High Dimensional Statistics

Yanbo Tang

Imperial College London

Feb. 2025

Content of this week

- Concentration for functions of random variables
- Performance bounds for linear regression
- LASSO

Sub-Gaussian random vectors

Definition

A random variable X is called sub-Gaussian with proxy variance σ^2 if there exists a $\sigma^2 > 0$:

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

for all $\lambda \in \mathbb{R}$. A random vector $X \in \mathbb{R}^d$ is sub-Gaussian with proxy variance σ^2 if $u^\top X$ is sub-Gaussian with proxy variance σ^2 for all $u \in S^{d-1}$.

Concentration for functionals

$f : \mathbb{R}^n \mapsto \mathbb{R}$

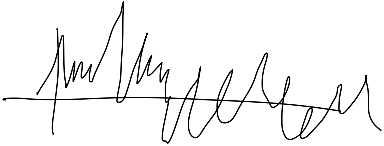
$X - \mathbb{E}(X)$

$$\frac{\sum_{i=1}^{n} X_i - \mathbb{E}(X_i)}{n}$$

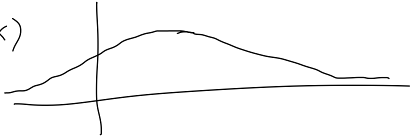
$$f(X_1, X_2, X_3, \dots, X_n) - \mathbb{E}[f(X_1, X_2, \dots, X_n)]$$

not good

$f(x_1 \quad)$



$f(x)$



Bounded differences

Suppose we have a function $f : \mathbb{R}^d \to \mathbb{R}$ for all $x_1, x_2, \ldots, x_d, x_1', x_2', \ldots, x_d' \in \mathbb{R}$

$$|f(x_1', x_2, \ldots, x_j, \ldots, x_d) - f(x_1, x_2, \ldots, x_j, \ldots, x_d)| \leq L_1 < \infty$$

$$\vdots$$

$$|f(x_1, x_2, \ldots, x_j', \ldots, x_d) - f(x_1, x_2, \ldots, x_j, \ldots, x_d)| \leq L_j$$

$$\vdots$$

$$|f(x_1, x_2, \ldots, x_j, \ldots, x_d') - f(x_1, x_2, \ldots, x_j, \ldots, x_d)| \leq L_d.$$

Bounded differences/McDiaramids

Proposition

(Bounded differences inequality/McDiaramids) *Suppose that f satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$ and that the random vector*

$$X = (X_1, X_2, \ldots, X_n)$$

has independent components. Then

$$\mathbb{P}\Big[|f(X) - \mathbb{E}[f(X)]| \geq t\Big] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}} \quad \text{for all } t \geq 0.$$

U-Statistics

"Canonical" $\frac{\sum_{i=1}^{n} X_i}{n} \approx E(X)$

$$E[|X_1 - X_2|] \approx \sum_{i<j} |X_i - X_j| \Big/ \binom{n}{2}$$

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} g(X_i, X_k)$$

Imagine $\|g\|_\infty < b$ (or $X_i \in [-b, b]$ a.s)

$$\forall j \in [1, \ldots, n] \; |U(X_1, \ldots, X_n) - U(X_1, \ldots, X_j', \ldots X_n)| \leq \frac{1}{\binom{n}{2}} \sum_{i \neq j} |g(X_i, X_j) - g(X_i, X_j')|$$

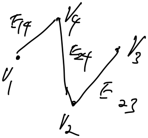
$$\leq \frac{(n-1) \cdot 2b}{\binom{n}{2}} = \frac{4b}{n}$$

$$\therefore L_j = \frac{4b}{n} \quad \forall j \in \{1, 2, \ldots, n\}$$

$$P(|U - E(U)| \geq t) \leq 2 \exp\left(\frac{-nt^2}{8b^2}\right)$$

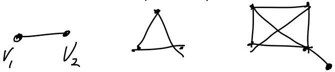
Asymptotic statistics
Van der Vaart

Clique sizes for random graphs



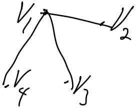
$G = (V, E)$

A "clique" is a set of fully connected vertices.



