Bootstrap Methods

MAST90083 Computational Statistics and Data Mining

Karim Seghouane

School of Mathematics & Statistics
The University of Melbourne

Outline

§6.1 Introduction

§6.2 Bootstrap principle

§6.3 Parametric vs Nonparametric

§6.4 Bias correction

Introduction

- Bootstrap methods use computer simulation to reveal aspects of the sampling distribution for an estimator $\hat{\theta}$ of interest.
- With the power of modern computers the approach has broad applicability and is now a practical and useful tool for applied statisticians and data scientists.

Introduction

- The bootstrap is a general tool for assessing statistical accuracy
- It is based on re-sampling strategy
- Having the estimated feature of the data that we compute based on the sample on hand, we are interested to understand how the estimate changes for a different sample
- Examples of features: prediction accuracy, the mean value, etc.
- But unfortunately we cannot use more than one sample
- Solution: bootstrap

Introduction

- The idea behind the bootstrap is an old one.
- Assume we wish to estimate a functional of a population distribution function F, such as the population mean

$$\theta = \int x dF(x)$$

- Consider employing the same functional of the sample (or empirical) distribution function $\hat{F}$, which in this case leads to the sample mean

$$\hat{\theta} = \int x d\hat{F}(x) = \bar{x}$$

Introduction

- One can use $\hat{\theta} = \bar{x}$ to estimate θ.
- Evaluating the variability in this estimation would require the sampling distribution of $\bar{x}$.

Empirical distribution

- The empirical distribution is that probability measure that assigns to a set a measure equal to the proportion of samples that lie in that set

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$$

- for a set $x_1, \ldots, x_n$ of i.i.d from F, where $\delta(x - x_i)$ represents a "point mass" at x_i (that assigns full probability to the point x_i and zero to all other points).
- $\hat{F}$ is the discrete distribution that assigns a mass $1/n$ to each point x_i, $1 \leq i \leq n$.
- By the L.L.N. $\hat{F} \to_p F$ as $n \to \infty$.

Sample and resample

- A sample $X = \{x_1, \ldots, x_n\}$ is a collection of n numbers (or vectors), without regard to order drawn at random from the population F.

- The x_i's are therefore i.i.d. random variables each having the population distribution function F
- A resample $X^* = \{x_1^*, \ldots, x_n^*\}$ is an unordered collection of n items randomly drawn from X with replacement

- It is known as a bootstrap sample and is a central step of the nonparametric bootstrap method

Resample

- Each x_i^* has probability $1/n$ of being equal to any given one x_j's

$$P(x_i^* = x_j | X) = \frac{1}{n}, \quad 1 \le i, j \le n$$

- The x_i^*'s are i.i.d. conditional on X.
- X^* is likely to contain repeats, all of which must be listed in X^*.

- Example: $X^* = \{1.5, 1.7, 1.7, 1.8\}$ is different from $\{1.5, 1.7, 1.8\}$ and X^* is the same as $\{1.5, 1.7, 1.8, 1.7\}$, $\{1.7, 1.5, 1.8, 1.7\}$.

Population and sample distribution

- F is the population distribution of X whereas $\hat{F}$ is its sample distribution.
- $\hat{F}$ on the other hand is the distribution function of the population from which X^* is drawn.
- $\left(F, \hat{F}\right)$ is generally written (F_0, F_1) in bootstrap iteration, where $i \geq 1$.
- F_i denotes the distribution function of a sample drawn from F_{i-1} conditional on F_{i-1}.
- The i^{th} application of the bootstrap is termed i^{th} iteration, not the $(i-1)^{th}$ iteration

Estimation as functional

- An estimate $\hat{\theta}$ is a function of the data and a functional of the sample distribution function $\hat{F}$
- Example: The sample mean

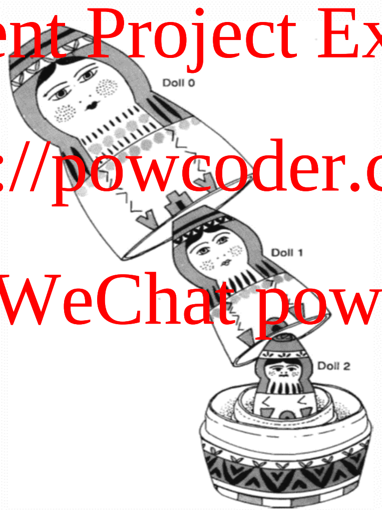
$$\hat{\theta} = \theta[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\theta} = \theta\left(\hat{F}\right) = \int x d\hat{F}(x)$$

