Risk-Premium PCA: Empirical Evaluation and Application to Risk-Premium Estimation

Bachelor Thesis

Quantitative Finance

Student: Hidde Hogenhout (500865)

Supervisor: Gustavo Freire

Second Assesor: Terri van der Zwan

Abstract

Risk-Premium PCA (RP-PCA) is a new method for estimating the latent asset pricing factors that fit both the cross-section and time series of expected returns. The estimator is a generalized version of principal component analysis (PCA) which includes a penalty on the pricing error of expected returns. This paper performs a large empirical analysis which finds the RP-PCA estimator able to detect weak factors more accurately than standard PCA. Implementation of the RP-PCA estimator into the three-pass estimator for risk premia leads to more accurate estimates than the standard three-pass estimator while retaining insensitivity to measurement error and omitted variable bias.

Erasmus School of Economics

Netherlands

May 2022

Contents

1 Introduction

Modern asset pricing theory predicts that expected returns can be explained by the exposure of an asset to risk factors. Investors are compensated for their exposure to these factors by way of a risk premium. Theory has proposed a multitude of factors, where Harvey et al. (2015) document more than 300 published proposed risk factors. An alternative to using a theoretical factor model is to use a statistical factor model. In a statistical factor model, the latent factors are found by applying principal component analysis (PCA) to the variance-covariance matrix of the returns, rather than choosing the factors through economic theory. The standard version of PCA incorporates information in the second moments of the data, but not the information contained in the means of the data. This leads factors based on PCA to capture comovement well, however they do not necessarily capture the differences in risk premia of the assets.

This paper will explore an alternative estimator, Risk-Premium PCA (RP-PCA), as introduced by Lettau and Pelger (2020), which can be seen as a generalized version of PCA that accounts for cross-sectional pricing errors. It incorporates information in both the means and the variances of the assets, thus identifying factors that can explain the comovement and the cross-section of expected returns simultaneously. The question the first part of this paper answers is: **To what extent can RP-PCA identify latent risk factors which capture the comovement and the cross-section of stock returns?**

The second part of this paper will focus on implementing the RP-PCA estimator into the three-pass estimator approach by Giglio and Xiu (2021). Their approach focuses on the estimation of risk-premia. Classical techniques such as Fama-MacBeth regressions and mimicking portfolio projections all suffer from omitted variable bias. The three-pass estimator, however, is designed to avoid omitted variable bias and measurement error bias, thus providing a more reliable way to estimate risk premia.

The premise of the three-pass estimator is that PCA, which is used in the first step, is able to consistently recover a rotation of the factor space. However, standard PCA is only able to do this when the latent factors are sufficiently strong. As shown in Lettau and Pelger (2020) RP-PCA is able to detect weak factors which cannot be detected by PCA, thus it is able to more consistently recover a rotation of the factor space. Therefore, I will use RP-PCA in the first step of the three-pass estimator to create a three-pass estimator that is more robust to weak factor structures. The central question in this second part of the paper is: **How much more robust to weak factors is the three-pass estimator based on RP-PCA?**

To test the performance of the RP-PCA estimator, two sets of data are used. One smaller dataset consisting of 25 double-sorted portfolios for various anomalies that have been shown to generate excess returns. The other is a larger dataset with single-sorted decile portfolios on 37 different anomaly characteristics.

This paper is at the crossroads between the literature which explores factor models and the literature surrounding the estimation of risk premia. The former is seeing advances by way of projected PCA that allows time-varying factor loadings in Fan et al. (2016), and in instrumented PCA (IPCA) by Kelly et al.

(2019). Furthermore, Bai and Ng (2019) lay the theoretical foundation for robust principal components. On the side of the latter, previous theory mostly focused on either two-pass regressions, such as in Black et al. (1972) and Fama and MacBeth (1973). The econometric techniques and theory driving the two-pass regression have been improved in recent years in papers like Connor et al. (2012) who explore a nonparametric regression methodology. The other side of the risk premium estimation literature focuses on mimicking portfolios, introduced by Breeden et al. (1989) and further popularized by Lamont (2001). This technique is being expanded by the use of boosting algorithms as in Lönn and Schotman (2018). The three-pass estimator by Giglio and Xiu (2021) can be interpreted as both an extension of the two-pass regression approach and the mimicking portfolio approach.

The paper follows the following structure. Section 2 discusses first the estimation of the RP-PCA estimator, thereafter the three-pass estimator and finally the way in which both can be combined to create a robust three-pass estimator. Section 3 gives an overview of the data and its sources. Section 4 presents the empirical results, it is divided into three parts. Firstly, in Subsection 4.1 the RP-PCA estimator is tested on eight double sorted portfolios, Subsection 4.2 evaluates the RP-PCA estimator based on a larger dataset of single sorted portfolios. Finally, Subsection 4.3 discusses the empirical performance of three-pass estimator based on RP-PCA. Section 5 presents an overview of results and some concluding remarks.

2 Methodology

In this section, I first introduce the Risk-Premium Principal Component Analysis (RP-PCA) estimator, which utilizes information contained in the means and the covariances to better identify the factors that generate excess returns. Second, the three-pass estimator by Giglio and Xiu (2021) is presented. This estimator deals with omitted variable bias and measurement error bias when estimating risk premia in linear factor models. Finally, I will show how the RP-PCA estimator can be combined with the three-pass estimator to form a more robust estimator of factor risk premia.

2.1 Risk-Premium PCA

RP-PCA can be seen as a generalised version of standard PCA, leading to an almost identical theoretical framework. In this framework, the excess return of an asset n at time t, X_{nt}, can be decomposed into two parts. There exists a systematic component which can be captured by K factors that are the same for all assets. The loadings Λ are asset specific and correspond to each asset's exposure to the risk factor F. The second part is an idiosyncratic part, it captures asset-specific risk and is thus unique for each asset. The excess returns of N assets over T time periods thus follow the mathematical structure:

$$\mathrm{X}_{nt} = \mathrm{F}_t \ \Lambda_n^\top + e_{nt} \tag{1}$$

$$\iff \underset{T\times N}{\mathrm{X}} = \underset{T\times K}{\mathrm{F}}\ \underset{K\times N}{\Lambda^{\top}} + \underset{T\times N}{\mathrm{e}} \tag{2}$$

In a statistical factor model such as PCA and RP-PCA, it is assumed that both the latent factors F and their loadings Λ are unknown. I consider cases where the number of cross-sectional observations N and the number of time series observations T goes to infinity, but N/T converges to a finite limit. Further, the assumption is made that the factors and residuals are uncorrelated, such that the covariance matrix of the returns consists of a systematic and an idiosyncratic part:

$$Var(\mathrm{X}) = \Lambda\, Var(\mathrm{F})\, \Lambda^{\top} + Var(\mathrm{e}) \tag{3}$$

From Equation 3 it follows that the largest eigenvalues of the matrix $Var(\mathrm{X})$ are driven by the factors. Principal component analysis makes use of this fact to estimate the loadings and factors. PCA thus only uses information that is contained in the variance of the excess returns and ignores any information that might be contained in the means.

For excess returns, however, we know that there is information contained in the means. Ross' arbitrage pricing theory (APT) implies that the mean excess return of a given asset is equal to its exposure to risk factors multiplied by the risk premium, plus some idiosyncratic component.

$$\mathrm{E}[\mathrm{X_n}] = \alpha_n + \Lambda_n\, \mathrm{E}[\mathrm{F}] \tag{4}$$

Since my analysis is based solely on diversified portfolios, I can assume that the idiosyncratic part α_n of the expected returns is equal to zero. Thus the strong form of Ross' APT applies here, it is given by:

$$\mathrm{E}[\mathrm{X_n}] = \Lambda_n\, \mathrm{E}[\mathrm{F}] \tag{5}$$

PCA is designed such that the factors it identifies explain as much time variation as possible. Conventional PCA can be applied to either the sample covariance matrix or the sample correlation matrix. The covariance matrix is given by $\frac{1}{\mathrm{T}}\mathrm{X}^{\top}\mathrm{X} - \overline{\mathrm{X}}\,\overline{\mathrm{X}}^{\top}$ where $\overline{\mathrm{X}}$ denotes the sample mean excess return. The correlation matrix is simply the case when the test assets are normalized by their standard deviation, such that the diagonal of the variance-covariance matrix of the returns is equal to one.

Standard PCA analysis proceeds by rotating the N test assets X using the eigendecomposition of the variance-covariance matrix of the excess returns (Σ_{X}), to obtain N orthogonal factors. The eigenvalues of the rotated factors correspond to their variances. The factors with the largest eigenvalues have the largest variance and thus explain the largest part of total variance, therefore the factors are sorted according to their eigenvalues. Then the K factors with the largest eigenvalues are chosen as estimates of the true factors F in Equation 2. PCA thus achieves a dimension reduction from N to K, which is why it is suitable for large datasets.

The factor loadings Λ in Equation 2 are proportional to the eigenvectors associated with the K largest eigenvalues. The K $\times$ N matrix of estimated loadings $\widehat{\Lambda}_{\mathrm{PCA}}$ is thus equal to the matrix of eigenvectors

that correspond to the K largest eigenvalues. The estimated T × K matrix of rotated orthogonal factors is then given by $\widehat{\mathrm{F}}_{\mathrm{PCA}} = \mathrm{X}\,\widehat{\Lambda}_{\mathrm{PCA}}\,(\widehat{\Lambda}_{\mathrm{PCA}}^{\top}\,\widehat{\Lambda}_{\mathrm{PCA}})^{-1}$.
The RP-PCA estimator is formed by applying PCA to the matrix

$$\Sigma_{\mathrm{RP}} = \frac{1}{\mathrm{T}}\mathrm{X}^{\top}\mathrm{X} + \gamma\,\overline{\mathrm{X}}\,\overline{\mathrm{X}}^{\top} \tag{6}$$

It is clear from Equation 6 that when $\gamma = -1$, RP-PCA is equal to standard PCA. Furthermore, RP-PCA is equal to PCA applied to the second moment matrix when γ is set to zero. This means that RP-PCA with the γ parameter set to zero is equivalent to PCA applied to demeaned data, which is often done in practice. RP-PCA with $\gamma > 0$ can be seen as a form of PCA that overweights the means of the data. As in standard PCA, the eigenvectors of the K largest eigenvalues of Σ_{RP} are proportional to the loadings $\widehat{\Lambda}_{\mathrm{RP}}$. However, unlike in PCA, the eigenvalues cannot be interpreted as factor variances. In RP-PCA they denote more of a general notion of signal strength of a factor. A regression of the excess returns on the estimated loadings gives the RP-PCA factors $\widehat{\mathrm{F}}_{\mathrm{RP}} = \mathrm{X}\,\widehat{\Lambda}_{\mathrm{RP}}\,(\widehat{\Lambda}_{\mathrm{RP}}^{\top}\,\widehat{\Lambda}_{\mathrm{RP}})^{-1}$.

The RP-PCA approach is compared against standard PCA and the Fama-French three- and five-factor models. The performance of each model is evaluated both in- and out-of-sample based on three different criteria: the maximum achievable Sharpe ratio, the root-mean-squared alpha, and the unexplained idiosyncratic variance.

The maximum achieveable Sharpe ratio from a linear combination of the factors is given by the Sharpe ratio of the mean-variance optimal portfolio of the factors. Given the estimated factors $\widehat{\mathrm{F}}$, their weights in the mean-variance optimal portfolio can be computed as:

$$\widehat{\mathrm{b}}_{\mathrm{MV}} = \Sigma_{\mathrm{F}}^{-1}\mu_{\mathrm{F}} \tag{7}$$

By multiplying the estimated factors $\widehat{\mathrm{F}}$ with their weights $\widehat{\mathrm{b}}_{\mathrm{MV}}$ in the mean-variance optimal portfolio we find the minimum variance SDF. The maximum possible Sharpe ratio is then computed by dividing the time-series SDF mean by the time-series SDF standard deviation. This maximum possible Sharpe ratio is thus an indication of how close to the true SDF the model is.

