Stats 210A, Fall 2024
Homework 5

Due on: Wednesday, Oct. 9

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

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

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).

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 (MCMC algorithms)**.** This problem considers MCMC sampling from a generic posterior density $\lambda(\theta \mid x)$ where $\theta \in \mathbb{R}^d$.

(a) The Metropolis–Hastings algorithm is a Markov chain using the following update rule: First, sample $\zeta \sim f(\cdot \mid \theta^{(t)})$ according to some "proposal distribution" $f(\zeta \mid \theta) : \Theta \times \Theta \to (0, \infty)$, where $f(\cdot \mid \theta)$ is a probability density for each θ (assume λ and $f(\cdot \mid \theta)$ are densities w.r.t. the same dominating measure). Next, compute the "accept probability" as

$$a(\zeta \mid \theta) = \min\left\{1, \ \frac{\lambda(\zeta \mid X)}{\lambda(\theta \mid X)} \frac{f(\theta \mid \zeta)}{f(\zeta \mid \theta)}\right\}.$$

Finally, let $\theta^{(t+1)} = \zeta$ with probability $a(\zeta \mid \theta^{(t)})$ and $\theta^{(t+1)} = \theta^{(t)}$ with probability $1 - a(\zeta \mid \theta^{(t)})$. Show that $\lambda(\theta \mid X)$ is stationary for the Metropolis–Hastings algorithm.

(b) Consider the version of the Gibbs sampler that updates a *single* random index $J^{(t+1)} \sim \text{Unif}\{1, \ldots, d\}$ at each step, so

$$\theta_j^{(t+1)} = \begin{cases} \zeta_j^{(t+1)} & \text{if } j = J^{(t+1)} \\ \theta_j^{(t)} & \text{if } j \neq J^{(t+1)} \end{cases},$$

with

$$\zeta_j^{(t+1)} \mid \theta^{(t)} \sim \lambda(\theta_j \mid \theta_{\setminus j} = \theta^{(t)}, X),$$

where λ above is the conditional density for the jth coordinate of θ given the others, and the data, in the full Bayes model. Show that this algorithm is a special case of the Metropolis–Hastings algorithm.

Note: The Metropolis-Hastings algorithm is computationally attractive because we can can always implement it using only the unnormalized posterior $p_\theta(X)\lambda(\theta)$ (or any function $g(\theta)$ that is proportional to it), which is often much easier to compute than the normalized posterior.

Problem 4 (Gamma-Poisson empirical Bayes)**.** Consider the Bayes model with

$$\theta_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k, \sigma), \quad i = 1, \ldots, n$$
$$X_{ij} \mid \theta_i \overset{\text{ind.}}{\sim} \text{Pois}(\theta_i), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m$$

Assume $k > 0$ (shape parameter) is known and $\sigma > 0$ (scale parameter) is unknown and estimated via the MLE. In addition, assume $\sum_{ij} X_{ij} > 0$ (though the formulae below would be basically correct in a limiting sense if the sum were zero, too).

(a) If $m = 1$, show that the empirical Bayes posterior mean for θ_i is

$$\frac{\overline{X}}{\overline{X} + k}(k + X_{i1}), \qquad \text{where } \overline{X} = n^{-1} \sum_i X_{i1}.$$

You may use without proof the fact that the marginal distribution of X_i is negative binomial.

(b) For general m, show that the empirical Bayes posterior mean for θ_i is

$$\frac{\overline{X}}{\overline{X}+k/m}(k/m+\overline{X}_i), \quad \text{where } \overline{X}_i = m^{-1}\sum_j X_{ij} \quad \text{and } \overline{X} = (nm)^{-1}\sum_{ij} X_{ij}.$$

Hint: Make a sufficiency reduction and remember that σ is a scale parameter.

