Statistical foundations of machine learning

INFO-F-422

Gianluca Bontempi

Machine Learning Group
Computer Science Department
mlg.ulb.ac.be

Estimation of arbitrary parameters

- Consider a set D_N of N data points sampled from a one-dimensional distribution of a r.v. $\mathbf{z}$.
- Suppose that our quantity of interest θ is the mean of $\mathbf{z}$. It is straightforward to define the estimate $\hat{\mu}$ of the mean and to assess its quality in terms of bias and variance:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} z_i, \qquad \text{Bias}[\hat{\boldsymbol{\mu}}] = 0, \qquad \text{Var}\,[\hat{\boldsymbol{\mu}}] = \frac{\sigma^2}{N}$$

- Moreover, if we are in a parametric setting (e.g. normal), we know the entire distribution of $\hat{\boldsymbol{\mu}}$.

Estimation of arbitrary parameters

- Consider now another quantity of interest θ, for example the skewness of the distribution or a bivariate correlation. While it is easy to define an estimator of these quantities (e.g. by plug-in or maximum likelihood), their accuracy is difficult to assess.
- In other terms, **given an arbitrary estimator $\hat{\theta}$ and a nonparametric setting, the analytical form of the variance Var $\left[\hat{\theta}\right]$ and the bias Bias$[\hat{\theta}]$ is typically not available.**
- Statistical methods that do not rely on any specific assumption about the form of the probability distribution are called **nonparametric** or also **distribution-free**.

Example: the patch data

- Let us consider an example from the Efron/Tibshirani book on bootstrap.
- A pharma company wishes to introduce in the market a new medical patch designed to infuse a certain hormone in the blood.
- An experimental medical study is carried out with eight subjects.
- Each subject has his hormone levels measured after wearing three different patches: a placebo, the old' patch and the new patch.
- The goal is to show bioequivalence. In other terms the Food and Drug Administration (FDA) will approve the new patch for sale only if the new patch is bioequivalent to the old one.

Example: the patch data

- The FDA criterion is

$$\theta = \frac{|E(\text{new}) - E(\text{old})|}{E(\text{old}) - E(\text{placebo})} \leq 0.2$$

- Let us choose the following estimator

$$\hat{\boldsymbol{\theta}} = \frac{|\hat{\boldsymbol{\mu}}_{\text{new}} - \hat{\boldsymbol{\mu}}_{\text{old}}|}{\hat{\boldsymbol{\mu}}_{\text{old}} - \hat{\boldsymbol{\mu}}_{\text{placebo}}}$$

- This is a ratio estimator: if both the estimator of the numerator and of the denominator are unbiased, the ratio estimator will be biased.

Example: the patch data (II)

subj	plac	old	new	z=old-plac	y=new-old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
...	...	...	...	...	...
8	18806	29044	26325	10238	-2719
mean:				6342	-452.3

The plug-in estimate is

$$\hat{\theta} = t(\hat{F}) = \frac{|\hat{\mu}_{\text{new}} - \hat{\mu}_{\text{old}}|}{\hat{\mu}_{\text{old}} - \hat{\mu}_{\text{placebo}}} = \frac{|\hat{\mu}_y|}{\hat{\mu}_z} = \frac{452.3}{6342} = 0.07$$

Does it satisfy the FDA criterion? What about its accuracy, bias, variance?

Example: vaccine data

In Nov 2020, Pfizer and BioNTech announce that they have performed an interim analysis of an ongoing Randomized Controlled Trial (RCT) with more than 43,000 volunteers from diverse backgrounds. Their vaccine was found to be more than 90% effective in preventing Covid-19.

