Stats 210A, Fall 2025
Homework 5

Due on: Wednesday, Oct. 8

Instructions: You may disregard measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, "all functions" vs. "all measurable functions," etc. (unless the problem is explicitly asking about such issues).

Problem 1 (Ridge regression)**.** Consider the *Gaussian linear model* where

$$y_i = x_i'\beta + \varepsilon_i, \quad \text{with } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \text{ for } i = 1, \dots n,$$

where $\beta \in \mathbb{R}^d$ is unknown, and the covariate vectors $x_i \in \mathbb{R}^d$ are fixed and known. Assume the error variance $\sigma^2 > 0$ is also known. We observe the response vector $y \in \mathbb{R}^n$.

(a) Assume that $d \leq n$, and the design matrix $\mathbf{X}$ (the $n \times d$ matrix whose ith row is x_i') has full column rank. Show that the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ is the UMVU estimator of β.

Note: Remember that the design matrix $\mathbf{X}$ is not data in the same sense y is; it is more like a known parameter.

(b) Now consider Bayesian estimation with the prior $\beta \sim N(\mu, \tau^2 I_d)$. Under the same prior as in part (b), find the posterior distribution of β. Does it matter whether $d > n$, or whether $\mathbf{X}$ has full column rank?

(c) Suppose that $\mathbf{X}\gamma = 0$ for some nonzero $\gamma \in \mathbb{R}^d$ whose entries are all nonzero. Show that no unbiased estimator exists for any β_j. In your opinion, should this give us any reason for concern about the Bayes estimator? (The subjective part of the question will be graded leniently).

Moral: While the humble ridge regression method was initially introduced just to deal with poorly conditioned regression problems, it has an appealing Bayesian interpretation. Moreover, using Bayesian methods often helps us sidestep concerns about identifiability — but we should think hard about the implications of doing so.

Problem 2 (Admissibility and Bayes estimators)**.** One of the frequentist motivations for Bayes estimators is their connection to admissibility.

(a) Suppose that the Bayes estimator δ_Λ for the prior Λ is unique up to $\mathcal{P}$-almost-sure equality. That is, for any other Bayes estimator $\tilde{\delta}_\Lambda$, we have $\delta_\Lambda(X) = \tilde{\delta}_\Lambda(X)$ almost surely, for every $P_\theta \in \mathcal{P}$. Show that δ_Λ is admissible.

(b) Now suppose that Θ is discrete (possibly countably infinite) and Λ has a strictly positive probability mass function λ, i.e. $\lambda(\theta) > 0$ for all $\theta \in \Theta$. Show that any Bayes estimator with finite Bayes risk is admissible.

(c) **Optional:** (Not graded, no extra points) If $\Theta \subseteq \mathbb{R}^d$ and Λ has a strictly positive density, can we likewise say that all Bayes estimators with finite Bayes risk are admissible? Prove or give a counterexample.

(d) A *randomized estimator* is an estimator that is a random function of the data. We can formalize it generically as $\delta(X, W)$ where $X \sim P_\theta$ as usual and W is some auxiliary random variable generated by the analyst. For this part, "admissible" and "Bayes" are defined with respect to all estimators including randomized ones.

Now consider a model with a finite parameter space, $|\Theta| = n < \infty$ and assume we are estimating some real-valued $g(\theta)$ using a bounded non-negative loss $L : \Theta \times \mathbb{R} \to [0, \infty)$. Show that every admissible estimator is a (possibly randomized) Bayes estimator for some prior.

Hint: consider the set $\mathcal{A}$ of all achievable risk functions, and the set $\mathcal{D}_\delta$ of all (possibly unachievable) risk functions that would dominate a given estimator δ. Recall the *hyperplane separation theorem*: for any two disjoint non-empty convex subsets $A, B \subseteq \mathbb{R}^n$ there exist $c \in \mathbb{R}$ and nonzero $\lambda \in \mathbb{R}^n$ such that $\lambda' a \geq c \geq \lambda' b$ for all $a \in A, b \in B$. It might help to draw pictures for $n = 2$.

Moral: Minimizing average-case risk is closely related to admissibility, though the general relationship is not quite as tight as what we've shown in this problem for finite parameter spaces. In more general parameter spaces, there is a more general result which roughly states that all admissible estimators are limits of Bayes estimators, under relatively mild conditions.

Problem 3 (Other loss functions)**.** Assume for each problem below that there exists an estimator with finite Bayes risk.

(a) Consider a Bayesian model with a discrete parameter θ. What is the Bayes estimator for the loss $L(\theta, d) = 1\{\theta \neq d\}$?

(b) Next consider a Bayesian model with a single real parameter θ, and assume that the posterior distribution of θ given $X = x$ is absolutely continuous (with respect to the Lebesgue measure) for all x. What is the Bayes estimator for the *absolute error loss* $L(\theta, d) = |\theta - d|$?

(c) Under the same assumptions as part (b), what loss function $L_\gamma(\theta, d)$ would give the posterior γ quantile as its Bayes estimator; that is, the estimator $\delta_\gamma(X)$ has $\mathbb{P}(\theta < \delta_\gamma(X) \mid X) = \gamma$.

Moral: Using different loss functions often gives other appealing summaries of the posterior distribution besides the posterior mean. Using these loss functions as the prediction error in regression problems can also lead to appealing methods such as quantile regression.

