

Asymptotics in nonparametric statistics: some recent developments

Recenti sviluppi dei metodi asintotici in statistica non parametrica

Pier Luigi Conti¹

Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma “La Sapienza”

Riassunto: In questo lavoro si fornisce una rassegna di alcuni recenti sviluppi dei metodi asintotici in statistica non parametrica. L'enfasi è posta principalmente sui processi empirici indicizzati da funzioni, di cui viene evidenziato, anche attraverso esempi, il fondamentale ruolo da essi svolto. Vengono poi discusse applicazioni alla consistenza di stimatori, alla valutazione della velocità di convergenza, e allo studio di distribuzioni limite. Infine, viene discusso il metodo delta funzionale.

Keywords: Weak convergence, empirical processes, consistency, asymptotic distribution, metric entropy, Vapnik-Chervonenkis dimension, delta-method.

1. Introduction

Asymptotic theory plays a primary role in statistics, for several important reasons.

(i) It provides, in many cases, a useful tool for *approximating* the actual sampling distribution of a statistical quantity of interest. As well-known, in many cases the exact probability distribution of an estimator (or a test-statistic) possesses a complicate form, and frequently it cannot be even expressed in a closed form. Even more importantly, it depends on the population distribution, which is unknown. This case is particularly important in nonparametric statistics, where the *form* of the population distribution is unknown. On the other hand, in several cases the *large sample* distribution of estimators can be obtained virtually effortless. Its form is frequently simple, and this is appealing for applications, because of the relative simplicity in finding out confidence intervals and/or rejection regions of a given size. Of course, asymptotic theory only provides an approximation of exact sampling distributions. Hence, it would be important to evaluate, even roughly, the quality of such an approximation.

(ii) At a more abstract level, asymptotic theory is important since it provides an answer to the following question: “What is the behaviour of a statistical procedure (an estimator or a test-statistic) when the sample size increases?” A fairly natural requirement is that the behaviour of a statistical procedure should be better and better as the sample size increases, *i.e.* the gain due to a larger sample size should be a greater precision. In other words, when the sample size increases, the risk of error should decrease. This is actually the well-known problem of the *consistency* of a statistical procedure.

(iii) A question connected with (ii), but somewhat deeper, is the evaluation of the *speed* at which the “error” of a statistical procedure decreases as the sample size increases. In

¹Address for correspondence: Dipartimento di Statistica, Probabilità e Stat. Appl.; Università di Roma “La Sapienza” - P.le A. Moro, 5; 00185 Roma; Italy. E-mail: pierluigi.conti@uniroma1.it

²This work is partially supported by a MIUR research grant.

technical terms, the problem is to evaluate the *rate of convergence* of an estimator.

(iv) A problem frequently encountered by statisticians is to choose among different statistical procedures as solutions of the same problem. For a finite sample size, such a choice is very difficult, and in many cases impossible. However, when the sample size is large the problem becomes simpler. As the sample size increases, the structure of the risk of error connected to a statistical procedure become simpler and more transparent, and this drastically simplifies the problem described above. At a more elaborated level, this is the problem of studying the *asymptotic efficiency* of an estimator.

Asymptotic theory in nonparametric statistics has received a considerable impulse in the last 15 years. This is basically due to two reasons. First of all, the use of nonparametric techniques is becoming very popular, due to their excellent flexibility, as well as to the availability of fast computers. In the second place, since the beginning of the '90s, excellent mathematical tools are available. They are (a) recent and important developments in the theory of empirical processes, and (b) the availability of an “efficient” differential calculus for maps defined on functions spaces.

In the present review, our first aim is to show the main ideas behind the techniques used in (i)-(iv). Although statistical problems are different, the theoretical treatment of the main asymptotic properties rests on a few, fairly simple ideas. This is undoubtedly the main merit of the modern asymptotic theory based on empirical processes, that provide a general setting for many different problems. Our main effort is to enucleate such ideas, and to present them in a unitary way.

2. Basic definitions and examples

A population is formalized as a d -dimensional random variable (r.v.) X , taking values in $\mathcal{X}$, a subset of $\mathbb{R}^d$. The *population probability distribution* is a *probability measure* P_0 on $\mathbb{R}^d$ (equipped by the Borel σ -field $\mathcal{B}(\mathbb{R}^d)$). Furthermore, we will denote by $F_0(x) = Pr_{P_0}(X \leq x) = P_0((-\infty, x])$ the *population distribution function* (d.f., for short). Given two vectors $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d)$, $y \leq x$ means that $x_i \leq y_i$ for every $i = 1, \dots, d$. Let $(X_n; n \geq 1)$ be a sequence of independent “copies” of the population X , i.e. a sequence of independent and identically distributed (*i.i.d.*) r.v.s (defined on a probability space $(\Omega, \mathcal{A}, \pi)$), such that $Pr_{P_0}(X_i \leq x) = F_0(x)$ for every $i \geq 1$. Of course, a sample of size n is the collection $X_1, \dots, X_n$ of r.v.s.

The population probability distribution P_0 is assumed unknown. Our only assumption, at this stage, is that P_0 is a member of the class $\mathcal{P}$ of all possible population distributions assumed as “reasonable” for the population X . The family $\mathcal{P}$ completely defines the *statistical model* we deal with. A statistical model is *parametric* when the members of $\mathcal{P}$ can be indexed by a Euclidean (finite-dimensional) parameter. A statistical model is *nonparametric* when $\mathcal{P}$ is the set of all probability distributions on $\mathcal{X}$ (with possibly some restrictions, such as continuity). A model is *semiparametric* when the members of $\mathcal{P}$ can be written in the form $P_{\theta, \nu}$, where θ is a Euclidean parameter and ν lies to an infinite-dimensional family of probability distributions. Although in the present review we mainly focus on nonparametric models, many technical tools extend to semiparametric models.

A “statistical quantity of interest”, conventionally a *parameter*, is essentially a function of the probability population distribution, and is denoted by $\theta = \theta(P)$. In particular, $\theta_0 = \theta(P_0)$ is the “true population parameter”. The set $\Theta = \{\theta = \theta(P); P \in \mathcal{P}\}$ is the *parameter space*. Since a probability distribution $P \in \mathcal{P}$ is completely specified by its

distribution function $F(x) = Pr_P(X \leq x) = P((-\infty, x])$, from now on we will use as equivalent the notations $\theta(P)$ and $\theta(F)$.

On the basis of a sample $X_1, \dots, X_n$, a “natural estimator” of the population probability distribution P_0 is the *empirical distribution* P_n , which is the discrete probability distribution giving mass $1/n$ to the sample points $x_1, \dots, x_n$. In symbols: $P_n(x_i) = 1/n$ for every $i = 1, \dots, n$. If B is a subset of $\mathcal{X}$, it is then immediate to see that

$$P_n(B) = \frac{\#(x_i \in B)}{n}.$$

The expectation of a real-valued function $g(\cdot)$ under P (P_n , resp.) is denoted by

$$Pg = E_P[g(X)] = \int g(x) dP, \quad P_n g = \frac{1}{n} \sum_{i=1}^n g(x_i) = \int g(x) dP_n. \quad (1)$$

In particular, $P_0 g = \int g(x) dP_0$ is the expectation of $g(X)$ under the “true population probability distribution”. Moreover, if we consider the indicator function of a set B (i.e. $I_B(x) = 1$ if $x \in B$, and $I_B(x) = 0$ otherwise) then it is easy to see that $PI_B = P(B)$, and $P_n I_B = P_n(B)$.

The reason why P_n is a natural estimator of the population probability distribution is simple. For every fixed set B , the r.v. $nP_n(B)$ = number of sample points in the set B possesses binomial distribution: $Pr_{P_0}(nP_n(B) = k) = \binom{n}{k} P_0(B)^k (1 - P_0(B))^{n-k}$, $k = 0, 1, \dots, n$. Hence, the expected value and the variance of $P_n B$, under P_0 , are $P_0(B)$, $P_0(B)(1 - P_0(B))/n$, respectively. For a fixed B , $P_n(B)$ is a *consistent* estimator of $P_0(B)$. In fact, from the strong law of large numbers it is easy to deduce that $P_n(B)$ converges almost surely (a.s., for short) to $P_0(B)$. In symbols: $P_n(B) \xrightarrow{a.s.} P_0(B)$ as n goes to infinity. Furthermore, from the central limit theorem, it is not difficult to deduce that $\sqrt{n}(P_n(B) - P_0(B))$ tends in distribution, as n goes to infinity, to a normal r.v. with mean zero and variance $P_0(B)(1 - P_0(B))$.

A *statistical estimate* of the parameter $\theta_0 = \theta(P_0)$ is a function $\hat{\theta}_n = \hat{\theta}_n(P_n)$ of the empirical distribution.

Example 1 (Moments) The simplest example of parameter is the population mean: $\theta(P_0) = P_0 x = \int x dF_0(x)$. More generally, the k th moment is given by: $\theta(P_0) = P_0 x^k = \int x^k dF_0(x)$.

A simple estimator is then $\hat{\theta}_n = P_n x^k = \int x^k dP_n(x) = n^{-1} \sum X_i^k$, the k th sample moment. If the k -th population moment $\theta_0 = P_0 x^k$ exists finite, from the strong law of large numbers it follows that $\hat{\theta}_n$ is a consistent estimator of θ_0 : $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ as $n \rightarrow \infty$. Furthermore, from the central limit theorem we have that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ tends in distribution, as the sample size increases, to a normal r.v. with mean zero and variance $P_0 x^{2k} - (P_0 x^k)^2$. $\square$

Example 2 (Population distribution function and density) The simplest example of a *functional* parameter is the population d.f., i.e. the whole function $\theta(P_0) = F_0(\cdot)$. Suppose that $\mathcal{X} \subseteq \mathbb{R}$, and that the d.f. $F_0(\cdot)$ is absolutely continuous. Then, there exist a function $f_0(\cdot)$ for which the relationships $F_0(x) = \int_{-\infty}^x f_0(t) dt$, and $f_0(x) = dF_0(x)/dx$ (almost everywhere w.r.t. the Lebesgue measure) hold true. The function $\theta(P_0) = f_0(\cdot)$ is (a version of) the *density function* corresponding to the probability distribution P_0 , and is a “parameter” of considerable interest in statistics.

