Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operations Research

Scaled PCA: Evaluating Different Scalars

Koen van Doorn (537135)

Supervisor:	Daan Opschoor
Second assessor:	Dr. Mikhail Zhelonkin
Date final version:	2nd July 2023

Abstract

Principal component analysis (PCA) is a widely used statistical method for dimensionality reduction. However, when used for forecasting, its equal weighting of predictors may overlook important information about the target variable. To address this limitation, scaled principal component analysis (sPCA) was introduced. sPCA incorporates target variable information into the dimensionality reduction process by scaling predictors with their respective predictive slopes of the target variable. This paper extends the application of sPCA by introducing the t-statistic and quantile regression coefficient as a scalar. The aim of the research is to evaluate the use of the slope (sPCA-slope), the t-statistic (sPCA-tstat), and the quantile regression coefficient (sPCA-quantile) as scalars to enhance the predictive power and robustness of the sPCA method. A simulation study compares the performance of the PCA and three sPCA techniques in different scenarios. The results highlight the accuracy of sPCA-stat in normal circumstances and the robustness of the sPCA-quantile to a small number of extreme outliers. On the other hand, sPCA-slope demonstrates consistent results in various conditions and robustness against larger errors. Empirical analysis using U.S. macroeconomic variables confirms the superior predictive power of sPCA techniques compared to PCA, with sPCA-tstat performing remarkable well in both in-sample and out-of-sample forecasting. Overall, this research provides valuable insights for practitioners in selecting the most appropriate sPCA method for different circumstances.

Contents

1 Introduction

Principal component analysis (PCA) is a well-established statistical method employed to reduce the dimensionality of datasets. Its objective is to transform high-dimensional data into lower-dimensional data while preserving the information contained in the original dataset. PCA is the oldest and most widely used dimension-reduction technique, introduced by (Pearson, 1901). PCA can be a useful tool for forecasting in various fields, including finance, marketing, and economics. The works by (Stock and Watson, 2002), (Stock and Watson, 2006), (Huang et al., 2019), and (Skittides and Früh, 2014) are examples of studies where PCA is used for forecasting. Although PCA can be useful as a dimensionality reduction technique, it might not be the most appropriate technique for forecasting. PCA puts equal weight on all predictors and therefore neglects the information of the target variable. Addressing this limitation, (Huang et al., 2022) proposed a technique called scaled principal component analysis (sPCA), which incorporates the information of the target variable into the dimensionality reduction process. This technique scales each predictor with its predictive slope on the target to be forecasted and then performs PCA on the transformed predictor set. By using the slope as a scalar, sPCA puts more weight on the predictors with stronger forecasting power. (Huang et al., 2022) demonstrate, through a simulation and an empirical study, that the sPCA outperforms the PCA forecast technique. The works by (He et al., 2021) [1], (Ma et al., 2022), and (Wang et al., 2022) explore the application of the sPCA forecast approach and provide evidence of its effectiveness in different fields.

This paper extends the application of sPCA by introducing new scalars and therefore new sPCA techniques. Literature shows that the OLS estimator is prone to outliers and by incorporating different scalars, this research aims to increase the predictive power and robustness of the sPCA method.

Firstly, the t-statistic of the predictive slope on the target variable is introduced as a scalar. The use of the t-statistic as a scalar can offer an advantage by incorporating statistical significance. While the regression slope provides information about the magnitude and direction of the relationship between predictor and target variable, the t-statistic considers the precision and reliability of this relationship. The t-statistic puts weight on predictors that have a higher level of significance and as a result potentially improves the overall predictive performance of the sPCA method. The works by (Clark and McCracken, 2007) and (Magee and Veall, 1991) demonstrate the use of t-statistics in aspects of forecasting.

Secondly, the quantile (median) regression coefficient of the target variable on the predictor is introduced as a scalar. Quantile regression is a type of regression analysis. Whereas OLS estimates the conditional mean of the target variable across values of the predictors, quantile regression estimates the conditional quantile of the target variable. In this paper, we exclusively use the median as a quantile, which represents the middle value of a dataset. Key references to quantile regression are (Koenker and Bassett Jr, 1978) and (Koenker and Hallock, 2001). Quantile (median) regression provides robustness against outliers and asymmetric data distributions, as it focuses on the centre of a distribution and therefore is less affected by extreme values. Incorporating the median quantile regression coefficient as a scalar in a sPCA method potentially provides a more robust forecasting technique. (Gaglianone and Lima, 2012), (Bremnes, 2004), (Ma and Pohlman, 2008), (Nielsen et al., 2006), (Liu et al., 2015), and (Taillardat et al., 2016) are researches that demonstrate an effective use of quantile regression for forecasting.

The sPCA forecasting approach consists of two steps. Firstly, we derive the scalars for the predictor set. This research evaluates the performance of three scalars: the slope, the t-statistic, and the quantile (median) regression coefficient. Secondly, each predictor is scaled with its corresponding scalar and PCA

[1] First paper by (Huang et al., 2022) about scaled PCA is published in 2019.

is applied to the adjusted predictor set. The derived principal components are used to create forecasts of the target variable. This research uses different measures to evaluate the prediction performance of the PCA and the three different sPCA techniques.

Theoretically, we perform a simulation study to evaluate the prediction power of the PCA and the sPCA approaches (sPCA-slope, sPCA-tstat, and sPCA-quantile). This study considers a partially relevant latent factor framework with strong and weak factors, representing relevant and irrelevant predictors, respectively. The simulation covers three different scenarios. Firstly, the standard error scenario where errors are generated from a normal distribution. Secondly, the larger error scenario, where errors are generated from a heavier tailed distribution, and lastly, the extreme outlier scenario where extreme outliers are added to the data. The aim of this study is to investigate the predictive power and the robustness of the sPCA techniques in these distinct scenarios. The findings indicate that sPCA-tstat exhibits superior forecasting accuracy but is less robust against larger errors and extreme outliers. In contrast, the sPCA-quantile proves to be the most robust against extreme outliers and the sPCA-slope method consistently performs well in the three different scenarios and shows robustness against larger errors.

Empirically, we apply the sPCA techniques to forecast U.S. inflation, industrial production, unemployment, and stock market returns (S&P 500), using 123 macroeconomic variables as predictors. We employ PCA and the sPCA techniques in an in-sample and out-of-sample environment and evaluate their forecast accuracy. The results demonstrate that all three sPCA approaches have superior predictive power compared to PCA. Notably, among the sPCA techniques, sPCA-tstat exhibits greater predictive power for both in-sample and out-of-sample forecasting.

In summary, this research highlights the characteristics of three different sPCA techniques: sPCA-slope, sPCA-stat, and sPCA-quantile. The findings provide valuable insights for practitioners in selecting the most appropriate method for specific circumstances.

The paper is structured as follows. Section 2 provides a detailed explanation of the methodology employed, describing the methods and tools used in the research. Section 3 presents the simulation study and its results, while Section 4 focuses on the empirical study, discussing the data, in-sample and out-of-sample results. The paper ends with Section 5, which summarizes the findings and provides a conclusion.

2 Methodology

In this section, we present the methods and techniques to conduct the three sPCA forecast approaches: sPCA-slope, sPCA-tstat, and sPCA-quantile. We select a number of target variables denoted by y_{t+h}, where h indicates the forecast horizon. In this research, the focus is exclusively on the one-step-ahead forecast $h = 1$. To create forecasts, we select a predictor set $(X_{1,t}, ..., X_{N,t})$ where N indicates the number of predictors and $t = 1, ..., T$ the number of observations.

In this paper, we consider a similar latent factor model structure as (Huang et al., 2022). For the target variable y_{t+h} and the N predictors X_t, this is given by

$$\begin{aligned} X_{i,t} &= \lambda_i' f_t + e_{i,t} \\ &= \phi_i' g_t + \psi_i' h_t + e_{i,t}, \\ y_{t+h} &= \alpha + \beta' g_t + \epsilon_{t+h}. \end{aligned}$$

In this structure the factors $f_t = (f_{1,t}, ..., f_{N,t}) = (g_t', h_t')'$ are divided into the relevant factors g_t an the irrelevant factors h_t. The relevant factors imply predictive power and are associated with the target

variable y_{t+h}. In contrast, irrelevant factors lack any meaningful connection with the target variable. $\lambda_i = (\phi_i', \psi_i')$ denote the loadings of the factors.

2.1 Principal Component Analysis (PCA)

An evident method to derive the factors is PCA, a dimension reduction technique that transforms high-dimensional data into lower-dimensional while preserving the most important patterns. To perform PCA on a predictor set, we consider the sample covariance matrix of $X_t = (X_{1,t}, ..., X_{N,t})'$ given by

$$\hat{V} = \frac{1}{T} \sum_{t=1}^{T} X_t^* (X_t^*)',$$

where $X_t^* = (X_{1,t}^*, ..., X_{N,t}^*)$. In this case $X_{i,t}^* = X_{i,t} - \bar{X}_i$ and $\bar{X}_i = \frac{1}{T} \sum_{t=1}^{T} X_{i,t}$. Note that we standardize the predictors before performing PCA. This helps to stabilize the variances of the predictors and it allows for a meaningful comparison by putting them on a common scale. PCA finds those linear combinations of X_t that are uncorrelated and have maximum variance which can be elaborated by the following steps:

1. The first principal component (PC) is the linear combination $f_{1,t} = a_1' X_t$, maximizing $\text{Var}(f_{1,t}) = a_1' \hat{V} a_1$ subject to $a_1' a_1 = \sum_{j=1}^{N} a_{1j}^2 = 1$.
2. The j-th PC $f_{j,t} = a_j' X_t$ maximizes $\text{Var}(f_{1,t}) = a_j' \hat{V} a_j$ subject to the constraints $a_j' a_j = 1$ and $\text{Cov}(f_{j,t}, f_{i,t}) = a_j' \hat{V} a_i = 0$ for $i = 1, ..., j-1$.

