Interpretable Proximate Factors for Large Dimensions

Markus Pelger [1] Ruoxuan Xiong [2]

[1]Stanford University

[2]Stanford University

February 1, 2018

Risk Management Seminar
UC Berkeley

Motivation: What are the factors?

Statistical Factor Analysis

- Factor models are widely used in big data settings
    - Reduce data dimensionality
    - Factors are traded extensively
    - Problem: Which factors should be used?
- Statistical (latent) factors perform well
    - Factors estimated from principle component analysis (PCA)
    - Weighted averages of all features/assets
    - Problem: Hard to interpret

Goals of this paper:

Create interpretable proximate factors

- Shrink most assets' weights to zero to get proximate factors

⇒ More interpretable

⇒ Significantly lower transaction costs when trading factors

Contribution of this paper

Contribution

- This Paper: Estimation of interpretable proximate factors
- Key elements of estimator:
  1. Statistical factors instead of pre-specified (and potentially miss-specified) factors
  2. Uses information from large panel data sets: Many assets with many time observations
  3. Proximate factors approximate latent factors very well with a few assets without sparse structure in population factors
  4. Only 5-10% of the cross-sectional observations with the largest exposure are needed for proximate factors

Contribution

Theoretical Results

- Asymptotic probabilistic lower bound for generalized correlations of proximate factors with population factors
- Guidance on how to construct proximate factors

Empirical Results

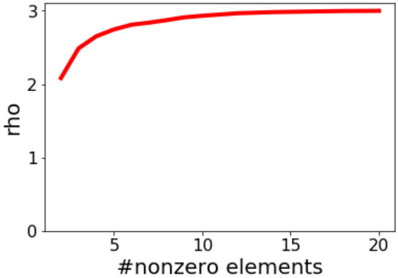
- Very good approximation to population factors with 5-10% portfolios, measured by generalized correlation, variance explained, pricing error and Sharpe-ratio
- Interpret statistical latent factors for
  - Double-sorted portfolio data
  - 370 single-sorted anomaly portfolios
  - High-frequency returns of S&P 500 companies

Literature (partial list)

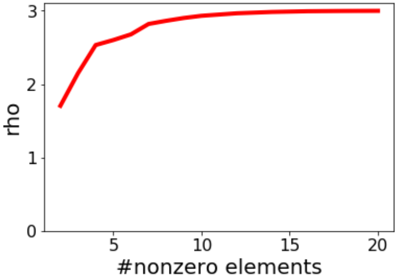
- Large-dimensional factor models with PCA
  - Bai (2003): Distribution theory
  - Fan et al. (2013): Sparse matrices in factor modeling
  - Fan et al. (2016): Projected PCA for time-varying loadings
  - Pelger (2016), Aït-Sahalia and Xiu (2015): High-frequency
- Large-dimensional factor models with penalty term
  - Bai and Ng (2017): Robust PCA with ridge shrinkage
  - Lettau and Pelger (2017): Risk-Premium PCA with pricing penalty
  - Zhou et al. (2006): Sparse PCA

Empirical example: Double-sorted portfolios

Daily data of 25 double-sorted Fama-French portfolios



(a) Size and Book-to-Market



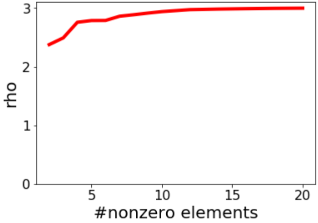
(b) Size and Investment

Figure: Sum of generalized correlation $\hat{\rho}$ between estimated 3 PCA factors and 3 proximate factors

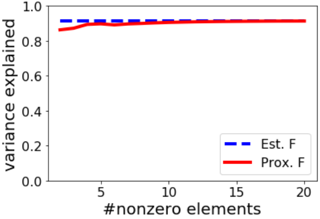
- Problem in interpreting factors: Factors only identified up to invertible linear transformations.
- Generalized correlation measures how many factors two sets have in common.

Empirical Application: Size and Book-to-market Portfolios

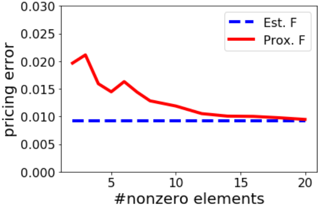
- 25 portfolios formed on size and book-to-market (07/1963-10/2017, 3 factors, daily data)



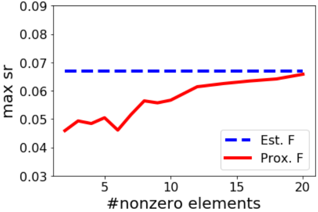
(a) Generalized correlation



(b) Variance explained



(c) RMS pricing error



(d) Max Sharpe Ratio

Empirical Application: Size and Book-to-market Portfolios

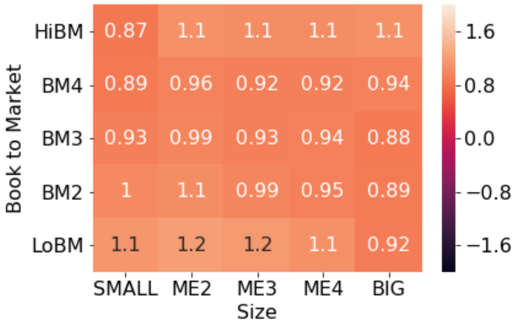


Figure: Portfolio weights of 1. statistical factor

⇒ Equally weighted market factor

Empirical Application: Size and Book-to-market Portfolios

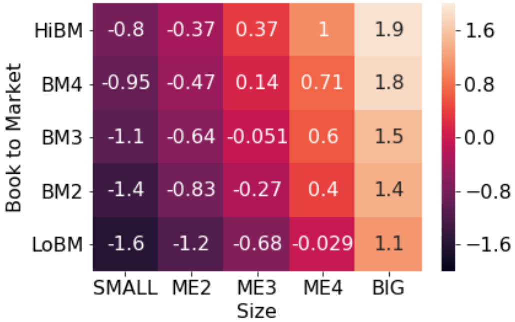


Figure: Portfolio weights of 2. statistical factor

$\Rightarrow$ Small-minus-big size factor

$\Rightarrow$ Proximate factor with 4 largest weights correlation 0.88 with size factor

