Factors that Fit the Time Series and Cross-Section of Stock Returns[☆]

Martin Lettau

Haas School of Business, University of California at Berkeley, Berkeley, CA 94720, lettau@berkeley.edu,NBER,CEPR

Markus Pelger

Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, mpelger@stanford.edu

Abstract

We propose a new method for estimating latent asset pricing factors that fit the time-series and cross-section of expected returns. Our estimator generalizes Principal Component Analysis (PCA) by including a penalty on the pricing error in expected returns. We show that our estimator strongly dominates PCA and finds weak factors with high Sharpe-ratios that PCA cannot detect. Studying a large number of characteristic sorted portfolios we find that five latent factors with economic meaning explain well the cross-section and time-series of returns. We show that out-of-sample the maximum Sharpe-ratio of our five factors is more than twice as large as with PCA with significantly smaller pricing errors. Our factors are based on only a subset of the stock characteristics implying that a significant amount of characteristic information is redundant.

Keywords: Cross section of returns, anomalies, expected returns, high-dimensional data, latent factors, weak factors, PCA

JEL classification: C14, C52, C58, G12

This draft: August 22, 2018

[☆]We thank Svetlana Bryzgalova, John Cochrane, Jianqing Fan, Kay Giesecke, Bob Hodrick, Per Mykland, Serena Ng, Viktor Todorov, Dacheng Xiu and seminar participants at Columbia, Chicago, Stanford, UC Berkeley, Zürich, Toronto, Boston University, Humboldt University, Frankfurt, Ulm, Bonn and the conference participants at the NBER-NSF Time-Series Conference, SoFiE, Western Mathematical Finance Conference and INFORMS.

1. Introduction

The fundamental insight of asset pricing theory is that the cross-section of expected returns should be explained by exposure to systematic risk factors. Finding the "right" factors has become the central question of asset pricing. Harvey et al. (2016) document that more than 300 published candidate factors have predictive power for the cross-section of expected returns. As argued by Cochrane (2011), this "factor zoo" leads to the questions of which risk factors are really important and which factors are subsumed by others.

This paper develops a new statistical method to find the most important factors for explaining asset returns and to bring order into the chaos of factors. Our methodology uses large financial data sets to identify factors that simultaneously explain the time series *and* cross-section of stock returns. The estimation approach is a generalization of the widely utilized Principal Component Analysis (PCA), *e.g.* in Connor and Korajczyk (1986, 1988). Statistical factor analysis based on PCA has problems identifying factors with a small variance that are important for capturing the cross-section of returns. Our estimator, Risk-Premium PCA (RP-PCA), can be interpreted as generalized PCA with a penalty term to account for cross-sectional pricing errors, thus combining PCA factor analysis with the Arbitrage Pricing Theory (APT) of Ross (1976).

The RP-PCA method considers three key elements concurrently: First, we find statistical factors instead of selecting from a set of pre-specified and potentially miss-specified factors. Second, we use information from large panel data sets; *i.e.* a large number of time observations for many financial assets. Using the information provided by a large number of anomaly-sorted portfolios allows identification of all relevant factors, but the "curse of dimensionality" also makes the statistical estimation significantly more challenging. Third, RP-PCA identifies factors with high Sharpe-ratios that fit the cross-section of expected returns and not only co-movement in the time series dimension: This is important as low Sharpe-ratio factors that can only explain time-series comovements, for example industry factors, do not seem to be important for the cross-section of expected returns and hence do not capture the risk-return trade-off.[1]

This paper focuses on the empirical estimation using RP-PCA while the econometric asymptotic theory for the estimator is developed in Lettau and Pelger (2018a). That paper shows that RP-PCA dominates conventional estimation based on PCA along a number of dimensions, especially when factors are "weak". We define "weak" factors as factors that affect only a subset of the underlying assets. Weak factors are harder to detect than "strong" factors that affect all assets (such as the "market" factor). Many anomaly-based factors are

[1] Bryzgalova (2016) shows that time- and cross-sectional information need to be considered together to avoid spurious factors.

more likely to be "weak" factors and RP-PCA can find weak factors with high Sharpe-ratios, which cannot be detected with PCA, even if an infinite amount of data is available.

Our estimator is grounded in economic theory. The objective of finding factors that can explain co-movement and the cross-section of expected returns simultaneously is based on the fundamental insight of of Ross' APT. This contrasts our estimator with other regularized estimation approaches that use ad-hoc penalty functions unrelated to the economic problem. We show theoretically and empirically that including the additional information of arbitrage pricing theory in the estimation of factors leads to factors that have better out-of-sample pricing performance.

Our empirical analysis is based on three data sets. First, we analyze factors from monthly returns of double-sorted portfolios for a sample of monthly returns from July 1963 to December 2017. Second, we study the monthly returns of 370 decile portfolios based on single-sorts on 37 characteristics. Lastly, we study the factor structure in individual stock returns. Our empirical findings are sixfold. First, PCA is not a reliable method to estimate latent asset pricing factors and is strongly dominated by RP-PCA. We show that even for the double-sorted portfolios that follow a clear factor structure, PCA can fail to detect the underlying factor structure, while RP-PCA reliably finds all asset-pricing factors. Second, we show that a small number of five latent factors can explain the covariance and expected return structure of the 370 anomaly portfolios. In contrast a large number of ad-hoc long-short factors is needed to achieve similar pricing properties. Third, we demonstrate that using the information in the mean of the data in the estimation is essential for discovering factors that can explain the risk-premium of assets. The maximum Sharpe-ratio of five RP-PCA factors is more than twice as large as the Sharpe-ratio of five PCA factors; a result that holds in- and out-of-sample. The cross-sectional pricing errors out-of-sample are sizable smaller for RP-PCA than for PCA factors. We conclude that the RP-PCA methodology captures the pricing information better while explaining the same amount of variation and co-movement in the data compared with standard PCA. Fourth, we demonstrate that the better pricing performance results are due to the detection of weak factors with high Sharpe-ratios that are not found by PCA. Fifth, RP-PCA factors are based on only a subset of the stock characteristics implying that a significant amount of characteristic information is redundant. The RP-PCA pricing factors can be interpreted as a market factor, value/value-interaction, momentum/momentum-interaction, profitability and a "high SR" factor that loads heavily on reversal. We also show that most of the pricing information is contained in the extreme deciles. Sixth, we verify that linear factor models estimated on portfolio data are stable over time, while this is not the case for estimations using individual stock return data. This suggests that factor models with constant factor weights are inappropriate for estimating factor models for individual stocks.

Our paper contributes to an emerging literature that uses new econometric techniques in

asset pricing for high-dimensional data, including machine learning methods. Kozak et al. (2017) also exploit economic restrictions relating expected returns to the covariance to extract the factors spanning the stochastic discount factor. Their estimator is based on an elastic net that shrinks the contributions of low-variance principal components of candidate factors. This is in contrast to the RP-PCA estimator that also penalizes large pricing errors but can actually strengthen low-variance principal components. It is important to understand that in our main study our estimator is applied to a large number of sorted portfolios and not individual stocks. A collection of portfolio sorts is in general not representing the covariance structure of the individual stocks. Hence, a factor that has a high Sharpe-ratio and is necessary to span the stochastic discount factor might only explain a small amount of the variation in a particular choice of sorted portfolios and hence leads to a low-variance principal component. Our approach is able to find this factor and hence is in this sense robust to the choice of basis assets. Kelly et al. (2017) introduce Instrumented-PCA to perform dimensionality reduction of the characteristic space. Their work is closely related to the Projected-PCA of Fan et al. (2016) that allows for time-varying factor loadings. Both methods first project the individual stock returns into managed portfolios based on observed characteristics and then apply PCA to the projected data. These approaches could be combined with our methodology by applying RP-PCA in the second step to identify the factors that can best "price" the managed portfolios. Kelly et al. (2017) and Kozak et al. (2017) argue that the SDF based on the dominant principal components can explain the cross-section of expected returns well. DeMiguel et al. (2017), Freyberger et al. (2017) and Feng et al. (2017) employ factor selection with Lasso-style L^1-norm penalties. These papers have in common that they assume that the stochastic discount factor has a sparse exposure to the characteristics. In contrast, we only assume a low-dimensional factor structure without imposing any sparsity constraints on how the characteristics can affect the factors and hence the stochastic discount factor. RP-PCA estimation can also be combined with other PCA-based methods, *e.g.* the 3-pass model in Giglio and Xiu (2017).

The rest of the paper is organized as follows. In Section 2 we introduce the model and provide an intuition for our estimators. The simulations in Section 3 verify the theoretical results. The empirical results are collected in Section 4. Section 5 concludes. The appendix contains all the tables and figures.

2. Methodology

2.1. A Factor Model

We assume that excess returns follow a standard approximate factor model and the assumptions of the arbitrage pricing theory are satisfied. This means that returns of an asset

n, X_{nt}, have a systematic component captured by K factors and a nonsystematic, idiosyncratic component capturing asset-specific risk. The approximate factor structure allows the non-systematic risk to be weakly dependent. We observe the excess return of N assets over T time periods:

$$X_{nt} = \mathbf{F}_t \, \mathbf{\Lambda}_n^\top + e_{nt} \qquad n = 1, ..., N \;\; t = 1, ..., T \tag{1}$$

$$\iff \underbrace{\mathbf{X}}_{T\times N} = \underbrace{\mathbf{F}}_{T\times K} \underbrace{\mathbf{\Lambda}^\top}_{K\times N} + \underbrace{\mathbf{e}}_{T\times N} \tag{2}$$

Our goal is to estimate the unknown latent factors $\mathbf{F}$ and loadings $\mathbf{\Lambda}$. We will work in a large dimensional panel, i.e. the number of cross-sectional observations N and the number of time-series observations T are both large and we study the asymptotics as they jointly go to infinity.

Under the assumption that the factors and residuals are uncorrelated, the covariance matrix of the returns consists of a systematic and idiosyncratic part:

$$\text{Var}(\mathbf{X}) = \mathbf{\Lambda}\text{Var}(\mathbf{F}) \, \mathbf{\Lambda}^\top + \text{Var}(\mathbf{e}).$$

Since the largest eigenvalues of $\text{Var}(\mathbf{X})$ are driven by the factors, Principal Component Analysis (PCA) can be used to estimate the loadings and factors. PCA estimators of latent factors only utilize the information contained in the second moment but ignore information that is contained in the first moment.

Factor models, *e.g.* Connor and Korajczyk (1988, 1993), Bai (2003), Bai and Ng (2002) and Stock and Watson (2002), typically assume that the mean of the data matrix $\mathbf{X}$ is equal to zero. As we will see below, this assumption is restrictive if the means contain information about the factor structure, as is the case in applications of financial data. The RP-PCA estimator exploits this information, so it is crucial to allow the means to be unrestricted.

In the case of asset return data, the role of means is explicitly given by Ross' Arbitrage-Pricing Theory (APT), which implies that expected excess returns are explained by the exposure to the risk factors multiplied by the risk-premium of the factors. If the factors are excess returns, the APT implies

$$\text{E}[\mathbf{X}_n] = \mathbf{\Lambda}_n \, \text{E}[\mathbf{F}].$$

We assume a strong form of APT, where residual risk has a risk-premium of zero. In its more general form APT requires only the risk-premium of the idiosyncratic part of well-diversified portfolios to go to zero. As most of our analysis will be based on portfolios, there is no loss of generality by assuming the strong form.

Factors identified by standard PCA explain as much time-variation as possible. Conven-

tional statistical factor analysis applies PCA to the sample covariance matrix $\frac{1}{T}\mathbf{X}^\top\mathbf{X} - \bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ where $\bar{\mathbf{X}}$ denotes the sample mean of excess returns. The eigenvectors of the largest eigenvalues are proportional to the loadings $\hat{\mathbf{\Lambda}}_{\text{PCA}}$ and factors are obtained from a regression on the estimated loadings. It can be shown that conventional PCA factor estimates are based on the objective function (Stock and Watson (2002)): [2]

$$\min_{\mathbf{\Lambda},\mathbf{F}} \frac{1}{NT} \sum_{n=1}^{N}\sum_{t=1}^{T}(X_{nt} - \mathbf{F}_t\,\mathbf{\Lambda}_n^\top)^2. \tag{3}$$

Our Risk-Premium-PCA (RP-PCA) estimator modifies the objective function so that cross-sectional pricing errors are taken into account. The RP-PCA objective function minimizes jointly the unexplained variation and cross-sectional pricing errors:

$$\min_{\mathbf{\Lambda},\mathbf{F}} \underbrace{\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}(X_{nt} - \mathbf{F}_t\,\mathbf{\Lambda}_n^\top)^2}_{\text{unexplained TS variation}} + \gamma \underbrace{\frac{1}{N}\sum_{n=1}^{N}\left(\bar{\mathbf{X}}_n - \bar{\mathbf{F}}\,\mathbf{\Lambda}_n^\top\right)^2}_{\text{XS pricing error}}, \tag{4}$$

where γ is the weight of the average cross-sectional pricing error relative to the times-series error in standard PCA. It is straightforward to show that minimizing (4) is equivalent to applying PCA to the matrix

$$\frac{1}{T}\mathbf{X}^\top\mathbf{X} + \gamma\,\bar{\mathbf{X}}\bar{\mathbf{X}}^\top. \tag{5}$$

Note that (5) is equal to the variance-covariance matrix of $\mathbf{X}$ if $\gamma = -1$. Thus RP-PCA with $\gamma > -1$ can be understood as PCA applied to a matrix that "over-weights" the means. As in standard PCA, the eigenvectors of the largest eigenvalues of (5) are proportional to the loadings $\hat{\mathbf{\Lambda}}_{\text{RP}}$. RP-PCA factors are estimated by a regression of the returns on the estimated loadings, i.e. $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{\Lambda}}_{\text{RP}}\left(\hat{\mathbf{\Lambda}}_{\text{RP}}^\top\hat{\mathbf{\Lambda}}_{\text{RP}}\right)^{-1}$.

There are four different interpretations of RP-PCA:

(a) *Variation and pricing objective functions:* As outlined above the RP-PCA estimator combines the time-series variation and cross-sectional pricing error criteria function. As such it only selects factors that are priced and hence have small cross-sectional alpha's. But at the same time it protects against spurious factors that have vanishing loadings as it requires the factors to explain a large amount of the variation in the data as well.[3]

[2] Stock and Watson (2002) assume that the data matrix $\mathbf{X}$ has a mean of zero. In this case PCA on the covariance matrix of $\mathbf{X}$ is equivalent to minimizing (3). In our model, we do not restrict the mean of $\mathbf{X}$ and the equations below reflect this more general case.

[3] A natural question to ask is why do we not just use the cross-sectional objective function for estimating latent factors, if we are mainly interested in pricing? First, the cross-sectional pricing objective function alone does not identify a set of factors. For example it is a rank 1 matrix and it would not make sense to apply PCA to

(b) *Penalized PCA:* RP-PCA is a generalization of PCA regularized by a pricing error penalty term. Factors that minimize the variation criterion need to explain a large part of the variance in the data. Factors that minimize the cross-sectional pricing criterion need to have a non-vanishing risk-premia. Our joint criteria is essentially looking for the factors that explain the time-series but penalizes factors with a low Sharpe-ratio. Hence the resulting factors usually have much higher Sharpe-ratios than those based on conventional factor analysis.

(c) *Information interpretation:* Conventional PCA of a covariance matrix only uses information contained in the second moment but ignores all information in the first moment. As using all available information in general leads to more efficient estimates, there is an argument for including the first moment in the objective function. Our estimator can be seen as combining two moment conditions efficiently.

(d) *Signal-strengthening:* The matrix $\frac{1}{T}\mathbf{X}^\top\mathbf{X} + \gamma\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ should converge to[4]

$$\mathbf{\Lambda}\left(\mathbf{\Sigma}_F + (1+\gamma)\,\boldsymbol{\mu}_F\boldsymbol{\mu}_F^\top\right)\mathbf{\Lambda}^\top + \mathrm{Var}(\mathbf{e}),$$

where $\mathbf{\Sigma}_F = \mathrm{Var}(\mathbf{F})$ denotes the covariance matrix of $\mathbf{F}$ and $\boldsymbol{\mu}_F = \mathrm{E}[\mathbf{F}]$ the mean of the factors. After normalizing the loadings, the strengths of the factors in the standard PCA of a covariance matrix are equal to their variances. Larger factor variances will result in larger systematic eigenvalues and a more precise estimation of the factors. In our RP-PCA the signal of weak factors with a small variance can be "pushed up" by their mean if γ is chosen accordingly. In this sense our estimator strengthens the signal of the systematic part.

Lettau and Pelger (2018a) develop a formal statistical theory that provides guidance on the optimal choice of the key parameter γ. The intuition is as follows. There are essentially two different factor model interpretations: a strong factor model and a weak factor model. In a strong factor model the factors provide a strong signal and lead to exploding eigenvalues in the covariance matrix. This is either because the strong factors affect a large number of assets and/or because they have large variances themselves. In a weak factor model the factors' signal is weak and the resulting eigenvalues are large compared to the idiosyncratic spectrum, but they do not explode. Weak factors only affect a small subset of assets or affect all assets only weakly. This intuition is formalized below. [5]

---

it. Second, there is the problem of spurious factor detection (see *e.g.* Bryzgalova (2016)). Factors can perform well in a cross-sectional regression because their loadings are close to zero. Thus "good" asset pricing factors need to have small cross-sectional pricing errors and explain the variation in the data.

[4] In this large-dimensional context the limit is more complicated and studied in Lettau and Pelger (2018a).

[5] The Arbitrage-Pricing Theory developed by Chamberlain and Rothschild (1983) assumes that only strong

In a strong factor model, PCA and RP-PCA yield consistent estimates of factors and loadings but RP-PCA estimates are more efficient. In a weak factor model RP-PCA strengthens the signal of the weak factors, which could otherwise not be detected and usually dominates PCA estimates. Depending on which framework is appropriate, the optimal choice of γ varies. A weak factor model usually suggests much larger choices for the optimal γ than a strong factor model. However, in strong factor models our estimator is consistent for any choice of γ and choosing a γ that is "too large" results in only minor efficiency losses. On the other hand, a γ tat is "too small" can prevent weak factors from being detected at all. Thus in our empirical analysis we opt for the choice of larger γ's.

The empirical spectrum of eigenvalues in equity data suggests a combination of strong and weak factors. In all equity data sets that we have considered, the first eigenvalue of the sample covariance matrix was large, typically around ten times the size of the rest of the spectrum. The second and third eigenvalues usually stand out, but have only magnitudes around twice or three times of the average of the residual spectrum, which would be more in line with a weak factor interpretation. The first statistical factor in our data sets is always very strongly correlated with an equally-weighted market factor. Hence, if we are interested in learning more about factors besides the market, the weak factor model is likely to be appropriate.

In this paper we consider a simplified factor model with the purpose to make the weak and strong factor models comparable. The assumptions of the strong factor models can be relaxed considerably without affecting the results, as shown in Lettau and Pelger (2018a).

Assumption 1: Factor Model

The factor model (2) holds with the following assumptions:

(a) **Convergence rates:** $T, N \longrightarrow \infty$, $N/T \longrightarrow c$ *with* $0 < c < \infty$.

(b) **Factors:** *The factors* $\mathbf{F}$ *are uncorrelated, are independent of* $\mathbf{e}$ *and* $\mathbf{\Lambda}$ *and have bounded first two moments:*

$$\hat{\boldsymbol{\mu}}_F = \frac{1}{T}\sum_{t=1}^{T} \mathbf{F}_t \xrightarrow{p} \boldsymbol{\mu}_F \qquad \hat{\mathbf{\Sigma}}_F = \frac{1}{T}\mathbf{F}_t\mathbf{F}_t^\top \xrightarrow{p} \mathbf{\Sigma}_F = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_K^2 \end{pmatrix}$$

(c) **Loadings:** *The loadings* $\mathbf{\Lambda}$ *are independent of the factors and residuals.*

factors are non-diversifiable and explain the cross-section of expected returns. As pointed out by Onatski (2012) a weak factors can be regarded as a finite sample approximation for strong factors, i.e. the eigenvalues of factors that are theoretically strong grow so slowly with the sample size that the weak factor model provides a more appropriate description of the data.

(d) **Residuals:** *The matrix of residuals can be represented as* $\mathbf{e} = \boldsymbol{\epsilon}\boldsymbol{\Sigma}_e$ *with* $\epsilon_{t,i} \sim i.i.d.\ N(0,1)$.

Assumption 1 (a) imposes that time and cross-sectional dimensions grow at the same proportional rate c. Assumption 1 (b) is relatively weak and the main constraint is the independence of the factors of the loadings and residuals. The assumption that the factors are uncorrelated is a normalization and without loss of generality. Assumptions (c) and (d) will be further specified for the strong and weak factor models. Both models require a form of weak dependence of the residuals but differ in the strength of the loadings.

In latent factor models only the product $\mathbf{F}\boldsymbol{\Lambda}^\top$ is identified. For any full rank $K \times K$ matrix $\mathbf{H}$ the factors $\mathbf{F}\mathbf{H}^{-1}$ and loadings $\boldsymbol{\Lambda}\mathbf{H}^\top$ yield the same factor model. We use the standard convention to normalize the loadings $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}/N = \mathbf{I}_K$ and assume that the factors are uncorrelated. That means that a factor with a variance of $\sigma_F^2 = 0.5$ could be interpreted as affecting only half of the assets with an average loadings of 1. Alternatively, we could normalize the covariance matrix of the factors to an identity matrix, so that the norm of the loading vectors measure the strength of the factors.