In order to find the first PC, we solve the maximization problem $\max_{a_1} a_1' \hat{V} a_1$ such that $a_1' a_1 = 1$. We solve this by forming the Langragian $L = a_1' \hat{V} a_1 - l(a_1' a_1 - 1)$, where l is the Lagrange multiplier. The derivative with respect to a_1 is equal to $2\hat{V}a_1$ - $2la_1$. Setting this equal to zero gives the first order condition which can be rewritten as $\hat{V}a_1 = la_1$. This shows that the solution a_1 is an eigenvector of $\hat{V}$. Now let $(\lambda_1, e_1), ..., (\lambda_N, e_N)$ be the eigenvalue-eigenvector pairs of $\hat{V}$ in order that $\lambda_1 \geq ... \geq \lambda_N > 0$. Given that the eigenvalues are the solution to the maximization problems, we can write the j-th principal component of X_t as

$$f_{j,t} = e_j' X_t = e_{j,1} X_{1,t} + e_{j,2} X_{2,t} + ... + e_{j,N} X_{N,t} \quad \text{for } j = 1, ..., N.$$

This gives, using PCA, that the variance of the factor is given by $\text{Var}(f_{j,t}) = e_j' \hat{V} e_j = \lambda_j$ for $j = 1, ..., N$ and the covariance of the factors are given by $\text{Cov}(f_{j,t}, f_{i,t}) = e_j' \hat{V} e_i = 0$ for all $i \neq j$. We denote the fraction $\frac{\lambda_j}{\sum_{i=1}^{N} \lambda_i}$ as the fraction of the total 'variance explained' in X_t by the j-th factor. The variance explained is a relevant measure for evaluating PCA techniques as it quantifies the proportion of data variability that can be accounted for by the extracted components. A high variance explained indicates that the PCA technique can capture a larger proportion of the total variability in the data. This suggests a more effective representation of the underlying patterns or structure.

2.2 Scaled PCA (sPCA)

When the model is specified as the latent factor model, PCA can overlook the influence of the target variable on the predictors. PCA is a technique that distributes equal forecasting power to each predictor, regardless of the information of the target variable. sPCA is a technique that addresses this issue by

multiplying each predictor by a scalar that puts more weight on predictors with more forecasting power. In this paper, we investigate the performance of three different scalars: the regression coefficient (slope), the t-statistic, and the median quantile regression coefficient. These sections explain the three different types of scalars and how to use the principal components to forecast the target variable.

2.2.1 Slope Scalar

The slope is the regression coefficient of the target variable on a predictor. The slope can be used as a scalar for sPCA because it serves as an indicator of the predictor's predictive power. It reveals the magnitude and direction of the predictor on the target variable. A larger absolute coefficient implies a stronger influence, indicating that small changes in the predictor's value can result in significant changes of the target variable. Furthermore, the sign of the coefficient provides information about the direction of the relationship. Positive coefficients indicate a positive association, while negative coefficients suggest an inverse relationship. This sPCA technique is explored by (Huang et al., 2022) and their findings demonstrate that incorporating the regression coefficient as a scalar improves predictive power. We employ sPCA with the slope as a scalar (sPCA-slope) to forecast the target variable. This approach involves incorporating the following steps into the prediction process:

1. Firstly, we derive the regression coefficient of the target variable on each predictor. y_{t+h} denotes the target variable ($h = 1$) and $(X_{1,t}, ..., X_{N,t})$ the set of standardized predictors. We obtain estimates of the slopes by regressing the target variable on each predictor i, $y_{t+h} = \alpha_i + \beta_i X_{i,t} + \epsilon_{t+h}$, where α_i and β_i represent the intercept and slope coefficients, respectively. Ordinary least squares (OLS) is used to obtain the estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$.

2. Secondly, we take the estimated regression coefficient, $\hat{\beta}_i$, and multiply it by the corresponding predictor $X_{i,t}$ for every $i = 1, ..., N$. This results in a set of scaled predictors $(\hat{\beta}_1 X_{1,t}, \hat{\beta}_2 X_{2,t}, ..., \hat{\beta}_N X_{N,t})$. We then apply PCA to this set of scaled predictors to extract r relevant factors. These factors are used to predict the target variable. The sPCA-slope forecast for y_{t+h} is given by

$$\tilde{y}_{t+h}^{sPCA_s} = \hat{\gamma}^{sPCA_s} + (\hat{\pi}^{sPCA_s})' \hat{g}_t^{sPCA_s},$$

   where we select $\hat{g}_t^{sPCA_s}$ as the relevant factors that are associated with the target variables and $(\hat{\gamma}^{sPCA_s}, (\hat{\pi}^{sPCA_s})')'$ are the OLS estimates.

2.2.2 T-statistic Scalar

The t-statistic of the regression coefficient provides valuable insights into the predictive power of a predictor. It measures the strength of the association between the predictor and the target variable, considering both the magnitude of the coefficient and its significance. A higher absolute t-statistic indicates a stronger and more significant impact of the predictor on the target variable, suggesting greater predictive power. By considering the t-statistic, we can quantitatively assess the significance of the relationship and prioritize predictors with stronger predictive capabilities in modeling and forecasting tasks. We employ sPCA with the t-statistic as a scalar (sPCA-tstat) to forecast the target variable. This approach involves incorporating the following steps into the prediction process:

1. Firstly, we derive the t-statistic of the regression coefficient of the target variable on each predictor. The estimation of $\hat{\beta}_i$ follows a similar approach as described in step 1 of the sPCA-slope technique

(Section 2.2.1). The t-statistic is denoted by $t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$, where the denominator indicates the standard error of the regression coefficient.

This research uses the Newey-West estimator to estimate the standard error of the regression coefficient. This estimator is particularly useful when the standard assumptions of the residuals, such as homoskedasticity and no autocorrelation, do not hold. By accounting for heteroskedasticity and/or autocorrelation in the residuals, the Newey-West estimator enhances the accuracy and reliability of the standard errors of the regression coefficients. It was devised by (Newey and West, 1986) and there are several lateral variants by (Andrews, 1991), (Newey and West, 1994) and (Smith, 2005). In OLS regression, the estimation of regression coefficients in the model $y = X\beta + \epsilon$ is given by the formula $\hat{\beta} = (X'X)^{-1}X'y$. The covariance matrix of the coefficient estimates is given by $\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$. It is important to note that under the assumption of homoskedasticity and no autocorrelation, the value σ^2 represents the common variance of the residuals. The covariance matrix is denoted by

$$\text{Cov}(\hat{\beta}) = (X'X)^{-1}X'SX(X'X)^{-1} \text{ , where } S = \sigma^2 I.$$

The Newey-West method follows a similar approach but with a different calculation for $X'SX$. The resulting standard errors are known as Heteroskedasticity and Autocorrelation Corrected (HAC) standard errors. In the presence of autocorrelation with lags up to $l > 0$, the term $X'SX$ is computed as follows:

$$X'SX = \frac{n}{n-k}\sum_{i=1}^{n} e_i^2 X_i' X_i + \frac{n}{n-k}\sum_{i=1}^{l}(1 - \frac{i}{l+1})\sum_{j=i+1}^{n} e_j e_{j-i}(X_j' X_{j-i} + X_{j-i}' X_j),$$

where n represents the number of observations, k is the number of variables, e_i denotes the i-th residual, and X_i represents the i-th row in the design matrix X. The first term in the formula corresponds to the value of $X'SX$ when there is no autocorrelation. The second term incorporates the Newey-West method to handle autocorrelation up to a lag of l. It assumes that lags beyond l can be disregarded. This research uses exclusively autocorrelation with lags up to $l = 1$.

2. Secondly, we take the t-statistic $t_{\hat{\beta}_i}$ and multiply it by its corresponding predictor $X_{i,t}$. This results in a set of scaled predictors $(t_{\hat{\beta}_1}X_{1,t}, ..., t_{\hat{\beta}_N}X_{N,t})$. Apply PCA to this set of scaled predictors to extract r relevant factors and use them to predict the target variable. The sPCA-tstat forecast for y_{t+h} is given by

$$\tilde{y}_{t+h}^{sPCA_t} = \hat{\gamma}^{sPCA_t} + (\hat{\pi}^{sPCA_t})'\hat{g}_t^{sPCA_t},$$

where we again select $\hat{g}_t^{sPCA_t}$ as the relevant factors and $(\hat{\gamma}^{sPCA_t}, (\hat{\pi}^{sPCA_t})')'$ are the OLS estimates.

2.2.3 Quantile Regression Scalar

Quantile regression is a statistical method used to estimate the relationship between variables at different quantiles of the conditional distribution. It extends the concept of traditional regression, which focuses on estimating the conditional mean. In quantile regression, instead of estimating the conditional mean, we estimate the conditional quantiles. The quantiles represent specific points in the distribution that divide the data into equal-sized portions. The median, which corresponds to the 50th percentile, is a commonly used quantile that we exclusively use in this paper. A significant coefficient suggests that the

predictor carries predictive power for predicting the target variable. This coefficient is more robust to outliers, making it useful in capturing predictive power. Section 2.3 provides a detailed description of quantile regression. We employ sPCA with the median quantile regression coefficient as a scalar (sPCA-quantile) to forecast the target variable. This approach involves incorporating the following steps into the prediction process:

1. Firstly, we derive the quantile regression coefficients of the target variable on each predictor using the median as the quantile. The quantile regression model can be expressed as $q_{y_{t+h}}(\tau) = \alpha_i(\tau) + \beta_i(\tau)X_{i,t}$, where we employ the median quantile with $\tau = 0.5$. Here, $q_{y_{t+h}}$ denotes the conditional median of the target variable.

2. Secondly, we take the estimated quantile regression coefficient, denoted as $\hat{\beta}_i(\tau)$, and multiply it by the corresponding predictor $X_{i,t}$ for every $i = 1, ...N$. This results in a set of scaled predictors $(\hat{\beta}_1(\tau)X_{1,t}, \hat{\beta}_2(\tau)X_{2,t}, ..., \hat{\beta}_N(\tau)X_{N,t})$. We then apply PCA to this set of scaled predictors to extract r relevant factors and use these to predict the target variable. The sPCA-quantile forecast for y_{t+h} is given by

$$\tilde{y}_{t+h}^{sPCA_q} = \hat{\gamma}^{sPCA_q} + (\hat{\pi}^{sPCA_q})'\hat{g}_t^{sPCA_q},$$

where we again select $\hat{g}_t^{sPCA_q}$ as the relevant factors and $(\hat{\gamma}^{sPCA_q}, (\hat{\pi}^{sPCA_q})')'$ are the OLS estimates.

