Modern Statistics 52311, 2020-21
Homework 2 - Multiple Hypothesis Testing, James-Stein Estimator

April 24, 2021

The grade for entire exercise is the sum of the 3 highest scores for individual questions - a complete and correct solution for each question is worth 33.3% of this exercise grade. To get full credit for a question, you need to solve it correctly and completely, and explain your answer clearly. For the computerized questions, you need to supply plots, explanations and your code.

You may submit in groups of size two or less. Submit your solutions by 6/5/2020.

Problems

1. In the next two questions we derive the risk of the Bayes and James-Stein estimators for the means of independent Gaussian random variables.

   Let $x_i \sim N(\mu_i, 1)$, $i = 1, .., n$ be independent Gaussian random variables (in matrix form: $x \sim N(\mu, I_n)$.) We use the squared loss. Recall that the risk under this loss of the MLE estimator $\hat{\mu}^{(MLE)} = x$ is n. For a given prior $P(\beta)$ define the Bayes estimator as the posterior mean, $\hat{\mu}^{(Bayes)} = E[\beta|x]$.

   (a) Suppose that the parameters μ_i have i.i.d. prior $\mu_i \sim N(0, \sigma^2)$. Prove that the Bayes estimator is $\hat{\mu}^{(Bayes)} = \frac{\sigma^2}{\sigma^2+1}x$.

   (b) Prove that if the true parameters vector is μ, then the risk of $\hat{\mu}^{(Bayes)}$ is:

   $$R_\mu(\hat{\mu}^{(Bayes)}) = \Big[1 - \frac{\sigma^2}{\sigma^2+1}\Big]^2 ||\mu||^2 + n\Big[\frac{\sigma^2}{\sigma^2+1}\Big]^2 \tag{1}$$

   (c) Suppose that the parameters μ_i have i.i.d. prior $\mu_i \sim N(0, \sigma^2)$. Prove that the overall Bayes risk of the Bayes estimator is:

   $$R^{(Bayes)}(\hat{\mu}^{(Bayes)}) = E_\mu\Big[R^{(Bayes)}(\mu)\Big] = \frac{\sigma^2}{\sigma^2+1}n \tag{2}$$

(d) Show that for *any* estimator (with finite moments) $\hat{\mu}$ we have:

$$E[||\hat{\mu} - \mu||^2] = E[||\hat{\mu} - x||^2] - n + 2\sum_{i=1}^{n} COV(x_i, \hat{\mu}_i) \tag{3}$$

2. Under the conditions of question 1, and assuming $n \geq 3$, define the James-Stein estimator $\hat{\mu}^{(JS)} = (1 - \frac{n-2}{||x||^2})x$.

   (a) Use integration by parts and the multivariate Gaussian density to show that for any continuously differentiable estimator (with finite moments):

   $$COV(x_i, \hat{\mu}_i) = E[\frac{\partial \hat{\mu}_i}{\partial x_i}] \tag{4}$$

   Hint: Here integration by parts for a vector function may look a bit different then the scalar formula you are used to. Consider two differentiable functions $f(x), g(x) : \mathbb{R}^p \to \mathbb{R}$. Then we have $\forall i = 1, .., p$:

   $$\int_x f(x)\frac{\partial g(x)}{\partial x_i}dx = [f(x)g(x)] - \int_x \frac{\partial f(x)}{\partial x_i}g(x)dx. \tag{5}$$

   (b) Use the previous result and 1d to show that the risk of the James-Stein estimator for any given parameter μ is:

   $$R_\mu(\hat{\mu}^{(JS)}) = n - E\Big[\frac{(n-2)^2}{||x||^2}\Big]\Big(< n = R_\mu(\hat{\mu}^{(MLE)})\Big) \tag{6}$$

   (c) Suppose that the parameters μ_i have i.i.d. prior $\mu_i \sim N(0, \sigma^2)$. Prove that the overall Bayes risk of the James-Stein estimator is:

   $$R^{(Bayes)}(\hat{\mu}^{(JS)}) = E_\mu\Big[R_\mu(\hat{\mu}^{(JS)})\Big] = \frac{n\sigma^2 + 2}{\sigma^2 + 1} \tag{7}$$

