Final Exam 12/16 In-class Exam (2 Hour, Openbook / open source / open website
Answer conceptual questions (10 questions)
Take-home Exam : Make a code for each question
=> Get an answer. (5 questions)

STA6171: Statistical Computing for DS 1

Bootstrap

Take-home (Optimization / Numerical Integration / EM : Rcpp

Ick Hoon Jin

Combinatorial Optimization / Bootstrap : R.

Yonsei University, Department of Statistics and Data Science

2020.11.04

Upload Exam question at Dec. 2nd
Due : Dec. 20. 11:59PM.
(1) Optimization.
(2) Combinatorial Optimization

EM Derivation
(3) EM algorithm 1
(4) EM algorithm 2. → EM Implementation
(5) Numerical Integration
(6) Bootstrap.

$X = (1, 2, 3, 4, 5)$

# of data point: small

Sample mean distribution

sampling with replacement.

$X^{rep}_{(1)} = (1, 1, 2, 3, 5) \Rightarrow \bar{X}^{rep}_{(1)} = 2.4$

sample (X . # of data point in X , replace = TRUE)

$X^{rep}_{(2)} = (2, 2, 2, 3, 3) \Rightarrow \bar{X}^{rep}_{(2)} = 2.4$

$\bar{X}^{rep}_{(1)}, \bar{X}^{rep}_{(2)}, \cdots, \bar{X}^{rep}_{(w)}$

$\Rightarrow$ make a distribution of statistic.

$\Rightarrow$ make an inference using distribution from bootstrapped sample

$\max(X^{rep}_{(1)}), \max(X^{rep}_{(2)}), \cdots, \max(X^{rep}_{(w)})$

Introduction

Resampling Simulate the distribution of a statistics.

- Bootstrapping is a computational intensive method that allows researchers to simulate the distribution of a statistic. (Sample from data with replacement)
- The idea is to repeatedly resample the observed data, each time producing an empirical distribution function from the resampled data.
- For each resampled data set—or equivalently each empirical distribution function—a new value of the statistic can be computed, and the collection of these values provides an estimate of the sampling distribution of the statistic of interest.
- In this manner, the method allows you to “pull yourself up by your bootstraps” (an old idiom, popularized in America, that means to improve your situation without outside help).
- Bootstrapping is nonparametric by nature, and there is a certain appeal to letting the data speak so freely.

The Bootstrap Principal

distribution function : F.
interested in $\theta = T(F)$.

- Let $\theta = T(F)$ be an interesting feature of a distribution function, F, expressed as a functional of F.
  mean of the distribution. $T(F) = \int z\, dF(z)$
- For example, $T(F) = \int z dF(z)$ is the mean of the distribution.
  data $x_1, \cdots, x_n$: realization of RV $X_1, \cdots, X_n \sim F$.
  - Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be data observed as a realization of the random variables $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim F$.
  - We use $\mathbf{X} \sim F$ to denote that $\mathbf{X}$ is distributed with density function f having corresponding cumulative distribution function F.
  - Let $\mathcal{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ denote the entire dataset.

The Bootstrap Principal

- If $\hat{F}$ is the empirical distribution function of the observed data, then an estimate of θ is $\hat{\theta} = T(\hat{F})$. For example, when θ is a univariate population mean, the estimator is the sample mean,

$$\hat{\theta} = \int z d\hat{F}(z) = \sum_{i=1}^{n} X_i / n$$

- Statistical inference questions are usually posed in terms of $T(\hat{F})$ or some $R(\mathcal{X}, F)$, a statistical function of the data and their unknown distribution function F. For example, a general test statistic might be

$$R(\mathcal{X}, F) = \frac{T(\hat{F}) - T(F)}{S(\hat{F})},$$

where S is a functional that estimates the standard deviation of $T(\hat{F})$.

