NBER WORKING PAPER SERIES

ESTIMATING LATENT ASSET-PRICING FACTORS

Martin Lettau
Markus Pelger

Working Paper 24618
http://www.nber.org/papers/w24618

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2018, Revised June 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, UC Berkeley, UC Irvine, Zürich, Toronto, Boston University, Humboldt University, Ulm, Bonn, Frankfurt and the conference participants at the NBER-NSF Time-Series Conference, SoFiE, Western Mathematical Finance Conference and INFORMS for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Estimating Latent Asset-Pricing Factors
Martin Lettau and Markus Pelger NBER
Working Paper No. 24618
May 2018, Revised June 2018
JEL No. C14,C38,C52,G12

ABSTRACT

We develop an estimator for latent factors in a large-dimensional panel of financial data that can explain expected excess returns. Statistical factor analysis based on Principal Component Analysis (PCA) has problems identifying factors with a small variance that are important for asset pricing. We generalize PCA with a penalty term accounting for the pricing error in expected returns. Our estimator searches for factors that can explain both the expected return and covariance structure. We derive the statistical properties of the new estimator and show that our estimator can find asset-pricing factors, which cannot be detected with PCA, even if a large amount of data is available. Applying the approach to portfolio data we find factors with Sharpe-ratios more than twice as large as those based on conventional PCA and with significantly smaller pricing errors.

Martin Lettau
Haas School of Business
University of California, Berkeley
545 Student Services Bldg. #1900
Berkeley, CA 94720-1900
and NBER
lettau@haas.berkeley.edu

Markus Pelger
312 Huang Engineering Center
Department of Management Science
& Engineering
Stanford University
Stanford, CA 94305
mpelger@stanford.edu

1. Introduction

Approximate factor models have been a heavily researched topic in finance and macroeconomics in the last years (see Bai and Ng (2008), Stock and Watson (2006) and Ludvigson and Ng (2010)). The most popular technique to estimate latent factors is Principal Component Analysis (PCA) of a covariance or correlation matrix. It estimates factors that can best explain the co-movement in the data. A situation that is often encountered in practice is that the explanatory power of the factors is weak relative to idiosyncratic noise. In this case conventional PCA performs poorly (see Onatski (2012)). In some cases economic theory also imposes structure on the first moments of the data. Including this additional information in the estimation turns out to significantly improve the estimation of latent factors, in particular for those factors with a weak explanatory power in the variance.

We suggest a new statistical method to find the most important factors for explaining the variation and the mean in a large dimensional panel. Our key application are asset pricing factors. The fundamental insight of asset pricing theory is that the cross-section of expected returns should be explained by exposure to systematic risk factors.[1] Hence, asset pricing factors should simultaneously explain time-series covariation as well as the cross-section of mean returns. Finding the "right" risk factors is not only the central question in asset pricing but also crucial for optimal portfolio and risk management.[2] Traditional PCA methods based on the covariance or correlation matrices identify factors that capture only common time-series variation but do not take the cross-sectional explanatory power of factors into account.[3] We generalize PCA by including a penalty term to account for the pricing errors in the means. Hence, our estimator Risk-Premium PCA (RP-PCA) directly includes the object of interest, which is explaining the cross-section of expected returns, in the estimation. It turns out, that even if the goal is to explain the covariation and not the mean, the additional information in the mean can improve the estimation significantly.

This paper develops the asymptotic inferential theory for our estimator under a general approximate factor model and shows that it dominates conventional estimation based on PCA if there is information in the mean. We distinguish between strong and weak factors in our model. Strong factors essentially affect all underlying assets. The market-wide return is an example of a strong factor in asset pricing applications. RP-PCA can estimate these factors more efficiently than PCA as it efficiently combines information in first and second moments of the data. Weak factors affect only a subset of the underlying assets and are harder to detect. Many asset-pricing factors fall into this category. 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.

We build upon the econometrics literature devoted to estimating factors from large dimensional panel data sets. The general case of a static large dimensional factor model is treated in Bai (2003)

[1] Arbitrage pricing theory (APT) formalized by Ross (1976) and Chamberlain and Rothschild (1983) states that in an approximate factor model only systematic factors carry a risk-premium and explain the expected returns of diversified portfolios. Hence, factors that explain the covariance structure must also explain the expected returns in the cross-section.

[2] 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) in his presidential address this leads to the crucial questions, which risk factors are really important and which factors are subsumed by others.

[3] PCA has been used to find asset pricing factors among others by Connor and Korajczyk (1988), Connor and Korajczyk (1993) and Kozak et al. (2017). Kelly et al. (2017) and Fan et al. (2016) apply PCA to projected portfolios.

and Bai and Ng (2002). Forni et al. (2000) introduce the dynamic principal component method. Fan et al. (2013) study an approximate factor structure with sparsity. Aït-Sahalia and Xiu (2017) and Pelger (2017) extend the large dimensional factor model to high-frequency data. All these methods assume a strong factor structure that is estimated with some version of PCA without taking into account the information in expected returns, which results in a loss of efficiency. We generalize the framework of Bai (2003) to include the pricing error penalty and show that it only effects the asymptotic distribution of the estimates but not consistency.

Onatski (2012) studies principal component estimation of large factor models with weak factors. He shows that if a factor does not explain a sufficient amount of the variation in the data, it cannot be detected with PCA. We provide a solution to this problem that renders weak factors with high Sharpe-ratios detectable. Our statistical model extends the spiked covariance model from random matrix theory used in Onatski (2012) and Benaych-Georges and Nadakuditi (2011) to include the pricing error penalty. We show that including the information in the mean leads to larger systematic eigenvalues of the factors, which reduces the bias in the factor estimation and makes weak factors detectable. The derivation of our results is challenging as we cannot make the standard assumption that the mean of the stochastic processes is zero. As many asset pricing factors can be characterized as weak, our estimation approach becomes particularly relevant.

Our work is part of the emerging econometrics literature that combines latent factor extraction with a form of regularization. Bai and Ng (2017) develop the statistical theory for robust principal components. Their estimator can be understood as performing iterative ridge instead of least squares regressions, which shrinks the eigenvalues of the common components to zero. They combine their shrinked estimates with a clean-up step that sets the small eigenvalues to zero. Their estimates have less variation at the cost of a bias. Our approach also includes a penalty which in contrast is based on economic information and does not create a bias-variance trade-off. The objective of finding factors that can explain co-movements and the cross-section of expected returns simultaneously is based on the fundamental insight of arbitrage pricing theory. 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 estimator depends on a tuning parameter that trades off the information in the variance and the mean in the data. Our statistical theory provides guidance on the optimal choice of the tuning parameter that we confirm in simulations and in the data.

Our work is closely related to the paper by Fan and Zhong (2018) which allows estimating latent factors based on an over-identifying set of moments. We combine the first and second moments to estimate factors while their approach allows the inclusion of additional moments. Their analysis is based on a generalized method of moment approach under the assumption of a finite cross-section. Our strong factor model formulation can be similarly related to a general method of moment problem. We consider a large number of assets and include the additional perspective of a weak factor model which we think is particularly relevant in the context of asset pricing factors.

We apply our methodology to monthly returns of 370 decile sorted portfolios based on relevant financial anomalies for 55 years. We find that five factors can explain very well these expected returns and strongly outperforms PCA-based factors. The maximum Sharpe-ratio of our five factors is more

than twice as large as those based on PCA; a result that holds in- and out-of-sample. The pricing errors out-of-sample are sizably smaller. Our method captures the pricing information better while explaining the same amount of variation and co-movement in the data. Our companion paper Lettau and Pelger (2018) provides a more in-depth empirical analysis of asset-pricing factors estimated with our approach.

The rest of the paper is organized as follows. In Section 2 we introduce the model and provide an intuition for our estimators. Section 3 discusses the formal objective function that defines our estimator. Section 4 provides the inferential theory for strong factors, while 5 presents the asymptotic theory for weak factors. Section 6 provides Monte Carlo simulations demonstrating the finite-sample performance of our estimator. In Section 7 we study the factor structure in a large equity data set. Section 8 concludes. The appendix contains the proofs.

2. 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 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[4] return of N assets over T time periods:

$$X_{t,i} = F_t \Lambda_i^\top + e_{t,i} \qquad i = 1, ..., N \;\; t = 1, ..., T.$$

In matrix notation this reads as

$$\underset{T\times N}{\underbrace{X}} = \underset{T\times K}{\underbrace{F}} \; \underset{K\times N}{\underbrace{\Lambda^\top}} + \underset{T\times N}{\underbrace{e}} .$$

Our goal is to estimate the unknown latent factors F and the loadings Λ. 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 for them jointly going to infinity.

Assume that the factors and residuals are uncorrelated. This implies that the covariance matrix of the returns consists of a systematic and idiosyncratic part:

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

Under standard assumptions the largest eigenvalues of $\text{Var}(X)$ are driven by the factors. This motivates Principal Component Analysis (PCA) as an estimator for the loadings and factors. Essentially all estimators for latent factors only utilize the information contained in the second moment, but ignore information that is contained in the first moment.

Arbitrage-Pricing Theory (APT) has a second implication: The expected excess return is explained by the exposure to the risk factors multiplied by the risk-premium of the factors. If the factors are

[4] Excess returns equal returns minus the risk-free rate.

excess returns APT implies

$$E[X_i] = \Lambda_i E[F].$$

Here 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 constructed by PCA explain as much common time-series variation as possible. Conventional statistical factor analysis applies PCA to the sample covariance matrix $\frac{1}{T}X^\top X - \bar{X}\bar{X}^\top$ where $\bar{X}$ denotes the sample mean of excess returns. The eigenvectors of the largest eigenvalues are proportional to the loadings $\hat{\Lambda}^{\text{PCA}}$. Factors are obtained from a regression on the estimated loadings. It can be shown that conventional PCA factor estimates are based on the time-series variation objective function: [5]

$$\min_{\Lambda,F} \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T}(X_{ti} - F_t\Lambda_i^\top)^2$$

We call our approach Risk-Premium-PCA (RP-PCA). It applies PCA to a covariance matrix with overweighted mean

$$\frac{1}{T}X^\top X + \gamma \bar{X}\bar{X}^\top$$

with the risk-premium weight γ. The eigenvectors of the largest eigenvalues are proportional to the loadings $\hat{\Lambda}^{\text{RP-PCA}}$. We show that RP-PCA minimizes jointly the unexplained variation and pricing error:

$$\min_{\Lambda,F} \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{ti} - F_t\Lambda_i^\top)^2}_{\text{unexplained variation}} + \gamma \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{X}_i - \bar{F}\Lambda_i^\top\right)^2}_{\text{pricing error}},$$

where $\bar{F}$ denotes the sample mean of the factors. Factors are estimated by a regression of the returns on the estimated loadings, i.e. $\hat{F} = X\hat{\Lambda}\left(\hat{\Lambda}^\top\hat{\Lambda}\right)^{-1}$.

We develop the statistical theory that provides guidance on the optimal choice of the key parameter γ. 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 very large number of assets and/or because they have very large variances themselves. In a weak factor model the factors' signals are weak and the resulting eigenvalues are large compared to the idiosyncratic spectrum, but they do not explode.[6] In both cases it is always optimal to choose $\gamma \neq -1$, i.e. it is better to use our estimator instead of PCA applied to the covariance matrix. In a strong factor model,

[5] The variation objective function assumes that the data has been demeaned.

[6] Arbitrage-Pricing Theory developed by Chamberlain and Rothschild (1983) assumes that only strong factors are non-diversifiable and explain the cross-section of expected returns. As pointed out by Onatski (2012), a weak factors can be

the estimates become more efficient. In a weak factor model it strengthens the signal of the weak factors, which could otherwise not be detected. Depending on which framework is more 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 too large γ results in only minor efficiency losses. On the other hand a too small γ 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 the equity data that we have tested the first eigenvalue of the sample covariance matrix is very 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 might provide better guidance.

3. Objective Function

This section explains the relationship between our estimator and the objective function that is minimized. We introduce the following notation: $\mathbb{1}$ is a vector $T \times 1$ of 1's and thus $F^\top \mathbb{1}/T$ is the sample mean estimator of F. The projection matrix $M_\Lambda = I_N - \Lambda(\Lambda^\top\Lambda)^{-1}\Lambda^\top$ annihilates the $K-$dimensional vector space spanned by Λ. I_N and I_T denote the N- respectively T-dimensional identity matrix.

The objective function of conventional statistical factor analysis is to minimize the sum of squared errors for the cross-section and time dimension, i.e. the estimator $\hat{\Lambda}$ and $\hat{F}$ are chosen to minimize the unexplained variance. This variation objective function is

$$\min_{\Lambda,F} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{ti} - F_t \Lambda_i^\top)^2 = \min_{\Lambda} \frac{1}{NT} \operatorname{trace}(XM_\Lambda)^\top (XM_\Lambda)) \quad \text{s.t. } F = X(\Lambda^\top\Lambda)^{-1}\Lambda^\top.$$

The second formulation makes use of the fact that in a large panel data set the factors can be estimated by a regression of the assets on the loadings, $F = X(\Lambda^\top\Lambda)^{-1}\Lambda^\top$, and hence the residuals equal $X - F\Lambda^\top = XM_\Lambda$. This is equivalent to choosing $\hat{\Lambda}$ proportional to the eigenvectors of the first K largest eigenvalues of $\frac{1}{NT}X^\top X$.[7] In most applications the data is first demeaned, which means that the estimator applies PCA to the estimated covariance matrix of X. Thus $\hat{\Lambda}$ is proportional to the eigenvectors of the first K largest eigenvalues of $\frac{1}{NT}X^\top \left(I_T - \frac{\mathbb{1}\mathbb{1}^\top}{T}\right) X$.

Arbitrage-pricing theory predicts that the factors should price the cross-section of expected excess

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.

[7] Factor models are only identified up to invertible transformations. Therefore there is no loss of generality to assume that the loadings are orthonormal vectors and that the inner product of factors is a diagonal matrix.

returns. This yields a pricing objective function which minimizes the cross-sectional pricing error:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}X_i^\top \mathbb{1} - \frac{1}{T}F_i^\top \mathbb{1}\Lambda_i^\top\right)^2 = \frac{1}{N}\,\text{trace}\left(\left(\frac{1}{T}\mathbb{1}^\top X M_\Lambda\right)\left(\frac{1}{T}\mathbb{1}^\top X M_\Lambda\right)^\top\right).$$

We propose to combine these two objective functions with the risk-premium weight γ. The idea is to obtain statistical factors that explain the co-movement in the data and produce small pricing errors:

$$\begin{aligned}&\min_{\Lambda,F}\frac{1}{NT}\,\text{trace}\left(((XM_\Lambda)^\top(XM_\Lambda)\right) + \gamma\frac{1}{NT}\,\text{trace}\left(\left(\frac{1}{T}\mathbb{1}^\top X M_\Lambda\right)\left(\frac{1}{T}\mathbb{1}^\top X M_\Lambda\right)^\top\right)\\ =&\min_{\Lambda}\frac{1}{NT}\,\text{trace}\left(M_\Lambda X^\top\left(I+\frac{\gamma}{T}\mathbb{1}\mathbb{1}^\top\right)XM_\Lambda\right)\qquad \text{s.t. } F = X(\Lambda^\top\Lambda)^{-1}\Lambda^\top.\end{aligned}$$

Here we have made use of the linearity of the trace operator. The objective function is minimized by the eigenvectors of the largest eigenvalues of $\frac{1}{NT}X^\top\left(I_T+\frac{\gamma}{T}\mathbb{1}\mathbb{1}^\top\right)X$. Hence the factors and loadings can be obtained by applying PCA to this new matrix. The estimator for the loadings $\hat{\Lambda}$ are the eigenvectors of the first K eigenvalues of $\frac{1}{NT}X^\top\left(I_T+\frac{\gamma}{T}\mathbb{1}\mathbb{1}^\top\right)X$ multiplied by $\sqrt{N}$. $\hat{F}$ are $\frac{1}{N}X\hat{\Lambda}$. The estimator for the common component $C = F\Lambda$ is simply $\hat{C} = \hat{F}\hat{\Lambda}^\top$. The estimator simplifies to PCA of the covariance matrix for $\gamma = -1$.

In practice conventional PCA is often applied to the correlation instead of the covariance matrix. This implies that the returns are demeaned and normalized by their standard-deviation before applying PCA to their inner product. Hence, factors are chosen that explain most of the correlation instead of the variance. This approach is particularly appealing if the underlying panel data is measured in different units. Usually estimation based on the correlation matrix is more robust than based on the covariance matrix as it is less affected by a few outliers with very large variances. From a statistical perspective this is equivalent to applying a cross-sectional weighting matrix to the panel data. After applying PCA to the inner product, the inverse of the weighting matrix has to be applied to the estimated eigenvectors. The statistical rationale is that certain cross-sectional observations contain more information about the systematic risk than others and hence should obtain a larger weight in the statistical analysis. The standard deviation of each cross-sectional observation serves as a proxy for how large the noise is and therefore down-weighs very noisy observations.

Mathematically, a weighting matrix means that instead of minimizing equally weighted pricing errors we apply a weighting function Q to the cross-section resulting in the following weighted combined objective function:

$$\begin{aligned}\min_{\Lambda,F}\frac{1}{NT}\;&\text{trace}(Q^\top(X-F\Lambda^\top)^\top(X-F\Lambda^\top)Q)\\ &+\gamma\frac{1}{N}\,\text{trace}\left(\mathbb{1}^\top(X-F\Lambda^\top)QQ^\top(X-F\Lambda^\top)^\top\mathbb{1}\right)\\ =\min_{\Lambda}\;&\text{trace}\left(M_\Lambda Q^\top X^\top\left(I+\frac{\gamma}{T}\mathbb{1}\mathbb{1}^\top\right)XQM_\Lambda\right)\\ &\text{s.t. } F = X(\Lambda^\top\Lambda)^{-1}\Lambda^\top.\end{aligned}$$

Therefore factors and loadings can be estimated by applying PCA to $Q^\top X^\top \left(I + \frac{\gamma}{T}\mathbb{1}\mathbb{1}^\top\right) XQ$. In our empirical application we only consider the weighting matrix Q which is the inverse of a diagonal matrix of standard deviations of each return. For $\gamma = -1$ this corresponds to using a correlation matrix instead of a covariance matrix for PCA.

