Rank Regularized Estimation of Approximate Factor Models

Jushan Bai Serena Ng

Columbia University

April 2018

Outline

1. Approximate Factor Models
   - APC vs PC
2. Rank Minimization: NP Hard
3. Approximate-Rank Minimization
   - RPC vs PC
4. Rank Regularized Factor Models
   - Number of Factors
   - Linear Restrictions

Overview: Model: $X = F\Lambda' + e$

- APC: asymptotic principal components $\widetilde{F} = \sqrt{T}U_r$
  - eigenvectors can be constructed by iterative OLS.
- What if we do iterative ridge regressions instead of OLS?
  - Singular value thresholding $\Rightarrow$ robust PC
  - Regularize rank of common component
  - Algorithmic view: finite sample error bounds.
- This paper: rank regularized factor analysis
  - Parametric analysis, asymptotic results for inference.
  - A new, conservative factor selection rule.
  - (*) Factor analysis under general linear restrictions.

Notation

- $X_{it} \sim (0,1)$, $i = 1, \ldots N, t = 1, \ldots T$.
- SVD: $X = U\widetilde{D}V' = U_r\widetilde{D}_r V_r' + U_{n-r}\widetilde{D}_{n-r} V_{n-r}'$
- Normalized Data: $Z = \frac{X}{\sqrt{NT}} = UDV'$, $D = \frac{\widetilde{D}}{\sqrt{NT}}$
- Unscaled model: $X = F^0\Lambda^{0'} + e$.
- Scaled model: $Z = F^*\Lambda^{*\prime} + e^*$
- $F^* = \frac{F^0}{\sqrt{T}}$, $\Lambda^* = \frac{\Lambda^0}{\sqrt{N}}$.

Asymptotic Principal Components: (APC)

- $\min_{F,\Lambda} \frac{1}{NT} \|X - F\Lambda'\|_F^2$ assuming strong factor structure
  - $\Sigma_F > 0, \Sigma_\Lambda > 0$; e weakly correlated.
- $(\widetilde{F}, \widetilde{\Lambda}) = (\sqrt{T}U_r, V_r D_r)$, $\frac{\widetilde{F}'\widetilde{F}}{T} = I_r$, $\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N} = D_r^2$.
- (Bai 2003): Under normalization $\frac{F'F}{T} = I_r$ or $\frac{\Lambda'\Lambda}{N} = I_r$,

$$\begin{aligned} \sqrt{N}(\widetilde{F}_t - \widetilde{H}'_{NT} F_t^0) &\xrightarrow{d} \mathcal{N}(0, \textsc{Avar}(\widetilde{F}_t)) \\ \sqrt{T}(\widetilde{\Lambda}_i - \widetilde{G}_{NT} \Lambda_i^0) &\xrightarrow{d} \mathcal{N}(0, \text{Avar}(\widetilde{\Lambda}_t)). \end{aligned}$$

with $\widetilde{G} = \widetilde{H}_{NT}^{-1}$.

Lemma

Rotation matrix: $\widetilde{H}_{NT} = \left(\frac{\Lambda^{0\prime}\Lambda^0}{N}\right)\left(\frac{F^{0\prime}\widetilde{F}}{T}\right)D_r^{-2}$.

- Let $\widetilde{H}_{1,NT} = (\Lambda^{0\prime}\Lambda^0)(\widetilde{\Lambda}'\Lambda^0)^{-1}$;

$$\widetilde{H}_{NT} = \widetilde{H}_{1,NT} + o_p(1).$$

- Let $\widetilde{H}_{2,NT} = (F^{0\prime}F^0)^{-1}(F^{0\prime}\widetilde{F})$;

$$\widetilde{H}_{NT} = \widetilde{H}_{2,NT} + o_p(1).$$

Results of independent interest:

Principal Components (PC)

- Recall APC: $(\widetilde{F}, \widetilde{\Lambda}) = (\sqrt{T}U_r, V_r D_r)$, $\frac{\widetilde{F}'\widetilde{F}}{T} = I_r$, $\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N} = D_r^2$.
- Many definitions of PC: e.g. $(\widetilde{F}, \widetilde{\Lambda}) = (\sqrt{T}U_r D_r, V_r)$.

Principal Components (PC)

- Recall APC: $(\widetilde{F}, \widetilde{\Lambda}) = (\sqrt{T}U_r, V_r D_r)$, $\frac{\widetilde{F}'\widetilde{F}}{T} = I_r$, $\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N} = D_r^2$.
- Many definitions of PC: e.g. $(\widetilde{F}, \widetilde{\Lambda}) = (\sqrt{T}U_r D_r, V_r)$.
- This paper defines PC:
  $(\widehat{F}, \widehat{\Lambda}) = (\sqrt{T}U_r D_r^{1/2}, \sqrt{N}V_r D_r^{1/2}) = (\sqrt{T}\widehat{F}_z, \sqrt{N}\widehat{\Lambda}_z)$.
- Normalization: $\frac{\widehat{F}'\widehat{F}}{T} = D_r$, $\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N} = D_r$.
- Why? $\frac{\widetilde{F}'\widetilde{F}}{T} = I_r$. Not convenient to put restrictions.

- Relation with APC: $\widehat{F} = \widetilde{F} D_r^{1/2} \quad \widehat{\Lambda} = \widetilde{\Lambda} D_r^{-1/2}$.

