Math 6262: Statistical Estimation
Homework 1

due on Sunday, Feb 11 at 11:59 pm

Please submit electronically directly to Canvas as a PDF file.

$\mathbf{0^o}$: **Warm-up** (not graded) – *expectation and covariance matrix in* $\mathbb{R}^d$.

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 $\mu \neq 0$, find the eigenvalues and eigenvectors of $\boldsymbol{I}_d + \mu\mu^\top$.

$\mathbf{1^o}$: *Tail bounds for the Gaussian distribution.*

Let $\phi(\cdot)$ be the p.d.f. of $\mathcal{N}(0,1)$, i.e. $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. For any $u \geqslant 0$, let $\Phi(u) := \int_{t \geqslant u} \phi(t) dt$.

(a) Prove the following bounds (holding for all $u \geqslant 0$):

$$\left(\frac{1}{u} - \frac{1}{u^3}\right)\phi(u) \leqslant \Phi(u) \leqslant \frac{1}{u}\phi(u).$$

Hint 1: Try to prove the upper bound first.

Hint 2: Integrate by parts – first to prove the upper bound, then again for the lower bound.

(b) Capitalizing on the trick you have just figured out to get the lower bound from the upper bound, prove a new upper bound:

$$\Phi(u) \leqslant \left(\frac{1}{u} - \frac{1}{u^3} + \frac{3}{u^5}\right)\phi(u).$$

Note that this bound is sharper than the previous one for large enough u.

$^*(c)$ **Bonus.** If we continue this approach iteratively, write down the bounds that we get further (at step k). You can omit a rigorous proof – just figure out the mechanism, and explain it.

2°: *Stein's paradox.*

Consider the problem of estimating the mean μ in the multivariate Gaussian location family

$$\mathbb{P}_\mu = \mathcal{N}(\mu, \boldsymbol{I}_d), \quad \mu \in \mathbb{R}^d, \tag{1}$$

where $\boldsymbol{I}_d$ is the $d \times d$ identity matrix, from a single observation $X \sim \mathbb{P}_\mu$. Note that here, X itself is the maximum likelihood estimator (MLE) for μ. Defining for any estimator $\hat{\mu} = \hat{\mu}(X)$ of μ the variance

$$\mathrm{Var}_\mu[\hat{\mu}] := \mathbb{E}_\mu[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2]$$

and the quadratic risk

$$\mathrm{Risk}_\mu[\hat{\mu}] := \mathbb{E}_\mu[\|\hat{\mu} - \mu\|^2],$$

where $\|x\| := (\sum_i x_i^2)^{1/2}$ is the Euclidean norm of $x = (x_1, ..., x_d) \in \mathbb{R}^d$, we see that for any $\mu \in \mathbb{R}^d$,

$$\mathrm{Risk}_\mu[X] = \mathrm{Var}_\mu[X] = d.$$

Intuitively, one can suspect that no better estimator of X can be found: really, what can be done with only a single observation of the mean? Yet, this turns out to be false: one may improve over the MLE uniformly on the family (1) when $d > 2$. This celebrated result was established by James and Stein in 1976, and our goal is to reproduce it. But first, let us establish the terminology.

Definition 1. An estimator $\hat{\mu}$ is *dominated* by some other estimator $\hat{\mu}'$ if $\mathrm{Risk}_\mu[\hat{\mu}'] \leqslant \mathrm{Risk}_\mu[\hat{\mu}]$ for any μ, and there exists a parameter value $\bar{\mu}$ such that $\mathrm{Risk}_{\bar{\mu}}[\hat{\mu}'] < \mathrm{Risk}_{\bar{\mu}}[\hat{\mu}]$.

Definition 2. An estimator $\hat{\mu}$ is called *admissible* if it is not dominated by any other estimator. Otherwise, it is called *inadmissible*.

As statisticians, ideally we would like to compare two estimators over the whole family at once, without specifying a value of μ. Two admissible estimators cannot be compared this way, but at the very least we can rule out any *inadmissible* estimator, as for it there exists a uniformly better one.

You will show that the MLE is inadmissible when $d \geqslant 3$, by constructing a dominating estimator.

(*a*) Consider *shrinkage estimators* $\hat{\mu} = sX$ with $s \in \mathbb{R}$, and compute their risks for any s. Show that one can restrict attention to $s \in [0, 1]$ (hence "shrinkage") by finding a dominating estimator for $\hat{\mu}$ with $s < 0$ or $s > 1$.

(*b*) Show that, for given μ, the best value of s—i.e., the one minimizing the risk—is given by

$$s^* = \frac{\|\mu\|^2}{d + \|\mu\|^2} = 1 - \frac{d}{d + \|\mu\|^2}.$$

(*c*) Unfortunately, $\hat{\mu}^* = s^* X$ is not a proper estimator. (*Why?*) Instead of it, one may consider

$$\left(1 - \frac{d}{\|X\|^2}\right) X,$$

which is an actual estimator. Can you explain the heuristic motivation behind this estimator?

