Chap 1 : Bootstrap(s) in the i.i.d. case.

(Running Title: **Bootstrap(s) in the i.i.d. case.**)

by

Patrice Bertail*

CREST, Laboratoire de Statistique

Key Words and Phrases: Bootstrap, subsampling, parametric, non parametric, semiparametric, asymptotic properties, distance, second order theory.

*email : Patrice.Bertail@ensae.fr

1 Introduction

The resampling methods, among those the Jackknife (Quenouille, 1949) and the bootstrap (Efron 1979,1982) have been originally introduced to estimate the bias and the dispersion of an estimator in nonparametric situations (see e.g. Efron, 1982, Efron and Tibshirani 1993 for general introductions). The general principle is to repeatedly use the original sample to obtain several values of the statistics and calculate their mean and variance.

The bootstrap method is now considered as a kind of universal tool to obtain approximation of the distribution of regular statistics. The now well known underlying idea is the following : consider a sample $X_1, \ldots, X_n$ of independent and identically distributed (i.i.d.) random variables (r.v.'s) with unknown probability measure (p.m.) P. Assume we are interested in approximating the distribution of a statistic $T_n(X_1, ..., X_n)$ estimating some parameter θ and define $P_n := n^{-1}\Sigma_{i=1}^{n}\delta_{X_i}$ the empirical probability of the observation where $\delta_x(s) = \begin{cases} 0\, if\ s \neq x \\ \\ 1\, if\ s = x \end{cases}$ denotes the Dirac function. Since in some sense P_n is close to P when n is large, if one samples $X_1^*, \ldots, X_m^*$ i.i.d. r.v.'s from P_n and builds the distribution of $T_m(X_1^*, ...X_m^*)$ conditionally on P_n (which is made easy by using computer intensive calculus) then this distribution should imitate that of T_n when n and m get large. It has been emphasized that resampling methods have been made valuable in practice by the progresses of computer science and technology.

This idea has lead to considerable investigations to see when this method leads to correct approximations. When it does not, one generally tries to diagnose why and to find if there is any way to adapt it. This is in particular very important when we drop the i.i.d. assumption and focus on times series or random fields.

In this first chapter we would like to recall the main results of the bootstrap in the i.i.d. case but also to recast the bootstrap methodology in a functional approach to explain why there is not one form of Bootstrap, but a lot of "bootstrap(s)" specifically adapted to the model under consideration. These aspects are very important when one wants to generalize resampling methods to time series. We will give the main ideas why the bootstrap can lead to very accurate approximation of the statistics of interest, but why and when, the methodology breaks down. In particular, one has to be very careful on the choice of the statistics to bootstrap.

We also give a general theorem (due to Politis and Romano, 1994) stating the universal validity of the bootstrap when one chooses m much smaller than n. We will later explain why this method known as "subsampling" (see Politis Romano and Wolf, 2001) or m out of n bootstrap (Bickel, Götze and van Zwet, 1997) may lead to bad approximations (the price of generality) and how it is possible to improve these results by using interpolation or extrapolation techniques (Bertail, 1997). These ideas may be quite straightforwardly adapted to time-series as we will see later.

2 The functional approach

2.1 Some notations

$(\mathcal{X}, \mathcal{A}, \mathcal{P})$ probabilized space, with $\mathcal{X}$ some separable Banach space, $\mathcal{A}$ some $\sigma-$algebra, $\mathcal{P}$ probability space.

$T(.)$, functional from $\mathcal{P} \to B$ (Banach space), $B = \mathbb{R}$ or $\mathbb{R}^k$ from paragraph 3.to 6.

$\mathbb{P}$, non-parametric form of $\mathcal{P}$, convex set of probability containing Dirac measures.

$\mathcal{P}_\Theta$, a special parametric form for $\mathcal{P}$, set of probability measures indexed by a set $\Theta \subset \mathbb{R}^k$.

$\mathcal{P}_{\Theta,H}$, a special semi-parametric form for $\mathcal{P}$, set of probability measures indexed by two sets, $\Theta \subset \mathbb{R}^k$ and H an infinite dimensional space (most of the time an Hilbert space).

μ, measure of reference.

T_n statistics based on n i.i.d. observations $(X_1, ..., X_n)$ i.i.d. P in $\mathcal{P}$.

S_n, statistics used to standardize T_n.

$\tau_n(P)$, rate of convergence of the statistic T_n under $P^{\ n}$. We write $\tau_n = \tau_n(P)$ when the rate is uniform over a subclass of probability.

$k_m(x, P) = \Pr_{P^{\ m}}\{T_m(Y_{1,...,}Y_n) \leq x\},\ m \geq 1$, distribution of a statistic T_m under $P^{\ m}$.

$K_m(x, P) = \Pr_{P^{\ m}}\{\tau_m^{1/2}(T_m(Y_{1,...,}Y_n) - T(P)) \leq x\}$, standardized (centered and renormalized) form (for "regular" T_m).

$L_m(x, P) = \Pr_{P^{\ m}}\{\tau_m(P)S_m^{-1}(T_m(Y_{1,...,}Y_n) - T(P)) \leq x\}$, studentized form.

$||\ .\ ||_\infty$ the sup-norm. For $f : \mathcal{X} \to \mathbb{R}$ $||f||_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

2.2 Some basic examples.

In statistics, many parameters of interest may be seen as functionals from a probability space $\mathcal{P}$ to some Banach space, say B.

- For instance the expectation of some random variable X taking its values in some probabilized space $(\mathcal{X},\mathcal{A},\mathcal{P})$ with probability law P is a linear function of P defined by
$$T(P) = E_P X = \int xP(dx)$$
It is a functional from $\mathcal{P}$ to $\mathcal{X}$.

- Other important examples on $\mathcal{X}$ =$\mathbb{R}$ are moments (centered or not)

  $$M_k(P) = \int (x - E_P X)^k P(dx),\ k \in \mathbb{N},$$

  but also

- "M-parameters" that is parameters $\theta = M(P)$ which are the solution of some implicit equations

  $$E_P \psi(X, \theta) = \int \psi(x, \theta) P(dx) = 0,$$

  for some smooth measurable function ψ.

- V or U-parameters will be encountered frequently in this short course and we recall briefly their definitions : let $\omega(x_1, ..., x_m)$ be a symmetric (invariant by permutation of its argument) measurable kernel of degree m,

  $$V_m(P) = \int \int ... \int \omega(x_1, x_2, ..., x_m) P(dx_1)...P(dx_m)$$

  defines a $V-$parameter of degree m. If in addition $\omega(x_1, x_2, ..., x_m) = 0$ $(a.s)$ for any $x_i = x_j$, $i \neq j$ then we call this parameter a U-parameter. For instance the variance

  $$V_P(X) = T_2(P) = \frac{1}{2} \int \int (x_1 - x_2)^2 P(dx_1) P(dx_2)$$

  is a V or U-parameter of degree 2. If we have two independent random variables X, Y with probability P, the parameter

  $$W(P) = \Pr{}_{P^2}(X < Y) = \int \int 1_{\{x_1 < x_2\}} P(dx_1) P(dx_2)$$

  is also a V-statistic. See Serfling (1980) for a general introduction to U-statistic.

- When $B = \mathbb{R}$, if we denote $F(x) = P(] - \infty, x[)$ the cumulative distribution function and $F^{-1}(s) = \inf\{x \in \mathbb{R},\ F(x) \geq s\}$ its (left continuous) pseudo-inverse, quantiles and more generally "$L-$parameters" defined as

  $$L(P) = \int w(F^{-1}(s)) J(s) \lambda(ds)$$

  where w and J are real functions and λ the uniform measure on $[0, 1]$, are also simple functionals of the probability P via simple composition theorem.

Such an approach, which was considered in the early works of Von Mises (1947) has been intensively used in the robustness literature (see Huber, 1981 and Rieder, 1994 for an up-to-date account). Econometric formulations and applications may be found in Manski (1988) under the name of "the analogy principle".

2.3 The plug-in rule

In the following, T is a statistical functional defined on a space of probability measures $\mathcal{P}$, taking its value on a locally separable Banach space B. At the price of generality but for simplicity, we will work with probability measures but one should remember that it is sometimes better to work with signed measures if the functional T may be naturally extended to a space $\mathcal{M}$ of signed measures (see Gill,1989, van der Vaart and Wellner, 1996).

Consider now a probability space $(\mathcal{X},\mathcal{A},\mathcal{P})$. Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be a sequence of i.i.d. r.v.'s with common probability $P \in \mathcal{P}$, defining a sequence of experiments $(\mathcal{X}^n,\mathcal{A}^n,\mathcal{P}^n)$ on the product space. The basic idea of the plug-in rule (or analogy principle) is simply that, if one is able to construct a convergent estimator of the underlying distribution $\widehat{P}_n$ then $T(\widehat{P}_n)$ may be close to $T(P)$ and defines a "natural" estimator $T(P)$. This is in particular the case if T is continuous according to some topologies compatible with the convergence notion used for the convergence of $\widehat{P}_n$ to P (generally weak convergence or to avoid measurability problems, Hoffmann-Jorgensen convergence; see Van der Vaart and Wellner,1996). Of course the adequate choice of the estimator $\widehat{P}_n$ for P highly depends on the form of $\mathcal{P}$. There are mainly three interesting forms for $\mathcal{P}$ that appears in the statistical literature.

2.3.1 Non parametric framework

Assume that $\mathcal{P}$ is a large convex set of probability or signed measures containing all the Dirac measures. We will denote it by $\mathbb{P}$ and we will say that the statistical model is non-parametric. In the finite dimensional case, $\mathbb{F}$ will denote the corresponding set of cumulative distribution function (cdf). We will also use the notation $\mathbb{P}^0$ (resp. $\mathbb{F}^0$) for the set of all standardized probability measure (resp. cdf) with expectation equal to 0 and variance equal to 1.

For instance in that setting, estimating $T_0(P) = P$ the true underlying probability may be of interest. If we define

$$P_n := n^{-1}\Sigma_{i=1}^{n}\delta_{X_i},$$

the empirical measure, then P_n belongs to $\mathbb{P}$ for any realization of the r.v. and defines a "natural" estimator of P. More generally if T is "smooth" in a neighborhood of P, a natural estimator of $T(P)$ is its empirical counterpart $T(P_n)$ (see von Mises, 1947). Of course these notions have only a meaning if we make precise the mathematical notion of convergence that we use to describe the proximity between P_n and P , and to indicate the topologies that we use on $\mathcal{P}$ and B to quantify the degree of "smoothness" (continuity, differentiability) of the functional T. We shall do this later. We just want now to point out that in a non-parametric setting of this form, without any convergence consideration, it is very easy to define the ”analog” or ”plug-in” estimators of the quantity $T(P)$ that appeared before, **provided that T is defined on $\mathbb{P}$.**

- For instance, the analog estimator of $T(P) = E_P X = \int x P(dx)$ will simply be the empirical mean

$$T(P_n) = \int x P_n(dx) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- The analog estimators of $M_k(P)$ are the empirical moments,

$$M_k(P_n) = \int (x - E_{P_n} X)^k P_n(dx) = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - n^{-1} \sum X_i \right)^2$$

- Similarly $M(P)$ defines an M-estimator as the empirical value $\widehat{\theta}_n = M(P_n)$ which satisfies

$$\sum_{i=1}^{n} \psi(X_i, \widehat{\theta}_n) = 0.$$

  Maximum likelihood estimators and robust estimators are typically defined this way.

- The analog estimator of $L(P)$ is simply

$$L(P_n) = \int w(F_n^{-1}(s)) J(s) \lambda(ds) = \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} w(F_n^{-1}(s)) J(s) \lambda(ds)$$
$$= \sum_{i=1}^{n} a_{i,n} w(X_{(i)})$$

  where $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq ... \leq X_{(n)}$ is the ordered statistic and $a_{i,n} = \int_{\frac{i-1}{n}}^{\frac{i}{n}} J(s) \lambda(ds)$. Trimmed mean, empirical fractile, range are particular example of these estimators.

- V-statistic and (renormalized) U-statistic for a symmetric kernel ω of degree m are respectively given by

$$V_n = V(P_n) = n^{-m} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} ... \sum_{i_m=1}^{n} \omega(X_{i_1}, ..., X_{i_m})$$

and after an adequate renormalization (taking into account the number of terms in the sum)

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 \ne i_2 \ne \ldots \ne i_m \le n}^{n} \omega(X_{i_1}, ..., X_{i_m}) = E_{P_n}(\omega(X_1, ..., X_m)|(X_{(1)}, ..., X_{(n)})) \tag{1}$$

The last equality has only a meaning on an ordered space, but has the advantage to show that U_n is the Rao Blackwell improve of V_n. For m fixed, this representation implies that U_n is a martingale (with respect to the filtration engendered by $(X_{(1)}, ..., X_{(n)})$).

2.3.2 Parametric framework.

The statistical model (for good or bad reasons) may be restricted to belong to some subset of $\mathbb{P}$ indexed by some parameter θ belonging to Θ, a subset of some finite dimensional space assimilated to $\mathbb{R}^k$. In that case, we will denote $\mathcal{P}$ as

$$\mathcal{P}_\Theta = \{P_\theta,\ \theta \in \Theta \subset \mathbb{R}^k\}.$$

The model is then a classical parametric model indexed by the set Θ. In that framework, we will be interested in functional parameters of the form

$$T(P_\theta) = g(\theta),$$

where $g(\theta)$ is a "smooth" function from $\mathbb{R}^k$ to B. If the model is identifiable and under regular conditions on the parametric family $\mathcal{P}_\Theta$, convergent estimators of θ are available (see LeCam (1986) for a modern introduction to estimation in parametric model or Borovkhov (1998) for a more accessible book...). Then, they define a random sequence of probability $P_{\widehat{\theta}_n}$ and a corresponding estimator

$$T(P_{\widehat{\theta}_n}) = g(\widehat{\theta}_n).$$

Notice that the introduction of the random probability measure $P_{\widehat{\theta}_n}$ is somehow totally artificial in this framework. This device will however make the different form of Bootstrap found in the literature clearer.

2.3.3 Semiparametric framework.

In many statistical models, only one part of the parameters of the model (mainly because of their interpretation) are of interest whereas the others do not : there are called nuisance parameter in that they make the inference much more difficult than for parametric models. Semiparametric models are studied at length in Bickel, Klaasen, Ritov and Wellner (1993), BKRW in the following. A nice account may also be found in van der Vaart (1998). In that case, the set $\mathcal{P}$ may be denoted

$$\mathcal{P}_{\Theta,H} = \{P_{\theta,\eta},\ \theta \in \Theta,\ \eta \in H\},$$

where typically Θ is a subset of $\mathbb{R}^k$ (in some case Θ may also be infinite dimensional) and H is a space of infinite dimension. Just like in the parametric case, most of the functionals of interest are defined on $\mathcal{P}_{\Theta,H}$ have the form

$$T(P_{\theta,\eta}) = g(\theta,\eta) \text{ or } g(\theta)$$

However finding an "adequate" analog estimator, which enjoys good properties may be much more difficult.

