# Supervised Principal Component Analysis for the Purpose of Forecasting

By WANJUN LI[1]

Rutgers University

October 2020

As a classic dimension reduction model, Principal Component Analysis (PCA) summarizes information based on the regressors, but it fails to incorporate the goal of forecasting in the factor extraction step. Current literature either preselects a subset of the regressors before PCA or select among all factors after PCA. I propose a novel supervised PCA (SPCA) to combine dimension reduction and factor selection. The convergence of the SPCA factors is established. Together with Monte Carlo simulations, I find that SPCA outperforms PCA when N is not so large under conventional assumptions, and when models are misspecified, for instance, when factors important in explaining regressors are not as important in predicting target variables. SPCA shows excellent improvement in predicting bond risk premia and some macro variables, with a relative Mean Squared Forecast Errors (MSFE) as low as 0.55 using PCA as a baseline. I also show that my supervised design is compatible with other types of PCA (Sieve-LS Covariance), and it is more powerful, allowing for nonlinear predictive relationships. The best performing model is combining SPCA and Neural Network, which significantly reduced the MSFE by 70% comparing to PCA.

**Keywords:** PCA, forecasting, bond risk premia, machine learning

**JEL Codes:** C22, C51, C52, C53, C55

[1]Department of Economics, Rutgers University, New Brunswick, NJ, 08901. Email: wl392@economics.rutgers.edu

# I. INTRODUCTION

My capabilities to acquire a large scale of data have brought many opportunities and challenges in forecasting. More information helps to explain the economic states comprehensively. However, it also contains irrelevant disturbances. I propose a method to summarize information for forecasting.

Traditional Ordinary Lease Square (OLS) has some pitfalls in large dimension. The forecast error is proportional to $\sigma_\varepsilon^2 N/T$, and thus it grows as N/T increases Forni *et al.* (2000). Moreover, in empirical studies, the regressors are rarely orthogonal, which makes $\hat{\beta}$ not unbiased anymore. As dimension increases, the collinearity problem becomes extreme. The matrix $X'X$ can suffer from near singularity. Besides, among a large number of regressors, many of them could be noises for predicting the target variable, adding more variability.

Because of these drawbacks of OLS, there are many studies in the literature to shrink the dimension. One popular way is by Principal Component Analysis (PCA) and factor model (Stock and Watson, 2002a; Forni *et al.*, 2005). The factor model decomposes a high dimensional series into two stationary, orthogonal parts. One is called the common factors, and the other one is idiosyncratic shocks. N is the number of cross-sectional units, and T is the time length.

### *(a) Existing Methods Based on PCA*

The Static factor model is as follows (see Stock and Watson (1998))

$$x_{it} = \lambda_i' F_t + e_{it}$$

$$i = 1, ..., N, t = 1, ..., T$$

And to write this in the matrix form,

$$X = F\Lambda' + e$$

where $X = (X_1', ..., X_N')$ is a T by N matix.

Then the factors are used to predict the future y according to

$$y_{t+h} = \beta' F_t + \varepsilon_{t+h},$$

To extract the factors, the objective is to solve:

$$\min_{\Lambda^k, F^k} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \lambda_i^{k'} F_t^k)^2$$

PCA begins with obtaining the eigenvalues and eigenvectors of $XX'$ and rank the eigenvalues in descending orders. Let $\hat{F}$ be the eigenvectors corresponding the first k eigenvalues with the normalization of $\hat{F}'\hat{F}/T = I_k$. Then $\hat{\Lambda}$ can be calculated from OLS $\hat{\Lambda} = (\hat{F}'\hat{F})^{-1}\hat{F}'X = \hat{F}'X/T$. When $N < T$ a more efficient way to solve this is to obtain $\bar{\Lambda}$ as the first k eigenvectors of $X'X$, and do the normalization $\bar{\Lambda}'\bar{\Lambda}/N = I_k$. And again from OLS, $\bar{F} = X\bar{\Lambda}/N$.

Stock and Watson (1998) construct factors from the weighted averages of the regressors. They show that the forecasts based on the estimated factors are efficient. Their empirical work shows that these forecasts win many baseline models. Many other researchers apply factor models to forecast macroeconomic variables (den Reijer *et al.*, 2005; Banerjee and Marcellino, 2006; Forni *et al.*, 2005; Schumacher and Breitung, 2008). Forni *et al.* (2000) relax the orthogonality assumption on the error terms and prove consistency when both time length and the number of series go to infinity.

While PCA is a classic dimension reduction method, it does not incorporate the purpose of forecasting. In the factor extraction process, it only condenses data according to the regressors covariance matrix and completely neglect how factors and the target variables are related.

*(b) Supervised PCA*

For PCA to work well in forecasting, the factors extracted need to be informative about the target variables.

Double shrinkage becomes a popular approach for empirical study, and recent literature has looked into two directions (Kim and Swanson, 2014, 2016). The first path is to get all the factors from PCA. It is then combined with statistical methods such as LASSO to shrink the dimension of these factors by minimizing the prediction errors. The other avenue is to do statistical shrinkage first on the regressors. Once the most relevant subset of regressors is selected, PCA is used to summarize the information. The selecting criterion can be model-free and use simply the correlations of the regressors with the predictor Bair *et al.* (2006); Bai and Ng (2008).

Another way is to impose supervision while reducing dimension. This direction has not been heavily explored yet.

A close analogy is Partial Least Square (PLS), which solves PLS solves

$$\max_{\|w\|} corr^2(y, Xw)var(Xw)$$

Put in another way, PLS explains the variations of the regressors and the correlations between the regressors and the target variable at the same time.

As for supervised PCA, Barshan *et al.* (2011) applied the Hilbert-Schmidt independence criterion to get a different covariance matrix that involves both X and Y. Their supervised PCA solves

$$\arg\max_{F} \quad \text{tr}\left(F^T XLX^T F\right)$$

subject to $F^T F = I$, where L is the kernel matrix of target variable y, and data is centered. The kernel parameter is obtained using cross-validation. Their supervised PCA show significant improvement over other approaches in visualization, classification and regression.

The three-pass regression filter. For each predictor, a time series regression is run on the proxies (or the target variable directly). Then the slope estimate represents how sensitive the predictor is to the latent factors that are relevant to the target variable. In the second pass, a cross-section regression of the predictors on the first pass slope estimates is run for each time period t. The coefficients in the second pass back out estimates of

the factors. The third pass is a time series regression of the target variable on the latent factors. The three-pass regressions can easily be applied to unbalanced data. If the panel is balanced, there's a closed-form solution as well for fast computation.

How to find proxies relevant to the target? The authors proposed an iterative way. First, initialize the proxy with the target variable itself. Then, run the 3PF and obtain an estimation of the target. Use the residuals of the target estimation as a new proxy and add them to the proxy list. Redo this process L times. The 3PF is improved by adding additional proxies in each iteration.

Another similar approach is the scaled PCA. The predictors are scaled by the coefficients from regressing the target on the predictors. Instead of using the panel of predictors directly to extract factors, this method uses the projections of the target on the predictors to extract factors. When a predictor is irrelevant to the target, the project of the target on this predictor is small.

*(c) Main Contributions*

My primary contributions are two-fold. First, I provide a new method of supervised PCA (SPCA) to combine dimension shrinkage and data selection for improved accuracy in forecasting. Under standard model assumptions (Bai and Ng, 2002, 2006; Bai, 2003), I establish the convergence of factors and identify that if N is not so large, but T is large, SPCA can outperform PCA. Simulations confirm this and provide more insights when SPCA dominates PCA in misspecified models.

Second, I apply SPCA together with other methods to predict bond risk premia and 10 macro variables data. Other methods cover two aspects. The first aspect is to use a different covariance matrix, so I can check if my supervised thinking applies to other modified PCA. The second aspect is to accommodate nonlinear predictive relationship between factors and the target variables. Both aspects add flexibility to SPCA and make it more powerful.

The remainder of this paper is organized as follows. In section 2, the estimation methodology is presented. Section 3 presents the assumptions. Section 4 demonstrates

the asymptotic theorem. Section 5 shows the Monte Carlo simulation results. Section 6 gives empirical applications. Section 7 concludes the paper. Proves are included in the Appendix.

## II. METHODOLOGY

I combine the dimension reduction and selection process together. Recall $X_{it} = \lambda_i' F_t + e_t$, and $y_{t+h} = \beta' F_t + u_t$.

### *(a) Estimated factors*

I use regressors $(X_1, ..., X_{T-h})$ and the target variable $(y_{1+h}, ...y_T)$ to get the $\widehat{\lambda}, \widehat{\beta}$ and $(\widetilde{F}_1, ..., \widetilde{F}_{T-h})$ in equation 1.

$$(1) \quad V(F, \Lambda, \beta) = \min \left\{ w \sum_{i=1}^{N} \sum_{t=1}^{T-h} \left(x_{it} - \lambda_i' F_t\right)^2 + (1-w) \sum_{t=1}^{T-h} \left(y_{t+h} - F_t \beta\right)^2 \right\}$$

To forecast $y_{T+h}$, I need an estimate for $F_T$. And this can be calculated from with OLS. For differentiation, I denote the estimates of Factors as $\widetilde{F}_t$ when $t = 1, ..., T-h$, and $\widehat{F}_t$ when $t = T-h+1, ...., T$. Then $\widehat{F}_T = X_T \widehat{\Lambda} \left(\hat{\Lambda}' \widehat{\Lambda}\right)^{-1}$

The solution for equation (1) is comparable to the conventional PCA. Use $X_u, y_u$ to denote the T-h periods of data used in equation (1). Solving equation (1) is equivalent to solving

$$(2) \quad arg\max \left\{ (w)\, tr\left[F'\left(X_u X_u'\right) F\right] + (1-w)\, tr\left[F'\left(y_u y_u'\right) F\right] \right\}$$

### *(b) Solving SPCA*

$$V(F, \Lambda, B) = \min \left\{ w \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \lambda_i' F_t\right)^2 + (1-w) \sum_{i=1}^{n} \sum_{t=1}^{T-h} \left(y_{i,t+h} - b_i' F_t\right)^2 \right\},$$

Denote $z_t = y_{t+h}$

$$V(F, \Lambda, B) = \min \left\{ w \|X - F\Lambda'\|^2 + (1-w) \|Z - F\beta\|^2 \right\},$$

$$F'F/T = I$$

$$\Lambda' = \left(F'F\right)^{-1} F'X = F'X/T$$

$$\beta = \left(Fu'Fu\right)^{-1} Fu'Z = F'Y/T$$

$$\|X - F\Lambda'\|^2$$

$$= tr\left[\left(X - F\Lambda'\right)'\left(X - F\Lambda'\right)\right], \ \textit{where tr represents for trace}$$

$$= tr\left[\left(X - F\left(F'F\right)^{-1}F'X\right)'\left(X - F\left(F'F\right)^{-1}F'X\right)\right], \textit{ denote } P_F = F\left(F'F\right)^{-1}F'$$

$$= tr\left[\left(X - P_F X\right)'\left(X - P_F X\right)\right]$$

$$= tr\left[X'\left(1 - P_F\right)'\left(1 - P_F X\right)\right]$$

$$= tr\left[X'X - X'P_F X\right]$$

$$= tr\left[X'\left(1 - P_F\right)X\right]$$

Similarly

$$\|Z - F\beta\|^2 = tr\left[Z'\left(1 - P_F\right)Z\right]$$

The original problem is equivalent to the following problem:

$$arg\max\left\{(w)\,tr\left[X'P_F X\right] + (1-w)\,tr\left[Z'P_F Z\right]\right\}$$

$$= arg\max\ \left\{(w)\,tr\left[X'FF'X\right] + (1-w)\,tr\left[Z'FF'Z\right]\right\}$$

$$= arg\max\left\{(w)\,tr\left[F'\left(XX'\right)F\right] + (1-w)\,tr\left[F'\left(ZZ'\right)F\right]\right\}$$

$\widetilde{F}$ is the $\sqrt{T-h}$ times the eigenvectors corresponding to the k largest eigenvalues of the matrix $[w\left(X_u X_u'\right) + (1-w)\left(y_u y_u'\right)]$. $\widetilde{F}$ is normalized such that $\widetilde{F}'\widetilde{F}/(T-h) = I_k$. $\widehat{\Lambda}' = \widetilde{F}'X_u/(T-h)$ and $\widehat{\beta} = \widetilde{F}y'/(T-h)$.

*(c) Estimated number of factors*

The number of factors k for supervised PCA is chosen by a modified BIC. The loss I am interested in is y instead of x: $loss(y) = \|\hat{y} - y\|^2/T$.

$$BIC(k) = log(loss(y)) + (k+1)\frac{log(y)}{T}.$$

*(d) Estimated w*

The weighting w is what pulls the extracted factors closer to the targeting predictions. When the big data covers accurate information about y, the original PCA works well. My Supervised PCA includes this case by setting the optimal w to 1. When the factors extracted from X only predict poorly about $y_{T+h}$, more weights will be put on the second term in equation (1). Notice w can be a very small positive number, but not 0. I still need some weights on the big data because it contains the most comprehensive information.

The optimal weight solves the equation (3)

$$(3) \quad \min_{w} \quad E\left(y_{T+h} - \widehat{y}_{T+h}\right)^2$$

To estimate the optimal weight, I can use cross validation. Call each window as sample $S_i, i = 1, ..., P$. Now divide each $S_i$ into an in-sample training period $RR_i$, and an in-sample validation period $V_i$.

$w_i$ for $S_i$ is estimated by minimizing the h period ahead MSFE within $S_i$.

$$(4) \quad \underset{w_i}{MSFE_h} = \frac{1}{V_i} \sum_{t=RR_i-h+1}^{TT-h} (y_{t+h} - \hat{y}_{t+h})^2$$

where TT is the length of $S_i$. $\hat{w}_i$ is the estimated weight for $S_i$ to make h ahead out of sample prediction. I set the first 90% time series data in each window as its training period, and the rest 10% as its validation period.

## III. Asymptotic Theorem

I follow the assumptions used in (Bai and Ng, 2002, 2006; Bai, 2003) for estimating factors and their loadings. Please refer to Appendix A for the assumptions I need.

THEOREM 1: For any fixed $k \geq 1$, there exists a $(r \times k)$ matrix $H^k$ with rank $\left(H^k\right) = \min\{k, r\}$, and $\delta^{-2} = O_p(w^2(\frac{1}{N} + \frac{1}{T})) + O_p(1-w)^2$ such that

$$\frac{\delta^2}{T} \sum_{s=1}^{T} \left\| \widetilde{F}_s - H' F_s^0 \right\|^2 = O_p(1)$$

Theorem 1 is proved as below.

*(a) Theorem 1*

Want to find if there is a $\delta$ such that LEMMA A.1:

$$\delta^2 \left( \frac{1}{T} \sum_{t=1}^{T} \| \widetilde{F}_t - H' F_t^0 \|^2 \right) = O_p(1).$$

$\widetilde{F}^k = N^{-1} X \widetilde{\Lambda}^k$, and $\widetilde{\Lambda}^k = T^{-1} X' \widetilde{F}^k$. $V_{NT}$ is the diagonal matrix of the top r largest eigenvalues of

$$\frac{w}{NT} XX' + \frac{1-w}{T} yy'$$

Let $H = w\left(\Lambda^{0'}\Lambda^0/N\right)\left(F^{0'}\tilde{F}/T\right) + (1-w)\left(\beta^0\beta^0\right)\left(F^{0'}\tilde{F}/T\right)$, I have

$$\begin{aligned} \widetilde{F}_t - H' F_t^0 = V_{NT}^{-1} & \left( \frac{w}{T} \sum_{s=1}^{T} \widetilde{F}_s \gamma_N(s,t) + \frac{w}{T} \sum_{s=1}^{T} \widetilde{F}_s \zeta_{st} + \frac{w}{T} \sum_{s=1}^{T} \widetilde{F}_s \eta_{st} + \frac{w}{T} \sum_{s=1}^{T} \widetilde{F}_s \xi_{st} \right) \\ + V_T^{-1} & \left( \frac{1-w}{T} \sum_{s=1}^{T} \widetilde{F}_s u_s u_t + \frac{1-w}{T} \sum_{s=1}^{T} \widetilde{F}_s F_s^{0'} \beta^0 u_t + \frac{1-w}{T} \sum_{s=1}^{T} \widetilde{F}_s F_t^{0'} \beta^0 u_s \right) \end{aligned}$$

where

$$\begin{aligned} \zeta_{st} &= \frac{e_s' e_t}{N} - \gamma_N(s,t), \\ \eta_{st} &= F_s^{0'} \Lambda^{0'} e_t / N, \\ \xi_{st} &= F_t^{0'} \Lambda^{0'} e_s / N. \end{aligned}$$

Notice that $\|H\| = O_p(1)$ is implied by assumptions A, B and $\widetilde{F}_k'\widetilde{F}_k/T = I_k$ . Because

$$\|H\| \leq w\|\Lambda^{0'}\Lambda^0/N\|\|\widetilde{F}'\widetilde{F}/T\|^{1/2}\|\|F^{0'}F^0/T\|^{1/2} + (1-w)\|\beta^0\beta^0\|\|\widetilde{F}'\widetilde{F}/T\|^{1/2}\|\|F^{0'}F^0/T\|^{1/2}$$

$(1/T)\sum_{t=1}^{T}\|\widetilde{F}_t - H'F_t^0\|^2 <= 1/T\sum_{t=1}^{T}(a_t+b_t+c_t+d_t+e_t+f_t+g_t)$, where

$$\begin{aligned}
a_t &= (\frac{w}{T})^2\|\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)\|^2 \\
b_t &= (\frac{w}{T})^2\|\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}\|^2 \\
c_t &= (\frac{w}{T})^2\|\sum_{s=1}^{T}\widetilde{F}_s\eta_{st}\|^2 \\
d_t &= (\frac{w}{T})^2\|\sum_{s=1}^{T}\widetilde{F}_s\xi_{st}\|^2 \\
e_t &= (\frac{1-w}{AT})^2\|\sum_{s=1}^{T}\widetilde{F}_s u_s' u_t\|^2 \\
f_t &= (\frac{1-w}{AT})^2\|\sum_{s=1}^{T}\widetilde{F}_s F_s^{0'}\beta^0 u_t\|^2 \\
g_t &= (\frac{1-w}{AT})^2\|\sum_{s=1}^{T}\widetilde{F}_s F_t^{0'}\beta^0 u_s\|^2
\end{aligned}$$