Empirical Application: Size and Book-to-market Portfolios

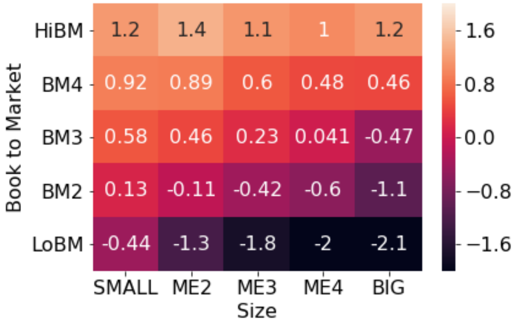


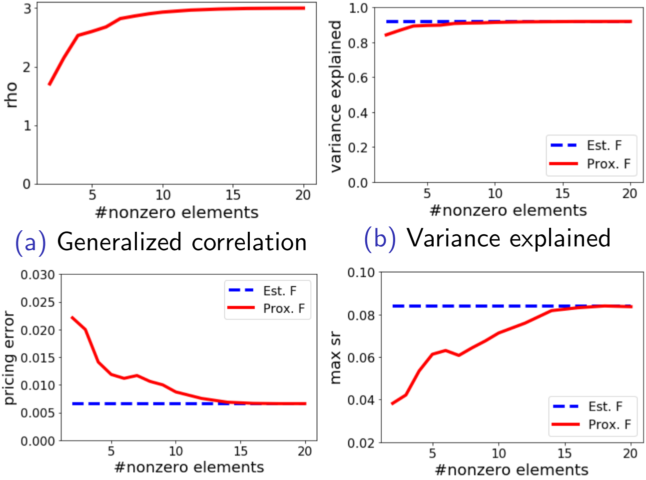
Figure: Portfolio weights of 3. statistical factor

⇒ High-minus-low value factor

⇒ Proximate factor with 4 largest weights correlation 0.91 with value factor

Empirical Application: Size and Investment Portfolios

- 25 portfolios formed on size and investment (07/1963-10/2017, 3 factors, daily data)



(a) Generalized correlation

(b) Variance explained

(c) RMS pricing error

(d) Max Sharpe Ratio

Empirical Application: Size and Investment Portfolios

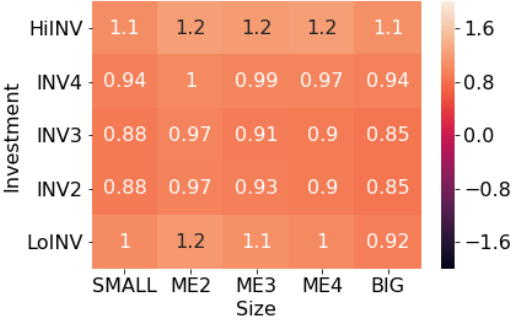


Figure: Portfolio weights of 1. statistical factor

$\Rightarrow$ Equally weighted market factor

Empirical Application: Size and Investment Portfolios

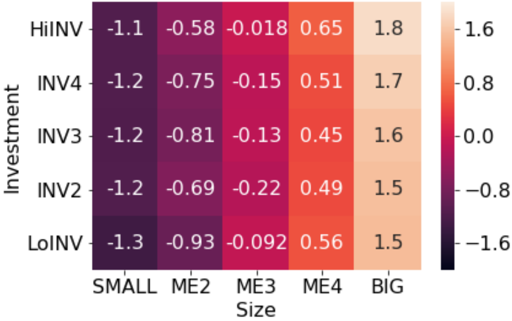


Figure: Portfolio weights of 2. statistical factor

⇒ Small-minus-big size factor

⇒ Proximate factor with 4 largest weights correlation 0.97 with size factor

Empirical Application: Size and Investment Portfolios

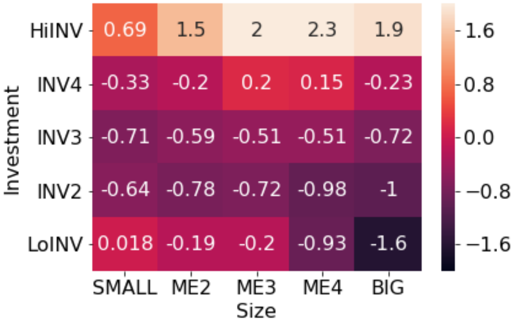


Figure: Portfolio weights of 3. statistical factor

⇒ High-minus-low value factor

⇒ Proximate factor with 4 largest weights correlation 0.79 with investment factor

The Model

Approximate Factor Model

- Observe excess returns of N assets over T time periods:

$$X_{t,i} = \underbrace{\underset{1\times K}{F_t}^{\top}}_{factors} \underbrace{\underset{K\times 1}{\Lambda_i}}_{loadings} + \underbrace{e_{t,i}}_{idiosyncratic} \qquad i = 1, ..., N \quad t = 1, ..., T$$

- Matrix notation

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

  - N assets (large)
  - T time-series observation (large)
  - K systematic factors (fixed)
- F, Λ and e are unknown

The Model

Approximate Factor Model

- Systematic and non-systematic risk (F and e uncorrelated):

$$Var(X) = \underbrace{\Lambda Var(F)\Lambda^\top}_{systematic} + \underbrace{Var(e)}_{non-systematic}$$

$\Rightarrow$ Systematic factors should explain a large portion of the variance

$\Rightarrow$ Idiosyncratic risk can be weakly correlated

Estimation: PCA (Principal component analysis)

- Apply PCA to the sample covariance matrix: $\frac{1}{T}X^\top X - \bar{X}\bar{X}^\top$ with $\bar{X}$ = sample mean of asset excess returns
- Eigenvectors of largest eigenvalues estimate loadings $\hat{\Lambda}$.
- $\hat{F}$ estimator for factors: $\hat{F} = \frac{1}{N}X\hat{\Lambda} = X\hat{\Lambda}^\top(\hat{\Lambda}^\top\hat{\Lambda})^{-1}$.

The Model

Proximate Factors

- Sparse loadings $\tilde{\Lambda}$ are obtained from
  - Select finitely many m_N loadings with largest absolute value from $\hat{\Lambda}_k$
  - Shrink estimated loadings $\hat{\Lambda}$ to 0 except for m_N largest values
  - Divide by column norms, i.e. $\tilde{\lambda}_k^\top \tilde{\lambda}_k = 1$
- Proximate factors $\tilde{F} = X^T \tilde{\Lambda}$

The Model

Closeness measure

- For 1-factor model: Correlation between $\tilde{F}$ and F.
- Problem for multiple factors: Factors are only identified up to invertible linear transformations $\Rightarrow$ Need measure for closeness between span of two vector spaces
- For multi-factor model: The "closeness" between $\tilde{F}$ and F is measured by generalized correlation:
  - Total generalized correlation measure:
    $$\rho = trace\left((F^T F/T)^{-1}(F^T \tilde{F}/T)(\tilde{F}^T \tilde{F}/T)^{-1}(\tilde{F}^T F/T)\right)$$
    - $\rho = 0$: $\tilde{F}$ and F are orthogonal
    - $\rho = K$: $\tilde{F}$ and F are span the same space
  - Alternative measure: Element-wise generalized correlations are eigenvalues instead of trace of above matrix
    - Element-wise generalized correlations close to 1 measure how many factors are well approximated

Intuition: Why picking largest elements in $\hat{\Lambda}$ works?

