Instrumented Principal Component Analysis
Kelly, Pruitt, Su (2018)

Nhi Truong, Yu Wang

September 30, 2019

Arbitrage Pricing Theory

- Let $r_{i,t}$ denote the excess return of some stock i at time t.
- Ross (1976): Arbitrage Pricing Theory: there exists SDF m_t such that

$$\mathbb{E}_t[r_{i,t+1}] = \underbrace{\frac{\mathrm{Cov}(m_{t+1}, r_{i,t+1})}{\mathrm{Var}_t(m_{t+1})}}_{\beta_{i,t}} \underbrace{\left(-\frac{\mathrm{Var}_t(m_{t+1})}{\mathbb{E}_t[m_{t+1}]}\right)}_{\lambda_t}.$$

- $\beta_{i,t}$: exposure to systematic risk
- λ_t: risk premium

Factor Modeling

- Recall from homework 3:

$$\underbrace{R}_{N\times T} = \underbrace{\Lambda}_{N\times K}\underbrace{F}_{K\times T} + \underbrace{e}_{N\times T}.$$

- Previous presentation on projected PCA: Λ is a function of observable characteristics (written in terms of basis functions).
- Projected RP-PCA: Λ is a function of observable characteristics, with basis functions being indicators of characteristics by deciles.
- **IPCA**: Λ is a **linear** function of characteristics.

IPCA Model

- N stocks, L characteristics, K factors
- Model:

$$r_{t+1} = \alpha_t + \beta_t \underbrace{f_{t+1}}_{K\times 1} + \epsilon_{t+1}$$

$$\alpha_t = \underbrace{Z_t}_{N\times L} \underbrace{\Gamma_\alpha}_{L\times 1} + \nu_{\alpha,t}, \quad \beta_{i,t} = Z_t \underbrace{\Gamma_\beta}_{L\times K} + \nu_{\beta,t}$$

- Z_t is matrix of stacked observable characteristics.
- Γ_β: linear map from characteristics to loading.
- Γ_α: linear map from chracteristics to stock's alphas.

Managed Portfolios

- Z_t is stock's characteristics ranked and normalized to be in $[-0.5, 0.5]$.
- Let $X_t = Z_t' r_{t+1}$, get L portfolios
- Suppose $\Gamma_\alpha = 0$. Model becomes:

$$X_t = Z_t' Z_t \Gamma_\beta f_{t+1} + \epsilon^*_{t+1}.$$

- If $Z_t' Z_t \approx$ constant, then minimizing squared residual is the same as doing PCA.

Identification

- Restricted case ($\Gamma_\alpha = 0$):

$$\Gamma_\beta' \Gamma_\beta = I_{K\times K}$$

$$Cov(f_t) = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_K \end{pmatrix}$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K$$

- Unrestricted Case ($\Gamma_\alpha \neq 0$):

$$\Gamma_\alpha' \Gamma_\beta = 0_{1\times K}.$$

Restricted Model: $\Gamma_\alpha = 0$

- Recall model:

$$r_{t+1} = Z_t \Gamma_\beta f_{t+1} + \epsilon^*_{t+1}.$$

- Objective: minimize sum of squared errors:

$$\min_{\Gamma_\beta, f_t} \quad \sum_{t=1}^{T-1} (r_{t+1} - Z_t \Gamma_\beta f_{t+1})'(r_{t+1} - Z_t \Gamma_\beta f_{t+1})$$

- Search strategy: alternating least square

Unrestricted Model: $\Gamma_\alpha \neq 0$

- Recall model:

$$r_{t+1} = Z_t \Gamma_\alpha + Z_t \Gamma_\beta f_{t+1} + \epsilon^*_{t+1}.$$

- Let $\tilde{\Gamma} = [\Gamma_\alpha, \Gamma_\beta]$, and $\tilde{f}_t = [1, f_t]^T$,

$$r_{t+1} = Z_t \tilde{\Gamma} \tilde{f}_{t+1} + \tilde{\epsilon}_{t+1}.$$

- Use the same search strategy as before, then back out Γ_α, Γ_β from $\tilde{\Gamma}$, and f_{t+1} from $\tilde{f}_{t+1}$.