According to Bai and Ng (2002), we have the following results:

$$\begin{aligned}
T^{-1}\sum_{t=1}^{T}a_t &= O_p(\frac{w^2}{T}) \\
T^{-1}\sum_{t=1}^{T}b_t &= O_p(\frac{w^2}{N}) \\
T^{-1}\sum_{t=1}^{T}c_t &= O_p(\frac{w^2}{N}) \\
T^{-1}\sum_{t=1}^{T}d_t &= O_p(\frac{w^2}{N})
\end{aligned}$$

For the remaining items, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} e_t &= \frac{1}{T}\sum_{t=1}^{T}\Big((\frac{1-w}{T})^2\|\sum_{s=1}^{T}\widetilde{F}_s u_s' u_t\|^2\Big) \\
&\leq \frac{1}{T}(1-w)^2\big(\frac{1}{T}\|\sum_{s=1}^{T}\widetilde{F}_s\|^2\big)\cdot\big(\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\|u_s' u_t\|^2\big) \\
&= O_p(\frac{(1-w)^2}{T})
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} f_t &= \frac{1}{T}\sum_{t=1}^{T}\Big((\frac{1-w}{T})^2\|\sum_{s=1}^{T}\widetilde{F}_s F_s^{0\prime}\beta^{0\prime} u_t\|^2\Big) \\
&\leq (1-w)^2\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\|u_t'\beta^0\|^2(\frac{1}{T}\sum_{s=1}^{T}\|\widetilde{F}_s\|^2)(\frac{1}{T}\sum_{s=1}^{T}\|F_s^0\|^2) \\
&= O_p(1-w)^2
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} g_t &= \frac{1}{T}\sum_{t=1}^{T}\Big((\frac{1-w}{AT})^2\|\sum_{s=1}^{T}\widetilde{F}_s F_t^{0\prime}\beta^{0\prime} u_s\|^2\Big) \\
&\leq (1-w)^2\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\|u_s'\beta^0\|^2\frac{1}{T}\sum_{1}^{T}\|F_t^{0\prime}\|^2\frac{1}{T}\|\sum_{s=1}^{T}\widetilde{F}_s\|^2 \\
&= O_p(1-w)^2
\end{aligned} \tag{7}
$$

Combing the results I have $(1/T)\sum_{t=1}^{T}\|\widetilde{F}_t - H'F_t^0\|^2 = O_p(\frac{w^2}{T}) + O_p(\frac{w^2}{N}) + O_p((1-w)^2)$

LEMMA A.1: Let $\delta = min\{\frac{\sqrt{T}}{w}, \frac{\sqrt{N}}{w}, \frac{1}{1-w}\}$

$$\delta^2\big(\frac{1}{T}\sum_{t=1}^{T}\|\widetilde{F}_t - H'F_t^0\|^2\big) = O_p(1).$$

THEOREM 2: Under Assumptions $A-G$, as $\sqrt{N}/T \to 0$, I get:

$$\sqrt{N}\Big(\widetilde{F}_t - \bar{H}'F_t^0\Big) = \frac{1}{\sqrt{N}}e_t'\bar{H}^{-1}\Lambda^0(\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}$$

$$+\left(\frac{1}{T}\left\|\sum_{t=1}^{T}(\widetilde{F}_t - H'F_t^0)\right\|^2\right)^{1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(e_{it}e_{ti}-E(e_{it}e_{ti})\right)+o_p(1)$$
$$\xrightarrow{d} N\left(0, V^{-1}Q\Gamma_t Q'V^{-1}\right)$$

$V$ and $Q$ are defined in Proposition 1, and $\Gamma_t$ is defined in $F3$;

(ii) if $\liminf \sqrt{N}/T \geq \tau > 0$, then

$$T\left(\widehat{F}_t - H'F_t^0\right) = O_p(1)$$

In Theorem 2, $\widehat{F}_t - \bar{H}'F_t = O_p(\frac{1}{T}) + O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1-w}{\sqrt{T}})$, and when $\sqrt{N}/T \to 0$, the dominating terms are the last two terms, as opposed to $O_p(\frac{1}{\sqrt{N}})$ with PCA. This indicates that SPCA can still outperform PCA when there is no model misspecification, if I do not have a very large N. Theorem 2 is proven in (c).

*(b) Factor Loadings*

Now I will going to calculate $\delta_\Lambda^2$ such that $\delta_\Lambda^2\left(\frac{1}{N}\sum_{n=1}^{N}\|\widehat{\lambda}_i - \bar{H}^{-1}\lambda_i^0\|^2\right) = O_p(1)$.

$$\begin{aligned}\widehat{\lambda}_i &= T^{-1}F'F^0\lambda_i^0 + T^{-1}\widetilde{F}'e_i \\ &= \bar{H}^{-1}\lambda_i^0 + T^{-1}\left(\widetilde{F} - F^0H\right)'e_i + T^{-1}H'F^{0\prime}e_i\end{aligned}$$

where $\bar{H}^{-1} = T^{-1}F'F^0$. Therefore,

$$\widehat{\lambda}_i - \bar{H}^{-1}\lambda_i^0 = T^{-1}H'F^{0'}e_i + T^{-1}(\widetilde{F} - F^0H)'e_i.$$

Let $a_i = \frac{1}{T^2}\left\|H'F^{0'}e_i\right\|^2$, $\quad b_i = \frac{1}{T^2}\left\|(\widetilde{F} - F^0H)'e_i\right\|^2$, then

$$\frac{1}{N}\sum_{i=1}^{N}(\left\|\widehat{\lambda}_i - \bar{H}^{-1}\lambda_i^0)\right\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\left(a_i + b_i\right)$$ Calculate the first piece:

$$\frac{1}{N}\sum_{i=1}^{N}a_i = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T^2}\left\|H'F^{0'}e_i\right\|^2\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}H_t'F_t^0e_{it}\Big\|^2\frac{1}{T}$$
$$= O_p(\frac{1}{T})$$

The second piece can be further broken up.

$$\begin{aligned}
&T^{-1}\sum_{t=1}^{T}(\widetilde{F}_t - H'F_t^0)e_{it}\\
&= V_{NT}^{-1}\Bigg(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)e_{it} + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}e_{it}\\
&\quad + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\eta_{st}e_{it} + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\xi_{st}e_{it} + \frac{1-w}{AT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_su_su_te_{it}\\
&\quad + \frac{1-w}{AT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_sF_s^{0'}\beta^0u_te_{it} + \frac{1-w}{AT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_tF_t^{0'}\beta^0u_se_{it}\Bigg)\\
&= V_{NT}^{-1}(I+II+III+IV+V+VI+VII)
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}I^2\\
=&\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)e_{it}\Big)^2\\
\le&\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}H'F_s^0\gamma_N(s,t)e_{it} + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)\gamma_N(s,t)e_{it}\Big)^2\\
\le&\frac{1}{N}\sum_{i=1}^{N}\frac{w^2}{T^3}\Bigg(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\Big|\gamma_N(s,t)\Big|^2\Big(E\left\|F_s^0\right\|^2\Big)\left(Ee_{it}^2\right)\Bigg)\\
&+\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w^2}{T}\sum_{s=1}^{T}\Big\|\widetilde{F}_s - H'F_s^0\Big\|^2\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\Big|\gamma_N(s,t)\Big|^2\frac{1}{T}\sum_{t=1}^{T}e_{it}^2\frac{1}{T}\Big)\\
=&O_p(\frac{w^2}{T^3}) + O_p(\frac{w^2}{\delta^2T})
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}II^2\\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}e_{it}\Big)^2
\end{aligned}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}HF_s^0\zeta_{st}e_{it}\Big)^2 + \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)\zeta_{st}e_{it}\Big)^2$$

Because $E\big\|\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{k=1}^{N}F_s^0[e_{ks}e_{kt}-E(e_{ks}e_{kt})]\big\|^2 \leq M$, and $E(e_{it})^2 \leq M$, we have $\frac{1}{T}\sum_{t=1}^{T}\Big(\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{k=1}^{N}F_s^0[e_{ks}e_{kt}-E(e_{ks}e_{kt})]e_{it}\Big) = O_p(1)$ The first piece is $O_p(\frac{w^2}{NT})$

$$\begin{aligned}
&\Big(\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)\zeta_{st}e_{it}\Big)^2 \\
&\leq \frac{1}{T}\sum_{s=1}^{T}\big\|\widetilde{F}_s - H'F_s^0\big\|^2\frac{1}{T}\sum_{s=1}^{T}\Big(\frac{1}{T\sqrt{N}}\sum_{t=1}^{T}\big(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}[e_{ks}e_{kt}-E(e_{ks}e_{kt})]\big)e_{it}\Big)^2 \\
&= O_p(\frac{1}{\delta^2 N})
\end{aligned}$$

The second piece is $O_p(\frac{w^2}{\delta^2 N})$ Therefore $\frac{1}{N}\sum_{i=1}^{N}II^2 = O_p(\frac{w^2}{NT}) + O_p(\frac{w^2}{\delta^2 N})$

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}III^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\eta_{st}e_{it}\Big)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}H'F_s^0\eta_{st}e_{it}\Big)^2 + \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)\eta_{st}e_{it}\Big)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(H'(\frac{1}{T}\sum_{s=1}^{T}F_s^0F_s^{0'})(\frac{1}{TN}\sum_{t=1}^{T}\sum_{k=1}^{N}\lambda_k e_{kt}e_{it})\Big)^2 \\
&+ \frac{1}{N}\sum_{i=1}^{N}\Big((\frac{1}{T}\sum_{s=1}^{T}\big\|\widetilde{F}_s - H'F_s^0\big\|^2)(\frac{1}{T}\sum_{s=1}^{T}(\frac{1}{T}\sum_{t=1}^{T}\eta_{st}e_{it})^2)\Big)^2 \\
&= O_p(\frac{w^2}{N^2}) + O_p(\frac{w^2}{NT}) + O_p(\frac{w^2}{\delta^2 N})
\end{aligned}$$

$\Big(H'(\frac{1}{T}\sum_{s=1}^{T}F_s^0F_s^{0'})(\frac{1}{TN}\sum_{t=1}^{T}\sum_{k=1}^{N}\lambda_k e_{kt}e_{it})\Big)^2 = O_p(\frac{1}{N^2}) + O_p(\frac{1}{NT})$ under weak cross-sectional dependence. If we assume cross-section independence then it's $O_p(\frac{1}{NT})$. For the last term, $\frac{1}{T}\sum_{t=1}^{T}\eta_{st}e_{it} = \frac{1}{\sqrt{N}}F_s^{0'}\frac{1}{T}\sum_{t=1}^{T}\big(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\lambda_k e_{kt}\big)e_{it} = O_p(\frac{1}{\sqrt{N}})$.

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}IV^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\xi_{st}e_{it}\Big)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}H'F_s^0\xi_{st}e_{it}\Big)^2+\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s-H'F_s^0)\xi_{st}e_{it}\Big)^2\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\Big(w^2\|H'\|^2\big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^{0'}e_{it}\big)^2\big(\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{k=1}^{N}F_s^0\lambda_k^{0'}e_{ks}\big)^2\frac{1}{NT^2}\Big)\\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{w^2\|H'\|^2}{T}\sum_{s=1}^{T}\big\|\widetilde{F}_s-H'F_s^0\big\|^2\big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^{0'}e_{it}\big)^2\frac{1}{T}\sum_{s=1}^{T}\big(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\lambda_k^{0'}e_{ks}\big)^2\frac{1}{NT}\Big)\\
&=O_p(\frac{w^2}{NT^2})+O_p(\frac{w^2}{\delta^2NT})
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{N}\sum_{i=t}^{N}V^2&=\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_su_su_te_{it}\Big)^2\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}HF_s^0u_su_te_{it}\Big)^2+\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s-HF_s^0)u_su_te_{it}\Big)^2\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\Big((\frac{1-w}{T})^2\big\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}u_te_{it}\big\|^2\quad\big\|\frac{1}{\sqrt{T}}\sum_{s=1}^{T}HF_s^0u_s\big\|^2\Big)\\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{(1-w)^2}{T}\big\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}u_te_{it}\big\|^2\quad\frac{1}{T}\sum_{s=1}^{T}\big\|(\widetilde{F}_s-HF_s^0)\big\|^2\frac{1}{T}\sum_{s=1}^{T}\big\|u_s\big\|^2\Big)\\
&=O_p(\frac{(1-w)^2}{T^2})+O_p(\frac{(1-w)^2}{T\delta^2})
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{N}\sum_{i=t}^{N}VI^2&=\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_sF_s^{0'}\beta^{0'}u_te_{it}\Big)^2\\
&\leq (1-w)^2\big\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_sF_s^{0'}\big\|^2\big\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\beta^{0'}u_te_{it}\big\|^2\frac{1}{T}\\
&=O_p(\frac{(1-w)^2}{T})
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{N}\sum_{i=t}^{N}VII^2&=\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1-w}{T^2}\sum_{t}^{T}\sum_{s}^{T}\widetilde{F}_sF_t^{0'}\beta^{0'}u_se_{it}\Big)^2\\
&=\frac{1}{N}\sum_{i=1}^{N}\Big((1-w)\big\|\frac{1}{T}\sum_{s}^{T}\widetilde{F}_su_s\big\|\big\|\frac{1}{T}\sum_{t}^{T}\beta^{0'}F_t^{0'}e_{it}\big\|\Big)^2
\end{aligned}
$$

The first term can be written as

$$\|\frac{1}{T}\sum_{s}^{T}\widetilde{F}_s u_s\|^2 \leq \|\frac{1}{T}\sum_{s}^{T}(\widetilde{F}_s - H'F_s^0)u_s\|^2 + \|\frac{1}{T}\sum_{s}^{T}H'F_s^0 u_s\|^2 = O_p(\frac{1}{\delta^2}) + O_p(\frac{1}{T})$$

$$\|\frac{1}{T}\sum_{t}^{T}\beta^{0'}F_t^{0'}e_{it}\|^2 = O_p(\frac{1}{T})$$

Thus $\frac{1}{N}\sum_{i=t}^{N} VII^2 = O_p(\frac{(1-w)^2}{T\delta^2}) + O_p(\frac{(1-w)^2}{T^2})$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} b^2 &= O_p(\frac{w^2}{T\delta^2}) + O_(\frac{w^2}{\delta^2 N}) + O_p(\frac{w^2}{T^3}) + O_p(\frac{w^2}{N^2}) \\ &\quad + O_p(\frac{(1-w)^2}{T}) + O_p(\frac{(1-w)^2}{T^2}) + O_p(\frac{(1-w)^2}{T}) \\ &= O_p(\frac{w^2}{T\delta^2}) + O_(\frac{w^2}{\delta^2 N}) + O_p(\frac{w^2}{N^2}) + O_p(\frac{(1-w)^2}{T})\end{aligned}$$

$$\begin{aligned}\delta_\Lambda^2 &= O_p(\frac{1}{T}) + O_p(\frac{w^2}{T\delta^2}) + O_(\frac{w^2}{\delta^2 N}) + O_p(\frac{w^2}{N^2}) + O_p(\frac{(1-w)^2}{T}) \\ &= O_p(\frac{1}{T}) + O_p(\frac{w^2}{\delta^2 N}) + O_p(\frac{w^2}{N^2})\end{aligned}$$

We can see that for PCA, $\delta_{\Lambda,NT}^2 = O_p(\frac{1}{T}) + O_p(\frac{1}{T\delta_{NT}^2}) + O_(\frac{1}{\delta_{NT}^2 N}) + O_p(\frac{1}{N^2})$. $\delta_\Lambda^2$ would be worse than that of PCA if the dominating term is $O_p(\frac{(1-w)^2}{T})$. $\delta_\Lambda$ can be better than that of PCA if $\frac{(1-w)N}{w\sqrt{T}} \to 0$ and $w < 1$, and N is small. When w = 1, $\delta_\Lambda$ equals to that of PCA.