2.3 Quantile Regression

Quantile regression is a technique that differs from traditional OLS regression. While OLS regression estimates the conditional mean of the target variable given the predictor variables, quantile regression estimates the conditional quantile of the target variable. A quantile is a specific value that divides a dataset into equal-sized subsets. It represents the threshold below which a certain proportion of the data falls. For example, when the data of U.S. inflation lies within the τ-th quantile, it means that it is higher than τ of the observations and lower than $(1 - \tau)$ of the observations. A common quantile is the median (exclusively used in this research), setting $\tau = 0.5$.

Quantile regression is useful when the assumptions of linear regression are not satisfied or when there is interest in understanding the relationship between predictors and specific quantiles of the dependent variable. An advantage of quantile regression relative to OLS regression is that the quantile regression estimates are more robust against outliers. Figure 3 in the Appendix displays a scatterplot of a variable y against x with the presence of an outlier. It illustrates the difference between OLS and the more robust median quantile regression method.

Suppose we have a dependent variable y_t and a set of predictors X_t for $t = 1, ..., T$, where X_t represents a vector of k predictors. Unlike OLS, which minimizes the sum of squared residuals, the quantile regression estimator minimizes the sum of absolute residuals. The regression model is specified as the equation $y_t = X_t\beta + \epsilon_t$, where β represents the vector of regression coefficients and ϵ_t represents the error term or residual for observation t. The τ-th sample quantile, $0 < \tau < 1$, is defined as any solution to the minimization problem described by (Koenker and Bassett Jr, 1978):

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^k} [\sum_{t \in t: y_t \geq X_t\beta} \tau|y_t - X_t\beta| + \sum_{t \in t: y_t < X_t\hat{\beta}} |(1-\tau)y_t - X_t\beta|].$$

The regression quantile minimization problem is equivalent to the linear program (P):

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^k} [\tau \iota' k^+ + (1-\tau)\iota' k^-],$$

where k^+ and k^- equals $|y_t - X_t\beta|$ when respectively $y_t \geq X_t\beta$ and $y_t < X_t\beta$, subject to

$$\begin{aligned} y &= X\beta + k^+ - k^-, \\ (\beta, k^+, k^-) &\in \mathbb{R}^k \text{ x } \mathbb{R}_+^{2T}, \end{aligned}$$

where $\iota = (1, 1, ..., 1)$, a T vector of ones. The dual linear program (D) is defined as:

$$\hat{\delta} = \max_{\delta} [y'\delta]$$

subject to

$$\begin{aligned} X'\delta &= 0 \\ \delta &\in [\tau - 1, \tau]^T, \end{aligned}$$

where $[\tau - 1, \tau]^T$ denotes the T-fold Cartesian product of the closed interval $[\tau - 1, \tau]$. In other words, it denotes the set of T-tuples, where each element of the tuple comes from the interval $[\tau - 1, \tau]$. The superscript T indicates the number of times the interval is repeated in the Cartesian product. For example, consider $\tau = 0.5$ and $T = 3$. The expression $[\tau - 1, \tau]^T$ would be: $[0.5 - 1, 0.5]^3 = [-0.5, 0.5] \times [-0.5, 0.5] \times [-0.5, 0.5]$ Expanding this Cartesian product, we obtain:

$$\begin{aligned} [-0.5, 0.5]^3 = \{&(-0.5, -0.5, -0.5), (-0.5, -0.5, 0.5), \\ &(-0.5, 0.5, -0.5), (-0.5, 0.5, 0.5), \\ &(0.5, -0.5, -0.5), (0.5, -0.5, 0.5), \\ &(0.5, 0.5, -0.5), (0.5, 0.5, 0.5)\}. \end{aligned} \tag{1}$$

(Koenker and Bassett Jr, 1978) demonstrate that it is convenient to make a small adjustment to the dual linear program. Therefore, we modify the dual variables to $\Delta = \delta + 1 - \tau$. This gives the adjusted dual linear program (D'):

$$\hat{\Delta} = \max_{\Delta} [y'\Delta]$$

subject to

$$\begin{aligned} X'\Delta &= (1-\tau)X'\iota \\ \Delta &\in [0, 1]^T. \end{aligned}$$

The modified dual formulation proves to be convenient for computational purposes. This paper uses standard linear programming algorithms to solve the adjusted dual linear program (D') and obtain the quantile regression estimators [2]. This paper focuses solely on employing the median quantile regression model, $\tau = 0.5$. Note that by adopting the median, the minimization model becomes more concise, as

[2]The MATLAB code for quantile regression referenced is developed by Roger Koenker and sourced from the website http://www.econ.uiuc.edu/roger/research/rq/rq.html,

indicated by $\min_{\beta \in \mathbb{R}^k} [\sum_t \tau |y_t - X_t \beta|]$.

3 Simulation Study

In this section, we perform Monte Carlo experiments to assess and compare the forecasting accuracies of the four distinct PCA techniques. Specifically, we consider the following approaches: traditional principal component analysis (PCA), scaled PCA with the slope as a scalar (sPCA-slope), scaled PCA with the t-statistic as a scalar (sPCA-tstat), and scaled PCA with the median quantile regression coefficient as a scalar (sPCA-quantile). In this study, the forecast accuracies of one-step-ahead ($h = 1$) forecasts are compared.

The simulation is conducted based on a two-latent-factor model design where one factor is specifically relevant to the target variable. See Section 2 for a description of the latent-factor model structure. The predictors are denoted by $X_{i,t} = \lambda_i' f_t + e_{i,t} = \phi_i' g_t + \psi_i' h_t + e_{i,t}$, for $i = 1, ..., N$ and $t = 1, ..., T$, where g_t and h_t indicate the relevant and irrelevant factor, respectively. The two factors are independently normally distributed with zero mean and unit variance, that is $g_t \sim \mathcal{N}(0, 1)$ and $h_t \sim \mathcal{N}(0, 1)$. The idiosyncratic noises $e_{i,t}$ are generated from a distribution with zero mean and standard deviation σ_i and are independent across predictors and over time. In this study, we vary the distribution of the idiosyncratic noises to explore the effects on forecast accuracy and check for robustness. The exact distributions are given and explained in the following sections. σ_i for $i = 1, ..., N$ are drawn independently from a uniform distribution with support [0,1], that is $\sigma_i \sim \mathcal{U}[0, 1]$.

In this paper, we simulate scenarios with strong and weak factors. We perform this by assigning different values to parameters ϕ_i and ψ_i. The strength of factors is indicative of their predictive power, with stronger factors demonstrating higher predictive capability compared to weaker factors. To simulate the strong factors, we randomly sample ϕ_i and ψ_i from an independent uniform distribution with support [0, 1], that is $\phi_i \sim \mathcal{U}[0, 1]$ and $\psi_i \sim \mathcal{U}[0, 1]$. To simulate the weak factors, we set ϕ_i and ψ_i to be zero. In this simulation study, we explore various scenarios by varying the number of strong factors, represented by the parameter n, along with the weak factors.

The target variable is expressed as the sum of the relevant factor and the disturbances, that is $y_{t+1} = g_t + \epsilon_{t+1}$. The factor g_t is incorporated into the model to capture its impact on the target variable. The disturbances ϵ_{t+1} are assumed to be independently and normally distributed, following a standard normal distribution with a mean of zero and a variance of one, that is $\epsilon_{t+1} \sim \mathcal{N}(0, 1)$.

In this study, we compare the accuracies of one-step-ahead forecasts. We focus on an out-of-sample environment and consider forecasts with the estimated PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile principal components. We create forecasts using one, two, and three principal components:

$$\begin{aligned}
\tilde{y}_{t+1} &= \hat{\gamma} + \hat{\pi}_1 z_{1,t}, \\
\tilde{y}_{t+1} &= \hat{\gamma} + \hat{\pi}_1 z_{1,t} + \hat{\pi}_2 z_{2,t}, \\
\tilde{y}_{t+1} &= \hat{\gamma} + \hat{\pi}_1 z_{1,t} + \hat{\pi}_2 z_{2,t} + \hat{\pi}_3 z_{3,t},
\end{aligned}$$

where $\hat{\gamma}$ indicates a constant, $z_{i,t}$ the i-th principal component and $\hat{\pi}_i$ its parameter. To compare the forecasts we use the mean squared forecast error (MSFE). The MSFE is a metric used to evaluate the accuracy of a forecasting model. It measures the average squared difference between the forecasted values and the actual observed values. The MSFE provides an overall evaluation of the model's predictive performance, with lower values indicating better accuracy. The MSFE is given by $MSFE = \frac{1}{T} \sum_{t=1}^{T} (\hat{y}_{t,forecast} - y_t)^2$, where T is the number of forecasts, $\hat{y}_{t,forecast}$ is the forecast and y_t the actual observed value on t.

This research investigates the results of three distinct scenarios to examine the impact of different error simulations and check for robustness. This results in this paper are structured as follows: Section 3.1 explores the scenario where standard errors are simulated, Section 3.2 presents the results of a scenario with increased error generation, and Section 3.3 focuses on a special scenario where standard errors are generated and a number of extreme outliers are added to the data.

3.1 Scenario I: Standard Errors

In scenario I, we simulate the data where errors are generated from a normal distribution with zero mean and standard deviation σ_i, that is $e_{i,t} \sim \mathcal{N}(0, \sigma_i)$. The dataset consists of $N = 200$ predictors observed over a period of $T = 250$ time points. To evaluate the model's performance, we split the data into two parts: the first 200 observations are used for sample training, and the remaining 50 observations are reserved for out-of-sample evaluation. Forecasts are created using one, two, or three factors. n represents the number of strong factors, where $\frac{n}{N}$ represents the strong factor ratio of a case. Note that $n = 10$ represents the weakest case, while $n = 200$, where all factors have higher predictive capability, represents the strongest case. Table 1 displays the median MSFEs obtained from 100 repetitions for the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile forecasts. The lowest median MSFE values are highlighted in bold.