There are four different interpretations of RP-PCA:

(1) *Variation and pricing objective functions:* Our estimator combines a variation and 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.[8]

(2) *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.

(3) *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. This interpretation drives the results for the strong factor model in Section 4.

(4) *Signal-strengthening:* The matrix $\frac{1}{T}X^\top X + \gamma \bar{X}\bar{X}^\top$ should converge to[9]

$$\Lambda\left(\Sigma_F + (1+\gamma)\mu_F\mu_F^\top\right)\Lambda^\top + \mathrm{Var}(e),$$

where $\Sigma_F = \mathrm{Var}(F)$ denotes the covariance matrix of F and $\mu_F = E[F]$ 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. This interpretation is the basis for the weak factor model studied in Section 5.

4. Strong Factor Model

In a strong factor model RP-PCA provides a more efficient estimator of the loadings than PCA. Both RP-PCA and PCA provide consistent estimator for the loadings and factors. In the strong factor

[8] 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 it. Second, there is the problem of spurious factor detection (see e.g. Bryzgalova (2017)). 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.

[9] In this large-dimensional context the limit will be more complicated and studied in the subsequent sections.

model, the systematic factors are so strong that they lead to exploding eigenvalues. This is captured by the assumption that $\frac{1}{N}\Lambda^\top\Lambda \to \Sigma_\Lambda$ where Σ_Λ is a full-rank matrix.[10] This could be interpreted as the strong factors affecting an infinite number of assets.

The estimator for the loadings $\hat{\Lambda}$ are the eigenvectors of the first K eigenvalues of $\frac{1}{N}\left(\frac{1}{T}X^\top X + \gamma\bar{X}\bar{X}^\top\right)$ multiplied by $\sqrt{N}$. Up to rescaling the estimators are identical to those in the weak factor model setup. The estimator for the common component $C = F\Lambda$ is $\hat{C} = \hat{F}\hat{\Lambda}^\top$.

Bai (2003) shows that under Assumption 1 the PCA estimator of the loadings has the same asymptotic distribution as an OLS regression of the true factors F on X (up to a rotation). Similarly, the estimator for the factors behaves asymptotically like an OLS regression of the true loadings Λ on $X^\top$ (up to a rotation). Under slightly stronger assumptions we will show that the estimated loadings under RP-PCA have the same asymptotic distribution up to rotation as an OLS regression of WF on WX with $W^2 = \left(I_T + \gamma\frac{\mathbb{1}\mathbb{1}^\top}{T}\right)$. Surprisingly, estimated factors under RP-PCA and PCA have the same distribution.

Assumption 1 is identical to Assumptions A-G in Bai (2003) plus the additional assumption in E.4 that relates to $\frac{1}{\sqrt{T}}\sum_{t=1}^T e_{t,i}$. See Bai (2003) for a discussion of the assumptions. The correlation structure in the residuals can be more general in the strong model than in the weak model. This comes at the cost of larger values for the loading vectors. The residuals still need to satisfy a form of sparsity assumption restricting the dependence. The strong factor model provides a distribution theory which is based on a central limit theorem of the residuals. This is satisfied for relevant processes, e.g. ARMA models.

Assumption 1: Strong Factor Model

A: *Factors:* $E[\|F_t\|^4] \leq M < \infty$ *and* $\frac{1}{T}\sum_{t=1}^T F_t F_t^\top \xrightarrow{p} \Sigma_F$ *for some* $K \times K$ *positive definite matrix* Σ_F *and* $\frac{1}{T}\sum_{t=1}^T F_t \xrightarrow{p} \mu_F$.

B: *Factor loadings:* $\|\Lambda_i\| \leq \bar{\lambda} < \infty$, *and* $\|\Lambda^\top\Lambda/N - \Sigma_\lambda\| \to 0$ *for some* $K \times K$ *positive definite matrix* Σ_Λ.

C: *Time and cross-section dependence and heteroskedasticity: There exists a positive constant* $M < \infty$ *such that for all* N *and* T*:*

1. $E[e_{t,i}] = 0$, $E[|e_{t,i}|^8] \leq M$.
2. $E[N^{-1}\sum_{i=1}^N e_{s,i}e_{t,i}] = \gamma(s,t)$, $|\gamma(s,s)| \leq M$ *for all* s *and for every* $t \leq T$ *it holds* $\sum_{s=1}^T |\gamma(s,t)| \leq M$
3. $E[e_{t,i}e_{t,j}] = \tau_{ij,t}$ *with* $|\tau_{ij,t}| \leq |\tau_{ij}|$ *for some* τ_{ij} *and for all* t *and for every* $i \leq N$ *it holds* $\sum_{i=1}^N |\tau_{ij}| \leq M$.
4. $E[e_{t,i}e_{s,j}] = \tau_{ij,ts}$ *and* $(NT)^{-1}\sum_{i=1}^N\sum_{j=1}\sum_{t=1}^T\sum_{s=1}^T |\tau_{ij,st}| \leq M$.
5. *For every* (t,s), $E\left[|N^{-1/2}\sum_{i=1}^N(e_{s,i}e_{t,i}) - E[e_{s,t}e_{t,i}]|^4\right] \leq M$.

[10] In latent factor models only the product $F\Lambda$ is identified. Hence without loss of generality we will normalize Σ_Λ to the identity matrix I_K and assume that the factors are uncorrelated.

D: *Weak dependence between factors and idiosyncratic errors:*
$E\left[\frac{1}{N}\sum_{i=1}^{N}\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t e_{t,i}\|^2\right] \leq M.$

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

1. *For each t, $E\left[\left\|\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{k=1}^{N}F_s(e_{s,k}e_{t,k} - E[e_{s,k}e_{t,k})]\right\|^2\right] \leq M$*
2. *The $K \times K$ matrix satisfies $E\left[\|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}F_t\Lambda_i^\top e_{t,i}\|^2\right] \leq M$*
3. *For each t as $N \to \infty$:*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Lambda_i e_{t,i} \xrightarrow{d} N(0,\Gamma_t),$$

where $\Gamma_t = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\Lambda_i\Lambda_j^\top E[e_{t,i}e_{t,j}]$

4. *For each i as $T \to \infty$:*

$$\begin{pmatrix}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t e_{t,i}\\ \frac{1}{\sqrt{T}}\sum_{t=1}^{T}e_{t,i}\end{pmatrix} \xrightarrow{D} N(0,\Omega_i) \qquad \Omega_i = \begin{pmatrix}\Omega_{11,i} & \Omega_{12,i}\\ \Omega_{21,i} & \Omega_{22,i}\end{pmatrix}$$

where $\Omega_i = p\lim_{T\to\infty}\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E\left[\begin{pmatrix}F_t F_s^\top e_{s,i}e_{t,i} & F_t e_{s,i}e_{t,i}\\ F_s^\top e_{s,i}e_{t,i} & e_{s,i}e_{t,i}\end{pmatrix}\right]$.

F: *The eigenvalues of the $K \times K$ matrix $\Sigma_\Lambda\Sigma_F$ are distinct.*

Theorem 1 provides a complete inferential theory for the strong factor model.

Theorem 1: Asymptotic distribution in strong factor model
Assume Assumption 1 holds. Then:

1. *If $\min(N,T) \to \infty$, then for any $\gamma \in [-1,\infty)$ the factors and loadings can be estimated consistently pointwise.*

2. *If $\frac{\sqrt{T}}{N} \to 0$, then the asymptotic distribution of the loadings estimator is given by*

$$\sqrt{T}\left(H^\top\hat{\Lambda}_i - \Lambda_i\right) \xrightarrow{D} N(0,\Phi_i)$$

$$\Phi_i = \left(\Sigma_F + (\gamma+1)\mu_F\mu_F^\top\right)^{-1}\left(\Omega_{11,i} + \gamma\mu_F\Omega_{21,i} + \gamma\Omega_{12,i}\mu_F + \gamma^2\mu_F\Omega_{22,i}\mu_F\right)\left(\Sigma_F + (\gamma+1)\mu_F\mu_F^\top\right)^{-1}$$
$$H = \left(\frac{1}{T}F^\top W^2 F\right)\left(\frac{1}{N}\Lambda\hat{\Lambda}\right)V_{TN}^{-1}$$

and V_{TN} is a diagonal matrix of the largest K eigenvalues of $\frac{1}{NT}X^\top W^2 X$ and $W^2 = \left(I_T + \gamma\frac{\mathbb{1}\mathbb{1}^\top}{T}\right)$. For $\gamma = -1$ this simplifies to the conventional case $\Sigma_F^{-1}\Omega_{11,i}\Sigma_F^{-1}$.

3. *If $\frac{\sqrt{N}}{T} \to 0$, then the asymptotic distribution of the factors is not affected by the choice of γ.*

4. *For any choice of $\gamma \in [-1,\infty)$ the common components can be estimated consistently if*

$\min(N,T) \to \infty$. *The asymptotic distribution of the common component depends on* γ *if and only if* T/N *does not go to zero. For* $T/N \to 0$

$$\sqrt{T}\left(\hat{C}_{t,i} - C_{t,i}\right) \xrightarrow{D} N\left(0, F_t^\top \Phi_i F_t\right).$$

Note that Bai (2003) characterizes the distribution of $\sqrt{T}\left(\Lambda_i - H^{\top -1}\hat{\Lambda}_i\right)$, while we rotate the estimated loadings $\sqrt{T}\left(H^\top \hat{\Lambda}_i - \Lambda_i\right)$. Our rotated estimators are directly comparable for different choices of γ. The proof of the theorem is essentially identical to the arguments of Bai (2003). The key argument is based on an asymptotic expansion. Under Assumption 1 we can show that the following expansions hold

1. $\sqrt{T}\left(H^\top \hat{\Lambda}_i - \Lambda_i\right) = \left(\frac{1}{T}F^\top W^2 F\right)^{-1} \frac{1}{\sqrt{T}} F^\top W^2 e_i + O_p\left(\frac{\sqrt{T}}{N}\right) + o_p(1)$
2. $\sqrt{N}\left(H^{\top -1}\hat{F}_t - F_t\right) = \left(\frac{1}{N}\Lambda^\top \Lambda\right)^{-1} \frac{1}{\sqrt{N}} \Lambda^\top e_t^\top + O_p\left(\frac{\sqrt{N}}{T}\right) + o_p(1)$
3. $\sqrt{\delta}\left(\hat{C}_{t,i} - C_{t,i}\right) = \frac{\sqrt{\delta}}{\sqrt{T}} F_t^\top \left(\frac{1}{T}F^\top W^2 F\right)^{-1} \frac{1}{\sqrt{T}} F^\top W^2 e_i + \frac{\sqrt{\delta}}{\sqrt{N}} \Lambda_i^\top \left(\frac{1}{N}\Lambda^\top \Lambda\right)^{-1} \frac{1}{\sqrt{N}} \Lambda^\top e_t^\top + o_p(1)$
   with $\delta = \min(N,T)$.

We just need to replace the factors and asset space by their projected counterpart WF and WX in Bai's (2003) proofs. Conventional PCA, i.e. $\gamma = -1$ is a special case of our result, which typically leads to inefficient estimation.

Lemma 1: *If* $\mu_F \neq 0$*, then it is not efficient to use the covariance matrix for estimating the loadings and common components, i.e. the choice of* $\gamma = -1$ *does not lead to the smallest asymptotic covariance matrix for the loadings and common components.*

In order to get a better intuition we consider an example with i.i.d. residuals over time. This simplified model will be more comparable to the weak factor model in the next section.

Example 1: Simplified Strong Factor Model

1. **Rate:** *Assume that* $\frac{N}{T} \to c$ *with* $0 < c < \infty$.
2. **Factors:** *The factors* F *are uncorrelated among each other and are independent of* e *and* Λ *and have bounded first two moments.*

$$\hat{\mu}_F := \frac{1}{T}\sum_{t=1}^{T} F_t \xrightarrow{p} \mu_F \qquad \hat{\Sigma}_F := \frac{1}{T} F_t F_t^\top \xrightarrow{p} \Sigma_F = \begin{pmatrix} \sigma_{F_1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{F_K}^2 \end{pmatrix}.$$

3. **Loadings:** $\Lambda^\top \Lambda / N \xrightarrow{p} I_K$ *and all loadings are bounded. The loadings are independent of the factors and residuals.*

4. **Residuals:** *Residual matrix can be represented as* $e = \epsilon\Sigma$ *with* $\epsilon_{t,i} \overset{i.i.d.}{\sim} N(0,1)$. *All elements and all row sums of* Σ *are bounded.*

Corollary 1: Simplified Strong Factor Model:
The assumptions of example 1 hold. The factors and loadings can be estimated consistently. The asymptotic distribution of the factors is not affected by γ. *The asymptotic distribution of the loadings is given by*

$$\sqrt{T}\left(H^\top\hat{\Lambda}_i - \Lambda_i\right) \xrightarrow{D} N(0,\Omega_i),$$

where $E[e_{t,i}^2] = \sigma_{e_i}^2$ *and*

$$\Omega_i = \sigma_{e_i}^2\left(\Sigma_F + (1+\gamma)\mu_F\mu_F^\top\right)^{-1}\left(\Sigma_F + (1+\gamma)^2\mu_F\mu_F^\top\right)\left(\Sigma_F + (1+\gamma)\mu_F\mu_F^\top\right)^{-1}.$$

The optimal choice for the weight minimizing the asymptotic variance is $\gamma = 0$. *Choosing* $\gamma = -1$, *i.e. the covariance matrix for factor estimation, is not efficient.*

The estimator in the strong factor model can be formulated as a GMM problem. Up to a remainder term that vanishes under appropriate rate conditions the loading estimator is given by

$$H^\top\hat{\Lambda}_i = \left(F^\top W^2 F\right)^{-1} F^\top W^2 X_i.$$

This is equivalent to combining the OLS and the pricing moment conditions with a weight γ. More specifically, we define the following $K+1$ population and sample moments

$$G(\Lambda_i) = E\left[\begin{pmatrix}(X_{t,i} - F_t\Lambda_i^\top)F_t\left(E[F_tF_t^\top]\right)^{-1/2}\\ E[X_i - F_t\Lambda_i^\top]\end{pmatrix}\right] \qquad \hat{G}(\Lambda_i) = \begin{pmatrix}\frac{1}{\sqrt{T}}\left(X_i - F\Lambda_i^\top\right)^\top F\left(F^\top F\right)^{-1/2}\\ \frac{1}{T}\left(X_i - F\Lambda_i^\top\right)^\top \mathbb{1}\end{pmatrix}$$

The first K moments are identical to the OLS first order condition of a regression of X on F. The last moment is the APT pricing moment equation. The GMM estimator

$$\operatorname{argmin}\ \hat{G}^\top\begin{pmatrix}I_K & 0\\ 0 & \gamma\end{pmatrix}\hat{G}$$

has the solution $H^\top\hat{\Lambda}_i$.

5. Weak Factor Model

If factors are weak rather than strong RP-PCA can detect factors that are not estimated by conventional PCA. Weak factors affect only a smaller fraction of the assets. After normalizing the loadings, a weak factor can be interpreted as having a small variance. If the variance of a weak factor is below a critical value, 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 their variance is below the critical detection value. Weak factors can only be estimated with a bias but the bias will generally be smaller for RP-PCA than for PCA.

In a weak factor model $\Lambda^\top\Lambda$ is bounded in contrast to a strong factor model in which $\frac{1}{N}\Lambda^\top\Lambda$ is bounded. The statistical model for analyzing weak factor models is based on spiked covariance models from random matrix theory. It is well-known that under the assumptions of random matrix 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.[11]

The estimator of the loadings $\hat{\Lambda}$ are the first K eigenvectors of $\frac{1}{T}X^\top X + \gamma\bar{X}\bar{X}^\top$. Conventional PCA of the sample covariance matrix corresponds to $\gamma = -1$.[12] The estimators of the factors are the regression of the returns on the loadings, i.e. $\hat{F} = X\hat{\Lambda}$.

5.1. Assumptions

We impose the following assumptions on the approximate factor model:

Assumption 2: Weak Factor Model

A: **Rate:** *Assume that* $N/T \to c$ *with* $0 < c < \infty$.

B: **Factors:** *The factors* F *are uncorrelated among each other and are independent of* e *and* Λ *and have bounded first two moments.*

$$\hat{\mu}_F := \frac{1}{T}\sum_{t=1}^{T} F_t \xrightarrow{p} \mu_F \qquad \hat{\Sigma}_F := \frac{1}{T}F_t F_t^\top \xrightarrow{p} \Sigma_F = \begin{pmatrix} \sigma_{F_1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{F_K}^2 \end{pmatrix}.$$

C: **Loadings:** $\Lambda^\top\Lambda \xrightarrow{p} I_K$ *and the column vectors of the loadings* Λ *are orthogonally invariant (e.g.* $\Lambda_{i,k} \sim N(0, 1/N)$ *and independent of the factors and residuals.*

D: **Residuals:** *The empirical eigenvalue distribution function of* Σ *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* Σ *converge to* b.

[11] 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, RP-PCA crucially depends on using non-zero means of random variables. Hence, we need to develop new arguments to overcome this problem.

[12] The properties of weak factor models based on covariances have already been studied in Onatski (2012), Paul (2007) and Benaych-Georges and Nadakuditi (2011). We replicate those results applied to our setup. They will serve as a benchmark for the more complex risk-premium estimator.

