Advanced Regression: Shrinkage estimates

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

25th April 2019

Bias-variance trade-off in estimation

- Consider an estimate $\hat{\theta}$ for a parameter θ.
- Examples:
  - ◇ Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ for the population mean.
  - ◇ Sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ for the population variance.

Bias of an estimate

$$Bias(\hat{\theta}) = E\left(\hat{\theta} - \theta\right)$$

Bias-variance trade-off in estimation

Mean squared error (MSE) of an estimate $\hat{\theta}$

MSE is the squared average difference between an estimate $\hat{\theta}$ and the true parameter θ.

$$
\begin{aligned}
MSE(\hat{\theta}) &= E\left(\hat{\theta} - \theta\right)^2 \\
&= (E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2) \\
&= (E(\hat{\theta}^2) - \underbrace{E(\hat{\theta})^2 + E(\hat{\theta})^2}_{0} - 2E(\hat{\theta})\theta + \theta^2) \\
&= (E(\hat{\theta}^2) - E(\hat{\theta})^2) + (E(\hat{\theta})^2 - 2E(\hat{\theta})\theta + \theta^2) \\
&= Var(\hat{\theta}) + \left(Bias(\hat{\theta})\right)^2
\end{aligned}
$$

Bias-variance trade-off in estimation

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left(Bias(\hat{\theta})\right)^2$$

where

- $Var(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$
- $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta = E(\hat{\theta} - \theta)$

- Many classical techniques are designed to be unbiased (BLUE) or consistent (Maximum Likelihood), but may have a large variance in high-dimensional settings.
- Shrinkage estimates trade unbiasedness against reduced variance.

Shrinkage estimate

Shrinkage estimate $\hat{\theta}^{\text{shrink}}$ for an unknown parameter vector θ of length p is based only on three components

$$\hat{\theta}^{\text{shrink}} = (1-\lambda)\hat{\theta} + \lambda\theta^{\text{target}}$$

with

- $\hat{\theta}$ as an unregularized estimate for θ,
- θ^{target} as the target, and
- λ shrinkage parameter $[0, 1]$

Bias-variance trade-off

- The unregularised estimate $\hat{\theta}$ has no or little bias, but large variance.
- The target θ^{target} has large bias, but no or little variance.

How to design a target?

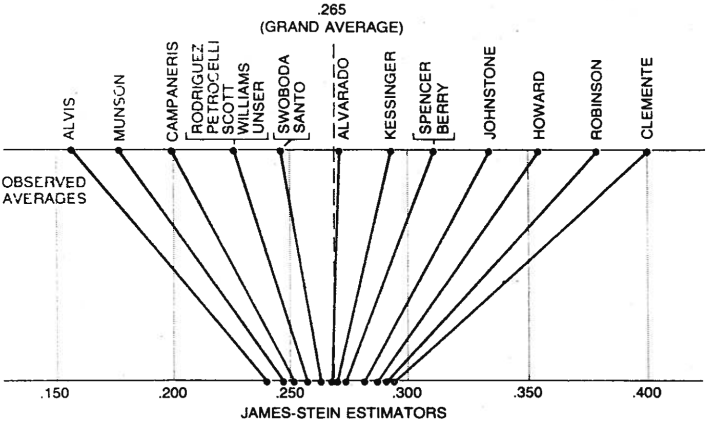
The design of the target depends on the data and our implicit prior belief.

Examples:

1. $\theta^{\text{target}} = 0$

   Pessimistic prior belief, we prefer to shrink all variables towards zero (James-Stein estimate, regularised regression).
2. $\theta^{\text{target}} = \bar{x}$
   - 'Regression to the mean': Large values will get smaller, small values will increase.
   - Implicit belief that we can learn general information from all variables (mean, median) and improve the estimate for a single variable by accounting for the behaviour of other variables → Borrowing strength across variables (Efron-Morris estimate).

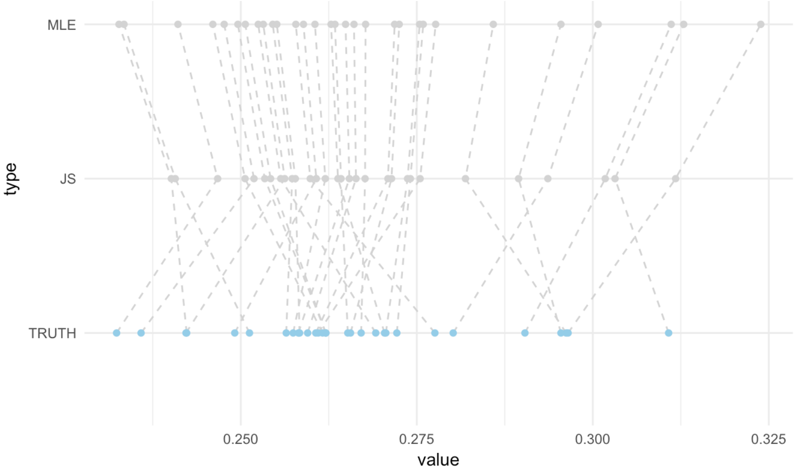
Efron-Morris estimate



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value.

Efron and Morris 1977

Efron-Morris estimate



https://bookdown.org/content/927/james-stein.html

Efron-Morris estimate

$$\begin{aligned}\hat{\theta}_j^{EM} &= \bar{x} + (1 - \frac{(p-3)}{\sum_{j=1}^{p} x_j^2})(x_j - \bar{x}) \\ &= \bar{x} + c(x_j - \bar{x})\end{aligned}$$

- where $\bar{x} = \frac{1}{p}\sum_{j=1}^{p} x_j$ is the grand mean over all variables, which is used to aggregate information across variables (borrowing strength)
- x_j unbiased estimate with large variance
- Shrinkage parameter

$$c = 1 - \frac{(p-3)}{\sum_{j=1}^{p} x_j^2} \rightarrow \begin{cases} 0 & \text{when } \sum_{j=1}^{p} x_j^2 \text{ is small} \\ 1 & \text{when } \sum_{j=1}^{p} x_j^2 \text{ is large} \end{cases}$$

Efron-Morris estimate

- $c \to 0$ when $\sum_{j=1}^{p} x_j^2$ (variability) is small (Shrinkage to $\bar{x}$)
- $c \to 1$ when $\sum_{j=1}^{p} x_j^2$ (variability) is large (No shrinkage)

$$\hat{\theta}_j^{EM} = \begin{cases} \bar{x} + 0 = \bar{x} & \text{if } c = 0 \\ \bar{x} + x_j - \bar{x} = x_j & \text{if } c = 1 \end{cases}$$

- If $c = 0$ there is an implicit assumption that all $\theta_1, ..., \theta_p$ might be similar and replaced with $\bar{x}$ which can be estimated much more precisely.