High Dimensional Statistics

Zhenbang Wang

Summer 2019 - Fall 2019

This notes is based on the course note of Advanced Statistical Theory I/II at CMU : http://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/schedule.html

Contents

1 Intro to High-dimensional statistical models

1.1 Recap of Parametric Statistical Models

Definition 1.1. *$P = \{P_\theta : \theta \in \Theta\}$ where $\Theta \in \mathbb{R}^d$ and P_θ is a probability distribution on $\mathbb{R}^n$.*

Example 1.1. *Normal case : $\Theta = \{(\mu, \sum) : \mu \in \mathbb{R}^k, \sum \in S_+^k\}$ where S_+^k is the cone of positive definite $k \times k$ matrices. Then $P_\theta \sim \mathcal{N}(\mu, \sum)$ and $dim(\Theta) = k + \frac{k(k+1)}{2} = \frac{k^2}{2} + \frac{3}{2}k$*

Example 1.2. *Linear Regression : $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $\beta \in \mathbb{R}^{d \times 1}$ and $\sigma > 0$.*

Model: $Y = X\beta + \epsilon$ where $\epsilon = (\epsilon_1, ..., \epsilon_n)$ i.i.d. from $\mathcal{N}(0, \sigma^2)$. Observe $X = (x_1, ..., x_n)$ i.i.d. from P_{θ_0}. Goal : Draw inference on θ_0.

Important Assumption : P, θ_0 are fixed as $n \to \infty$. In high dimensional statistics, we assume $d \to \infty$ as $n \to \infty$. In non-parametric statistics, we assume P grows as $n \to \infty$.

1.2 High-dimensional statistical models

Definition 1.2. *A high-dimensional parametric statistical model is a sequence of parametric statistical models $\{P_n\}_{n=1}^{\infty}$ where for each n, the sample space has size S_n and the parameter space has dimension d_n, where S_n, d_n are allowed to grow with n.*

Example 1.3.

1.3 Different Types of Parametric Models

1. Fixed d models
2. d_n is allowed to change but $d_n \in o(n)$
3. $d_n >> n$; Not generally possible without additional structural assumptions (sparsity, data near a low - dimensional manifold, etc.)

2 Examples of high dimensional statistical models

2.1 Covariance Estimation

In the problem setting, we obtain vector samples $X_1, ..., X_n$ i.i.d. $(0, \Sigma)$ in $\mathbb{R}^d$ where Σ is a $d \times d$ matrix. We want to estimate Σ using the empirical covariance matrix, given by $\hat{\Sigma_n} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$. Note that the empirical covariance matrix is an unbiased estimator of the covariance matrix.

We are interested in finding $||\hat{\Sigma}_n - \Sigma||_\infty$, to quantify the goodness of the estimator. If this is fairly small, we could possibly say we have a good estimator. But we can't be sure if the estimate is positive definite or not. How do we measure this?

There are two cases that we need to consider. Case 1, where d is fixed and Case 2 where the dimension of the problem d grows with n.

2.1.1 Fixed d

For a given pair (i, j) in $< 1, ..., d >$, let $\widehat{\Sigma}_{n(i,j)} = \frac{1}{n}\sum_{k=1}^{n} Z_k^{(i,j)}$ where $Z_k^{(i,j)} = X_{k,i}X_{k,j}$. This implies that every entry is an average of product of two things. In particular, $Z_1^{(i,j)}, ..., Z_n^{(i,j)}$ are i.i.d. with $\mathbf{E}[\widehat{\Sigma}_{n(i,j)}] \to \Sigma_{(i,j)}$. By WLLN, $\widehat{\Sigma}_{n(i,j)} \xrightarrow{P} \Sigma_{(i,j)} \; \forall \; (i, j)$. Following this, we see that

$$||\hat{\Sigma}_n - \Sigma||_\infty \leq \sum_{i,j} |\widehat{\Sigma}_{n(i,j)} - \Sigma_{(i,j)}| \tag{1}$$

Since $|\widehat{\Sigma}_{n(i,j)} - \Sigma_{(i,j)}| \xrightarrow{P} 0 \; \forall \; (i, j)$, each term can be expressed as $o_p(1)$.

Notes: We defined $o(n)$. In particular, if $x_n = o(1)$, this is equivalent to saying that, $x_n \to 0$ as $n \to \infty$. Here, x_n represents a deterministic sequence. What if we had random sequences?

Definition 2.1. *If $\{X_n\}_{n=1,2,...}$ is a sequence of random vectors and $\{y_n\}_{n=1,2,...}$ is a sequence of positive numbers, then $X_n = o_p(1) \iff X_n \xrightarrow{P} 0$.*

This tells us that (1) can be expressed as

$$||\hat{\Sigma}_n - \Sigma||_\infty \leq \sum_{i,j} o_p(1) = \frac{d(d+1)}{2} o_p(1)$$

If d is fixed as n goes to infinity, $||\hat{\Sigma}_n - \Sigma||_\infty \leq o_p(1)$ since the rest can be written of as a constant.