*(c) Theorem 2*

The $\widetilde{F}_t$ I discussed above is for t=1,...,T-h. To make predictions of $y_{T+1}, ..., y_{T+h}$, I use OLS so $\widehat{F} = X\widehat{\Lambda}\left(\widehat{\Lambda}'\widehat{\Lambda}\right)^{-1}$, for t=T-h+1, ..., T. Let's use the subscript 'new' to represent t=T-h+1, ..., T. In this section I am finding $\widehat{F}_{new} - \bar{H}'F_{new}$.

$$\widehat{F}_{new} = (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'X_{new}$$

$$
\begin{aligned}
&= (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'(\Lambda^0 F_{new} + e_{new}) \\
&= (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'\Lambda^0 F_{new} + (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}' e_{new} \\
&= (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Lambda}\bar{H}' F_{new} + (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'(\Lambda^0 - \widehat{\Lambda}\bar{H}')F_{new} + (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}' e_{new}
\end{aligned}
$$

We obtain

$$
\widehat{F}_{new} - \bar{H}' F_{new} = (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'(\Lambda^0 - \widehat{\Lambda}\bar{H}')F_{new} + (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}' e_{new}
$$

$$
\begin{aligned}
&(\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'(\Lambda^0 - \widehat{\Lambda}\bar{H}')F_{new} \\
&= (\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}\frac{1}{N}\widehat{\Lambda}'\big(\Lambda^0(\bar{H}^{-1})' - \widehat{\Lambda}\big)\bar{H}' F_{new} \\
&= (\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}\frac{1}{N}(\widehat{\Lambda} - \Lambda^0(\bar{H}^{-1})')'\big(\Lambda^0(\bar{H}^{-1})' - \widehat{\Lambda}\big)\bar{H}' F_{new} \\
&+ (\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}\frac{1}{N}\bar{H}^{-1}\Lambda^{0\prime}\big(\Lambda^0(\bar{H}^{-1})' - \widehat{\Lambda}\big)\bar{H}' F_{new} \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\big\|\lambda_i^0 - \bar{H}^{-1}\widehat{\lambda}_i\big\|^2 + \frac{1}{N}\big\|\sum_{i=1}^{N}(\lambda_i^0 - \bar{H}^{-1}\widehat{\lambda}_i)\lambda_i^{0\prime}\big\| \\
&= \frac{1}{N}\sum_{i=1}^{N}\big\|\lambda_i^0 - \bar{H}^{-1}\widehat{\lambda}_i\big\|^2 + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}H' F^{0\prime} e_i \lambda_i^{0\prime} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}(\widetilde{F} - F^0 H')e_{it}\lambda_i^{0\prime} \\
&= \frac{1}{\delta_\Lambda^2} + \quad \frac{1}{\sqrt{NT}} + \frac{1}{\delta^2 W}
\end{aligned}
$$

Because $\frac{1}{T}\sum_{t=1}^{T}(F_t - F_t H)\,e_{it} = w\left(\frac{1}{\sqrt{T}\delta} + \frac{1}{T} + \frac{1}{N} + \frac{1}{\sqrt{N}\delta} + \frac{1}{\sqrt{NT}}\right) + \frac{1-w}{A\sqrt{T}} = \frac{1}{\delta\sqrt{T}} + \frac{1}{\delta\sqrt{N}} + \frac{1}{\delta^2} + \frac{1-w}{A\sqrt{T}} = \frac{1}{\delta^2 W} + \frac{1-w}{A\sqrt{T}} = \frac{1}{\delta^2 W}$

$$
(\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}' e_{new} = \frac{1}{N}(\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}(\widehat{\Lambda} - \Lambda^0(\bar{H}^{-1})')' e_{new} + \frac{1}{N}(\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}\bar{H}^{-1}\Lambda^{0\prime} e_{new}
$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}(\widehat{\lambda}_i - \bar{H}^{-1}\lambda_i^0)e_{newi} + \frac{1}{N}\sum_{i=1}^{N}\lambda_i^0 e_{newi}$$

To calculate the first piece,

$$\frac{1}{N}\sum_{i=1}^{N}(\widehat{\lambda}_i - \bar{H}^{-1}\lambda_i^0)e_{newi} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}H'F^{0'}e_i e_{newi} + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{N}(\widetilde{F}_t - F_t^0 H)e_{it}e_{newi}$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}H'F_t^0 e_{it}e_{newi} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}H'F_t^0\big(e_{it}e_{newi} - E(e_{it}e_{newi})\big) + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}H'F_t^0 E(e_{it}e_{newi}$$

Assume

$$E\big\|\sum_{t=1}^{T}F_t^0 E(\frac{1}{N}\sum_{i=1}^{N}e_{it}e_{newi})\big\| \leq E\big\|F_t^0\big\|\sum_{t=1}^{T}\big\|E\frac{1}{N}\big(\sum_{i=1}^{N}e_{it}e_{newi}\big)\big\| \leq M^{1/4+1}$$

also assume

$$E\big\|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}F_t^0[e_{it}e_{newi} - E(e_{it}e_{newi})]\big\|^2 \leq M$$

Then

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}H'F_t^0 e_{it}e_{newi} = O_p(\frac{1}{\sqrt{NT}}) + O_p(\frac{1}{T})$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}(\widetilde{F}_t - F_t^0 H)e_{it}e_{newi} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}(\widetilde{F}_t - F_t^0 H)\big(e_{it}e_{newi} - E(e_{it}e_{newi})\big) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{N}(\widetilde{F}_t - F_t^0 H)E(e_{it}e_{newi}) \\
&= \big(\frac{1}{T}\big\|\sum_{t=1}^{T}(\widetilde{F} - H'F_t^0)\big\|^2\big)^{1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\big(e_{it}e_{newi} - E(e_{it}e_{newi})\big) \\
&+ \big(\frac{1}{T}\big\|\sum_{t=1}^{T}(\widetilde{F} - H'F_t^0)\big\|^2\big)^{1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}E(e_{it}e_{newi}) \\
&= O_p(\frac{1}{\delta\sqrt{N}}) + O_p(\frac{1}{\delta\sqrt{T}})
\end{aligned}$$

Assume

$$\frac{1}{T}\sum_{t=1}^{T}\Big\|\frac{1}{N}\sum_{i=1}^{N}\big(e_{it}e_{newi}-E(e_{it}e_{newi})\big)\Big\|^2=O_p(\frac{1}{N})$$

$$\frac{1}{T}\sum_{t=1}^{T}\Big\|\frac{1}{N}\sum_{i=1}^{N}\big(E(e_{it}e_{newi})\big)\Big\|^2=O_p(\frac{1}{T})$$

Therefore, $\frac{1}{N}\sum_{i=1}^{N}(\widehat{\lambda}_i-\bar{H}^{-1}\lambda_i^0)e_{newi}=O_p(\frac{1}{\sqrt{NT}})+O_p(\frac{1}{T})+O_p(\frac{1}{\delta\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{T}})$.

Moreover,

$$\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0 e_{newi}=O_p(\frac{1}{\sqrt{(N)}})$$

Hence,

$$(\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'e_{new}=O_p(\frac{1}{T})+O_p(\frac{1}{\sqrt{NT}})+O_p(\frac{1}{\delta\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{T}}+O_p(\frac{1}{\sqrt{N}}))$$

And

$$\begin{aligned}&\widehat{F}_{new}-\bar{H}'F_{new}\\&=O_p(\tfrac{1}{T})+O_p(\tfrac{w^2}{\delta^2N})+O_p(\tfrac{w^2}{N^2})+w\left(\tfrac{1}{\sqrt{T}\delta}+\tfrac{1}{T}+\tfrac{1}{N}+\tfrac{1}{\sqrt{N}\delta}+\tfrac{1}{\sqrt{NT}}\right)+\tfrac{1-w}{A\sqrt{T}}\\&+O_p(\tfrac{1}{\sqrt{NT}})+O_p(\tfrac{1}{\sqrt{N}})+O_p(\tfrac{1}{\delta\sqrt{N}})+O_p(\tfrac{1}{\delta\sqrt{T}})\end{aligned}$$

$$\begin{aligned}&\widehat{F}_{new}-\bar{H}'F_{new}\\&=O_p(\frac{1}{T})+O_p(\frac{w^2}{\delta^2N})+O_p(\frac{w^2}{N^2})+O_p(\frac{1}{w\sqrt{N}\delta^2})+O_p(\frac{1-w}{A\sqrt{T}})\\&\quad+O_p(\frac{1}{\sqrt{NT}})+O_p(\frac{1}{\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{T}})\end{aligned}$$

$$\widehat{F}_{new}-\bar{H}'F_{new}=\text{terms with w + terms without w}$$

terms with w are

in $1/\left(N\sum_{i=1}^{N}\left\|\lambda_i^0-\bar{H}^{-1}\widehat{\lambda}_i\right\|^2:\frac{1}{N}\sum_{i=1}^{N}\left(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\widetilde{F}_s-H'F_s^0)\zeta_{st}e_{it}\right)^2=O_p(\frac{w^2}{\delta^2 N})\frac{1}{N}\sum_{i=1}^{N}\left(H'(\frac{1}{T}\sum_{s=1}^{T}F_s^0F_s^{0\prime})(\frac{1}{TN}\sum_{t=1}^{T}\sum_{k=1}^{N}\lambda_k\right.\right.$

where

$\zeta_{st}=\frac{1}{N}\sum_{k=1}^{N}[e_{ks}e_{kt}-E(e_{ks}e_{kt})]$

$\frac{1}{T}\sum_{t=1}^{T}\eta_{st}e_{it}=\frac{1}{\sqrt{N}}F_s^{0\prime}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\lambda_k e_{kt}\right)e_{it}=O_p(\frac{1}{\sqrt{N}})$

Recall that $\delta=min\{\frac{\sqrt{N}}{w},\frac{\sqrt{T}}{w}\frac{1}{1-w}\}$. The dominating terms from above are $O_p(\frac{1}{\delta\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{T}})=O_p(\frac{1}{w\delta^2})$

terms without w

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}H'F^{0\prime}e_i\lambda_i^{0\prime}=\cancel{O_p(\frac{1}{\sqrt{NT}})}$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}H'F^{0\prime}E(e_ie_{newi})=O_p(\frac{1}{T})$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{N}H'F_t^0(e_{it}e_{newi}-E(e_{it}e_{newi}))=\cancel{O_p(\frac{1}{\sqrt{NT}})}$$

$$\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0e_{newi}=O_p(\frac{1}{\sqrt{N}})$$

in $\frac{1}{N}\sum_{i=1}^{N}\left\|\lambda_i^0-H^{-1}\widehat{\lambda}_i\right\|^2:\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T^2}\left\|H'F^{0\prime}e_i\right\|^2\right)=O_p(\frac{1}{T})$

The dominating term from above is $O_p(\frac{1}{T})+O_p(\frac{1}{\sqrt{N}})$. Therefore

$$\begin{aligned}&\widehat{F}_{new}-\bar{H}'F_{new}\\&=O_p(\frac{1}{T})+O_p(\frac{1}{\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{N}})+O_p(\frac{1}{\delta\sqrt{T}})\\&=O_p(\frac{1}{T})+O_p(\frac{1}{\sqrt{N}})+O_p(\frac{w}{N})+O_p(\frac{w}{T})+O_p(\frac{(1-w)}{\sqrt{N}})+O_p(\frac{(1-w)}{\sqrt{T}})\end{aligned}$$

Recall that for PCA, $\widetilde{F}_t-H'F_t^0=O_p\left(\frac{1}{\sqrt{T}\delta_{NT}}\right)+O_p\left(\frac{1}{\sqrt{N}\delta_{NT}}\right)+O_p\left(\frac{1}{\sqrt{N}}\right)+O_p\left(\frac{1}{\sqrt{N}\delta_{NT}}\right)$.

**Case 1** ($O_p(\frac{1}{\sqrt{N}})$dominates): When $\sqrt{N}/T \to 0$ and $\sqrt{N}(1-w)/\sqrt{T} \to 0$, the dominating term is $O_p(\frac{1}{\sqrt{N}})$. In this case, as $\sqrt{N} \to \infty$ , $w \to 1$. And when $w = 1$, we only need $\sqrt{N}/T \to 0$.

$$\sqrt{N}\left(\widetilde{F}_{new} - \bar{H}'F^0_{new}\right) = \frac{1}{\sqrt{N}}(\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N})^{-1}\bar{H}^{-1}\Lambda^{0\prime}e_{new} + o_p(1)$$
$$= \frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_i\hat{\lambda}_i{}')^{-1}\frac{1}{T}\sum_{s=1}^{T}(\widetilde{F}_s F^{0\prime}_s)\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda^0_i e_{it} + o_p(1)$$

$$(\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_i\hat{\lambda}_i{}') \to \Sigma_\Lambda$$

$$\left(F^{0\prime}\tilde{F}/T\right) = \frac{1}{T}\sum_{t=1}^{T}F^0_t\widetilde{F}_t = \frac{\widetilde{F}'F^0}{T} \to Q$$

I assumed $(1/\sqrt{N})\sum_{i=1}^{N}\lambda^0_i e_{it} \xrightarrow{d} N(0,\Gamma_t)$. Therefore $\sqrt{N}\left(\widetilde{F}_t - H'F^0_t\right) \xrightarrow{d} N\left(0,\Sigma_\Lambda^{-1}Q\Gamma_t Q'\Sigma_\Lambda^{-1}\right)$

Where Q comes from

PROPOSITION 1 : $\text{plim}_{T,N\to\infty}\frac{\widetilde{F}'F^0}{T} = Q$

The matrix $Q$ is invertible and is given by $Q = V^{1/2}T'\Sigma_A^{-1/2}$, where $V = \text{diag}\,(v_1, v_2, \ldots, v_r)$,

$v_1 > v_2 > \cdots > v_r > 0$ are the eigenvalues of $\Sigma_A^{1/2}\Sigma_F\Sigma_A^{1/2}$, and $T$ is the corresponding eigenvector matrix such that $T'T = I_r$.

**Case 2** $O_p(\frac{1}{\sqrt{N}})$ and $O_p(\frac{1}{T})$ dominates: If $\liminf\sqrt{N}/T \geq \tau > 0$, and $w \to 1$, the dominating terms are $O_p(\frac{1}{T}) + O_p(\frac{1}{\sqrt{N}})$. We have the same result as in PCA:

$$T\left(\widetilde{F}_{new} - H'F^0_{new}\right) = O_p(1) + O_p(T/\sqrt{N})) = O_p(1)$$

in view of $\limsup(T/\sqrt{N}) \leq 1/\tau < \infty$.

**Case 3** $O_p(\frac{1}{\sqrt{N}}$ and $O_p(\frac{1-w}{\sqrt{T}}))$ dominates: If $\liminf N/T \geq \tau > 0$, and $1-w \geq M > 0$, the dominating terms are $O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1-w}{\sqrt{T}})$.

$$\sqrt{T}\left(\widetilde{F}_{new} - H'F^0_{new}\right) = O_p(1) + O_p(\sqrt{T}/\sqrt{N})) = O_p(1)$$

in view of $\limsup(\sqrt{T}/\sqrt{N}) \leq 1/\tau < \infty$.

Case 4 ($O_p(\frac{1-w}{\sqrt{T}})$ dominates): If $\frac{\sqrt{T}}{\sqrt{N}(1-w)} \to 0$, then the dominating term is $O_p(\frac{1-w}{\sqrt{T}})$

$$\sqrt{T}\left(\widetilde{F}_{new} - H'F^0_{new}\right) = o_p(1) + O_p(\sqrt{T}/\sqrt{N})) = O_p(1)$$

Since $\frac{1}{\sqrt{T}}$ dominates $\frac{1}{T}$, if $\frac{1}{\sqrt{T}}$ also dominates $\frac{1}{N}$, that is when T is small and $w \neq 1$, then sPCA will be worse than PCA. That is for $F_{new}$ of sPCA to be better than PCA, we need $\frac{(1-w)N}{w\sqrt{T}} \to 0$ and N is small.

THEOREM 3 - Estimation: Suppose Assumptions A-E hold. If $\sqrt{T}/N \to 0$, then

$$\sqrt{T}(\widehat{\beta} - \beta) \xrightarrow{d} N\left(0, \Sigma_\beta\right)$$

where $\Sigma_\beta = \Phi_0'^{-1}\Sigma^{-1}_{FF,u}\Sigma^{-1}_{FF}\Phi_0^{-1}$ with $\Phi_0 = \text{diag}\left(V^{-1}Q\Sigma_\Lambda, I\right)$ being block diagonal, $V = \text{plim}\ \tilde{V}, Q = \text{plim}\ \tilde{F}'F/T$, and $\Sigma_\Lambda$ defined in Assumption B. A consistent estimator for $\Sigma_\beta$, denoted by $Avar(\widehat{\beta}\ )$, i s

$$\widehat{\text{Avar}(\hat{\beta})} = T^{-1}\sum_{t=1}^{T}\widehat{u}^2_{t+h} \times \widetilde{F}_t\widetilde{F}_t'$$

This estimator is robust to heteroskedasticity. If I assume homoskedasticity then $E\left(u^2_{t+h}|F_t\right) = \sigma^2_u, \forall t$. Denote $\widehat{\sigma}^2_u = \frac{1}{T}\sum_{t=1}^{T-h}\widehat{u}^2_{t+h}$,

$\text{Avar}(\widehat{\beta})$ can be consistently estimated by

$$\widehat{\text{Avar}(\widehat{\beta})} = \widehat{\sigma}^2_u\left[\frac{1}{T}\sum_{t=1}^{T-h}\widetilde{F}_t\widetilde{F}_t'\right]^{-1} = \widehat{\sigma}^2_u$$

Theorem 3 is proven in (d).

*(d) Theorem 3*

$\widehat{\beta} = \left(\widetilde{F}'\widetilde{F}\right)^{-1}\widetilde{F}'y = \widetilde{F}'y/T$. Replace $y$ with $y = F^0\beta^0 + u$, I have $\widehat{\beta} = T^{-1}\widetilde{F}'F^0\beta^0 + T^{-1}\widetilde{F}'u$ . Let $\bar{H}^{-1} = \frac{1}{T}\sum_{t=1}^{T} F_t^0\widetilde{F}_t$. Then

$$\beta - \bar{H}^{-1}\beta^0 = T^{-1}H'F^{0\prime}u + T^{-1}\left(\widetilde{F} - F^0H\right)'u$$

$$\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta) = H'F^{0'}u/\sqrt{T} + \left(\widetilde{F} - F^0H\right)'u/\sqrt{T}$$

The second piece on the right side will be analyzed. The calculation of the second piece is similar to $T^{-1}(\widetilde{F} - F^0H)'e_i$.

$$\begin{aligned}
&T^{-1}\left(\widetilde{F} - F^0H\right)'u \\
&= T^{-1}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)u_t \\
&= V_{NT}^{-1}\left(\frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)u_t + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}u_t\right. \\
&\quad \left. + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\eta_{st}u_t + \frac{w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_s\xi_{st}u_t\right) + V_{NT}^{-1}\left(\frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_su_su_tu_t\right. \\
&\quad \left. + \frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_sF_s^{0\prime}\beta^0u_tu_t + \frac{1-w}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\widetilde{F}_sF_t^{0\prime}\beta^0u_su_t\right) \\
&= V_{NT}^{-1}(I + II + III + IV + V + VI + VII)
\end{aligned}$$

$$\begin{aligned}
&T^{-1}\left(\widetilde{F} - F^0H\right)'u \\
&= O_p(\frac{w}{\sqrt{T}\delta}) + O_(\frac{w}{\sqrt{N}\delta}) + O_p(\frac{w}{T}) + O_p(\frac{w}{N}) + O_p(\frac{(1-w)}{\sqrt{T}}) + O_p(\frac{(1-w)}{T}) + O_p(\frac{(1-w)}{\sqrt{T}}) \\
&= O_p(\frac{w}{\sqrt{T}\delta}) + O_(\frac{w}{\sqrt{N}\delta}) + O_p(\frac{w}{T}) + O_p(\frac{w}{N}) + O_p(\frac{(1-w)}{\sqrt{T}})
\end{aligned}$$

**Case 1:** Dominating term is $O_p(w/\min(N,T))$. Then $\left(\widetilde{F} - F^0H\right)'u/\sqrt{T} = O_p(w\sqrt{T}/\min(N,T)) = o_p(1)$, when $\sqrt{T}/N \to 0$ and $w \to 1$.