2.2. Strong Factor Model

Traditional factor models as in Connor and Korajczyk (1988, 1993), Bai (2003), Bai and Ng (2002) and Stock and Watson (2002) are based on the assumption that factors are "strong" in the following sense. The "strengths" of systematic factors are given by their corresponding eigenvalues of the matrix $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$. The eigenvalues of strong factors diverge to infinity (while in the weak factor models considered below the eigenvalues are bounded). This is captured by the assumption that $\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda} \to \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}$ where $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}$ is a full-rank matrix. Without loss of generality we will use the normalization $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}} = \mathbf{I}_K$. The interpretation of the assumption $\frac{1}{N}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda} \to \mathbf{I}_K$ is that strong factors affect an infinite number of assets with non-vanishing loadings. One example of this factor structure is $\Lambda_{nk} = N(1, 1/\sqrt{N})\ \forall n,k$, i.e. the loading of each asset on each factor converge to 1 as N grows. The results below show that in a strong factor model RP-PCA and PCA provide consistent estimators for the loadings and factors but the RP-PCA estimator of the loadings is more efficient than the PCA estimator. Furthermore, the assumptions required for the strong factor model are weaker than those imposed in the weak factor model below.

Assumption 2: Simplified Strong Factor Model

Assume Assumption 1 holds and in addition:

(a) **Loadings:** $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}/N \xrightarrow{p} \mathbf{I}_K$ *and all loadings are bounded.*

(b) **Residuals:** *All elements and all row sums of* $\boldsymbol{\Sigma}_e$ *are bounded.*

The following result is proven in Lettau and Pelger (2018a):

Proposition 1: Simplified Strong Factor Model

Suppose that Assumption 2 holds. The RP-PCA estimator that minimizes the objective function (4) has the following properties:

(a) The factors and loadings estimators are consistent.

(b) The asymptotic distribution of the factor estimates is independent of γ.

(c) The asymptotic distribution of the estimates of the loadings of factor k is given by

$$\sqrt{T}\left(\mathbf{H}^{\top}\hat{\mathbf{\Lambda}}_k - \mathbf{\Lambda}_k\right) \longrightarrow N(0, \mathbf{\Omega}_k).$$

Closed-form expressions of $\mathbf{\Omega}_k$ and the rotation matrix $\mathbf{H}$ are given in Lettau and Pelger (2018a).

(d) The optimal γ that minimizes the asymptotic variance is $\gamma = 0$. Choosing $\gamma = -1$, i.e. PCA with covariance matrix for factor estimation, is not efficient.

Simulations results in section 3 and Lettau and Pelger (2018a) show that efficiency gains of choosing $\gamma = 0$ are relatively minor. We conclude that strong factors are relatively easy to estimate by PCA-based methods. We now turn to the more interesting case when factors "weak".

2.3. Weak Factor Model

In contrast to strong factors, weak factors affect only a smaller fraction of the assets (or have a small variance). Formally, in a weak factor model, the matrix $\mathbf{\Lambda}^{\top}\mathbf{\Lambda}$ is bounded and converges to a fixed limit, in contrast to a strong factor model where $\frac{1}{N}\mathbf{\Lambda}^{\top}\mathbf{\Lambda}$ converges. Without loss of generality, we use the normalization $\mathbf{\Lambda}^{\top}\mathbf{\Lambda} \xrightarrow{p} \mathbf{I}_K$. Lettau and Pelger (2018a) give a general characterization of weak factors. In this paper, we focus on an easily interpretable class of weak factors of the form $\Lambda_{nk} = N(0, 1/\sqrt{m})$ with probability m/N, $\Lambda_{nk} = 0$ with probability $1 - m/N$. For $m = N$, the factor k affect all assets but the loadings converge to zero. If m is small, fewer assets are affected by the factors (in expectation) but with larger weights. A weak factor can thus either be interpreted as a factor with only a weak effect on many assets or a factor with strong effects on a few assets. Moreover for $m > 1$, factors of this form have the interpretation of long-short portfolios since the mean of all Λ_{nk} is zero, so that in expectation half the loadings are positive and half are negative.

The statistical model for analyzing weak factor models is considerably more complicated than that of the strong factor model and is based on spiked covariance models from random

matrix theory (see Lettau and Pelger (2018a) for details). [6] The distinction of strong and weak factors matters in practice since the estimation of weak factors is more difficult than the estimation of strong factors. Recall that strong factors and their loadings can be estimated consistently. In contrast, PCA-based estimators of weak factors are generally biased, even in the limit as T and N grow. For example, the correlation of an estimated weak factor with the "true" weak factor typically converges to a number less than one (see Proposition 2 below). Lettau and Pelger (2018a) derive a set of results that show that RP-PCA dominates standard PCA in a number of dimension. The optimal choice of γ is generally larger than -1 and in most cases larger than 0. We present some simulation evidence in section 3 and refer to Lettau and Pelger (2018a) for more results. The bias reduction of choosing $\gamma > -1$ is often substantial.

Moreover, in the weak factor model RP-PCA can detect factors which are missed by conventional PCA. If a weak factor affects "too few" assets, if the loadings are "too small" or if the variance of the factor is "too low", it cannot be detected by PCA. However, the signal of RP-PCA depends on the mean and the variance of the factors. Thus, RP-PCA can detect weak factors with a high Sharpe-ratio even if the factor is below the critical detection value for PCA. The results in Lettau and Pelger (2018a) are summarized in the following proposition.

Assumption 3: Weak Factor Model

Assume Assumption 1 holds and in addition:

(a) **Loadings:** $\mathbf{\Lambda}^\top \mathbf{\Lambda} \xrightarrow{p} \mathbf{I}_K$ *and the column vectors of the loadings* $\mathbf{\Lambda}$ *are orthogonally invariant, e.g.* $\Lambda_{1k} = N(0, 1/\sqrt{m})$ *with prob.* m/N, $\Lambda_{nk} = 0$ *with prob.* $1 - m/N$.[7]

(b) **Residuals:** *The empirical eigenvalue distribution function of* $\mathbf{\Sigma}_e$ *converges almost surely weakly to a non-random spectral distribution function with compact support. The supremum of the support is* b *and the largest eigenvalues of* $\mathbf{\Sigma}_e$ *converge to* b.

[6] The intuition is as follows. Under the assumptions of random matrix theory the eigenvalues of a sample covariance matrix separate into two areas: (1) the bulk spectrum with the majority of the eigenvalues that are clustered together and (2) some spiked large eigenvalues separated from the bulk. Under appropriate assumptions the bulk spectrum converges to the generalized Marchenko-Pastur distribution. The largest eigenvalues are estimated with a bias which is characterized by the Stieltjes transform of the generalized Marchenko-Pastur distribution. If the largest population eigenvalues are below some critical threshold, a phase transition phenomena occurs. The estimated eigenvalues will vanish in the bulk spectrum and the corresponding estimated eigenvectors will be orthogonal to the population eigenvectors. Onatski (2012) studies weak factor models and shows the phase transition phenomena for weak factors estimated with PCA. Our paper provides a solution to this factor detection problem. It is important to notice that essentially all models in random matrix theory work with processes with mean zero. However, this assumption is not appropriate for factor models of asset returns. Lettau and Pelger (2018a) develop new methods when this assumption is relaxed.

[7] The normality assumption is not essential and can be relaxed to allow for wider class of distributions (*e.g.* all spherical distributions).

As mentioned above, factors of the form considered in Assumption 3 (a) have the interpretation of long-short portfolios that either affect many assets weakly, or few assets more strongly. Assumption 3 (b) is a standard assumption in random matrix theory. The assumption allows for non-trivial weak cross-sectional correlation in the residuals, but excludes serial-correlation.

The following proposition conveys the intuition of the results that are formally stated and proven in Lettau and Pelger (2018a).

Proposition 2: Risk-Premium PCA under weak factor model

Suppose that Assumption 3 holds.

Let $\theta_1, \dots, \theta_K$ be the first K largest eigenvalues of an appropriately defined "signal matrix" given in Lettau and Pelger (2018a). The eigenvalues $\theta_k = \theta_k(\gamma)$ are strictly increasing functions of γ and have the interpretation of "signal strengths" of individual factors k.

Let $\hat{F}_k$ denote the RP-PCA estimator of factor F_k. Then the correlation of $\hat{F}_k$ with the true factor F_k converges to

$$Corr(\hat{F}_k, F_k) \xrightarrow{p} \begin{cases} \tau_k(\gamma) > 0 & \text{if } \theta_k(\gamma) > \theta_{crit} \\ 0 & \text{otherwise.} \end{cases}$$

θ_{crit} is a threshold that does not depend on γ or the factors but only on the noise distribution.

$\tau_k(\gamma)$ has the following properties:

(a) For generic cases $\tau_k(\gamma) < 1$.

(b) Typically, $\tau_k(\gamma)$ has a unique finite maximum $\gamma^* = \text{argmax}\, \tau_k(\gamma)$ that yields the highest correlation of the estimated factor with the true factor.

(c) If $\boldsymbol{\mu}_F \neq 0$, then $\gamma^* > -1$ and RP-PCA yields factor estimates that are more highly correlated with the true factors than PCA with $\gamma = -1$. If $\boldsymbol{\mu}_F = 0$ then $\gamma^* = -1$.

Proposition 2 states that weak factors can only be estimated with a bias even as N and K grow. If a factor is too weak, *i.e.* its signal strength $\theta_k(\gamma)$ is below the threshold θ_{crit}, then it cannot be detected at all. Since $\theta_k(\gamma)$ is increasing in γ, RP-PCA is more likely to detect factors with lower signal strength than PCA with $\gamma = -1$.

γ also affects the correlations of estimated factors with true factors. Typically, there is a finite γ that maximizes the correlation of detected factors with true factors. In particular, RP-PCA with $\gamma > -1$ yields estimated factors that are more highly correlated with true factors

than PCA with $\gamma = -1$.[8] The factor means $\boldsymbol{\mu}_F$ play a crucial role for this result. If the factors have means of zero, PCA with $\gamma = -1$ is optimal. If $\boldsymbol{\mu}_F \neq 0$, then RP-PCA with $\gamma > -1$ is optimal. If $\boldsymbol{\mu}_F = 0$, as in traditional factor models, then PCA with $\gamma = -1$ is optimal.This distinction is important since asset pricing factors have non-zero means.

Figure 1 plots the function $\tau_k(\gamma)$ for parameters used in the simulation exercise described below.[9] The variance of the factor in the top panel is set to 0.05 and to 0.1 in the bottom panel. *Ceteris paribus*, a higher factor variance increases the signal strength of the factor. We also consider three positive values for the Sharpe-ratio (or, equivalently, the mean) of the factor, as well as a case with mean-0 factor. The vertical lines indicate the γ that maximizes $\tau_k(\gamma)$[10]. In all cases with positive Sharpe-ratios, $\tau_k(\gamma)$ increases sharply with γ for small γ and then flattens out as γ increases. For fixed factor variance, the gains in the correlation of the estimated and true factors increase with the Sharpe-ratio (and mean) of the factor. Comparing the two panel shows that the optimal γ is higher if the signal strength of the factor is lower (*i.e.* the factor variance decreases) and if the Sharpe-ratio increases. If the Sharpe-ratio (or the mean) of a factor is zero, then $\tau_k(\gamma)$ is decreasing in γ and the optimal weight $\gamma = -1$, as stated in Proposition 2. Note that in this case $\tau_k(\gamma) = 0$ if γ is large enough. The reason is that the factor mean does not contain any useful information but more and more weight is put on the uninformative cross-sectional errors. Hence, in this case RP-PCA performs strictly worse than PCA.

How should the optimal γ be chosen? Based on our asymptotic theory, there are two possible criteria. First, a very large γ maximizes the probably of detecting a weak factor but lowers the correlation with the true factor. Alternatively, a moderately high γ maximizes this correlation. Another aspect of choosing γ is the out-of-sample performance of the RP-PCA estimator. As we will show in the next section, RP-PCA estimation deteriorates out-of-sample if γ is chosen too high.

2.4. Number of Factors

We use different diagnostic criteria to determine the number of factors. The basic idea for estimating the number of factors is that the systematic eigenvalues of the covariance matrix are "large" while the idiosyncratic eigenvalues are "small". Traditional methods designed for strong factor models, such as the information criterion in Bai and Ng (2002) and the eigenvalue ratio estimator of Ahn and Horenstein (2010), are exploit the fact that strong

[8] Intuitively, the phase transition phenomena that hides weak factors can be avoided by putting some weight on the information captured by the risk-premium.

[9] We have estimated the sparse residual correlation matrix with a thresholding approach based on $N = 370$ deciles sorted portfolios as described in the empirical Section 4. The other parameters are the same as in the simulation section.

[10] We calculate the smallest value of γ such that $\tau_k(\gamma)$ is less than 0.5% from its maximum value as $\tau_k(\gamma)$ can be essentially flat for large values of γ and there is practically no gain in increasing γ.

factors lead to exploding eigenvalues while the idiosyncratic eigenvalues are bounded. Estimating the number of factor in weak factor models is more complicated since eigenvalues of weak factors are bounded just as the idiosyncratic eigenvalues. Onatski (2010) shows that the pattern of successive eigenvalues can be used to determine the number of systematic factors and proposes an eigenvalue difference estimator. He argues that idiosyncratic eigenvalues cluster around a single point, *i.e.* successive eigenvalues are similar in magnitudes. In contrast, successive systematic eigenvalues are different in magnitude. Hence, the number of weak factors can be estimated by determining when, starting from the smallest eigenvalues, the clustering in the residual eigenvalues stops. However, only weak factors that are detectable, i.e. whose eigenvalues separate from the idiosyncratic spectrum, can be identified. When the signal strength of weak factors is too low, *i.e.* below the detection threshold in Proposition 2, their eigenvalues can not be separated form the eigenvalues of idiosyncratic factors. However, if we apply the Onatski (2010) argument to the eigenvalues of the RP-PCA covariance matrix with overweighted mean, we can detect these weak factors.[11]

3. Simulation

In this section we use a Monte Carlo simulation to study the behavior of the RP-PCA estimator in samples of the size typically encountered in practice. We focus on the effects of γ, the signal strength and excess returns of factors (Lettau and Pelger (2018a) conduct a more comprehensive simulation study).

We consider a two factor model, where the first factor mimics a strong "market" factor that affects all assets with an average loading of one. The second factor is a weak long-short factor with an average loading of zero: $\Lambda_{n,1} = 1, \Lambda_{n,2} \sim N(0,1)$. The mean and standard deviation of the market factor are $\mu_1 = 0.5\%, \sigma_1 = 4.5\%$ and the Sharpe ratio is 0.11. The signal strength of the weak second factors depends on its variance σ_2^2 and Sharpe ratio. We consider a variety of different values for σ_F^2 and SR_F to assess the RP-PCA estimation of weak factors.

The performance of RP-PCA is evaluated by the following criteria: (*i*) the correlation between the estimated and "true" factors, $\rho(\hat{F}_k, F_k)$, (*ii*) the Sharpe-ratios of the estimated factors, $\widehat{\text{SR}}_k$, and (*iii*) the root-mean-squared pricing errors of time series regressions of returns on estimated factors, $\text{RMSE} = \sqrt{1/N \sum_n \alpha_n^2}$. We normalize the RMSE of the model with estimated factors by the RMSE for the true factors. To conserve space we only report out-of-sample results where we first estimate the loading vector in-sample and then obtain

[11] The mean operator with the weight γ can lead to an additional "spurious" spike in the eigenvalues of $\frac{1}{T}X^\top X + \gamma \bar{X}\bar{X}^\top$. Hence, if we observe that the first q eigenvalues do not cluster, we must have at least $q-1$ strong or weak factors, but it is theoretically possible that the qth spike is spurious. For realistic values in our simulations this spurious phenomena did not occur.

the out-of-sample factor estimates by projecting the out-of-sample returns on the estimated loadings. The sample size of $N = 74$ and $T = 650$ observations is based on our empirical study below. Finally, the correlation matrix of our simulated residuals is set to the empirical correlation that we observe in the data. [12]

We first consider estimation of the strong first factor. Figure 2 shows the evaluation criteria as function of γ. Panel A shows that the correlation of the estimated factor and the true factor $\rho(\widehat{F}_1, F_1)$ is close to one indicating that unobservable strong factor is estimated with high precision for all values of γ. Panel B shows that the estimated Sharpe ratio $\widehat{\text{SR}}_1$ is close to the true SR of the strong factor (0.11) for all values of γ. This confirms the theoretical results in Proposition 1 that stated that strong factors can be estimated consistently and that γ only affects the efficiency of the estimator. We conclude that strong factors are relatively easy to estimate by PCA-based methods. We next consider the more interesting case of weak factors.

Figure 3 has the same layout as Figure 2 but considers a range of values for the variance and Sharpe ratio of the weak factor: $\sigma_2^2 = 0.05, 0.1, 0.5$ and $\text{SR}_2 = 0.1, 0.3, 0.5, 0.6$. The panel columns have different values of σ_2^2 and each panel has four lines corresponding to four different values of SR_2. Recall that the signal strength of a factor increases in σ_2^2 and SR_2, so a factor with a small variance but a high Sharpe ratio is detectable by RP-PCA as long as γ is sufficiently high. Consider a factor with a low variance $\sigma_2^2 = 0.05$ (left column in Figure 3). The top panel shows that the correlation $\rho(F_2, \widehat{F}_2)$ is essentially zero if the Sharpe ratio is also low ($\text{SR}_2 = 0.1$). This is an example of an undetectable factor that is below the signal threshold in Proposition 2. As the Sharpe ratio of the weak factor increases, the factor becomes detectable as long as γ is set sufficiently high. The intuition is that a non-zero Sharpe ratio contains information that can be exploiting by putting more weight on the mean, i.e. increasing γ. Hence a factor that is missed by standard PCA with $\gamma = -1$ can be detected by RP-PCA. However, $\rho(F_2, \widehat{F}_2)$ is below one in all cases showing that in contrast to strong factors, weak factors are not fully identifiable as shown theoretically in Proposition 2.

The panel below shows that the estimated Sharpe ratio $\widehat{\text{SR}}_2$ of the second factor increases with γ. This is intuitive since a higher γ allows for a better identification of the weak factor.

[12] In more detail, we have estimated the residual correlation matrix based on $N = 74$ extreme deciles sorted portfolios as described in the empirical Section 4. In each case we have first regressed out the systematic factors and then estimated the residual covariance matrix with a hard thresholding approach setting small values to zero, see Bickel and Levina (2008) and Fan, Liao and Mincheva (2013). This provides a consistent estimator of the residual population covariance matrix. We have regressed out the first 6 PCA factors. Our results remain unchanged when we calculate residuals based on more PCA factors or using RP-PCA factors. The additional results are available upon request. The remaining correlation structure in the residuals is sparse. In particular the estimated eigenvalues of the simulated residuals coincide with the empirical estimates of the eigenvalues. The average idiosyncratic noise level is estimated from the data and set to $\sigma_e^2 = 4$.

Finally, the bottom panel shows that the RMSE time-series error decreases slightly with γ if SR $= 0.6$. If the Sharpe ratio is lower, coupled with a low variance of the second factor relative to the variance of the first factor (0.05 vs. 0.21), the true model is dominated by the market factor and identifying the second factor decreases the RMSE only marginally.

The middle column of Figure 3 considers a larger σ_2^2 of 0.1. The patterns are similar as for $\sigma_2^2 = 0.05$ but the effect of a higher γ is more pronounced. The effect of a higher γ is especially pronounced if the Sharpe ratio of the weak factor is high. For $\text{SR}_2 = 0.6$, the improvements along all three evaluation criteria are substantial.

Since the signal strength of the second factor is higher, $\rho(F_2, \hat{F}_2)$ and $\widehat{\text{SR}}_1$ are larger than in the left column as long as γ is sufficiently high. The effect of a higher γ on the normalized RMSE is larger than in the left panel with smaller factor variance. Note that standard PCA misses to detect this factor completely and thus RP-PCA strictly dominates PCA for $\gamma \gtrsim 5$.

In the right columns σ_2^2 is increased further to 0.5. Now, the signal strength is sufficiently high that the second factor is better characterized as strong rather than weak and thus can be estimated regardless of γ. This example shows that RP-PCA is particularly powerful for factors in the intermediate range between very weak factors that are undetectable and factors that are strong enough to be detectable by standard PCA. We will argue in the empirical section below that in the data the market index is a strong factor while any additional factors are better characterized as weak, so that they might missed by standard PCA but not too weak, so that they can be estimated by RP-PCA.

In Figure 4, we plot sample paths of one representative simulation run. Each panel shows the true factor as well as the estimated RP-PCA factors for a range of γ's. The estimated sample paths of the strong factor in the first panel are close to the sample path of the true factor, which is not surprising since the correlations of the estimated factors with the true factor is essentially equal to one, as discussed above. The pattern is different for the second (weak) factor in the second panel since the sample paths of the factor estimates diverge from the path of the true factor. As predicted by Proposition 2, the paths are closer to the path of the true factor for higher γ. In particular, the estimated PCA factor with $\gamma = -1$ has no resemblance to the path of the true factor and thus the weak factor is unidentified by PCA.

4. Empirical Results

We proceed with the empirical application in two steps. First, we consider two cases with 25 double-sorted portfolios. The RP-PCA is designed to handle much larger cross-sections but the cases with smaller N are helpful in demonstrating the methodology. Next, we consider the larger cross-section of anomaly portfolios used in Kozak et al. (2017) that is based on single-sorts of 37 different characteristics. [13] The sample span is July 1963 to December

[13] We thank the authors for sharing their data.

2017 in all cases.

As mentioned in section 2, only the product of factors and loadings, $\mathbf{F}\mathbf{\Lambda}^\top$, is identified. For the theoretical derivations in section 2, it was convenient to normalize the loadings in terms of $\mathbf{\Lambda}^\top\mathbf{\Lambda}$. When presenting empirical results, it turns out to be more useful to normalize the variance-covariance matrix of the factors instead, hence we set $\mathbf{\Sigma}_F = \mathbf{I}_K$, unless otherwise noted. The reason is that we want to compare different factor model and their composition in terms of the original portfolios. To do so, the "units" of estimated loadings must be comparable across models. Normalizing $\mathbf{\Sigma}_F = \mathbf{I}_K$ achieves this objective. For some results, normalizing the loadings, i.e. $\mathbf{\Lambda}^\top\mathbf{\Lambda} = \mathbf{I}_k$, is more informative. Of course, the overall model is not affected by any normalization.