Since $F_0(x)$ is nothing more than the probability of the half-line $(-\infty, x]$, a natural estimator of $F_0(x)$ is the *empirical distribution function* (e.d.f., for short):

$$F_n(x) = P_n((-\infty, x]) = \frac{\# \text{ of } X_i \leq x}{n} = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x)}.$$

For every fixed x $nF_n(x)$ possesses a binomial distribution: $Pr_{P_0}(nF_n(x) = k) = \binom{n}{k} F_0(x)^k (1 - F_0(x))^{n-k}$, $k = 0, 1, \dots, n$. By the strong law of large numbers, $F_n(x)$ converges almost surely to $F_0(x)$, the true population d.f.: $F_n(x) \xrightarrow{a.s.} F_0(x)$ as $n \rightarrow \infty$, $\forall x \in \mathbb{R}^k$. Furthermore, by the central limit theorem, $F_n(x)$ is asymptotically normal with mean $F_0(x)$ and variance $F_0(x)(1 - F_0(x))$. In symbols: $\sqrt{n}(F_n(x) - F_0(x)) \xrightarrow{d} N(0, F_0(x)(1 - F_0(x)))$ as $n \rightarrow \infty$, $\forall x \in \mathbb{R}^k$, where $\xrightarrow{d}$ denotes convergence in distribution, and $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 . $\square$

Example 3 (Monotone dependence measures) Let $g(\cdot)$ be a real valued, convex and non-decreasing function, such that $g(0) = 0$. Suppose further that the population (X, Y) takes values in $\mathbb{R}^2$, and denote by $F_0(x, y)$, $F_{01}(x)$, $F_{02}(y)$ the joint and marginal d.f.s, respectively. The following (non-normalized) general measure of monotone dependence is considered in Cifarelli *et al.* (Cifarelli *et al.* (1996)):

$$\theta(F_0) = \int \int \{g(|F_{01}(x) + F_{02}(y) - 1|) - g(|F_{01}(x) - F_{02}(y)|)\} dF_0(x, y). \quad (2)$$

With obvious notation, a “natural” estimator is then:

$$\hat{\theta}_n = \int_{\mathbb{R}^2} \{g(|F_{1n}(x) + F_{2n}(y) - 1|) - g(|F_{1n}(x) - F_{2n}(y)|)\} dF_n(x, y). \quad (3)$$

It is not difficult to see that $\hat{\theta}_n$ can be expressed in terms of sample ranks. Up to normalization factors, and up to terms $O(n^{-1})$, when $g(z) = z$ it reduces to Gini’s cograduation index, and when $g(z) = z^2$ it reduces to Spearman’s rank correlation coefficient. Asymptotic properties of the estimator (3) are studied in Cifarelli *et al.* (1996). $\square$

Example 4 (Quantile function) Suppose that the r.v.s X_i s are real valued, and consider a real number $0 < y < 1$. The y th population quantile is then defined as $Q_0(y) = F_0^{-1}(y) = \inf\{x : F_0(x) \geq y\}$. The *quantile population function* is then the function $Q_0(\cdot)$. In this example, we take as $\theta(P_0)$ the whole function $Q_0(\cdot)$. As far as an estimator of $\theta(F_0)$ is concerned, take a fixed $0 < y < 1$, and let $Q_n(y) = F_n^{-1}(y) = \inf\{x : F_n(x) \geq y\}$ be the y th *sample quantile*. It easily seen that $Q_n(y) = X_{n:[ny]}$, where $X_{n:1} \leq \dots \leq X_{n:n}$ are the ordered sample observations. The *sample quantile function* is then $\theta_n = Q_n(\cdot)$. $\square$

Example 5 (Cumulative hazard function) The cumulative hazard function is of primary interest in survival analysis, reliability, etc.. Suppose that T_i s are *i.i.d* “failure times”, taking non-negative values, and let F_{T0} be their d.f.. Denote by F_{T0}^- its left-continuous version: $F_{T0}^-(t) = \lim_{\epsilon \downarrow 0} F_{T0}(t - \epsilon)$ (if F_{T0} is continuous, then F_{T0}^- coincides with F_{T0}). The cumulative hazard at point the t is defined as:

$$\Lambda_{T0}(t) = \int_0^t \frac{1}{1 - F_{T0}^-(x)} dF_{T0}(x).$$

In survival analysis, the failure times T_i s are not properly observed. What are commonly observed are *right censored data*, *i.e.* pairs (X_i, Δ_i) s, where $X_i = \min(T_i, C_i)$, C_i being a censoring time, and $\Delta_i = 1$ if $T_i \leq C_i$, $\Delta_i = 0$ otherwise. Denote by P_{X0} the “true” probability distribution of X_i s, and by $F_{X0}(x) = Pr_{X0}(X_i \leq x)$ the d.f. of X_i s. Assume further that C_i s are *i.i.d.* r.v.s independent of T_i s, and define the “sub-distribution function” $F_{10}(x) = Pr_{X0}(X_i \leq x, \Delta_i = 1)$. Taking into account that $1 - F_{X0}^- = (1 - F_{T0}^-)(1 - F_C^-)$ (where F_C is the d.f. of C_i s), and that $dF_{10} = (1 - F_C^-) dF_{X0}$, it is

not difficult to see that the relationship

$$\Lambda_{T_0}(t) = \int_0^t \frac{1}{1 - F_{X_0}^-(x)} dF_{10}(x)$$

holds. The quantity of interest $\theta(F_{X_0}, F_{10})$ is here the cumulative hazard function $\Lambda_{T_0}(\cdot)$.

On the basis of sample data, F_{X_0} and F_{10} can be estimated *via* their empirical counterparts: $F_{X_n}(x) = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x)}$, $F_{1n}(x) = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x, \Delta_i=1)}$. In this way, we obtain the *Nelson-Aalen estimator* of the cumulative hazard function

$$\hat{\theta}_n = \Lambda_n(t) = \int_0^t \frac{1}{1 - F_{X_n}^-(x)} dF_{1n}(x). \square \quad (4)$$

Example 6 (Nonparametric regression) The regression model takes the form $Y = g_0(x) + U$, where Y is the *response variable*, x is a *covariate*, and U is the error. Conditionally on x , the error U is a zero mean r.v.: $E[U|x] = 0$, so that $g_0(x) = E[Y|x]$. The function $g_0(x)$ is the *regression function* of Y w.r.t. x . When the covariate x is fixed in advance, before observing the corresponding response Y , the regression model is a *fixed design* model. On the other hand, when x is a realization of a r.v. X , then the regression model is a *stochastic design* model. Denote by $P_{U|x}$ the probability distribution of the error U given the covariate x , and (in the case of a stochastic design model) by P_X the probability distribution of X . In both cases the quantity of interest is the regression function g_0 . In symbols: $\theta(P_{Y|\cdot}) = g_0(\cdot)$. A sample of size n , in the present case, consists of n observations (Y_i, x_i) , linked by the model

$$Y_i = g_0(x_i) + U_i, \quad i = 1, \dots, n$$

where the errors U_i s are assumed to be independent r.v.s, but not necessarily identically distributed. In this case, the parameter of interest is the function $g_0(\cdot)$. $\square$

Example 7 (Ruin probability in collective risk theory) An insurance company receives claims of size $C_1, C_2, \dots$ at random times $T_1, T_1 + T_2, \dots$. The most common assumption (Andersen (1957)) is that the two sequences $(T_i; i \geq 1)$, $(C_i; i \geq 1)$ are independent, and that each of them is composed by *i.i.d.* r.v.s. Denote by F_{T_0}, F_{C_0} the d.f.s of T_i s and C_i s, respectively. For every $t \geq 0$, let $N(t)$ be the number of claims in the time interval $(0, t]$. If the company possesses an initial capital $K > 0$, and receives premiums linearly at a rate $p > 0$, then its reserve at time t is equal to $S(t) = K + pt - \sum_{i=1}^{N(t)} C_i$. The company is *ruined* as soon as its reserve becomes negative. Its *ruin probability* is $\pi_0 = \Pr(\inf_{t \geq 0} S(t) < 0)$. Now, π_0 can be computed in a closed form only in very special cases. However, a convenient approximation can be obtained under additional assumptions. Let $W_i = -pT_i$, denote by F_{W_0} its d.f., assume that $G_{C_0}(u) = \int_0^\infty e^{ux} dF_{C_0}(x)$, $G_{W_0}(u) = \int_0^\infty e^{ux} dF_{W_0}(x)$ exist when u lies in a neighbourhood of the origin, and let $G_0(u) = G_{C_0}(u) G_{W_0}(u)$. Define further R_0 as the *largest non-negative root of the equation* $G_0(u) = 1$. Then, the inequality $\pi_0 \leq e^{-R_0 K}$ holds (Grandell (1991)). The quantity R_0 is the *adjustment coefficient*, and plays a primary role in risk theory, as well as in queueing models. We consider R_0 as our parameter of interest. An estimator of R_0 is the *empirical adjustment coefficient*: $\theta_n = R_n =$ *largest non-negative root of the equation* $G_n(u) = 1$, where $G_n(u) = G_{C_n}(u) G_{W_n}(u)$, and

$$G_{C_n}(u) = \int_0^\infty e^{ux} dF_{C_n}(x) = \frac{1}{n} \sum_{i=1}^n e^{uC_i}, \quad G_{W_n}(u) = \int_0^\infty e^{ux} dF_{W_n}(x) = \frac{1}{n} \sum_{i=1}^n e^{uW_i}. \square$$

Examples 1, 3, 4, 7 are special cases of *statistical functionals*. Suppose that the population parameter of interest is a real or vector parameter, expressed in the form $\theta_0 = \theta(F_0)$. A simple (and natural, as well) idea to construct an estimate of θ_0 is to replace the “true” population d.f. F_0 by the e.d.f. F_n in $\theta(\cdot)$; in symbols: $\hat{\theta}_n = \theta(F_n)$.

As already said, asymptotic statistics mainly deals with consistency, rates of convergence, and limiting distributions. In particular, consistency means that the estimator $\hat{\theta}_n$, as the sample size increases, with “high probability” becomes “very close” to the unknown population parameter. To express this notion in general, we must define a distance on Θ , i.e. a function $d(\cdot, \cdot) : \Theta \times \Theta \rightarrow [0, \infty)$ such that (a) $d(\theta_1, \theta_2) = 0$ if and only if $\theta_1 = \theta_2$, (b) $d(\theta_1, \theta_2) = d(\theta_2, \theta_1)$; (c) $d(\theta_1, \theta_2) \leq d(\theta_1, \theta_3) + d(\theta_2, \theta_3)$. Many of the distances used in asymptotic analysis are induced by norms. A *norm* on (the linear space) Θ is a function $\|\cdot\| : \Theta \rightarrow [0, \infty)$ such that (i) $\|\theta\| = 0$ if and only if θ is null; (ii) $\|c\theta\| = |c| \|\theta\|$; (iii) $\|\theta_1 + \theta_2\| \leq \|\theta_1\| + \|\theta_2\|$. The quantity $d(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$ is then a *distance* on Θ . When (i) ((a)) does not hold, $\|\cdot\|$ ($d(\cdot, \cdot)$) is a pseudo-norm (pseudo-distance).

