Finding the Factors Explaining Industry Portfolios

Jesper Lauteslager 510887

Supervisor: Gustavo Freire

Second assessor: Terri van der Zwan

This paper evaluates the newly proposed risk-premium Principal Component Analysis (RP-PCA) method by Lettau and Pelger [2020a]. Next to this, the interpretation of the factors is researched. RP-PCA is a method to extract the factors explaining asset returns by combining standard PCA with the Arbitrage Pricing Theory. We evaluate the performance of RP-PCA on a small set of double-sorted portfolios, a larger set of single-sorted anomaly portfolios, and a set of industry portfolios. RP-PCA outperforms PCA and Fama-French (FF) when applied to the double-sorted or single-sorted anomaly portfolios. Also, RP-PCA can detect factors that PCA is not able to detect. For the industry portfolios, a model with RP-PCA factors 1, 2, 4, and 5 results in the smallest cross-sectional pricing errors and captures the common time-series movement best. However, the Fama-French five-factor model results in the highest Sharpe ratio. The first RP-PCA factor can be interpreted as a market factor, the second RP-PCA factor is related to the mining industry, and the fifth factor can be linked to the High-Tech industry.

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Quantitative Finance

3 July 2022

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics, or Erasmus University Rotterdam.

Contents

1 Introduction

Pricing an asset is essential for many parties in the economy. Pension funds, asset managers, private investors, central banks, etc. are all interested in the "true" price of an asset. Lettau and Pelger [2020a] state that "finding the "right" factors has become the central question of asset pricing." A lot of research on finding these factors has been conducted. Currently, hundreds of potential risk factors are proposed. According to Feng, Giglio, and Xi [2017], this leads to a "zoo" of factors in which it is hard to distinguish between useful and redundant pricing factors that appear significant.

Lettau and Pelger [2020a] propose a new method to find the most important factors explaining asset returns. This new method proposed is called risk-premium Principal Component Analysis (RP-PCA). Standard PCA only uses information contained in the second moment of asset returns, whereas RP-PCA uses information contained in both the first and second moments. Lettau and Pelger [2020a] apply the RP-PCA method to a small set of double-sorted portfolios with 25 assets. Also, RP-PCA is applied to a larger set of anomaly portfolios used in Kozak, Nagel, and Santosh [2020]. Lettau and Pelger [2020a] find that RP-PCA factor models provide smaller pricing errors and higher Sharpe Ratios. Also, a clear economic interpretation can be given to the RP-PCA factors.

In this paper, we apply the RP-PCA and PCA methods to the excess returns of the same double-sorted and single-sorted portfolios as in Lettau and Pelger [2020a]. To extend the paper, we apply both methods to a set of 48 industry portfolios. We evaluate the differences between RP-PCA and PCA and compare both with the Fama-French three-factor model proposed in Fama and French [1993], which is used as a benchmark model. It is interesting to see, if also for industry portfolios, RP-PCA provides better results than PCA and the Fama-French three-factor model. As the RP-PCA method is relatively new, it is relevant to test the performance on a different dataset. Currently, there is no literature comparing the RP-PCA method with other methods, except for the article in which the RP-PCA method is proposed [Lettau and Pelger, 2020a].

More importantly, the economic interpretation of the factors is researched. According to Cavaglia, Brightman, and Aked [2000], the relative importance of industry factors has been growing over time due to globalization. Can the factors found be related to (co-moving) industries, or do other aspects play a role? This leads to the following research question:

What are the most important factors explaining the asset returns of industry portfolios?

We use a set of monthly returns of 48 different industry portfolios. Because of the high number of industries, each of the 48 industries is rather specific. The 48 specific industries can be divided into five broader industries. It could be the case that the factors explaining the industry portfolios can be related to these broader industries. It is also possible that we find factors that explain certain industries that are unrelated to each other via those five broader industries, which could lead to new insights. This can be useful for investors because they can use that information to reduce risk when they know which industries are co-moving.

From an academic point of view, this research can also be useful. Lettau and Pelger state that RP-PCA results in smaller pricing errors and higher Sharper ratios (SR) than PCA. Also, RP-PCA is computationally not much more complex than PCA. Therefore RP-PCA could be used for different applications when it turns out that RP-PCA produces better results than PCA in different samples.

The fit of RP-PCA, PCA, and Fama-French are evaluated using the maximum Sharpe ratio, cross-sectional pricing errors, and the unexplained time-series variation. We find that, consistent with Lettau and Pelger [2020a], RP-PCA outperforms PCA and Fama-French for the double-sorted portfolios. Also, RP-PCA can detect factors that PCA is not able to detect. For the single-sorted anomaly portfolios, RP-PCA again outperforms both PCA and Fama-French. More specifically, the RP-PCA model with five factors is the most preferred. A model with RP-PCA factors 1, 2, 4, and 5 is the preferred model for the industry portfolios. This model results in the smallest cross-sectional pricing errors and explains best the common time-series movement. However, in terms of the Sharpe ratio, this model is outperformed by the Fama-French five-factor model. RP-PCA and PCA generate very similar results in terms of factors and factor loadings. When increasing the RP-PCA parameter γ substantially, RP-PCA and PCA still select the same factors. The first factor found can be interpreted as a market factor. The second factor is related to the mining industry. The fifth factor can be linked to one of the five broader industries, the High-Tech industry. No interpretation can be found for the third and fourth factors.

2 Relevant Literature

A lot of research has been conducted on finding the factors explaining asset returns. Well-known research was conducted by Fama and French [1993]. They found that the returns of assets can be described by three factors. The first factor is the market risk. The second factor is the return of a portfolio long in high book-to-market stocks and short in low book-to-market stocks. The third factor is the return of a portfolio long in small-cap companies and short in large-cap companies. Research has also been conducted on finding the factors that explain the returns of industry portfolios in particular. In most papers researching this, factors are constructed based on economic theory. Cavaglia, Hodrick, Vadim, and Zhang [2002], for example, compare different models based on economic theory, including the Fama-French three-factor model, to price the returns on 36 industry portfolios. They find that the Fama-French three-factor model performs best compared to the other methods used in the paper. We use this model as a benchmark and compare the PCA method and the newly proposed RP-PCA method to it. Our paper has some similarities with Elhadary [2021]. This paper tries to explain industry portfolios using the traditional Fama-French three-factor model and an industry-based Fama-French three-factor model. For this industry-based model, the Fama-French factors are constructed for each industry group. The industry portfolios are divided into industry groups based on their Standard Industrial Classification (SIC) code, which is the same approach as in our paper. The SIC code defines a company's core business and, thus, the industry it belongs to. Elhadary [2021] shows that the industry-based Fama-French three-factor model outperforms the traditional Fama-French three-factor model.

This research contributes to the emerging literature that uses econometric methods to explain and forecast asset prices. In recent years, numerous new techniques for asset pricing have been proposed. Fan, Liao, and Wang [2016] developed a model that allows for time-varying factor loadings: projected-PCA. Kelly, Pruitt, and Su [2020] came up with the idea of instrumented-PCA, which performs dimensionality reduction of the characteristics space. Also, this research is part of the growing field of econometrics literature that merges regularization and latent factor extraction. Principal components can be constructed by iteratively applying least squares regressions. Bai and Ng [2019] replace least squares with ridge regressions. When extreme outliers are present or some factors have very small loadings, it is useful to apply this method. This method decreases the eigenvalues of the common components to zero. This approach leads to smaller variation but at the cost of a higher bias.

Our research adds to the existing literature by applying a newly proposed method, RP-PCA, to a different dataset and evaluating its performance. Also, we try to give an economic interpretation to the factors found, which can be useful for investors and asset managers.

3 Methodology

In this section, we explain the idea and intuition behind RP-PCA and PCA. We denote the excess return of asset n at time t as X_{nt}. The excess return of an asset is equal to the "raw" return minus the risk-free rate. Therefore, we first subtract the risk-free rate from the raw asset returns. The excess returns can be explained by systematic risk factors F_t and idiosyncratic components e_{nt}. We observe the excess returns X_{nt} of N assets over T time periods as

$$X_{nt} = F_t \Lambda_n^{\mathrm{T}} + e_{nt}, n = 1, ..., N, t = 1, ..., T, \tag{1}$$

where F_t are the factors, Λ_n are the factor loadings, and e_{nt} are the idiosyncratic components. The essential idea is to describe the variation in the N excess returns with a limited number of k factors, where $k < N$.

3.1 PCA

According to Lettau and Pelger [2020b], PCA aims to find the factors that explain as much common time-series variation as possible. As shown by Stock and Watson [2002], the PCA factors and factor loadings are solutions to the following function:

$$\min_{\Lambda, F} \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} ((X_{nt} - \bar{X}_n) - (F_t - \bar{F}) \Lambda_n^T)^2, \tag{2}$$

where $\bar{X}_n$ and $\bar{F}$ represent the mean of the excess returns and the mean of the factors, respectively.

The loadings, Λ, are eigenvectors of the largest eigenvalues of the sample covariance matrix of the excess returns X:

$$\Sigma = \frac{1}{T} X^T X - \overline{X} \overline{X}^T, \tag{3}$$

where $\overline{X}$ denotes the sample mean of the excess returns.

The factors can be obtained from a regression of X on the estimated loadings. The j^{th} factor explaining the excess returns can be found as follows:

$$F_{jt} = \Lambda_j' X_t, \tag{4}$$

where Λ_j is the j^{th} factor loading, corresponding to the j^{th} eigenvector of the sample covariance matrix of the excess returns.

3.2 RP-PCA

First, the idea of the Arbitrary Pricing Theory (APT) is explained. According to Ross [2013], the APT suggests that the excess returns can be explained by the exposure to systematic risk factors (Λ_n) multiplied by the risk premium of the factors ($E[F_t]$):

$$E[X_{n,t}] = E[F_t] \Lambda_n^T. \tag{5}$$

So, according to ATP, the systematic factors should explain the cross-section of expected returns. Therefore, the objective of the APT is to minimize the cross-sectional pricing error:

$$\frac{1}{N} \sum_{n=1}^{N} (\overline{X}_n - \overline{F} \Lambda_n^T)^2. \tag{6}$$

RP-PCA combines PCA and APT. Therefore we can say that RP-PCA explains as much common time-series variation as possible **and** considers cross-sectional pricing errors. This results in the following objective function, which is a combination of the objective functions of both PCA and APT:

$$\min_{\Lambda,F} \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}(\tilde{X}_{nt} - \tilde{F}_t\Lambda_n^T)^2 + \gamma\frac{1}{N}\sum_{n=1}^{N}(\overline{X}_n - \overline{F}\Lambda_n^T)^2, \tag{7}$$

where γ determines how much weight is given to the cross-sectional pricing errors from APT relative to the time-series errors from PCA.

RP-PCA applies PCA to a matrix different from the sample covariance matrix. RP-PCA applies PCA to the matrix

$$\Sigma_{RP} = \frac{1}{T}X^TX + \gamma\overline{XX}^T. \tag{8}$$

Formulas (3) and (8) again show the difference between PCA and RP-PCA. RP-PCA gives a higher weight to the mean than PCA. Note that they are equal when $\gamma = -1$. In essence, PCA is a more specific/restricted case of RP-PCA because, in RP-PCA, one can choose the parameter γ to be equal to any number, as opposed to PCA, in which γ is always equal to minus one.

3.3 Performance Evaluation

Next, we evaluate the performance of RP-PCA, PCA, and the Fama-French models. The performance is evaluated both in-sample and out-of-sample based on three different measures, being:

- Maximum Sharpe ratio obtained from estimated factors (SR)
- Root-mean-squared pricing errors (RMS_α)
- Idiosyncratic variance ($\bar{\sigma}_e^2$)

3.3.1 In-Sample

We want to find the model that best describes the asset returns. The stochastic discount factor (SDF) is used to describe asset prices. The closer the model is to the true SDF of the assets, the higher the maximum Sharpe ratio. Therefore, we prefer a high maximum Sharpe ratio. The maximum Sharpe ratio can be obtained from the estimated factors. These factors are estimated using either RP-PCA, PCA, or are equal to the factors implied by Fama-French. First, compute $b = \Sigma_F^{-1}\mu_F$, where Σ_F^{-1} and μ_F are equal to the covariance matrix and the mean of the estimated factors, respectively. The in-sample maximum Sharpe ratio can then be obtained as

$$SR = \frac{E[b^T\hat{F}_t]}{\sigma(b^T\hat{F}_t)}. \tag{9}$$

To compute the other two performance measures, we regress the excess returns on the estimated factors and a constant using Ordinary Least Squares (OLS):

$$X_{nt} = \alpha_n + \hat{F}_t B_n^T + e_{nt}, \qquad n = 1, ..., N. \tag{10}$$

The value of α_n can be seen as the magnitude of the pricing errors of the model. This α_n is used to describe the cross-sectional pricing errors. The values of e_{nt} are used to compute the amount of unexplained time-series variation. This results in the following two performance measures:

$$RMS_\alpha = \sqrt{\frac{\hat{\alpha}^T \hat{\alpha}}{N}} \tag{11}$$

$$\bar{\sigma}_e^2 = \frac{\frac{1}{N}\sum_{n=1}^N Var(\hat{e}_n)}{\frac{1}{N}\sum_{n=1}^N Var(X_n)}. \tag{12}$$

Ideally, you want both RMS_α and $\bar{\sigma}_e^2$ as low as possible.

