Grabski Franciszek

Załęska-Fornal Agata
Naval University, Gdynia, Poland

Applications of bootstrap and resampling methods in empirical Bayes estimation of reliability parameters

Keywords

bootstrap method, method, estimate, bootstrap replicates

Abstract

Bootstrap and resampling methods are the computer methods used in applied statistics. It is a type of Monte Carlo method based on observed data. Bradley Efron described it in 1979 and he has written a lot about the method and its generalizations since then. Here we apply these methods in an empirical Bayes estimation using bootstrap or resampling copies of the data to obtain an empirical prior distribution.

1. Introduction

The bootstrap is a data-based method of simulation for assessing statistical accuracy. The term bootstrap derives from the phrase *'to pull oneself up by one's bootstrap'* which can be found in the eighteenth century Adventures of Baron Munchausen by Raspe. The method was proposed by Efron. The main goal of the bootstrap method is a computer-based fulfilling of basic statistical ideas.

2. Bootstrap and resampling copies of the data

Suppose we observe independent data points $x_1, x_2, ..., x_n$, which we denoted as a vector $\mathbf{x}_n = (x_1, x_2, ..., x_n)$. This vector is a value of random vector $\mathbf{X}_n = (X_1, X_2, ..., X_n)$, where the random variables $X_1, X_2, ..., X_n$ are mutually independent and identically distributed (i.i.d.) with probability cumulative distribution function $F_\theta(\cdot)$, where $\theta \in \Theta$ is true but unknown parameter. Suppose that we are able to estimate this parameter by using estimator $\hat{\boldsymbol{\theta}}_n = T(\mathbf{X}_n)$. A number $\hat{\theta}_n = T(\mathbf{x}_n)$ is its value. After that we can use a distribution $F_{\hat{\theta}}(\cdot)$ to simulate so-called *bootstrap copies*

$$\mathbf{x}_n^{*(b)} = (x_1^{*(b)}, x_2^{*(b)}, ..., x_n^{*(b)}),\ b = 1,2,...., B$$

of data $\mathbf{x}_n = (x_1, x_2, ..., x_n)$. The bootstrap copies of data are the values of the random vectors $\mathbf{X}_n^{*(b)} = (X_1^{*(b)}, X_2^{*(b)}, ..., X_n^{*(b)}),\ b = 1,2,...., B$, which are called the *bootstrap samples.* The function $F_{\hat{\theta}_n}(\cdot)$ is a cumulative probability distribution of the i.i.d. random variables $X_1^{*(b)}, X_2^{*(b)}, ..., X_n^{*(b)}$.

If we have vector of the observation $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ of size n, we can define the empirical cumulative distribution function $\hat{F}$ as

$$\hat{F}(x; \mathbf{x}_n) = \frac{\#\{x_i : x_i \le x\}}{n}$$

that is equivalent to the discrete distribution

$$\hat{p}_k = \frac{n_k}{n}, \quad k = 1,2,..., l,$$

where

$$n_k = \#\{i : x_i = x_k\}.$$

This distribution can be expressed as a vector of frequencies $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_l)$.
Vectors of the data

$$\mathbf{x}_n^{\circ(r)} = (x_1^{\circ(r)}, x_2^{\circ(r)}, ..., x_n^{\circ(r)}),\ r = 1,2,...., R$$

coming from the distribution $\hat{F}(x;\mathbf{x})$ are said to be *resampling copies* of the data $\mathbf{x}_n = (x_1, x_2, ..., x_n)$. In other words a resampling copy $\mathbf{x}_n^{\circ} = (x_1^{\circ}, x_2^{\circ}, ..., x_n^{\circ})$ of the data $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ is generated by randomly sampling $n-$times with replacement from the original data points $\mathbf{x}_n = (x_1, x_2, ..., x_n)$. The randomly sampling means the random choice of an element among $x_1, x_2, ..., x_n$ in each of n drawings. The resampling copy of the data $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ is composed of the elements of the original sample, some of them can be taken zero times, some of them can be taken ones or twice etc. Notice that in the resampling copy $\mathbf{x}_n^{\circ} = (x_1^{\circ}, x_2^{\circ}, ..., x_n^{\circ})$, the elements are repeated as a rule.

The typical number of the bootstrap B or resampling copies of the data R, range from 50 to 1000.