Table 1: The MSFEs of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile forecasts in scenario I, where errors are generated from a normal distribution

	PCA			sPCA - slope		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
n = 200	1.489	1.012	1.030	1.248	**0.994**	**1.008**
n = 150	1.554	**0.996**	1.009	1.284	1.004	1.011
n = 100	1.458	1.014	1.015	1.230	1.027	1.043
n = 50	1.509	1.077	1.086	1.250	1.026	1.045
n = 25	1.554	1.664	1.648	1.246	1.107	1.115
n = 10	1.741	1.603	1.748	1.317	1.280	**1.280**

	sPCA - tstat			sPCA - quantile		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
n = 200	**1.212**	1.032	1.037	1.247	1.024	1.014
n = 150	**1.249**	1.004	**1.003**	1.277	1.011	1.026
n = 100	**1.175**	**0.993**	**0.994**	1.254	1.025	1.040
n = 50	**1.217**	**1.009**	**1.029**	1.255	1.037	1.039
n = 25	**1.202**	1.115	1.114	1.250	**1.088**	**1.097**
n = 10	**1.296**	1.289	1.310	1.313	**1.287**	1.310

A first observation is that the accuracy of the four models improves as the number of strong factors increases. This implies that all (s)PCA methods make more accurate predictions when there is a greater number of factors with higher predictive power. This can be explained due to the inherent nature of PCA, which seeks to identify underlying patterns in the data. The strong factor case provides more structure which makes it easier to find such patterns, improving the forecast accuracies of the models. Secondly, the forecast accuracy of the four models improves using two factors and/or three factors. This

can be explained by the fact that the model can capture more variability in the data when including more principal components.

The MSFE values of the PCA method for one factor range from 1.46 to 1.74, while the MSFE values of the sPCA methods range from 1.21 to 1.32. When considering two factors, the PCA method yields MSFE values ranging from 1.00 to 1.66, while the sPCA methods yield values ranging from 0.99 to 1.29. For three factors, the MSFE values of the PCA method range from 1.03 to 1.75, while the sPCA methods range from 0.99 to 1.31. It is observed that, in general, the sPCA methods exhibit lower MSFE values compared to the PCA method, indicating more accurate forecasting performance.

The comparison of sPCA methods reveals that the sPCA-stat approach outperforms both the sPCA-slope and the sPCA-quantile method. The MSFE values obtained by using sPCA-stat are consistently the lowest using one, two or three factors across various strong/weak cases. When using one factor, the MSFE values for sPCA-tstat are the lowest in all cases. These results indicate that the use of sPCA-tstat leads to the most accurate forecasts in the scenario where standard errors are generated. It demonstrates that using the t-statistic as a scalar gives the most predictive power in a normal situation. This can be attributed to the characteristics of the t-statistic, that not only captures the relationship between the predictor and the target variable but also its statistical significance.

3.2 Scenario II: Larger Errors

In scenario II, we simulate the data by introducing errors generated from a t-distribution with standard deviation σ_i. Compared to a normal distribution, a t-distribution exhibits heavier tails (Appendix Figure 4), allowing for the occurrence of much larger values. We introduce larger errors to examine how different the (s)PCA forecast techniques handle these and to asses the robustness of these approaches. The errors are generated from t-distributions with varying degrees of freedom: $v = 3$ and $v = 1$, where a lower degree of freedom signifies heavier tails. That is $e_{i,t} \sim t(0, v, \sigma_i)$. To evaluate the model's performance, we again use the first 200 observations for sample training and the remaining 50 observations for out-of-sample evaluation. Table 2 and Table 3 displays the median MSFEs obtained from 100 repetitions for the PCA, sPCA-slope, sPCA-stat and sPCA-quantile forecasts. The lowest median MSFE values are highlighted in bold.

Firstly, we obtain interesting results about the sPCA-tstat approach. When generating errors from the heavier tailed t-distribution with $v = 3$, we observe a decrease in forecast accuracy compared to the standard error scenario. This is indicated by the decreasing number of cases where sPCA-stat outperforms the other techniques. When generating larger errors (t-distribution with degrees of freedom $v = 1$) we observe an even poorer forecast performance of the sPCA-stat, indicated by the highest MSFEs in almost all cases, ranging from 1.52 to 2.01. These results indicate the non-robustness of the sPCA-tstat technique in terms of forecast performance. It suggests that the sPCA-tstat approach demonstrates reduced forecast accuracy when more larger errors are generated and lacks robustness in comparison to PCA, sPCA-slope and sPCA-quantile. This can be explained due to the fact that the t-statistic becomes less reliable in the scenario with larger errors. The t-statistic is calculated by the estimated parameter divided by its standard deviation. Very large errors can increase the variability of the parameters, which results in larger standard deviations. As a result, the t-statistic may become smaller which makes it more difficult to detect statistically significant relationships.

Table 2: The MSFEs of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile forecasts in scenario II, where errors are generated from a t-distribution with df = 3

	PCA			sPCA - slope		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
n = 200	1.540	1.090	1.093	1.243	**1.052**	**1.057**
n = 150	1.494	1.086	1.095	1.248	**1.080**	**1.094**
n = 100	1.437	1.135	1.145	1.226	**1.105**	**1.112**
n = 50	1.497	1.486	1.503	1.274	1.146	**1.139**
n = 25	1.555	1.567	1.580	**1.303**	1.273	1.269
n = 10	1.917	1.946	1.938	**1.418**	**1.393**	**1.399**

	sPCA - tstat			sPCA - quantile		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
n = 200	**1.236**	1.054	1.068	1.237	1.057	1.062
n = 150	**1.230**	1.097	1.105	1.250	1.135	1.135
n = 100	**1.218**	1.115	1.125	1.228	1.113	1.113
n = 50	**1.252**	**1.121**	1.144	1.273	1.161	1.155
n = 25	1.328	1.290	1.282	1.314	**1.261**	**1.254**
n = 10	1.422	1.397	1.413	1.421	1.409	1.424

Secondly, there are significant findings regarding the sPCA-slope and sPCA-quantile approaches. In both cases of generating errors (t-distribution with degrees of freedom $v = 1$ and $v = 3$) these two methods outperform PCA. Comparing sPCA-slope and sPCA-quantile, we observe that sPCA-slope outperforms sPCA-quantile, particularly for $v = 1$, demonstrating the lowest MSFEs in sixteen out of eighteen cases. These results suggest that while the sPCA-quantile offers a robust alternative for forecasting, it falls short to the robustness exhibited by the sPCA-slope approach. To conclude, the sPCA-slope technique consistently achieves superior forecast accuracy when generating larger errors. This means that in this scenario, the use of sPCA-slope, which incorporates the mean relationship of the target variable and the predictors, is more robust than the than use of sPCA-quantile, which incorporates the median relationship.

Table 3: The MSFEs of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile forecasts in scenario II, where errors are generated from a t-distribution with df = 1

	PCA			sPCA - slope		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
n = 200	1.549	1.552	1.581	1.294	**1.194**	**1.213**
n = 150	1.530	1.597	1.660	**1.243**	**1.168**	**1.172**
n = 100	1.586	1.617	1.629	**1.252**	**1.236**	**1.225**
n = 50	1.950	1.950	1.967	**1.361**	**1.396**	1.459
n = 25	1.974	1.978	2.080	**1.470**	**1.465**	**1.496**
n = 10	1.897	1.920	1.956	**1.679**	**1.711**	**1.735**

	sPCA - tstat			sPCA - quantile		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
n = 200	1.802	1.529	1.520	**1.285**	1.264	1.273
n = 150	1.769	1.773	1.748	1.304	1.302	1.300
n = 100	1.910	1.886	1.833	1.333	1.324	1.352
n = 50	1.953	2.013	2.090	1.428	1.435	**1.434**
n = 25	1.968	1.999	2.009	1.661	1.603	1.581
n = 10	1.942	1.988	2.009	1.793	1.758	1.802

3.3 Scenario III: Extreme Outliers

In Scenario III, we investigate the case where standard errors are generated and a set of extreme outliers are added to the dataset. By adopting this approach, we can explore the impact of extreme outliers on the sPCA techniques. It allows us to gain insights into the performance and robustness of these techniques in scenarios where the data exhibits unusual observations.

To simulate this scenario, errors are generated from a standard normal distribution with a mean of zero and a standard deviation of σ_i, denoted as $e_{i,t} \sim \mathcal{N}(0, \sigma_i)$. Subsequently, a set of extreme outliers is introduced for each predictor i, specifically with the value of $10^4 \times \sigma_i$. These outliers are randomly assigned to each predictor across the time period $T = 1, ..., 250$ and can be either positive or negative. In this analysis, we examine the effects of incorporating 1 to 5 extreme outliers per predictor throughout the time period T. The number of outliers added to the data is denoted by Q. In contrast with scenario I, all factors are set to be strong factors ($n = 200$). This gives the best insights of the impact of the outliers as it assumes no noise of the weak factors in this case. To evaluate the model's performance, we again use the first 200 observations for sample training and the remaining 50 observations for out-of-sample evaluation. Table 4 displays the median MSFEs obtained from 100 repetitions for the PCA, sPCA-slope, sPCA-stat and sPCA-quantile forecasts.