A sequence $(\hat{\theta}_n; n \geq 1)$ of estimators of θ_0 is (weakly) *consistent* if $d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0$ as n goes to infinity (the symbol $\xrightarrow{p}$ indicates “convergence in probability”).

The *rate of convergence* of a consistent estimator $\hat{\theta}_n$ indicates the speed at which $d(\hat{\theta}_n, \theta_0)$ goes to zero in probability as n increases. Let $t(n)$ be a positive, increasing function of n , tending to infinity as n does. $\hat{\theta}_n$ converges to θ_0 at the rate $t(n)^{-1}$ if $t(n) d(\hat{\theta}_n, \theta_0)$ is *bounded in probability*, i.e. if for every positive ϵ there exists $K_\epsilon > 0$ such that $Pr_{P_0}(t(n) d(\hat{\theta}_n, \theta_0) \leq K_\epsilon) \geq 1 - \epsilon$ for every n . In this case, we will write $d(\hat{\theta}_n, \theta_0) = O_p(1/t(n))$. Consider Ex. 1 and suppose that $k = 1$, so that $\hat{\theta}_n$ and θ_0 reduce to the sample and population mean, respectively. Using the distance $d(\hat{\theta}_n, \theta_0) = |\hat{\theta}_n - \theta_0|$, from Chebyshev’s inequality it follows that $\hat{\theta}_n$ converges to θ_0 at the rate $1/\sqrt{n}$.

In some cases, if $\hat{\theta}_n$ possesses a rate of convergence $t(n)^{-1}$, then the probability law $t(n)(\hat{\theta}_n - \theta_0)$ converges to the probability law of a proper random variable. For instance, consider again Ex. 1 with $k = 1$. From the central limit theorem it follows that the probability law of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $P_0 x^2 - (P_0 x)^2$. This kind of result holds true in more general cases, even when the parameter θ is a function. This is actually the most mathematically difficult part of our programme. Roughly speaking, the estimator $\hat{\theta}_n(\cdot)$ may be viewed as a “rule”, or better as a *map*, that associates an element of Θ to every realization $x_1, \dots, x_n$ of the random sample $X_1, \dots, X_n$. If the elements of Θ are functions (as in Examples 2, 4, 5, 6), then $\hat{\theta}_n$ is a *random function*, i.e. a *stochastic process*.

Modern theory of empirical processes offers very powerful tools to study the consistency, rates of convergence, and asymptotic distributions of estimators. As it will be clear in the sequel, the problems consistency, rates of convergence and the existence of an asymptotic distribution can be attacked essentially by using a unique tool: the *entropy*. The basic advantage consists in reducing the study of asymptotic properties of estimators to a study of asymptotic properties of empirical processes. In the sequel, a few examples motivating the use of the empirical processes theory are given. Next, the foundations of the modern theory of weak convergence and empirical processes are shortly reviewed.

3. Motivations for using empirical processes techniques

As already said, an estimator $\hat{\theta}_n$ of θ_0 is consistent if $d(\hat{\theta}_n, \theta_0)$ converges in probability to zero as the sample size increases. Now, in many cases the distance $d(\hat{\theta}_n, \theta_0)$ can be studied via a quantity of the form $\sup_{g \in \mathcal{G}_n} |P_n g - P_0 g|$, where $\mathcal{G}_n$ is a class of functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$ (we are using the notation introduced in (1)). The class $\mathcal{G}_n$ may depend on n . More precisely, in some cases it is possible to prove an inequality of the form

$$d(\hat{\theta}_n, \theta_0) \leq \text{const} \sup_{g \in \mathcal{G}_n} |P_n g - P_0 g| \quad (5)$$

which of course implies that

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{P} 0 \text{ if } \sup_{g \in \mathcal{G}_n} |P_n g - P_0 g| \xrightarrow{P} 0. \quad (6)$$

In other cases, an inequality of the form (5) does not hold. However, the consistency of $\hat{\theta}_n$ can be still proved via (6).

Example 8 (Population distribution function, Ex. 2 contd.) Consider again Ex. 2, and suppose that the goal is to estimate the whole population d.f $F_0(\cdot)$. The simplest estimator of F_0 is the e.d.f. $F_n(\cdot)$, defined in Ex. 2. Of course, the parameter space is the set $\mathcal{F}$ of all d.f.'s on $\mathcal{X}$. According to the programme described in the previous section, we must define a distance on $\mathcal{F}$. The most widely used distance is the *Kolmogorov metric* $d_K(F_n, F_0) = \sup_{x \in \mathbb{R}^d} |F_n(x) - F_0(x)|$. For every x in $\mathbb{R}^d$, define the function $g_x : \mathbb{R}^d \rightarrow \mathbb{R}$ as $g_x(y) = I_{(-\infty, x]}(y)$, and consider the class $\mathcal{G} = \{g_x(\cdot); x \in \mathbb{R}^d\}$. From $F_n(x) = P_n g_x$, $F_0(x) = P_0 g_x$, we conclude that

$$d_K(F_n, F_0) = \sup_{x \in \mathbb{R}^d} |P_n g_x - P_0 g_x| = \sup_{g \in \mathcal{G}} |P_n g - P_0 g|. \square$$

Example 9 (MLE of a density, Ex. 2 contd.) Suppose that the population X possesses a density (w.r.t. the Lebesgue measure, for the sake of simplicity), and let $\mathcal{F}$ be the class of all possible density functions for X . Denote as usual by f_0 the “true” population density. As a distance on $\mathcal{F}$ take the *Hellinger metric*, defined as:

$$h(f_1, f_2) = \left(\frac{1}{2} \int \left(\sqrt{f_1(x)} - \sqrt{f_2(x)} \right)^2 dx \right)^{1/2}$$

The maximum likelihood estimator (MLE, for short) $\hat{f}_n$ of f_0 is the function $f \in \mathcal{F}$ that maximizes the likelihood function $L(f) = \prod_i f(x_i)$. With obvious symbols:

$$\hat{f}_n = \operatorname{argmax}_{f \in \mathcal{F}} \int \log f(x) dP_n.$$

Let $\bar{f}_n = (\hat{f}_n + f_0)/2$, and let A_0^+ be the set of all x s such that $f_0(x) > 0$. It can be shown (see Birgé (1983), Birgé and Massart (1993), van de Geer (2000)) that the inequalities

$$\begin{aligned} h(\hat{f}_n, f_0)^2 &\leq 16h(\bar{f}_n, f_0)^2 \\ &\leq 8 \left\{ \frac{1}{n} \sum_{i=1}^n \log \frac{\bar{f}_n(x_i)}{2f_0(x_i)} I_{A_0^+}(x_i) - \int \log \frac{\bar{f}_n(x)}{2f_0(x)} I_{A_0^+}(x) dP_0 \right\} \\ &= 8 \left(P_n \log \frac{\bar{f}_n}{2f_0} I_{A_0^+} - P_0 \log \frac{\bar{f}_n}{2f_0} I_{A_0^+} \right) \end{aligned} \quad (7)$$

hold. To obtain an inequality similar to (5), denote by $\mathcal{G}$ the class of functions $\{g(x) = \log\{(f(x) + f_0(x))/2f_0(x)\} I_{A_0^+}(x); f \in \mathcal{F}\}$. From (7) it is immediate to see that

$$h(\hat{f}_n, f_0)^2 \leq 8 \sup_{g \in \mathcal{G}} |P_n g - P_0 g|.$$

If, in addition, the class $\mathcal{F}$ of densities is *convex*, then a result better (and simpler, as well) than (7) can be obtained. In fact, if $\mathcal{F}$ is convex, then the inequality

$$h(\hat{f}_n, f_0)^2 \leq 2 \left(P_n \frac{\hat{f}_n}{\hat{f}_n + f_0} - P_0 \frac{\hat{f}_n}{\hat{f}_n + f_0} \right) \quad (8)$$

holds (see van de Geer (1993)). Hence, by considering the class of functions $\mathcal{G}^c = \{g(x) = f(x)/(f(x) + f_0(x)); f \in \mathcal{F}\}$, from (8) we obtain the inequality

$$h(\hat{f}_n, f_0)^2 \leq 2 \sup_{g \in \mathcal{G}^c} |P_n g - P_0 g|. \square \quad (9)$$

Example 10 (Mixture models) An important application of (9) is to mixture models. We consider here only a very simple case. Let Y be a real r.v. with unknown d.f. $F_0 \in \mathcal{F}$, the set of all d.f.s on the real line. The r.v. Y is not observable. What is observable is a r.v. X whose density function, conditionally on $Y = y$, is $f_X(x|y)$. For the sake of simplicity, assume that $f_X(x|y)$ is known. Define the density $f_{FX}(x) = \int f_X(x|y) dF(y)$, and denote by $\mathcal{F}_X = \{f_{FX}(\cdot); F \in \mathcal{F}\}$ the set of all possible f_{FX} s. When $F = F_0$ the density $f_{0X} = \int f_X(x|y) dF_0(y)$ is the “true” density of the r.v. X . The problem is to estimate F_0 on the basis of a sample $X_1, \dots, X_n$. Our basic assumption is that F_0 is *identifiable*, i.e. that f_{FX} coincides with f_{0X} if and only if F is equal to F_0 . Using the same notation as in Ex. 9, this is equivalent to assume that $h(f_{FX}, f_{0X}) = 0$ if and only if $F \equiv F_0$. Consider then the map $T(\cdot)$ defined as $T(F) = f_{FX}$. If F_0 is identifiable, then the map T is *invertible*, and it is possible to write $F = T^{-1}(f_{FX})$. The form of T , and hence that of its inverse T^{-1} , is in principle *known*. Invertibility of T allows us to produce an estimator of F_0 . The basic idea is to (i) estimate first f_{0X} by its MLE: $\hat{f}_{nX}(\cdot) = \operatorname{argmax}_{f \in \mathcal{F}_X} \sum \log f(x_i)$; (ii) estimate F_0 by inverting T : $\hat{F}_n(\cdot) = T^{-1}(\hat{f}_{nX}(\cdot))$.