Maximal clique size $\in \{1, \dots, d\}$

Erdős-Rényi Random graph



$X_{ij} = \text{Bernoulli}(p)$

$$C(G) = C(X_{11}, X_{12}, \dots, X_{ij}, \dots X_{dd})$$

$$\left| C(X_{11}, \dots, X_{ij}', \dots, X_{dd}) - C(X_{11}, \dots, X_{dd}) \right| \le 1$$

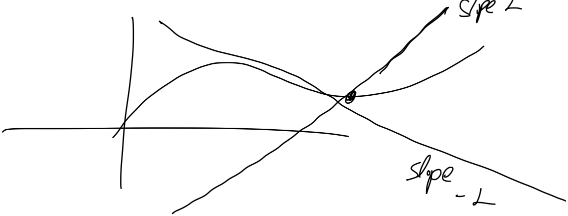
$$P\left[\frac{1}{n} \left| C(G) - E[C(G)] \right| \ge t \right] \le 2e^{-nt^2}$$

$$n = \binom{d}{2}$$

Lipschitz property

We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm if:

$$|f(x) - f(y)| \leq L\|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^n.$$



Gaussian concetration

n

Theorem

Let $(X_1, \dots X_n)$ be a vector of IID standard Gaussian random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz function with respect to the Euclidean norm. Then $f(X) - E[f(X)]$ is sub-Gaussian with parameter at most L and

$$\mathbb{P}[|f(X) - E[f(X)]| \geq t] \leq 2\exp\left(\frac{-t^2}{2L^2}\right) \text{ for all } t \geq 0.$$

Singular value decomposition

As a reminder for a real matrix $A \in \mathbb{R}^{n \times d}$, the singular value decomposition is:

$$A = \sum_{i=1}^{r} s_i(A) u_i v_i^\top, \text{ where } r = rank(A).$$

The non negative numbers $s_i(A)$ are called the singular values of A, the vectors $u_i \in \mathbb{R}^n$ are the left singular vectors of A, and $v_i \in \mathbb{R}^d$ are the right singular vectors of A.

Singular values are eigenvalues $\sqrt{\lambda(A^TA)}$, $\sqrt{\lambda(AA^T)}$

Wely's theorem

$s_1 \geq s_2 \geq \cdots \geq s_\ell$

Theorem

Given two matrices X and Y in $\mathbb{R}^{n\times d}$, we have

$$\max_{i=1,\ldots,d} |s_k(X) - s_k(Y)| \leq s_1(X-Y) \leq \|X-Y\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm of a $\mathbb{R}^{n\times d}$ matrix:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} a_{ij}^2} = \sqrt{Trace(A^\top A)} = \sqrt{\sum_{i=1}^{\min(n,d)} s_i(A)^2}.$$

Singular values of Gaussian random matrices

$X \in \mathbb{R}^{n \times d}$ where $x_{ij} \sim N(0,1)$

$$\max_{i=1,\dots,d} |S_k(X) - S(Y)| \le \|X - Y\|_F = 1 \cdot \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - y_{ij})^2}$$

$$\forall_{k=1,\dots,d} \quad P(|S_k(X) - E[S_k(X)]| \ge \delta) \le \exp\left(\frac{-\delta^2}{2}\right)$$

Terrence Tao notes on Random matrix theory

What about non-Gaussians?

We need some other assumptions:

Definition

A distribution supported in $\mathbb{R}^n$ with density $p(x) = \exp(-\psi(x))$ is said to be γ strongly log concave if there exists a $\gamma > 0$ such that:

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|x-y\|_2^2,$$

for all $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^n$.

Cont.

Theorem

Let $\mathbb{P}$ be any strongly log-concave distribution with parameter $\gamma > 0$. Then for any L-Lipschitz function with respect to the Euclidean norm:

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq t\right] \leq 2\exp\left(\frac{-\gamma t^2}{4L^2}\right).$$

Concentration inequalities
by MASSART

Linear Regression

Prediction

Low MSE

for new obs.