The Bootstrap Principal

$R(\mathcal{X}, F)$ → we want to know the distribution of $R(\mathcal{X}, F)$

may depend on F

- The distribution of the random variable $R(\mathcal{X}, F)$ may be intractable or altogether unknown. This distribution also may depend on the unknown distribution F.
- The bootstrap provides an approximation to the distribution of $R(\mathcal{X}, F)$ derived from the empirical distribution function of the observed data.
- Let $\mathcal{X}^*$ denote a bootstrap sample of *pseudo-data*, which we will call a *pseudo-dataset*. The elements of are iid random variables with distribution $\hat{F}$.
- The bootstrap strategy is to examine the distribution of $R(\mathcal{X}^*, \hat{F})$, that is, the random variable formed by applying R to $\mathcal{X}^*$.

from the empirical distribution of the observed data → an approximation to the distribution of $R(\mathcal{X}, F)$

$\mathcal{X}^*$: bootstrap sample of pseudo-data $\overset{iid}{\sim}$ $\hat{F}$

Bootstrap => Examine the distribution of $R(\mathcal{X}^*, \hat{F})$

Simple Illustration

$n = 3$ $\{x_1, x_2, x_3\} = \{1, 2, 6\}$ x_1, x_2, x_3 observed as an iid sample from a distribution F with mean θ

mass $1/3$ $1/3$ $1/3$

- Suppose $n = 3$ univariate data points, namely $\{x_1, x_2, x_3\} = \{1, 2, 6\}$, are observed as an i.i.d. sample from a distribution F that has mean θ.
- At each observed data value, $\hat{F}$ places mass $\frac{1}{3}$.

We want to estimate sample mean $\hat{\theta}$ with bootstrap.

- Suppose the estimator to be bootstrapped is the sample mean $\hat{\theta}$, which we may write as $T(\hat{F})$ or $R(\mathcal{X}, F)$, where R does not depend on F in this case.

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

Simple Illustration

$\mathcal{X}^* = \{X_1^*, X_2^*, X_3^*\}$ iid sampled from F.

$3^3 = 27$ possible outcomes for $\mathcal{X}^*$.

- Let $\mathcal{X}^* = \{X_1^*, X_2^*, X_3^*\}$ consist of elements drawn iid from $\hat{F}$.
- There are $3^3 = 27$ possible outcomes for $\mathcal{X}^*$.
- Let $\hat{F}^*$ denote the empirical distribution function of such a sample, with corresponding estimate $\hat{\theta}^* = T(\hat{F}^*)$. $\hat{\theta}^* = T(\hat{F}^*)$
- Since $\hat{\theta}^*$ does not depend on the ordering of the data, it has only 10 distinct possible outcomes.

$\{1,2,6\}$ $\{2,1,6\}$

$\{2,6,1\}$ $\{1,6,2\}$

$\{6,1,2\}$ $\{6,2,1\}$

Simple Illustration

27 possible outcomes → ignore order. → 10 possible outcomes

TABLE 9.1 Possible bootstrap pseudo-datasets from {1, 2, 6} (ignoring order), the resulting values of $\widehat{\theta}^* = T(\widehat{F}^*)$, the probability of each outcome in the bootstrapping experiment ($P^*[\widehat{\theta}^*]$), and the observed relative frequency in 1000 bootstrap iterations.

$\mathcal{X}^*$	$\widehat{\theta}^*$	$P^*[\widehat{\theta}^*]$	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

$\widehat{\theta}^* \to T(\widehat{F})$; $P^*[\widehat{\theta}^*]$ → prob. of each outcome in bootstrapped example; Observed Frequency → real implementation

Nonparametric Bootstrap

Above illustration : sample size small ⇒ we can completely enumerate all possible cases.
⇒ For real data : complete enumeration is not possible.
↳ n^n then remove order.

- For realistic sample sizes the number of potential bootstrap pseudo-datasets is very large, so complete enumeration of the possibilities is not practical.

⇒ Generate B independent random bootstrap pseudo-dataset $\sim F$.

- Instead, B independent random bootstrap pseudo-datasets are drawn from the empirical distribution function of the observed data, namely F.

$\mathcal{X}_i^* = \{X_{i1}^*, X_{i2}^*, \cdots, X_{in}^*\}$ for $i = 1, \cdots, B$

- Denote these $\mathcal{X}_i^* = \{\mathbf{X}_{i1}^*, \cdots, \mathbf{X}_{in}^*\}$ for $i = 1, \cdots, B$.
- The empirical distribution of the $R(\mathcal{X}_i^*, \hat{F})$ for $i = 1, \cdots, B$ is used to approximate the distribution of $R(\mathcal{X}, F)$, allowing inference.