- Consider one factor and one nonzero element in $\tilde{\Lambda}$:
  $F = [f_{1t}] \in \mathbb{R}^{T\times 1}$, $\Lambda = [\lambda_{1,i}] \in \mathbb{R}^{N\times 1}$
- $\tilde{\Lambda} = [\tilde{\lambda}_{1,i}]$ is sparse. Assume nonzero element in $\tilde{\lambda}_{1,i}$ is $\tilde{\lambda}_{1,1}$.

$$
\begin{aligned}
\tilde{F} &= X^T\tilde{\Lambda} = F\Lambda^T\tilde{\Lambda} + e^T\tilde{\Lambda} \\
&= f_1\lambda_{1,1} + e_1
\end{aligned}
$$

- Assume

$$
f_{1,t} \sim (0, \sigma_f^2), \qquad e_{1,t} \overset{iid}{\sim} (0, \sigma_e^2)
$$

$$
\frac{f_1^T f_1}{T} \to \sigma_f^2, \qquad \frac{e_1^T e_1}{T} \to \sigma_e^2
$$

- Define signal-to-noise ratio $s = \frac{\sigma_f}{\sigma_e}$

Intuition: Why pick the largest elements in $\hat{\Lambda}$?

$$
\begin{aligned}
\rho &= tr\left((F^T F/T)^{-1}(F^T \tilde{F}/T)(\tilde{F}^T \tilde{F}/T)^{-1}(\tilde{F}^T F/T)\right) \\
&= \left(\frac{f_1^T(f_1\lambda_{1,1}+e_1)/T}{(f_1^T f_1/T)^{1/2}((f_1\lambda_{1,1}+e_1)^T(f_1\lambda_{1,1}+e_1)/T)^{1/2}}\right)^2 \\
&\to \frac{\lambda_{1,1}^2}{\lambda_{1,1}^2+1/s^2}
\end{aligned}
$$

- (Generalized) correlation increases in size of loading $|\lambda_{1,1}|$.
- (Generalized) correlation increases in signal-to-noise ratio s.
- No sparsity in population loadings assumed!

Asymptotic results

- Proximate factors $\tilde{F}$ are in general not consistent.

$$\tilde{F} = X^T \tilde{\Lambda} = F \Lambda^T \tilde{\Lambda} + e^T \tilde{\Lambda}$$

  - Idiosyncratic component not diversified away
  - Assume $e_{i,l} \overset{iid}{\sim} (0, \sigma^2_{e.,l})$, then each element in $e^T \tilde{\Lambda}$ has

$$Var\left(\sum_{i=1}^{m_N} \tilde{\lambda}_{j,j_i} e_{j_i,l} \right) = \sum_{i=1}^{m_N} \tilde{\lambda}^2_{j,j_i} \sigma^2_{e.,l} = \sigma^2_{e.,l} \not\to 0$$

- Instead we provide probabilistic lower bound for (generalized) correlation ρ given a target correlation level ρ_0:

$$P(\rho > \rho_0)$$

Assumptions

Assumptions

1. **Factors:** Uncorrelated and demeaned factors:

$$E[F] = 0 \qquad \frac{F^\top F}{T} \to \Sigma_F = diag(\sigma^2_{f_1}, \sigma^2_{f_2}, \cdots, \sigma^2_{f_r})$$

2. **Loadings:** Random variables $\lambda_{i,j} = O_p(1)$ and $\Lambda^\top \Lambda \to \Sigma_\Lambda$
3. **Systematic factors:** Eigenvalues of $\Sigma_\Lambda \Sigma_F$ bounded away from 0.
4. **Residuals:** Weak Dependency
   - $E[e_{i,l}] = 0$ and $Var(e_{i,l}) \leq \sigma_e^2 \; \forall i, l$
   - e independent from F and Λ
   - $\frac{1}{\sqrt{T}} e_{(i)}^T e_{(k)} = O_p(1) \; \forall i, k$ and $i \neq k$
5. **Consistent estimator:**

$$\hat{f}_j - Hf_j = O_p\left(\frac{1}{\sqrt{N}}\right) \qquad \hat{\lambda}_i - H^{-1}\lambda_i = O_p\left(\frac{1}{\sqrt{T}}\right) \qquad N, T \to \infty$$

Sufficient conditions in Bai (2003) and Bai and Ng (2002)

One factor case

Theorem

Assume $K = 1$ factor and population loadings $\lambda_{1,i}$ are i.i.d for all i. For any ρ_0 we have for $N, T \to \infty$

$$P(\rho > \rho_0) \geq 1 - \sum_{j=0}^{m_N - 1} \binom{N}{j} (1 - \mathbb{F}_{|\lambda_{1,i}|}(y_{m_N}))^j \mathbb{F}_{|\lambda_{1,i}|}(y_{m_N})^{N-j} \quad (1)$$

where

$$y_{m_N} = \sqrt{\frac{1}{m_N} \frac{\sigma_e^2}{\sigma_{f_1}^2} \frac{\rho_0}{1 - \rho_0}}$$

$$\mathbb{F}_{|\lambda_{1,i}|}(y) = P(|\lambda_{1,i}| \leq y)$$

One factor case

- Denote the lower probability bound for $P(\rho > \rho_0)$ by $\underline{p} = 1 - \sum_{j=0}^{m_N-1} \binom{N}{j}(1 - \mathbb{F}_{|\lambda_{1,i}|}(y_{m_N}))^j \mathbb{F}_{|\lambda_{1,i}|}(y_{m_N})^{N-j}$
- It holds,

$$\frac{\partial \underline{p}}{\partial \mathbb{F}_{|\lambda_{1,i}|}(y_{m_N})} < 0$$

- $\underline{p}$ is decreasing in $\mathbb{F}_{|\lambda_{1,i}|}(y_{m_N})$. Hence $\underline{p}$ is
    - decreasing in ρ_0
    - increasing in $s = \sigma_{f_1}/\sigma_e$
    - increasing in m_N
    - increasing in the dispersion of the distribution of $|\lambda_{1,i}|$

Multiple Factors

Multiple Factor: Simple Case

- Denote by $\{j_1, j_2, \cdots, j_{m_N}\}$ indexes of nonzero entries in $\tilde{\lambda}_j$ (i.e. largest m_N entries in $\hat{\lambda}_j$ in absolute value).
- Let U be the "sparse" rotated population loadings $\Lambda H \in R^{N\times k}$ with non-zero entries $\{j_1, j_2, \cdots, j_{m_N}\}$.
- Assume U columns do not overlap
- Let $v_{j,(m_N)} = \min(|u_{j,j_1}|, |u_{j,j_2}|, \cdots, |u_{j,j_{m_N}}|)$ to be the m_N-th order statistic of $|u_j|$