Nevertheless, if one can construct convergent estimators estimator of $\widehat{\theta}_n$ and $\widehat{\eta}_n$, one may define a sequence $P_{\widehat{\theta}_n\,\widehat{\eta}_n}$ and the corresponding plug-in estimator

$$T(P_{\widehat{\theta}_n\,\widehat{\eta}_n}) = g(\widehat{\theta}_n,\widehat{\eta}_n) \text{ or } g(\widehat{\theta}_n).$$

Somehow, one of the main challenges in the semiparametric literature is to find efficient sequences (in a semiparametric sense, see BKRW) $\widehat{\theta}_n$, $\widehat{\eta}_n$ such that $P_{\widehat{\theta}_n\,\widehat{\eta}_n}$ stays a.s. inside the model $\mathcal{P}_{\Theta,H}$... Unfortunately this is not always possible. An other way (which is almost equivalent) to solve this problem is to try to extend the functional T to a much more general space, $\mathbb{P}$ for instance, by defining a pseudo-projection into the set $\mathcal{P}_{\Theta,H}$. Indeed the main practical problem in these models is actually that $T(P_n)$ is not defined or does not make sense. In many problems, it is possible to extend the functional defined on $\mathcal{P}_{\Theta,H}$ to the whole space $\mathbb{P}$, that is, to consider, for any $P \in \mathbb{P}$, a functional $\widetilde{T}(P) = T \circ \Pi(P)$ (or a sequence of extensions $\widetilde{T}_m(P) = T \circ \Pi_m(P)$) with $\Pi(P) = P$, on $\mathcal{P}_{\Theta,H}$ (resp. $\Pi_m(P) \to_{m\to\infty} P$) . Π may be a projector for convex model, a pseudo-projector (not uniquely defined) or a convolution operator etc... Then $\Pi(P_n)$ defines a sequence $P_{\widehat{\theta}_n\,\widehat{\eta}_n}$ (see Bertail, 2003).

Let us give some simple examples.

Example A. Let μ be some probability measure such that $\int x^2\mu(dx) < \infty$. Consider the semiparametric model

$\mathcal{P}_H = \{P_\eta$ such that $P_\eta << \mu$ and $\frac{dP_\eta}{d\mu}(x) = \eta(x),\ \eta \in H\}$, where $H = L^2(\mu)$.

We may be interested in estimating for instance $T(P_\eta) = \int \eta^2 d\mu$. In that case, P_n does not belong to $\mathcal{P}_H$ and the functional $T(P_n)$ is not defined. If we define $\Pi_m(P) = K_{h_m} * P$ as the convolution operator, with some Kernel probability K_{h_m} dominated by μ, with smoothing parameter h_m (the variance of K_{h_m}) for

$P \notin \mathcal{P}_H$, then the sequence $\Pi_m(P_n)$ belongs to $\mathcal{P}_H$ and is simply a smoothed version of the empirical distribution.

If we are interested in $T(P_\eta) = E_{P_\eta}X$ notice that T may be trivially extended to $\mathbb{P}$ and then $T(P_n)$ has a meaning.

Example B. The symmetric position model

$$\mathcal{P}_{\Theta,H} = \{P_{\theta,\eta} \text{ such that } P_{\theta,\eta} << \mu \text{ and } \frac{dP_{\theta,\eta}}{d\mu}(x) = \eta(x-\theta),\ \eta \in H,\ x \in \mathbb{R}\}$$

where H is a specific subset of symmetric density (such that $\eta(x) = \eta(-x)$). Of course P_n is generally not dominated by μ and does not belong to $\mathcal{P}_{\Theta,H}$. However, under $P_{\theta,\eta}$, the $X_i - \theta$ are i.i.d. η and it is easy to build a symmetric kernel estimator $\widehat{\eta}_{\theta,n}(x)$ of η. Then it can be shown that the maximum likelihood estimator (m.l.e.) $\widehat{\theta}_n$ using this estimator is a $n^{1/2}-$asymptotically efficient estimator for θ, leading to an estimator $\eta_n(x) = \widehat{\eta}_{\widehat{\theta}_n,n}(x)$ for $\eta(x)$ (see BKRW).

Example C. Consider the regression model based on some real valued components (Y_i, X_i) $i = 1, ...n,$ *i.i.d.*, with $P_{\theta,\eta}$ in the set

$$\mathcal{P}_{\Theta,H} = \{P_{\theta,\eta},\ P_{\theta,\eta}(y,x) = P_\varepsilon(\frac{y - x\beta}{\sigma})P_X(x),\ \theta \in \Theta,\ \eta \in H\}$$

$$\Theta = \{\theta = (\beta,\sigma) \in \mathbb{R} \times \mathbb{R}^+\},\ H = \{\eta = (P_X,P_\varepsilon),\ P_X \in \mathbb{P},\ P_\varepsilon \in \mathbb{P}^0\ \}.$$

This simply means that

$$Y_i = X_i\beta + \sigma\varepsilon_i,$$

with X_i are i.i.d. P_X independent of ε_i i.i.d. P_ε with $E_{P_\varepsilon}\varepsilon_i = 0,\ V_{P_\varepsilon}(\varepsilon_i) = 1$. Typically the parameter of interest will be

$$T(P_{\theta,\eta}) = \beta.$$

However we may also be interested in estimating the whole model $T(P_{\theta,\eta}) = P_{\theta,\eta}$.

A simple way to retrieve all the components of the model is first to estimate β by the least square estimator (l.s.e.)

$$\widehat{\beta}_n = \frac{Cov_{P_n}(Y_i, X_i)}{V_{P_n}(X_i)},$$

which is nothing else than the empirical estimator of β (defined on the whole space $\mathbb{P}$). An obvious estimator of the P_X is simply given by the empirical probability $P_{X,n} = n^{-1}\sum_{i=1}^n \delta_{X_i} \in \mathbb{P}$. Now let $\widehat{\varepsilon}_i = Y_i - X_i\widehat{\beta}_n$ be the residuals of the model. $\widehat{\sigma}_n = (n^{-1}\sum_i(\widehat{\varepsilon}_i - n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i)^2)^{1/2}$ is a convergent estimator of σ. If we now define $\widetilde{\varepsilon}_i = (\widehat{\varepsilon}_i - n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i)/\widehat{\sigma}_n$ the recentered standardized residuals, then it can be easily shown that $\widehat{P}_{\varepsilon,n} = n^{-1}\sum_{i=1}^n \delta_{\widetilde{\varepsilon}_i}$ is a convergent estimator of P_ε belonging to $\mathbb{P}^0$. That is if we define $P_{\widehat{\theta},\widehat{\eta}}$, with $\widehat{\theta}=(\widehat{\beta}_n,\widehat{\sigma}_n)$ and $\widehat{\eta}_n=(\widehat{P}_{X,n},\widehat{P}_{\varepsilon,n})$, then we have constructed a (random) convergent sequence $P_{\widehat{\theta},\widehat{\eta}}$ which belongs to $\mathcal{P}_{\Theta,H}$. The plug-in rule consists in this framework to use $T(P_{\widehat{\theta},\widehat{\eta}})$ to estimate $T(P_{\theta,\eta})$. This is a common rule which is used for instance to estimate the variance of $\widehat{\theta}_n$.

3 A functional view of the Bootstrap(s).

One the main goal of statistics is to obtain approximation of the distribution of some statistic of interest estimating some parameters. Either because we want to compute some particular quantities of interest, its variance, its skewness etc... or because we want to construct confidence intervals for the parameters. We will focus on the case when $T(P)\in\mathbb{R}$ or $\mathbb{R}^p$ to avoid a complicated exposition (and measurability problem) but one should have in mind that all the ideas exposed here have also extensions for parameters to general Banach space.

On $(\mathcal{X},\mathcal{A},\mathcal{P})$, let again $X,X_1,\ldots,X_n$ be a sequence of i.i.d.r.v.'s with common probability $P\in\mathcal{P}$, and consider some parameter of interest $T(P)$ defined on $\mathcal{P}$ estimated by some estimator $T_n=T_n(X_1,...,X_n)$ (may be a plug-in estimator, but this may be a more general statistic).

Actually when one wants to determine the distribution of T_n and its asymptotic properties, one is now interested in the sequence of parameter $\{k_m(.,P)\}_{m\in\mathbb{N}}$

$$\begin{aligned}k_m(x,P)&=\Pr{}_{P^{\ m}}\{T_m(Y_{1,\ldots,}Y_m)\le x\}\\&=\int_{\mathcal{X}^m}1_{\{T_m(y_1,\ldots,y_m)\le x\}}P(dy_1)...P(dy_m)\end{aligned}$$

or $\{K_m(.,P)\}_{m\in\mathbb{N}}$

$$\begin{aligned}K_m(x,P)&=\Pr{}_{P^{\ m}}\{\tau_m(T_m(Y_{1,\ldots,}Y_m)-T(P)\le x\}\\&=\int_{\mathcal{X}^m}1_{\{\tau_m(T_m(y_1,\ldots,y_m)-T(P)\le x\}}P(dy_1)...P(dy_m),\end{aligned}$$

where $Y_1,....,Y_m$ is an independent copy of $(X_1,....X_m)$, $m\ge1$, *i.i.d. P*. Here τ_m is an asymptotic rate of convergence typically $\tau_m=m^{1/2}$ in regular case. Of course, in most non parametric, semiparametric cases (and even in a parametric framework) exact computations of these values are impossible.

Notice that these quantities may be seen as V-parameters with kernel of order m given respectively by

$$\mathbb{I}_{\{T_m(x_1,\ldots,x_m)\le x\}}$$

and

$$\mathbb{I}_{\left\{m^{1/2}(T_m(x_{1,\dots,}x_m)-T(P))\le x\right\}}.$$

More generally, if $S_m(Y_{1,\dots,}Y_m)$ is some random standardization (converging to some constant) then in the spirit of constructing pivotal asymptotic distribution, we may also be interested in the sequence of V-parameters $\{L_m(.,P)\}_{m\in\mathbb{N}}$

$$L_m(x,P)=\Pr{}_{P^{\ m}}\left\{\tau_m(P)S_m^{-1}(Y_1,\dots,Y_m)\left(T_m(Y_{1,\dots,}Y_m)-T(P)\right)\le x\right\},$$

where $\tau_m=\tau_m(P)$ is some rate of convergence of the statistic $T_m(Y_{1,\dots,}Y_m)$. Notice that the rate of convergence of a statistic may itself depend on the true underlying probability. In the following Φ and ϕ are respectively the cumulative distribution function and the density of the normal distribution N(0,1).

When one is unable to calculate the exact distribution of the statistic of interest, we can use at least two different approaches to get approximation of these parameters:

- the first one is a traditional asymptotic approach. That is to try to study the asymptotic behavior for instance of $K_m(.,P)$ or $L_m(.,P)$.

- the second one is to estimate these parameters with the available data by a plug-in approach : this is the Bootstrap approach.

To illustrate the properties of the bootstrap we will frequently use the following simple examples.

Example 1

If $T(P)=E_PX$ and $T_n=T(P_n)=n^{-1}\sum_{i=1}^n X_i$, then we know that under $0<E_P(X-E_PX)^2=\sigma^2<\infty$, by the CLT

$$K_m(x,P)-\Phi(x/\sigma)\to 0 \text{ as } m\to\infty, \text{ uniformly in } x,$$

so that one may use $\Phi(./\sigma)$ as a deterministic approximation for $K_n(.,P)$. We have as well with $S_m^2=m^{-1}\sum_{i=1}^m(Y_i-\overline{Y}_m)^2$ and $\tau_m(P)=m^{1/2}$,

$$||L_m(.,P)-\Phi(x)||_\infty\to 0$$

Example 2

If $T(P)=(E_PX)^2$ and $T_n=T(P_n)=(n^{-1}\sum_{i=1}^n X_i)^2$ the situation is a bit different

If $E_P(X)\ne 0$ then by Slutsky theorem (a simple Taylor expansion)

$$L_m(x,P)=\Pr{}_{P^{\ m}}\{m^{1/2}(T_m(Y_{1,\dots,}Y_m)-T(P)/(2|\overline{Y}_n|S_m)\le x\}\to\Phi(x)$$

but if $E_P(X)=0$

$$L_m(x,P)=\Pr{}_{P^{\ m}}\{m(T_m(Y_{1,\dots,}Y_m)-T(P)/S_m^2\le x\}\to\chi^2(1)$$

That is $\tau_m(P) = m^{1/2}$ for $P \in \mathbb{P} \backslash \mathbb{P}^0$ and $\tau_m(P) = m$ for $P \in \mathbb{P}^0$.

The classical approach is to use the asymptotic distribution $K(.,P)$ as an approximation of the true underlying distribution: this is a deterministic approximation. However, it is tempting to choose a plug-in approach and to use the empirical counterparts of $\{K_m(.,P)\}_{m \in \mathbb{N}}$, provided that one has as estimator $\widehat{P}_n$ of P in the model of interest $\mathcal{P}$. The plug-in estimators $\{K_m(.,\widehat{P}_n)\}_{m \in \mathbb{N}}$ of the parameters. $\{K_m(.,P)\}_{m \in \mathbb{N}}$ defines random measures, which are nothing else than Bootstrap distributions.

We begin to illustrate this point of view in several framework : nonparametric, parametric and semiparametric framework. The next questions in paragraph 4 will be

1) Do we have consistency of the Bootstrap distribution, that is under which conditions, do we have (uniformly in the argument or for some other norms, L_2 or L_p norm)

$$K_m(.,\widehat{P}_n) - K(.,P) \to 0$$

or

$$K_m(.,\widehat{P}_n) - K_m(.,P) \to 0$$

or

$$K_m(.,\widehat{P}_n) - K_n(.,P) \to 0$$

(resp. for $k(.,P),\ L(.,P)$) either a.s. or in probability either for $m = n$ or $m \neq n$, as $n \to \infty$ and/or $m \to \infty$.

2) What are the rates of convergence (in probability, a.s. minimax rates) of these estimators? or their asymptotic distributions?