- Define $\widehat{H}_{NT} = \widetilde{H}_{NT} D_r^{1/2}$. From identities:

$$\begin{aligned} \sqrt{N}(\widehat{F}_t - \widehat{H}_{NT}' F_t^0) &= \sqrt{N} D_r^{1/2} (\widetilde{F}_t - \widetilde{H}_{NT}{}' F_t^0), \\ \sqrt{T}(\widehat{\Lambda}_i - \widehat{H}_{NT}^{-1} \Lambda_i^0) &= \sqrt{T} D_r^{-1/2} (\widetilde{\Lambda}_i - \widetilde{H}_{NT}^{-1} \Lambda_i^0). \end{aligned}$$

- Asymptotic properties:

(i) $\sqrt{N}(\widehat{F}_t - \widehat{H}_{NT}' F_t^0) \xrightarrow{d} N\left(0,\ \mathbb{D}_r^{1/2}\, \textsc{Avar}(\widetilde{F}_t)\, \mathbb{D}_r^{1/2}\right)$;

(ii) $\sqrt{T}(\widehat{\Lambda}_i - \widehat{G}_{NT} \Lambda_i^0) \xrightarrow{d} N\left(0,\ \mathbb{D}_r^{-1/2}\, \textsc{Avar}(\widetilde{\Lambda}_i)\, \mathbb{D}_r^{-1/2}\right)$.

with $\widehat{G}_{NT} = \widehat{H}_{NT}^{-1}$.

Let A be a $n \times n$ matrix with eigenvalues in $D = \text{diag}(d)$:

- Trace norm: $\sum_{k=1}^{n} A_{kk} = \sum_{k=1}^{n} d_k$
- Nuclear norm: $||A||_* = \sum_{k=1}^{n} d_k$
- Frobenius norm: $||A||_F^2 = \sum_{ij} A_{ij}^2 = \text{trace}(A'A)$.
- ℓ_1 norm: $||A||_1 = \sum_{ij} |A_{ij}|$
- Spectral norm: $||A||_2 = \max_k |d_k|$.

Spark vs Rank

For $A \in \mathbb{R}^{m \times n}$, $n < m$

$$\begin{aligned} \text{spark}(A) &= \min_{x \neq 0} ||x||_0 \quad \text{s.t. } Ax = 0 \\ \text{rank}(A) &= ||D||_0 = nnz(D). \end{aligned}$$

- $\text{spark}(A) =$ size of smallest set of lin. dep. columns.
- $\text{rank}(A) =$ size of largest set of lin. indep. columns.
- $\text{spark}(A) = n + 1 \Leftrightarrow \text{rank}(A) = n$.
- If $\text{spark}(A) \neq n + 1$:
  - $\text{spark}(A) \leq \text{rank}(A)$.
  - $\text{spark}(A) \geq 1 + \frac{1}{\mu(A)}$, $\mu(A) = \max_{m \neq n}(a_m, a_n)|$.
- Computing $\text{spark}(A)$ is NP-hard: Tillmann/Pfetsch IEEE-14

NP Hard

- NP problems: decision problems in which the answer "yes" can be efficiently verified using deterministic computations performed in polynomial time.
- An NP hard problem is one that admits no general computational solution that is significantly faster than a brute force search.

1. Minimum Rank Factor Analysis

- Early factor analysis: decompose $\Sigma_X = \Sigma_C + \Sigma_e$ s.t.
  - i commonality matrix Σ_C has smallest rank
  - ii Σ_C: a non-negative definite,
  - iii Σ_e: diagonal positive definite matrix (Haywood cases).
- Rank minimization is NP hard (non-convexity),
- Evidence in 1950s suggest many non-zero eigenvalues. Questioned usefulness of the concept of minimum rank.

1980s: Decompose Σ_X by solving surrogate problems s.t.

$$(i)\ \Sigma_X - \Sigma_e \geq 0 \qquad (ii)\ \Sigma_e \geq 0.$$

(i) CMTFA: min trace$(\Sigma_X - \Sigma_e) = \sum_{i=1}^{N} D_{ii}^C$

(ii) MARFA: $C = C^* + C^-$.

- C^* is best minimum rank approximation of C.
- rank$(C^*) = r$, min $\sum_{i=r+1}^{N} D_{ii}^C$

Approximate Minimum Rank: ten Berge-Kiers (1991)

$$\min_r \sum_{i=r+1}^{N} D_{ii}^C \leq \delta, \quad \text{s.t. (i)+(ii).} \qquad (*).$$

- δ: tolerance for max unexplained common variance.
- The approximate minimum rank of Σ_C is the smallest r that solves (*) for some $\delta \geq 0$.
- Minimum rank: special case of $\delta = 0$.
- Sum of eigenvalues is convex.

2. Matrix Completion

- Complete the matrix Z with missing values.
- $\Omega =$ index set of positions of observed data.
- Underdetermined without some structure.
- Assume the latent matrix L is low rank.
- Netflix challenge: $L = AB'$, A=movie genres, B=taste
- Hard problem:

$$\min \text{ rank}(L) \quad \text{with } L_{ij} = Z_{ij} \quad (i,j) \in \Omega.$$

- Surrogate problem:

$$\min \|L\|_* \quad \text{with } L_{ij} = Z_{ij}, \quad (i,j) \in \Omega.$$

Z can be recovered if (i) there are not too many missing values, and (ii) they are missing at random.

3. Low Rank Decomposition

- Eckart-Young: Best rank r approximation of Z : $U_r D_r V_r'$.
- SVD is sensitive to noise corruption.

$$Z = \underbrace{L}_{\text{low rank}} + \underbrace{S}_{\text{sparse, big noise}}$$

  - Compressed sensing: solve underdetermined systems, recover sparse signals
  - Computer vision: S=background noise.
- Hard problem:

$$\min_{L,S} \ \text{rank}(L) + \overline{\gamma} \underbrace{||S||_0}_{\text{sparsity constraint}}.$$

- Objective function and constraint both non-convex.