4.1. Double-Sorted Portfolio, $N = 25$

We start with two well-studied data sets of 25 double-sorted portfolio: Size/accruals and size/short-term reversal. We compare RP-PCA and PCA with three factors ($K = 3$). We also consider a three-factor model with ad hoc long-short portfolios where the factors are long in the equally weighted top quintiles and short in the bottom quintiles. We evaluate models by the same criteria as in section 3: The Sharpe ratio, cross-sectional pricing errors and unexplained time-series variation.[14]

Figure 5 plots the Sharpe ratio in the top panel, the RMS time series pricing errors α_n, and the unexplained time series variation (as a fraction of total variation) as a function of the RP-PCA weight γ for the size/accruals portfolios. The left panels are in-sample and the right panels are out-of-sample statistics. To illustrate the effect of each factor, we consider models with one, two and three factors. The figure shows that γ has little effect on the Sharpe ratios, α's and unexplained variances in models with one and two factors. Hence RP-PCA and PCA are essentially equivalent, which suggests that the first two factors are "strong". This is not the case when a third factor is added (solid green lines). As long as γ is less than 5, adding a third factor does not change the in-sample SR and α materially. For γ larger than 5, the SR increases and the RMS α decreases. For $\gamma > 15$, the in-sample SR is about twice as high

[14] We calculate the maximum Sharpe ratio that can be obtained by a linear combination of the factors, i.e. it combines the factors with the weights $\Sigma_F^{-1}\mu_F$. This is the Sharpe ratio of the stochastic discount factor implied by the factors. The root-mean-squared pricing error ($RMS\alpha$) equals $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\alpha_i^2}$, where the pricing error α_i is the intercept of a time-series regression of the excess return of asset i on the factors. The idiosyncratic variation is the average variance of the residuals after regressing out the factors. The in-sample analysis is based on the whole time horizon of $T = 650$ months. The out-of-sample analysis estimates the loadings with a rolling window of 20 years ($T = 240$). With these estimated loadings including information up to time t we predict the systematic return and obtain a pricing error out-of-sample at $t + 1$. This corresponds to a cross-sectional pricing regression with out-of-sample loadings. The mean and variance of the out-of-sample errors are used to calculate the average pricing error and the idiosyncratic variation. We use the optimal portfolio weights for the maximum Sharpe-ratio portfolio estimated in the rolling window period to create an out-of-sample optimal return giving us the maximum Sharpe-ratio portfolio out-of-sample.

as for $\gamma = -1$ and the RMS α is cut in half. Note that the increase in residual variance rises only very little compared relative to the PCA case with $\gamma = -1$. The out-of-sample statistics are similar but the effect of higher γ starts at smaller values.

What is the cause for the difference between PCA and RP-PCA when a third factor is added? Figure 6 shows a heat map of the loadings of the three factors for PCA with $\gamma = -1$ and RP-PCA with $\gamma = 10$. Positive loadings are in green and negative loadings in red. The figure shows that the loading of the first two factors are similar for PCA and RP-PCA. The first factor is a "long" factor with positive weights on all portfolios with a tilt towards small-stock portfolios. The second factor is a long-short factor with positive weights on small-stock portfolios and short in large-stock portfolios. These factors are similar to the market and SMB factors in Fama-French models. The composition of the third factor is different for PCA and RP-PCA. The PCA factor has no clear pattern and adds very little information, as shown in Figure 5. In contrast, the third RP-PCA factor has positive loadings on low-accrual stocks and negative loadings on high-accrual stocks, similar to a Fama-French-type factor. This pattern explains the higher Sharpe ratio and lower α's in the three factor RP-PCA model shown in Figure 5.

The top panel of Table 1 reports in-sample and out-of-sample Sharpe ratios, RMS α's and unexplained variance of three factor PCA, RP-PCA with $\gamma = 10$ and the long/short Fama-French-style model with ad-hoc weights. In-sample, the ad hoc long/short model performs almost as well as the RP-PCA model but the OOS SR of the RP-PCA is almost twice as high as the OOS SR of the ad hoc model. The SR of the PCA model is lower and its RMS α is higher than RP-PCA both in-sample and out-of-sample. Note that the unexplained variation of RP-PCA is only slightly higher than that of PCA.

The evidence suggests that the underlying factor model for 25 size/accruals portfolio consists of two "strong" factors that are easily identifiable and one "weak" factor that is missed by PCA but can be identified by RP-PCA.

Consider next the results for the 25 size/short-term reversal portfolios shown in Figure 7. The SR and RMS α for models with one factors are almost unaffected by γ and change only moderately for the the three factor model. In contrast, the statistics of two-factor models change substantially for different γ. The IS-SR rises and IS-RMSE α decreases once γ exceeds 18 and settle down at $\gamma \approx 20$. Figure 8 compares the loadings for the three factors for PCA and RP-PCA with $\gamma = 20$. As in the case of the size/accruals portfolios, the loadings of the first factor is very similar for both models. The second PCA factor has positive loadings of small-stock portfolios and negative ones of large-stock portfolios, it is hence a "small-minus-big" portfolios. The third PCA factor is long in low-reversal portfolios and short in high-reversal stocks and is therefore a "long-minus-short" reversal factor. These two factors are reversed in RP-PCA. The second RP-PCA factor is a "long-minus-short" reversal factor

while the third factor is "small-minus-big" factor. The reason for this switch is that the return spread is much larger along the reversal dimension than along the size dimension since RP-PCA gives high Sharpe-ratio factors additional weight. A "small-minus-big" factor contributes more towards the common time-series variation and is this favored by PCA, whereas the "long-minus-short" reversal factor contributes more to the cross-sectional dimension and is thus favored by RP-PCA.

4.2. Large Cross-section of Single-Sorted Portfolio

Next, we apply the RP-PCA estimator to a larger cross-section of portfolios. We select 37 characteristics from the data set used in Kozak et al. (2017) that are available as of July 1963. Table 3 shows the mean return, the standard deviation, the SR of the long-short portfolios as well as the mean return of the extreme portfolios. We consider two cases: (*i*) The two extreme 1 and 10 portfolios of each characteristic sort, so that $N = 74$, (*ii*) all deciles with $N = 370$.

It will be useful to group related anomalies into categories defined as follows:

value: *value*, *valuem*, *divp*, *ep*, *cfp*, *sp*

value interactions: *valmom*, *valmomprof*, *valprof*

momentum: *mom*, *mom12*, *indmom*

reversal: *lrrev*, *strev*, *momrev*

industry reversal: *indmomrev*, *indrrev*, *indrrevlv*;

growth: *inv*, *invcap*, *igrowth*, *growth*, *sgrowth*

profitability: *prof*, *roaa*, *roea*, *noa*, *gmargins*, *aturnover*

other: All other portfolios are in categories of their own

Number of factors

We start the factor analysis by investigating how many factors are "systematic" and how many are "idiosyncratic". We use two approaches to determine the number of factors: First, we apply a diagnostic criterion to the eigenvalues as described in section 2.4. Second, we study the out-of-sample performance based on the number of factors.

Figure 9 plots the differences of consecutive eigenvalues of $\frac{1}{T}\mathbf{X}^\top\mathbf{X} + \gamma\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ for different values of γ for the sample with only deciles 1 and 10 in Panel A and sample with all deciles in Panel B. The red line indicates the critical values of the Onatski-criterion. For $\gamma < 10$, the fifth eigenvalue difference is below the critical value in both samples indicating that the first four factors are systematic. However, for $\gamma \geq 10$, the fifth eigenvalue difference is above the critical value indicating that there are five systematic factors. This suggests that the fifth

factor is weak and below the detection threshold for PCA and RP-PCA with $\gamma < 10$. Once γ is chosen high enough, RP-PCA is able to detect this factor.

Our second diagnostic criterion concerns the out-of-sample performance of the estimator. If a spurious factor overfits the noise in the data, we expect it to have a poor out-of-sample performance. As we will see below, when adding incrementally factors and determining the maximum Sharpe-ratio, pricing errors and amount of unexplained variation, we observe a significant change when including five factors in the RP-PCA estimation in-sample as well as out-of-sample. A fifth factor does add much to the explanatory power of the RP-PCA model confirming the evidence based on the Onatski criterion.

Estimation Results: RP-PCA vs. PCA

Figure 10 compares the SR, RMS α and unexplained variance in-sample and out-of sample for both samples ($N = 74$ in the left column and $N = 370$ in the right). For each model we start with one factor and successively add factors (up to a total of six factors). The RP-PCA weight γ is set to 10. The top two panels show the Sharpe-ratios of the stochastic discount factors that are implied by the estimated factors. Note that the SRs for the case with only the extreme deciles ($N = 74$) are very similar to those of the case with all deciles ($N = 370$) suggesting that estimated factors for the full model are mostly comprised of the extreme 1 and 10 decile portfolios. We will study the portfolio weights of the estimated factors in more detail below. As expected, adding factors increases the in-sample Sharpe ratios of the SDFs for all models but the magnitudes of the incremental results differ. First, the overall SRs are significantly higher for RP-PCA than those for PCA. The IS-SR for models with five factors for RP-PCA is 0.57, about twice as high as the IS-SR for PCA. The OOS-SRs exhibit a similar pattern but are somewhat lower than the IS-SRs. Second, higher order factors add substantially to the SRs, especially for RP-PCA. Consider the case with $N = 74$. The SR of the 1-factor RP-PCA model is 0.12; adding factors two and three raises the SR to about 0.4. The fourth factor has little effect on the SR but the fifth factor raises the SR by 0.2 to around 0.6. Adding a sixth factor has a minimal effect. In contrast, the incremental effects of additional factors is much smaller in the PCA model. Note that the OOS-SRs in the right panel are only slightly lower than the IS-SRs suggesting that the RP-PCA model does not suffer from excessive overfitting.

The middle row shows the root-mean-squared cross-sectional pricing errors. First, note that adding incremental factors in all models reduces pricing errors not only in-sample but also out-of-sample, again confirming that the factor structure is stable and the models are not overfitting. Second, RP-PCA pricing errors are smaller than those for PCA in all cases but the incremental effects of the factors are different. In the sample with only the one and ten deciles, the second and fifth RP-PCA factors have a particularly large incremental effect on the pricing errors. These are the same factors that also have a large incremental effect

on the Sharpe-ratios shown in the top panel suggesting that these are important for pricing cross-sectional returns. Since the corresponding PCA factors have much smaller effects on the Sharpe-ratio, they also do not improve the cross-sectional fit of the PCA model by much. Finally, the RMS pricing errors are smaller for the large sample with all portfolios than for the sample with only the 1-10 deciles. This is because the extreme 1 and 10 portfolios are the most mispriced while the pricing errors of the middle portfolios are smaller resulting in a smaller average error. If we plotted only the pricing errors of 1 and 10 portfolios in the large sample, the patterns would look as in the case with $N = 74$.

The plots for the unexplained idiosyncratic variations in the bottom row are almost identical for comparable RP-PCA and PCA models. Since PCA factors minimize unexplained variation, RP-PCA sacrifices some power along this dimension but the plots show that the effect is minute.

Panel A of Table 4 compares the Sharpe-ratios, pricing errors and idiosyncratic variance for RP-PCA and PCA models with three and five factors. The in- and out-of-sample RP-PCA Sharpe ratios are up to twice as high than those for PCA in three and five factor models and for both samples. RP-PCA also implies smaller pricing errors in all cases while idiosyncratic variances for RP-PC is only slightly larger than those for PCA. Properties for individual factors are shown in Table 5. To make factors comparable, we normalize all factor loadings to a norm of one. Not surprisingly, the variance of the first factor is a magnitude larger than the variances of the other factors but for RP-PCA its Sharpe-ratio is smaller than for factors with smaller variances. Moreover, the Sharpe-ratios of RP-PCA factors are significantly larger than the corresponding PCA factors (with the exception of the fourth factor for $N = 74$). Note the Sharpe-ratio of the fifth RP-PCA factor is almost twice as large as the Sharpe-ratios of the first four RP-PCA factors. Recall that the fifth factor is significant for RP-PCA but not for PCA.

In summary, RP-PCA outperforms PCA in terms of pricing errors and extracting factors with higher Sharpe ration and sacrifices little in terms of time series variation. Next, we analyze these results in more detail.

Figure 11 shows the mean-squared cross-sectional pricing errors (α's) for each anomaly based on models with six factors. We include a fifth factor for PCA to enable a comparison with 5-factor RP-PCA model. The anomalies are ranked by the Sharpe-ratio of the decile-10 return minus the return of decile-1 ('10-1'). The dark blue bars are RP-PCA α's and the light blue bars are those for PCA. In almost all cases the RP-PCA pricing errors are smaller than those of PCA. Only the momentum (*mom*) portfolios α's are consistently larger for RP-PCA than for PCA. RP-PCA markedly outperforms PCA for a number of portfolios, including *ivol*, *valuem*, *mom12*, and *indrrev*. The figure also shows that the anomalies with the highest Sharpe-ratio are the ones with the largest pricing errors. RP-PCA reduces these pricing errors

relative to PCA, especially in the small sample with 74 portfolios.

Next, we study the composition of the estimated SDF and individual factors. Each model implies an SDF that is a linear combination of the portfolios in the sample. Figure 12 shows the compositions of the RP-PCA and PCA SDF loadings for the case with the first and 10th deciles ($N = 74$). The anomalies are ranked by Sharpe-ratio. The dark bars represent the portfolio weights of the decile-10 portfolios and the light bars show the the decile-1 weights. First, note that the high-returns decile-10 weights are mostly of the opposite sign of the weights of the low-return decile-1 portfolios for RP-PCA as well as PCA. Hence, the SDF is composed of long-short anomaly portfolios but the weights do not necessarily sum to zero, as is the case for Fama-French style factors. Next, note that the largest loadings (in absolute value) of the RP-PCA SDF are related to anomalies with the highest Sharpe-ratio (*e.g. indrrev* and *indmomrev*) while the magnitudes of the PCA weight are not linked to Sharpe-ratios. This is due to the fact that RP-PCA puts more weight on the cross-sectional pricing errors than PCA and confirms the validity of the simulations results when RP-PCA is applied to actual data.

Factor composition

The SDF of each model is linear combination of the five individual factors; we study composition of individual factors next. Figure 13a shows heatmaps of the RP-PCA and PCA portfolio weights in each factors for the sample with only the 1 and 10 deciles. The color range is dark blue (largest) to dark brown (most negative) while white corresponds to a loading of zero. Each panel shows the RP-PCA and PCA decile-10 weights for a given factor in the top two rows and their decile-1 loadings in rows three and four. The factors are normalized to have a variance of one in order to make the loadings comparable across different factors. The portfolios on the x-axis in the first four factors are sorted by categories while the heatmap for the fifth factor is sorted by SR.

Typically, the first factor in factor models of asset returns is a "level" factor that is "long" in all portfolios. The top panel of Figure 13a shows that this is the case here as well. The loadings range from 2.9 to 7.6 with an average of 4.99 for RP-PCA and 4.97 for PCA. Moreover, the weights of first RP-PCA and PCA factors are almost identical and the factors have a correlation of 0.999 suggesting that the first factor is "strong" and relatively easy to identify by PCA-based methods. The correlations of the first RP-PCA and PCA factor with the CRSP-VW index return is 0.99 confirming the interpretation of "level" factors and essentially proxies for the market return.

The second panel of Figure 13a shows the composition of the second factor. In contrast to the long-only structure of the first factor, the second factor is composed of long and short positions of individual portfolios. Moreover, the decile-10 and decile-1 loadings are (mostly)

of different signs, hence this factor is akin to long-short factors in which the high-return factors are of opposite signs of the low-return factors. For most portfolios the high-return decile 10 has a positive weight while the weight of the low-return decile is negative but there are several exceptions. The average RP-PCA portfolio weight is -0.41, much smaller than the mean loading of 4.99 of the first factor, and hence represents an almost zero-investment strategy similar to the Fama-French factor portfolios. The heatmap also shows that the second factor has large (in absolute value) loadings of value/growth related portfolios. The second PCA factor loads strongly on traditional multiples (book/market, dividend/price, earnings/price, cash-flow/price and sales/price) and suggests an interpretation of a "value" factor. The second RP-PCA factor also loads positively on these multiple portfolios, but in contrast to the PCA factor, the loadings of value-related multi-dimensional anomalies are also large (Value-Momentum, Value-Momentum-Profitability, Value-Profitability). Hence this factor can be interpreted as a "value/value-interaction" factor.

The interpretation of the third factor is less clear. It has a similar long-short structure as the second factor with a mean holding of close to zero. The RP-PCA factor loads most on momentum and profitability-related portfolios but the PCA factor is not linked to particular categories. In contrast, the heatmaps in the top panel of Figure 13b show that the fourth RP-PCA and PCA factors load most heavily on momentum and momentum-interaction portfolios. The loadings of other portfolios are significantly smaller (with some exceptions). For both estimation methods, the fourth factor represents a "momentum/momentum-interaction" factor.

The portfolios on the x-axis of the heatmaps for the fifth factor in the bottom panel of Figure 13b are sorted by the SR of the 10-1 returns instead of categories. Recall that the fifth RP-PCA factor has a SR of 0.45 while the SR of the fifth PCA factor is only 0.18 (see Table 5). The heatmap shows the reason why this is the case. The loadings of the fifth RP-PCA factor are linked to the SR of portfolios while this is not the case for the fifth PCA factor. The RP-PCA weights of the five anomaly portfolios with the highest SR (*indrrev*, *indmomrev*, *indrrev*, *season*and *valprof*) are all positive while theses PCA weights are close to zero. Thus, the fifth RP-PCA factor is a "high SR" factor while the fifth PCA factor does not have a clear interpretation, which is not surprising since it is insignificant (see Figure 9).

While the estimation is agnostic about the types and "names" of portfolios, it is noteworthy that most of the identified factors have clear economic interpretations. The first RP-PCA and PCA factor is a market proxy, the second factor is a "value/value-interaction" factor (RP-PCA) or a pure "value" factor (PCA) and the third factor is a "momentum/momentum-interaction" factor. The fourth RP-PCA factor is linked to momentum and profitability while the fifth RP-PCA factor is a "high SR" factor.

Proxy factors

In order to interpret latent factor models Pelger and Xiong (2018) propose the use of proxy factors. The proxy factors use only the largest portfolio weights of the latent factors and set the smaller portfolio weights to zero. Pelger and Xiong (2018) show that the largest factor portfolio weights already contain most of the information signal even if the true factor itself is not sparse. We approximate the first latent factor by an equally weighted portfolio of all assets and use the 8 largest portfolio weights for $N = 74$ respectively the 10% largest portfolio weights for $N = 370$ to approximate the latent second to fifth RP-PCA and PCA factors.

Based on the portfolio weights of the proxy factors in Table 6 we confirm our previous interpretation about which characteristics describe the latent factors. The first RP-PCA proxy is a market factor. The second to fourth RP-PCA proxy factors load heavily on value, momentum, value- and momentum-interaction. The fifth RP-PCA proxy factor can be interpreted as a LS-factor based on reversal-interaction, which are high SR anomalies. The second to fourth PCA proxy factors are based on value, momentum and momentum-interaction. Interestingly, the fifth PCA proxy factor has a clear interpretation as an asset turnover/profitability LS-factor, which has a relatively low SR.

One of the major problems when comparing two different sets of factors is that a factor model is only identified up to invertible linear transformations. Two sets of factors represent the same factor model if the factors span the same vector space. When trying to interpret estimated factors by comparing them with economic factors, we need a measure to describe how close two vector spaces are to each other. As proposed by Bai and Ng (2006) the generalized correlation is a natural candidate measure. Intuitively, we calculate the correlation between the latent and candidate factors after rotating them appropriately. Generalized correlations close to 1 measure of how many factors two sets have in common.[15]

Table 7 shows the generalized correlations of the original factors with the proxy factors, i.e. the correlations after rotating the latent factors. The generalized correlation of the first four proxy factors with the estimated RP-PCA and PCA factors is 0.94 and higher confirming that the proxy factors provide a good approximation to the latent factors. The fifth GC is somewhat lower; 0.71 for RP-PCA and 0.86 for PCA. The next question is how well the proxy

[15] Let F be the K latent factors and G are K_G candidate factors. We want to test if a linear combination of the candidate factors G can replicate some or all of the factors F. The first generalized correlation is the highest correlation that can be achieved through a linear combination of the factors F and the candidate factors G. For the second generalized correlation we first project out the subspace that spans the linear combination for the first generalized correlation and then determine the highest possible correlation that can be achieved through linear combinations of the remaining $K-1$ respectively K_G-1 dimensional subspaces. This procedure continues until we have calculated the $\min(K, K_G)$ generalized correlation. Mathematically the generalized correlations are the square root of the $\min(K, K_G)$ largest eigenvalues of the matrix $\text{Cov}(F,G)\text{Var}(F)^{-1}\text{Cov}(F,G)\text{Var}(G)^{-1}$. If $K = K_G = 1$ it is simply the correlation.

factors price the portfolios. Panel B of Table 4 shows the pricing statistics for models with proxy factors in comparison with the statistics of the original factors in Panel A. Overall, the results are very similar confirming that the proxy factors retain most of the pricing information of the the original factors. Note that the proxy models tend to outperform the models with original factors out-of-sample. This is due to the more parsimonious representation with fewer portfolios that have non-zero weights that comprise the proxy factors.

The full sample, $N = 370$

Next, we study the loadings for the full sample with all ten deciles of each anomaly ($N = 370$). Instead of plotting individual factor loadings we focus on the composition of the SDF of the estimated RP-PCA and PCA models. Figure 14 shows the heatmap of all 370 individual weights in the SDFs of RP-PCA (Panel A) and PCA (Panel B). All loadings are multiplied by 10 for ease of display. The anomalies on the x-axis are sorted by the anomaly Sharpe ratios. Several patterns emerge from the SDF loadings. First, the RP-PCA SDF weights are related to the Sharpe ratios of anomalies. The largest loadings are the high-return decile 10 portfolios of the anomalies with the highest Sharpe-ratios in the top left corner of the heatmap. The most negative loadings are the low-return decile 1 portfolios of these anomalies in the bottom left. Aside from the high SR anomalies, some profitability portfolio loadings are also large. Second, the SDF is mostly comprised of the 1 and 10 deciles as the weights of the middle portfolios tend to be small.

