Stats 210A, Fall 2025
Homework 6

Due on: Wednesday, Oct. 15

Instructions: See the standing homework instructions on the course web page

Problem 1 (Gamma-Poisson empirical Bayes)**.** Consider a hierarchical Bayes model with

$$\begin{aligned}
\sigma &\sim \lambda_0 \\
\theta_i \mid \sigma &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(k, \sigma) =, \quad i = 1, \ldots, n \\
X_{ij} \mid \sigma, \theta &\overset{\text{ind.}}{\sim} \text{Pois}(\theta_i) = \frac{\theta_i^x e^{-\theta_i}}{x!}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,
\end{aligned}$$

where

$$\text{Gamma}(k, \sigma) = \frac{1}{\Gamma(k)\sigma^k} x^{k-1} e^{-x/\sigma} \quad \text{on } x > 0,$$

and

$$\text{Pois}(\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \text{on } x = 0, 1, 2, \ldots.$$

Assume $k > 0$ (shape parameter) is known. We'll consider three possible ways an analyst could handle σ:

(i) a point mass on a fixed, known value σ (no learning across problem instances)

(ii) a point mass on a fixed, estimated value $\hat{\sigma}^{-1}$, estimated by maximum likelihood (empirical Bayes)

(iii) a hyperprior $\sigma^{-1} \sim \text{Exp}(1)$ (hierarchical Bayes)

All Bayes estimators will be calculated with respect to squared error loss (i.e., use the posterior mean).

Here, the maximum likelihood estimator of σ is the estimator that maximizes $p_\sigma(X)$, where we implicitly marginalize over the latent parameters $\theta_1, \ldots, \theta_n$ (called *random effects* in the frequentist model). You may use that the distribution of X_{ij} given σ is negative binomial:

$$X_{ij} \mid \sigma \sim \text{NB}\left(k, \frac{1}{\sigma + 1}\right),$$

where the $\text{NB}(k, p)$ distribution has pmf

$$\binom{k + x - 1}{x} (1 - p)^x p^k, \quad \text{on } x = 0, 1, 2, \ldots,$$

mean $\frac{k(1-p)}{p}$, and variance $\frac{k(1-p)}{p^2}$.

Note: X_{i1} and X_{i2} are *not* independent conditional on σ because they are correlated through θ_i.

(a) First, consider the estimator with fixed and known σ. Show that the Bayes estimator for θ_i is

$$\delta_{i,\sigma}(X) = \frac{\sigma}{m\sigma + 1}\left(k + \sum_j X_{ij}\right) = \frac{\overline{X}_i + k/m}{1 + (\sigma m)^{-1}}, \quad \text{where } \overline{X}_i = m^{-1} \sum_j X_{ij}.$$

(b) 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: To calculate the MLE, it may help to start with $m = 1$, then make a sufficiency reduction remembering that if $Y \sim \text{Gamma}(k, \sigma)$ then $cY \sim \text{Gamma}(k, c\sigma)$ (σ is a scale parameter).

You may assume $\sum_{ij} X_{ij} > 0$ (though the formulae below would be basically correct in a limiting sense if the sum were zero, too).

(c) Now, consider the hierarchical Bayes problem where $\sigma^{-1} \sim \text{Exp}(1)$. Give an explicit algorithm for one full iteration of the Gibbs sampler, with closed-form updates. It may be helpful to look up the inverse gamma distribution on Wikipedia.

(d) 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 `gibbspoisson.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, and

(iii) the UMVU estimator for θ_1 in the model where θ is fixed and unknown.

(e) 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 as an analyst who did not know the value of σ would, and just as you did in part (d).

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 analogous plots for the scenarios $(m, n) = (30, 100)$, $(m, n) = (3, 10)$, and $(m, n) = (30, 10)$. I recommend using a log scale for the horizontal and vertical axis.

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.