The other performance measures used in this paper find their origin in the way the RP-PCA loadings $\widehat{\Lambda}$ can be related to ordinary least squares (OLS) betas of the time series regressions of excess returns on factors. The factor model in Equation 2 implies that if the factors F are already known, the loadings of the assets Λ can be estimated by OLS without a constant. By using this model the no-constant structure is thus imposed onto the data. Alternatively, we can estimate:

$$\mathrm{X}_{nt} = \alpha_n + \widehat{\mathrm{F}}_t\mathrm{B}_n^{\top} + e_{nt} \tag{8}$$

The model can then be evaluated by the estimated $\widehat{\alpha}_n$ and $\widehat{e}_{nt}$. The RMS_{α}, or root-mean-square error, is computed by taking the root of the mean of the squares of $\widehat{\alpha}_n$. It corresponds to the average pricing error across assets, a lower value thus indicates a better performing model. Another measure we can compute is the unexplained idiosyncratic variance, or $\overline{\sigma}_e^2$. This is calculated by taking the mean of the variance of the residual for each asset. If we further divide by the average variance of the assets X_n, we get the

unexplained variance expressed as a percentage of total variance. This measure is by definition minimized by PCA in-sample and shows how much of the total variance cannot be explained by the factor model, a lower number thus indicates a better performing model.

To test the out-of-sample (OOS) performance of the models the same measures are used. The factors and loadings are estimated using a rolling window of twenty years (T = 240). To calculate the maximum possible Sharpe ratio OOS, the first data point outside the window is used in combination with the estimated $\widehat{\Lambda}$ inside the window, to obtain the forecasted value of the factors outside the window. By multiplying this with the optimal mean-variance weights in the window as calculated in Equation 7 we arrive at the maximum possible Sharpe ratio for each window. The total OOS Sharpe ratio is then simply the average Sharpe ratio over all possible windows.

The RMSE and unexplained idiosyncratic variance are computed similarly. First, a regression is performed to estimate the B in Equation 8 within the window. This estimated B is then used in combination with the next data point outside the window, to calculate the OOS alpha for that window. The time-series average of this alpha is then taken to find the total OOS alpha. From the total OOS alpha, the OOS RMS_α and $\overline{\sigma}_e^2$ are then calculated the same way as in-sample.

2.2 Three-pass estimator

The three-pass estimator is a method to estimate the risk premium of a risk factor g_t developed by Giglio and Xiu (2021). The estimator deals with measurement error and missing variable bias, which are problems that can occur when using the two most common ways of estimating risk premia: Fama-MacBeth two-pass regressions and the mimicking portfolio approach. These techniques are explained in more detail at the end of this section and the theoretical framework of why they are sensitive to omitted variable bias and measurement error is laid out in Giglio and Xiu (2021).

The three-pass estimator is named appropriately so, because it estimates the risk premium of a factor g_t in three steps, or passes. The first step is to use PCA on a set of test asset returns, such that some rotation of the true factor space is recovered. Here it is important that one: the test assets represent the aggregate stock market well, they must be influenced by the same risk factors as the market. Second, it is important that the underlying factor structure is strong enough, such that a rotation of the priced risk factors can be detected by PCA. Finally, the number of factors to use has to be selected, this paper uses $k = 7$ factors following Giglio and Xiu (2021). Further research might focus on techniques to determine the correct number of factors, or examine the robustness to the chosen number of factors.

Once the factors and the loadings have been estimated by PCA, the second step of the three-pass procedure, the cross-sectional regression step, can commence. To obtain the risk premia of the estimated factors, the average excess return for each asset $\overline{\text{X}}$ is regressed onto the estimated factor loadings $\widehat{\Lambda}$ using ordinary least squares (OLS):

$$\widehat{\gamma} = (\widehat{\Lambda}^\top \widehat{\Lambda})^{-1} \widehat{\Lambda}^\top \overline{\text{X}} \tag{9}$$

The third step of the approach by Giglio and Xiu (2021) is to run a time series regression using

OLS. The dependent variable is the factor of which the risk premium is to be estimated g_t, in case the risk premium of a combination of factors is to be estimated, we take the cross-sectional mean $\overline{G}$. The independent variables are the factors we estimated in step one, denoted by $\widehat{F}$. The variable $\widehat{\eta}$ which is estimated here can be interpreted as the loading of g_t onto the unobservable fundamental factors.

$$\widehat{\eta} = \overline{G}\,\widehat{F}^\top(\widehat{F}\,\widehat{F}^\top)^{-1} \tag{10}$$

Finally, the risk premium of the observable factor g_t can be estimated by combining the estimators from steps two and three.

$$\widehat{\gamma}_g = \widehat{\eta}\,\widehat{\gamma} \tag{11}$$

All previous steps can be combined to write the estimator in a more compact form.

$$\widehat{\gamma}_g = \overline{G}\,\widehat{F}^\top(\widehat{F}\,\widehat{F}^\top)^{-1}(\widehat{\Lambda}^\top\widehat{\Lambda})^{-1}\widehat{\Lambda}^\top\overline{X} \tag{12}$$

2.3 Robust three-pass estimator

The idea of the robust three-pass estimator is based on the assumption by Giglio and Xiu (2021) that PCA will consistently recover a rotation of the factor space if the factors are sufficiently strong. Lettau and Pelger (2020) show that not all factors driving stock returns are strong factors and that PCA will not recover some of these weak but informative factors. I will use RP-PCA instead of standard PCA in the first step of the three-pass estimator in order to more consistently recover the factor space, thus more accurately identifying risk premia when weak factors are present in the dataset.

To evaluate the performance of the robust three-pass estimator I compare it against a variety of risk premium estimation models. The models will be used to estimate the risk premium of traded factors, such as the five Fama-French factors and the 37 anomalies in Kozak et al. (2020), see Section 3 for a detailed description of the data. Since these are all traded factors their risk premium can be reliably estimated by taking the mean return of the factor over the sample period. The models are then compared by evaluating the difference between the mean factor return and the estimated risk premium. The closer the estimated risk premium is to the mean return, the better the model. As a general measure of model performance the root-mean-square error over multiple factors, such as the five Fama-French factors, is then computed for all models. This means the difference between the mean and the estimated risk premium is squared, then the mean over all factors g_t is taken, and finally, the root of this number is the RMSE of the model.

The first model used for comparison is the Fama and MacBeth (1973) two-pass regression. Two-pass regressions estimate the risk premium γ_g by first doing a time-series regression of the test assets' excess return onto the factor of which the risk premium is to be estimated [1]. This gives the estimated risk exposure of the assets to the risk factor. Second, a cross-sectional regression of the mean asset returns onto the asset risk exposure gives the estimated risk premium $\widehat{\gamma}_g$.

[1] These test assets are also referred to as base assets in this paper. They should span the entire factor space as best as possible, therefore I take the entire set of N=370 decile-1 through decile-10 anomaly portfolios unless otherwise noted. Robustness checks to this choice of base assets are available in the appendix

In the literature, two-pass regression models are often augmented by adding control factors to the first regression. Rather than regressing the assets' excess returns on only the risk factor itself, they are regressed onto the risk factor and the control factor(s). This makes it so any risk premium related to the controls is not seen as risk belonging to the risk factor. In this paper, three different two-pass models are considered: one with no controls at all, one with only the market factor as control, and lastly one with the three Fama-French factors as controls.

The second model used to compare the robust three-pass estimator is the mimicking portfolio model. Using this approach, the risk premium is estimated by projecting the risk factor onto a set of tradable excess asset returns. This constructs a portfolio that is maximally correlated with the risk factor. The risk premium can then simply be estimated by taking the mean return of the mimicking portfolio. I consider three versions of this approach: the first one projects the risk factor only onto the market factor, thus achieving some CAPM-like model without a constant. The second model projects the risk factor onto the three Fama-French factors. Lastly, I consider an approach where the risk factor is projected onto the set of 74 decile-1 and decile-10 anomaly portfolios from Kozak, Nagel and Santosh.

Lastly, the robust three-pass estimator is compared against its regular counterpart as described in Subsection 2.2. For both the regular three-pass estimator and the robust three-pass estimator the set of base assets used should span the factor space of the returns. Furthermore, these techniques are designed to work best in large N environments. Therefore, the entire set of N = 370 decile-1 through decile-10 anomaly returns are used as test assets.

3 Data

Firstly, the performance of the RP-PCA approach will be tested on a smaller sample containing double-sorted anomaly portfolios. The dataset consists of eight sets of double-sorted 5x5 portfolios on size and book-to-market, accruals, investment, profitability, momentum, short-term reversal, variance, and residual variance. Excess returns are calculated by subtracting the risk-free rate from the simple returns. The portfolio returns and the risk-free rate are downloaded from the Kenneth French data library.
Thereafter, results will be examined for a larger cross-section of single-sorted decile portfolios. The dataset consists of all 37 characteristics that are used in Kozak et al. (2020). Following Lettau and Pelger (2020), I decide to show results only for the extreme first and tenth deciles, rather than for the whole dataset. This is possible since most relevant information is contained in these extreme deciles. While this reduction in portfolios is useful to allow for more clear presentation of the results, it is not necessary since PCA and RP-PCA are well-suited to handle large cross-sections of data. Again the risk-free rate is subtracted from the returns to arrive at excess returns. The single-sorted characteristic portfolio returns are downloaded from the Serhiy Kozak website.
The RP-PCA method is tested against standard PCA and the Fama and French (1993) three-factor model in the smaller dataset, while it will also be tested against the Fama and French (2015) five-factor model in the larger dataset. The factor returns for both models are downloaded from the Kenneth French library.

[2] For all data I consider monthly data in the sample period from November 1963 until December 2017, leading to T=650 observations. Table 1 gives an overview of all different datasets and their sources.

Table 1: Overview of the data sources

Name	**Source**	**Description**	**N**	**T**
Double sorts	Kenneth French library	Eight sets of double sorted portfolios	25	650
Single sorts	Serhiy Kozak's website	37 decile portfolios sorted on anomaly characteristics	370	650
Extremes	Serhiy Kozak's website	First and tenth deciles of the single sorts	74	650
3-factor FF	Kenneth French library	Factor returns in the 3-factor Fama-French model	3	650
5-factor FF	Kenneth French library	Factor returns in the 5-factor Fama-French model	5	650
Risk-free rate	Kenneth French library	1-month treasury bill return	1	650

4 Empirical results

In this section, the empirical application of RP-PCA to return data will be discussed. This is done in two parts: firstly results for RP-PCA applied to eight double-sorted portfolios with 25 test assets each will be shown. Thereafter, the method is applied to a larger cross-section of single-sorted portfolios for various anomaly characteristics. For a detailed description of the data see section 3. All reported Out-Of-Sample (OOS) results are estimated using a rolling window of 20 years ($T = 240$). With the estimated loadings and factors over this period the OOS pricing error at $t + 1$ is estimated. The mean and variance of this OOS pricing error correspond to the average pricing error and the unexplained idiosyncratic variation. Furthermore, the factor weights in the mean-variance optimal portfolio are used to estimate the maximum Sharpe ratio OOS. This thus gives three performance metrics: the root-mean-square pricing error (RMS), average unexplained variance ($\overline{\sigma}_e^2$) and the maximum Sharpe ratio that can be obtained by a linear combination of the factors.