↑

this week

Attribution

sparsity pattern

recovery

Inference

Confidence interval

for θ,

Introduction

Most regression models can be written in the form of:

$$Y_i = f(x_i) + \epsilon_i, i = 1, \ldots, n,$$

where $f(\cdot)$ is some functional relationship and ϵ_i are some centred error terms.

Here, we assume the data generating model is:

$$Y_i = x_i^\top \theta^\star + \epsilon_i, i = 1, \ldots, n,$$

$\varepsilon \sim subG(\sigma^2)$

MSE

We first consider the performance of our estimated models in terms of the *Mean Square Error* (MSE), for a general regression problem this is:

$$MSE(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_n(x_i) - f(x_i)\right)^2,$$

for us this will simply to

$$MSE(X\hat{\theta}) = \frac{1}{n}\|X(\hat{\theta}_n - \theta^\star)\|_2^2,$$

where $\hat{\theta}_n$ is some estimated value for the regression parameter and $\theta^\star$ is the true data-generating value of θ.

Unconstrained estimation

Proposition

The least squares estimator $\hat{\theta}^{LS} \in \mathbb{R}^d$ satisfies

$$X^\top X \hat{\theta}^{LS} = X^\top Y.$$

Moreover, $\hat{\theta}^{LS}$ can be chosen to be

$$\hat{\theta}^{LS} = (X^\top X)^\dagger X^\top Y,$$

where $(X^\top X)^\dagger$ denotes the Moore-Penrose pseudoinverse of $X^\top X$.

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

$Ax = b$

$A^\dagger$ is such that

$\forall x \in \mathbb{R}^n \quad \|Ax - b\|_2 \geq \|Az - b\|_2$

$z = A^\dagger b$

Proof: $\theta \mapsto \|Y - X\theta\|_2^2$ is convex $\therefore$ all minima

satisfies $\nabla_\theta \|Y - X\theta\|_2^2 = 0$

$$\nabla_\theta \|Y - X\theta\|_2^2 = -2(X^T Y - \theta^T X^T X)^T = 0$$

$$\Rightarrow \quad X^T X \theta = X^T Y$$

Cont.

$\lesssim$ $\quad$ $5\log(n)\cdot p \lesssim \log(n)\cdot p$

Theorem

Assume that the linear model holds where $\varepsilon \sim subG_n(\sigma^2)$. Then the least squares estimator $\hat{\theta}^{LS}$ satisfies

$$\mathbb{E}[MSE(X\hat{\theta}^{LS})] = \frac{1}{n}\mathbb{E}|X\hat{\theta}^{LS} - X\theta^*|_2^2 \lesssim \sigma^2 \frac{r}{n},$$

where $r = rank(X^\top X)$. Moreover, for any $\delta > 0$, with probability at least $1 - \delta$,

$$MSE(X\hat{\theta}^{LS}) \lesssim \sigma^2 \frac{r + \log(1/\delta)}{n}.$$

Small fact that we'll need:

Moments of sub-Gaussian random variables. Let X be any random variable such that

$$\mathbb{P}[|X| > t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

then for any positive integers $k \geq 1$,

$$E[|X|^k] \leq (2\sigma^2)^{k/2} k\Gamma(k/2).$$

Proof

$$\|Y - X\hat{\theta}\|_2^2 \le \|Y - X\theta^*\|_2^2 = \|\varepsilon\|_2^2$$

$$\|\varepsilon\|_2^2 \ge \|Y - X\hat{\theta}\|_2^2 = |X\hat{\theta} - X\theta^*|_2^2 - 2\varepsilon^T X(\hat{\theta} - \theta^*) + |\varepsilon|_2^2$$

$$= |X\theta^* + \varepsilon - X\hat{\theta}|$$

$$\Rightarrow \quad 0 \ge |X\hat{\theta} - X\theta^*|_2^2 - \varepsilon^T X(\hat{\theta} - \theta^*)$$

$$|X\hat{\theta} - X\theta^*|_2^2 \le 2\varepsilon^T X(\hat{\theta} - \theta^*) = 2|X\hat{\theta} - X\theta^*|_2 \cdot \left(\frac{\varepsilon^T X(\hat{\theta} - \theta^*)}{|X(\hat{\theta} - \theta^*)|_2}\right)$$

Let $\Phi = [\phi_1, \dots, \phi_r] \in \mathbb{R}^{r \times n}$ be an orthonormal basis for X.