The heatmap of the PCA SDF in Panel B shows a different pattern. First, there is no link between loadings and Sharpe ratios. This is consistent with the findings for the smaller sample with only the 1 and 10 deciles discussed above. For example, the largest PCA loading is *lev* portfolio 10 with 0.19, while its RP-PCA loading is only 0.02 (recall that the loading in the heatmap are multiplied by 10). The weights of the high SR anomaly portfolios on the left are (mostly) small and hence do not enter in the PCA in any significant way. Hence it is not surprising that the SR of the RP-PCA model is about twice as high as that of the PCA SDF (cf. Table 4).

Figure 17 shows the cumulative weights by decile portfolios. For example, the "Decile 1" bars show the sum of the 37 decile-1 RP-PCA and PCA SDF weights. The pattern is similar for RP-PCA and PCA estimations: The SDF is mostly comprised of the extreme 1 and 10 deciles while the middle decile loadings are smaller. In other words, most of the pricing information is contained in the extreme deciles.

How do the statistical RP-PCA and PCA factors relate to simple long-short factors (i.e. with weights of 1 for decile 10, -1 of decile 1 and 0 otherwise)? To answer this question, we compute generalized correlations of statistical factors with long-short portfolios. The first LS-factor is the market factor and LS-factors are added incrementally based on the largest

accumulative absolute loading of the anomaly in the portfolio weights of the statistical factors. The results are shown in Figure 16. Each panel shows five lines for the five statistical factors in each model/sample. The number of incremental factors is on the x-axis and the generalized correlation is on the y-axis. Consider first the RP-PCA model using the small sample with only the deciles 1 and 10 in the top left panel. The purple line corresponding to the first RP-PCA factor shows a generalized correlation of close to one with the first LS-factor (the market return) confirming the results described above that the first RP-PCA factor is essentially a market proxy. This is true for the first factors in all cases displayed in Figure 15. The generalized correlation of the second to fifth statistical factors with LS-portfolios is generally significantly lower than one. For example, five LS-portfolios are required to generate a generalized correlation with the second RP-PCA factor of 0.8 while about 10 LS-portfolios are required for factors three and four. The fifth factor is the most difficult to proxy for with LS portfolios; a combination of 18 LS-portfolios yields a generalized correlation of 0.8. These patterns are broadly similar for PCA factors and the complete panel of 370 portfolios. Statistical factors clearly yield more parsimonious representations of pricing factors than simple LS-factors.

Stability over time

Next, we investigate whether the RP-PCA and PCA factor estimations are stable over time and whether the assumption of constant loadings is justified. Using a rolling window with 240 months we estimate the loadings for RP-PCA and PCA locally for each month and compare them with the loadings estimated on the whole sample of $T = 650$. We are interested if the space spanned by factors estimated over different time horizons is stable. If we have the same set of latent factors but their order changes over time, it still describes the same factor model. Figure 18 plots the generalized correlations of the five RP-PCA and PCA factor loadings estimated using rolling windows with those estimated on the whole sample. Generalized correlations close to one indicate that the loading and hence also factor space stays stable. We observe that RP-PCA is more stable over time than PCA and its smallest generalized correlations is on average larger than 80%. In Figure 19 we plot the loadings for each portfolio (on the x-axis) and each rolling sample. The loadings are rotated to match the loadings of the whole time horizon. If the loadings were constant over time, the plot would show a single line with the constant loading for each portfolios. If loadings were to vary substantially, the plots would show many different lines and look more "fuzzy". All five RP-PCA factor loadings appear to be quite stable with little variation over time. This is true also for the first four PCA factors but the loadings of fifth PCA factor are significantly less stable over time indicating that PCA is unable to detect this factor.

Individual stocks

Next, we briefly explore the ability of RP-PCA and PCA to capture returns of individual stocks rather than returns of portfolios. Recall that both methods assume that factor loadings are constant over time and we showed that this assumptions is satisfied in the case of portfolio returns. Intuitively, the assumption of constant loadings appears problematic in the case of individual stocks. For example, the turnover in sorts of some anomaly characteristics is very high. For example, a given stock might be in the highest momentum portfolio at some period but in a lower momentum portfolio in another period. If momentum is a priced factor, then the momentum loading of stocks is likely to change over time. To explore whether this is indeed the case, we estimate RP-PCA and PCA on a panel of large stocks.

Our sample consists of monthly excess returns of a balanced panel of stock returns from 1/1972 to 12/2014, which results in $N = 270$ stocks with $T = 500$ monthly returns. The data is obtained from CRSP. We include only stocks that have been constituents of the S&P 500 index at some point during the sample span and have no missing values during the time period that we consider. Extending to a longer time period would drastically reduce the number of stocks in our panel. By construction the stocks in our sample are mainly large cap stocks and the results are not driven by small stocks.

Figures 20 to 22 show the generalized correlations and loadings for the RP-PCA and PCA estimations using rolling windows, and in-sample and out-of-sample pricing statistics. The plots show that, in contrast to the estimation for portfolios, the estimation for individual stocks is much less robust and exhibits instability in loadings and pricing performance. Generalized correlations of loadings of subsamples with loadings for the entire sample are significantly lower than in the case of portfolios (*cf.* Figures 20 and 18), especially at the start of the sample. Moreover, loadings for rolling windows vary more for stocks than for portfolios (*cf.* Figures 19 and 21).

We study the maximum SR, pricing error and amount of unexplained variation in Figure 22. While adding factors increasing SRs and lowers pricing error in-sample for RP-PCA and PCA, the performance completely deteriorates out-of-sample. This is an indication for the time-variation in the factor structure for individual stocks. We conclude that a factor model with constant loadings is not appropriate for modeling data of individual stock returns. Hence the development of methods that allow for time-varying loadings is essential for estimating factor models on firm-level data, see *e.g.* Kelly et al. (2017), Fan et al. (2016), and Lettau and Pelger (2018b).

5. Conclusion

We propose a new estimator for latent asset pricing factors to bring order into the "factor zoo" and find the asset pricing factors that provide independent information for expected

returns. Our estimator is PCA generalized with a penalty term on the pricing error and as easy to use as conventional PCA. We show that our estimator RP-PCA strongly dominates conventional PCA. We can detect weak factors with high Sharpe-ratios which are undetectable with PCA. Strong factors are estimated more efficiently with RP-PCA compared to PCA.

We have four main conclusions. First, by estimating latent factors instead of using a pre-specified (and potentially misspecified) set of observable factors, we find a small number of factors that can explain the covariance and expected return structure in the data. Second, we demonstrate that using the information in the mean of the data in the estimation is essential for discovering factors that can explain the risk-premium of assets. Our model outperforms the conventional PCA approach. Our factors achieve a higher out-of-sample mean-variance efficiency and smaller pricing errors than alternative models in- and out-of-sample. Third, our factors are based on only a subset of the stock characteristics implying that a significant amount of characteristic information is redundant. Our asset pricing factors can be interpreted as a market factor, value/value-interaction, momentum/momentum-interaction, profitability and a "high SR" factor that loads heavily on reversal. Fourth, most of the pricing information is contained in the extreme deciles.

The key idea for successfully extracting asset pricing factors is to use the insight of arbitrage pricing theory in the estimation. Factors that fit the time-series have to simultaneously fit the cross-section of expected returns. RP-PCA achieves this goal by incorporating this insight in the estimation, while neither observable nor PCA based factors reach this goal. A distinguishing advantage of our estimator is that it finds high Sharpe-ratio factors that are weak and only affect a subset of the assets. These factors are crucial for spanning the stochastic discount factor, but hard or impossible to detect when using PCA. Essentially all latent factor estimators work on projected data which can be interpreted as managed portfolios of stocks based on observable firm characteristics. This projection is necessary to control for the time-variation in the loadings of individual stocks. However, factors that are strong on the original stocks can become weak when considering the sorted portfolios and hence cannot be detected by PCA. RP-PCA however is robust to the choice of managed portfolios and will detect all asset pricing factors.

Last but not least, our model has a direct practical benefit for optimal portfolio choice. As our asset pricing factors provide a better span of the stochastic discount factor compared to PCA or observable factor models, their linear combinations obtains a higher Sharpe-ratio out-of-sample. This suggests that an optimal investment portfolio should only be composed of our RP-PCA factors.

References

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Bryzgalova, S., 2016. Spurious factors in linear asset pricing models. Technical report, Stanford University .

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51, 1281–1304.

Cochrane, J.H., 2011. Presidential address: Discount rates. Journal of Finance 66, 1047–1108.

Connor, G., Korajczyk, R., 1988. Risk and return in an equilibrium apt: Application to a new test methodology. Journal of Financial Economics 21, 255–289.

Connor, G., Korajczyk, R., 1993. A test for the number of factors in an approximate factor model. Journal of Finance 58, 1263–1291.

Connor, G., Korajczyk, R.A., 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. Journal of Financial Economics 15, 373–394. URL: https://www.sciencedirect.com/science/article/pii/0304405X86900279, doi:10.1016/0304-405X(86)90027-9.

DeMiguel, V., Garlappi, L., Nogales, F., Uppal, R., 2017. A portfolio perspective on the multitude of firm characteristics. Working paper .

Fan, J., Liao, Y., Wang, W., 2016. Projected principal component analysis in factor models. The Annals of Statistics 44, 219–254.

Feng, G., Giglio, S., Xiu, D., 2017. Taming the factor zoo. Technical Report, Chicago Booth .

Freyberger, J., Neuhierl, A., Weber, M., 2017. Dissecting characteristics nonparametrically. Technical Report, Chicago Booth .

Giglio, S., Xiu, D., 2017. Asset pricing with omitted factors. Working paper .

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. Review of Financial Studies 29, 5–68.

Kelly, B., Pruitt, S., Su, Y., 2017. Instrumented principal component analysis. Working Paper .

Kozak, S., Nagel, S., Santosh, S., 2017. Shrinking the cross section. Working Paper, Chicago Booth .

Lettau, M., Pelger, M., 2018a. Estimating latent asset pricing factors. Working paper .

Lettau, M., Pelger, M., 2018b. Factor models with time-varying loadings. Working paper .

Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics , 244–258.

Pelger, M., Xiong, R., 2018. Interpretable proximate factors for large dimensions. Working paper .

Ross, S.A., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory 13, 341–360.

Stock, J.H., Watson, M.W., 2002. Forecasting Using Principal Components from a Large Number of Predictors. URL: https://www.jstor.org/stable/3085839, doi:10.2307/3085839.

Table 1: RP-PCA vs. PCA for 25 Double-sorted Portfolios

	In-sample			Out-of-sample		
	SR	RMS α	Idio. Var.	SR	RMS α	Idio. Var.
Size and accrual ($\gamma = 10$)						
RP-PCA	0.24	0.12	6.11	0.21	0.11	6.75
PCA	0.13	0.14	5.92	0.11	0.14	6.72
FF-long/sort	0.21	0.12	7.90	0.11	0.12	7.11
Size and short-term reversal ($\gamma = 20$)						
RP-PCA	0.22	0.15	6.48	0.15	0.19	8.40
PCA	0.19	0.16	6.38	0.10	0.19	8.41
FF-long/sort	0.18	0.17	8.47	0.08	0.20	8.42

Note: Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation. $K = 3$ statistical factors.

Table 2: Largest Eigenvalues

	Size and accrual		Size and short-term reversal	
	PCA	RP-PCA ($\gamma = 10$)	PCA	RP-PCA ($\gamma = 20$)
σ_1^2	19.01	19.01	21.19	21.19
σ_2^2	1.03	1.03	1.07	1.07
σ_3^2	0.24	0.24	0.98	0.98
σ_4^2	0.15	0.15	0.35	0.34
σ_5^2	0.12	0.12	0.24	0.23

Note: Variance signal for different factors: Largest eigenvalues of $\mathbf{\Lambda \Sigma}_F \mathbf{\Lambda}^\top$ normalized by the average idiosyncratic variance $\sigma_e^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_{e,i}^2$.

Table 3: Single-sorted Portfolios

Portfolio	Abbreviation	Mean	SD	SR	Mean		RP-PCA		PCA	
					low	high	low	high	low	high
Ind. Rel. Rev. (L.V.)	indrrevlv	1.28	3.05	0.42	0.00	1.28	-0.29	0.20	0.07	0.09
Industry Mom. Rev.	indmomrev	1.17	3.44	0.34	0.07	1.24	-0.18	0.12	0.07	0.06
Industry Rel. Reversals	indrrev	1.01	4.12	0.24	0.08	1.09	-0.29	0.22	-0.07	-0.05
Seasonality	season	0.79	3.94	0.20	0.22	1.01	-0.19	0.15	0.04	-0.01
Value-Profitability	valprof	0.75	3.81	0.20	0.26	1.01	-0.17	0.15	-0.17	0.28
Momentum (12m)	mom12	1.25	6.88	0.18	-0.16	1.08	-0.17	0.10	-0.14	0.07
Value-Mom-Prof.	valmomprof	0.83	4.83	0.17	0.35	1.18	-0.07	0.15	-0.19	0.20
Investment/Assets	inv	0.47	3.06	0.15	0.40	0.86	-0.06	0.04	0.02	0.06
Composite Issuance	ciss	0.48	3.30	0.14	0.30	0.78	-0.12	0.00	0.02	0.07
Investment Growth	igrowth	0.38	2.71	0.14	0.33	0.71	-0.05	0.02	-0.08	0.00
Sales/Price	sp	0.52	4.28	0.12	0.41	0.93	0.04	0.06	-0.03	0.24
Earnings/Price	ep	0.57	4.69	0.12	0.33	0.90	-0.07	-0.03	-0.10	0.18
Net Operating Assets	noa	0.38	3.28	0.12	0.26	0.64	-0.12	0.10	-0.11	0.00
Accrual	Accrual	0.35	3.14	0.11	0.31	0.66	-0.02	0.05	0.06	-0.08
Value (A)	value	0.48	4.56	0.11	0.45	0.93	0.09	-0.01	0.04	0.10
Gross Profitability	prof	0.36	3.38	0.11	0.38	0.75	-0.17	0.19	-0.23	0.19
Asset Turnover	Aturnover	0.41	3.83	0.11	0.30	0.71	-0.20	0.12	-0.32	0.24
Value-Momentum	valmom	0.51	5.05	0.10	0.40	0.91	0.07	-0.06	-0.04	0.12
Cash Flows/Price	cfp	0.43	4.37	0.10	0.42	0.85	0.06	-0.05	0.04	0.12
Momentum-Reversals	momrev	0.46	4.84	0.10	0.42	0.88	-0.06	0.09	-0.06	0.08
Asset Growth	growth	0.29	3.46	0.08	0.41	0.70	-0.02	-0.03	-0.04	0.02
Long Run Reversals	lrrev	0.41	5.07	0.08	0.47	0.89	0.00	0.11	0.00	0.18
Industry Momentum	indmom	0.47	6.21	0.08	0.34	0.81	0.03	-0.06	-0.06	0.10
Idiosyncratic Volatility	ivol	0.54	7.16	0.08	-0.01	0.53	-0.18	-0.04	-0.19	0.07
Value (M)	valuem	0.40	5.86	0.07	0.52	0.92	0.08	0.14	0.06	-0.05
Short-Term Reversals	strev	0.36	5.27	0.07	0.26	0.62	-0.26	0.12	-0.06	-0.02
Size	size	0.29	4.80	0.06	0.46	0.75	-0.03	0.09	0.00	0.16
Momentum (6m)	mom	0.35	6.25	0.06	0.58	0.93	0.07	0.00	-0.12	0.07
Leverage	lev	0.26	4.63	0.06	0.48	0.73	0.10	-0.09	0.04	0.22
Return on Assets (A)	roaa	0.21	4.08	0.05	0.37	0.58	-0.08	0.12	-0.16	0.09
Dividend/Price	divp	0.18	5.08	0.04	0.49	0.67	0.02	-0.11	0.08	0.13
Investment/Capital	invcap	0.12	5.01	0.02	0.57	0.68	0.15	-0.09	-0.06	-0.01
Return on Book Equity (A)	roea	0.08	4.39	0.02	0.51	0.59	-0.07	0.08	-0.12	0.13
Sales Growth	sgrowth	0.05	3.64	0.01	0.58	0.53	0.03	-0.08	-0.07	-0.06
Gross Margins	gmargins	0.01	3.37	0.00	0.55	0.56	-0.10	0.10	-0.01	-0.01
Share Volume	shvol	0.02	5.96	0.00	0.49	0.47	-0.03	-0.05	-0.08	0.06
Price	price	0.01	6.80	0.00	0.50	0.49	0.07	-0.03	-0.05	0.01

Note: Long-short portfolios of extreme deciles of 37 single-sorted portfolios from 07/1963 to 12/2016: Mean, standard deviation and Sharpe-ratio, mean of low and high portfolios and contribution to the SDF based on RP-PCA and PCA with 6 factors applied to the extreme decile portfolios ($N = 74$).

Table 4: Fit of RP-PCA vs. PCA

	In-sample			Out-of-sample		
	SR	RMS α	Idio. Var.	SR	RMS α	Idio. Var.
Panel A: Estimated Factors						
$N = 74$						
RP-PCA 3 factors	0.37	0.24	13.94%	0.30	0.22	15.32%
PCA 3 factors	0.22	0.27	14.00%	0.14	0.25	16.00%
RP-PCA 5 factors	0.57	0.17	10.40%	0.50	0.15	12.06%
PCA 5 factors	0.30	0.22	10.30%	0.24	0.20	11.98%
$N = 370$						
RP-PCA 3 factors	0.23	0.17	12.75%	0.18	0.15	14.57%
PCA 3 factors	0.17	0.17	12.68%	0.14	0.15	14.66%
RP-PCA 5 factors	0.53	0.14	10.76%	0.45	0.12	12.70%
PCA 5 factors	0.24	0.14	10.66%	0.17	0.14	12.56%
Panel B: Proxy Factors						
$N = 74$						
Proxy RP-PCA 3 factors	0.31	0.25	14.52%	0.26	0.22	16.01%
Proxy PCA 3 factors	0.24	0.28	14.46%	0.22	0.24	15.72%
Proxy RP-PCA 5 factors	0.58	0.17	10.40%	0.50	0.15	11.97%
Proxy PCA 5 factors	0.33	0.22	11.09%	0.27	0.18	12.10%
$N = 370$						
Proxy RP-PCA 3 factors	0.26	0.18	13.07%	0.22	0.15	14.77%
Proxy PCA 3 factors	0.19	0.17	12.97%	0.15	0.15	14.86%
Proxy RP-PCA 5 factors	0.56	0.12	11.29%	0.47	0.12	13.06%
Proxy PCA 5 factors	0.29	0.14	11.12%	0.22	0.13	13.23%

Note: Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for panels of single-sorted portfolios: 74 decile 1 and decile 10 portfolios and all 370 deciles portfolios of 37 anomalies.

Table 5: Individual RP-PCA and PCA Factors

	RP-PCA			PCA		
	Mean	Variance	SR	Mean	Variance	SR
			$N = 74$			
1. Factor	5.35	1892.53	0.12	5.22	1940.31	0.12
2. Factor	2.46	67.51	0.30	0.68	98.30	0.07
3. Factor	1.80	96.56	0.18	1.40	73.67	0.16
4. Factor	0.28	66.73	0.03	1.13	70.96	0.13
5. Factor	2.09	21.72	0.45	0.81	19.85	0.18
			$N = 370$			
1. Factor	11.67	7295.35	0.14	11.56	7387.22	0.13
2. Factor	2.65	222.62	0.18	1.66	241.03	0.11
3. Factor	0.46	213.34	0.03	0.23	207.49	0.02
4. Factor	2.40	125.92	0.21	1.52	132.57	0.13
5. Factor	2.76	39.10	0.44	0.78	49.30	0.11

Note: Properties of individual factors in estimation of panels of single-sorted portfolios: 74 decile 1 and decile 10 portfolios and all 370 deciles portfolios of 37 anomalies.