- whereas the population mean

$$\theta = \theta(F) = \int x dF(x).$$

Bootstrap principle



Bootstrap principle

- Assume we can't observe "doll 0" $\rightarrow$ it represents the population in a sampling scheme
- We wish to estimate the number n_0 of freckles on its face.
- Let n_i denotes the number of freckles on the face of "doll i"
- Assuming the ration of n_1/n_2 close to the ratio n_0/n_1, we have $\hat{n}_0 \simeq n_1^2/n_2$
- The key feature of this argument is our hypothesis that the relationship between n_2 and n_1 should closely resemble that between n_1 and the unknown n_0

Bootstrap principle

- Statistical inference amounts to describing the relationship between a sample and the population from which the sample is drawn
- Formally: Given a functional f_t from a class $\{f_t : t \in \boldsymbol{\tau}\}$, we aim to find t_0 such that

$$\mathbb{E}\left\{f_t\left(F_0, F_1\right) \middle| F_0\right\} = \mathbb{E}_{F_0}\left\{f_t\left(F_0, F_1\right)\right\} = 0$$

- where $F_0 = F$ (population distribution) and $F_1 = \hat{F}$ (sample distribution)
- we want to find t_0 the solution of the population equation (because we need properties of the population to solve this equation exactly)

Bootstrap principle

Example:

- Let $\theta_0 = \theta(F) = \theta(F_0)$ be the true parameter value, such as the r^{th} power of a mean

$$\theta_0 = \left\{ \int x dF_0(x) \right\}^r$$

- Let $\hat{\theta} = \theta(F_1)$ be the bootstrap estimate of θ_0

$$\hat{\theta} = \left\{ \int x dF_1(x) \right\}^r = \bar{x}^r$$

- where F_1 is the empirical distribution function

Example: Bias correction

- Correcting $\hat{\theta}$ for bias is equivalent to finding t_0 that solves

$$\mathbb{E}_{F_0}\{f_t(F_0, F_1)\} = 0$$

- where

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t$$

- and the bias corrected estimate is $\hat{\theta} + t_0$

Example: Confidence interval

- To construct a symmetric, $(1-\alpha)$ confidence interval for θ_0 is equivalent to using

$$f_t(F_0, F_1) = I\{\theta(F_1) - t \le \theta(F_0) \le \theta(F_1) + t\} - (1-\alpha)$$

- where $I(.)$ denotes the indicator of the event that the true parameter value $\theta(F_0)$ lies in the interval

$$[\theta(F_1) - t, \theta(F_1) + t] = \left[\hat{\theta} - t, \hat{\theta} + t\right]$$

- minus the nominal coverage $1-\alpha$ of the interval. Asking that

$$\mathbb{E}\{f_t(F_0, F_1) | F_0\} = 0$$

- is equivalent to insisting that t be chosen so that the interval has zero coverage error.

Bootstrap principle

- The equation

$$\mathbb{E}\left\{f_t\left(F_0, F_1\right) | F_0\right\} = 0$$

- provides an explicite description of the relationship between F_0 and F_1 we are trying to determine.
- The analogue in the case of the number of freckles problem is

$$n_0 - tn_1 = 0$$

- where n_i is the number of freckles on doll "i"

Bootstrap principle

- If we had $t = t_0$ solving the equation, then $n_0 = t_0 n_1$.
- The estimation of t_0 is obtained from the pair (n_1, n_2) we know

$$n_1 - tn_2 = 0$$

- we obtain the solution $\hat{t}_0$ of this equation and thereby

$$\hat{n}_0 = \hat{t}_0 n_1 = \frac{n_1^2}{n_2}$$

- is the estimate of n_0

Bootstrap principle

- Similarly, the population equation

$$\mathbb{E}\{f_t(F_0, F_1)\,|F_0\} = 0$$

- is solved via the sample equation

$$\mathbb{E}\{f_t(F_1, F_2)\,|F_1\} = 0$$

- where F_2 is the distribution function of a sample drawn from F_1 is the analogue of n_2.

Bootstrap principle

- The solution $\hat{t}_0$ is a function of the sample values

- The idea is that the solution of the sample equation should be a good approximation of the solution of the population equation
- The population equation is not obtainable in practice
- $\rightarrow$ this is the bootstrap principle.

Bootstrap principle

- We call $\hat{t}_0$ and $\mathbb{E}\left\{f_t\left(F_1, F_2\right) | F_1\right\}$ "the bootstrap estimates" of t_0 and $\mathbb{E}\left\{f_t\left(F_0, F_1\right) | F_0\right\}$.