	infected	non infected
vaccinated	8	21500-8
no vaccine	86	21500-86

Efficacy rate$= 1 - 8/86 \approx 90\%$

Bootstrap

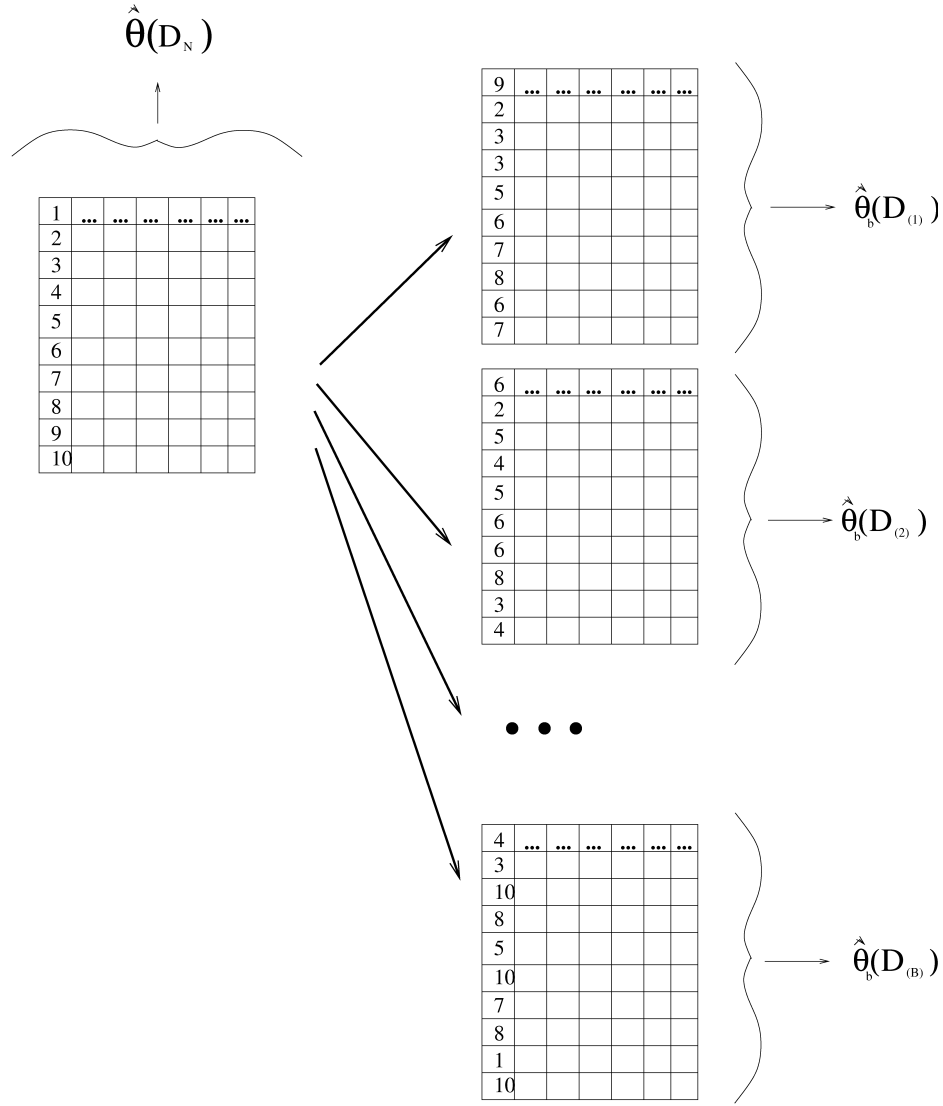
- Consider a random sample D_N from an unknown probability distribution $F_{\mathbf{z}}(\cdot, \theta)$. We wish to estimate a parameter of interest θ of the distribution. For this purpose we calculate an estimate $\hat{\theta} = g(D_N)$. How accurate is the estimator $\hat{\boldsymbol{\theta}}$? What about its sampling distribution?
- For some specific parameter (e.g. mean) or if the distribution of $\mathbf{z}$ is known, the accuracy can be estimated in analytical form.
- In most of the cases, however, we have not access to the underlying distribution. What to do?
- The method of **bootstrap** was proposed by Efron in 1979 as a computer-based technique to estimate the accuracy of $\hat{\boldsymbol{\theta}}$.

Bootstrap: why this name?

- Bootstrap is a data-based simulation method for statistical inference.
- The use of the term *bootstrap* derives from the phrase *to pull oneself up by one's bootstrap*, widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by R.E. 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.
- It is not the same as the term "bootstrap" used in computer science to "boot" a computer from a set of core instructions (though the derivation is similar).

Bootstrap (II)

- The idea of bootstrap is very simple, namely that in absence of any other information, the sample itself offers the best guide of the sampling distribution.
- The method is completely automatic, requires no theoretical calculation, and is available no matter how mathematically complicated the estimator $\hat{\boldsymbol{\theta}}$ is.
- By *resampling with replacement* from D_N we can build a set of B datasets $D_{(b)}, b = 1, \ldots, B$ all of size equal to N.
- From the empirical distribution of the statistics $g(D_{(b)})$ we can construct the sampling distribution, confidence intervals and tests for significance.


