Stats 210A, Fall 2024
Homework 6

Due on: Wednesday, Oct. 16

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

Problem 1 (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 2 (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 3 (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) **Optional:** (Not graded, no extra points) If we make a change of variables to a certain $Z = f(X)$ with $Z \sim N_d(\mu, I_d)$, then δ_{JS_2} 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 $Z_{-1} \sim N_{d-1}(\mu_{-1}, I_{d-1})$. Find such a transformation f and use this construction to repeat part (c).

Problem 4 (Tweedie's formula)**.** Besides James–Stein, another well-known empirical Bayes method is *Tweedie's formula* for doing Bayes estimation of natural parameters in exponential family models.

Assume that the data come from a common 1-parameter exponential family with a different parameter for each observation:

$$X_i \overset{\text{ind.}}{\sim} p_{\eta_i}(x) = e^{\eta_i x - A(\eta)} h(x),$$

Additionally, assume $\eta_i \overset{\text{i.i.d.}}{\sim} \lambda(\eta)$ where λ is an unknown density on $\mathbb{R}$ (so this is a non-parametric model for the prior). Define the marginal

$$q(x) = \int p_\eta(x)\lambda_0(\eta),$$

(a) Show that the posterior distribution $\lambda(\eta_i \mid x_i)$ follows a one-parameter exponential family model with sufficient statistic η_i and normalizing constant $B(x_i) = \log(q(x_i)/h(x_i))$.

(b) Use part (a) to find the Bayes posterior mean of η_i given X_i.

Moral: There are a variety of methods (beyond the scope of this course) to obtain nonparametric density estimators for the marginal density $q(x)$ when we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} q$. This problem shows that such an estimator leads directly to *nonparametric* empirical Bayes estimators for η_i.