Bootstrap methods for nonparametric curve estimation

Ricardo Cao (Universidade da Coruña, Spain)
Masaryk University (Brno, Czech Republic)

April 13-14, 2011

Contents

1 Introduction to the bootstrap

The basic ideas about the bootstrap methods are introduced in this section.

1.1 Motivation

Let us consider a simple random sample (SRS), $(X_1, X_2, \ldots, X_n)$, from a distribution function F. Consider the problem of constructing a confidence interval for the mean, μ, of F, with known standard deviation, σ. The classical statistic used for this aim is the well-known standardized diference of the sample mean and the population mean

$$T = \frac{n^{\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)}{\sigma}.$$

If F is a normal distribution then $T \overset{\mathsf{d}}{=} N(0,1)$. Since the distribution of T is known (and it is tabulated) it can be used to construct exact confidence intervals for μ. Even when the distribution of F is not normal then $T \xrightarrow{\mathsf{d}} N(0,1)$, but for samples of moderate or small size the approximation of the distribution of T by a standard normal distribution may be poor.

If F is the exponential distribution with parameter λ ($F(x) = 1 - \exp(-\lambda x)$) then

$$S = \sum_{i=1}^{n} X_i \overset{\mathsf{d}}{=} \Gamma(\lambda, n),$$

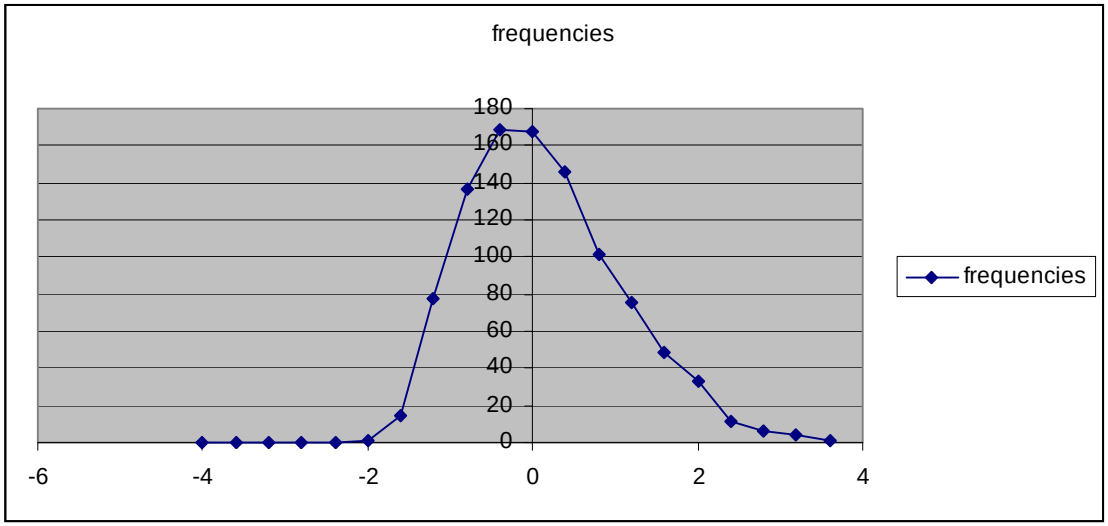

Figure 1: Distribution of T with $n = 5$ for an exponential distribution with $\lambda = 1$.

so

$$T = \frac{n^{\frac{1}{2}}\left(\frac{1}{n}S - \frac{1}{\lambda}\right)}{\frac{1}{\lambda^2}} = n^{\frac{1}{2}}\left(\frac{\lambda^2}{n}S - \lambda\right),$$

which is essentially a recentered and rescaled Gamma distribution with mean zero and variance one. However the distribution of T it is quite different from a normal if n is small.

The relevant question is: Can we use a different approximation for the distribution of T?

Key idea: mimic the data generating process that gave rise to T using the empirical distribution function, F_n, instead of F.

1.2 The bootstrap method

Let us consider a simple random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a distribution function F, and a statistic of interest, $\mathbf{R} = R(\mathbf{X}, F)$. An example of such a statistic is that one in the previous section:

$$R(\mathbf{X}, F) = \frac{n^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)}{\sigma} = \frac{n^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \int x dF(x)\right)}{\left[\int x^2 dF(x) - \left(\int x dF(x)\right)^2\right]^{1/2}}.$$

The main idea of the the bootstrap method (see Efron (1979)) is to approximate the sampling distribution of $\mathbf{R}$ by the resampling distribution of

$\mathbf{R}^* = R\left(\mathbf{X}^*, \hat{F}\right)$, where $\hat{F}$ is some estimator of the underlying cdf, F, and $\mathbf{X}^*$ is a random resample obtained from $\hat{F}$, i.e., $\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_n^*\right)$ where, conditionally on $\mathbf{X}$, the X_i^* are iid observations coming from $\hat{F}$.

A general algorithm for the bootstrap method proceeds as follows:

1. Given the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, construct some estimator, $\hat{F}$, of the true cdf, F.

2. Draw bootstrap resamples $\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_n^*\right)$ from $\hat{F}$.

3. Construct $\mathbf{R}^* = R\left(\mathbf{X}^*, \hat{F}\right)$.

4. Approximate the sampling distribution of $\mathbf{R} = R\left(\mathbf{X}, F\right)$ by the resampling distribution of $\mathbf{R}^*$.