Assumption 2.C can be interpreted as considering only well-diversified portfolios as factors. It essentially assumes that the portfolio weights of the factors are random with a variance of $1/N$. The orthogonally invariance assumption on the loading vectors is satisfied if for example $\Lambda_{i,k} \overset{i.i.d.}{\sim} N(0, 1/N)$. This is certainly a stylized assumption, but it allows us to derive closed-form solutions that are easily interpretable.[13] Assumption 2.D is a standard assumption in random matrix theory.[14] The assumption allows for non-trivial weak cross-sectional correlation in the residuals, but excludes serial-correlation. It implies clustering of the largest eigenvalues of the population covariance matrix of the residuals and rules out that a few linear combinations of idiosyncratic terms have an unusually large variation which could not be separated from the factors. It can be weakened as in Onatski (2012) when considering estimation based on the covariance matrix. However, when including the risk-premium in the estimation it seems that the stronger assumption is required. Many relevant cross-sectional correlation structures are captured by this assumption e.g. sparse correlation matrices or an ARMA-type dependence.

5.2. Asymptotic Results

In order to state the results for the weak factor model, we need to define several well-known objects from random matrix theory. We define the average idiosyncratic noise as $\sigma_e^2 := \text{trace}(\Sigma)/N$, which is the average of the eigenvalues of Σ. If the residuals are i.i.d. distributed σ_e^2 would simply be their variance. Our estimator will depend strongly on the dependency structure of the residual covariance matrix which can be captured by their eigenvalues. Denote by $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$ the ordered eigenvalues of $\frac{1}{T}e^\top e$. The Cauchy transform (also called Stieltjes transform) of the eigenvalues is the almost-sure limit:

$$G(z) = a.s. \lim_{T\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{z-\lambda_i} = a.s. \lim_{T\to\infty} \frac{1}{N} \text{trace}\left((zI_N - \frac{1}{T}e^\top e)\right)^{-1}.$$

This function is well-defined for z outside the support of the eigenvalues. This Cauchy transform is a well-understood object in random matrix theory. For simple cases analytical solutions exist and for general Σ it can easily be simulated or estimated from the data.

A second important transformation of the residual eigenvalues is

$$B(z) = a.s. \lim_{T\to\infty} \frac{c}{N}\sum_{i=1}^{N} \frac{\lambda_i}{(z-\lambda_i)^2} = a.s. \lim_{T\to\infty} \frac{c}{N} \text{trace}\left(\left((zI_N - \frac{1}{T}e^\top e)\right)^{-2}\left(\frac{1}{T}e^\top e\right)\right).$$

The function $B(z)$ is proportional to the derivative of $G(z)$. For special cases a closed-form solution is available and for the general case it can be easily estimated.

The crucial tool for understanding RP-PCA is the concept of a "signal matrix" M. The signal matrix essentially represents the largest true eigenvalues. For PCA estimation based on the sample covari-

[13] Onatski (2012) does not impose orthogonally invariant loadings, but requires the loadings to be the eigenvectors of $\frac{1}{T}e^\top e$. In order to make progress we need to impose some kind of assumption that allows us to diagonalize the residual covariance matrix without changing the structure of the systematic part.

[14] Similar assumptions have been imposed in Onatski (2010), Onatski (2012), Harding (2013) and Ahn and Horenstein (2013).

ance matrix the signal matrix M_{PCA} equals:

$$M_{\text{PCA}} = \Sigma_F + c\sigma_e^2 I_K = \begin{pmatrix} \sigma_{F_1}^2 + c\sigma_e^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{F_K}^2 + c\sigma_e^2 \end{pmatrix}$$

and the "signals" are the K largest eigenvalues $\theta_1^{\text{PCA}}, .., \theta_K^{\text{PCA}}$ of this matrix. The "signal matrix" for RP-PCA $M_{\text{RP-PCA}}$ is defined as

$$M_{\text{RP-PCA}} = \begin{pmatrix} \Sigma_F + c\sigma_e^2 & \Sigma_F^{1/2}\mu_F(1+\tilde{\gamma}) \\ \mu_F^\top \Sigma_F^{1/2}(1+\tilde{\gamma}) & (1+\gamma)(\mu_F^\top \mu_F + c\sigma_2^2) \end{pmatrix}.$$

We define $\tilde{\gamma} = \sqrt{\gamma+1} - 1$ and note that $(1+\tilde{\gamma})^2 = 1+\gamma$. The RP-PCA "signals" are the K largest eigenvalues $\theta_1^{\text{RP-PCA}}, .., \theta_K^{\text{RP-PCA}}$ of $M_{\text{RP-PCA}}$. Intuitively, the signal of the factors is driven by $\Sigma_F + (1+\gamma)\mu\mu^\top$, which has the same eigenvalues as

$$\begin{pmatrix} \Sigma_F & \Sigma_F^{1/2}\mu_F(1+\tilde{\gamma}) \\ \mu_F^\top \Sigma_F^{1/2}(1+\tilde{\gamma}) & (1+\gamma)(\mu_F^\top \mu_F) \end{pmatrix}.$$

This is disturbed by the average noise which adds the matrix $\begin{pmatrix} c\sigma_e^2 & 0 \\ 0 & (1+\gamma)c\sigma_e^2 \end{pmatrix}$. Note that the disturbance also depends on the parameter γ. We denote the corresponding orthonormal eigenvectors of M_{PCA} by $\tilde{U}$:

$$\tilde{U}^\top M_{\text{RP-PCA}} \tilde{U} = \begin{pmatrix} \theta_1^{\text{RP-PCA}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_{K+1}^{\text{RP-PCA}} \end{pmatrix}.$$

Unlike the conventional case of the covariance matrix with uncorrelated factors we cannot link the eigenvalues of $M_{\text{RP-PCA}}$ with specific factors. The rotation $\tilde{U}$ tells us how much the first eigenvalue contributes to the first K factors, etc..

Theorem 2: Risk-Premium PCA under weak factor model

Assume Assumption 2 holds. We denote by $\theta_1, ..., \theta_K$ the first K largest eigenvalues of the signal matrix $M = M_{PCA}$ or $M = M_{RP\text{-}PCA}$. The first K largest eigenvalues $\hat{\theta}_i$ $i = 1, ..., K$ of $\frac{1}{T}X^\top \left(I_T + \gamma \frac{\mathbb{1}\mathbb{1}^\top}{T}\right) X$ satisfy

$$\hat{\theta}_i \xrightarrow{p} \begin{cases} G^{-1}\left(\frac{1}{\theta_i}\right) & \text{if } \theta_i > \theta_{crit} = \lim_{z \downarrow b} \frac{1}{G(z)} \\ b & \text{otherwise.} \end{cases}$$

The correlation of the estimated with the true factors[15] *converges to*

$$\widehat{Corr}(F,\hat{F}) = \underbrace{\tilde{Q}}_{rotation} \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \rho_K \end{pmatrix} \underbrace{\tilde{R}}_{rotation}$$

with

$$\rho_i^2 \xrightarrow{p} \begin{cases} \frac{1}{1+\theta_i B(\hat{\theta}_i))} & \text{if } \theta_i > \theta_{crit} \\ 0 & \text{otherwise} \end{cases}$$

For $\theta_i > \theta_{crit}$ the correlation ρ_i is strictly increasing in θ_i. If $\mu_F \neq 0$, then for any $\gamma > -1$ RP-PCA has higher correlations ρ_i than PCA and RP-PCA strictly dominates PCA in terms of detecting factors, i.e. $\rho_i > 0$.

The rotation matrices satisfy $\tilde{Q}^\top \tilde{Q} \leq I_K$ and $\tilde{R}^\top \tilde{R} \leq I_K$. Hence, the correlation $\widehat{Corr}(F_i, \hat{F}_i)$ is not necessarily an increasing function in θ. For $\gamma > -1$ the rotation matrices equal:

$$\tilde{Q} = \begin{pmatrix} I_K & 0 \end{pmatrix} \tilde{U}_{1:K} \qquad \tilde{R} = D_K^{1/2} \hat{\Sigma}_{\hat{F}}^{-1/2},$$

where $\tilde{U}_{1:K}$ are the first K columns of $\tilde{U}$ and

$$\hat{\Sigma}_{\hat{F}} = D_K^{1/2} \left(\begin{pmatrix} \rho_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_K \\ 0 & \cdots & 0 \end{pmatrix}^\top \tilde{U}^\top \begin{pmatrix} I_K & 0 \\ 0 & 0 \end{pmatrix} \tilde{U} \begin{pmatrix} \rho_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_K \\ 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} 1-\rho_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1-\rho_K^2 \end{pmatrix} \right) D_K^{1/2}$$

$$D_K = diag\left(\begin{pmatrix} \hat{\theta}_1 & \cdots & \hat{\theta}_K \end{pmatrix} \right)$$

For PCA ($\gamma = -1$) the rotation matrices simplify to $\tilde{Q} = \tilde{R} = I_K$.

Theorem 2 states that the asymptotic behavior of the estimator can be completely explained by the signals of the factors for a given distribution of the idiosyncratic shocks. The theorem also states that weak factors can only be estimated with a bias. If a factor is too weak then it cannot be detected at all. Weak factors can always be better detected using Risk-Premium-PCA instead of covariance PCA. The phase transition phenomena that hides weak factors can be avoided by putting some weight on the information captured by the risk-premium. Based on our asymptotic theory, we can choose the optimal weight γ depending on our objective, e.g. to make all weak factors detectable or achieving the largest correlation for a specific factor. Typically the rotation matrices $\tilde{U}$ and $\tilde{V}$ are decreasing in γ while ρ_i is strictly increasing in γ, yielding an optimal value for the largest correlation.

[15] $\widehat{\text{Corr}}(F,\hat{F}) = \left(\frac{1}{T}F^\top\left(I - \frac{\mathbb{1}\mathbb{1}^\top}{T}\right)F\right)^{-1/2}\left(\frac{1}{T}F^\top\left(I - \frac{\mathbb{1}\mathbb{1}^\top}{T}\right)\hat{F}\right)\left(\frac{1}{T}\hat{F}^\top\left(I - \frac{\mathbb{1}\mathbb{1}^\top}{T}\right)\hat{F}\right)^{-1/2}$.

5.3. Examples

In order to obtain a better intuition for the problem we consider two special cases. First, we analyze the effect of γ in the case of only one factor. Second, we study PCA for the special case of cross-sectionally uncorrelated residuals.

Example 2: One-factor model

Assume that there is only one factor, i.e. $K = 1$. We introduce the following notation

- *Noise-to-signal ratio:* $\Gamma_e = \frac{c \cdot \sigma_e^2}{\sigma_F^2}$
- *Sharpe-ratio:* $SR = \frac{\mu_F}{\sigma_F}$.
- $\Phi(\theta_i) := B(\hat{\theta}_i(\theta_i))$.

The signal matrix M_{RP-PCA} simplifies to

$$M_{RP\text{-}PCA} = \sigma_F^2 \begin{pmatrix} 1+\Gamma_e & SR\sqrt{1+\gamma} \\ SR\sqrt{1+\gamma} & (SR^2+\Gamma_e)(1+\gamma) \end{pmatrix}$$

and has the largest eigenvalue:

$$\begin{aligned}\theta =& \frac{1}{2}\sigma_F^2(1+\Gamma_e+(SR^2+\Gamma_e)(1+\gamma) \\ &+\sqrt{(1+\Gamma_e+(SR^2+\Gamma_e)(1+\gamma))^2-4(1+\gamma)\Gamma_e(1+SR^2+\Gamma_e)}.\end{aligned}$$

Corollary 2: One-factor model

Assume Assumption 2 holds and $K = 1$. The correlation between the estimated and true factor has the following limit:

$$\widehat{Corr}(F,\hat{F})^2 \xrightarrow{p} \frac{1}{1+\theta\Psi(\theta)\left(\frac{\left(\frac{\theta}{\sigma_F^2}-(1+\Gamma_e)\right)^2}{SR^2(1+\gamma)}+1\right)}$$

and the estimated Sharpe-ratio converges to

$$\widehat{SR} \xrightarrow{p} \frac{\frac{\theta}{\sigma_F^2}-(1+\Gamma_e)}{SR(1+\gamma)}\widehat{Corr}(F,\hat{F}).$$

For $\gamma \to \infty$ these limits converge to

$$\begin{aligned}\widehat{Corr}(F,\hat{F})^2 &\xrightarrow{p} \frac{1}{1+\Gamma_e+\frac{\Gamma_e^2}{SR^2}} \\ \widehat{SR} &\xrightarrow{p} \left(SR+\frac{\Gamma_e}{SR}\right)\frac{1}{\sqrt{1+\Gamma_e+\frac{\Gamma_e^2}{SR^2}}}.\end{aligned}$$

In the case of PCA, i.e. $\gamma = -1$ *the expression simplifies to*

$$\widehat{Corr}(F, \hat{F})^2 \xrightarrow{p} \frac{1}{1 + \theta\Psi(\theta)}$$

with $\theta_{PCA} = \sigma_F^2(1 + \Gamma_e)$.

A smaller noise-to-signal ratio Γ_e and a larger Sharpe-ratio combined with a large γ lead to a more precise estimation of the factors. In the simulation section we find the optimal value of γ to maximize the correlation. Note that a larger value of γ decreases $\theta\Psi(\theta)$, while it increases $\frac{\left(\frac{\theta}{\sigma_F^2} - (1+\Gamma_e)\right)^2}{SR^2(1+\gamma)}$, creating a trade-off. In all our simulations $\gamma = -1$ was never optimal.

Now we study PCA for the special case of cross-sectionally uncorrelated residuals but many factors[16]

Example 3: PCA for model with independent residuals

Assume that $e_{t,i}$ *i.i.d.* $N(0, \sigma_e^2)$, *i.e.* $\Sigma = \sigma_e^2 I_N$. *In this case the residual eigenvalues follow the well-known Marcenko-Pasteur Law. For simplicity assume that* $\frac{N}{T} \to c$ *with* $c > 1$. *The results can be easily extended to the case* $0 < c < 1$.

The maximum residual eigenvalue equals $b = \sigma_e^2(1 + \sqrt{c})^2$. *The Cauchy transform takes the form*

$$G(z) = \frac{z - \sigma_e^2(1 - c) - \sqrt{(z - \sigma_e^2(1 + c))^2 - 4c\sigma_e^2}}{2cz\sigma_e^2}.$$

Hence, the critical value for detecting factors is now $\theta_{crit} = \frac{1}{G(b^+)} = \sigma_e^2(c + \sqrt{c})$. *The inverse of the Cauchy transform and the B-function are given explicitly by*

$$G^{-1}\left(\frac{1}{z}\right) = z\left(\frac{1 + \frac{\sigma_e^2(1-c)}{z}}{1 - \frac{c\sigma_e^2}{z}}\right)$$

$$B(z) = \frac{z - \sigma_e^2(1 + c)}{2\sigma_e^2\sqrt{z^2 - 2(1 + c)\sigma_e^2 z + (c - 1)^2\sigma_e^4}} - \frac{1}{2\sigma_e^2}.$$

Corollary 3: PCA for model with independent residuals

Assumption 2 holds and $e_{t,i}$ *i.i.d.* $N(0, \sigma_e^2)$. *The largest K eigenvalues of the sample covariance matrix have the following limiting values:*

$$\hat{\lambda}_i \xrightarrow{p} \begin{cases} \sigma_{F_i}^2 + \frac{\sigma_e^2}{\sigma_{F_i}^2}(c + 1 + \sigma_e^2) & \text{if } \sigma_{F_i}^2 + c\sigma_e^2 > \theta_{crit} \Leftrightarrow \sigma_F^2 > \sqrt{c}\sigma_e^2 \\ \sigma_e^2(1 + \sqrt{c})^2 & \text{otherwise.} \end{cases}$$

[16] These results have already been shown in Onatski (2012), Paul (2007) and Benaych-Georges and Nadakuditi (2011). We present them to provide intuition for the model.

The correlation between the estimated and true factors converges to

$$\widehat{Corr}(F,\hat{F}) \xrightarrow{p} \begin{pmatrix} \rho_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_K \end{pmatrix}$$

with

$$\rho_i^2 \xrightarrow{p} \begin{cases} \frac{1-\frac{c\sigma_e^4}{\sigma_{F_i}^4}}{1+\frac{c\sigma_e^2}{\sigma_{F_i}^2}+\frac{\sigma_e^4}{\sigma_{F_i}^4}(c^2-c)} & \text{if } \sigma_{F_i}^2 + c\sigma_e^2 > \theta_{crit} \\ 0 & \text{otherwise.} \end{cases}$$

Note that for $\sigma_{F_i}^2$ going to infinity, we are back in the strong factor model and the estimator becomes consistent.

6. Simulation

Next, we illustrate the performance of RP-PCA and its ability to detect weak factors with high Sharpe-ratios using a simulation exercise. We simulate factor models that try to replicate moments of the data that we are going to study in section 7. The parameters of the factors and idiosyncratic components are based on our empirical estimates. We analyze the performance of RP-PCA for different values of γ, sample size and strength of the factors. Conventional PCA corresponds to $\gamma = -1$. In a factor model only the product $F\Lambda^\top$ is well-identified and the strength of the factors could be either modeled through the moments of the factors or the values of the loadings. Throughout this section we normalize the loadings to $\Lambda^\top\Lambda/N \xrightarrow{p} I_K$ and vary the moments of the factors. The factors are uncorrelated with each others and have different means and variances. The variance of the factor can be interpreted as the proportion of assets affected by this factor. With this normalization a factor with a variance of $\sigma_F^2 = 0.5$ could be interpreted as affecting 50% of the assets with an average loading strength of 1. The theoretical results for the weak factor model are formulated under the normalization $\Lambda^\top\Lambda \xrightarrow{p} I_K$. The PCA signal in the weak factor framework corresponds to $\sigma_F^2 \cdot N$ under the normalization in the simulation.

