1 Chapter 1: Empirical Bayes & the James-Stein Estimator

1.1 Bayes Rule & Multivariate Normal Estimation

Suppose that we wish to make inferences about the mean parameter μ following the prior distribution $\mu \sim g(\cdot)$. Then, after processing a sample z, we assume that $z|\mu \sim f_{\mu}(z)$. That is z follows some distribution that is parametrized by μ. With this information we can form the posterior distribution as follows

$$g(\mu|z) = \frac{f_{\mu}(z)g(\mu)}{f(z)} \quad \text{where} \quad f(z) = \int f(z,\mu)d\mu = \int f_{\mu}(z)g(\mu)d\mu \tag{1}$$

The most difficult part of this calculation is calculating the marginal distribution of z, $f(z)$. However, we can usually avoid this by seeing that $g(\mu|z) \propto f_{\mu}(z)g(\mu)$ and finding characterizing factors of the posterior distribution. Moreover, we note that the odds ratio is invariant to the marginal distribution of the data $f(z)$

$$\frac{g(\mu_1|z)}{g(\mu_2|z)} = \frac{g(\mu_1)}{g(\mu_2)}\frac{f_{\mu_1}(z)}{f_{\mu_2}(z)} \tag{2}$$

If we specifically focus on the example $\mu_i \sim N(0, A)$ and $z_i|\mu_i \sim N(\mu_i, 1)$ then we see that for the vector version of these quanities we have

$$\vec{\mu} \sim \mathcal{N}(0, AI) \qquad \vec{z}|\vec{\mu} \sim \mathcal{N}(\vec{\mu}, I) \tag{3}$$

We can find the posterior distribution of $\vec{\mu}|\vec{z}$ as follows. First we start with a single element of this vector.

$$\begin{aligned} g(\mu_i|z_i) &\propto \frac{1}{\sqrt{2\pi}\sqrt{A}}\exp\left\{-\frac{1}{2A}\mu_i^2\right\} \times \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(z_i-\mu_i)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left\{\left(\frac{A+1}{A}\right)\mu_i^2 - 2z_i\mu_i\right\}\right\} \end{aligned}$$

From this we see that

$$\mathbb{E}(\mu_i|z_i) = \frac{A}{A+1}z_i \qquad \text{Var}(\mu_i|z_i) = \frac{A}{A+1} \tag{4}$$

For easy of notation, let $B = \frac{A}{A+1}$ then we can write the full posterior distribution of $\vec{\mu}|\vec{z}$ as follows

$$\vec{\mu}|\vec{z} \sim \mathcal{N}(B\vec{z}, BI) \tag{5}$$

Our goal is to now compare this estimator $\hat{\mu}^{(Bayes)} = B\vec{z}$ against the maximum likelihood estimate $\hat{\mu}^{(MLE)} = \vec{z}$. Well considering the L^2-Risk and Bayes risk for the MLE we see that

$$\begin{aligned} R^{(MLE)} &= \mathbb{E}_{\mu}||\hat{\mu}^{(MLE)} - \mu|| = \mathbb{E}_{\mu}\left[\sum_{i=1}^{n}\left(\hat{\mu}_i^{(MLE)} - \mu\right)^2\right] = \sum_{i=1}^{n}\text{Var}(\hat{\mu}_i^{(MLE)}) = N \\ BR^{(MLE)} &= \mathbb{E}_{\pi}[R(\hat{\mu}^{(MLE)})] = \mathbb{E}_{\pi}[N] = N \end{aligned}$$

Now calculating the same quantities for the Bayes-Estimator we have

$$\begin{aligned}
\text{Bias}(\hat{\mu}^{(Bayes)}) &= \mathbb{E}_{\mu}[B\vec{z}] - \vec{\mu} = B\vec{\mu} - \vec{\mu} = (B-1)\vec{\mu} \\
\text{Var}_{\mu}(\hat{\mu}^{(Bayes)}) &= B^2\text{Var}(\vec{z}|\vec{\mu}) = B^2 I \\
R^{(Bayes)} &= \sum_{i=1}^{n}\left\{(B-1)^2\mu_i^2 + B^2\right\} = (B-1)^2||\vec{\mu}||^2 + NB^2 \\
BR^{(Bayes)} &= \mathbb{E}\left\{(B-1)^2||\vec{\mu}||^2 + NB^2\right\} = (B-1)^2\sum_{i=1}^{n}\text{Var}(\mu_i) + NB^2 = N\frac{A}{A+1}
\end{aligned}$$

The key point to consider here is the fact that $\frac{BR^{(Bayes)}}{BR^{(MLE)}} = \frac{A}{A+1} < 1$ for all values of N. That is the Bayes estimator out preforms the classic MLE for any sample size.