Table 4: The MSFEs of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile forecasts in scenario III, where errors are generated from a normal distribution with the addition of extreme outliers

	PCA			sPCA - slope		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
Q = 1	1.798	1.634	1.756	1.464	1.568	1.426
Q = 2	2.046	2.249	2.123	1.761	1.773	1.730
Q = 3	1.946	2.051	2.131	1.987	2.075	2.044
Q = 4	1.935	1.999	2.183	2.036	2.080	2.215
Q = 5	2.056	2.002	2.001	1.982	2.089	2.120

	sPCA - tstat			sPCA - quantile		
	One factor	Two factors	Three factors	One factor	Two factors	Three factors
Q = 1	1.611	1.706	1.865	*1.259*	*1.298*	*1.285*
Q = 2	1.956	2.012	2.067	*1.631*	*1.571*	1.806
Q = 3	2.040	2.029	2.043	1.951	2.089	2.083
Q = 4	1.931	2.041	2.216	1.957	2.006	2.175
Q = 5	2.007	2.084	1.987	2.069	2.079	2.071

The results show interesting outcomes for the sPCA-quantile for $Q = 1$ and $Q = 2$, which are italicized in Table 4. For $Q = 1$, we obtain significantly lower MSFEs utilizing the sPCA-quantile approach in comparison with the other approaches. The sPCA-quantile MSFEs of the predictions using one, two, or three factors range between 1.26 and 1.30. In comparison, the MSFEs of PCA, sPCA-slope, and sPCA-stat range from 1.63 to 1.80, 1.43 to 1.57, and 1.61 to 1.89, respectively. This suggest that the sPCA-quantile technique has the most predictive power when one extreme outlier is added to the data. For $Q = 2$, the sPCA-quantile approach also yields significantly lower MSFEs in the cases where one or two factors are used for forecasting. However, these outcomes are comparatively less significant than in the case of a single extreme outlier ($Q = 1$). The presence of more than two extreme outliers ($Q = 3$, $Q = 4$ and $Q = 5$) leads to similar MSFEs along the four different techniques and there is no advantage of the sPCA-quantile technique obtained in these cases.

The analysis of the results reveals an interesting insight. When introducing a small number of extreme outliers, the sPCA-quantile technique demonstrates the best forecast accuracy, especially in the case of single extreme outlier. This suggests that sPCA-quantile, which incorporates the median quantile regression coefficient as a scalar, provides a robust approach when encountering an extreme outlier. The effectiveness of this technique can be attributed to the nature of median quantile regression, which focuses on estimating the conditional median of the target variable instead of the conditional mean as in OLS. As a result, median quantile regression is less sensitive to extreme outliers because these values have less impact on median estimation compared to mean estimation. By incorporating the median quantile regression coefficients as scalars, the sPCA-quantile technique enhances robustness and produces a set of scaled predictors that is less influenced by extreme outliers. As a result, the sPCA-quantile model exhibits optimal performance in this scenario. However, when confronted with an increased number of extreme outliers, the model's stability diminishes, leading to lower out-of-sample performance

Overall, this scenario highlights the importance of considering the sPCA-quantile method as the most accurate technique for forecasting in datasets with individual extreme outliers.

4 Empirical Study

In this section, we apply the (s)PCA techniques for macroeconomic forecasting with real data. We compare the forecast accuracies of the PCA, sPCA-slope, sPCA-tstat and the sPCA-quantile method. We evaluate these methods through in-sample and out-of-sample forecasting. The section is structured as follows: we begin by introducing the dataset, followed by presenting the in-sample results, and ending with the out-of-sample results.

4.1 Data

In this paper, we consider 123 U.S. macroeconomic variables introduced by (McCracken and Ng, 2016) which we use as predictors. The variables are from the FRED-MD database, covering monthly data from January 1960 to December 2019. The FRED-MD database, also known as the Federal Reserve Economic Data-Macroeconomic Data, is a comprehensive collection of macroeconomic and financial time series data maintained by the Federal Reserve Bank of St. Louis. It consists of a wide range of economic indicators which includes variables related to GDP, inflation, employment, interest rates, and financial markets. The FRED-MD database is widely used by economists, researchers, and policymakers to analyze and study macroeconomic trends and dynamics. The 123 predictors cover the following economic categories: output (No. 1-16), labor (No. 17-47), housing (No. 48-64), money (No. 65-78), interest and exchange rates (No. 79-99) and prices (No. 100-123). Table 6 in the Appendix shows a detailed description of the predictor set (and their transformations).

This research uses the following monthly data as target variables: inflation, industrial production (IP), unemployment rate, and stock market returns from the U.S. spanning from January 1960 to December 2019. We specify inflation as the 'Consumer Price Index for All Urban Consumers: All Items in U.S. City Average' and stock market returns as the monthly returns of the 'S&P 500'. For $y_t \in \{$CPI, IP, stock market prices$\}$, we use the transformation $\dot{y}_t = ln(y_t) - ln(y_{t-1})$ and $\dot{y}_t = y_t - y_{t-1}$ for when y_t equals the unemployment rate.

4.2 In-Sample Results

This section examines the in-sample performance of the four methods: PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile. The evaluation consists of two main parts. Firstly, a comparison is made between PCA and the sPCA techniques. Secondly, the in-sample forecasting results are presented.

4.2.1 Comparison PCA and sPCA's

To evaluate the performance of PCA and the three distinct sPCA techniques, we compare the variance explained of each method. The variance explained for principal component j denotes the proportion of the total variation of the predictors explained by the j-th principal component. Table 5 reports, in descending order, the variance explained for the first 10 principal components of each (s)PCA technique. It shows the results for the PCA and sPCA techniques for each target variable: inflation, industrial production (IP), unemployment rate and stock market returns.

Table 5: Variance explained by the principal components of PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile

	Inflation				IP			
	PCA	sPCA-s	sPCA-t	sPCA-q	PCA	sPCA-s	sPCA-t	sPCA-q
1st	0.15	0.38	0.45	0.46	0.15	0.36	0.33	0.40
2nd	0.07	0.11	0.10	0.09	0.07	0.10	0.10	0.10
3rd	0.07	0.09	0.06	0.08	0.07	0.06	0.07	0.08
4th	0.05	0.06	0.06	0.04	0.05	0.06	0.07	0.06
5th	0.04	0.05	0.05	0.05	0.04	0.06	0.06	0.05
6th	0.03	0.05	0.04	0.04	0.03	0.03	0.03	0.03
7th	0.03	0.04	0.03	0.03	0.03	0.03	0.03	0.03
8th	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02
9th	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
10th	0.02	0.02	0.01	0.01	0.02	0.02	0.02	0.02

	Unemployment				Stock Market			
	PCA	sPCA-s	sPCA-t	sPCA-q	PCA	sPCA-s	sPCA-t	sPCA-q
1st	0.15	0.41	0.40	0.69	0.15	0.31	0.32	0.27
2nd	0.07	0.10	0.10	0.13	0.07	0.21	0.21	0.19
3rd	0.07	0.06	0.07	0.07	0.07	0.08	0.08	0.09
4th	0.05	0.05	0.05	0.03	0.05	0.07	0.06	0.05
5th	0.04	0.04	0.04	0.02	0.04	0.05	0.05	0.05
6th	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.04
7th	0.03	0.02	0.02	0.01	0.03	0.02	0.02	0.03
8th	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.02
9th	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.02
10th	0.02	0.02	0.02	0.00	0.02	0.01	0.01	0.02

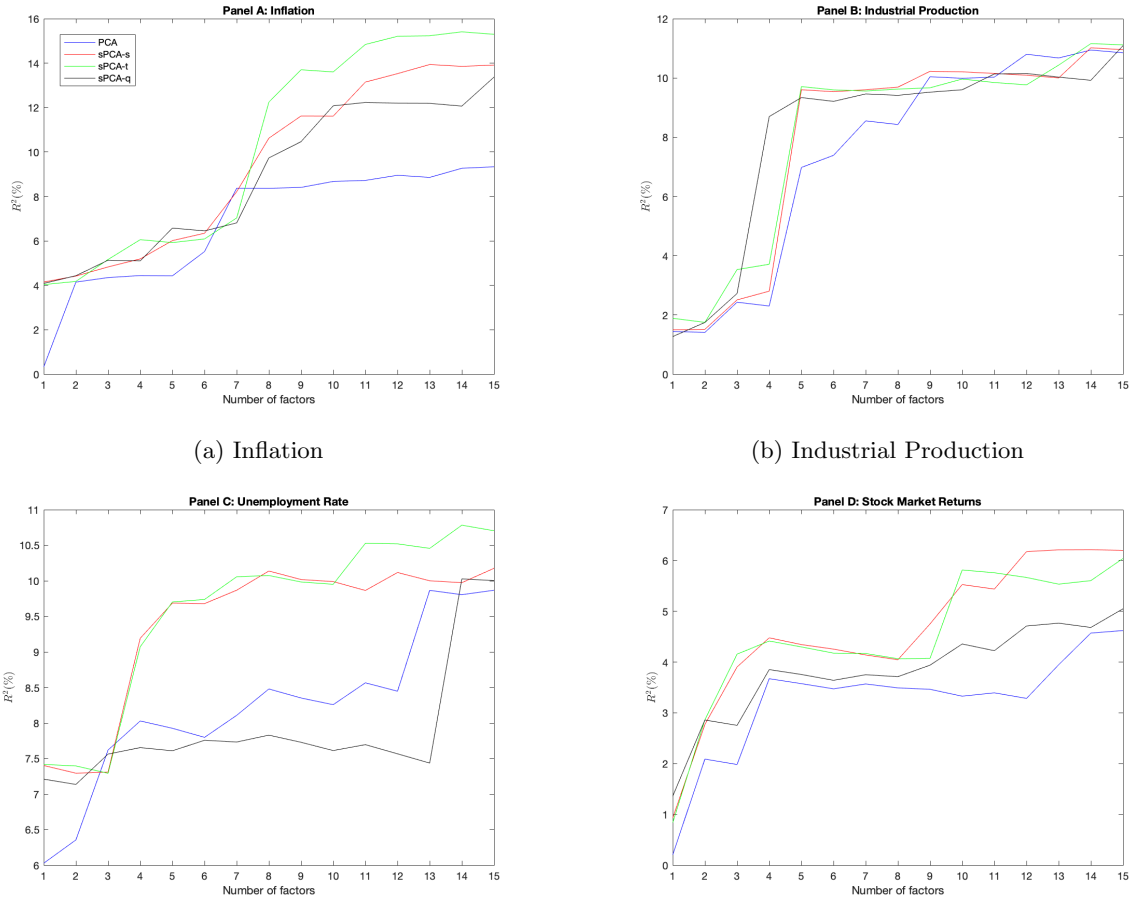
The comparison between PCA and the three sPCA methods reveals a remarkable difference in the variance explained. We observe that the variance explained is much higher around the first principal components of the sPCA methods compared to PCA. This difference can be attributed to the specific characteristics and objectives of the sPCA techniques. Unlike PCA, the sPCA methods are designed to capture the variation of the predictor set, which is modified to increase predictive power. By incorporating the information from the target variable into the dimensionality reduction process, the sPCA methods can extract the principal components that are most relevant for explaining the patterns and structure of the data specific to the target variable.