Table 6: Portfolio Weights of Proxy Factors 2 to 5

RP-PCA							
divp 10	1.53	mom12 10	2.04	size 10	2.14	valuem1 0	1.93
growth 1	-1.46	mom 10	1.99	ivol 1	2.13	indrrev 10	1.39
igrowth 1	-1.51	indmomrev 10	1.90	valmomprof 10	1.89	price 1	1.31
ep 1	-1.53	mom 1	-2.29	mom12 10	1.84	indrrevlv 10	1.26
invcap 1	-1.69	valuem 10	-2.32	mom 10	1.82	lrrev 10	1.25
shvol 1	-1.72	ivol 1	-2.93	price 1	1.69	strev 1	-1.22
mom12 1	-2.32	price 1	-3.51	shvol 1	1.65	indrrevlv 1	-1.34
ivol 1	-2.48	mom12 1	-4.00	indmomrev 1	-1.57	indrrev 1	-1.37
PCA							
valuem 10	2.91	divp 10	1.74	indmom 10	2.42	valprof 10	1.25
price 1	2.52	ivol 10	1.69	mom 10	2.39	Aturnover 10	1.15
divp 10	2.26	roea 1	-1.64	valmom 10	2.18	prof 10	0.95
value 10	2.24	mom12 1	-1.65	mom12 10	2.12	sp 10	0.95
lrrev 10	2.06	size 10	-1.82	valmomprof 10	2.12	lrrev 10	0.86
sp 10	1.98	shvol 1	-1.90	indmom 1	-2.38	valprof 1	-0.98
cfp 10	1.92	ivol 1	-3.16	mom12 1	-2.70	prof 1	-1.51
mom12 1	1.88	price 1	-3.21	mom 1	-2.71	Aturnover 1	-1.89

Note: Portfolio composition of second to fifth proxy factors based on $N = 74$ extreme deciles. The portfolio weights of the proxy factors are the 8 largest loadings of the latent factors. RP-weight $\gamma = 10$

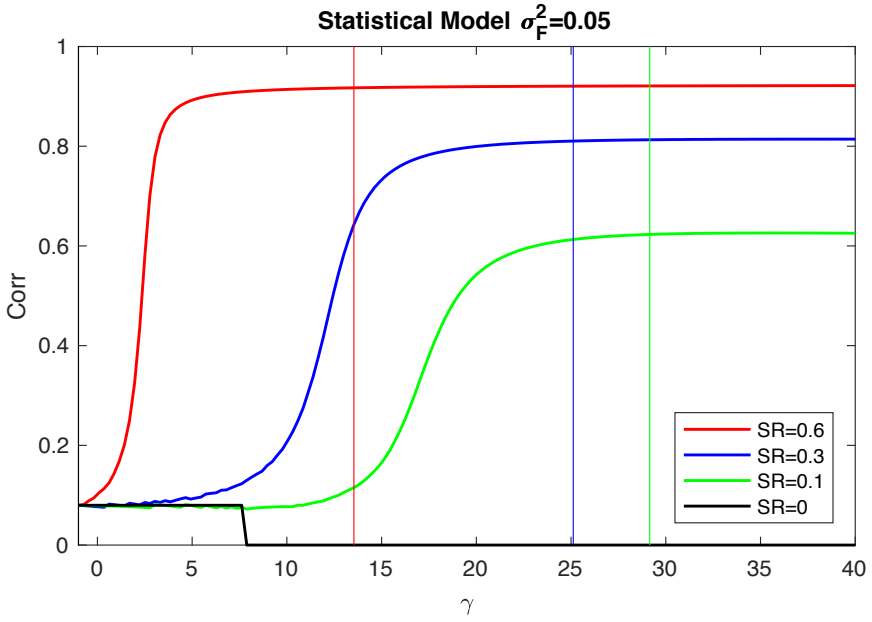
Table 7

	Proxy factors		37 Long-short factors	
	RP-PCA	PCA	RP-PCA	PCA
		$N = 74$		
1.GC	1.00	1.00	1.00	1.00
2.GC	0.99	0.99	1.00	1.00
3.GC	0.95	0.97	1.00	1.00
4.GC	0.95	0.94	0.99	0.99
5.GC	0.71	0.86	0.65	0.65
		$N = 370$		
1.GC	1.00	1.00	0.99	0.99
2.GC	0.99	0.99	0.99	0.99
3.GC	0.98	0.99	0.98	0.98
4.GC	0.94	0.94	0.93	0.91
5.GC	0.77	0.89	0.64	0.64

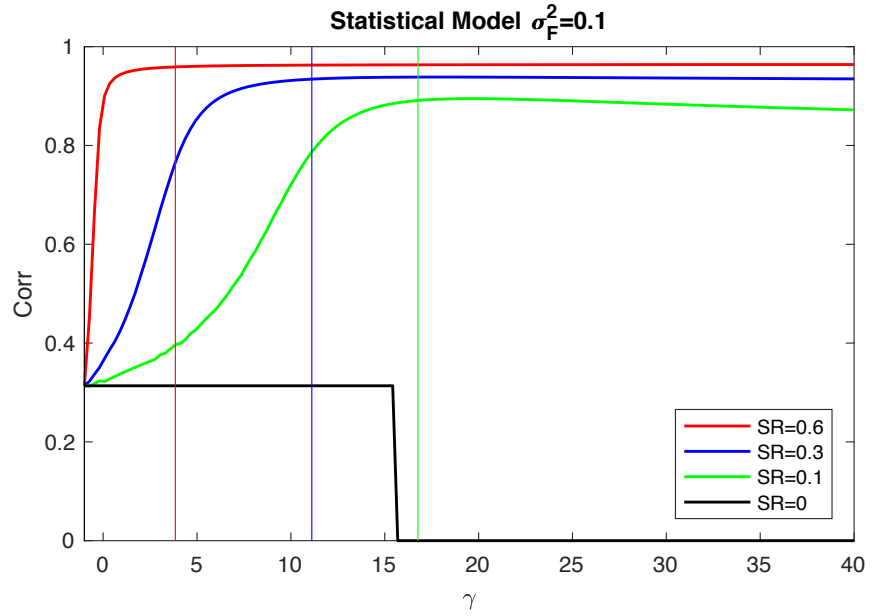
Note: Top: First and last decile of 37 single-sorted portfolios ($N = 74$): Bottom: Deciles of 37 single-sorted portfolios ($N = 370$): Generalized correlations of statistical factors with proxy factor (portfolios of 8 assets for $N = 74$ and portfolios of 10% of assets for $N = 370$) and 37 long-short anomaly factors. $K = 5$ statistical factors and RP-weight $\gamma = 10$

Figure 1: Weak Factors – Correlation of estimated Factor with the true Factor

Panel A: Low Factor Variance



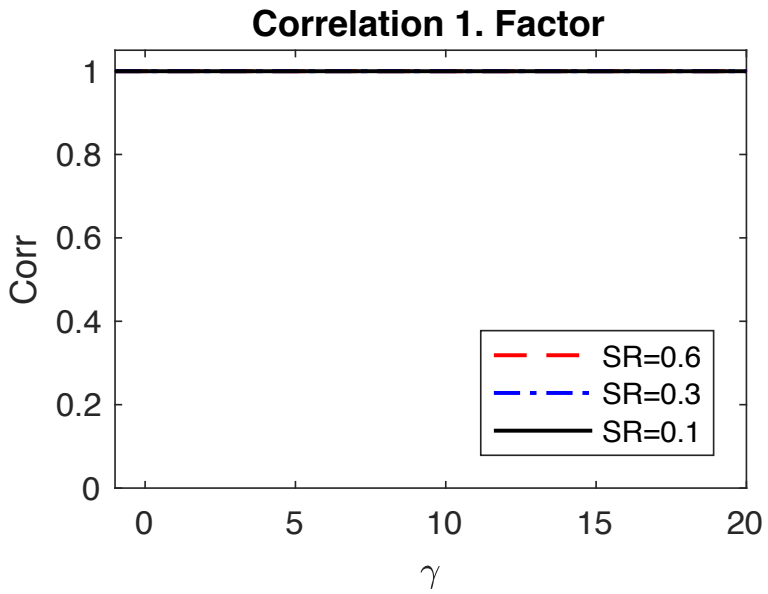
Panel B: High Factor Variance



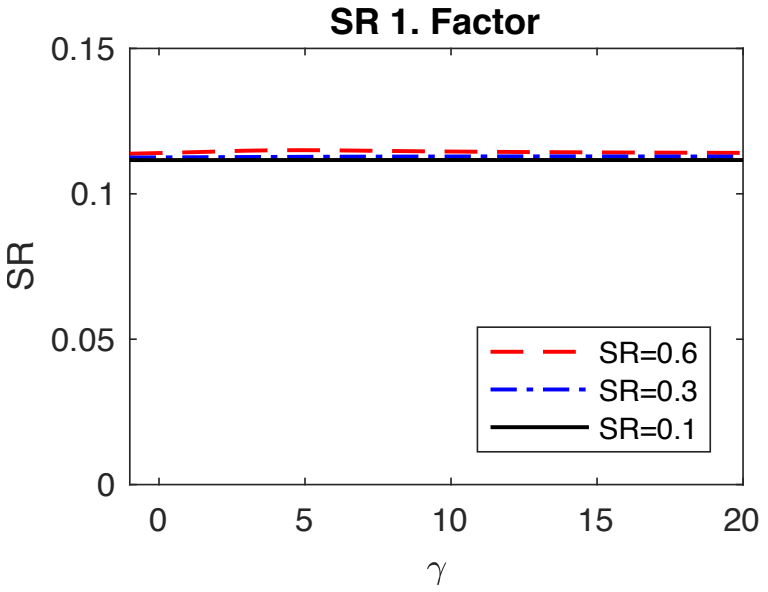
Note: This plot shows $\tau_k(\gamma)$ as a function of γ for different parameter settings ($N = 370$ and $T = 650$).

Figure 2: Simulation - First Factor

Panel A: Correlation with True Factor

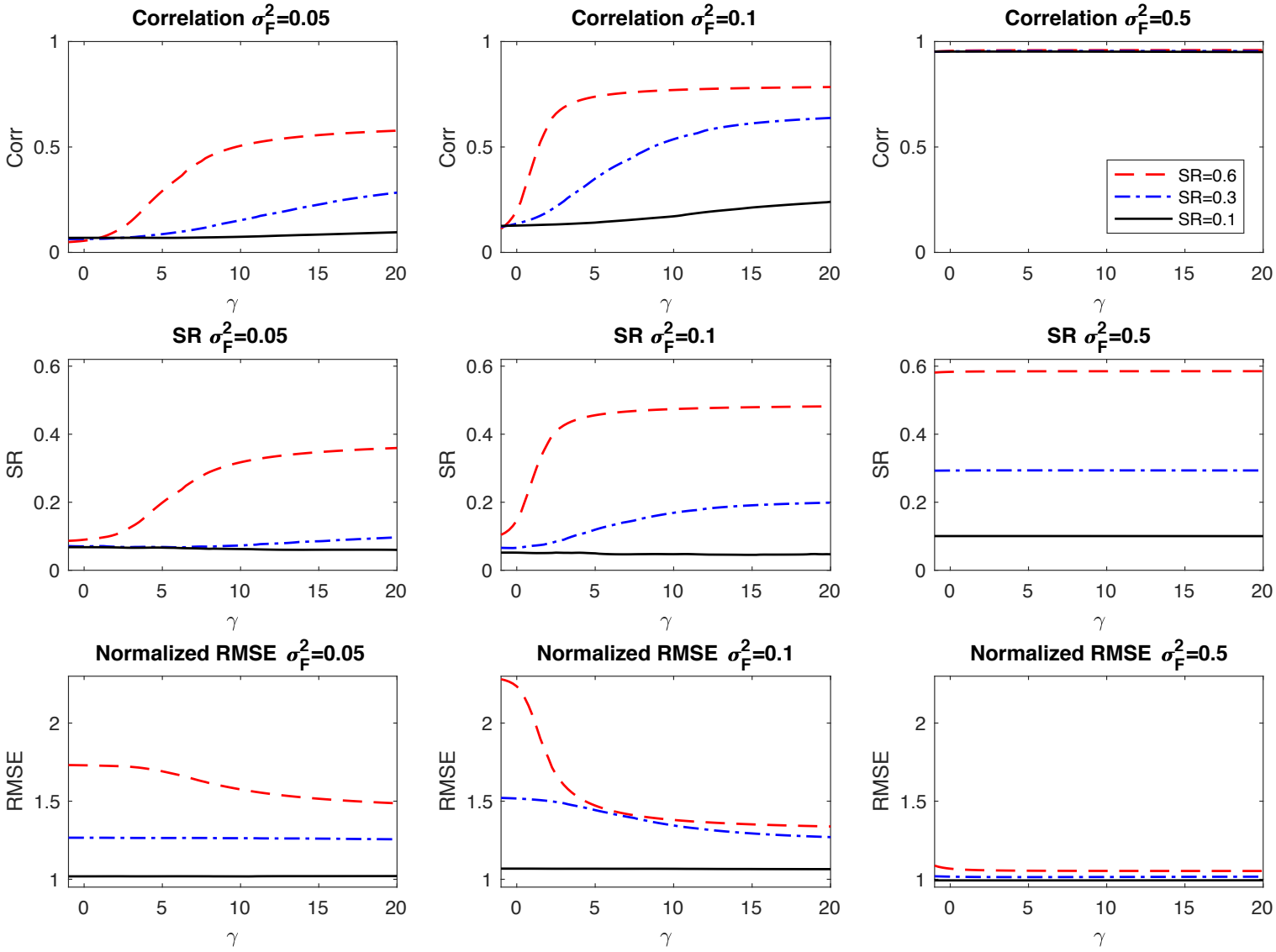


Panel B: SR



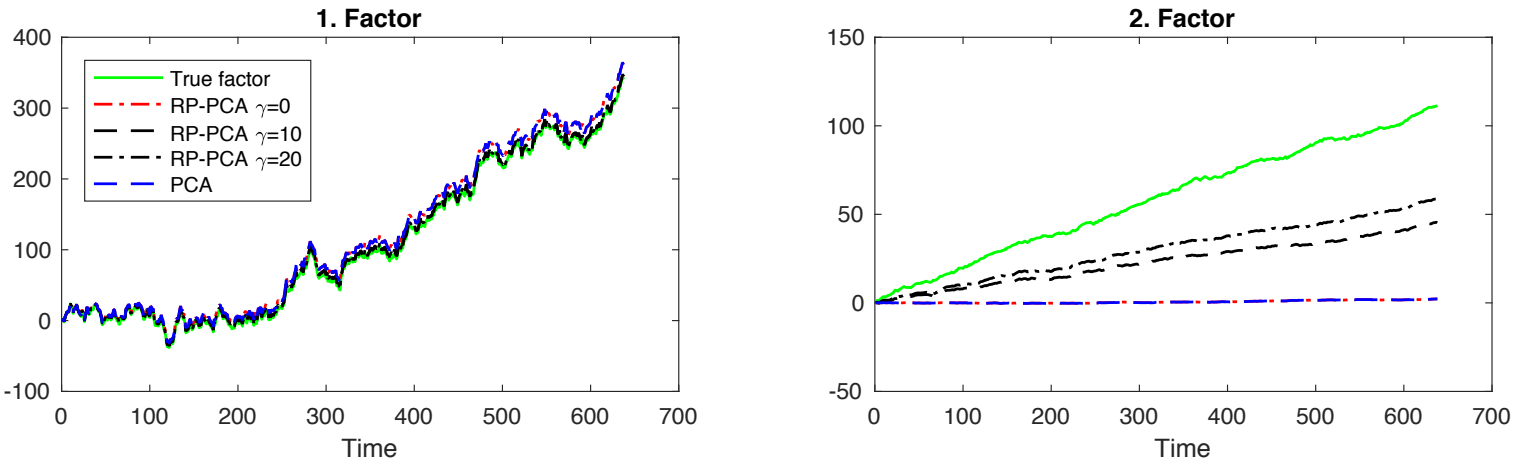
Note: Correlations and Sharpe-ratios as a function of γ for different variances and Sharpe ratios for the first (strong) factor.

Figure 3: Simulation - Second Factor



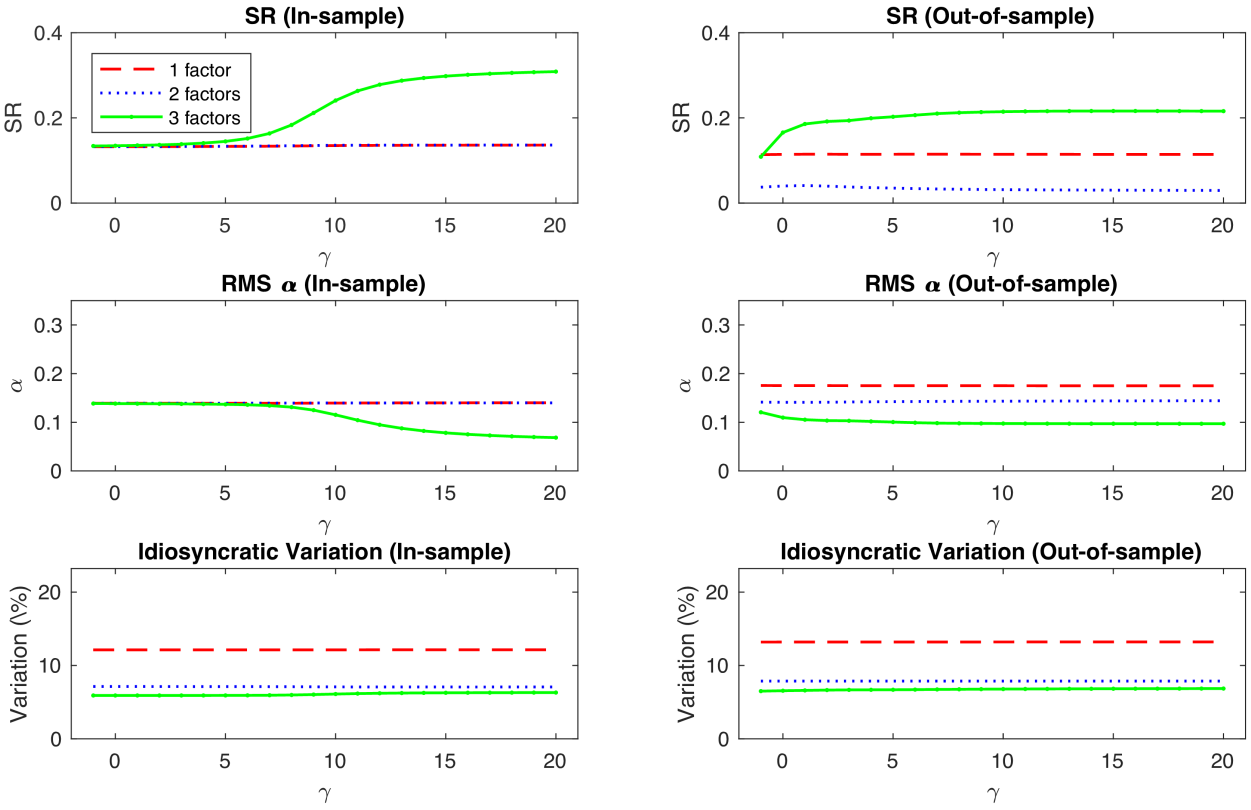
Note: Out-of-sample correlations, Sharpe-ratios and normalized RMSE for the second (weak) factor as a function of γ for different variances and Sharpe ratios.

Figure 4: Simulation - Factor realizations



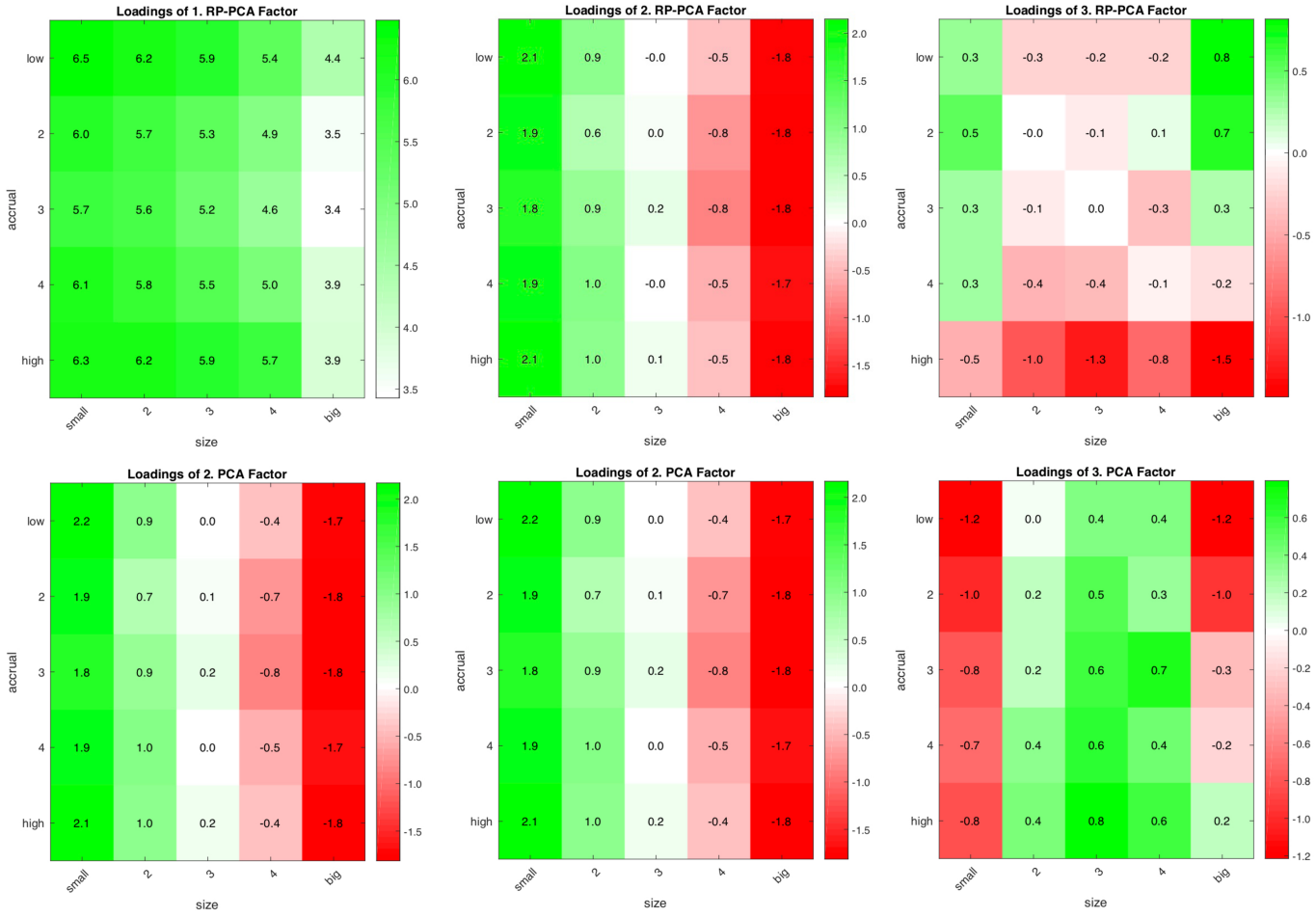
Note: Sample path of one representative simulation for the true factor as well as estimated factors for different values of γ.