4.1 Double-sorted portfolios

Table 2 reports the Sharpe Ratio, RMS and idiosyncratic variance for eight double-sorted portfolios on size and book-to-market, accruals, investment, operating profitability, residual variance, variance, momentum and reversal. I compare results for the RP-PCA model and PCA model with three factors with a modified version of the Fama-French three-factor model. The Fama-French model consists of a market factor, small-minus-big (SMB) factor and a high-minus-low (HML) factor. Both the SMB factor and the HML factor are constructed from the appropriate double-sorted portfolios. So for example

[2]The factor returns are constructed from one double-sorted portfolio on size and book-to-market for the three-factor model, while for the five-factor model they are constructed from three sets of double-sorted portfolios on size and book-to-market, profitability, and investment. Therefore, the returns of the SMB factor are different in the three- and five-factor models.

for accruals, the SMB factor corresponds to the average return of the five portfolios with the smallest companies minus the average return of the five portfolios with the largest companies. The HML factor corresponds with the average return of the five portfolios containing the companies with the highest accruals, minus the average return of the five portfolios with the companies with the lowest accruals. Furthermore, $\gamma = 20$ in the RP-PCA model, however, the results are not sensitive to this choice, as will be shown later in this paper.

An inspection of the Sharpe ratios for the various models reveals that in seven out of eight cases the Sharpe ratio is highest for RP-PCA. The RP-PCA technique is specifically developed to be able to detect weak factors with high Sharpe ratios, these results give a preliminary indication that it is able to perform this task better than the PCA and Fama-French models. Furthermore, the RMS for RP-PCA is lowest in all cases, and the unexplained variation is also lowest in five out of eight cases. This last result is interesting because PCA, by definition, minimizes in-sample idiosyncratic variation, however, as Table 2 shows, OOS this property of PCA often does not hold anymore and RP-PCA is in practice better able to minimize OOS idiosyncratic variation.

Table 2: OOS Sharpe ratio, RMS and unexplained idiosyncratic variation of RP-PCA, PCA and the Fama-French three-factor model for eight double-sorted portfolios.

	Sharpe Ratio			RMS			$\overline{\sigma}_e^2$		
	RP-PCA	PCA	FF	RP-PCA	PCA	FF	RP-PCA	PCA	FF
Book-to-Market	**0.20**	0.18	0.18	**0.17**	0.17	0.18	7.97	7.91	**7.87**
Accruals	**0.21**	0.12	0.14	**0.09**	0.11	0.11	6.74	**6.44**	6.90
Investment	**0.26**	0.23	0.20	**0.13**	0.15	0.16	**6.95**	7.00	7.20
Operating profitability	0.13	0.14	**0.15**	**0.09**	0.10	0.10	**6.94**	7.08	7.08
Residual Variance	**0.29**	0.23	0.24	**0.16**	0.17	0.18	**6.22**	6.24	6.35
Variance	**0.27**	0.21	0.24	**0.18**	0.19	0.20	**6.27**	6.30	6.41
Momentum	**0.21**	0.18	0.17	**0.20**	0.21	0.22	**8.30**	8.40	8.48
Reversal	**0.16**	0.11	0.11	**0.18**	0.19	0.19	7.89	7.86	**7.82**

OOS maximal Sharpe ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for double sorted portfolios on size and the shown characteristic. Bold numbers indicate the best performing models, the bold numbers have been determined with an accuracy higher than two decimals.

How is it possible that RP-PCA outperforms standard PCA in most cases? There are three ways in which the RP-PCA model can differ from the standard PCA model. Firstly, it can detect a different set of factors. Second, the compositions of the factors that are detected can be different. Lastly, the order in which the factors are detected can be different. To see the effects of these differences, two cases will be inspected in more detail: size/accruals and size/reversal. These cases illustrate the differences between RP-PCA and PCA most clearly.

The three most prominent factors in these portfolios are the market factor, the size factor and the

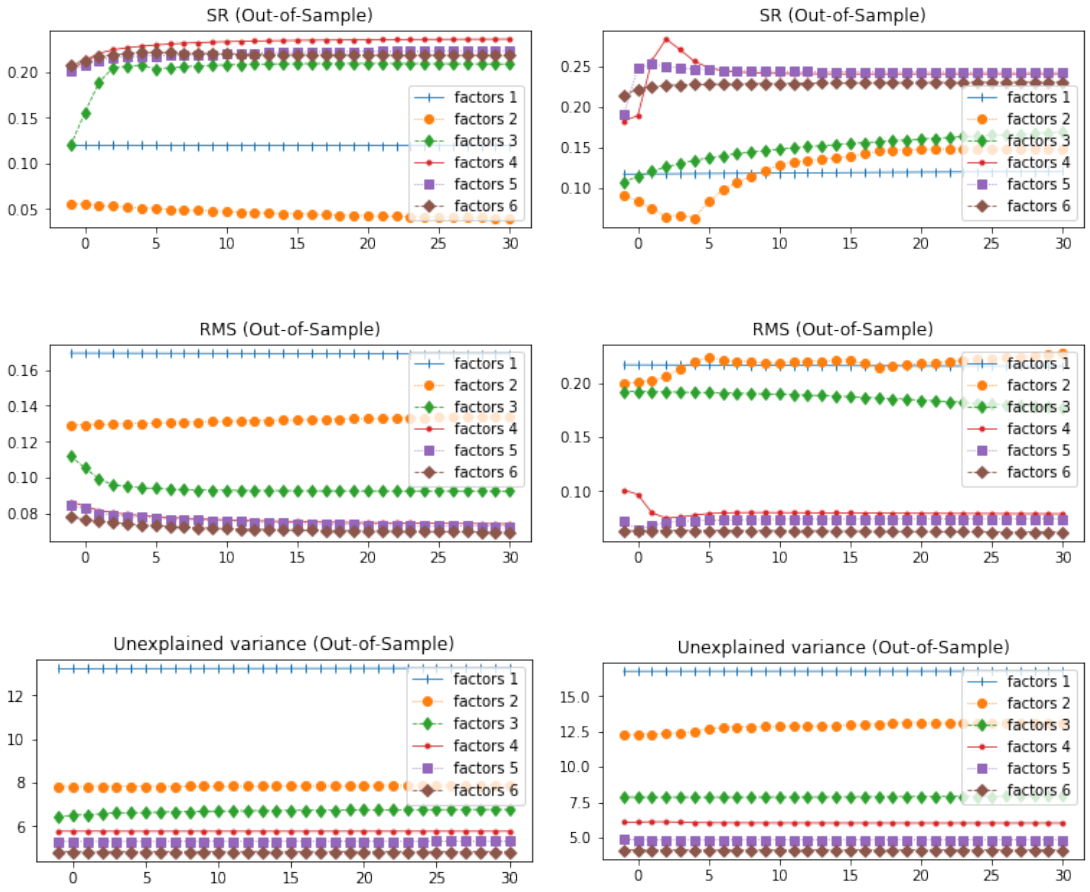
accruals/reversal factor, since this is how the portfolios have been sorted. As is shown in Figures 1, 2 and 3, both RP-PCA and PCA extract the market factor and a size factor. However, standard PCA is not able to detect the long/short factor related to accruals, while RP-PCA is able to pick up on this factor. Furthermore, in the case of the portfolio sorted on size and short-term reversal, the order of the factors changes. For RP-PCA the second factor is the high/low reversal factor, while the second factor for standard PCA is the one related to size. This is because the reversal factor captures more return differences than the size factor and is thus given a higher weight in the RP-PCA estimation.

Figure 1 plots the Sharpe ratio, root-mean-square error and idiosyncratic variation as a function of gamma. Firstly, note that a model with only one factor is completely insensitive to the choice of gamma, indicating that standard PCA as well as RP-PCA are able to identify the market factor correctly regardless of the choice of gamma. Similarly, the choice of gamma seems to have little effect on the SR, RMS and unexplained variance for a model with two factors in the size/accruals portfolio. However, for the size/short-term reversal portfolio, we see an increase in the Sharpe ratio in a two-factor model starting from around $\gamma = 5$, indicating that RP-PCA selects a higher Sharpe ratio factor as the second factor than standard PCA. A similar effect can be seen in the size/accruals portfolio once a third factor is added to the model. The Sharpe ratio and RMS start improving as soon as the gamma parameter is higher than -1.

To find out what causes these differences, we will inspect the factor loadings of the first three factors for the size/accrual and size/reversal portfolios. Figure 2 shows the factor loadings for both PCA and RP-PCA for the size/accruals portfolio and Figure 3 shows the same for the size/reversal portfolio. Note here that unlike in Lettau and Pelger (2020), I show the loadings in the case where the eigenvectors have been normalized to have unit length, $\Lambda^\top \Lambda = \mathrm{I}$. This means the sum of the squares of the loadings in the heatmap will thus add up to one. The normalization only affects the numerical value of the loadings and the factors, not the composition or detection of factors.

A closer inspection of Figure 2 reveals that for the size/accruals portfolio the loadings of the first factor are all positive and similar in size, this corresponds with an average return, or market, factor. The second factor also looks very similar for RP-PCA and PCA, with negative loadings for small stock portfolios and positive loadings for big ones. Hence, this factor corresponds with a long/short small minus big portfolio in both cases. Interestingly, we observe very different loadings for the third factor. RP-PCA gives negative weights to high-accrual stock portfolios and positive weights to low ones, thus successfully detecting the low minus high accrual long/short factor. PCA on the other hand shows no clear pattern, referring back to Figure 1 we see that indeed this factor adds very little information.

Figure 3 shows a clear example of the way in which the ordering of the factors can differ between RP-PCA and PCA. Both methods select the market factor first, indicated by the all-positive weights for the first factor. The second factor, however, corresponds to a long/short portfolio based on size for PCA, while for RP-PCA it is given by a long/short portfolio based on reversal. Consequently, the third factor in the PCA model is the reversal factor, while it is the size factor in the RP-PCA model. In the top-right



(a) Size/Accruals

(b) Size/Short-Term Reversal

Figure 1: Out of sample results as a function of γ

panel of Figure 1 it is shown that the Sharpe ratio increases as gamma increases. This shows that the second factor in the RP-PCA model, the reversal factor, has a higher Sharpe ratio than the size factor. Since the RP-PCA model overweights the means in the covariance matrix, it will prioritize this factor over the size factor.