3.1 The Nonparametric or Naive Bootstrap : $\mathcal{P} = \mathbb{P}$.

In that case we may choose $\widehat{P}_n = P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ then we have an explicit form for the Bootstrap distributions

$$\begin{aligned} k_m(.,P_n) &= \Pr{}_{P_n{}^m}\{T_m(Y_{1,\dots,}Y_m) \le x\} \\ &= E_{P_n{}^m}\left(\mathbb{I}_{\{T_m(Y_{1,\dots,}Y_m)\le x\}}\right) \\ &= n^{-m}\sum_{i_1=1}^{n}\dots\sum_{i_m=1}^{n}\mathbb{I}_{\left\{T_m(X_{i_1,\dots,}X_{i_m})\le x\right\}}, \end{aligned}$$

that is the bootstrap distribution of the statistic T_m is simply the empirical distribution of all the values that can be computed by forming subsets (with replacement) of size m from the sample. It is also the V-statistic associated to the V-parameter $k_m(.,P)$

We similarly have

$$\begin{aligned} K_m(.,P_n) &= \Pr{}_{P_n{}^m}\{\tau_m\left(T_m(Y_{1,\dots,}Y_m) - T(P_n)\right) \le x\} \\ &= n^{-m}\sum_{i_1=1}^{n}\dots\sum_{i_m=1}^{n}\mathbb{I}_{\left\{\tau_m\left(T_m(X_{i_1,\dots,}X_{i_m})-T(P_n)\right)\le x\right\}} \end{aligned}$$

In that case, the values are recentered by the value of the statistic computed on the whole sample. Thus $K_m(.,P_n)$ is not exactly a V-statistic. Actually in some cases the recentering will cause some problem for the bootstrap (an adequate recentering is fundamental for time-series)...

And finally

$$\begin{aligned} L_m(x,P_n) &= \Pr{}_{P_n{}^m}\left\{\tau_m(P_n)S_m^{-1}((T_m(Y_{1,\dots,}Y_m) - T(P_n)) \le x\right\} \\ &= n^{-m}\sum_{i_1=1}^{n}\dots\sum_{i_m=1}^{n}\mathbb{I}_{\left\{\tau_m(P_n)S_m^{-1}\left(T_m(X_{i_1,\dots,}X_{i_m})-T(P_n)\right)\le x\right\}} \end{aligned}$$

Notice that this last expression questions the meaning of $\tau_m(P_n)$. If the behavior of T_m is uniform over a large class of probability as is the case in example 1, then we do not have any problem in standardizing T_m. In example 1, we have $\tau_m(P) = m^{1/2}$, for i.i.d. r.v.'s having moments of order 2 (as well as for triangular arrays satisfying the Lindeberg's condition). In general, it is assumed that $\tau_m(P) = \tau_m$ is fixed and known at least on a subset of probability in $\mathcal{P}$.

However when the behavior of T_m is not uniform as in example 2, we may have some real problem even in defining the plug-in estimator. Actually the lack of uniformity will be the main source of problem for this form of bootstrap.

The principle of the **"naive" bootstrap** considered by Efron is to choose $m = n$ and to estimate the distribution of $k_n(., P)$ (resp. $K_n(., P)$, $L_n(., P)$) by the plug-in estimators $k_n(., P_n)$ (resp. $K_n(., P_n)$, $L_n(., P_n)$). Of course if n is large an exact computation of $K_n(., P_n)$ may be time-consuming. For these reasons, one generally prefers to consider a Monte-Carlo approximation of $K_n(., P_n)$, which justifies the algorithm considered in the introduction and more frequently used in the literature.

Let $\mathcal{X}_j = (X_{i_1}, ..., X_{i_m})$, $j = 1, ...B$ be some subsets of size m taken uniformly with replacement from the original data $(X_1, ..., X_n)$ then a Monte-Carlo approximation of $K_n(., P_n)$ is given by

$$K_m^B(., P_n) = \frac{1}{B} \sum_{j=1}^{B} \mathbb{I}_{\{\tau_m(T_m(\mathcal{X}_j) - T(P_n)) \leq x\}}$$

Monte Carlo approximations for the other distribution may be defined similarly.

Notice that by the LLN, we have as $B \to \infty$,

$$K_m^B(., P_n) - K_m(., P_n) \to 0.$$

By the Berry Esséen Theorem, we may even get that

$$||K_m^B(., P_n) - K_m(., P_n)||_\infty \leq C B^{-1/2}, \text{ for some constant } C > 0.$$

Thus provided that B is chosen large enough, we can get an approximation of the bootstrap distributions as precise as we want.

3.2 m out of n Bootstrap and subsampling distribution : non-parametric case

When m is smaller than n, the non-parametric Bootstrap $K_m(., P_n)$ is called the m out of n bootstrap. A variant of this bootstrap distribution is the subsampling distribution or bootstrap without replacement defined as

$$k_m^U(., P_n) = \binom{n}{m}^{-1} \sum_{\sigma \in S} \mathbb{I}_{\left\{T_m(X_{\sigma(1)}, ..., X_{\sigma(m)}) \leq x\right\}},$$

where S is the set of all injections from $\{1, ..., m\}$ to $\{1, ..., n\}$. That is, the subsampling distribution is simply the empirical distribution of all the values of the statistic that may be computed on all the subsamples taken without replacement from the original data.. This is actually exactly the U-statistic with kernel $\mathbb{I}_{\{T_m(x_1, ..., x_m) \leq x\}}$ associated to the V–statistic $k_m(., P_n)$.

For $m \leq n,$ we similarly define the subsampling distribution associated to $K_m(.,P_n)$ and $L_m(x,P_n)$ by

$$K^U_m(.,P_n) = \binom{n}{m}^{-1} \sum_{\sigma \in S} \mathbb{I}_{\left\{\tau_m\left(T_m(X_{\sigma(1)},\dots,X_{\sigma(m)})-T(P_n)\right)\leq x\right\}}$$

and

$$L^U_m(x,P_n) = \binom{n}{m}^{-1} \sum_{\sigma \in S} \mathbb{I}_{\left\{\tau_m S_m^{-1}(X_{\sigma(1)},\dots,X_{\sigma(m)})(T_m(X_{\sigma(1)},\dots,X_{\sigma(m)})-T(P_n))\leq x\right\}}.$$

In the particular case when $m = n-1$, this is actually the so-called Jackknife empirical distribution (or Jackknife histogram). More generally the subsampling distribution may be seen as an $(n-m)$ delete jackknife distribution.

Our first theorem due to Politis and Romano (1994), although historically one of the oldest, will show that the m out of n bootstrap or subsampling can yield asymptotically convergent distribution of the statistic of interest under minimal assumptions. The proof of the theorem is simply based on the fact that the subsampling distribution is (up to a recentering factor) a U-statistic of degree $m.$ Almost sure versions of these results are also available in the book by Politis, Romano, Wolf (2001). We give a complete proof of this result because the underlying ideas are also interesting for time series.

Theorem 1 *Assume that*

H_0 : *there exists a rate τ_n, such that $\tau_n\left(T_n(X_{1,\dots,}X_n)-T(P)\right)$ converges to a non degenerate distribution K_P as $n->\infty$.*

H_1 : *K_P is supposed to have a non-degenerate continuous distribution function $K(.,P)$ almost everywhere.*

If in addition we have

H_2 : *$\frac{m}{n} \to 0$ and $\frac{\tau_m}{\tau_n} \to 0$ then*

$$||K^U_m(.,P_n) - K_n(.,P)||_\infty \underset{n\to\infty}{\to} 0 \text{ Pr}.$$

If in addition, we have $\frac{m}{n^{1/2}} \to 0$ then the m out of n bootstrap is also universally consistent

$$||K_m(.,P_n) - K_n(.,P)||_\infty \underset{n\to\infty}{\to} 0 \text{ Pr}.$$

Proof :
The proof is based on Hoeffding inequality for U-statistic with bounded kernel.

Lemma 2 *Let U_m be a U-statistic of degree m with symmetric kernel $\omega(x_1,...,x_m)$ bounded by M*

$$U_m = \binom{n}{m}^{-1} \sum_{\sigma \in S} \omega(X_{\sigma(1),...,}X_{\sigma(m)})$$

then

$$\Pr\left\{|U_m - E_{P^{m}}\omega(X_1,...,X_m)| \geq x\right\} \leq 2\exp\left\{-[\frac{n}{m}]x^2/(8M^2)\right\}.$$

Proof : See Annex 1

Now introduce $U_m(x,P_n) = \binom{n}{m}^{-1} \sum_{\sigma \in S} \mathbb{I}_{\left\{\tau_m\left(T_m(X_{\sigma(1),...,}X_{\sigma(m)}-\theta\right) \leq x\right\}}$. This is exactly a U-statistic of order m with a kernel bounded by 1 with expectation

$$E_{P^{n}}U_m(x,P_n) = Pr_{P^{m}}\left\{\tau_m\left(T_m(X_{\sigma(1),...,}X_{\sigma(m)}) - \theta\right) \leq x\right\}$$

$$= K_m(.,P)$$

We thus have, for any $x \in \mathbb{R}$, for any $\varepsilon > 0$,

$$\Pr\left\{|U_m(x,P_n) - K_m(x,P)| \geq \varepsilon\right\} \leq 2\exp\left\{-[\frac{n}{m}]\varepsilon^2/8\right\}.$$

It follows by a classical chaining argument that

$$\Pr\left\{||U_m(x,P_n) - K_m(x,P)||_\infty \geq \varepsilon\right\} \leq C\exp\left\{-[\frac{n}{m}]\varepsilon^2/8\right\}.$$

Thus $||U_m(x,P_n) - K_m(x,P)||_\infty \to 0$ pr. as soon as $\frac{m}{n} \to 0$ (even for a fixed m).

Now, for any $\eta > 0$,

$$K_m^U(x,P_n) = U_m(x - \tau_m(\theta - T(P_n)), P_n)\ I_{\{\tau_m|T(P_n)-\theta| \leq \eta\}}$$

$$+ U_m(x - \tau_m(\theta - T(P_n)), P_n)\ I_{\{\tau_m|T(P_n)-\theta| \geq \eta\}}$$

It follows that for any $\eta > 0$,

$$||K_m^U(x,P_m) - U_m(x,P_n)||_\infty \leq ||U_m(x+\eta,P_n) - U_m(x,P_n)||_\infty$$

$$+ ||U_m(x-\eta,P_n) - U_m(x,P_n)||_\infty + I_{\{\tau_m|T(P_n)-\theta| \geq \eta\}}$$

The results follow by noticing that if $\tau_m/\tau_n \to 0$ then $I_{\{\tau_m|T(P_n)-\theta|\geq\eta\}} - > 0$ *pr.* Combining all these results and the bound

$$||K_m^U(x,P_m) - K_n(x,P)||_\infty \leq ||K_m^U(x,P_m) - U_m(x,P_n)||_\infty + ||U_m(x,P_n) - K_m(x,P)||_\infty$$

$$+ ||K_m(x,P) - K(x,P)||_\infty + ||K_n(x,P) - K(x,P)||_\infty$$

we can easily get the result. Now the subsampling distribution and the m out of n distribution only differ in probability by a factor $\frac{n!/(n-m)!}{n^m}$ which converges to 1 iff $\frac{m^2}{n} \to 0$, by the Stirling formula. To see this, simply notice that the probability to draw (with replacement) a sample of size m without any repetition is $\frac{n}{n}\frac{n-1}{n}...\frac{n-m+1}{n} = \frac{n!/(n-m)!}{n^m}$.

Remark : The choice $\frac{m^2}{n} \to 0$ can not be improved, if we work with general statistics. Consider for instance that P is continuous with respect to Lebesgue measure, and that we have a U-statistic with kernel $g(x_1,x_2) = \begin{cases} \frac{1}{2}(x_1-x_2)^2 \text{ if } x_1 \neq x_2 \\ \infty \;\; else \end{cases}$, that does not allow for redoubling. There is no problem in studying the asymptotic behaviour of this U-statistic (because the probability that 2 points are equal in the sample is 0). However, the bootstrap approximation gets into trouble because of the discrete nature of P_n. The choice $m = o(n^{1/2})$ is the price to pay for not seeing these redoublings. This does not mean that for a particular class of statistic that this is the optimal choice! Most of of the time it is possible to take m close to $n/log(n)^\alpha$ $\alpha > 0$, (see for instance Bretagnolle, 1983). Many results on subsampling are summarized in the book by Politis, Romano and Wolf (2001).

3.3 The parametric bootstrap

When $\mathcal{P} = \mathcal{P}_\Theta = \{P_\theta,\ \theta \in \Theta \subset \mathbb{R}\}$, assume that the model is regular (in LeCam, 1986 sense), then it can be shown that one can construct an efficient regular estimator of θ say $\widehat{\theta}_n$.

If we are interested in estimating the distribution $k_n(.,P_\theta)$ of

$$T(P_{\widehat{\theta}_n}) = g(\widehat{\theta}_n).$$

a simple solution is just to use $k(.,P_{\widehat{\theta}_n})$:

$$k_m(x,P_{\widehat{\theta}_n}) = \Pr{}_{P_{\widehat{\theta}_n}^m}\{g(\widehat{\theta}_m(Y_1,...,Y_m)) \leq x\}$$

Notice for instance that estimating the variance of $\widehat{\theta}_m(Y_1,...,Y_m)$ by $V_{P_{\widehat{\theta}_n}^{\ m}}\widehat{\theta}_m(Y_1,...,Y_m)$ is exactly using a parametric bootstrap variance estimator.

In that case, the exact computation of $k_m(x, P_{\widehat{\theta}_n})$ is not available so that a Monte-Carlo approximation is necessary.

We illustrate this parametric bootstrap with the following example (left as an exercise to the reader).

Example 3 : Assume that $(X_1,...,X_n)$ *i.i.d.* $LN(m,\sigma^2)$. It can be shown that

$$E_PX = \exp(m+\sigma^2/2)$$

$$V_PX = \exp(2m+\sigma^2)\{\exp(\sigma^2)-1\}.$$

We may be interested in constructing confidence intervals for $\xi = V_PX$ or $\psi = \frac{V_PX}{E_PX}$

The m.l.e. of the parameters (m,σ^2) are given by

$$\widehat{m}_n = n^{-1}\sum_{i=1}^{n}\log(X_i)$$

$$\widehat{\sigma}^2_n = n^{-1}\sum_{i=1}^{n}(\log(X_i)-\widehat{m}_n)^2$$

and we have $(\widehat{m}_n, \widehat{\sigma}^2_n) \rightsquigarrow N((m,\sigma),\quad)$ (to be calculated). The plug-in estimators for E_PX and V_PX are

$$\mu_n = \exp(\widehat{m}_n + \widehat{\sigma}^2_n/2)$$

$$S^2_n = \exp(2\widehat{m}_n + \widehat{\sigma}^2)\{\exp(\widehat{\sigma}^2)-1\}$$

The exact distribution of μ_n is easy to obtain with the preceding results but the distribution of S^2_n or of S^2_n/μ_n are intractable. An asymptotic Gaussian approximation in this framework does not seem very appealing... The parametric bootstrap is another alternative : as can be seen from simulation results (and for well understood theoretical reasons) it has the advantage to take into account, the bias and asymmetry induced by the non-linear transformation. Even in this totally parametric framework, the parametric bootstrap may be of some use. Other examples are given in Efron (1985).

