High Dimensional Statistics

Yanbo Tang

Imperial College London

March. 2025

Content of this week

- Covariance estimation
- PCA
- Matrix Regression

Some useful matrix inequalities

$\lambda \to$ singular values.

$A = \sum \lambda_i u_i^T v_i$

$\|A\|_F^2 = \sum_{i=1}^{n} \lambda_i(A)^2$ (1)

Weyl's inequality $\max_i |\lambda_i(A) - \lambda_i(B)| \le \|A - B\|$ (2)

Hoffman: $\sum_i (\lambda_i(A) - \lambda_i(B))^2 \le \|A - B\|_F^2$ (3)

$\langle A, B \rangle = tr(A^T B)$

Hölder's inequality $|\langle A, B \rangle| \le \|A\|_p \|B\|_q$ $\frac{1}{q} + \frac{1}{p} = 1$ (4)

where $\|A\|_p = \left(\sum_{i=1}^{r} \lambda_i(A)^p \right)^{1/p}$

$|\langle A, B \rangle| = |tr(A^T B)| \le \sum |\lambda_i(B)| \cdot \|A\|$

Matrix Bernstein

Theorem

(Matrix Bernstein's inequality). Let $X_1, \ldots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N} X_i\right\| \geq t\right\} \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Here $\sigma^2 = \left\|\sum_{i=1}^{N} \mathbb{E}X_i^2\right\|$ is the norm of the matrix variance of the sum.

Expectation

Lemma

(Matrix Bernstein's inequality: expectation). Let $X_1, \ldots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i.

$$\mathbb{E}\left\|\sum_{i=1}^{N} X_i\right\| \lesssim \left\|\sum_{i=1}^{N} \mathbb{E}X_i^2\right\|^{1/2} \sqrt{1+\log n} + K(1+\log n).$$

Covariance estimation

We can estimate the second moment matrix $\Sigma = \mathbb{E}XX^T$ by its sample version

$$\hat{\Sigma}_m = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T.$$

Recall that if X has zero mean, then Σ is the covariance matrix of X and Σ_m is the sample covariance matrix of X.

Covariance Estimation sub-Gaussian

Theorem

Let $Y \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}[Y] = 0$, $\mathbb{E}[YY^\top] = I_d$ and $Y \sim subG_d(1)$. Let $X_1, \ldots, X_n$ be n independent copies of sub-Gaussian random vector $X = \Sigma^{1/2} Y$. Then $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^\top] = \Sigma$ and $X \sim subG_d(\|\Sigma\|_{op})$. Moreover,

$$\|\hat{\Sigma} - \Sigma\|_{op} \lesssim \|\Sigma\|_{op} \left(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n} \right),$$

with probability $1 - \delta$.

General covariance estimation

Theorem

(General covariance estimation). Let X be a random vector in $\mathbb{R}^n$, $n \geq 2$. Assume that for some $K \geq 1$,

$$\|X\|_2 \leq K(\mathbb{E}\|X\|_2)^{1/2} \quad \text{almost surely.} \tag{5.16}$$

Then, for every positive integer m, we have

$$\mathbb{E}\|\Sigma_m - \Sigma\| \lesssim \left(\sqrt{\frac{K^2 d \log d}{n}} + \frac{K^2 d \log d}{n} \right) \|\Sigma\|.$$

Proof

$$\mathbb{E}[\|X\|_2^2] = \mathrm{tr}(\Sigma)$$

$$\|X\|_2^2 \le k^2 \mathrm{tr}(\Sigma)$$

$$\mathbb{E}[\|\hat{\Sigma} - \Sigma\|] = \frac{1}{n}\mathbb{E}[\|\sum_{i=1}^{n}(X_i X_i^T - \Sigma)\|] \le \frac{1}{n}(\sigma\sqrt{\log(d)} + M\log(d))$$

$$\sigma^2 = n\|\mathbb{E}(XX^T - \Sigma)^2\| = \|\sum_{i=1}^{n}\mathbb{E}(X_i X_i^T - \Sigma)\|$$

M : $\|XX^T - \Sigma\| \le M$ a.s.

