High Dimensional Statistics

Yanbo Tang

Imperial College London

March. 2025

Content of this week

- LASSO sparse set recovery (PDW)
- Matrix concentration $\|A\|_{op}$
- Spectral clustering on Random graphs
- Matrix Bernstein $\|\sum A_i\|_{op}$
- Covariance estimation $\frac{\sum_{i=1}^{n} X_i^T X}{n}$

LASSO

Theorem

Consider an S-sparse linear regression model for which the design matrix satisfies Assumptions 2 and 3. Then for any choice of regularization parameter such that

$$\lambda_n \geq \frac{2}{1-\alpha} \left\| X_S^T \Pi_{S^\perp}(\mathbf{X}) \frac{\epsilon}{n} \right\|_\infty, \tag{1}$$

the Lasso program has the following properties:

(a) ***Uniqueness:*** *There is a unique optimal solution* $\hat{\theta}$.

(b) ***No false inclusion:*** *This solution has its support set* $\hat{S}$ *contained within the true support set* S.

(c) ℓ_∞***-bounds:*** *The error* $\hat{\theta} - \theta^*$ *satisfies*

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} X_S^T \frac{\epsilon}{n} \right\|_\infty + \left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right\|_\infty \lambda_n, \tag{2}$$

where $\|A\|_\infty = \max_{i=1,\ldots,s} \sum_j |A_{i,j}|$ *is the matrix* ℓ_∞*-norm.*

(d) ***No false exclusion:*** *The Lasso includes all indices* $i \in S$ *such that* $|\theta_i^*| > B(\lambda_n; \mathbf{X})$*, and hence is variable selection consistent if* $\min_{i \in S} |\theta_i^*| > B(\lambda_n; \mathbf{X})$.

Corollary

For a S-sparse linear model based on a noise vector ϵ with zero-mean i.i.d. σ-sub-Gaussian entries, and a deterministic design matrix X that satisfies Assumptions 2 and 3, as well as the C-column normalization condition $\max_{j=1,\ldots,d} \|\mathsf{X}_j\|_2/\sqrt{n} \leq C$. Suppose that we solve the Lasso program with regularization parameter

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2\log(d-k)}{n}} + \delta\right\}$$

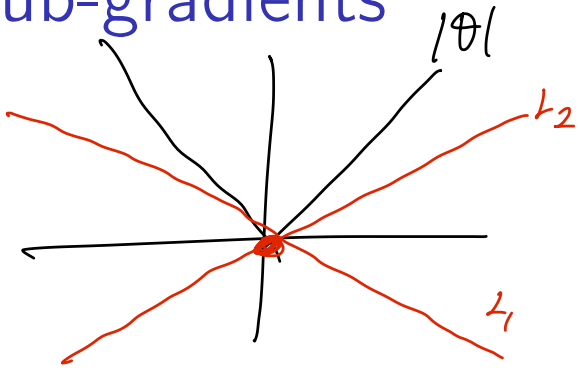
for some $\delta > 0$. Then the optimal solution $\widehat{\theta}$ is unique with its support contained within S, and satisfies the ℓ_∞-error bound

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{2\log s}{n}} + \delta\right) + \left\|\left(\frac{\mathsf{X}_S^T\mathsf{X}_S}{n}\right)^{-1}\right\|_\infty \lambda_n, \qquad (3)$$

all with probability at least $1 - 4e^{-n\delta^2/2}$.

Proof

Sub-gradients



Convex function $f: \mathbb{R}^d \mapsto \mathbb{R}$

we say $z \in \mathbb{R}^d$ is a sub-gradient of f at θ

$$f(\theta+\Delta) \geq f(\theta) + \langle z, \Delta \rangle \quad \forall \Delta \in \mathbb{R}^d$$

then $z \in \partial f(\theta)$

We say a pair $(\hat{\theta}, \hat{z})$ is primal-Dual optimal

if 1) $\hat{\theta}$ is a minimizer

2) $\hat{z} \in \partial \|\hat{\theta}\|_1$

$z_i \in [-1, 1]$

and

$$\frac{1}{n} X^T (X\hat{\theta} - Y) + \lambda \hat{z} = 0$$

Primal Dual Witness

Definition

Primal–dual witness (PDW) construction:

1. Set $\widehat{\theta}_{S^c} = 0$.
2. Determine $(\widehat{\theta}_S, \widehat{z}_S) \in \mathbb{R}^s \times \mathbb{R}^s$ by solving the *oracle subproblem*

$$\widehat{\theta}_S \in \arg\min_{\theta_S \in \mathbb{R}^s} \left\{ \underbrace{\frac{1}{2n}\|y - X_S\theta_S\|_2^2 + \lambda_n\|\theta_S\|_1}_{=:f(\theta_S)} \right\}, \tag{4}$$

   and then choosing $\widehat{z}_S \in \partial\|\widehat{\theta}_S\|_1$ such that $\nabla f(\theta_S)\big|_{\theta_S=\widehat{\theta}_S} + \lambda_n \widehat{z}_S = 0$.
3. Solve for $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ via the zero-subgradient equation, and check whether or not the *strict dual feasibility* condition $\|\widehat{z}_{S^c}\|_\infty < 1$ holds.

Witness for LASSO $\lambda_{\min}\left(\frac{X_S^T X_S}{n}\right) > C_{\min} > 0$

Lemma

If the lower eigenvalue condition holds, then success of the PDW construction implies that the vector $(\widehat{\theta}_S, 0) \in \mathbb{R}^d$ is the unique optimal solution of the Lasso.

sketch if $\tilde{\theta}$ $\|\tilde{\theta}\|_1 \le \langle \hat{z}, \tilde{\theta} \rangle \le \|\hat{z}\|_\infty \|\tilde{\theta}\|_1 = \|\tilde{\theta}\|_1$

$\hat{z}_i = -1, 1$ if $i \in S$

$\langle \hat{z}, \tilde{\theta} \rangle = \|\tilde{\theta}\|_1$

$\langle \hat{z}_S, \tilde{\theta}_S \rangle = \|\tilde{\theta}\|_1$

$\Rightarrow \tilde{\theta}_{S^c} = 0$

Proof

Proof of main theorem

$$\hat{z}_{S^c} = \frac{-1}{\lambda_n n} X_{S^c}^T X_S (\hat{\theta}_S - \theta_S^*) + X_{S^c}^T \left(\frac{\varepsilon}{\lambda_n n}\right) \quad (1)$$

$$\hat{\theta}_S - \theta_S^* = (X_S^T X_S)^{-1} X_S^T \varepsilon - \lambda_n n (X_S^T X_S)^{-1} \hat{z}_S \quad (2)$$

$$\hat{z}_{S^c} = \underbrace{X_S^T X_S (X_S^T X_S)^{-1} \hat{z}_S}_{u} + \underbrace{X_{S^c}^T \left(I - X_S (X_S^T X_S)^{-1} X_S^T\right) \frac{\varepsilon}{\lambda_n n}}_{V_{S^c}}$$

mutual incoherence gives $\|V_{S^c}\|_\infty \le \alpha$

$$\|\hat{z}_S\|_\infty \le 1$$

$$\|\hat{z}_{S^c}\|_\infty \le \|u\|_\infty + \|V_{S^c}\|_\infty$$

$\|\cdot\|^2$ + penalty

$\lambda_n \|\theta\|_1 \quad \|\theta\|_1 < \delta$

$\lambda_n \|f\|_{\mathcal{H}} \quad \|f\|_{\mathcal{H}} < \delta$

Matrix Concentration

Some review

The Courant Fisher min-max theorem which states that:

$$\lambda_i(A) = \max_{dim(E)=i} \min_{x \in S(E)} x^\top A x$$

where the maximum is taken over all i-dimensional subspaces E of $\mathbb{R}^n$.

Some review

The Courant Fisher min-max theorem which states that:

$$\lambda_i(A) = \max_{dim(E)=i} \min_{x \in S(E)} x^\top A x$$

where the maximum is taken over all i-dimensional subspaces E of $\mathbb{R}^n$.Using this we can characterize the operator norm or the maximum singular value as follows:

$$\|A\|_{op} := \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in S^{n-1}} \|Ax\|_2.$$

Equivalently, the operator norm of A can be computed by maximizing the quadratic form $\langle Ax, y \rangle$ over all unit vectors x, y:

$$\|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} \langle Ax, y \rangle.$$

Controlling the operator norm

Lemma

Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any ε-net $\mathcal{N}$ of the sphere S^{n-1}, we have

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1-\varepsilon} \sup_{x \in \mathcal{N}} \|Ax\|_2.$$

