Multivariate regression

. <u>Simple linear regression</u> : $\underset{\in\mathbb{R}}{y^{(i)}} = \beta^T x^{(i)} + \varepsilon^{(i)}$, $i=1,\dots,m$

($\beta, x^{(i)} \in \mathbb{R}^p$)

. <u>Multivariate regression</u>: $\underset{\in\mathbb{R}^T}{y^{(i)}} = \underset{\mathbb{R}^{p\times T}}{A^T} \underset{\in\mathbb{R}^p}{x^{(i)}} + \varepsilon^{(i)}$, $i=1,\dots,m$

<u>matrix formulation</u>:

$$\underbrace{\begin{bmatrix} (y^{(1)})^T \\ \vdots \\ (y^{(m)})^T \end{bmatrix}}_{=:Y\in\mathbb{R}^{m\times T}} = \underbrace{\begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}}_{=X\in\mathbb{R}^{m\times p}} A + \underbrace{\begin{bmatrix} (\varepsilon^{(1)})^T \\ \vdots \\ (\varepsilon^{(m)})^T \end{bmatrix}}_{=:E\in\mathbb{R}^{m\times T}}.$$

① Maximum Likelihood Estimation

. <u>Statistical model</u>: $Y = XA^* + E$ with $E_{ij} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$.

. $\text{Likelihood}(A) = \prod_{i=1}^{m} \frac{1}{(2\pi\sigma^2)^{T/2}} e^{-\frac{1}{2\sigma^2}\|y^{(i)} - A^T x^{(i)}\|^2}$

So

$$-\log \text{Likelihood}(A) = \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{m} \|y^{(i)} - A^T x^{(i)}\|^2}_{=\|Y - XA\|_F^2} + \frac{mT}{2}\log(2\pi\sigma^2)$$

where $\|M\|_F^2 = \sum_{i,j} M_{ij}^2 = \mathrm{Tr}(M^T M)$

So $\hat{A}^{MLE} \in \underset{A\in\mathbb{R}^{p\times T}}{\operatorname{argmin}} \|Y - XA\|_F^2$

Remark: we denote by A_k the k-th column of A: $A_k := A[\,,k]$

The MLE optimisation is separable since $\|Y-XA\|_F^2 = \sum_{k=1}^{T} \|Y_k - XA_k\|^2$

So $\hat{A}_k^{MLE} \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|Y_k - X\beta\|^2$, for $k=1,\dots,T$

$\iff$ T simple regressions.

Estimation with hidden low dimensional structures?

2 Sparse estimation

a/ Coordinate sparsity

- Assume that $|A^*|_0 = \mathrm{Card}\{(i,j) : A^*_{ij} \neq 0\}$ small.
- ℓ^1-penalisation:
  - $\hat{A}^{\ell^1} \in \underset{A\in\mathbb{R}^{p\times T}}{\operatorname{argmin}} \{\|Y-XA\|_F^2 + \lambda |A|_1\}$

    with $|A|_1 = \sum_{j,k} |A_{jk}| = \sum_k |A_k|_1$ ← separable
  - T Lasso problems:

    $\hat{A}_k^{\ell^1} \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \{\|Y_k - X\beta\|^2 + \lambda|\beta|_1\}$, $k=1,\dots,T$

b/ Row sparsity

$y^{(i)} = (A^*)^T x^{(i)} + \varepsilon^{(i)} = \sum_{j=1}^{p} (A^*_{j:})^T x_j^{(i)} + \varepsilon^{(i)}$

variable selection $\iff$ row sparsity of A^*: $\mathrm{card}\{j : A^*_{j:} \neq 0\}$ small

$\leadsto$ $\hat{A}^{RS} \in \underset{A\in\mathbb{R}^{p\times T}}{\operatorname{argmin}} \{\|Y-XA\|_F^2 + \lambda \sum_{j=1}^{p} \|A_{j:}\|\}$

"I" Looks like group Lasso?