For any threshold ρ_0 and for $N, T \to \infty$ we have

$$P(\rho > \rho_0) \geq P\left(\sum_{j=1}^{k} \frac{1}{s_j v_{j,(m_N)}^2} \leq \frac{m_N(K - \rho_0)}{\sigma_e^2}\right)$$

Multiple Factors

Multiple Factor: Threshold and then rotate

- Denote by $\{j_1, j_2, \cdots, j_{m_N}\}$ indices of nonzero entries in $\tilde{\lambda}_j$
- Let $\breve{\Lambda}$ be the "sparse" population loadings ΛH with non-zero entries $\{j_1, j_2, \cdots, j_{m_N}\}$.
- Assume there exists orthonormal matrix P s.t. $\breve{\Lambda}P$ columns do not overlap
- Signal matrix S is diagonal matrix of the eigenvalues of $\Sigma_\Lambda \Sigma_F$ in decreasing order
- Define $[w^P_{M,1}, w^P_{M,2}, \cdots, w^P_{M,k}]$ as normalized elements of $\breve{\Lambda}S^{1/2}P$
- Let $w^P_{j,(m_N)} = \min(|w^P_{j,j_1}|, |w^P_{j,j_2}|, \cdots, |w^P_{j,j_{m_N}}|)$ to be the m_N-th order statistic of $|w^P_j|$

For any threshold ρ_0 and for $N, T \to \infty$ we have

$$P(\rho > \rho_0) \geq P\left(\sum_{j=1}^{K} \frac{1}{(w^P_{j,(m_N)})^2} \leq \frac{m_N(1-\gamma)(K-\rho_0)}{\sigma_e^2(1+\epsilon)^4}\right)$$

with known constants c and ϵ and γ.

Multiple Factors

Multiple Factor: Rotate and threshold

- Similar to previous theorem, but first find a rotation of the data and then threshold such that columns of sparse loadings to not overlap

For any threshold ρ_0 and for $N, T \to \infty$ we have

$$P(\rho > \rho_0) \geq P\left(\sum_{j=1}^{K} \frac{1}{(w_{j,(m_N)}^P)^2} \leq \frac{m_N(1-\gamma)(K-\rho_0)}{\sigma_e^2} \right)$$

with known constants c and ϵ and γ.

Multiple Factors

- Denote the lower probability bound for $P(\rho > \rho_0)$ by $\underline{p}$
- It holds (very similar to the one factor case) that $\underline{p}$ is
  - decreasing in ρ_0
  - increasing in $s = \sigma_{f_1}/\sigma_e$
  - increasing in m_N
  - increasing in the dispersion of the distribution of $|\lambda_{1,i}|$

Relationship with Lasso

Alternative approach with Lasso:

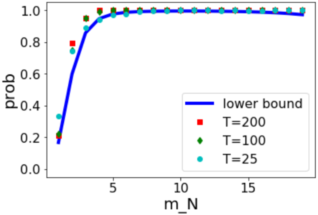
1. Estimate factors by PCA, i.e $X^T X \hat{F} = \hat{F} V$ with V matrix of eigenvalues.
2. Estimate loadings by $\left\| X - \Lambda \hat{F}^T \right\|_F^2 + \alpha \left\| \Lambda \right\|_1$. Divide the minimizer by its column norm (standardize each loading) to obtain $\bar{\Lambda}$
3. Proximate factors from Lasso approach are $\bar{F} = X^T \bar{\Lambda} (\bar{\Lambda}^T \bar{\Lambda})^{-1}$

$\Rightarrow$ Same selection of non-zero elements (for one factor case) but different weighting

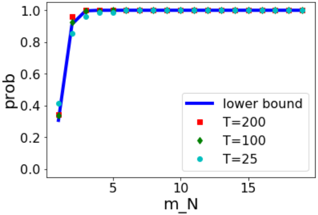
$\Rightarrow$ Under certain conditions worse performance than thresholding approach

- Tuning parameter less transparent

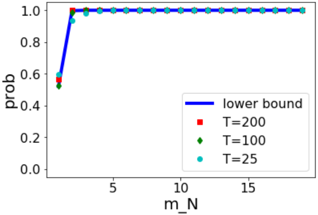
Simulation: One Factor ($\sigma_e = 1, \lambda \sim N(0,1)$, 500 MCs)



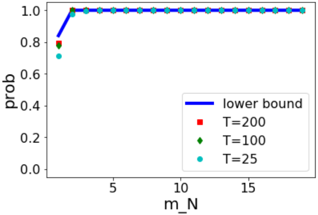
(a) N=50



(b) N=100



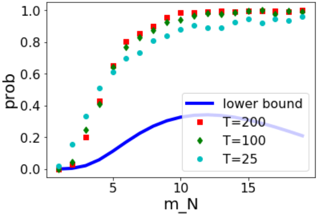
(c) N=200



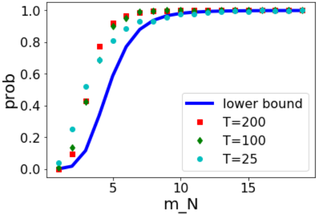
(d) N=500

Figure: $\sigma_f = 1.5$, $\rho_0 = 0.95$

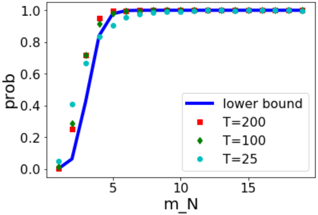
Simulation: One Factor ($\sigma_e = 1, \lambda \sim N(0,1)$, 500 MCs)



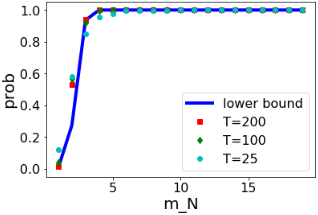
(a) N=50



(b) N=100



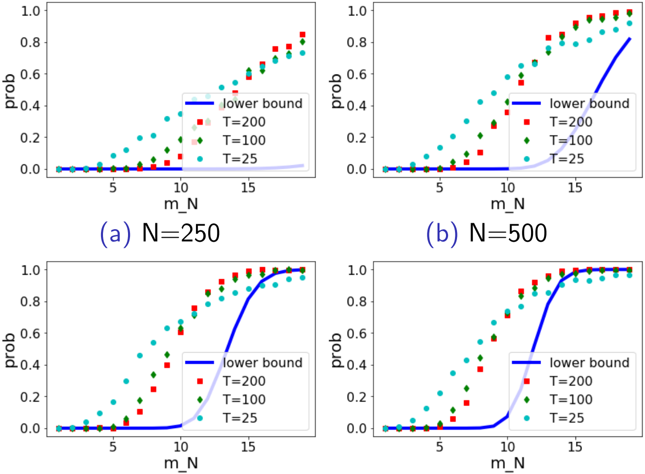
(c) N=200



(d) N=500

Figure: $\sigma_f = 1.0$, $\rho_0 = 0.95$

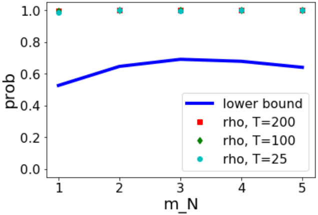
Simulation: One Factor ($\sigma_e = 1, \lambda \sim N(0, 1)$, 500 MCs)



