Part 3: Bootstrap and EM algorithm

TMA4300: Computer Intensive Statistical Methods
(Spring 2019)

Sara Martino

*

*Slides are based on lecture notes kindly provided by Håkon Tjelmeland and Andrea Riebler.

Last part of this course

⇒ Not closely related to the two first parts

- no more MCMC
- mostly non-Bayesian perspective

⇒ Two topics (not closely related to each other):

- Bootstrapping
- Expectation-Maximization algorithm

Bootstrap

http://tradingconsequences.blogs.edina.ac.uk/files/2013/10/Dr_Martens_black_old.jpg

. . . pull oneself up by one's bootstraps

To begin an enterprise or recover from a setback without any outside help; to succeed only on one's own effort or abilities.

Wiktionary

The term is sometimes attributed to Rudolf Erich Raspe's story "The Surprising Adventures of Baron Munchausen", where the main character pulls himself (and his horse) out of a swamp by his hair

Münchhausen O. Herrfurth pinx

http://redstateeclectic.typepad.com/redstate_commentary/2010/11/sustainability-isnt-sustainable.html

Bootstrapping in statistics

Bootstrap is a computer-based technique for doing statistical inference (usually with a minimum of assumptions). It is not Bayesian.

An example for introduction

Group Treatment	Survival Time	Sample size	Mean	Estimated SE
	94,197,16,38	7	86.86	25.24
Control	99,141,23	9		
	52,104,146,10,5146		56.22	14.14
	30,40,27,46			
		Differance:	30.63	28.93

- Is the difference in mean significant?
- What if we want to compare the medians instead?
  Show code Bootstrap_into.R

Bootstrap principle

Assume we have iid observations from an (unknown) distribution F:

$$F \to (x_1, \ldots, x_n)$$

The empirical distribution function $\hat{F}$ is the CDF that puts mass $1/n$ at each data point x_i:

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

where $1(\cdot)$ denotes the indicator function.

For iid samples $\hat{F}$ is a sufficient estimator for F.

Bootstrap principle

Let θ be an interesting feature of F, $\theta = T(F)$.

For example:

$$\theta = \mathrm{E}(X) = \int x f(x) dx$$

$$\theta = \mathrm{Var}(X) = \int (x - \mathrm{E}(X))^2 f(x) dx$$

The plug-in estimator for θ is defined by:

$$\hat{\theta} = t(\hat{F})$$

The plug-in principle is quite good, if the only information about F, comes from the sample x.

Examples

Thus

$$\theta = \mathrm{E}(X) \Rightarrow \hat{\theta} = \mathrm{E}_{\hat{F}}(X) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

$$\theta = \mathrm{Var}(X) \Rightarrow \hat{\theta} = \mathrm{Var}_{\hat{F}}(X) = \mathrm{E}_{\hat{F}}[(X - \mu_{\hat{F}})^2]$$
$$= \sum_{i=1}^{n} (x_i - \mu_{\hat{F}})^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\theta = \mathrm{SD}(X) \Rightarrow \hat{\theta} = \mathrm{SD}_{\hat{F}}(X) = \sqrt{\mathrm{Var}_{\hat{F}}(X)}$$
$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Setting

Assume we have :

$$F \rightarrow (x_1, \dots, x_n)$$

Thus $\hat{F}$ gives mass $\frac{1}{n}$ to each observed value.

A bootstrap sample is defined to be a random sample of size n from $\hat{F}$, say $x^\star = (x_1^\star, \dots, x_n^\star)$

$$\hat{F} \rightarrow (x_1^\star, \dots, x_n^\star)$$

Simple illustration

Suppose $n = 3$ univariate data points, namely

$$\{x_1, x_2, x_3\} = \{1, 2, 6\}$$

are observed as an iid sample from F that has mean θ . At each observed data value, $\hat{F}$ places mass $1/3$. Suppose the estimator to be bootstrapped is the sample mean $\hat{\theta}$.

There are $3^3 = 27$ possible outcomes for $\mathcal{X}^\star = \{X_1^\star, X_2^\star, X_3^\star\}$.

Simple illustration (II)

$\mathcal{X}^\star$	$\hat{\theta}^\star$	$P^\star(\hat{\theta}^\star)$	Observed frequency
1 1 1	3/3	1/27	36/1000
1 1 2	4/3	3/27	101/1000
1 2 2	5/3	3/27	123/1000
2 2 2	6/3	1/27	25/1000
1 1 6	8/3	3/27	104/1000
1 2 6	9/3	6/27	227/1000
2 2 6	10/3	3/27	131/1000
1 6 6	13/3	3/27	111/1000
2 6 6	14/3	3/27	102/1000
6 6 6	18/3	1/27	40/1000

Bootstrap estimate for standard error

- Parameter of interest: $\theta = T(F)$
- Our estimator for θ: $\hat{\theta} = s(x)$
- Want (to estimate) $\mathrm{SD}_F(\hat{\theta})$.

A bootstrap replication of $\hat{\theta}$ is

$$\hat{\theta}^\star = s(x^\star)$$

Use plug-in principle to estimate $\mathrm{SD}_F(\hat{\theta})$.

The bootstrap estimate of the standard error of $\hat{\theta} = s(x)$ is $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$.

This is called the ideal bootstrap estimate of standard error of $\hat{\theta}$.