Thus,

$$\text{(A.1)} \quad \sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta) = T^{-1/2} H' F^{0'} u + o_p(1)$$

Since $F^{0\prime}u/\sqrt{T} \xrightarrow{d} N(0, \Sigma_{FF,u})$ by Assumption, the above is asymptotically normal. The asymptotic variance matrix is the probability limit of

$$\text{(A.2)} \quad H' \left( \frac{1}{T} \sum_{t=1}^{T} u_{t+h}^2 F_t^0 F_t^{0\prime} \right) H,$$

Because $HF_t = \widetilde{F}_t + o_p(1)$, I have $H\left(\frac{1}{T}\sum_{t=1}^{T} u_{t+h}^2 F_t^0 F_t^{0\prime}\right) H' = \left(\frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{t+h}^2 \widetilde{F}_t \widetilde{F}_t'\right) + o_p(1)$. Therefore,

$$\widehat{Avar(\widehat{\beta})} = T^{-1} \sum_{t=1}^{T} \widehat{u}_{t+h}^2 \times \widetilde{F}_t \widetilde{F}_t'$$

is a consistent estimator for $\Sigma_\beta$. If I further assume homoskedasticity such that $\forall t, E\left(u_{t+h}^2 | z_t\right) = \sigma_u^2$ and letting $\widehat{\sigma}_u^2 = \frac{1}{T}\sum_{t=1}^{T-h} \widehat{u}_{t+h}^2$, a consistent estimate of Avar$(\widehat{\beta})$ is

$$\widehat{\text{Avar}(\widehat{\beta})} = \widehat{\sigma}_u^2 \left[ \frac{1}{T} \sum_{t=1}^{T-h} \widetilde{F}_t \widetilde{F}_t' \right]^{-1} = \widehat{\sigma}_u^2$$

**Case 2** The dominating term is $\frac{(1-w)}{\sqrt{T}}$ when $\sqrt{T}/N \to 0$ and w < 1. In this case,

$$\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta) = O_p(1)$$

**Case 3** The dominating terms are $\frac{(1-w)}{\sqrt{T}}$ and $\frac{w}{N}$ when $liminf \sqrt{T}/N \geq \tau > 0$ and w < 1. In this case,

$$\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta) = O_p(1)$$

Since $\frac{1}{\sqrt{T}}$ dominates $\frac{1}{T}$,if $\frac{1}{\sqrt{T}}$ also dominates $\frac{1}{N}$, sPCA cannot improve PCA. That is for $\beta$ of sPCA to be better than PCA, we need $\frac{(1-w)N}{w\sqrt{T}} \to 0$ and N is small.

THEOREM 4: Let $\hat{y}_{T+h|T} = \widehat{\beta}'\widehat{F}_T$. Under the assumptions of Theorem 3 and $\sqrt{N}/T \to 0$,

$$\frac{\left(\widehat{y}_{T+h|T} - y_{T+h|T}\right)}{\sqrt{\text{var}\left(\widehat{y}_{T+h|T}\right)}} \xrightarrow{d} N(0,1)$$
$$\text{where var}\left(\widehat{y}_{T+h|T}\right) = \frac{1}{T}\widehat{F}_T' \text{Avar}(\widehat{\beta})\widehat{F}_T + \frac{1}{N}\widehat{\beta}' \text{Avar}\left(\widetilde{F}_T\right)\widehat{\beta}$$

Theorem 4 is proven in IV.. Under conventional assumptions, SPCA outperforms PCA if N is not very large, and T is large. Moreover the real world problems rarely fit all those conventions. SPCA is designed to approximate the real patterns. Now I will use simulations to cover both situations.

## IV. THEOREM 4

$$\begin{aligned}
&\widehat{y}_{T+h|T} - y_{T+h|T} \\
&= \widehat{\beta}'\widehat{F}_T - \beta' F_T \\
&= \left(\widehat{\beta} - \bar{H}^{-1}\beta\right)'\widehat{F}_T + \beta'\bar{H}^{-1}\left(\widehat{F}_T - \bar{H}F_T\right) \\
&= T^{-1/2}[\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta)]'\widehat{F}_T + N^{-1/2}\beta'\bar{H}^{-1}\left[\sqrt{N}\left(\widehat{F}_T - \bar{H}F_T\right)\right]
\end{aligned}$$

If $\frac{\sqrt{T}}{N} \to 0$ and $\frac{\sqrt{N}}{T} \to 0$ and $w \to 1$, then $\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta)$ and $\sqrt{N}\left(\widehat{F}_T - \bar{H}F_T\right)$ are asymptotically normal. The variance of $T^{-1/2}[\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta)]'\widehat{F}_T$ can be estimated by $\frac{1}{T}F_T'\text{Avar}(\widehat{\beta})'F_T$. Since $T^{-1/2}\left(\widehat{F}_T - \bar{H}F_T\right) = o_p(1)$, the variance could be estimated using $\frac{1}{T}\widehat{F}_T$ Avar $(\widehat{\beta})\widehat{F}_T$. Likewise, the variance of the second piece can be estimated with $\frac{1}{N}\beta'\bar{H}^{-1}\text{Avar}\left(\widehat{F}_T\right)\bar{H}^{-1\prime}\beta$, which is then estimated by $\frac{1}{N}\widehat{\beta}'\text{Avar}\left(\widehat{F}_T\right)\widehat{\beta}$ . As the limit of $\sqrt{T}(\widehat{\beta} - \bar{H}^{-1}\beta)$ is determined by $(u_1, \ldots, u_T)$ and that of $\sqrt{N}\left(\widehat{F}_T - \bar{H}F_T\right)$ is determined by $e_{iT}$ for $i = 1, 2, \ldots, N.$, they are also asymptotically independent. Therefore an estimate for the forecasting error variance is $\text{var}\left(\widehat{y}_{T+h|T}\right) = \frac{1}{T}\widehat{F}_T'$ Avar $(\widehat{\beta})\widehat{F}_T + \frac{1}{N}\widehat{\beta}'\text{Avar}\left(\widehat{F}_T\right)\widehat{\beta}$, and $\left(\widehat{y}_{T+h|T} - y_{T+h|T}\right) / \text{var}\left(\widehat{y}_{T+h|T}\right)^{1/2} \xrightarrow{d} N(0,1)$.

## V. SIMULATIONS

This part attempts to find out the circumstances where SPCA outperforms the conventional PCA. In the first set of the simulations, I want to test how SPCA performs while other parameters vary and it helps us to understand the theoretical issues of w. k is

fixed to its true value here. In the second set of simulations, I want to use more realistic experiments where models are misspecified and the k is estimated. In practice, both fixed k and window vary k are popular choices. $\hat{w}$ is estimated by leaving the last 10% historical data out as a validation set.

*(a) Simulations I*

**Data Generating Process (DGP)**.

The DGP is designed as the following :

$$F_t = F_{t-1}\Phi + v_t$$

$$x_t = \Lambda F_t + e_t$$

$$y_t = F_t\beta + u_t, t = 1,2,...,T.$$

The covariance matrix of $F_t$ is a identity matrix. $\Phi$ is a $r \times r$ diagonal matrix with the diagonal elements equaling to 0.6. $v_t \sim \sqrt{0.64}N(0,1)$, and is uncorrelated with factors. r=2. $\beta = (-1,2)'$. $u_t$ is a white noise process from N(0,1).

$\Lambda \sim N(0,1)$.When errors are not cross-sectionally correlated, $e_t \sim N(0,1)$. $e_t$ is drawn from a normal distribution with mean 0s and a covariance matrix $C$. If I allow for cross-sectional correlation, C=cc', where c is a $N \times 1$ vector drawing from U(0,1). If I assume there is no cross-sectional correlation among errors, then C is an identity matrix.

When I vary the variance of $\Lambda$, then $\Lambda \sim \sqrt{(var(\Lambda))}N(0,1)$ and other parameters are still generated as above. When I change the variance of errors, then the covariance matrix for $e_t$ is $var(e)C$.

**Forecast evaluation**.

Let T be the time length of the data. The first T/2 of it is the training period R, and the rest part is the prediction period P. Using moving window, the data used in $i^{th}$ window is $(X_i', ..., X_{R+i-1}')'$.

The MSFE corresponds to the out-of-sample testing set

$$MSFE = \frac{1}{P}\sum_{i}^{P}\left(y_{s,R+i-1+h} - \hat{y}_{s,R+i-1+h}(w_i)\right)^2$$

w can be the infeasible optimal weight $w*$ or the feasible weight $\hat{w}$. Ten weights 0.1, 0.2, ..., 1 are used to get 10 estimates of $\hat{y}_{R+i-1+h}(w)$ as well as errors. The infeasible optimal weight minimized the squared forecast errors $(y_{R+i-1+h} - \hat{y}_{R+i-1+h}(w))^2$ at each prediction, as if the future data is known. Feasible weight on the other hand is based on historical data. As mentioned before, in each window, I leave out the last 10% observations as a validation set. Then I use the first 90% data to predict the validation set using different weight. $\hat{w}$ is the one that minimize MSFE corresponding to the in-sample validation

After repeating this 1000 times, I can calculate the mean of the MSFE across simulations, which are reported in the tables. The relative MSFE is calculated as

$$rMSFE(method) = \frac{MSFE(method)}{MSFE(PCA)}$$

**N and T**

Table 1 illustrates how the ratios of MSFE change as N and T vary, while holding other settings fixed. SPCA generate much smaller MSFE than PCA when N is small and T is large and allowing errors to be cross-sectionally correlated. In this table, the ratio with w* is as low as 0.50 when N=10, T=400. For a fixed T, say, T=400, when N increases from 10 to 200, the feasible MSFE ratio of sup-PCA1 increases from 0.50 to 0.69. Similarly, if N is fixed to 200, and T increases from 100 to 400, the relative MSFE decreases from 0.78 to 0.69.

When I use $\hat{w}$, SPCA generate ratios between 0.84 to 0.89. When using a fixed weight for all predictions, smaller weights have better outcomes. With bigger weights, SPCA gets closer to PCA.

TABLE 1 RELATIVE MSFE AS N AND T VARY

| N | T | rMSFE | | | |
|---|---|---|---|---|---|
| | | w* | $\hat{w}$ | w=0.5 | w=1 |
| 10 | 100 | 0.60 | 0.84 | 0.89 | 1.03 |
| 50 | 100 | 0.64 | 0.89 | 1.01 | 1.07 |
| 100 | 100 | 0.76 | 0.95 | 1.05 | 1.10 |
| 200 | 100 | 0.88 | 1.02 | 1.09 | 1.12 |
| 10 | 200 | 0.61 | 0.82 | 0.88 | 1.01 |
| 50 | 200 | 0.63 | 0.85 | 0.96 | 1.04 |
| 100 | 200 | 0.68 | 0.89 | 1.01 | 1.05 |
| 200 | 200 | 0.81 | 0.95 | 1.03 | 1.06 |
| 10 | 400 | 0.58 | 0.80 | 0.88 | 1.01 |
| 50 | 400 | 0.60 | 0.85 | 0.93 | 1.02 |
| 100 | 400 | 0.63 | 0.86 | 0.98 | 1.03 |
| 200 | 400 | 0.72 | 0.91 | 1.00 | 1.05 |

TABLE 2RELATIVE MSFE AS THE VARIANCE OF $\lambda$ VARIES

| var($\lambda$) | rMSFE | | | | |
|---|---|---|---|---|---|
| | w* | $\hat{w}$ | w=0.1 | w=0.5 | w=1 |
| 0.1 | 0.45 | 0.79 | 0.54 | 0.93 | 1.02 |
| 0.3 | 0.50 | 0.81 | 0.59 | 0.95 | 1.03 |
| 0.5 | 0.51 | 0.80 | 0.59 | 0.96 | 1.03 |
| 1 | 0.61 | 0.85 | 0.69 | 0.97 | 1.03 |
| 3 | 0.93 | 0.99 | 0.97 | 1.00 | 1.00 |

If errors are not cross-sectionally correlated, For SPCA, the best possible case has a relative MSFE of 0.92. However, when errors are independent of each other, it is hard to estimate the weight that improves PCA.

This leads to the conclusion that as N/T is small, SPCA improves PCA in terms of forecasting.

**The variance of $\Lambda$**

Now move to table 2, I am analyzing the effect of the variance of lambda on MSFE. As the variance of $\Lambda$ gradually rises from 0.1 to 3, the MSFE of SPCA tend to get closer to that of PCA. If the errors are cross-sectionally correlated, when lambda have small variances as 0.1, SPCA with w* have the relative MSFE ratios of 0.48. When the variance is 3, SPCA with w* have the relative MSFE of 0.96.

If the errors are not cross-sectionally correlated, SPCA and the PCA have close or the same MSFE.

**The variance of errors**

Table 3 and Table 4 show the MSFE ratios when the variance of errors changes. The column $var(e)$ shows the variance of the errors from $e_t = X - F\Lambda'$, and $var(u)$ corresponds to the variance of the errors from $u_t = y - F\beta$.

TABLE 3 RELATIVE MSFE AS THE VARIANCE OF ERRORS VARY PART I

| var(e) | var(u) | rMSFE | | | |
|---|---|---|---|---|---|
| | | w* | $\hat{w}$ | w=0.5 | w=1 |
| Allow cross-section correlation in errors | | | | | |
| 0.1 | 0.1 | 0.99 | 0.99 | 1.00 | 1.00 |
| 0.1 | 0.5 | 0.99 | 1.00 | 1.00 | 1.00 |
| 0.1 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.1 | 3 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.5 | 0.1 | 0.72 | 0.88 | 0.96 | 1.02 |
| 0.5 | 0.5 | 0.85 | 0.95 | 0.98 | 1.01 |
| 0.5 | 1 | 0.90 | 0.96 | 0.99 | 1.01 |
| 0.5 | 3 | 0.95 | 0.99 | 1.00 | 1.00 |
| 1 | 0.1 | 0.43 | 0.74 | 0.93 | 1.05 |
| 1 | 0.5 | 0.54 | 0.81 | 0.95 | 1.04 |
| 1 | 1 | 0.63 | 0.86 | 0.97 | 1.03 |
| 1 | 3 | 0.79 | 0.94 | 0.99 | 1.02 |
| 3 | 0.1 | 0.38 | 0.71 | 0.91 | 1.04 |
| 3 | 0.5 | 0.50 | 0.79 | 0.94 | 1.03 |
| 3 | 1 | 0.57 | 0.83 | 0.96 | 1.02 |
| 3 | 3 | 0.74 | 0.91 | 0.99 | 1.01 |

TABLE 4 RELATIVE MSFE AS THE VARIANCE OF ERRORS VARY PART II

| e | u | rMSFE | | | |
|---|---|---|---|---|---|
| | | w* | $\hat{w}$ | w=0.5 | w=1 |
| Errors are not crossly correlated | | | | | |
| 0.1 | 0.1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.1 | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.1 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.1 | 3 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.5 | 0.1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.5 | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.5 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.5 | 3 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1 | 0.1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1 | 0.5 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1 | 3 | 1.00 | 1.00 | 1.00 | 1.00 |
| 3 | 0.1 | 1.00 | 1.00 | 1.00 | 1.01 |
| 3 | 0.5 | 0.99 | 1.00 | 1.00 | 1.00 |
| 3 | 1 | 0.99 | 1.00 | 1.00 | 1.00 |
| 3 | 3 | 0.99 | 1.00 | 1.00 | 1.00 |

When errors are cross-sectionally correlated, C is not a diagonal matrix. With w*, the best estimation happens when $var(u)$ is small (0.1) and $var(e)$ is large (3) with the relative MSFE to be 0.36. If $var(e)$ is very small already, then supervised PCA does not really improve the usual PCA with w*. That is the case when X can recover very precisely the latent factors of y, and accordingly, a supervised method will not improve much. If $var(e)$ get bigger, then there is room for supervised PCA to improve with small $var(u)$.

**Summary of the main findings in simulations I**

Here is a summary of what I have found. In this set of simulations, I fix k and vary other parameters:

(1) *When N is small and T is large*, SPCA has better prediction performance than the usual PCA. X contains comprehensive information, and if N is very large, this will already cover enough about y. However if N is not so big, both X and the history of y are important in prediction y.

(2) *When cross-sectional correlations in $e_t$ are present* it is hard for PCA to detect the errors as noise. SPCA can improve the forecasting performance, especially if the $e_t$ has large variances, while $u_t$ has small variances.

(3) *If the variance of factor loadings $\Lambda$ is small*, SPCA is better than the usually PCA. It's because the factors do not have a strong impact on the variance of X.

*(b) Misspecified Models*

Recall $F_t = F_{t-1}\Phi + u_t, \quad x_t = \Lambda F_t + e_t, \quad y_t = F_t\beta + u_t, t = 1,2,\ldots,T$. Now I will consider misspecified models. And k is estimated from modified BIC.

**DGP All Factors Structure.** Under this structure, r=N. $\Lambda = I_r$ and $e_t = 0$ Therefore $x_t = F_t$ and $y_t = x_t\beta + u_t$. The $r \times r$ matrix $\Phi$ is a diagonal matrix with value $\phi = 0.6$ on the diagonal. The errors $u_t$ are from iid N(0,0.64). Then $\text{var}(F_t) = I_r$. $\beta$ is a $r \times 1$ vector and can be the following: $\beta = (1,0,\ldots,0)'$, only the first factor is related; $\beta = (0,0,\ldots,1)'$, only the first factor is related; $\beta \sim iidN(0,1)$, all of the factors are related.