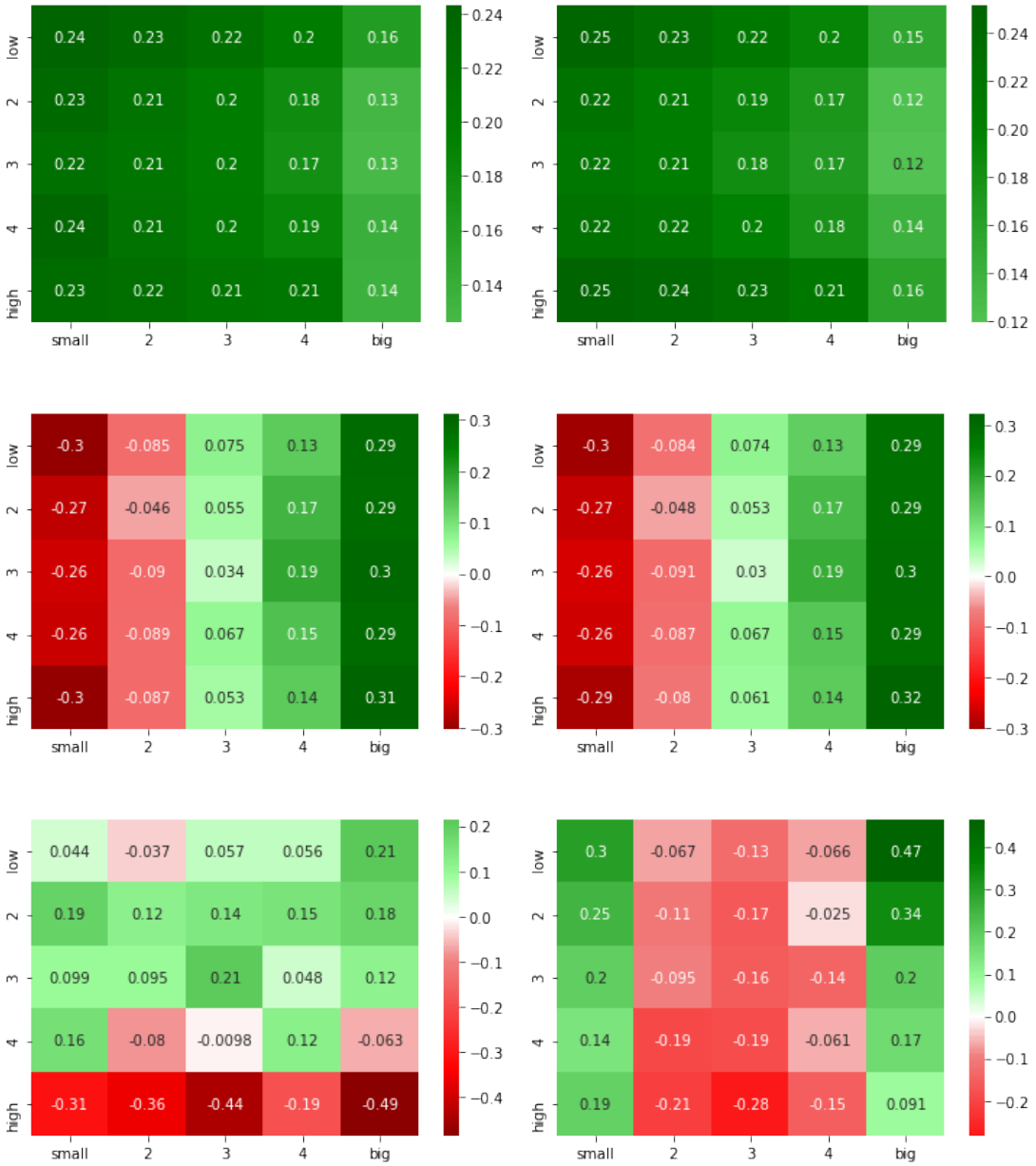
4.2 Single-sorted anomaly portfolios

To continue the empirical examination of RP-PCA, I will apply the estimator to a larger cross-section of portfolios. All 37 anomaly characteristics from the data set in Kozak, Nagel and Santosh are considered. The complete dataset contains decile sorts for each anomaly making for a total of $N = 370$ portfolios. While RP-PCA and PCA are well-suited to handle large datasets like this, I will perform most of the empirical analysis only on the 74 decile-1 and decile-10 portfolios. This makes the presentation of results more feasible and clear to the reader, while most of the significant information is contained in these extreme deciles, so there is very limited loss of information. All results are robust to the use of the 74 extreme deciles compared to the 370 total deciles. For more specific information on the data used see

Section 3.

4.2.1 Number of systematic factors

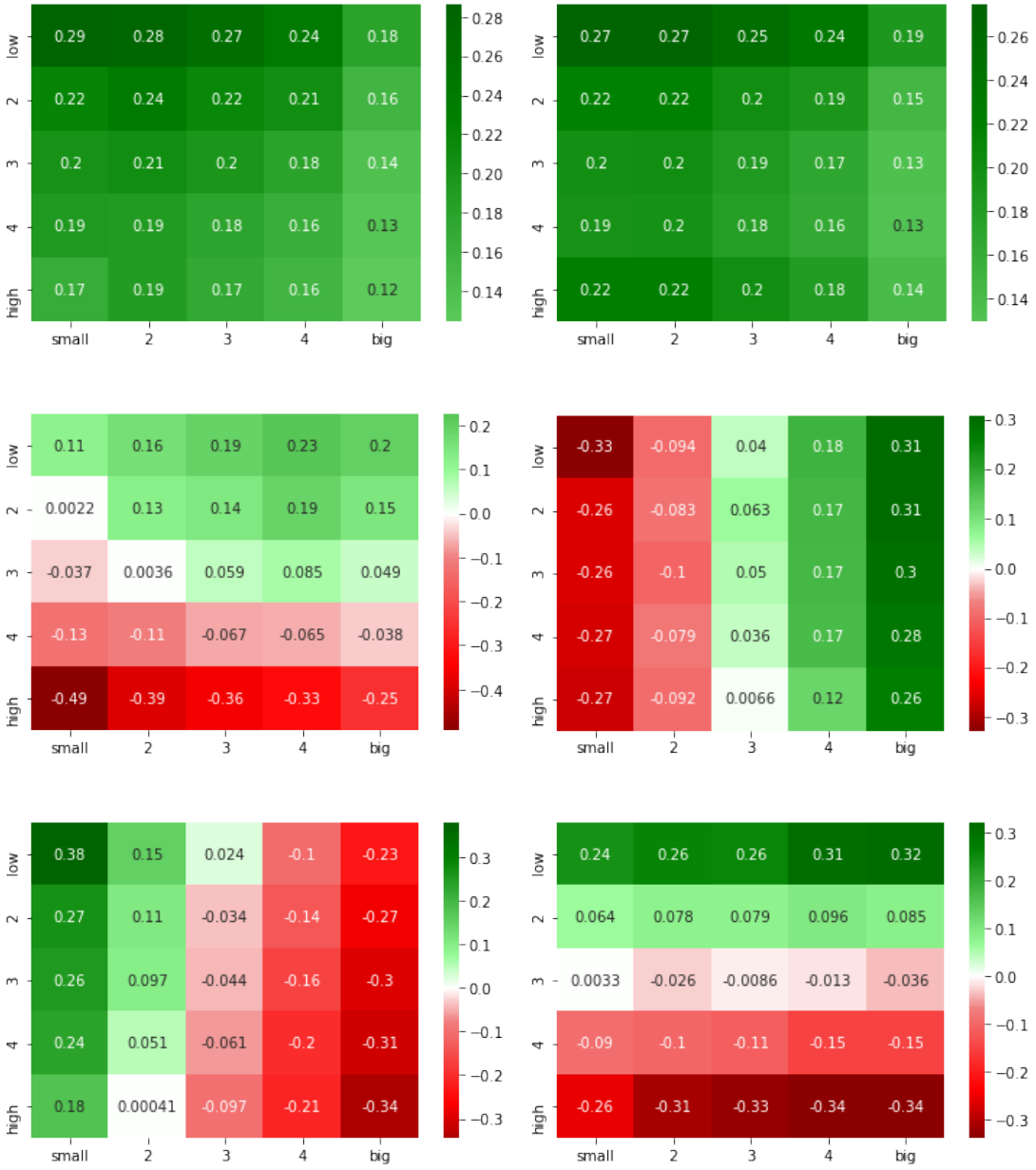
Figure 4 gives the difference of consecutive eigenvalues of the matrix $\frac{1}{T}X^\top X + \gamma \overline{X}\,\overline{X}^\top$ for the decile-1 and decile-10 portfolios as well as the full set of portfolios. We consider first the top figure, where RP-PCA is applied to the 74 decile-1 and decile-10 portfolios. The first eigenvalue difference is an order of magnitude



(a) Size/Accruals: RP-PCA (b) Size/Accruals: PCA

Figure 2: Loadings of the first three PCA ($\gamma = -1$) and RP-PCA ($\gamma = 20$) factors for the Size/Accruals portfolio

greater than the others. The figure is most interesting around the fifth difference, the difference between the fifth and sixth eigenvalue, because for the low-gamma RP-PCA we see this difference is small, but for high-gamma RP-PCA the difference is still quite large. This indicates that RP-PCA will generally include the fifth factor, while standard PCA might not. The same conclusions can be drawn from the bottom figure, indicating that these results are robust. From now on we will thus consider five-factor PCA and RP-PCA models.



(a) Size/Reversal: RP-PCA

(b) Size/Reversal: PCA

Figure 3: Loadings of the first three PCA ($\gamma = -1$) and RP-PCA ($\gamma = 20$) factors for the Size/Reversal portfolio

4.2.2 Estimation results RP-PCA versus PCA

Figure 5 gives an overview of the Sharpe ratio, root-mean-square error, and the residual variance expressed as a percentage of the total variance, for RP-PCA and standard PCA. Considering the in-sample results, expressed in the left-hand column, first. We observe that for any number of factors greater than one RP-PCA is able to achieve a higher Sharpe ratio and RMSE. Note that PCA by definition minimizes in-sample residual variance, but achieves similar or only slightly better results compared to RP-PCA. Further, we note that the ordering of the factors seems to differ between RP-PCA and PCA. Where the second RP-PCA factor achieves a large increase in Sharpe ratio and a decrease in the RMSE, the second factor in the PCA model achieves almost no improvement in these areas. Conversely, the third factor in the RP-PCA model improves the Sharpe ratio and the RMSE only marginally, while the third PCA factor achieves much larger improvements.

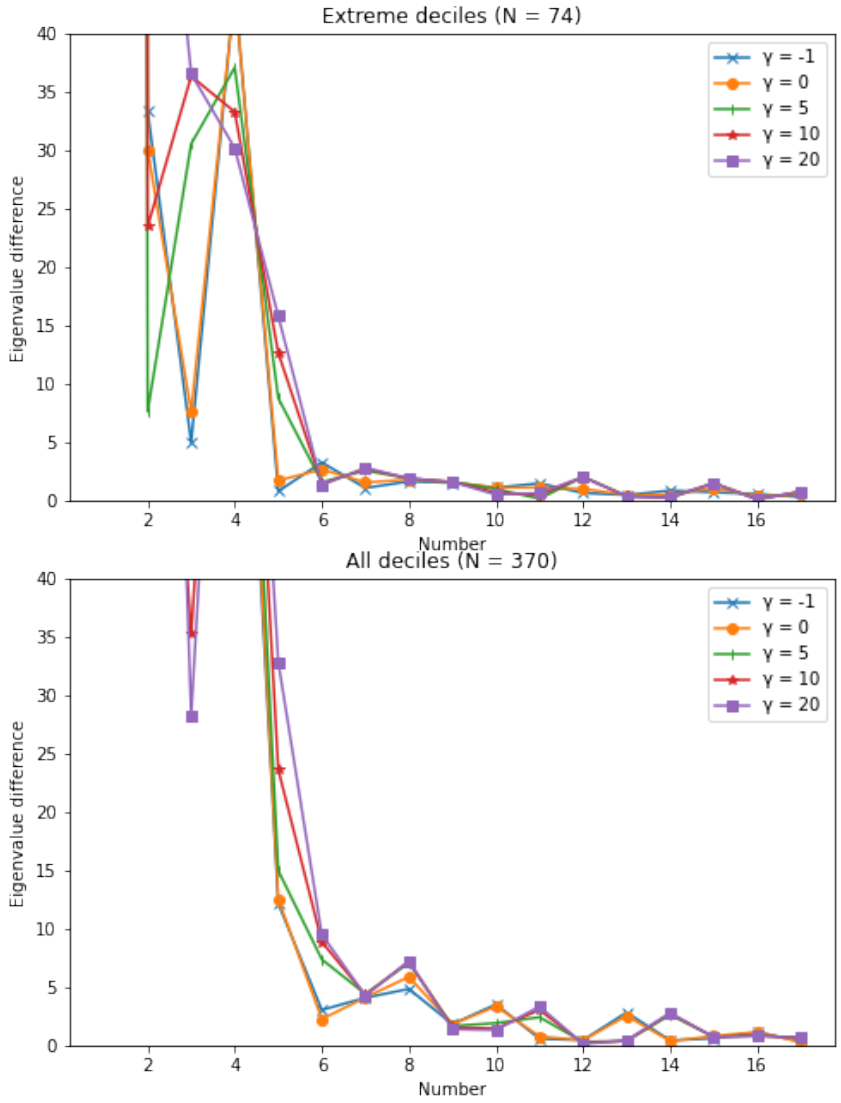


Figure 4: Difference of consecutive eigenvalues of the matrix $\frac{1}{T}\mathrm{X}^\top\mathrm{X} + \gamma\,\overline{\mathrm{X}}\,\overline{\mathrm{X}}^\top$ for the extreme deciles (N = 74) and the full set of single sorted portfolios (N = 370).