1.2 Empirical Bayes Estimation

One fair objection to this conclusion, and most Bayesian inference, is the fact that this estimator was assumed to have known prior variance - a fact that is rarely known in practice. However, if we could instead replace this known assumption with an estimate informed by the sample, we could remove some of these objections.[1] This the central task of *Empirical Bayes Estimation.*

Assuming that $\vec{z}|\vec{\mu}$ and $\vec{\mu}$ follow the previous discussed distribution then

$$\vec{z} \sim \mathcal{N}(0, (A+1)I) \tag{6}$$

Now, if we define $S = ||\vec{z}||^2$ then $S \sim (A+1)\chi_n^2$. From here one can show that $\frac{1}{S}$ has an inverted χ^2 distribution and by normalizing by $N-2$ we see that

$$\mathbb{E}\left[\frac{N-2}{S}\right] = \frac{1}{A+1} \tag{7}$$

Now, using this quantity we can define an estimator that uses the Bayes-estimator but instead estimates the prior parameters. This estimator is called the James-Stein estimator. We give it and compare it to the original Bayes estimate

$$\hat{\mu}^{(Bayes)} = \frac{A}{A+1}\vec{z} = \left(1 - \frac{1}{A+1}\right)\vec{z} \qquad \hat{\mu}^{(JS)} = \left(1 - \frac{N-2}{S}\right)\vec{z} \tag{8}$$

The name “Empirical” Bayes here refers to the use of empirical estimate of the Bayesian estimator. One can show that this estimator has L^2-Risk

$$R^{(JS)} = N\frac{A}{A+1} + \frac{2}{A+1} \tag{9}$$

This is of course larger than the Bayesian estimator but the penalty is only of order $O_p(\frac{1}{N})$. That is

$$\frac{R^{(JS)}}{R^{(Bayes)}} = 1 + \frac{2}{NA} \tag{10}$$

While these results are quite impressive in their own right, a more general result was given by James & Stein where $\vec{\mu}$ was allowed to be centered around *any* quantity.

[1] Not all however. We will still assume the parameters follow a certain parametric distribution and simply estimate this distributions parameters.

Theorem 1. *For $N \geq 3$ the James-Stein estimator dominates the MLE $\hat{\mu}^{(MLE)}$ in terms of expected L^2 loss. That is*

$$\mathbb{E}_{\mu}\left\{||\hat{\mu}^{(JS)} - \mu||^2\right\} < \mathbb{E}_{\mu}\left\{||\hat{\mu}^{(MLE)} - \mu||^2\right\} \tag{11}$$

for every choice of $\vec{\mu}$.

The strength of the result states that regardless of one's prior beliefs of the parameter, the James-Stein estimator dominates the MLE for moderately small sample sizes (i.e. $N \geq 3$). While this specific version of the JS estimator shrinks the data z_i toward zero we don't necessarily need to choose this as a shrinking point. A more general analysis of this problem is focused on the following generalized prior and likelihood functions

$$\mu_i \overset{ind.}{\sim} N(M, A) \qquad z_i|\mu_i \overset{ind.}{\sim} N(\mu_i, \sigma_0^2) \tag{12}$$

which results in the following marginal distribution and posterior distribution

$$z_i \overset{ind.}{\sim} N(M, A + \sigma_0^2) \qquad \mu_i|z_i \overset{ind.}{\sim} N(M + B(z_i - M), B\sigma_0^2) \tag{13}$$

where $B = \frac{A}{A+\sigma_0^2}$. Here we see that the posterior is shrinking the data z_i towards the prior mean M. From here, the empirical James-Stein estimator is given by

$$\hat{\mu}_i^{(JS)} = \overline{z} + \left(1 - \frac{(N-3)\sigma_0^2}{S}\right)(z_i - \overline{z}) \tag{14}$$

Here we see that instead of shrinking the data z_i towards zero, we instead center the estimate around $\overline{z}$ and shrink the deviation of the data from this sample mean $(z_i - \overline{z})$ towards the mean. Here, theorem 1 also holds but instead we require $N \geq 4$.

An interesting note to be made here is the independence assumption. Note that for the general JS estimator, we see that each estimate is centered around the sample mean $\overline{z}$, *even though the observations are independent.* In Bayesian estimation, we assume that all of these samples come from a common prior distribution implying that there is evidence of an indirect dependence structure.

1.3 Learning from the Experience of Others

This general method of shrinkage can be extended past just shrinkage towards the sample mean. Instead we can choose to shrink towards an original estimate. For instance, if we look to center our estimate around a regression estimate we can consider the model

$$\mu_i \overset{ind.}{\sim} N(\beta^T x_i, A) \qquad z_i \sim N(\mu_i, \sigma_0^2) \tag{15}$$

the James-Stein estimator becomes

$$\hat{\mu}_i^{(JS)} = \hat{\beta}^T x_i + \left(1 - \frac{(N-4)\sigma_0^2}{S}\right)(z_i - \hat{\beta}^T x_i) \tag{16}$$

Here we see that we can introduce a dependence structure that attempts to the model the indirect dependence through the covariates. This can be see as extending the "borrowing strength from the mean" idea to "borrowing strength from the other estimate."

1.4 Empirical Bayes Confidence Intervals

Based on the posterior distribution we can construct reliability intervals for a given entry μ_i. In particular $\mu_0|z_0 \sim N(Bz_0, B)$ which corresponds to the interval

$$\mu_0 \in Bz_0 \pm z_{\alpha/2}\sqrt{B} \tag{17}$$

For the Empirical Bayes estimate, however, we estimate A and by extension B. For this reason, we need to include the variability in $\hat{B}$ in the size of the interval. An accurate interval can be given as follows

$$\mu_0 \in \hat{B}z_0 \pm z_{\alpha/2}\left\{\hat{B} + \frac{2}{N-2}\left[z_0(1-\hat{B})\right]^2\right\}^{1/2} \tag{18}$$