3. Bootstrap and resampling estimators

There is given a following situation: a random sample $\mathbf{x}_n = (x_1, x_2, ..., x_n)$ has been observed from an unknown probability distribution $F(\cdot) = F_\theta(\cdot),\ \theta \in \Theta$ and our interest is to estimate a true parameter θ, that satisfies equation $\theta = \tilde{T}(F)$. The estimate (the value of the estimator) $\hat{\theta}$ of the parameter θ satisfies the equation

$$\hat{\theta} = T(\mathbf{x}_n) = \tilde{T}(\hat{F}). \tag{1}$$

Let us consider, as an example, an estimation of the expectation

$$\theta = m = \int_{\mathbf{R}} x dF(x).$$

An estimate of this parameter is a number

$$\hat{\theta} = \bar{x} = \int_{\mathbf{R}} F(x;\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

that is the value of a statistics

$$\hat{\boldsymbol{\theta}} = T(\mathbf{X}) = \bar{X} = \int x d\hat{F}(x;\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

Similarly the standard deviation and the quintiles and their estimates have the required representation.

Let $\mathbf{X}_n^* = (X_1^*, X_2^*, ..., X_n^*)$ be a bootstrap sample for the given vector of data $\mathbf{x}_n = (x_1, x_2, ..., x_n)$.

A random variable $\boldsymbol{\theta}_n^* = T(\mathbf{X}_n^*)$ is said to be a bootstrap estimator of the parameter θ.

A random variable $\boldsymbol{\theta}_n^{\circ} = T(\mathbf{X}_n^{\circ})$, where $\mathbf{X}_n^{\circ} = (X_1^{\circ}, X_2^{\circ}, ..., X_n^{\circ})$ is called a resampling estimator of the parameter θ.

The distribution of the statistics $\boldsymbol{\theta}_n^* - \hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n^{\circ} - \hat{\boldsymbol{\theta}}_n$ for the bootstrap sample with the fixed data values $x_1, x_2, ..., x_n$ is close to the distribution of the statistics $\hat{\boldsymbol{\theta}}_n - \theta$.

From that rule it follows that the shapes of the distributions of the statistics $\boldsymbol{\theta}_n^*, \boldsymbol{\theta}_n^{\circ}, \hat{\boldsymbol{\theta}}_n$ are similar. To obtain an empirical distribution of the random variable $\boldsymbol{\theta}_n^*$ we have to simulate bootstrap copies

$$\mathbf{x}_n^{*(b)} = (x_1^{*(b)}, x_2^{*(b)}, ..., x_n^{*(b)}),\ b = 1,2,...., B$$

of data $\mathbf{x}_n = (x_1, x_2, ..., x_n)$. After that we calculate the values of statistics

$$\theta_n^{*(b)} = T(\mathbf{x}_n^{*(b)}),\ b = 1,2,...., B$$

We can use a nonparametric kernel estimator to obtain the estimate of the probability density of the bootstrap estimate of $\boldsymbol{\theta}_n^*$. The value of this estimator with the Gaussian kernel is given by

$$\hat{g}(\vartheta) = \frac{1}{Bh}\sum_{b=1}^{B} K\left(\frac{\vartheta - \theta_n^{*(b)}}{h}\right)$$

where

$$K(\vartheta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\vartheta^2}{2}},\ \vartheta \in (-\infty, \infty),$$

and

$$h = 1.06\, s\, B^{-0.2},$$

s – standard deviation of $\theta_n^{*(b)},\ b = 1,2,...., B$.

The above-mentioned consideration can be presented on the following diagram:

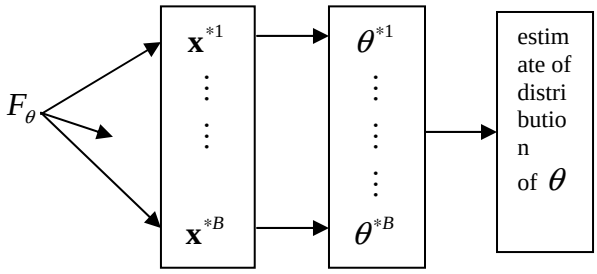

Figure 1. Diagram illustrating the realization of the bootstrap method.