**DGP Many Factors Structure.** Here I consider r=N/2. $F_t$ follows the same DGP as before.

$$\Lambda \sim \begin{bmatrix} 1, & 1,\ldots, & 1 \\ 1/2, & 1/2,\ldots, & 1/2 \\ \ldots & & \ldots \\ 1/N & 1/N & 1/N \end{bmatrix} \odot \xi$$

where $\xi$ is an $(N \times N)$ matrix consisting random numbers from iid standard normal distributions, and $\odot$ denotes the element-wise product. This is when many factors are

TABLE 5DGP-ALL FACTORS

| DGP Parameters | | | | | MSFE | rMSFE | | Mean Weight | | Mean k | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | N | T | $\beta$ | var(u) | PCA | SPCA | SPCA($\hat{w}$) | w | $\hat{w}$ | PCA | SPCA | SPCA($\hat{w}$) |
| 10 | 10 | 400 | first | 0.1 | 0.35 | 0.3 | 0.32 | 0.11 | 0.23 | 6.63 | 6.18 | 6.28 |
| 10 | 10 | 400 | first | 1 | 1.31 | 0.81 | 0.84 | 0.16 | 0.4 | 5.94 | 5.9 | 5.95 |
| 10 | 10 | 400 | all | 0.1 | 0.6 | 0.18 | 0.19 | 0.1 | 0.21 | 6.67 | 6.37 | 6.47 |
| 10 | 10 | 400 | all | 1 | 1.56 | 0.68 | 0.72 | 0.15 | 0.37 | 6.25 | 5.48 | 5.85 |
| 50 | 50 | 400 | first | 0.1 | 0.33 | 0.42 | 0.45 | 0.11 | 0.23 | 32.51 | 32.57 | 32.69 |
| 50 | 50 | 400 | first | 1 | 1.75 | 0.76 | 0.82 | 0.51 | 0.57 | 9.21 | 31.22 | 25.69 |
| 50 | 50 | 400 | all | 0.1 | 0.51 | 0.27 | 0.29 | 0.1 | 0.18 | 33.26 | 33.15 | 33.24 |
| 50 | 50 | 400 | all | 1 | 1.99 | 0.68 | 0.74 | 0.35 | 0.46 | 20.39 | 31.76 | 30.02 |
| 100 | 100 | 400 | first | 0.1 | 0.44 | 0.48 | 0.52 | 0.29 | 0.36 | 53.4 | 64.26 | 63.46 |
| 100 | 100 | 400 | first | 1 | 1.93 | 0.87 | 0.92 | 0.68 | 0.68 | 4.6 | 31.63 | 33.9 |
| 100 | 100 | 400 | all | 0.1 | 0.57 | 0.38 | 0.41 | 0.18 | 0.25 | 62.47 | 65.54 | 65.72 |
| 100 | 100 | 400 | all | 1 | 2.62 | 0.68 | 0.75 | 0.59 | 0.61 | 10.97 | 52.56 | 43.51 |

$\beta$ and $u$ come from $y_t = F_t\beta + u$. 'first' means $\beta = (1,0,...,0)$ and 'all' means $\beta \sim N(0,1)$. As $\Lambda \sim iid(N,1)$, $\beta = (0,0,..,1)$ and $\beta = (1,0,...,0)$ gives almost identical results, so we don't report the case 'last' here.

TABLE 6DGP-MANY FACTORS

| DGP Parameters | | | | | MSFE | rMSFE | | Mean Weight | | Mean k | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | N | T | $\beta$ | var(u) | PCA | SPCA(w*) | SPCA($\hat{w}$) | w | $\hat{w}$ | PCA | SPCA(w*) | SPCA($\hat{w}$) |
| 5 | 10 | 400 | first | 0.1 | 0.22 | 0.76 | 0.84 | 0.17 | 0.49 | 4.82 | 4.87 | 4.85 |
| 5 | 10 | 400 | first | 1 | 1.16 | 0.95 | 0.98 | 0.44 | 0.55 | 4.5 | 4.87 | 4.79 |
| 5 | 10 | 400 | all | 0.1 | 0.58 | 0.59 | 0.71 | 0.1 | 0.46 | 4.89 | 4.86 | 4.88 |
| 5 | 10 | 400 | all | 1 | 1.5 | 0.85 | 0.9 | 0.16 | 0.51 | 4.74 | 4.85 | 4.83 |
| 5 | 10 | 400 | last | 0.1 | 0.21 | 0.7 | 0.8 | 0.15 | 0.48 | 4.77 | 4.74 | 4.78 |
| 5 | 10 | 400 | last | 1 | 1.15 | 0.94 | 0.97 | 0.42 | 0.56 | 4.33 | 4.82 | 4.78 |
| 25 | 50 | 400 | first | 0.1 | 0.13 | 0.97 | 1.03 | 0.73 | 0.58 | 24.66 | 24.74 | 24.67 |
| 25 | 50 | 400 | first | 1 | 1.29 | 0.91 | 1 | 0.8 | 0.67 | 18.46 | 22.99 | 22.11 |
| 25 | 50 | 400 | all | 1 | 1.58 | 0.94 | 0.99 | 0.54 | 0.55 | 24.78 | 24.67 | 24.68 |
| 25 | 50 | 400 | all | 0.1 | 0.54 | 0.72 | 0.82 | 0.22 | 0.48 | 24.9 | 24.78 | 24.8 |
| 25 | 50 | 400 | last | 0.1 | 0.13 | 0.97 | 1.03 | 0.72 | 0.57 | 24.6 | 24.49 | 24.56 |
| 25 | 50 | 400 | last | 1 | 1.3 | 0.9 | 1 | 0.81 | 0.67 | 18.53 | 22.92 | 23.04 |
| 50 | 100 | 400 | first | 0.1 | 0.15 | 0.99 | 1.14 | 0.9 | 0.63 | 49.27 | 49.14 | 48.96 |
| 50 | 100 | 400 | first | 1 | 1.91 | 0.78 | 0.93 | 0.78 | 0.71 | 5.13 | 39.43 | 36.51 |
| 50 | 100 | 400 | all | 0.1 | 0.62 | 0.81 | 0.93 | 0.49 | 0.53 | 49.89 | 49.85 | 49.86 |
| 50 | 100 | 400 | all | 1 | 1.83 | 0.97 | 1.09 | 0.8 | 0.6 | 49.78 | 49.79 | 49.73 |
| 50 | 100 | 400 | last | 0.1 | 0.15 | 0.77 | 1.14 | 0.92 | 0.63 | 49.27 | 49.31 | 49.22 |
| 50 | 100 | 400 | last | 1 | 1.94 | 0.92 | 0.97 | 0.77 | 0.71 | 4.87 | 40.52 | 36.28 |

$\beta$ and $u$ come from $y_t = F_t\beta + u$. 'first' means $\beta = (1,0,...,0)$ and 'all' means $\beta \sim N(0,1)$. The variance of $\Lambda$ is decaying as N gets larger.

relevant but the strength of them decays. I also include $\Lambda \sim iidN(0,1)$, so each factor has similar strength on X. $\beta$ is the same as DGP all factors structure.

**DGP Few Factors Structure.** Here I consider r=2 and other things are the same as DGP all factors structure.

**Summary of main findings in simulations II**

Here I summarized the situations when SPCA outperforms PCA under misspecified models. 1) *Number of factors are large*. The best rMSFE I get are from Table 5 when r=N, ranging from 0.19 to 0.92. And almost 50% of the time the MSFE is reduced by more than half. The second best rMSFE lie in Table 6 when r=N/2. Not always rMSFE

## TABLE 7DGP-MANY FACTORS 2

| DGP Parameters | | | | | MSFE | rMSFE | | Mean Weight | | Mean k | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | N | T | $\beta$ | var(u) | PCA | SPCA | SPCA($\hat{w}$) | w | $\hat{w}$ | PCA | SPCA | SPCA($\hat{w}$) |
| 5 | 10 | 400 | first | 0.1 | 0.28 | 0.93 | 0.94 | 0.1 | 0.52 | 5.67 | 6.81 | 6.3 |
| 5 | 10 | 400 | first | 1 | 1.18 | 0.99 | 1 | 0.1 | 0.68 | 4.67 | 6.88 | 5.93 |
| 5 | 10 | 400 | all | 0.1 | 0.31 | 0.95 | 0.97 | 0.1 | 0.5 | 5.67 | 6.88 | 6.4 |
| 5 | 10 | 400 | all | 1 | 1.24 | 0.99 | 1 | 0.1 | 0.67 | 4.73 | 6.89 | 5.93 |
| 25 | 50 | 400 | first | 0.1 | 0.18 | 0.99 | 1.02 | 0.9 | 0.79 | 24.52 | 25.09 | 28.01 |
| 25 | 50 | 400 | first | 1 | 1.52 | 0.95 | 0.97 | 0.9 | 0.84 | 15.28 | 34.42 | 24.89 |
| 25 | 50 | 400 | all | 1 | 1.44 | 0.99 | 1.02 | 1 | 0.85 | 15.49 | 15.49 | 24.46 |
| 25 | 50 | 400 | all | 0.1 | 0.18 | 1 | 1.02 | 1 | 0.81 | 24.61 | 24.61 | 27.91 |
| 50 | 100 | 400 | first | 0.1 | 0.18 | 0.99 | 1.02 | 1 | 0.89 | 48.73 | 48.73 | 53.75 |
| 50 | 100 | 400 | first | 1 | 1.81 | 0.99 | 1.07 | 1 | 0.9 | 6.72 | 6.72 | 27.51 |
| 50 | 100 | 400 | all | 0.1 | 0.18 | 1 | 1.04 | 1 | 0.91 | 48.78 | 48.78 | 53 |
| 50 | 100 | 400 | all | 1 | 1.85 | 1.01 | 1.04 | 1 | 0.9 | 6.13 | 6.13 | 27.8 |

$\beta$ and $u$ come from $y_t = F_t\beta + u$. 'first' means $\beta = (1,0,...,0)$ and 'all' means $\beta \sim N(0,1)$. $\Lambda \sim iidN(0,1)$. As $\Lambda \sim iid(N,1)$, $\beta = (0,0,..,1)$ and $\beta = (1,0,...,0)$ gives almost identical results, so we don't report the case 'last' here.

## TABLE 8DGP-FEW FACTORS

| DGP Parameters | | | | | MSFE | rMSFE | | Mean Weight | | Mean k | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | N | T | $\beta$ | var(u) | PCA | rSPCA | rSPCA($\hat{w}$) | w | $\hat{w}$ | PCA | SPCA | SPCA($\hat{w}$) |
| 2 | 10 | 400 | first | 0.1 | 0.25 | 0.92 | 0.94 | 0.21 | 0.52 | 2.86 | 6.17 | 7.06 |
| 2 | 10 | 400 | first | 1 | 1.17 | 1 | 1.01 | 0.63 | 0.59 | 2.03 | 7.4 | 7.35 |
| 2 | 10 | 400 | all | 0.1 | 0.4 | 0.9 | 0.91 | 0.16 | 0.52 | 3.04 | 5.55 | 6.83 |
| 2 | 10 | 400 | all | 1 | 1.33 | 0.99 | 0.99 | 0.49 | 0.57 | 2.2 | 7.19 | 7.29 |
| 2 | 50 | 400 | first | 0.1 | 0.13 | 0.99 | 1.03 | 0.4 | 0.53 | 2 | 2.21 | 7.87 |
| 2 | 50 | 400 | first | 1 | 1.03 | 1 | 1.07 | 1 | 0.75 | 1.92 | 1.92 | 7.6 |
| 2 | 50 | 400 | all | 1 | 1.08 | 1.02 | 1.12 | 0.96 | 0.52 | 1.97 | 8.08 | 12.25 |
| 2 | 50 | 400 | all | 0.1 | 0.16 | 0.99 | 1.01 | 0.28 | 0.53 | 2.07 | 8.73 | 14.15 |
| 2 | 100 | 400 | first | 0.1 | 0.11 | 1.01 | 1.07 | 0.78 | 0.54 | 2.01 | 8.2 | 20.75 |
| 2 | 100 | 400 | first | 1 | 1.03 | 1.01 | 1.3 | 0.99 | 0.54 | 1.93 | 8 | 15.19 |
| 2 | 100 | 400 | all | 0.1 | 0.13 | 1.01 | 1.03 | 0.58 | 0.55 | 2.02 | 8.29 | 17.72 |
| 2 | 100 | 400 | all | 1 | 1.04 | 1.01 | 1.27 | 0.98 | 0.52 | 1.96 | 8 | 15.88 |

$\beta$ and $u$ come from $y_t = F_t\beta + u$. 'first' means $\beta = (1,0,...,0)$ and 'all' means $\beta \sim N(0,1)$. As $\Lambda \sim iid(N,1)$, $\beta = (0,0,..,1)$ and $\beta = (1,0,...,0)$ gives almost identical results, so we don't report the case 'last' here.

is smaller than 1, but about 67% of the time the rMSFE is smaller than 0.94, and 50% chances the rMSFE is smaller than 0.83.

2) *The variance of factor loadings are different.* Table 5 and Table 6 show the comparison.

3) *Factors not important in explaining x are important in forecasting y.* This is illustrated in Table 5, where the first factor has the strongest strength in explaining x, while the last factor has the weakest strength. When $beta \sim N(0,1)$, every factor contributes to y, and I get the smallest rMSFE as small as 0.71. Interestingly, $\beta = (1,...,0)$ and $\beta = (0,...,1)$ have similar rMSFEs. This is very likely caused by the estimated k. In these two situation, y only needs 1 k. But the estimated k is still close to its truth number. Therefore, both situation contains more than enough factors, and generate similar rMFSE.

4) *N is small.* Within the same DGP , rMSFE increases with N. For example in Table 5 line 1,5 and 9, rMSFE rise from 0.32, to 0.45 then to 0.52 as N increases . In all tables, when N equals to 10, I observed reduced MSFE.

5) *u in $y_{t+h} = F_t\beta + u$ is close to 0.* Within the same DGP, when $u$ $0.1N(0,1)$, the rMSFE is always smaller than its counterpart where only u is changed to $u$ $N(0,1)$ product

## VI. Empirical Applications

In the empirical study, I aim to find ways to improve SPCA further. The first approach is to use Sieve-LS Covariance matrix. Referring to Fan *et al.* (2016), I project X on additional observables under additive Fourier basis. The benefit of using proxies are extra explanatory power, less strictly identification conditions, and faster convergence rate. The second approach is to accommodate nonlinear combinations of factors in the forecasting process. I included the classic Index Model and the relatively newer Neural Networks (NN). In this section, I want to test if our supervised design is compatible with other types of PCA (Sieve-LS Covariance) and if nonlinear prediction models enhance the performance of SPCA factors.

*(a) Model Specifications*

**Neural Networks**

Derived nodes $N_m$ are functions of linear combinations of the inputs. Those functions are called active functions and the commonly used ones are sigmoid, ReLU, TanH and ect. This step can be repeat several times and each time I create new a layer of nodes. At last, the target $Y_k$ is modeled from linear combinations of nodes in $Y_k$'s previous layer.

$$\begin{aligned} N_m =& \sigma\left(\alpha_{0m} + \alpha_m^\top X\right), m = 1, \ldots, M \\ T_k =& \beta_{0k} + \beta_k^\top N, k = 1, \ldots, K \\ f_k(X) =& g_k(T), k = 1, \ldots, K \end{aligned}$$

The nonlinearity generates from simple activation functions, and therefore NN usually needs several layers to recover the rich informations contained in X.

The use of NN has be demonstrated in the literature. Babikir and Mwambi (2016) demonstrate that combination methods with nonlinear neural network offer more accurate forecasts than combination methods with linear specifications in forecasting return indexes. Gu *et al.* (2018) investigate 30,000 stocks over 60 years then constructed portfolios with various methods: Elastic Net, PCR, PLS, Generalized Linear, Boosting, Random Forests and Neural Networks. In both Monte Carlo simulations and empirical study, they find that Neural Network provides significantly outperform other methods. NN is arguably the 'universal appropriator' and the most competitive device in the machine learning world (Hornik *et al.*, 1989). But at the same time NN is only as good as the data it's been given, therefore data cleaning is essential (Azoff, 1994). This give us the idea that NN can be further improved if it's been given the supervised factors instead of the big data set.

**Index Model**

Ichimura (1993) defines the single-index model and presents $1/\sqrt{(}n)$ consistency and asymptotic normality. Single Index Model can be expanded to more indexes. In my paper a double index model is used. Index model covers censored Tobit and bi-

nary choice models. The conditional expectation is estimated with a kernel estimator (Watson, 1964).

The index model does not assume the conditional mean function is known, it is more flexible. It is potentially superior to nonparametric model in the following aspects. First as the dimension of X increases, the estimation error of conditional mean increases rapidly (Stone, 1980). But double index model achieves the convergence rate as if there were only a two dimensional variable.

At each time i,

$$(8) \quad Y_i = M_i(\theta_0, \gamma_0) + \varepsilon_i$$

where $i = 1, ..., T,\ E(\varepsilon|X) = 0,\ E(\varepsilon\varepsilon'|X) = \sigma^2 I$

$$(9) \quad M_i(\theta_0, \gamma_0) = E(Y_i|X_i) = E[Y_i|X_{1i} + \sum_{j=2}^{N} X_{ji}\theta_0, X_{1i} + \sum_{j=2}^{N} X_{ji}\gamma_0]$$

Then the estimator solves the following problem:

$$(10) \quad (\hat{\theta}, \hat{\gamma}) = argmin\hat{Q}(\theta, \gamma) = 1/N \sum_{i=1}^{T} \left\{ Y_i - \hat{M}(\theta, \gamma) \right\}^2 \tau_i$$

where $\tau_i$ trims the outliers. And $\hat{M}$ is the nonparametric expectation of Y conditional on the index $V(X; \theta, \gamma)$. With a Gaussian kernel,

$$(11) \quad \hat{M}(\theta, \gamma) = \hat{E}\big(Y|(V(X; \theta; \gamma) = v)\big) = \frac{\frac{1}{T}\sum_{i=1}^{T} y_i \frac{\phi[(v - v_i)/h]}{h}}{\frac{1}{T}\sum_{i=1}^{T} \frac{\phi[(w - v_i)/h]}{h}}$$

The first order condition of equation 10 is

$$(12) \quad \frac{1}{T}\sum_{i=1}^{T} \left[ Y_i - \hat{M}_i(\theta; \gamma) \right] \tau_i \triangledown \hat{M}_i$$

To solve the derivative free minimization problem, Quasi-Newton method is used.