It is very uncommon that the resampling distribution of $\mathbf{R}^*$ could be computed in a closed form. Nevertheless, one can approximate this resampling distribution by Montecarlo just repeating Steps 2-3 above a large number of times (say B) and considering B replications of the bootstrap version of our statistic: $\mathbf{R}^{*1}, \mathbf{R}^{*2}, \ldots, \mathbf{R}^{*B}$. The empirical cdf of these bootstrap values

$$\frac{1}{B} \sum_{j=1}^{B} \mathbf{1}_{\{\mathbf{R}^{*j} \leq x\}}$$

is then an approximation of the resampling distribution of $\mathbf{R}^*$, which is an approximation of the sampling distribution of $\mathbf{R}$.

A classical way to estimate F (in a nonparametric way) is to use F_n, the empirical cdf. This leads to the wellknown naïve bootstrap given by $P^*(X^* = X_i) = \frac{1}{n}$, for $i = 1, 2, \ldots, n$. In parametric setups (i.e. $F \in \{F_\theta / \theta \in \Theta\}$) it is natural to resample from $F_{\hat{\theta}}$, where $\hat{\theta}$ is an estimator of $\hat{\theta}$. This is the wellknown parametric bootstrap.

Some relevant monographs on the bootstrap are Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995) and Davison and Hinkley (1997).

1.3 Use of the bootstrap

The bootstrap method is often used for

- estimating the sampling distribution of a statistic.
- estimating the bias, variance or MSE of some estimator.
- compute confidence intervals for some parameter θ.
- perform hypothesis tests for some parameter θ.

1.4 An example: the median

Let us consider a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a distribution F with median $\theta(F) = F^{-1}\left(\frac{1}{2}\right)$. Suppose that the sample size is odd: $n = 2m - 1$. By considering the ordered statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$, the sample median is just the central ordered statistic: $\hat{\theta} = X_{(m)}$. In order to construct confidence intervals for θ we consider the statistic $\mathbf{R} = R(\mathbf{X}, F) = X_{(m)} - \theta$. The bootstrap version of $\mathbf{R}$ is $\mathbf{R}^* = R(\mathbf{X}^*, F_n) = X^*_{(m)} - \theta(F_n) = X^*_{(m)} - X_{(m)}$, where $\mathbf{X}^*$ is a random resample coming from F_n.

In this setup it is possible to obtain a closed expression for the bootstrap distribution (this is very exceptional):

$$\begin{aligned} P^*\left(X^*_{(m)} \leq x\right) &= P^*\left(\#\left\{X^*_i \leq x / i = 1, 2, \ldots, n\right\} \geq m\right) \\ &= \sum_{j=m}^{n} \binom{n}{j} F_n(x)^j \left(1 - F_n(x)\right)^{n-j}. \end{aligned}$$

Consequently,

$$\begin{aligned} P^*\left(X^*_{(m)} = X_{(i)}\right) &= P^*\left(X^*_{(m)} \le X_{(i)}\right) - P^*\left(X^*_{(m)} < X_{(i)}\right) \\ &= \sum_{j=m}^{n} \binom{n}{j} \left(\frac{i}{n}\right)^j \left(\frac{n-i}{n}\right)^{n-j} \\ &\quad - \sum_{j=m}^{n} \binom{n}{j} \left(\frac{i-1}{n}\right)^j \left(\frac{n-i+1}{n}\right)^{n-j} \\ &= \sum_{j=m}^{n} \binom{n}{j} a_{ijn}, \end{aligned}$$

where

$$a_{ijn} = \left(\frac{i}{n}\right)^j \left(\frac{n-i}{n}\right)^{n-j} - \left(\frac{i-1}{n}\right)^j \left(\frac{n-i+1}{n}\right)^{n-j}.$$

So, the bootstrap distribution of $\mathbf{R}^*$ is

$$P^*\left(\mathbf{R}^* = X_{(i)} - X_{(m)}\right) = \sum_{j=m}^{n} \binom{n}{j} a_{ijn}.$$

2 Introduction to nonparametric curve estimation

In the following subsections some nonparametric methods for density and regression estimation are introduced.

2.1 Nonparametric density estimation

2.1.1 Background

Let $(X_1, X_2, \ldots, X_n)$ be a SRS from a population with distribution function F, absolutelly continuous, density function f. The kernel density estimator proposed by Parzen (1962) and Rosenblatt (1956) is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$, K is a kernel function (typically a symmetric density) and $h > 0$ is a smoothing parameter, often called bandwidth, that regulated the size of the neighbour used for the estimation. This estimator generalizes the wellknown histogram, more precisely the moving histogram. Choosing K as the density of a $U(-1,1)$, the Parzen-Rosenblatt estimator results in:

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\mathbf{1}\left\{\frac{x-X_i}{h}\in(-1,1)\right\} &= \frac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}\left\{X_i\in(x-h,x+h)\right\}\\ &= \frac{\#\left\{X_i\in(x-h,x+h)\right\}}{2nh},\end{aligned}$$

which is the relative frequency of data X_i in the interval $(x-h,x+h)$ divided by the length of this intervalo $(2h)$.

It is typically assumed that the kernel function, K, is nonnegative and integrates out to one:

$$K(u)\geq 0,\ \forall u,\ \int_{-\infty}^{\infty}K(u)\,du = 1.$$

It is also common to assume that K is a symmetric function: $K(-u) = K(u)$.

The choice of the function K does not have a big impact in the properties of the estimator (just in its regularity: continuity, differentiability, etc.) but the

choice of the smoothing parameter is crucial for a correct estimation. In other words, the size of the neighbourhood for the nonparametric estimation should be adecuate (not to large, not too small).

