Table of Contents

About

This Shiny app was developed by Don van den Bergh (donvdbergh<at>hotmail.com) for the "Tools for Teaching Quantative Thinking" (TquanT) Erasmus programme. The goal of this shiny app is to visualize the effect of shrinkage estimators and compare their performance to other estimators. This app is intended for educational purposes; the different tabs contain explanations of the used datasets and estimators.

The sections below contain all information present in the app.

Datasets

Stein's Paradox

This classical dataset discussed in Stein's Paradox (Efron & Morris 1977) contains data on the batting average of 18 players in the baseball season of 1970. The batting average is how often a player managed to hit a baseball with a bat divided by the number of oppertunities a player had to hit the ball. In this dataset, at the time of the first measurement, only players that had exactly 45 oppertunities to hit a ball were recorded. The batting averages of these players were recored again at the end of the season. Unintuitively, the batting average of the first 45 matches is not the best predictor for the batting average at the end of the season. Instead, these averages can be shrunk towards the group mean to obtain a better estimate (see estimator tab for details).

References:

Efron, B., & Morris, C. N. (1977). Stein's paradox in statistics. WH Freeman. Link

Olympics (2012 & 2016)

This dataset contains the population / medal for each country that participated in the olympics of London 2012 and Rio de Janeiro 2016. A big difference with this dataset compared to the other datasets is that we do not know the true scores in this dataset. Hence we cannot calculated the RMSE in the standard way to evaluate prediction error. Instead, we looked at the Olympics of London 2012, used a shrinkage estimator, and attempted to predict the outcome of the olympics in Rio de Janeiro 2016.

To facilitate the analysis only countries that competed in both the olympics of London 2012 and Rio de Janeiro 2016 were included in the analysis. We also excluded some countries that were outliers. This biases results, but for evaluating the estimators this is good enough. For more accurate results, one could use data from more olympics years.

One of the nice aspects of this dataset is that it shows that there is a better predictor for the performance of countries during the next olympic games than just their performance on the previous olympic games! This dataset is also an example where the James Stein estimator does not perform better than the observed data, but a hierarchical bayesian model does.

The data for both olympics can be found at medalspercapita.

Eredivisie (2015 - 2016)

Data from the Dutch soccer competition in 2015/2016 taken from Wikipedia.

Estimators

James Stein

The James stein estimator is a specific case of the empirical Bayes estimator. The Empirical Bayes estimator for the mean is:

$$\mu_{EB} = (1 - \frac{A}{A+1})\mu_{ML}$$

Here μ_{ML} denotes the maximum likelihood estimator for the mean ($\mu_{ML} = N^{-1}\sum_{i=1}^{N} x_i$) and μ_{EB} denotes the empirical Bayes estimator. In the case of the James Stein Estimator this equation becomes:

$$\mu_{JS} = \bar{y} + c(y - \bar{y})$$

Here y denotes the individual mean of a group and $\bar{y}$ denotes the grand mean of all averages. Subsequently, the James stein estimator shrinks every mean y towards the grand mean $\bar{y}$ with a shrinkage factor c. The shrinkage factor c is defined as:

$$c = 1 - \frac{(k-3)\cdot\sigma^2}{\sum_i (y_i - \bar{y})^2}$$

Above, k denotes the number of unknown means, σ^2 is the variance of the observed means and $\sum_i (y_i - \bar{y})^2$ is simply the sum of squares of the observed means.

References:

Efron, B., & Morris, C. N. (1977). Stein's paradox in statistics. WH Freeman. Link

Efron, B. (2012). Large-scale inference: empirical Bayes methods for estimation, testing, and prediction (Vol. 1). Cambridge University Press. Link

Hierarchical Bayes

In hierarchical Bayes we explicitly model the data with Bayes theorem:

$$P(Z, \theta|x) = \frac{P(x|\theta, Z)P(\theta)P(Z)}{\int\int P(x|\theta, Z)P(\theta)P(Z)\ \mathrm{d}\theta\ \mathrm{d}Z}$$

Here, x represents the observed data, θ represents the model parameters and Z represents the group level parameter(s). Furthermore, $P(x|\theta, Z)$ is the likelihood of the data and $P(\theta)$ and $P(Z)$ are priors for their respective parameters. The double integral in the denominator is intractible, thus there is no way to analytically calculate the posterior distribution (denoted $P(Z, \theta|x)$). The posterior above has no closed form solution and therefore it needs to be evaluated with an iterative method; This shiny app uses Gibbs sampling.

References:

Wikipedia