Proof: fix x for which $\|A\| = \|Ax\|_2$

Choose $x_0 \in \mathcal{N}$ where $\|x - x_0\|_2 \leq \varepsilon$

$$\|Ax - Ax_0\|_2 \leq \|A\| \underbrace{\|x - x_0\|_2}_{\leq \varepsilon} \leq \varepsilon \|A\|$$

$$\|Ax_0\|_2 \geq \|Ax\|_2 - \|Ax - Ax_0\|_2 \geq \|A\| - \varepsilon\|A\| = (1-\varepsilon)\|A\|$$

$$\|A\| \leq \frac{\|Ax_0\|}{1-\varepsilon} \leq \sup_{x \in \mathcal{N}} \frac{\|Ax\|}{1-\varepsilon}$$

Proof

The actual lemma

Lemma

Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any ε-net $\mathcal{N}$ of the sphere S^{n-1} and any ε-net $\mathcal{M}$ of the sphere S^{m-1}, we have

$$\sup_{x \in \mathcal{N}} \sup_{y \in \mathcal{M}} y^\top A x \leq \|A\|_{\text{op}} \leq \frac{1}{1 - 2\varepsilon} \sup_{x \in \mathcal{N}} \sup_{y \in \mathcal{M}} y^\top A x.$$

Random sub-Gaussian matrices

Theorem

Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, σ-sub-gaussian random variables. Then, for any $t > 0$ we have

$$\|A\| \leq C\sigma\left(\sqrt{m} + \sqrt{n} + t\right)$$

with probability at least $1 - 2\exp(-t^2)$ for some constant $C > 0$.

$\|D\| \leq \|D\|_F = O(\sqrt{m \cdot n})$

Step 1: Approximation.

Choose $\varepsilon = 1/4$ then have ε-net $\mathcal{N}(X)$

ε-net $\mathcal{M}(Y)$

$|\mathcal{N}| \le 9^n$ $\quad$ $|\mathcal{M}| \le 9^m$ $\quad$ $\left(1+\frac{2}{\varepsilon}\right)^d$

$$\|A\| \le 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle$$

Step 2: Concentration.

$\left(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} x_i x_j\right)$

fix $x \in N$ and $y \in M$

$\langle Ax, y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij} x_i y_j$ is a sum of independent sub-Gaussians.

$\sigma^2 \le C \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma^2 x_i^2 y_j^2 \le C\sigma^2$

$P(\langle Ax, y\rangle \ge u) \le \exp\left(-\frac{u^2}{C\sigma^2}\right)$

Step 3: Union bound.

if $\max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \geq u$

$$P\left(\max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \geq u\right) \leq \sum_{x \in \mathcal{N}, y \in \mathcal{M}} P\left(\langle Ax, y \rangle \geq \epsilon\right)$$

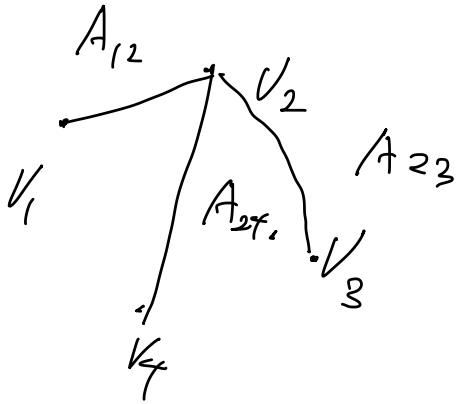
$$\leq 9^{n+m} \cdot 2 \exp\left(-\frac{u^2}{C\sigma^2}\right)$$

$$u = C\sigma\left(\sqrt{n} + \sqrt{m} + \epsilon\right)$$

then solve to get

$$\leq 2 \exp\left(-\epsilon^2\right)$$

Application stochastic block model



Erdős-Rényi Random graph $G(n,p)$

$A_{ij} \sim \text{Bern}(p)$

$G(n,p,q)$ $\quad E[A] = \begin{bmatrix} p & p & q & q \\ p & p & q & q \\ q & q & p & p \\ q & q & p & p \end{bmatrix}$ $p > q$

$n/2$ per group

2 groups

$A = E[A] + R$ $\quad \|E[A]\| \approx n$, $\|R\| \le c\sqrt{n}$

$E[A]$ $\quad \lambda_1 = \frac{(p+q)}{2} \cdot n$

$\lambda_2 = \frac{(p-q)}{2} \cdot n$

$\lambda_3 = 0$

$\vdots$

$u_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{bmatrix}$, $u_2 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{bmatrix}$