2.1.2 Bias, variance and mean squared error

Straight forward calculations lead the bias of the Parzen-Rosenblatt estimator:

$$\begin{aligned} Bias\left(\hat{f}_h(x)\right) &= E\left(\hat{f}_h(x)\right) - f(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)\, dy - f(x) \\ &= (K_h * f)(x) - f(x), \end{aligned}$$

where $*$ denotes the convolution operator:

$$(f * g)(x) = \int f(x-y)\, g(y)\, dy.$$

Using the bias expression an asymptotic expression for the bias can be obtained:

$$E\left(\hat{f}_h(x)\right) - f(x) = \frac{d_K}{2} h^2 f''(x) + O\left(h^4\right),$$

with $d_K = \int t^2 K(t)\, dt$.

The variance can be handled similarly:

$$\begin{aligned} Var\left(\hat{f}_h(x)\right) &= \frac{1}{nh^2} Var\left(K\left(\frac{x-X_1}{h}\right)\right) \\ &= \frac{1}{nh^2}\left[\int K\left(\frac{x-y}{h}\right)^2 f(y)\,dy - \left(\int K\left(\frac{x-y}{h}\right) f(y)\,dy\right)^2\right] \\ &= \frac{1}{n}\left[\left((K_h)^2 * f\right)(x) - \left((K_h * f)(x)\right)^2\right] \\ &= \frac{1}{nh}\left[\left(K^2\right)_h * f\right](x) - \frac{1}{n}\left[(K_h * f)(x)\right]^2. \end{aligned}$$

Its asymptotic expression results in:

$$Var\left(\hat{f}_h(x)\right) = \frac{c_K}{nh} f(x) - \frac{1}{n} f(x)^2 + O\left(\frac{h}{n}\right),$$

with $c_K = \int K(t)^2\,dt$.

Consequently the mean squared error of the estimator is:

$$\begin{aligned} MSE\left(\hat{f}_h(x)\right) &= \int \left(\hat{f}_h(x) - f(x)\right)^2 dx = Sesgo\left(\hat{f}_h(x)\right)^2 + Var\left(\hat{f}_h(x)\right) \\ &= \left[(K_h * f)(x) - f(x)\right]^2 + \frac{1}{nh}\left[\left(K^2\right)_h * f\right](x) \\ &\quad -\frac{1}{n}\left[(K_h * f)(x)\right]^2. \end{aligned}$$

Its asymptotic expression is:

$$MSE\left(\hat{f}_h(x)\right) = \frac{d_K^2}{4}h^4 f''(x)^2 + \frac{c_K}{nh}f(x) - \frac{1}{n}f(x)^2 + O\left(h^6\right) + O\left(\frac{h}{n}\right).$$

2.1.3 Mean integrated squared error (MISE)

A global error measure (not for a particular x) of the estimator is them mean integrated squared error:

$$\begin{aligned} MISE\left(\hat{f}_h(x)\right) &= \int E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] dx = \int MSE\left(\hat{f}_h(x)\right) dx = \\ &\int \left[(K_h * f)(x) - f(x)\right]^2 dx + \frac{c_K}{nh} - \frac{1}{n}\int \left[(K_h * f)(x)\right]^2 dx. \end{aligned}$$

An asymptotic expression for it is the following:

$$MISE\left(\hat{f}_h(x)\right) = \frac{d_K^2}{4}h^4\int f''(x)^2\,dx + \frac{c_K}{nh} - \frac{1}{n}\int f(x)^2\,dx + O\left(h^6\right) + O\left(\frac{h}{n}\right).$$

The negative effect of choosing a too large or too small bandwidth (h) is evident from this expression.

2.2 Nonparametric regression estimation

Let $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ de a SRS from a two-dimensional population (X, Y), with $E(|Y|) < \infty$. We would like to estimate the regression function of Y given X: $m(x) = E(Y|_{X=x})$. The regression function can be written as:

$$\begin{aligned} m(x) &= \int y f_{2|1}(y|_x)\,dy = \int y\frac{f(x,y)}{f_1(x)}dy = \frac{\int y f(x,y)\,dy}{f_1(x)} \\ &= \frac{\int y f_{1|2}\left(x|_y\right) f_2(y)\,dy}{f_1(x)} = \frac{\Psi(x)}{f_1(x)}, \end{aligned}$$

where $f_1(x)$ is the marginal density function of X and

$$\Psi(x) = \int y f_{1|2}\left(x|_y\right) f_2(y)\, dy = E\left(Y f_{1|2}\left(x|_Y\right)\right).$$

The functions $\Psi(x)$ and $f_1(x)$ can be estimated using the kernel method:

$$\begin{aligned} \hat{f}_{1,h}(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \\ \hat{\Psi}_h(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i, \end{aligned}$$

which give the Nadaraya-Watson kernel estimator (see Nadaraya (1964) and Watson (1964)):

$$\hat{m}_h(x) = \frac{\hat{\Psi}_h(x)}{\hat{f}_{1,h}(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\, Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}.$$

Similar properties to those mentioned for the Parzen-Rosenblatt kernel density estimator can be proved for the regression estimator.

3 Bootstrap methods for density estimation

In the next two subsections two different problems in nonparametric density estimation are considered: approximating the sampling distribution of the Parzen-Rosenblatt kernel estimator (to obtain confidence intervals for the density, for instance) and bandwidth selection. The bootstrap is used for these two aims.

3.1 Bootstrap approximation for the sampling distribution of the Parzen-Rosenblatt kernel estimator

Before introducing the bootstrap in this setup the limit distribution of the Parzen-Rosenblatt estimator will be presented. Other approximations will be also considered (see Cao (1990) for futher details on these results).

