UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Lecture 16 Model Inference: Bootstrap

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

Outline

The bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

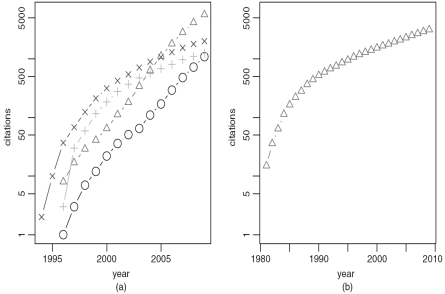


Fig. 2. Cumulative citation counts (on a log-scale) from the Thomson ISI *Web of Knowledge* (the largest abscissa on the *x*-axis corresponds to August 31st, 2010): (a) the lasso (○) (Tibshirani, 1996), false discovery rate (Δ) (Benjamini and Hochberg, 1995), reversible jump Markov chain Monte Carlo sampling (+) (Green, 1995) and wavelet shrinkage (×) (Donoho and Johnstone, 1994), published between 1994 and 1996; (b) the bootstrap (Δ) (Efron, 1979), published earlier

The name of bootstrap

- The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstraps, widely thought to be based on one of the eighteenth century "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe:
- *The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.*

- The Adventures of Baron Munchausen (1988) by John Neville (the X Files)

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

A simple Example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y , respectively, where X and Y are random quantities.
- We will invest a fraction of our money in X, and will invest the remaining $1-\alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $\text{Var}(\alpha X + (1-\alpha)Y)$.
- One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

where $\sigma_X^2 = \text{Var}(X)$, $\sigma_Y^2 = \text{Var}(Y)$, and $\sigma_{XY}^2 = \text{Cov}(X, Y)$.

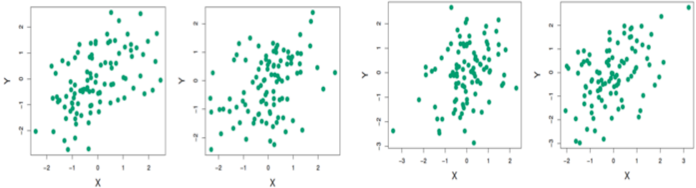
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

A simple Example

- The values σ_X^2, σ_Y^2, and σ_{XY}^2 are unknown, which can be estimated from the data, denoted by $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\hat{\sigma}_{XY}^2$.
- We can then estimate the value of α that minimizes the variance of our investment using

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

- Each panel displays 100 simulated returns for investments X and Y . From left to right, the resulting estimates for α are 0.576, 0.532, 0.657, and 0.651.



A simple Example

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y, and estimating $1,000$ times.
- We thereby obtained $1,000$ estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_{1000}$.
- For these simulations the parameters were set to $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$, and $\sigma_{XY}^2 = 0.5$. So the true value of $\alpha = 0.6$.
- The mean over all $1,000$ estimates for α is

$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996,$$

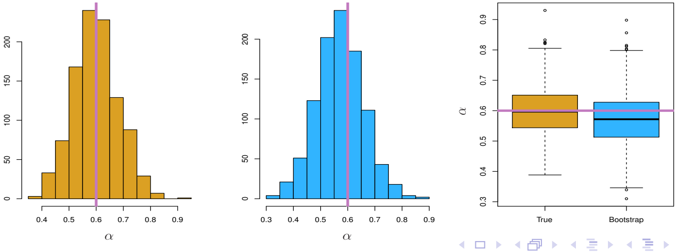
very close to $\alpha = 0.6$, and the the standard deviation of the estimates is

$$\sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})} = 0.083.$$

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

A simple Example

- This gives a very good idea of the accuracy of $\hat{\alpha}$, $\mathrm{SE}(\hat{\alpha}) \approx 0.083$.
- So roughly speaking, for a random sample from the population, we would expect $\hat{\alpha}$ from α by approximately 0.08, on average.
- Left: A histogram of the estimates of α obtained by generating $1,000$ simulated data sets from the true population.
- Right: ... from $1,000$ bootstrap samples from a single data set.
- Right: Boxplots from True and *bootstrap*, The pink line indicates the true value of α.