Furthermore, if $Z_k^{(i,j)}$ has a second moment (that is entries of the random vector have a fourth moment) then, by CLT

$$||\hat{\Sigma}_n - \Sigma||_\infty \leq O_p(\frac{1}{\sqrt{n}})$$

Notes : The Big - O notation may be familiar, and is defined for deterministic sequences, say $\{x_n\}$, $\{y_n\}$. If $x_n = O(y_n), \exists\, c > 0$, and $n_0 = n_0(c)$ such that $\forall n > n_0 : |\frac{x_n}{y_n}| < c$.

Definition 2.2. *For a sequence of random vectors $\{X_n\}$ and a sequence of positive numbers y_n where $X_n = O_p(y_n)$, $\forall \epsilon > 0$, $\exists\, c = c(\epsilon)$ such that $\forall n > n_0 : P(|\frac{x_n}{y_n}| > c) < \epsilon$.*

This implies that the sequence of random vectors is bounded in probability.

Continuing with our covariance estimation problem, let $X_1, ..., X_n \sim (\mu, \sigma^2)$.

Then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu$, $\bar{X}_n = \mu + o_p(1)$. By central limit theorem, $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0,1)$, $\bar{X}_n = \mu + O_p(\frac{1}{\sqrt{n}})$.

We are ignoring σ here because it is a constant. The statement obtained through CLT implies the first statement and also gives us a rate.

2.1.2 d increases with n

If d is a function of n, we need different tools. In HW1, you will show that with probability at least $1 - \frac{1}{n}$,

$$||\hat{\Sigma}_n - \Sigma||_\infty \leq C(\frac{\log d_n + \log n}{n})^{\frac{1}{2}} = O_p(\frac{\log d_n}{n})^{\frac{1}{2}}$$

The increased rate of convergence shows the price you pay for the growing dimension. This may be a misleading result because it seems to imply you can do well for $d >> n$ but you should recall that the metric under study is not a good one to begin with.

2.2 High Dimensional Probability Distributions

Commonly known probability distributions do not look similar in a high dimensional space. However, the good part is that they tend to concentrate.

Definition 2.3. *Euclidean ball :* $B_d(0,r) = \{x \in \mathbb{R}^d : ||x||_2 \leq r\}$
Cube : $C_d(0,r) = \{x \in \mathbb{R}^d : ||x||_\infty \leq r\}$

In two dimensions the Euclidean unit ball, $B_2(0,1)$ is a circle with radius 1 and the unit cube $C_2(0,1)$ is a square symmetric about the origin with each side = 2.

Let's look at the volume of the sets considered above. Volume of the Euclidean norm ball $B_d(0,r) = r^d v_d$, where

$$v_d = \text{Vol}(B_d(0,1)) = \frac{\pi^{\frac{d}{2}}}{\Gamma(d/2+1)} \sim (\frac{2\pi e}{d})^{\frac{d}{2}}$$

The gamma function is given by $\Gamma(x) = \int_0^{+\infty} \exp(-z) z^{x-1} dz$. Note that the volume of the Euclidean unit ball goes to zero really fast in high dimensions. Although, this doesn't hold for $C_d(0,1)$ which is equal to 2^d even in higher dimensions.

Assume X is uniformly distributed over $B_d(0,1)$, $\mathbf{E}||X|| = \frac{d}{d+1}$. Now, pick $\epsilon \in (0,1)$

$$P(||X|| \geq 1-\epsilon) = \frac{v_d - (1-\epsilon)^d v_d}{v_d} = 1 - (1-\epsilon)^d \geq 1 - \exp(-\epsilon d)$$

The probability that $||X||$ is close to 1 goes to 1 exponentially fast in d. Similarly, for the normal distribution, if $X \sim \mathcal{N}_d(0, I_d)$, then with high probability $||X|| \sim \sqrt{d}$. This implies that if you distribute points according to the normal distribution, the whole space never gets filled in.

Let's go back to the unit cube, $c_d(0,1) = \{x \in \mathbb{R}^d : ||x||_\infty \leq 1\}$. It turns out that

$$\lim_{d\to\infty} \mathbf{P}\left(\frac{\sqrt{d}}{3}(1-\epsilon) \leq ||X|| \leq \frac{\sqrt{d}}{3}(1+\epsilon)\right) \forall \epsilon \in (0,1).$$

The main idea is that if $X_1, ..., X_n$ are independent random variables and $f : \mathbb{R}^n \to \mathbb{R}$ such that it doesn't depend too much on any of its coordinates, then $f(X_1, ..., X_n)$ is very close to $\mathbf{E}[f(X_1, ..., X_n)]$.

3 Sub-Gaussian random variables

3.1 Basic concentration inequalities

Let $X_1, ..., X_n \sim (\mu, \sigma^2)$. By central limit theorem, $\bar{X}_n = \frac{1}{n}\sum_i X_i = \mu + O_p(\frac{1}{\sqrt{n}})$. Note that this is a purely asymptotic statement and doesn't tell us about the behaviour for intermediate values of n, say n = 30. We would like to know $P(|\bar{X}_n - \mu| \geq t)$ for some $t > 0$.