Figure 5: RP-PCA vs. PCA Fit for Size/Accrual Portfolios



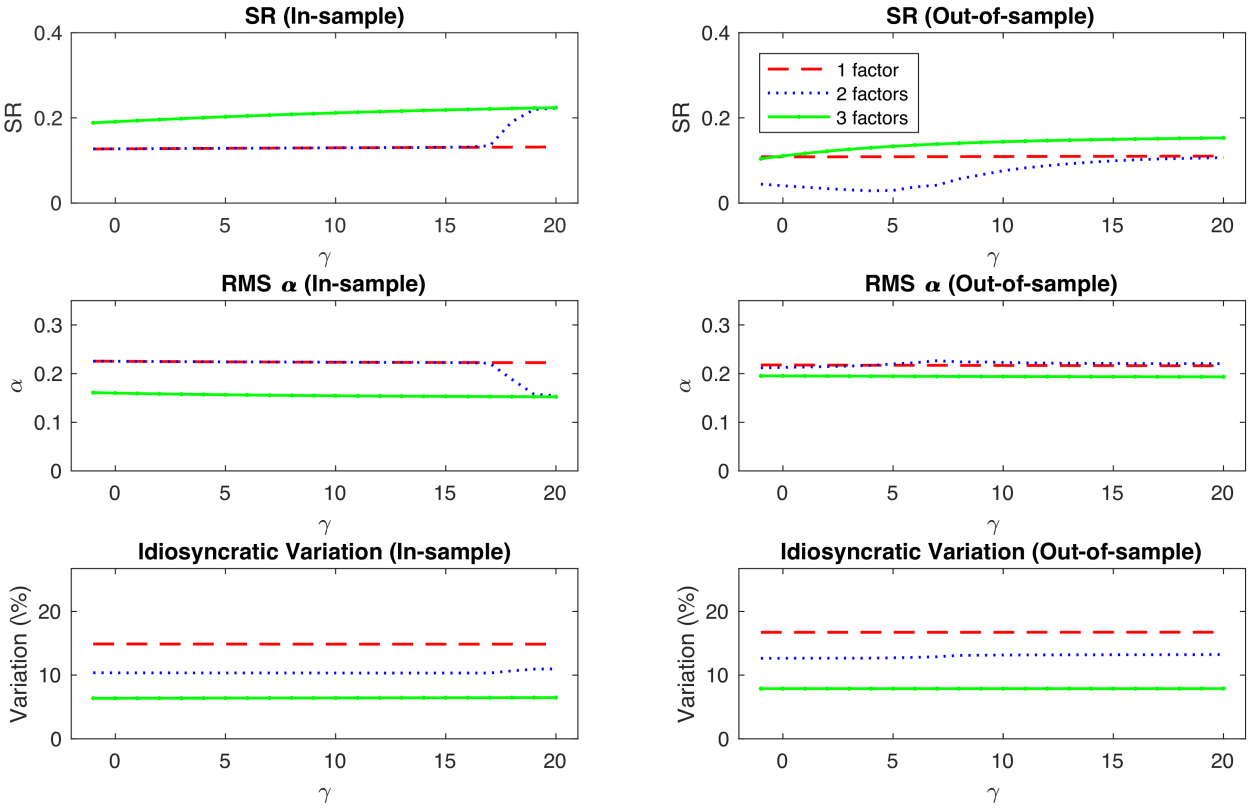
Note: Size and Accrual: Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for different values of γ.

Figure 6: Heatmap of Factor Loadings for Size/Accrual Portfolios



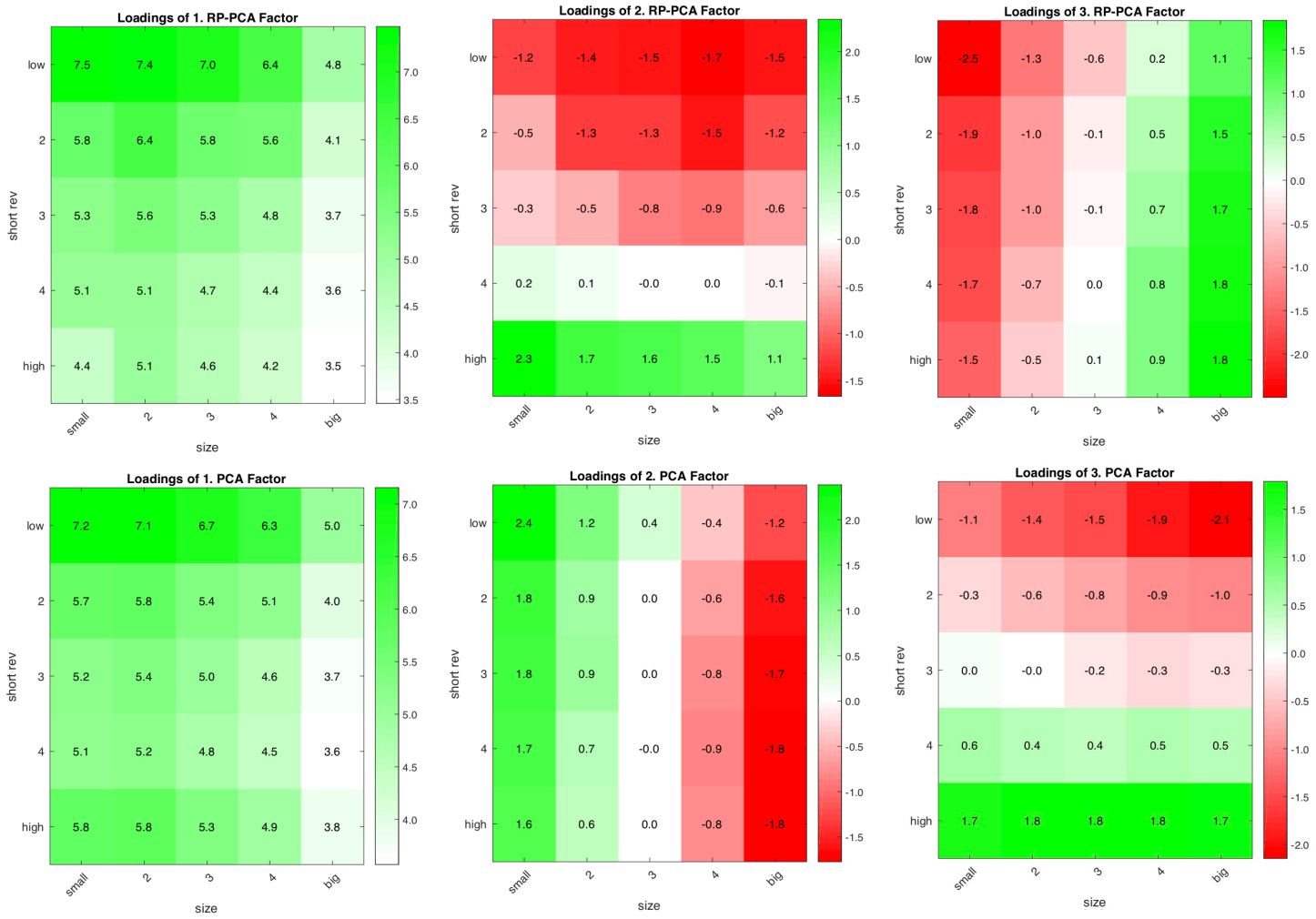
Note: Size and Accrual: Heatmap of loadings. $K = 3$ statistical factors and RP- weight $\gamma = 10$.

Figure 7: RP-PCA vs. PCA Fit for Size/ST Reversal Portfolios



Note: Size and short-term reversal: Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for different values of γ.

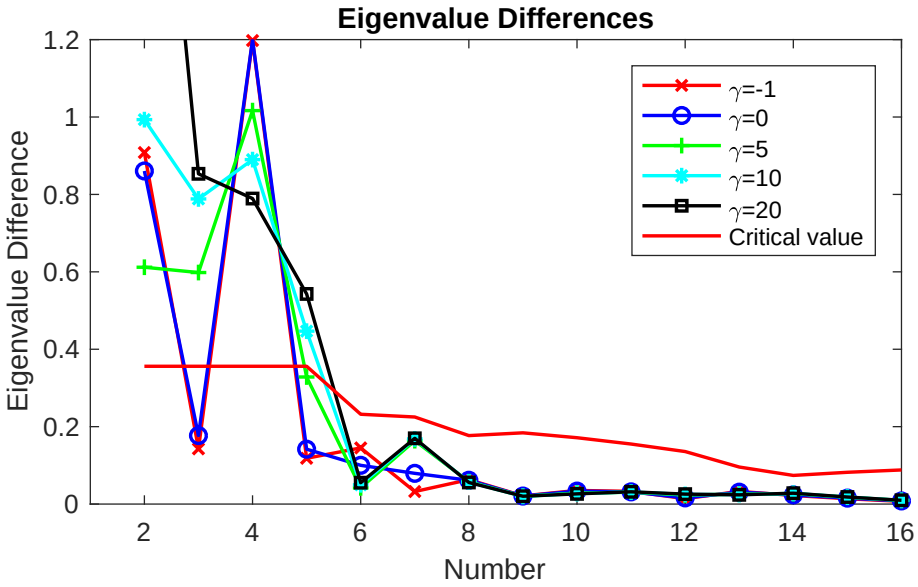
Figure 8: Heatmap of Factor Loadings for Size/ST Reversal Portfolios



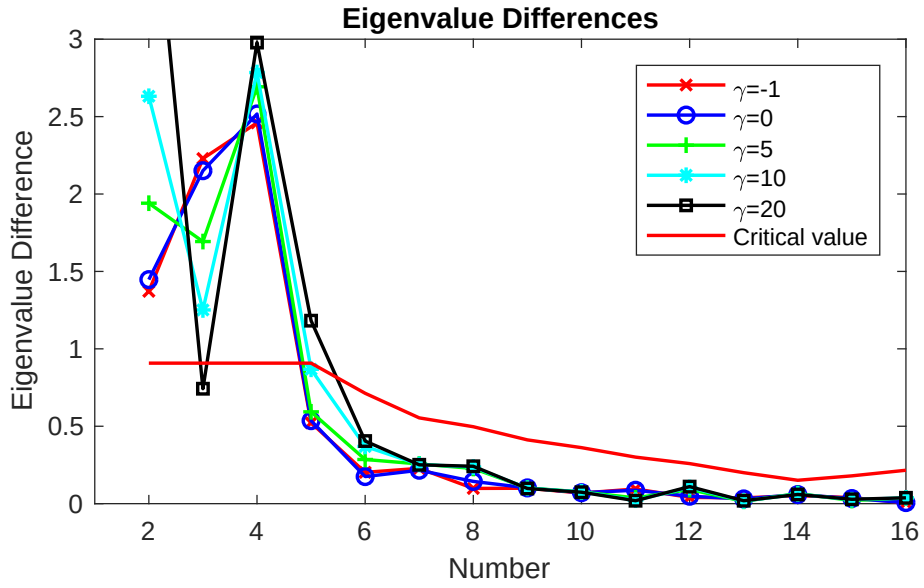
Note: Size and short-term reversal: Heatmap of loadings. $K = 3$ statistical factors and RP- weight $\gamma = 20$.

Figure 9: Eigenvalue Differences for Single-sorted Portfolios

Panel A: $N = 74$

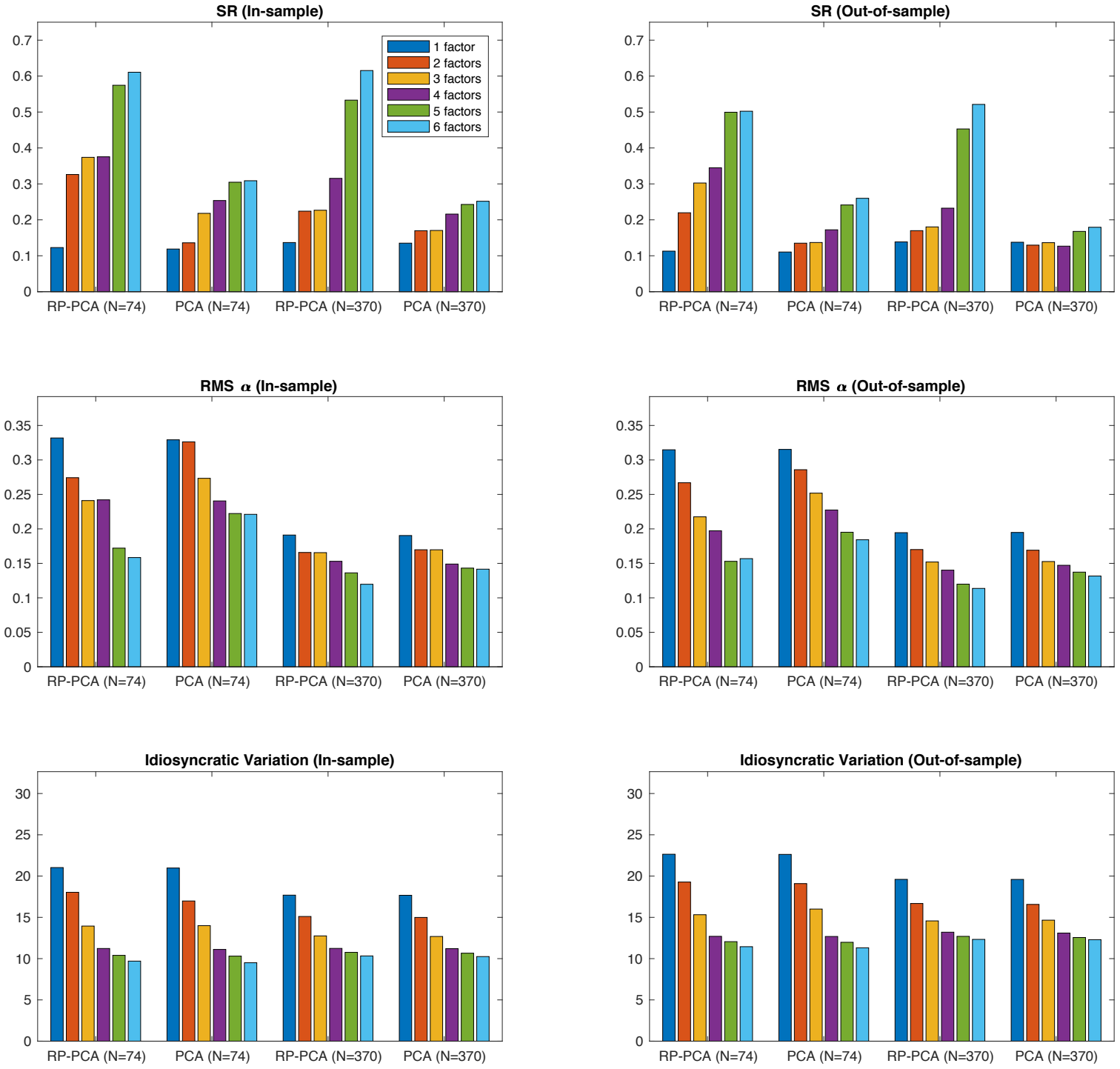


Panel B: $N = 370$



Note: Differences of consecutive eigenvalues of the matrix $\left(\frac{1}{T}\mathbf{X}^\top\mathbf{X} + y\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\right)$ for different RP-weights y. The red line indicates the critical value of the Onatski-test. Top panel: First and last decile of 37 single-sorted portfolios ($N = 74$). Bottom panel: Deciles of 37 single-sorted portfolios ($N = 370$).

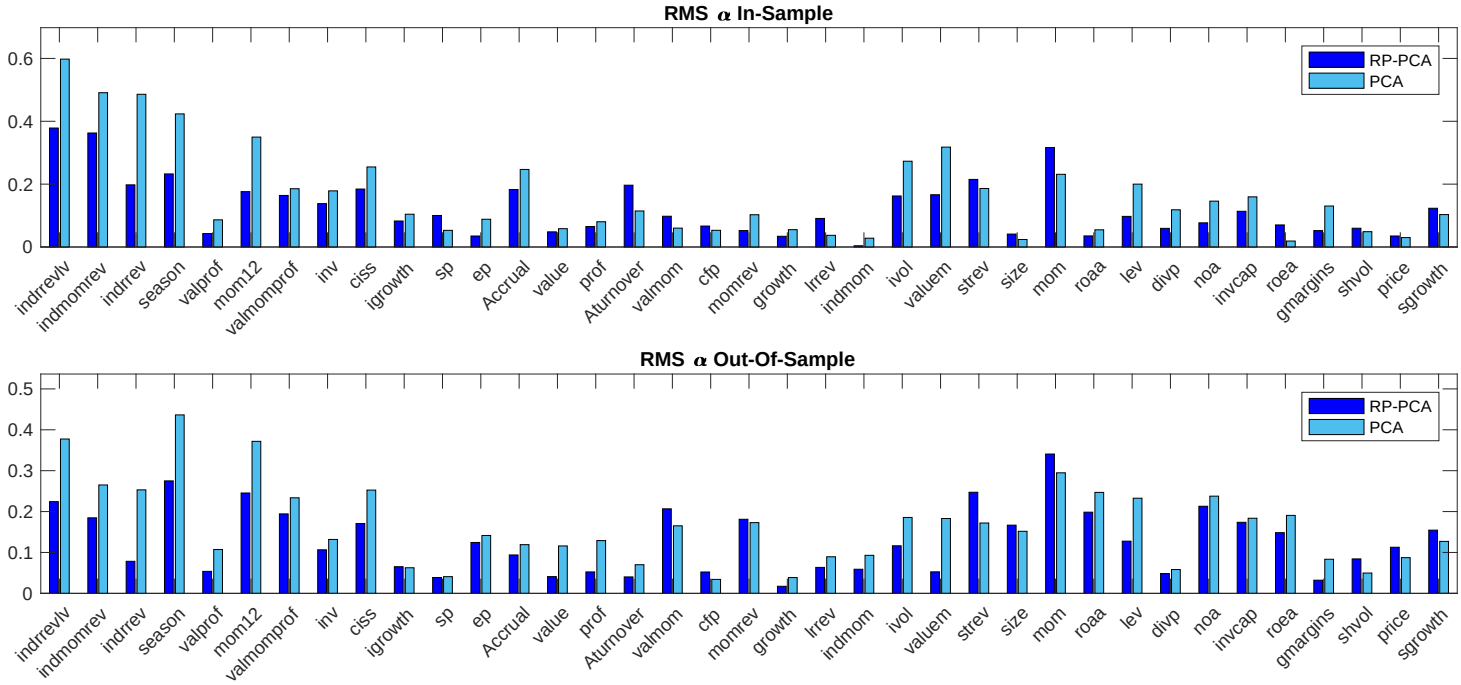
Figure 10: RP-PCA vs. PCA Fit for Single-sorted Portfolios



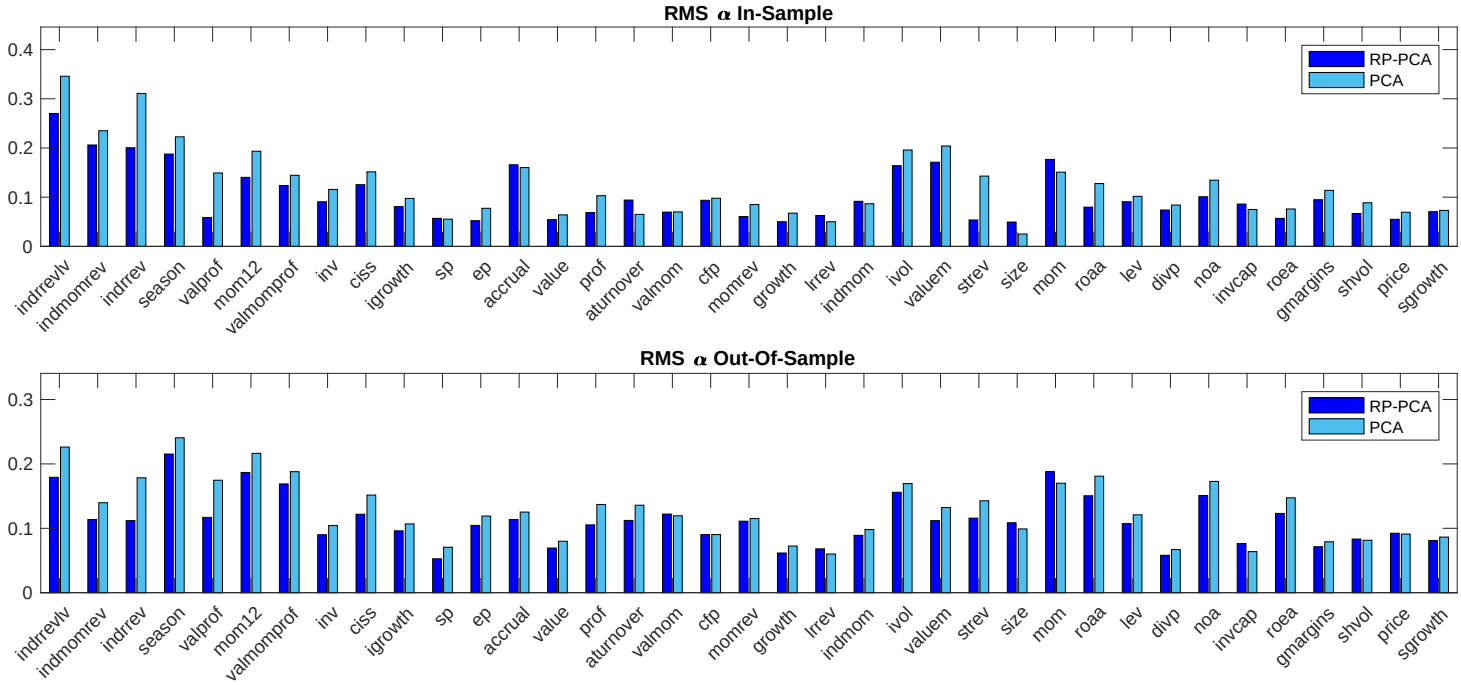
Note: Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for different number of factors. RP-weight $\gamma = 10$, extreme deciles ($N = 74$) or all deciles ($N = 370$).