$\hat{\theta}(D_N)$
$\hat{\theta}_b(D_{(1)})$
$\hat{\theta}_b(D_{(2)})$
$\hat{\theta}_b(D_{(B)})$

Bootstrap sampling

- Consider a data set D_N.
- A **bootstrap data set** $D_{(b)}$, $b = 1, \ldots, B$ is created by randomly selecting N points from the original set D_N **with replacement**.
- Since D_N itself contains N points there is nearly always duplication of individual points in a bootstrap data set.
- The bootstrap idea is to replace the unknown distribution $F(\cdot)$ of $\mathbf{z}$ with the known empirical distribution $\hat{F}_N(\cdot)$: each sample has equal probability $1/N$ of being chosen on each draw.
- Hence, the probability that a point is chosen exactly k times in a dataset $D_{(b)}$ is given by the binomial distribution

$$\text{Prob}\{k\} = \frac{N!}{k!(N-k)!}\left(\frac{1}{N}\right)^k\left(\frac{N-1}{N}\right)^{N-k} \qquad 0 \leq k \leq N$$

Bootstrap sampling (II)

- There is a total of $\binom{2N-1}{N}$ distinct bootstrap datasets. The number is quite large already for $N > 10$ so the probability of repeating a particular resample is quite small. For instance for $N = 20$ and $B = 2000$ there is a 0.95 probability of no repetition.
- For example, if $N = 3$ and $D_N = \{a, b, c\}$, we have 10 different bootstrap sets:
{a,b,c},{a,a,b},{a,a,c},{b,b,a},{b,b,c},
{c,c,a},{c,c,b},{a,a,a},{b,b,b},{c,c,c}.
- Under **balanced bootstrap sampling** the B bootstrap sets are generated in such a way that each original data point is present exactly B times in the entire collection of bootstrap samples.

Bootstrap estimate of the variance

- Corresponding to a bootstrap dataset $D_{(b)}$, $b = 1, \ldots, B$, we can define a **bootstrap replication**

$$\hat{\theta}_{(b)} = g(D_{(b)}) \qquad b = 1, \ldots, B$$

that is the value of the statistic for the specific bootstrap sample.

- The **bootstrap estimate of the variance of the estimator** $\hat{\boldsymbol{\theta}}$, is the variance of the set $\hat{\theta}_{(b)}$, $b = 1, \ldots, B$.

$$\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)} \quad \text{where} \quad \hat{\theta}_{(\cdot)} = \frac{\sum_{b=1}^{B}\hat{\theta}_{(b)}}{B}$$

- $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ is the variance of $\hat{\boldsymbol{\theta}}$ if the distribution of $\mathbf{z}$ is $\hat{F}$.
- If $\hat{\theta} = \hat{\mu}$, for $B \to \infty$ the bootstrap estimate $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ converges to the variance $\text{Var}[\hat{\boldsymbol{\mu}}]$.

A tongue-twister

Let

- $\theta = \sigma^2$.
- $\hat{\boldsymbol{\theta}}$ the estimator of θ.
- Var $\left[\hat{\boldsymbol{\theta}}\right]$ the variance of the estimator of θ.
- $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ the bootstrap estimate of Var $\left[\hat{\boldsymbol{\theta}}\right]$

then $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ is the bootstrap estimate of the variance of the estimator of the variance....

Bootstrap estimate of bias

- Let $\hat{\boldsymbol{\theta}}$ be the plug-in estimator based on the original sample D_N and

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

- Since $\text{Bias}[\hat{\boldsymbol{\theta}}] = E[\hat{\boldsymbol{\theta}}] - \theta$, the **bootstrap estimate of the bias of the plug-in estimator** $\hat{\boldsymbol{\theta}}$ is

$$\text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\theta}_{(\cdot)} - \hat{\theta}$$

- Then, since

$$\text{Bias}[\hat{\boldsymbol{\theta}}] = E[\hat{\boldsymbol{\theta}}] - \theta \Rightarrow \theta = E[\hat{\boldsymbol{\theta}}] - \text{Bias}[\hat{\boldsymbol{\theta}}]$$

the **bootstrap bias corrected** estimate is

$$\hat{\theta}_{\text{bs}} = \hat{\boldsymbol{\theta}} - \text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\boldsymbol{\theta}} - (\hat{\theta}_{(\cdot)} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{(\cdot)}$$

- See R file `patch.R` for the estimation of bias and variance in the case of the patch data.