$\hat{u}_2$

Illustration

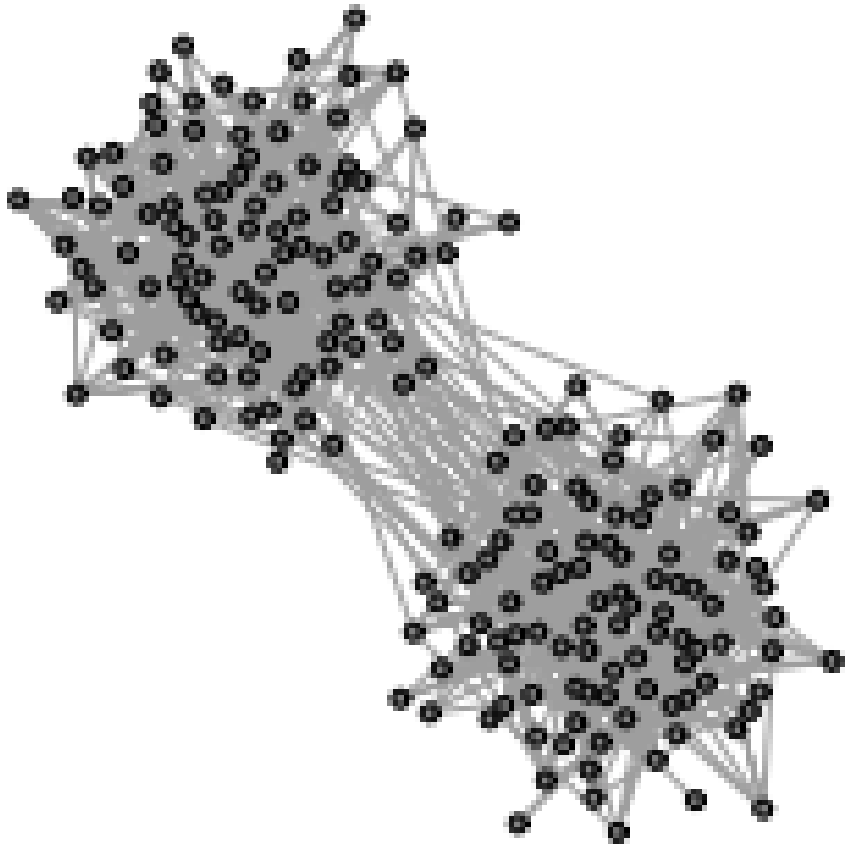


Figure 1: Taken from High Dimensional Probability by Roman Vershynin, with $n = 200$, $p = 1/20$ and $q = 1/200$.

Cont.

Davis · kahan.

Theorem

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

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

The conclusion of the Davis-Kahan theorem implies that the unit eigenvectors $v_i(S)$ and $v_i(T)$ are close to each other up to a sign, namely

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

Cont.

$$\delta := \max\left(\frac{p-q}{2}, q\right)\cdot n =: \mu\cdot n$$

$$\exists \theta \in \{-1,1\} : \quad \|V_2(\Xi(A)) - \theta V_2(A)\|_2 \le \frac{C\sqrt{n}}{\mu\cdot n} = \frac{C}{\mu\sqrt{n}}$$

$$1-\exp(-u)$$

$$\|V_2(\Xi(A)) - \theta V_2(A)\|_2 \le \frac{C}{\mu}$$

$$\Rightarrow \quad \sum_{j=1}^{n} |V_2(\Xi(A))_j - \theta V_2(A)_j|^2 \le \frac{C}{\mu^2}$$

total number of mistake $\le \frac{C}{\mu^2}$

Bernstein inequality for matrices

$X_i \sim N(0, \Sigma)$

Recall that we were able to use the fact that

$$E[\exp(t(X+Y))] = E[\exp(tX)\exp(tY)],$$

and use Chernoff's method if X and Y are real valued random variables. To generalize this approach we need to define what a *matrix exponential* means:

Bernstein inequality for matrices

Recall that we were able to use the fact that

$$E[\exp(t(X+Y))] = E[\exp(tX)\exp(tY)],$$

and use Chernoff's method if X and Y are real valued random variables. To generalize this approach we need to define what a *matrix exponential* means:

Definition

For a function $f : \mathbb{R} \to \mathbb{R}$ and an $n \times n$ symmetric matrix