$\exists v: \; X(\hat{\theta} - \theta) = \Phi v$

$$\frac{\varepsilon^T X(\hat{\theta} - \theta^*)}{|X(\hat{\theta} - \theta^*)|_2} = \frac{\varepsilon^T \Phi v}{|\Phi v|_2} = \frac{\varepsilon^T \Phi v}{|v|_2} = \tilde{\varepsilon}^T \frac{v}{|v|_2} \le \sup_{u \in B_2^r} \tilde{\varepsilon}^T u$$

$\tilde{\varepsilon} = \Phi\varepsilon$ is Sub-Gaussian (σ^2) (Exercise)

$$E\left[|X\hat{\theta} - X\theta^*|_2^2\right] \le E\left[4 \sup_{u \in B_2} (\tilde{\varepsilon}^T u)^2\right] = 4 \sum_{i=1}^{n} E[\tilde{\varepsilon}_i^2] \le 16\sigma^2 r$$

Proof Cont.

$\sup (\hat{\varepsilon}^{T} u)^2 \leq 8 \log(d) \sigma^2 \cdot r + 8\sigma^2 \log(1/\delta)$

solve

L^1 constraint

$$\hat{\theta}_K \in \arg\min_{\theta \in K} |y - X\theta|_2^2$$

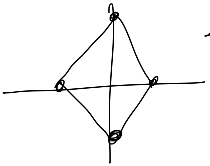
$$|X\hat{\theta}_K - X\theta^*|_2^2 \le 2\varepsilon^T X(\hat{\theta}_K - \theta^*)$$

$$\le 2 \sup_{\theta \in K-K} (\varepsilon^T X\theta)$$

$K - K = \{x - y,\ x, y \in K\}$

if have K symmetric

$$= 4 \sup_{v \in XK} (\varepsilon^T v)$$



L^1 ball has $2 \cdot d$ vertices

$$V = \{e_1, -e_1, e_2, -e_2, \dots\}$$

$$XV = \{X_1, -X_1, X_2, -X_2, \dots\}$$

Sub-Gaussian width of L^1 ball

Theorem

Let P be a polytope with N vertices $v^{(1)}, \dots, v^{(N)} \in \mathbb{R}^d$ and let $X \in \mathbb{R}^d$ be a random vector such that $[v^{(i)}]^\top X, i = 1, \dots, N$, are sub-Gaussian random variables with variance proxy σ^2. Then

$$\mathbb{E}\left[\max_{\theta \in P} \theta^\top X\right] \leq \sigma\sqrt{2\log(N)},$$

and

$$\mathbb{E}\left[\max_{\theta \in P} |\theta^\top X|\right] \leq \sigma\sqrt{2\log(2N)}.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\left(\max_{\theta \in P} \theta^\top X > t\right) \leq N e^{-\frac{t^2}{2\sigma^2}},$$

and

$$\mathbb{P}\left(\max_{\theta \in P} |\theta^\top X| > t\right) \leq 2N e^{-\frac{t^2}{2\sigma^2}}.$$

Performance

Theorem

Let $\mathcal{B}_1$ be the unit ℓ_1 ball of $\mathbb{R}^d$, $d \geq 2$ and assume that $\theta^* \in \mathcal{B}_1$. Moreover, assume the conditions of Theorem 6 and that the columns of $\mathbb{X}$ are normalized such that $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$. Then the constrained least squares estimator $\hat{\theta}^{LS}_{\mathcal{B}_1}$ satisfies

$$\mathbb{E}[MSE(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1})] = \frac{1}{n}\mathbb{E}|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma\sqrt{\frac{\log d}{n}}.$$

Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$MSE(\mathbb{X}\hat{\theta}^{LS}_{B_1}) \lesssim \sigma\sqrt{\frac{\log(d/\delta)}{n}}.$$