3.1.1 Asymptotic distribution of the Parzen-Rosenblatt estimator

The minimal requirements for the bias and variance to tend to zero when the sample size tends to infinity are $h \to 0$, $nh \to \infty$. Under these assumptions

$$(nh)^{1/2} \left(\hat{f}_h(x) - f(x) \right) \xrightarrow{d} N(B, V).$$

On the other hand, it can be proved that the asymptotically optimal value of h, in the sense of MSE, is $h = c_0 n^{-1/5}$, with

$$c_0 = \left(\frac{c_K f(x)}{d_K^2 f''(x)^2} \right)^{1/5}.$$

This choice for h leads to the following values for the mean and the variance of the normal limit distribution:

$$\begin{aligned} B &= \frac{1}{2} c_0^{5/2} d_K f''(x), \\ V &= c_K f(x). \end{aligned}$$

In order to use the limit distribution to construct confidence intervals for $f(x)$ one can ...

1. estimate B and V and use them in the normal distribution (**plug-in method**).

2. design a resampling plan and use the **bootstrap method**.

3.1.2 Plug-in approximation

It consists in estimating B and V by means of

$$\begin{aligned} \hat{B} &= \frac{1}{2}\hat{c}_0^{5/2} d_K \hat{f}_g''(x), \\ \hat{V} &= c_K \hat{f}_h(x), \end{aligned}$$

where g is a suitable bandwidth to estimate the second derivative of the density function. Using the Berry-Esséen inequality the following rate of convergence can be obtained:

$$\sup_{z \in \boldsymbol{R}} \left| P\left[(nh)^{1/2} \left(\hat{f}_h(x) - f(x) \right) \leq z \right] - \Phi\left(\frac{z - \hat{B}}{\hat{V}} \right) \right| = O_P\left(n^{-1/5} \right).$$

This rate is worse than that of the theoretical normal approximation, based on the exact mean and variance (B_n and V_n):

$$\sup_{z \in \boldsymbol{R}} \left| P\left[(nh)^{1/2} \left(\hat{f}_h(x) - f(x) \right) \leq z \right] - \Phi\left(\frac{z - B_n}{V_n} \right) \right| = O\left(n^{-2/5} \right),$$

but is not worse than that of the asymptotic normal, $N(B, V)$, which has a rate of order $O_P\left(n^{-1/5}\right)$.

3.1.3 Bootstrap approximation

The bootstrap resampling plan is as follows:

1. Use the sample $(X_1, X_2, \ldots, X_n)$ and a **pilot bandwidth**, g, to compute the Parzen-Rosenblatt estimator, $\hat{f}_g$.

2. Draw bootstrap resamples $\left(X_1^*, X_2^*, \ldots, X_n^*\right)$ from the density $\hat{f}_g$.

3. Construct the bootstrap version of the Parzen-Rosenblatt estimator

$$\hat{f}_h^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right).$$

4. Approximate the sampling distribution of $(nh)^{1/2}\left(\hat{f}_h(x) - f(x)\right)$ by means of the resampling distribution of $(nh)^{1/2}\left(\hat{f}_h^*(x) - \hat{f}_g(x)\right)$.

If we were interested in the bias or the variance of $\hat{f}_h(x)$ (rather than in its asymptotic distribution) then Step 4 in the previous algorithm will be replaced by computing the bootstrap version of the bias, $E^*\left(\hat{f}_h^*(x) - \hat{f}_g(x)\right)$, or the variance, $Var^*\left(\hat{f}_h^*(x)\right)$.

In the previous algorithm, the bandwidth g has to be asymptotically larger than h. In fact, a reasonable choice for g is the minimizer of $E\left[\left(\hat{f}_g''(x) - f''(x)\right)^2\right]$. Asymptotically, this minimizer has the form

$$g \simeq \left(\frac{5f(x)\int K''(t)^2\,dt}{d_K^2 f^{(4)}(x)^2\, n}\right)^{1/9}.$$

The convergence rate for the bootstrap aproximation is given by

$$\begin{aligned} & \sup_{z \in \boldsymbol{R}} \left| P\left[(nh)^{1/2}\left(\hat{f}_h(x) - f(x)\right) \leq z\right] - P^*\left[(nh)^{1/2}\left(\hat{f}_h^*(x) - \hat{f}_g(x)\right) \leq z\right] \right| \\ = \; & O_P\left(n^{-2/9}\right), \end{aligned}$$

which is better than those of the theoretical normal aproximation and the plug-in method.

3.2 Bootstrap methods for bandwith selection

3.2.1 Asymptotic expression for the optimal bandwidth

The $MISE$ has an asymptotic expression that may be used as a criterion to obtain an optimal value for the smoothing parameter:

$$MISE(h) = AMISE(h) + O\left(h^6\right) + O\left(\frac{h}{n}\right),$$

with

$$AMISE\left(h\right)=\frac{d_K^2}{4}h^4\int f''\left(x\right)^2dx+\frac{c_K}{nh}-\frac{1}{n}\int f\left(x\right)^2dx.$$

The smoothing parameter that minimizes $AMISE$ is

$$h_{AMISE}=\left(\frac{c_K}{nd_K^2\int f''\left(x\right)^2dx}\right)^{1/5}.$$

There exist plenty of methods devoted to the problem of bandwidth selection. Among them we mention plug-in methods, cross validation methods (smoothed or not) and, of course, bootstrap methods (see, for instance, Marron (1992)).