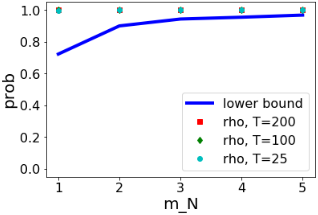
(a) N=250 (b) N=500 (c) N=750 (d) N=1000

Figure: $\sigma_f = 0.5$, $\rho_0 = 0.95$

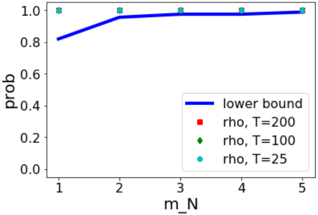
Simulation: Two Factors ($\sigma_e = 1$, $\lambda \sim N(0,1)$, 500 MCs)



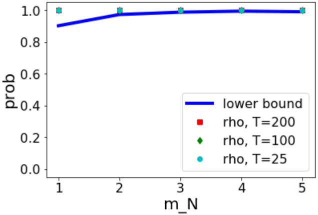
(a) N=50



(b) N=100



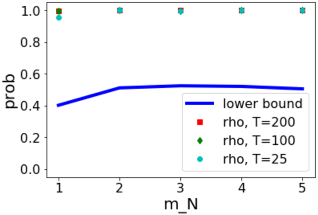
(c) N=200



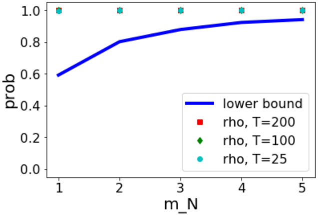
(d) N=500

Figure: $\sigma_f = 2.0$, $\rho_0 = 1.8$

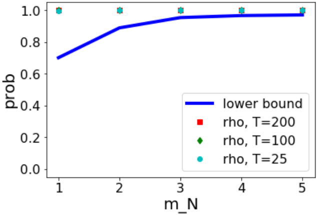
Simulation: Two Factors ($\sigma_e = 1$, $\lambda \sim N(0,1)$, 500 MCs)



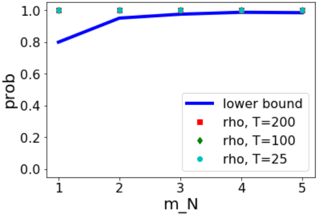
(a) N=100



(b) N=200



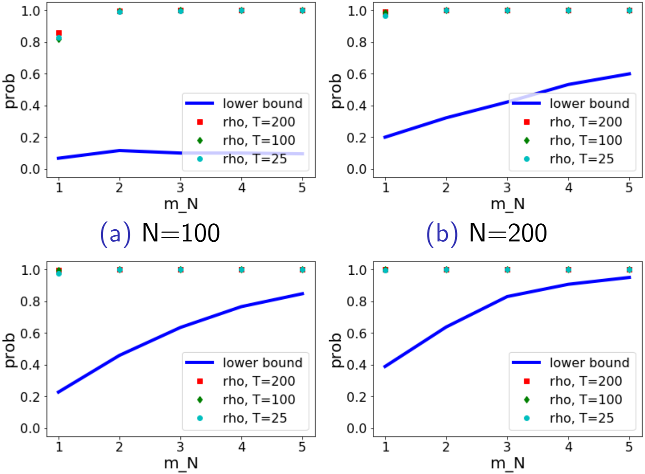
(c) N=300



(d) N=500

Figure: $\sigma_f = 1.5$, $\rho_0 = 1.7$

Simulation: Two Factors ($\sigma_e = 1$, $\lambda \sim N(0,1)$, 500 MCs)



(a) N=100 (b) N=200

(c) N=300 (d) N=500

Figure: $\sigma_f = 1.0$, $\rho_0 = 1.6$

Extreme deciles of single-sorted portfolios

Portfolio Data

- Kozak, Nagel and Santosh (2017) data: 370 decile portfolios sorted according to 37 anomalies
- Monthly return data from 07/1963 to 12/2016 ($T = 638$)
- First only lowest and highest decile portfolio for each anomaly ($N = 74$).
- Risk-Premium PCA (RP-PCA) from Lettau and Pelger (2017) applies PCA to $\frac{1}{T}X^\top X + \gamma \bar{X}\bar{X}^\top \Rightarrow$ penalty for pricing error
- Factors:
  1. **RP-PCA**: $K = 6$ and $\gamma = 100$.
  2. **PCA**: $K = 6$
  3. **Fama-French 5:** The five factor model of Fama-French (market, size, value, investment and operating profitability).
  4. **Proxy factors**: RP-PCA and PCA factors approximated with 8 largest positions.

Extreme Deciles

	In-sample			Out-of-sample		
	SR	RMS α	Idio. Var.	SR	RMS α	Idio. Var.
RP-PCA	**0.64**	**0.18**	**3.59**	**0.53**	**0.15**	**4.23**
PCA	0.35	0.22	3.57	0.28	0.19	4.24
RP-PCA Proxy	0.62	0.19	4.08	0.48	0.17	4.19
PCA Proxy	0.37	0.22	3.77	0.315	0.18	4.20
Fama-French 5	0.32	0.30	7.31	0.31	0.262	6.40

Table: First and last decile of 37 single-sorted portfolios from 07/1963 to 12/2016 ($N = 74$ and $T = 638$): Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation. $K = 6$ statistical factors.

- Proximate factors approximate latent factors very well
- Results hold out-of-sample.

Interpreting factors: Generalized correlations with proxies

	RP-PCA	PCA
1. Gen. Corr.	1.00	1.00
2. Gen. Corr.	1.00	1.00
3. Gen. Corr.	0.98	0.99
4. Gen. Corr.	0.96	0.97
5. Gen. Corr.	0.88	0.95
6. Gen. Corr.	0.72	0.89

Table: Generalized correlations of statistical factors with proxy factors (portfolios of 8 assets).

- Generalized correlations close to 1 measure of how many factors two sets have in common.
- Total generalized correlation ρ sum of element-wise generalized correlations

$\Rightarrow$ Proxy factors approximate statistical factors well.