3.3.2 Out-of-Sample

For the out-of-sample performance evaluation, we use a rolling window of 20 years (240 observations). Using these 240 observations, we estimate the factor loadings. To obtain the parameters needed for calculating the performance measures, we use the following procedure:

- Using the excess returns at time $t + 1$ and the estimated factor loadings up to time t, we predict the factor $\hat{F}_{t+1}$.
- Compute $b_t = \Sigma_F^{-1} \mu_F$ in the estimation window.
- Calculate the estimated out-of-sample portfolio return $b_t^T \hat{F}_{t+1}$.
- Compute B_n in the equation $X_{nt} = \alpha_n + \hat{F}_t B_n^T + e_{nt}, \quad$ for $n = 1, ..., N$, using data contained in the estimation window.
- Calculate the out-of-sample pricing error as $\hat{\alpha}_{n,t+1} = X_{n,t+1} - \hat{F}_{t+1} B_n^T$.

After covering the whole testing sample, one can obtain the performance measures similarly to in-sample, using the following formulas:

$$SR = \frac{E[b^T \hat{F}_{t+1}]}{\sigma(b^T \hat{F}_{t+1})} \tag{13}$$

$$RMS_\alpha = \sqrt{\frac{\hat{\alpha}^T \hat{\alpha}}{N}}, \text{ where } \bar{\alpha}_n = \frac{1}{T}\sum_{t=1}^{T} \hat{\alpha}_{n,t+1} \tag{14}$$

$$\bar{\sigma}_e^2 = \frac{\frac{1}{N}\sum_{n=1}^N Var(\hat{\alpha}_n)}{\frac{1}{N}\sum_{n=1}^N Var(X_n)}. \tag{15}$$

4 Data

For replicating part of the research conducted by Lettau and Pelger [2020a], we use monthly portfolio returns ranging from November 1963 until December 2017 (T=650). First, we look at the returns of a small cross-section of double-sorted portfolios with 25 test assets. These portfolios are from the Kenneth R. French Data Library [French, 2022]. Next, we consider a larger cross-section of single-sorted anomaly portfolios. We use the same 370 anomaly decile portfolios as Kozak, Nagel, and Santosh [2020]. The portfolio returns can be found on the personal website of Serhiy Kozak, [Kozak, 2013]. An in-depth explanation of the used anomaly portfolios can be found in Kozak, Nagel, and Santosh [2020].

For the extension part of this paper, we look at the returns of 48 different industry portfolios. We use data ranging from January 1970 until April 2022 (T=628). The returns of the industry portfolios can be obtained from the Kenneth R. French Data Library, [French, 2022]. The industry portfolios consist of stocks included in the NYSE, AMAX, and NASDAQ. Every year, each stock is assigned to an industry portfolio at the end of June based on its SIC code. RP-PCA, PCA, and Fama-French are all applied to excess returns. Therefore, we have to subtract the risk-free rate from the returns of double-sorted portfolios, the single-sorted anomaly portfolios, and the industry portfolios. The risk-free rate is also obtained from the Kenneth R. French Data Library [French, 2022].

5 Results

5.1 Double-Sorted Portfolios

We compare the performance of the RP-PCA, PCA, and the Fama-French three-factor model for the double-sorted portfolios. We use eight sets of 25 double-sorted portfolios. The portfolios are sorted on size and book-to-market value, accruals, investment, profitability, short-term reversal, momentum, volatility, and idiosyncratic volatility. We set the RP-PCA parameter γ equal to 20, but the results are similar for different values of γ.

Table 1: Out-of-sample performance of RP-PCA, PCA and Fama-French models

	SR			RMS_α			$\bar{\sigma}_e$		
	RP-PCA	PCA	FF	RP-PCA	PCA	FF	RP-PCA	PCA	FF
SIZE&BM	**0.20**	0.18	0.15	**0.17**	0.17	0.18	7.97%	**7.91%**	7.97%
SIZE&ACC	**0.21**	0.12	0.15	**0.09**	0.11	0.10	6.74%	**6.44%**	7.17%
SIZE&INV	**0.26**	0.22	0.15	**0.13**	0.15	0.13	**6.95%**	7.00%	7.06%
SIZE&OP	0.13	0.14	**0.15**	**0.09**	0.10	0.11	**6.94%**	7.08%	8.54%
SIZE&ST-REV	**0.16**	0.11	0.15	**0.18**	0.11	0.19	7.89%	**7.86%**	10.88%
SIZE&MOM	**0.21**	0.18	0.15	**0.20**	0.21	0.30	**8.30%**	8.40%	13.76%
SIZE&IVOL	**0.29**	0.23	0.15	**0.16**	0.17	0.22	**6.22%**	6.24%	7.11%
SIZE&VOL	**0.27**	0.21	0.15	**0.18**	0.19	0.23	**6.27%**	6.30%	7.04%

Note. Eight sets of 25 double-sorted portfolios are used. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 20$. All models use three factors. The best-performing models are marked in bold.

As can be seen in Table 1, in terms of the Sharpe ratio, RP-PCA outperforms PCA in seven out of eight cases. What stands out is that for all different portfolios, the out-of-sample Sharpe ratio for Fama-French is equal. The out-of-sample Sharpe ratio is calculated solely based on the estimated factors, the mean, and the covariance matrix of the estimated factors. Factors of the Fama-French three-factor model are known and equal for different portfolios. Therefore, the out-of-sample Sharpe ratio for the Fama-French three-factor model is the same for all eight portfolios. This is not the case for RP-PCA or PCA, because those methods extract the factors from the excess portfolio returns. Different portfolios have different returns, resulting in different factors and Sharpe ratios. In terms of cross-sectional errors (RMS_α), RP-PCA outperforms PCA and Fama-French for all eight portfolios. The idiosyncratic variance is smaller for RP-PCA in five of the eight cases than for the other two methods. In the other three cases, PCA results in the smallest out-of-sample idiosyncratic variance. When considering all three performance measures, we can state that RP-PCA has better out-of-sample performance than PCA and Fama-French for the double-sorted portfolios.

But what causes this difference in terms of performance between RP-PCA and PCA? To better understand the difference, we take a closer look at the Size/Accruals and Size/Short-term reversal portfolios. Figure A.1 in Appendix A shows the out-of-sample Sharpe ratios, cross-sectional pricing errors, and the unexplained time-series variation as a function of the RP-PCA weight γ. This visualizes the behavior of RP-PCA for different values of γ. We consider using one factor up to six factors to illustrate the effect of each factor added.

Observable is that adding more factors results in better out-of-sample performance. For the Size-Accrual portfolios, the value of γ does not affect the performance when one or two factors are included. However, when adding a third factor, the Sharpe ratio increases and the cross-sectional pricing error decreases as γ increases. The idiosyncratic variation stays almost constant for different values of γ. So, RP-PCA outperforms PCA in terms of Sharpe ratio and cross-sectional pricing errors in a three-factor model. But what exactly causes RP-PCA to outperform PCA? To investigate this further, we look at the factor loadings of the first three factors for $\gamma = -1$ and $\gamma = 20$, corresponding to PCA and RP-PCA, respectively. Figure A.2 in Appendix A shows the heatmap of the factor loadings. The loadings of the first two factors of the Size/Accrual portfolios are almost identical for RP-PCA and PCA. This is in line with the fact that for a one- and two-factor model, the value of γ does not affect the performance of the model. The first factor has positive weights on all portfolios. This factor can thus be interpreted as a market factor. The second factor has negative weights for small-size stocks and positive weights for big-size stocks. This factor can be linked to the Small minus Big (SMB) factor from the Fama-French three-factor model. In this case, the factor is a Big minus Small (BMS) factor, but a different way of identification can cause this difference. The third RP-PCA factor shows a different pattern than the third PCA factor. The third RP-PCA factor has positive weights for high-accrual stocks and negative weights for low-accrual stocks. This factor is thus built similar to a Fama-French type factor and could be interpreted as a "high-minus-low accrual factor". The third PCA factor has no clear pattern and does not add much information. This illustrates why the three-factor RP-PCA model has a higher Sharpe ratio and lower cross-sectional pricing error than the three-factor PCA model.

Figure A.1 in Appendix A shows that for the Size/Short-term reversal portfolios, the value of γ does not affect the performance of RP-PCA when one factor is included. When three factors are included, the performance of RP-PCA is only slightly affected by the value of γ. When two factors are included in the model, the performance of RP-PCA is considerably affected by the value of γ. More specifically, when $\gamma > 5$, the performance increases in terms of Sharpe ratio and cross-sectional errors. Figure A.2 in Appendix A again compares the factor loadings for $\gamma = -1$ and $\gamma = 20$. Once more, the first RP-PCA and PCA factors have positive loadings for all portfolios and thus can be interpreted as market factors. The second and third factors are different for RP-PCA and PCA. The second RP-PCA factor is similar to the third PCA factor, and the third RP-PCA factor is similar to the second PCA factor. A possible explanation is that the difference in returns between high short-term reversal and low short-term reversal portfolios is larger than the difference in returns between big-size and small-size portfolios. Because of the larger return differences in the reversal dimension, the reversal factor better explains the cross-section of returns. RP-PCA tries to minimize the cross-sectional pricing errors, while PCA does not. Therefore, this factor is preferred by RP-PCA.

So, RP-PCA outperforms both PCA and Fama-French when applied to double-sorted portfolios. Furthermore, interpretation can be given to the RP-PCA factors. The first RP-PCA factor can be interpreted as a market factor. The second and third RP-PCA factors are built similarly to Fama-French type factors.

5.2 Single-Sorted Portfolios

Subsequently, we evaluate the performance of the RP-PCA method on a large cross-section of single-sorted anomaly portfolios from the Serhiy Kozak website, Kozak [2013]. The single-sorted anomaly portfolios are decile portfolios constructed from 37 anomaly characteristics, which results in a total of 370 portfolios. Lettau and Pelger [2020a] show that most information is contained in the extreme decile portfolios, the first and tenth deciles. Therefore, we use the set of 74 first and tenth decile portfolios for the major part of the analysis. As mentioned earlier, the sample ranges from November 1963 until December 2017 (T=650).

5.2.1 RP-PCA Versus PCA

We compare the fit of RP-PCA and PCA. We use the performance measures explained in the Methodology to compare the fit: maximum Sharpe ratio, cross-sectional pricing errors, and the unexplained time-series variation. We evaluate the fit both in-sample and out-of-sample using the set of 74 extreme decile portfolios. The fit is evaluated using only one factor, and we successively add factors up to ten factors. The results are shown in Figure B.1 in Appendix B. For each amount of factors considered, the in-sample and out-of-sample Sharpe ratios are higher for RP-PCA than for PCA. Also, both the in-sample and out-of-sample Sharpe ratios increase as you add more factors. The RP-PCA in-sample Sharpe ratio especially increases when adding the second and the fifth RP-PCA factor. After adding more than five factors, the in-sample Sharpe ratio stays around a constant level. The out-of-sample RP-PCA Sharpe ratio grows more steadily when adding the first five factors. Again, adding more factors does not increase the out-of-sample Sharpe ratio by much after adding the first five factors. The out-of-sample Sharpe ratios are marginally lower than the in-sample Sharpe ratios, indicating that the RP-PCA model is not overfitted.

Next, we look at the root-mean-squared cross-sectional pricing errors, RMS_{α}. The more factors you add, the lower the RMS_{α}. This again suggests that the model does not suffer from overfitting. Also, the second and fifth RP-PCA factors have the largest effect on the in-sample RMS$_\alpha$. This is not the case for the out-of-sample RMS$_\alpha$, which decreases more gradually. After adding the first five factors, adding more factors does not change the RP-PCA RMS$_\alpha$ notably. Lastly, we look at the idiosyncratic variation. The idiosyncratic variation decreases both in-sample and out-of-sample when adding more factors. The plots of the idiosyncratic variation are nearly identical for comparable RP-PCA and PCA models. The objective functions (2) and (7) of PCA and RP-PCA show that both methods try to explain as much common time-series variation as possible and thus minimize the idiosyncratic variation. Therefore the amount of unexplained idiosyncratic variation is similar for both methods. RP-PCA tries to explain as much cross-sectional time-series variation as possible, whereas PCA does not. Therefore RP-PCA gives lower root-mean-squared cross-sectional pricing errors than PCA.