3.2.2 Bootstrap version of $MISE$

The basic idea consists in providing a smoothed bootstrap resampling plan to estimate $MISE$. We will follow the proposal by Cao (1993). It consists of the following steps:

1. Given the sample $(X_1, X_2, \ldots, X_n)$ a pilot bandwidth, g, is used to compute the Parzen-Rosenblatt kernel estimator, $\hat{f}_g$.

2. Bootstrap resamples $\left(X_1^*, X_2^*, \ldots, X_n^*\right)$ are drawn from the density $\hat{f}_g$.

3. For every $h > 0$, the bootstrap version of the Parzen-Rosenblatt estimator is computed

$$\hat{f}_h^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right).$$

4. The bootstrap version of $MISE$ is constructed:

$$MISE^*(h) = \int E^*\left[\left(\hat{f}_h^*(x) - \hat{f}_g(x)\right)^2\right] dx.$$

5. $MISE^*(h)$ is minimized in $h > 0$ and the bootstrap selector is obtained:

$$h_{MISE}^* = \arg\min_{h>0} MISE^*(h)$$

3.2.3 Closed expression for $MISE^*$

It is possible to obtain a closed expression for the bootstrap version of $MISE$:

$$\begin{aligned} MISE^*(h) &= \int \left[\left(K_h * \hat{f}_g\right)(x) - \hat{f}_g(x)\right]^2 dx \\ &\quad + \frac{c_K}{nh} - \frac{1}{n}\int \left[\left(K_h * \hat{f}_g\right)(x)\right]^2 dx \\ &= \frac{c_K}{nh} - \frac{1}{n^3}\sum_{i,j=1}^{n} \left[(K_h * K_g) * (K_h * K_g)\right]\left(X_i - X_j\right) \\ &\quad + \frac{1}{n^2}\sum_{i,j=1}^{n} \left[(K_h * K_g - K_g) * (K_h * K_g - K_g)\right]\left(X_i - X_j\right). \end{aligned}$$

Such a property is not very common for the bootstrap in other setups.

3.2.4 Pilot bandwidth choice

The bandwidth selection problem for the piloto smoothing parameter, g, is close related to minimizing the mean squared error of the density curvature:

$$E\left[\left(\int \hat{f}_g''(x)^2\,dx - \int f''(x)^2\,dx\right)^2\right].$$

The asymptotical value of the pilot bandwidth, g, minimizing the previous expression is

$$g_0 = \left(\frac{\int K''(t)^2\,dt}{n d_K \int f^{(3)}(x)^2\,dx}\right)^{1/7}.$$

3.2.5 Asymptotic results

Using any deterministic piloto bandwidth with the following property $\frac{g-g_0}{g_0} = O\left(n^{-1/14}\right)$, it holds

$$\begin{aligned} \frac{h^*_{MISE} - h_{MISE}}{h_{MISE}} &= O_P\left(n^{-5/14}\right), \\ \frac{MISE\left(h^*_{MISE}\right) - MISE\left(h_{MISE}\right)}{MISE\left(h_{MISE}\right)} &= O_P\left(n^{-5/7}\right). \end{aligned}$$

Using somewhat more sofisticated techniques (that let the pilot bandwidth, g, depend on h), a slightly better rates can be obtained:

$$\frac{h^*_{MISE} - h_{MISE}}{h_{MISE}} = O_P\left(n^{-1/2}\right).$$

3.2.6 Gaussian kernel case

If the kernel function, K, is Gaussian (the density la function of a $N(0,1)$), then:

- K_h is the density of a $N\left(0, h^2\right)$
- K_g is the density of a $N\left(0, g^2\right)$
- $K_h * K_g$ is the density of a $N\left(0, h^2 + g^2\right)$
- $(K_h * K_g) * (K_h * K_g)$ is the density of a $N\left(0, 2h^2 + 2g^2\right)$
- $(K_h * K_g) * K_g$ is the density of a $N\left(0, h^2 + 2g^2\right)$
- $K_g * K_g$ is the density of a $N\left(0, 2g^2\right)$

Consequently,

$$
\begin{aligned}
MISE^{*}(h) \;=\; & \frac{c_K}{nh} - \frac{1}{n^3}\sum_{i,j=1}^{n} K_{\left(2h^2+2g^2\right)^{1/2}}\left(X_i - X_j\right) \\
& + \frac{1}{n^2}\sum_{i,j=1}^{n}\Big[K_{\left(2h^2+2g^2\right)^{1/2}}\left(X_i - X_j\right) \\
& - 2K_{\left(h^2+2g^2\right)^{1/2}}\left(X_i - X_j\right) + K_{\left(2g^2\right)^{1/2}}\left(X_i - X_j\right)\Big].
\end{aligned}
$$

3.2.7 Comparison with other bandwidth selectors

The bootstrap method presented here is very similar to the smoothed cross validation method proposed by Hall, Marron and Park (1992). In comparative simulation studies (see Cao, Cuevas and González-Manteiga (1993)) it can be observed that this bootstrap method is very competitive with other methods for bandwidth selection. In general this bootstrap method is the one which presents a better behaviour, together with the solve-the-equation plug-in method by Sheather and Jones (1991) and the smooth cross validation.

There exist other bootstrap bandwidth selectors that exhibit a much worse behaviour. Among them we include:

- Hall (1990), that resamples from the empirical cdf, so it does not mimic the bias.

- Faraway and Jhun (1990), that choose g as the cross validation bandwidth, that results to be too small.

