1 Agenda

- Go over Exam 2 solutions
- Problem 1, on James-Stein estimation (Decision Theory)
- Problem 2, on power calculations (Hypothesis Testing)

Next section, we will spend all of the time on hypothesis testing.

2 Stein's Paradox

2.1 Multivariate Loss

Let θ be a $p \times 1$ parameter with estimator $\hat{\theta}$. For this multivariate setting, we may extend the squared error loss as

$$L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||^2 = \sum_{j=1}^{p} (\theta_j - \hat{\theta}_j)^2,$$

where θ_j and $\hat{\theta}_j$ are the j-th components of these variables. The corresponding (frequentist) risk then is

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}_{\theta} L(\theta, \hat{\theta}) \\ &= \mathbb{E}\left(||\theta - \mathbb{E}(\hat{\theta})||^2\right) + tr(\mathrm{Cov}(\hat{\theta})) \end{aligned}$$

where the second line was shown in Lecture 6 and gives a multivariate bias-variance tradeoff. Note too that the bias part of the risk can be shown to be

$$\mathbb{E}\left(||\theta - \mathbb{E}(\hat{\theta})||^2\right) = \sum_{j=1}^{p} (\theta_j - \mathbb{E}(\hat{\theta}_j))^2,$$

giving a convenient formula for the risk:

$$R(\theta, \hat{\theta}) = \sum_{j=1}^{p} (\theta_j - \mathbb{E}(\hat{\theta}_j))^2 + tr(\mathrm{Cov}(\hat{\theta}))$$

2.2 Estimation of Multivariate Normal Mean

Let $X \sim N_p(\theta, \sigma^2 I_p)$. We have then that

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix},$$

$$\theta = \mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_p) \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$$

$$\sigma^2 I_p = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix}$$

Suppose also that each of the components of X are independent, i.e., that X_j is independent of X_k for any $j \neq k$.

Recall that the MLE is $\hat{\theta}^{MLE} = \bar{X} = X$, and it is unbiased and minimax for squared error loss.

It turns out that if $p \geq 3$, then $\hat{\theta}^{MLE}$ is inadmissible for squared error loss. The James-Stein estimator dominates it and is defined as

$$\hat{\theta}^{JS} = \left(1 - (p-2)\frac{\sigma^2}{\sum_{k=1}^p X_k^2}\right)^+ X.$$

In vector form, this is

$$\hat{\theta}^{JS} = \begin{pmatrix} \left(1 - (p-2)\frac{\sigma^2}{\sum_{k=1}^p X_k^2}\right)^+ \mathbb{E}(X_1) \\ \vdots \\ \left(1 - (p-2)\frac{\sigma^2}{\sum_{k=1}^p X_k^2}\right)^+ \mathbb{E}(X_j) \end{pmatrix}.$$

It seems surprising that the MLE is inadmissible for $p \geq 3$, considering that the MLE is typically viewed as a good estimator. And it also seems surprising that the James-Stein estimator would do better for $p \geq 3$, since for each j, $\hat{\theta}_j^{JS}$ *uses information from other components of* X...which we assumed to be independent!

However, the result seems less surprising when we consider the fact that the risk is measured over the *full vector* X (the squared error loss sums the individual squared error losses for each component)– the James-Stein estimator wouldn't necessarily do better for *each individual element of* θ. Also we assumed that each component X_j has common variance σ^2. Additionally, the James-Stein estimator can be interpreted in terms of a broader class of estimators called *shrinkage estimators* (see Problem 1).

Problem 1. Shrinkage Estimator for Mean

Let $X \in \mathbb{R}^p$, and suppose $X \sim N(\theta, \sigma^2 I_p)$ with $\sigma^2 = 1$ and where the components of X are independent.

(a) Calculate the (frequentist) risk of the MLE $\hat{\theta}$.

(b) Consider a shrinkage estimator $\tilde{\theta} = cX$ for some $c \in [0, 1]$. Calculate the (frequentist) risk of $\tilde{\theta}$.

(c) Discuss what values of c will tend to give better risk for $\tilde{\theta}$, depending on the size of the bias and the variance.

(d) Choose c to give the James-Stein estimator (where still we are assuming $\sigma^2 = 1$, for simplicity here). Interpret the risk.