σ^2

$$\mathbb{E}[(XX^T - \Sigma)^2] = \mathbb{E}[(XX^T)^2] - \Sigma^2 \preceq \mathbb{E}[(XX^T)^2]$$

$$(XX^T)^2 = \|X\|^2 XX^T \preceq k^2 \mathrm{tr}(\Sigma) XX^T$$

$$\Rightarrow \mathbb{E}[(XX^T)^2] \preceq k^2 \mathrm{tr}(\Sigma)\|\Sigma\|$$

$$\Rightarrow \sigma^2 \le k^2 n \,\mathrm{tr}(\Sigma)\|\Sigma\|$$

Proof cont.

$$\|XX^T - \Sigma\| \le \|X\|_2^2 + \|\Sigma\|$$

A probability bound

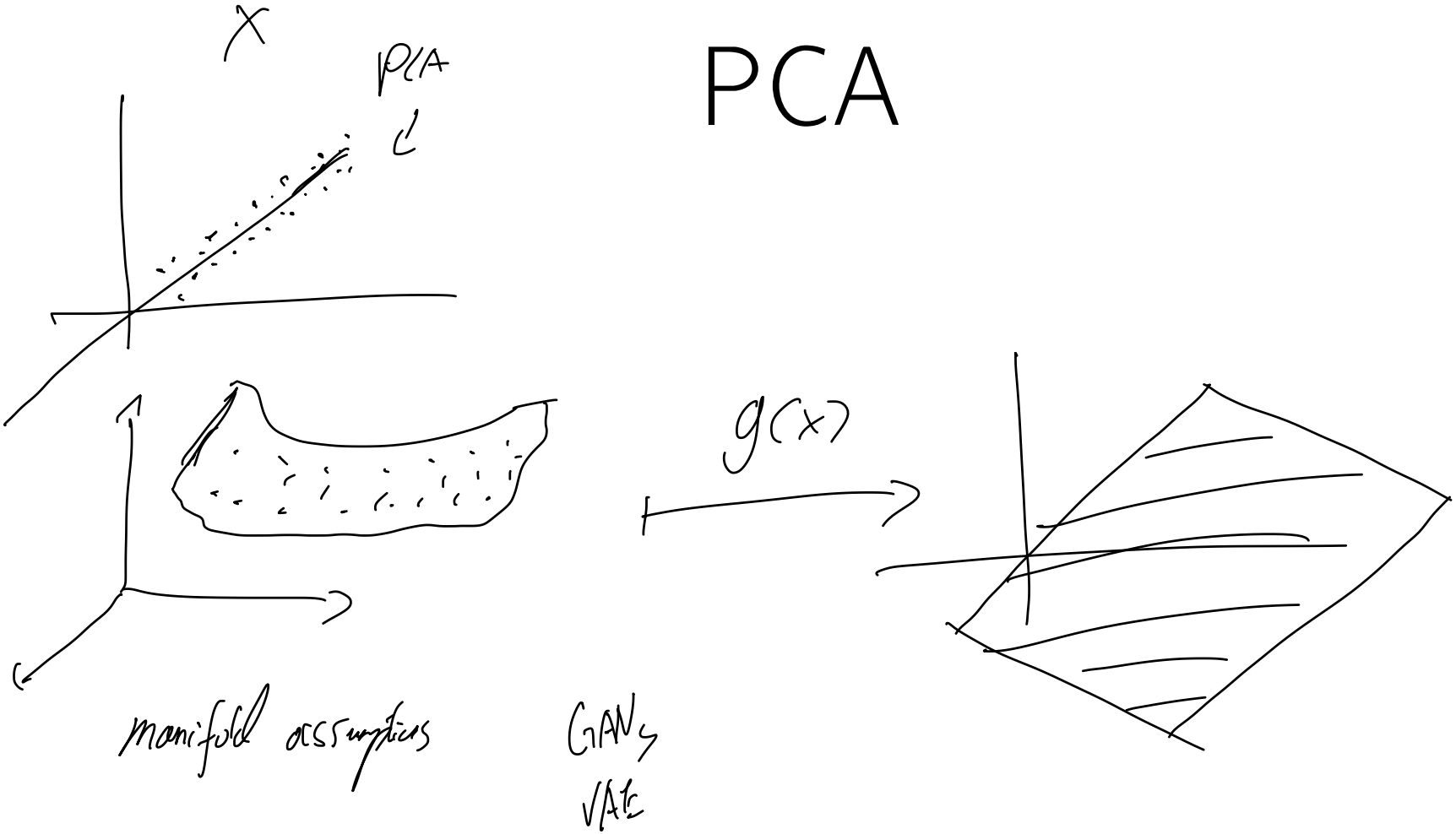
Lemma

Under the same assumption as the previous theorem,we have

$$\|\hat{\Sigma} - \Sigma\| \lesssim \left(\sqrt{\frac{K^2 d(\log d + \log(2/\delta))}{n}} + \frac{K^2 d(\log d + \log(2/\delta))}{n} \right) \|\Sigma\|$$

with probability at least $1 - \delta$.

PCA



Eckart-Young-Mirsky

Lemma

Let A be a rank-r matrix with singular value decomposition

$$A = \sum_{i=1}^{r} \lambda_i u_i v_i^\top,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ are the ordered singular values of A. For any $k < r$, let $A_k = \sum_{i=1}^{k} \lambda_i u_i v_i^\top$. Then for any matrix B such that $\mathrm{rank}(B) \leq k$, it holds

$$\|A - A_k\|_F \leq \|A - B\|_F.$$

Moreover,

$$\text{(i)} \quad \|A - A_k\|_F^2 = \sum_{j=k+1}^{r} \lambda_j^2.$$

Proof

(1) is by definition

since $A - A_k = \sum_{j=k+1}^{r} \lambda_j u_j v_j^T$

so $\|A - A_k\|_F^2 = \sum_{j=k+1}^{r} \lambda_j^2$

$\forall B$ of rank $\le k$

$\sigma_1 \ge \sigma_2 \ge \sigma_3$

are singular value of B

$$\|A - B\|_F^2 = \sum_{j=1}^{r} (\lambda_j - \sigma_j)^2$$

$$= \sum_{j=1}^{k} (\lambda_j - \sigma_j)^2 + \sum_{j=k+1}^{r} \lambda_j^2$$

$$\ge \sum_{j=k+1}^{r} \lambda_j^2$$

Spiked covariance

For a fixed direction $v \in S^{d-1}$, and consider the sequence $Y_1, \ldots, Y_n \sim N(0, I_d)$, then the vectors $v^\top Y_i v$ all live in the one dimensional space spanned by v. Typically we would have some noise in our observations so that they will be closer to:

$$X_i = v^\top Y_i v + Z_i,$$

for a noise vector $Z_i \sim N(0, \sigma^2 I_d)$. Note that the covariance matrix of X_i will be:

$$\Sigma = E\left[XX^\top\right] = vv^\top + \sigma^2 I_d.$$

Spiked covariance

Definition

A covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is a spiked covariance matrix if:

$$\Sigma = \theta v v^\top + I_d,$$

for $\theta > 0$ and $v \in S^{d-1}$. The vector v is called the spike.

$V_1(\hat{\Sigma}) = \hat{v}$

Davis Kahan reminder

Theorem

(Davis-Kahan) Let S and T be symmetric matrices with the same dimensions. Fix i and assume that the i-th largest eigenvalue of S is well separated from the rest of the spectrum:

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0.$$

Then the angle between the eigenvectors of S and T corresponding to the i-th largest eigenvalues (as a number between 0 and $\pi/2$) satisfies

$$\sin \angle(v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}.$$

which implies

$$\exists \theta \in \{-1, 1\} : \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{2^{3/2}\|S - T\|}{\delta}.$$

Guarantees

Corollary

Let $Y \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}[Y] = 0$, $\mathbb{E}[YY^T] = I_d$ and $Y \sim subG_d(1)$. Let $X_1, \ldots, X_n$ be n independent copies of sub-Gaussian random vector $X = \Sigma^{1/2} Y$ so that $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^T] = \Sigma$ and $X \sim subG_d(\|\Sigma\|_{op})$. Assume further that $\Sigma = \theta vv^T + I_d$ satisfies the spiked covariance model. Then, the largest eigenvector $\hat{v}$ of the empirical covariance matrix $\hat{\Sigma}$ satisfies,

$$\min_{e \in \{\pm 1\}} \|e\hat{v} - v\|_2 \lesssim \frac{1+\theta}{\min(\theta, 1)} \left(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n} \right)$$

with probability $1 - \delta$.