- Taylor (1989), that chooses $g = h$, so $MISE^*(h) \to 0$, when $h \to \infty$, which produces a global minimum of $MISE^*$ that is inconsistent with h_{MISE}.

4 Bootstrap methods for nonparametric regression estimation

In this section, two bootstrap methods are presented for nonparametric regression. The aim is to approximate the sampling distribution of the Nadaraya-Watson estimator. The asymptotic results show the behaviour of these bootstrap methods, conditionally on the sample of the explanatory variable as well as in an unconditional sense.

4.1 Asymptotic distribution of the Nadaraya-Watson estimator

Before embarking on presenting the bootstrap in this context, it is useful to present the limit distribution of the Nadaraya-Watson estimator, given by

$$\hat{m}_h(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\, Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}.$$

As in the density case, it can be proved that the minimal conditions required for the consistency of the estimator are $h \to 0$, $nh \to \infty$, when $n \to \infty$. Under these assumptions,

$$(nh)^{1/2} \left(\hat{m}_h (x) - m (x) \right) \xrightarrow{d} N (B, V) .$$

On the other hand, it can be proved that the asimptotically optimal value for h, in the sense of MSE, is of the form $h = c_0 n^{-1/5}$. For such a bandwidth, the mean and variance for the normal limit distribution are

$$\begin{aligned} B &= \frac{1}{2} c_0^{5/2} d_K \frac{m'' (x) f (x) + 2 m' (x) f' (x)}{f (x)}, \\ V &= c_K \frac{\sigma^2 (x)}{f (x)}, \end{aligned}$$

where $f (x)$ is the marginal density function of X and $\sigma^2 (x) = Var (Y|_{X=x})$ is the conditional la variance of Y given $X = x$.

As for the density case, to construct confidence intervals for $m (x)$ we may

1. Estimate B and V and use them in the corresponding normal distribution (**plug-in method**).

2. Design a resampling plan using the **bootstrap method**.

The rates of convergence for the approximation of the (conditional or unconditional) distribution of the statistic to the normal limit distribution are:

$$\sup_{z\in\boldsymbol{R}}\left|P^{Y|_X}\left[(nh)^{1/2}\left(\hat{m}_h(x)-m(x)\right)\leq z\right]-\Phi\left(\frac{z-B}{V}\right)\right| = O_P\left(n^{-1/5}\right),$$

$$\sup_{z\in\boldsymbol{R}}\left|P\left[(nh)^{1/2}\left(\hat{m}_h(x)-m(x)\right)\leq z\right]-\Phi\left(\frac{z-B}{V}\right)\right| = O\left(n^{-2/5}\right),$$

where $P^{Y|_X}(B)$ denotes $P\left(B|_{X_1,X_2,\ldots,X_n}\right)$.

4.2 Plug-in approximation

The plug-in approximation consists in estimating B and V using suitable estimators of $f(x)$, $f'(x)$, $m(x)$, $m'(x)$, $m''(x)$ and $\sigma^2(x)$. For any of these functions one could use bandwidth selectors intended to approximate the optimal bandwidths for every one (this is a rather tedious process). Using such

an aproach, estimators for the bias, $\hat{B}$, and the variance, $\hat{V}$, can be obtained. It may be proved for these estimators that $\hat{B} - B = O_P\left(n^{-2/9}\right)$ and $\hat{V} - V = O_P\left(n^{-2/5}\right)$. Consequently the following (conditional and unconditional) convergence rates can be proved for the plug-in approximation:

$$\sup_{z \in \boldsymbol{R}} \left| P^{Y|X}\left[(nh)^{1/2}\left(\hat{m}_h(x) - m(x)\right) \leq z\right] - \Phi\left(\frac{z - \hat{B}}{\hat{V}}\right)\right| = O_P\left(n^{-1/5}\right),$$

$$\sup_{z \in \boldsymbol{R}} \left| P\left[(nh)^{1/2}\left(\hat{m}_h(x) - m(x)\right) \leq z\right] - \Phi\left(\frac{z - \hat{B}}{\hat{V}}\right)\right| = O_P\left(n^{-2/9}\right).$$

The first rate is worse than and the second one is the same as the rate for the theoretical normal limit approximation (see Cao (1991)).

4.3 Wild bootstrap

This bootstrap resampling method, proposed by Wu (1986) and estudied by Härdle and Marron (1991), proceeds as follows:

1. Using the Nadaraya-Watson estimator of $m(x)$ and using the initial smoothing parameter, h, the residuals are constructed $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$, $i = 1, 2, \ldots, n$.

2. For every index $i = 1, 2, \ldots, n$, and conditionally on the observed sample, $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, a bootstrap residual, $\hat{\varepsilon}_i^*$, is drawn from a probability distribution fulfilling $E^*(\hat{\varepsilon}_i^*) = 0$, $E^*\left(\hat{\varepsilon}_i^{*2}\right) = \hat{\varepsilon}_i^2$ and $E^*\left(\hat{\varepsilon}_i^{*3}\right) = \hat{\varepsilon}_i^3$. Although the third moment condition is not strictly necessary, it is useful to prove the asymptotic validity of the method.

3. Using a pilot bandwidth, g, asymptotically larger than h (i.e. $g/h \to \infty$), bootstrap versions of the observations from the response variable are drawn: $Y_i^* = \hat{m}_g(X_i) + \hat{\varepsilon}_i^*$, $i = 1, 2, \ldots, n$.