The bottom-left image in Figure 5 provides us with some intuition as to why the ordering of the factors might be switched. The third factor in the RP-PCA model contributes more to a reduction in residual variance than the second factor. Since PCA is defined to minimize residual variance it will always prioritize this factor, therefore it is the second factor in the PCA model. RP-PCA on the other hand considers not only the reduction in residual variance but also the improvement in the maximal achievable Sharpe ratio. Since the second RP-PCA factor achieves a sizeable reduction in residual variance, though not as large as the third factor, as well as a drastic increase in the Sharpe ratio, RP-PCA will prioritize this factor.

Out-of-sample results are largely similar, with RP-PCA outperforming standard PCA in terms of Sharpe ratio and RMSE in all cases with more than one factor. Furthermore, the differences in residual variance seem even smaller than in-sample and the RP-PCA and PCA residual variance is now almost indistinguishable. This indicates that while RP-PCA also takes into account the means of the data and achieves a sizeable improvement in terms of Sharpe ratio and RMSE, it does not compromise the ability of PCA to explain a large part of the variance in the data, especially OOS.

Both out-of-sample and in-sample an almost monotonic increase in Sharpe ratio and decrease in RMSE are visible. The fact that this pattern holds up out-of-sample indicates that neither model suffers from excessive overfitting. Interestingly, it seems like the roles of the second and third factors in the PCA model are a lot less defined out-of-sample than in-sample. Where in-sample the second PCA factor contributed very little to the Sharpe ratio and the third factor made for a big increase, OOS the second factor makes for a small increase and the addition of the third factor actually causes a decrease in the maximal Sharpe ratio OOS. A similar pattern is visible in the IS and OOS RMSE for PCA. For RP-PCA we also see a smoothening of the roles of the factors OOS. The third and fourth factors contribute little to the Sharpe ratio and RMSE IS, but OOS they do make for an increase in the Sharpe ratio and a decrease in RMSE. Furthermore, similar to the third factor in PCA, the seventh RP-PCA factor adds to the maximum Sharpe ratio in-sample, but actually leads to a decrease OOS. In conclusion, RP-PCA improves the maximum Sharpe ratio and root-mean-square error in- and out-of-sample while leaving almost the same amount of unexplained variance.

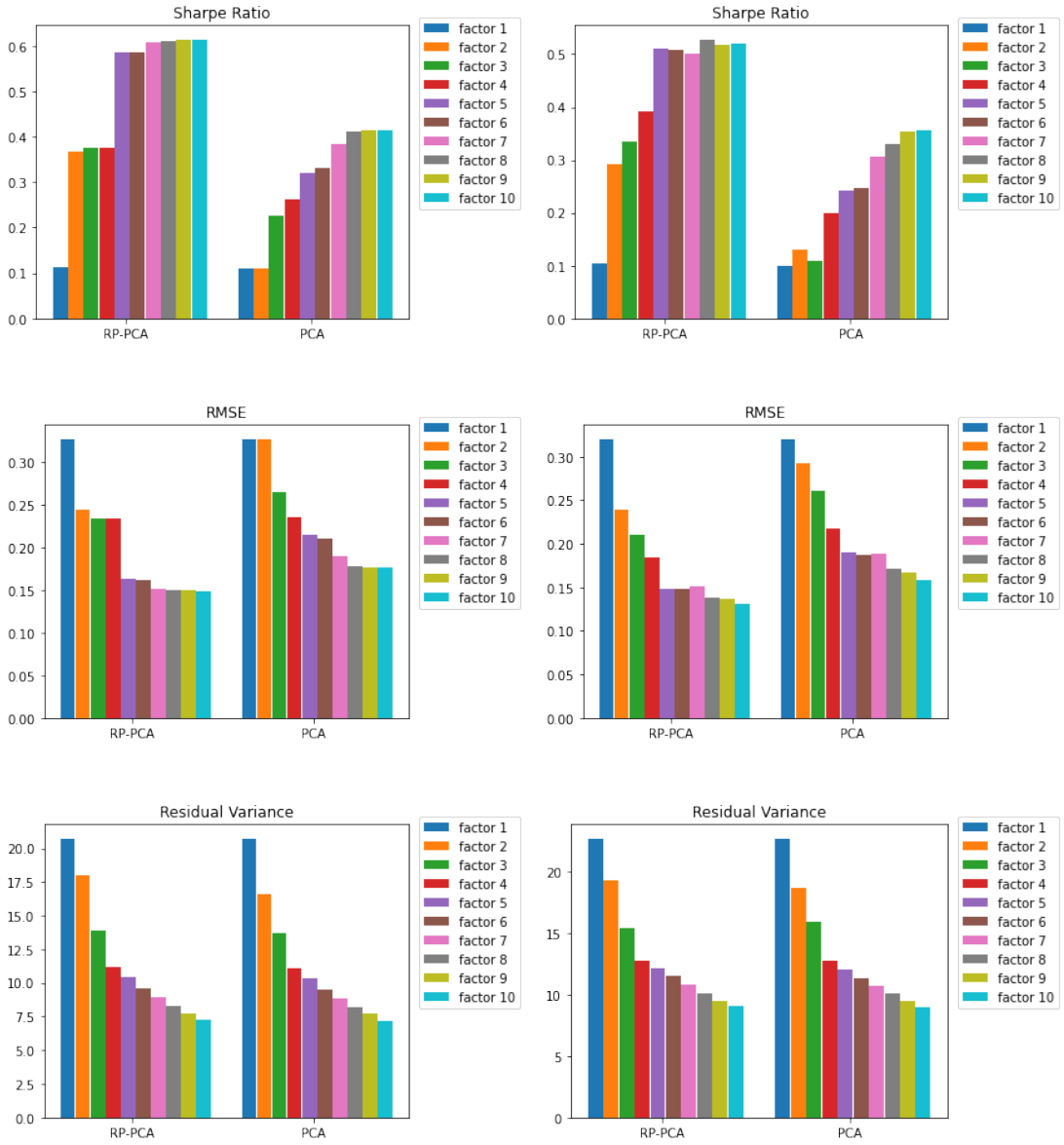
Table 3 shows in- and out-of-sample RMSE, unexplained idiosyncratic variation and Sharpe ratio for standard PCA and RP-PCA. The Fama-French three-factor model, based on market, SMB and HML, and the Fama-French five-factor model, which additionally includes RMW and CMA, are also considered for comparison. The first three rows show results for three-factor models, while the last three rows present empirics for five-factor models.

In the case of three factors, RP-PCA performs best in every metric with the exception of in-sample unexplained idiosyncratic variation, which is to be expected because PCA minimizes in-sample unexplained idiosyncratic variation by definition. OOS however, we notice that RP-PCA is also able to have the lowest idiosyncratic variation, indicating that the RP-PCA model might be less susceptible to the risk of overfitting than standard PCA. The Fama-French model performs considerably worse than both

PCA and RP-PCA in-sample, but out-of-sample it has comparable performance to PCA, though still strictly worse than RP-PCA.

The five-factor models display similar results to the three-factor models. Yet again RP-PCA has a better Sharpe ratio and RMSE both in- and out-of-sample, even achieving almost double the Sharpe ratio as the PCA and Fama-French models. However, in the five-factor case, PCA also leaves less residual variance OOS. The difference between 12.04% for PCA and 12.13% for RP-PCA is not very large, while the differences in RMSE and Sharpe ratios are much larger.

Comparing the five-factor RP-PCA model with the three-factor RP-PCA models, we see that the



(a) In-Sample

(b) Out-of-Sample

Figure 5: Fit for 74 decile-1 and decile-10 portfolios. Maximal Sharpe ratios, root-mean-squared pricing errors, and unexplained idiosyncratic variation for different number of factors. RP weight $\gamma = 10$.

out-of-sample performance of the five-factor model is between 20% and 50% better across the various metrics, with the Sharpe ratio seeing the most improvement and the unexplained idiosyncratic variation seeing the least improvement. Furthermore, if we compare in-sample performance with out-of-sample performance for both models we see that the deterioration is similar for both models. This indicates that the risk of overfitting is not much worse for the five-factor model. We conclude that a five-factor RP-PCA model is the preferred model for all three datasets.

Table 3: Performance measures RP-PCA, PCA and Fama-French models

	In-Sample			Out-of-Sample		
	RMSE	$\overline{\sigma}_e^2$	Sharpe	RMSE	$\overline{\sigma}_e^2$	Sharpe
RP-PCA(3)	**0.24**	13.95	**0.41**	**0.21**	**15.40**	**0.34**
PCA(3)	0.27	**13.74**	0.23	0.26	15.88	0.11
Fama-French(3)	0.31	17.49	0.21	0.25	16.54	0.16
RP-PCA(5)	**0.16**	10.45	**0.61**	**0.15**	12.13	**0.51**
PCA(5)	0.21	**10.30**	0.32	0.19	**12.04**	0.24
Fama-French(5)	0.26	16.05	0.32	0.19	13.91	0.31

In-sample and out-of-sample root-mean-square error (RMSE), idiosyncratic variation ($\overline{\sigma}_e^2$), and maximal Sharpe ratio. Bold numbers indicate the best performing models. Results are for the 74 decile-1 and decile-10 portfolios.

Figure 6 shows in- and out-of-sample root-mean-squared cross-sectional pricing errors (α's) for each anomaly based on models with five factors. The alphas are first computed for each asset, in the OOS case the time-series mean alphas per asset. Thereafter all alphas are squared and the mean per anomaly, so the average of the decile-1 and decile-10 alpha, is taken. Lastly, the square root of the alphas is taken. The anomalies in Figure 6 are sorted by the Sharpe ratio of the decile-10 minus decile-1 return. The blue bars represent the RP-PCA alphas and the orange bars represent the PCA alphas.

Both in- and out-of-sample the PCA pricing errors for high Sharpe ratio anomalies are much higher than the ones for RP-PCA. Especially for the portfolios where PCA produces the highest pricing errors RP-PCA is able to find a lot of improvement. However, the dispersion seems to be a bit smaller OOS compared to IS. In-sample we see that PCA produces much smaller pricing errors for some low-Sharpe ratio portfolios, but OOS this difference has almost entirely disappeared. In conclusion, the RP-PCA model lowers pricing errors of most portfolios compared to standard PCA, especially for high Sharpe ratio portfolios the reduction is grand.

4.2.3 Time-series versus cross-sectional factors

In this section, we will consider the properties of the individual factors in the models, whereas so far we have focused more on the performance of various models as a whole. Table 4 shows the mean, variance and Sharpe ratio of the first ten factors as estimated by RP-PCA and PCA. All factors have

been normalized such that their mean return is positive. For both PCA and RP-PCA the variance of the first factor is a magnitude larger than the other factors. The factors are sorted in descending order based on their variance for PCA, as PCA by construction minimizes the amount of residual variance. For RP-PCA however, we observe that this ordering is broken briefly because the second factor has lower variance than the third factor. RP-PCA is designed to not only minimize residual variance but also to detect weaker but high Sharpe ratio factors. This aspect of RP-PCA is reflected in its prioritizing of the higher-Sharpe, lower-variance factor over the lower-Sharpe higher-variance factor, whereas PCA does the exact opposite.

Upon closer inspection of the first five RP-PCA factors, we find that the first, second and fifth factors all have relatively high means and Sharpe ratios. On the other hand, the third and fourth factors have low means but do explain a significant part of the total variation. This indicates that the second and fifth factors might correspond with priced factors that capture the cross-section of returns, while the third and fourth factors are not priced but capture the co-movement of returns.