The strength of a factor has to be put into relationship with the noise level. Based on our theoretical results the signal to noise ratio $\frac{\sigma_F^2}{\sigma_e^2}$ with $\sigma_e^2 = \frac{1}{N}\sum_{i=1}^N \sigma_{e,i}^2$ determines the variance signal of a factor. Our empirical results suggest a signal to noise ratio of around 5-7 for the first factor which is essentially a market factor. The remaining factors in the different data sets seem to have a variance signal between 0.04 and 0.8. Based on this insight we will model a four-factor model with variances $\Sigma_F = diag(5, 0.3, 0.1, \sigma_F^2)$. The variance of the fourth factor takes the values $\sigma_F^2 \in \{0.03, 0.1\}$. The first factor is a dominant market factor, while the second is also a strong factor. The third factor is weak, while the fourth factor varies from very weak to weak. We normalize the factors to be uncorrelated with each other. The Sharpe-ratios are defined as $SR_F = (0.12, 0.1, 0.3, sr)$, where the

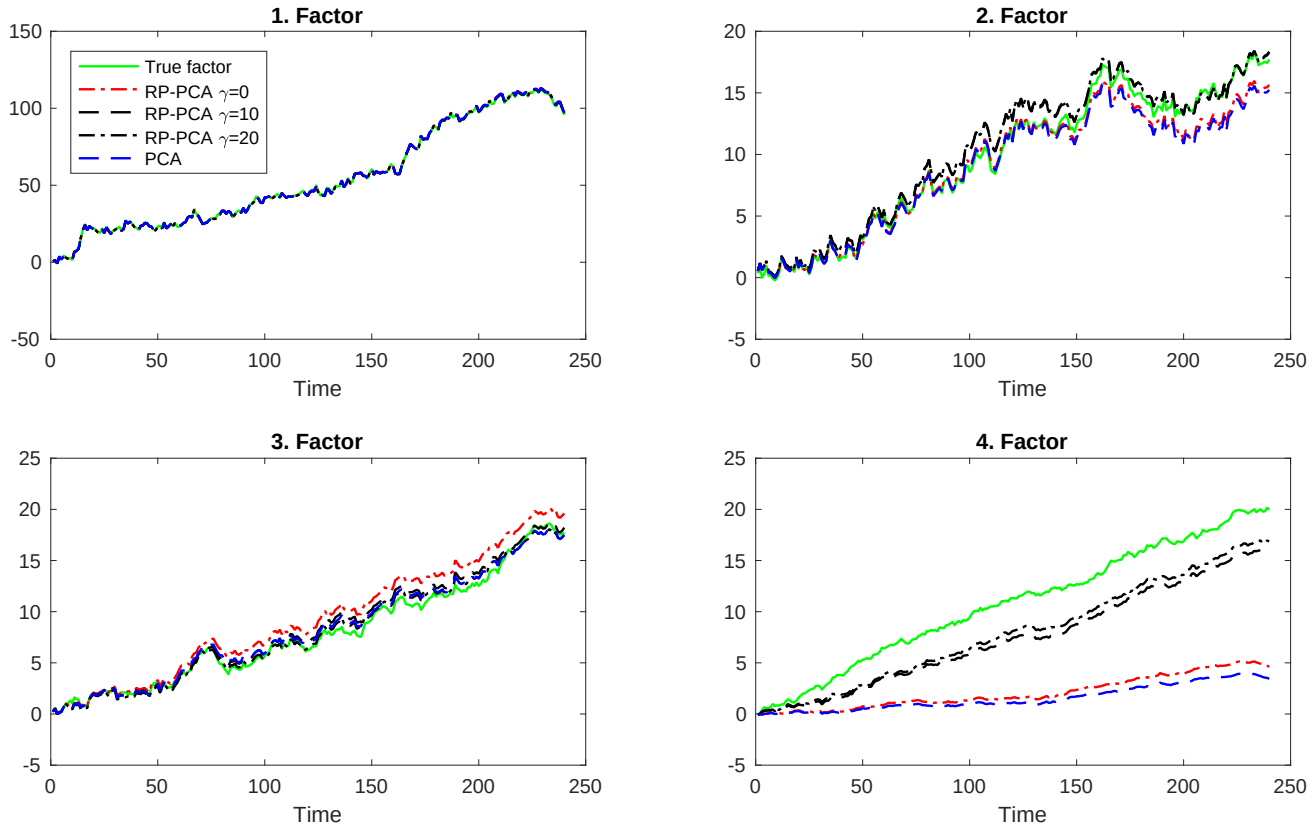

Figure 1: Sample paths of the cumulative returns of the first four factors and the estimated factor processes. The fourth factor has a variance $\sigma_F^2 = 0.03$ and Sharpe-ratio $sr = 0.5$. $N = 74$ and $T = 250$.

Sharpe-ratio of the fourth factor varies between the following values $sr \in \{0.2, 0.3, 0.5, 0.8\}$. These parameter values are consistent with our data sets.

The properties of the estimation approach depend on the average idiosyncratic variance and dependency structure in the residuals. We normalize the average noise variance $\sigma_e^2 = 1$, which implies that the factor variances can be directly compared to the variance signals in the data.[17] We use two different set of residual correlation matrices.

First, the correlation matrix of our simulated residuals is set to the empirical correlation that we observe in the data. In more detail, we have estimated the residual correlation matrix based on $N = 25$ size and value double-sorted portfolios, $N = 74$ extreme deciles sorted portfolios and $N = 370$ decile sorted portfolios as described in the empirical Section 7.[18] In each case we have first regressed out the systematic factors and then estimated the residual covariance matrix with a hard

[17] For the empirical data sets with $N = 370$ assets the average noise variance is around $\sigma_e^2 = 2.5$. Instead of normalizing $\sigma_e^2 = 1$ we could also multiply Σ_F by 2.5 and obtain the same factor model that is consistent with the data.

[18] We use the same data set as Kozak, Nagel and Santosh (2017) to construct $N = 370$ decile-sorted portfolios of monthly returns from 07/1963 to 12/2017 (T=650). We use the lowest and highest decile portfolio for each anomaly to create a data set of $N = 74$ portfolios. The $N = 25$ double-sorted portfolios are from Kenneth-French website for the same time period.

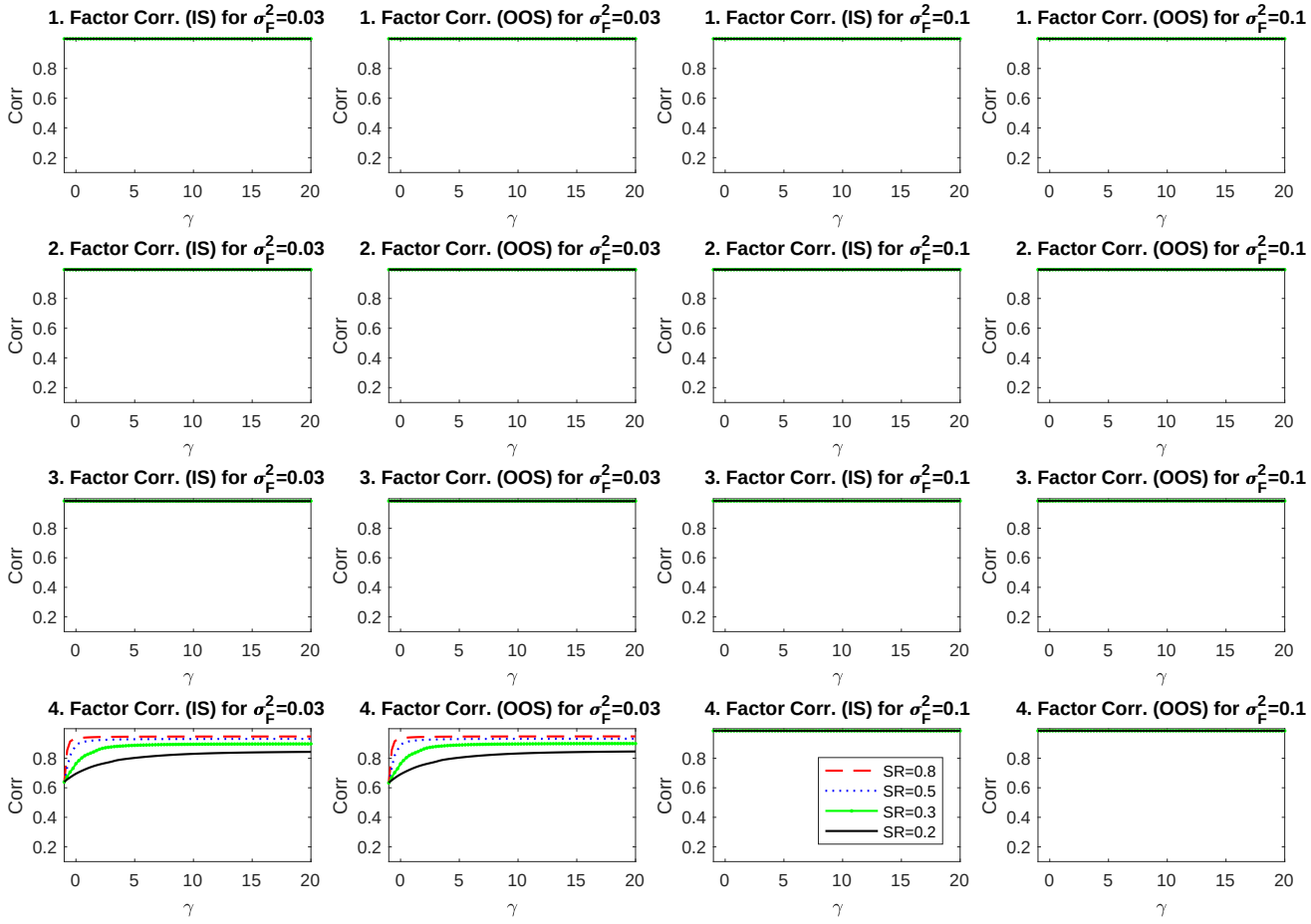

Figure 2: $N = 370, T = 650$: Correlation of estimated rotated factors in-sample and out-of-sample for different variances and Sharpe-ratios of the fourth factor and for different RP-weights γ. We use the empirical residual correlation matrix.

thresholding approach setting small values to zero.[19]. This provides a consistent estimator of the residual population covariance matrix. We have regressed out the first 3 PCA factors for the first data set and the first 6 PCA factors for the last two data sets.[20] 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. Second, for $N = 370$ assets we create a sparse residual correlation matrix based on $\Sigma = CC^\top$, where C is a matrix with where the first 13 off-diagonal elements take the value 0.7. The resulting covariance matrix is normalized to the corresponding correlation matrix. The residuals are then generated as $e_t = \epsilon\Sigma$ where ϵ_t are i.i.d. draws from a multivariate standard normal distribution.

In the main part we consider only the cross-sectional dimension $N = 370$ and time dimension $T = 650$, but in the appendix we also study the combinations $\{N = 74, T = 650\}$ and $\{N = 25, T = 240\}$ motivated by our empirical analysis. The loadings are i.i.d draws from a standard multivariate normal distribution. The factors are i.i.d. draws from a multivariate normal distribution with means

[19] See Bickel and Levina (2008) and Fan, Liao and Mincheva (2013))

[20] 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.

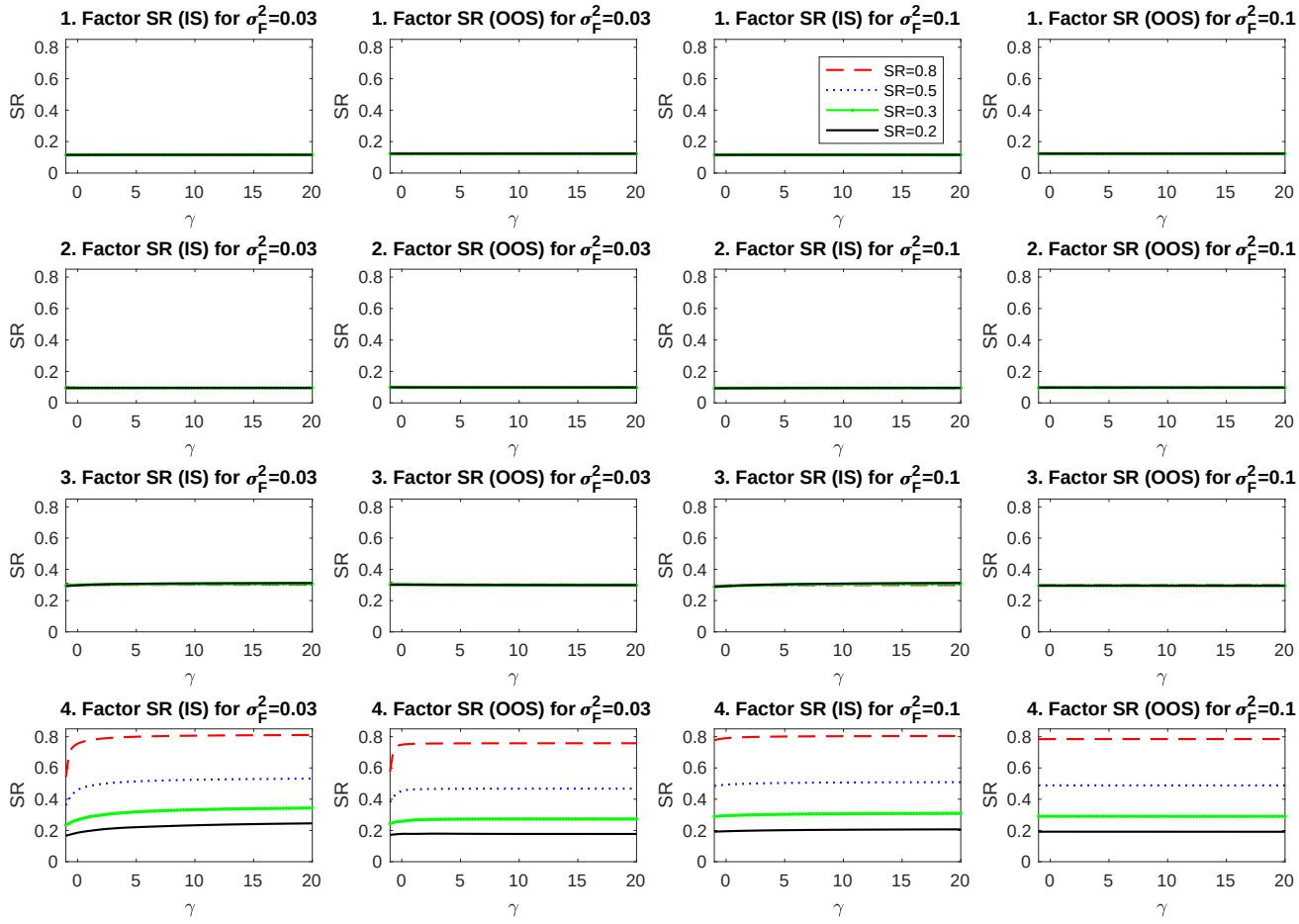

Figure 3: $N = 370, T = 650$: Sharpe ratios of estimated rotated factors in-sample and out-of-sample for different variances and Sharpe-ratios of the fourth factor and for different RP-weights γ. We use the empirical residual correlation matrix.

and variances specified as above. The idiosyncratic components are i.i.d. draws from a multivariate normal distribution with mean zero and covariance matrix based on a consistent estimation of the empirical residual correlation matrix respectively the parametric band-diagonal matrix. For each setup we run 100 Monte-Carlo simulations. For the out-of-sample results we first estimate the loading vector in-sample and then obtain the out-of-sample factor estimates by projecting the out-of-sample returns on the estimated loadings.

Figure 1 provides some intuition for our estimator. It illustrates the sample path estimates for different values of γ. If the fourth factor is weak with a high Sharpe-ratio, then conventional PCA or RP-PCA with a too small value of γ cannot detect it while RP-PCA with a sufficiently large γ is able to detect the factor.

Figures 2 and 3 show correlations and Sharpe-ratios in the four-factor model for $N = 370$ and $T = 650$ based on the empirical residual correlation structure. A.10 and A.11 show the results for $N = 74$.[21] The risk-premium weight γ has the largest effect on estimating the fourth factor if it is weak ($\sigma_F^2 = 0.03$) and has a high Sharpe ratio ($sr \geq 0.3$). The second takeaway is that the estimates

[21] All simulation results in the appendix are based on the empirical residual correlation matrix.

of the strong factors is essentially not affected by the properties of the weak factors and vice versa. Hence, one could first estimate the strong factors and project them out and then estimate the weak factors from the projected data. Motivated by this finding we will study a one-factor model in more detail.

Figure 4 compares the prediction of our weak factor model theory with the Monte-Carlo simulation for the empirical and the band-diagonal residual correlation matrix. We consider one factor with Sharpe-ratio 0.8, but increasing variance. The prediction of our statistical model is confirmed by the Monte-Carlo simulation. It convincingly shows how weak factors can be better estimated with RP-PCA with a large γ when the Sharpe-ratio is high. In Figure 5 we plot the value of ρ_i^2 in the weak factor model which determines the detection and correlation of the factors. We vary the signal θ which among others depends on the choice of γ. We compare uncorrelated residuals with our weak dependency structures. It is apparent that increasing the signal strength for detecting weak factors becomes more relevant for correlated residuals.