- Define $\text{Vect}(\underset{\in \mathbb{R}^{d\times T}}{M}) := \begin{bmatrix} M_1 \\ \vdots \\ M_T \end{bmatrix} \in \mathbb{R}^{dT}.$

Then,

$$\text{Vect}(Y) = \underbrace{\begin{bmatrix} X & 0 & & 0 \\ 0 & X & & \vdots \\ 0 & & \ddots & 0 \\ \vdots & & & \\ 0 & \cdots & 0 & X \end{bmatrix}}_{=: \tilde{X} \in \mathbb{R}^{mT\times Tp}} \text{Vect}(A) + \text{Vect}(E) .$$

Setting $G_j = \{k : k \equiv j \ [p]\}$ ← indices corresponding to $A_{j:}$

$$\text{Vect}(\hat{A}^{RS}) \in \underset{\beta \in \mathbb{R}^{Tp}}{\text{argmin}} \left\{ \|\text{vect}(Y) - \tilde{X}\beta\|^2 + \lambda \sum_{j=1}^{p} \|\beta_{G_j}\| \right\}$$

⇝ group-Lasso in dimension pT.

- Look at Theorem 8.6 for a risk bound

③ Low rank regression

Other structures?

- a common situation is that the signal $A^T x$ remains close to some linear span $\underset{\text{unknown}}{V} \subset \mathbb{R}^T$, for all x.
- if $A^T x \in V \ \forall x \in \mathbb{R}^p$ and $\dim(V) \ll T$, then
  - $\text{range}(A^T) \subset V$
  - $\text{rank}(A) = \text{rank}(A^T)$ small.

⇝ estimation with rank constraint. ← non linear!

a/ Refresher on SVD : Appendix C

- SVD: $\underset{\in \mathbb{R}^{m \times p}}{M} = \sum_{k=1}^{r} \sigma_k u_k v_k^T$ with
  - $\sigma_1 \geq \dots \geq \sigma_r > 0$
  - $r = \mathrm{rank}(M)$
  - $(u_1, \dots, u_r)$ orthonormal family in $\mathbb{R}^m$
  - $(v_1, \dots, v_r)$ ———————— $\mathbb{R}^p$

  - $MM^T u_k = \sigma_k^2 u_k$
  - $M^T M v_k = \sigma_k^2 v_k$

- Best low rank approximation (Theorem C.5)

  Define $(M)_{(d)} := \sum_{k=1}^{d} \sigma_k u_k v_k^T$ for $d \leq \mathrm{rank}(M)$. Then

  $$(M)_{(d)} \in \underset{B:\, \mathrm{rank}(B) \leq d}{\mathrm{argmin}} \|M - B\|_F^2$$

  $$\|M - (M)_{(d)}\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2$$

  $$\Rightarrow \|M\|_F^2 = \|(M)_{(d)}\|_F^2 + \|M - (M)_{(d)}\|_F^2$$

- Ky-Fan $(2,q)$ norm (Theorem C.5)

  $$\|M\|_{(2,d)}^2 := \sum_{k=1}^{d} \sigma_k(M)^2 = \|(M)_{(d)}\|_F^2$$

  improved Cauchy-Schwartz: for $d = \mathrm{rank}(A) \wedge \mathrm{rank}(B)$

  $$\langle A, B \rangle_F \leq \|A\|_{(2,d)} \|B\|_{(2,d)}$$

- Weyl inequality (Theorem C.6)

  $A \to \sigma_k(A)$ is 1-Lipschitz with respect to the operator norm.

b/ Some results on random matrices

- $W_{m\times T} \in \mathbb{R}^{m\times T}$ with $[W_{m\times T}]_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$

- classical asymptotic : T fixed, $m \to \infty$

$$\left[\frac{1}{m} W^T W\right]_{ab} = \frac{1}{m}\sum_{i=1}^{m} W_{ia} W_{ib} \xrightarrow[m\to\infty]{} \mathbb{1}_{a\neq b} \quad \text{a.s.} \qquad \text{(L.L.N.)}$$

i.e. $\frac{1}{m} W_{m\times T}^T W_{m\times T} \xrightarrow[m\to\infty]{} I_T$ a.s.

and $\sigma_k\left(\frac{1}{\sqrt{m}} W_{m\times T}\right) \xrightarrow[m\to\infty]{} 1$ a.s. for $k=1,\dots,T$