Extreme Deciles: Maximal Sharpe-ratio

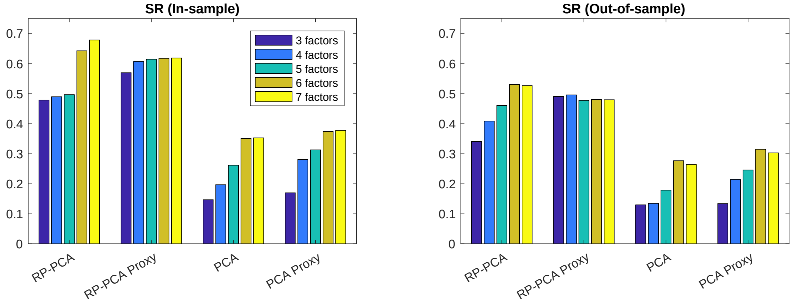


Figure: Maximal Sharpe-ratios.

⇒ Spike in Sharpe-ratio for 6 factors

⇒ Proximate factors capture similar Sharpe-ratio pattern

Extreme Deciles: Pricing error

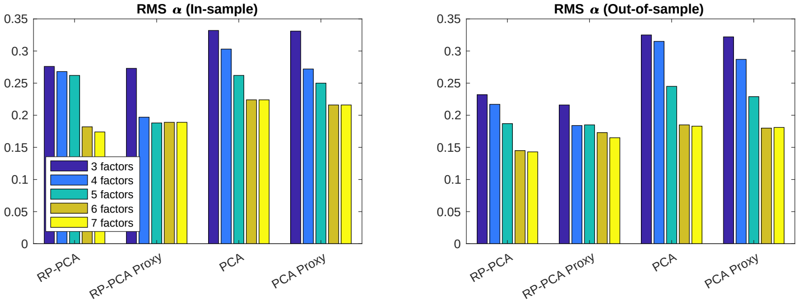


Figure: Root-mean-squared pricing errors.

$\Rightarrow$ RP-PCA has smaller out-of-sample pricing errors

$\Rightarrow$ Proximate factors have similar pricing errors

Extreme Deciles: Idiosyncratic Variation

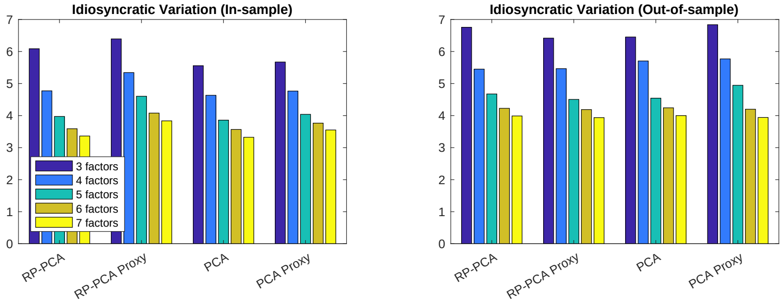


Figure: Unexplained idiosyncratic variation.

$\Rightarrow$ Unexplained variation similar for RP-PCA and PCA

$\Rightarrow$ Proximate factors explain the same variation

Interpreting factors: Composition of proxies

2. Proxy (RP-PCA)		3. Proxy (RP-PCA)		4. Proxy (RP-PCA)		5. Proxy (RP-PCA)		6. Proxy (RP-PCA)	
indrrevlv10	0.54	valmomprof10	0.17	mom1210	0.28	mom1210	-0.28	price1	0.38
indmomrev10	0.52	indmomrev10	-0.20	mom10	0.26	mom10	-0.28	mom1	0.36
ivol10	0.24	ivol10	-0.21	valuem1	0.25	valmomprof10	-0.29	valuem10	0.34
Accrual1	-0.21	mom121	-0.23	lrrev10	-0.24	roea1	-0.32	indrrev10	0.32
shvol1	-0.22	indrrevlv10	-0.26	mom1	-0.30	shvol1	-0.33	indrrev1	-0.26
ep1	-0.22	indmomrev1	-0.40	valuem10	-0.44	price1	-0.37	valmom10	-0.27
indrrev1	-0.25	indrrevlv1	-0.41	price1	-0.45	size10	-0.42	indmom10	-0.29
mom121	-0.42	ivol1	-0.67	mom121	-0.49	noa10	-0.47	ivol1	-0.53
2. Proxy (PCA)		**3. Proxy (PCA)**		**4. Proxy (PCA)**		**5. Proxy (PCA)**		**6. Proxy (PCA)**	
ivol1	0.59	valuem10	0.46	mom10	0.36	divp10	0.30	valprof10	0.33
indrrevlv10	0.43	price1	0.38	indmom10	0.35	roea1	-0.25	Aturnover10	0.32
indmomrev10	0.37	divp10	0.37	mom1210	0.34	shvol1	-0.27	sp10	0.27
indrrevlv1	0.36	value10	0.36	valmomprof10	0.33	size10	-0.28	prof10	0.24
indmomrev1	0.36	sp10	0.32	valmom1	-0.31	mom1	-0.29	valprof1	-0.25
ivol10	0.27	lrrev10	0.31	indmom1	-0.35	noa10	-0.37	prof1	-0.40
mom121	0.03	cfp10	0.31	mom121	-0.39	mom121	-0.38	ivol1	-0.42
indmom1	0.03	valuem1	-0.29	mom1	-0.39	price1	-0.57	Aturnover1	-0.51

Table: Portfolio-composition of proxy factors for first and last decile of 37 single-sorted portfolios: First proxy factors is an equally-weighted portfolio.

Interpreting factors: Cumulative absolute proxy weights

RP-PCA Proxy		PCA Proxy	
Momentum (12m)	1.70	Idiosyncratic Volatility	1.28
Idiosyncratic Volatility	1.65	Momentum (12m)	1.14
Industry Rel. Rev. (L.V.)	1.21	Momentum (6m)	1.04
Momentum (6m)	1.21	Price	0.95
Price	1.21	**Asset Turnover**	0.83
Industry Mom. Reversals	1.11	Industry Rel. Rev. (L.V.)	0.79
Value (M)	1.03	Value (M)	0.75
Industry Rel. Reversals	0.84	Industry Momentum	0.73
Share Volume	0.55	Industry Mom. Reversals	0.73
Net Operating Assets	0.47	Dividend/Price	0.67
Value-Momentum-Prof.	0.46	**Gross Profitability**	0.64
Size	0.42	**Sales/Price**	0.58
Return on Book Equity (A)	0.32	Value-Profitability	0.58
Industry Momentum	0.29	Net Operating Assets	0.37
Value-Momentum	0.27	Value (A)	0.36
Long Run Reversals	0.24	Value-Momentum-Prof.	0.33
Earnings/Price	0.22	Cash Flows/Price	0.31

Single-sorted portfolios

Portfolio Data

- Monthly return data from 07/1963 to 12/2016 ($T = 638$) for $N = 370$ portfolios
- Kozak, Nagel and Santosh (2017) data: 370 decile portfolios sorted according to 37 anomalies
- Risk-Premium PCA (RP-PCA) from Lettau and Pelger (2017) applies PCA to $\frac{1}{T}X^\top X + \gamma \bar{X}\bar{X}^\top \Rightarrow$ penalty for pricing error
- Factors:
  1. **RP-PCA**: $K = 6$ and $\gamma = 100$.
  2. **PCA**: $K = 6$
  3. **Fama-French 5:** The five factor model of Fama-French (market, size, value, investment and operating profitability, all from Kenneth French's website).
  4. **Proxy factors**: RP-PCA and PCA factors approximated with 5% of largest position.

