Lecture 39: Stein's paradox

Ciaran Evans

Some key parts of the course

- Maximum likelihood estimation
  - univariate and multivariate problems
  - maximum likelihood estimation for regression models

- Properties of MLEs (under regularity conditions)
  - consistency
  - asymptotic normality
  - asymptotic efficiency (asymptotic variance is CRLB)

- Hypothesis testing
  - Neyman-Pearson test (involves likelihoods)
  - Wald test (can be used for asymptotically normal estimators, like MLEs)
  - Likelihood ratio test

Today: Maximum likelihood estimation isn't *always* best

The problem

Suppose we observe a single observation $X \sim N(\mu, \mathbf{I})$ from a d-dimensional multivariate normal distribution. We wish to estimate

$$\mu = (\mu_1, \mu_2, ..., \mu_d)^T$$

Question: What do you think is the MLE $\widehat{\mu}_{MLE}$?

MSE

Suppose we observe a single observation $X \sim N(\mu, \mathbf{I})$ from a d-dimensional multivariate normal distribution. We wish to estimate

$$\mu = (\mu_1, \mu_2, ..., \mu_d)^T$$

MLE: $\widehat{\mu}_{MLE} = X$

$$MSE(\widehat{\mu}_{MLE}) = \mathbb{E}[||\widehat{\mu}_{MLE} - \mu||^2] =$$

Another estimator

Suppose we observe a single observation $X \sim N(\mu, \mathbf{I})$ from a d-dimensional multivariate normal distribution. The **James-Stein estimator** of μ is

$$\widehat{\mu}_{JS} = \left(1 - \frac{d-2}{||X||^2}\right) X$$

Activity

$$\widehat{\mu}_{MLE} = X \qquad \widehat{\mu}_{JS} = \left(1 - \frac{d-2}{||X||^2}\right) X$$

Compare $MSE(\widehat{\mu}_{MLE})$ to $MSE(\widehat{\mu}_{JS})$

Activity

$$\widehat{\mu}_{MLE} = X \qquad \widehat{\mu}_{JS} = \left(1 - \frac{d-2}{||X||^2}\right) X$$

Question: How does $MSE(\widehat{\mu}_{MLE})$ compare to $MSE(\widehat{\mu}_{JS})$?

Comparing MSE

- We know that $MSE(\widehat{\mu}_{MLE}) = d$
- It turns out that $MSE(\widehat{\mu}_{JS}) = d - (d-2)^2 \mathbb{E}\left[\frac{1}{||X||^2}\right] < d$

Stein's paradox

Suppose we observe a single observation $X \sim N(\mu, \mathbf{I})$ from a d-dimensional multivariate normal distribution.

Regardless of the true value of μ, if $d \geq 3$ then

$$MSE(\widehat{\mu}_{JS}) < MSE(\widehat{\mu}_{MLE})$$

So: in this situation, if our goal is to minimize MSE we should *never* use the MLE. That is, the MLE is **inadmissible**

Shrinkage estimators

The James-Stein estimator beats the MLE by shrinking towards 0.
Other examples of shrinkage estimators are important in regression.

Linear regression:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

- Least squares estimates: $\widehat{\beta} = \text{argmin}_\beta ||\mathbf{y} - \mathbf{X}\beta||^2$
- Ridge regression: $\widehat{\beta} = \text{argmin}_\beta ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||^2$
- Lasso: $\widehat{\beta} = \text{argmin}_\beta ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1$