We know that $\lim_{n\to+\infty} P(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) > t) = P(Z \geq t)$, $(\frac{1}{t} - \frac{1}{t^3})\phi(t) \leq 1 - \phi(t) \leq \frac{1}{t}\phi(t) \leq \frac{1}{2}\exp\{-\frac{t^2}{2}\}$, Following this, we may be tempted to conclude that $P(|\bar{X}_n - \mu| \geq t) \leq \exp\left(-\frac{nt^2}{2\theta^2}\right)$. Although this is good for intuition, this isn't exactly correct. We now look at the finite version of CLT, also known as Berry Essen Bound.

Theorem 3.1. *(Berry Essen Bound) Let $X_1, ..., X_n \sim (\mu, \sigma^2)$, third moments exist then*

$$\sup_{x\in\mathbb{R}} |P(\frac{\sum_i (X_i - \mu)}{\sqrt{n}\sigma} \leq x) - P(Z \leq x)| \leq C\frac{\gamma}{n};\ \gamma = \frac{\mathbf{E}[|X_i - \mu|^3]}{\sigma^3}, C \leq \frac{1}{2}$$

We know that Gaussian random variables concentrate around their mean, i.e. for $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$, it holds $P(|\bar{X}_n - \mu| \geq t) \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right)$ for every $t \geq 0$. Thus, the probability that the sample average $\bar{X}_n$ is far away from the mean μ decays rapidly. We want to replicate this type of behavior for other random variables in a manner that allows us to (1) obtain finite samples guarantees (i.e. for every n), and (2) circumvent the need for too many distributed assumptions on $X_1, ..., X_n$.

Goal : Given some $X \sim P$ with mean μ, we want to derive an upper bound on $P(|X - \mu| \geq t)$ which holds for all $t \geq 0$.

3.1.1 Markov Inequality

We make a first attempt at bounding the above probability in terms of moments of X based on Markov's inequality.

Theorem 3.2. *(Markov's Inequality) Let X be a random variable and $E(X) = \mu$. Then,*

$$P(|X - \mu| \geq t) \leq \min_{q\in\mathbb{N}} \frac{\mathbf{E}[|X - \mu|^q]}{t^q}$$

This procedure often yields an analytically sharp bound. However, it requires us to compute the centered moments of X which is often infeasible or computationally expensive.

3.1.2 Chernoff bound

For a second approach to bounding of the above probability, we draw on the moment generating function of the centered version of X, i.e. $\psi_X(\lambda) = \log(E[e^{\lambda(X-\mu)}])$, which is well-defined for all $\lambda \in (-b, b)$ for some $0 \leq b \leq \infty$. Assuming a $0 \leq \lambda \leq b$, we get with Markov's inequality that

$$\begin{aligned} P(X - \mu \geq t) &= P(e^{X-\mu} \geq e^t) \\ &= P(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \\ &\leq \frac{E[e^{\lambda(X-\mu)}]}{e^{\lambda t}} = \exp(\psi_X(\lambda) - \lambda t) \end{aligned}$$

which results in the following bound.

Theorem 3.3. *(Chernoff Bound) Let X be a random variable and $E(X) = \mu$. Then,*

$$P(X - \mu \geq t) \leq \exp(-\psi_X^*(t))$$

where $\psi_X^*(t) = \sup_{\lambda \in (0,b)}(\lambda t - \psi_X(\lambda))$.

In some sense, deriving a Chernoff bound does not require less knowledge about a distribution than a Markov - based bound since we need the moment generating function of $X - \mu$. In fact, we have to assume the existence of infinity many moments. A main advantage is that these moments do not have to be painstakingly calculated, and in turn, Chernoff bounds are usually the way to go when having enough knowledge about the distribution although they are not as sharp as the Markov-based bounds.

Example 3.1. *(Chernoff bound for Gaussian) Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbf{E}[e^{\lambda X}] = e^{\mu\lambda + \sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. So, we have that*

$$\sup_{\lambda > 0}\left(\lambda t - \log(\mathbf{E}[e^{\lambda(X-\mu)}])\right) = \sup_{\lambda > 0}(\lambda t - \frac{\lambda^2\sigma^2}{2}) = \frac{t^2}{2\sigma^2},$$

which yields the bound

$$P(X - \mu \geq t) \leq e^{-\frac{t^2}{2\sigma^2}} \text{ for all } t > 0.$$

Theorem 3.4. *(Two - sided Chernoff bound) Let X be a random variable and $\mathbf{E}[X] = \mu$. Then,*

$$P(|X - \mu| \geq t) \leq 2\exp(-\psi_X^*(t)),$$

where $\psi_X^*(t) = \sup_{\lambda \in (-b,b)}(\lambda t - \psi_X(\lambda))$

3.2 Sub-Gaussian random variables

In order to be able to derive Chernoff bounds, we need a bound for $\psi_X(\lambda)$ which is not always easily attainable. A sufficient condition in this setting is that the random variable is sub-Gaussian, i.e. its tails decay faster than the tails of some Gaussian. An extensive overview over sub-Gaussian random variables can be found in Metric Characterization of Random Variables and Random Processes.