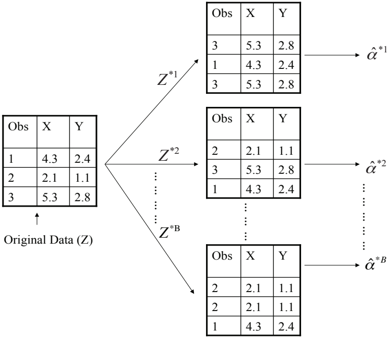
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

The Bootstrap

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement.
- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Example with just 3 observations



- A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations.
- Each bootstrap data set contains n observations, sampled with replacement from the original data set.
- Each bootstrap data set is used to obtain an estimate of α.

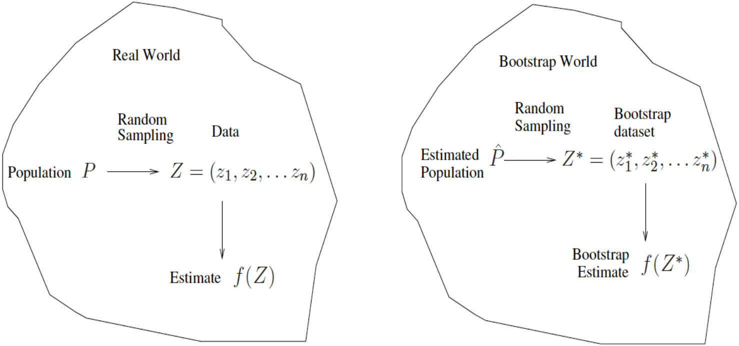
The Bootstrap

- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α which we call α^{*1}.
- This procedure is repeated B times for some large value of B (say 100 or 1000), in order to produce B different bootstrap data sets, $Z^{*1}, \cdots, Z^{*B}$ and B corresponding α estimates, $\alpha^{*1}, \cdots, \alpha^{*B}$.
- We estimate the standard error of these bootstrap estimates using the formula

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}_r - \bar{\alpha})}.$$

- This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set.
- The Bootstrap result for this example is $\mathrm{SE}_B(\hat{\alpha}) = 0.087$.

A general picture for the bootstrap



UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

The Bootstrap

- In more complex data situations, figuring out the appropriate way to generate bootstrap samples can require some thought.
- For example, if the data is a time series, we can't simply sample the observations with replacement (why not?).
- We can instead create blocks of consecutive observations, and sample those with replacements. Then we paste together sampled blocks to obtain a bootstrap dataset.
- This is called Block Bootstrap. There are many other variations, say, Wild Bootstrap and so on.

Other uses

- Primarily used to obtain standard errors of an estimate.
- Also provides approximate confidence intervals for a population parameter.
- For example, looking at the histogram in the middle panel of the Figure on slide 8, the 5% and 95% quantiles of the 1000 values is $(0.43, 0.72)$.
- This represents an approximate 90% confidence interval for the true α. How do we interpret this confidence interval?
- The above interval is called a Bootstrap Percentile confidence interval.
- It is the simplest method (among many approaches) for obtaining a confidence interval from the bootstrap.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

The bootstrap can not be used for prediction errors

- The bootstrap can not be used for prediction errors!!!
- In cross-validation, each of the K validation folds is distinct from the other $K-1$ folds used for training: there is no overlap. This is crucial for its success. Why?
- To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.
- But each bootstrap sample has significant overlap with the original data. About two-thirds of the original data points appear in each bootstrap sample.
- This will cause the bootstrap to seriously underestimate the true prediction error. Why?
- The other way around — with original sample = training sample, bootstrap dataset = validation sample — is worse!

Summary and Remark

- Introduction
- The bootstrap
- Related topics
- Read textbook Chapter 8 and `R` code
- Do `R` lab