Math 542: Analysis of Variance and Regression
Homework 1

due on Sunday, March 5 at 11:59 pm

Please submit electronically directly to Blackboard in a PDF file.

0^o (Warm-up: expectation and covariance for random vectors).

Let $X \in \mathbb{R}^d$ be a random vector with $\mathbb{E}[X] = \mu$ and covariance matrix $\mathrm{Cov}(X) = \mathbf{\Sigma}$. Show that:

(a) For the second-moment matrix of X is $\mathbb{E}[\|X\|^2] = \mu\mu^\top + \mathbf{\Sigma}$.

(b) $Z := \mathbf{\Sigma}^{-1/2}(X - \mu)$ has zero mean and identity covariance $\boldsymbol{I}_d$.

(c) Find the mean, covariance matrix, and the second-moment matrix of $W := \Sigma^{-1/2} X$.

(d) Assuming that $d > 1$ and $\mu \neq 0$, show that the eigenvalues of $\boldsymbol{I}_d + \mu\mu^\top$ are $\|\mu\|^2 + 1$ and 1. What are the corresponding eigenvectors?

1^o (Fixed-design linear regression).

Now, consider the linear regression model we analyzed in class: observed are pairs (x_i, y_i) where

$$y_i = x_i^\top \theta^* + \sigma\xi_i, \quad i \in \{1, ..., n\};$$

the *predictors* (or *covariates*) x_i's are deterministic (non-random), and $\theta^* \in \mathbb{R}^d$ is fixed, but unknown; finally, $\xi_i \sim \mathcal{N}(0,1)$ are i.i.d. Recall that this can be equivalently written in a matrix-vector form:

$$Y = \boldsymbol{X}\theta^* + \sigma\xi \tag{1}$$

where $Y, \xi \in \mathbb{R}^n$, and

$$\boldsymbol{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}$$

is the design matrix. Define $\mu^* := \boldsymbol{X}\theta^*$, the mean of Y. Assume that $n \geqslant d$, and X has full column rank, so that $X^\top X$ is invertible. Recall, from what we have seen in class, that $\widehat{\theta} := \boldsymbol{X}^+ Y$ and $\widehat{\mu} = \mathbf{\Pi}_{\boldsymbol{X}} Y$ are the least-squares estimates of θ^* and μ^* correspondingly; here

$$\boldsymbol{X}^+ := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

is the *left pseudoinverse* of $\boldsymbol{X}$ (that is, $\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{I}$), while

$$\begin{aligned} \mathbf{\Pi}_{\boldsymbol{X}} &:= \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \\ &= \boldsymbol{X}\boldsymbol{X}^+ \end{aligned}$$

is the projector on $\mathrm{Col}(\boldsymbol{X})$, the column space of $\boldsymbol{X}$.

Prediction:

(a) Recap of in-class material: show that $\widehat{\mu}$ is unbiased, and $\mathrm{Cov}(\widehat{\mu}) = \sigma^2 \boldsymbol{\Pi}_{\boldsymbol{X}}$. (You don't need to assume $\xi \sim \mathcal{N}(0, \boldsymbol{I}_n)$—only $\mathbb{E}[\xi] = 0$ and $\mathrm{Cov}(\xi) = \boldsymbol{I}_n$.) Conclude that $\mathbb{E}[\|\widehat{\mu} - \mu^*\|^2] = \sigma^2 d$, and compare this with the mean-squared error $\mathbb{E}[\|Y - \mu^*\|]$ of Y—the "trivial estimate" of μ^*.

(b) Using the previous result, show that for any fixed unit vector $u \in \mathbb{R}^n$ (i.e., such that $\|u\| = 1$),

$$\mathbb{E}[\langle u, \widehat{\mu} - \mu^* \rangle] = 0 \quad \text{and} \quad \mathrm{Var}(\langle u, \widehat{\mu} - \mu^* \rangle) = \sigma^2 \|\boldsymbol{\Pi}_{\boldsymbol{X}} u\|^2 \leqslant \sigma^2.$$

Give a geometric-statistical interpretation of these two identities (*what is* $\langle u, \widehat{\mu} - \mu^* \rangle$*?*). Using the properties of multivariate Gaussian, show that $\langle u, \widehat{\mu} - \mu^* \rangle \sim \mathcal{N}(0, \sigma_u^2)$ with appropriate σ_u^2.

(c) Using (*a*)–(*b*), show that $\frac{1}{\sigma^2}\|\widehat{\mu} - \mu^*\|^2 \sim \chi_d^2$. (*Hint: select d vectors* $u^{(1)}, ..., u^{(d)}$ *appropriately.)*

Estimation:

For the remaining part of this exercise, define $\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$. (The factor $\frac{1}{n}$ might look unwarranted here, but it will become natural in the context of random-design regression.)

(*d*) Show that $\mathbb{E}[\widehat{\theta}] = \theta^*$ and $\mathrm{Cov}(\widehat{\theta}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}$. Explain (in a few words) why $\widehat{\theta}$ has a multivariate Gaussian distribution.