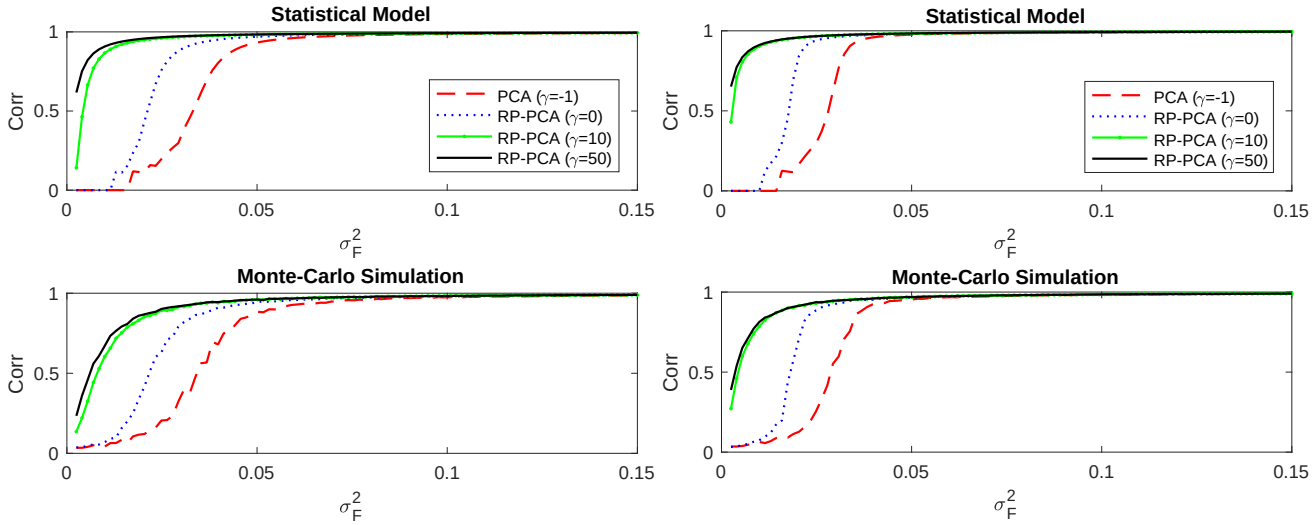

Figure 4: Correlations between estimated and true factor based on the weak factor model prediction and Monte-Carlo simulations for different variances of the factor. Left plots: The residuals have cross-sectional correlation defined by the band-diagonal matrix. Right plots: The residuals have the empirical residual correlation matrix. The Sharpe-ratio of the factor is 0.8, i.e. the mean equals $\mu_F = 0.8 \cdot \sigma_F$. We have $T = 650$ and $N = 370$, i.e. the normalized variance of the factors in the weak factor model corresponds to $\sigma_F^2 \cdot N$.

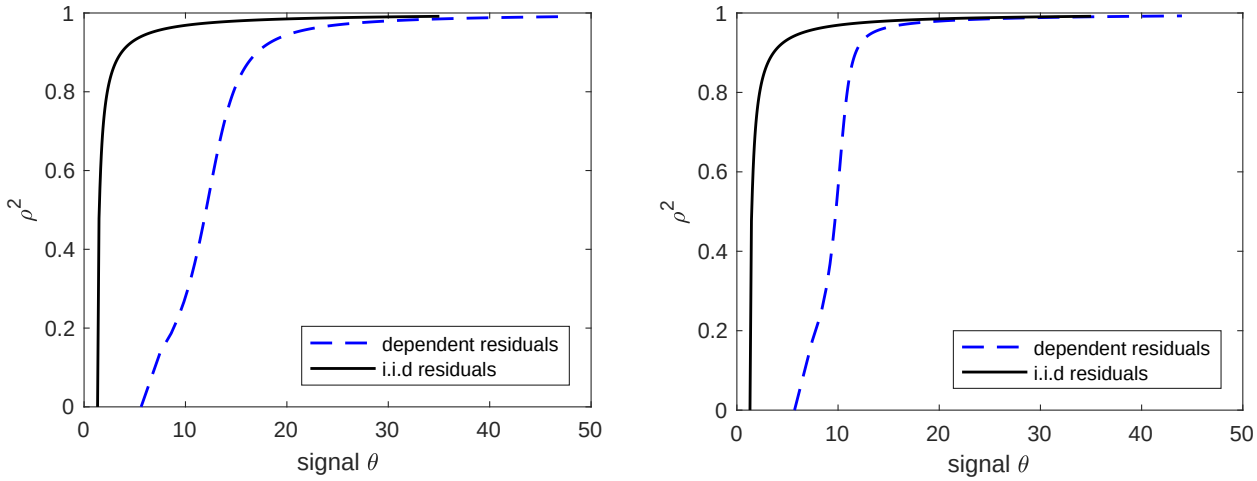

Figure 5: Model-implied values of ρ_i^2 ($\frac{1}{1+\theta_i B(\hat{\theta}_i))}$ if $\theta_i > \sigma_{crit}^2$ and 0 otherwise) for different signals θ_i. The average noise level is normalized in both cases to $\sigma_e^2 = 1$. Left plots: The residuals have cross-sectional correlation defined by the band-diagonal matrix. Right plots: The residuals have the empirical residual correlation matrix.

Figures 6 and 7 provide more refined results for the one-factor model for $N = 370$ and $T = 650$ for the empirical and band-diagonal residual correlation matrix. We consider a factor variance $\sigma_F^2 \in \{0.03, 0.05, 0.1, 0.3, 1.0\}$ which ranges from weak to strong factors. Figures A.12 to A.16 show the results for $N = 74$ and $N = 25$ and include estimates of the root-mean-squared pricing errors. The risk-premium weight γ has the largest effect on correlations, Sharpe-ratios and pricing errors if the factors are weak ($\sigma_F^2 = 0.03$ or 0.05) and have a high Sharpe ratio ($sr \geq 0.3$). Note, that if there is not much information in the mean, i.e. the Sharpe-ratio of the factor is low, a too high value $\gamma > 10$ can lead to an overestimation of the Sharpe-ratio in-sample. This makes sense as if too much weight is given to an uninformative mean, the estimator will pick up some of the non-zero residuals. Note, that the out-of-sample results provide reliable estimates that are not affected by overfitting issues. Our estimator has a larger effect for smaller values of N as this implies a weaker signal for the factors.

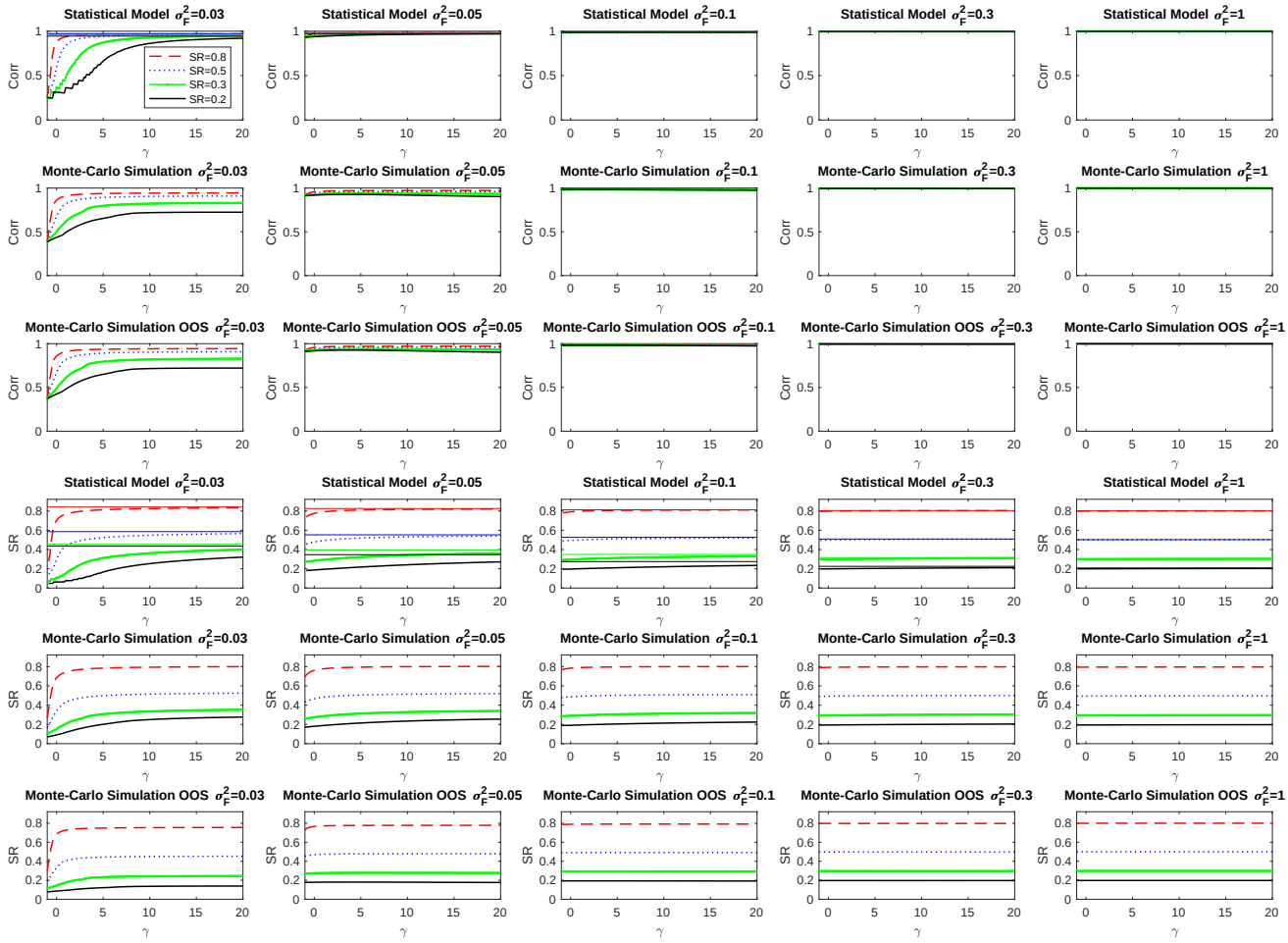

Figure 6: $N = 370, T = 650$: Correlations and Sharpe-ratios as a function of the RP-weight γ for different variances and Sharpe-ratios. The residuals have cross-sectional correlation defined by the band-diagonal matrix.

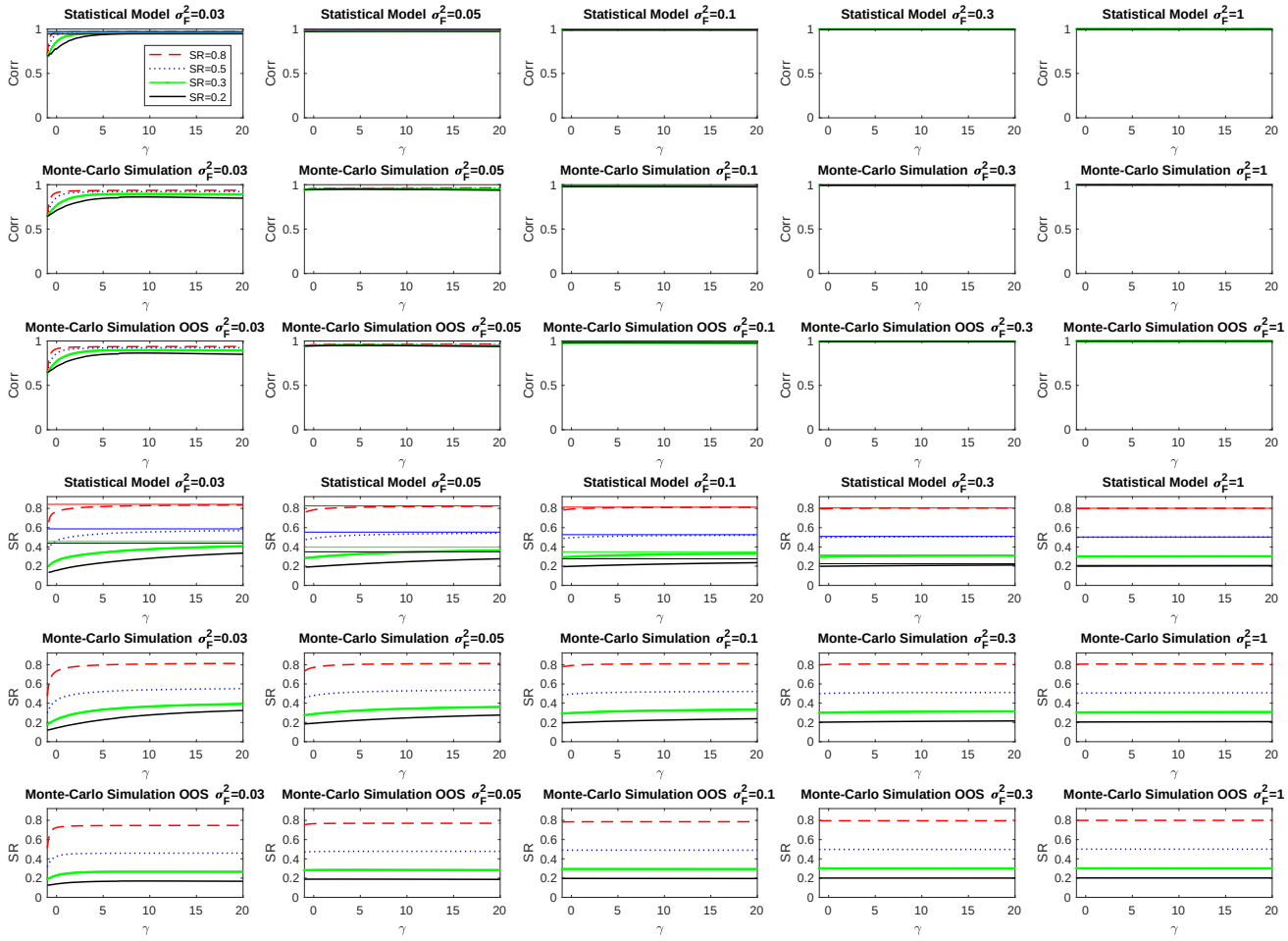

Figure 7: $N = 370, T = 650$: Correlations and Sharpe-ratios as a function of the RP-weight γ for different variances and Sharpe-ratios. The residuals have the empirical residual correlation matrix.

7. Empirical Application

We apply our estimator to a large number of anomaly sorted portfolios. The same data is studied in more detail in our companion paper Lettau and Pelger (2018). Based on the universe of U.S. firms in CRSP, we consider 37 anomaly characteristics following standard definitions in Novy-Marx and Velikov (2016), McLean and Pontiff (2016) and Kogan and Tian (2015). We use the same data set as Kozak, Nagel and Santosh (2017)[22] who have sorted the stock returns in yearly rebalanced decile portfolios. This gives us a total cross-section of $N = 370$ portfolios of monthly returns from 07/1963 to 12/2017 (T=650).[23] The risk-free rate to obtain excess returns is from Kenneth French's website. We estimate statistical factors for different choices of γ and evaluate the maximum Sharpe-ratio, average pricing error and explained variation in- and out-of-sample.

Table 1 reports the results for $K = 3$ and $K = 5$ factors for RP-PCA with $\gamma = 10$ and PCA ($\gamma = -1$). SR denotes the maximum Sharpe-ratio that can be obtained by a linear combination of the factors, i.e.

[22] We thank the authors for sharing the data.

[23] Kozak, Nagel and Santosh (2017) create a data set based on 50 anomalies, but 13 of these anomalies are only available for a significantly shorter time horizon. We choose only those anomalies that are available for the whole time horizon of $T = 650$ observations.

it combines the factors with the weights $\Sigma_F^{-1}\mu_F$. It measures how well the factors can approximate the stochastic discount factor. 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.

	In-sample			Out-of-sample		
	SR	RMS α	Idio. Var.	SR	RMS α	Idio. Var.
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%

Table 1: Maximal Sharpe-ratios, root-mean-squared pricing errors and idiosyncratic variation for different number of factors. RP-weight $\gamma = 10$.

RP-PCA and PCA differ the most in terms of the maximum Sharpe-ratio. For $K = 5$ factors the in- and out-of-sample Sharpe-ratio of RP-PCA is twice as large as for PCA. For $K = 3$ factors there is still a sizeable difference in Sharpe-ratios, but it is less pronounced than for a larger number of factors. A possible reason is that the 4th or 5th factor is weak with a high Sharpe-ratio and only picked up by RP-PCA, while the first four factors are stronger and hence can be detected by PCA. Surprisingly, the pricing errors and the unexplained variation are very close for the two methods. Only the out-of-sample pricing error of RP-PCA is smaller than for PCA. It seems that RP-PCA selects high Sharpe-ratio factors with smaller out-of-sample pricing errors without sacrificing explanatory power for the variation.

Figure 8 analyzes the effect of γ and the number of factors on the three criteria maximum Sharpe-ratio, pricing error and variation. The Sharpe-ratio and pricing error change significantly when including the 5th factor. This 5th factor is also strongly affected by the choice of γ and seems to require $\gamma > 5$ to be detected by RP-PCA. Adding the 6th factor has only a very minor effect on the three criteria. That is why we opt for a 5-factor model. The figure illustrates that the amount of unexplained variation is insensitive to the choice of γ. Hence, our factors capture more pricing information while explaining the same amount of variation in the data.