Definition 3.1. *(Sub - Gaussian) A random variable X is sub-Gaussian with parameter σ if*

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

for all $\lambda \in \mathbb{R}$. In that case, we write $X \in SG(\sigma^2)$.

A first simple observation is given by $X \in \mathrm{SG}(\sigma^2)$ iff $-X \in \mathrm{SG}(\sigma^2)$.

Now, if $X \in \mathrm{SG}(\sigma^2)$, then the mgf of X can be bounded by the Gaussian mgf which yields the same Chernoff bound as in Example 3.1.

Proposition 3.1. *We observe several properties of sub-Gaussian random variables.*
(1) Let $X \in SG(\sigma^2)$, then $\mathbf{V}(X) \leq \sigma^2$ with $\mathbf{V}[X] = \sigma^2$ if X is Gaussian.
(2) If there are $a, b \in \mathbb{R}$, such that $a \leq X - \mu \leq b$ almost everywhere, then $X \in SG(\frac{(b-a)^2}{2})$.
(3) Let $X \in SG(\sigma^2)$ and $Y \in SG(\tau^2)$, then

1. *$\alpha X \in SG(\alpha^2\sigma^2)$ for all $\alpha \in \mathbb{R}$ with $\alpha \neq 0$.*
2. *$X + Y \in SG((\tau + \sigma)^2)$, and*
3. *if $X \perp\!\!\!\perp Y$, $X + Y \in SG(\tau^2 + \sigma^2)$.*

Proof. (1) It holds by assumption that $\mathbf{E}[e^{\lambda(X-\mathbf{E}[X])}] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for all $\lambda \in \mathbb{R}$, and applying the Taylor expansion on both side,

$$1 + \lambda \underbrace{\mathbb{E}[X-\mu]}_{=0} + \lambda^2 \frac{\mathbb{E}\left[(X-\mu)^2\right]}{2} + o\left(\lambda^2\right) \leq 1 + \frac{\lambda^2\sigma^2}{2} + o\left(\lambda^2\right)$$

We divide both sides of this inequality by λ^2(and assume $\lambda \neq 0$), and let $\lambda \to 0$.

(2) Without Loss Of Generality, let $\mu = 0$. We show that $\log(\mathbf{E}[e^{\lambda X}]) \leq \frac{(b-a)^2\lambda^2}{8}$ for all $\lambda \in \mathbb{R}$. First, notice that $\mathbf{V}(X) \leq (\frac{b-a}{2})^2$. For any $\lambda \in \mathbb{R}$, let X_λ be a RV with distribution that has density of the form $x \mapsto e^{\lambda x}e^{-\psi_X(\lambda)}f_X(x)$ if $a \leq x \leq b$. Then, $\mathbf{V}[X_\lambda] = \psi_X^{''}(\lambda) \leq (\frac{b-a}{2})^2$. Since $\psi_\lambda(0) = \psi_\lambda^{'}(0) = 0$, we have with the fundamental theorem of calculus that

$$\psi_X(\lambda) = \int_0^\lambda \psi_X'(u)du = \int_0^\lambda \int_0^\mu \psi_X''(w)dwdu \leq \int_0^\lambda \int_0^\mu \lambda^2 \frac{(b-a)^2}{4} dwdu = \lambda \frac{(b-a)^2}{8}$$

(3) We prove (ii) and (iii) and assume that $\mathbf{E}[X] = \mathbf{E}[Y]$. If $X \perp\!\!\!\perp Y$, the proof is immediate. If not, it holds for every $\lambda \in \mathbb{R}$ that $\mathbf{E}[e^{\lambda(X+Y)}] = \mathbf{E}[e^{\lambda X}e^{\lambda Y}]$, then apply Hölder's inequality and obtain

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X+Y)}\right] &= \mathbb{E}\left[e^{\lambda X}e^{\lambda Y}\right] \leq \left(\mathbb{E}\left[e^{\lambda p X}\right]\right)^{1/p}\left(\mathbb{E}\left[e^{\lambda q Y}\right]\right)^{1/q} \leq \exp\left(\frac{\lambda^2 p^2 \sigma^2}{2}\frac{1}{p} + \frac{\lambda^2 q^2 \tau^2}{2}\frac{1}{q}\right) \\ &= \exp\left(\frac{\lambda^2}{2}\left(p\sigma^2 + q\tau^2\right)\right) = \exp\left(\frac{\lambda^2}{2}(\sigma+\tau)^2\right)\end{aligned}$$

where we set $p = \tau/\sigma + 1$ in the last step. □

3.2.1 Hoeffding inequality

Theorem 3.5. *(Hoeffding inequality) Let $X_1, ..., X_n$ be independent random variables such* *that $X_i \in SG$ for all i. Then,*

$$P\left(\left|\sum_{i=1}^{n} \frac{X_i - \mathbf{E}[X_i]}{n}\right| \geq t\right) \leq 2\exp\left(\frac{-n^2t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right)$$