3.4 The semi-parametric Bootstrap

These last types of models yield very different Bootstrap methods each adapted to the particular models of interest. The idea once again is just to use the plug-in estimator of $k(.,P)$ (resp. $K(.,P)$, $L(.,P)$), for P in $\mathcal{P}_{\Theta,H}$ and some estimator $P_{\widehat{\theta}_n\ \widehat{\eta}_n}$ in $\mathcal{P}_{\Theta,H}$.

Example A.: The smooth Bootstrap (see DiCiccio, Hall, Romano, 1989).

Take $\mathcal{X}$ =$\mathbb{R}$, $\mathcal{P}_H = \{P_\eta$ such that $P_\eta << \mu$ and $\frac{dP_\eta}{d\mu}(x) = \eta(x),\ \eta \in H\}$, where $H = L^2(\mu)\}$, put $F_\eta(x) = P_\eta(] - \infty, x[)$.

In this model, the natural plug-in estimator is a smooth kernel estimator $\widehat{P}_n = K_{h_n} * P_n$. The smooth bootstrap consists then in studying the distribution of the statistic of interest under the distribution of $\widehat{P}_n{}^n$ of the observations. Observing that this distribution can be obtain by perturbing the observation X_i by some Gaussian noise with variance h_n. A simple Monte-Carlo algorithm for the smooth bootstrap is then the following : generate B times an i.i.d. sample $(\varepsilon_1, ..., \varepsilon_n) \rightsquigarrow K_{h_n}$, take a sample with replacement of size m from $(X_1 + \varepsilon_1, ..., X_n + \varepsilon_n)$; for each generated sample calculate the value of the statistic of interest. The empirical distribution of the obtained value is a Monte-Carlo approximation of the bootstrap distribution $k_m(.,\widehat{P}_n)$.

Assume that we are interested in $T_\alpha(F_\eta) = F_\eta^{-1}(\alpha),\ \alpha \in]0,1[$ where $F_\eta^{-1}(t) = \inf\{x,\ F_\eta(x) \geq t\}$.

Actually $T_\alpha(.)$ may be extended on the whole space $\mathbb{F}$. However if we estimate $T_\alpha(F_\eta)$ by $T_\alpha(F_n)$, we may be interested in the limiting behavior of $T_\alpha(F_n)$ and its asymptotic variance is given on $\mathcal{P}_H$ by

$$V_{as}(T(F_n)) = \frac{\alpha(1-\alpha)}{\eta(T_\alpha(F_\eta))^2}$$

and explicitly depends on the behavior of η at $T_\alpha(.)$. It is thus more reasonable in this context to work on $\mathcal{P}_H$...

Example B. The symmetric Bootstrap is left as an exercise. If the model is symmetric in some sense, a naive bootstrap (consisting in resampling residuals) will not performed well in that there is no reason for Q_n to be symmetric. If we have some information on the symmetry, it has to be incorporated into the bootstrap procedure for it to perform efficiently.

Example C. Residual resampling bootstrap

Reconsider the regression model

$$\mathcal{P}_{\Theta,H} = \{P_{\theta,\eta},\ P_{\theta,\eta}(y,x) = P_\varepsilon(\frac{y - x\beta}{\sigma})P_X(x),$$

$$\theta = (\beta, \sigma) \in \mathbb{R} \times \mathbb{R}^+,\ P_X \in \mathbb{P},\ P_\varepsilon \in \mathbb{P}^0,\ \eta = (P_X, P_\varepsilon)\}.$$

and recall that we have some estimators of $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\sigma}_n)$ (the least square estimators) and $\widehat{\eta}_n = (\widehat{P}_{X,n}, \widehat{P}_{\varepsilon,n})$ (respectively the empirical distribution of the X_i and the empirical distribution of the recentered residuals).

The bootstrap distribution of $\widehat{\beta}_n$ is simply in that framework the plug-in distribution

$$k_m(., P_{\widehat{\theta}_n, \widehat{\eta}_n})$$

Notice that generating $(Y_i^{(n)}, X_i^{(n)})$ $i = 1, ..., n$ i.i.d. with distribution $P_{\widehat{\theta}_n, \widehat{\eta}_n}(y, x) = \widehat{P}_{\varepsilon,n}(\frac{y - X\widehat{\beta}_n}{\widehat{\sigma}_n}|X = x)\ \widehat{P}_{X,n}(x)$ simply means that we have observations satisfying

$$Y_i^{(n)} = X_i^{(n)} \widehat{\beta}_n + \widehat{\sigma}_n \varepsilon_i^{*(n)}$$

where the $\varepsilon_i^{*(n)}$ $'s$ are i.i.d. $\widehat{P}_{\varepsilon,n}$, $X_i^{(n)}$ i.i.d. $\widehat{P}_{X,n}$. This is the usual presentation of the residual resampling method (see Freedman, 1981). Such artificial construction is mainly a simple Monte-Carlo device to generate approximation in an easy way. Fundamentally, the residual resampling method is simply a semiparametric plug-in bootstrap method. The fact that the residuals are recentered ensuring that $P_{\widehat{\theta}_n, \widehat{\eta}_n}$ is in $\mathcal{P}_{\Theta,H}$ is fundamental for the method to work in this case.

Exercise: (Freedman, 1984): How would you bootstrap the two stage least square estimator in a "linear" model where the X_i are correlated with the ε_i $'s$ but some instruments V_i orthogonal to the residuals are available. Assume that $E(X_i V_i) \neq 0$, $E(V_i \varepsilon_i) = 0$ and $(X_i, V_i, Y_i, \varepsilon_i)$ *i.i.d.*

Exercise : same exercise if in addition, ε_i is assumed to have a symmetric distribution around 0 (without assuming its distribution is dominated).

Exercise : choose your own (complex) model and propose the corresponding complex-bootstrap method. Try it by simulation and compare with the naive bootstrap or other methods.

4 Another way to see the bootstrap : resampling plans.

In the preceeding part, we have seen that Bootstrap distributions are essentially plug-in estimators (random distributions) of the distribution of a statistic. However the original presentation of Efron as well as the fact that we are considering the distribution of some r.v. $T_m(Y_{1,...,}Y_m)$ under $\widehat{P}_n{}^m$ plead in favor of considering the Bootstrap distribution as a conditional distribution. To be in agreement with the bootstrap literature the variables $(Y_1, ..., Y_m)$ will be now denoted by $\mathcal{X}_{m,n}^* = (X_{1,n}^*, ...X_{m,n}^*)$. $\mathcal{X}_{m,n}^*$ is called a bootstrap sample of size m (most of the time denoted by $(X_1^*, ...X_m^*)$). As n varies, the sequence $\mathcal{X}_{m,n}^*$ actually forms a triangular array of i.i.d. r.v.s with probability law $\widehat{P}_n$ conditionally to the data. The star should recall us that we are working with arrays, conditionally to $(X_1, ...X_n) = \mathbf{X}_n$.

For the naive bootstrap, $\widehat{P}_n = P_n$, the conditional distribution (denoted by P_r^* in the bootstrap literature) of the $X_{i,n}^*$ is characterized by

$$\Pr{}^*(X_{1,n}^* = X_i \; |\mathbf{X}_n) = \frac{1}{n} \;,\; i = 1, ..., n.$$

That is (and this is how practically one generate bootstrap samples), we have for an i.i.d. sample $(U_1, ..., U_m)$ uniformly distributed over $[0,1]$ (defined on some underlying, generally unspecified, probability space),

$$X_{i,n}^* = \sum_{j=1}^{n} X_j \; 1_{\{U_i \in [\frac{j-1}{n}, \frac{j}{n}]\}}$$

It follows that the empirical bootstrap distribution has the following representation

$$\begin{aligned} P_m^* &= \frac{1}{m} \sum_{i=1}^{m} \delta_{X_{i,n}^*} = \frac{1}{m} \sum_{j=1}^{n} \delta_{X_j} \left(\sum_{i=1}^{m} 1_{\{U_i \in [\frac{j-1}{n}, \frac{j}{n}]\}} \right) \\ &= \frac{1}{m} \sum_{j=1}^{m} W_{j,m,n} \delta_{X_j}, \end{aligned}$$

where $(W_{1,m,n}, W_{2,m,n}, ...W_{m,m,n})$ is a vector with multinomial distribution $Mult\left(m, (\frac{1}{n}, ..., \frac{1}{n})\right)$. The vector $\mathbf{W}_{m,n} = (W_{1,m,n}, W_{2,m,n}, ...W_{m,m,n})$ is called a **resampling plan** (Efron, 82) in that it tells us how one resamples the original data to create new artificial data.

The bootstrap distribution of a general functional $T(P)$ may now be interpreted as the distribution of $T(P_m^*)$ under the law of the resampling plan $\mathbf{W}_{m,n}$ with distribution $\mathcal{W}_{m,n}$, conditionally to the original sample that is

$$\begin{aligned} k_m(., P_n) &= P_r^*(T(P_m^*) \leq x \; |\mathbf{X}_n) \\ &= E_{\mathcal{W}_{m,n}}\{1_{\{T(P_m^*) \leq x\}} |\mathbf{X}_n\} \end{aligned}$$

and

$$K_m(., P_n) = \Pr{}^*(\tau_m(T(P_m^*) - T(P_n) \leq x \; |\mathbf{X}_n).$$

In the same way, one can artificially define resampling plans corresponding to jackknife or subsampling distributions : one may check that subsampling corresponds to weights with distribution defined by

$$P_W(W_{1,m,n} = i_1, W_{2,m,n} = i_2, ..., W_{n,m,n} = i_n) = \binom{n}{m}^{-1} \text{ with } \sum_{k=1}^{n} i_k = m, \;\; i_k \in \{0,1\}$$

The jackknife histogram may be obtained by choosing $m = n - 1$.

A large number of different type of weights have been considered in the Bootstrap literature. In particular, Bayesian statistician have given a non parametric Bayesian interpretation of this kind of weighted Bootstrap distribution. Based on some non parametric Bayesian considerations by Ferguson(1973), when the weights are chosen with Dirichlet distribution $D(m, \frac{1}{n}, ..., \frac{1}{n})$, the Bootstrap distribution may be interpreted as the posterior distribution of $T(P)$, when a Dirichlet prior is put on P. Such bootstrap method is called the **Bayesian bootstrap** (Rudin, 1981).

Bayesian Bootstrap clones have been investigated by Lo (1991) : this consists in generating weights of the form $W_{i,m,n} = \frac{nY_i}{\sum_{i=1}^{m} Y_i}$, with $(Y_1, ..., Y_m)$ i.i.d or by considering simply independent weights $W_{i,m,n} = Y_i$. A particular case is the so called **wild Bootstrap** which simply consists in choosing the Y_i as independent Rademacher random variable $\varepsilon_i = \left\{ \begin{smallmatrix} -1 \\ +1 \end{smallmatrix} \right.$ probability 1/2 (if we work with recentered values).

More generally, Mason and Newton (1992) proved that under fairly general conditions, if one only assumes that the $W_{i,m,n}\ 's$ are exchangeable then the main properties of the naive bootstrap remain asymptotically valid for the mean, under simple moment conditions on the weights. It can be shown that randomly weighted empirical p.m.'s lead to the same type of consistency results. So, there is a priori no reason to prefer Efron's scheme to almost any arbitrary random weighted scheme. This approach has leads to two basic questions. How well does the generalized bootstrap work? What are the differences between all the different weighted schemes? This is the subject of the monograph by Barbe and Bertail (1995). We will not pursue this approach here.

It should be however stressed that the mathematical tools used to study such quantities highly rely on probabilistic tools from Banach space probability theory (see Ledoux and Talagrand, 1992) : in particular proving some central limit theorem for the bootstrap somehow amounts in using Multiplicative Central Limit Theorems, that is conditional central limit theorem for randomly weighted empirical means (see van der Vaart and Wellner,1996).

5 Asymptotic validity

It is known in the robustness literature that the consistency of the plug-in rule is essentially a consequence of

1) the weak convergence of $\widehat{P}_n$ to P.

2) the continuity of the functional T for a topology compatible with the weak convergence.

Such results have similar translations in term of estimating sequence of distributions.

5.1 Some "adhoc" theorems.

Assume thus that $\mathcal{P}$ the set of probability measures on a measurable space $(\mathcal{X},\ \mathcal{A})$ is equipped with a topology $\mathcal{O}$ metrizing weak convergence. $\mathcal{O}$ may be generated by some metric denoted by h (typically in applications the uniform distance, Hellinger distance or Wasserstein distance, see these metrics below).

$(\mathcal{D},\rho)$ denote a separable metric space equipped with some metric ρ. In the study of the bootstrap, $(\mathcal{D},\rho)$ will be a space of distribution with a metric ρ metrizing weak convergence on $\mathcal{D}$.

Let us first recall the definition of equicontinuity for a sequence of functionals μ_n taking its value in $(\mathcal{D},\rho)$.

Definition : $\mu_n :\ (\mathcal{P},\mathcal{O}) \ \rightarrow (\mathcal{D},\rho)$ is said to be equicontinuous iff for some $N_0>0$

$$\forall P\in\mathcal{P},\ \forall\varepsilon>0,\ \exists V_\varepsilon(P) \text{ neighborhood of P } \subset\mathcal{P}\ ,$$

$$\forall Q\in V_\varepsilon,\ \forall n\geq N_0,\ \rho(\mu_n(P),\mu_n(Q))<\varepsilon$$

We then have actually the following somehow "ad hoc" result, which yields the asymptotic validity of the Bootstrap method.

Theorem 3 *If μ_n from $(\mathcal{P},\mathcal{O})$ to (D,ρ) is equicontinuous and $\widehat{P}_n$ is a $\mathcal{O}-$convergent estimator of $P\in\mathcal{P}$ in probability (resp. a.s.) then the corresponding plug-in estimator is convergent in probability (resp. a.s)*

$$\mu_n(\widehat{P}_n)-\mu_n(P)\underset{n\rightarrow\infty}{\rightarrow}0 \text{ Pr. } (resp.\ a.s.).$$

Proof : For any P, for any $\varepsilon>0$, by equicontinuity, there exists a neighborhood V_ε such that for any n,

$$\Pr(\rho(\mu_n(\widehat{P}_n),\mu_n(P))>\varepsilon)\leq\Pr(\widehat{P}_n\notin V_\varepsilon(P))\rightarrow 0.$$

Of course the main problem lies on the adequate choices of the distances (or topologies) which ensure the equicontinuity of μ_n, that is in other terms the stability of μ_n in the neighborhood of P. See the relevant literature on ideal metric, for instance in Rachev (1991).