Table 2 compares RP-PCA, PCA, and Fama-French models with three and five factors in terms of Sharpe Ratio, root-mean-squared cross-sectional pricing errors, and idiosyncratic variation. We use the sample including the 74 extreme first and tenth decile portfolios. The full sample (using all 370 portfolios) and another sample using 98 extreme first and tenth decile portfolios based on a smaller sample (November 1973-December 2017, T=530) are added for robustness in Table B.1 in Appendix B. For the three-factor models, RP-PCA dominates PCA and Fama-French in terms of Sharpe ratio and cross-sectional pricing errors. A similar result is obtained for the five-factor models. Overall, the results show that RP-PCA outperforms both PCA and Fama-French. More specifically, the RP-PCA model with five factors is the preferred model, resulting in the highest Sharpe ratio and the lowest cross-sectional pricing errors.

Table 2: In- and out-of-sample performance of RP-PCA, PCA, and Fama-French models.

	In-sample			Out-of-sample		
Model(k)	SR	RMS$_\alpha$	$\bar{\sigma}_e$	SR	RMS$_\alpha$	$\bar{\sigma}_e$
RP-PCA(3)	**0.37**	**0.23**	13.88%	**0.31**	**0.22**	**15.42%**
PCA(3)	0.23	0.27	**13.74%**	0.11	0.26	15.88%
FF(3)	0.21	0.31	17.49%	0.15	0.25	16.64%
RP-PCA(5)	**0.59**	**0.16**	10.43%	**0.50**	**0.15**	12.11%
PCA(5)	0.32	0.21	**10.30%**	0.24	0.19	**12.04%**
FF(5)	0.32	0.26	16.05%	0.32	0.19	13.91%

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$. The best-performing models are marked in bold.

5.2.2 Time-Series Versus Cross-Sectional Factors

After focussing on the performance of the different models, we now look at the individual factors. Again we use the sample of 74 extreme first and tenth decile portfolios. Table 3 shows the mean, variance, Sharpe ratio, and mean rank of the first ten factors obtained by RP-PCA and PCA.

Table 3: Individual factors obtained by using RP-PCA and PCA.

	RP-PCA				PCA			
Factor	Mean	Variance	SR	Mean Rank	Mean	Variance	SR	Mean Rank
1	5.02*	1932.09	0.11	1	4.83*	1941.87	0.11	1
2	2.32*	66.22	0.29	3	0.18	102.39	0.02	9
3	0.30	100.83	0.03	2	1.65*	69.03	0.20	2
4	0.10	65.34	0.01	4	1.05*	64.06	0.13	3
5	0.73*	26.30	0.14	5	0.83*	20.18	0.18	4
6	0.03	19.52	0.01	6	0.34*	19.39	0.08	7
7	0.14	17.93	0.03	7	0.79*	16.14	0.20	5
8	0.05	15.40	0.01	8	0.57*	15.08	0.15	6
9	0.04	13.53	0.01	9	0.20	13.46	0.01	8
10	0.03	11.95	0.01	10	0.04	11.95	0.01	10

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$. Significant means are marked with a star.

For both RP-PCA and PCA, the first factor has a way larger variance than the other factors, which is a common result for asset returns. The ranking of the PCA factors is solely based on their variance. This is not the case for RP-PCA. The variance of the second RP-PCA factor is lower than that of the third RP-PCA factor. However, the mean of the second RP-PCA factor is high and significant. Therefore, it is selected by RP-PCA before the factor with a higher variance. The second RP-PCA factor is priced higher than the third, but the third captures more co-movement and is not / less priced. Factors with significant means and high Sharpe ratios (factors 2 and 5) capture the cross-section of returns, while factors with low means and high variance (factors 3 and 4) capture co-movement. To elaborate further on the individual RP-PCA factors, we look at the fit for models using different sets of RP-PCA factors. The results are presented in Table 4.

Table 4: Fit for RP-PCA model with subset of factors.

	In-Sample			Out-of-sample		
Factors	SR	RMS	$\bar{\sigma}_e$	SR	RMS	$\bar{\sigma}_e$
[1,2,3,4,5]	0.59	0.16	10.43%	0.38	0.15	11.84%
[1]	0.11	0.33	20.75%	0.12	0.32	22.64%
[1,2,5]	0.57	0.23	17.07%	0.36	0.18	18.11%
[2,5]	0.41	1.69	74.72%	0.20	0.34	21.97%
[1,3,4]	0.12	0.33	13.93%	0.12	0.31	15.89%
[3,4]	0.03	0.66	93.06%	0.01	0.52	55.88%

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$.

The fit of the model using factors 1, 2, and 5 in terms of Sharpe ratio and cross-sectional pricing errors is pretty similar to the fit of the model using all five factors. In terms of explained common time-series variation, the model with all five factors outperforms the model using only factors 1, 2, and 5. This indicates that adding factors 3 and 4, having a low mean and high variance, increases the amount of explained common time-series variation.

The model using factors 1, 3, and 4 has a low Sharpe ratio and high cross-sectional pricing errors compared to the model using all five factors. This can be declared by the fact that factors 3 and 4, having a low mean and high variance, are not priced but capture co-movement. Therefore, the cross-section of the asset returns is not explained well. Because factors 3 and 4 capture co-movement well, the explained time-series variation of the model using factors 1, 3, and 4 is close to the model using all five factors (roughly 86% compared to 90% in-sample and 84% compared to 88% out-of-sample).

So, adding RP-PCA factors 2 and 5 to the model improves the Sharpe ratio and the cross-sectional fit but not the time-series fit. Adding RP-PCA factors 3 and 4 improves the time-series fit but not the Sharpe ratio and the cross-sectional fit. So, a model using factors 1, 2, and 5 can be used to capture the cross-section of the returns, and a model using factors 1, 3, and 4 can be used to capture the common time-series movement of the returns.

The results for the double-sorted and single-sorted portfolios are similar to the results obtained in Lettau and Pelger [2020a]. We find that RP-PCA outperforms PCA and Fama-French in terms of Sharpe ratio and pricing errors. Also, we can give interpretation to the factors. When applied to the double-sorted portfolios, the first RP-PCA factor can be interpreted as a market factor. The second and third RP-PCA factors are built similarly to Fama-French type factors. When applied to the single-sorted portfolios, the first factor again can be interpreted as a market factor. The second and fifth RP-PCA factors have high means and relatively low variance. The third and fourth factors have low means and relatively high variance. A model using RP-PCA factors 1, 2, and 5 can be used to capture the cross-section of the returns, and a model using RP-PCA factors 1, 3, and 4 can be used to capture the common time-series movement.

5.3 Industry portfolios

5.3.1 RP-PCA Versus PCA

After replicating part of the paper by Lettau and Pelger [2020a], we evaluate the RP-PCA method on industry portfolios. We compare the fit of RP-PCA with the fit of PCA and Fama-French models. We use a set of 48 industry portfolios from the Kenneth R. French Data Library French [2022]. The data ranges from January 1970 until April 2022 (T=628). Table 5 shows the in- and out-of-sample fit of RP-PCA, PCA, and Fama-French models using three and five factors.

Table 5: In- and out-of-sample performance of RP-PCA, PCA, and Fama-French models.

	In-sample			Out-of-sample		
Model(k)	SR	RMS_α	$\bar{\sigma}_e$	SR	RMS_α	$\bar{\sigma}_e$
RP-PCA(3)	0.15	**0.21**	32.98%	0.01	0.29	39.00%
PCA(3)	0.13	0.22	**32.92%**	0.05	0.27	**38.79%**
FF(3)	**0.19**	0.25	44.09%	**0.18**	**0.23**	43.76%
RP-PCA(5)	0.21	**0.16**	26.82%	0.05	0.23	**31.77%**
PCA(5)	0.17	0.18	**26.75%**	0.08	0.23	31.95%
FF(5)	**0.33**	0.31	41.75%	**0.33**	**0.20**	38.38%

Note. The sample consists of 48 industry portfolios. The sample ranges from January 1970 until April 2022 (T=628). RP-PCA parameter $\gamma = 10$. Best-performing models are marked in bold.

Table 5 shows that in terms of Sharpe ratio, the Fama-French model outperforms RP-PCA and PCA both in-sample and out-of-sample. Especially the out-of-sample Sharpe ratios for RP-PCA and PCA are very low compared to the Sharpe ratio implied by Fama-French. The in-sample cross-sectional pricing errors are the lowest for RP-PCA models. However, Fama-French outperforms RP-PCA and PCA in terms of out-of-sample cross-sectional pricing errors. PCA explains more time-series variation compared to RP-PCA and Fama-French. So, when only looking at Table 5, one would suggest that the Fama-French five-factor model is the preferred model. Remarkable is the fact that the out-of-sample cross-sectional pricing errors are lower for Fama-French than for RP-PCA. When looking at the objective function of RP-PCA, RP-PCA aims to find factors that minimize the cross-sectional pricing errors. The Fama-French factors, on the other hand, are based on economic theory. To further investigate the fit of RP-PCA and PCA, we evaluate the fit using only one factor and successively add factors up to ten factors. The results are shown in Figure C.1 in Appendix C.

When looking at the in-sample performance of both RP-PCA and PCA, the performance increases when more factors are added. Also, RP-PCA performs better than PCA in terms of in-sample Sharpe ratio and cross-sectional pricing errors. After adding the sixth factor, the in-sample RP-PCA Sharpe ratio and the in-sample RP-PCA cross-sectional pricing error do not increase much when adding additional factors. The in-sample unexplained time-series variation is equal for RP-PCA and PCA. This can again be declared by the fact that both methods try to minimize the unexplained time-series variation in their objective function. So, in-sample, RP-PCA has a better fit than PCA. When looking at the out-of-sample fit of RP-PCA and PCA, the patterns of the Sharpe ratio and the cross-sectional pricing errors look rather odd when adding factors. A model with more than one RP-PCA factor results in a lower Sharpe ratio than a model with only the first RP-PCA factor. The third RP-PCA factor, in particular, has a big negative impact on the Sharpe ratio. Also, for the out-of-sample cross-sectional pricing error, adding the second and third RP-PCA factors increases the cross-sectional pricing error. In terms of Sharpe ratio and cross-sectional pricing errors, PCA outperforms RP-PCA out-of-sample.

The out-of-sample maximum Sharpe ratio consists of two components (recall Formula (13) in the Methodology). To get a better insight into the out-of-sample RP-PCA Sharpe ratio, we look at the behavior of the two components of the Sharpe ratio separately when additional factors are added. The standard deviation of the expected returns (denominator of the Sharpe ratio) can be seen as a measure of volatility. The two separate components, calculated using only one up to ten factors, are shown in Figure C.2 in Appendix C. This figure shows that adding more factors gradually increases the volatility. However, the expected return decreases when more factors are added, which decreases the Sharpe ratio. Especially adding the third RP-PCA factor is disastrous for the expected returns of the industry portfolios. To get a better insight into the RP-PCA (and PCA) factors, we try to interpret them by looking at the individual factors and factor loadings.

5.3.2 Interpretation of the Factors

Bro and Smilde [2014] state that "loadings define what a principal component represents". Therefore, by looking at the factor loadings, we can see which industries are represented by which factors. This way, we can give interpretation to the RP-PCA and PCA factors. The 48 industry portfolios can be divided into five broader industries based on their SIC code. We order the industry portfolios in such a way that the first x industries belong to broader industry 1, the next y industries belong to broader industry 2, etc. We consider the same five broader industries as on the website of Kenneth R. French [French, 2022]: Consumer goods, Manufacturing, High-technology, Health, and Other. The number of industry portfolios belonging to each of the five broader industries is shown in Table 6. The exact ranking of the industry portfolios, including the portfolio names, can be found in Table C.1 in Appendix C.

Table 6: The number of industry portfolios belonging to each of the five broader industries.

Industries	Consumer Goods	Manufacturing	High-Tech	Health	Other
Portfolios included	13	16	4	3	12

Note. The total number of industry portfolios is equal to 48. Based on their SIC code, the industry portfolios are assigned to one of the broader industries.

It could be the case that the factors explaining the industry portfolios can be related to the five broader industries. Therefore, we construct heatmaps in which each column corresponds to one of the five broader industries. Figure C.3 in Appendix C shows heatmaps of the first five RP-PCA factor loadings with $\gamma = 10$ and PCA factor loadings with $\gamma = -1$. We consider the first five factors because the fifth factor is the last factor that has a substantial incremental effect on the out-of-sample Sharpe ratio for either RP-PCA or PCA.