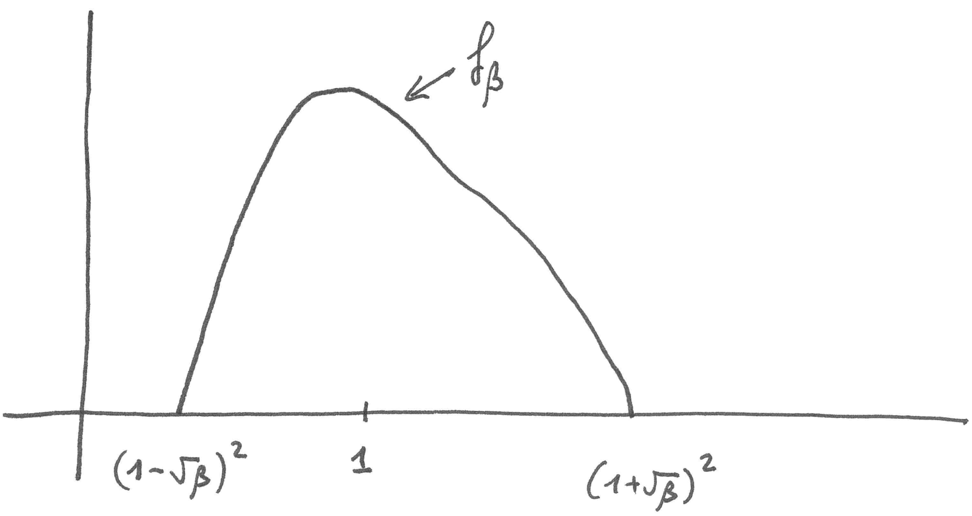
- Marchenko-Pastur asymptotic : $T \sim \beta m$, with $\beta \leq 1$, $\beta > 0$

  - no convergence of $\frac{1}{m} W_{m\times T}^T W_{m\times T}$: we look at the empirical distribution of the singular values

$$d\mu_\omega(x) = \frac{1}{T}\sum_{k=1}^{T} \delta_{\sigma_k^2\left(\frac{1}{\sqrt{m}} W_{m\times T}(\omega)\right)} \Longrightarrow f_\beta(x)\,dx \quad \text{a.s.}$$

where $f_\beta(x) = \frac{1}{2\pi\beta x}\sqrt{\left(x-(1-\sqrt{\beta})^2\right)\left((1+\sqrt{\beta})^2 - x\right)}\;\mathbb{1}_{[(1-\sqrt{\beta})^2,(1+\sqrt{\beta})^2]}(x)$

i.e. For all $F \in C_b(\mathbb{R})$: $\int F(x)\,d\mu_\omega(x) \overset{a.s.}{\to} \int F(x) f_\beta(x)\,dx$



Marchenko-Pastur distribution

- Non-asymptotic : Weyl + Gaussian concentration inequality :
  - There exists $\zeta, \zeta' \sim \mathrm{Exp}(1)$ such that

$$\mathbb{E}[\sigma_1(W_{m\times T})] - \sqrt{2\zeta'} \leq \sigma_1(W_{m\times T}) \leqslant \mathbb{E}[\sigma_1(W_{m\times T})] + \sqrt{2\zeta}$$

- Lemma 8.3 Davidson - Szarek

$$\boxed{\mathbb{E}[\sigma_1(W_{m\times T})] \leqslant \sqrt{m} + \sqrt{T}}$$

We will prove the weaker bound $\mathbb{E}[\sigma_1(W_{m\times T})] \leqslant \sqrt{m} + 5\sqrt{T} + \frac{2}{\sqrt{T}}$

Lemma : There exists $\zeta \sim \mathrm{Exp}(1)$ such that

- $|W^T_{m\times T} W_{m\times T} - m I_T|_{op} \leqslant 2\sqrt{18mT + 8m(1+\zeta)} + 9T + 4(1+\zeta)$
- $\sigma_1(W_{m\times T}) \leqslant \sqrt{m} + 5\sqrt{T} + \frac{1+\zeta}{\sqrt{T}}$