Proof

By Davis-Kahan $\|V_1(\Sigma) - V_1(\hat{\Sigma})\|_2 \le 2^{3/2} \frac{\|\Sigma - \hat{\Sigma}\|}{\delta}$

If $\theta, \theta, 0$ $\quad \delta := \min(1, \theta)$

$\|\Sigma - \hat{\Sigma}\|$

Sparse PCA

If we assume that v in the spiked covariance model is k-sparse: $|v|_0 = k$. Therefore, a natural candidate to estimate v is given by $\hat{v}$ defined by

$$\hat{v}^\top \hat{\Sigma} \hat{v} = \max_{u \in S^{d-1}, |u|_0 = k} u^\top \hat{\Sigma} u.$$

It is the case that $\lambda_{\max}^k(\hat{\Sigma}) = \hat{v}^\top \hat{\Sigma} \hat{v}$ is the largest of all leading eigenvalues among all $k \times k$ sub-matrices of $\hat{\Sigma}$ so that the maximum is indeed attained.

Sparse PCA

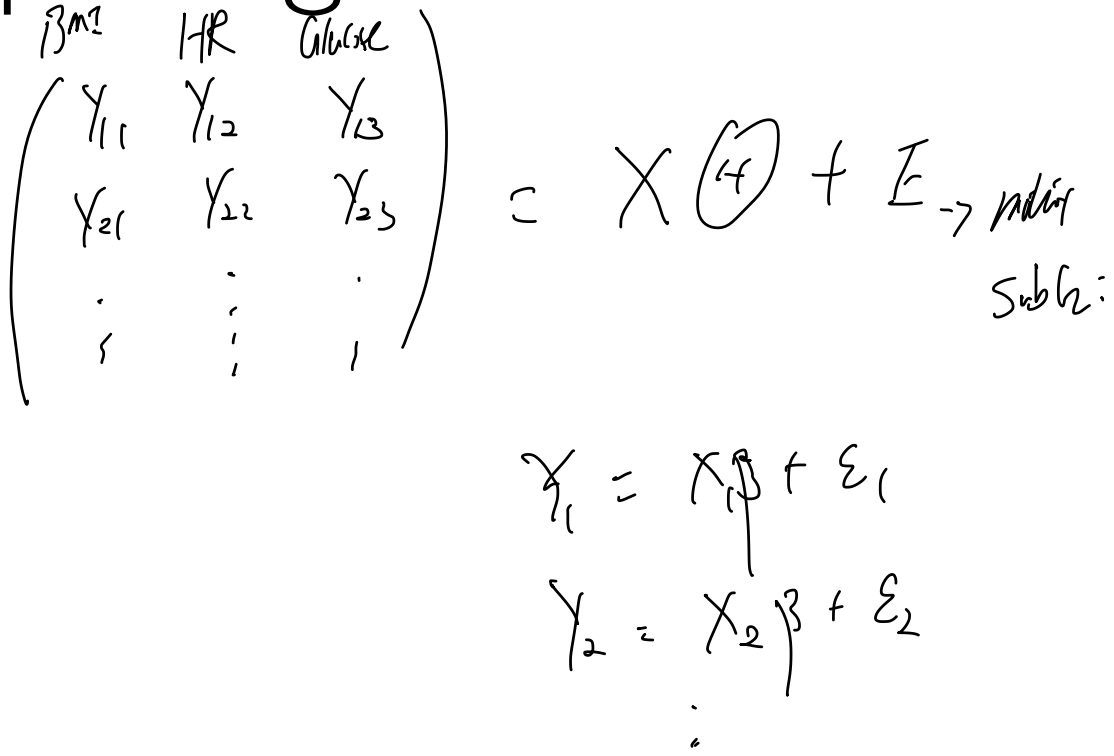
Theorem

Let $Y \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}[Y] = 0$, $\mathbb{E}[YY^\top] = I_d$ and $Y \sim subG_d(1)$. Let $X_1, \ldots, X_n$ be n independent copies of sub-Gaussian random vector $X = \Sigma^{1/2} Y$ so that $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^\top] = \Sigma$ and $X \sim subG_d(\|\Sigma\|_{op})$. Assume further that $\Sigma = \theta vv^\top + I_d$ satisfies the spiked covariance model for v such that $|v|_0 = k \leq d/2$. Then, the k-sparse largest eigenvector $\hat{v}$ of the empirical covariance matrix satisfies,

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \hat{v} - v\|_2 \lesssim \frac{1+\theta}{\min(\theta, 1)} \left(\sqrt{\frac{k \log(ed/k) + \log(1/\delta)}{n}} \vee \frac{k \log(ed/k) + \log(1/\delta)}{n} \right)$$

with probability $1 - \delta$.