Problem 5 (Gibbs Sampler for Gamma-Poisson model)**.** Consider a hierarchical Bayes model instead, where

$$\begin{aligned} \sigma^{-1} &\sim \text{Exp}(1) \\ \theta_i \mid \sigma &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(k,\sigma), \quad i=1,\dots,n \\ X_{ij} \mid \sigma,\theta &\overset{\text{ind.}}{\sim} \text{Pois}(\theta_i), \quad i=1,\dots,n, \quad j=1,\dots,m \end{aligned}$$

where σ is a scale parameter, and k, n, m, are fixed and known.

Note: For parts (b) and (c) below, be sure to read the instructions on coding problems at the top of this problem set.

(a) Give an explicit algorithm for one full iteration of the Gibbs sampler. It may be helpful to look up the inverse gamma distribution on Wikipedia.

(b) Implement the Gibbs sampler in a programming language of your choice (R is recommended since it is easy to draw random draws from standard distributions; Python or Matlab will probably also work fine). For $k = m = 3$ and $n = 100$, download the matrix $X \in \mathbb{R}^{n\times m}$, in `hw5.csv` from the course website and implement the Gibbs sampler (the standard version where you update all variables in every round; use 100 rounds of burn-in and take the next 10,000 rounds of sampling, without thinning). Make a trace plot of your draws from σ and θ_1 and include them in your homework submission. Report the following three estimators of θ_1, to three significant digits:

  (i) the hierarchical Bayes estimator (for squared error loss),

  (ii) the empirical Bayes estimator from Problem 4, and

  (iii) the UMVU estimator (in the model where θ is fixed and unknown).

(c) Next, carry out a Monte Carlo simulation to estimate the Bayes risk conditional on σ, for four estimators: (i–iii) from part (b), plus the "oracle Bayes" estimator where the value of σ is known. That is, for each estimator $\delta_1^{(\ell)}(X)$ of θ_1, approximately evaluate:

$$R^{(\ell)}(\sigma) = \mathbb{E}[(\delta_1^{(\ell)}(X)-\theta_1)^2 \mid \sigma] = \mathbb{E}\left[n^{-1}\sum_i(\delta_i^{(\ell)}(X)-\theta_i)^2 \mid \sigma\right],$$

where the expectation is taken over θ and X (but *not* σ, since we are conditioning on that). The second equality follows from the exchangeability over different values of i (you do not need to prove it yourself, but you should use it to save yourself computation). **Note:** for the hierarchical Bayes estimator, this does *not* mean you should hold σ fixed in your MCMC chain: you should compute it just as you did in part (b). Use the values $\sigma = 0.1, 0.2, 0.5, 1, 2, 5, 10$ and include a 4×7 table of risk values, each reported to at least 3 significant figures, in your answer.

For each of the three non-oracle estimators, plot the relative excess risk

$$\frac{R^{(\ell)}(\sigma)}{R^{(\text{oracle})}(\sigma)} - 1$$

against σ for the values above. Make an analogous plot for $m = 30, n = 100$ and another for $m = 3, n = 10$. I recommend using a log scale for the horizontal and vertical axis but it is not required.

Note: This exercise should not take you an absurd amount of computer time; using 100 MC runs per value of σ and the 7 values of σ above, takes my three-year-old laptop computer less than three minutes to produce each of the three plots requested above. If it is taking your computer much much longer you are probably doing something very inefficiently.

Moral: The hierarchical Bayes and oracle Bayes do almost the same thing: get a highly precise estimate for σ by pooling all n problems, and then carrying out the Bayes rule at the estimated value. They perform almost as well as the oracle Bayes rule because they are effective at figuring out what the true value of σ is. This is least true for $m = 3, n = 10$, because there simply aren't that many θ_i values from which to estimate σ. Note that from the perspective of estimating σ, increasing n has a bigger effect on the accuracy of $\hat{\sigma}$ (empirical or hierarchical Bayes estimate) than increasing m: getting to see more θ_i values helps more than just getting to estimate each one more accurately. But increasing m has a large effect on the absolute risk, because we observe each θ_i with a great deal of accuracy even before shrinking them toward the average θ value. This is also when the UMVU estimator does almost as well as oracle Bayes.