The first thing that stands out when looking at Figure C.3 is that the factor loadings of RP-PCA and PCA display similar patterns for the first five factors. For RP-PCA and PCA, the first factor has positive loadings for all portfolios. This factor can thus be interpreted as a market factor. The loadings of the second factor are mostly somewhere in the range of [-0.1, 0.1]. However, the factor loadings of industry portfolios 17, 25, 26, 39, and 40 have relatively high positive values compared to the others. These portfolios correspond to the industries of steel, coal, oil, gold, and mines, respectively. Especially the factor loadings of the coal and the gold industry are very high (0.51 and 0.66). These industries are not categorized together into the same broader industry based on their SIC code. However, the industries can be linked to each other based on (economic) theory. All five industries are related to mining or subtracting finite goods from the earth. For the production of steel, large amounts of iron ore are required, according to ArcelorMittal [2022], one of the largest steel producers in the world. Iron ore is excavated in mines and therefore related to the mining industry. Also, the gold industry depends on the amount of gold excavated in gold mines. Coal also comes from coal mines and is thus strongly related to the mining industry. Oil does not come from mines. However, it is extracted from underground reservoirs. Because the second factor represents industries related to mining and sourcing of raw materials, this factor can be interpreted as a "mining factor" or a "sourcing of raw materials factor".

The loadings of the third RP-PCA and PCA factor are again similar. Most loadings are somewhere in the range of [-0.2, 0.2]. However, two factor loadings stand out. The factor loadings of the coal and the gold industry portfolios are equal to about -0.6 and 0.6, respectively. This factor could explain the difference between the gold and coal industries. However, we can not give clear interpretation to the factor. Also, no literature about a potential negative relationship between these two industries can be found. The loadings of the fourth RP-PCA factor display a similar pattern to those of the fourth PCA factor. However, the loadings do not show a clear pattern. Also, the loadings can not be related to co-moving industries or different behavior of industries. The loadings of the fifth RP-PCA factor display a similar pattern as the loadings of the fifth PCA factor. The fifth RP-PCA factor is heavily loaded on the third industry category. This is the High-Tech industry. Therefore, we can say that the fifth factor is linked to the High-Tech industry.

After evaluating the factor loadings, we conclude that RP-PCA can not detect any factors that PCA is not able to detect. Also, we find that we can give interpretation to the first, second, and fifth RP-PCA factors. The first factor has positive weights for all portfolios and thus can be interpreted as a market factor. The second factor can be related to mining or the sourcing of raw materials from the earth. The fifth factor can be linked to the High-Tech industry. No clear interpretation can be found for the third and fourth RP-PCA factors. Next, we consider the properties of the individual factors.

5.3.3 Time-Series vs Cross-Sectional Factors

Table 7 shows the mean, variance, Sharpe ratio and mean rank of the first five RP-PCA and PCA factors. The variance of the first factor is much larger than the variance of the other factors, which is a typical result for asset returns. As expected, the PCA factors are ranked based on their variance. However, the RP-PCA factors are also ranked based on their variance. The RP-PCA factor with the lowest mean is selected as the second factor by RP-PCA. This is not as expected, as RP-PCA also takes into account the mean of the factor. It could be the case that the RP-PCA parameter $\gamma = 10$ is too low. Therefore, we also consider the individual factors when γ is equal to 50 and 100. The results are shown in Table C.2 in Appendix C.

Table 7: Individual factors obtained by using RP-PCA and PCA.

	RP-PCA				PCA			
Factor	Mean	Variance	SR	Mean Rank	Mean	Variance	SR	Mean Rank
1	4.46	1214.88	0.13	1	4.38	1218.37	0.13	1
2	0.01	146.53	0.00	5	0.01	146.53	0.00	5
3	0.54	87.17	0.06	4	0.31	88.85	0.03	4
4	0.67	74.18	0.07	2	0.52	74.97	0.06	2
5	0.63	59.88	0.08	3	0.71	58.73	0.09	3

Note. The sample consists of 48 industry portfolios. The sample ranges from January 1970 until April 2022 (T=628). RP-PCA parameter $\gamma = 10$.

Table C.2 in Appendix C shows that, even when the RP-PCA parameter γ is increased by a factor of five or ten, RP-PCA is still not able to select factors with a high mean. The factor with mean rank five is still selected as the second factor by RP-PCA when the RP-PCA weight γ is increased. So, RP-PCA can not detect factors that PCA is not able to detect, even when γ is increased. This suggests that the RP-PCA method is not an improvement compared to PCA for explaining industry portfolios. To elaborate further on the individual RP-PCA factors, we look at the fit for models using different sets of RP-PCA factors. The results are presented in Table 8.

Table 8: Fit for RP-PCA model with subset of factors.

	In-Sample			Out-of-sample		
Factors	SR	RMS	$\bar{\sigma}_e$	SR	RMS	$\bar{\sigma}_e$
[1]	0.13	0.22	43.77%	0.17	0.24	50.21%
[1,2]	0.13	0.22	37.01%	0.17	0.25	42.71%
[1,2,3]	0.15	0.21	32.98%	0.01	0.29	38.79%
[1,2,4]	0.16	0.21	33.57%	0.18	0.24	38.47%
[1,2,5]	0.16	0.21	34.22%	0.21	0.22	38.96%
[1,4,5]	0.19	0.19	37.57%	0.22	0.21	42.89%
[1,2,4,5]	0.19	0.19	30.81%	0.22	0.20	35.14%
[1,2,3,4,5]	0.21	0.16	26.82%	0.05	0.23	31.77%

Note. The sample consists of 48 industry portfolios. The sample ranges from January 1970 until April 2022 (T=628). RP-PCA parameter $\gamma = 10$.

Table 8 shows that the in-sample performance measures gradually improve when adding more factors. The third RP-PCA factor does not harm the in-sample performance. However, the out-of-sample Sharpe ratio heavily decreases when adding the third RP-PCA factor. This can be seen by the difference in out-of-sample Sharpe ratio between the models with factors [1,2] and [1,2,3] and the models with factors [1,2,4,5] and [1,2,3,4,5]. Also, in terms of cross-sectional pricing errors, the out-of-sample performance decreases when adding the third RP-PCA factor. Therefore, we prefer a model without the third RP-PCA factor. More specifically, the preferred model is a model with the first, second, fourth, and fifth RP-PCA factor. This model results in the highest Sharpe ratio, the smallest cross-sectional pricing errors, and performs well in terms of explained common time-series movement. When comparing this model with the Fama-French five-factor model in Table 5, Fama-French outperforms this model in terms of Sharpe ratio. However, in terms of cross-sectional pricing errors and explained common time-series movement RP-PCA outperforms the Fama-French five-factor model. So, the RP-PCA model with factors 1, 2, 4, and 5 is the best fit for industry portfolios in terms of explaining the cross-section and the common time-series movement but does not lead to the highest Sharpe ratio.

6 Conclusion and Discussion

In this paper, we compare the newly proposed RP-PCA method with PCA and Fama-French. RP-PCA tries to capture as much common time-series movement as possible and minimizes the cross-sectional pricing errors. PCA only tries to capture as much common time-series movement as possible. When applied to double-sorted portfolios, RP-PCA outperforms PCA and Fama-French. Also, interpretation can be given to the RP-PCA factors. The first factor can be interpreted as a market factor. The second and third RP-PCA factors are built similarly to Fama-French type factors. For the set of single-sorted anomaly portfolios, RP-PCA outperforms both PCA and Fama-French in terms of Sharpe ratio and root-mean-squared pricing errors. We find that the first five factors found by RP-PCA are sufficient to explain the returns of 370 single-sorted decile portfolios. Next to this, we can give interpretation to the factors found. The first factor has positive weights for all portfolios and thus can be interpreted as a market factor. The second and fifth factors have a high mean and low variance and, thus, a high Sharpe ratio. Therefore, a model using the first, second, and fifth factors best explains the cross-sectional return differences. The third and fourth factors have high variance, a low mean, and thus a low Sharpe ratio. A model using the first, third, and fourth factors captures most of the common time-series movement. The analysis of double-sorted and single-sorted anomaly portfolios is a replication of the paper by Lettau and Pelger [2020a]. The results obtained and the conclusions drawn are very similar.

We extend the paper of Lettau and Pelger by applying the RP-PCA method to a set of 48 industry portfolio returns. We find that the out-of-sample performance of RP-PCA is relatively weak. This is mainly due to the presence of the third RP-PCA factor. When removing the third RP-PCA factor, the performance of RP-PCA increases. A model with RP-PCA factors 1, 2, 4, and 5 performs best among all RP-PCA models. This model has the best fit for industry portfolios. Also, there is not much difference between RP-PCA and PCA in terms of factors and factor loadings. It could be the case that there is not much risk-premium contained in industry portfolios. Another possible explanation could be that the cross-section is too small. However, the fact that RP-PCA does well on the set of 74 extreme first and tenth decile anomaly portfolios weakens this declaration.

RP-PCA takes into account the mean of a factor when selecting factors. However, even when increasing the RP-PCA parameter γ by a factor of ten, RP-PCA is still unable to detect other factors than PCA. The first factor found by RP-PCA and PCA can be interpreted as a market factor. The second factor can be interpreted as a factor related to mining or extracting finite goods from the earth. The fifth factor can be linked to the High-Tech industry. For the third and fourth RP-PCA factors, we can find no interpretation.

So, RP-PCA does well for the double-sorted and single-sorted anomaly portfolios. A model with RP-PCA factors 1, 2, 4 and 5 is the best fit for industry portfolios as it results in the lowest cross-sectional pricing errors and best explains the common time series variation. However, the Fama-French five-factor model results in the highest Sharpe ratio. Also, applying RP-PCA and PCA to industry portfolios results in similar factors and factor loadings, even when the RP-PCA parameter is increased. The first RP-PCA factor can be interpreted as a market factor, the second RP-PCA factor can be related to mining or the extraction of finite goods from the earth, and the fifth RP-PCA factor can be linked to the High-Tech industry.

References

ArcelorMittal. Making steel. `https://corporate.arcelormittal.com/about/making-steel`, 2022. Accessed: 03-06-2022.

Bai and Ng . Rank Regularized Estimation of Approximate Factor Models. *Journal of Econometrics*, 212(1):78–96, 2019.

Bro and Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014.

Cavaglia, Brightman, and Aked. The Increasing Importance of Industry Factors. *Financial Analysts Journal*, 56(5):41–54, 2000.

Cavaglia, Hodrick, Vadim, and Zhang. Pricing the Global Industry Portfolios, 2002.

Elhadary. Using the Industry-Based Fama-French Model to Evaluate Industry Portfolios. *Journal of Accounting and Finance*, 21(2):24–40, 2021.

Fama and French. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

Fan, Liao, and Wang. Projected Principal Component Analysis in Factor Models. *Annals of Statistics*, 44(1):219, 2016.

Feng, Giglio, and Xi. Taming the Factor Zoo. *Fama-Miller Working Paper*, 24070, 2017.

French. Kenneth R. French Data Library, 2022. URL `mba.tuck.dartmouth.edu/pages/faculty/ken.french/datalibrary.html`. Accessed: 10/05/2022.

Kelly, Pruitt, and Su. Instrumented Principal Component Analysis. *Available at SSRN 2983919*, 2020.

Kozak. Serhiy Kozak Data, 2013. URL `https://sites.google.com/site/serhiykozak/data`. Accessed: 10/05/2022.

Kozak, Nagel, and Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 135 (2):271–292, 2020.

Lettau and Pelger. Factors That Fit the Time Series and Cross-Section of Stock Returns. *The Review of Financial Studies*, 33(5):2274–2325, 2020a.