When comparing the three distinct sPCA methods, we denote a notable difference between the sPCA-quantile approach and the sPCA-tstat and sPCA-slope approaches. We obtain that the variances explained by the first principal component is higher for sPCA-quantile than sPCA-slope and sPCA-tstat, especially when forecasting unemployment rate (69%). In the case of sPCA-quantile, it specifically focuses on capturing the relationship at the median of the target variable's conditional distribution. This approach allows sPCA-quantile to potentially capture more information about the central tendency of the target variable, resulting in a higher variance explained by the first principal component. On the other hand, sPCA-slope and sPCA-tstat consider the slope coefficients and t-statistics, respectively, in their approaches. These methods emphasize the relationship between the predictors and the mean of the target

variable. They may not fully capture the details of the conditional distribution of the target variable, resulting in a lower variance explained by the first principal component compared to sPCA-quantile.

4.2.2 In-Sample Forecasting

In this section, we assess the in-sample forecasting performance of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile methods. We evaluate this performance by comparing the adjusted-R^2 of the one-month-ahead forecasts of inflation (Figure 1a), IP (Figure 1b), unemployment (Figure 1c) and stock market returns (Figure 1d), using the PCA and sPCA techniques. To create the forecasts, we consider a forecasting model with a number of factors ranging from 1 to 15.



(a) Inflation

(b) Industrial Production

(c) Unemployment Rate

(d) Stock Market Returns

Figure 1: In-sample forecasting performance of PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile

To evaluate the forecasting performance, we include previous observations of the target variable as lagged values in our analysis. The selection of the lag structure is based on the Bayesian Information Criterion (BIC). The BIC is a common statistical criterion for model selection. We limit the inclusion of lagged values to a maximum of three. The model for one-month-ahead forecasts is given by

$$\tilde{y}_{t+1} = \hat{\gamma} + \hat{\pi}_1 z_{1,t} + ... + \hat{\pi}_k z_{k,t} + \hat{\phi}_1 y_t + ... + \hat{\phi}_p y_{t+1-p},$$

where k denotes the number of principal components, p the number of lagged values ($p_{max} = 3$), $\hat{\gamma}$ the

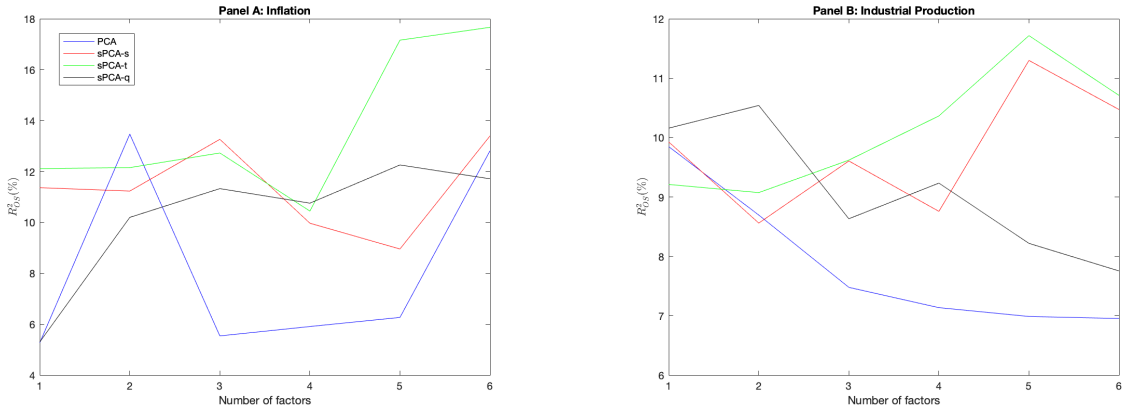
constant, $z_{i,t}$ the i-th principal component and $\hat{\pi}_i$ its parameter, and $\hat{\phi}_j$ the parameter of the j-th lagged variable. We use OLS to derive the parameters for this model.

Figure 1a demonstrates the performance of the methods for predicting inflation. It is observed that the sPCA methods consistently outperform the PCA method, as indicated by the lower R^2 values obtained for (almost) all factors. This suggest that the sPCA techniques exhibits greater in-sample predictive power for inflation compared to other methods. Increasing the number of factors leads to improved forecast accuracy for the sPCA-tstat approach. This means this approach has the most in-sample predictive power for forecasting inflation. The sPCA-slope and sPCA-quantile techniques display similar values of R^2. The performance of the methods for predicting industrial production (IP) is shown in Figure 1b. It is observed that all three sPCA techniques outperform PCA in terms of R^2 values when considering the first 9 principal components. This indicates that the sPCA techniques demonstrate a better in-sample predictive capability for IP. Figure 1c shows the in sample R^2 for the unemployment rate. The sPCA-slope and sPCA-tstat methods exhibit higher values compared to the sPCA-quantile and PCA methods. Among these, the sPCA-stat method exhibits the highest values for R^2. The sPCA-quantile method shows substantially lower performance, indicating its limited suitability for forecasting the unemployment rate. This observation suggests that using the median regression relationship of the unemployment rate and the predictors leads to a reduced effectiveness of the sPCA-quantile method. Figure 1d illustrates the R^2 values for the stock market return. The results in this figure indicate that the three sPCA techniques consistently outperform PCA, with sPCA-tstat and sPCA-slope demonstrating the highest values for all principal components.

Overall, the findings suggest that using the sPCA techniques, particularly sPCA-tstat and sPCA-slope, can enhance the in-sample predictive capability. The sPCA-tstat method demonstrates the highest performance among the sPCA techniques. This can be attributed to the fact that this method considers the statistical significance of the predictors by using t-statistic as a scalar. This means that it not only captures the relationships between variables but also assesses the significance and reliability of these relationships. By giving more weight to statistically significant predictors, the sPCA-tstat method can effectively focus on the most informative variables, which leads to improved in-sample predictive performance.

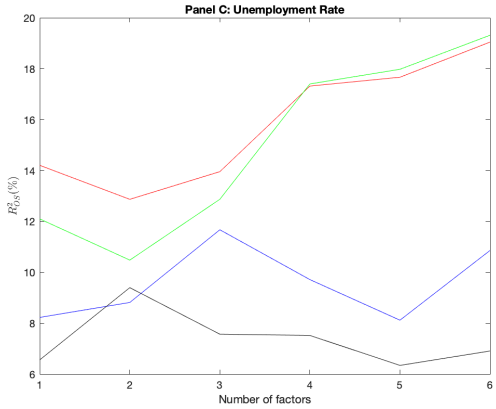
4.3 Out-Of-Sample Results

This section assesses the out-of-sample forecasting performance of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile methods. We compare the techniques by evaluating the out-of-sample R^2, a measurement of how well a model predicts new data. It indicates the proportion of variation in the target variable that the model can explain for this new data. A high out-of-sample R^2 means good predictive performance, while a low value suggests poor prediction ability. The the out-of-sample R^2 is calculated as follows: $R^2_{OS} = 1 - \frac{\sum_{t=1}^{T}(y_t - \tilde{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$, where $\bar{y}$ indicates the mean of the target variable, $\tilde{y}_t$ the forecasts, y_t the actual values at t.

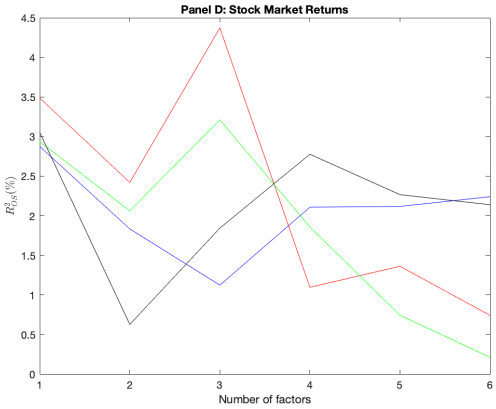


(a) Inflation

(b) Industrial Production



(c) Unemployment Rate



(d) Stock Market Returns

Figure 2: Out-of-sample forecasting performance of PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile

The factors and predictive regressions are recursively estimated with an expanding window scheme. The estimation window ranges from January 1960 to December 1994 (420 observations). The out-of-sample evaluation period is from January 1995 to December 2019 (300 observations).

To create the forecasts, we consider a forecasting model with a number of factors ranging from 1 to 6. In the model, we incorporate lagged values of the target variable by selecting the number of lags based on the Bayesian information criterion (BIC). Therefore, we use the same prediction model as described in Section 4.2.2 for in-sample forecasting. This approach ensures the consistency of our prediction model for out-of-sample forecasting.

Figure 2a demonstrates that the three sPCA techniques generate higher R^2_{OS} values for forecasting inflation, with the exception of using 2 principal components, where PCA outperforms the three sPCA techniques. On average, the sPCA-tstat approach exhibits a better performance compared to both the sPCA-slope and sPCA-quantile approaches. Figure 2b illustrates the R^2_{OS} for predicting IP. Once again, the sPCA techniques exhibit higher values compared to PCA. Notably, the sPCA-tstat approach has the highest out-of-sample prediction performance, when forecasting with multiple principal compontents. The out-of-sample forecast performance for unemployment rate is shown in Figure 1c. We obtain the highest performance of the sPCA-tstat and sPCA-slope techniques. The sPCA-quantile approach displays the lowest forecast accuracy for all principal components. A similar result is shown in 1c, where this techniques also exhibits the lowest R^2.

Figure 1d displays the R^2_{OS} for stock market returns. The findings in this figure suggest that, for the initial three principal components, sPCA-slope and sPCA-tstat exhibit greater out-of-sample forecasting performance compared to PCA and sPCA-quantile. However, when using more than three principal components, the sPCA-quantile and PCA demonstrate higher out-of-sample performance.