Candes et al (2009)

Surrogate problem is convex:

$$\min_{L,S} ||L||_* + \overline{\gamma}||S||_1,$$

- L, S can be recovered with high probability under
  - *incoherence conditions*: L not sparse, S not low rank,
- General problem

$$Z = \underbrace{L}_{\text{low rank}} + \underbrace{S}_{\text{sparse, big noise}} + \underbrace{W}_{\text{small noise}}$$

$$\min_{L,S} ||L||_* + \overline{\gamma}||S||_1, \quad ||W||_F \leq \delta.$$

Overview: Good to Relax

- Hard problems: rank function.
- Surrogate problems: nuclear norm

Cai et al (2008, Theorem 1):

$$U_r D_r^{\gamma} V_r' = \text{argmin}_L \gamma \|L\|_* + \frac{1}{2} \|Z - L\|_F^2 .$$

- SVT=Singular-value thresholding operator:

$$D_r^{\gamma} = \begin{pmatrix} (D_{11} - \gamma)_+ & & \\ & \cdots & \\ & & (D_{rr} - \gamma)_+ \end{pmatrix}$$

- SVT is the proximal operator of the nuclear norm:
- Optimal low rank approx. under rank constraint: $U_r D_r^{\gamma} V_r$.

Relation to Factor Models

We have low rank solution

$$U_r D_r^{\gamma} V_r' = \min_L \gamma \|L\|_* + \frac{1}{2}\|Z - L\|_F^2 . \qquad (1).$$

L (of rank r) can be factorized: $L = AB'$,

$$\min_{A,B} \gamma \|AB'\|_* + \frac{1}{2}\|Z - AB'\|_F^2 \qquad (2)$$

- Theorem: $(\overline{A}, \overline{B})$ solves (2) iff $\overline{L} = \overline{A}\,\overline{B}'$ solves (1).
- Solution: Robust Principal Components (RPCA)

$$\overline{A} = U_r(D_r^{\gamma})^{1/2} \quad \overline{B} = V_r(D_r^{\gamma})^{1/2}.$$

- Sketch of idea, $\gamma = 0$: $AB' = U_r D_r V_r^T$:

$$\text{trace}(D_r) \quad = \quad \text{trace}(U_r' AB' V_r) \leq \|A\|_F \|B\|_F \leq \frac{1}{2}\left(\|A\|_F^2 + \|B\|_F^2 \right).$$

- $L = AB'$, $\|L\|_* = \text{trace}(D_r) \leq \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$.

- Put, $A = UD_r^{1/2}, B = D_r^{1/2} V$,

$$\frac{1}{2}(\|A\|_F^2 + \|B\|_F^2) = \frac{1}{2}(||D_r^{1/2}||_F^2 + ||D_r^{1/2}||_F^2) = \|D_r\|_1 .$$

- Bound holds with equality: $\|D_r\|_1 = \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$.

FOC View

- FOC: (i) $-(Z - \overline{A}\overline{B}')\overline{B} + \gamma\overline{A} = 0$, (ii) $-(Z - \overline{A}\overline{B}')'\overline{A} + \gamma\overline{B} = 0$.
- Left multiplying (i) by $\overline{A}'$ and (ii) by $\overline{B}'$: $\overline{A}'\overline{A} = \overline{B}'\overline{B}$. Rearrange

$$\begin{pmatrix} -\gamma I & Z \\ Z' & -\gamma I \end{pmatrix} \begin{pmatrix} \overline{A} \\ \overline{B} \end{pmatrix} = \begin{pmatrix} \overline{A} \\ \overline{B} \end{pmatrix} \overline{A}'\overline{A}.$$

  This has the generic structure $\mathbb{Z}\mathbb{V} = \mathbb{V}\mathbb{X}$.

- Eigenvalues of $\mathbb{X}$ are those of $\mathbb{Z}$, $\mathbb{V}$ are corresponding eigenvectors.

$$\overline{A} = U_r(D_r^\gamma)^{1/2}, \quad \overline{B} = V_r(D_r^\gamma)^{1/2}.$$

- A particular normalization.

Factor Analysis and RPC

With $\overline{A}'\overline{A} = \overline{B}'\overline{B} = \overline{D}_r^{\gamma}$.

- RPC of Z: $(\overline{A}, \overline{B}) = (U_r(D_r^{\gamma})^{1/2}, V_r(D_r^{\gamma})^{1/2})$
- RPC of X: $(\overline{F}, \overline{\Lambda}) = (\sqrt{T}U_r(D_r^{\gamma})^{1/2}, \sqrt{N}V_r(D_r^{\gamma})^{1/2})$.
- PC of X: $(\widehat{F}, \widehat{\Lambda}) = (\sqrt{T}U_r(D_r)^{1/2}, \sqrt{N}V_r(D_r)^{1/2})$.
- Relation between RPC and PC:

$$\overline{F} = \widehat{F}\left(D_r^{\gamma}D_r^{-1}\right)^{1/2} \quad \overline{\Lambda} = \widehat{\Lambda}\left(D_r^{\gamma}D_r^{-1}\right)^{1/2}.$$

- Even big factors will be shrunk
- Small factors can be killed since $\text{rank}(D_r^{\gamma}) \leq r$
- Sparse large noise not treated as factors
- Smaller common component: $\text{var}(\overline{C}) \leq \text{var}(\widehat{C})$.

Effects of Regularization: $\Delta_{NT}^2 = D_r^{\gamma} D_r^{-1}$.

- $\overline{H}_{NT} = \widehat{H}_{NT}\Delta_{NT}$.

$$\begin{aligned} \overline{F}_t - \overline{H}'_{NT}F_t^0 &= \Delta_{NT}(\widehat{F}_t - \widehat{H}'_{NT}F_t^0) \\ \overline{\Lambda}_i - \overline{G}_{NT}\Lambda_i^0 &= \Delta_{NT}(\widehat{\Lambda}_i - \widehat{H}_{NT}^{-1}\Lambda_i^0) \end{aligned}$$

Proposition

(i) $\sqrt{N}(\overline{F}_t - \overline{H}'_{NT}F_t^0) \xrightarrow{d} N\left(0, \Delta_{\infty}\text{AVAR}(\widehat{F})\Delta_{\infty}\right)$;

(ii) $\sqrt{T}(\overline{\Lambda}_i - \overline{G}_{NT}\Lambda_i^0) \xrightarrow{d} N\left(0, \Delta_{\infty}\text{AVAR}(\widehat{\Lambda})\Delta_{\infty}\right)$.