Figure 11: RMS of Time-series α's by Characteristic

A: $N = 74$

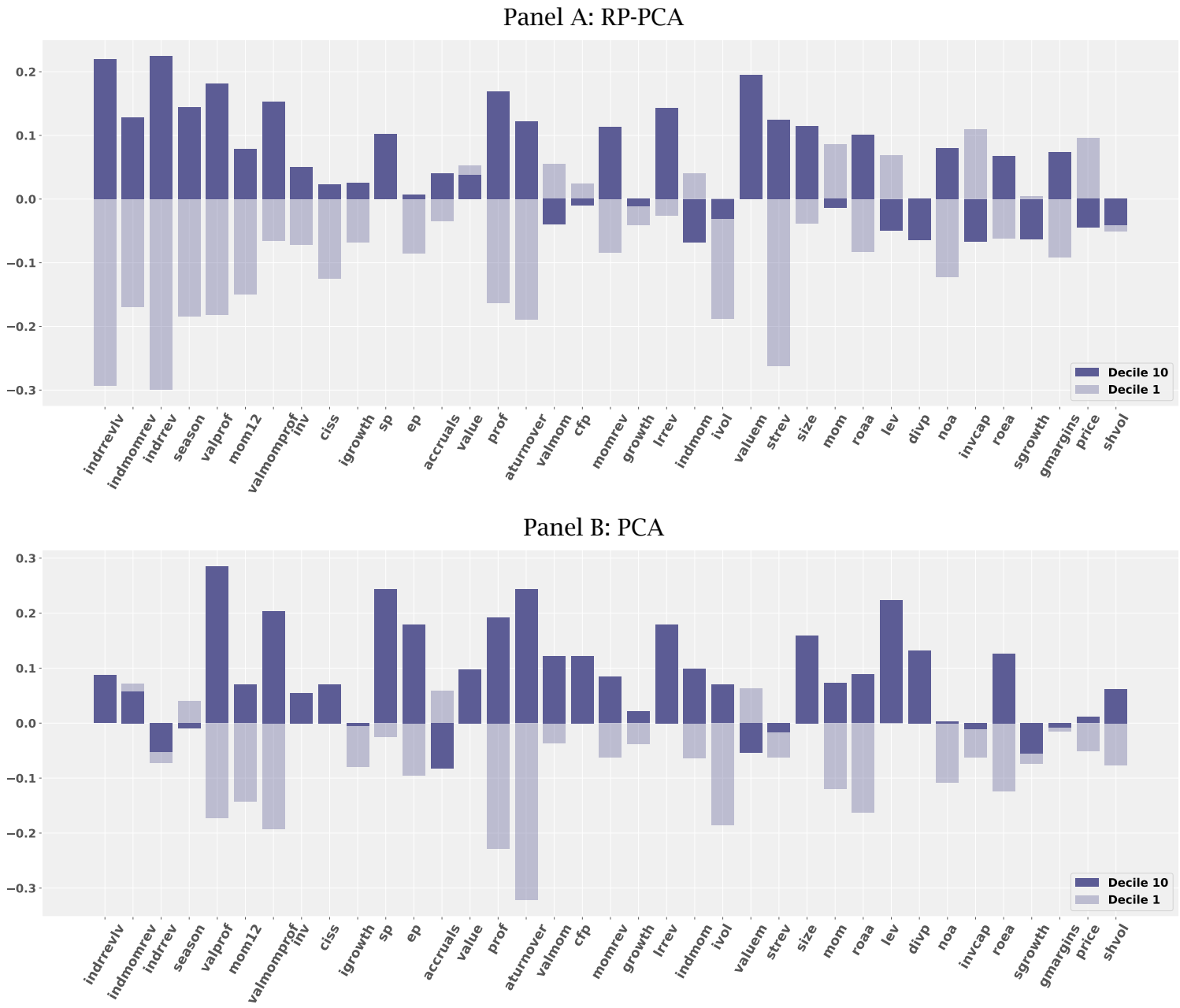


B: $N = 370$



Note: Top: First and last decile of 37 single-sorted portfolios ($N = 74$ and $T = 638$). Bottom: Deciles of 37 single-sorted portfolios ($N = 370$ and $T = 638$). Root-mean-squared pricing errors in- and out-of-sample for 6 RP-PCA and PCA factors.

Figure 12: Portfolio Weights in RP-PCA and PCA SDFs



Note: Portfolio loadings of SDF estimated by RP-PCA and PCA for first and last decile of 37 single-sorted portfolios. The anomalies on the x-axis are sorted by their SR.

Figure 13a: Heatmap of Factor Loadings

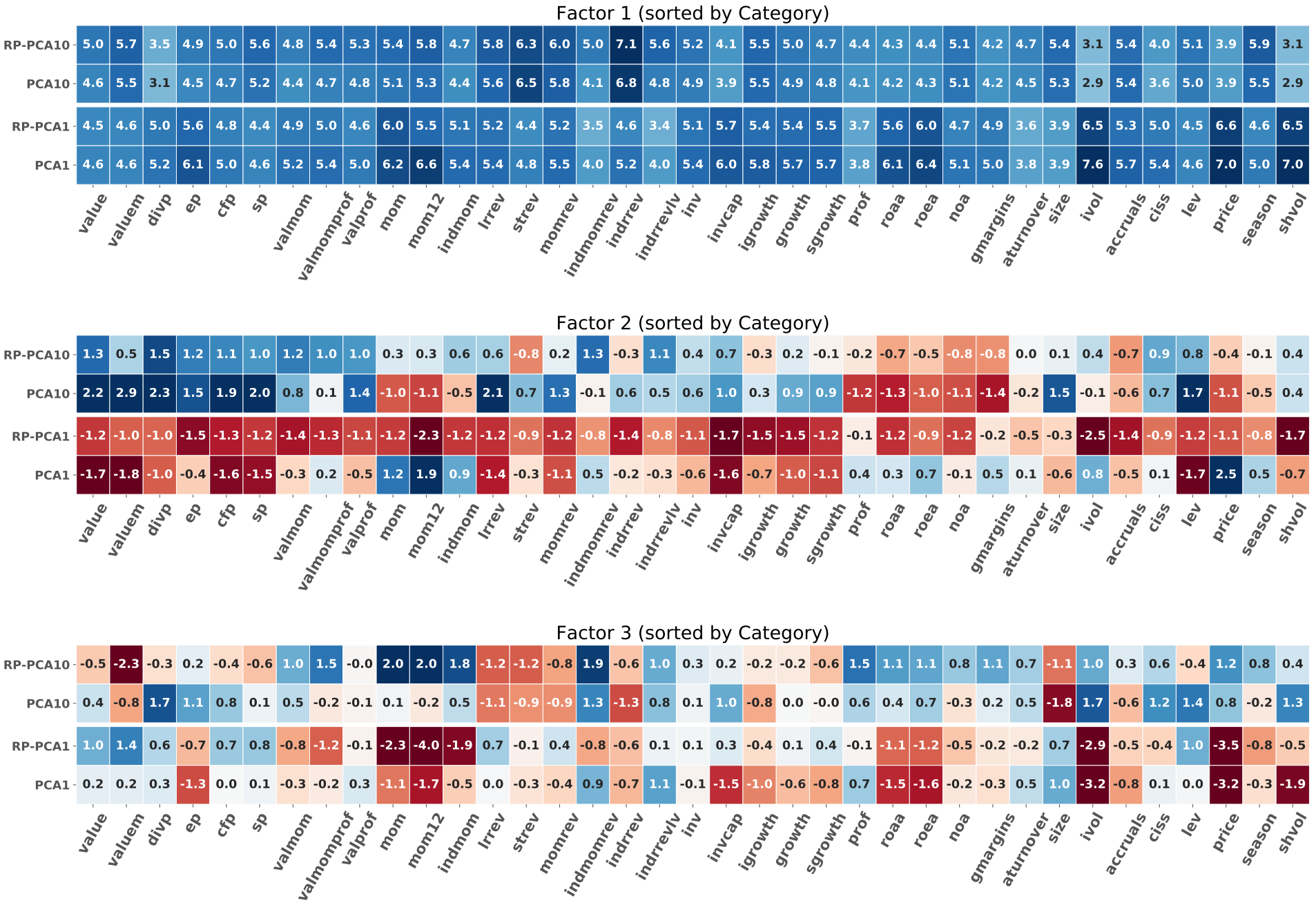
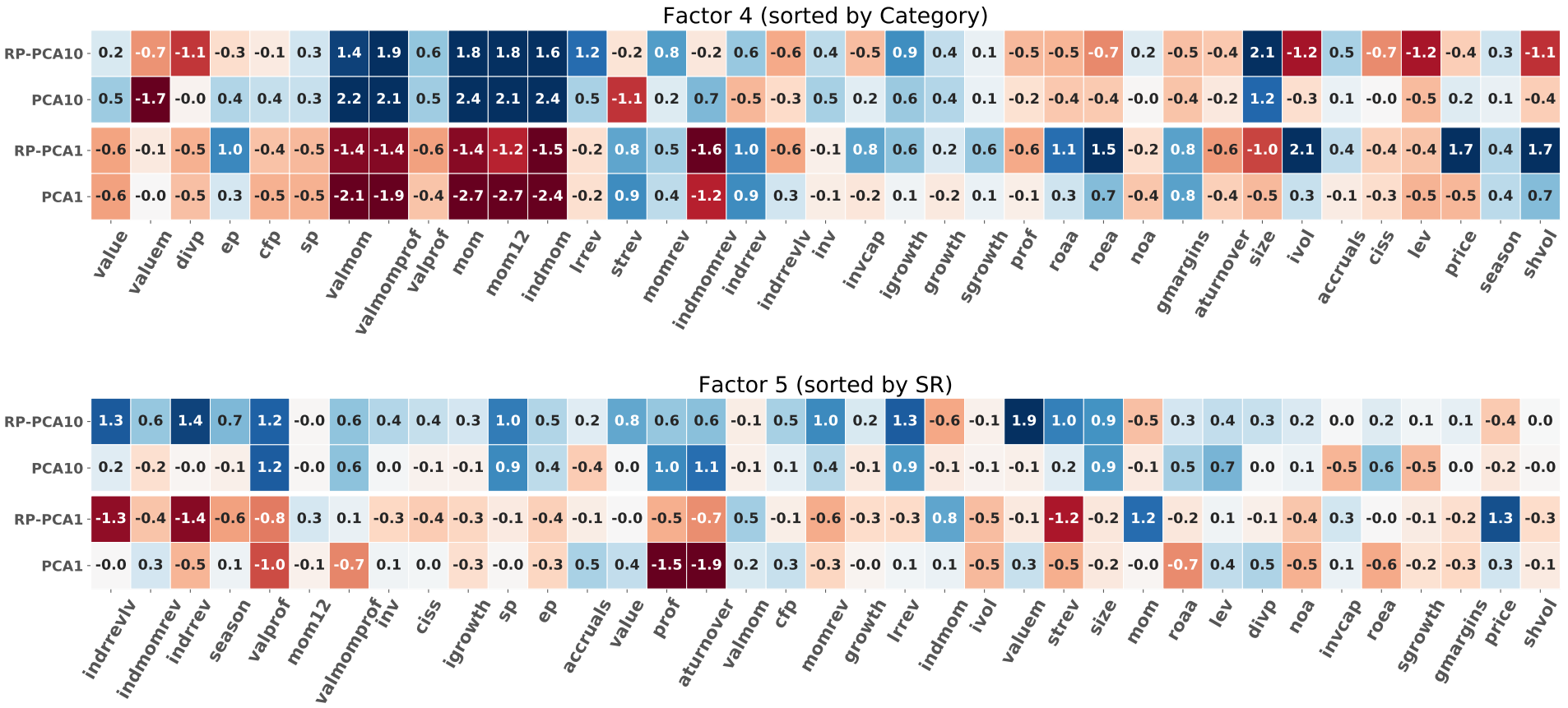
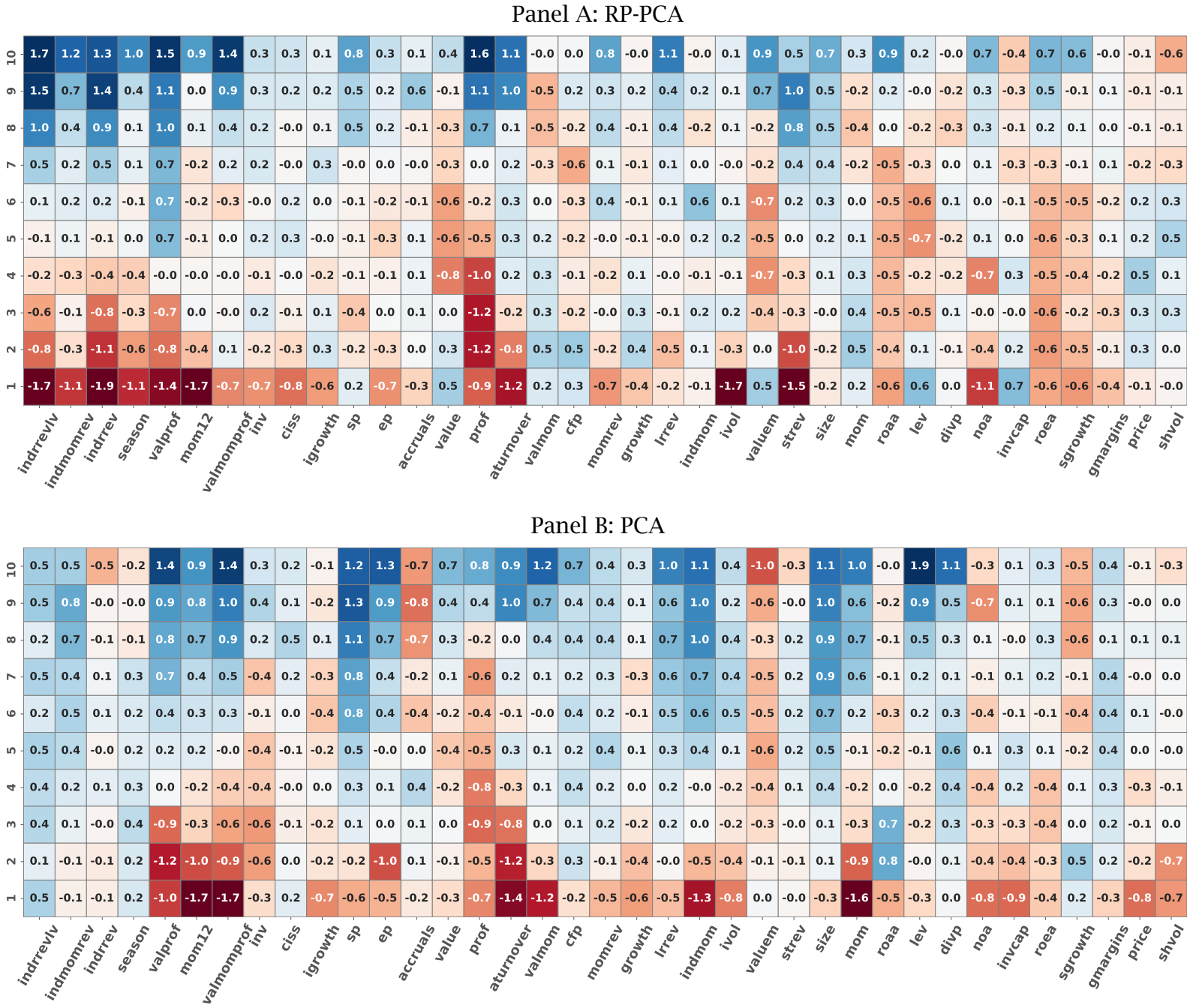


Figure 13b: Heatmap of Factor Loadings, contd.



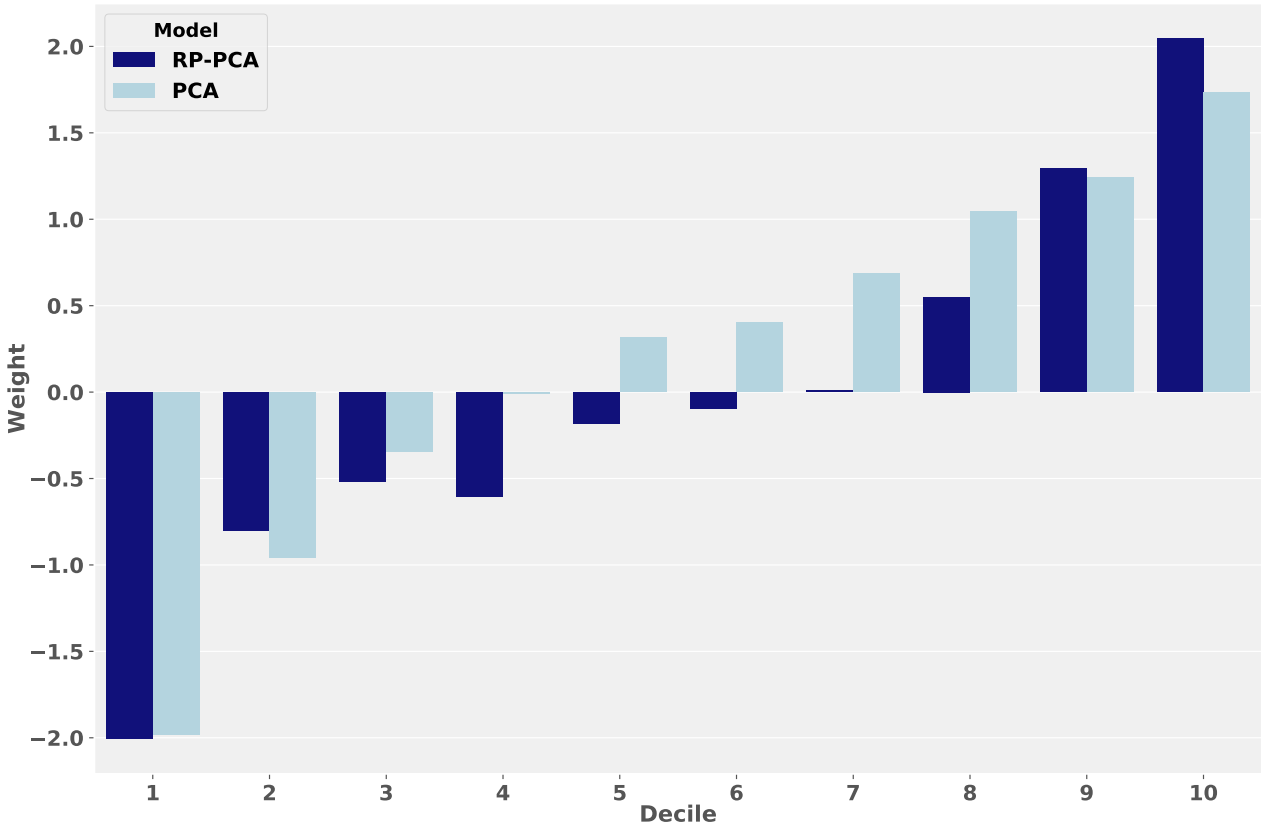
Note: Heatmaps of portfolio weights in six factors for first and last decile of 37 single-sorted portfolios ($N = 74$).

Figure 14: Heatmap of Factor Loadings, $N = 370$



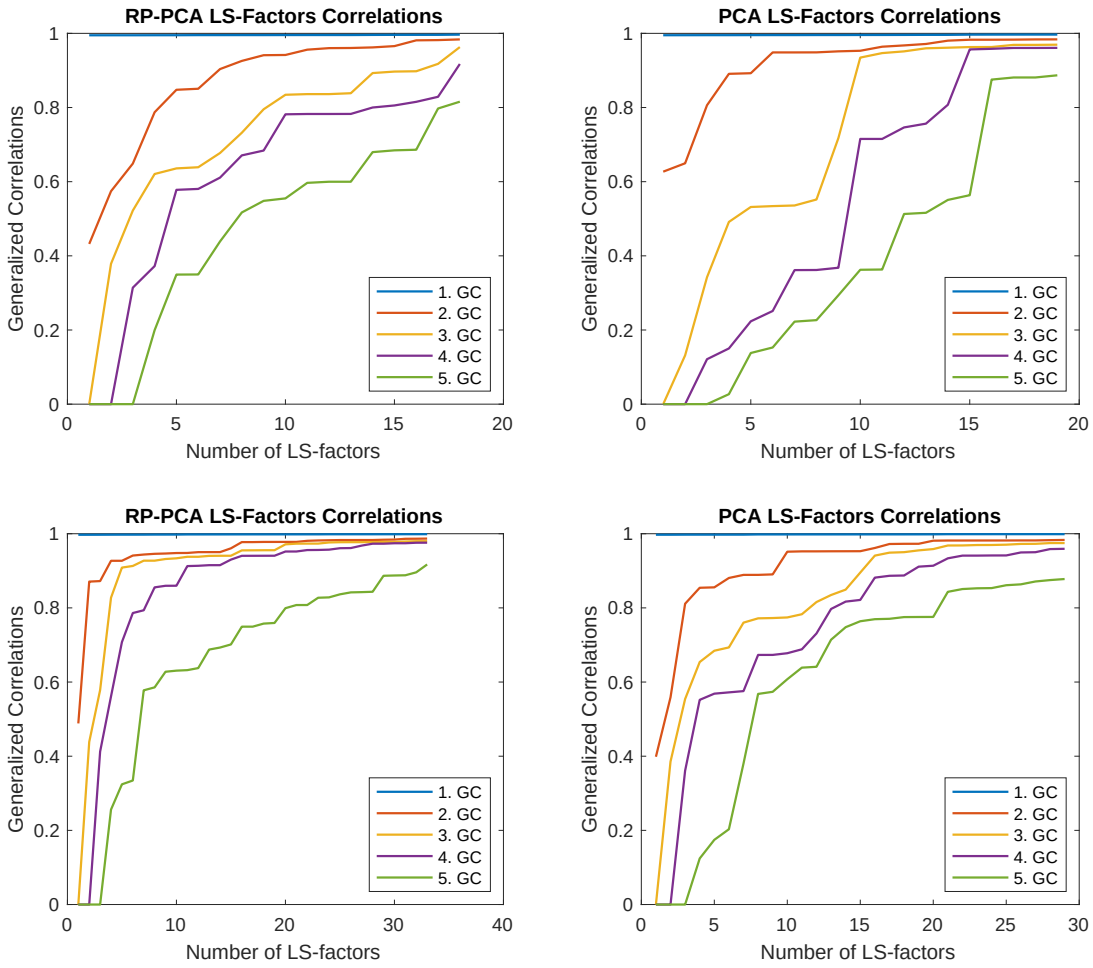
Note: Heatmaps of portfolio weights RP-PCA and PCA SDFs for all decile portfolios ($N = 370$). The loadings are multiplied by 10.