3. This question is meant to explain the empirical Bayes approach to FDR in a simplified manner.

   Assume that we test m hypotheses, each with a z-score test statistic X_i, such that the statistics have the following joint Gaussian distribution:

   $$X \sim (\vec{\mu}, \Sigma) \tag{8}$$

   with $\mu_i = 0$ for $i = 1, .., m_0$ (corresponding to H_0) and $\mu_i = \mu$ for $i = m_0 + 1, .., m$ (corresponding to H_1), and with $\Sigma_{ii} = 1, \Sigma_{ij} = \rho$ for $i \neq j$.

   Assume that we perform a one-sided test, and for each test statistic we reject the null hypothesis if $X_i \geq C$, where C is determined by the procedure.

Simulate tests statistics $X_1, .., X_m$ from the above joint distribution with the following parameters: $m = 5000, m_0 = 4000, \mu = 2.5, \rho = 0$.

Assume a prior of $\pi_0 = 0.8$ for each hypotheses to be true null, and imagine that we estimated from the data the distribution of the z-scores under the null and alternative hypotheses to be $F_0 = N(0, 1)$ and $F_1 = N(2.5, 1)$ respectively (in reality we'll have to estimate π_0 and F_1 from the z-values themselves).

(a) Draw a histogram of the X_i with 100 equally spaced bins. Draw on the same plot the densities for $\pi_0 F_0, (1-\pi_0)F_1$ and the mixture $\pi_0 F_0 + (1-\pi_0)F_1$, scaled to match the number of observations (i.e. the areas under the histogram, F_0, F_1 and $\pi_0 F_0 + (1-\pi_0)F_1$ should all be the same). Explain your results.

(b) Give an expression for the $FDR(z)$ and the local $fdr(z)$ as a function of z. Plot for $z \in [-4, 4]$ the expressions for both the $FDR(z)$ and local $fdr(z)$ a function and compare them. Explain your results.

(c) Convert each X_i to a right-tail p-value P_i (assuming X_i is a z-score), and compute the q-value for each P_i - this is the minimal α for which the BH procedure will reject the i-th hypothesis. Plot on the same plot the values $\pi_0 \times q_i$ as a function of the X_i's (again, treating them as z-values) and compare to the empirical Bayes calculation of $FDR(z)$. Explain your results.

4. In this question we study empirically the effect of dependency on the FDR. Simulate test statistics from the same model as in question 3, but with parameters $m = 1000, m_0 = 750, \mu = 2.5$ and different values of ρ (to be specified in the sub-questions). For each set of parameters simulate $N = 5,000$ independent realizations of all m test statistics, and for each realization, apply the BH procedure with parameter $\alpha = 0.1$ and record the number of rejections R and false rejections V.

(a) Draw a histogram of the total number of rejections, R for $\rho = 0, 0.95$. Explain your results.

(b) Draw a histogram of the total number of false rejections, V for $\rho = 0, 0.95$. Explain your results.

(c) Draw a histogram of the false discovery proportion, $Q = \frac{V}{R^+}$ for $\rho = 0, 0.95$. Explain your results.

(d) For each $\rho = 0, 0.05, 0.1, .., 0.95, 1$ simulate N realizations and estimate the mean (FDR) and standard deviation of $Q = \frac{V}{R^+}$.

Plot the estimators you got as a function of ρ . Compare the resulting estimated FDR to the theoretical guarantees according on the BH procedure we learned in class. Explain your results for both the mean and variance.

Instructions: one way to simulate positively correlated Gaussians is as follows:

First, simulate a set of independent Gaussians, $Y_0, Y_1, ...Y_m \sim N(0,1)$.
Then, for a parameter $0 \leq \rho \leq 1$ and for all $i = 1, .., m$ set $X_i = \rho Y_0 + (1-\rho)Y_i + \mu_i$.