(*e*) Reflect on the formula $\mathrm{Cov}(\widehat{\theta}) = \frac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}$ assuming $\boldsymbol{\Sigma}$ is a diagonal matrix, i.e. $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$ with

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_d).$$

In this case, $\mathrm{Var}(\widehat{\theta}_i) = \frac{\sigma^2}{\lambda_i n}$ for each coordinate $i \in \{1, ..., d\}$—in particular, the smaller λ_i, the larger the error of estimating the correponding θ_i^*. (E.g., if $\lambda_1 = 0.01$ and $\lambda_2 = ... = \lambda_d = 1$, then $\mathrm{Var}(\widehat{\theta}_1) = 100\frac{\sigma^2}{n}$ but $\mathrm{Var}(\widehat{\theta}_i) = \frac{\sigma^2}{n}$ for $i > 1$.) The next part of the problem explains this!

*(f) Denote $\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$. I claim that the problem of estimating θ^* from "indirect" observations Y, cf. (1), can be reformulated as estimating the same vector θ^* but from "direct" observations,

$$\omega = \theta^* + \sigma\varepsilon, \tag{2}$$

with "colored" noise $\varepsilon \sim \mathcal{N}(0, \frac{1}{n}\boldsymbol{\Sigma}^{-1})$.

(f.1) Describe—rigorously—how to pass from (1) to (2).[1]

(f.2) Verify that $\widehat{\theta} = \boldsymbol{X}^+ Y$ is precisely ω, and is also the (trivial) least-squares estimate of θ^* from observations ω in (2). *(Hint: we can treat* (2) *as a specific case of* (1)*, can't we?)*

2^o (Right tail bound for χ_d^2, a.k.a. Bernstein's inequality).

Let $X \sim \chi_{2d}^2$ (chi-squared distribution with $2d$ degrees of freedom), that is $X = \|Z\|^2 = Z_1^2 + ... + Z_{2d}^2$ where $Z \sim \mathcal{N}(0, \boldsymbol{I}_d)$ (equivalently, $Z_i \sim \mathcal{N}(0,1)$ are i.i.d.). Define $M_{2d}(\cdot)$ as the moment generating function (MGF) of $X \sim \chi_{2d}^2$, i.e.

$$M_{2d}(t) := \mathbb{E}[e^{tX}], \quad t \in \mathbb{R};$$

[1] Model (2) is called Gaussian sequence model (GSM). In fact, even in the case $\boldsymbol{\Sigma} = \boldsymbol{I}$—trivial in our context—GSM gives rise to a rich theory as soon as θ^* is allowed to vary over some set $\Theta \subseteq \mathbb{R}^d$, instead of being fixed. This theory goes way beyond our course—see, e.g., the books `https://imjohnstone.su.domains//GE_08_09_17.pdf` and [Tsy09].

in particular, $M_2(t) = \mathbb{E}\big[e^{t(Z_1^2+Z_2^2)}\big]$. Our ultimate goal here is to prove that, with probability $\geqslant 1-\delta$,

$$X - 2d \leqslant \sqrt{Cd\log\left(\frac{1}{\delta}\right)} + c\log\left(\frac{1}{\delta}\right) \tag{3}$$

for some numerical constants $C, c > 0$. This bound is, in fact, optimal (see, e.g., [LM00, Lemma 1]).

(i) Derive the explicit form of $M_2(t)$:

$$M_2(t) = \frac{1}{1-2t}, \quad t < \frac{1}{2},$$

and $M_2 = +\infty$ for $t \geqslant \frac{1}{2}$. (To take the integral, pass to polar coordinates $(z_1, z_2) \mapsto (r, \theta)$ with $r = \sqrt{z_1^2 + z_2^2}$—and don't forget the Jacobian, which equals r.) Claim that, as a corollary,

$$M_{2d}(t) = \frac{1}{(1-2t)^d}, \quad t < \frac{1}{2}.$$

(ii) Using Chernoff's method, bound the tail function $\mathbb{P}(X > x)$, for any $x > 2d$, as follows:

$$\mathbb{P}(X > x) = \inf_{t<\frac{1}{2}} \frac{e^{-tx}}{(1-2t)^d} = \exp\left(d\log\left(\frac{2d}{x}\right) - \frac{x-2d}{2}\right).$$

(Hint: it is convenient to take the logarithm, and use that $u \mapsto \log(u)$ *on* $\mathbb{R}$ *is increasing.)* Note that, in terms of the deviation $z = x - 2d > 0$ above $2d$, this is equivalent to

$$\mathbb{P}(X - 2d > z) = \exp\left(d\log\left(\frac{2d}{2d+z}\right) - \frac{z}{2}\right).$$

*(iii) *Bear with me: this part is a bit delicate, but we need it to reach the conclusion.* Use that

$$\log(u) \leqslant u - 1 \quad (\forall u \in \mathbb{R}),$$

along with some simple algebra, to show that

$$\mathbb{P}(X - 2d > z) \leqslant \begin{cases} \exp\left(-\dfrac{z^2}{8d}\right) & \text{for } \ 0 \leqslant z \leqslant 2d, \\ \exp\left(-\dfrac{z}{4}\right) & \text{for } \ z > 2d. \end{cases}$$

It is also fine if you get some worse pair of constants $C > 8, c > 4$ (leading to a weaker bound). Finally, reformulating the last bound as

$$\mathbb{P}(X - 2d > z) \leqslant \exp\left(-\min\left\{\frac{z^2}{8d}, \frac{z}{4}\right\}\right)$$

and letting $\mathbb{P}(X - 2d > z) = \delta$, "invert" the last inequality to get (3) with $C = 8$ and $c = 4$ (or with worse constants if in (iii) you got a weaker bound). (*Hint:* $\max\{a, b\} \leqslant a + b$ *for* $a, b \geqslant 0$.)

References

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.