Even if the functional fails to be equicontinuous, but if we have some pointwise convergence of $\mu_n(P)$; then the number of points at which the Bootstrap will fail is of "small" size. More precisely consider the following

Definition : N is a set of first category (a meager set) in $(\mathcal{P}, \mathcal{O})$ if it is the countable union of nowhere dense sets S_i

$$N = \overset{\infty}{\underset{i=1}{\cup}} S_i, \ \overset{\circ}{\overline{S_i}} = \emptyset.$$

Then the following theorem due to Putter and van Zwet(1996) shows that the set of probability at which the bootstrap fails is a set of first category.

Theorem 4 *If*

H_1 : *for n fixed sufficiently large,* $\mu_n(P)$ *is continuous as a functional of* P : $(\mathcal{P}, \mathcal{O}) \to (\mathcal{D}, \rho)$,

H_2 : $\mu_n(P) \to \mu(P)$, *as* $n \to \infty$,

H_3 : $\widehat{P}_n$ *is an O-convergent estimator of* P *in* $\mathcal{P}$

then there exists a set N *of first category in* $\mathcal{P}$ *such that for any* $P \in \mathcal{P}/N$

$$\mu_n(\widehat{P}_n) - \mu_n(P) \underset{n \to \infty}{\to} 0.$$

This theorem applies to all categories of bootstrap(s) and any bootstrap distributions that we have considered. If $(\mathcal{P}, \mathcal{O})$ is not complete, then N may be large (even in a parametric framework, in the sense that the Lebesgue measure of the set N may be strictly positive). If $(\mathcal{P}, \mathcal{O})$ is complete then one may isolate the submodels, for which the bootstrap fails: we will see then that it is possible to propose a semi-parametric bootstrap for the subclasses in question. We refer to the very interesting paper by Putter and van Zwet (1996) for other results in that direction. Beran (1997) in a parametric framework has also an interesting characterization (necessary and sufficient conditions) of the points at which the parametric bootstrap fails : these are exactly the points at which the statistic is not locally regular, yielding for instance the invalidity of the parametric Bootstrap at hyperefficiency points. These results strongly support the fact that one should be very carefull in Bootstrapping statistics, which have different limiting behaviours according to the value of the underlying probability (or parameter).

5.2 Some metrics

Before giving some particular examples, we introduce some metric that may be very useful in our context.

Let X, Y defined on the same probability space, taking their value on some separable Banach space $(B, ||.||)$. Denote the joint distribution of (X, Y) by Λ, and their respective marginal distributions $\Lambda(X, \infty)$, $\Lambda(\infty, Y)$.

Wasserstein or Mallows distance of order p.

$$DM_p(P, Q) = \inf_{\substack{(X,Y) \rightsquigarrow \Lambda \\ \Lambda(X,\infty)=P \\ \Lambda(\infty,Y)=Q}} \left(\{E_\Lambda ||X - Y||^p\}^{1/p}\right)$$

$L_p(\mathcal{X}) = \{P,\ E_P X^p < \infty\}$, equipped with $DM_p(P, Q)$ is a Banach space. These metric have the following interesting properties

$$DM_p(P^{*m}, Q^{*m}) \leq m DM_p(P, Q), \qquad (2)$$

where P^{*m} is the m convoluted distribution under P. Moreover, if we have a sequence of probabilities $P^{(n)}$ converging to P for DM_p i.e.

$$DM_p(P^{(n)}, P) \longrightarrow 0,$$

then $P^{(n)}$ converges weakly to P and

$$\int x^p dP^{(n)} \to \int x^p dP.$$

These convergences are a.s. if the original convergence is a.s., in probability if the convergence of the metric holds in probability.

Hellinger distance (parametric and semiparametric bootstrap).

On $\mathcal{P}_H = \{P_\eta$, such that $P_\eta << \mu$ and $\frac{dP_\eta}{d\mu}(x) = \eta(x)\}$, the Hellinger distance between P and Q is given by

$$H(P, Q) = \left(\int \left(\frac{dP}{d\mu}^{1/2} - \frac{dQ^{1/2}}{d\mu} \right)^2 d\mu \right)^{1/2}$$

The set $\mathcal{P}_H$ equipped with $H(P, Q)$ may be identified to $L^2(\mu)$, via the transformation $\eta \longrightarrow s = \eta^{1/2}$ (so that $\int s^2 d\mu = 1$).

Zolotarev metrics (semi-metric indexed by classes of functions).

The distance commonly used in the robustness literature to control the continuity of functionals on $\mathcal{X}$=$\mathbb{R}$ or $\mathbb{R}^p$ is Kolmogorov's distance (see e.g. Huber (1981)). The main reason for this, is that the Dvoretsky-Kiefer-Wolfowitz (1956) theorem provides a very precise control on $||P_n(]-\infty,.[) - P(]-\infty,.[)||_\infty = O_P(n^{-1/2})$. Nevertheless, this distance is not always strong enough to even ensure the continuity of simple functionals. For instance, the mean ($T(P) = \int x dP(x)$) is not continuous for this distance, even if we restrict P to belong to the set of all probability measures with finite mean.

For these reasons, consider the following generalization of Kolmogorov's distance: let $\mathcal{H}$ be a class of real valued measurable functions (in fact measurability is not necessary, in the sense that we can use outer measure and Hoffmann-Jorgensen's (1991) convergence: see 5.4), and define :

$$d_{\mathcal{H}}(P,Q) := \sup_{h\in\mathcal{H}} \mid \int hd(P-Q) \mid .$$

For $\mathcal{H}_K := \{\mathbb{I}\{. \leq y\}; y \in \mathbb{R}\}$, this defines Kolmogorov's distance in the real valued case. Taking $\mathcal{H}_{K,p} := \{| \ . \ |^p \ \mathbb{I}\{. \leq y\}; y \in \mathbb{R}\}$ on the set of probabilities with a finite p-th moment, ensures the continuity and differentiability of functionals of the form $T(P) = \int |x|^p dP(x)$ for $d_{\mathcal{H}_{K,p}}$. Many statistical functionals are continuous and even Fréchet-differentiable (see later) for $d_{\mathcal{H}}$ with a suitable $\mathcal{H}$.

Taking $\mathcal{H}_{BL} = \{f, \ f \ , \ ||f||_\infty \leq 1, \ f$ 1-lipschitz$\}$ defines the Bounded-Lipschitz distance which coincides with DM_1 on $L_1(\mathcal{X})$ $=\{P/ \ E_P X < \infty\}$. Another interesting case for studying the second order properties of the bootstrap (see below) is the metric indexed by the class

$$\mathcal{H}_K = \{f, \ ||f||_\infty \leq 1, \ ||f^{(i)}||_\infty \leq 1, \ i \geq K \geq 3\}.$$

$d_{\mathcal{H}_3}$ is called Dudley's metric. This metric with $K = 4$ is for instance considered in Giné (1996) to prove some second order results for the bootstrap. These metrics also satisfy a property similar to (2).

5.3 Asymptotic validity of the naive bootstrap in regular case

Most of the effort to understand the properties of the Bootstrap have been done on linear functionals $T(P) = E_P f(X)$ for smooth measurable functions f and on the empirical process that is by considering $T(P) = P$ (which may also be seen as a mean in an infinite dimensional space $T(P) = E_P \delta_X$). Validity of the Bootstrap for these quantities leads to the asymptotic validity of a very large class of regular functionals (Fréchet and/or continuously Hadamard differentiable functionals).

By asymptotic validity of the naive bootstrap it is generally meant that either

$$K_n(.,\widehat{P}_n) - K_n(.,P) \to 0$$

or

$$L_n(.,\widehat{P}_n) - L_n(.,P) \rightarrow 0$$

either in probability or a.s (generally for the uniform metric but other metric may be of interest). For this it is sufficient to prove that

$$K_n(.,\widehat{P}_n) - K(.,P) \rightarrow 0$$

or

$$L_n(.,\widehat{P}_n) - L(.,P) \rightarrow 0$$

Because of the conditional representation

$$K_n(.,\widehat{P}_n) = \Pr{}^*(\tau_n(T(P_n^*) - T(P_n) \leq x \ |\mathbf{X}_n)$$

the convergence of the Bootstrap statistic $\tau_n(T(P_n^*) - T(P_n)$ may then be interpreted as "convergence in law conditionally on the sample in probability" (if convergence of $K_n(.,\widehat{P}_n)$ is studied in probability) or "convergence in conditional law a.s. along the sequence", if the convergence of the bootstrap distribution is a.s.: this is denoted

$$\tau_n(T(P_n^*) - T(P_n) \ |\mathbf{X}_n \xrightarrow{L^*} K_P \ , \ in \ P-pr. \ (resp. \ a.s.) \text{ as } n \rightarrow \infty.$$

These notions are usefull but sometimes a bit confusing in the bootstrap literature.

5.3.1 The case of the mean

Let us begin with examples 1 and 2 given before

Example 1 : Let $\theta = E_P X$ take its value in $(B, ||.||)$ a separable Banach Space. $\mathcal{P} = \{P, \ E_P||X||^2 < \infty\}$ is equipped with Mallows metric DM_2 and $\rho = DM_2$ then with $\tau_m = m^{1/2}$. It follows from property (2) that any bootstrap method such that $\widehat{P}_n- > P$ a.s. and such that $\int x^2 \widehat{P}_n(x) \rightarrow \int x^2 P(dx) \ a.s.$ (resp. in $P-$Pr) is asymptotically valid for the mean (Bickel and Freedman,1981).

Then $DM_2(K_m(.,\widehat{P}_n), K_m(.,P)) \rightarrow 0$ implies pointwise convergence of $K_m(.,\widehat{P}_n)$ to $K(.,P)$, but also by the false Dini Theorem (also known as a Polya theorem), uniform convergence over $\mathbb{R}$ (this theorem states that pointwise convergence to a continuous increasing bounded function implies uniform convergence).

This is true in particular for the naive bootstrap $(m = n)$ but also for the smooth bootstrap (provided that the smoothing parameter is chosen adequately), the parametric bootstrap etc...

The next example shows further why we may run into trouble at some particular points. This is actually a particular case of degenerate V-statistics for which the naive bootstrap does not work (see Bretagnolle, 1983).

Example 2 : Let $\theta = (E_PX)^2 \in \mathbb{R}$ on $\mathcal{P} = \{P,\ \sigma^2(P) = E_P(X - E_PX)^2 < \infty\}$ endowed with DM_2, for $T_n = (\overline{X}_n)^2$,we have the decomposition
$T_n - \theta = 2E_PX\ n^{-1}\sum(X_i - E_PX) + (\overline{X}_n - \theta)^2$

If $E_PX \neq 0$ choose $\tau_n(P) = n^{1/2}$, and we have $K_n(.,P) \to \Phi(./S(P))$, with $S(P)^2 = 4(E_PX)^2\sigma^2(P)$ and it is immediate using the result in example 1 to see that the naive Bootstrap works (apply Slutsky Theorem).

If $E_PX = 0$, choose $\tau_n(P) = n$, then $K_n(.,P) \to \Pr(\sigma^2(P)\ \chi^2(1) \leq x)$ and the equicontinuity condition fails, if we do not stay in $\mathcal{P}_0$. The naive bootstrap fails as well as can be seen from the following decomposition : we have under P_n

$$T_n(X_1^*,...,X_n^*) - T_n = 2T_n\ n^{-1}\sum_{i=1}^{n}(X_i^* - \overline{X}_n)\ +\ (\overline{X}_n^* - \overline{X}_n)^2.$$

So that we have

$$\Pr{}^*\{n(T_n^* - T_n) \leq x|\ \mathbf{X}_n\} \underset{n\to\infty}{\to} \Pr\{2N_1N_2 + N_2^2 \leq x/\sigma^2(P)\}\ \ a.s.,$$

where N_1 and N_2 are two independent normalized Gaussian r.v.

Notice that if we choose $m \neq n$ then

$$m(T_m(Y_1,...,Y_m) - T_n) = 2m^{1/2}T_n\ \{m^{-1/2}\sum_{i=1}^{m}(Y_i - T_n)\}\ +\ (m^{1/2}(\overline{Y}_n - T_n))^2$$

Thus if $\frac{m}{n} \to 0$, then the first part of the r.h.s degenerates to 0 (because $m^{1/2}T_n \to 0$ Pr) and the m out of n bootstrap works in probability.

Moreover since we know by the law of iterated logarithm (LIL) that

$$\overline{lim}\ n^{1/2}/\log(\log(n))^{1/2}T_n\ \leq \sqrt{2},$$

then if $\frac{m}{n}\log\log(n) \to 0$ then the m out of n bootstrap works also a.s.

Notice that this model is in some aspect semiparametric. On $\mathcal{P}^0$= $\{P,\ E_PX = 0,\ E_PX^2 \prec \infty\}$ the adequate (semiparametric) estimator P belonging to $\mathcal{P}^0$ is

$$\widehat{P}_n = n^{-1/2}\sum\delta_{X_i - \overline{X}_n}.$$

The equicontinuity problem mentioned before can be easily solved in that case by choosing a metric which discriminates between $\mathcal{P}/\mathcal{P}^0$ and $\mathcal{P}^0$ and makes $K_n(.,P)$ equicontinuous everywhere. We then can try to build a compatible convergent estimator $\widehat{P_n}$.

For instance, take here

$$DM(P,Q) = \begin{cases} DM_2(P,Q)\ if \text{ both } P \text{ and } Q \text{ belongs to } \mathcal{P}/\mathcal{P}^0\ or \text{ to } \mathcal{P}^0 \\ \\ \qquad\qquad\qquad 1\ else \end{cases}$$

and the threshold estimator

$$\widehat{P}_n = n^{-1} \sum \delta_{X_i - \overline{X}_n 1_{\{|\overline{X}_n| \leq n^{-1/4}\}}}.$$

We then have the validity of the bootstrap for any P. This is an idea that can also be used in time-series for bootstrapping AR model possibly with unit roots.

The following theorem due to Giné and Zinn (1989) establish a necessary and sufficient condition for the naive bootstrap of the mean. The moral of this theorem is essentially that the naive bootstrap of the mean works a.s. iff $E_P X^2 < \infty$.