Bootstrap percentile confidence interval

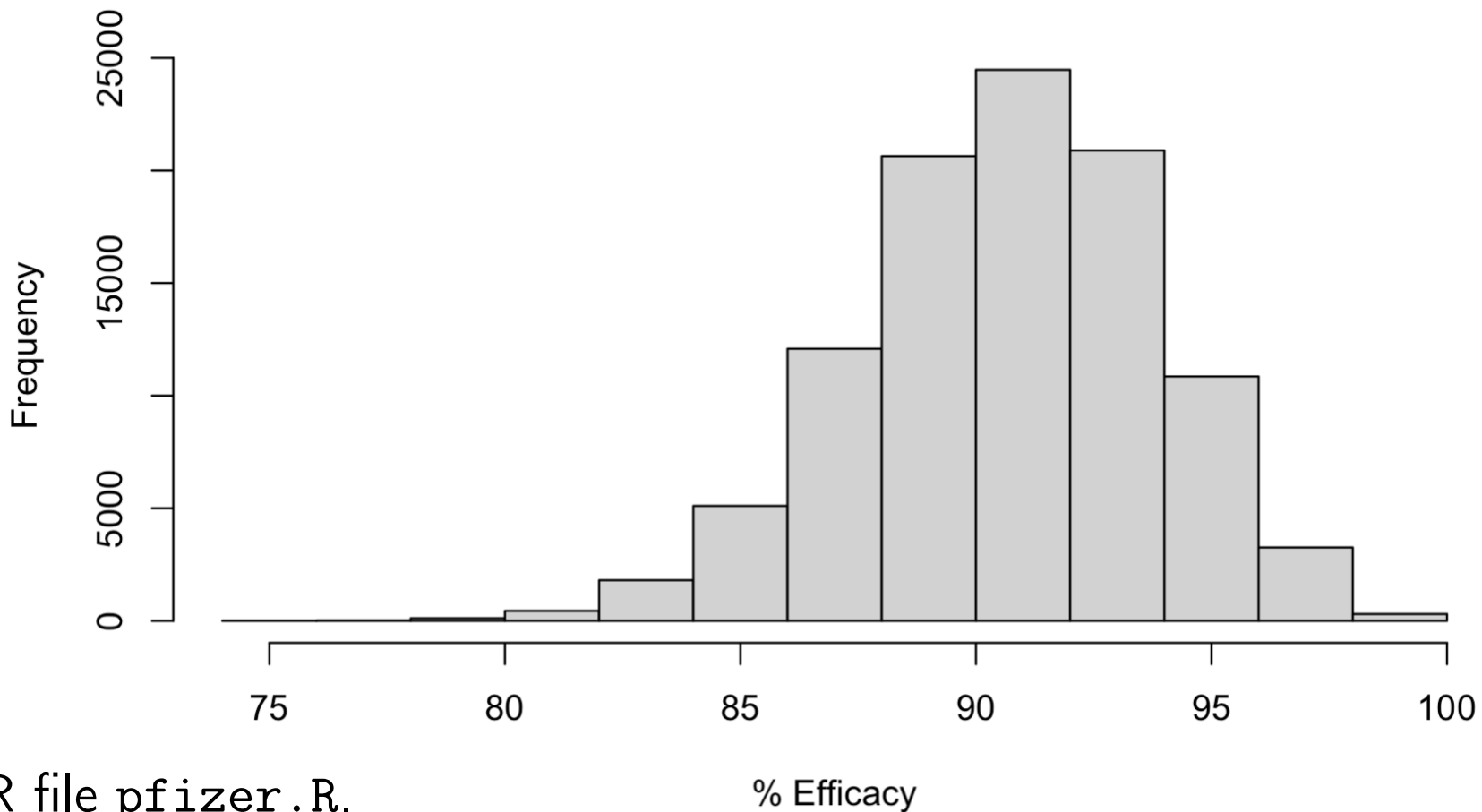
- The bootstrap approach for constructing a $100(1-\alpha)\%$ confidence interval is to use the upper and lower $\alpha/2$ values of the bootstrap distribution.
- The approaches using the full bootstrap distribution are often referred to as **percentile confidence limits**.
- If $\hat{\theta}_{L,\alpha/2}$ denotes the value such that only a fraction $\alpha/2$ of all bootstrap estimates are inferior to it, and likewise $\hat{\theta}_{H,\alpha/2}$ is the value exceeded by only $\alpha/2$ of all bootstrap estimates, then an approximate confidence interval is given by

$$[\hat{\theta}_{L,\alpha/2}, \hat{\theta}_{H,\alpha/2}]$$

also called the **Efron's percentile confidence limit**.

