16.1 Outline

Agenda: e-BH procedure; Estimation of a multivariate normal mean.

1. FDR control with e-values (e-BH)
2. Stein’s Phenomenon
3. The James-Stein estimator
4. Stein’s Unbiased Risk Estimate (SURE)

16.2 FDR control with e-values (e-BH)

Last week, we talked about how to construct e-values and the advantages of e-values when dealing with certain problems. The main advantages of e-values include:

- Validity for arbitrary dependence (expectations are easier)
- Validity for optional stopping time (with martingale theories)
- Easy to combine
- Flexible to stop/continue (online; unfixed sample size)
- Non-asymptotic and often model-free

One example of using e-values is to test multiple hypotheses in multi-armed bandit problems with $K$ arms, where the null hypothesis $k$ is that arm $k$ has a mean reward at most 1. In such problems, at time $t \geq 1$, one would pull arm $k_t$ and obtain an i.i.d. reward $X_{k_t,t} \geq 0$, and the aim is to quickly detect arms with mean $> 1$, to maximize profit, or to minimize regret. There is usually a complicated dependence structure due to exploration/exploitation, so classical ways of dealing with the tests are non-trivial. However, it is easy to construct e-values by considering the running reward $M_{k,t} = \prod_{j=1}^{t} X_{k_j,j} \mathbb{1}\{k_j=k\}$, and $M_{1,\tau}, \ldots, M_{K,\tau}$ would be e-values for any stopping time $\tau$.

To see how dealing with expectation is considerably easier than dealing with tail probability, consider the following E-BH procedure proposed by [3].
Definition 1. Let \( e_1, \ldots, e_n \) be the \( n \) realized e-values associated to \( H_1, \ldots, H_n \), and denote the order statistics as \( e_{(1)} \geq \cdots \geq e_{(n)} \). To control the FDR at level \( \alpha \in (0, 1) \), the E-BH procedure reject hypotheses with the largest \( \hat{k} \) e-values, where

\[
\hat{k} = \max \left\{ i : \frac{i e_{(i)}}{n} \geq \frac{1}{\alpha} \right\}.
\]

This is almost the BH procedure, except that now \( 1/e_{(i)} \) plays the role of \( p_{(i)} \). It turns out that the e-BH procedure also controls the FDR under arbitrary dependence.

**Theorem 1** (Wang & Ramdas ’20). The e-BH procedure has FDR at most \( n_0 \alpha / n \).

**Proof.** With notations from previous lectures,

\[
FDP = \frac{V}{R \vee 1} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{R \vee 1}.
\]

If \( R = 0 \), \( FDP = 0 \), so it suffices to study the case where \( R > 0 \). For any \( i \) rejected, by definition of the e-BH procedure,

\[
\frac{1}{R} \leq \frac{1}{i} \leq \frac{\alpha e_{(i)}}{n}.
\]

Then

\[
FDP \leq \sum_{i \in \mathcal{H}_0} \frac{V_i}{n} \leq \sum_{i \in \mathcal{H}_0} \frac{\alpha e_{(i)}}{n} = \frac{\alpha}{n} \sum_{i \in \mathcal{H}_0} e_{(i)},
\]

and it follows that

\[
FDR = \mathbb{E}[FDP] \leq \frac{\alpha}{n n_0}.
\]

Note that we don’t require independence for the proof to work, but the procedure might not be as powerful as the original BH procedure.

In summary, e-values allow for safe (i.e., Type I error guarantee) testing under optional continuation, but using an e-value is less likely to produce a “rejection”, which leads to underperformance when a 0-1 decision needs to be made.

### 16.3 Stein’s Phenomenon

#### 16.3.1 Estimation of a multivariate normal population

Suppose we are interested in estimating the mean \( \mu \) in the \( p \)-dimensional multivariate normal model

\[
X \sim N_p(\mu, \sigma^2 I),
\]

which can be equivalently written as

\[
X_i = \mu_i + \sigma z_i, \quad z_{i.i.d} \sim N(0, 1), \quad i = 1, \ldots, p.
\]
Note that $X_i$ can also be sample averages, e.g., $X_i = \bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$. Our primary focus is to find an estimator $\hat{\mu}$ that performs well in terms of quadratic loss, defined as

$$l(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|_2^2 = \sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2,$$

which is a random variable that depends on the data. The corresponding risk function, usually called MSE, is defined as the expected loss (a function of $\mu$) and is given by

$$R(\hat{\mu}, \mu) = \mathbb{E}_\mu \|\hat{\mu} - \mu\|_2^2 = \mathbb{E}_\mu l(\hat{\mu}, \mu).$$

A natural estimator of $\mu$ is the MLE $\hat{\mu}_{\text{MLE}} = X$ (the sample mean), which has risk

$$R(\hat{\mu}_{\text{MLE}}, \mu) = \mathbb{E}_\mu \|X - \mu\|_2^2 = \sigma^2 \mathbb{E} \|z\|_2^2 = p\sigma^2.$$

For a long time, the MLE was thought to be ‘the best’ estimate for a multivariate mean. No estimator achieving a lower MSE for all values of $\mu$ was believed to exist.