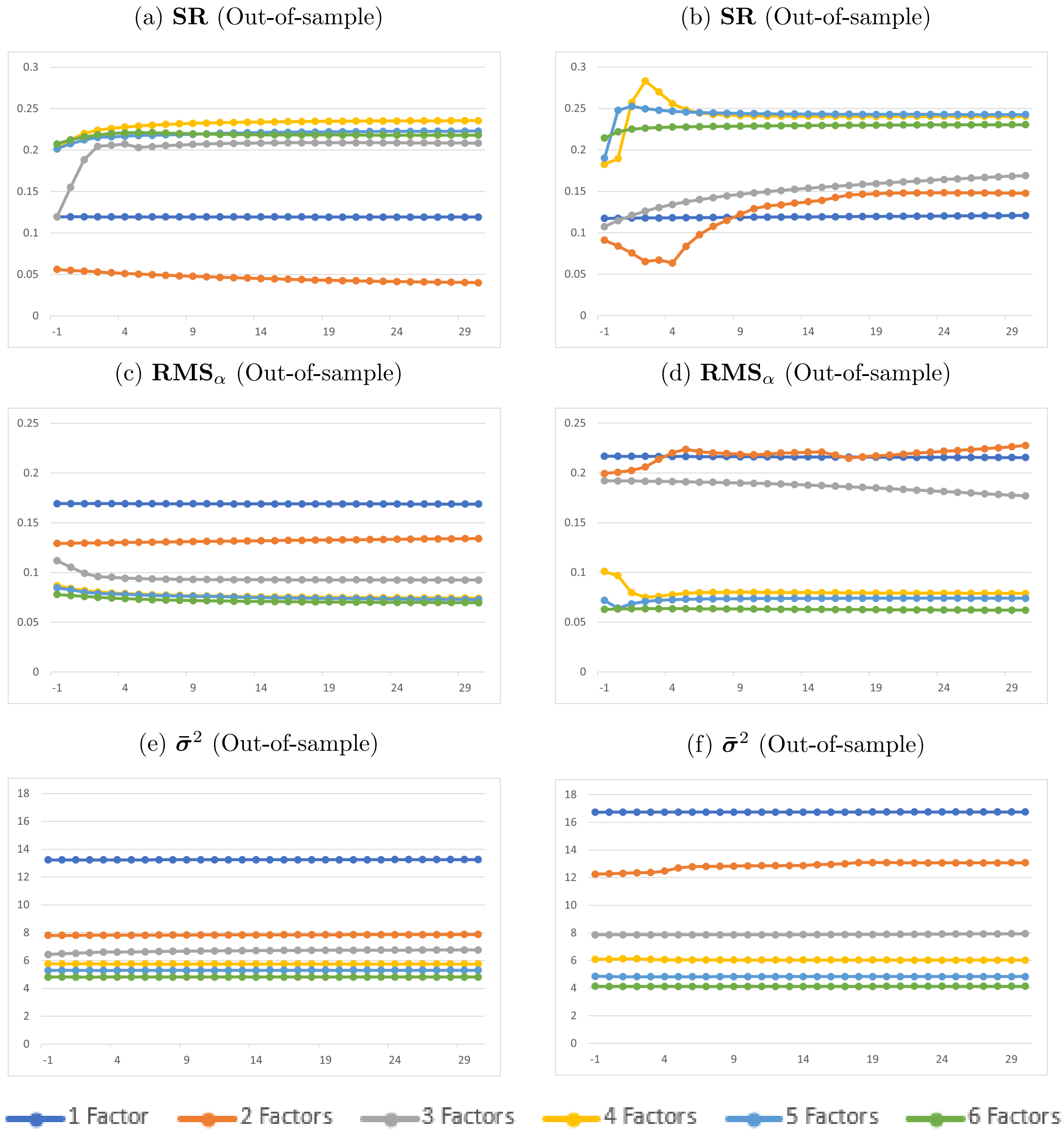
Lettau and Pelger. Estimating Latent Asset-Pricing Factors. *Journal of Econometrics*, 218(1): 1–31, 2020b.

Ross. The Arbitrage Theory of Capital Asset Pricing. In *Handbook of the Fundamentals of Financial Decision Making: Part I*, pages 11–30. World Scientific, 2013.

Stock and Watson. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

Appendix A. Double-Sorted Portfolios

Figure A.1: Out-of-sample performance as a function of γ



(a) **SR** (Out-of-sample)

(b) **SR** (Out-of-sample)

(c) $\mathbf{RMS}_{\alpha}$ (Out-of-sample)

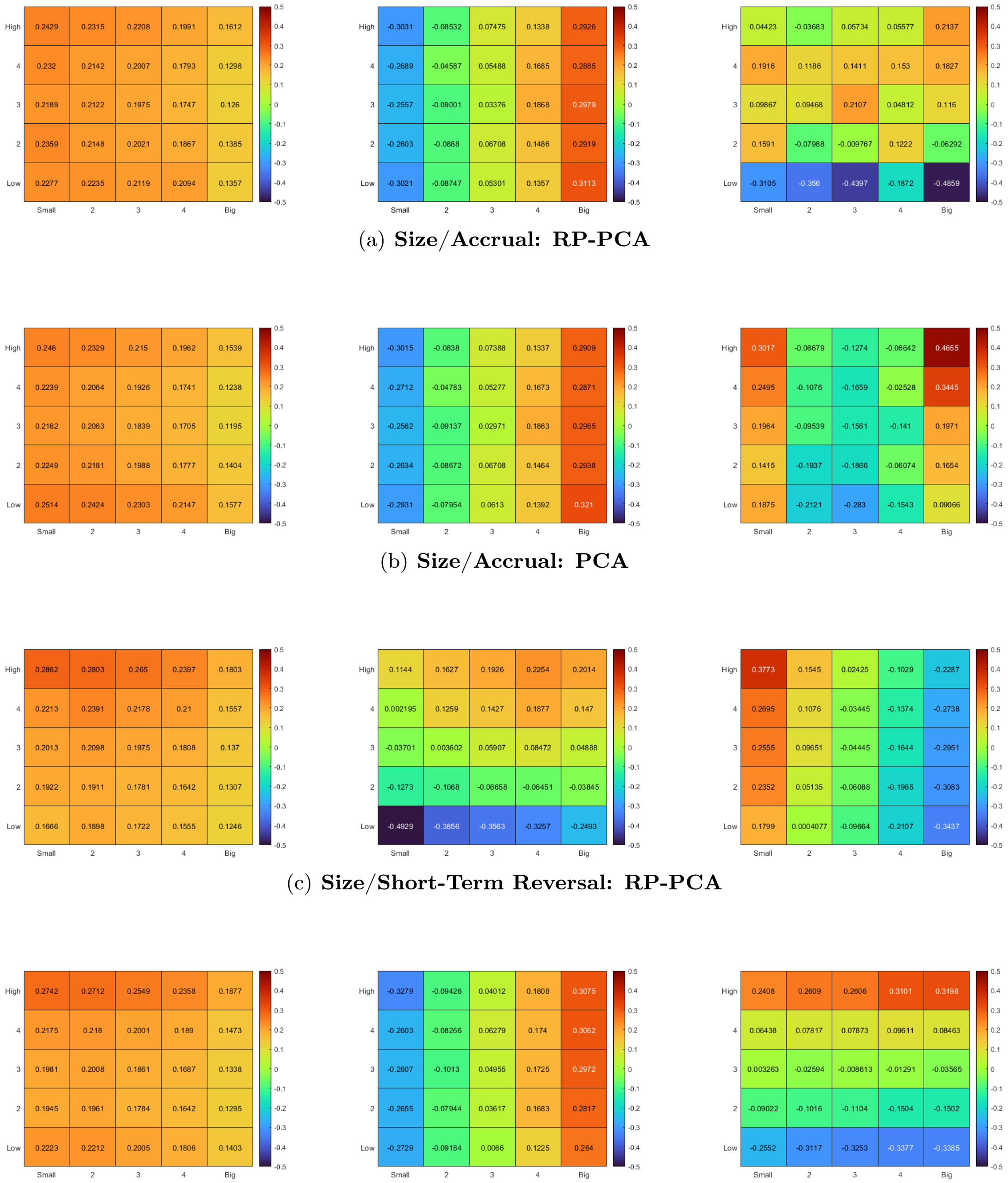
(d) $\mathbf{RMS}_{\alpha}$ (Out-of-sample)

(e) $\bar{\boldsymbol{\sigma}}^2$ (Out-of-sample)

(f) $\bar{\boldsymbol{\sigma}}^2$ (Out-of-sample)

Note. The left-sided figures show the results for the Size/Accrual portfolios. The right-sided figures show the results for the Size/Short-Term Reversal portfolios. The sample ranges from November 1963 until December 2017 (T=650).

Figure A.2: Heatmap of the first three factor loadings using RP-PCA and PCA for Size/Accrual portfolios and Size/Short-term reversal portfolios.



	Small	2	3	4	Big
High	0.2429	0.2315	0.2208	0.1991	0.1612
4	0.232	0.2142	0.2007	0.1793	0.1298
3	0.2189	0.2122	0.1975	0.1747	0.126
2	0.2359	0.2148	0.2021	0.1867	0.1385
Low	0.2277	0.2235	0.2119	0.2094	0.1357

	Small	2	3	4	Big
High	-0.3031	-0.08532	0.07475	0.1338	0.2926
4	-0.2689	-0.04587	0.05488	0.1685	0.2885
3	-0.2557	-0.09001	0.03376	0.1868	0.2979
2	-0.2603	-0.0888	0.06708	0.1486	0.2919
Low	-0.3021	-0.08747	0.05301	0.1357	0.3113

	Small	2	3	4	Big
High	0.04423	-0.03683	0.05734	0.05577	0.2137
4	0.1916	0.1186	0.1411	0.153	0.1827
3	0.09867	0.09468	0.2107	0.04812	0.116
2	0.1591	-0.07988	-0.009767	0.1222	-0.06292
Low	-0.3105	-0.356	-0.4397	-0.1872	-0.4859

(a) **Size/Accrual: RP-PCA**

	Small	2	3	4	Big
High	0.246	0.2329	0.215	0.1962	0.1539
4	0.2239	0.2064	0.1926	0.1741	0.1238
3	0.2162	0.2063	0.1839	0.1705	0.1195
2	0.2249	0.2181	0.1988	0.1777	0.1404
Low	0.2514	0.2424	0.2303	0.2147	0.1577

	Small	2	3	4	Big
High	-0.3015	-0.0838	0.07388	0.1337	0.2909
4	-0.2712	-0.04783	0.05277	0.1673	0.2871
3	-0.2562	-0.09137	0.02971	0.1863	0.2965
2	-0.2634	-0.08672	0.06708	0.1464	0.2938
Low	-0.2931	-0.07954	0.0613	0.1392	0.321

	Small	2	3	4	Big
High	0.3017	-0.06679	-0.1274	-0.06642	0.4655
4	0.2495	-0.1076	-0.1659	-0.02528	0.3445
3	0.1964	-0.09539	-0.1561	-0.141	0.1971
2	0.1415	-0.1937	-0.1866	-0.06074	0.1654
Low	0.1875	-0.2121	-0.283	-0.1543	0.09066

(b) **Size/Accrual: PCA**

	Small	2	3	4	Big
High	0.2862	0.2803	0.265	0.2397	0.1803
4	0.2213	0.2391	0.2178	0.21	0.1557
3	0.2013	0.2098	0.1975	0.1808	0.137
2	0.1922	0.1911	0.1781	0.1642	0.1307
Low	0.1666	0.1898	0.1722	0.1555	0.1246

	Small	2	3	4	Big
High	0.1144	0.1627	0.1926	0.2254	0.2014
4	0.002195	0.1259	0.1427	0.1877	0.147
3	-0.03701	0.003602	0.05907	0.08472	0.04888
2	-0.1273	-0.1068	-0.06658	-0.06451	-0.03845
Low	-0.4929	-0.3856	-0.3563	-0.3257	-0.2493

	Small	2	3	4	Big
High	0.3773	0.1545	0.02425	-0.1029	-0.2287
4	0.2695	0.1076	-0.03445	-0.1374	-0.2738
3	0.2555	0.09651	-0.04445	-0.1644	-0.2951
2	0.2352	0.05135	-0.06088	-0.1985	-0.3083
Low	0.1799	0.0004077	-0.09664	-0.2107	-0.3437

(c) **Size/Short-Term Reversal: RP-PCA**

	Small	2	3	4	Big
High	0.2742	0.2712	0.2549	0.2358	0.1877
4	0.2175	0.218	0.2001	0.189	0.1473
3	0.1981	0.2008	0.1861	0.1687	0.1338
2	0.1945	0.1961	0.1784	0.1642	0.1295
Low	0.2223	0.2212	0.2005	0.1806	0.1403

	Small	2	3	4	Big
High	-0.3279	-0.09426	0.04012	0.1808	0.3075
4	-0.2603	-0.08266	0.06279	0.174	0.3062
3	-0.2607	-0.1013	0.04955	0.1725	0.2972
2	-0.2655	-0.07944	0.03617	0.1683	0.2817
Low	-0.2729	-0.09184	0.0066	0.1225	0.264

	Small	2	3	4	Big
High	0.2408	0.2609	0.2606	0.3101	0.3198
4	0.06438	0.07817	0.07873	0.09611	0.08463
3	0.003263	-0.02594	-0.008613	-0.01291	-0.03565
2	-0.09022	-0.1016	-0.1104	-0.1504	-0.1502
Low	-0.2552	-0.3117	-0.3253	-0.3377	-0.3385

(d) **Size/Short-Term Reversal: PCA**

Note. The Size/Accrual and Size/Short-Term Reversal portfolios are used. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 20$.