Example: vaccine data

Bootstrap distribution of efficacy rate:



R file `pfizer.R`.

The number of bootstrap replicates

- In bootstrap the number of replicates B can be adjusted to the computer resources.
- It can be shown that for $B \to \infty$ the bootstrap estimate of variance converge to the plug-in estimate
- In practice two "rules of thumb" are typically used:
  1. Even a small number of bootstrap replications, e.g. $B = 25$, is usually informative. $B = 50$ is often enough to give a good estimate of Var $\left[\hat{\boldsymbol{\theta}}\right]$.
  2. Very seldom are more than $B = 200$ replications needed for estimating Var $\left[\hat{\boldsymbol{\theta}}\right]$. Much bigger values of B are required for bootstrap confidence intervals.

The bootstrap principle

- What we would like to know in an estimation problem is the distribution of $\hat{\boldsymbol{\theta}} - \theta$.
- What we have in bootstrap is a Monte Carlo approximation to the distribution $\hat{\boldsymbol{\theta}}_{(b)} - \hat{\theta}$.
- The key idea of the bootstrap is that for N sufficiently large we expect the two distributions to be nearly the same.
- In other terms the variability of $\hat{\boldsymbol{\theta}}_{(b)}$ (based on the empirical distribution) around $\hat{\theta}$ is expected to be similar (or mimic) the variability of $\hat{\boldsymbol{\theta}}$ (based on the true distribution) around θ.
- There is good reason to believe this will be true for large N, since as N gets larger and larger, the empirical $\hat{F}(\cdot)$ converge to $F(\cdot)$ (see the Glivenko-Cantelli theorem for iid samples).
- This idea is sometimes referred to as the *bootstrap principle*.

Glivenko Cantelli

It is a consistency result ($N \to \infty$) for the empirical distribution

$$\sup_{-\infty<z<\infty} |\hat{\mathbf{F}}_{\mathbf{z}}(z) - F_{\mathbf{z}}(z)| \overset{N\to\infty}{\to} 0 \quad \text{with probability one}$$

Error in resampling methods

- The error in resampling methods is generally a combination of **statistical error** and **simulation error**.
- **Statistical error** is due to the difference between the underlying distribution $F(\cdot)$ and the empirical distribution $\hat{F}(\cdot)$. The magnitude of this error depends on the choice of $t(F)$.
- The use of rough statistics $t(F)$ (e.g. unsmooth or unstable) can make the resampling approach behave wildly. Example of nonsmooth statistics are sample quantiles and the median.
- **Simulation error** is due to the use of empirical (Monte Carlo) properties of $t(\hat{F})$ rather than exact properties.
- Simulation error decreases by increasing the number B of bootstrap replications.

$$\mathrm{Var}_F[\hat{\theta}] \underbrace{\approx}_{\text{depends on N}} \mathrm{Var}_{\hat{F}}[\hat{\theta}] \underbrace{\approx}_{\text{depends on B}} \mathrm{Var}_{\mathrm{bs}}[\hat{\theta}]$$

Convergence of bootstrap estimate

In general terms for iid observations, the following conditions are required for the convergence of the bootstrap estimate

1. the uniform convergence with probability one of $\hat{F}$ to F (stated by the Glivenko-Cantelli theorem) for $N \to \infty$;
2. a plug-in estimator such that the estimate $\hat{\theta}$ is the corresponding functional of the empirical distribution.

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

This is satisfied for sample means, standard deviations, variances, medians and other sample quantiles.

3. a smoothness condition on the functional. This is not true for extreme order statistics such as the minimum and the maximum values.

When might the bootstrap fail?

So far we have assumed that the dataset D_N is iid sampled from a distribution F.
In some non conventional configurations, bootstrap might fail. For example

- Too few samples ($N \leq 30$).
- Incomplete data (survival data, missing data).
- Dependent data (e.g. variance of a correlated time series).
- Dirty data (outliers)

For a critical view on bootstrap, see the publication *Exploring the limits of bootstrap* edited by Le Page and Billard which is a compilation of the papers presented at a special conference of the Institute of Mathematical Statistics held in Ann Arbor, Michigan, 1990.