To make the problem of estimating F_0 precise, we must adopt a distance on $\mathcal{F}$. The most convenient distance for our problem is the *Levy distance* (see Schweizer and Sklar (1983)). Given two d.f.s F_1, F_2 on the real line, their Levy distance $d_L(F_1, F_2)$ is equal to $\inf\{\epsilon > 0 : F_1(x - \epsilon) - \epsilon \leq F_2(x) \leq F_1(x + \epsilon) + \epsilon \forall x \in \mathbb{R}\}$. As shown in Pfanzagl (1988), identifiability of F_0 implies that $d_L(\hat{F}_n, F_0)$ goes to zero in probability as n increases, whenever $h(\hat{f}_{nX}, f_{0X})$ does. Hence, to prove that $\hat{F}_n$ is a consistent estimator of F_0 , it is enough to show that $\hat{f}_{nX}$ is a consistent estimator of f_{0X} , and this essentially depends on inequality (9). $\square$

Example 11 (Nonparametric regression, Ex. 6 contd.) Consider the nonparametric regression model introduced in Ex. 6, and let $\mathcal{G}$ be the parameter space, i.e. the set of all regression functions of the assumed statistical model. For instance, in parametric regression $\mathcal{G}$ takes the form $\{\beta_1 \phi_1(x) + \dots + \beta_k \phi_k(x)\}$, where $\phi_j(x)$ s are known functions, and β_j s are unknown parameters. The *least squares estimator* of g_0 is:

$$\hat{g}_n = \operatorname{argmin}_{g \in \mathcal{G}} \sum_{i=1}^n (Y_i - g(x_i))^2 \quad (10)$$

Of course, when $\mathcal{G} = \{\beta_1\phi_1(x) + \dots + \beta_k\phi_k(x)\}$, the estimator (10) takes the form $\hat{\beta}_1\phi_1(x) + \dots + \hat{\beta}_k\phi_k(x)$, where $\hat{\beta}_j$ s are the least squares estimators of β_j s.

The r.v.s Y_i s, in this case, are *not i.i.d.*. Let $\mathbf{y}$ be the vector $(y_1 \dots y_n)$, and let $\mathbf{g}$ be the vector $(g(x_1) \dots g(x_n))$. Moreover, define the scalar product $\langle \mathbf{y}, \mathbf{g} \rangle_n = (y_1g(x_1) + \dots + y_ng(x_n))/n$, and the norm $\|\mathbf{y}\|_n = \sqrt{(y_1^2 + \dots + y_n^2)/n}$. With this notation we have $\hat{\mathbf{g}}_n = \operatorname{argmin}_{\mathbf{g} \in \mathcal{G}} \|\mathbf{y} - \mathbf{g}\|_n$. The parameter space $\mathcal{G}$ is equipped with the distance $d_n(g_1, g_2) = \|\mathbf{g}_1 - \mathbf{g}_2\|_n$.

In this example, unless to impose very stringent conditions, we cannot obtain an inequality of the form (5). However, the argument in (6) still works. Let K be a positive real, and consider (with obvious notation) the class of functions $\mathcal{G}_n(K) = \{g \in \mathcal{G} : \|\mathbf{g} - \mathbf{g}_0\|_n \leq K\}$. As shown is van de Geer and Wegkamp (1997), the inequality

$$\begin{aligned} Pr_{P_0} (\|\hat{\mathbf{g}}_n - \mathbf{g}_0\|_n > \epsilon) &\leq Pr_{P_0} \left(\sup_{g \in \mathcal{G}_n(K)} \frac{1}{n} \sum_{i=1}^n (U_i I_{(|U_i| \leq M)} (g(x_i) - g_0(x_i))) \geq \frac{\epsilon^2}{4} \right) \\ &+ 2 Pr_{P_0} \left(\frac{1}{n} \sum_{i=1}^n U_i^2 I_{(|U_i| \leq M)} \geq \frac{\epsilon^4}{16K^2} \right) \end{aligned} \quad (11)$$

holds, for every positive K, M, ϵ . Under simple regularity conditions on U_i s, by letting M large and K small the last term in (11) can be made arbitrarily small. Hence, the consistency of $\hat{\mathbf{g}}_n$ essentially depends on whether the quantity $\sup_{g \in \mathcal{G}_n(K)} \sum_i U_i I_{(|U_i| \leq M)} (g(x_i) - g_0(x_i))/n$ converges in probability to zero for small values of K . $\square$

The examples above show two important facts. (i) In many nonparametric problems the consistency of estimators depends on the asymptotic behaviour of quantities of the form $\sup_{g \in \mathcal{G}_n} |P_n g - P g|$. (ii) The speed of convergence of an estimator $\hat{\theta}_n$ to the “true parameter” θ_0 depends on the speed of convergence (to zero) of $\sup_{g \in \mathcal{G}_n} |P_n g - P g|$. Both these points, roughly speaking, are related to the “dimension” of the class $\mathcal{G}$. If $\mathcal{G}_n$ is “too large”, then $\sup_{g \in \mathcal{G}_n} |P_n g - P g|$ could not tend to zero in probability. Furthermore, the smaller $\mathcal{G}_n$, the fastest $\sup_{g \in \mathcal{G}_n} |P_n g - P g|$ tend to zero (in probability), and hence the fastest $\hat{\theta}_n$ converges to θ_0 .

As already said in Section 2, an important problem is to approximate (if possible) the probability distribution of an estimator $\hat{\theta}_n$. More formally, the problem consists in showing that, for some sequence $(t(n); n \geq 1)$, the probability law of $t(n)(\hat{\theta}_n - \theta_0)$ converges to the probability law of a proper random variable, that should be (at least approximately) computed. We stress again that when $\hat{\theta}_n$ takes values in a functional space, $t(n)(\hat{\theta}_n - \theta_0)$ is a *random function* (a *stochastic process*), and the study of its limiting distribution requires some basic tools of the theory of weak convergence.

Example 12 (Elementary empirical process and Kolmogorov statistic, Ex. 8 contd.)

Consider again Ex. 8, and suppose that X_i s are real r.v.s (univariate data). Let $W_n(x) = \sqrt{n}(F_n(x) - F_0(x))$, $x \in \mathbb{R}$. For every fixed x $W_n(x)$ is a r.v., such that $nF_n(x) = n(W_n(x)/\sqrt{n} + F_0(x))$ possesses binomial distribution. The *random function* $(W_n(x); x \in \mathbb{R})$ is the simplest example of empirical process, and will be termed “elementary empirical process”. When X_i s are uniformly distributed on $(0, 1)$, the elementary empirical process is called “uniform empirical process”.

The “classical” *Glivenko-Cantelli theorem* establishes the uniform a.s. convergence on $F_n(\cdot)$ to $F_0(\cdot)$: $\sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \xrightarrow{a.s.} 0$ as $n \rightarrow \infty$. A stronger result holds when one considers the asymptotic distribution of the random function $W_n(\cdot) = (W_n(x); x \in \mathbb{R})$.

$\mathbb{R}$). For the sake of simplicity, assume that the population d.f. F_0 is continuous. As a consequence of the central limit theorem, for every fixed x the r.v. $W_n(x)$ converges in distribution to a normal r.v. with mean zero and variance $F_0(x)(1 - F_0(x))$. As a simple extension of this result, it is not difficult to see that for every $m \geq 1$ and for every $x_1, \dots, x_m$, the m -dimensional r.v. $\sqrt{n}(W_n(x_1), \dots, W_n(x_m))$ converges in distribution, as n goes to infinity, to a m -variate normal r.v. $(Y_1, \dots, Y_m)$, with mean vector zero and covariance matrix $E[Y_j Y_h] = F_0(\min(x_j, x_h)) - F_0(x_j)F_0(x_h)$. This result suggests that the probability law of the whole random function $W_n(\cdot)$ should converge, in some sense, to the probability law of a random function $G(\cdot)$ such that, for every $m \geq 1$ and for every $x_1, \dots, x_m$, the r.v. $(G(x_1), \dots, G(x_m))$ possesses a m -variate normal distribution with $E[G(x_j)] = 0$ and $E[G(x_j)G(x_h)] = F_0(\min(x_j, x_h)) - F_0(x_j)F_0(x_h)$. Clearly, $G(\cdot)$ defines a *Gaussian stochastic process*. Consider now a Gaussian stochastic process $B(\cdot) = (B(t); 0 < t < 1)$, with $E[B(t)] = 0$, $E[B(t)B(u)] = \min(t, u) - tu$. The process $B(\cdot)$ is a *Brownian bridge*. Among its properties, we mention the (a.s.) continuity of its trajectories: with probability one, realizations of $B(\cdot)$ are continuous functions, with $B(0) = B(1) = 0$. Now, it is not difficult to see that the probability law of $(G(x); x \in \mathbb{R})$ coincides with the probability law of $(B(F_0(x)); x \in \mathbb{R})$. In other terms, it is possible to write $G(\cdot) \stackrel{d}{=} B(F_0(\cdot))$, where $\stackrel{d}{=}$ denotes “equality in distribution”. Hence, the probability law of the elementary empirical process $W_n(\cdot)$ converges, in some sense, to the probability law of the Gaussian stochastic process $B(F_0(\cdot))$. As a consequence, the probability of the Kolmogorov statistic $\sqrt{n} \sup_x |F_n(x) - F_0(x)|$ should converge, as n increases, to the probability law of $\sup_x |B(F_0(x))|$. Since (by a continuity argument) $\sup_x |B(F_0(x))|$ is a.s. equal to $\sup_{0 < t < 1} |B(t)|$, we conclude that:

$$\sqrt{n} \sup_{x \in \mathbb{R}} |W_n(x)| = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \xrightarrow{d} \sup_{0 < t < 1} |B(t)| \text{ as } n \rightarrow \infty. \quad (12)$$

Taking into account that $P(\sup_t |B(t)| \leq x) = 1 - 2 \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 x^2}$ (Billingsley (1999)), relationship (12) gives the asymptotic distribution of the Kolmogorov statistic. $\square$

The heuristic reasoning of Ex. 12 was first used by Doob (1949), with an evident lack of rigorousness. Its main weak-points are two.

(i) What is the exact meaning of the sentence “the probability law of the empirical process $W_n(\cdot)$ converges to the probability law of the Gaussian stochastic process $B(F(\cdot))$ ”? Elementary probability theory deals with convergence in distribution of real-valued or vector-valued r.v.s. The objects involved here are considerably more complicated, because their realizations are not real numbers or vectors, but *functions*. Hence, we need an extension of the elementary theory of convergence in distribution.

(ii) Why should the probability law of the functional $\theta(W_n) = \sup_x |W_n(x)|$ converge in distribution to $\sup_t |B(t)|$? Elementary theory of probability tells us that if $(Y_n; n \geq 1)$ and Y are (real-valued or vector valued) r.v.s such that $Y_n \xrightarrow{d} Y$ as n tends to infinity, and if $g(\cdot)$ is a continuous function, then $g(Y_n) \xrightarrow{d} g(Y)$ as n increases. In other words, convergence in distribution is preserved under continuous transformations. We need an extension of this argument to stochastic processes.

The first rigorous answer to these questions was given by Donsker (1952), who clarified the meaning of the convergence of the probability law of the empirical process $\sqrt{n}(F_n(\cdot) - F(\cdot))$; such a kind of convergence was termed since then “weak convergence”. This originates a series of important studies, that settled the theory of weak convergence for Borel-measurable random functions. An excellent exposition of this theory (with some extensions to non-Borel-measurable random functions) is in Billingsley

(1999). Statistical applications motivated further advances to the whole theory of weak convergence. We present in the sequel a very short account of the weak convergence theory due to Hoffmann-Jørgensen (1991) (see also van der Vaart and Wellner (1996)).

