Handbook of Economic Forecasting*, vol. 1 (Elsevier, Amsterdam, Netherlands), 515–554.