4. The bootstrap estimate of the standard error

The bootstrap replication of the statistics values

$$\theta_n^{*(b)} = T(\mathbf{x}_n^{*(b)}),\ b = 1,2,....,B \tag{2}$$

correspond to the bootstrap data

$$\mathbf{x}_n^{*(b)} = (x_1^{*(b)}, x_2^{*(b)},..., x_n^{*(b)}),\ b = 1,2,....,B$$

The bootstrap estimate of the standard error of $\hat{\theta}$ is defined by the following formula

$$se_{\hat{\theta}^*} = \sqrt{\frac{\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \overline{\theta}^*\right)^2}{B-1}}, \tag{3}$$

where

$$\overline{\theta}^* = \frac{\sum_{b=1}^{B}\hat{\theta}^{*i}}{B}.$$

The bootstrap algorithm for estimating standard errors goes as follows:
- get B independent bootstrap samples

$$\mathbf{x}_n^{*(b)} = (x_1^{*(b)}, x_2^{*(b)},..., x_n^{*(b)}),\ b = 1,2,....,B$$

for estimating a standard error, (the number of B should be in the range 30-200).
- compute the bootstrap replication correspond each bootstrap sample,

$$\theta_n^{*(b)} = T(\mathbf{x}_n^{*(b)}),\quad b = 1,2,....,\ B$$

- compute the standard error $se_{\hat{\theta}^*}$ by the sample standard deviation of B replications according to (3).

5. Empirical Bayes estimation

The recent works dealing with the empirical Bayes estimation have been stimulated by the work of Robbins (1955), although early examples of the empirical Bayes approach are given by Von Mises (1942) and Von Neuman (1946).
Let $f_\theta(x)$ be a density function of a random variable X with an unknown parameter $\theta \in \Theta$ It is well known that the value of the Bayes estimator $\hat{\theta}_B$ of the parameter θ under the squared-loss function is an expectation in posterior distribution

$$\hat{\theta}_B = E(\theta \mid \mathbf{x}) = \frac{\int_\Theta \theta \mathbf{f}_\theta(\mathbf{x}) g(\theta) d\nu(\theta)}{\int_\Theta \mathbf{f}_\theta(\mathbf{x}) g(\theta) d\nu(\theta)}$$

where $\mathbf{f}_\theta(\mathbf{x}) = l(\mathbf{x};\theta)$ is a likelihood function and ν denotes a discrete counting measure or the Lebesgue measure and $g(\theta)$ is a prior density function of the parameter with respect to the measure ν. If $\hat{\theta}$ is a value of a sufficient statistics for the parameter θ then the value of the Bayes estimator $\hat{\theta}_B$ of the parameter θ is

$$\hat{\theta}_B = E(\theta \mid \hat{\theta}) = \frac{\int_\Theta \theta \tilde{\mathbf{f}}(\hat{\theta} \mid \theta) g(\theta) d\nu(\theta)}{\int_\Theta \tilde{\mathbf{f}}(\hat{\theta} \mid \theta) g(\theta) d\nu(\theta)}\ .$$

We suppose that a prior density function of the parameter mentioned above is unknown. In classical empirical Bayesian procedure a prior distribution is assessed from the *past dat*a. Very often the only date we have is the small sample $\mathbf{x} = (x_1, x_2,..., x_n)$. In those cases instead of the past data, we can use vectors

$$\mathbf{x}_n^{*(b)} = (x_1^{*(b)}, x_2^{*(b)},..., x_n^{*(b)}),\ b = 1,2,....,B,$$

that are the values of the *bootstrap samples* or the *resampling copies* of the data

$$\mathbf{x}_n^{\circ(r)} = (x_1^{\circ(b)}, x_2^{\circ(r)},..., x_n^{\circ(r)}),\ r = 1,2,....,R.$$

The resampling copies are generated independently from the empirical distribution corresponding to an unknown distribution $F(x \mid \theta)$ of a random variable X denoting (for example) time to failure. The bootstrap copies are generated from the distribution $F_{\hat{\theta}}(\cdot)$, where $\hat{\theta}_n = T(\mathbf{x}_n)$. To estimate the unknown parameter θ we have to calculate the values of the sufficient bootstrap statistics

$$\theta_n^{*(b)} = T(\mathbf{x}_n^{*(b)}),\ b = 1,2,....,B$$

or the resampling statistics

$$\theta_n^{\circ(r)} = T(\mathbf{x}_n^{\circ(r)}),\ r = 1,2,....,R$$

of that one. As a empirical prior we propose a discrete density function

$$g(\theta) = \frac{m_i}{m}\delta(\theta, \theta_n^{*(i)}),$$
$$i \in \{j_1, j_2, ..., j_w\} \subseteq \{1,...,m\}, \quad m = B$$

where

$$m_i = \#\{k: \ \theta_n^{*(k)} = \theta_n^{*(i)}\}$$

denotes the number of observations equal to $\theta_n^{*(i)}$.