4. Some basic theory of weak convergence

In this section we provide a short sketch of the theory of weak convergence due to Hoffmann-Jørgensen (1991). Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let (S, d) be a metric space. The *Borel σ -field* $\mathcal{B}(S)$ on (S, d) is the smallest σ -field containing all the open sets w.r.t. the topology induced by the distance d . Of course, when S is the k -dimensional Euclidean space (i.e. $\mathbb{R}^k$ equipped by the Euclidean metric), $\mathcal{B}(S)$ reduces to the “usual” Borel σ -field. A map $Y : \Omega \rightarrow S$ is *Borel-measurable* if the inverse image $Y^{-1}(B) = \{\omega : Y(\omega) \in B\}$ is in $\mathcal{A}$ for every B in the Borel σ -field $\mathcal{B}(S)$. A Borel-measurable map is usually referred to as a *random element*. The very first developments of the theory of weak convergence of stochastic processes were based on the assumption of Borel-measurability. The “prototype” is essentially the elementary empirical process $W_n(\cdot) = \sqrt{n}(F_n(\cdot) - F_0(\cdot))$, that can be viewed as a random function taking values in the set $D[-\infty, +\infty]$ of right-continuous functions with left-hand limits. The set $D[-\infty, +\infty]$ can be equipped by an appropriate metric (the Skorokhod metric) that makes $W_n(\cdot)$ Borel-measurable. However, in many cases of interest in statistics, one has to consider random functions that are *not* Borel-measurable. This has forced modern developments of the theory of weak convergence of random functions.

Let $(Y_n; n \geq 1)$ be a sequence of random functions, i.e. a sequence of maps $Y_n : \Omega_n \rightarrow S$, where (S, d) is an arbitrary metric space. Although the notion of weak convergence is just the infinite-dimensional counterpart of the usual convergence in distribution of random variables or vectors, its definition is considerably more involved. Consider a random function Y taking values in S , and let $g : S \rightarrow \mathbb{R}$ be a real function. Unless Y is a random element, $g(Y)$ is not a random variable, and the expected value $E[g(Y)]$ does not make any sense. To overcome this inconvenient, we take its *outer expectation*:

$$E^*[g(Y)] = \inf\{E[U] : U : \Omega \rightarrow \mathbb{R} \text{ is measurable, } U \geq g(Y), \text{ and } E[U] \text{ exists}\}.$$

In particular, when $g(x) = I_A(x)$ is the indicator function of a set A , then the outer expectation reduces to the *outer probability* of A .

A sequence $(Y_n; n \geq 1)$ of random functions *converges weakly* to a random element Y if $\lim_{n \rightarrow \infty} E^*[f(Y_n)] = E[f(Y)]$ for every continuous and bounded function $f : S \rightarrow \mathbb{R}$. This definition *does not* require that Y_n s are random elements, but only that Y is Borel-measurable. Weak convergence possesses nice mathematical properties. Among them, we mention here the *continuous mapping theorem*: if $f : S \rightarrow E$ is a continuous function from the metric space (S, d) to the metric space (E, ρ) , and if Y_n converges weakly to Y , then $f(Y_n)$ converges weakly to $f(Y)$.

Generally speaking, it is a very difficult task to verify whether a given sequence of random functions converges weakly to a random element. However, the effort required reduces considerably in the simple but important case (at least, very important for statistical applications) of empirical processes. Before giving general results, let us consider the simple case of the uniform empirical process $(E_n(x) = \sqrt{n}(F_n(x) - x); 0 \leq x \leq 1)$, arising when the sample observations X_i s are uniformly distributed on $(0, 1)$ (cfr. Ex. 12). We have already said that $E_n(\cdot)$ converges weakly to the Brownian bridge $B(\cdot)$. This assertion is equivalent to the following two:

- (i) the m -variate r.v. $(E_n(x_1), \dots, E_n(x_m))$ converges in distribution to $(B(x_1), \dots, B(x_m))$, for every $m \geq 1$ and $x_1, \dots, x_m$ in $[0, 1]$;
(ii) for every positive ϵ, δ , there exist a partition of $[0, 1]$ into finitely many intervals $I_1, \dots, I_k$ such that $\limsup_{n \rightarrow \infty} P^* \left(\sup_{1 \leq j \leq k} \sup_{x, y \in I_j} |E_n(x) - E_n(y)| \geq \epsilon \right) \leq \delta$.

This criterion reduces the verification of the weak convergence of the *random function* E_n to a simpler problem, involving only a *finite* number of real-valued random variables. It can be extended to many random functions arising in asymptotic statistics. Let T be an arbitrary set, and let $l^\infty(T)$ be the set of all *bounded*, real-valued functions $f : T \rightarrow \mathbb{R}$. Consider further a sequence $(Y_n; n \geq 1)$ of random functions, and a random element Y , taking values in $l^\infty(T)$. In the case of the uniform empirical process, $T = [0, 1]$. Clearly, the random functions Y_n s, Y , can be written in the form $(Y_n(t); t \in T)$, $(Y(t); t \in T)$. As shown in van der Vaart and Wellner (1996), Y_n converges weakly to Y if and only if:
(i) the m -variate r.v. $(Y_n(t_1), \dots, Y_n(t_m))$ tends in distribution to $(Y(t_1), \dots, Y(t_m))$, for every $m \geq 1$ and $t_1, \dots, t_m$ in T ;
(ii) for every positive ϵ, δ , there exist a partition of T into finitely many sets $T_1, \dots, T_k$ such that $\limsup_{n \rightarrow \infty} P^* \left(\sup_{1 \leq j \leq k} \sup_{u, t \in T_j} |Y_n(t) - Y_n(u)| \geq \epsilon \right) \leq \delta$.

Weak convergence of random elements in $l^\infty(T)$ can be also characterized as follows. Let $BL(l^\infty(T))$ be the set of all functions $h : l^\infty(T) \rightarrow [-1, 1]$ that are *uniformly Lipschitz*: $|h(u_1) - h(u_2)| \leq c \sup_t |u_1(t) - u_2(t)|$. Then, $(Y_n(t); t \in T)$ converges weakly to $(Y(t); t \in T)$ if and only if (van der Vaart and Wellner (1996))

$$\sup_{h \in BL(l^\infty(T))} |E^*[h(Y_n)] - E[h(Y)]| \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (13)$$

5. Empirical processes

5.1 Basic problems

The results established so far are of considerable utility when applied to empirical processes. The general setting is the same as in Sections 2, 3. Let $(X_n; n \geq 1)$ be a sequence of r.v.s, defined on a probability space $(\Omega, \mathcal{A}, P)$, and let $F_0(x) = Pr_{P_0}(X_i \leq x)$ be the population d.f.. Denote by $\mathcal{X}$ the set of values that each X_i may take. Of course, the most common case is $\mathcal{X} = \mathbb{R}^d$ (d -variate observations). Consider further a function $g : \mathcal{X} \rightarrow \mathbb{R}$, lying in a class $\mathcal{G}$ of functions. Using the same notation of Section 2, let P_n be the empirical measure corresponding to a sample $X_1, \dots, X_n$ of size n , and let $P_n g = \sum g(X_i)/n$, $P_0 g = \int g(x) dP_0$. The *empirical process* indexed by g is defined as:

$$G_n(\cdot) = (G_n(g); g \in \mathcal{G}) = (\sqrt{n}(P_n g - P_0 g); g \in \mathcal{G}).$$

For every *fixed* g in the class $\mathcal{G}$, the quantity $G_n(g)$ is a real-valued random variable. Hence, the whole *family* $(G_n(g); g \in \mathcal{G})$ is a stochastic process.

As shown in the examples of Section 3, to study the large sample properties of an estimator $\hat{\theta}_n$ we have to study first the large sample behaviour of the empirical process $G_n(\cdot)$. Since $(g(X_i); i \geq 1)$ is a sequence of *i.i.d.* r.v.s, the limiting behaviour of $G_n(g)$ for a fixed g can be easily studied. Assuming that $E_{P_0}|g(X)|$ is finite, from the strong law of large numbers it follows that

$$P_n g = \frac{1}{n} \sum_{i=1}^n g(X_i) \xrightarrow{a.s.} E_{P_0}[g(X)] = P_0 g \text{ as } n \rightarrow \infty \quad (14)$$

and hence $|P_n g - P_0 g| \xrightarrow{a.s.} 0$ as n increases. Assuming further that the r.v.s $g(X_i)$ s possess a finite variance (i.e. $E_{P_0}[g(X_i)^2] = P_0 g^2 < \infty$), from the central limit theorem for sums of i.i.d. r.v.s, it follows that $G_n(g)$ converges in distribution to a normal r.v. $G(g)$ with mean zero and variance $E_{P_0}[g(X_i)^2] - E_{P_0}[g(X_i)]^2 = P_0 g^2 - (P_0 g)^2$. In symbols:

$$G_n(g) \xrightarrow{d} G(g) \sim N(0, P_0(g^2) - (P_0 g)^2) \text{ as } n \rightarrow \infty. \quad (15)$$

As an easy generalization of (15), it is not difficult to prove that, for every positive integer m and for every $g_1, \dots, g_m$ in $\mathcal{G}$, the m -variate r.v. $(G_n(g_1), \dots, G_n(g_m))$ converges in distribution, as n increases, to a m -variate normal distribution $(G(g_1), \dots, G(g_m))$ with $E_{P_0}[G(g_j)] = 0$ and $E_{P_0}[G(g_j)G(g_h)] = E_{P_0}[g_j(X_i)g_h(X_i)] - E_{P_0}[g_j(X_i)]E_{P_0}[g_h(X_i)] = P_0(g_j g_h) - (P_0 g_j)(P_0 g_h)$. As in the case of the elementary empirical process, the family of r.v.s $(G(g); g \in \mathcal{G})$ defines a random function taking values in $l^\infty(\mathcal{G})$, i.e. a *stochastic process* indexed by g . Unfortunately, (14) and (15) are not enough to characterize the asymptotic behaviour of the stochastic process $(G_n(g); g \in \mathcal{G})$. In the sequel, we will answer to two basic questions.

- (i) Does (14) hold *uniformly* in the class $\mathcal{G}$? In other words, does the quantity $\sup_{g \in \mathcal{G}} |P_n g - P_0 g|$ tend a.s. to zero as n increases?
- (ii) Does the probability law of the stochastic process $(G_n(g); g \in \mathcal{G})$ converge to the law of a stochastic process $(G(g); g \in \mathcal{G})$, as n goes to infinity? More formally, does $(G_n(g); g \in \mathcal{G})$ converge weakly to $(G(g); g \in \mathcal{G})$ as n increases?