Problem 2 (Effective degrees of freedom)**.** We can write a standard Gaussian sequence model in the form

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

with $\mu \in \mathbb{R}^n$ and $\sigma^2 > 0$ possibly unknown. 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

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

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

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

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 $\hat{\mu}$ is differentiable with $\mathbb{E}_\mu \|D\hat{\mu}(Y)\|_F < \infty$ 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 singular value decomposition of X).

Moral: When we do estimation with no shrinkage or other regularization, there is a real sense in which just counting the number of free parameters we estimate gives us a useful picture of how hard our estimator has fit (or overfit) to the data. For estimators that do a lot of regularization, however, naive parameter counting is not a good measure of overfitting. In this context, the effective degrees of freedom as defined above is a more natural generalization of the parameter dimension.

Problem 3 (Soft thresholding)**.** Consider the *soft thresholding operator* with parameter $\lambda \geq 0$, defined as

$$\eta_\lambda(x) = \begin{cases} x - \lambda & x > \lambda \\ 0 & |x| \leq \lambda \\ x + \lambda & x < -\lambda \end{cases}$$

Note that, although we didn't prove it in class, Stein's lemma applies for continuous functions $h(x)$ which are differentiable except on a measure zero set; you can apply it here without worrying.

Assume $X \sim N_d(\theta, I_d)$ for $\theta \in \mathbb{R}^d$, which we will estimate via $\delta_\lambda(X) = (\eta_\lambda(X_1), \ldots, \eta_\lambda(X_d))$. Soft thresholding is sometimes used when we expect *sparsity*: a small number of relatively large θ_i values. λ here is called a *tuning parameter* since it determines what version of the estimator we use, but doesn't have an obvious statistical interpretation.

(a) Show that $|\{i : |X_i| > \lambda\}|$ is an unbiased estimator of the degrees of freedom of δ_λ (so, in a sense, the DF is the expected number of "free variables").

(b) Show that

$$d + \sum_i \min(X_i^2, \lambda^2) - 2\,|\{i : |X_i| \leq \lambda\}|$$

is an unbiased estimator for the MSE of δ_λ.

(c) Show that, if some $\theta_i \neq 0$, the risk-minimizing value λ^* solves

$$\lambda \sum_i \mathbb{P}_{\theta_i}(|X_i| > \lambda) = \sum_i \phi(\lambda - \theta_i) + \phi(\lambda + \theta_i),$$

where $\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$ is the standard normal density.

Hint: To show that there is a minimum in $(0, \infty)$, it may help to recall the Gaussian tail bound

$$\left(\frac{1}{z} - \frac{1}{z^3}\right)\phi(z) \leq \mathbb{P}(Z > z) \leq \frac{1}{z}\phi(z),$$

for $Z \sim N(0,1)$. It might also help to show that $\frac{\phi(\lambda-\theta_2)}{\phi(\lambda-\theta_1)} \to 0$ as $\lambda \to \infty$, if $\theta_1 > \theta_2$.

(d) Consider a problem with $\theta_1 = \cdots = \theta_{20} = 10$ and $\theta_{21} = \cdots = \theta_{500} = 0$. Compute λ^* numerically. Then simulate a vector X from the model and use it to automatically tune the value of λ by minimizing SURE. Call the automatically tuned value $\hat{\lambda}(X)$ and report both λ^* and $\hat{\lambda}(X)$. Finally plot the true MSE of δ_λ along with its SURE estimate against λ for a reasonable range of λ values. Add a horizontal line for the risk of the UMVU estimator.

(e) Compute and report the squared error loss $\|\delta(X) - \theta\|^2$ for the following four estimators:

(i) the UMVU estimator $\delta_0(X) = X$,

(ii) the optimally tuned soft-thresholding estimator $\delta_{\lambda^*}(X)$,

(iii) the automatically tuned soft-thresholding estimator $\delta_{\hat{\lambda}(X)}(X)$, and

(iv) the James-Stein estimator.

You do not need to compute the MSE. Intuitively, what do you think accounts for the good performance of soft-thresholding in this example?