Alternating Least Squares

Algorithm 1: Alternating Least Square

initialization: $\hat{\Gamma}_\beta$ as eigenvectors corresponding to K largest eigenvalues of $\sum_{t=1}^{T} X_t X_t'$.

while *not convergent* **do**

$$\hat{f}_{t+1} = (\hat{\Gamma}_\beta' Z_t' Z_t \hat{\Gamma}_\beta)^{-1} \hat{\Gamma}_\beta' Z_t' r_{t+1} \text{ for all } t$$

$$\mathrm{vec}(\hat{\Gamma}_\beta) = \left(\sum_{t=1}^{T-1} Z_t' Z_t \otimes \hat{f}_{t+1} \hat{f}_{t+1}' \right) \left(\sum_{t=1}^{T-1} [Z_t \otimes \hat{f}_{t+1}']' r_{t+1} \right)$$

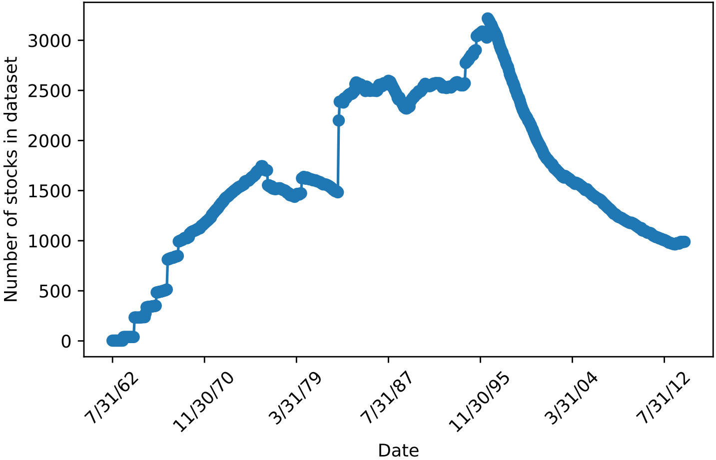
end

Rotate Γ_β and F

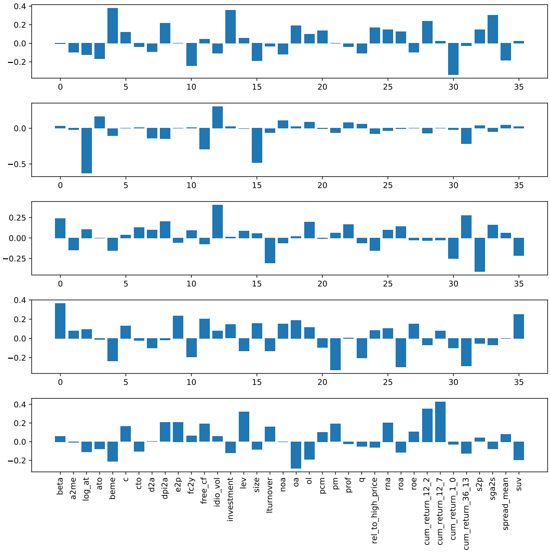
Data

- From Freyberger, Neuhierl, Weber (2017). Original: CRSP.
- Monthly observations of $N = 7593$ stocks from 01/71 to 05/14. $\mathrm{T} = 521$
- Average of 2000 stocks/month.
- L=36 characteristics.

Number of Stocks over Time



Dependence of Loadings on Characteristics



Performance Measurement

- Total R^2:

$$1 - \frac{\sum_{i,t}(r_{i,t+1} - z'_{i,t}(\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{f}_{t+1}))^2}{\sum_{i,t} r^2_{i,t+1}}$$

- Predictive R^2:

$$1 - \frac{\sum_{i,t}(r_{i,t+1} - z'_{i,t}(\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{\lambda}))^2}{\sum_{i,t} r^2_{i,t+1}},$$

where $\hat{\lambda}$ is the average of $\hat{f}_{t+1}$.