5.2 Entropy and Vapnik-Cervonenkis dimension

The answers to questions (i), (ii), as well as the evaluation of the rate of convergence of estimators, are largely based on a fundamental mathematical tool: the *metric entropy* of the class $\mathcal{G}$. Roughly speaking, the entropy of $\mathcal{G}$ provides a measure of the “width” of $\mathcal{G}$. Given a class $\mathcal{G}$ of functions $g : \mathcal{X} \rightarrow \mathbb{R}$, let Q be a probability measure on $\mathcal{X}$, and let m be a positive integer. Define the (semi) norm

$$\|g\|_{Q,m} = \{Qg^m\}^{1/m} = \sqrt[m]{\int g(x)^m dQ}. \quad (16)$$

Clearly, $\|g_1 - g_2\|_{Q,m}$ is a (semi) distance between g_1 and g_2 . An *open ball* of center g_0 and radius $\epsilon > 0$ is the set of functions $B_{Q,m}(g_0, \epsilon)$ of all functions in $\mathcal{G}$ having distance from g_0 smallest than ϵ : $B_{Q,m}(g, \epsilon) = \{g \in \mathcal{G} : \|g - g_0\|_{Q,m} < \epsilon\}$. The ϵ -covering number $\mathcal{N}_m(\epsilon, \mathcal{G}, Q)$ is the smallest number of balls of radius ϵ covering $\mathcal{G}$. In symbols:

$$\mathcal{N}_m(\epsilon, \mathcal{G}, Q) = \text{smallest } k \text{ for which } \exists g_1, \dots, g_k \text{ such that } \mathcal{G} \subseteq \bigcup_{j=1}^k B_{Q,m}(g_j, \epsilon).$$

The ϵ -entropy of $\mathcal{G}$ is the logarithm of its ϵ -covering number. In symbols: $\mathcal{H}_m(\epsilon, \mathcal{G}, Q) = \log \mathcal{N}_m(\epsilon, \mathcal{G}, Q)$. In particular, when the measure Q is the *empirical measure* P_n the entropy is termed *random entropy*, and it is denoted by $\mathcal{H}_m(\epsilon, \mathcal{G}, P_n)$.

The computation of the entropy of a class $\mathcal{G}$ of functions could be, in principle, complicated. A simplification can be obtained by using the so-called “entropy with bracketing”. The basic idea of covering numbers is to make a partition of $\mathcal{G}$ by balls of radius ϵ covering $\mathcal{G}$. Bracketing is based on the same idea, but it does not use balls. Given a positive real ϵ , an ϵ -bracket of $\mathcal{G}$ is a set of N pairs of functions, $\{(g_1^L, g_1^U), \dots, (g_N^L, g_N^U)\}$ such

that: (a) $\|g_j^L - g_j^U\|_{Q,m} < \epsilon$ for every $j = 1, \dots, N$; (b) for every $g \in \mathcal{G}$ there exists an index $1 \leq j \leq N$ such that $g_j^L \leq g \leq g_j^U$. The ϵ -bracketing number of $\mathcal{G}$ is the smallest N such that there exists an ϵ -bracket $\{(g_1^L, g_1^U), \dots, (g_N^L, g_N^U)\}$. In symbols:

$$\mathcal{N}_{Bm}(\epsilon, \mathcal{G}, Q) = \inf \{N \text{ such that } \exists \text{ a } \epsilon\text{-bracket } \{(g_1^L, g_1^U), \dots, (g_N^L, g_N^U)\} \text{ of } \mathcal{G}\}.$$

The ϵ -entropy with bracketing of $\mathcal{G}$ is the logarithm of its ϵ -bracketing number. In symbols: $\mathcal{H}_{Bm}(\epsilon, \mathcal{G}, Q) = \log \mathcal{N}_{Bm}(\epsilon, \mathcal{G}, Q)$. Bracketing numbers are easier to compute than covering numbers. Furthermore, they are bigger than covering numbers, since for every positive ϵ the inequality $\mathcal{H}_m(\epsilon, \mathcal{G}, Q) \leq \mathcal{H}_{Bm}(\epsilon, \mathcal{G}, Q)$ holds.

The entropy of several classes of functions interesting for statistical applications is computed in van der Vaart and Wellner (1996), Vapnik (1998), and in references therein. In many cases, entropy can be computed by elegant combinatorial arguments, that form the so-called *Vapnik-Chervonenkis theory*; cfr. Vapnik and Chervonenkis (1971), Vapnik (1998). Let $x_1, \dots, x_n$ be n points of $\mathcal{X}$, let $\mathcal{C}$ be a class of subsets of $\mathcal{X}$, and let $\mathcal{S}_{\mathcal{C}}(x_1, \dots, x_n)$ be the class of all subsets of the form $\{x_1, \dots, x_n\} \cap C$, with C in $\mathcal{C}$:

$$\mathcal{S}_{\mathcal{C}}(x_1, \dots, x_n) = \{\{x_1, \dots, x_n\} \cap C; C \in \mathcal{C}\}.$$

Finally, denote by $S_{\mathcal{C}}(x_1, \dots, x_n)$ the number of sets in $\mathcal{S}_{\mathcal{C}}(x_1, \dots, x_n)$. Clearly, $1 \leq S_{\mathcal{C}}(x_1, \dots, x_n) \leq 2^n$. The largest value of $S_{\mathcal{C}}(x_1, \dots, x_n)$:

$$S_{\mathcal{C}}(n) = \max_{x_1, \dots, x_n} S_{\mathcal{C}}(x_1, \dots, x_n)$$

is the *shatter coefficient* of $\mathcal{C}$. The *Vapnik-Chervonenkis dimension* (VC-dimension, for short) $V_{\mathcal{C}}$ of $\mathcal{C}$ is the largest n for which the equality $S_{\mathcal{C}}(n) = 2^n$ holds. In symbols: $V_{\mathcal{C}} = \sup\{n : S_{\mathcal{C}}(n) = 2^n\}$. If $S_{\mathcal{C}}(n) = 2^n$ for every integer n , then we set $V_{\mathcal{C}} = \infty$. A celebrated inequality (Sauer's lemma) establishes that the relationships

$$S_{\mathcal{C}}(n) \leq \sum_{k=0}^{V_{\mathcal{C}}} \binom{n}{k} \leq (n+1)^{V_{\mathcal{C}}}.$$

hold. Hence, $S_{\mathcal{C}}(n)$ is either identically equal to 2^n , or bounded by a polynomial in n .

Let now $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \rightarrow [a, b]$, denote by SG_g its *subgraph*: $SG_g = \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : t \leq g(x)\}$, and consider the class of sets $\mathcal{G}^+ = \{SG_g; g \in \mathcal{G}\}$. The following inequality, due to Haussler (1992), holds:

$$\mathcal{N}_m(\epsilon, \mathcal{G}, Q) \leq e(V_{\mathcal{G}^+} + 1) \left(\frac{2e(b-a)}{\epsilon} \right)^{mV_{\mathcal{G}^+}} \quad \forall \epsilon > 0, m \geq 1. \quad (17)$$

Inequality (17) shows the importance of VC-dimension in bounding the entropy of classes of functions, since $\mathcal{H}_1(\epsilon, \mathcal{G}, Q)$ (and in particular $\mathcal{H}_1(\epsilon, \mathcal{G}, P_n)$) is finite if $V_{\mathcal{G}^+} < \infty$. For instance, if $\mathcal{G} = \{\theta_1 \phi_1(\cdot) + \dots + \theta_k \phi_k(\cdot)\}$, with $\phi_j(\cdot)$ s known real-valued functions and θ_j s unknown real parameters, then $V_{\mathcal{G}^+} \leq k+1$ (van der Vaart and Wellner (1996)).

5.3 Glivenko-Cantelli and Donsker theorems

A class $\mathcal{G}$ of functions is a *Glivenko-Cantelli class* if the *uniform law of large numbers* holds: $\sup_{g \in \mathcal{G}} |P_n g - P_0 g| \xrightarrow{a.s.} 0$ as n goes to infinity. Entropy allows one to establish regularity conditions under which a class $\mathcal{G}$ is Glivenko-Cantelli. A function $\bar{g} : \mathcal{X} \rightarrow \mathbb{R}$ is an *envelope function* for $\mathcal{G}$ if $|g(x)| \leq \bar{g}(x)$ for every $g \in \mathcal{G}$ and $x \in \mathcal{X}$.

Theorem 1 (Glivenko-Cantelli) Suppose that the envelope $\bar{g}$ of $\mathcal{G}$ possesses finite mean value $P_0\bar{g} = E_{P_0}[\bar{g}(X)] < \infty$, and that one of the following assumptions is met.

- (i) $\mathcal{H}_{B1}(\epsilon, \mathcal{G}, P_0) < \infty$ for every positive ϵ .
 - (ii) $\mathcal{H}_1(\epsilon, \mathcal{G}, P_n)/n$ tends to zero in (outer) probability, as $n \rightarrow \infty$, for every positive ϵ .
- Then, the uniform law of large numbers holds:

$$\sup_{g \in \mathcal{G}} |P_n g - P_0 g| = \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(X_i) - E_{P_0}[g(X)] \right| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \square$$

A proof of Theorem 1 is in van der Vaart and Wellner (1996). Condition (ii) is also necessary for the uniform law of large numbers; see Talagrand (1996). A “uniform” Glivenko-Cantelli theorem, involving $\sup_{P_0 \in \mathcal{P}} \sup_{g \in \mathcal{G}} |P_n g - P_0 g|$, is studied in Alon *et al.* (1997).

In view of (17), condition (ii) is fulfilled whenever $V_{\mathcal{G}^+}$ is finite. In some cases of statistical interest (e.g. nonparametric regression) condition (i) is too restrictive. On the other hand, (ii) is less stringent, and more suitable for statistical applications.