Theorem 5 *Under $E_P X^2 < \infty$, then*

$$K_n(., P_n) -> \Phi(./\sigma) \ a.s.$$

Inversely if there exists a normalization τ_n such that $K_n(., P_n)$ converges a.s. to a non degenerate distribution then $E_P X^2 < \infty$, $\tau_n n^{-1/2} \rightarrow C > 0$ and the asymptotic distribution is gaussian.

Deep theorems in the same spirit as well as important results in the case of Banach valued r.v. may be found in the monograph of Giné(1996). We will not treat this aspects in these lectures. However, the validity of the bootstrap of empirical process (seen as a mean in infinite dimensional Banach space) is essential for many statistical applications and we will give a few elements about it in the next paragraph.

5.4 Empirical processes (*).

Modern empirical process theory is closely connected to probability in Banach space (see Ledoux and Talagrand, 1991). We give here very basic elements about bootstrap of empirical process indexed by classes of functions, that somehow generalize to infinite dimensional space the result of the previous part. Interested reader should definitely read the monograph by Giné (1996) and chap. 3.6 of van der Vaart and Wellner (1996).

In empirical processes indexed by some class of real functions $\mathcal{F}$, we are interested on the behavior of the (infinite) dimensional vectors of the form

$$\{n^{1/2}(\int f dP_n - \int f dP), \ f \in \mathcal{F}\}$$

but also in the rate of uniform convergence

$$\sup_{f\in\mathcal{F}}\{|n^{1/2}(\int fdP_n-\int fdP)|\}=n^{1/2}d_{\mathcal{F}}(P_n,P).$$

For example, if $\mathcal{F}=\mathcal{H}_K:=\{\mathbb{I}\{.\leq y\};y\in\mathbb{R}\}$, this means that we are interested in $n^{1/2}(F_n(t)-F(t))$, seen as an empirical process (in $t\in\mathbb{R}$), as well as in the uniform convergence of this quantity. We will assume, that the class $\mathcal{F}$ admits a square integrable envelop, that is there exists $H>\eta>0$ such that for any $f\in\mathcal{F}$,

$$|f|\leq H$$

and

$$E_PH(X)^2<\infty.$$

To better see that a signed measure (for instance P_n-P) acts as a linear function on $\mathcal{F}$, it is now common to use the notation $\mu f=\int fd\mu$. We thus have $(P_n-P)f=\int fdP_n-\int fdP$ and since the envelop is finite a.s. everywhere, it is bounded. An empirical process indexed by $\mathcal{F}$ may thus be seen as a random element of $l^\infty(\mathcal{F})$ (the space of all bounded functions from $\mathcal{F}$ to $\mathbb{R}$) equipped with the metric, $||z||_{\mathcal{F}}=\sup_{f\in\mathcal{F}}|z(f)|$. $(l^\infty(\mathcal{F}),\ ||z||_{\mathcal{F}})$ is a (generally non separable) Banach space and $\{n^{1/2}(\int fdP_n-\int fdP),\ f\in\mathcal{F}\}$ is essentially a standardized mean $n^{-1/2}(\sum_{i=1}^n(\delta_{X_i}-P))$ in $(l^\infty(\mathcal{F}),\ ||z||_{\mathcal{F}})$. One of the main problem when dealing with sums in infinite dimensional (non separable) space is measurability of events. For instance, the empirical process $n^{1/2}(F_n-F)$ seen as a random sequence in the Skorokhod space D of cadlag functions distribution endowed with the supremum norm is not Borel measurable. To solve these problems in the case of D, the uniform topology in $\mathcal{D}$ may be replaced by the Skorokhod metric (see other solutions in Pollard, 1984). The ideas developed by Hoffmann-Jorgensen (1991) have lead to a general solution of this kind of problems. The idea is that if one works with outer probabilities (resp. outer expectation) that is if one extends the initial probability to non measurable events, by setting $\mathbb{P}(\mathbb{A})=\inf\{P(B),\ A\subset B,\ B$ measurable$\}$ (resp. outer expectation $\mathbb{E}_PT=\inf\{E_PU,\ U$ measurable and $U\geq T\}$, then one may somehow forget the measurability assumptions. However because we take infinums in the definitions, this may cause some problems with permutation of limits and integrals. The whole theory has been rebuilt with these tools : most of the results of empirical process obtained on D are still valid (it should be noticed however that the standard Fubini theorem does not hold in this context). We refer to van der Vaart and Wellner(1996) for details and general results. We need however the following basic definitions. Expectations and probabilities are now to be understood as outer expectations for non-measurable events.

Definitions : 1) A P-Brownian Bridge G_P is a centered gaussian process in $L_2(P)$ such that

$$cov(G_P f, G_P g) = cov_P(f, g).$$

2) A class $\mathcal{F}$ is said to be pregaussian if the restriction of the process G_P to $l^\infty(\mathcal{F})$ induces a tight Borel probability measure on $l^\infty(\mathcal{F})$.

3) A class $\mathcal{F}$ is said to be P-Donsker (or satisfies a Central Limit Theorem under P) if $\mathcal{F}$ is pregaussian and

$$n^{1/2}(P_n - P) \to_w G_P$$

in $l^\infty(\mathcal{F})$. Here weak convergence is metrized by the bounded Lipchitz metric on the space $l^\infty(\mathcal{F})$ that is for $\mathbf{X}$ and $\mathbf{Y}$ in $l^\infty(\mathcal{F})$

$$d_{BL}(\mathbf{X}, \mathbf{Y}) = \sup_{b \in BL_1(l^\infty(\mathcal{F}))} |\mathbb{E}b(X) - \mathbb{E}b(Y)|,$$

where $BL_1(l^\infty(\mathcal{F}))$ is the set of all 1-lipchitz function on $l^\infty(\mathcal{F})$. Then we have the following theorem, which is the empirical process version of Theorem 5 due to Giné and Zinn (1990).

Theorem 6 *If $\mathcal{F}$, with square integrable envelop F, is P-Donsker, then $\mathcal{F}$ is also bootstrap P-Donsker in probability and a.s. that is*

$$d_{BL}(n^{1/2}(P_n^* - P_n), G_P) \to 0 \ \textit{a.s. or Pr. (in outer probability)}$$

Inversely if $\mathcal{F}$ with envelop F is image admissible Suslin (see Dudley 1984, 10.3) and if there exist a centered gaussian process G (with Radon law in $l^\infty(\mathcal{F})$) such that $d_{BL}(n^{1/2}(P_n^* - P_n), G) \to 0$ a.s. then $\mathcal{F}$ is P-Donsker, $E_P F(X)^2 < \infty$ and $G = G_P$.

Praestgaard and Wellner (1993) have obtained similar results for general weighted Bootstrap (that is for general resampling plans). These results pave the way to obtain the validity of the bootstrap(s) of many statistical functionals. In particular the following paragraph shows how it is possible to get the validity of the bootstrap for Fréchet differentiable functionals with respect to some metric indexed by classes of function or for functionals, which are Hadamard differentiable (tangentially to some well chosen set of functions which are Donsker).

5.5 Bootstrap of Fréchet and Hadamard Differentiable functionals

Differentiability of functionals on some probability space is a tool to generalize Slutsky theorem (obtained in a finite dimensional setting). See Von Mises (1947), Filippova (1961) for early works, Rieder (1994), van der Vaart and Wellner (1996) for modern accounts. The idea is essentially to try to linearize the functional of interest, with some control on the remainder terms: this is the so called **Delta-method**. However, from a practical point of view, one has to be very careful with the choice of the metrics, which makes the functional differentiable.

Let us first recall a few basic notions of differentiability (see e.g. Flett, 1980, Gill, 1989 for details). The linear space of measures engendered by $\mathcal{P}$, say $<\mathcal{P}>$ is assumed to be endowed with a metric d and we assume that we may extend T to $<\mathcal{P}>$. If $\mathcal{C}$ is a class of subsets of $<\mathcal{P}>$, we say that T is m-times $\mathcal{C}$-differentiable at $P \in \mathcal{P}$, if there exists some p-linear functions $D^pT_P, 0 \leq p \leq m$ such that for any $t \in [0,1]$, and for any $Q \in <\mathcal{P}>$,

$$T(P+tQ)) - T(P) = \Sigma_{p=1}^{m}(p!)^{-1}t^pD^pT_P.Q^p + R^{(m)}(t,P,Q),$$

with for any $C \in \mathcal{C}$,

$$\lim_{t\to 0}\sup_{Q\in C}|t|^{-m} \mid R^{(m)}(t,P,Q) \mid = 0$$

Special cases of interest are

(i) $\mathcal{C}$ is the set of all bounded subsets of $<\mathcal{P}>$, T is said to be m-times Fréchet-differentiable.

(ii) $\mathcal{C}$ is the set of all compact subsets of $<\mathcal{P}>$, T is m-times compact-differentiable or Hadamard differentiable.

(iii) $\mathcal{C}$ is the set of all subsets of $<\mathcal{P}>$ with only one element, T is said to be m-times Gâteaux-differentiable.

We have the following (strict) implications between the different notions of differentiability: $(i) \Rightarrow (ii) \Longrightarrow (iii)$.

On a finite dimensional space or a Hilbert space, Riesz Theorem ensures that the p-linear functions D^pT admit an integral representation (else we will assume that such representation exists)

$$D^pT_P(Q-P)^p = \int\ldots\int T^{(p)}(x_1,\ldots,x_p,P)d(Q-P)(x_1)\ldots d(Q-P)(x_p),$$

where the functions $T^{(p)}$ are not necessarily unique. In this chapter, we will be essentially interested in $T^{(1)}$ and $T^{(2)}$. Because they are not uniquely defined, we can normalize them by setting

$$E_P T^{(1)}(X,P)=0,$$
$$T^{(2)}(X,Y,P)=T^{(2)}(Y,X,P)\ \text{(symmetry)},$$
$$E_P T^{(2)}(X,y,P)=0 \text{ for every } y\ a.s.$$

Then $\{1, T^{(1)}, T^{(2)}\}$ are orthogonal for the scalar product cov_P. Notice that $T^{(1)}$ is the influence function as defined in Huber (1981). $T^{(1)}(x,P)$ and $T^{(2)}(x,y,P)$ are also called the first and second order canonical gradient. A very convenient way to obtain them is to calculate the Gâteaux-derivatives

$$T^{(1)}(x,P)=\frac{d}{dt}T((1-t)P+t\delta_x)|_{t=0},$$
$$T^{(2)}(x,y,P)=\frac{\partial^2}{\partial t\partial t'}T((1-t-t')P+t\delta_x+t'\delta_y)|_{t=t'=0}.$$

As far as Fréchet differentiability is concerned, it is convenient to use a distance $d(.,.)$ which satisfies (as the Kolmogorov distance does)

$$d(P+t(Q-P),P)=\mid t\mid d(P,Q)$$

Then , if T is m-times Fréchet-differentiable at P for such distance d, it is easy to show that there exists a function $\epsilon^{(m)}(.,P)$ such that $\epsilon^{(m)}(.,P)$ is continuous at 0 with $\epsilon^{(m)}(0,P)=0$ and such that for any probability $Q\in\mathcal{P}$

$$T(Q)-T(P)=\Sigma_{p=1}^{m}D^{(p)}T_P(Q-P)/p!+R^{(m)}(Q,P)$$

with

$$R^{(m)}(Q,P)=d(P,Q)^m\epsilon^{(m)}(d(P,Q),P).$$

It has often been pointed out in the robustness literature, that Fréchet-differentiability with respect to Kolmogorov's distance easily yields central limit theorem (CLT). But very few functionals are Fréchet-differentiable with respect to Kolmogorov's distance, whereas they are often Gâteaux-differentiable or Hadamard-differentiable. Assuming continuous compact differentiability of T in a neighborhood of P, and using the validity of the bootstrap empirical process, Gill (1989) and Pons and Turkheim (1989) have established the strong consistency of Efron's bootstrap. However continuous compact differentiability is quite a strong assumption and does not allow for a precise control of the rate of convergence of Bootstrap distributions. The choice of an adequate metric which makes a functional Fréchet differentiable is not an easy task. However in many cases, it is possible to find a class of functions $\mathcal{F}$, which makes the functional Fréchet differentiable for the metric $d_{\mathcal{F}}$ (see Dudley, 1990, Barbe and Bertail, 1995).

If T is 1-Fréchet differentiable for a metric $d_{\mathcal{F}}$, then we have the following representation

$$T(P_n) - T(P) = n^{-1} \sum T^{(1)}(X_i, P) + d_{\mathcal{F}}(P_n, P)\epsilon^{(1)}(d_{\mathcal{F}}(P_n, P), P)$$

If $\mathcal{F}$ has a strictly positive envelop H, Fréchet differentiability at P for $d_{\mathcal{F}}$ implies that $|T^{(1)}(x, P)| \leq C_P \ H(x)$, for some constant $C_P > 0$ (Barbe and Bertail, 1995). Thus if we can control that $d_{\mathcal{F}}(P_n, P) = O_P(n^{-1/2})$, which is the case if the class $\mathcal{F}$ is P-Donsker then it is immediate to see that the limiting distribution of $n^{1/2}(T(P_n) - T(P))$ is given by the first term in this development, which is gaussian as soon as $0 < E_P T^{(1)}(X_i, P)^2 < \infty$.

A similar representation holds then for the bootstrap version

$$T(P_n^*) - T(P) = n^{-1} \sum T^{(1)}(X_i^*, P) + d_{\mathcal{F}}(P_n^*, P)\epsilon^{(1)}(d_{\mathcal{F}}(P_n^*, P), P).$$

But the results on bootstrap empirical process essentially ensure that $d_{\mathcal{F}}(P_n^*, P) \leq d_{\mathcal{F}}(P_n^*, P_n) + d_{\mathcal{F}}(P_n, P) = O_P(n^{-1/2})$ if $\mathcal{F}$ is P-Donsker. We thus have

Theorem 7 *It there exists a class of function $\mathcal{F}$ with square integrable envelop H, such that*

(1) T is Frechet differentiable at P for $d_{\mathcal{F}}$ with $0 < E_P T^{(1)}(X, P)^2$

(2) $\mathcal{F}$ is P-Donsker

then

$$K_n(x, P_n) - K_n(x, P) \to 0 \ pr.$$

(with $\tau_n = n^{1/2}$).