In-sample R^2

		$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
Total	$\Gamma_\alpha = 0$	0.00105	0.00181	0.00251	0.00316	0.00379
	$\Gamma_\alpha \neq 0$	0.00109	0.00186	0.00254	0.00320	0.00382
Pred.	$\Gamma_\alpha = 0$	-9e-6	-7e-6	-9e-6	-1e-5	-1e-5
	$\Gamma_\alpha \neq 0$	4e-5	4e-5	3e-5	4e-5	2e-5

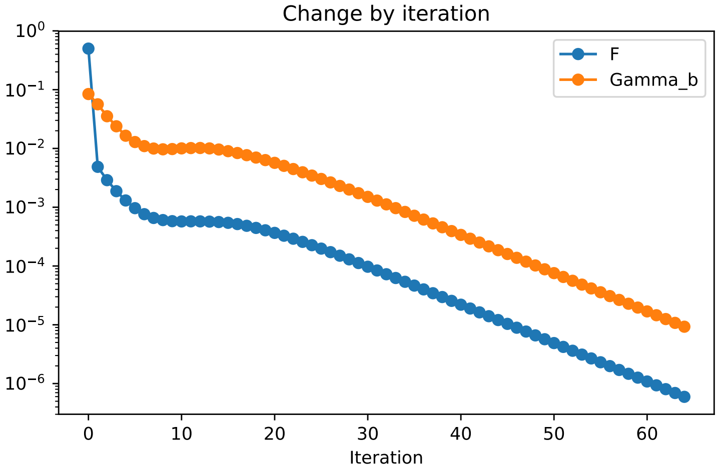
Out-of-sample R^2: $\Gamma_\alpha = 0$

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
Total R^2	0.0004	0.0007	0.0013	0.0015	0.0021
Pred. R^2	-0.0002	-0.0004	-0.0007	-0.0009	-0.0015

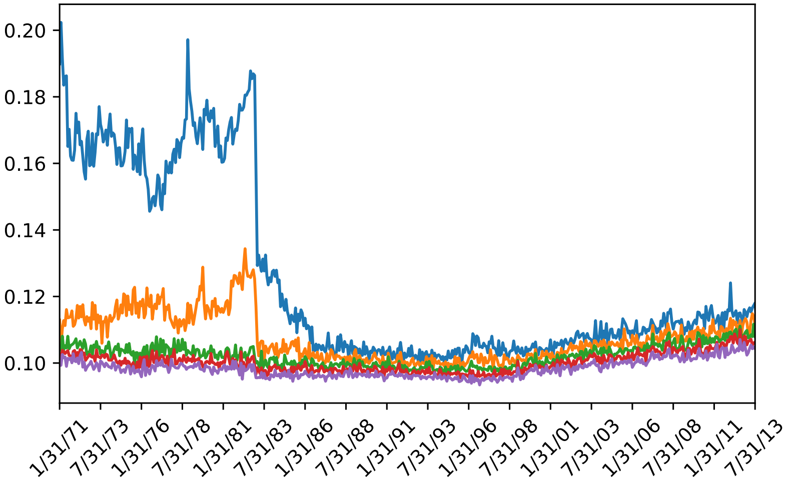
Sharpe Ratio Comparison

	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
IPCA	0.223	0.185	0.265	0.291	0.316
PCA	0.131	0.141	0.252	0.279	0.328

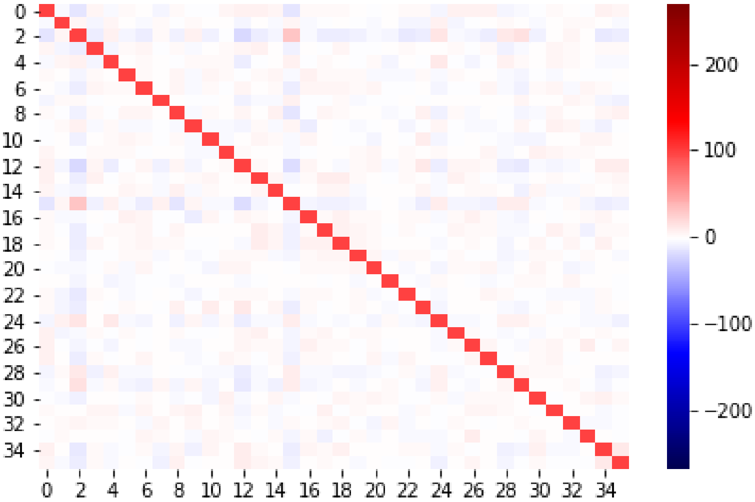
Convergence Plot



Singular Values of $Z_t'Z_t$



Heatmap of $Z_t'Z_t$



Conclusion

- There are many discrepancies in replication and paper's original report.
- Long run time impedes hypothesis testing.
- There is need to understand $Z_t' Z_t$.