Figure 15: Loadings by Deciles



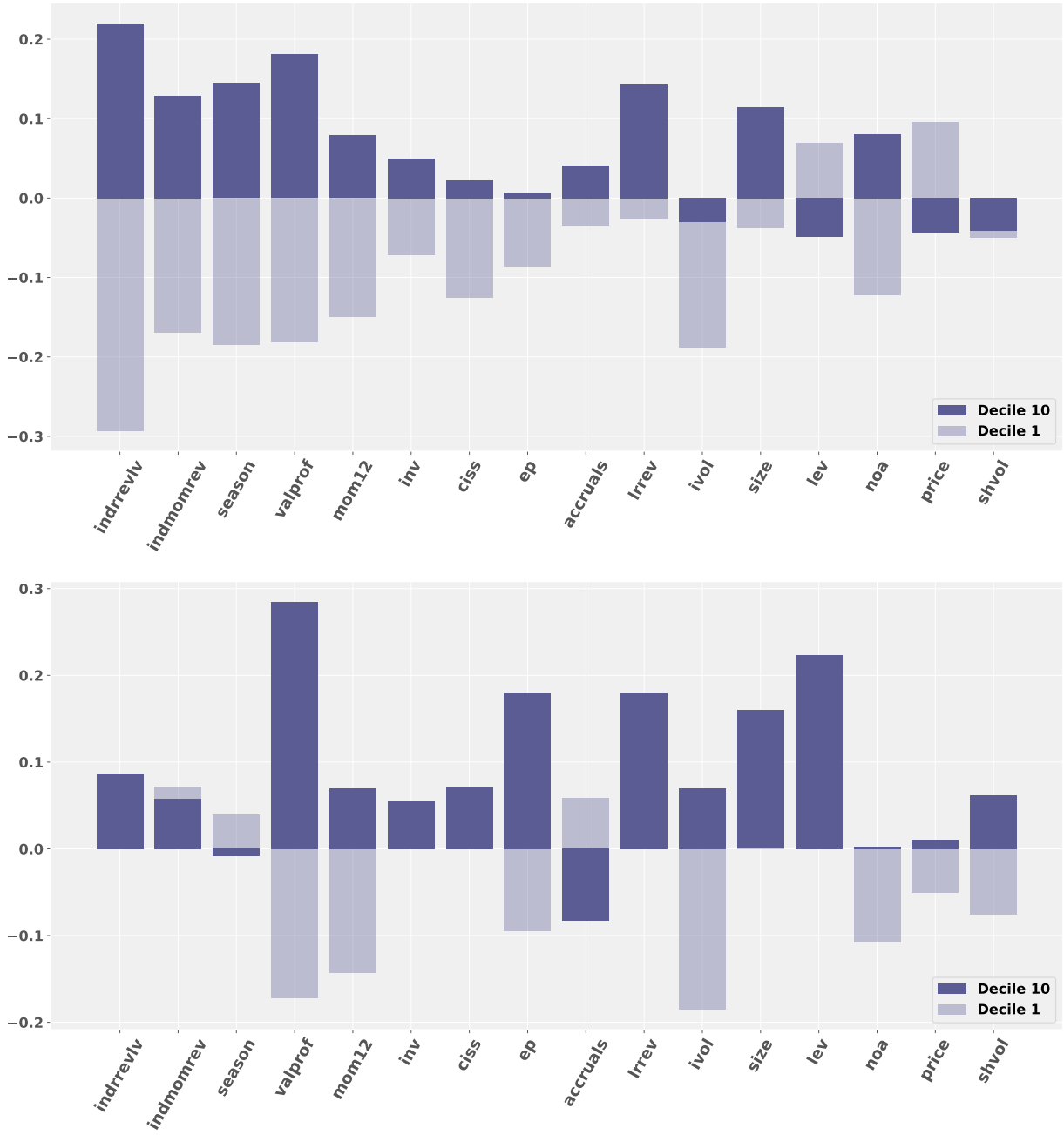
Note: The figure shows combined loadings by deciles for RP-PCA and PCA for $N = 370$.

Figure 16: Generalized Correlations or RP-PCA and PCA Factors with Long-short Portfolios



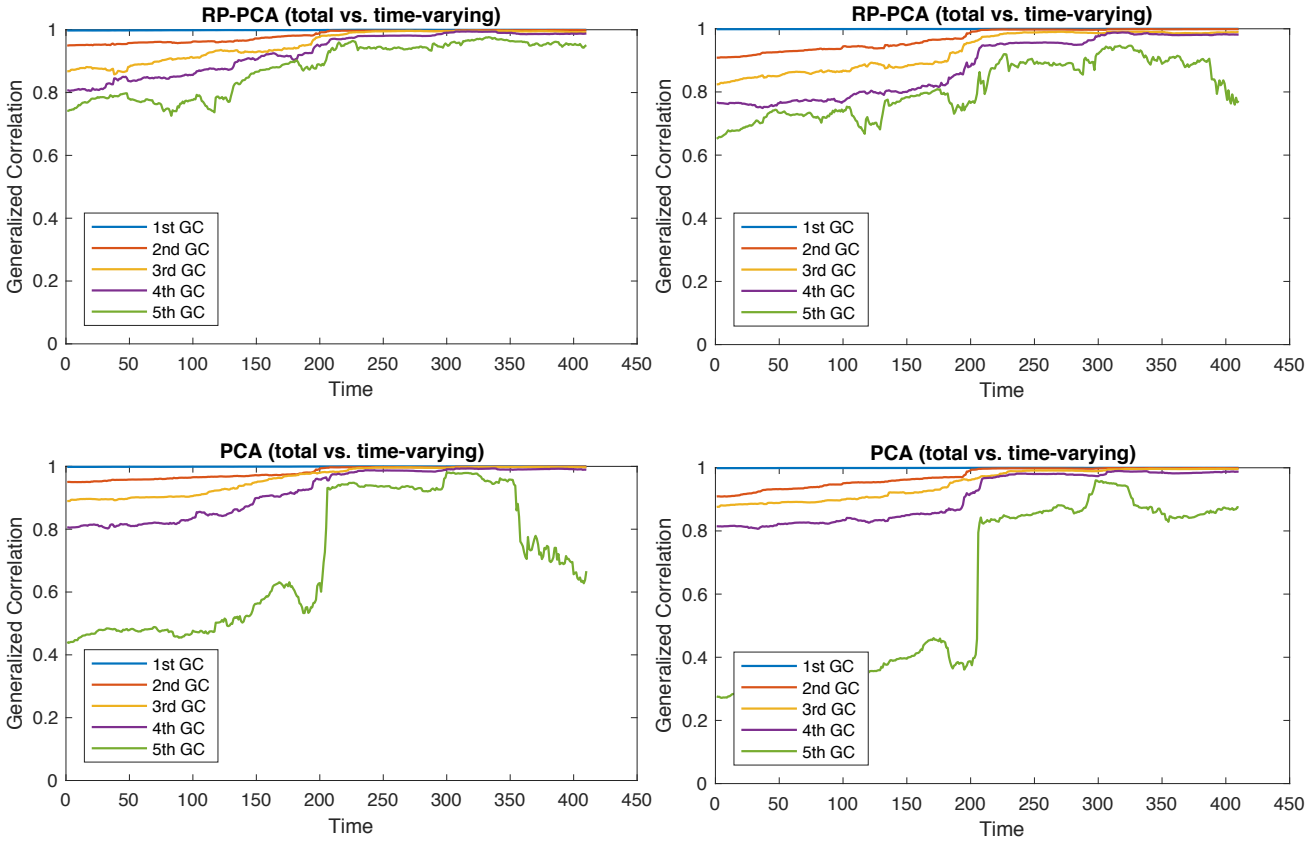
Note: Top: First and last decile of 37 single-sorted portfolios ($N = 74$ and $T = 638$). Bottom: Deciles of 37 single-sorted portfolios ($N = 370$ and $T = 638$). Generalized correlation of $K = 5$ statistical factors ($\gamma = 10$) and an increasing number of long-short anomaly factors. The first LS-factor is the market factor and LS-factors are added incrementally based on the largest accumulative absolute loading of the anomaly in the portfolio weights of the statistical factors.

Figure 17: Portfolio Weights in RP-PCA and PCA SDFs: Categories with Single Portfolios



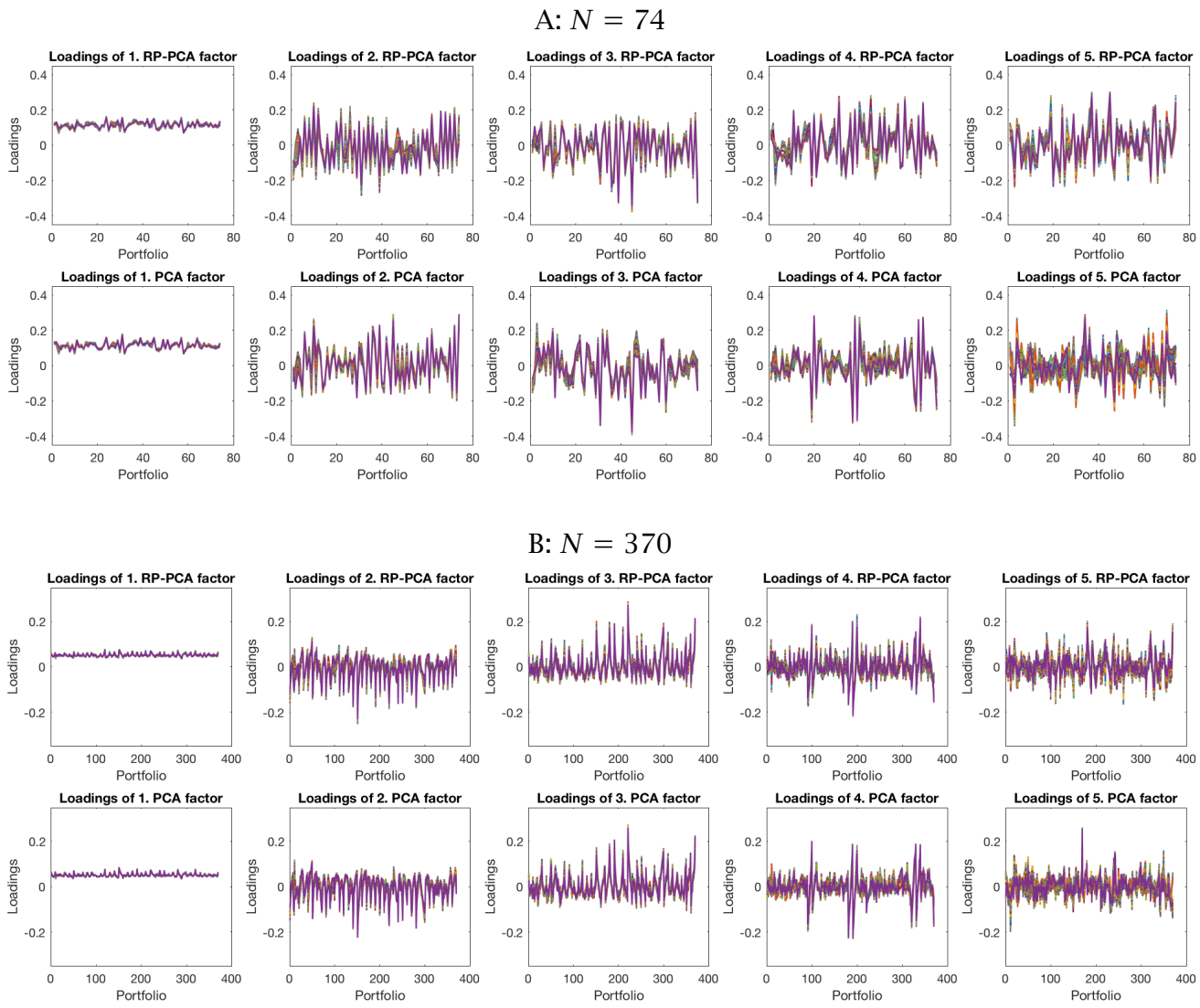
Note: The figure shows RP-PCA and PCA SDF weights for a sample with one portfolio per category. For each category we pick the the anomaly with the highest SR: value: *ep*, value interactions: *valprof*, momentum: *mom12*, reversal: *lrrev*, momentum/reversal: *indmomrev*, relative industry reversal: *indrrevlv*, growth: *inv*, profitability: *noa*. The sample also includes the eight anomalies that form their own categories: *size*, *ivol*, *accruals*, *ciss*, *lev*, *price*, *season* and *shvol*.

Figure 18: Generalized Correlations of Local Loading Estimates



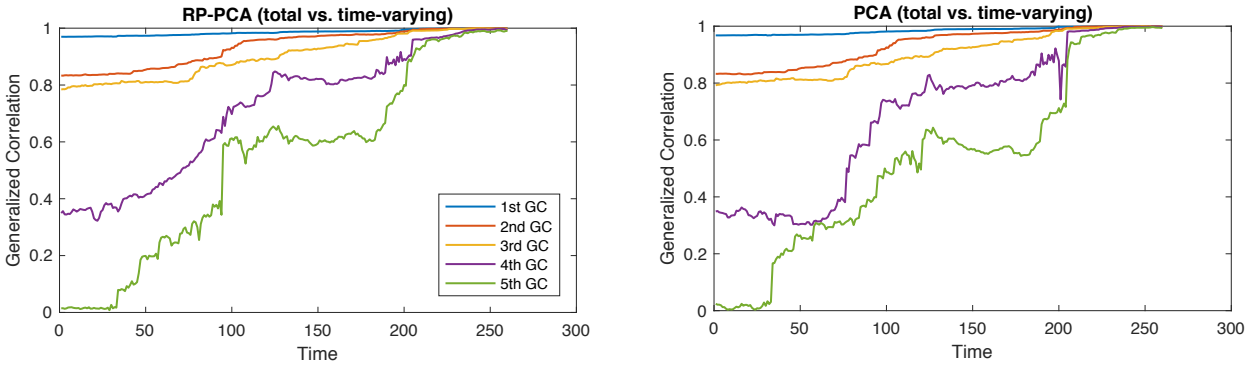
Note: Generalized correlations between loadings estimated on the whole time horizon $T = 650$ and a rolling window with 240 months for $K = 5$ factors. Left panel: First and last decile of 37 single-sorted portfolios ($N = 74$). Right panel: Deciles of 37 single-sorted portfolios ($N = 370$).

Figure 19: Rolling Loadings Estimates



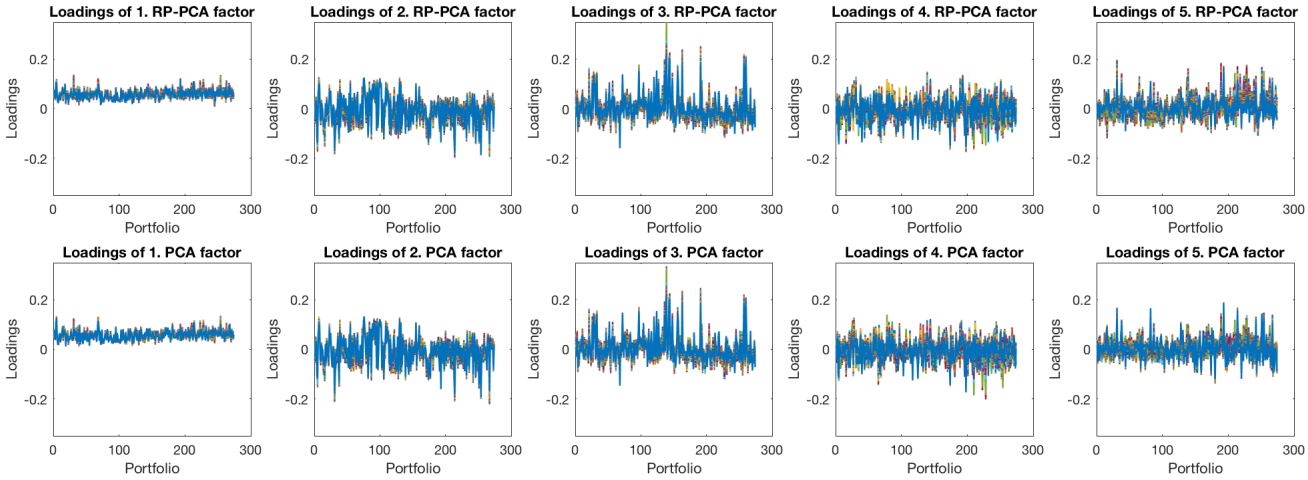
Note: Time-varying rotated loadings for the first five factors. Loadings are estimated on a rolling window with 240 months. Different lines corresponds to the loading estimated for different time windows. The rotation is calculated at each point in time to minimize the distance between time-varying loadings and those calculated over the whole sample. Top panel: Extreme deciles ($N = 74$). Bottom panel: All deciles ($N = 370$).

Figure 20: Generalized Correlations of Local Loading Estimates for Stocks



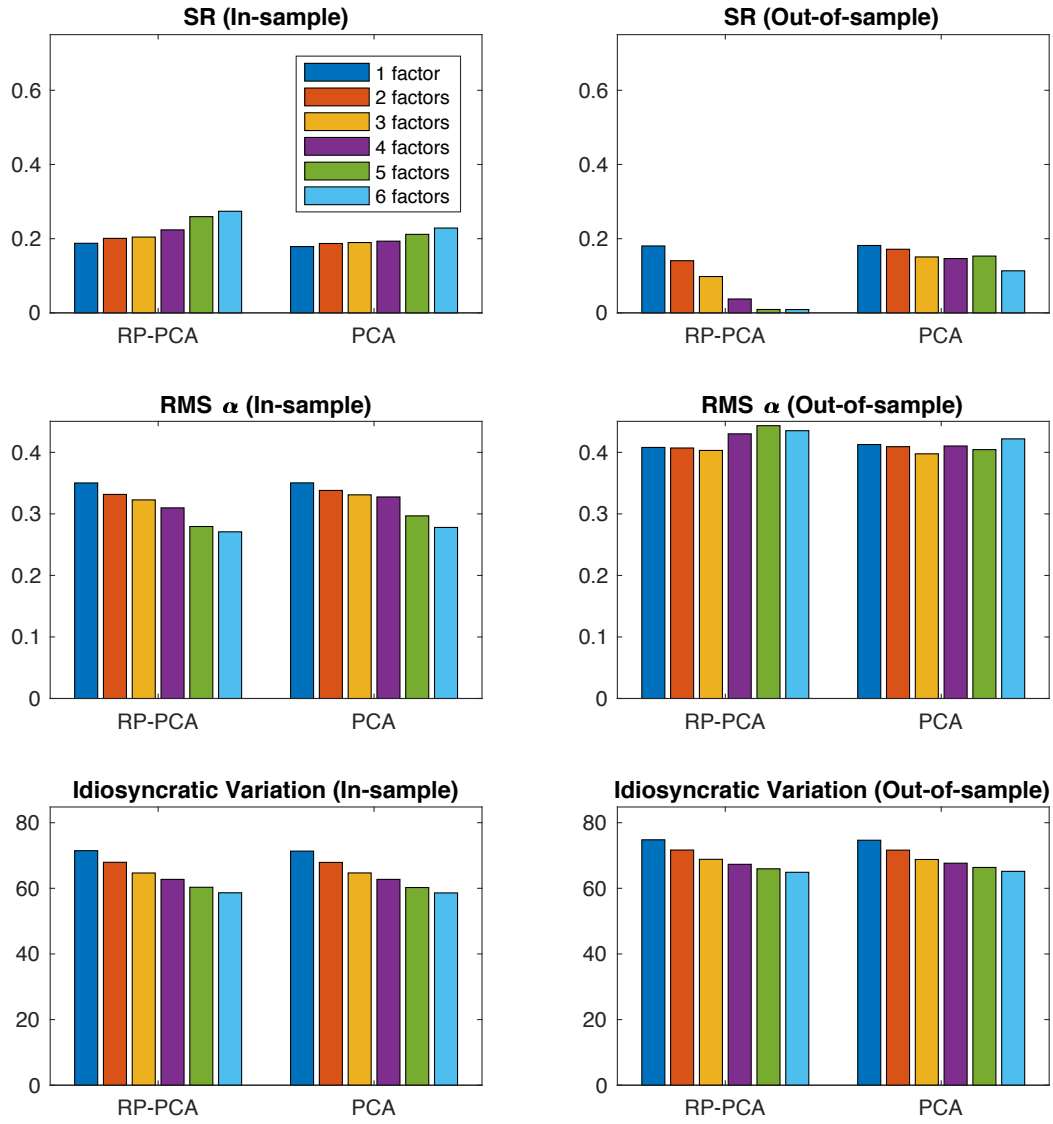
Note: Generalized correlations between loadings estimated on the whole time horizon $T = 500$ and a rolling window with 240 months for stock price data ($N = 270$ and $T = 500$)

Figure 21: Rolling Loadings Estimates for Stocks



Note: Time-varying rotated loadings for the first five factors. Loadings are estimated on a rolling window with 240 months. Different lines corresponds to the loading estimated for different time windows. The rotation is calculated at each point in time to minimize the distance between time-varying loadings and those calculated over the whole sample. Stock price data ($N = 270$ and $T = 500$)

Figure 22: RP-PCA vs. PCA Fit for Stocks



Note: Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for different number of factors. RP-weight $\gamma = 10$, Stock price data ($N = 270$ and $T = 500$)