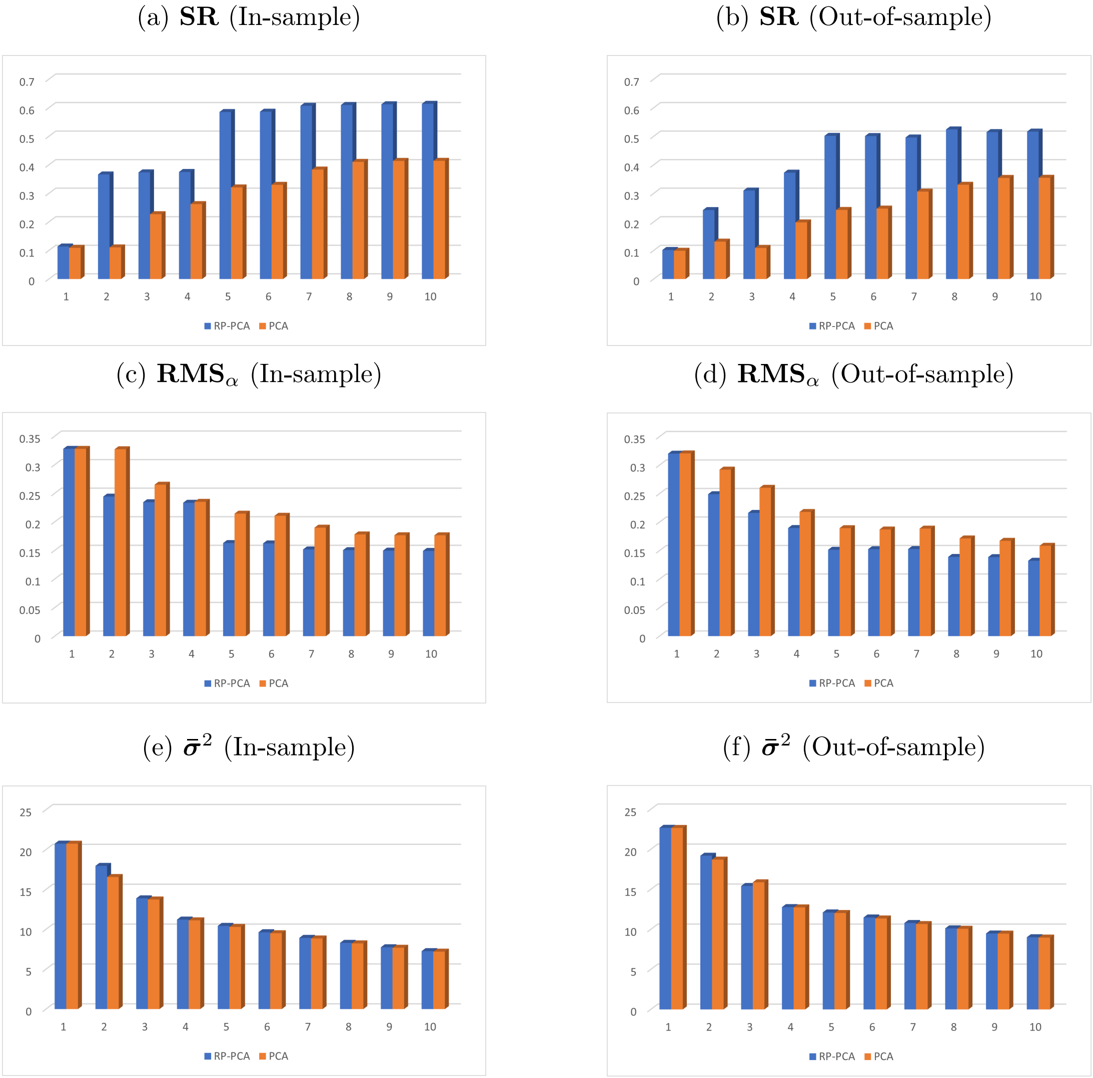
Appendix B. Single-Sorted Anomaly Portfolios

Table B.1: In- and Out-of-sample performance of RP-PCA, PCA, and Fama-French models.

Model(k)	SR	RMS_α	$\bar{\sigma}_e$	SR	RMS_α	$\bar{\sigma}_e$
		In-sample			Out-of-sample	
			Panel 1: 370 portfolios			
RP-PCA(3)	**0.24**	**0.17**	12.77%	**0.20**	**0.15**	**14.41%**
PCA(3)	0.17	0.17	**12.68%**	0.13	0.15	14.49%
FF(3)	0.21	0.18	14.61%	0.15	0.16	14.89%
RP-PCA(5)	**0.56**	**0.13**	10.80%	**0.47**	**0.12**	12.72%
PCA(5)	0.25	0.14	**10.68%**	0.18	0.14	**12.59%**
FF(5)	0.32	0.16	13.60%	0.32	0.13	13.74%
			Panel 2: 98 portfolios			
RP-PCA(3)	**0.36**	**0.22**	13.56%	**0.20**	**0.23**	**16.37%**
PCA(3)	0.25	0.25	**13.42%**	0.14	0.24	16.60%
FF(3)	0.21	0.32	17.04%	0.15	0.26	17.67%
RP-PCA(5)	**0.54**	**0.17**	10.38%	**0.42**	**0.15**	13.06%
PCA(5)	0.33	0.20	**10.27%**	0.22	0.19	**12.91%**
FF(5)	0.34	0.23	15.27%	0.32	0.20	14.91%

Note. The sample of Panel 1 consists of 370 decile portfolios and ranges from November 1963 until December 2017 (T=650). The sample of Panel 2 consists of 98 extreme decile portfolios and ranges from November 1973 until December 2017 (T=530). RP-PCA parameter $\gamma = 10$. The best-performing models are marked in bold.

Figure B.1: Maximum Sharpe ratios, root-mean-squared pricing errors, and unexplained idiosyncratic variation for a different number of factors.



(a) **SR** (In-sample)

(b) **SR** (Out-of-sample)

(c) $\mathbf{RMS}_\alpha$ (In-sample)

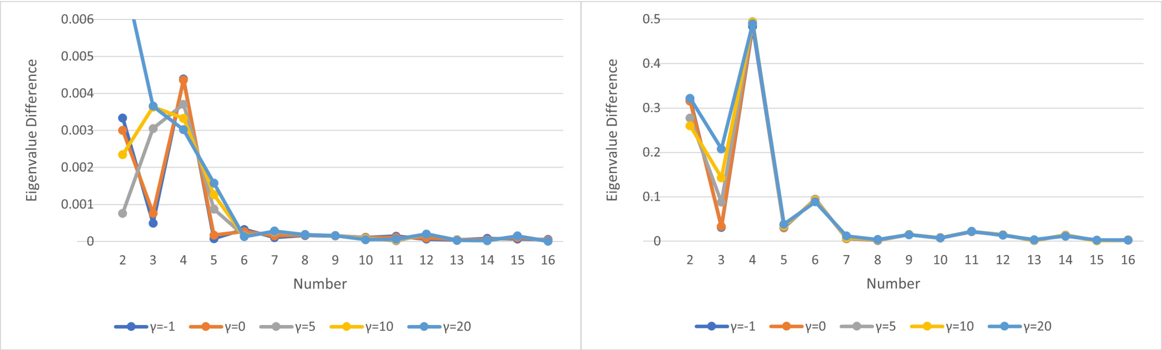
(d) $\mathbf{RMS}_\alpha$ (Out-of-sample)

(e) $\bar{\boldsymbol{\sigma}}^2$ (In-sample)

(f) $\bar{\boldsymbol{\sigma}}^2$ (Out-of-sample)

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$. The number of used factors can be found on the x-axis.

Figure B.2: Successive eigenvalue differences for different values of γ

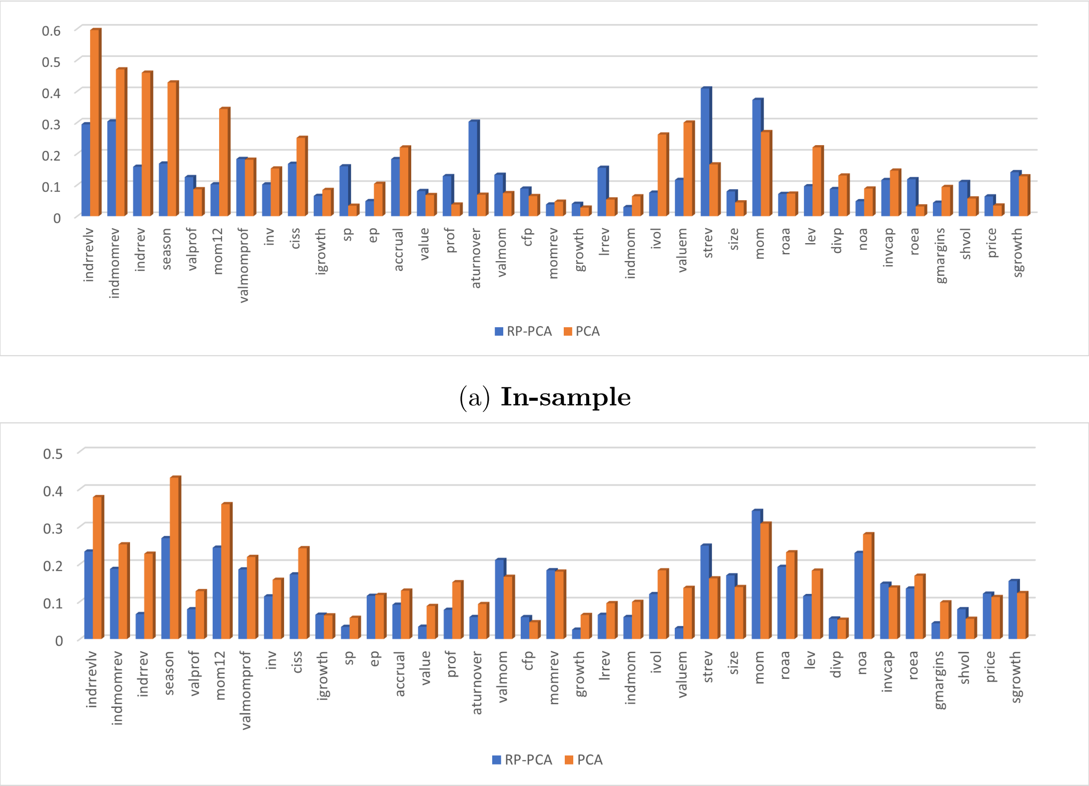


(a) **N=74**

(b) **N=370**

Note. Consecutive eigenvalue differences of $\frac{1}{T}X^TX + \gamma\overline{X}\overline{X}^T$ for different values of γ. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650).

Figure B.3: In- and out-of-sample mean-squared cross-sectional pricing errors of RP-PCA and PCA.

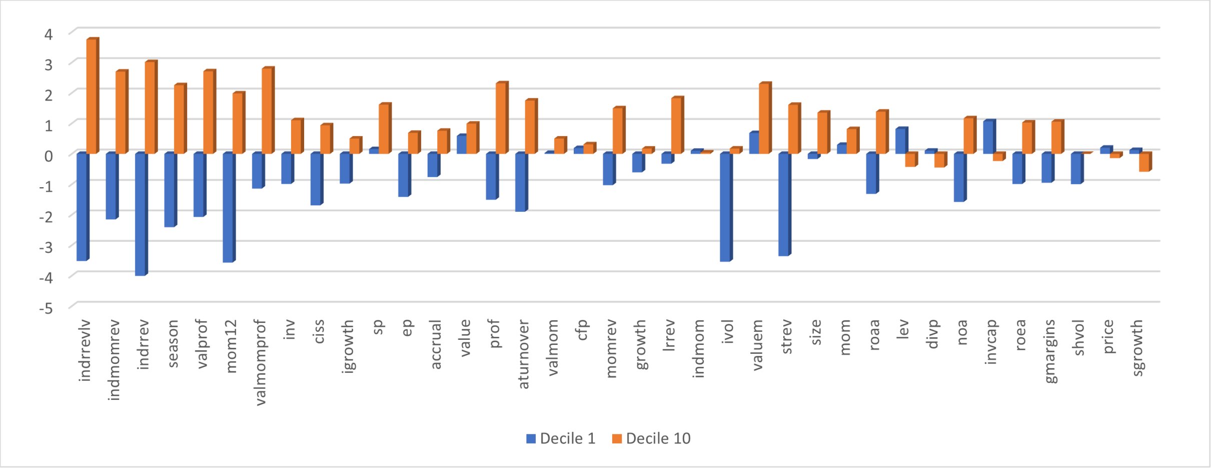


(a) **In-sample**

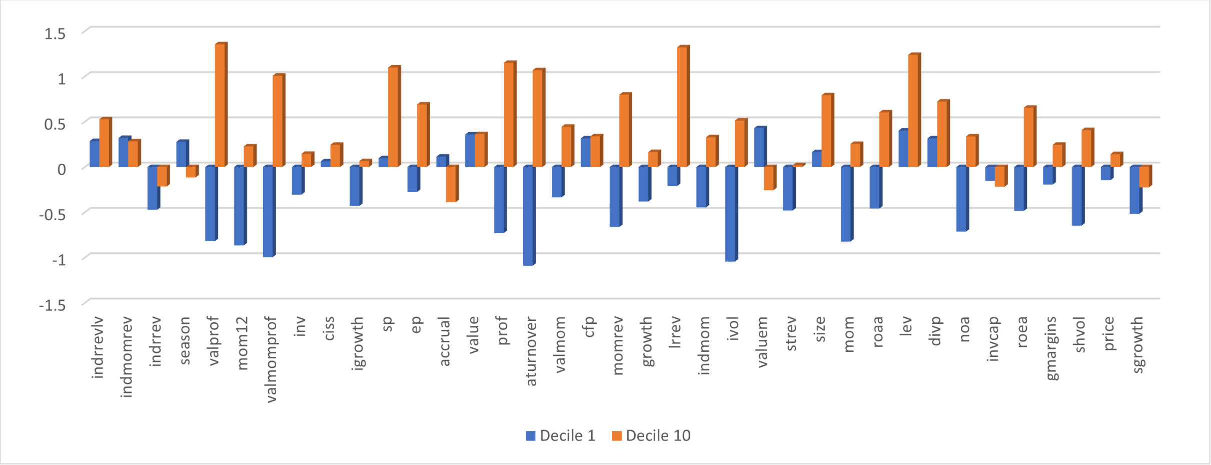
(b) **Out-of-sample**

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$. Both RP-PCA and PCA use five factors. The portfolios are ranked based on their Sharpe ratio.