Table 2 shows that the variance signal for different factors suggests the existence of weak factors. Here we extract the first 6 factors with RP-PCA ($\gamma = 10$) and PCA. In addition, we include the pop-

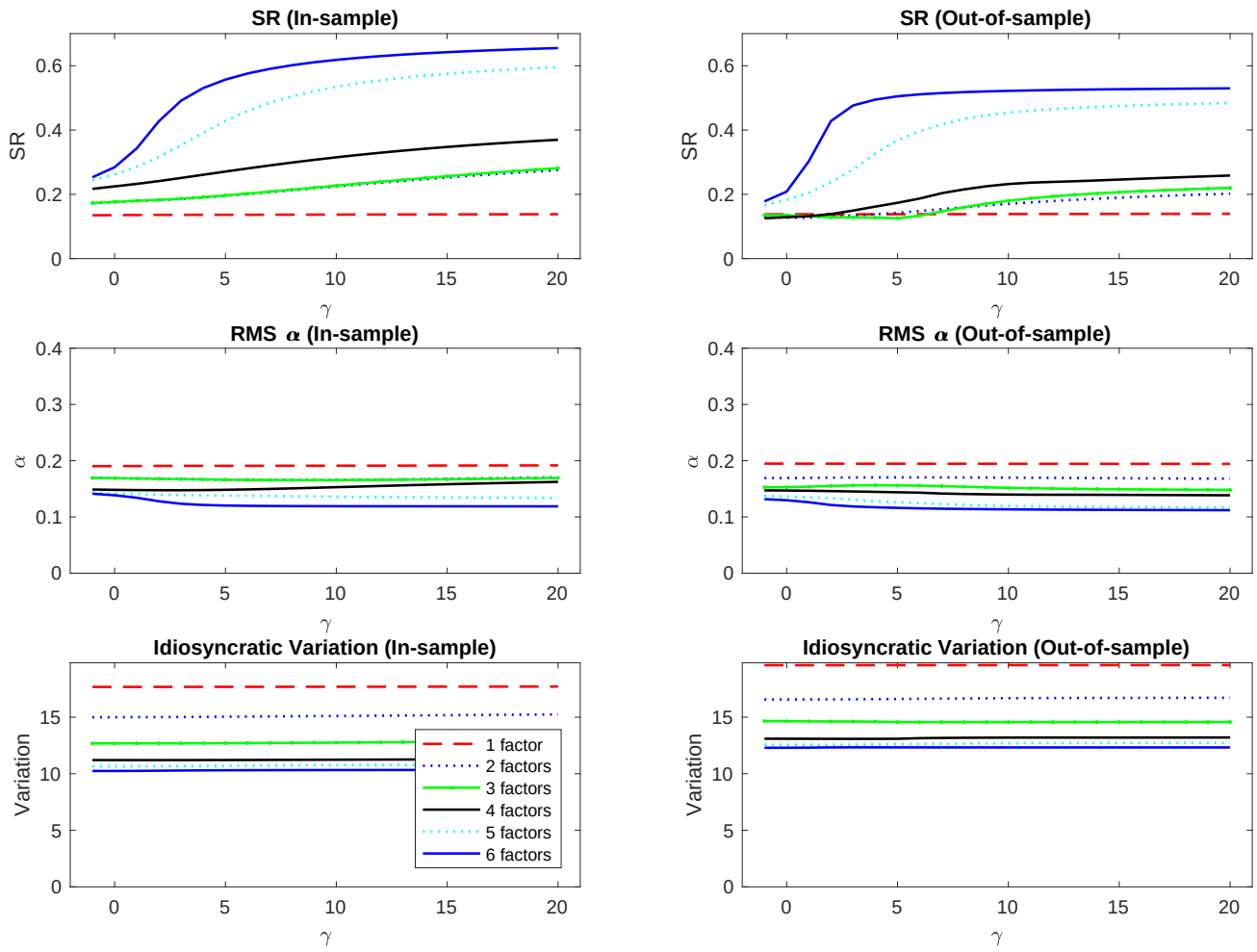

Figure 8: Deciles of 37 single-sorted portfolios from 07/1963 to 12/2016 ($N = 370$ and $T = 650$): Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation for different values of γ.

ular Fama-French 5 factors (marke, size, value, profitability and investment) from Kenneth French's website. The variance signal is defined as the largest eigenvalues of $\Lambda\Sigma_F\Lambda^\top$. We normalize these eigenvalue by the same constant $\sigma_e^2 = \frac{1}{N}\sum_{i=1}^N \sigma_{e,i}^2$ based on the residuals from 6 PCA factors.[24] This makes the variance signals comparable to our simulation design. The 5th factor has a variance signal around 0.05 which based on our simulation is well described by a weak factor model. The simulations also predict that these weak factors can be better estimated by RP-PCA if they have a large Sharpe-ratio. This is exactly what we observe in the data.

The left plot in Figure 9 shows the eigenvalues of the matrix $\frac{1}{N}\left(\frac{1}{T}X^\top X + \gamma\bar{X}\bar{X}^\top\right)$ normalized by the average idiosyncratic variance. Our weak factor model predicts that the signal of this matrix should be larger for RP-PCA compared to PCA. The eigenvalue curves confirm that the signal for the weaker factors clearly separates from the PCA signal. $\gamma = 10$ seems to be sufficient for strengthening the signal. The right plot in Figure 9 normalizes the eigenvalues by the corresponding PCA eigenvalues. In particular the signal for the 6th factor is strengthened.

[24] The results do not change if we regress out more PCA or RP-PCA factors and are available upon request.

	PCA	RP-PCA ($\gamma = 10$)	FF5
σ_1^2	8.05	8.05	8.00
σ_2^2	0.27	0.27	0.21
σ_3^2	0.21	0.21	0.17
σ_4^2	0.14	0.14	0.03
σ_5^2	0.05	0.05	0.02
σ_6^2	0.03	0.04	0.00

Table 2: Deciles of 37 single-sorted portfolios: Variance signal for different factors: Largest eigenvalues of $\Lambda\Sigma_F\Lambda^\top$ normalized by the average idiosyncratic variance $\sigma_e^2 = \frac{1}{N}\sum_{i=1}^N \sigma_{e,i}^2$.

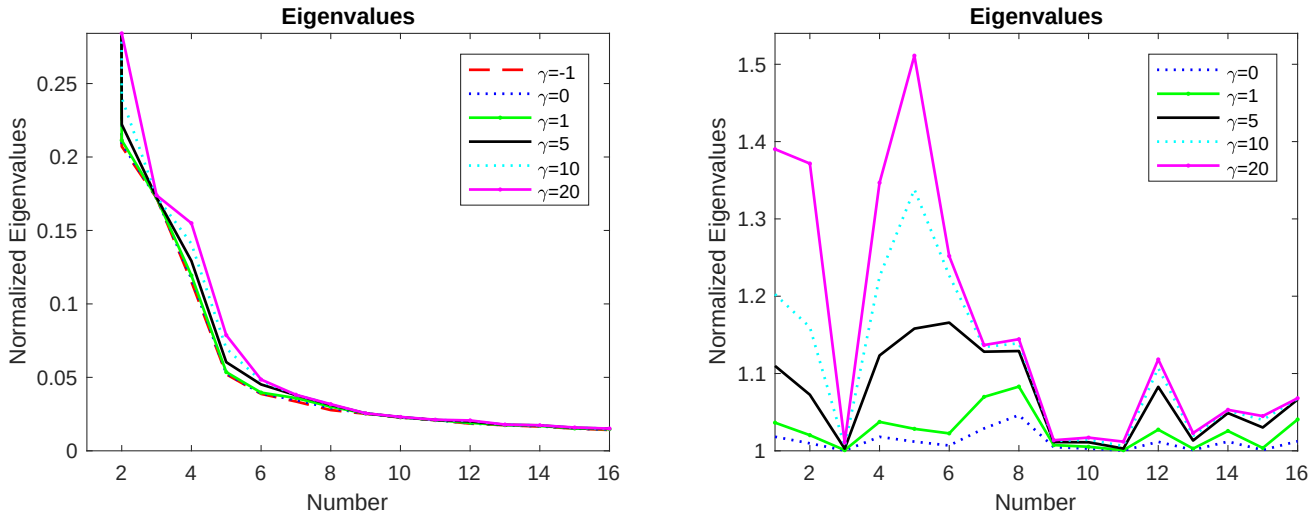

Figure 9: Deciles of 37 single-sorted portfolios from 07/1963 to 12/2016 ($N = 370$ and $T = 650$): Largest normalized eigenvalues of the matrix $\frac{1}{N}\left(\frac{1}{T}X^\top X + \gamma\bar{X}\bar{X}^\top\right)$ for different RP-weights γ. Left plot: Eigenvalues are normalized by division through the average idiosyncratic variance $\sigma_e^2 = \frac{1}{N}\sum_{i=1}^N \sigma_{e,i}^2$ estimated by the average of the non-systematic PCA eigenvalues. Right plot: Eigenvalues are normalized by the corresponding PCA ($\gamma = -1$) eigenvalues.

8. Conclusion

We develop a new estimator for latent asset pricing factors from large data sets. Our estimator is essentially a regularized version of PCA that puts a penalty on the pricing error. We derive the asymptotic distribution theory under weak and strong factor model assumptions and 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.

Appendix A. Simulation

Appendix A.1. Multi-Factor Model

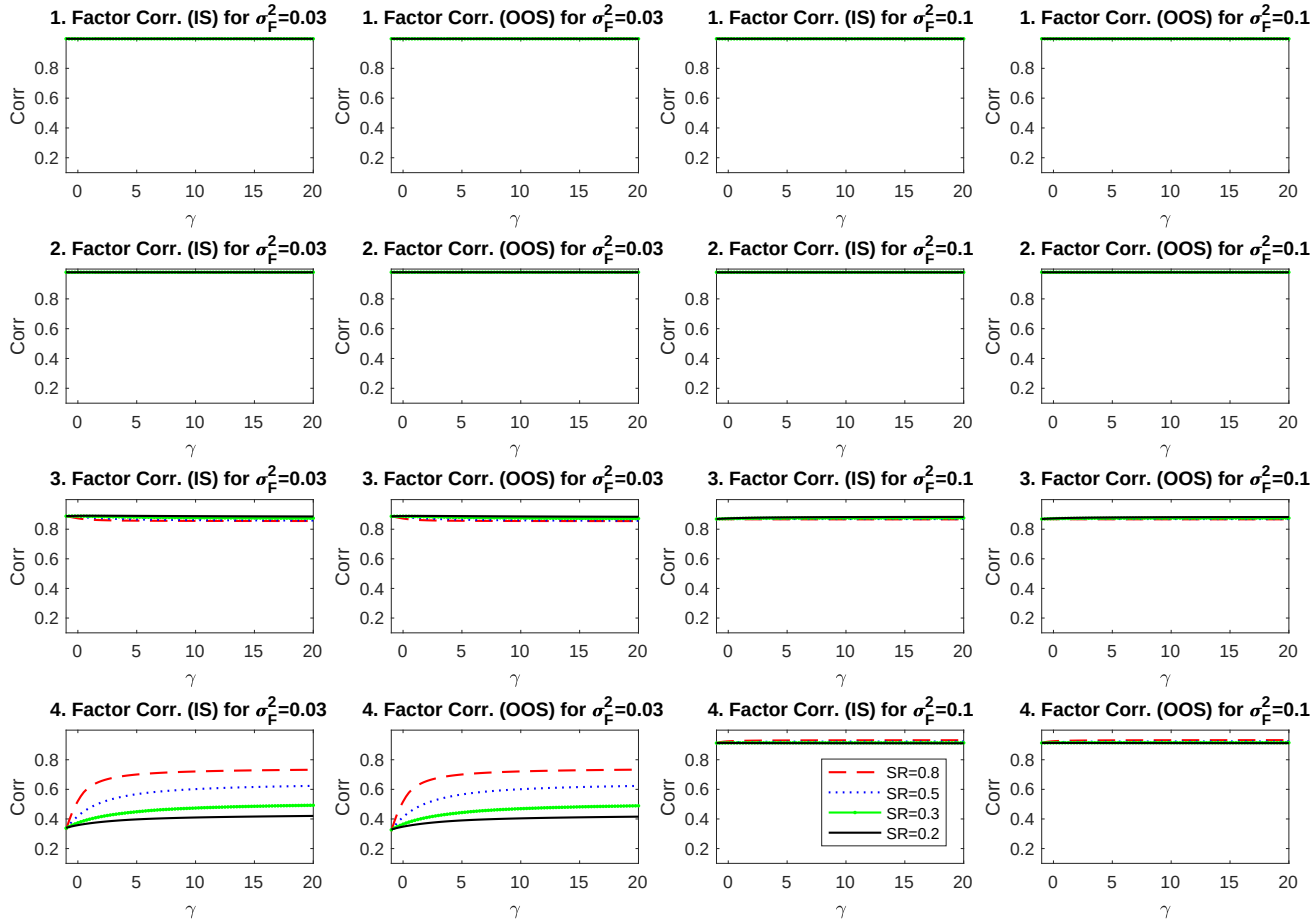

Figure A.10: $N = 74, T = 650$: Correlation of estimated rotated factors with true factors in-sample and out-of-sample for different variances and Sharpe-ratios of the fourth factor and for different RP-weights γ.

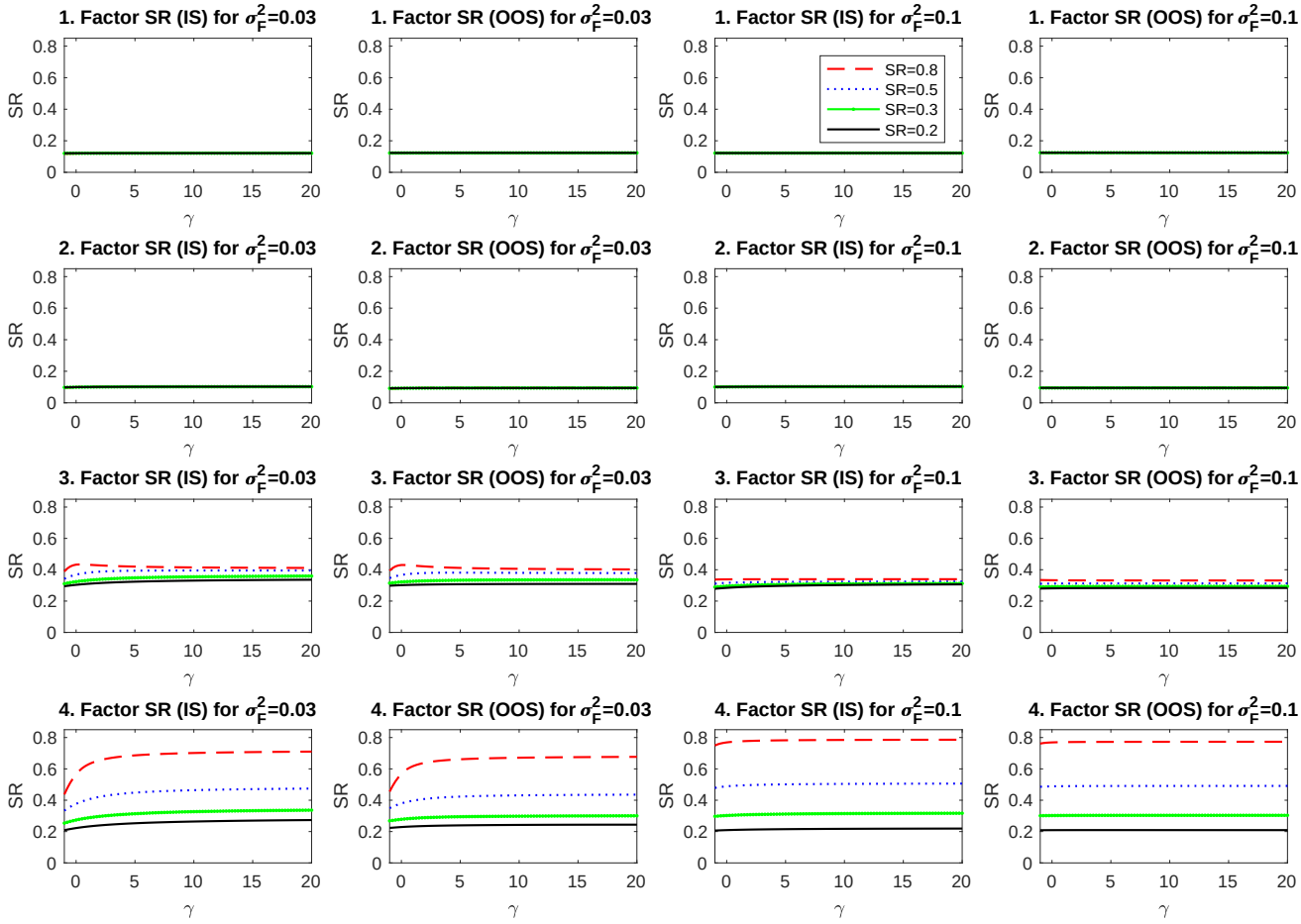

Figure A.11: $N = 74, T = 650$: Sharpe ratios of estimated rotated factors in-sample and out-of-sample for different variances and Sharpe-ratios of the fourth factor and for different RP-weights γ.

Appendix A.2. Single-Factor Model with $N = 74$ and $T = 650$

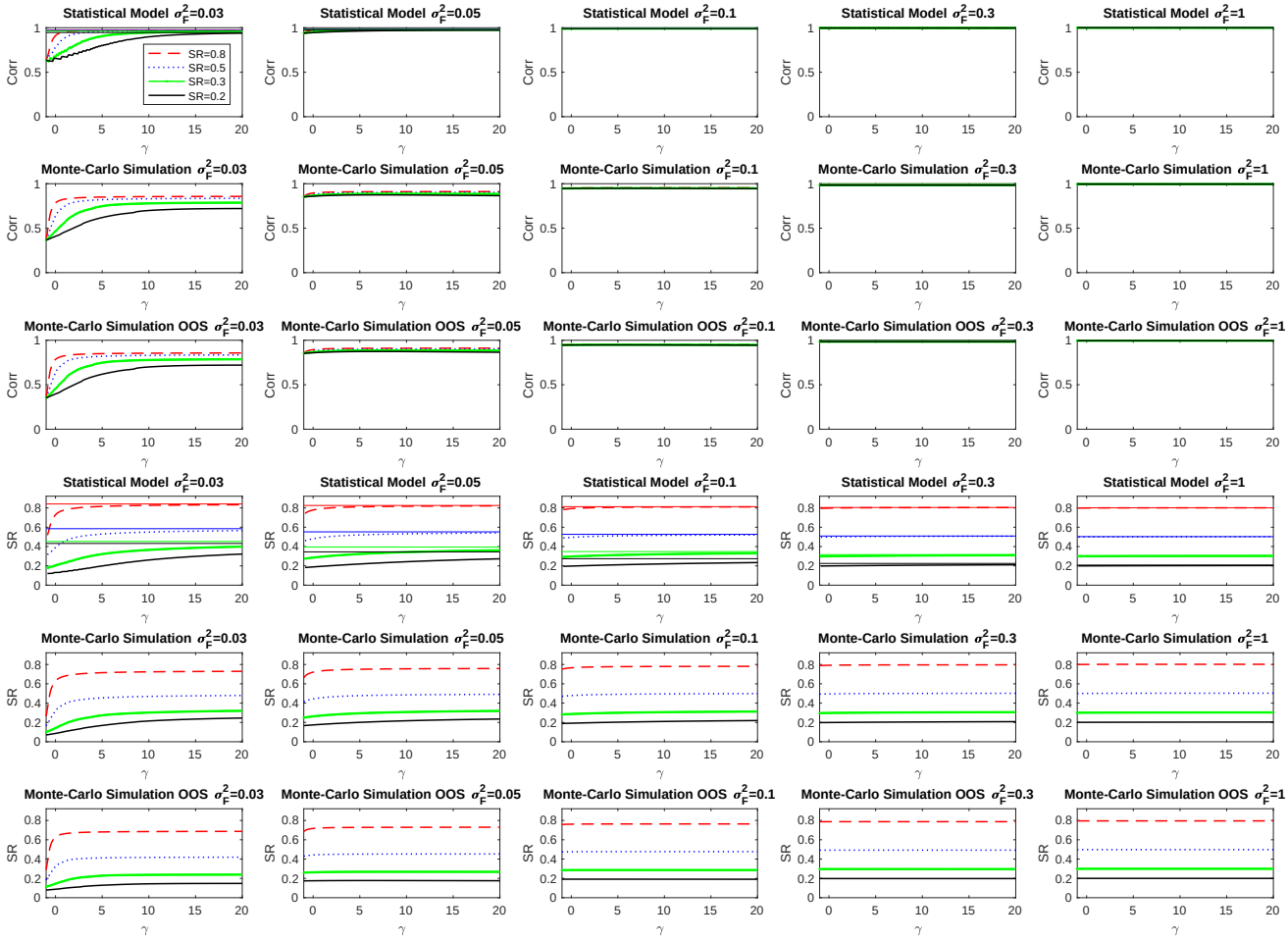

Figure A.12: $N = 74, T = 650$: Correlations and Sharpe-ratios as a function of the RP-weight γ for different variances and Sharpe-ratios.