Example 3.2. *(Hoeffding for Bernoulli RV) Let $X_1, ..., X_n$ be independent RV with $X_i \sim Bernoulli(p_i)$ for some $p_i \in (0,1)$. Then, $X_i \in SG(1/4)$ and thus,*

$$P(|\bar{X}_n - \bar{p}_n| \geq t) \leq 2\exp(-2nt^2)$$

Thus, we have that

$$P\left(|\bar{X}_n - \bar{p}_n| \leq \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}\right) \geq 1-\delta.$$

3.2.2 Comparing Hoeffding and Chernoff Bounds

3.2.3 Equivalent Definitions of Sub-Gaussian Random Variables

Sub - Gaussianity can equivalently be characterized using Orlicz norms, as will be explored in the second assignment. It turns out that Sub - Gaussian random variables are also uniquely characterized by their moments.

Proposition 3.2. *Let $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ be the Gamma function. If $X \in SG(\sigma^2)$, then $\mathbb{E}\left[|X|^p\right] \leq p2^{p/2}\sigma^p\Gamma(p/2), \quad \forall p > 0$, In particular, there exists $C > 0$ not depending on p such that $\left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \leq C\sigma\sqrt{p}$.*

Proof. We have that

$$\begin{aligned}\mathbf{E}[|X|^p] &= \int_0^{+\infty} P(|X|^p \geq u)du = \int_0^{+\infty} P(|X| \geq u^{\frac{1}{p}})du \leq 2\int_0^\infty \exp\left\{-\frac{u^2}{2\sigma^2}\right\}du \\ &\leq (2\sigma^2)^{\frac{p}{2}}p\int_0^\infty \exp\left\{\frac{u^{2/p}}{2\sigma^2}\right\}\left(\frac{u^{2/p}}{2\sigma^2}\right)^{\frac{p}{2}-1} d\frac{u^{2/p}}{2\sigma^2} = (2\sigma^2)^{\frac{p}{2}}p\Gamma(\frac{p}{2})\end{aligned}$$

□

4 Sub-Exponential random Variables

In this section, we consider a broader class of distributions than the Sub - Gaussian family, call the Sub - Exponential family. We will see that interesting tail bounds can still be derived for random variables belonging to this collection. One motivation for its definition is that Sub - Gaussian random variables are not closed under taking squares, in the sense that $X \in \mathrm{SG}(\sigma^2)$ does not imply X^2 is Sub - Gaussian. For example, the square of a standard Gaussian is a Chi - Squared random variable, which cannot be Sub - Gaussian since its moment generating function is not defined on the entire real line.

Example 4.1. *Let $X \sim$ Laplace(b) for $b > 0$. Then it can be shown that*

$$\mathbb{P}(|X| \geq t) \leq \exp(-tb), \quad \forall t > 0$$

This is a different tail behaviour than what we are used to for Sub - Gaussian random variables, and indeed, we note that $X \notin SG(\sigma^2)$ since its moment generating funciton is only defined on a subset of the real line.

$$\mathbb{E}\left[e^{\lambda X}\right] = \frac{1}{1-\lambda^2} b^2, \quad \forall |\lambda| < \frac{1}{b}$$

Definition 4.1. *(Sub - Exponential Random Variable) We say that a random variable X is Sub - Exponential with parameters $v, \alpha > 0$, and we write $X \in SE(v^2, \alpha)$, if*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}(X))}\right] \leq \exp\left(\frac{\lambda^2 \nu^2}{2}\right), \quad \forall |\lambda| < \frac{1}{\alpha}$$

Observe that the moments of X are still well defined since they can be found as the derivative of the MGF (moment generating function) at zero. An immediate consequence of the definition is that SG$(\sigma^2) \subset$ SE$(\sigma^2, 0)$. Thus, all Sub - Gaussian random variables are also Sub - Exponential.

Example 4.2. *Let $Z \sim \mathcal{N}(0,1)$, and $X = Z^2 \sim \mathcal{X}^2_{(1)}$, $E(X) = 1$. Let $\lambda < \frac{1}{2}$. Then, $E\left[e^{\lambda(X-1)}\right] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq \exp\left\{\frac{\lambda^2}{1-2\lambda}\right\} \leq \exp\left\{\frac{4\lambda^2}{2}\right\}$. Thus, $X \in SE(4,4)$. Note that above follows from the following inequality: $-\log(1-u) - u \leq \frac{u^2}{2(1-u)}, \quad \forall u \in (0,1)$ with $u = 2\lambda$.*

Proposition 4.1. *(1) Squares and products of centered Sub - Gaussian are Sub - Exponential: $X \in SG(\sigma^2) \Rightarrow X^2 \in SE(256\sigma^2, 16\sigma^2)$*

(2) Suppose X is a random variable with $Var[X] = \sigma^2$ and $|X - E[X]| \leq b$ almost everywhere, for some $b > 0$. Then, $X \in SE(2\sigma^2, 2b)$. Unlike Sub - Gaussian bounded random variables, the variance of X appears in the Sub - Exponential parameters.