$^*(d)$ **Bonus.** Assuming that $d \geqslant 2$, derive the *James-Stein estimator*

$$\hat{\mu}^{JS} = \left(1 - \frac{d-2}{\|X\|^2}\right) X \tag{2}$$

by minimizing over $\delta \in \mathbb{R}$ the risk of the estimator

$$\hat{\mu}^{\delta} = \left(1 - \frac{\delta}{\|X\|^2}\right) X$$

for a fixed μ. In order to show that $R(\delta) = \mathrm{Risk}_{\mu}[\hat{\mu}^{\delta}]$ is minimized at $d-2$, use Stein's lemma:

Lemma 1. *Let $X \sim \mathcal{N}(\mu, I)$ and $g(x)$ be a function on $\mathbb{R}^d$ differentiable almost everywhere, and such that $\mathbb{E}_{\mu}\left[|\frac{\partial}{\partial x_i} g(X)|\right] < \infty$ and $\mathbb{E}_{\mu}[|(X_i - \mu_i)g(X)|] < \infty$ for any $i \in [d] := \{1, 2, ..., d\}$. Then*

$$\mathbb{E}_{\mu}[(X_i - \mu_i)g(X)] = \mathbb{E}_{\mu}\left[\frac{\partial}{\partial x_i} g(X)\right], \quad i \in [d].$$

When applying Stein's lemma to the right function $g(X)$, please do check the absolute integrability conditions in its premise, and explain why the argument does not work for $d = 1$.

Finally, verify that $R(\delta)$ is strictly convex when $d \geqslant 3$ (thus $\hat{\mu}^{JS}$ indeed dominates the MLE).

3^o^: *Right tail bound for χ^2_d, a.k.a. Bernstein's inequality.*

Let $X \sim \chi^2_{2d}$ (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 $X \sim \chi^2_{2d}$, i.e.

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

in particular, $M_2(t) = \mathbb{E}\left[e^{t(Z_1^2+Z_2^2)}\right]$. 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]).

(a) 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,

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

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

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

(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

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

*(*c*) **Bonus.** *Bear with me, this part is a bit delicate – but we need it to reach the conclusion.*

(*c.i*) Show that

$$\mathbb{P}(X - 2d > z) \leqslant \begin{cases} \exp\left(-\dfrac{z^2}{16d}\right) & \text{for } 0 \leqslant z \leqslant 2d, \\ \exp\left(-\dfrac{z}{8}\right) & \text{for } z > 2d. \end{cases}$$

It is OK if you get some worse pair of constants $C > 16, c > 8$ leading to a weaker bound. *Hint: first show, using calculus, that*

$$\log(1+u) \leqslant u - \tfrac{1}{4}\min\{u, u^2\} \quad \forall u \geqslant 0$$

(*c.ii*) *Reformulating the last bound as*

$$\mathbb{P}(X - 2d > z) \leqslant \exp\left(-\min\left\{\frac{z^2}{16d}, \frac{z}{8}\right\}\right)$$

and letting $\mathbb{P}(X - 2d > z) = \delta$, "invert" the last inequality to get (3) with $C = 16$ and $c = 8$ (or with some worse constants). Hint: $\max\{a, b\} \leqslant a + b$ for $a, b \geqslant 0$.

*$\mathbf{4^o}$ **Bonus:** *Planar Venn diagrams.*
A (congruent) *Venn diagram* in $\mathbb{R}^d$ for n sets is the following object: you choose a "base" set $A \subset \mathbb{R}^d$ and n locations $a_1, ..., a_n \in \mathbb{R}^d$ such that the shifted sets $A_1, A_2, ..., A_n$, where $A_j := \{a + a_j, a \in A\}$, intersect in all possible combinations: for any subset of indices $I \subseteq \{1, 2, ..., n\}$, the set $A_I := \cap_{i \in I} A_i$ must be nonempty. Prove the following result:

One cannot draw a planar ($d = 2$) Venn diagram for $n \geqslant 5$ sets by shifting a circle.

Use **Euler's formula**: any planar graph with V vertices, E edges, and F faces (subsets in which $\mathbb{R}^2$ is partitioned by the graph) satisfies

$$V - E + F = 2.$$

For example, in the case of a triangle $V = E = 3$ and $F = 2$.
Hint: estimate V_n, E_n, F_n in a Venn diagram for n sets in terms of $V_{n-1}, E_{n-1}, F_{n-1}$ respectively.[1]

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.*

[1] In fact, $n = 4$ is also impossible, but I am not aware of a purely combinatorial (and elegant) proof.