4. The bootstrap resample $\left\{\left(X_1, Y_1^*\right), \left(X_2, Y_2^*\right), \ldots, \left(X_n, Y_n^*\right)\right\}$ is used to construct a bootstrap version of the Nadaraya-Watson estimator:

$$\hat{m}_h^*(x) = \frac{\frac{1}{n}\sum_{i=1}^n K_h(x - X_i) Y_i^*}{\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)}.$$

5. The sampling distribution of $(nh)^{1/2}\left(\hat{m}_h(x) - m(x)\right)$ is approximated by the resampling distribution of $(nh)^{1/2}\left(\hat{m}_h^*(x) - \hat{m}_g(x)\right)$.

Step 2 is usually carried out by considering a random variable, V^*, fulfilling $E^*(V^*) = 0$, $E^*\left(V^{*2}\right) = 1$ and $E^*\left(V^{*3}\right) = 1$, drawing a sample of size n from it, $\left(V_1^*, V_n^*, \ldots, V_n^*\right)$, and then defininining $\hat{\varepsilon}_i^* = \hat{\varepsilon}_i V_i^*$ for $i = 1, 2, \ldots, n$.

A common choice for the distribution of V^* is the discrete distribution that gives positive probability to only two points: $P^*(V^* = a) = p$ and $P^*(V^* = b) = 1 - p$. The distribution can be obtained as the solution of the system of three equations given by the first three moments:

$$\begin{aligned} ap + b(1-p) &= 0, \\ a^2 p + b^2(1-p) &= 1, \\ a^3 p + b^3(1-p) &= 1. \end{aligned}$$

This gives rise to the so called wild bootstrap (or golden section bootstrap), with $a = \frac{1-5^{1/2}}{2}$, $b = \frac{1+5^{1/2}}{2}$, $p = \frac{5+5^{1/2}}{10}$, i.e.

$$\begin{aligned} P^*\left(V^* = \frac{1-5^{1/2}}{2}\right) &= \frac{5+5^{1/2}}{10} \\ P^*\left(V^* = \frac{1+5^{1/2}}{2}\right) &= \frac{5-5^{1/2}}{10} \end{aligned}$$

The selection of the pilot bandwidth g, appearing in Step 3, is strongly linked to the estimation of $m''(x)$, since this is the critical term to estimate B and V. Taking a pilot bandwidth of optimal order in this sense, $g_0 \simeq d_0 n^{-1/9}$, the following (conditional and unconditional) rates of convergence for the wild

bootstrap approximation can be obtained:

$$\begin{aligned} & \sup_{z\in\boldsymbol{R}} \left| P^{Y|_X}\left[(nh)^{1/2}\left(\hat{m}_h(x)-m(x)\right)\le z\right]- \right. \\ & \left. -P^*\left[(nh)^{1/2}\left(\hat{m}_h^*(x)-\hat{m}_g(x)\right)\le z\right]\right| \\ = \; & O_P\left(n^{-2/9}\right), \\ & \sup_{z\in\boldsymbol{R}} \left| P\left[(nh)^{1/2}\left(\hat{m}_h(x)-m(x)\right)\le z\right]- P^*\left[(nh)^{1/2}\left(\hat{m}_h^*(x)-\hat{m}_g(x)\right)\le z\right]\right| \\ = \; & O_P\left(n^{-1/5}\right). \end{aligned}$$

4.4 Smoothed bootstrap in the explanatory variable

The idea of this resampling method is to consider the variability coming from the explanatory variable (in the wild bootstrap this part of the resampling is kept fixed) and also to allow the resampling distribution of $Y^*|_{X^*=X_i}$ not to be degenerate (as it is in the two-dimensional naive bootstrap).

The resampling plan, proposed by Cao and González-Manteiga (1993), consists of the following steps:

1. Given the sample $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, some estimator of the joint distribution function of (X, Y) is constructed:

$$\hat{F}_g(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq y\}} \int_{-\infty}^{x} K_g(t - X_i)\, dt.$$

   It resembles the empirical cumulative distribution function in the response variable and the smoothed distribution function in the expanatory variable.

2. Bootstrap resamples, $\left\{\left(X_1^*, Y_1^*\right), \left(X_2^*, Y_2^*\right), \ldots, \left(X_n^*, Y_n^*\right)\right\}$, are drawn from the distribution $\hat{F}_g(x, y)$.

3. A bootstrap version of the Nadaraya-Watson estimator is constructed:

$$\hat{m}_h^*(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i^*\right) Y_i^*}{\frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i^*\right)}.$$

4. The resampling distribution of $(nh)^{1/2}\left(\hat{m}_h^*(x) - \hat{m}_g(x)\right)$ is used to approximate sampling distribution of the statistic: $(nh)^{1/2}\left(\hat{m}_h(x) - m(x)\right)$.

The optimal pilot bandwidth, g, is also of order $n^{-1/9}$, i.e. asymptotically larger than h.

The two-dimensional distribution used for the resampling mechanismo in Step 2, $\hat{F}_g(x,y)$, can be replaced by a smoothed distribution in the two variables:

$$\tilde{F}_g(x,y) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{y} K_g(s-Y_i)\,ds \int_{-\infty}^{x} K_g(t-X_i)\,dt.$$

This is equivalent to resample from the two-dimensional density

$$\hat{f}_g(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_g(x-X_i)\,K_g(y-Y_i),$$

which is the Parzen-Rosenblatt kernel estimator of the two-dimensional variable (X,Y).

Straight forward calculations can be used to prove that if (X^*,Y^*) has distribution $\hat{F}_g(x,y)$, then,

- X^* has bootstrap marginal density $\hat{f}_g(x)$.