Proof: Since $W^TW - mI_T$ is symmetric, we have

$$|W^TW - mI_T|_{op} = \sup_{u\in\partial B_{\mathbb{R}^T}(0,1)} |\langle (W^TW - mI_T)u, u\rangle|$$

$$= \sup_{u\in\partial B_{\mathbb{R}^T}(0,1)} \big|\, \|Wu\|^2 - m \underbrace{\|u\|^2}_{=1} \big|$$

- concentration of $\|Wu\|^2 - m$: $[Wu]_i = (W_{i:})^T u \overset{iid}{\sim} \mathcal{N}(0, \underbrace{\|u\|^2}_{=1})$

So $Wu \sim \mathcal{N}(0, I_m)$ and from Exercise 1.6.6 we have

$\exists\, \zeta_u, \zeta'_u \sim \mathrm{Exp}(1)$ such that

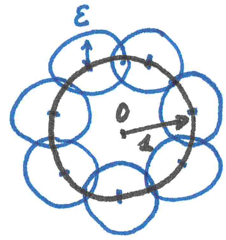
$$-2\sqrt{m\zeta'_u} \leqslant \|Wu\|^2 - m \leqslant \sqrt{8m\zeta_u} + 2\zeta_u$$

so $\big|\|Wu\|^2 - m\big| \leqslant \sqrt{8m\, \zeta_u \vee \zeta'_u} + 2\zeta_u$

- How can we handle $\sup_{u \in \partial B_{\mathbb{R}^T}(0,1)} \cdots$?

discretization of $\partial B_{\mathbb{R}^T}(0,1)$ + union bound.

i) discretization: For any A symmetric



$$|A|_{op} \leqslant \frac{1}{1-2\varepsilon} \sup_{u \in \mathcal{N}_\varepsilon} |\langle Au, u\rangle|$$

where $\mathcal{N}_\varepsilon$ is an ε-net of $\partial B_{\mathbb{R}^T}(0,1)$, i.e.

→ $\mathcal{N}_\varepsilon \subset \partial B_{\mathbb{R}^T}(0,1)$

→ $\forall x \in \partial B_{\mathbb{R}^T}(0,1)$, $\exists y \in \mathcal{N}_\varepsilon$ such that $\|y - x\| \leqslant \varepsilon$

Proof: $|A|_{op} = |\langle Au^*, u^*\rangle|$

$$\underset{\substack{y \in \mathcal{N}_\varepsilon \\ \|y-u^*\| \leqslant \varepsilon}}{=} |\langle Ay, y\rangle + \langle A(u^*-y), y\rangle + \langle Au^*, u^*-y\rangle|$$

$$\leqslant |\langle Ay, y\rangle| + |A|_{op}\,\varepsilon + |A|_{op}\,\varepsilon \qquad \square$$

ii) union bound:

Lemma: There exists $\mathcal{N}_\varepsilon$ an ε-net of $\partial B_{\mathbb{R}^T}(0,1)$ with cardinality

$$|\mathcal{N}_\varepsilon| \leqslant \left(1 + \frac{2}{\varepsilon}\right)^T$$

Proof:

- Take $x_1 \in \partial B_{\mathbb{R}^T}(0,1)$, then $x_2 \in \partial B_{\mathbb{R}^T}(0,1) \setminus B_{\mathbb{R}^T}(x_1, \varepsilon)$, ...

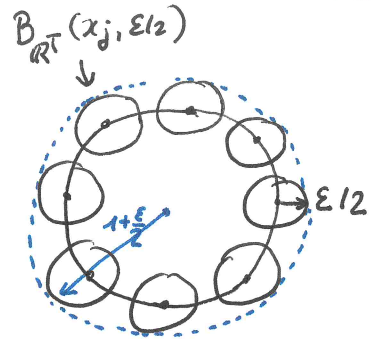
  then $x_k \in \partial B_{\mathbb{R}^T}(0,1) \setminus \bigcup_{j \leqslant k-1} B_{\mathbb{R}^T}(x_j, \varepsilon)$, ..., until impossible.

- $\mathcal{N}_\varepsilon = \{x_1, x_2, \dots\}$

- by construction: • $\mathcal{N}_\varepsilon$ is an ε net

  • $\|x - y\| \geqslant \varepsilon \quad \forall x, y \in \mathcal{N}_\varepsilon,\ x \neq y$.