$R(\mathcal{X}_i^*, \hat{F})$ $\xrightarrow{\text{approximate}}$ $R(\mathcal{X}, F)$
$i = 1, \cdots, B$

Nonparametric Bootstrap

Complete enumeration of all possible datasets.
→ Use bootstrap simulation error ↓ B↑
simulation error

- The simulation error introduced by avoiding complete enumeration of all possible pseudo-datasets can be made arbitrarily small by increasing B.
- Using the bootstrap frees the analyst from making parametric assumptions to carry out inference, provides answers to problems for which analytic solutions are impossible, and can yield more accurate answers than given by routine application of standard parametric theory.

Free from parametric assumptions ⇒ Give solutions for the case which analytic solution is impossible.

- A fundamental requirement of bootstrapping is that the data to be resampled must have originated as an iid sample. If the sample is not iid, the distributional approximation of $R(\mathcal{X}, F)$ by $R(\mathcal{X}^*, \hat{F})$ will not hold.

Bootstrap Requirement : the data should be an iid sample.

Parametric Bootstrap

$\mathcal{X}^* = (X_1^*, \cdots, X_n^*) \sim \hat{F}$ nonparametric bootstrap.

- The ordinary nonparametric bootstrap described above generates each pseudo-dataset $\mathcal{X}^*$ by drawing $\mathbf{X}_1^*, \cdots, \mathbf{X}_n^*$ iid from $\hat{F}$.
- When the data are modeled to originate from a parametric distribution, so $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim$ iid $F(\mathbf{x}, \theta)$, another estimate of F may be employed.
- Suppose that the observed data are used to estimate θ by $\hat{\theta}$.
- Then each parametric bootstrap pseudo-dataset $\mathcal{X}^*$ can be generated by drawing $\mathbf{X}_1^*, \cdots, \mathbf{X}_n^*$. When the model is known or believed to be a good representation of reality, the parametric bootstrap can be a powerful tool, allowing inference in otherwise intractable situations and producing confidence intervals that are much more accurate than those produced by standard asymptotic theory.

When we have a model for data $\mathcal{X} = (X_1, \cdots, X_n) \sim^{iid} F(x, \theta)$,

we can generate $\mathcal{X}^* = (X_1^*, \cdots, X_n^*)$ from $F(x, \hat{\theta})$

Generate many replicates $\mathcal{X}_1^*, \cdots, \mathcal{X}_B^* \longrightarrow \theta$

Check model validation

Bootstrapping Regression

$Y_i = X_i^T \beta + \varepsilon_i$ where $i = 1, \cdots, n$.

X_i : predictor

β : parameter

- Consider the ordinary multiple regression model, $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ where the $i = 1, \cdots, n$ are assumed to be i.i.d. mean zero random variables with constant variance.
- $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are p-vectors of predictors and parameters, respectively.
- A naive bootstrapping mistake would be to resample from the collection of response values a new pseudo-response, say Y_i^*, for each observed $\mathbf{x}_i$, thereby generating a new regression dataset.

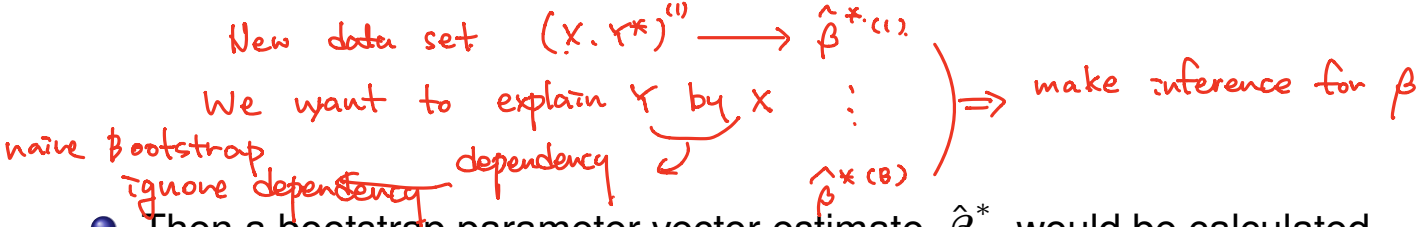
Naive approach: Each observation $X_i \longrightarrow$ resample Y_i

$$\begin{pmatrix} X_1, & Y_1^* \\ X_2, & Y_2^* \\ & \vdots \\ X_n, & Y_n^* \end{pmatrix}$$

new regression dataset.