Table 5 shows the root-mean-square error (RMS), unexplained idiosyncratic variation ($\overline{\sigma}_e^2$), and the

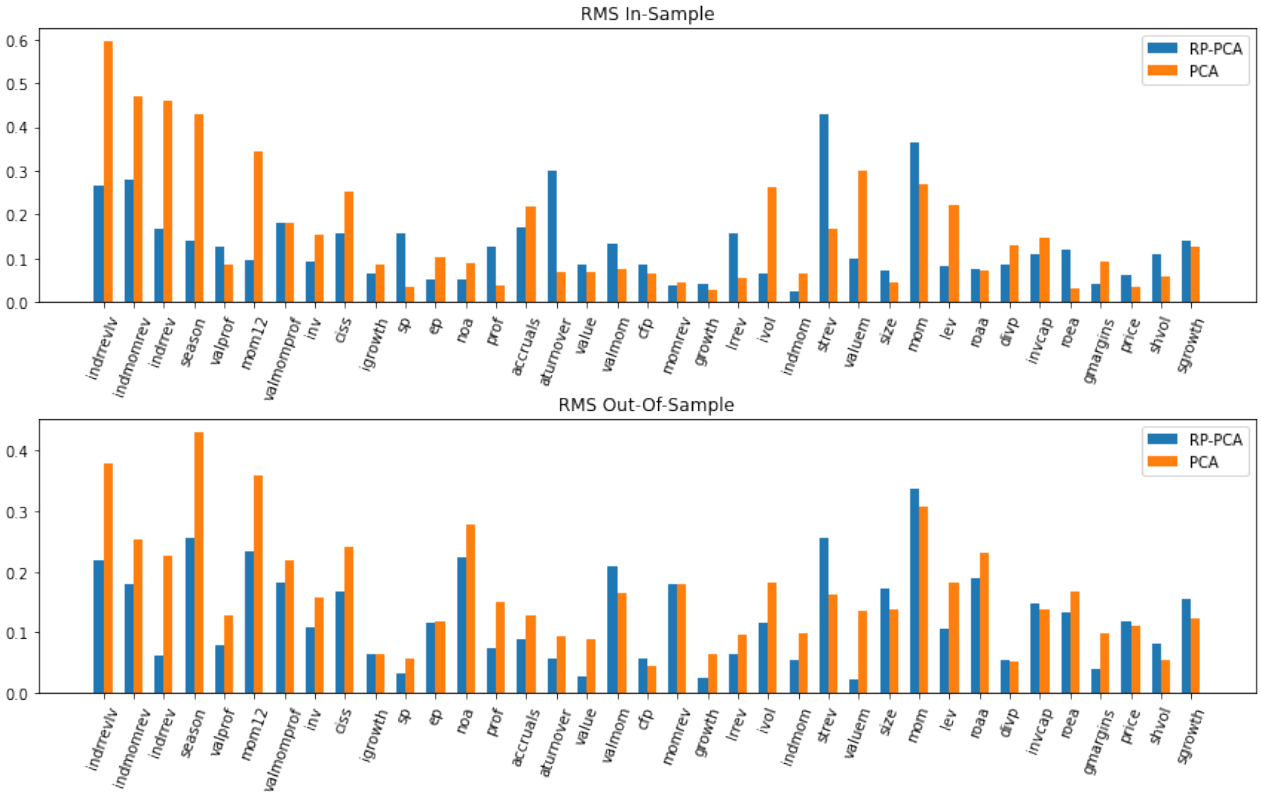


Figure 6: RMS of time-series α's by characteristic. First and last deciles of 37 single-sorted anomaly portfolios. Shown is the square root of the average of the squared decile-1 and decile-10 alpha per anomaly, anomalies are sorted on the Sharpe ratio of the decile-10 minus decile-1 return. Alphas are calculated based on five-factor PCA and RP-PCA models, where the OOS alpha corresponds to the mean alpha over the entire period.

Table 4: Individual factors

	RP-PCA			PCA		
Factor	Mean	Variance	Sharpe	Mean	Variance	Sharpe
1	5.14	1911.94	0.12	4.83	1941.87	0.11
2	2.14	79.11	0.24	0.18	102.39	0.02
3	0.09	101.90	0.01	1.65	69.03	0.20
4	0.06	65.45	0.01	1.05	64.06	0.13
5	0.43	31.48	0.08	0.83	20.18	0.18
6	0.02	19.53	0.00	0.34	19.39	0.08
7	0.07	18.13	0.02	0.79	16.14	0.20
8	0.02	15.42	0.01	0.57	15.08	0.15
9	0.02	13.54	0.01	0.20	13.46	0.05
10	0.02	11.96	0.01	0.04	11.95	0.01

Mean, variance and Sharpe ratio for the first ten factors estimated by RP-PCA and PCA. The RP-PCA factors are estimated using $\gamma = 20$.

Sharpe ratio for various subsets of the five-factor PCA model. A model containing only the first factor is able to produce an RMS of 0.32, $\overline{\sigma}_e^2$ of 22.70, and a Sharpe ratio of 0.1 OOS. In terms of RMS and idiosyncratic variation, this is about half of the performance of the full five-factor model, while the Sharpe ratio is five times lower than the full model. A subset containing only factors one, two and five is able to achieve almost the same Sharpe ratio and RMS IS as the full model, however, OOS the difference is larger. Furthermore, the full model leaves significantly less residual variation, indicating that a model containing only factors one, two and five is able to capture priced factors well, but misses some time-series co-movement in the data.

The subset containing only factors one, three and four on the other hand produces only a slightly better RMS and Sharpe ratio than the one-factor model. However, it is able to reduce unexplained idiosyncratic variation significantly compared to the one-factor model. This shows that the one, three and four-factor model is not able to adequately capture all priced risk factors, but it does capture time-series co-movement in the data.

When examining the models where the first factor is not included, this pattern becomes more clear. The subset of factors two and five make for a high Sharpe ratio, but also high RMS and high residual variance IS. Interestingly, OOS the RMS and unexplained idiosyncratic variation are much smaller than IS. The reason for this is that the factors are orthogonal in-sample. So higher order factors can only capture variation that has not yet been captured by the lower order factors. Since the first factor already captures a significant part of the total variation there is not much variation left to be captured by the higher order factors in-sample. Out-of-sample however, the factors are not necessarily orthogonal so the

higher order factors can capture some part of the variation that the first factor would normally capture. The model containing only factors three and four has high RMS and residual variation both IS and OOS. The Sharpe ratio is only 0.01 in-sample, the lowest out of all considered subsets. OOS this Sharpe ratio improves to 0.15, likely due to the factors not necessarily being orthogonal OOS and thus the third and fourth factors can account for a portion of the Sharpe ratio that would be due to the first and second factors in-sample.

To conclude this section we identify the first RP-PCA factor as important both for the cross-sectional and time-series fit of the model, likely representing a market factor. The second and fifth factors correspond to high-Sharpe ratio factors which contribute to a decrease in RMS, but less to the residual variance. These factors are priced factors important for explaining the cross-section of returns. Lastly, we identify the third and fourth factors as low-Sharpe ratio factors which are not priced. They contribute most to decreasing residual variance and are important for explaining the cross-section of returns.

Table 5: Model fit for subsets of factors

	In-sample			Out-of-Sample		
	RMS	$\overline{\sigma}_e^2$	Sharpe	RMS	$\overline{\sigma}_e^2$	Sharpe
(1,2,3,4,5)	0.16	10.45	0.61	0.15	12.13	0.51
(1)	0.33	20.76	0.12	0.32	22.70	0.10
(1,2,5)	0.19	17.22	0.60	0.21	17.17	0.34
(2,5)	1.75	51.77	0.30	0.30	19.78	0.08
(1,3,4)	0.33	13.92	0.12	0.30	16.20	0.14
(3,4)	0.65	93.10	0.01	0.70	88.69	0.15

Root-mean-square error (RMS), unexplained idiosyncratic variation ($\overline{\sigma}_e^2$), and the Sharpe ratio for subsets of the five-factor RP-PCA model. The factors are estimated using $\gamma = 20$.

4.2.4 Composition of the SDF

In this section, we will study the composition of the SDF. The implied SDF is perfectly negatively correlated with the tangency portfolio of the estimated factors. Since the estimated factors themselves are linear combinations of the test assets, we can express the SDF as a linear combination of the test assets. Figure 7 shows the SDF weights for each of the 74 test portfolios, sorted by the Sharpe ratio of the decile-10 minus decile-1 return. The top figure shows the weights as estimated by RP-PCA, while the bottom figure shows the weights estimated by PCA. The orange bars indicate the decile-10 weights and the blue bars the decile-1 weights. For both estimation methods, most decile-10 weights are positive and decile-1 weights are negative, thus creating long-short portfolios. However, the weights do not necessarily sum to one and we see in some cases that both the decile-1 and decile-10 weights are positive (value), or both the weights are negative (indrrev for PCA).

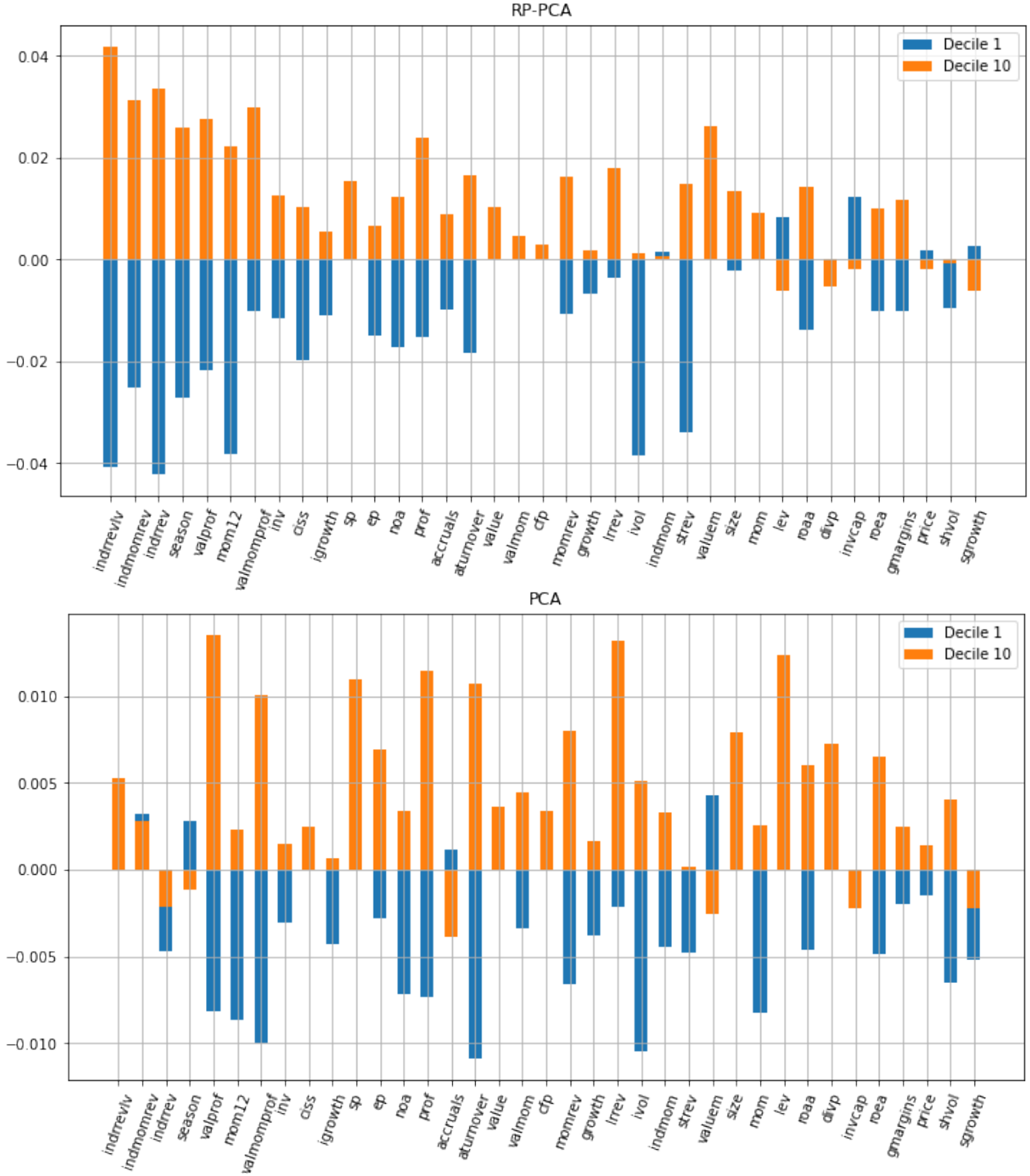


Figure 7: Asset weights in RP-PCA and PCA SDFs. Loadings of the SDF for each of the 37 anomalies as estimated by RP-PCA (top) and PCA (bottom). The anomalies are sorted by there Sharpe ratio.