Parametric bootstrap

- What we discussed so far is also called the nonparametric bootstrap since no assumption about the distribution $F_{\mathbf{z}}$ is made.
- Suppose now to know the parametric shape of the distribution $F_{\mathbf{z}}(\cdot, \theta)$.
- **Parametric bootstrap** differs from nonparametric bootstrap since the samples $D_{(b)}$ are not obtained by sampling D_N with replacement but by sampling $F_{\mathbf{z}}(\cdot, \hat{\theta})$.

Exercise

Let $\mathbf{z}$ such that $E[\mathbf{z}] = \mu$ and $\mathrm{Var}[\mathbf{z}] = \sigma^2$. Suppose we want to estimate from i.i.d. dataset D_N the parameter $\theta = \mu^2$.
Let us consider three estimators:

1. $$\hat{\theta}_1 = \left(\frac{\sum_{i=1}^{N} z_i}{N} \right)^2$$
2. $$\hat{\theta}_2 = \frac{\sum_{i=1}^{N} z_i^2}{N}$$
3. $$\hat{\theta}_3 = \frac{(\sum_{i=1}^{N} z_i)^2}{N}$$

Estimate by bootstrap the bias of the three estimators and compare this estimation with the one in the previous chapter.

Combination of two estimators

Consider two unbiased estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ of the same parameter θ

$$E[\hat{\boldsymbol{\theta}}_1] = \theta \qquad E[\hat{\boldsymbol{\theta}}_2] = \theta$$

having the same variance

$$\text{Var}\left[\hat{\boldsymbol{\theta}}_1\right] = \text{Var}\left[\hat{\boldsymbol{\theta}}_2\right] = v$$

and being uncorrelated, i.e. $\text{Cov}[\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2] = 0$.

Combination of two estimators (II)

Let $\hat{\theta}_{\text{cm}}$ be the combined estimator

$$\hat{\theta}_{\text{cm}} = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$$

This estimator has the nice properties of being unbiased

$$E[\hat{\theta}_{\text{cm}}] = \frac{E[\hat{\theta}_1] + E[\hat{\theta}_2]}{2} = \theta$$

and with a reduced variance

$$\text{Var}\left[\hat{\theta}_{\text{cm}}\right] = \frac{1}{4}\text{Var}\left[\hat{\theta}_1 + \hat{\theta}_2\right] = \frac{\text{Var}\left[\hat{\theta}_1\right] + \text{Var}\left[\hat{\theta}_2\right]}{4} = \frac{v}{2}$$

This trivial computation shows that the simple average of two unbiased estimators with a non zero variance returns a combined estimator with reduced variance.

Regularisation of an estimator

Let $\hat{\boldsymbol{\theta}}$ be un unbiased estimator of θ and

$$\hat{\boldsymbol{\theta}}_r = \lambda\theta_0 + (1-\lambda)\hat{\boldsymbol{\theta}}$$

its *regularised* version where $\theta_0 \neq \theta$ is a constant a-priori estimator and $0 < \lambda < 1$ is the regularisation parameter.

$$\text{Bias}[\hat{\boldsymbol{\theta}}_r] = E_D[\hat{\boldsymbol{\theta}}_r] - \theta = \lambda(\theta_0 - \theta) \neq 0 = \text{Bias}[\hat{\boldsymbol{\theta}}]$$

$$\text{Var}\left[\hat{\boldsymbol{\theta}}_r\right] = (1-\lambda)^2\text{Var}\left[\hat{\boldsymbol{\theta}}\right] < \text{Var}\left[\hat{\boldsymbol{\theta}}\right]$$

If

$$\lambda^2(\theta_0-\theta)^2+(1-\lambda)^2\text{Var}\left[\hat{\boldsymbol{\theta}}\right] \leq \text{Var}\left[\hat{\boldsymbol{\theta}}\right] \Leftrightarrow \lambda \leq 2\frac{\text{Var}\left[\hat{\boldsymbol{\theta}}\right]}{(\theta_0-\theta)^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right]}$$

then

$$\text{MSE}[\hat{\boldsymbol{\theta}}_r] \leq \text{MSE}[\hat{\boldsymbol{\theta}}]$$