Ideal bootstrap estimate of standard error

- For the sample mean it can be computed analytically
- For (very) small sample sizes it can be computed using all the possible bootstrap replicates. (Number of possible bootstrap sample: n^n.)
- In other cases it can be approximated via Monte Carlo techniques

Computational way of obtaining a good estimate

We can estimate $\text{SD}_{\hat{F}}(\hat{\theta}^\star)$ by simulation:

1. Generate B bootstrap samples $x^{1\star}, \dots, x^{B\star}$.
2. Evaluate the corresponding parameter estimates

$$\hat{\theta}^\star(b) = s(x^{b\star}), \quad b = 1, 2, \dots, B$$

3. Estimate $\text{SD}_{\hat{F}}(\hat{\theta}^\star)$ by

$$\widehat{\text{SE}}_B = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\theta}^\star(b) - \hat{\theta}^\star(\cdot))^2}{B - 1}}$$

where

$$\hat{\theta}^\star(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^\star(b)$$

Note

$$\lim_{B \to \infty} \widehat{\text{SE}}_B = \widehat{\text{SE}}_\infty = \widehat{\text{SD}}_{\hat{F}}(\hat{\theta}^\star)$$

Example

Setting

$$\theta = \mathrm{E}(X)$$

$$\hat{\theta} = s(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\theta}^\star = s(x^\star) = \frac{1}{n}\sum_{i=1}^{n} x_i^\star = \bar{x}^\star$$

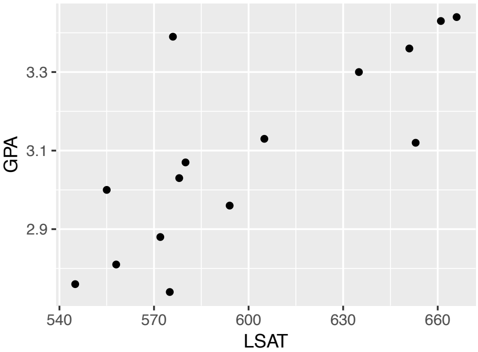
Here, the ideal bootstrap estimate exists

see blackboard

Example: The correlation coefficient

Scores for 15 law schools in the USA

$$y_i = (LSAT_i, GPA_i), \ \ t = i \dots, 15$$



The correlation between the two scores is estimated to be 0.78, but what is its standard error?

Example: The correlation coefficient

- 1000 bootstrap replicates

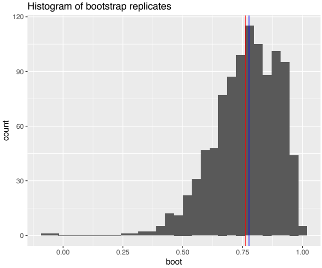
$$y^{1\star}, \ldots, y^{1000\star}$$

- For each replicates compute

$$\hat{\theta}^{i\star} = s(y^{i\star})$$

- Estimate bootstrap SE

$$\hat{SD}_{\hat{F}}(\theta) = 0.121$$



How large do we need B?

Intuitively we understand that the $\widehat{\mathsf{SE}}_B$ has larger standard deviation than $\widehat{\mathsf{SE}}_\infty$.

Theory, not to be discussed here, gives the following rules of thumb:

1. Even a small B is informative, say $B = 25$ or $B = 50$ is often enough to get a good estimate of $\mathsf{SE}_F(\hat{\theta})$.
2. Very seldomly more than $B = 200$ is necessary to estimate $\mathsf{SE}_F(\hat{\theta})$.

The parametric bootstrap

When data are modeled to originate from a parametric distribution, so

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x, \xi),$$

another estimate of F may be employed.

Suppose that the observed data are used to estimate ξ by $\hat{\xi}$. Then each parametric bootstrap pseudo-dataset $\mathcal{X}^\star$ can be generated by drawing $X_1^\star, \ldots, X_n^\star \overset{\text{iid}}{\sim} F(x, \hat{\xi}) = \hat{F}_{\text{par}}$.

Again . . .

. . . we can/must estimate $\mathrm{SD}_{\hat{F}_{\mathbf{par}}}(\hat{\theta}^\star)$ by simulation:

1. Generate B bootstrap samples $x^{1\star}, \ldots, x^{B\star}$, where

$$x^{b\star} = (x_1^{b\star}, \ldots, x_n^{b\star})$$

   with $x_1^{b\star}, \ldots, x_n^{b\star} \overset{\text{iid}}{\sim} \hat{F}_{\text{par}}$.

2. Evaluate the corresponding parameter estimates

$$\hat{\theta}^\star(b) = s(x^{b\star}), \quad b = 1, 2, \ldots, B$$

3. Estimate $\mathrm{SD}_{\hat{F}_{\mathbf{par}}}(\hat{\theta}^\star)$ by

$$\widehat{\mathrm{SE}}_B = \sqrt{\frac{\sum_{b=1}^B (\hat{\theta}^\star(b) - \hat{\theta}^\star(\cdot))^2}{B-1}}$$

   where

$$\hat{\theta}^\star(\cdot) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^\star(b)$$

Example: Correlation coefficients

We assume now that

$$y_i = (LSAT_i, GPA_i) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ i.i.d}$$

where $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{smallmatrix}\right)$ Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and obtain:

$$\hat{F}_{(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})}$$

Example: The correlation coefficient

- 1000 bootstrap replicates

$$y^{1\star}, \ldots, y^{1000\star} \sim \hat{F}_{(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})}$$

- For each replicates compute

$$\hat{\theta}^{i\star} = s(y^{i\star})$$

- Estimate bootstrap SE

$$\hat{SD}_{\hat{F}}(\theta)$$