Appendix A.3. Single-Factor Model with $N = 25$ and $T = 240$

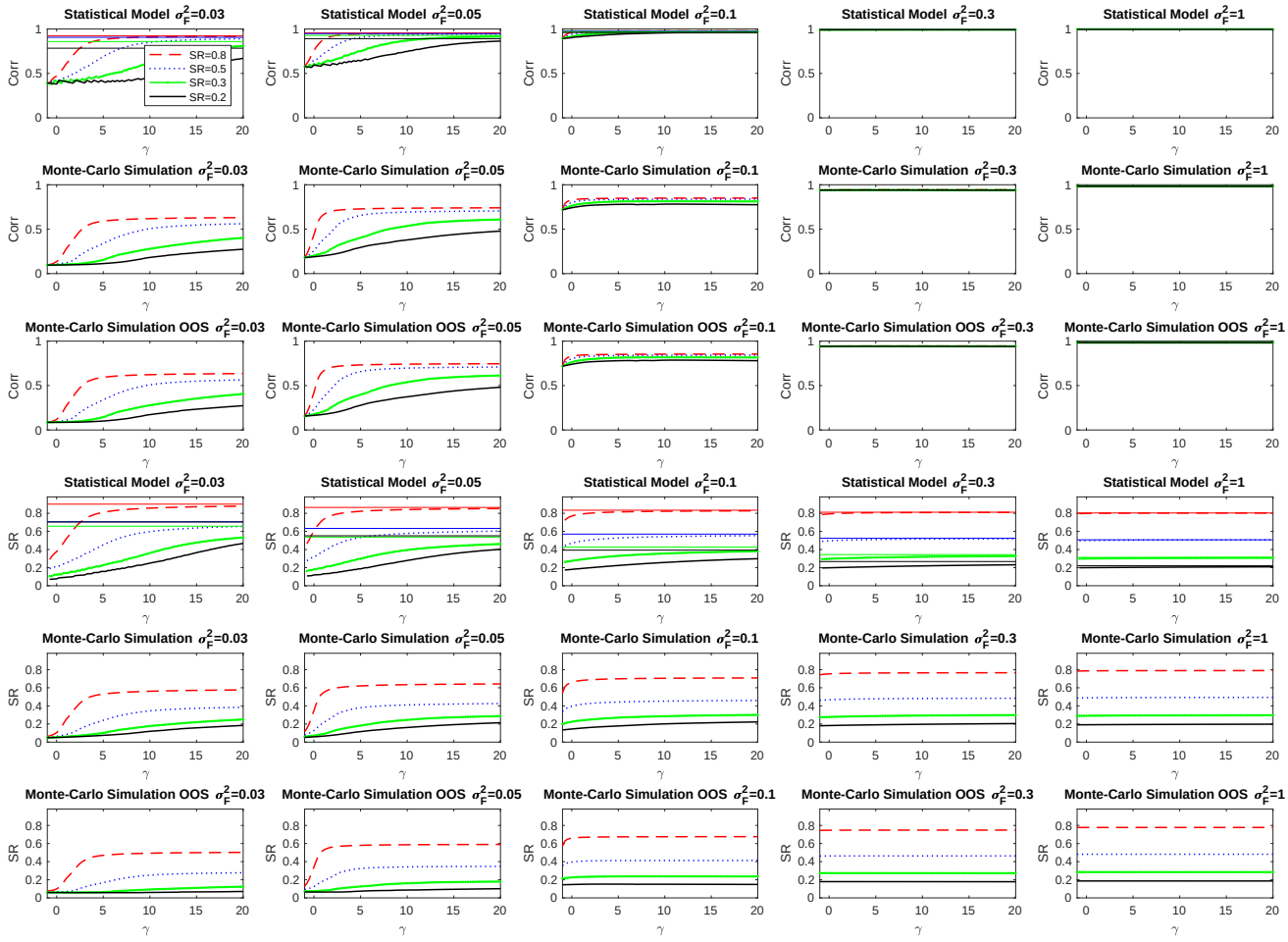

Figure A.13: $N = 25, T = 240$: Correlations and Sharpe-ratios as a function of the RP-weight γ for different variances and Sharpe-ratios.

Appendix A.4. Pricing Errors for Single-Factor Model

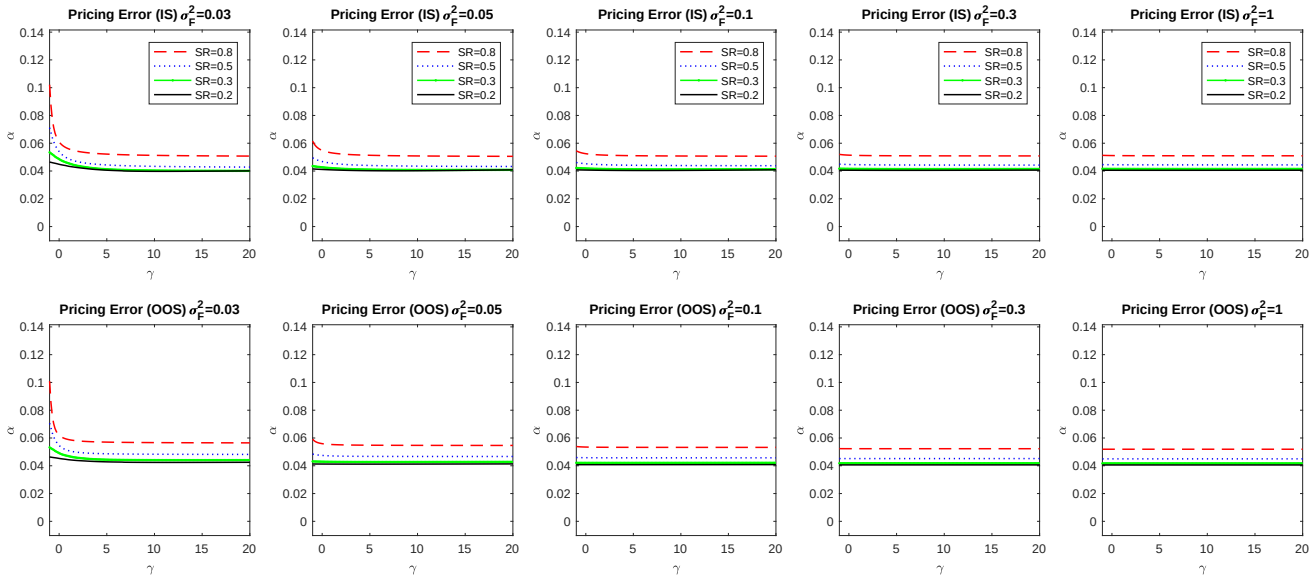

Figure A.14: $N = 370, T = 650$: Root-mean-squared pricing errors as a function of the RP-weight γ for different variances and Sharpe-ratios.

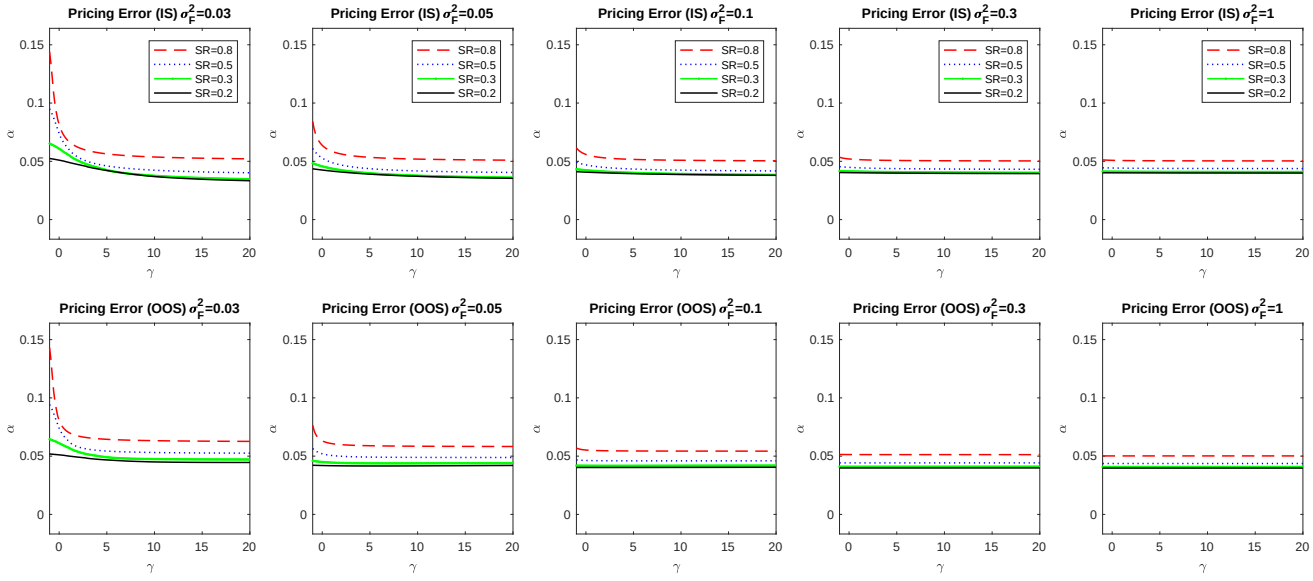

Figure A.15: $N = 74, T = 650$: Root-mean-squared pricing errors as a function of the RP-weight γ for different variances and Sharpe-ratios.

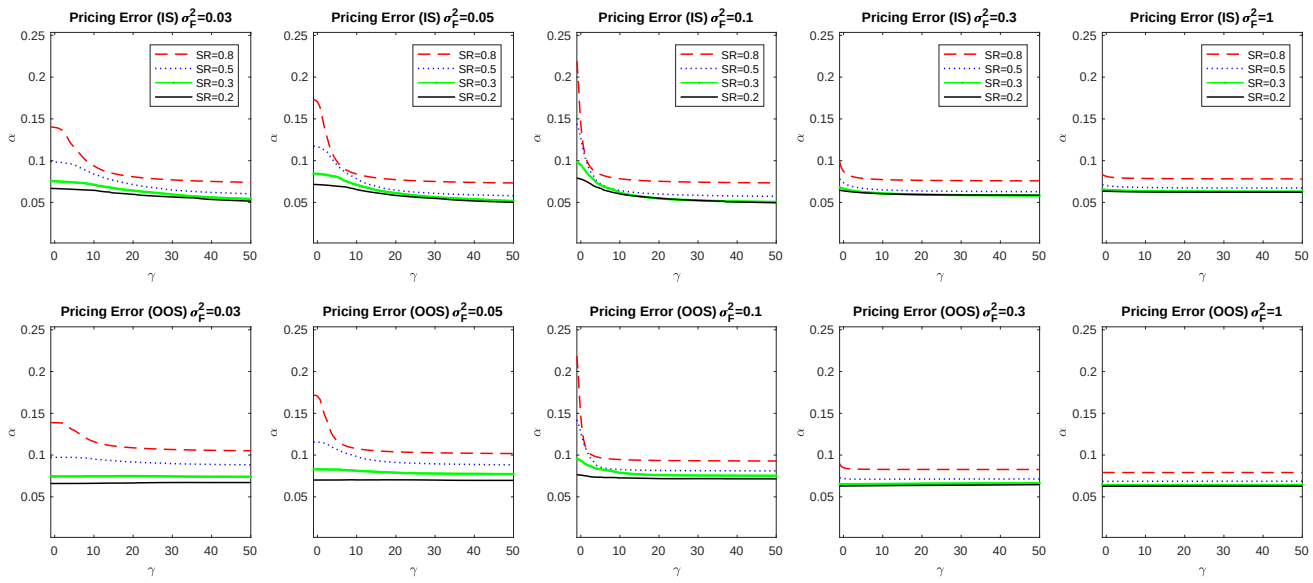

Figure A.16: $N = 25, T = 240$: Root-mean-squared pricing errors as a function of the RP-weight γ for different variances and Sharpe-ratios.

Appendix B. Proofs for the Weak Factor Model

We only prove the statements for RP-PCA. The statements for the conventional PCA based on the covariance matrix are a special case. Given an $N \times N$ matrix A we denote the sorted eigenvalues by $\lambda_1(A) \geq ... \geq \lambda_N(A)$. Let $\phi_A(z)$ be the empirical eigenvalue distribution, i.e. the probability measure defined as $\phi_A(z) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i(A)}$ where δ_x is the Dirac measure. In our case the probability measure ϕ_A converges almost surely weakly for $T \to \infty$ (and therefore also $N \to \infty$ as $\frac{N}{T} \to c > 0$ and N and T are asymptotically proportional).

Proof of Theorem 2:

Instead of using $\frac{1}{T}X^\top W^2 X$ we study $\frac{1}{T}WXX^\top W$ with $W = I_T + \frac{\tilde{\gamma}}{T}\mathbb{1}\mathbb{1}^\top$ and $\tilde{\gamma} = \sqrt{\gamma+1} - 1$. Define the orthonormal matrix $U = (U_1, U_2)$ consisting of the $T \times K+1$ matrix U_1 and the $T \times T-K-1$ matrix U_2 by

$$U_1 = \left(\left(I_T - \tfrac{1}{T}\mathbb{1}\mathbb{1}^\top\right)\frac{F}{\sqrt{T}} \quad \frac{\mathbb{1}}{\sqrt{T}}\right)\begin{pmatrix}(F^\top(I_T - \frac{1}{T}\mathbb{1}\mathbb{1}^\top)F)^{-1/2} & 0\\ 0 & 1\end{pmatrix}\tilde{U},$$

where the $K+1 \times K+1$ matrix $\tilde{U}$ consists of the orthonormal eigenvectors of the "signal matrix" $M_{\text{RP-PCA}}$:

$$\tilde{U}^\top\begin{pmatrix}\Sigma_F + c\sigma_e^2 & \Sigma_F^{1/2}\mu_F(1+\tilde{\gamma})\\ \mu_F^\top\Sigma_F^{1/2}(1+\tilde{\gamma}) & (1+\gamma)(\mu_F^\top\mu + c\sigma_2^2)\end{pmatrix}\tilde{U} = \begin{pmatrix}\theta_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \theta_{K+1}.\end{pmatrix}$$

U_2 are orthonormal vectors orthogonal to U_1, i.e. $U_1^\top U_2 = 0$ and $U_2^\top U_2 = I_{T-K-1}$.

We now analyze the spectrum of $S := \frac{1}{T}U^\top WXX^\top WU$, which has the same eigenvalues as $\frac{1}{T}X^\top W^2X$.

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{T}U_1^\top W(F\Lambda^\top + e)(F\Lambda^\top + e)^\top WU_1 & \frac{1}{T}U_1^\top W(F\Lambda^\top + e)e^\top WU_2 \\ \frac{1}{T}U_2^\top We(\Lambda F^\top + e)WU_1 & \frac{1}{T}U_2^\top Wee^\top WU_2 \end{pmatrix}.$$

An eigenvalue of S that is not an eigenvalue of S_{22} satisfies

$$0 = \det(\lambda I_T - S) = \det(\lambda I_{T-K-1} - S_{22})\det(\lambda I_{K+1} - \kappa_T(\lambda))$$

with

$$\kappa_T(\lambda) = S_{11} + S_{12}(\lambda I_{T-K-1} - S_{22})^{-1}S_{21}.$$

For sufficiently large T it holds $\det(\lambda I_{T-K-1} - S_{22}) \neq 0$ for the first $K+1$ eigenvalues. Therefore the first $K+1$ eigenvalues satisfy

$$\det(\lambda I_{K+1} - \kappa_T(\lambda)) = 0.$$

We want to study the limiting behavior of $\kappa_T(\lambda)$ for $T \to \infty$.

$$\begin{aligned} \kappa_T(\lambda) =& \frac{1}{T}\left(U_1^\top W(F\Lambda^\top + e)\right)\left(I_N + \frac{1}{T}e^\top WU_2\left(\lambda_{T-K-1} - \frac{1}{T}U_2^\top Wee^\top WU_2\right)^{-1}U_2^\top eW\right) \\ & \cdot \left(U_1^\top W(F\Lambda^\top + e)\right)^\top \\ =& \frac{\lambda}{T}\left(U_1^\top W(F\Lambda^\top + e)\right)\left(\lambda_N - \frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\left(U_1^\top W(F\Lambda^\top + e)\right)^\top, \end{aligned}$$

where we have used the identify that for $\lambda \neq 0$ which is not an eigenvalue of $A^\top A$ it holds $I_T + A(\lambda I_N - A^\top A)^{-1}A^\top = \lambda(\lambda I_N - AA^\top)^{-1}$.