Scholars have used semiparametric for forecasting. The performances are mixed. Chan and Liu (2017) propose a semiparametric approach to model the online auction

TABLE 9MODEL DESCRIPTIONS

| Model Name | Factor Extraction | Forecasting |
|---|---|---|
| PCA | Principle Component Analysis k=8 | OLS on 8 factors of PCA |
| SPCA | Supervised Principle Component Analysis k=8 | OLS on 8 factors of SPCA |
| PCA_SIEVE | Use PCA on the Sieve-LS Covariance matrix | OLS on 8 factors of PCA_SIEVE |
| PCA_IDX | Principle Component Analysis k=8 | Single Index Model on 8 factors of PCA |
| SPCA_IDX | Supervised Principle Component Analysis k=8 | Single Index Model on 8 factors of SPCA |
| NN | NAN | Neural Networks on all X |
| PCA_NN | Principle Component Analysis k=8 | Neural Networks on 8 factors of PCA |
| SPCA_NN | Supervised Principle Component Analysis k=8 | Neural Networks on 8 factors of SPCA |

process. Except for the pricing curves, their model also captures the arrival rates of the bidding. Then forecasts perform well with Xbox data. Weron and Misiorek (2008) apply semiparametric in forecasting the spot electricity prices. They use the smoothed nonparametric ML estimator (SN) and iterated Hsieh-Manski estimator (IHM). The semiparametric models generally stand out among other models. On the other hand, McAleer and Da Veiga (2008) compare the performance of the single-index model with portfolio models in forecasting value-at-risk (VaR) thresholds. The portfolio model approach leads to superior forecasts compared with their single-index analogue based on $R^2$.

**Sieve_LS Covariance Matrix**

$w_t$ is some external data other than X. Fan *et al.* (2016) explained the process of incorporating Sieve_LS covariance matrix with PCA. To incorporate the explanatory power of $w_t$, first they project $\{X_t\}$ on the external variables $\{w_t\}$ and obtain fitted value $\left\{\widehat{X}_t\right\}$. Then they conduct PCA on $\left(\widehat{X}_1,\ldots,\widehat{X}_T\right)$.

There are two main benefits of using proxies. First, the loadings can be identified given any N, not asymptotically (e.g., as $N\to\infty$). Secondly when $\gamma_t$ is near zero, using proxies can estimated factors well with a relatively smaller sample size, as the convergence rate is faster at $O_P\left(T^{-1}+(NT)^{-1/2}\right)$.

Here I project X on addictive Fourier basis. Specifically, $M(w_t) = (m_1(w_t)$ $\ldots, m_L(w_t))'$ is a $L\times 1$ vector of Fourier basis. Let $P = M'\left(MM'\right)^{-1}M, (T\times T), \quad M = (M(w_1),\ldots,M(w_T)), (L\times T), \quad X = (x_1,\ldots,x_T), (T\times N)$ Then, the linear projection

$E(X|w_1,\ldots,w_T)$ is $PX$ and $\Sigma$ is $\tilde{\Sigma} = \frac{1}{T}X'PX$ is the estimated covariance matrix. Here L = 5 Fan *et al.* (2016).

The models being used are described in Table 9. The MSFEs are calculated with both moving windows and recursive windows. Relative MSFE are reported together with (Diebold and Mariano, 1995) predictive accuracy test (DM-test). The null hypothesis of DM test is that the predictive accuracy of two models are the same: $E\{\varepsilon^2_{1,t+h|t}\} - E\{\varepsilon^2_{2,t+h|t}\} = 0$, where $\varepsilon_{i,t+h|t}$ is model i's h period ahead prediction error. The test statistic DM-test=$\frac{1}{P}\sum_{i=1}^{P} d_t \hat{\sigma}_{\bar{d}}$. Where $d_t = \hat{\varepsilon}^2_{1,t+h|t} - \hat{\varepsilon}^2_{2,t+h|t}$,$\bar{d}$ is the average of the estimated difference $d_t$. And $\hat{\sigma}_{\bar{d}}$ is the estimated standard deviation of $\bar{d}$. significance level. When DM-test is significantly smaller than 0, then Model 1 is more accurate in prediction than Model 2.

*(b) Data*

Following Ludvigson and Ng (2009); Jurado *et al.* (2015), I use 131 monthly series, and 4 bond series between 1964 till the end of 2003. The excess bond returns are calculated with respect to U.S zero coupon Treasure bond Cochrane and Piazzesi (2005). Please refer to Ludvigson and Ng (2009) for descriptions of data transformation. To be analogous to the study of Ludvigson and Ng (2009) and Bai and Ng (2002), the number of factor is fixed at 8.

Here is how I calculate bond risk premium. I buy an n year bond at time t then sell it next year. The excess return comparing to the one-year bond yield is what I define the bond risk premium. Let $p_t^{(n)}$ be the discounted log price of an $n$ -year bond at time t. The log yield is $z_t^{(n)} \equiv -\frac{1}{n}p_t^{(n)}$. The log return for holding the bond for 1 year is $r_{t+1}^{(n)} \equiv p_{t+1}^{(n-1)} - p_t^{(n)}$. At last, I get the log excess return

$$y_{t+1}^{(n)} = r_{t+1}^{(n)} - z_t^{(1)}, \quad t = 1,\cdots,T$$

. By definition, 1 period ahead means 1 year ahead so I set h=1 year.

Moreover, I picked 10 macro variables referring to Kim and Swanson (2014). 1 month ahead and 1 year ahead forecasts are being made. And the descriptions are in Table 10.

TABLE 10TARGET FORECASTING VARIABLE

| Series | Description | Transformation |
|---|---|---|
| UR | Unemployment Rate: All | $\Delta lv$ |
| PI | RPI ex. Transfers | $\Delta ln$ |
| 10 T | 10 year T-bond | $\Delta lv$ |
| CPI | CPI all Items | $\Delta lv$ |
| PPI | Producer Price Index: Finished Goods | $\Delta^2 ln$ |
| EMP | Employees on Nonfarm Payrolls | $\Delta ln$ |
| 5 T | 5 year T-bond | $\Delta ln$ |
| HOUST | Housing Starts: Non farm | $ln$ |
| IP | Industrial Production Index | $\Delta ln$ |
| M2 | Money Supply real | $\Delta^2 ln$ |
| S&P 500 | S&P's Common Stock Price Index | $\Delta ln$ |

*(c) Empirical Findings*

Now I will go through some details of the results first and summarize the main findings at the end.

By definition, the one period ahead for bond risk premium means one year ahead. In Table 11, MSFE0 is the MSFE using PCA. The following rows are relative MSFE using PCA as a baseline. Let us first look at the sample where I have 240 training and testing months. PCA has smaller MSFE when the maturity time is shorter. The 2 year bond has 3.044 MSFE while the 5 year bond has almost 10 times the MSFE. SPCA also forecast the best for the 2 year bond risk premium at a relative MSFE of 0.686. The relative MSFEs for 3 to 5 year bonds are all below 0.75. Then I incorporate both PCA and SPCA with the single-index model. rMSFE for PCA_IDX are all significantly below 1 now, with the smallest being 0.876 and the largest being 0.943. The rMSFE for

TABLE 11BOND RISK PREMIA RMSFE H=1 YEAR

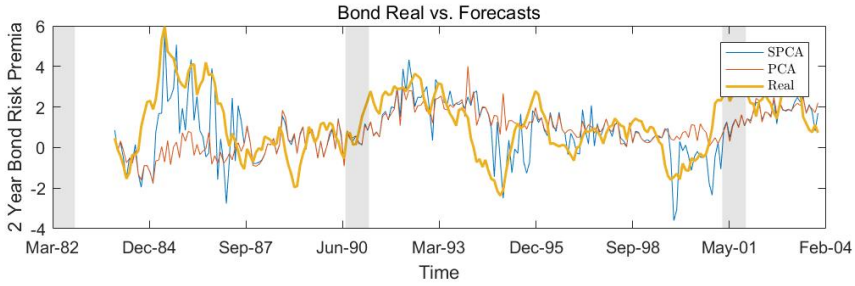
| Moving | | | | | Recursive | | | | |
|---|---|---|---|---|---|---|---|---|---|
| h=1 year | yr2 | yr3 | yr4 | yr5 | h= 1 year | yr2 | yr3 | yr4 | yr5 |
| Jan-64:Dec-03 (R=P=240) | | | | | Jan-64:Dec-03 (R=P=240) | | | | |
| MSFE0 | 3.044 | 10.532 | 21.453 | 32.771 | MSFE0 | 2.948 | 10.279 | 20.848 | 31.776 |
| SPCA | 0.686** | 0.744** | 0.720** | 0.747** | SPCA | 0.674** | 0.715** | 0.711** | 0.810** |
| PCA_SIEVE | 0.728 | 0.746 | 0.732** | 0.742** | PCA_SIEVE | 0.772** | 0.784** | 0.763** | 0.774** |
| SPCA_SIEVE | 0.689** | 0.706** | 0.665*** | 0.686** | SPCA_SIEVE | 0.711** | 0.724** | 0.684*** | 0.694*** |
| PCA_IDX | 0.876*** | 0.905** | 0.926** | 0.943** | NN | 0.644** | 0.642*** | 0.672** | 0.718** |
| SPCA_IDX | 0.616*** | 0.692*** | 0.681*** | 0.706*** | PCA_IDX | 1.019** | 0.994 | 1.005 | 1.018** |
| NN | 0.751** | 0.765** | 1.128 | 0.772** | SPCA_IDX | 0.746** | 0.758** | 0.740** | 0.777** |
| PCA_NN | 0.688*** | 0.967 | 0.814** | 0.750** | PCA_NN | 0.649** | 0.667** | 0.673** | 0.734** |
| SPCA_NN | 0.568*** | 0.655*** | 0.622*** | 0.571*** | SPCA_NN | 0.407*** | 0.446*** | 0.376*** | 0.433*** |
| Jan-64:Dec-83 (R=P=180) | | | | | Jan-64:Dec-83 (R=P=180) | | | | |
| MSFE0 | 4.865 | 16.546 | 31.666 | 47.619 | MSFE0 | 4.911 | 16.073 | 28.893 | 43.387 |
| SPCA | 0.794** | 0.800** | 0.818** | 0.828* | SPCA | 0.784** | 0.805** | 0.730** | 0.669** |
| PCA_SIEVE | 0.841** | 0.852* | 0.835** | 0.851* | SIEVE | 0.752 | 0.785** | 0.757* | 0.772** |
| SPCA-SIEVE | 0.776** | 0.707** | 0.666*** | 0.702** | SPCA-SIEVE | 0.715** | 0.682*** | 0.630*** | 0.666*** |
| PCA_IDX | 0.969 | 1.005 | 1.027 | 1.01 | PCA_IDX | 0.943 | 0.981 | 0.956 | 0.966 |
| SPCA_IDX | 0.815** | 0.828** | 0.867** | 0.854** | SPCA_IDX | 0.698** | 0.711** | 0.778** | 0.723* |
| NN | 0.826** | 0.817** | 0.829** | 0.888 | NN | 0.541** | 0.668** | 0.485*** | 0.513** |
| PCA_NN | 0.904** | 0.854** | 0.715** | 0.846** | PCA_NN | 0.605** | 0.699* | 0.655** | 0.501** |
| SPCA_NN | 0.669** | 0.811** | 0.523*** | 0.785** | SPCA_NN | 0.488*** | 0.495*** | 0.514** | 0.358*** |
| Jan-84:Dec-03 (R=P=180) | | | | | Jan-84:Dec-03 (R=P=180) | | | | |
| MSFE0 | 3.323 | 12.561 | 24.673 | 38.458 | MSFE0 | 3.692 | 12.983 | 26.076 | 39.084 |
| SPCA | 0.906** | 0.957 | 0.948 | 0.957 | SPCA | 0.962** | 0.961** | 0.952* | 0.957* |
| PCA_SIEVE | 1.103** | 1.050** | 1.038** | 1.041** | SIEVE | 1.215** | 1.160** | 1.147** | 1.142* |
| SPCA-SIEVE | 0.816** | 0.738** | 0.701** | 0.692** | SPCA-SIEVE | 1.052 | 0.966 | 0.94 | 0.913** |
| PCA_IDX | 0.835** | 0.813** | 0.802** | 0.828** | PCA_IDX | 1.025 | 1.051** | 1.043** | 1.040** |
| SPCA_IDX | 0.866* | 0.851* | 0.803** | 0.846* | SPCA_IDX | 0.952** | 0.966 | 0.972 | 0.969 |
| NN | 0.893** | 1 | 1.009 | 0.905** | NN | 0.847** | 0.881** | 1.001 | 0.823** |
| PCA_NN | 0.928** | 0.995 | 0.885** | 0.793** | PCA_NN | 0.862* | 0.947** | 0.866* | 0.771** |
| SPCA_NN | 0.823** | 0.937** | 0.934 | 0.810** | SPCA_NN | 0.864** | 0.845* | 0.823** | 0.922** |

Note: yr2-yr5 are the risk premium on the n-year Treasury bond. MSFE0 is the MSFE for PCA. Other entires are relative MSFE with PCA as the baseline. *, **, *** indicates the significance level of 0.1, 0.05, and 0.001 respectively.

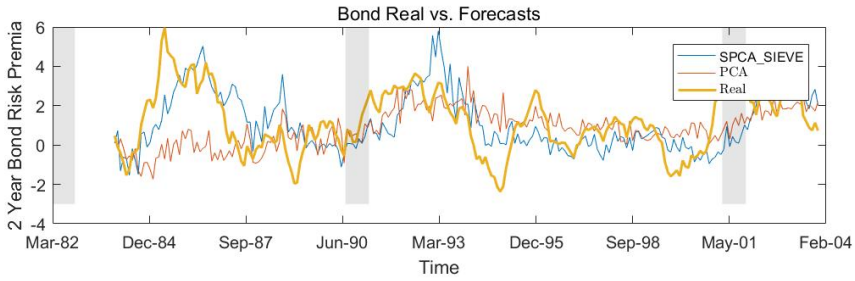
SPCA_IDX are all smaller than SPCA. This means non-linearity of single-index model improves the forecasting power of both PCA and SPCA, while SPCA is still superior. The next type of non-linear model is Neural Networks. Using NN alone generates similar MSFE for yr2 and yr4 to SPCA, and slightly better MSFE for yr3 and yr5. NN also improves both PCA and SPCA. Now PCA_NN has comparable MSFE as SPCA. And SPCA_NN generates the smallest rMSFE for each bond, varying from 0.386 to 0.408. Note that even for the 5 year bond,SPCA_NN is still very good.

The next two sub-samples have 180 months of training and testing period. MSFE for PCA are bigger than that from the previous longer sample. All the rMSFE are also bigger than those using longer sample. This time I see some entries greater than 1. The appeared at PCA_IDX and PCA_NN. That tells us that non-linear model is worse than simple OLS for PCA when sample size is small. In this sample, the best rMSFE for 2 year and 3 year bond comes from SPCA_NN, at 0.703 and 0.583 respectively. SPCA gives the best rMSFE for 4 year bond. SPCA and NN both gives similar small rMSFE for 5 year bond.

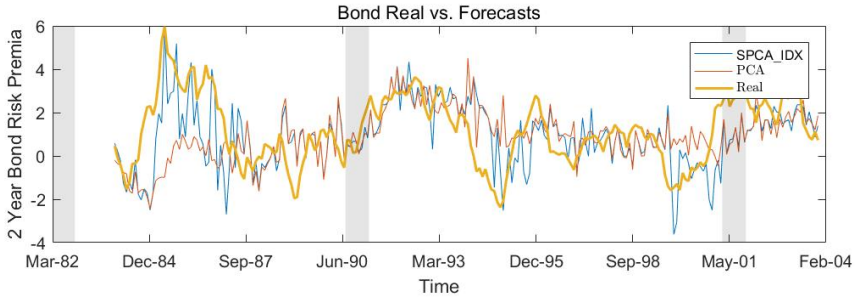
In another 180 months sub-sample, it is again clear that SPCA_NN generates the best forecasting performance. SPCA alone is no better than any combination model.




(A) SPCA




(B) SPCA_SIEVE




(C) SPCA_IDX

From the timeline of SFE in Figures **??**, I can see that PCA has several bad forecasts starting at December 1984. When combining with non-linear model, PCA still have many worse estimates in non-recession periods. SPCA had one or two bad forecasts toward then end of the 2001 recession. I can say that in general SPCA works better, but may not be good when recession ends.