- They are obtained by replacing F_0 and F_1 in the formulae for t_0
- The bootstrap version of the bias corrected estimate is $\hat{\theta} + \hat{t}_0$

- The bootstrap confidence interval is $\left[\hat{\theta} - \hat{t}_0, \hat{\theta} + \hat{t}_0\right]$ called the symmetric percentile method confidence interval for θ_0

Parametric vs Nonparametric

- In both parametric and nonparametric problems, the inference is based on a sample X of size n (n i.i.d. observations of the population)

- In nonparametric case F_1, is the empirical distribution function of X
- Similarly F_2 is the empirical distribution function of a sample drawn at random from the population F_1

Nonparametric

- It is the empirical distribution of a sample X^* drawn randomly with replacement from X

- If we denote the population by X_0, then we have a nest of sampling operations
  - X is drawn at random from X_0
  - X^* is drawn at random from X

Parametric

- In this case F_0 is assumed completely known up to a finite vector $\boldsymbol{\lambda}_0$ of unknown parameters.
- $F_0 = F_{(\boldsymbol{\lambda}_0)}$ is an element of a class $\left\{F_{(\boldsymbol{\lambda})}, \boldsymbol{\lambda} \in \Lambda\right\}$ of possible distributions

- Then $F_1 = F_{(\hat{\boldsymbol{\lambda}})}$, the distribution function obtained using the sample estimate $\hat{\boldsymbol{\lambda}}$ obtained from X often (but not necessary) using maximum likelihood estimate.

- Let X^* denotes the sample drawn at random from $F_{(\hat{\boldsymbol{\lambda}})}$ and $F_2 = F_{(\hat{\boldsymbol{\lambda}}^*)}$
- In both cases, X^* is obtained by resampling from a distribution determined by the original sample X

Example

- Estimate of the MSE

$$\tau^2 = \mathbb{E}\left(\hat{\theta} - \theta_0\right)^2 = \mathbb{E}\left\{[\theta(F_1) - \theta(F_0)]^2 \,|\, F_0\right\}$$

- has bootstrap estimate

$$\hat{\tau}^2 = \mathbb{E}\left\{\left(\hat{\theta}^* - \hat{\theta}\right)^2 \,|\, X\right\} = \mathbb{E}\left\{[\theta(F_2) - \theta(F_1)]^2 \,|\, F_1\right\}$$

- where $\hat{\theta}^* = \theta[X^*]$ is an estimate version of $\hat{\theta}$ obtained using X^* instead of X

Bias correction

- Here we have

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t$$

- and the sample equation

$$\mathbb{E}\{f_t(F_1, F_2)|F_1\} = \mathbb{E}\{\theta(F_2) - \theta(F_1) + t|F_1\} = 0$$

- whose solution is

$$t = \hat{t}_0 = \theta(F_1) - \mathbb{E}\{\theta(F_2)|F_1\}$$

Bias correction

- The bootstrap bias-reduced estimate is this

$$\hat{\theta}_1 = \hat{\theta} + \hat{t}_0 = \theta(F_1) + \hat{t}_0 = 2\theta(F_1) - \mathbb{E}\{\theta(F_2)|F_1\}$$

- Note that the estimate $\hat{\theta} = \theta(F_1)$ is also a bootstrap functional since it is obtained by replacing F_1 for F_0 in the functional formula $\theta_0 = \theta(F_0)$

- the expectation $\mathbb{E}\{\theta(F_2)|F_1\}$ is computed (or approximated) by Monte Carlo simulation

Bias correction

- Draw B resamples $\{X_b^*, 1 \leq b \leq B\}$ independently from the distribution function F_1
- In the nonparametric case F_1 is the empirical distribution of the the sample X
- Let F_{2b} denotes the empirical distribution function of X_b^*
- In the parametric case, $\hat{\lambda}_b^* = \lambda(X_b^*)$ is the estimate of λ_0 obtained from X_b^* and $F_{2b} = F_{(\hat{\lambda}_b^*)}$

Bias correction

- Define $\hat{\theta}_b^* = \theta(F_{2b})$, then

$$\hat{u}_B = \frac{1}{B}\sum_{b=1}^{B}\theta(F_{2b}) = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^*$$

- converge to (as $B \to \infty$)

$$\hat{u} = \mathbb{E}\{\theta(F_2)|F_1\} = \mathbb{E}\left\{\hat{\theta}^*|X\right\}$$

Example (1)

- Let

$$\mu = \int x dF_0(x) \text{ and assume } \theta_0 = \theta(F_0) = \mu^3$$

- $X = \{x_1, ..., x_n\}$ and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- In nonparametric approach

$$\hat{\theta} = \theta(F_1) = \bar{x}^3$$

Example (2)

- In nonparametric approach

$$
\begin{aligned}
\mathbb{E}\left\{\theta(F_1)|F_0\right\} &= \mathbb{E}_{F_0}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^3\right\} \\
&= \mathbb{E}\left\{\left(\mu + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right)^3\right\} \\
&= \mu^3 + n^{-1}3\mu\sigma^2 + n^{-2}\gamma
\end{aligned}
$$