Proof. Let $|\lambda| < \frac{1}{2b}$. Then,

$$\begin{aligned}
\mathbf{E}[e^{\lambda(X-\mathbf{E}(X))}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{\lambda^n \, \mathbf{E}[(X-\mathbf{E}(X))^k]}{k!} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} \sum_{k=0}^{\infty} (|\lambda| b)^k \leq 1 + \frac{\lambda^2\sigma^2}{2} \frac{1}{1-|\lambda| b} \\
&\leq \exp\left\{\frac{\lambda^2\sigma^2}{1-|\lambda| b}\right\} \leq \exp\{\lambda^2\sigma^2\}
\end{aligned}$$

□

4.1 Tail behavior for Sub-Exponential Random Variables

Theorem 4.1. *(Tail Bounds for Sub - Exponential Random Variables) Let $X \in SE(v^2, \alpha)$, and $t > 0$. Then, $\mathbb{P}\{|X - \mathbb{E}(X)| \geq t\} \leq 2\exp\left\{-\frac{1}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right)\right\}$*

Proof. Assume that $\mu = 0$. Then repeating Chernoff argument, one obtains :

$$\mathbf{P}(X \geq t) \leq \exp\left\{-\lambda t + \frac{\lambda^2 v^2}{2}\right\} = \exp\{g(\lambda, t)\},\ \forall \lambda \in (0, \frac{1}{\alpha})$$

To obtain the tightest bound one needs to find: $g^*(t) = \inf\limits_{\lambda \in (0, \frac{1}{\alpha})} g(\lambda, t) = \inf\limits_{\lambda \in (0, \frac{1}{\alpha})} -\lambda t + \frac{\lambda^2 v^2}{2}$

Consider two cases: 1. $0 < t \leq \frac{v^2}{\alpha}$, $\lambda^* = \frac{t}{v^2}$, $g(t) = -\frac{t^2}{2v^2}$, we obtains the bound describing sub-Gaussian behavior. 2. $t > \frac{v^2}{\alpha}$, $\lambda^* = \frac{1}{\alpha}$, $g(t) = -\frac{t}{\alpha} + \frac{v^2}{2\alpha^2} \leq -\frac{t}{2\alpha}$ □

Recall that sufficient conditions for a random variable to be a Sub-Gaussian include:

- Boundedness of a random variable.
- Condition on the moments $(E|X|^k)^{1/k}$

One would like to obtain a similar condition allowing unbounded random variables to behave sub-exponentially. One such condition is called Bernstein condition.

Definition 4.2. (Bernstein condition) *Let X be a random variable with mean μ and variance σ^2. Assume that $\exists b > 0 : \mathbf{E}\,|X - \mu|^k \leq \frac{1}{2}k!\sigma^2 b^{k-2}, \quad k = 3, 4, \ldots.$ Then one says that X satisfies Bernstein condition.*

Lemma 1. *If random variable X satisfies Bernstein condition with parameter b, then : $\mathbf{E}\, e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}}, \quad \forall |\lambda| < \frac{1}{b}$. Additionally, from the bound on moment generating function one can obtain the following tail bound(also known as a* **Bernstein inequality***)*

$$\mathbf{P}(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right),\ \forall t > 0$$

Proof. Pick $\lambda : |\lambda| < \frac{1}{b}$ (allowing interchanging summation and taking expectation) and expand the MGF in a Taylor series :

$$\begin{aligned}\mathbf{E}\exp\lambda(X-\mu) = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\mathbf{E}\,|X-\mu|^k}{k!}\lambda^k &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda| b)^{k-2} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|} \\ &\leq \exp\left\{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}\right\}\end{aligned}$$

where we used $1+x \leq e^x$. To show the final bound, take $\lambda : |\lambda| < \frac{1}{2b}$. Then the bound becomes :

$$\exp\left\{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}\right\} \leq \exp \lambda^2\sigma^2 = \exp\frac{\lambda^2(2\sigma^2)}{2}$$

implying that $X \in \mathrm{SE}(2\sigma^2, 2b)$. The concentration result then follow by taking $\lambda = \frac{t}{bt+\sigma^2}$. □

4.2 Composition property of Sub - Exponential random variables

Let $X_1, ..., X_n$ be independent random variables such that $\mathbf{E}\, X_i = \mu_i$ and $X_i \in \mathrm{SE}(v_i^2, \alpha_i)$. Then

$$\sum_{i=1}^n (X_i - \mu_i) \in \mathrm{SE}\left(\sum_{i=1}^n v_i^2, \max_i \alpha_i\right)$$