HW 9.1, 9.4(a)
R (No Rcpp)

Bootstrapping Regression



- Then a bootstrap parameter vector estimate, $\hat{\boldsymbol{\beta}}^*$, would be calculated from these pseudo-data.
- After repeating the sampling and estimation steps many times, the empirical distribution of $\hat{\boldsymbol{\beta}}^*$ would be used for inference about $\boldsymbol{\beta}$.
- The mistake is that the $Y_i \mid \mathbf{x}_i$ are not iid - they have different conditional means. Therefore, it is not appropriate to generate bootstrap regression datasets in the manner described.

Bootstrapping Regression

Correct bootstrap we want to make iid → error term has iid assumption
↳ bootstrap from error term.

We must ask what variables are iid in order to determine a correct bootstrapping approach. The ϵ_i are iid given the model. Thus a more appropriate strategy would be to bootstrap the residuals as follows.

(Y_i, X_i) → Fit the regression $(\hat{y}_i, \hat{\epsilon}_i)$ $(\epsilon_1, \cdots \epsilon_n)$ iid $N(0, \sigma_\epsilon^2)$

- Start by fitting the regression model to the observed data and obtaining the fitted responses $\hat{y}_i$ and residuals $\hat{\epsilon}_i$.
  → bootstrap $\{\hat{\epsilon}_1^*, \cdots, \hat{\epsilon}_n^*\}$ from the set of fitted $\{\hat{y}_1, \cdots, \hat{y}_n\}$
- Sample a bootstrap set of residuals, $\{\hat{\epsilon}_1^*, \cdots, \hat{\epsilon}_n^*\}$, from the set of fitted residuals, completely at random with replacement.
- Create a bootstrap set of pseudo-responses, $Y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$, for $i = 1, \cdots, n$
  $Y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$
- Regress Y^* on $\mathbf{x}$ to obtain a bootstrap parameter estimate $\hat{\beta}^*$.
  Regress Y^* on X. ⇒ $\hat{\beta}^*$
- Repeat this process many times to build an empirical distribution for $\hat{\beta}^*$ that can be used for inference.
  $\hat{\beta}_{(1)}^*$ ... $\hat{\beta}_{(B)}^*$ → make Bootstrap inference

Bootstrapping Regression

experiment. (residual)

- This approach is most appropriate for designed experiments or other data where the **x**$_i$ values are fixed in advance.
- The strategy of bootstrapping residuals is at the core of simple bootstrapping methods for other models such as autoregressive models, nonparametric regression, and generalized linear models.
- Bootstrapping the residuals is reliant on the chosen model providing an appropriate fit to the observed data, and on the assumption that the residuals have constant variance.

constant variance
appropriate fit

- Without confidence that these conditions hold, a different bootstrapping method is probably more appropriate.

Bootstrapping Regression

- Suppose that the data arose from an observational study, where both response and predictors are measured from a collection of individuals selected at random.
- In this case, the data pairs $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ can be viewed as values observed for iid random variables $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ drawn from a joint response-predictor distribution.
- To bootstrap, sample $\mathbf{Z}_1^*, \cdots, \mathbf{Z}_n^*$ completely at random with replacement from the set of observed data pairs, $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$.
- Apply the regression model to the resulting pseudodataset to obtain a bootstrap parameter estimate $\hat{\beta}^*$.
- Repeat these steps many times, then proceed to inference as in the first approach. This approach of bootstrapping the cases is sometimes called the paired bootstrap.

Bootstrapping Regression

(1) Bootstrap error terms
(2) Bootstrap pairs

- If you have doubts about the adequacy of the regression model, the constancy of the residual variance, or other regression assumptions, the paired bootstrap will be less sensitive to violations in the assumptions than will bootstrapping the residuals.
- The paired bootstrap sampling more directly mirrors the original data generation mechanism in cases where the predictors are not considered fixed.