- where $\sigma^2 = E(x-\mu)^2$ and $\gamma = E(x-\mu)^3$ denote the population variance and skewness

Example

- In the nonparametric case

$$\mathbb{E}\left\{\theta(F_2)|F_1\right\} = \bar{x}^3 + n^{-1}3\bar{x}\hat{\sigma}^2 + n^{-2}\hat{\gamma}$$

- where $\hat{\sigma}^2 = n^{-1}\sum(x_i - \bar{x})$ and $\hat{\gamma} = n^{-1}\sum(x_i - \bar{x})^3$ denote the sample variance and skewness
- Therefore the bootstrap bias reduced estimate is

$$\hat{\theta}_1 = 2\theta(F_1) - \mathbb{E}\left\{\theta(F_2)|F_1\right\} = 2\bar{x}^3 - \left(\bar{x}^3 + n^{-1}3\bar{x}\hat{\sigma}^2 + n^{-2}\hat{\gamma}\right)$$

$$= \bar{x}^3 - n^{-1}3\bar{x}\hat{\sigma}^2 - n^{-2}\hat{\gamma}$$

Example

- If the population is normal $N(\mu, \sigma^2)$, $\gamma = 0$ and

$$\mathbb{E}\left\{\theta(F_1)|F_0\right\} = \mu^3 + n^{-1} 3\mu\sigma^2$$

- The maximum likelihood could be used to estimate $\hat{\lambda} = (\bar{x}, \hat{\sigma}^2)$
- $\theta(F_2)$ is the statistic $\hat{\theta}$ computed for a sample from a normal $(\bar{x}, \hat{\sigma}^2)$ distribution and in direct analogy we have

$$\mathbb{E}\left\{\theta(F_2)|F_1\right\} = \bar{x}^3 + n^{-1} 3\bar{x}\hat{\sigma}^2$$

- Therefore

$$\hat{\theta}_1 = 2\theta(F_1) - \mathbb{E}\left\{\theta(F_2)|F_1\right\} = \bar{x}^3 - n^{-1} 3\bar{x}\hat{\sigma}^2$$

Example

- If the population is exponential with mean μ and

$$f_\mu(x) = \mu^{-1} \exp\left(-\frac{x}{\mu}\right) \quad \text{for } x > 0$$

- here $\sigma^2 = \mu^2$ and $\gamma = 2\mu^3$

$$\mathbb{E}\left\{\theta(F_1)|F_0\right\} = \mu^3\left(1 + 3n^{-1} + 2n^{-2}\right)$$

- Taking the maximum likelihood estimate $\bar{x}$ for μ

$$\mathbb{E}\left\{\theta(F_2)|F_1\right\} = \bar{x}^3\left(1 + 3n^{-1} + 2n^{-2}\right)$$

- Therefore

$$\hat{\theta}_1 = 2\theta(F_1) - \mathbb{E}\left\{\theta(F_2)|F_1\right\} = \bar{x}^3\left(1 - 3n^{-1} - 2n^{-2}\right)$$

Example

- The estimate $\hat{\theta}_1$ represent improvement in the sense of bias reduction on the basic bootstrap estimate $\hat{\theta} = \theta(F_1)$
- To check the bias reduction observe that for general distributions with finite third moments

$$\mathbb{E}(\bar{x}^3) = \mu^3 + n^{-1}3\mu\sigma^2 + n^{-2}\gamma$$

$$\mathbb{E}(\bar{x}\hat{\sigma}^2) = \mu\sigma^2 + n^{-1}(\gamma - \mu\sigma^2) - n^{-2}\gamma$$

$$\mathbb{E}(\hat{\gamma}) = \gamma\left(1 - 3n^{-1} + 2n^{-2}\right)$$

Example

- In the case of general population

$$\mathbb{E}(\hat{\theta}_1) - \theta_0 = n^{-2}3(\mu\sigma^2 - \gamma) + n^{-3}6\gamma - n^{-4}2\gamma$$

- In the case of normal population

$$\mathbb{E}(\hat{\theta}_1) - \theta_0 = n^{-2}3\mu\sigma^2$$

- In the case of exponential population

$$\mathbb{E}(\hat{\theta}_1) - \theta_0 = -\mu^3(9n^{-2} + 12n^{-3} + 4n^{-4})$$

Remarks

- Therefore bootstrap bias reduction has diminished bias to at most $O(n^{-2})$ in each case.

- This is compared with the bias of $\hat{\theta}$ which is of size n^{-1} unless $\mu = 0$.
- Bootstrap bias correction reduces the order of magnitude of the bias by the factor n^{-1}.