Note: It is not difficult to improve on the MLE at a single point; e.g., the estimator $\hat{\mu} = 0$ outperforms $\hat{\mu}_{\text{MLE}}$ at $\mu = 0$. Rather, we are interested in knowing whether MSE is an admissible estimator, i.e., whether there exists another estimator $\hat{\mu}$ such that $R(\hat{\mu}, \mu) \leq p\sigma^2$ for all $\mu$, and $R(\hat{\mu}, \mu) < p\sigma^2$ for some $\mu$.

### 16.3.2 The James-Stein estimator

For $p = 1, 2$, this belief that the MLE is admissible is correct. However, for $p \geq 3$, it is no longer correct. A result of Stein 1956 hinted at this. A proof was eventually provided in 1961 by James & Stein [1].

**Definition 2.** The James-Stein estimator is defined as

$$\hat{\mu}_{\text{JS}} = \left[1 - \frac{p - 2}{\|X\|^2}\right] X.$$

This estimator is nonlinear, biased, and shrinks the MLE towards 0.

**Theorem 2** (James, Stein 1961). $\hat{\mu}_{\text{JS}}$ dominates the MLE everywhere in terms of MSE. More precisely, for all $\mu \in \mathbb{R}^p$,

$$\mathbb{E}_\mu \|\hat{\mu}_{\text{JS}} - \mu\|^2 < \mathbb{E}_\mu \|\hat{\mu}_{\text{MLE}} - \mu\|^2.$$

Note that the James-Stein estimator combines independent information. This estimator would only improve the overall estimation accuracy, but not the estimation accuracy for every single $\mu_i$.

This result proves the inadmissibility of the sample mean as an estimator of the mean for $p \geq 3$. However, the James-Stein estimator is not admissible either.
16.3.3 Stein’s original argument (1956)

In Stein’s original work, he argued that a good estimate should obey \( \hat{\mu}_i \approx \mu_i \) for every \( i \). Thus we should also have \( \hat{\mu}_i^2 \approx \mu_i^2 \), which further implies that \( \sum_i \hat{\mu}_i^2 \approx \sum_i \mu_i^2 \).

Consider the estimator \( \hat{\mu}_{\text{MLE}} = X \). For this estimator, we have

\[
\mathbb{E} \sum_i X_i^2 = \mathbb{E} \left[ \sum_i (\mu_i + \sigma z_i)^2 \right]
= \sum_i (\mu_i^2 + \sigma^2)
= \|\mu\|^2 + \sigma^2 p.
\]

This suggests that for large \( p \), \( \|X\|^2 \) is likely to be considerably larger than \( \|\mu\|^2 \), and hence we may be able to obtain a better estimator by shrinking the estimator \( \hat{\mu}_{\text{MLE}} \) towards 0. (See Figure 16.1 for a graphical illustration.)

In [1], the authors considered a family of estimators indexed by \( c \) of the form

\[
\hat{\mu}_c = \left( 1 - c \frac{\sigma^2}{\|X\|^2} \right) X,
\]

and showed that for all \( c \in (0, 2(p-2)) \), \( R(\hat{\mu}_c, \mu) < R(\hat{\mu}_{\text{MLE}}, \mu) \).
16.4 Stein’s Unbiased Risk Estimate (SURE)

16.4.1 Stein’s identity

Before we prove the James-Stein theorem, let’s first take a detour and consider the Stein’s unbiased risk estimate \[^2\]. Suppose, as before, that $X \sim N_p(\mu, \sigma^2 I)$, and consider the estimator of the form

$$\hat{\mu} = X + g(X),$$

where $g$ is almost differentiable, and the partial derivatives of $g$ are integrable, i.e.,

$$\mathbb{E} \sum_{i=1}^{p} |\partial_i g_i(X)| < \infty.$$

If the above two assumptions don’t hold, Stein’s identity would no longer be correct.

Almost differentiability means that there exists $h_i$ so that we can write

$$g_i(x + z) - g_i(x) = \int_{0}^{1} \langle h_i(x + tz), z \rangle dt.$$

Usually, we write $h_i = \nabla g_i$. In particular, a discontinuous estimator wouldn’t be almost differentiable, and thus the Stein’s formula doesn’t apply to estimators that are discontinuous.

The main result that we will use in order to compute the risk of $\hat{\mu}$ in this setup is Stein’s identity.

**Theorem 3** (Stein’s identity 1981).

$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = p\sigma^2 + \mathbb{E}\left[\|g(X)\|^2 + 2\sigma^2 \sum_i \partial_i g_i(X)\right].$$

If we remove the expectation in the above identity, we will end up with a statistic that we can actually compute. An important consequence of Stein’s identity is Stein’s Unbiased Risk Estimate

$$\text{SURE}(\hat{\mu}) = p\sigma^2 + \|g(X)\|^2 + 2\sigma^2 \cdot \text{div} g(X),$$

where $\text{div} g(X) = \sum_i \partial_i g_i(X)$ is the divergence of $g(X)$. Note that SURE($\hat{\mu}$) is an unbiased statistic for the risk of the estimator $\hat{\mu}$.