- The bootstrap marginal distribution of Y^* is the empirical cdf of the Y_i: $\hat{F}_n^Y(y) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i \leq y\}}$.

- The regression function of the bootstrap resampling plan coincides with the Nadaraya-Watson estimation with bandwidth g:

$$E\left(Y^*|_{X^*=x}\right) = \hat{m}_g(x).$$

- In fact, the conditional distribution of $Y^*|_{X^*=x}$ is

$$\hat{F}_g\left(y|_x\right) = \frac{\frac{1}{n}\sum_{i=1}^n K_g(x - X_i)\,\mathbf{1}_{\{Y_i \leq y\}}}{\frac{1}{n}\sum_{i=1}^n K_g(x - X_i)},$$

  which is the Nadaraya-Watson kernerl estimador of the conditional distribution function.

As a consequence of the last remark it is easy to design a method to simulate bootstrap values of (X^*, Y^*), as required in Step 2. To do this, it is enough to simulate X^* from the Parzen-Rosenblatt estimator constructed with the

sample of the explnatory variable (this is just the classical smooth bootstrap) and then simulate Y^* from the discrete distribution that gives, to every datum Y_i, the following probability

$$w_i\left(X^*\right) = \frac{K_g\left(X^* - X_i\right)}{\frac{1}{n}\sum_{j=1}^{n} K_g\left(x - X_j\right)},\ i = 1, 2, \ldots, n.$$

The convergence rates of the bootstrap aproximation given by this method are:

$$\begin{aligned}
& \sup_{z \in \boldsymbol{R}} \left| P^{Y|_X}\left[(nh)^{1/2}\left(\hat{m}_h\left(x\right) - m\left(x\right)\right) \leq z\right] - \right. \\
& \left. - P^{Y^*|_{X^*}}\left[(nh)^{1/2}\left(\hat{m}_h^*\left(x\right) - \hat{m}_g\left(x\right)\right) \leq z\right] \right| \\
= \ & O_P\left(n^{-2/9}\right), \text{ en probabilidad } P, \\
& \sup_{z \in \boldsymbol{R}} \left| P\left[(nh)^{1/2}\left(\hat{m}_h\left(x\right) - m\left(x\right)\right) \leq z\right] - P^*\left[(nh)^{1/2}\left(\hat{m}_h^*\left(x\right) - \hat{m}_g\left(x\right)\right) \leq z\right] \right| \\
= \ & O_P\left(n^{-2/9}\right).
\end{aligned}$$

4.5 Comparison of convergence rates

The following table collects a summary of the convergence rates obtained with all the approximations considered:

Approximation	conditional	unconditional
Normal limit (theoretical)	$O_P\left(n^{-1/5}\right)$	$O\left(n^{-2/5}\right)$
Plug-in	$O_P\left(n^{-1/5}\right)$	$O_P\left(n^{-2/9}\right)$
Wild bootstrap	$O_P\left(n^{-2/9}\right)$	$O_P\left(n^{-1/5}\right)$
Smooth bootstrap in the explanatory variable	$O_{P^*}\left(n^{-2/9}\right)$ in probability P	$O_P\left(n^{-2/9}\right)$

Apart from the theoretical normal limit (useless in practice), the convergence rates of the smooth bootstrap in the explanatory variable are equal to or better than (both conditionally and unconditionally) the rest of the methods. In the conditional setting the two bootstrap resmpling plans are the ones which provide the best convergence rate ($n^{-2/9}$, compared to $n^{-1/5}$ for the plug-in approximation). In the unconditional setup the smooth bootstrap in the

explanatory variable and the plug-in approximation exhibit the best rate ($n^{-2/9}$, compared to $n^{-1/5}$ for the wild bootstrap).

References

Cao, R. (1990). Órdenes de convergencia para las aproximaciones normal y bootstrap en la estimación no paramétrica de la función de densidad. *Trabajos de Estadística*, **5**, 23-32.

Cao, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist* **19**, 2226-2231.

Cao, R. (1993). Bootstrapping the mean integrated squared error. *Jr. Mult. Anal.* **45**, 137–160.

Cao, R., Cuevas, A. and González-Manteiga, W. (1993). A comparative study of several smoothing methods in density estimation. *Comp. Statist. Data Anal.* **17**, 153–176.

Cao, R. and González-Manteiga, W. (1993). Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics* **2**, 379-388.

Davison, A. C. and Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge University Press, Cambridge.

Efron, B. (1979). Bootstrap methods - Another look at the jackknife. *Ann. Statist* **7**, 1-26.

Efron, B. and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman and Hall, New York.

Faraway, J.J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *Jr. Amer. Statist. Assoc.* **85**, 1119–1122.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.* **32**, 177–203.

Hall, P. (1992). The bootstrap and Edgeworth expansion. Springer-Verlag, New York.

Hall, P., Marron, J.S. and Park, B. (1992). Smoothed cross-validation. *Probab. Theor. Rel. Fields* **92**, 1–20.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778–796.

Marron, J.S. (1992). Bootstrap bandwidth selection. In *Exploring the limits of the bootstrap*, LePage, R. and Billard, L. eds., pp. 249–262. New York: Wiley.

Nadaraya, E.A. (1964). On estimating regression. *Theor. Probab. Appl.* **9,** 141-142.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. *Ann. Math. Statist.* **27**, 832–837.

Shao, J. and Tu, D.S. (1995). The jackknife and bootstrap. Springer-Verlag, New York.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Jr. Royal Statist. Soc. Ser. B* **53**, 683–690.

Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76**, 705–712.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā* Ser. A **26**, 359-372.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–1350.