A class $\mathcal{G}$ of functions is a *Donsker class* if, when n increases, $(G_n(g); g \in \mathcal{G})$ converges weakly a Gaussian stochastic process $(G(g); g \in \mathcal{G})$ with mean function $E_{P_0}[G(g)] = 0$ and covariance kernel $E_{P_0}[G(g_1)G(g_2)] = P_0 g_1 g_2 - (P_0 g_1)(P_0 g_2)$. Conditions on entropy are also the basic ingredient to establish whether $\mathcal{G}$ is a Donsker class, i.e. to prove the weak convergence of the empirical process $(G_n(g); g \in \mathcal{G})$. Regularity conditions are essentially two. First of all, the class $\mathcal{G}$ should possess a squared integrable envelope function. Secondly, it is apparent from the definition of entropy that $\mathcal{H}_m(\epsilon, \mathcal{G}, Q)$ goes to infinity as ϵ tends to zero. In order to have weak convergence to a limiting process, entropy should not increase “too fast”. More precisely, consider again the envelope function $\bar{g}(\cdot)$, and take its norm $\|\bar{g}\|_{Q,m}$, as defined in (16). The *uniform covering numbers* are defined as

$$\bar{\mathcal{N}}_m(\epsilon, \mathcal{G}) = \sup_Q \mathcal{N}_m(\epsilon \|\bar{g}\|_{Q,m}, \mathcal{G}, Q)$$

where the supremum is taken over all probability measures Q for which $\|\bar{g}\|_{Q,m} > 0$. The *uniform entropy integral* and the *bracketing integral* are defined as

$$\mathcal{J}_m(\delta, \mathcal{G}) = \int_0^\delta \sqrt{\log \bar{\mathcal{N}}_m(\epsilon, \mathcal{G})} d\epsilon, \quad \mathcal{J}_{Bm}(\delta, \mathcal{G}, P_0) = \int_0^\delta \sqrt{\log \mathcal{N}_{Bm}(\epsilon, \mathcal{G}, P_0)} d\epsilon \quad (18)$$

respectively. The following result holds (a proof is in van der Vaart and Wellner (1996)).

Theorem 2 (Donsker) Suppose that the envelope $\bar{g}$ of $\mathcal{G}$ possesses finite second moment $P_0\bar{g}^2 = E_{P_0}[\bar{g}(X)^2] < \infty$, and that one of the following assumptions is met.

- (i) The uniform entropy $\mathcal{J}_2(\delta, \mathcal{G})$ is finite for some $\delta > 0$.
- (ii) The bracketing integral $\mathcal{J}_{B2}(\delta, \mathcal{G}, P_0)$ is finite for some $\delta > 0$.
- (iii) There is a decreasing function H such that $\int_0^\delta \sqrt{H(t)} dt < \infty$ for some $\delta > 0$, and

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left(\sup_{0 < \epsilon < \delta} \frac{\mathcal{H}_2(\epsilon, \mathcal{G}, P_n)}{H(\epsilon)} > M \right) = 0.$$

Then, $((G_n(g); g \in \mathcal{G}); n \geq 1)$ converges weakly to a Gaussian process $(G(g); g \in \mathcal{G})$ with $E_{P_0}[G(g)] = 0$, $E_{P_0}[G(g_1)G(g_2)] = P_0(g_1 g_2) - (P_0 g_1)(P_0 g_2)$. $\square$

The main difficulty in applying condition (i) is that the computation of the uniform entropy could be very difficult. However, as a consequence of (17), if $\mathcal{G}$ possesses finite VC-dimension then (i) is fulfilled. The computation of the bracketing integral does only require the computation of an expectation w.r.t. P_0 , the population probability distribution. Hence, condition (ii) is apparently simpler than (i). The price to pay is a loss of generality, since the bracketing integral is larger than the uniform entropy.

5.4 Examples

Example 13 (Ex. 8 contd.) Consider again Ex. 8 and suppose, for the sake of simplicity, that $d = 1$ (unidimensional data). If $x < z$, then it is easy to see that $\|I_{(-\infty, x]}(\cdot) - I_{(-\infty, z]}(\cdot)\|_{P_n, 1} = (\#x < x_i \leq z)/n$. Hence, taking $g_i(y) = I_{(\infty, x_i]}(y)$, $i = 1, \dots, n$, we have $\mathcal{N}_1(\epsilon, \mathcal{G}, P_n) \leq 1/n$ for every $\epsilon > 1/n$. This result clearly shows that $\mathcal{H}_1(\epsilon, \mathcal{G}, P_n)/n$ tends to zero in probability as n goes to infinity, for every fixed positive ϵ . Hence, we have proved the classical *Glivenko-Cantelli theorem*:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \square$$

Examples 9, 10 offer other simple cases where the theory developed in Subsection 5.3 applies. To prove the consistency of the involved estimators, it is enough to show that $\mathcal{H}_1(\epsilon, \mathcal{G}, P_n)/n$ tends to zero in (outer) probability as the sample size increases.

Example 14 (Isotonic regression) Consider Ex. 11, and denote by Q_n the probability measure giving mass $1/n$ to each x_i , $i = 1, \dots, n$. From (11) and a result by van de Geer and Wegkamp (1997), $Pr_{P_0}(\|\hat{g}_n - g_0\|_n > \epsilon)$ tends to zero as n increases provided that

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E_{P_0} [U_i^2 I_{|U_i| > M}] = 0, \quad \frac{1}{n} \mathcal{H}_1(\epsilon, \mathcal{G}_n(K), Q_n) \xrightarrow{P} 0 \text{ as } n \rightarrow \infty$$

for every positive ϵ and K . In particular, suppose that $\mathcal{G}$ is the class of all real valued, increasing function. This is the case of *isotonic regression*. If the functions in $\mathcal{G}$ are uniformly bounded, i.e. if there exists $R > 0$ such that $\sup_x |g(x)| \leq R$ for every g in $\mathcal{G}$, then, using a result by Birman and Solomjak (1967), it can be shown that $\mathcal{H}_1(\epsilon, \mathcal{G}_n(K), Q_n) \leq C(\sqrt{n}K + R)/\epsilon$, C being a positive constant. Hence, in this case $\mathcal{H}_1(\epsilon, \mathcal{G}_n(K), Q_n)/n$ tends to zero in (outer)probability as n increases, from which the consistency of the least squares estimator follows. $\square$

Further applications to density estimation, penalized maximum likelihood and least squares, sieves, M-estimators, as well as to the evaluation of rates of convergence, are in the books by van der Vaart and Wellner (1996), van de Geer (2000), Györfi *et al.* (2002).

6. Functional delta method

6.1 Introduction and motivations

The second important step in developing general tools for asymptotic statistics is the so-called functional delta method. At an elementary level, delta method is widely used by statisticians. Suppose that $(Y_n; n \geq 1)$, V are r.v.s, and that $a_n; n \geq 1$, α are constants such that $a_n \rightarrow \infty$ and $a_n(Y_n - \alpha) \xrightarrow{d} V$ as n increases. Consider next a differentiable function g , with continuous first derivative such that $g'(\alpha) \neq 0$. Then, $a_n(g(Y_n) - g(\alpha))$

asymptotically behaves like $g'(\alpha)a_n(Y_n - \alpha)$. In symbols: $a_n(g(Y_n) - g(\alpha)) \xrightarrow{d} g'(\alpha)V$ as n tends to infinity. Delta method is a very powerful method in asymptotic statistics, since it provides for free the asymptotic distribution of $a_n(g(Y_n) - g(\alpha))$ when (i) the function g is smooth enough, and (ii) the asymptotic distribution of $a_n(Y_n - \alpha)$ is known.

As remarked in Section 2, many estimators are expressed as functionals of the empirical measure (or the empirical d.f., which is essentially the same). In symbols: $\hat{\theta}_n = \theta(F_n)$. Since the asymptotic behaviour of $\sqrt{n}(F_n(\cdot) - F(\cdot))$ is known (Ex. 12), a question naturally arises: Is it possible to obtain the asymptotic distribution of $\sqrt{n}(\theta(F_n) - \theta(F))$ on the basis of the asymptotic law of the random function (empirical process) $\sqrt{n}(F_n(\cdot) - F(\cdot))$? What we need is a general tool, similar in principle to the delta method illustrated above. Of course, there is a considerable mathematical difficulty. The random function $\sqrt{n}(F_n(\cdot) - F(\cdot))$ takes values in an infinite-dimensional function space, and what we need is actually to extend the notion of derivative to infinite-dimensional spaces.

6.2 Hadamard derivatives in normed spaces

Let $\mathcal{U}$, $\mathcal{V}$ be two linear normed spaces. To avoid confusion, we denote by $\|\cdot\|_{\mathcal{U}}$ the norm on $\mathcal{U}$, and by $\|\cdot\|_{\mathcal{V}}$ the norm on $\mathcal{V}$. A map $\psi : \mathcal{U} \rightarrow \mathcal{V}$ is *Gateaux-differentiable* at $u_0 \in \mathcal{U}$ if for every fixed $u \in \mathcal{U}$ there exists a linear, continuous map $\psi'_{u_0}(u)$ such that

$$\lim_{h \rightarrow 0} \left\| \frac{1}{h}(\psi(u_0 + hu) - \psi(u_0)) - \psi'_{u_0}(u) \right\|_{\mathcal{V}} = 0. \quad (19)$$

The quantity $\psi'_{u_0}(u)$ is the *Gateaux derivative* of ψ at u_0 , along the direction u . When both $\mathcal{U}$, $\mathcal{V}$ are the real line, clearly $\psi'_{u_0}(u) = u(d\psi/du|_{u_0})$. When $\mathcal{U} = \mathbb{R}^k$, $\mathcal{V} = \mathbb{R}$, $\psi'_{u_0}(u)$ reduces to the directional derivative of ψ , along the direction of the vector u .

The notion of Gateaux-differentiability is insufficient for our purposes. What is needed is a strengthening (19) based on a kind of uniformity in u . If (19) is required to hold uniformly for u in every compact subset of $\mathcal{U}$, then $\psi'_{u_0}(u)$ becomes the *Hadamard derivative* of ψ . Equivalently, a map ψ is *Hadamard differentiable* at u_0 if there exists a linear, continuous map $\psi'_{u_0} : \mathcal{U} \rightarrow \mathcal{V}$ such that

$$\lim_{h \rightarrow 0} \left\| \frac{1}{h}(\psi(u_0 + hu_h) - \psi(u_0)) - \psi'_{u_0}(u) \right\|_{\mathcal{V}} = 0 \quad (20)$$

for every $u_h \rightarrow u$ as $h \rightarrow 0$ (i.e. for every $\|u_h - u\|_{\mathcal{U}} \rightarrow 0$ as $h \rightarrow 0$).

The map ψ that appear in (20) is not necessarily required to be defined in the whole set $\mathcal{U}$. It suffices that the domain of ψ is a subset $\mathcal{U}_0$ of $\mathcal{U}$ containing u_0 , and that (20) holds for every $u_h \rightarrow u$ such that $u_0 + hu_h$ is in $\mathcal{U}_0$ for every h small enough. On the other hand, the map $\psi'_{u_0}(\cdot)$ is required to be defined on the whole $\mathcal{U}$. When $\psi'_{u_0}(\cdot)$ only exists on a subset $\mathcal{U}_1$ of $\mathcal{U}$, and the sequences $u_h \rightarrow u$ are restricted to converge to $u \in \mathcal{U}_1$, then $\psi'_{u_0}(\cdot)$ is the Hadamard derivative of ψ at u_0 , and *tangentially to $\mathcal{U}_1$* .