Because of the orthonormality we have $U_2e =: \tilde{e}$ with $\tilde{e}_t \overset{i.i.d.}{\sim} N(0,\Sigma)$. Note that $U_2W = U_2$ by construction. For any matrix C independent of $U_1^\top We$ we have

$$\begin{aligned} E\left[U_1^\top WeCe^\top WU_1\right] =& \ \operatorname{trace}(\Sigma)\cdot\ \operatorname{trace}(C)\cdot U_1^\top WU_1 \\ =& \ \operatorname{trace}(\Sigma)\cdot\ \operatorname{trace}(C)\cdot \tilde{U}^\top\begin{pmatrix} I_K & 0 \\ 0 & 1+\gamma \end{pmatrix}\tilde{U}. \end{aligned}$$

By the law of large numbers and Lemma A.2 in Benaych-Georges and Nadakuditi (2011) it holds first

$$\begin{aligned} & \frac{\lambda}{T}\left(U_1^\top W(F\Lambda^\top)\right)\left(\lambda I_N - \frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\left(U_1^\top W(F\Lambda^\top)\right)^\top \\ =& \lambda\left(\frac{1}{T}U_1^\top WFF^\top WU_1\right)\frac{1}{N}\operatorname{trace}\left(\left(\lambda I_N - \frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\right) + o_p(1), \end{aligned}$$

second

$$\begin{aligned}&\frac{\lambda}{T}\left(U_1^\top e\right)\left(\lambda I_N-\frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\left(U_1^\top We\right)^\top\\ =&\lambda\left(U_1^\top WU_1\right)\cdot\frac{\operatorname{trace}(\Sigma)}{N}\frac{N}{T}\frac{1}{N}\operatorname{trace}\left(\left(\lambda I_N-\frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\right)+o_p(1)\end{aligned}$$

and last but not least

$$\frac{\lambda}{T}\left(U_1^\top W(F\Lambda^\top)\right)\left(\lambda I_N-\frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\left(U_1^\top We\right)^\top=o_p(1).$$

Note that $\frac{1}{\sqrt{N}}\epsilon^\top$ has orthogonally invariant column vectors by the properties of the normal distribution and Lemma A.2 in Benaych-Georges and Nadakuditi (2011) applies. In summary the limit value of κ_T is described by

$$\begin{aligned}\kappa_T(\lambda)=&\lambda\tilde{U}^\top\left(\begin{pmatrix}\Sigma_F & \Sigma_F^{1/2}\mu_F(1+\tilde{\gamma})\\ \mu_F^\top\Sigma_F^{1/2}(1+\tilde{\gamma}) & \mu_F^\top\mu_F(1+\gamma)\end{pmatrix}+\frac{c\cdot\operatorname{trace}(\Sigma)}{N}\begin{pmatrix}I_K & 0\\ 0 & 1+\gamma\end{pmatrix}\right)\tilde{U}\\ &\cdot\frac{1}{N}\operatorname{trace}\left(\left(\lambda_N-\frac{1}{T}e^\top WU_2U_2^\top We\right)^{-1}\right)+o_p(1).\end{aligned}$$

As $U_2We=U_2e=\tilde{e}$ with $\tilde{e}_t\overset{i.i.d.}{\sim}N(0,\Sigma)$ for $t=1,...,T-K-1$ we have

$$\kappa_T(\lambda)\xrightarrow{p}\kappa(\lambda)=\lambda\tilde{U}^\top M_{\text{RP-PCA}}\tilde{U}G(\lambda).$$

Therefore λ is eigenvalue of $\lambda\begin{pmatrix}\theta_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \theta_{K+1}\end{pmatrix}G(\lambda)$ which is equivalent to

$$G(\lambda)=\frac{1}{\theta_i}\qquad\text{respectively}\qquad\lambda=G^{-1}\left(\frac{1}{\theta_i}\right).$$

If a solution outside the support of the spectrum of S_{22} exists, then it must satisfy the equation $G(\lambda)=\frac{1}{\theta_i}$ for some $i=1,...,K+1$. Otherwise by Weil's inequality and the same arguments as in Benaych-Georges and Nadakuditi (2011) $\lambda\xrightarrow{b}b$. For $z>b$ we have $G'(z)<0$. Therefore if $\theta_i>\frac{1}{G(b)}$ then a solution exists. If $\theta_i<\frac{1}{G(b)}$ then no solution exists and $\lambda\xrightarrow{p}b$.

Recall that the estimators for the loadings and factors are defined as follows: $\hat{\Lambda}$ are the first K eigenvectors of $\frac{1}{T}X^\top W^2X$ and $\hat{F}=X\hat{\Lambda}$. For the proofs we will use an equivalent formulation. Denote by V the first K eigenvectors of $\frac{1}{T}U^\top WXX^\top WU$. Then $\hat{\Lambda}=X^\top WUVD_K^{-1/2}$, where D_K is a diagonal matrix with the first K largest eigenvalues of $\frac{1}{T}U^\top X^\top W^2XU$, i.e.

$$\frac{1}{T}V^\top U^\top WXX^\top WUV=D_K.$$

The factors estimator takes the form $\hat{F}=X\hat{\Lambda}=\sqrt{T}W^{-1}UVD_K^{1/2}$.

We analyze the $K+1$ eigenvectors of $\frac{1}{T}U^\top WXX^\top WU$. Assume u_i is an eigenvector of S associated with λ_i:

$$\begin{pmatrix} \lambda_i I_{K+1} - S_{11} & -S_{12} \\ -S_{21} & \lambda_i I_{T-K-1} - S_{22} \end{pmatrix} \begin{pmatrix} u_{i,1} \\ u_{i,2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where $u_{i,1}$ and $u_{i,2}$ are the first $K+1$ respectively last $T-K-1$ components of the vector u_i. Hence

$$\begin{aligned} u_{2,i} &= \left(\lambda_i I_{T-K-1} - S_{22}\right)^{-1} S_{21} u_{i,1} \\ 0 &= \left(\lambda_i I_{K+1} - \kappa_T(\lambda_i)\right) u_{i,1}. \end{aligned}$$

Assume that $\theta_i > \theta_{crit}$, i.e. $\lambda_i I_{K+1} - \kappa_T(\lambda_i) = 0$ has a solution. Consequently

$$\left(I_{K+1} - \theta_i^{-1} \begin{pmatrix} \theta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_{K+1} \end{pmatrix} \right) u_{i,1} = o_p(1).$$

As a consequence the vector $u_{i,1}$ has all elements equal to zero except at the ith position:

$$u_{i,1}^\top = \begin{pmatrix} 0 & \cdots 0 & \|u_{i,1}\| & 0 & \cdots & 0 \end{pmatrix}$$

where $\|u_{i,1}\|$ denotes the length of the vector which is completely determined by the ith element. The vector $u_{i,2}$ satisfies

$$\begin{aligned} u_{i,2}^\top u_{i,2} &= u_{i,1}^\top S_{12} \left(\lambda_i I_{T-K-1} - S_{22}\right)^{-2} S_{21} u_{i,1} \\ &= u_{i,1}^\top \frac{1}{T} U_1^\top W \left(F\Lambda^T + e\right) \left(e^\top W U_2 \left(\lambda_i I_{T-K-1} - S_{22}\right)^{-2} U_2^\top W e\right) \left(F\Lambda^T + e\right)^\top W U_1. \end{aligned}$$

By similar arguments as in the first part of the proof showing the convergence of $\kappa_T(\lambda)$ it follows that

$$\begin{aligned} u_{i,2}^\top u_{i,2} =& u_{i,1}^\top \begin{pmatrix} \theta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_{K+1} \end{pmatrix} u_{i,1} \\ & \cdot \operatorname{trace}\left(e^\top W U_2 \left(\lambda_i I_{T-K-1} - \frac{1}{T} U_2^\top W e e^\top W U_2 \right)^{-2} U_2^\top W e \right) + o_p(1). \end{aligned}$$

Recall that $U_2^\top We = \tilde{e}$ can be interpreted as $T-K-1$ independent draws of a $N(0,\Sigma)$. Denote the eigenvalue distribution function of $\frac{1}{T}\tilde{e}^\top\tilde{e}$ by $\phi_T(z)$ and of $\frac{1}{T}\tilde{e}\tilde{e}^\top$ by $\tilde{\phi}_T(z)$. By assumption both converge to limit spectral distribution functions that are related through $\tilde{\phi}(z) - c\phi(z) = (1-c)\delta_0$ where δ_0 is the Dirac-measure with point-mass at zero.[25] By the properties of the trace operator

$$\operatorname{trace}\left(e^\top W U_2 \left(\lambda_i I_{T-K-1} - \frac{1}{T} U_2^\top W e e^\top W U_2 \right)^{-2} U_2^\top W e \right) = \int \frac{z}{(\lambda_i - z)^2} d\tilde{\phi}_T(z)$$

[25] See Chapter 2 in Yao, Zheng and Bai (2015).

which converges almost surely to

$$\begin{aligned}\int \frac{z}{(\lambda_i - z)^2} d\tilde{\phi}(z) &= \int \frac{z}{(\lambda_i - z)^2} d(c\phi(z) + (1-c)\delta_0) \\ &= c \int \frac{z}{(\lambda_i - z)^2} d\phi(z) = B(\lambda_i).\end{aligned}$$

Consequently

$$1 = \|u_{i,1}\|^2 + \|u_{i,2}\|^2 = u_{i,1}^\top u_{i,1} \left(1 + \theta_i B(\lambda_i)\right) + o_p(1)$$

and therefore

$$\|u_{i,1}\|^2 \xrightarrow{p} \frac{1}{1 + \theta_i B(\lambda_i)}.$$

Assume that $\theta_i < \theta_{crit}$, i.e. $\lambda_i I_{K+1} - \kappa_T(\lambda_i) = 0$ has no solution. It still holds

$$u_{i,2}^\top u_{i,2} = u_{i,1}^\top \begin{pmatrix} \theta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_{K+1} \end{pmatrix} u_{i,1} \lim_{z \downarrow b} B(z)$$

as λ_i converges in probability to b. If $\lim_{z \downarrow b} B(z) = -\infty$, then $\|u_{i,1}\| \xrightarrow{p} 0$ and

$$u_{i,1}^\top = \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix}.$$

All we need to show is that $\theta_i < \theta_{crit}$ implies $\lim_{z \downarrow b} B(z) = -\infty$. This follows for the largest eigenvalue λ_1 by the same argument as in the proof of theorem 2.3 in Benaych-Georges and Nadakuditi (2011). If $K > 1$ we need in addition eigenvalue repulsion to show the result for λ_i for $i = 2, ..., K$ (see Nadakuditi (2014), appendix 7). Assume that the distance between the largest eigenvalues of the matrix $\frac{1}{T} e^\top e$ decays with a certain rate

$$\left| \lambda_{i+1}\left(\frac{e^\top e}{T}\right) - \lambda_i \left(\frac{e^\top e}{T}\right) \right| \leq O_p\left(\frac{\log(N)}{N^{2/3}}\right).$$

This is satisfied for normally distributed residuals as in our case (see Onatski (2012)). Hence,

$$\begin{aligned} B(\lambda_i) &= c \int \frac{z}{(\lambda_i - z)^2} d\tilde{\phi}_T(z) + o_p(1) \\ &\leq O_p\left(\frac{1}{N}\right) \cdot \frac{1}{(\lambda_1(S_{22}) - \lambda_{K+1}(S_{22}))^2} + o_p(1) \\ &\leq O_p\left(\frac{N^{1/3}}{\log(N)^2}\right). \end{aligned}$$

which satisfies the explosion condition.

We can now go back to the original problem: Define

$$\rho_i = \begin{cases} \frac{1}{\sqrt{1+\theta_i B(G^{-1}(\theta_i))}} & \text{if } \theta_i > \theta_{crit} \\ 0 & \text{otherwise.} \end{cases}$$

The estimator for the factors can now be written as

$$\begin{aligned} \hat{F} &= \sqrt{T} W^{-1} U V D_K^{1/2} \\ &= \sqrt{T} W^{-1} U_1 \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \rho_K \\ 0 & \cdots & & 0 \end{pmatrix} D_K^{1/2} \end{aligned}$$

with

$$D_K = \begin{pmatrix} \hat{\theta}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\theta}_K \end{pmatrix} \quad , \quad \hat{\theta}_i = \begin{cases} G^{-1}\left(\frac{1}{\theta_i}\right) & \text{if } \theta_i > \theta_{crit} \\ b & \text{otherwise.} \end{cases}$$

The calculation for $\widehat{\text{Corr}}(F, \hat{F})$ is straightforward. Note that the mean can be estimated by

$$\hat{\mu}_{\hat{F}} = \frac{1}{1+\tilde{\gamma}} \begin{pmatrix} O_K & \mathbb{1}_K \end{pmatrix} \tilde{U} \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \rho_K \\ 0 & \cdots & & 0 \end{pmatrix} D_K^{1/2}.$$

Here we used that $W^{-1} = I_T - \frac{\tilde{\gamma}}{1+\tilde{\gamma}} \mathbb{1}\mathbb{1}^\top$ and $(1+\tilde{\gamma})^2 = 1+\gamma$.

Proof for i.i.d. residuals:

For the special case where $e_{t,i}$ i.i.d. $N(0, \sigma_e^2)$, i.e. $\Sigma = \sigma_e^2 I_N$, the matrix $\frac{1}{T} e^\top e$ follows the Marcenko-Pasteur law:

$$d\phi(z) = \frac{1}{2\pi c \sigma_e^2 z} \sqrt{(b-z)(z-a)} \mathbb{1}_{\{z \in (a,b)\}} dz + \max\left(0, 1 - \frac{1}{c}\right) \delta_0$$

with

$$\begin{aligned} a &= \sigma_e^2 (1 - \sqrt{c})^2 \\ b &= \sigma_e^2 (1 + \sqrt{c})^2 \end{aligned}$$

a and b are the smallest respectively largest eigenvalue. For simplicity take $c > 1$, but the results can be easily extended to the case $0 < c < 1$. The object of interest is the Cauchy transform of the

eigenvalue distribution function. Calculations as outlined in Bai and Silverstein (2010) lead to

$$G(z) = \frac{z - \sigma_e^2(1-c) - \sqrt{(z - \sigma_e^2(1+c))^2 - 4c\sigma_e^2}}{2cz\sigma_e^2}.$$

Simple but tedious calculations show that

$$G^{-1}(z) = \frac{z\sigma_e^2(1-c) + 1}{z - c\sigma_e^2 z^2}.$$

Proof of Corollary 2: Plugging the eigenvalues and eigenvector formulas into Theorem 2 yields:

$$\widehat{\text{Corr}}(F, \hat{F}) \xrightarrow{p} \begin{pmatrix} 1 & 0 \end{pmatrix} \tilde{U} \begin{pmatrix} \rho_1 \\ 0 \end{pmatrix} \hat{\theta}_1^{1/2} \widehat{\text{Var}}(\hat{F})^{1/2}$$

$$\widehat{\text{Var}}(\hat{F}) \xrightarrow{p} \hat{\theta}_1 \left(\tilde{U}_{1,1}^2 \|u_{1,1}\|^2 + \|u_{1,2}\|^2 \right)$$

$$\hat{\mu}^2 \xrightarrow{p} \frac{1}{1+\gamma} \tilde{U}_{1,2}^2 \rho_1 \hat{\theta}_1.$$

The proof for the limit for $\gamma \to \infty$ is based on the insight that

$$\lim_{\theta \to \infty} B(\theta)\theta^2 \to c\sigma_e^2.$$

Lemma 2: Detection of weak factors
If $\gamma > -1$ and $\mu_F \neq 0$, then the first K eigenvalues of $M_{RP\text{-}PCA}$ are strictly larger than the first K eigenvalues of M_{PCA}, i.e.

$$\theta_i^{RP\text{-}PCA} > \sigma_{F_i}^2 + c\sigma_e^2.$$

For $\theta_i > \theta_{crit}$ it holds that

$$\frac{\partial \hat{\theta}_i}{\partial \theta_i} > 0 \qquad \frac{\partial \rho_i}{\partial \theta_i} > 0 \qquad i = 1, ..., K.$$

Thus, if $\gamma > -1$ and $\mu_F \neq 0$, then $\rho_i^{RP\text{-}PCA} > \rho_i^{PCA}$.

Proof of Lemma 2:
See result (12) on page 75 in Lütkepohl (1996) and straightforward calculations.

References

Ahn, S. C., Horenstein, A. R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.

Aït-Sahalia, Y., Xiu, D., 2017. Principal component estimation of a large covariance matrix with high-frequency data. Journal of Econometrics 201, 384–399.

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.

Bai, J., Ng, S., 2008. Large dimensional factor analysis. Foundations and Trends in Econometrics 3 (2), 89–163.

Bai, J., Ng, S., 2017. Principal components and regularized estimation of factor models. Working Paper.

Benaych-Georges, F., Nadakuditi, R. R., 2011. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Advances in Mathematics 227, 494–521.

Bryzgalova, S., 2017. 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.

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.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society 75 (4), 603–680.

Fan, J., Liao, Y., Wang, W., 2016. Projected principal component analysis in factor models. The Annals of Statistics 44 (1), 219–254.

Fan, J., Zhong, Y., 2018. Optimal subspace estimation using overidentifying vectors via generalized method of moments. Working paper.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. Review 82, 540–554.

Harding, M., 2013. Estimating the number of factors in large dimensional factor models. Working paper.

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

Kozak, S., Nagel, S., Santosh, S., 2017. Shrinking the cross section. Technical Report, Chicago Booth.

Lettau, M., Pelger, M., 2018. Factors that fit the time series and cross-section of stock returns. Working paper.

Ludvigson, S., Ng, S., 2010. A factor analysis of bond risk premia. Handbook of the Economics of Finance.

Lütkepohl, H., 1996. Handbook of Matrices. John Wiley & Sons.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Review of Economic and Statistics 92, 1004–1016.

Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics (168), 244–258.

Paul, D., 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica 17 (4), 1617–1642.

Pelger, M., 2017. Large-dimensional factor modeling based on high-frequency observations large-dimensional factor modeling based on high-frequency observations. Working paper.

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

Stock, J., Watson, M., 2006. Macroeconomic Forecasting Using Many Predictors. Handbook of Economic Forecasting. North Holland.