Hence



$$\bigsqcup_{x\in\mathcal{N}_\varepsilon} B_{\mathbb{R}^T}(x, \varepsilon/2) \subset B_{\mathbb{R}^T}(0, 1+\tfrac{\varepsilon}{2})$$

comparing the volumes

$$|\mathcal{N}_\varepsilon| \times \left(\frac{\varepsilon}{2}\right)^T v_T(1) \leq \left(1+\frac{\varepsilon}{2}\right)^T v_T(1)$$

$\square$

- In particular, we can choose $\mathcal{N}_{1/4}$ with $|\mathcal{N}_{1/4}| \leq 9^T$

and

$$|W^TW - mI_T|_{op} \leq 2 \max_{u\in\mathcal{N}_{1/4}} \left(\sqrt{8m\, \zeta_u' v \zeta_u'} + 2\zeta_u\right)$$

- union bound

$$\mathbb{P}\left[\max_{u\in\mathcal{N}_{1/4}} \zeta_u' v \zeta_u' \geq \log(2|\mathcal{N}_{\frac{1}{4}}|)+t\right] \leq 2|\mathcal{N}_{1/4}|\, e^{-\log(2|\mathcal{N}_{1/4}|)-t} = e^{-t}$$

So $\exists \zeta \sim \mathcal{E}xp(1)$:

$$|W^TW - mI_T|_{op} \leq 2\sqrt{8m\left(\log(2\times 9^T)+\zeta\right)} + 4\log(2\times 9^T) + 4\zeta$$

$$\leq 2\sqrt{18mT + 8m(1+\zeta)} + 9T + 4(1+\zeta) \,.$$

- In addition:

$$|W^TW|_{op} \leq \left(\sqrt{m} + \sqrt{18T + 8(1+\zeta)}\right)^2$$

So
$$\sigma_1(W) = |W^TW|_{op}^{1/2} \leq \sqrt{m} + \sqrt{18T+8(1+\zeta)}$$

$$\leq \sqrt{m} + \sqrt{18T} + \frac{8(1+\zeta)}{2\sqrt{18T}}$$

$$\leq \sqrt{m} + \sqrt{18T} + \frac{1+\zeta}{\sqrt{T}}$$

$\square$

Corollary:

$$\boxed{\begin{array}{l}\text{Let } P \in \mathbb{R}^{m\times m} \text{ be a projector on a linear span of dimension } d.\\ \text{Then } \mathbb{E}\left[\,|P W_{m\times T}|_{op}\right] \leq \sqrt{d} + \sqrt{T}\end{array}}$$

Proof: . $P = UU^T$ with $U \in \mathbb{R}^{m\times d}$ with orthonormal columns

$\|PWx\| = \|U^T W x\|$ so $|PW|_{op} = |U^T W|_{op}$

. $[U^T W]_j = \underbrace{U^T W_j}_{\sim \mathcal{N}(0, I_m)} \sim \mathcal{N}(0, \underbrace{U^T U}_{I_d})$

So $U^T W_{m\times T} \overset{(d)}{=} W_{d\times T}$ and the result follows from Davidson-Szarek lemma $\square$

c/ Estimation with known rank

Reminder: we have in mind that $\mathrm{range}(A^T) \subset V$ with V a linear span of dimension $r \ll T$.

$\to S_V = \{A \in \mathbb{R}^{p\times T} : \mathrm{range}(A^T) \subset V\}$ is a linear span.

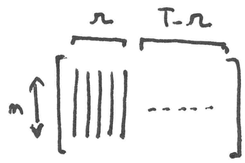
but the family $\{S_V : \dim(V) = r\}$ is uncountable, so we cannot apply model selection on it.