In particular, denote $v_*^2 = \sum_{i=1}^n v_i^2$, $\alpha_* = \max_i \alpha_i$. Then :

$$\mathbf{P}\left(\frac{1}{n}|\sum_{i=1}^n (X_i - \mu_i)| \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{nt^2}{2v_*^2}\right), & 0 < nt \leq \frac{v_*^2}{\alpha_*} \\ 2\exp\left(-\frac{nt}{2\alpha_*}\right), & \text{otherwise} \end{cases}$$

or, equivalently,

$$\mathbf{P}\left(\frac{1}{n}|\sum_{i=1}^n (X_i - \mu_i)| \geq t\right) \leq 2\exp\left\{-\frac{n}{2}\min\left(\frac{t^2}{\nu_*^2}, \frac{t}{\alpha_*}\right)\right\}$$

Example 4.3. *Let $X \sim \mathcal{X}_n^2$ i.e.$X = \sum_{i=1}^n Z_i^2$ where $Z_i \sim \mathcal{N}(0,1)$. Then $X \in SE(4n, 4)$ and thus,*

$$\mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^n (Z_i^2 - 1)\right| \geq t\right) \leq 2\exp\left\{-\frac{n}{2}\min\left(\frac{t^2}{4}, \frac{t}{4}\right)\right\}$$

4.3 Orlicz norms

Everything said so far can be handled in more general way using Orlicz norms.

Definition 4.3. *(ψ - Orlicz norm) Let function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ satisfy the following properties : 1. $\psi(X)$ is strictly increasing function 2. $\psi(X)$ is a convex function 3. $\psi(0) = 0$. Then the ψ - Orlicz norm of a random variable X is defined as : $||X||_\psi = \inf\{t > 0 : \mathbf{E}\, \psi\left(\frac{|X|}{t}\right) \leq 1\}$*

Let us look at several examples :

1. Let $\psi(x) = x^p,\ p \geq 1$. Then : $||X||_\psi = ||X||_p = (\mathbf{E}\, |X|^p)^{\frac{1}{p}}$
2. Let $\psi_p(x) = e^{x^p} - 1,\ p \geq 1$
   (a) p = 1: then $||X||_{\psi_1} < \infty$ is equivalent to X belonging to the class of Sub - Exponential random variables.

(b) p = 2: then $||X||_{\psi_2} < \infty$ is equivalent to X belonging to the class of Sub - Gaussian random variables.

It is easy to show that : $||X^2||_{\psi_1} = (||X||_{\psi_2})^2$, $||XY||_{\psi_1} \leq ||X||_{\psi_2}||Y||_{\psi_2}$

Theorem 4.2. *(***Concentration of a sub-Gaussian random vector***) Let $X = (X_1, ..., X_d)^T \in \mathbb{R}^d$ be such that : $\mathbf{E}\, X_i = 0, \mathbf{V}(X_i) = 1$ and assume that $X_i \in SG(\sigma^2)$. Then we can show that $||X||_2 = \sqrt{\sum_{i=1}^d X_i^2}$ concentrated around $\sqrt{d}$.*

Proof. Since $||X||_2^2 = \sum_{i=1}^d X_i^2$, so $X_i^2 - 1 \in \text{SE}(v^2, \alpha)$ where both parameters are determined by σ^2. Thus, by the property of Composition property of Sub - Exponential, we have that

$$\mathbf{P}\left(\frac{1}{d}\left|\sum_{i=1}^{d}(X_i^2 - 1)\right| \geq t\right) \leq 2\exp\left\{-\frac{d}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right)\right\}, \quad \forall t > 0$$

We will need to use the following fact : fix $c > 0$. Then for any numbers $z > 0$:

$$|z-1| \geq c \xrightarrow{\text{implies}} z^2 - 1 \geq \max\{c, c^2\}$$

Using this fact allows to conclude that :

$$\mathbf{P}\left(\left|\frac{\|X\|_2}{\sqrt{d}} - 1\right| \geq u\right) = \mathbf{P}\left(\left|\frac{\|X\|_2^2}{d} - 1\right| \geq \max\{u, u^2\}\right) \leq 2\exp\left(-\frac{du^2}{2C}\right)$$

□

4.4 Hoeffding vs Bernstein

Denote $\mu = \mathbf{E}(X)$ and $\sigma^2 = \mathbf{V}(X)$. Assume that $|X - \mu| \leq b$ a.e. Then :

$$\mathbf{P}(|X-\mu| \geq t) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2b^2}\right) & \text{Hoeffding} \\ 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) & \text{Bernstein} \end{cases}$$

For small t (meaning $bt \ll \sigma^2$) Bernstein's inequality gives rise to a bound of the order:

$$\mathbf{P}(|X-\mu| \geq t) \leq 2\exp\left(-\frac{t^2}{c\sigma^2}\right)$$

while Hoeffding's gives:

$$\mathbf{P}(|X-\mu| \geq t) \leq 2\exp\left(-\frac{t^2}{cb^2}\right)$$

But $\sigma^2 \leq b^2$ and thus, Bernstein's bound is better / tighter.

