UC Berkeley
Department of Statistics
Fall 2016

STAT 210A: Introduction to Mathematical Statistics

Problem Set 5

Student: Ivana Malenica **Due:** Thursday, Oct. 13

Note: All measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, "all functions" vs. "all measurable functions," etc. are disregarded for the time being.

1. Empirical Bayes for exponential families

Consider an n-parameter exponential family model in canonical form:

$$p_\theta(x) = e^{\theta' x - A(\theta)} h(x)$$

where $x = (x_1, \ldots, x_n)$ and the random vector Θ has prior density $\lambda_\gamma(\theta)$, indexed by an unknown real hyperparameter $\gamma \in \Omega$, where $\Omega \subseteq \mathbb{R}$ is open. Let $\lambda_\gamma(\theta \mid x)$ and $q_\gamma(x)$ denote the posterior and marginal, respectively.

Let $\hat{\gamma}(X)$ denote the maximum likelihood estimator (MLE) of γ based on the observed data:

$$\hat{\gamma}(X) = \arg\max_{\gamma \in \Omega} q_\gamma(X)$$

Show that the empirical posterior mean of Θ, using $\hat{\gamma}$ to estimate γ, is

$$\mathbb{E}_{\hat{\gamma}} [\Theta \mid X = x] = \nabla \log q_{\hat{\gamma}}(x) - \nabla \log h(x).$$

Assume all relevant quantities are suitably differentiable. Hint: recall from calculus that if $f(\cdot, \cdot)$ and $g(\cdot)$ are differentiable functions then

$$\frac{d}{du} f(u, g(u)) = g'(u) \frac{\partial}{\partial v} f(u, v)\Big|_{v=g(u)} + \frac{\partial}{\partial u} f(u, v)\Big|_{v=g(u)}.$$

2. Gamma-Poisson empirial Bayes model

Consider the empirical Bayes model with

$$\Theta_i \sim \text{Gamma}(k, \sigma)$$
$$X_i \mid \Theta_i = \theta_i \sim \text{Pois}(\theta_i),$$

independently for $i = 1, \ldots, n$, and assume k (shape parameter) is known and σ (scale parameter) is unknown and estimated via the MLE. Show that the empirical Bayes posterior mean for Θ_i is

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

(Hint: you may use without proof the fact that the marginal distribution of X_i is negative binomial.)

3. Effective degrees of freedom

We can write a standard normal means model in the form

$$Y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n$$

with $\mu \in \mathbb{R}^n$ and σ^2 possibly unknown (this is a common setup for signal processing and denoising applications). If we estimate μ by some estimator $\hat{\mu}(Y)$, we can compute the residual sum of squares (RSS):

$$\text{RSS}(\hat{\mu}, Y) = \|\hat{\mu}(Y) - Y\|^2 = \sum_{i=1}^{n} (\hat{\mu}_i(Y) - Y_i)^2.$$

If we were to observe the same signal with independent noise $Y^* = \mu + \varepsilon^*$, the expected prediction error (EPE) is defined as

$$\text{EPE}(\hat{\mu}, \mu) = \mathbb{E}\left[\|\hat{\mu}(Y) - Y^*\|^2\right] = \mathbb{E}\left[\|\hat{\mu}(Y) - \mu\|^2\right] + n\sigma^2.$$

Because $\hat{\mu}$ is typically chosen to make RSS small for the observed data Y (i.e., to fit Y well), the RSS is usually an optimistic estimator of the EPE, especially if $\hat{\mu}$ tends to overfit. To quantify how much $\hat{\mu}$ overfits, we can define the *effective degrees of freedom* (or simply the *degrees of freedom*) of $\hat{\mu}$ as

$$\text{DF}(\hat{\mu}, \mu) = \frac{1}{2\sigma^2} \mathbb{E}\left[\text{EPE} - \text{RSS}\right],$$

which uses optimism as a proxy for overfitting.

For the following questions assume we also have a predictor matrix $X \in \mathbb{R}^{n \times d}$, which is simply a matrix of fixed real numbers. Suppose that $d \leq n$ and X has full column rank.

(a) Show that if $\mathbb{E}\|D\hat{\mu}(Y)\|_F < \infty$ then

$$\sum_{i=1}^{n} \frac{\partial \hat{\mu}_i(Y)}{\partial Y_i}$$

is an unbiased estimator of the DF. (Recall $D\hat{\mu}(Y)$ is the Jacobian matrix from class).

(b) Suppose $\hat{\mu} = X\hat{\beta}$, where $\hat{\beta}$ is the ordinary least squares estimator (i.e., chosen to minimize the RSS). Show that the DF is d. (This confirms that DF generalizes the intuitive notion of degrees of freedom as "the number of free variables").

(c) Suppose $\hat{\mu} = X\hat{\beta}$, where $\hat{\beta}$ minimizes the penalized least squares criterion:

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \rho\|\beta\|_2^2,$$

for some $\rho \geq 0$. Show that the DF is $\sum_{j=1}^{d} \frac{\lambda_j}{\rho + \lambda_j}$, where $\lambda_1 \geq \cdots \geq \lambda_d > 0$ are the eigenvalues of $X'X$ (counted with multiplicity) (Hint: use the SVD of X).

4. Stein's lemma for exponential families

There is a generalization of Stein's lemma to exponential family models. Consider an s-dimensional exponential family density on $\mathbb{R}$ with

$$p_\theta(x) = e^{\sum_j \theta_j T_j(x) - A(\theta)} h(x),$$

where $h(x)$ is positive and differentiable, and $T(x)$ is differentiable for all $x \in \mathbb{R}$. Suppose $g(x)$ is a differentiable function for which $\mathbb{E}|g'(X)| < \infty$ and $e^{\sum_j \theta_j T_j(x)} h(x) g(x) \to 0$ as $x \to \pm\infty$. Then show that

$$\mathbb{E}\left[\left(\frac{h'(X)}{h(X)} + \sum_j \theta_j T_j'(X)\right) g(X)\right] = -\mathbb{E}g'(X)$$

5. Likelihood ratio test for Cauchy

This question concerns hypothesis testing in the Cauchy location family:

$$p_\theta(x) = \frac{1}{\pi(1 + (x - \theta)^2)}$$

(a) Derive the likelihood ratio test for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$ (you can give the cutoff implicitly in terms of a solution to an integral).

(b) For $\theta_0 = 0, \theta_1 = 1$, and $\alpha = 0.05$, numerically compute the rejection region of the likelihood ratio test (show your code).

(c) Let θ_0, θ_1 be any two real numbers. Show that for some $\alpha^*(\theta_0, \theta_1)$, the rejection region for any $\alpha \in (0, \alpha^*)$ is a bounded interval (Note: you do not need to find an explicit expression for $\alpha^*(\theta_0, \theta_1)$ but if you are interested, recall that $\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2}$. My original statement of this problem incorrectly assumed that $\alpha^*(\theta_0, \theta_1) \geq 1/2$ when, in fact, that is never true.)

(d) It is somewhat unusual for the rejection region to be a bounded interval; give a heuristic explanation of why that is the case here. (Note: there is no unique right answer to this question; grading will be accordingly lenient).