**Proof.** (from last year’s notes) Assume without loss of generality that $\sigma = 1$. Then the risk of $\hat{\mu}$ is

$$\mathbb{E}\|X + g(X) - \mu\|^2 = \mathbb{E}\|X - \mu\|^2 + 2\mathbb{E}[(X - \mu)^T g(X)] + \mathbb{E}\|g(X)\|^2.$$

We just need to show that $\mathbb{E}[(X - \mu)^T g(X)] = \mathbb{E}\text{div} g(X)$. This follows easily from integration by parts.
Let $\phi$ denote the $N(0, I)$ pdf. Then we can write
\[
E(X_i - \mu_i)g_i(X) = \int (x_i - \mu_i)g_i(x)\phi(x - \mu)dx.
\]
Since
\[
\partial_i\phi(x - \mu) = -(x_i - \mu_i)\phi(x - \mu),
\]
we have
\[
E(X_i - \mu_i)g_i(X) = \int \partial_i g_i(x)\phi(x - \mu)dx = E\partial_i g_i(X).
\]

16.4.2 Applying SURE to $\hat{\mu}_{JS}$ ($\sigma = 1$)

Now we can apply the Stein’s identity to show that the JS estimator dominates the MLE. Note that
\[
\hat{\mu}_{JS} = X - \frac{p - 2}{\|X\|^2}X.
\]
Thus $\hat{\mu}_{JS}$ is of the form $X + g(X)$ where $g(X) = -(p - 2)x/\|x\|^2$. This gives
\[
\|g(X)\|^2 = \frac{(p - 2)^2}{\|X\|^2}
\]
\[
\partial_i g_i(x) = \partial_i \left\{ -(p - 2)\frac{x_i}{\|x\|^2} \right\} = -\frac{p - 2}{\|x\|^2} + \frac{2(p - 2)x_i^2}{\|x\|^4}
\]
\[
\Rightarrow \text{div } g(x) = \sum_i \partial_i g_i(x) = -\frac{(p - 2)^2}{\|x\|^2}.
\]
By Stein’s identity, putting everything together gives
\[
E_{\mu}\|\hat{\mu}_{JS} - \mu\|^2 = p + E \left[ \|g(X)\|^2 + 2\text{div } g(X) \right]
\]
\[
= p - E \left[ \frac{(p - 2)^2}{\|X\|^2} \right]
\]
\[
< p
\]
Therefore, the risk of the JS estimator is always less than the risk of the MLE.

Remark: We can be even more precise if we want to get some approximations. Note that $\|X\|^2$ follows a non-central $\chi^2$ distribution, and thus
\[
E_{\mu} \frac{1}{\|X\|^2} \geq \frac{1}{(p - 2) + \|\mu\|^2}
\]
with equality if $\mu = 0$. We can then bound the risk of the James-Stein estimator by
\[
E_{\mu}\|\hat{\mu}_{JS} - \mu\|^2 \leq p - \frac{p - 2}{1 + \|\mu\|^2}.
\]
It is interesting to consider a few special cases.
Figure 16.2. Comparison of the risk of $\hat{\mu}_{\text{MLE}}$ (red) to the upper bound derived for risk of $\hat{\mu}_{\text{JS}}$ (black).

- Under the global null, $\|\mu\|^2 = 0$, in which case $R(\hat{\mu}_{\text{JS}}, \mu) = 2$;

- If the signal to noise ratio is approximately 1, $\|\mu\|^2 = p - 2$, and $R(\hat{\mu}_{\text{JS}}, \mu) \leq p/2 + 1$;

- As $\|\mu\|^2 \to \infty$, $R(\hat{\mu}_{\text{JS}}, \mu) \to p$.

Figure 16.2 shows a plot of the upper bound obtained for the risk of $\hat{\mu}_{\text{JS}}$ compared to the risk of $\hat{\mu}_{\text{MLE}}$. For the risk of $\hat{\mu}_{\text{JS}}$ to be close to the risk of $\hat{\mu}_{\text{MLE}}$, $\|\mu\|^2$ needs to be quite large.

If the shrinkage in $\hat{\mu}_{\text{JS}}$ is too large, it is possible that the estimator switches to the other sign. By precluding the possibility of a sign reversal, the positive JS estimator

$$\hat{\mu}_{\text{JS}}^+ = \left(1 - \frac{p - 2}{\|X\|^2}\right)_{+} X$$

further improves upon the JS estimate, i.e., $R(\hat{\mu}_{\text{JS}}^+, \mu) < R(\hat{\mu}_{\text{JS}}, \mu)$ for all $\mu$. However, this estimator is not admissible either.
References