Theorem 4.3. *(Classic Bernstein inequality) Let $X_1, \cdots, X_n$ be independent random variables such that $|X_i - \mathbf{E}\, X_i| \le b$, a.e and $\max_i \mathbf{V}(X_i) \le \sigma^2$. Then*

$$\mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - \mathbf{E}\, X_i)\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2bt}{3}}\right), \quad \forall t > 0$$

Theorem 4.4. *(Laurent-Massart bounds for $\mathcal{X}^2$) Let $Z_1, \cdots, Z_d \sim \mathcal{N}(0,1)$ and $a = (a_1, \cdots, a_d)$ with $a_i \ge 0, \forall i \in \{1, \cdots, n\}$. Let $X = \sum_{i=1}^{n} a_i(X_i^2 - 1)$. Then for right-tail behavior is described by*

$$\mathbf{P}(X \ge 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t) \le e^{-t}, \forall t > 0$$

and for left - tail behavior:

$$\mathbf{P}(X \le -2\|a\|_2\sqrt{t}) \le e^{-t}, \forall t > 0$$

5 The bounded differences inequality

5.1 Bounded Difference Property

5.2 Application

6 Bound for Sub-gaussian vector and Covariance matrix

6.1 SG vectors and bound for the their norm

Definition 6.1. *(Sub-Gaussian random vectors) A random vector $X \in \mathbb{R}^d$ is a sub - Gaussian random vector with parameter σ^2 if $v^T X \in SG(\sigma^2)$, $v \in \mathbb{S}^{d-1}$ where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ is d - 1 unit sphere. We denote $X \in SG_d(\sigma^2)$.*

Claim : If $X \sim \mathcal{N}(0, \Sigma)$, then $X \in \mathbb{R}^d$ is a sub - Gaussian random vector with parameter $\|\Sigma\|_{op}$.

Proof. For any $v \in \mathbb{S}^{d-1}, v^T\Sigma v \le \|\Sigma\|_{op}$. Take MGF : $\mathbf{E}[e^{\lambda v^T X}] = e^{\lambda^2 v^T \Sigma v/2} \le e^{\lambda^2\|\Sigma\|_{op}/2}$. □

Theorem 6.1. *Let $X \in SG_d(\sigma^2)$, then $\mathbf{E}\|X\|_2 \le 4\sigma\sqrt{d}$. Moreover, with probability at least $1 - \delta$ for $\delta \in (0,1)$, $\|X\|_2 \le 4\sigma\sqrt{d} + 2\sigma\sqrt{\log(\frac{1}{\delta})}$.*

Proof. □

6.2 Covariance matrix estimation in the operator norm.

Theorem 6.2. *Let $X_1, \cdots, X_n$ be iid samples from a distribution with mean 0 and covariance matrix Σ. Assume $X_i \in SG_d(\sigma^2)$ and are centered. Let $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$. Then there exists a universal constant $C > 0$ s.t.*

$$\mathbf{P}\left(\frac{\|\widehat{\Sigma}_n - \Sigma\|_{op}}{\sigma^2} \geq C \max\left\{\sqrt{\frac{d+\log(\frac{2}{\delta})}{n}}, \frac{d+\log(\frac{2}{\delta})}{n}\right\}\right) \leq \delta, \delta \in (0,1)$$

Proof. □

7 Matrix Concentration Inequalities

7.1 Matrix Bernstein Inequality

Theorem 7.1.

Proof. □

7.2 Matrix Hoeffding Inequality

7.3 Application

8 Ordinary and Penalized regression

8.1 OLS regression in high dimension

Recall : $Y = X\beta^* + \epsilon$ where X is the fixed design matrix, $\epsilon \in \mathrm{SG}_n(\sigma^2)$. We have that $\beta^* = (X^TX)^{-1}X^TY$ as the OLS solution (which can be one of infinitely many solutions).

Theorem 8.1. *There exists universal constants $C > 0$ s.t. $\frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2 \leq C\sigma^2\left(\frac{r+\log(1/d)}{n}\right)$ where $r = rank(X^TX)$.*

Proof. □

8.2 Penalized regression

$$\widehat{\beta} \in \min_{\beta \in \mathbb{R}^d} \|Y - X\beta\|_2^2 + \lambda_n f(\beta)$$

A classic penalty term is $f(\beta) = \|\beta\|_2^2$ (ridge regression) :

$$\beta_{\text{ridge}} = (X^T X + \lambda_n I)^{-1} X^T Y$$

which is always unique even if $n > d$.

The interpretation is, consider the SVD decomposition of $X = U\Lambda U^T$. Plugging this into $X\widehat{\beta}_{\text{ridge}}$, we have

$$X\widehat{\beta}_{\text{ridge}} = X(X^T X + \lambda_n I)^{-1} X^T Y = UHU^T Y = \sum_{j=1}^{r} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \langle u_j, Y \rangle$$

where H is a diagonal matrix with $H_{jj} = \frac{\sigma_j^2}{\sigma_j^2 + \lambda}$. We can see that ridge gives higher weight to directions u_j with large σ_j^2 and may be considered a smarter projection, whereas for OLS, all basis u_j is weighted the same amount.

8.3 Slow and Fast rates for Lasso

9 Principle Component Analysis

10 Uniform Law of Large number