TABLE 12 MACRO VARIABLES RELATIVE MSFE H=1 MONTH RECURSIVE WINDOWS

| h=1 Month | UR | PI | 10 TB | CPI | PPI | EMP | HOUST | IP | M2 | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|---|
| Jan-64 : Dec-03 (R=P=240) | | | | | | | | | | |
| MSFE0 | 0.666 | 1.403 | 0.763 | 0.778 | 0.815 | 0.237 | 0.148 | 0.426 | 0.924 | 1.022 |
| SPCA | 1.102** | 1.006** | 1.143** | 1.170** | 1.046** | 0.949** | 0.753** | 1.175** | 0.783** | 1.293** |
| PCA_SIEVE | 1.040** | 0.997** | 1.090** | 1.112** | 1.038 | 0.954*** | 1.741*** | 1.142** | 0.787* | 1.061** |
| SPCA_SIEVE | 1.416** | 1.115** | 1.236*** | 1.013 | 0.994 | 2.336*** | 5.602*** | 1.690** | 0.958 | 1.117** |
| PCA_IDX | 0.973 | 0.959** | 0.943** | 1.023 | 1.043** | 0.973** | 1 | 1.034** | 1.009 | 0.966** |
| SPCA_IDX | 1.176* | 1.086** | 1.005 | 1.135** | 1.069** | 1.230** | 0.664*** | 1.146** | 0.745*** | 1.046 |
| NN | 0.315*** | 0.559** | 0.464*** | 0.424*** | 0.289*** | 0.597** | 0.432*** | 0.564* | 0.338*** | 0.368** |
| PCA_NN | 0.292*** | 0.115** | 0.482*** | 0.518*** | 0.367*** | 0.584** | 0.325*** | 0.656** | 0.349*** | 0.352*** |
| SPCA_NN | 0.237*** | 0.339** | 0.264*** | 0.183*** | 0.233*** | 0.336*** | 0.091*** | 0.220*** | 0.178*** | 0.349** |
| Jan-64 : Dec-83 (R=P=180) | | | | | | | | | | |
| MSFE0 | 1.3 | 0.563 | 1.789 | 1.2 | 0.957 | 1.03 | 0.224 | 1.187 | 0.772 | 1.069 |
| SPCA | 1.129** | 1.087** | 1.145* | 1.783** | 0.970** | 1.015 | 1.028 | 1.032** | 1.046 | 1.01 |
| PCA_SIEVE | 1.030*** | 1.049** | 1.060** | 1.12 | 1.104** | 1.147*** | 0.963*** | 1.120*** | 0.812 | 0.983 |
| SPCA_SIEVE | 1.393*** | 1.219** | 0.935 | 1.105** | 1.180** | 3.091*** | 4.945*** | 1.717*** | 0.857** | 0.972 |
| PCA_IDX | 1.099** | 0.98 | 1.012 | 1.068** | 1.089 | 1.072** | 1.156* | 1.277** | 0.780** | 1.102** |
| SPCA_IDX | 0.984 | 1.069 | 1.257** | 1.441* | 1.07 | 1.087** | 1.281** | 1.094** | 0.745** | 1.048 |
| NN | 0.468*** | 0.435** | 0.389** | 1.183 | 0.497** | 1.01 | 0.783** | 0.589** | 0.973 | 0.383*** |
| PCA_NN | 0.541** | 0.505** | 0.406** | 1.489 | 0.455** | 0.483** | 0.717** | 0.660** | 0.610* | 0.475** |
| SPCA_NN | 0.421*** | 0.534** | 0.574** | 1.15 | 0.460** | 0.834 | 0.384*** | 0.640** | 0.404*** | 0.349*** |
| Jan-84 : Dec-03 (R=P=180) | | | | | | | | | | |
| MSFE0 | 0.746 | 1.527 | 0.922 | 0.624 | 0.791 | 0.213 | 0.259 | 0.424 | 0.723 | 0.887 |
| SPCA | 1.007 | 1.381** | 1.057** | 1.267** | 1.078** | 0.982** | 0.735** | 1.008 | 0.807** | 1.012** |
| PCA_SIEVE | 1.040** | 0.955*** | 1.137 | 1.100** | 1.124 | 1.118*** | 1.956*** | 1.176* | 0.92 | 1.074** |
| SPCA_SIEVE | 1.235** | 1.076* | 1.104** | 0.993 | 1.106** | 2.267*** | 4.438*** | 1.344*** | 0.972 | 1.002 |
| PCA_IDX | 0.955** | 0.977** | 0.982 | 1.006 | 1.094** | 0.977 | 1.114** | 0.897** | 1.012 | 1.003 |
| SPCA_IDX | 1.043 | 1.162 | 1.03 | 1.142** | 1.123** | 0.954 | 0.911** | 0.967 | 0.900** | 0.948** |
| NN | 0.281*** | 0.843 | 0.486*** | 0.522** | 0.493** | 0.646** | 0.453*** | 0.354*** | 0.623** | 0.345** |
| PCA_NN | 0.372*** | 0.409** | 0.618** | 0.552* | 0.456** | 0.901 | 1.467 | 0.560** | 0.481*** | 0.352** |
| SPCA_NN | 0.401*** | 0.98 | 0.435*** | 0.272*** | 0.321*** | 0.840** | 0.171*** | 0.373*** | 0.207*** | 0.445** |

Note: MSFE0 is the MSFE for PCA. Other entires are relative MSFE with PCA as the baseline. *, **, *** indicates the significance level of 0.1, 0.05, and 0.001 respectively. R and P are the length of training and testing period respectively.

TABLE 13 MACRO VARIABLES RECURSIVE WINDOWS RELATIVE MSFE H=12 MONTHS

| h= 12 Months | UR | PI | 10 TB | CPI | PPI | EMP | HOUST | IP Index | M2 | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|---|
| Jan-64 : Dec-03 (R=P=240) | | | | | | | | | | |
| MSFE0 | 0.721 | 1.448 | 0.801 | 0.758 | 0.886 | 0.542 | 0.505 | 0.548 | 1.062 | 1.096 |
| SPCA | 1.079** | 0.987 | 1.041** | 1.053** | 1.084** | 0.758*** | 0.742** | 1.028** | 0.976 | 1.145** |
| PCA_SIEVE | 0.952** | 1.113 | 0.985** | 1.021** | 0.718** | 1.037 | 1.009 | 0.879* | 1.034 | 1.018** |
| SPCA_SIEVE | 1.027** | 1.065** | 0.935** | 1.078** | 0.846** | 1.018 | 1.032 | 1.041 | 1.037** | 1.050** |
| NN | 0.284*** | 0.504* | 0.258*** | 0.435*** | 0.469*** | 0.479*** | 0.594*** | 0.436*** | 0.596** | 0.449** |
| PCA_IDX | 0.982** | 0.947** | 0.984 | 0.946* | 0.933** | 0.920** | 0.822** | 0.923** | 0.966** | 0.866** |
| SPCA_IDX | 1.063** | 1.165 | 1.115** | 0.967** | 1.017 | 0.843** | 0.735** | 1.02 | 1.043 | 0.927 |
| PCA_NN | 0.570** | 0.348* | 0.367*** | 0.360*** | 0.389*** | 0.489*** | 0.583*** | 0.470*** | 0.551*** | 0.343*** |
| SPCA_NN | 0.214*** | 0.461** | 0.337*** | 0.253*** | 0.296*** | 0.213*** | 0.371*** | 0.349*** | 0.280*** | 0.330*** |
| Jan-64 : Dec-83 (R=P=180) | | | | | | | | | | |
| MSFE0 | 1.839 | 0.671 | 2.205 | 1.324 | 1.116 | 1.63 | 1.14 | 1.759 | 1.238 | 1.422 |
| SPCA | 0.993 | 0.982** | 1.016 | 0.957** | 0.950** | 0.972** | 1.219** | 0.977** | 1.218* | 1.005 |
| SIEVE | 1.002** | 1.038** | 1.076 | 1.092 | 0.774*** | 1.022 | 0.988** | 0.822** | 1.017** | 1.014 |
| SPCA-SIEVE | 1.050** | 1.143* | 1.088** | 1.108** | 1.344* | 1.027** | 1.106** | 0.662** | 1.052** | 1 |
| PCA_IDX | 1.081** | 0.931** | 0.968 | 0.770** | 0.733** | 0.886** | 0.852** | 0.829* | 0.843** | 0.888** |
| SPCA_IDX | 1.078** | 0.935** | 0.942** | 0.792* | 0.857 | 0.839** | 0.880** | 0.908** | 1.035 | 1.024 |
| NN | 0.444*** | 0.517** | 0.438** | 0.456*** | 0.824** | 0.572** | 1.093 | 0.547*** | 0.752** | 0.338*** |
| PCA_NN | 0.383*** | 0.464** | 0.341** | 1.136 | 0.895 | 0.761** | 1.041 | 0.304*** | 0.826** | 0.417*** |
| SPCA_NN | 0.386*** | 0.395*** | 0.281** | 0.708** | 1.359 | 0.448** | 0.938 | 0.538** | 0.624** | 0.368*** |
| Jan-84 : Dec-03 (R=P=180) | | | | | | | | | | |
| MSFE0 | 0.838 | 1.612 | 0.965 | 0.723 | 0.813 | 0.545 | 0.77 | 0.536 | 0.921 | 0.954 |
| SPCA | 0.992** | 0.817** | 0.989 | 0.990** | 1.002 | 1.047 | 0.795** | 0.967** | 0.937** | 1 |
| SIEVE | 0.939** | 1.121** | 1.011 | 1.035 | 0.655*** | 1.039** | 1.046 | 1.088*** | 1.025 | 1.009** |
| SPCA-SIEVE | 1.031** | 1.032 | 0.974 | 1.076** | 0.395*** | 0.986 | 1.06 | 1.622*** | 1.035 | 1.058** |
| PCA_IDX | 1.003 | 0.992 | 0.963 | 0.862** | 0.981 | 0.961 | 0.721*** | 1.126** | 0.907** | 1.001 |
| SPCA_IDX | 1.364** | 0.854** | 0.914** | 0.928** | 0.989 | 1.160** | 0.793** | 1.198* | 0.882** | 0.97 |
| NN | 0.330*** | 0.096* | 0.370*** | 0.194*** | 0.186*** | 0.850** | 0.654** | 0.495** | 0.753* | 0.190*** |
| PCA_NN | 0.315*** | 0.455** | 0.509** | 0.282*** | 0.249*** | 0.769** | 0.691** | 0.512** | 0.632** | 0.296*** |
| SPCA_NN | 0.433*** | 0.380** | 0.550*** | 0.190*** | 0.183*** | 0.584** | 0.424*** | 0.393*** | 0.399*** | 0.196*** |

Note: MSFE0 is the MSFE for PCA. Other entires are relative MSFE with PCA as the baseline. *, **, *** indicates the significance level of 0.1, 0.05, and 0.001 respectively. R and P are the length of training and testing period respectively.

The results actually don't change using recursive windows from using moving window. SPCA is still better than PCA. And SPCA_NN is still the best method.

Next I test my model with Macro variables. First let us concentrate on R=P=240 in Table 14. SPCA only improves Housing Starts for the 1 step ahead forecasts, where the rMSFE is almost half of that using PCA. PCA_IDX and SPCA_IDX both are significantly better than PCA for about half of the macro variables. Among which SPCA_IDX is the best model for Housing Starts. Other than Housing Starts, NN and the hybrid models with NN generates much lower rMSFE than PCA. SPCA_NN is rMSFE best for PI, 10TB, PPI,M2REAL and S&P 500. Moreover the best rMSFE for UR and PAYEMS are very close to those from SPCA_NN.

Then I move to the smaller sub-samples. Again SPCA improves Housing Starts. But the relative improvements are smaller than the first larger sample. This is true when SPCA combined with Index model or NN in most cases. For instance, the rMSFE for PAYEMS using SPCA_NN are 0.584, 0.788 and 0.639 respectively for the sample 1,2 and 3.

TABLE 14 MOVING WINDOWS RELATIVE MSFE H=1 MONTH

| h= 1 Month | UR | PI | 10 TB | CPI | PPI | EMP | HOUST | IP | M2 | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Jan-64 : Dec-03 (R=P=240) | | | | | |
| MSFE0 | 0.686 | 1.363 | 0.772 | 0.753 | 0.828 | 0.225 | 0.245 | 0.431 | 0.914 | 1.014 |
| SPCA | 1.117** | 1.114** | 1.103* | 1.099* | 1.100** | 1.106* | 0.553*** | 1.951** | 0.951** | 1.102** |
| PCA_SIEVE | 0.993*** | 1.064** | 1.061** | 1.005** | 1.080** | 1.039*** | 1.023*** | 1.203*** | 0.788** | 1.011** |
| SPCA-SIEVE | 1.304** | 1.189** | 1.154** | 1.046** | 1.130* | 3.035*** | 3.474*** | 1.676*** | 1.005 | 1.067** |
| PCA_IDX | 0.993*** | 1.064** | 1.061** | 1.005** | 1.080** | 1.039*** | 1.023*** | 1.203*** | 0.788** | 1.011** |
| SPCA_IDX | 1.304** | 1.189** | 1.154** | 1.046** | 1.130* | 3.035*** | 3.474*** | 1.676*** | 1.005 | 1.067** |
| NN | 0.324*** | 0.704** | 0.355*** | 0.233*** | 0.498*** | 0.887 | 0.843** | 0.506*** | 0.661** | 0.175*** |
| PCA_NN | 0.502*** | 0.181** | 0.382*** | 0.348*** | 0.411*** | 0.577** | 0.802** | 0.374*** | 0.669** | 0.356*** |
| SPCA_NN | 0.344*** | 0.111** | 0.286*** | 0.309*** | 0.396*** | 0.584** | 0.741*** | 0.502*** | 0.525*** | 0.261*** |
| | | | | | Jan-64 : Dec-83 (R=P=180) | | | | | |
| MSFE0 | 0.930 | 1.385 | 1.683 | 0.920 | 0.760 | 0.545 | 0.260 | 0.634 | 0.748 | 0.915 |
| SPCA | 1.072** | 1.322** | 1.041** | 1.011 | 1.058** | 1.027 | 0.866* | 1.117** | 0.947** | 1.015** |
| PCA_SIEVE | 0.990** | 1.080** | 1.123** | 1.199** | 1.041 | 0.951*** | 0.941*** | 1.074*** | 0.910** | 1.019* |
| SPCA-SIEVE | 1.481** | 1.177** | 1.210** | 1.059** | 1 | 2.420*** | 5.206*** | 1.791*** | 1.136** | 1.174* |
| PCA_IDX | 0.990** | 1.080** | 1.123** | 1.199** | 1.041 | 0.951*** | 0.941*** | 1.074*** | 0.910** | 1.019* |
| SPCA_IDX | 1.481** | 1.177** | 1.210** | 1.059** | 1 | 2.420*** | 5.206*** | 1.791*** | 1.136** | 1.174* |
| NN | 0.327*** | 0.967 | 0.601** | 0.99 | 0.266*** | 1.227** | 1.029 | 0.644** | 0.854** | 0.242** |
| PCA_NN | 0.656* | 1.181 | 0.649** | 1.071 | 0.436** | 0.955 | 0.815** | 1.18 | 0.675** | 0.354** |
| SPCA_NN | 0.314*** | 1.043 | 0.566** | 0.838 | 0.229*** | 0.788** | 0.783** | 0.552** | 0.434*** | 0.446** |
| | | | | | Jan-84 : Dec-03 (R=P=180) | | | | | |
| MSFE0 | 0.925 | 2.515 | 1.022 | 1.160 | 1.255 | 0.283 | 0.291 | 0.628 | 1.405 | 1.501 |
| SPCA | 0.983** | 1.265** | 1.060** | 1.166** | 1.097** | 1.024** | 0.877** | 1.121** | 0.846** | 1.187** |
| PCA_SIEVE | 1.030** | 1.063* | 1.072** | 1.035 | 1.024** | 0.986*** | 0.653*** | 1.161** | 0.838** | 1.109** |
| SPCA-SIEVE | 1.412** | 1.176** | 0.951 | 1.076** | 1.184** | 3.478*** | 4.365*** | 1.598*** | 0.928 | 1.013 |
| PCA_IDX | 1.030** | 1.063* | 1.072** | 1.035 | 1.024** | 0.986*** | 0.653*** | 1.161** | 0.838** | 1.109** |
| SPCA_IDX | 1.412** | 1.176** | 0.951 | 1.076** | 1.184** | 3.478*** | 4.365*** | 1.598*** | 0.928 | 1.013 |
| NN | 0.401*** | 1.049 | 0.288*** | 1.163 | 0.429** | 0.654** | 0.726** | 0.335*** | 0.708** | 0.298*** |
| PCA_NN | 0.325*** | 0.724** | 0.402*** | 1.265 | 1.738** | 0.820** | 0.766** | 0.436** | 0.797** | 0.360*** |
| SPCA_NN | 0.460** | 0.951 | 0.442*** | 0.969 | 0.620** | 0.639** | 0.660** | 0.423*** | 0.505* | 0.261*** |



Note: MSFE0 is the MSFE for PCA. Other entires are relative MSFE with PCA as the baseline. *, **, *** indicates the significance level of 0.1, 0.05, and 0.001 respectively. R and P are the length of training and testing period respectively.

I also conducted the same experiment for 1 year step ahead forecast in Table **??** and Table 13. SPCA_NN is rMSFE best for UR, 10TB, PAYEMS,HOUST, IP, M2REAL and S&P 500. All together 7/10 cases when R=P=240. macroh12mv

*(d) Portfolio Excess Returns*

In the classical financial models, the underlying fundamentals to the stock returns. CAPM(Sharpe, 1964), for example, identifies the market risk premium as a factor. Fama and French added two more factors, size, and book-to-market value on top of the CAPM to represent the underlying forces for stock returns. They later included more factors. With easier access to a large amount of data, Principle Component Analysis can extend the traditional factor models by offering a way to summarize a large number of variables into a small number of factors.

I obtained monthly stock returns from January 1978 to December 2019 on CRSP, which spans 42 years. The whole sample is further divided into 15 years of training sample (1978-1992), 6 years of validation sample(1993-1998) and 21 years of testing sample (1999-2019).One month Treasury-bill rate is also obtained to calculate stock excess returns. Green et al. (2017) identified 94 firm characteristics, such as bid-ask spread, cash flow to debt and percent accruals. I include the same collection of characteristics. There variables are obtained or constructed from CRSP or Compustat database. I excluded observations when firm characteristics are missing. In accordance with Gu et al. (2018), Freyberger et al. (2020) among others, what matters for the firm characteristics is its relevant rank among other firms in each month. At each month, I cross-sectionally rank each firm characteristic variable and transform the ranks falling into the range of [-0.5, 0.5].

To form a portfolio, I first forecast the one-month, or 20 trading days ahead returns. Using a zero investment strategy, I long the Stocks in which the predicted returns are among the top 90% Quantiles. And short stocks in which the predicted returns are in the lowest 10% quantiles. These stocks can be either evaluated or equal-weighted. To check the robustness of our strategies, I report the portfolio performance for longing

TABLE 15 PERFORMANCE OF EQUAL-WEIGHTED PORTFOLIOS

| | PCA | | | | sPCA | | | |
|---|---|---|---|---|---|---|---|---|
| Deciles | RET_PRED% | RET_REAL% | STD | SR | RET_PRED% | RET_REAL% | STD | SR |
| 1 | 0.188 | 0.122 | 0.083 | 0.051 | -0.613 | -0.125 | 0.076 | -0.057 |
| 2 | 0.619 | 0.678 | 0.073 | 0.322 | 0.091 | 0.492 | 0.066 | 0.257 |
| 3 | 0.842 | 0.785 | 0.069 | 0.392 | 0.455 | 0.777 | 0.061 | 0.440 |
| 4 | 1.013 | 0.765 | 0.066 | 0.404 | 0.743 | 0.823 | 0.058 | 0.492 |
| 5 | 1.160 | 0.931 | 0.061 | 0.529 | 1.000 | 1.001 | 0.056 | 0.618 |
| 6 | 1.298 | 1.005 | 0.056 | 0.625 | 1.247 | 1.104 | 0.057 | 0.675 |
| 7 | 1.437 | 1.263 | 0.053 | 0.827 | 1.502 | 1.183 | 0.059 | 0.693 |
| 8 | 1.584 | 1.245 | 0.051 | 0.844 | 1.789 | 1.248 | 0.060 | 0.718 |
| 9 | 1.762 | 1.380 | 0.053 | 0.903 | 2.152 | 1.364 | 0.065 | 0.728 |
| 10 | 2.109 | 1.413 | 0.054 | 0.911 | 2.849 | 1.722 | 0.075 | 0.800 |
| High - Low | 1.921 | 1.291 | 0.044 | 1.028 | 3.462 | 1.847 | 0.056 | 1.141 |

Note: Monthly test sample from 2000 to 2018. Stocks in deciles 1 to 10 are sorted according to the predicted returns. High-Low is the zero investment portfolio. RET_PRED%, RET_REAL%, STD, SR display the predicted returns, realized returns, standard deviation, and annualized Sharpe Ratio respectively.

stocks in each quantile. I expect to see the lower quantiles corresponds to the smaller realized returns, and the top quantiles correspond to larger realized returns. I will report the average realized returns, standard deviation, and the Sharp Ratio.