Unlike APC and PC, $\overline{G}_{NT} = \Delta_{NT}\widehat{H}_{NT}^{-1} \neq \widehat{H}_{NT}^{-1}$.

Bias/Variance Tradeoff

$\text{diag}(\Delta_\infty) = \delta$, $\delta_i < 1$. Proposition implies

- $\text{AVAR}(\overline{F}) \leq \text{AVAR}(\widehat{F})$, and $\text{AVAR}(\overline{\Lambda}) \leq \text{AVAR}(\widehat{\Lambda})$.
- Regularization bias since $\widehat{C} = U_r D_r V_r' \neq \overline{C} = U_r D_r^\gamma V_r'$.
- Case $r = 1$: $\delta_1 = \frac{(D_{11} - \gamma)_+}{D_{11}}$, $\overline{C}_{it} = \delta_1 \widehat{C}_{it}$.

$$\begin{aligned} \text{ABIAS}(\overline{C}_{it}) &= (\delta_1 - 1) C_{it}^0 \\ \text{AVAR}(\overline{C}_{it}) &= \delta_1^2 \text{AMSE}(\widehat{C}_{it}) \\ \text{AMSE}(\overline{C}_{it}) &= (\delta_1 - 1)^2 (C_{it}^0)^2 + \delta_1^2 \text{AMSE}(\widehat{C}_{it}). \end{aligned}$$

Relative MSE < 1 when $\text{AMSE}(\widehat{C}_{it})$ large:

$$\frac{\text{AMSE}(\overline{C}_{it})}{\text{AMSE}(\widehat{C}_{it})} = (\delta_1 - 1)^2 \frac{(C_{it}^0)^2}{\text{AMSE}(\widehat{C}_{it})} + \delta_1^2.$$

Asymptotic vs. Finite Sample Results

$Z = L + S$ consistent with many probabilistic structure

Econometric theory: $X = F^0\Lambda^{0'} + e$, $Z = \frac{X}{\sqrt{NT}}$

- Strong factor structure $\Sigma_F > 0, \Sigma_\Lambda > 0$
- r population eigenvalues diverge with N
- Estimation: choose F, Λ with e residually determined.
- $\min(\sqrt{N}, \sqrt{T})(\widehat{C}_{it} - C^0_{it}) \xrightarrow{d} N(0, Avar(\widehat{C}_{it}))$.

Machine Learning Results:

- Solve problem given data (finite sample).
- Choose L and S simultaneously.
- Netflix/noiseless problems: no reference to eigenvalues.
- Incoherence condition: L is not sparse. Details
- S is selected uniformly at random and not low rank.
- For $\gamma = \frac{1}{\sqrt{\max(m,n)}}$, $(\widehat{L}, \widehat{S}) = (L, S)$ with prob. $1 - \frac{c_0}{n^{10}}$ if $||S||_0 < c_1 mn$, and rank $(L) \leq c_1 \frac{\min(m,n)}{\mu} \log(\max(m,n))^{-2}$.

Agarwal. Negahban and Wainwright (2012, Annals of Statistics)

M estimation based on regularized nuclear norm. Assume restricted strong convexity of loss function.

- With noisy data, cannot exactly recovery L.
- What matters are eigenvectors of largest singular values.
- $\text{err}^2 = ||\widehat{L} - L||_F^2 + ||\widehat{S} - S||_F^2$
- if $||L||_\infty < \frac{c}{\sqrt{m\,n}}$, with high probability, $\text{err}^2 \leq c\left(\frac{N+T}{NT}\right)$.

- $||L||_\infty = \max_{it} |L_{it}|$, $||L||_F^2 = \sum_{i=1}^r d_i^2$.
- $||L||_\infty < \frac{c}{mn}$ is a constraint on sum of eigenvalues.
- $\frac{N+T}{NT} \approx \min(N, T)^{-1}$.
- Econometric theory: $\min(\sqrt{N}, \sqrt{T})(\widehat{C}_{it} - C_{it}) = O_p(1)$.
- Different objective, results broadly agree.
- Also related: Bertisimas, Copenhaver, Mazumder (2016), Lettau and Pelger (2017).

Number of Factors: min rank+model complexity

$$\text{BaiNg-02} \quad \widehat{r} = \min_k \log(\widehat{\text{SSR}}_k) + kg(N,T), \quad \widehat{\text{SSR}}_k = \left\| Z - \widehat{F}_k \widehat{\Lambda}_k' \right\|_F^2$$

$$\text{BaiNg-17} \quad \overline{r} = \min_k \log(\overline{\text{SSR}}_k) + kg(N,T), \quad \overline{\text{SSR}}_k = \left\| Z - \overline{F}_k \overline{\Lambda}_k' \right\|_F^2$$

$$\widehat{\text{SSR}}_k = 1 - \sum_{j=1}^{k} d_j^2, \quad \overline{\text{SSR}}_k = 1 - \sum_{j=1}^{k} (d_j - \gamma)_+^2$$

$$\overline{IC_k} \approx \widehat{IC}_k + \gamma \frac{\sum_{j=1}^{k} (2d_j - \gamma)}{\widehat{\text{SSR}}_k}.$$

- A data dependent, heavier penalty.
- $r \geq r^*$: sparse outliers or weak factors.
- $||Z||_F = 1$. $\gamma = .05$ reduces contribution of factor i by $(d_i - .05)^2$. Effect on small factors proportionally larger.