$\to$ instead we directly look at

$$\mathcal{S}_r := \{A \in \mathbb{R}^{p\times T} : \mathrm{rank}(A) = r\}$$

⚠ it is not a linear span

It is a submanifold of dimension $= rm + (T-r)r$
$= r(m+T) - r^2$



Constrained MLE:

$$\hat{A}_r \in \underset{\mathrm{rank}(A) \leq r}{\mathrm{argmin}} \ \|Y - XA\|_F^2 \quad \longleftarrow \quad ⚠ \text{ non convex !}$$

Computation: (Lemma 8.1)

Set $P := X(X^TX)^+X^T = \mathrm{Proj}_{\mathrm{range}(X)}$ - Then

- $X\hat{A}_r = (PY)_{(r)}$
- $\hat{A}_r = (X^TX)^+X^T(PY)_{(r)}$

Proof:

- $$\|Y - XA\|_F^2 = \sum_{k=1}^{T} \|Y_k - XA_k\|^2 \overset{\text{Pythagore}}{=} \sum_{k=1}^{T} \left(\|Y_k - PY_k\|^2 + \|PY_k - XA_k\|^2\right)$$
$$= \|Y - PY\|_F^2 + \|PY - XA\|_F^2$$

- We observe that

$\rightarrow$ $\|Y - (PY)_{(r)}\|_F^2 \leq \|Y - X\hat{A}_r\|_F^2$ since $\mathrm{rank}(X\hat{A}_r) \leq r$

$\rightarrow$ $\|PY - (PY)_{(r)}\|_F^2 \overset{\text{Pyth.}}{=} \|PY - P(PY)_{(r)}\|_F^2 + \|(I-P)(PY)_{(r)}\|_F^2$

with $\mathrm{rank}(P(PY)_{(r)}) \leq r$, so

$$(PY)_{(r)} = P(PY)_{(r)} = X \underbrace{(X^TX)^+X^T(PY)_{(r)}}_{\mathrm{rank} \leq r}$$

. **Conclusion:** $X\hat{A}_r = (PY)_{(r)}$ and $\hat{A}_r = (X^TX)^+X^T(PY)_{(r)}$ □

☺ . $\hat{A}_r$ can be computed from a partial SVD of PY

. $\{\hat{A}_r : r = 1, .., \mathrm{rank}(PY)\}$ can be computed from a <u>single</u> SVD of PY

Risk bound

Proposition 8.2 - Corollary 8-4:

(i) $\|X\hat{A}_r - XA^*\|_F^2 \leq 9 \min_{\text{rank}(A)\leq r} \|XA - XA^*\|_F^2 + 12 r |PE|_{op}^2$

(ii) Setting $q = \text{rank}(X)$

$$\mathbb{E}\left[\|X\hat{A}_r - XA^*\|_F^2\right] \leq 9 \min_{\text{rank}(A)\leq r} \|XA - XA^*\|_F^2 + 36 r (\sqrt{q}+\sqrt{T})^2 \sigma^2$$

Proof: . (ii) follows from (i) and

$$\mathbb{E}\left[|PE|_{op}^2\right] \underset{\substack{\uparrow \\ \text{Gaussian concentration}}}{\leq} \mathbb{E}\left[\left(\mathbb{E}[|PE|_{op}] + \sigma\sqrt{2\xi}\right)^2\right] \leq 2 \underbrace{\mathbb{E}\left[|PE|_{op}\right]^2}_{\leq \sigma^2(\sqrt{q}+\sqrt{T})^2} + 4\sigma^2 \underbrace{\mathbb{E}[\xi]}_{=1}$$

$$\leq 3\sigma^2(\sqrt{q}+\sqrt{T})^2$$

. Let us prove (i). Let B_r be such that $(XA^*)_{(r)} = XB_r$ with $\text{rank}(B_r) \leq r$.
Starting from $\|Y - X\hat{A}_r\|_F^2 \leq \|Y - XB_r\|_F^2$ and $Y = XA^* + E$ we get

$$\|XA^* - X\hat{A}_r\|_F^2 \leq \|XA^* - XB_r\|_F^2 + 2\underbrace{\langle E, X\hat{A}_r - XB_r\rangle_F}$$

$$= \langle PE, X\hat{A}_r - XB_r\rangle_F$$

$$\text{rank}(X\hat{A}_r - XB_r) \leq 2r \longrightarrow \quad \leq \|PE\|_{(2,2r)} \|X\hat{A}_r - XB_r\|_{(2,2r)}$$

$$\leq \sqrt{2r}\, |PE|_{op} \underbrace{\|X\hat{A}_r - XB_r\|_F}_{\leq \|X\hat{A}_r - XA^*\|_F + \|XA^* - XB_r\|_F}$$

$2xy \leq ax + \frac{1}{a}y$

$$\longrightarrow \leq \left(1+\frac{1}{b}\right)\|XA^* - XB_r\|_F^2 + \frac{1}{a}\|X\hat{A}_r - XA^*\|_F^2 + (a+b)\times 2r|PE|_{op}^2$$

Set $a = 3/2$ and $b = 1/2$ to conclude.