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^\top,$$

define

$$f(X) := \sum_{i=1}^{n} f(\lambda_i) u_i u_i^\top.$$

Matrix power series

For a convergent power series expansion of f about x_0:

$$f(x) = \sum_{k=1}^{\infty} a_k (x - x_0)^k.$$

It is the case that series of matrix terms converges, and

$$f(X) = \sum_{k=1}^{\infty} a_k (X - x_0 I)^k.$$

As an example, for each $n \times n$ symmetric matrix X we have

$$e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots$$

$A \cdot B \neq B \cdot A$

$\exp(A+B) \overset{?}{=} \exp(A) \cdot \exp(B)$

Generalization of exponential inequality

Theorem

(Golden-Thompson inequality). For any $n \times n$ symmetric matrices A and B, we have

$$tr(e^{A+B}) \leq tr(e^A e^B).$$

Unfortunately, Golden-Thompson inequality does not hold for three or more matrices: in general, the inequality $\mathrm{tr}(e^{A+B+C}) \leq \mathrm{tr}(e^A e^B e^C)$ may fail.

Theorem

(Lieb's inequality). Let H be an $n \times n$ symmetric matrix. Define the function on matrices

$$f(X) := tr \exp(H + \log X).$$

Then f is concave on the space on positive definite $n \times n$ symmetric matrices.

Cont.

Lemma

(Lieb's inequality for random matrices). Let H be a fixed $n \times n$ symmetric matrix and Z be a random $n \times n$ symmetric matrix. Then

$$\mathbb{E}tr\exp(H+Z) \leq tr\exp(H+\log \mathbb{E}Z).$$

This follows by using Jensen's inequality.

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.

Moment generating function of random matrices

Lemma

(Moment generating function). Let X be an $n \times n$ symmetric mean zero random matrix such that $\|X\| \leq K$ almost surely. Then

$$\mathbb{E}\exp(\lambda X) \preceq \exp\left(g(\lambda)\mathbb{E}X^2\right) \quad \textit{where} \quad g(\lambda) = \frac{\lambda^2/2}{1-|\lambda|K/3},$$

provided that $|\lambda| < 3/K$.

Proof Reduction to MGF

Matrix Bernsetein - Step 1: Reduction to MGF

$$S := \sum_{i=1}^{N} X_i$$

$$\|S\| = \max_{i=1,\dots,n} |\lambda_i(S)| = \max\left(\lambda_{\max}(S), -\lambda_{\min}(S)\right)$$

do $\max \lambda$

$$\mathbb{P}\left(\lambda_{\max}(S) \geq t\right) = \mathbb{P}\left(e^{\lambda \cdot \lambda_{\max}(S)} \geq e^{\lambda t}\right)$$

$$\leq e^{-\lambda t}\, \underbrace{\mathbb{E}\left[e^{\lambda \cdot \lambda_{\max}(S)}\right]}_{\mathbb{E}}$$

$$\mathbb{E} = \mathbb{E}\left[\lambda_{\max}(e^{\lambda S})\right] \leq \mathbb{E}\left[\operatorname{tr} \exp(\lambda S)\right]$$

Step 2: Application of Lieb's inequality

$\mathbb{E}[XA]$

$\mathbb{E}\,\mathbb{E}[XA\,|\,B]]$

$$\mathbb{E} \le \mathbb{E}\left[\operatorname{tr}\exp\Big(\sum_{i=1}^{N-1}\lambda X_i + \lambda X_N\Big)\right]$$

Condition on $(X_i)_{i=1}^{N-1}$ apply Lemma with $H := \sum_{i=1}^{N-1}\lambda X_i$ and $A := \lambda X_N$

$$\le \mathbb{E}\left[\operatorname{tr}\exp\Big(\sum_{i=1}^{N-1}\lambda X_i + \log \mathbb{E}[e^{\lambda X_N}]\Big)\right]$$

$$\le \operatorname{tr}\exp\left[\sum_{i=1}^{N}\log\big(\mathbb{E}[e^{\lambda X_i}]\big)\right]$$

$$\le \operatorname{tr}\exp\big(g(\lambda)\,Z\big) \qquad Z := \sum_{i=1}^{N}\mathbb{E}[X_i^2]$$

$$\le n\cdot\lambda_{\max}\big(\exp(g(\lambda)Z)\big)$$

$$= n\cdot\exp\big(g(\lambda)\,\lambda_{\max}(Z)\big)$$

$$= n\cdot\exp\big(g(\lambda)\,\sigma^2\big)$$

then plug back into (*) and minimize get the result.

Step 3: Using the MGF bound

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.

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

General covariance estimation

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

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T.$$

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.

General covariance estimation

Theorem

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

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

Then, for every positive integer m, we have

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

Proof

Proof cont.