Implications for Factor Augmented Regressions

$$y_{t+h} = \alpha' F_t + \beta' W_t + \epsilon_{t+h}.$$

- Replace F by $\widetilde{F}$, $\widehat{F}$, or $\overline{F}$ will give identical fit! They are all spanned by U_r, hence perfectly correlated.
- The estimates of α will simply adjust for scale difference.
- For $\overline{F}$ to have effect, do ridge regressions. Given κ,

$$\begin{aligned}
\overline{\alpha}_{ols} &= (\overline{F}'\overline{F})^{-1}\overline{F}y = (D_r^{\gamma})^{-1/2}U'y/\sqrt{T} \\
\overline{\alpha}_{R} &= (\overline{F}'\overline{F} + \kappa I_r)^{-1}\overline{F}y \\
&= (D_r^{\gamma} + \kappa_T I_r)^{-1} D_r^{\gamma}\overline{\alpha}_{OLS} = (I_r + \kappa_T (D_r^{\gamma}))^{-1}\,\overline{\alpha}_{OLS} \\
&\approx (I_r - \kappa_T D_r^{\gamma})\overline{\alpha}_{OLS}.
\end{aligned}$$

RPC by SVT via Iterative Ridge

Given a $m \times n$ matrix Z, initialize a $m \times r$ matrix $F = \mathbb{U}\mathbb{D}$ where $\mathbb{U}$ is orthonormal and $\mathbb{D} = I_r$.

A. Repeat till convergence

i. (solve Λ given F): $\widetilde{\Lambda} = Z'F(F'F + \gamma I_r)^{-1}$.

ii $\text{SVD}(\widetilde{\Lambda}) = \widetilde{\mathbb{U}}_\Lambda \widetilde{\mathbb{D}}_\Lambda \widetilde{\mathbb{V}}_\Lambda'$, $\Lambda = \widetilde{\mathbb{U}}_\Lambda \widetilde{\mathbb{D}}_\Lambda$. $\mathbb{D} = \widetilde{\mathbb{D}}_\Lambda$.

iii (solve F given Λ): $\widetilde{F} = Z\Lambda(\Lambda'\Lambda + \gamma I_r)^{-1}$.

iv $\text{SVD}(\widetilde{F}) = \widetilde{\mathbb{U}}_F \widetilde{\mathbb{D}}_F \widetilde{\mathbb{V}}_F'$, let $F = \widetilde{\mathbb{U}}_F \widetilde{\mathbb{D}}_F$ and $\mathbb{D} = \widetilde{\mathbb{D}}_F$.

B. (Cleanup) From $\text{SVD}(Z\mathbb{U}_\Lambda) = U_r D_r \mathbb{V}_r'$, let $V_r' = \mathbb{V}'\widetilde{\mathbb{U}}_r$, $D_r^\gamma = (D_r - \gamma I_r)_+$.

- Useful when T, N are large and direct SVD is expensive.
- Iterative ridge regressions implement SVT.
- Cleanup to take care of nuermical precision problem.

Generalized Ridge

General regularized problem :

$$(\overline{F}_{\gamma_1,\gamma_2,\tau}, \overline{\Lambda}_{\gamma_1,\gamma_2,\tau}) = \text{argmin}_{F,\Lambda} \frac{1}{2}||Z - F\Lambda'||_F^2 + \frac{\gamma_1}{2}||F||_F^2 + \frac{\gamma_2}{2}||\Lambda||_F^2.$$

- Let $\overline{D}_r^{\gamma} = (D_r - \sqrt{\gamma_1\gamma_2}\, I_r)_+$. Solution is