Problem 4 (Motivation for the Jeffreys prior)**.** Recall that the *Kullback–Leibler (KL) Divergence* from a density p to an approximating density q (on a sample space $\mathcal{X}$ with respect to base measure μ) is

$$D_{\mathrm{KL}}(p \,\|\, q) = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\, d\mu(x),$$

where the integrand is treated as zero wherever $p(x) = 0$. For a density p and a family $\mathcal{P}$, define the KL ball of radius ε around p as

$$B_\varepsilon(p) = \{q \in \mathcal{P} : \; D_{\mathrm{KL}}(p \,\|\, q) < \varepsilon\}$$

Now, assume we have a family $\mathcal{P}$ of positive densities p_θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$, with enough regularity for us to freely take derivatives under the integral, and for those derivatives to be continuous. In particular, assume the Fisher information $J(\theta)$ is continuous and (strictly) positive definite, so $|J(\theta)| > 0$.

In this problem you will show that the Jeffreys prior Λ spreads its prior mass evenly on $\mathcal{P}$, in the sense that for distinct values $\theta_1, \theta_2 \in \Theta^\circ$ (the interior of Θ), we have

$$\lim_{\varepsilon \to 0} \frac{\Lambda(B_\varepsilon(p_{\theta_1}))}{\Lambda(B_\varepsilon(p_{\theta_2}))} = 1.$$

Here Λ could refer either to the probability distribution with density proportional to $|J(\theta)|^{1/2}$ (if the Jeffreys prior is proper) or an infinite measure with density equal to $|J(\theta)|^{1/2}$. Note the slight abuse of notation where we use Λ to refer both to the Jeffreys prior on Θ and to the corresponding prior on $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$.

To rule out perversities, assume also that for each θ there is some radius r_θ such that $d_\theta(\zeta) = D_{\mathrm{KL}}(p_\theta \,\|\, p_\zeta)$ is convex for $\|\zeta - \theta\| \leq r$, and $\inf_{\|\zeta-\theta\|>r} d_\theta(\zeta) > 0$ (in particular, this implies identifiability). That is, small KL balls in $\mathcal{P}$ are not going to include p_ζ with faraway values of ζ. You don't need to explicitly use this condition; you can just not pay attention to what's happening away from infinitesimal neighborhoods around θ_1 and θ_2 and assume everything is fine.

(a) For $\theta \in \Theta^\circ$, and any vector $a \in \mathbb{R}^d$, show that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+\varepsilon a}) = \frac{1}{2} a' J(\theta) a, \tag{1}$$

or equivalently $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+a}) = \frac{1}{2} a' J(\theta) a + o(\|a\|^2)$.

Hint 1: Try differentiating the KL divergence and recall that, for $f : \mathbb{R}^d \to \mathbb{R}$, where $x, a \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}$, we have

$$\frac{d}{d\varepsilon} f(x + \varepsilon a) = a' \nabla f(x + \varepsilon a)$$
$$\frac{d^2}{d\varepsilon^2} f(x + \varepsilon a) = a' \nabla^2 f(x + \varepsilon a) a$$

Hint 2: You are welcome to jump to the general case $d \geq 1$, but it may help to warm up with the case $d = 1$, in which case the above boils down to

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+\varepsilon a}) = \frac{a^2}{2} J(\theta)$$

or more simply

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+\varepsilon}) = \frac{1}{2} J(\theta)$$

(b) Next, show that for any $\theta \in \Theta^\circ$, we have

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^{d/2}} \Lambda\left(\{\zeta : (\zeta - \theta)' J(\theta)(\zeta - \theta) < \varepsilon\}\right) = \begin{cases} V_d / C_\Lambda & \text{if } \Lambda \text{ is proper and } \lambda(\theta) \propto_\theta |J(\theta)|^{1/2} \\ V_d & \text{if } \Lambda \text{ is improper and } \lambda(\theta) = |J(\theta)|^{1/2} \end{cases}.$$

where $V_d = \int_{\mathbb{R}^d} 1\{\|u\| < 1\} du$ is the volume of a unit ball in $\mathbb{R}^d$ and $C_\Lambda = \int_\Theta |J(\theta)|^{1/2} \, d\theta$ is the normalizing constant for the Jeffreys prior, when it is finite.

Hint: Try the change of variables $\zeta(a) = \theta + J(\theta)^{-1/2} a$.

(c) Now, finish the argument by showing that we have

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^{d/2}} \Lambda(B_{\varepsilon/2}(p_\theta)) = \begin{cases} V_d / C_\Lambda & \text{if } \Lambda \text{ is proper and } \lambda(\theta) \propto_\theta |J(\theta)|^{1/2} \\ V_d & \text{if } \Lambda \text{ is improper and } \lambda(\theta) = |J(\theta)|^{1/2} \end{cases}.$$

Hint: Try to relate the sublevel sets of $f_\theta(\zeta) = D_{\mathrm{KL}}(p_\theta \,\|\, p_\zeta)$ to sublevel sets of $\frac{1}{2}(\zeta - \theta)' J(\theta)(\zeta - \theta)$.

Moral: Whereas the flat prior spreads mass evenly over the parameter space, the Jeffreys prior spreads mass evenly over the statistical model itself, by assigning equal probability to small neighborhoods of equal "KL radius." Since the KL divergence exists prior to the parametrization, the Jeffreys prior therefore implements the "principle of indifference" without reference to any parametrization of the model.