Multiple regression



Introduction

A simple question to ask is can we extend the classical regression set up to matrices? Specifically, is it useful to consider a model such that:

$$\mathbb{Y} = \mathbb{X}\Theta^{\star} + E,$$

where $Y \in \mathbb{R}^{n \times T}$, $\mathbb{X} \in \mathbb{R}^{n \times d}$ and $\Theta \in \mathbb{R}^{d \times T}$ is the unkown matrix of coefficents for some noise matrix $E \sim subG_{n \times T}(\sigma)$.

Definition

We call a $n \times m$ matrix A $subG_{n \times m}(\sigma)$ if for every $x \in S^{m-1}$ and $y \in S^{n-1}$ if:

$$y^{\top} A x \sim subG(\sigma^2).$$

Direct observation model

We are going to make our life simple (at first) and assume that we have an orthogonal design (ORT condition) for our design matrix, i.e., $X^\top X = nI_d$. Under the ORT assumption,

$$\frac{1}{n}X^\top Y = \Theta^* + \frac{1}{n}X^\top E.$$

Which can be written as an equation in $\mathbb{R}^{d\times T}$ called the *sub-Gaussian matrix model (sGMM)*:

$$y = \Theta^* + F,$$

where $y = \frac{1}{n}X^\top Y$ and $F = \frac{1}{n}X^\top E \sim \mathrm{subG}_{d\times T}(\sigma^2/n)$.

If $\Theta^\star$ is sparse then it is possible to estimate it from a single observation. Consider the SVD of Θ^*:

$$\Theta^* = \sum_j \lambda_j u_j v_j^\top.$$

and define $\|\Theta^*\|_0 := |\lambda|_0$. Therefore, if we knew u_j and v_j, we could simply estimate the λ_js thresholding. It turns out that estimating the eigenvectors by themselves is sufficient. Consider the SVD of the observed matrix y:

$$y = \sum_j \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top.$$

Singular value tresholding

Definition The **singular value thresholding** estimator with threshold $\tau_n \geq 0$ is defined by

$$\hat{\Theta}^{\mathrm{SVT}} = \sum_j \hat{\lambda}_j \mathbb{I}(|\hat{\lambda}_j| > \tau_n) \hat{u}_j \hat{v}_j^\top .$$

Singular value tresholding

Definition The **singular value thresholding** estimator with threshold $2\tau \geq 0$ is defined by

$$\hat{\Theta}^{\mathsf{SVT}} = \sum_j \hat{\lambda}_j \mathbb{I}(|\hat{\lambda}_j| > \tau_n) \hat{u}_j \hat{v}_j^\top .$$

Lemma

Let A be a $d \times T$ random matrix such that $A \sim subG_{d\times T}(\sigma^2)$. Then

$$\|A\|_{\not op} \leq 4\sigma\sqrt{\log(12)(d \vee T)} + 2\sigma\sqrt{2\log(1/\delta)}$$

with probability $1 - \delta$.

Theorem

Consider the multivariate linear regression model under the assumption ***ORT*** *or, equivalently, the sub-Gaussian matrix model. Then, the singular value thresholding estimator* $\hat{\Theta}^{SVT}$ *with threshold*

$$\tau_n = 8\sigma\sqrt{\frac{\log(12)(d \vee T)}{n}} + 4\sigma\sqrt{\frac{2\log(1/\delta)}{n}},$$

satisfies

$$\frac{1}{n}\|\bar{X}\hat{\Theta}^{SVT} - \bar{X}\Theta^*\|_F^2 = \|\hat{\Theta}^{SVT} - \Theta^*\|_F^2 \leq 36\,\mathrm{rank}(\Theta^*)\tau_n^2$$

$$\lesssim \frac{\sigma^2\,\mathrm{rank}(\Theta^*)}{n}\left(d \vee T + \log(1/\delta)\right).$$

with probability $1 - \delta$.