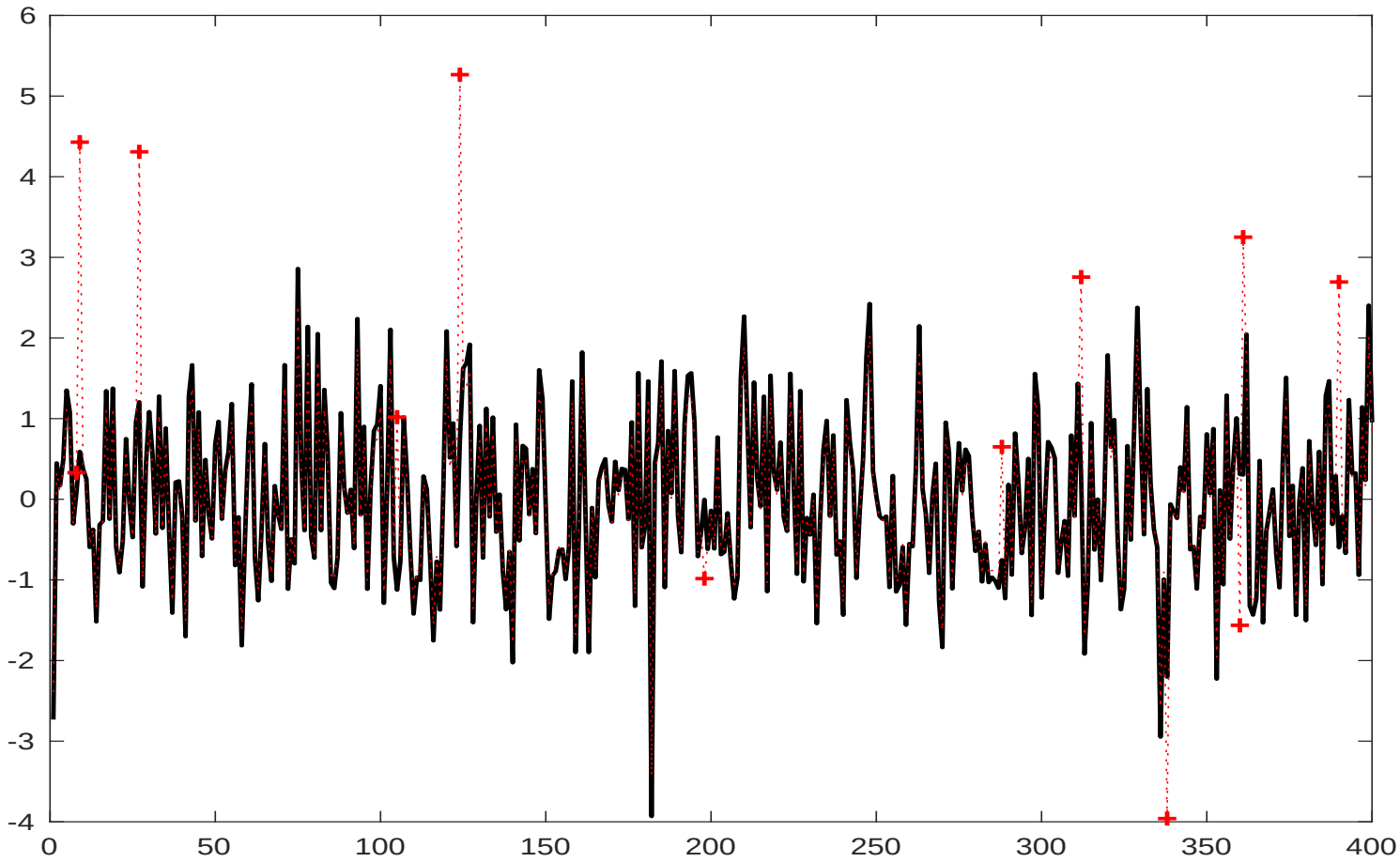
$$\begin{aligned}
\overline{F}_{\gamma_1,\gamma_2} &= \left(\frac{\gamma_2}{\gamma_1}\right)^{1/4} U_r(\overline{D}_r^{\gamma})^{1/2} \\
\overline{\Lambda}_{\gamma_1,\gamma_2} &= \left(\frac{\gamma_1}{\gamma_2}\right)^{1/4} (\overline{D}_r^{\gamma})^{1/2} \\
\overline{C}_{\gamma_1,\gamma_2} &= U_r \overline{D}_r^{\gamma} V_r'.
\end{aligned}$$

Monte Carlo

$$X_{it} = F_t^{0\prime}\Lambda_i^0 + e_{it} + s_{it}, \quad e_{it} \sim (0,1)$$

- sparse error $s_{it} \sim N(\mu, \omega^2)$ if $(i,t) \in \Omega$.
- $[\kappa_N N]$ units have outliers in $[\kappa_T T]$ of sample.
  - $(\kappa_N, \kappa_T) = (0.1, 0.03)$, $\omega \in (5, 10, 20)$
  - $\mu = 5$, $r = 5$.
- DGP1 (outliers) : $F_t^0 \sim N(0, I_r)$, $\Lambda_i^0 \sim N(0, I_r)$.
- DGP2 (weak loadings): $F^0 = U_r D_r^{1/2}$, $\Lambda^0 = V_r D_r^{1/2}$
  - $\text{diag}(D_r) = [1, 0.8, 0.5, 0.3, 0.2\theta]$, and $\omega = 5$.
    - θ $(1, 0.75, 0.5)$.