Remark: If we have a precise control of the remainder (for instance, the remainder may be of order $O_P(d_{\mathcal{F}}(P_n, P)^2)$ then the Donsker hypothesis is too strong since we just have to control that $n^{1/2} d_{\mathcal{F}}(P_n, P)\epsilon^{(1)}(d_{\mathcal{F}}(P_n, P), P)$ converge to 0 (and that we have bootstrap version of this control). For instance, if the remainder is of order $O_P(d_{\mathcal{F}}(P_n, P)^2)$, we just need to show that $d_{\mathcal{F}}(P_n, P) = o_P(n^{-1/4})$, as well as $d_{\mathcal{F}}(P_n^*, P_n) = o_P(n^{-1/4})$. See Barbe and Bertail (1995) for results in that direction in the case of general weighted bootstrap(s).

In many situation, it is even easier to linearize the functional of interest (for instance by calculating the first Gâteaux derivative) and to show that the remainder is small for the original statistic and its bootstrap version.

Fréchet differentiability of higher order will permit a better control of the approximation and allow to obtain Edgeworth expansions of differentiable functionals (see Pfanzagl, J., 1985), which are fundamental to obtain the rate of convergence of the bootstrap distributions.

6 Second order validity of the bootstrap of mean functionals

Until now, there is absolutely no reason to prefer a bootstrap type approximation (a plug-in approximation) of $k(.,P)$, $K(.,P)$ or $L(.,P)$ (which are convergent but random and generally approximated by some Monte-Carlo calculus) over the asymptotic approximation, when it is tractable. The only justification is in term of ease or laziness: getting a bootstrap approximation is effortless. One does not even need to choose a standardization for the original statistic. Subsampling , the m out of n bootstrap and the bootstrap (when it is asymptotically valid) can provide valid approximation, even if we do not standardize the statistic of interest and even if the limiting distribution is very complicated.

In this paragraph, we will explain why the naive Bootstrap can yield very accurate approximations of the distributions of interest , when this statistic is adequately standardized. The accuracy of the distribution is simply measured by the uniform rate of convergence of the plug-in estimators $k(.,\widehat{P}_n)$, $K(.,\widehat{P}_n)$, $L(.,\widehat{P}_n)$. Having an idea of this rate of convergence is not only important from a theoretical point of view but also is the only real justification for using a random approximation rather than a deterministic or simpler approximation. To compare these rates with the rate of the asymptotic distribution, we need to analyze more precisely the rate of convergence of $K_n(.,P)$ to $K(.,P)$. For simplicity we focus here on the case of the mean.

6.1 Berry-Esséen Lemma and Edgeworth expansions.

Berry Esséen and Edgeworth expansions give a precise control of the rate between the true distribution and the asymptotic distribution. We have in particular the following results for the functional

$$T(P) = E_P f(X) = Pf$$

and

$$T(P_n) = n^{-1} \sum_{i=1}^{n} f(X_i) = P_n f.$$

We will denote the variance of f

$$S^2(P) = E_P(f(X) - E_P f(X))^2$$

and, for a function g, its skewness is given by

$$k_{3,P}(g) = \frac{E_P(g(X) - E_P g(X))^3}{[E_P(g(X) - E_P g(X))^2]^{3/2}}.$$

In the following $C_1, ...C_{7....}$ denote universal constants. The Berry Esséen theorem is a benchmark to understand how well the asymptotic distribution approximate the true distribution.

Proposition 8 ***Berry Esséen and Edgeworth expansions***
If $E_P|f(X)|^3 < \infty$, *we have*

$$||K_n(.,P) - \Phi(./S(P))||_\infty \leq C_1\, k_{3,P}(|f|)/\sqrt{n}$$

If the following Cramer condition holds

$$\overline{\lim}_{t\to\infty}|E_P \exp(itf(X))| < 1)$$

and if $E_P|f(X)|^4 < \infty$,*then we have also have the following Edgeworth expansions of order 2, uniformly in* x,

$$K_n(x,P) = \Phi(x/S(P)) - \frac{k_{3,P}(f)}{6n^{1/2}}(\{x/S(P)\}^2 - 1)\phi(x/S(P)) + O(n^{-1})$$

and

$$L_n(x,P) = \Phi(x) + \frac{k_{3,P}(f)}{6n^{1/2}}(2x^2+1)\phi(x) + O(n^{-1}), \tag{3}$$

These results also hold for triangular arrays of random variables provided that $E_P|f(X)|^{4+\eta} < \infty$, *for* $\eta > 0$.

These expansions essentially mean that at the first order the statistic is Gaussian, but the second order takes into account the possible statistical asymmetry of the problem by taking into account the skewness of the distribution. Edgeworth expansions up to any order K (that is up to an order $O(n^{-K/2})$) may be obtained under additional moments conditions (See Bhattacharya, Rao, 1986).

An approximation of the distribution $K_n(.,P)$ *or* $L_n(.,P)$ is said to be **second order correct** if it improves over the asymptotic distribution in term of rate of convergence. As an illustration, if we have an estimator of $k_{3,P}(f)$, for instance, its empirical counterpart $k_{3,P_n}(f)$ (obtained by the plug-in rule) then it is immediate to see that (uniformly in x)

$$\Delta_n^E(x) = L_n(x,P) - \left(\Phi(x) + \frac{k_{3,P_n}(f)}{6n^{1/2}}(2x^2+1)\phi(x)\right) = o(n^{-1/2})\ a.s.$$

This means that "an empirical Edgeworth expansion" is second order correct a.s. for estimating $L_n(x,P)$. However one should remark, that it is not a distribution function (it may be negative or larger than 1). If in addition, we have $E_P|f(X)|^6 < \infty$, then

$k_{3,P_n}(f) - k_{3,P}(f) = O_P(n^{-1/2})$ and we can get the better rate $\Delta_n^E(x) = O_P(n^{-1})$. In this case we will say that the empirical Edgeworth expansion is **second order correct up to** $O(n^{-1})$. It is easy to see that, if one has a LIL for $k_{3,P_n}(f) - k_{3,P}(f)$, it is easy to get an a.s. rate of convergence for this approximation.

Marcinkiewicz Sygmund LLN shows that according to the hypothesis made on the moments of $f(X)$, many a.s. rate may be obtained for $k_{3,P_n}(f)$. For further use, we recall it now, in a useful form (see Chow and Teicher,1988, Theorem 2, p. 125, for a proof).

Lemma 9 *LLN Marcinkiewicz-Zygmund*

If $X_1, ..., X_n$, are i.i.d P, put $R_n = \sum_{i=1}^n f(X_i)$, then

for any $p \in]0, 2[$, there exits $-\infty < c < +\infty$, such that

$$\frac{R_n - nc}{n^{1/p}} \to 0 \;\; a.s.$$

iff $E_P|f(X)|^p < \infty$, in which case $c = E_P f(X)$ if $2 > p \geq 1$ and c is arbitrary (in particular c=0) for $p \in]0, 1[$.

For instance this result implies that, under the condition, $E_P f(X)^4 < \infty$, which is the minimal condition for the existence of the Edgeworth expansion of $L_n(., P)$ to hold up to $O(n^{-1})$, we have the precise rate $\Delta_n^E(x) = o(n^{-3/4})$ (take $f(x) = x^3$ and $p = 4/3$ in the lemma).

6.2 Bootstrap versions or empirical Edgeworth expansions

Let us first begin by a second simple proof of the validity of the naive Bootstrap using Berry-Esséen Lemma. Assume that $E_P f(X)^2 < \infty$. Notice that we have $E_{P_n}|f(X)|^3 < \infty$, thus by the Berry-Esséen theorem

$$||K_n(., P_n) - \Phi(./S(P_n))||_\infty \leq C_1 \; k_{3,P_n}(|f|)/\sqrt{n}.$$

Now we have

$$S(P_n) \to S(P) \; a.s. \text{ as } n- > \infty$$

and using Marcinkiewicz-Zygmund LLN

$$k_{3,P_n}(|f|)/\sqrt{n} - > 0, \; a.s. \text{ as } n- > \infty.$$

This yields directly the asymptotic validity of the Bootstrap. Such a proof is easy to adapt in many setting in which we have a Berry Esséen Bound and some L_p control.

From this result, we also immediately get that under $E_P f(X)^p < \infty,\ p \geq 4$,

$$||K_n(.,P_n) - K_n(.,P)||_\infty = O_P(n^{-1/2})$$

Thus there is no real improvement over the asymptotic in this case. This is due to the fact that, if the statistic of interest is not standardized, the main contribution of the difference between the bootstrap and the true distribution is of order $||\Phi(./S(P_n)) - \Phi(./S(P))||_\infty = O(S(P_n) - S(P))$, which is typically of order $n^{-1/2}$ under $E_P f(X)^4 < \infty$. This is actually a general rule that one should have in mind : for statistics which are not standardized, there is no reason to prefer the bootstrap distribution to the asymptotic distribution, if it is available and tractable. To get better approximation we have to standardize adequately the statistic.

Under $E_P|f(X)|^{4+\eta} < \infty$, it is easy to check that the same Edgeworth expansions hold for the bootstrap distributions. To show this, we have essentially to check that $\overline{\lim}_{t\to\infty}|E_{P_n}\exp(itf(X))| < 1$. Indeed, all the empirical moments and all the quantities appearing in the Edgeworth expansion are finite and converges by the L.L.N. to the true value under P.

The fact that $\overline{\lim}_{t\to\infty}|E_{P_n}\exp(itf(X))| < \infty$ is also a consequence of the L.L.N., with uniform convergence over $[0,M]$ for any $M > 0$. We then have

$$L_n(x,P_n) = \Phi(x) + \frac{k_{3,P_n}(f)}{6n^{1/2}}(2x^2+1)\phi(x) + O_P(n^{-1}).$$

that is a Bootstrap distribution is similar to an empirical Edgeworth expansion up to $O_P(n^{-1})$. This yields, for $K_0 = 2e^{-3/4}\sqrt{\frac{2}{\pi}} \simeq 0.76$

$$||L_n(x,P_n) - L_n(x,P)||_\infty = \frac{|k_{3,P_n}(f) - k_{3,P}(f)|}{6n^{1/2}}K_0 + O_P(n^{-1})$$

If $E_P f(X)^6 < \infty$, we obtain

$$||L_n(x,P_n) - L_n(x,P)||_\infty = O_P(n^{-1}).$$

Notice that from this (and from refined bound on the remainder $O(n^{-1})$) we can get L.I.L., asymptotic distributions as well as exact bounds (with constants depending on P in the $f-$unbounded case) for the bootstrap distribution, etc... A few practical results are synthesized in the following theorem. See Singh (1981) for the first results in that direction.

Theorem 10 *Assume that f is a measurable function satisfying the Cramer condition $\overline{\lim}_{t\to\infty}|E_P\exp(itf(X))| < \infty$ then if $E_P f(X)^p < \infty,\ p \geq 4$ we have*

$$||K_n(.,P_n) - K_n(.,P)||_\infty = O_P(n^{-1/2})$$

and if in addition $E_P f(X)^6 < \infty$, *then*

$$||L_n(x, P_n) - L_n(x, P)||_\infty = O_P(n^{-1}).$$

Let $t_n(1-\alpha) = L_n^{-1}(1-\alpha, P_n)$ *be the quantile of order* $1-\alpha$ *of the studentized bootstrap distribution then we have*

$$L_n(t_n(1-\alpha), P) = 1 - \alpha + O(n^{-1})$$

The last result (refined in Hall 1986, 1992) means, that we can construct second order correct unilateral confidence intervals by using the quantile of the studentized bootstrap distribution. For comparison if $k_3(P) \neq 0$ then the asymptotic distribution yields a quantile $u_{1-\alpha} = \Phi^{-1}(1-\alpha)$, which is such that $L_n(u_{1-\alpha}, P) = 1 - \alpha + O(n^{-1/2})$. See further developments of the importance of pivoting and "prepivoting" in Beran (1987,1988).

Remark 1 : on symmetric distribution.

When the skewness of P is equal to 0, one has to be very careful on how to construct an "efficient" bootstrap. The model under consideration is rather $\{P,\ P \in \mathbb{P},\ k_3(P) = 0\}$. Following the ideas exposed in the introduction, it is much more natural in that case to find an estimator $\widehat{P}_n$ which is such that $k_3(\widehat{P}_n) = 0$ for instance by symmetrizing P_n,

$$\widehat{P}_n = \frac{1}{2n} \sum_{i=1}^{n} (\delta_{X_i} + \delta_{-X_i}).$$

then, in that case, under $E_P X^8 < \infty$ it can be shown that $||L_n(x, P_n) - L_n(x, P)||_\infty = O_P(n^{-3/2})$ (This is left as an exercise).

Remark 2 : on bilateral confidence intervals.

A similar remark holds if one is interested in constructed bilateral confidence interval. In the case $T(P) = E_P f(X)$, the quantity of interest is rather $|T(P_n) - T(P)|$, which, under the Cramer conditions and $Ef(X)^6 < \infty$, has for symmetry reasons, an Edgeworth expansion of the form

$$\begin{aligned}\widetilde{L}_n(x, P) &= \Pr\left(\frac{n^{1/2}|T(P_n) - T(P)|}{S_n} \leq x\right) \\ &= \Phi(x) - \Phi(-x) + n^{-1}\{k_{4,P}(f)p_1(x) + k_{3,P}(f)^2 p_2(x)\}\phi(x) + O(n^{-2}),\end{aligned}$$

where p_1 and p_2 are polynomials and $k_{4,P}(f)$ is the kurtosis. In that case, if $Ef(X)^8 < \infty$, we have (exercise)

$$||\widetilde{L}_n(.,P_n) - \widetilde{L}_n(.,P)||_\infty = O_P(n^{-3/2}).$$

But if $\widetilde{t_n}(\alpha)$ is the quantile of order α of $\widetilde{L}_n(.,P_n)$ then it is possible to show with more refined arguments (see Hall 1992, 1986) that

$$\widetilde{L}_n(\widetilde{t_n}(\alpha),P) = \alpha + O(n^{-2}).$$

This means that the two sided symmetric bootstrap confidence interval

$$[T(P_n) \pm \widetilde{t_n}(1-\alpha)S_n n^{-1/2}],$$

is asymptotically correct up to an error of size $O(n^{-2})$. Notice that using the asymptotic distribution (a symmetric confidence interval based on $u_{1-\alpha/2}$) would yield in that case an error of size $O(n^{-1})$.

Remark 3 : it was argued in the introduction that a smooth bootstrap is more adapted for fractile. Indeed it can be shown theoretically (Falk and Kaufmann, 1991) that the naive bootstrap is worse than the asymptotic distribution (with an error of size $O(n^{-1/4})$) and that the smooth bootstrap is second order correct in that case. However, the best error size is typically $O_P(n^{-3/4})$ instead of $O_P(n^{-1})$.