$$\delta(\theta, \theta_n^{*(i)}) = \begin{cases} 1 & \text{for} \quad \theta = \theta_n^{*(i)} \\ 0 & \text{for} \quad \theta \neq \theta_n^{*(i)} \end{cases}$$

and

$$m = \sum_{i=1}^{w} m_{j_i}$$

Notice that a prior distribution is constructed on the basis on the bootstrap samples. Since

$$\hat{\theta}_B = E(\theta \mid \hat{\theta}) = \frac{\sum_{i=1}^{w} m_i \hat{\theta}_i^* f(\hat{\theta} \mid \theta_n^{*(i)})}{\sum_{i=1}^{w} m_i f(\hat{\theta} \mid \theta_n^{*(i)})} \tag{4}$$

is a value of the bootstrap empirical Bayes estimator.

6. Example

Suppose that we wish to estimate a mean time to failure $E(X) = \theta$ in the exponential distribution given by *pdf*

$$f(x \mid \theta) = \frac{1}{\theta} e^{-\frac{1}{\theta} x}, \quad x \geq 0, \quad \theta > 0.$$

In this case a likelihood function is

$$\mathbf{f}(\mathbf{x} \mid \theta) = l(\mathbf{x}; \theta) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}(x_1 + ... + x_n)} \tag{5}$$

a likelihood function for the vector of data $\mathbf{x} = (x_1, x_2, ..., x_n)$ is

$$\mathbf{f}(\mathbf{x} \mid \theta) = l(\mathbf{x}; \theta) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}(x_1 + ... + x_n)}$$

A value of the maximum likelihood estimator of θ is a value of a sufficient statistics

$$\hat{\theta} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Hence, the function (5), we can replace with a function

$$\tilde{\mathbf{f}}(\hat{\theta} \mid \theta) = \frac{1}{\theta^n} e^{-\frac{n\hat{\theta}}{\theta}}$$

In the next step we generate the bootstrap samples

$$\mathbf{x}_n^{*(k)} = (x_1^{*(k)}, x_2^{*(k)}, ..., x_n^{*(k)}), \quad k = 1,2,...,m$$

and calculate the values of the bootstrap sufficient statistics

$$\theta_n^{*(i)} = \frac{x_1^{*(i)} + x_2^{*(i)}, ..., x_n^{*(i)}}{n}, \quad i = 1,2,...,m$$

From (4) we obtain, a value of the bootstrap empirical Bayes estimator of a mean time to failure $\theta = E(X)$:

$$\hat{\theta}_B = \frac{\sum_{i=1}^{w} \frac{m_i}{(\hat{\theta}_i^*)^{n-1}} e^{-\frac{n\hat{\theta}}{\hat{\theta}_i^*}}}{\sum_{i=1}^{w} \frac{m_i}{(\hat{\theta}_i^*)^n} e^{-\frac{n\hat{\theta}}{\hat{\theta}_i^*}}}$$

By repetition we can obtain a sequence of values of a Bayes estimator that we can use to construct its empirical distribution.

7. Conclusions

There is possibility to apply the bootstrap and resampling methods in an empirical Bayes estimation. The bootstrap and resampling copies of the data are used to construct an empirical prior distribution.

References

[1] Belyaev Yu, K. (2001). Resampling and bootstrap methods in analysis of reliability data. *Proc. Safety & Reliability ESREL 2001*,1877-1882.

[2] Efron, B. & Tibshirani, R. J. (1993). *An introduction to the Bootstrap.* Chapman & Hall, New York, London.

[3] Koronacki, J. & Mielniczuk, J. (2001). *Statystyka dla studentów kierunków technicznych i przyrodniczych.* Wydawnictwa Naukowo-Techniczne, Warszawa.