Proof

Θ^* $\lambda_1 \geq \lambda_2 \geq \cdots$

Y $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \geq \cdots$

define $S = \{j : |\lambda_j| > \tau_n\}$ $\|F\| \leq \tau_n/2$ w.p. $1-\delta$

$|\hat{\lambda}_j - \lambda_j| \leq \|F\| \leq \tau_n/2$ by Weyl's

$S \subset \{j : |\lambda_j| > \tau_n/2\}$ $S^c \subset \{j : |\lambda_j| \leq 3\tau_n/2\}$

$|\hat{\lambda}_j| = |\hat{\lambda}_j - \lambda_j + \lambda_j| \geq |\lambda_j| - |\lambda_j - \hat{\lambda}_j| \geq \tau_n - \tau_n/2 \geq \tau_n/2$

$\bar{\Theta} = \sum_{j \in S} \lambda_j u_j v_j^T$

$\|\hat{\Theta} - \Theta^*\|_F^2 \leq 2\|\hat{\Theta} - \bar{\Theta}\|_F^2 + 2\|\bar{\Theta} - \Theta^*\|_F^2$ Young's inequ.

(1) (2)

(1) $\|\hat{\Theta} - \bar{\Theta}\|_F^2 \leq \operatorname{rank}(\hat{\Theta} - \bar{\Theta})\|\hat{\Theta} - \bar{\Theta}\|^2 \leq 2|S|\,\|\hat{\Theta} - \bar{\Theta}\|^2$

$\|\hat{\Theta} - \bar{\Theta}\| \leq \|\hat{\Theta} - Y\| + \|Y - \Theta^*\| + \|\Theta^* - \bar{\Theta}\| \leq \max_{j \in S^c} |\hat{\lambda}_j| + \tau_n/2 + \max_{j \in S^c} |\lambda_j|$

$\leq \tau_n$ $3\tau_n/2$

(2) $\leq \sum_{j \in S^c} \lambda_j^2$

Without ORT

$\leq 3\mathcal{Y}$

We can then consider penalizing the rank of the matrix instead of direct truncation in cases when we don't have the ORT assumption. Let $\hat{\Theta}^{\mathrm{RK}}$ be any solution to the following minimization problem:

$$\min_{\Theta \in \mathbb{R}^{d \times T}} \left\{ \frac{1}{n} \| Y - X\Theta \|_F^2 + \tau_n^2 \, \mathrm{rank}(\Theta) \right\}.$$

Guarantees

Theorem

Consider the multivariate linear regression model (5.1). Then, the estimator by rank penalization $\hat{\Theta}^{RK}$ with regularization parameter τ_n^2, where τ_n is defined in as in the previous theorem, satisfies

$$\frac{1}{n}\|X\hat{\Theta}^{RK} - X\Theta^*\|_F^2 \leq 2\,\mathrm{rank}(\Theta^*)\tau^2 \lesssim \frac{\sigma^2\,\mathrm{rank}(\Theta^*)}{n}\left(d\sqrt{T} + \log(1/\delta)\right),$$

with probability $1-\delta$.

Proof

Why can we solve this problem efficiently

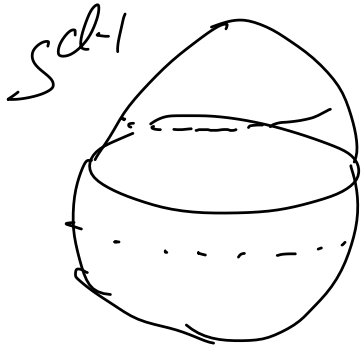
Concluding remarks

- Concentration inequalities → Isoperimetric inequalities

e.g. Chernoff bound

Talagrand

Talagrand. Massart.



$d \to \infty$

Concentration on spheres.

on gaussian processes.

- Complexity $\sup \|X\varepsilon\|_\infty$

$\sup \langle g, f \rangle$

$f \in F$.

Vershynin

- Random matrix theory Terrence Tao.

- minimax rates minimize the worst case Rigollet

$\inf \sup$

Chapter 4/