Moral: SURE gives us a reasonable way of selecting a tuning parameter for estimation problems, and can help us choose a tuning parameter that achieves the near optimal performance. Also, regularization methods that set a lot of parameters to zero can substantially reduce the MSE in sparse problems, by eliminating all the variance for most of the coordinates.

Problem 4 (Shrinking toward the average)**.** Assume we observe data from a Gaussian sequence model $X \sim N_d(\theta, I_d)$ with $d \geq 4$, and we want to estimate $\theta \in \mathbb{R}^d$ with low mean-squared error loss. Instead of shrinking toward zero, however, we want to shrink toward $\overline{X}$. This implements an inductive bias that the θ_i values should be close to each other, as opposed to assuming they should be close to zero.

We can use the estimator whose ith coordinate is

$$\delta_{gamma,i}(X) = \gamma\overline{X} + (1-\gamma)X_i = \overline{X} + (1-\gamma)(X_i - \overline{X}),$$

leading to

$$\delta_\gamma(X) = \overline{X}1_d + (1-\gamma)(X - \overline{X}1_d),$$

where $1_d = (1, 1, \ldots, 1) \in \mathbb{R}^d$. The course reader calculated the SURE for this model when we have a fixed γ.

We will instead consider a popular version of the James–Stein estimator, which uses an adaptive choice

$$\hat{\gamma}(X) = \frac{d-3}{\|X - \overline{X}1_d\|^2} = \frac{d-3}{\sum_i (X_i - \overline{X})^2},$$

leading to

$$\delta_{\text{JS}_2}(X) = \overline{X}1_d + \left(1 - \frac{d-3}{\|X - \overline{X}1_d\|^2}\right)(X - \overline{X}1_d)$$

(a) As with the previous James–Stein estimator, we can motivate this estimator in a similar way by empirical Bayes in a model with $\theta_i \overset{\text{i.i.d.}}{\sim} N(\mu, \tau^2)$. If we want we can write $\zeta = (1+\tau^2)^{-1}$ as before. Show that δ_{JS_2} is the empirical Bayes estimator for this prior, where we estimate the hyperparameters (μ, ζ) by UMVU.

(b) Derive an unbiased estimator for the risk $\text{MSE}(\theta; \delta_{\text{JS}_2})$. Your estimator should be a function of the data X, and should not involve any unknown parameters like μ, ζ, or θ.

(c) Find an expression for the MSE of δ_{JS_2} as a function of θ, and show that it dominates the MSE of $\delta_0(X) = X$ for all $\theta \in \mathbb{R}^d$. Evaluate your expression in the case where $\theta_1 = \theta_2 = \cdots = \theta_d$.

(d) Now make a change of variables to $Z = Q'X$, where $q_1 = d^{-1/2}1_d$ and $q_2, \ldots, q_d$ are any completion of an orthonormal basis for $\mathbb{R}^d$, and $Q = [q_1 \cdots q_d] \in \mathbb{R}^{d \times d}$. Show that

$$Z \sim N_d(\mu, I_d), \quad \text{where } \mu_1 = d^{-1/2}\sum_i \theta_i.$$

Show that $Q'\delta_{\text{JS}_2}(X)$, as an estimator of μ, could be characterized as estimating μ_1 as Z_1 (without any shrinkage), and estimating $\mu_{-1} = (\mu_2, \ldots, \mu_d)$ via the original James–Stein estimator on the $(d-1)$-variate normal $Z_{-1} \sim N_{d-1}(\mu_{-1}, I_{d-1})$. Use this construction to re-derive the results in part (c).

Moral: If we think of the James–Stein estimator as implementing an *inductive bias* where we believe $\|\theta\|$ to be small, then different variants of the James-Stein estimator allow for implementing different inductive biases, for example that the θ_i values should be near each other, or (in a later problem) that they should lie close to a regression line $w_i'\beta$ for covariates w_i.