The out-of-sample performance of equal-weighted portfolios is reported in table 15 . The stocks are sorted into 10 deciles according to the predicted stocks returns. The realized average returns increase with the deciles generally. Decile 1 is supposed to give the lowest realized return, however, the number is still positive at 0.122% for PCA. sPCA on the other hand, generates a negative return of -0.125% for decile 1. Looking at decile 10, which is supposed to give the highest realized return, sPCA offers higher return of 1.722% than PCA (1.413%). Regarding the zero investment strategy, sPCA has an average realized return of 1.847% per month (22.2% annually) and an average volatility of 5.6%. The risk-adjusted performance is measured by the annualized Sharpe Ratio, which is 1.14. There is an 10.99% increase of SR for sPCA compared to PCA for equal-weighted portfolios.

TABLE 16 PERFORMANCE OF EQUAL WEIGHTED PORTFOLIOS

| | PCA | | | | sPCA | | | |
|---|---|---|---|---|---|---|---|---|
| Deciles | PRED_RET% | AVG_RET% | STD | SR | PRED_RET% | AVG_RET% | STD | SR |
| 1 | 0.188 | 0.229 | 0.079 | 0.101 | -0.614 | 0.047 | 0.072 | 0.023 |
| 2 | 0.619 | 0.736 | 0.070 | 0.365 | 0.090 | 0.614 | 0.062 | 0.341 |
| 3 | 0.842 | 0.781 | 0.065 | 0.416 | 0.455 | 0.857 | 0.058 | 0.511 |
| 4 | 1.014 | 0.764 | 0.062 | 0.428 | 0.742 | 0.848 | 0.055 | 0.533 |
| 5 | 1.159 | 0.903 | 0.058 | 0.543 | 0.999 | 1.007 | 0.054 | 0.648 |
| 6 | 1.298 | 0.986 | 0.053 | 0.646 | 1.246 | 1.086 | 0.055 | 0.689 |
| 7 | 1.437 | 1.220 | 0.051 | 0.836 | 1.502 | 1.116 | 0.056 | 0.685 |
| 8 | 1.583 | 1.158 | 0.049 | 0.815 | 1.788 | 1.163 | 0.059 | 0.688 |
| 9 | 1.761 | 1.274 | 0.051 | 0.871 | 2.149 | 1.220 | 0.064 | 0.662 |
| 10 | 2.089 | 1.289 | 0.051 | 0.877 | 2.822 | 1.545 | 0.074 | 0.724 |
| High - Low | 1.901 | 1.059 | 0.042 | 0.876 | 3.436 | 1.498 | 0.057 | 0.913 |

Note: Monthly test sample from 2000 to 2018. Stocks in deciles 1 to 10 are sorted according to the predicted returns. High-Low is the zero investment portfolio. RET_PRED%, RET_REAL%, STD, SR display the predicted returns, realized returns, standard deviation, and annualized Sharpe Ratio respectively.

Table16 reports value-weighted performance. The realized excess returns and the Sharpe Ratios are lower than equal-weighted portfolios. Again sPCA outperforms PCA in zero investment strategy with 4.2% higher SR. The cause of a decrease in the SR gap using value-weighted method is that sPCA portfolio has higher volatility in decile 10. But sPCA still produces lower realized returns in lower deciles and higher realized returns in higher deciles.

Interestingly, although sPCA outperforms PCA for equal-weighted and value-weighted zero investment portfolios, PCA generates higher Sharpe Ratios in decile 7-10 than sPCA. This indicates that sPCA is better in isolating stocks with low or negative returns.

From Table 17, the maximum drawdown, maximum loss and turnover of PCA and sPCA generated portfolios generally varied by less than 1%. Value-weighted sPCA had the highest Max Loss at 27.22% among all portfolios. Turnover steadily remained around 98% for each portfolio. The maximum drawdown and maximum loss are not considered high. In unexpected market shocks, PCA and sPCA portfolios are less likely

TABLE 17 DRAWDOWN AND TURNOVER

| | Equal Weighted | | Value Weighted | |
|---|---|---|---|---|
| | PCA | sPCA | PCA | sPCA |
| MaxDrawdown(%) | 14.119 | 15.267 | 13.977 | 13.900 |
| MaxLoss(%) | 23.392 | 22.873 | 21.566 | 27.221 |
| Turnover(%) | 98.562 | 97.939 | 98.640 | 98.066 |

to suffer huge loss and are considered to be safe strategies. The calculations of these two terms are as followed:

$$\text{MaxDD} = \max_{0 \le t_1 \le t_2 \le T} (Y_{t_1} - Y_{t_2})$$

where $Y_i$ is the cumulative log return from date 0 through $t$.

Turnover for positions with positive and negative weights for the SDF factor portfolio. It is defined as

$$\frac{1}{T} \sum_{t-1}^{T} \left( \sum_i \mid (1 + R_{P,t+1}) w_{i,t+1} - (1 + R_{i,t+1}) w_{i,t} \mid \right)$$

, where $w_{i,t}$ is the portfolio weight of stock $i$ at time $t$, and $R_{P,t+1} = \sum_i R_{i,t+1} w_{i,t}$ is the corresponding portfolio return. Long and short positions are calculated separately..

With the long and short strategy, sPCA generates higher returns and Sharpe ratio.It's Sharpe Ratio (0.92) is better than shallow learning method such as elastic net(0.39), generalized linear model with group lasso (0.79) from Gu et al. (2018). However it does not beat their 4 layers Neural Network.

**Summary of Empirical Findings**

1. *SPCA exhibits exceptional promise for forecasting bond risk premia.* SPCA outperforms PCA for all bonds in both moving and recursive windows. Using longer training set is better than shorter one. When R=240, rMSFE for SPCA ranges

from 0.67 to 0.81. SPCA combined with other models also outperforms PCA combined with other models.

2. *My Supervised device can be applied to other types of PCA.* SPCA on SIEVE-LS Covariance matrix is better than PCA on SIEVE-LS Covariance for all bond series and better than SPCA on standard covariance matrix for longer maturity bonds.

3. *Allowing for nonlinearities substantially reduced forecast errors.* I accommodate nonlinear predictive relationships via Single Index Model and Neural Networks. Single Indel Model reduced the MSFE for PCA by as low as 20%. It reduced the MSFE for SPCA less frequently by as low as 15%. Neural Network gives the biggest improvement.

4. *Generally the best performing method is SPCA_NN.* SPCA_NN generates rMSFE around 0.35 in two sub samples. It gives the lowest rMSFE using longer training period and recursive windows.

5. *The most successfully predicted macro variables are Housing Starts and M2.* Macro data is harder to predict. The out-of-sample $R^2$ using PCA in most cases are very low, except Housing Starts and M2. Despite supervision, SPCA still only predicts Housing Starts and M2 well, with 0.553 and 0.846 as their smallest rMSFE respectively. Nonlinearity still improves performance. SPCA_IDX reduced the rMSFE from 0.94 to 0.84 in one subsample of HOUST for instance. And again SPCA_NN is the best method.

## VII. Conclusion

I proposed a novel supervised version of PCA (SPCA). At the broadest level, my findings demonstrate that SPCA can help improve the predictions of bond risk premia and some Macro variables. I established the distribution of SPCA factors and forecast error. I found that $\widehat{F}_t - \bar{H}'F_t = O_p(\frac{1}{T}) + O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1-w}{\sqrt{N}})$. When $\sqrt{N}/T \to 0$, the dominating terms are the last two terms, as opposed to $O_p(\frac{1}{\sqrt{N}})$ with PCA. This indicates that

under conventional assumptions, SPCA can outperform PCA if you have a very large N. As for misspecified models, Monte Carlo simulations help us understand several situations when SPCA dominates PCA: the number of factors is large; factors important in predicting y are not that important in explaining X; N is small; u in $y_{t+h} = F_t\beta + u$ is close to 0. I track down the best performing model is to combine SPCA with Neural Network. It generates a rMSFE as low as 0.37 for bond risk premia. I also show that my supervised design is compatible with other types of PCA (Sieve-LS Covariance), and it is more powerful allowing for nonlinear predictive relationships.

## REFERENCES


ARTIS, M. J., BANERJEE, A. and MARCELLINO, M. (2005). Factor forecasts for the uk. *Journal of forecasting*, **24** (4), 279–298.

AZOFF, E. M. (1994). *Neural network time series forecasting of financial markets*. John Wiley & Sons, Inc.

BABIKIR, A. and MWAMBI, H. (2016). Evaluating the combined forecasts of the dynamic factor model and the artificial neural network model using linear and nonlinear combining methods. *Empirical Economics*, **51** (4), 1541–1556.

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

— and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70** (1), 191–221.

— and — (2006). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, **131** (1-2), 507–537.

— and — (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, **146** (2), 304–317.

—, — *et al.* (2008). Large Dimensional Factor Analysis. *Foundations and Trends® in Econometrics*, **3** (2), 89–163.

BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101** (473), 119–137.

BALDI, P. and HORNIK, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, **2** (1), 53–58.

BANERJEE, A. and MARCELLINO, M. (2006). Are there any reliable leading indicators for us inflation and gdp growth? *International Journal of Forecasting*, **22** (1), 137–151.

BARSHAN, E., GHODSI, A., AZIMIFAR, Z. and JAHROMI, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, **44** (7), 1357–1371.

CHAN, N. H. and LIU, W. W. (2017). Modeling and forecasting online auction prices: a semiparametric regression analysis. *Journal of Forecasting*, **36** (2), 156–164.

COCHRANE, J. H. and PIAZZESI, M. (2005). Bond risk premia. *American Economic Review*, **95** (1), 138–160.

CRISTADORO, R., FORNI, M., REICHLIN, L. and VERONESE, G. (2005). A core inflation indicator for the euro area. *Journal of Money, credit, and Banking*, **37** (3), 539–560.

DEN REIJER, A. H. *et al.* (2005). *Forecasting Dutch GDP using large scale factor models*. Tech. rep., Netherlands Central Bank, Research Department.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Com paring predictive accu racy. *Journal of Business and Economic Statistics*, **13** (3), 253–263.

EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R. *et al.* (2004). Least angle regression. *The Annals of statistics*, **32** (2), 407–499.

FAN, J., KE, Y. and LIAO, Y. (2016). Augmented factor models with applications to validating market risk factors and forecasting bond risk premia. *Available at SSRN 2753404*.

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, **82** (4), 540–554.

—, —, — and — (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, **100** (471), 830–840.

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

GU, S., KELLY, B. and XIU, D. (2018). *Empirical asset pricing via machine learning*. Tech. rep., National Bureau of Economic Research.

HADI, A. S. and LING, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, **52** (1), 15–19.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12** (1), 55–67.

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, **2** (5), 359–366.

ICHIMURA, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, **58** (1-2), 71–120.

JURADO, K., LUDVIGSON, S. C. and NG, S. (2015). Measuring uncertainty. *American Economic Review*, **105** (3), 1177–1216.

KIM, H. H. and SWANSON, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, **178**, 352–367.

— and — (2016). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*.

— and — (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, **34** (2), 339–354.

LUDVIGSON, S. C. and NG, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies*, **22** (12), 5027–5067.

MCALEER, M. and DA VEIGA, B. (2008). Single-index and portfolio models for forecasting value-at-risk thresholds. *Journal of Forecasting*, **27** (3), 217–235.

MCCRACKEN, M. W. and NG, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, **34** (4), 574–589.

SCHUMACHER, C. and BREITUNG, J. (2008). Real-time forecasting of german gdp based on a large factor model with monthly and quarterly data. *International Journal of Forecasting*, **24** (3), 386–398.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *Journal Of Computational And Graphical Statistics*, **22** (2), 231–245.

STOCK, J. H. and WATSON, M. (2011). Dynamic factor models. *Oxford Handbook on Economic Forecasting*.

— and WATSON, M. W. (1998). *Diffusion indexes*. Tech. rep., National bureau of economic research.

— and — (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, **97** (460), 1167–1179.

— and — (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, **20** (2), 147–162.

— and — (2006). Forecasting with many predictors. *Handbook of economic forecasting*, **1**, 515–554.

— and — (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, **30** (4), 481–493.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372.

WERON, R. and MISIOREK, A. (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International journal of forecasting*, **24** (4), 744–763.

WOLD, S., SJÖSTRÖM, M. and ERIKSSON, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, **58** (2), 109–130.

## Appendix

### APPENDIX A: ESTIMATION AND ASSUMPTIONS

I follow the assumptions used in Bai (2002, 2003 and 2006) for estimating factors and their loadings.

Assumption A - Factors: $E\left\|F_t^0\right\|^4 \leq M < \infty$ and $T^{-1}\sum_{t=1}^{T} F_t^0 F_t^{0\prime} \xrightarrow{p} \Sigma_F$ for some $r \times r$ positive definite matrix $\Sigma_F$.

Assumption B - Factor Loadings: $\|\lambda_i\| \leq \bar{\lambda} < \infty$, and $\left\|\Lambda^{0\prime}\Lambda^0/N - \Sigma_N\right\| \to 0$ for some $r \times r$ positive definite matrix $\Sigma_A$.

Assumption C− Time and Cross-Section Dependence and Heteroskedasticity: There exists a positive constant $M < \infty$ such that for all $N$ and $T$ :

1. $E(e_{it}) = 0, E|e_{it}|^8 \leq M_{i=1}$
2. $E(e_s^t e_t/N) = E\left(N^{-1}\sum_{i=1}^{N} e_{is}e_{it}\right) = \gamma_N(s,t), |\gamma_N(s,s)| \leq M$ for all $s$,

$$T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$$

3. $E\left(e_{it}e_{jt}\right) = \tau_{ij,t}$ with $\left|\tau_{ij,t}\right| \leq \left|\tau_{ij}\right|$ for some $\tau_{ij}$ and for all $t$. In addition,

$$N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\tau_{ij}\right| \leq M$$

4. $E\left(e_{it}e_{js}\right) = \tau_{ij,ts}$ and $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{t=1}^{T}\sum_{t=1}^{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\left|\tau_{ij,ts}\right| \leq M$

5. For every $(t,s), E\left|N^{-1/2}\sum_{i=1}^{N}\left[e_{is}e_{it} - E\left(e_{is}e_{it}\right)\right]\right|^4 \leq M$

Assumption D-Weak dependence between factors and idiosyncratic errors:

$$E\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^0 e_{it}\right\|^2\right) \leq M$$

Assumption A allows factors to be dynamic. Assumption B ensures that each factor has a nontrivial contribution to the variance of Xt . Assumption C allows for Heteroskedasticities and limited dependence in both the time and cross-section dimensions. With Assumptions A-C, Assumption D is implied.

Assumption E-Weak Dependence: There exists $M < \infty$ such that for all $T$ and $N$, and for every $t \leq T$ and every $i \leq N$ :

1. $\sum_{s=1}^{T} |\gamma_N(s,t)| \leq M$
2. $\sum_{k=1}^{N} |\tau_{ki}| \leq M$

Assumption F - Moments and Central Limit Theorem: There exists an $M < \infty$ such that for all $N$ and $T$ :

1. for each $t$
$$E \left\| \frac{1}{\sqrt{NT}} \sum_{s=1}^{T} \sum_{k=1}^{N} F_s^0 \left[ e_{ks} e_{kt} - E\left( e_{ks} e_{kt} \right) \right] \right\|^2 \leq M$$
2. the $r \times r$ matrix satisfies
$$E \left\| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{k=1}^{N} F_t^0 \lambda_k^{0\prime} e_{kt} \right\|^2 \leq M$$
3. for each $t$, as $N \to \infty$
$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i^0 e_{it} \xrightarrow{d} N\left(0, \Gamma_t\right)$$
where $\Gamma_t = \lim_{N\to\infty} (1/N) \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i^0 \lambda_j^{0\prime} E\left( e_{it} e_{jt} \right)$
4. for each $i$, as $T \to \infty$,
$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} F_t^0 e_{it} \xrightarrow{d} N\left(0, \Phi_i\right)$$
where $\Phi_i = \text{plim}_{T\to\infty} (1/T) \sum_{s=1}^{T} \sum_{t=1}^{T} E\left[ F_t^0 F_s^{0\prime} e_{is} e_{it} \right]$

Assumption G: The eigenvalues of the $r \times r$ matrix $\left( \Sigma_\Lambda \cdot \Sigma_F \right)$ are distinct.

Proposition 1: Under Assumptions $A - D$ and $G$,

$$\text{plim}_{T,N\to\infty} \frac{\widetilde{F}' F^0}{T} = Q$$

The matrix $Q$ is invertible and is given by $Q = V^{1/2}\Upsilon'\Sigma_A^{-1/2}$, where $V = \text{diag}(v_1, v_2, \ldots, v_r), v_1 > v_2 > \cdots > v_r > 0$ are the eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, and $r$ is the corresponding eigenvector matrix such that $\Upsilon'\Upsilon = I_r$

Assumption E strengthens C2 and C3. Assumption F is not stringent.

The same as PCA, $F^0$ and $\Lambda^0$ can be estimated up to an invertible $r \times r$ matrix transformation. In the appendix, I used two such matrices. For $\widetilde{F}_t, t = 1, ..., T-h$ $\tilde{F}_t$ is an estimator of $HF_t^0$. To simplify calculation, I define another invertible matrix $\bar{H}$ so that $\widehat{F}_t, t = T-h+1, ..., T$ is an estimator of $\bar{H}F_t^0$ and $\widehat{\Lambda}$ is an estimator of $\Lambda^0 (\bar{H}')^{-1}$. Note when I do predictions of $y_{T+h}$, $\widehat{\Lambda}$ is used in OLS to get $\widehat{F}_T$. I only need to ensure as both of these estimators share the same invertible matrix $\bar{H}$ so that $\widehat{F}\widehat{\Lambda}'$ is an estimator of $F^0\Lambda^{0\prime}$. Moreover, $F^0\bar{H}$ and F give the same predicted value of y.