Case 1: Outlier, $\omega = 5$



Case 2: Small Eigenvalue, $\theta = 0.75$

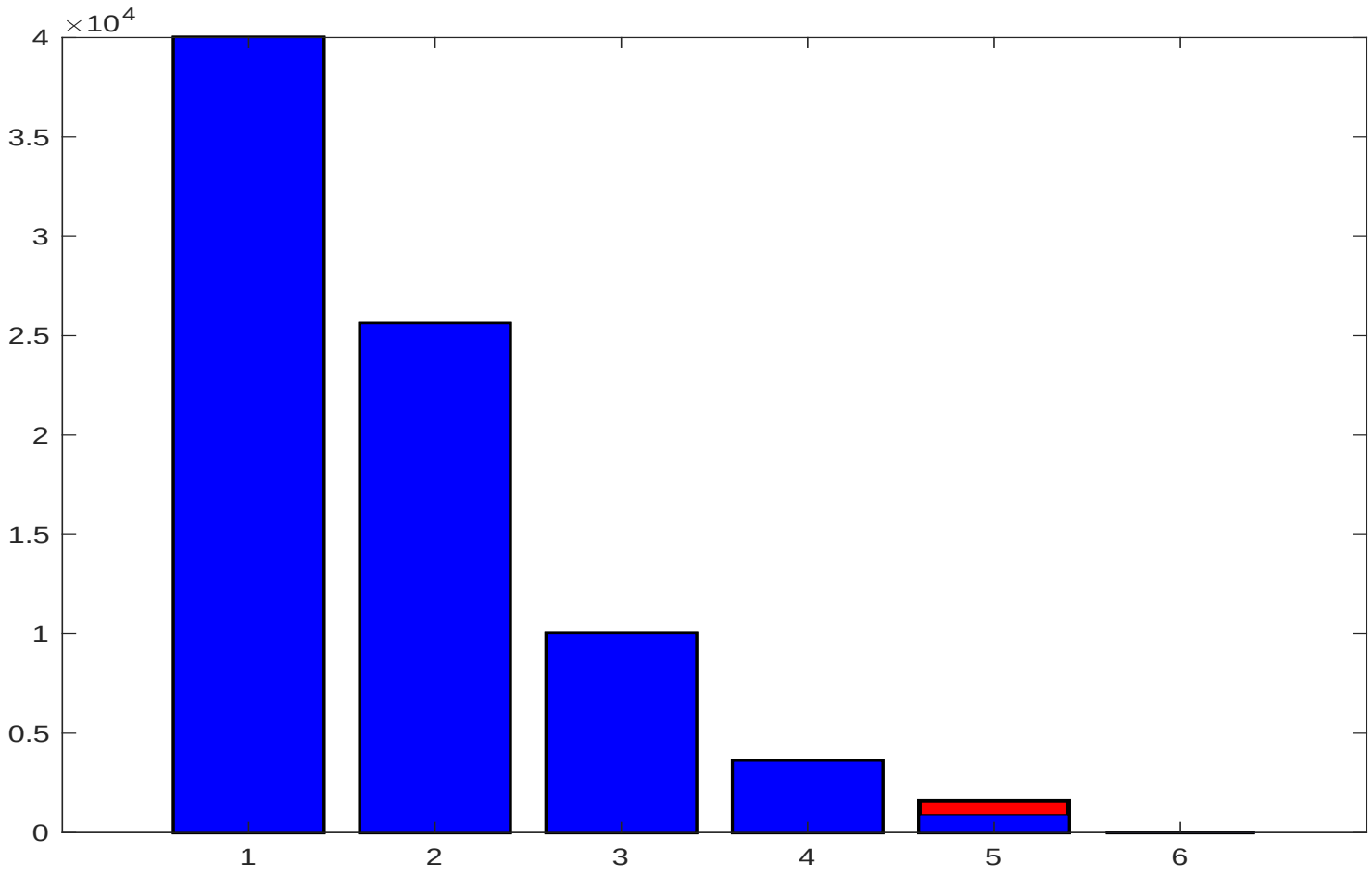


Table 1: DGP 1, $N = 100, r = 5, r^* = 5$

params	signal		noise	mean		span F^0			
T,ω	C^r	C_r	S	$\widehat{r}$	$\overline{r}$	$\widehat{C}_r$	$\overline{C}_r$	$\widehat{C}$	$\overline{C}$
100, 5	0.83	0.12	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100, 10	0.83	0.12	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100, 20	0.83	0.12	0.00	5.00	5.00	0.98	0.98	0.98	0.98
200, 5	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
200, 10	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
200, 20	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
400, 5	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
400, 10	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
400, 20	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100 , 5	0.81	0.12	0.02	5.36	5.00	0.63	0.98	0.92	0.98
100 ,10	0.78	0.12	0.06	5.79	5.00	0.28	0.98	0.85	0.97
100 ,20	0.69	0.12	0.17	6.81	5.00	0.01	0.97	0.72	0.97
200 , 5	0.81	0.13	0.02	5.67	5.00	0.32	0.98	0.87	0.98
200 ,10	0.78	0.13	0.06	5.91	5.00	0.19	0.98	0.84	0.98
200 ,20	0.69	0.13	0.17	7.13	5.00	0.00	0.98	0.69	0.98
400 , 5	0.81	0.13	0.02	5.88	5.00	0.12	0.98	0.84	0.98
400 ,10	0.78	0.13	0.06	5.90	5.00	0.16	0.98	0.84	0.98
400 ,20	0.69	0.13	0.18	7.15	5.00	0.00	0.98	0.69	0.98

Table 2: DGP 2, $N = 100, r = 5, r^* = 3, \omega = 5$

params	signal		noise	mean		span F^0			
T, ω	C^r	C_r	S	$\widehat{r}$	$\overline{r}$	$\widehat{C}_r$	$\overline{C}_r$	$\widehat{C}$	$\overline{C}$
100, 1.00	0.67	0.02	0.00	3.94	3.00	0.07	0.95	0.74	0.96
100, 0.75	0.67	0.01	0.00	3.95	3.00	0.05	0.95	0.73	0.96
100, 0.50	0.67	0.01	0.00	3.97	3.00	0.04	0.95	0.73	0.96
200, 1.00	0.67	0.02	0.00	4.01	3.00	0.00	0.95	0.73	0.97
200, 0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
200, 0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
400, 1.00	0.67	0.02	0.00	4.26	3.00	0.00	0.95	0.69	0.97
400, 0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
400, 0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
100 ,1.00	0.60	0.02	0.11	4.81	2.93	0.01	0.93	0.61	0.96
100 ,0.75	0.59	0.01	0.11	4.84	2.95	0.01	0.93	0.60	0.96
100 ,0.50	0.59	0.01	0.11	4.86	2.96	0.01	0.93	0.60	0.96
200 ,1.00	0.60	0.02	0.11	5.01	3.00	0.01	0.93	0.58	0.96
200 ,0.75	0.59	0.01	0.11	5.00	3.01	0.01	0.93	0.58	0.96
200 ,0.50	0.59	0.01	0.11	5.00	3.01	0.01	0.93	0.58	0.96
400 ,1.00	0.60	0.02	0.11	5.21	3.10	0.00	0.84	0.56	0.94
400 ,0.75	0.59	0.01	0.11	5.00	3.12	0.00	0.83	0.58	0.94
400 ,0.50	0.59	0.01	0.11	5.00	3.13	0.00	0.82	0.58	0.93

FRED-MD Data

Eigenvalues

F	Balanced Panel		Non-Balanced Panel	
	$\widehat{d}_1^2$	$\overline{d}_1^2$	$\widehat{d}_1^2$	$\overline{d}_1^2$
1	0.1828	0.1426	0.1493	0.1131
2	0.0921	0.0643	0.0709	0.0468
3	0.0716	0.0473	0.0682	0.0446
4	0.0604	0.0384	0.0561	0.0349
5	0.0453	0.0265	0.0426	0.0245
6	0.0416	0.0237	0.0341	0.0182
7	0.0301	0.0152	0.0317	0.0164
8	0.0287	0.0143	0.0268	0.0129
$(\widehat{r}, \overline{r})$	8	3	8	3