For RP-PCA the magnitude (in absolute value) of the asset weights in the SDF seems to correlate greatly to the Sharpe ratio of the asset. The largest absolute weights are given to the anomalies with the highest Sharpe ratio, while the anomalies with low Sharpe ratios also receive low absolute weights in the SDF. This correlation between the Sharpe ratio and SDF weight seems to not be present in the weights

estimated by PCA at all.

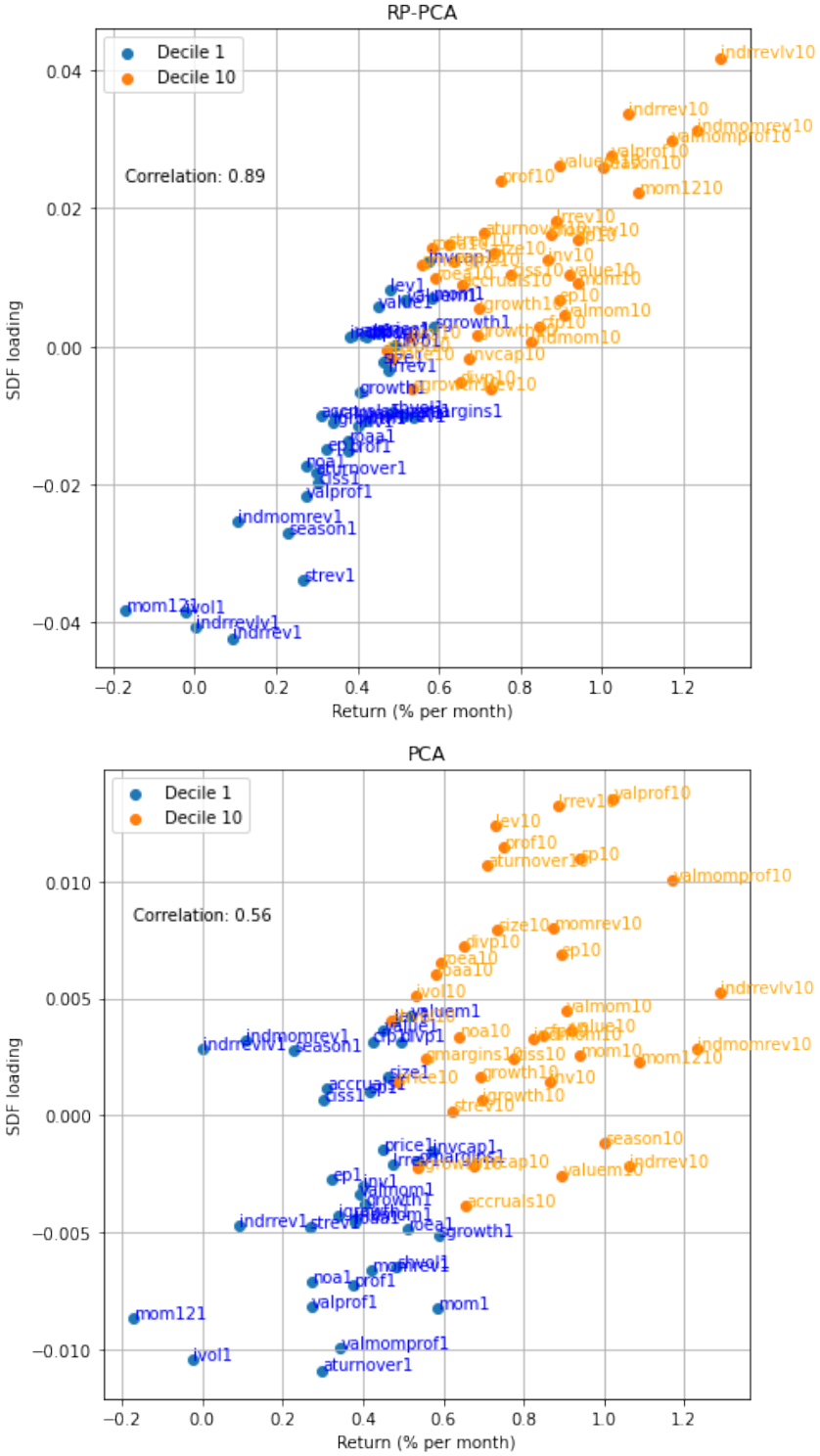


Figure 8: Portfolio returns and SDF loadings. Average portfolio return and weight in the SDF for 74 decile-1 and decile-10 portfolios. Top figure shows the RP-PCA SDF loadings while the bottom figure shows the loadings for standard PCA. Decile-1 returns are given in blue and decile-10 returns in orange.

Figure 8 shows the average return of each portfolio together with its weight in the SDF for the RP-PCA (top) and PCA (bottom) models. The correlation between average portfolio return and SDF weight is also printed in the figure. The SDF weights for RP-PCA are strongly correlated with the average portfolio returns, with the correlation being 0.89. The scatterplot shows that portfolios with high average returns, such as indrrevlv decile-10, have large positive weights, while low-return portfolios, such as indrrevlv decile-1, receive large negative weights.

For the PCA model, there is a similar correlation, but it is not nearly as strong. The overall correlation coefficient between average returns and SDF weights is only 0.56 in this case. We see that some portfolios with relatively high returns, such as indrrev decile-10, receive negative weights in the SDF. On the other hand, some low-return portfolios, such as indrrevlv decile-1, receive a positive weight in the SDF.

The difference in correlation between average returns and SDF weights between the PCA and RP-PCA models gives an indication of why the maximum possible Sharpe ratio is so much higher for RP-PCA than PCA, as is shown in Table 3.

4.3 Robust three-pass estimator

In this section risk-premia of various traded factors will be estimated using a variety of techniques. We will examine whether a version of the three-pass estimator by Giglio and Xiu (2021) which is based on RP-PCA instead of PCA is able to add to the current selection of risk-premium estimation techniques.

The standard three-pass estimator uses PCA in the first step to recover a rotation of the factor space of the returns. The estimator is based on the assumption that the factors driving returns are strong enough to be detected by PCA. We have seen in this paper that RP-PCA is able to detect weak factors more consistently than PCA, therefore it is a good replacement for PCA in the first step of the three-pass estimator, making the assumption that a rotation of the whole factor space is discovered in the first step more plausible. From now on the version of the three-pass estimator which uses RP-PCA in the first step will be referred to as robust three-pass estimator.

The robust three-pass estimator will be compared against its regular counterpart as described in Giglio and Xiu (2021), as well as three different two-pass regression models, and three mimicking portfolio approaches. The two-pass regressions consist of a model with no control variables in the regression, a model with only the market as a control variable, and a model with the three Fama-French factors as controls (market, SMB, and HML). For the mimicking portfolio approach the risk factor will be projected onto only the market factor, the three Fama-French factors (market, SMB and HML), and onto the set of 74 decile-1 and decile-10 anomaly portfolios. The complete set of 370 decile-1 through 10 portfolios is used to run the two- and three-pass regressions. Note that it would technically be possible to project the risk factor onto the entire set of 370 portfolios using the mimicking portfolio approach since $N = 370 < 650 = T$, but this approach is inefficient since N is still very large relative to T. For a more detailed description of the auxiliary models see section 2.3.

The robust three-pass estimator approach will be tested by estimating the risk premia of the five

factors in the Fama and French (2015) model. Since these are all traded factors, the risk premium of each factor in the sample period can be calculated by taking the mean return of the factor in this period. The mean return can thus be used as a target and various models can be compared by how close the risk premium estimate is to the mean return.

Table 6 gives the average and estimated risk-premia for the five Fama-French factors using three versions of the two-pass estimator, three versions of the mimicking portfolio approach and the normal and robust three-pass estimator. Following Giglio and Xiu (2021) the three-pass method is implemented using seven factors. Note first that any version of the mimicking portfolio approach where the risk factor is projected onto itself will deliver a risk-premium estimate exactly equal to the mean. This is because it can perfectly project the risk factor onto itself and take zero weights for any other possible factors, the average return of the projection will then be equal to the average return of the factor itself. Because of this fact, the mimicking portfolio approach with three Fama-French factors delivers perfect estimates of the risk-premium for the first three Fama-French factors. However, performance is significantly worse for the last two.

The last row of table 6 shows the root-mean-square error (RMSE) which is calculated by taking the root of the mean squared difference between the factor's average return and the estimated risk premium. It gives an indication of a model's overall performance, where all estimates have to be close to the average return and large differences are punished harder. The best performing models according to this criterion are the mimicking portfolio approach based on the 74 extreme deciles and the robust three-pass estimator, with an RMSE of 0.05 and 0.06 respectively.

Among the second best models is the regular three-pass estimator, having an RMSE of 0.11. When comparing the regular three-pass estimator directly with the robust three-pass estimator, we observe that the robust three-pass estimator produces equal or better risk-premium estimates for every risk factor. Actually, when considering more than two decimals, the robust three-pass estimator produces strictly better estimates than the regular three-pass estimator. Specifically, in the cases where the regular three-pass estimate is far off, such as for RMW and CMA, the robust three-pass estimator is able to find large improvements.

Tables 7 and 8 in the appendix provide robustness checks for these conclusion. Firstly, Table 7 checks the robustness of the mimicking portfolio, regular three-pass, and robust three-pass models with regard to the set of assets the risk factor is being projected onto. We calculate the estimated risk premium for each of the five Fama-French factors, using the eight different sets of double-sorted portfolios, for each model. Thereafter, the RMSE of these five risk premium estimates is calculated for each model, this is the figure given in the table. The mean and standard deviation of the eight RMSEs is also given. We see that the robust three-pass estimator still outperforms the standard three-pass estimator in six out of eight cases. Furthermore, the average RMSE is lowest for mimicking portfolio and robust three-pass estimator at 0.14, while the standard deviation of the RMSE of the three-pass estimator is slightly higher at 0.05. Note here that the (robust) three-pass estimator is designed to work best in large N environments, while

the mimicking portfolio approach becomes less efficient as N increases. With this notion, we can conclude that the (robust) three-pass estimator is not sensitive to the choice of base assets.

Table 8 shows average and estimated risk premia, as in Table 6, but for a much larger set of risk factors. We consider all 37 anomalies from the Kozak, Nagel and Santosh dataset, the risk factor is defined as the decile-10 return minus decile-1 return, and anomalies in Table 8 have been sorted by their Sharpe ratio [3]. The last row reports the RMSE, which is defined the same way as for Table 6. The RMSE for the robust three-pass estimator is the lowest out of all models considered, even being three times as low as its non-robust counterpart.

This indicates that the robust three-pass estimator is able to detect weaker factors which are important for determining the risk premium of some factors. We conclude that the robust RP-PCA estimator reduces the risk of not identifying factors that are important in determining risk-premia and is therefore able to produce more accurate estimates than the standard three-pass estimator.