Overall, the results obtained from the analysis using PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile techniques reveal interesting findings regarding their out-of-sample forecasting performance. We can conclude that the three sPCA techniques provide a better out-of-sample forecasting accuracy than PCA. This highlights the effectiveness of the sPCA methods in generating more accurate and reliable forecasts. In general, the sPCA-tstat method generally demonstrates superior out-of-sample performance compared to other sPCA techniques, which is similar in the in-sample forecasting case in Section 4.2.2. This can be attributed to its consideration of statistical significance, using the t-statistic as a scalar, which leads to improved out-of-sample forecasting. The superior out-of-sample forecasting performance of the sPCA-stat technique is further substantiated in our simulation study conducted in Section 3.1. These findings demonstrate that, in a standard scenario, the sPCA-technique exhibits the highest level of forecast accuracy.

5 Conclusion

The objective of this research is to investigate the accuracy of three different scaled PCA techniques: sPCA-slope, sPCA-tstat, and sPCA-quantile. In a simulation and an empirical study, we evaluate and compare the forecasts of the distinct techniques in various circumstances. The simulation study simulates three different scenarios. A first scenario where standard errors are generated, a second which incorporates larger errors and a third, where extreme outliers are added to the dataset. The results of this study give insights in the behaviour and robustness of the distinct techniques in different scenarios. In the empirical study, we perform the sPCA techniques on 123 macroeconomic variables to forecast U.S. inflation, industrial production, unemployment rate and stock market returns. We examine the in-sample and the out-of-sample performance of the methods.

The findings of these studies indicate that the sPCA-tstat technique proves to have the most predictive power in a normal situation. This is demonstrated by its superior forecast accuracy in scenario I of the simulation study, where standard errors are generated. Furthermore, the in-sample and out-of-sample results of the empirical study substantiate these findings. The results in this study reveal the highest forecast accuracy of the sPCA-tstat technique.

In contrast, scenario II and III of the simulation study reveal different outcomes of the sPCA-tstat technique. When larger outliers and/or extreme outliers are introduced, the method becomes significantly less accurate and exhibits the lowest forecast performance in comparison with the other techniques. This highlights the non-robustness of the sPCA-tstat technique and the limitations of using the t-statistic as a scalar. This behavior can be attributed to the characteristics of the t-statistic, which is calculated by dividing the parameter estimate by its standard deviation. With the introduction of larger errors or extreme outliers, the variability of the parameter estimates increases, leading to less accurate t-statistic scalars.

The results from Scenario III of the simulation study reveal interesting findings of the sPCA-quantile approach. It demonstrates that with the addition of some extreme outliers to the data, the sPCA-quantile technique exhibits superior forecast accuracy. Specifically, in the case of one extreme outlier, the sPCA-quantile technique proves to be more robust and outperforms other methods. This can be attributed to the robust nature of median quantile regression, which is better equipped to handle outliers compared

to OLS regression. The use of this method leads to less sensitive scalars which results in an optimal performance of the sPCA-quantile technique in this scenario.

Among the three techniques, sPCA-slope consistently performs well in the simulation and empirical study. It has constant results in scenario I of the simulation study and it also shows in scenario II its robustness to larger errors. Empirically, sPCA-slope demonstrates consistent forecast accuracy in both in-sample and out-of-sample analyses. This makes the sPCA-slope technique a reliable choice for forecasting purposes, as it maintains its performance across different circumstances.

Overall, this research highlights the characteristics of three different sPCA techniques: sPCA-slope, sPCA-stat, and sPCA-quantile. The findings provide valuable insights for practitioners in selecting the most appropriate method for specific circumstances.

References

D. W. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

J. B. Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 7(1):47–54, 2004.

T. E. Clark and M. W. McCracken. Forecasting with small macroeconomic vars in the presence of instabilities. 2007.

W. P. Gaglianone and L. R. Lima. Constructing density forecasts from quantile regressions. *Journal of Money, Credit and Banking*, 44(8):1589–1607, 2012.

M. He, Y. Zhang, D. Wen, and Y. Wang. Forecasting crude oil prices: A scaled pca approach. *Energy Economics*, 97:105189, 2021.

D. Huang, F. Jiang, K. Li, G. Tong, and G. Zhou. Scaled pca: A new approach to dimension reduction. *Management Science*, 68(3):1678–1695, 2022.

Y. Huang, L. Shen, and H. Liu. Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in china. *Journal of Cleaner Production*, 209: 415–423, 2019.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

B. Liu, J. Nowotarski, T. Hong, and R. Weron. Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, 8(2):730–737, 2015.

F. Ma, Y. Guo, J. Chevallier, and D. Huang. Macroeconomic attention, economic policy uncertainty, and stock volatility predictability. *International Review of Financial Analysis*, 84:102339, 2022.

L. Ma and L. Pohlman. Return forecasts and optimal portfolio construction: a quantile regression approach. *The European Journal of Finance*, 14(5):409–425, 2008.

L. Magee and M. R. Veall. Selecting regressors for prediction using press and white t statistics. *Journal of Business & Economic Statistics*, 9(1):91–96, 1991.

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

W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance matrix. 1986.

W. K. Newey and K. D. West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994.

H. A. Nielsen, H. Madsen, and T. S. Nielsen. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 9(1-2):95–108, 2006.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

C. Skittides and W.-G. Früh. Wind forecasting using principal component analysis. *Renewable Energy*, 69:365–374, 2014.

R. J. Smith. Automatic positive semidefinite hac covariance matrix and gmm estimation. *Econometric Theory*, 21(1):158–170, 2005.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179, 2002.

J. H. Stock and M. W. Watson. Forecasting with many predictors. *Handbook of economic forecasting*, 1: 515–554, 2006.

M. Taillardat, O. Mestre, M. Zamo, and P. Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

J. Wang, F. Ma, E. Bouri, and J. Zhong. Volatility of clean energy and natural gas, uncertainty indices, and global economic conditions. *Energy Economics*, 108:105904, 2022.

6 Appendix

6.1 A

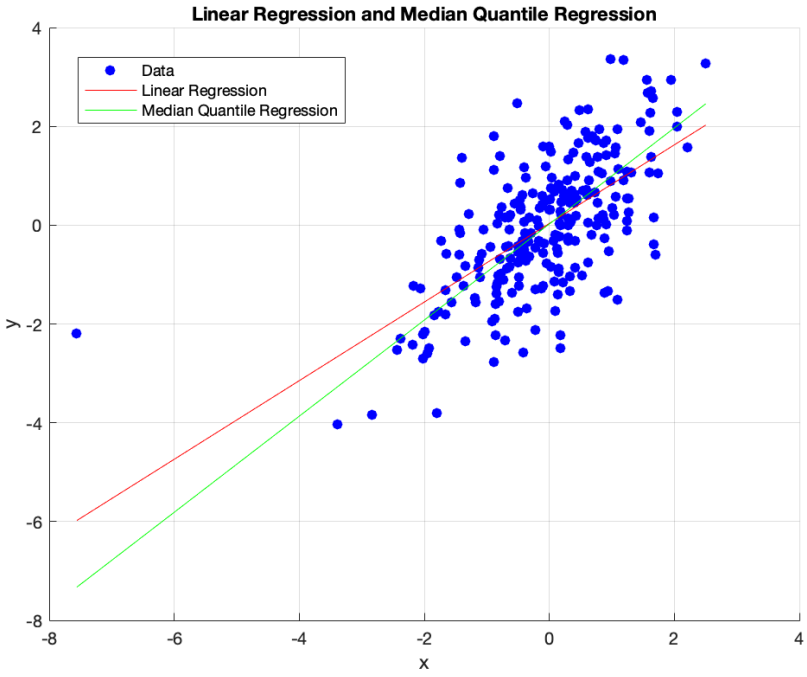


Figure 3: Median quantile regression and OLS

6.2 B

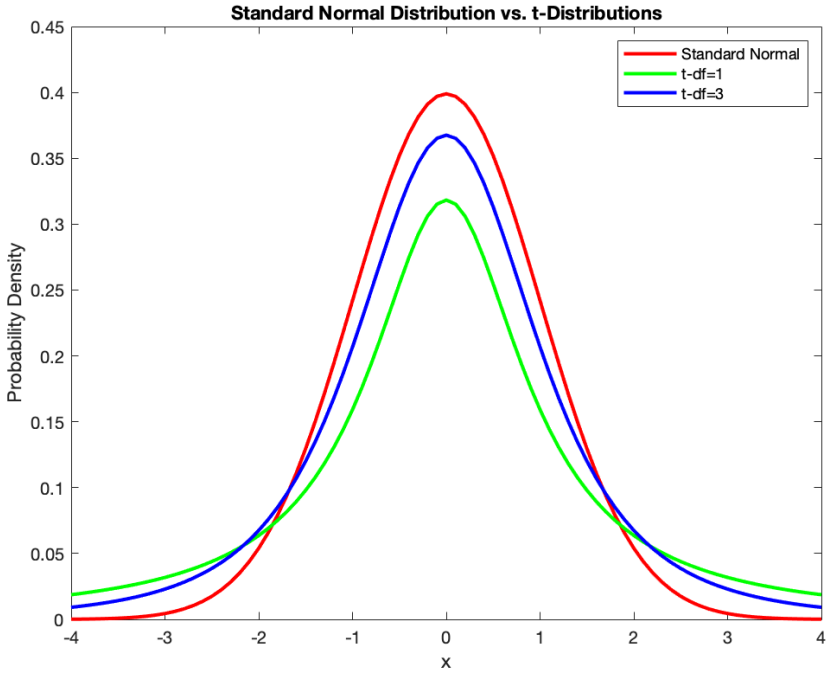


Figure 4: Normal and t-distribution for df=1 and df=3

6.3 C

Table 6.3 provides a list of 123 macroeconomic variables that we obtained from the Federal Reserve for Economic Research (FRED-MD). For each variable, we include its code (Mnemoic), a description (Variable Description), and the method used to transform the data to ensure it is stationary (trcode). This list of variables and their corresponding transformations aligns with the variables used in the study conducted by (Huang et al., 2022). Let $X_{i,t}$ and $\dot{X}_{i,t}$ denote the untransformed and the transformed predictors, respectively, for predictor i at time t. We apply the following transformations:

1. $\dot{X}_{i,t} = X_{i,t}$
2. $\dot{X}_{i,t} = X_{i,t} - X_{i,t-1}$
3. $\dot{X}_{i,t} = \Delta^2 X_{i,t}$
4. $\dot{X}_{i,t} = ln(X_{i,t})$
5. $\dot{X}_{i,t} = ln(X_{i,t}) - ln(X_{i,t-1})$
6. $\dot{X}_{i,t} = \Delta^2 ln(X_{i,t})$
7. $\dot{X}_{i,t} = \Delta(\frac{X_{i,t}}{X_{i,t-1}})$

Table 6: FRED-MD Macroeconomic Variables List

No	Mnemoic	Variable Description	trcode
1	RPI	Real Personal Income	5
2	W875RX1	Real personal income ex transfer receipts	5
3	INDPRO	IP Index	5
4	IPFPNSS	IP: Final Products and Nonindustrial Supplies	5
5	IPFINAL	IP: Final Products (Market Group)	5
6	IPCONGD	IP: Consumer Goods	5
7	IPDCONGD	IP: Durable Consumer Goods	5
8	IPNCONGD	IP: Nondurable Consumer Goods	5
9	IPBUSEQ	IP: Business Equipment	5
10	IPMAT	IP: Materials	5
11	IPDMAT	IP: Durable Materials	5
12	IPNMAT	IP: Nondurable Materials	5
13	IPMANSICS	IP: Manufacturing (SIC)	5
14	IPB51222S	IP: Residential Utilities	5
15	IPFUELS	IP: Fuels	5
16	CUMFNS	Capacity Utilization: Manufacturing	2
17	HWI	Help-Wanted Index for United States	2
18	HWIURATIO	Ratio of Help Wanted/No. Unemployed	2
19	CLF16OV	Civilian Labor Force	5
20	CE16OV	Civilian Employment	5
21	UNRATE	Civilian Unemployment Rate	2
22	UEMPMEAN	Average Duration of Unemployment (Weeks)	2
23	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	5
24	UEMP5TO14	Civilians Unemployed for 5-14 Weeks	5
25	UEMP15OV	Civilians Unemployed - 15 Weeks & Over	5
26	UEMP15T26	Civilians Unemployed for 15-26 Week	5
27	UEMP27OV	Civilians Unemployed for 27 Weeks and Over	5
28	CLAIMSx	Initial Claims	5
29	PAYEMS	All Employees: Total nonfarm	5
30	USGOOD	All Employees: Goods-Producing Industries	5
31	CES1021000001	All Employees: Mining and Logging: Mining	5
32	USCONS	All Employees: Construction	5
33	MANEMP	All Employees: Manufacturing	5
34	DMANEMP	All Employees: Durable goods	5
35	NDMANEMP	All Employees: Nondurable goods	5
36	SRVPRD	All Employees: Service-Providing Industries	5
37	USTPU	All Employees: Trade, Transportation & Utilities	5
38	USWTRADE	All Employees: Wholesale Trade	5
39	USTRADE	All Employees: Retail Trade	5
40	USFIRE	All Employees: Financial Activities	5
41	USGOVT	All Employees: Government	5

Continued on next page

Table 6 – continued from previous page

No	Mnemoic	Variable Description	trcode
42	CES0600000007	Avg Weekly Hours: Goods-Producing	1
43	AWOTMAN	Avg Weekly Overtime Hours: Manufacturing	2
44	AWHMAN	Avg Weekly Hours: Manufacturing	1
45	CES0600000008	Avg Hourly Earnings: Goods-Producing	6
46	CES2000000008	Avg Hourly Earnings: Construction	6
47	CES3000000008	Avg Hourly Earnings: Manufacturing	6
48	HOUST	Housing Starts: Total New Privately Owned	4
49	HOUSTNE	Housing Starts, Northeast	4
50	HOUSTMW	Housing Starts, Midwest	4
51	HOUSTS	Housing Starts, South	4
52	HOUSTW	Housing Starts, West	4
53	PERMIT	New Private Housing Permits (SAAR)	4
54	PERMITNE	New Private Housing Permits, Northeast (SAAR)	4
55	PERMITMW	New Private Housing Permits, Midwest (SAAR)	4
56	PERMITS	New Private Housing Permits, South (SAAR)	4
57	PERMITW	New Private Housing Permits, West (SAAR)	4
58	DPCERA3M086SBEA	Real personal consumption expenditures	5
59	CMRMTSPLx	Real Manu. and Trade Industries Sales	5
60	RETAILx	Retail and Food Services Sales	5
61	AMDMNOx	New Orders for Durable Goods	5
62	AMDMUOx	Unfilled Orders for Durable Goods	5
63	BUSINVx	Total Business Inventories	5
64	ISRATIOx	Total Business: Inventories to Sales Ratio	2
65	M1SL	M1 Money Stock	6
66	M2SL	M2 Money Stock	6
67	M2REAL	Real M2 Money Stock	5
68	AMBSL	St. Louis Adjusted Monetary Base	6
69	TOTRESNS	Total Reserves of Depository Institutions	6
70	NONBORRES	Reserves Of Depository Institutions	7
71	BUSLOANS	Commercial and Industrial Loans	6
72	REALLN	Real Estate Loans at All Commercial Banks	6
73	NONREVSL	Total Nonrevolving Credit	6
74	CONSPI	Nonrevolving consumer credit to Personal Income	2
75	MZMSL	MZM Money Stock	6
76	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	6
77	DTCTHFNM	Total Consumer Loans and Leases Outstanding	6
78	INVEST	Securities in Bank Credit at All Commercial Banks	6
79	FEDFUNDS	Effective Federal Funds Rate	2
80	CP3Mx	3-Month AA Financial Commercial Paper Rate	2
81	TB3MS	3-Month Treasury Bill	2
82	TB6MS	6-Month Treasury Bill	2
83	GS1	1-Year Treasury Rate	2

Continued on next page

Table 6 – continued from previous page

No	Mnemoic	Variable Description	trcode
84	GS5	5-Year Treasury Rate	2
85	GS10	10-Year Treasury Rate	2
86	AAA	Moody's Seasoned Aaa Corporate Bond Yield	2
87	BAA	Moody's Seasoned Baa Corporate Bond Yield	2
88	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS	1
89	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	1
90	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	1
91	T1YFFM	1-Year Treasury C Minus FEDFUNDS	1
92	T5YFFM	5-Year Treasury C Minus FEDFUNDS	1
93	T10YFFM	10-Year Treasury C Minus FEDFUNDS	1
94	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	1
95	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	1
96	EXSZUSx	Switzerland/U.S. Foreign Exchange Rate	5
97	EXJPUSx	Japan/U.S. Foreign Exchange Rate	5
98	EXUSUKx	U.S./U.K. Foreign Exchange Rate	5
99	EXCAUSx	Canada/U.S. Foreign Exchange Rate	5
100	PPIFGS	PPI: Finished Goods	6
101	PPIFCG	PPIFCG	6
102	PPIITM	PPIITM	6
103	PPICRM	PPI: Crude Materials	6
104	OILPRICEx	Crude Oil, spliced WTI and Cushing	6
105	PPICMM	PPI: Metals and metal products:	6
106	CPIAUCSL	CPI: All Items	6
107	CPIAPPSL	CPI: Apparel	6
108	CPITRNSL	CPI: Transportation	6
109	CPIMEDSL	CPI: Medical Care	6
110	CUSR0000SAC	CPI: Commodities	6
111	CUUR0000SAD	CPI: Durables	6
112	CUSR0000SAS	CPI: Services	6
113	CPIULFSL	CPI: All Items Less Food	6
114	CUUR0000SA0L2	CPI: All items less shelter	6
115	CUSR0000SA0L5	CPI: All items less medical care	6
116	PCEPI	Personal Cons. Expend.: Chain Index	6
117	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	6
118	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	6
119	DSERRG3M086SBEA	Personal Cons. Exp: Services	6
120	S&P 500	S&P's Common Stock Price Index: Composite	5
121	S&P: indust	S&P's Common Stock Price Index: Industrials	5
122	S&P div yield	S&P's Composite Common Stock: Dividend Yield	2
123	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio	2

6.4 Programming Code

This section explains the programming code and the data conducted in this research. The MATLAB code for simulation study in Section 3 is named *SimulationWeak*, the code for the in-sample empirical results in Section 4.2 is named *sPCA_InSample*, and the code for the out-of-sample empirical results in Section 4.3 *sPCA_OutSample*. These codes use a number of functions which are added in the zip-files. The most important functions are: *linear_reg* and *QuantileRegression* which compute the slope, t-statistic, and quantile regression scalars. The files *TestRsquared1* and *macro_nm2* contain the data of the target variables and the predictors, respectively. Descriptions of the MATLAB codes and the data are as follows:

- *SimulationWeak*: This code calculates the MSFEs of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile technique for 100 repitions. The first part of the code contains the data generating process which can be adjusted for scenario I, II, and III. The second and third part of the code contain the parameter estimation and the creation of forecasts. The last part of the code provides a calculation of the MSFEs for each technique.

- *sPCA_InSample*: This code calculates the variance explained and the in-sample R^2 of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile technique. The number of lags selected for the prediction model is computed by the *Select_AR_lag_SIC* function. The estimation of the parameters of the lagged variables is calculated by *Estimate_AR_res*. The first part of the code is the derivation and use of the scalars and the second part of the code calculates the variance explained and the R^2 for each principal component, denoted by the arrays *out_Rsquared* and *out_VarianceExplained*.

- *sPCA_OutSample*: This code provides the out-of-sample R^2 of the PCA, sPCA-slope, sPCA-tstat, and sPCA-quantile. This code has the same structure for the calculation of the lagged variables and scalars as *sPCA_InSample*. The parameters of the forecast model are calculated by the function *Estimate_ARDL_multi*, when nummer of selected lagged variables is greater than zero. The *linear_reg* function is used when the number of selected lags equals zero. The out-of-sample R^2 is caculated by the function *R2oostest*.

- *TestRsquared1*: This file contains the data of the target variables. The first, second, third, and fourth column contain the data of inflation, industrial production, unemployment rate, and stock market returns, respectively.

- *macro_nm2*: This file contains the data of the 123 macroeconomic variables (predictors). Each column represents the data of the corresponding variable.