Single-sorted portfolios

	In-sample			Out-of-sample		
	SR	RMS α	Idio. Var.	SR	RMS α	Idio. Var.
RP-PCA	**0.66**	**0.15**	**2.73**	**0.53**	**0.11**	**3.19**
PCA	0.28	0.15	2.70	0.22	0.14	3.19
Fama-French 5	0.32	0.23	4.97	0.31	0.21	4.62
RP-PCA Proxy 6	0.57	0.16	2.84	0.46	0.13	3.15
PCA Proxy 6	0.34	0.14	2.80	0.28	0.13	3.12

Table: Deciles of 37 single-sorted portfolios from 07/1963 to 12/2016 ($N = 370$ and $T = 638$): Maximal Sharpe-ratios, root-mean-squared pricing errors and unexplained idiosyncratic variation. $K = 6$ statistical factors.

- Proximate factors approximate latent factors very well
- Results hold out-of-sample.

Interpreting factors: Generalized correlations with proxies

	RP-PCA	PCA
1. Gen. Corr.	1.00	1.00
2. Gen. Corr.	1.00	1.00
3. Gen. Corr.	0.99	0.99
4. Gen. Corr.	0.98	0.99
5. Gen. Corr.	0.92	0.94
6. Gen. Corr.	0.78	0.89

Table: Generalized correlations of statistical factors with proxy factors (portfolios of 5% of assets).

- Generalized correlations close to 1 measure of how many factors two sets have in common.
- Total generalized correlation ρ sum of element-wise generalized correlations

$\Rightarrow$ Proxy factors approximate statistical factors well.

Single-sorted portfolios: Maximal Sharpe-ratio

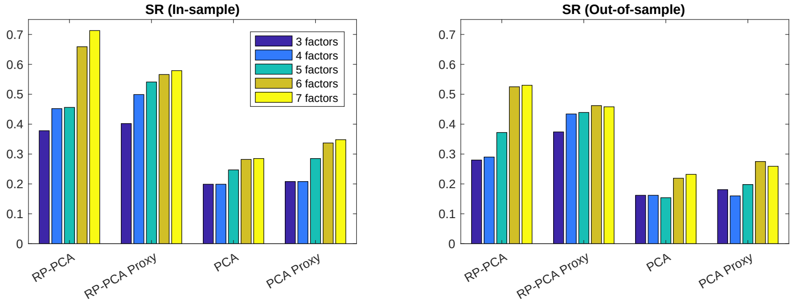


Figure: Maximal Sharpe-ratios.

⇒ Spike in Sharpe-ratio for 6 factors

⇒ Proximate factors capture similar Sharpe-ratio pattern

Single-sorted portfolios: Pricing error

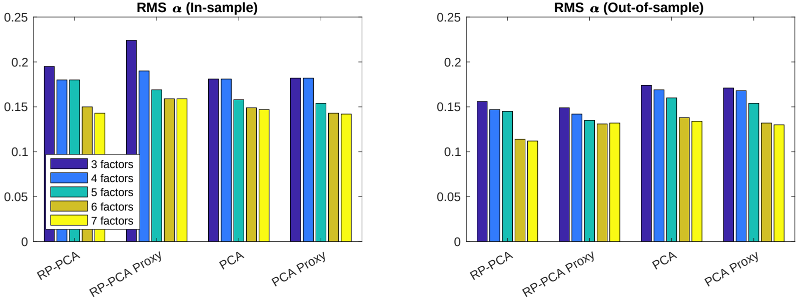


Figure: Root-mean-squared pricing errors.

⇒ RP-PCA has smaller out-of-sample pricing errors

⇒ Proximate factors have similar pricing errors

Single-sorted portfolios: Idiosyncratic Variation

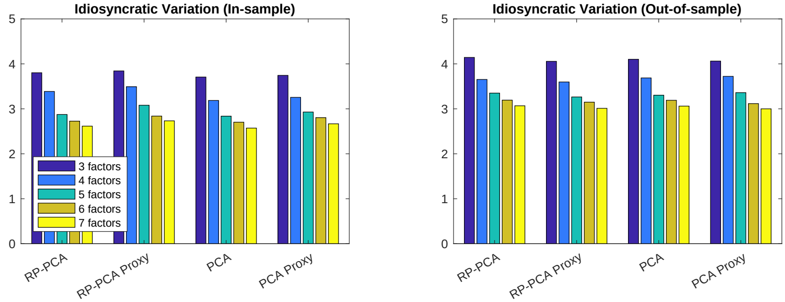


Figure: Unexplained idiosyncratic variation.

⇒ Unexplained variation similar for RP-PCA and PCA

⇒ Proximate factors explain the same variation

Interpreting factors: 6th proxy factor

6. Proxy RP-PCA	Weights	6. Proxy PCA	Weights
Momentum (6m) 1	0.28	Leverage 10	0.33
Momentum (6m) 2	0.25	Asset Turnover 10	0.25
Value (M) 10	0.25	Value-Profitability 10	0.25
Value-Momentum 1	0.23	Profitability 10	0.22
Industry Momentum 1	0.20	Asset Turnover 9	0.22
Industry Reversals 9	0.19	Sales/Price 10	0.20
Industry Momentum 2	0.19	Sales/Price 9	0.18
Momentum (6m) 3	0.18	Size 10	0.17
Idiosyncratic Volatility 2	-0.18	Value-Momentum-Profitability 1	-0.19
Industry Mom. Reversals	-0.18	Profitability 2	-0.19
Value-Momentum 8	-0.20	Value-Profitability 1	-0.20
Momentum (6m) 10	-0.21	Profitability 4	-0.20
Value-Momentum 9	-0.23	Value-Profitability 2	-0.20
Value-Momentum 10	-0.23	Profitability 1	-0.23
Short-Term Reversals 1	-0.24	Idiosyncratic Volatility 1	-0.24
Industry-Momentum 10	-0.24	Profitability 3	-0.25
Industry Rel. Reversals 1	-0.28	Asset Turnover 2	-0.28
Idiosyncratic Volatility 1	-0.38	Asset Turnover 1	-0.35

