A Generalized Machine Learning Framework for Linear Factor Model Test

Christopher Jones
University of Southern California
Jinchi Lv
University of Southern California
Kuntara Pukthuanthong
University of Missouri
Junbo Wang
Louisiana State University

LSU

Motivation– Data availability

- There are 400+ factors
- There are many testing assets

  Portfolios, individual stocks
- **Big data time!**
- Classical models: PCA or Lasso

LSU

What we do in this paper

We introduce SOFAR (sparse orthogonal factor regression, by Uematsu et al(2019)) into the linear factor test.

Advantage of the SOFAR:

- It can be applied to big data.
- It can take advantage of previous machine learning methods (sparsity from variable selection, common factor from PCA).

Introduction to SOFAR

- Time-series return model:

$$\boldsymbol{R} = \boldsymbol{FB} + \boldsymbol{E}$$

$\boldsymbol{R}$ (T × N) asset return (T: total periods, N number of assets)

$\boldsymbol{F}$ (T × M) factors (M number of factors)

$\boldsymbol{B}$ (M × N) factor loading

$\boldsymbol{E}$ (T × N) error term

- Classical method: estimation seemingly unrelated regressions.
- What if both N and M are large?

LSU

Introduction to SOFAR

- Singular value decomposition of B matrix: $\boldsymbol{B} = \boldsymbol{UDV'}$

$\boldsymbol{D} = diag(d_1, \cdots, d_L)$: L by L diagonal matrix, with $L = min(M, N)$

$\boldsymbol{U}$ (M times L) and $\boldsymbol{V}$ (N times L): orthonormal matrices (i.e.$\boldsymbol{U'U} = I$ and $\boldsymbol{V'V} = I$, with $\boldsymbol{V'}$ or $\boldsymbol{U'}$ the transpose of matrix $\boldsymbol{V}$ or $\boldsymbol{U}$)

- The model can be written as

$\boldsymbol{R} = \boldsymbol{FUDV'} + \boldsymbol{E}$

Define $\boldsymbol{FUD} = \overline{\boldsymbol{F}}$ as latent factors, $\boldsymbol{R} = \overline{\boldsymbol{F}}\boldsymbol{V'} + \boldsymbol{E}$

- If $\boldsymbol{D}$ has rank K, there are only K latent factors.
- If each of the $\overline{\boldsymbol{F}}$ only depends on a few of $\boldsymbol{F}$, $\boldsymbol{U}$ is sparse.
- If each of the $\overline{\boldsymbol{F}}$ is only correlated with a few of $\boldsymbol{R}$, $\boldsymbol{V}$ is sparse.

Introduction to SOFAR

- What should be the number K?
- Are each of K factors depend on all M candidate factors?
- Are each of K factors correlated with all N assets?

SOFAR can resolve these issues together.

$$\left(\hat{\boldsymbol{D}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}\right) = \underset{\boldsymbol{D},\boldsymbol{U},\boldsymbol{V}}{argmin} \left\{\frac{1}{2}(\| \boldsymbol{R} - \boldsymbol{FUDV'} \|_F + \lambda_d \| \boldsymbol{D} \|_1 + \lambda_a \rho_a(\boldsymbol{UD}) + \lambda_b \rho_b(\boldsymbol{VD}))\right\}$$

subject to $\boldsymbol{U'U} = I$ and $\boldsymbol{V'V} = I$.

$\|\cdot\|_F$ measures the distance

$\| \boldsymbol{D} \|_1 = \sum_{i,j} |D_{i,j}|$, penalty for D matrix

ρ_a and ρ_b are penalty functions for U and V matrices.

- We can choose any combination of these penalties, making the selection very flexible.

Application: Reduced Rank Approach Extension

- Huang, Li and Zhou (2020)

  The model can be written as $\boldsymbol{R} = \overline{\boldsymbol{F}}\boldsymbol{V}' + \boldsymbol{E}$ where $\boldsymbol{FUD} = \overline{\boldsymbol{F}}$.

  The goal is to find U, D and V matrices.

- Applying SOFAR:

  $$\left(\hat{\boldsymbol{D}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}\right) = \underset{\boldsymbol{D},\boldsymbol{U},\boldsymbol{V}}{argmin}\left\{\frac{1}{2}(\| \boldsymbol{R} - \boldsymbol{FUDV}' \|_F)\right\}.$$

Add penalties to select (1) latent factors, (2) candidate factors related with latent factors, and (3) assets correlated with latent factors

Data

- Portfolios: 202 portfolios Giglio and Xu (2020)
  - 25 portfolios sorted by size and book-to-market ratio
  - 17 industry portfolios
  - 25 portfolios sorted by operating probability and investment
  - 25 portfolios sorted by size and variance
  - 35 portfolios sorted by size and net issuance
  - 25 portfolios sorted by size and accruals
  - 25 portfolios sorted by size and beta
  - 25 portfolio sorted by size and momentum
- Candidate factors
  - Construct 219 candidate factors following Hou, Xue and Zhang (2019)

LSU

Application: RRA with selection of latent factors, candidate factors and correlated asset

	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalues	15.94	2.47	0.56	0.59
Correlated Assets	202	162	135	125

Factor 2		Factor 3		Factor 4	
F.6.25 Liquidity betas (net)	-0.24	E.5.46 Alm, asset liquidity	-0.42	E.5.8 Operating leverage	-0.53
B.2.3 Quarterly book-to-market equity	-0.17	F.6.25 Liquidity betas (net)	-0.12	F.6.17 Lm11, Turnover-adjusted number of zero daily volume	-0.37
F.6.17. Lm121, Turnover-adjusted number of zero daily volume	-0.09	C.3.20 Changes in book equity	-0.09	F.6.17. Lm121, Turnover-adjusted number of zero daily volume	-0.28
F.6.25 Liquidity betas illiquidity-illiquidity	-0.01	D.4.18 Cash-based operating profitability	-0.02	D.4.12 Operating profits to equity	-0.15
E.5.46 Alm, asset liquidity.	0.00	E.5.51 Average returns Rn[2,5]	0.25	F.6.1 Market equity	-0.14
E. 5.51 Average returns Rn[2,5]	0.00	E.5.30 Financial constraints (the Kaplan-Zingales index)	0.36	D.4.32 Book leverage	-0.11
E. 5.51 Average returns Ra[2,5]	0.01	E.5.4 R&D expense-to-market	0.39	E.5.4 R&D expense-to-market	-0.06
E.5.30 Financial constraints (the Kaplan-Zingales index)	0.07	E.5.11 RCA, Capital-to-assets	0.42	C.3.20 Changes in book equity	-0.02
F.6.14 Coefficient of variation of dollar trading volume.	0.07	E.5.10 Hiring rate	0.52	F.6.8 Market beta	0.03
F.6.25 Liquidity betas (return-return)	0.27			E.5.51 Average returns Ra[6,11]	0.07
C.3.20 Changes in book equity	0.58			D.4.5 Assets turnover	0.12
C.3.12 Composite equity issuance	0.70			E.5.30 Financial constraints (the Kaplan-Zingales index)	0.18
				E.5.46 Ala asset liquidity	0.18
				F.6.25 Liquidity betas (return-return)	0.21
				F.6.3 Idiosyncratic volatility	0.27
				F.6.11 Share turnover	0.35
				F.6.25 Liquidity betas (net)	0.36

Conclusions

- We introduce SOFAR

- We show that SOFAR can extend existing methods

- Applying SOFAR, we find

➢ Four factors to represent the covariance structure

LSU

Thank you !