Financial Data

Eigenvalues

F	Balanced Panel		Non-Balanced Panel	
	$\widehat{d}_1^2$	$\overline{d}_1^2$	$\widehat{d}_1^2$	$\overline{d}_1^2$
1	0.6896	0.6090	0.6800	0.6001
2	0.0464	0.0274	0.0447	0.0261
3	0.0341	0.0181	0.0337	0.0178
4	0.0138	0.0045	0.0141	0.0047
5	0.0114	0.0032	0.0133	0.0043
6	0.0092	0.0021	0.0109	0.0030
7	0.0072	0.0012	0.0090	0.0020
8	0.0066	0.0010	0.0075	0.0013
$(\widehat{r}, \overline{r})$	8	3	8	3

Linear Restrictions: : $R\text{vec}(\Lambda) = \phi$

$$(\overline{F}_{\gamma,\tau}, \overline{\Lambda}_{\gamma,\tau}) = \text{argmin}_{F,\Lambda} \frac{1}{2} \|Z - F\Lambda'\|_F^2 + \frac{\gamma}{2}\left(\|F\|_F^2 + \|\Lambda\|_F^2 \right) + \frac{\tau}{2} \|R\,\text{vec}(\Lambda) - \phi\|_2^2 .$$

- Vector form: $\|Z - F\Lambda'\|_F^2 = \|\text{vec}(Z') - (F \otimes I_N)\text{vec}(\Lambda)\|_2^2$.
- Allow cross-equation restrictions.
- F given Λ: $\overline{F}_{\gamma,\tau} = Z\Lambda(\Lambda'\Lambda + \gamma I_r)^{-1}$ (standard ridge)
- Λ given F: (generalized ridge)

$$\text{vec}(\overline{\Lambda}_{\gamma,\tau}) = \left((F'F \otimes I_N) + \gamma I_{Nr} + \tau R'R\right)^{-1}\left[\text{vec}(Z'F) + \tau R'\phi\right]$$

Implementation when constraints bind

Let $W_F = (F'F + \gamma I_r)^{-1}$.

$$\text{vec}(\overline{\Lambda}_{\gamma,\infty}) = \text{vec}(\overline{\Lambda}_{\gamma,0}) - (W_F \otimes I_N)R' \cdot \left[R(W_F \otimes I_N)R'\right]^{-1}\left(R\,\text{vec}(\overline{\Lambda}_{\gamma,0}) - \phi\right.$$

Two Step Approach

1. Estimate without linear restrictions, $\tau = 0$:

$$\overline{\Lambda}_{\gamma,0} = Z'F^k(F^{k^T}F^k + \gamma I_r)^{-1}.$$

2. Impose binding linear restrictions :

$$\text{vec}(\overline{\Lambda}_{\gamma,\infty}) = \text{vec}(\overline{\Lambda}_{\gamma,0}) - W_F^k \otimes I_N R' \cdot \left[R(W_F^k \otimes I_N)R'\right]^{-1}\left(R\,\text{vec}(\overline{\Lambda}_{\gamma,0}) - \phi\right)$$

Note: $\overline{F}'_{\gamma,\infty}F_{\gamma,\infty}$ and $\overline{\Lambda}'_{\gamma,\infty}\Lambda_{\gamma,\infty}$ will not , in general, be diagonal.

Conclusion

- Iterative least squares: PC
- Iterative ridge: implements SVT
- SVT solves surragate of minimum rank problem.
- min rank +parsimony: $\Rightarrow \overline{IC}$, a data dependent penalty.
- FRED-MD, Finance data: $\widehat{r} = 8$, $\overline{r} = 3$.
- Factor estimation under linear restrictions
- Missing values problem: in progress

Incoherence Conditions

$U \in \mathbb{R}^{T \times r}$, $V \in R^{N \times r}$

- Single incoherence: singular vectors not too skewed: $\max_{i=1,\ldots T} ||U' e_i||^2 \leq \frac{\mu_0 r}{T}$, and $\max_{j=1,\ldots N} ||V' e_j||^2 \leq \frac{\mu_0 r}{N}$
- Joint incoherence: singular vectors not too correlated: $\max_{i,j} ||(UV^T)_{ij}|| \leq \sqrt{\frac{\mu_1 r}{NT}}$
- Singular vectors are reasonably spread out for small μ.
- $\sum_i (UV')_{ij} = ||V' e_j||_2^2$ and $\sum_j (UV')_{ij}^2 = ||U' e_i||_2^2$.
- μ_1 dominates.

Example when incoherence condition fails:

$$Z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} [1] \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$

- Z too sparse, and singular vectors too sparse.
- Completion requires many entries of Z to be observed.

Back to Main Text

- Rank easy to compute, spark needs combinatorial search.
- Spark(A)$\leq m+1$ =rank(A)+1.
- Donoho and Elad (2003): spark$(A) \geq 1+\mu^{-1}(A)$.
- Stable L_1 recovery: $\min ||x||_1$ s.t. $||Ax-b||_2 \leq \epsilon$.
- Coherence-base guarantee: if A has normalized columns and $Ax=b$ has solution satisfying $||x||_0 < \frac{(1+\mu^{-1}(A))}{2}$, then x is the unique sparse solution.

- Restricted Isometry Property: a $m \times n$ matrix A satisfies the RIP of order k if

$$(1-\delta_k)||z||_2^2 \leq ||Az||_2^2 \leq (1+\delta_k)||z||_2^2, \quad ||z||_0 \leq k.$$

  RIP ensures that the matrix is property scaled.

- Statistical RIP property: $P(||Ax||^2 - ||x||^2) \geq 1-\epsilon$ with respect to a uniform distribution of vector x among all k sparse in $\mathbb{R}^n$.