Interpreting factors: Cumulative absolute proxy weights

RP-PCA Proxy		PCA Proxy	
Idiosyncratic Volatility	3.23	Idiosyncratic Volatility	2.35
Momentum (12m)	1.64	Momentum (12m)	1.47
Industry Mom. Reversals	1.56	**Asset Turnover**	1.11
Industry Rel. Reversals (L.V.)	1.50	**Gross Profitability**	1.09
Price	1.45	Industry Rel. Rev. (L.V.)	1.07
Momentum (6m)	1.44	Size	1.04
Value-Momentum	1.25	Industry Mom. Reversals	1.01
Size	1.09	Net Operating Assets	1.00
Industry Momentum	1.00	Momentum (6m)	0.99
Net Operating Assets	0.95	Price	0.92
Industry Rel. Reversals	0.88	Value-Momentum	0.86
Value (M)	0.75	Value-Profitability	0.82
Value-Momentum-Prof.	0.51	Value-Momentum-Prof.	0.80
Share Volume	0.46	Industry Momentum	0.73
Investment/Capital	0.41	Value (M)	0.67
Earnings/Price	0.40	Sales/Price	0.56
Short-Term Reversals	0.40	Dividend/Price	0.45

High-Frequency price data

Data

- High-frequency factor analysis from Pelger (2017)
- Time period: 2003 to 2012
- $X_i(t)$ is the log-return from the TAQ database
- N between 500 and 600 firms from the S&P 500
- 5-min sampling: on average 250 days with 77 increments each
- Estimator for number of factors indicate 4 latent factors
- Create factors for continuous (normal) movements and for jumps (rare large) movements
- **Question:** What are the factors?

Identification of factors

Interpretation of continuous factors

- Approach: Rotate and threshold
- Non-zero elements are almost all in specific industries
- 4 economic candidate factors:
  - Market (equally weighted)
  - Oil and gas (40 equally weighted assets)
  - Banking and Insurance (60 equally weighted assets)
  - Electricity (24 equally weighted assets)

Main result: Interpretation of factors

4 continuous factors with industry continuous factors			
1.00	**0.98**	**0.95**	**0.80**
4 jump factors with industry jump factors			
0.99	0.75	0.29	0.05
4 continuous factors with Fama-French Carhart Factors			
0.95	0.74	0.60	0.00

Table: Generalized correlations of first four largest statistical factors for 2007-2012 with economic factors

- Element-wise generalized correlations close to 1 measure of how many factors two sets have in common
- Economic industry factors: Market, oil, finance, electricity

⇒ Jump structure different from continuous structure

⇒ Size, value, momentum do not explain factors

Interpretation of continuous factors

2007-2012	2007	2008	2009	2010	2011	2012
1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.98	0.98	0.97	0.99	0.97	0.98	0.93
0.95	0.91	0.95	0.95	0.93	0.94	0.90
0.80	0.87	0.78	0.75	0.75	0.80	0.76

Generalized correlation of market, oil, finance and energy factors with first four largest statistical factors for 2007-2012

- $\Rightarrow$ Stable continuous factor structure
- $\Rightarrow$ Proximate factors approximate latent factors well

Interpretation of continuous factors

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.97	0.99	1.00	1.00	0.99	0.97	0.98	0.96	0.98	0.95
0.57	0.75	0.77	0.89	0.85	0.92	0.95	0.92	0.93	0.83
0.10	**0.23**	**0.16**	**0.35**	0.82	0.74	0.72	0.68	0.78	0.78

Generalized correlation of market, oil, finance and energy factors with first four largest statistical factors for 2003-2012

$\Rightarrow$ Finance factor disappears in 2003-2006

Conclusion

Methodology

- Proximate factors (portfolios of a few assets) for latent population factors (portfolios of all assets)
- Simple thresholding estimator based on largest loadings
- Proximate factors approximate population factors well without sparsity assumption
- Asymptotic probabilistic lower bound for (generalized) correlation
- Future work: Sharpen bounds based on extreme value theory

$\Rightarrow$ Few observations summarize most of the information

Empirical Results

- Good approximation to population factors with 5-10% portfolios
- Interpretation of RP-PCA and high-frequency PCA factors

Extreme Deciles

Anomaly	Mean	SD	Sharpe-ratio	Anomaly	Mean	SD	Sharpe-ratio
Accruals - accrual	0.37	3.20	0.12	Momentum (12m) - mom12	1.28	6.91	0.19
Asset Turnover - aturnover	0.40	3.84	0.10	Momentum-Reversals - momrev	0.47	4.82	0.10
Cash Flows/Price - cfp	0.44	4.38	0.10	Net Operating Assets - noa	0.15	5.44	0.03
Composite Issuance - ciss	0.46	3.31	0.14	Price - price	0.03	6.82	0.00
Dividend/Price - divp	0.2	5.11	0.04	Gross Profitability - prof	0.36	3.41	0.11
Earnings/Price - ep	0.57	4.76	0.12	Return on Assets (A) - roaa	0.21	4.07	0.05
Gross Margins - gmargins	0.02	3.34	0.01	Return on Book Equity (A) - roea	0.08	4.40	0.02
Asset Growth - growth	0.33	3.46	0.10	Seasonality - season	0.81	3.94	0.21
Investment Growth - igrowth	0.37	2.69	0.14	Sales Growth - sgrowth	0.05	3.59	0.01
Industry Momentum - indmom	0.49	6.17	0.08	Share Volume - shvol	0.00	6.00	0.00
Industry Mom. Reversals - indmomrev	1.18	3.48	0.34	Size - size	0.29	4.81	0.06
Industry Rel. Reversals - indrrev	1.00	4.11	0.24	Sales/Price sp	0.53	4.26	0.13
Industry Rel. Rev. (L.V.) - indrrevlv	1.34	3.01	0.44	Short-Term Reversals - strev	0.36	5.27	0.07
Investment/Assets - inv	0.49	3.09	0.16	Value-Momentum - valmom	0.51	5.05	0.10
Investment/Capital - invcap	0.13	5.02	0.03	Value-Momentum-Prof. - valmomprof	0.84	4.85	0.17
Idiosyncratic Volatility - ivol	0.56	7.22	0.08	Value-Profitability - valprof	0.76	3.84	0.20
Leverage - lev	0.24	4.58	0.05	Value (A) - value	0.50	4.57	0.11
Long Run Reversals - lrrev	0.46	5.02	0.09	Value (M) - valuem	0.43	5.89	0.07
Momentum (6m) - mom	0.35	6.27	0.06				

Table: Long-Short Portfolios of extreme deciles of 37 single-sorted portfolios from 07/1963 to 12/2016: Mean, standard deviation and Sharpe-ratio.