$\square$

d/ Rank selection

$$\hat{r} \in \underset{r=1,\dots,q\wedge T}{\operatorname{argmin}} \left\{ \|Y - X\hat{A}_r\|_F^2 + \lambda r \right\}$$

where $\lambda = K(\sqrt{T}+\sqrt{q})^2\sigma^2$ with $K>1$ and $q = \operatorname{rank}(X)$

Theorem 8.5: Oracle risk bound

$$\mathbb{E}\left[\|X\hat{A}_{\hat{r}} - XA^*\|_F^2\right] \leq c_K \min_{r=1,\dots,q\wedge T} \left\{ \mathbb{E}\left[\|X\hat{A}_r - XA^*\|_F^2\right] + r(\sqrt{T}+\sqrt{q})^2\sigma^2 \right\}$$

$$\leq c'_K \min_{A\in\mathbb{R}^{p\times T}} \left\{ \|XA - XA^*\|_F^2 + \operatorname{rank}(A)\,(T+q)\,\sigma^2 \right\}.$$

Proof: Same arguments as before: since $\|Y - X\hat{A}_{\hat{r}}\|_F^2 + \lambda\hat{r} \leq \|Y - X\hat{A}_r\|_F^2 + \lambda r$

$$\|X\hat{A}_{\hat{r}} - XA^*\|_F^2 \leq \|X\hat{A}_r - XA^*\|_F^2 + \lambda r + 2\underbrace{\langle PE, X\hat{A}_{\hat{r}} - X\hat{A}_r\rangle_F}_{} - \lambda\hat{r}$$

$$\leq \|PE\|_{(2,r+\hat{r})} \|X\hat{A}_{\hat{r}} - X\hat{A}_r\|_{(2,r+\hat{r})}$$

$$\leq \sqrt{r+\hat{r}}\,|PE|_{op}\left(\|X\hat{A}_{\hat{r}} - XA^*\|_F + \|XA^* - X\hat{A}_r\|_F\right)$$

with $2xy \leq ax + \frac{1}{a}y$

$$\left(1-\frac{1}{a}\right)\|X\hat{A}_{\hat{r}} - XA^*\|_F^2 \leq \left(1+\frac{1}{b}\right)\|X\hat{A}_r - XA^*\|_F^2 + \lambda r + \underbrace{(a+b)(r+\hat{r})|PE|_{op}^2 - \lambda\hat{r}}_{Z_r}$$

It remains to check that $\mathbb{E}[Z_r] \leq c(\sqrt{q}+\sqrt{T})^2\sigma^2 r$

- Since $\hat{r} \leq q\wedge T$ and $|PE|_{op} \leq (\sqrt{T}+\sqrt{q})\sigma + \sigma\sqrt{2\xi}$, with $a = \frac{3+K}{4}$ and $b = \frac{K-1}{4}$

$$(a+b)\,\hat{r}\,|PE|_{op}^2 - \lambda\hat{r} \leq \hat{r}\left(\frac{1+K}{2}(\sqrt{T}+\sqrt{q}+\sqrt{2\xi})^2 - K(\sqrt{T}+\sqrt{q})^2\right)\sigma^2$$

$$\xrightarrow[\text{with } a=\frac{K-1}{K+1}]{2xy\leq ax+\frac{1}{a}y} \leq (q\wedge T)\cdot\frac{2K(1+K)}{K-1}\,\xi\,\sigma^2$$

Hence $\mathbb{E}[Z_r] \leq 3\,\frac{K+1}{2}\,r\,(\sqrt{q}+\sqrt{T})^2\sigma^2 + \frac{2K(1+K)}{K-1}(q\wedge T)\,\sigma^2$

$$\leq C_K\, r\,(\sqrt{q}+\sqrt{T})^2\sigma^2$$

□

③ Low rank and row sparse matrices

. Can we handle matrices simultaneously low rank and row sparse?