Table 6: Average and estimated risk premia for the five Fama-French factors

		Two-Pass			Mimicking Portfolio			Three-pass	
	Mean	No Controls	Market	FF3	Market	FF3	Extremes	Regular	Robust
Market	0.53	0.58	0.58	0.56	0.53	0.53	0.52	0.54	0.52
SMB	0.21	0.98	-0.01	-0.01	0.11	0.21	0.20	0.19	0.23
HML	0.34	-1.11	0.32	0.31	-0.08	0.34	0.30	0.23	0.26
RMW	0.26	-1.01	0.23	0.39	-0.06	-0.08	0.22	0.16	0.24
CMA	0.30	-0.66	0.27	0.33	-0.09	0.11	0.20	0.12	0.20
RMSE	-	1.02	0.11	0.12	0.30	0.18	0.05	0.11	0.06

Mean return and estimated risk premia for the five Fama-French factors using various estimation methods. The two-pass regressions use no controls, only the market factor as control, and the three Fama-French factors as control. Mimicking portfolio approach is applied to only the market portfolio, the three Fama-French factors and 74 decile-1 and decile-10 anomaly portfolios. The three-pass estimator is implemented in the standard way as described in Giglio and Xiu (2021) and a robust version where RP-PCA is used in the first step, rather than standard PCA. Both three-pass estimators use seven factors and the robust version uses a gamma parameter in the RP-PCA step of 20. The root-mean-square error (RMSE) is the root of the mean of the squared difference between the average factor return and the risk premium estimate.

5 Conclusion

This paper first performs a large empirical analysis of the risk premium PCA (RP-PCA) estimator by Lettau and Pelger (2020). This estimator aims to identify latent factors in asset pricing by not only

[3]Note that in this table a mimicking portfolio approach with five Fama-French factors is used rather than the decile-1 and decile-10 portfolios as before. This is because the risk factors are linear combinations of these decile-1 and decile-10 returns. Projecting the risk factor onto these returns would thus always result in a risk premium estimate equal to the mean return.

incorporating information in the second moment of asset returns, like standard PCA but also using information contained in the first moment. Consequently, the objective of PCA is to extract factors that both capture time-series comovement as well as reduce cross-sectional pricing errors. RP-PCA can be seen as a generalized version of PCA and can be simply implemented by using standard eigenvalue decomposition after adding a variable to the variance-covariance matrix of asset returns.

Second, the RP-PCA estimator is integrated into the three-pass risk estimation framework by Giglio and Xiu (2021). The aim of the standard three-pass framework is to estimate risk premia without the risk of omitted variable or measurement error bias. However, it relies on the assumption that the factors driving stock returns are sufficiently strong enough to be detected by standard PCA. RP-PCA is shown to detect weak factors with higher accuracy than standard PCA, therefore I use RP-PCA over standard PCA in the three-pass estimator, which is named the robust three-pass estimator.

The empirical findings can be summarized as follows. For eight sets of double-sorted portfolios with N = 25, a three-factor RP-PCA model is able to reduce pricing errors and unexplained variation while increasing the maximum possible Sharpe ratio out-of-sample compared to standard PCA and the three-factor Fama-French model. The difference arises because RP-PCA is able to identify factors that are too weak to be detected by standard PCA, or because the ordering of the factors is different. A five-factor RP-PCA model is sufficient to capture the time-series and cross-sectional characteristics of 370 single-sorted decile portfolios based on 37 anomaly characteristics. The RP-PCA model is able to produce lower pricing error and higher Sharpe ratio than standard PCA and the five-factor Fama-French model, both in- and out-of-sample. Furthermore, while standard PCA by definition minimizes in-sample unexplained idiosyncratic variance, RP-PCA produces similar results in-sample and is even able to have less residual variance in some cases out-of-sample. Analysis of the roles of the individual factors reveals that the first factor in both PCA and RP-PCA can be seen as a market factor. It is important for explaining the time-series comovement of returns, as well as the cross-section of returns. The second and fifth RP-PCA factors have high means and high Sharpe ratios but contribute relatively little to the reduction of idiosyncratic variance. These factors are important for explaining the cross-section of returns, while their roll in explaining time-series comovement is less defined. Contrarily, factors 3 and 4 have low means and Sharpe ratios, but contribute more to the reduction in idiosyncratic variance. They can be seen as important for time-series comovement, but less important for the cross-section of returns. RP-PCA is able to achieve a twice as high Sharpe ratio out-of-sample compared to PCA. This is possible because the cross-sectional penalty imposed by RP-PCA makes for a strong correlation between mean asset returns and weight in the SDF. This correlation is much lower for standard PCA.

The implementation of the RP-PCA estimator into the three-pass estimator produces very promising results. For all five Fama-French factors the robust three-pass estimator produces strictly better risk premium estimates than the standard three-pass estimator. The root-mean-square error of the robust three-pass estimator is twice as low as standard PCA over the five Fama-French factors and even three times as low over the full set of 37 anomaly high-minus-low factors. The results are robust to the choice

of base assets.

The RP-PCA approach provides a reliable way of estimating factor models for asset returns. The implementation into the three-pass estimator allows for risk-premium estimation robust to measurement error, omitted variables, and weak factors. Future research can focus on the implementation of RP-PCA into other methods that rely on standard PCA, such as the SDF estimation by Kozak et al. (2020) and the instrumented-PCA estimator in Kelly et al. (2019). Furthermore, the robust three-pass estimator can be tested on its sensitivity to the choice of the number of factors used in the first step.

The objective of this paper is to show the merit of using RP-PCA in the three-pass estimator, to do so only the risk premia of traded factors are estimated since they can be compared with the mean return of the factor to provide a measure of accuracy. Now that the concept of the robust three-pass estimator has been shown to work in this paper, it can be applied in future research to different, possibly non-traded, factors.

6 Appendix

Table 7: RMSE of the estimation of the risk premium of five Fama-French factors using different double-sorted portfolios as base assets.

	Mimicking Portfolio	Regular Three-Pass	Robust Three-Pass
Book-to-Market	0.10	0.10	**0.09**
Accruals	**0.23**	0.27	0.26
Investment	**0.10**	0.12	0.11
Operating profitability	0.17	0.17	**0.16**
Residual Variance	0.11	**0.08**	0.09
Variance	0.11	**0.09**	0.10
Momentum	0.16	0.18	**0.16**
Reversal	**0.16**	0.18	0.17
Mean	0.14	0.15	0.14
Standard Deviation	0.04	0.06	0.05

RMSE when estimating the risk premium of the five factors in the Fama-French model. Left column shows which assets are used as the base assets, the assets onto which the risk factor is projected. The last two rows give the mean and standard deviation of the RMSE over the 8 double-sorted portfolios. Best performing models are given in bold.

Table 8: Average and estimated risk premia for 37 high minus low anomaly portfolios

		Two-Pass			Mimicking Portfolio			Three-pass	
	Mean	No Controls	Market	FF3	Market	FF3	FF5	Regular	Robust
indrrevlv	1.29	1.39	0.76	0.60	0.11	0.22	0.25	0.09	0.96
indmomrev	1.12	3.37	0.99	1.03	0.04	0.04	0.14	0.08	0.77
indrrev	0.97	1.32	0.22	0.09	0.19	0.28	0.22	0.12	1.07
season	0.77	3.22	0.45	1.42	0.07	-0.02	0.01	0.04	0.57
valprof	0.75	-0.98	0.62	1.06	-0.04	0.32	0.43	0.03	0.94
mom12	1.26	-3.01	0.69	0.87	-0.13	-0.32	0.02	-0.01	1.08
valmomprof	0.83	-2.39	0.70	0.83	-0.08	0.10	0.20	0.03	0.98
inv	0.46	-2.00	0.88	0.92	-0.06	0.13	0.20	-0.01	0.37
ciss	0.47	-0.87	0.64	1.02	-0.18	-0.07	0.12	-0.13	0.42
igrowth	0.36	-2.19	0.87	0.92	-0.03	0.13	0.25	0.01	0.25
sp	0.52	2.27	0.53	0.37	0.05	0.58	0.69	0.17	0.67
ep	0.57	-1.86	0.74	1.06	-0.16	0.23	0.49	0.01	0.65
noa	0.37	-0.54	0.09	0.89	0.01	-0.08	-0.05	-0.02	0.34
prof	0.38	2.43	-0.10	0.57	0.04	-0.12	0.10	0.01	0.34
accruals	0.35	-3.44	0.93	0.75	-0.03	-0.01	-0.06	-0.02	0.19
aturnover	0.41	2.36	0.59	0.75	0.08	0.07	0.40	0.10	0.68
value	0.47	1.39	0.52	0.39	-0.02	0.58	0.62	0.13	0.61
valmom	0.52	-2.27	0.85	0.81	-0.10	0.24	0.32	0.08	0.68
cfp	0.42	-0.22	0.58	0.41	-0.04	0.50	0.58	0.11	0.52
momrev	0.45	3.46	0.37	-0.29	0.02	0.36	0.32	0.07	0.46
growth	0.29	-1.71	0.65	0.63	-0.09	0.20	0.41	-0.00	0.27
lrrev	0.41	1.96	0.35	-0.09	-0.00	0.49	0.48	0.06	0.61
ivol	0.55	-1.50	0.12	0.44	-0.46	-0.46	0.09	-0.31	0.43
indmom	0.45	-3.05	0.76	0.83	-0.11	-0.13	0.07	0.04	0.35
strev	0.36	1.91	0.02	-0.17	0.20	0.27	0.19	0.12	0.99
valuem	0.38	2.68	0.14	-0.11	0.07	0.66	0.45	0.10	0.45
size	0.27	1.93	-0.02	-0.19	0.10	0.38	0.28	0.09	0.27
mom	0.36	-2.66	0.51	0.66	-0.12	-0.24	-0.08	0.00	0.55
lev	0.25	2.54	0.56	-0.05	0.05	0.63	0.57	0.20	-0.02
roaa	0.20	-1.20	0.06	0.56	-0.18	-0.36	-0.02	-0.16	0.21
divp	0.16	-1.77	0.49	-0.20	-0.22	0.30	0.31	-0.07	0.18
invcap	0.10	-1.55	0.48	0.43	-0.23	0.18	0.51	-0.04	0.07
roea	0.08	-1.26	0.05	0.41	-0.19	-0.38	-0.00	-0.17	0.13
gmargins	0.02	-1.58	-0.30	-0.04	-0.07	-0.34	-0.33	-0.15	-0.00
price	0.03	-1.87	0.10	0.29	-0.28	-0.65	-0.29	-0.20	-0.08
shvol	-0.01	-1.21	0.13	0.31	-0.42	-0.22	0.13	-0.30	0.10
sgrowth	-0.05	-1.71	0.45	-0.25	-0.10	0.19	0.36	-0.02	0.08
RMSE	-	2.18	0.33	0.42	0.59	0.55	0.45	0.54	0.18

References

Bai, J. and Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of econometrics*, 212(1):78–96.

Black, F., Jensen, M. C., Scholes, M., et al. (1972). The capital asset pricing model: Some empirical tests.

Breeden, D. T., Gibbons, M. R., and Litzenberger, R. H. (1989). Empirical tests of the consumption-oriented capm. *The Journal of Finance*, 44(2):231–262.

Connor, G., Hagmann, M., and Linton, O. (2012). Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica*, 80(2):713–754.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.

Fan, J., Liao, Y., and Wang, W. (2016). Projected principal component analysis in factor models. *Annals of statistics*, 44(1):219.

Giglio, S. and Xiu, D. (2021). Asset pricing with omitted factors. *Journal of Political Economy*, 129(7):1947–1990.

Harvey, C. R., Liu, Y., and Zhu, H. (2015). ... and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1):5–68.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

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

Lamont, O. A. (2001). Economic tracking portfolios. *Journal of Econometrics*, 105(1):161–184.

Lettau, M. and Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, 33(5):2274–2325.

Lönn, R. and Schotman, P. C. (2018). Empirical asset pricing with many assets and short time series. *Available at SSRN 3278229*.