Solution

Part (a)

We have

$$\begin{aligned}
R_{MLE} &= \mathbb{E}L(\theta, \hat{\theta}) \\
&= \sum_{j=1}^{p} (\theta_j - \mathbb{E}(\hat{\theta}_j))^2 + tr(\mathrm{Cov}(\hat{\theta}) \\
&= \sum_{j=1}^{p} (\theta_j - \mathbb{E}(X))^2 + tr(\mathrm{Cov}(X)) \\
&= \sum_{j=1}^{p} (\theta_j - \theta_j)^2 + tr(\mathrm{Cov}(I_p)) \\
&= 0 + p = p.
\end{aligned}$$

Part (b)

Applying the same formula as above but for the shrinkage estimator, we have

$$\begin{aligned}
R_S = \mathbb{E}L(\theta, \tilde{\theta}) &= \sum_{j=1}^{p} (\theta_j - \mathbb{E}(\tilde{\theta}))^2 + tr(\text{Cov}(\tilde{\theta})) && (1)\\
&= \sum_{j=1}^{p} (\theta_j - c\mathbb{E}(X))^2 + tr(\text{Cov}(cX)) && (2)\\
&= \sum_{j=1}^{p} (\theta_j - c\theta_j)^2 + tr(c^2\text{Cov}(X)) && (3)\\
&= (1-c)^2 \sum_{j=1}^{p} \theta_j^2 + c^2 \cdot tr(\text{Cov}(X)) && (4)\\
&= (1-c)^2 ||\theta||^2 + c^2 p. && (5)
\end{aligned}$$

Part (c)

1) If $c = 1$, then $\tilde{\theta} = X = \hat{\theta}$ (the same as the MLE). This gives a risk of $R_S = p$, which only includes the variance portion and no bias. But this is the maximum possible variance portion that $\tilde{\theta}$ could have, since $p \geq c^2 p$ for $c \in [0, 1]$.

2) If $c = 0$, then $\tilde{\theta} = 0$, and $R_s = ||\theta||^2$. So it will include the bias portion but no variance. But this is the maximum possible bias part that $\tilde{\theta}$ could have for a fixed θ, since $||\theta||^2 \geq (1-c)^2 ||\theta||^2$ for $c \in [0, 1]$.

3) For values of $c \in (0, 1)$, neither the bias nor the variance part of the risk will be 0.

Intuitively, if $||\theta||^2$ is large, then to minimize the risk, we would want to choose c close to 1. On the other hand, if p is large, then we would want to choose c close to 0.

This can show how, depending on the sizes of θ and p, a shrinkage estimator might perform better than the MLE here.

Part (d)

The James-Stein estimator here is

$$\begin{aligned}\hat{\theta}^{JS} &= \left(1 - (p-2)\frac{1}{\sum_{j=1}^{p}||X||^2}\right)^{+} X \\ &= cX,\end{aligned}$$

where

$$c = \left(1 - (p-2)\frac{1}{\sum_{j=1}^{p}||X||^2}\right)^{+}.$$

When $||X||^2$ is large, c will be close to 1. This makes sense intuitively, because on average, X will be large when θ is "large". On the other hand, when p is large, the value of c will be close to 0. Therefore, this estimator aligns with the idea* we had for choosing c in part (c) of this problem. The James-Stein estimator is "automatically" choosing a value of c to potentially improve the risk, depending on the size of the terms that contribute to the bias vs. the variance.

*Note that there are other reasonable ways of choosing c, besides what's used in the James-Stein estimator.

Problem 2. Power Calculations for Binomial

(Exercise on p. 10 of Lecture 7 notes)

Let $X \sim Bin(5, p)$. Consider testing $H_0 : p \leq 1/2$ versus $H_1 : p > 1/2$.

(a) Consider two different rejection regions:

$$\begin{aligned} R_1 &= \{x : x = 5\} \\ R_2 &= \{x : x \geq 3\}. \end{aligned}$$

Plot and compare the corresponding power functions $\beta_1(p)$ and $\beta_2(p)$.

(b) Consider a rejection region of the form $R = \{x : x \geq c\}$.

- What values of c do we need to consider?
- For each of these, find the size of the corresponding test.
- What c should we choose if we want a probability of Type-I error of no more than 10%?

Solution

See the Section 8 R code for the computations and plotting.