. With model selection: yes, but prohibitive computational cost.

Benchmark: if $r^* = \text{rank}(A^*)$ and $k^* = \text{card}\{j : A^*_{j:} \neq 0\}$

$$\mathbb{E}\left[\|X\hat{A}^{ms} - XA^*\|_F^2\right] \leq c\Big(\underbrace{r^*(T+k^*)}_{\text{low rank with } k^* \text{ rows}} + \underbrace{k^* \log \frac{ep}{k^*}}_{\text{complexity of rows identification}}\Big)\sigma^2 \quad \text{(Theorem 8.7)}$$

. Convex relaxation?

with group lasso:

$$\underset{A:\,\text{rank}(A) \leq r}{\text{argmin}} \left\{ \|Y - XA\|_F^2 + \lambda \sum_{j=1}^{p} \|A_{j:}\| \right\}$$

→ non convex :(

relaxation?

$$\text{rank}(A) = \sum_k \mathbb{1}_{\sigma_k(A) \neq 0} \quad \rightsquigarrow \quad \|A\|_* = \sum_k \sigma_k(A)$$

(nuclear norm).

Does nuclear norm penalization works?

$$\hat{A}_\lambda^{NN} \in \underset{A \in \mathbb{R}^{p \times T}}{\text{argmin}} \left\{ \|Y - XA\|_F^2 + \lambda |A|_* \right\}$$

Theorem 8.8

For $\lambda = 2K\sigma_1(X)(\sqrt{T} + \sqrt{q})\sigma$, with $K > 1$, $q = \text{rank}(X)$, we have with $\mathbb{P} \geq 1 - \exp\left(-(K-1)^2 \frac{T+q}{2}\right)$

$$\|X\hat{A}_\lambda - XA^*\|_F^2 \leq \inf_A \left\{ \|XA - XA^*\|_F^2 + 9K^2 \underbrace{\left(\frac{\sigma_1(X)}{\sigma_q(X)}\right)^2}_{\text{price for convexification}} (\sqrt{T} + \sqrt{q})^2 \sigma^2 \,\text{rank}(A) \right\}$$

(proof similar as for Lasso)

Convex criterion for Low-rank and row sparse

$$\hat{A}^{cvx} \in \underset{A \in \mathbb{R}^{p \times T}}{\operatorname{argmin}} \Big\{ \underbrace{\|Y - XA\|_F^2 + \lambda \sum_{j=1}^{p} \|A_{j:}\| + \mu |A|_*}_{\text{convex}} \Big\}$$

- can be computed
- no improvement % low rank or row sparse alone
- why ?

⤳ bias cumulate

Iterative algorithm ?

idea 1: decompose $A = UV$ with $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{r \times T}$

problem $UV = (\alpha U)(\frac{1}{\alpha}V)$ so the sizes of U and V must be stabilized

idea 2: Consider $F(U,V) = \underbrace{\|Y - XUV\|_F^2}_{\text{data fit}} + \underbrace{\frac{1}{2}\|U^T U - V^T V\|_F^2}_{\text{scale stabilization}}$

idea 3: proximal iterations related to

$$\min_{U,V} F(U,V) + \lambda |J(U)| \qquad \text{where } |J(U)| = \operatorname{card}\{j : A_{j:} \neq 0\}$$

$$\begin{pmatrix} U^{t+1} \\ V^{t+1} \end{pmatrix} \leftarrow \begin{pmatrix} H_\lambda^G(U^t - \eta \nabla_U F(U^t, V^t)) \\ V^t - \eta \nabla_V F(U^t, V^t) \end{pmatrix} \quad \text{with } H_\lambda^G = \text{group thresholding.}$$

Good properties: for t large enough ($\geq C \log n$) and under some restricted isometry property + initialisation with $\hat{A}^{RS}$

$$\|XU^tV^t - XA^*\|_F^2 \leq C\left(r^*(T + k^*) + k^* \log p\right)\sigma^2$$

Conclusion:

Convex relaxation	Iterative algorithm
$\boxed{0}\boxed{0}$	$\boxed{0}\ \boxed{1}$