Figure B.4: Portfolio weights in the RP-PCA and PCA SDF.



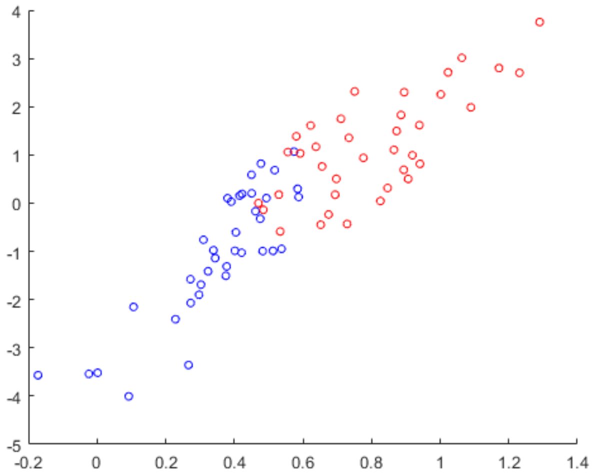
(a) **RP-PCA**



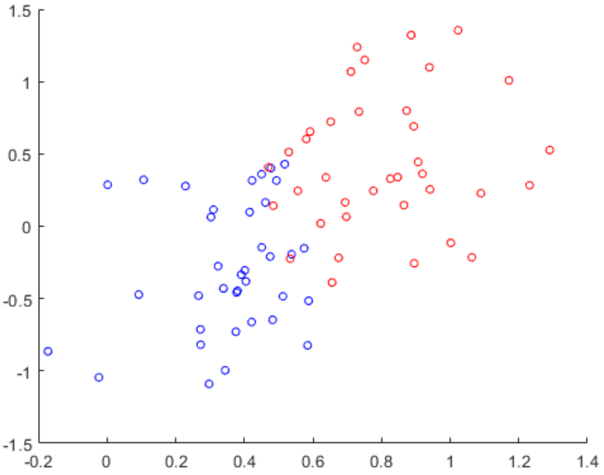
(b) **PCA**

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$. Both RP-PCA and PCA use five factors. The portfolios are ranked based on their Sharpe ratio.

Figure B.5: SDF weights plotted against the average returns (in % per month)



(a) **RP-PCA. Corr = 0.89**



(b) **PCA. Corr = 0.56**

Note. The sample consists of 74 extreme decile portfolios. The sample ranges from November 1963 until December 2017 (T=650). RP-PCA parameter $\gamma = 10$. Both RP-PCA and PCA use five factors. Blue dots correspond to the first decile portfolios. Red dots correspond to the last decile portfolios.

Appendix C. Industry Portfolios

Table C.1: The 48 different industry portfolios and the category they belong to.

	Portfolio	Category		Portfolio	Category
1	Agric	1	25	Coal	2
2	Food	1	26	Oil	2
3	Soda	1	27	Util	2
4	Beer	1	28	Paper	2
5	Smoke	1	39	Boxes	2
6	Toys	1	30	Telcom	3
7	Books	1	31	Comps	3
8	Hshld	1	32	Chips	3
9	Clths	1	33	Labeq	3
10	Txtls	1	34	Hlth	4
11	Whlsl	1	35	MedEq	4
12	Rtail	1	36	Drugs	4
13	Meals	1	37	Fun	5
14	Chems	2	38	Cnstr	5
15	Rubbr	2	39	Gold	5
16	BldMt	2	40	Mines	5
17	Steel	2	41	PerSv	5
18	FabPr	2	42	BusSv	5
19	Mach	2	43	Trans	5
20	ElcEq	2	44	Banks	5
21	Autos	2	45	Insur	5
22	Aero	2	46	RlEst	5
23	Ships	2	37	Fin	5
24	Guns	2	48	Other	5

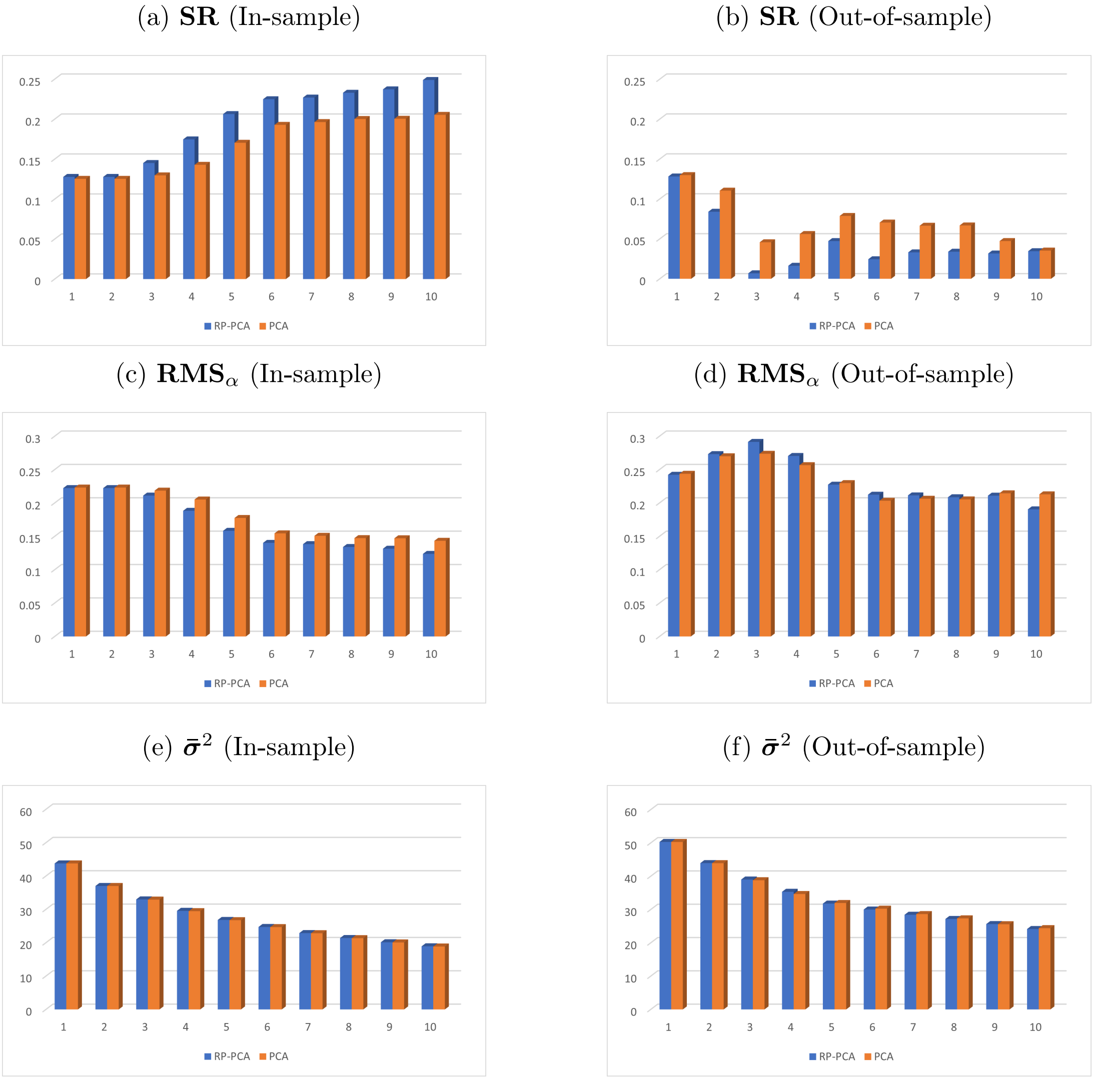
Note. The industry portfolios can be divided into five categories: 1. Consumer goods, 2. Manufacturing, 3. High-Tech, 4. Health and 5. Other. All industry portfolios are assigned to a category based on their SIC code. A more detailed explanation of the industry portfolios can be found in the Kenneth R. French Data Library [French, 2022].

Table C.2: Individual factors obtained by RP-PCA for different values of γ.

	$\gamma = 50$				$\gamma = 100$			
Factor	Mean	Variance	SR	Mean Rank	Mean	Variance	SR	Mean Rank
1	4.57	1187.13	0.13	1	4.61	1161.24	0.14	1
2	0.01	146.52	0.00	5	0.02	146.50	0.00	5
3	0.74	86.39	0.08	2	0.52	106.59	0.05	2
4	0.16	84.29	0.02	4	0.07	85.54	0.01	4
5	0.17	67.66	0.02	3	0.08	69.10	0.01	3

Note. The sample consists of 48 industry portfolios. The Sample ranges from January 1970 until April 2022 (T=628).

Figure C.1: Maximum Sharpe ratios, root-mean-squared pricing errors, and unexplained idiosyncratic variation for a different number of factors.



(a) **SR** (In-sample)

(b) **SR** (Out-of-sample)

(c) $\mathbf{RMS}_{\alpha}$ (In-sample)

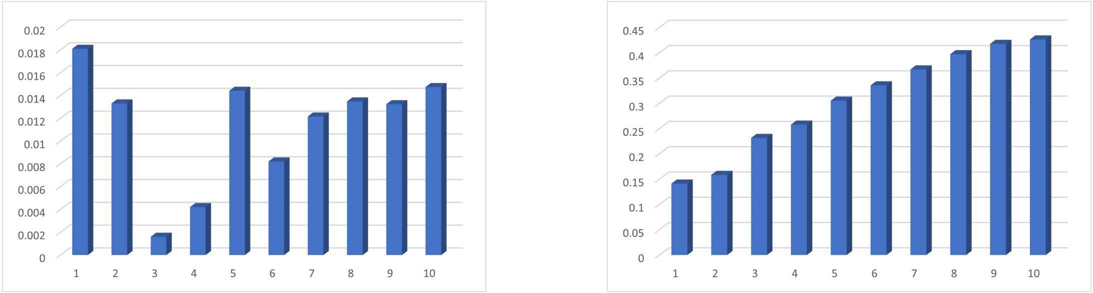
(d) $\mathbf{RMS}_{\alpha}$ (Out-of-sample)

(e) $\bar{\boldsymbol{\sigma}}^2$ (In-sample)

(f) $\bar{\boldsymbol{\sigma}}^2$ (Out-of-sample)

Note. The sample consists of 48 industry portfolios. The sample ranges from January 1970 until April 2022 (T=628). RP-PCA parameter $\gamma = 10$. The number of factors can be found on the x-axis.

Figure C.2: The out-of-sample expected return minus the risk-free rate and the out-of-sample standard deviation of the expected return (volatility) for the RP-PCA model.