Proof

$$\|X\hat{\theta} - X\theta^*\|_2^2 \le 4 \sup_{v \in XK} (\varepsilon^T v)$$

suppose $\varepsilon \sim \text{subG}(\sigma^2)$ $\forall x_j$ $\|x_j\|_2^2 \le n$ $\varepsilon^T x_j \sim \text{subG}(n\sigma^2)$

since $\varepsilon^T x_j = \|x_j\|_2 \frac{\varepsilon^T x_j}{\|x_j\|_2}$

apply our previous bounds on polytope

$$P(|\text{MSE}(\hat{\theta}_{l_1})| > t) \le P\left(\sup_{v \in XK}(\varepsilon^T v) > \frac{nt}{4}\right)$$

$$\le 2d \exp(-nt^2/32\sigma^2)$$

$$2d e^{-nt^2/32\sigma^2} \le \delta \iff$$

$$t^2 \ge 32\frac{\log(2d)}{n} + 32\sigma^2\frac{\log(1/\delta)}{n}$$

L^0 performance

Denote by $B_0(k)$ the ℓ_0 "ball" of $\mathbb{R}^d$, i.e., the set of k-sparse vectors, defined by

$$B_0(k) = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}.$$

Our goal is to control the MSE of $\hat{\theta}_K^{IS}$ when $K = B_0(k)$. Note that computing $\hat{\theta}_{B_0(k)}^{IS}$ defined as:

$$\hat{\theta}_{B_0(k)}^{\text{LS}} \in \underset{\theta \in B_0(k)}{\text{argmin}} \, |Y - \mathbb{X}\theta|_2^2,$$

but this would require computing $\binom{d}{k}$ least squares estimators since this loss is no longer smooth due to the constraint

Best subset selection

Theorem

Fix a positive integer $k \leq d/2$. Let $K = B_0(k)$ be set of k-sparse vectors of $\mathbb{R}^d$ and assume that $\theta^* \in B_0(k)$. Moreover, assume the conditions of Theorem 6. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$MSE(\mathbb{X}\hat{\theta}^{IS}_{B_0(k)}) \lesssim \frac{k\sigma^2}{n} \log\left(\frac{ed}{2k}\right) + \log(6)\frac{\sigma^2 k}{n} + \frac{\sigma^2}{n}\log(1/\delta).$$

Matching rates for LASSO?

Assumption INC(k) We say that the design matrix $\mathbb{X}$ has incoherence k for some integer $k > 0$ if

$$\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right|_\infty \leq \frac{1}{32k}$$

where the $|A|_\infty$ denotes the largest element of A in absolute value. Equivalently,

1. For all $j = 1, \ldots, d$,
$$\left|\frac{\|\mathbb{X}_j\|_2^2}{n} - 1\right| \leq \frac{1}{32k}.$$
2. For all $1 \leq i, j \leq d, i \neq j$, we have
$$\frac{|\mathbb{X}_i^T\mathbb{X}_j|}{n} \leq \frac{1}{32k}.$$

Technical Lemma

Lemma

Fix a positive integer $k \le d$ and assume that $\mathbb{X}$ satisfies assumption INC(k). Then, for any $S \in \{1, \dots, d\}$ such that $|S| \le k$ and any $\theta \in \mathbb{R}^d$ that satisfies the cone condition

$$|\theta_{S^c}|_1 \le 3|\theta_S|_1,$$

it holds

$$|\theta|_2^2 \le 2\frac{|\mathbb{X}\theta|_2^2}{n}.$$

We will interpret the cone condition more carefully next week when we consider sparse recovery.

LASSO

Theorem

Fix $n \geq 2$. Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume that $\|\theta^*\|_0 \leq k$ and that X satisfies assumption INC(k). Then the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter defined by

$$2\tau = 8\sigma\sqrt{\frac{\log(2d)}{n}} + 8\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

satisfies

$$MSE(\mathrm{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathrm{X}\hat{\theta}^{\mathcal{L}} - \mathrm{X}\theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}.$$

and

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}.$$

with probability at least $1 - \delta$.

Proof

Proof Cont.