Part (a)

In general, a power function is defined as

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

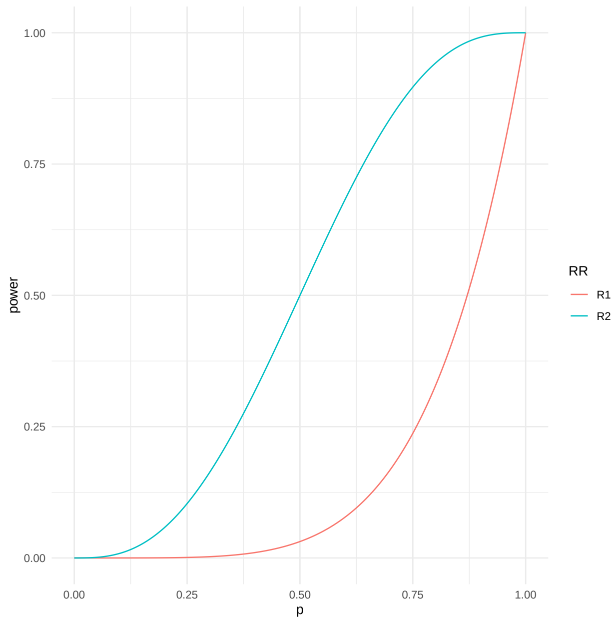
Therefore, the first power function is

$$\begin{aligned} \beta_1(p) &= \mathbb{P}_p(X \in R_1) \\ &= \mathbb{P}_p(X = 5) \\ &= \binom{5}{5} p^5 (1-p)^{5-0} = p^5, \end{aligned}$$

and the second one is

$$\begin{aligned}\beta_2(p) &= \mathbb{P}_p(X \geq 3)\\ &= \sum_{k=3}^{5} \mathbb{P}_p(X = k)\\ &= \sum_{k=3}^{5} \binom{5}{k} p^k (1-p)^{5-k}.\end{aligned}$$

From the below plot from R, we see that $\beta_2(p) > \beta_1(p)$ for all p. This makes sense, since $\beta_2(p) = \beta_1(p) + \sum_{k=3}^{4} \binom{5}{k} p^k (1-p)^{5-k}$. That is, there is a higher probability of rejecting H_0 if we reject whenever $X = 3, 4, 5$ vs. if we only reject when $X = 5$.



Part (b)

- X is discrete and takes values in $\{0, 1, \ldots, 5\}$. It follows then that c should take values in the support $\{0, 1, \ldots, 5\}$, since for other values the power would automatically be 0.

- In general, the size of a test is defined as

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \in R) = \underset{\theta \in \Theta}{\beta(\theta)},$$

which here is

$$\alpha = \sup_{p \in [0,1/2]} \mathbb{P}_p(X \in R) = \underset{p \in [0,1/2]}{\beta(p)} .$$

Hypothetically, for each value of c, one could calculate the power then maximize it over $p \in [0, 1/2]$ (using analytical methods or calculus).

Here is a helpful fact that we can use instead: If $Y \sim Bin(m, r)$, then for any k in the support, the CDF

$$F_r(k) = \mathbb{P}_r(Y \leq k)$$

is decreasing in p.

Let c be a value in the support that is not 0. The power is

$$\mathbb{P}_p(X \geq c) = 1 - \mathbb{P}_p(X < c) = 1 - \mathbb{P}_p(X \leq c - 1) = 1 - F_p(c - 1).$$

This is maximized with respect to p when $F_p(c-1)$ is as small as possible, which is when p is as large as possible. The maximum value of p under H_0 is $1/2$. So the size of the test is

$$\alpha = \mathbb{P}_{p=1/2}(X \geq c).$$

If $c = 0$, then $\mathbb{P}_p(X \geq 0) = 1$, so the power is trivially 1.

The powers are calculated in R and shown as follows:

c	α
0	1
1	0.9688
2	0.8125
3	0.5000
4	0.1875
5	0.0313.

• To obtain a Type-I error rate of no more than 10%, we would choose $c = 5$ since $\alpha = 0.0313 < 0.1$. Note that this is considered a "conservative" test, since the Type-I error rate α is *lower* than the desired Type-I error rate .1. This often happens for discrete RVs such as the Binomial.