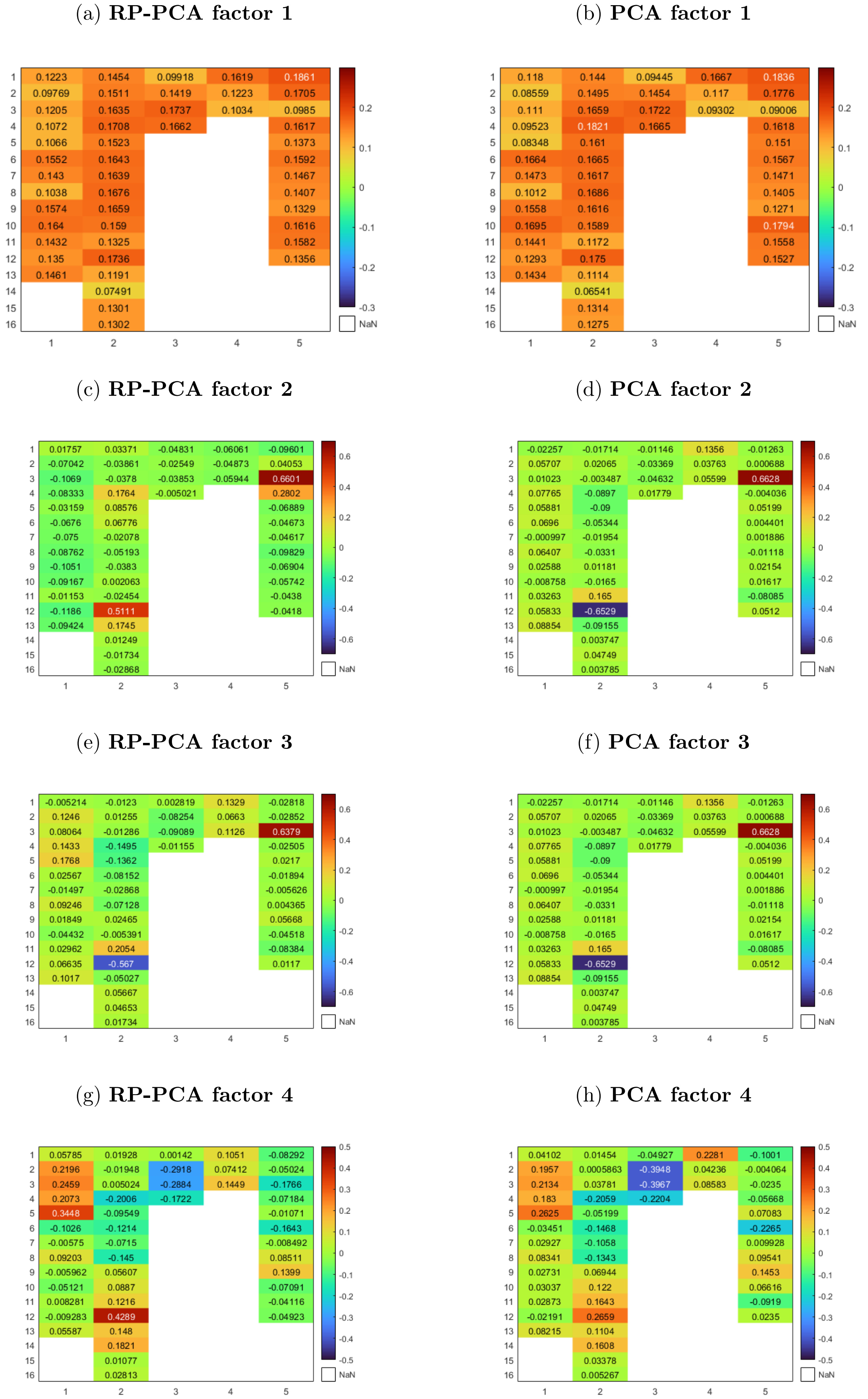


(a) **Expected return minus risk-free rate**

(b) **Volatility**

Note. The sample consists of 48 industry portfolios. The sample ranges from January 1970 until April 2022 (T=628). RP-PCA parameter $\gamma = 10$. Number of factors can be found on the x-axis.

Figure C.3: Heatmaps of the first five factors obtained by RP-PCA and PCA.



(a) **RP-PCA factor 1**

	1	2	3	4	5
1	0.1223	0.1454	0.09918	0.1619	0.1861
2	0.09769	0.1511	0.1419	0.1223	0.1705
3	0.1205	0.1635	0.1737	0.1034	0.0985
4	0.1072	0.1708	0.1662		0.1617
5	0.1066	0.1523			0.1373
6	0.1552	0.1643			0.1592
7	0.143	0.1639			0.1467
8	0.1038	0.1676			0.1407
9	0.1574	0.1659			0.1329
10	0.164	0.159			0.1616
11	0.1432	0.1325			0.1582
12	0.135	0.1736			0.1356
13	0.1461	0.1191			
14		0.07491			
15		0.1301			
16		0.1302			

(b) **PCA factor 1**

	1	2	3	4	5
1	0.118	0.144	0.09445	0.1667	0.1836
2	0.08559	0.1495	0.1454	0.117	0.1776
3	0.111	0.1659	0.1722	0.09302	0.09006
4	0.09523	0.1821	0.1665		0.1618
5	0.08348	0.161			0.151
6	0.1664	0.1665			0.1567
7	0.1473	0.1617			0.1471
8	0.1012	0.1686			0.1405
9	0.1558	0.1616			0.1271
10	0.1695	0.1589			0.1794
11	0.1441	0.1172			0.1558
12	0.1293	0.175			0.1527
13	0.1434	0.1114			
14		0.06541			
15		0.1314			
16		0.1275			

(c) **RP-PCA factor 2**

	1	2	3	4	5
1	0.01757	0.03371	-0.04831	-0.06061	-0.09601
2	-0.07042	-0.03861	-0.02549	-0.04873	0.04053
3	-0.1069	-0.0378	-0.03853	-0.05944	0.6601
4	-0.08333	0.1764	-0.005021		0.2802
5	-0.03159	0.08576			-0.06889
6	-0.0676	0.06776			-0.04673
7	-0.075	-0.02078			-0.04617
8	-0.08762	-0.05193			-0.09829
9	-0.1051	-0.0383			-0.06904
10	-0.09167	0.002063			-0.05742
11	-0.01153	-0.02454			-0.0438
12	-0.1186	0.5111			-0.0418
13	-0.09424	0.1745			
14		0.01249			
15		-0.01734			
16		-0.02868			

(d) **PCA factor 2**

	1	2	3	4	5
1	-0.02257	-0.01714	-0.01146	0.1356	-0.01263
2	0.05707	0.02065	-0.03369	0.03763	0.000688
3	0.01023	-0.003487	-0.04632	0.05599	0.6628
4	0.07765	-0.0897	0.01779		-0.004036
5	0.05881	-0.09			0.05199
6	0.0696	-0.05344			0.004401
7	-0.000997	-0.01954			0.001886
8	0.06407	-0.0331			-0.01118
9	0.02588	0.01181			0.02154
10	-0.008758	-0.0165			0.01617
11	0.03263	0.165			-0.08085
12	0.05833	-0.6529			0.0512
13	0.08854	-0.09155			
14		0.003747			
15		0.04749			
16		0.003785			

(e) **RP-PCA factor 3**

	1	2	3	4	5
1	-0.005214	-0.0123	0.002819	0.1329	-0.02818
2	0.1246	0.01255	-0.08254	0.0663	-0.02852
3	0.08064	-0.01286	-0.09089	0.1126	0.6379
4	0.1433	-0.1495	-0.01155		-0.02505
5	0.1768	-0.1362			0.0217
6	0.02567	-0.08152			-0.01894
7	-0.01497	-0.02868			-0.005626
8	0.09246	-0.07128			0.004365
9	0.01849	0.02465			0.05668
10	-0.04432	-0.005391			-0.04518
11	0.02962	0.2054			-0.08384
12	0.06635	-0.567			0.0117
13	0.1017	-0.05027			
14		0.05667			
15		0.04653			
16		0.01734			

(f) **PCA factor 3**

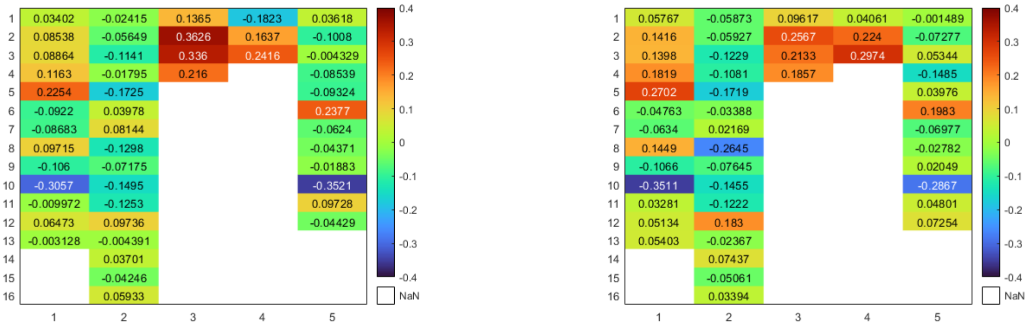
	1	2	3	4	5
1	-0.02257	-0.01714	-0.01146	0.1356	-0.01263
2	0.05707	0.02065	-0.03369	0.03763	0.000688
3	0.01023	-0.003487	-0.04632	0.05599	0.6628
4	0.07765	-0.0897	0.01779		-0.004036
5	0.05881	-0.09			0.05199
6	0.0696	-0.05344			0.004401
7	-0.000997	-0.01954			0.001886
8	0.06407	-0.0331			-0.01118
9	0.02588	0.01181			0.02154
10	-0.008758	-0.0165			0.01617
11	0.03263	0.165			-0.08085
12	0.05833	-0.6529			0.0512
13	0.08854	-0.09155			
14		0.003747			
15		0.04749			
16		0.003785			

(g) **RP-PCA factor 4**

	1	2	3	4	5
1	0.05785	0.01928	0.00142	0.1051	-0.08292
2	0.2196	-0.01948	-0.2918	0.07412	-0.05024
3	0.2459	0.005024	-0.2884	0.1449	-0.1766
4	0.2073	-0.2006	-0.1722		-0.07184
5	0.3448	-0.09549			-0.01071
6	-0.1026	-0.1214			-0.1643
7	-0.00575	-0.0715			-0.008492
8	0.09203	-0.145			0.08511
9	-0.005962	0.05607			0.1399
10	-0.05121	0.0887			-0.07091
11	0.008281	0.1216			-0.04116
12	-0.009283	0.4289			-0.04923
13	0.05587	0.148			
14		0.1821			
15		0.01077			
16		0.02813			

(h) **PCA factor 4**

	1	2	3	4	5
1	0.04102	0.01454	-0.04927	0.2281	-0.1001
2	0.1957	0.0005863	-0.3948	0.04236	-0.004064
3	0.2134	0.03781	-0.3967	0.08583	-0.0235
4	0.183	-0.2059	-0.2204		-0.05668
5	0.2625	-0.05199			0.07083
6	-0.03451	-0.1468			-0.2265
7	0.02927	-0.1058			0.009928
8	0.08341	-0.1343			0.09541
9	0.02731	0.06944			0.1453
10	0.03037	0.122			0.06616
11	0.02873	0.1643			-0.0919
12	-0.02191	0.2659			0.0235
13	0.08215	0.1104			
14		0.1608			
15		0.03378			
16		0.005267			



(i) **RP-PCA factor 5**

	1	2	3	4	5
1	0.03402	-0.02415	0.1365	-0.1823	0.03618
2	0.08538	-0.05649	0.3626	0.1637	-0.1008
3	0.08864	-0.1141	0.336	0.2416	-0.004329
4	0.1163	-0.01795	0.216		-0.08539
5	0.2254	-0.1725			-0.09324
6	-0.0922	0.03978			0.2377
7	-0.08683	0.08144			-0.0624
8	0.09715	-0.1298			-0.04371
9	-0.106	-0.07175			-0.01883
10	-0.3057	-0.1495			-0.3521
11	-0.009972	-0.1253			0.09728
12	0.06473	0.09736			-0.04429
13	-0.003128	-0.004391			
14		0.03701			
15		-0.04246			
16		0.05933			

(j) **PCA factor 5**

	1	2	3	4	5
1	0.05767	-0.05873	0.09617	0.04061	-0.001489
2	0.1416	-0.05927	0.2567	0.224	-0.07277
3	0.1398	-0.1229	0.2133	0.2974	0.05344
4	0.1819	-0.1081	0.1857		-0.1485
5	0.2702	-0.1719			0.03976
6	-0.04763	-0.03388			0.1983
7	-0.0634	0.02169			-0.06977
8	0.1449	-0.2645			-0.02782
9	-0.1066	-0.07645			0.02049
10	-0.3511	-0.1455			-0.2867
11	0.03281	-0.1222			0.04801
12	0.05134	0.183			0.07254
13	0.05403	-0.02367			
14		0.07437			
15		-0.05061			
16		0.03394			

Note. The sample consists of 48 industry portfolios. The sample ranges from January 1970 until April 2022 (T=628). RP-PCA parameter $\gamma = 10$. Positive loadings are shown in red and negative loadings are shown in blue.