6.3 Subsampling distribution and extrapolation

We have seen before that subsampling or m out of n bootstrap is universally valid. This universality has of course a price : a less accurate approximation of the true distribution. The following arguments are taken from Bertail (1997). To give an idea of the deficiency of this approximation consider again the case of the mean. Babu and Singh(1985) 's Edgeworth expansion for sampling without replacement from a finite population, with $\frac{b_n}{n} \underset{n\to\infty}{\longrightarrow} 0$, yields the Edgeworth expansion for the subsampling studentized distribution

$$L_m^U(.,P_n) = \binom{n}{m}^{-1} \sum_{\sigma\in S} \mathbb{I}_{\left\{\tau_m\left(T_m(X_{\sigma(1),\ldots,}X_{\sigma(m)})-T(P_n)\right)\leq x\right\}}$$

$$= \Phi\left(x(1-\frac{m}{n})^{-1/2}\right) + m^{-1/2}\frac{k_{3,P_n}(f)}{6}(2x^2+1)\phi(x) + o(m^{-1/2}), \qquad (4)$$

If $\frac{m}{n} = o(m^{-1/2})$ and since, under $E_P X^6 < \infty$, a Taylor expansion of (4) immediately leads to

$$L_m^U(.,P_n) = \Phi\left(x\right) + m^{-1/2}\frac{k_3}{6}(2x^2+1)\phi(x) + o_p(m^{-1/2}). \qquad (5)$$

Notice that the condition $\frac{m}{n} = o(m^{-1/2})$ simply means that, to take into account the error induced by the undersampling scheme, m should be small enough, typically of order $o(n^{2/3})$. Of course the two Edgeworth expansions (3) and (5) do not match in contrast to the usual bootstrap. It is thus clear that $L_m^U(.,P_n)$ is not a good approximation. The order of the error $O(m^{-1/2})$ is worse (at the best $O(n^{-1/3})$) than the usual Gaussian approximation of order $O(n^{-1/2})$.

However a combination of these two different approximations leads to a better one. Indeed, put

$$L_m^{Int}(.) \equiv \left(\frac{m}{n}\right)^{1/2} L_m^U(.,P_n) + \left(1 - \left(\frac{m}{n}\right)^{1/2}\right)\Phi(.),$$

which is a particular case of Richardson's extrapolation, we immediately see that

$$\sup_x |L_m^{Int}(x) - K_n(x,P)| = o(n^{-1/2}). \tag{6}$$

Some refinements of Edgeworth expansions and a correct standardization of the subsampling distributions taking into account the finite sample population factor $(1-\frac{m}{n})$ shows that one may choose $m = O(n^{2/3})$ and get an optimal error of size $O(n^{-5/6})$ in (6), (see Bertail, 1997). The interpolated subsampling approximation is thus second order correct and yields second order correct unilateral confidence intervals up to a rate $O(n^{-5/6})$ which is not so far from the optimal rate $O(n^{-1})$ of the bootstrap in the case of the mean. One advantage of subsampling and extrapolation is also computational, since calculating a statistic on much smaller subsample may be less computationally expensive than a regular bootstrap. However the meaning of these results is that the naive Bootstrap is better...when it works.

This interpolation or extrapolation result is in fact very general. If one knows a pivotal asymptotic distribution for a studentized statistic, the interpolation of the asymptotic distribution and the subsampling distribution yields second order correct approximations (Bertail,1997). It can even be shown that extrapolation of several undersampling distribution correctly standardized are also second order correct (Bertail and Politis, 2001). However, despite some interesting attempts, the main problem is the adequate choice of the subsampling size, which has actually its analog in time-series and random fields. When we have a precise control of Edgeworth expansions, then it is somehow possible to get some optimal choice but adaptative choice of m is still difficult.

These results also implies that just proving that a Bootstrap method is second order correct (i.e. up to $o(n^{-1/2})$) is somehow insufficient to have a good appreciation of the technique. An exact rate or even better an exact bound is the only way to distinguish between procedures (just like when we study a regular estimator...). This fact becomes particularly important when we move toward time series.

7 Summary

In this chapter, we have seen that

-There is a lot of bootstraps : each model can generate its own bootstrap.

-Undersampling (m out of n bootstrap or subsampling) is a universal tool to get asymptotically correct approximations and confidence intervals. Deficiency of the method may be compensated by interpolation techniques.

-The validity of the naive bootstrap $(m = n)$ relies on some equicontinuity properties of the sequence of distributions of interest. When it works it can enjoy second order properties that is it can yield very accurate approximations in its studentized version.

-The rates of convergence of the bootstrap(s) distribution(s) to the true distribution are of prime importance to evaluate the potentiality of the method.

Annex 1 : Hoeffding inequality for U-statistics.

First the recentered U-statistic $U_m - E_P\ {}_m\omega(X_1,...,X_m)$ has centered kernel $\widetilde{\omega}(x_1,...,x_m)$ bounded by $2M$. Now define the average of the kernel of this centered kernel over the $q=[\frac{n}{m}]$ disjoint blocks of size m that can be constructed with the data

$$h_m(x_1,...,x_n)=\frac{1}{q}\sum_{k=0}^{q}\widetilde{\omega}(x_{km+1},x_{km+2},....,x_{k(m+1)})$$

Then one gets (for $\mathcal{P}_n^n$ the set of all permutation in $\{1,...,n\}$)

$$\begin{aligned}U_m - E_P\ {}_m\omega(X_1,...,X_m) &= E\{\widetilde{\omega}(x_1,x_2,....,x_m)|(X_{(1)},.....,X_{(n)})\}\\ &= E\{h_m(x_1,...,x_n)|(X_{(1)},.....,X_{(n)})\}\\ &= \frac{1}{n!}\sum_{\sigma\in\mathcal{P}_n^n} h_m(X_{\sigma(1)},...,X_{\sigma(n)})\end{aligned}$$

By convexity we have

$$\begin{aligned}E_P\exp(\lambda|U_m - E_P\ {}_m\omega(X_1,...,X_m)|) &\leq \frac{1}{n!}\sum_{\sigma\in\mathcal{P}_n^n} E_P\exp(\lambda|h_m(X_{\sigma(1)},...,X_{\sigma(n)})|)\\ &= E_P\exp(\lambda|h_m(X_1,...,X_n)|)\\ &\leq E_P\exp(\lambda h_m(X_1,...,X_n)) + E_P\exp(-\lambda h_m(X_1,...,X_n))\\ &\leq \left\{E_P\exp(\frac{\lambda}{q}\widetilde{w}_m(X_1,...,X_m))\right\}^q\\ &+\left\{E_P\exp(-\frac{\lambda}{q}\widetilde{w}_m(X_1,...,X_m))\right\}^q\end{aligned}$$

Now we have for any centered bounded r.v. Y, $|Y|\leq 1$

$$E_P\exp(tY)\leq\exp(t^2/2)$$

(write $ty=(\frac{1-y}{2})t+\frac{(1+y)}{2}(-t)$ and use convexity of the exponential as well as the bound $cosh(x)\leq\exp(x^2/2)$).

It follows that

$$E_P\exp(\lambda|U_n|)\leq 2\exp(\lambda^2/(8qM^2)$$

Now apply the standard Chebyshev's arguments (exponentiation+Markov inequality) to get for any $\lambda > 0$

$$\Pr\{|U_m| \geq t\} \leq 2\exp(-\lambda t + \lambda^2/(8qM^2))$$

A straightforward optimization in λ leads to choose $\lambda = t/(4qM^2)$ and gives the result.

References

Arcones, M.A., Giné, E.. (1989). The bootstrap of the mean with arbitrary bootstrap sample size, *Ann. Inst. H. Poincare Probab. Statist.*, **25**, 457-481.

Barbe, Ph. and Bertail, P. (1995). *The Weighted Bootstrap.* Lecture Notes in Statistics, **98**, Springer Verlag, New-York.

Babu, G. and Singh, K. (1985). Edgeworth expansions for sampling without replacements from finite populations. *J. Multivariate Analysis*, **17**, 261-278.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets, *Biometrika*, **74**, 457-468.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements, *J. Amer. Statist. Assoc.*, **83**, 687-697.

Beran, R. (1997). Diagnosing bootstrap success. *Ann. Inst.Statist. Math.*, **49**, 1–24.

Bertail, P. (2003). Empirical likelihood in some semiparametric model, preprint CREST, submitted.

Bertail, P. (1997). Second order properties of an extrapolated bootstrap without replacement: the i.i.d. and the strong mixing cases, *Bernoulli*, **3**, 149-179.

Bertail, P., Politis , D.N. (2000.). Extrapolation of subsampling distribution estimators : the i.i;i. and strong mixing cases, *Canad. Journ. Stat.* , **29**, 4, 1-16.

Bhattacharya, R.N., Qumsiyeh, M. (1989). Second order comparisons between the bootstrap and empirical edgeworth expansions, *Ann. Statist.*, **17**, 160-169.

Bhattacharya, R.N., Rao, R. (1986). *Normal Approximation and Asymptotic Expansions*, Krieger, Melbourn.

Bickel, P.J., Freedman, D.A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196-1217.

Bickel, P., Götze, F. and van Zwet, W.R. (1997). Resampling fewer than n observations: Gains, losses and remedies for losses, *Statist. Sinica*, **7**, 1-31.

Bickel, P.J., Klaasen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient Estimation for semiparametric Models.* Johns Hopkins Univ. Press.

Borovkov, A.A. (1998). *Mathematical Statistics.* Gordon and Breach publishers.

Bretagnolle, J. (1983). Lois limites du Bootstrap de certaines fonctionnelles, *Ann. Inst. H. Poincaré, Statist. Probab.*, **XIX**, 281-296.

Chow, Y., Teicher, H. (1988) : *Probability Theory, Independence, Interchangeability, Martingales*, Springer, New York.

DiCiccio, T.J., Hall, P., Romano, J.P. (1989). On smoothing the bootstrap, *Ann. Statist.*, **17**,692-704.

DiCiccio, T.J., Romano, J. (1988) : A review of bootstrap confidence intervals (with discussions), *J. Roy. Statist. Soc.*, **50**, 338-370.

Dudley, R.M. (1984) : *A Course On Empirical Process*, Ecole d'été de Saint-Flour, *Lecture Notes in Math.*, **1097**, Springer-Verlag, New-York.

Dudley, R.M. (1990) : Non linear functionals of empirical measures and the bootstrap, in *Probability in Banach Spaces*, 7, E. Eberlein, J., Kielbs, M.B. Marcus eds., Birkhauser, Boston.

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

Efron, B. (1982) : *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS−NF 38, S.I.A.M., Philadelphia.

Efron, B. (1985) : Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, 42-58.

Efron, B., Tibshirani, R. (1993) : *An Introduction to the Bootstrap*, Chapman and Hall.

Falk, M., Kaufmann, E. (1991) Coverage probabilities of bootstrap confidence intervals for quantiles, *Ann. Statist.*, **19**, 485-495.

Filippova, A.A. (1961) : Mises' theorem and the asymptotic behaviour of functionals of empirical distribution function and its statistical applications, *Theor. Probab. Appl.*, **7**, 25-57.

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, **1**, 209-230.

Flett, T.M. (1980). *Differential Analysis*, Cambridge university press, London.

Freedman D.(1981). Bootstrapping Regression Models, *Ann.Statist.*, **9**,1218-1228.

Freedman D. (1984). On Bootstrapping Two-stage Least Squares Estimates in Stationary Linear Model , *Ann. Statist.*,**12**, 827-842.

Gill, R.D. (1989). Non and semi-parametric maximum likehood estimators and the Von Mises method , *Scand. J. Statist.*, **16**, 97-128.

Giné, E. (1996). *Lectures on some aspects of the Bootstrap*, in Ecole d'été de Calcul des probabilités de St Flour, Springer.

Giné, E., Zinn, J. (1989). Necessary conditions for the bootstrap of the mean, *Ann. Statist.*, **17**, 684-691.

Giné, E., Zinn, J. (1990). Bootstrapping general empirical functions, *Ann. Probab.*, **18**, 851-869.

Gray, H., Schucany, W. and Watkins, T. (1972). *The Generalized Jackknife Statistics.* Marcel Dekker, New-York.

Hall, P. (1986). On the bootstrap and confidence intervals, *Ann. Statist.*, **14**, 1431-1452.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer, New-York

Hampel, F. (1974). The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.*, **69**, 383-393.

Hoffmann−Jorgensen J. (1991). *Stochastic Processes on Polish Spaces*, PhD. Aarhus universitet.

Huber, P.J. (1981). *Robust Statistics*, Wiley, New-York.

LeCam , L. (1986). *Asymptotic methods in statistical decision theory.* Springer Verlag.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces.* Springer Berlin.

Lo, A. Y. (1991). Bayesian bootstrap clones and a biometry function, *Sankhya, Ser. A*, **53**, 320-333

Manski, C.F. (1988). *Analog Estimation Methods in Econometrics*, Chapman and Hall, New-York.

Mason, D.M., Newton, M. A. (1992). A rank statistic approach to the consistency of a general bootstrap, *Ann. Statist.*, **20**, 1611-1624.

von Mises, R. (1947). On the asymptotic distribution of differential functions, *Ann. Math. Statist.*, **18**, 309-348.

Pfanzagl, J. (1985). *Asymptotic Expansion for General Statistical Models*, *Lecture Notes in Statist.*, **31**, Springer-Verlag, Berlin.

Politis, D.N. and Romano, J.P. (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, **22**, 2031-2050.

Politis, D.N., Romano, J.P. , Wolf, M. (1999). *Subsampling*, Springer.

Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer Verlag New-York.

Pons, O., Turkheim, E. (1989). Méthodes de Von Mises, Hadamard diff,rentiabilité et bootstrap dans un modèle non paramétrique sur un espace métrique, *C. R. Acad. Sci. Paris*, **308**, 369-372.

Praestgaard, J., Wellner, J.A. (1993). Exchangeably weighted bootstrap of the general empirical process, *Ann. Probab.*, **21**, 2053-2086.

Putter, H., van Zwet, W.R. (1996). Resampling: consistency of substitution estimators, *Ann. Statist.*, **24**, 2297-2318

Quenouille, M.H. (1949). Approximate tests of correlation in time-series, *J. Roy. Statist. Soc.*, *Ser.* B, **11**, 68-84.

Rachev, S.T. (1991). *Probability Metrics and the Stability of Stochastic Models*, Wiley, New York.

Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer N.Y.

Rubin, D. (1981). The Bayesian bootstrap, *Ann. Statist.*, **9**, 130-134.

Serfling, J. (1980). *Approximation Theorem of Mathematical Statistics*, Wiley, New-York.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187-1195.

van der Vaart, AW, and Wellner, JA (1996). *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge series in Statistical and Probabilistic Mathematics.