The notion of Hadamard-differentiability is stronger than that of Gateaux-differentiability: if ψ is Hadamard-differentiable, then it is also Gateaux-differentiable, and the two derivatives coincide. On the other hand, Gateaux-differentiability does not necessarily imply Hadamard-differentiability.

One of the most important properties of the “usual” derivative is the formula of derivatives of composite functions, or *chain rule*: $df(g(x))/dx = f'(g(x))g'(x)$. The same kind of formula also holds for general Hadamard derivatives. More precisely, let $\psi : \mathcal{U}_0 \rightarrow \mathcal{V}_0$,

$\phi : \mathcal{V}_0 \rightarrow \mathcal{W}$ be two maps defined on subsets $\mathcal{U}_0, \mathcal{V}_0$ of two normed spaces $\mathcal{U}, \mathcal{V}$, respectively. If (i) ψ is Hadamard-differentiable at u_0 tangentially to $\mathcal{U}_1$, and (ii) ϕ is Hadamard-differentiable at $\psi(u_0)$ tangentially to $\psi'_{u_0}(\mathcal{U}_1)$, then $\phi(\psi(u))$ is Hadamard-differentiable at u_0 tangentially to $\mathcal{U}_1$, with Hadamard derivative $\phi'_{\psi(u_0)}(\psi'_{u_0}(u))$. The chain rule makes the Hadamard derivative really useful, since in many cases of interest for applications the computation of the derivative of $\phi(\psi(u))$ is made easier. In fact, if $\phi'_{\psi(u_0)}(\cdot)$ and $(\psi'_{u_0}(\cdot))$ are already known, then the Hadamard derivative of $\phi(\psi(u))$ can be evaluated virtually effortlessly, without any *ex novo* calculation.

6.3 Main results and examples

To expose the functional delta method, we consider first the special but important case of the elementary empirical process. For the sake of simplicity, assume (although it is not necessary) that the population d.f. F_0 is continuous on $\mathbb{R}$. As already seen, the elementary empirical process $(\sqrt{n}(F_n(x) - F_0(x)); -\infty < x < +\infty)$ converges weakly, as n goes to infinity, to a Gaussian process $(G(x); -\infty < x < +\infty) = (B(F_0(x)); -\infty < x < +\infty)$, $(B(t); 0 \leq t \leq 1)$ being a Brownian bridge. To estimate a real parameter $\theta = \theta(F)$, consider the statistical functional $\hat{\theta}_n = \theta(F_n)$. The following result holds.

Theorem 3 (Functional delta method for the elementary empirical process) *Suppose that the functional $\theta(\cdot)$ is Hadamard differentiable at F_0 , with Hadamard derivative $\theta'_{F_0}(\cdot)$. Then, $\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}(\theta(F_n) - \theta(F_0))$, as n goes to infinity, converges in distribution to $\theta'_{F_0}(G(\cdot)) = \theta'_{F_0}(B(F_0(\cdot)))$. $\square$*

The limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is expressed as the probability law of $\theta'_{F_0}(B(F_0(\cdot)))$, i.e. as the probability law of a functional of a Gaussian process. Since the Hadamard derivative $\theta'_{F_0}(\cdot)$ is a *linear*, real-valued functional, using the Riesz representation theorem, it can be written in the form $\theta'_{F_0}(H) = \int H(x) d\phi_{F_0}(x)$, for an appropriate bounded variation function ϕ_{F_0} (depending on F_0). Hence, we may write:

$$\theta'_{F_0}(G) = \int_{-\infty}^{+\infty} G(x) d\phi_{F_0}(x). \quad (21)$$

Relationship (21) shows that $\theta'_{F_0}(G)$ is essentially an integral of the Gaussian process $G(\cdot)$. As a consequence, it possesses *normal* distribution, with mean

$$E_{P_0} \left[\int_{-\infty}^{+\infty} G(x) d\phi_{F_0}(x) \right] = \int_{-\infty}^{+\infty} E_{P_0}[G(x)] d\phi_{F_0}(x) = 0$$

and variance (which is the *asymptotic variance* of $\hat{\theta}_n$)

$$V \left[\int G(x) d\phi_{F_0}(x) \right] = \int \int (\min(F_0(x), F_0(y)) - F_0(x) F_0(y)) d\phi_{F_0}(x) d\phi_{F_0}(y) \quad (22)$$

provided that the integral in (22) exists. In other words, we have obtained, under fairly general conditions, the asymptotic normality of statistical functionals. The asymptotic variance (22) depends on the population d.f. F_0 , which is unknown. General techniques for estimating (22) are based on the jackknife, the infinitesimal jackknife and the bootstrap methods. Their consistency is studied in Shao (1993), Shao and Tu (1995).

The functional delta method can be expressed in a form considerably more general than Theorem 3, and useful for statistical applications. Its proof is, for instance, in Gill and van der Vaart (1993).

Theorem 4 (Functional delta method, general version) Let $\mathcal{U}, \mathcal{V}$ be two normed linear spaces, and let $\psi : \mathcal{U}_0 \rightarrow \mathcal{V}$ (with $\mathcal{U}_0 \subseteq \mathcal{U}$) be Hadamard differentiable at u_0 , tangentially to $\mathcal{U}_1$, and with Hadamard derivative $\psi'_{u_0}(u)$. Assume further that $(W_n; n \geq 1)$ is a sequence of random functions taking values in $\mathcal{U}_0$, and $(a_n; n \geq 1)$ is a sequence of constants such that $a_n(W_n - u_0)$ converges weakly to a random element W taking values in $\mathcal{U}_1$. Then $a_n(\psi(W_n) - \psi(u_0))$ converges weakly to $\psi'_{u_0}(W)$. $\square$

Example 14 (Ex. 3 contd.) Consider the monotone dependence measure $\theta(F_0)$ introduced Ex. 3, and its estimator $\hat{\theta}_n = \theta(F_n)$. Assume that the joint d.f. $F_0(x, y)$ is continuous, and make the transformation $U = F_{01}(X), V = F_{02}(Y)$, so that the corresponding joint d.f. reduces to the copula $M_0(u, v) = F_0(F_{01}^{-1}(u), F_{02}^{-1}(v))$, $u, v \in (0, 1)^2$. Denote by $F_{01|2}(x|y)$ ($F_{02|1}(y|x)$) the conditional d.f. of X given Y (of Y given X), and let $\phi_{12}(u, v) = F_{01|2}(F_{02}^{-1}(v)|F_{01}^{-1}(u)) + F_{01|2}(F_{01}^{-1}(1-u)|F_{02}^{-1}(v))$ and $\phi_{21}(u, v) = F_{02|1}(F_{02}^{-1}(v)|F_{01}^{-1}(u)) + F_{02|1}(F_{02}^{-1}(1-v)|F_{01}^{-1}(u))$. As a by-product of results in Cifarelli *et al.* (1996), θ is Hadamard differentiable tangentially to the set $\mathcal{H}$ of all continuous functions on $[0, 1]^2$ with $h(u, 0) = h(0, v) = 0$, and its Hadamard derivative is equal to:

$$\begin{aligned} \theta'_{F_0}(h) = & \int_0^1 \int_0^1 \{h(u, v) + h(u, 1-v) \\ & - h(u, 1)\phi_{21}(u, v) - h(1, v)\phi_{12}(u, v)\} dg(|u+v-1|). \end{aligned}$$

The sequence of random functions $(B_n(x, y) = \sqrt{n}(F_n(x, y) - F_0(x, y)); x, y \in \mathbb{R})$ converges weakly, as n increases, to a Gaussian process $(G(x, y); x, y \in \mathbb{R})$ that can be represented as $G(x, y) = B(F_{01}(x), F_{02}(y))$, where $(B(u, v); u, v \in [0, 1])$ is a *Brownian sheet*, i.e. a two-parameter Gaussian process with a.s. continuous trajectories, $B(u, 0) = B(0, v) = 0$, $E[B(u, v)] = 0$, $E[B(u_1, v_1)B(u_2, v_2)] = M_0(\min(u_1, u_2), \min(v_1, v_2)) - M_0(u_1, v_1)M_0(u_2, v_2)$. Since integrals of Gaussian processes possess normal distribution (provided it exists), from Theorem 3 it follows that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ tends in distribution, as n increases, to a normal r.v. with mean zero and variance

$$\sigma_0^2 = \int_0^1 \int_0^1 \int_0^1 \int_0^1 E[\xi(u_1, v_1)\xi(u_2, v_2)] dg(|u_1 + v_1 - 1|) dg(|u_2 + v_2 - 1|)$$

where $\xi(u, v) = B(u, v) + B(u, 1-v) - B(u, 1)\phi_{21}(u, v) - B(1, v)\phi_{12}(u, v)$. $\square$

Example 15 (Ex. 4 contd.) Consider again Ex. 4 (quantile function), and assume that the population d.f. F_0 is continuously differentiable on $[a, b] = [Q_0(u) - \epsilon, Q_0(v) + \epsilon]$ for some $\epsilon > 0$ and $0 < u < v < 1$, $Q_0(y) = F_0^{-1}(y)$ being the y th population quantile. Denote by $f_0(x) = F'_0(x)$ the population density function, and assume that f_0 is positive on $[a, b]$. The map $\theta(F_0) = (Q_0(y); u \leq y \leq v)$ (quantile function restricted to $[u, v]$) is Hadamard differentiable at F_0 tangentially to $C[a, b]$, the set of continuous functions on $[a, b]$, with derivative $\theta'_{Q_0}(g(\cdot)) = -g(\cdot)/f_0(\cdot)$. On the other hand, the sequence of random functions $\sqrt{n}(F_n(x) - F_0(x); x \in \mathbb{R})$ tends weakly to a Gaussian process $(G(x) = B(F_0(x)); x \in \mathbb{R})$, where $(B(u); 0 \leq u \leq 1)$ is a Brownian bridge (see Ex. 12). Since the random function $G(\cdot)$ possesses trajectories a.s. continuous, from Theorem 4 we conclude that the sequence of random functions $(\sqrt{n}(Q_n(y) - Q_0(y)); u \leq y \leq v)$ converges weakly to a Gaussian process $(G(y); u \leq y \leq v)$ that can be written in the form $G(y) = B(y)/f_0(Q_0(y))$. In particular, for a fixed y in $[u, v]$ $\sqrt{n}(Q_n(y) - Q_0(y))$ tends in distribution to a normal r.v. with mean zero and variance $y(1-y)/f_0(Q_0(y))^2$. $\square$

Example 16 (Ex. 5 contd.) In Ex. 5 we have defined the Nelson-Aalen estimator $\Lambda_n(t)$ (4). Now, F_{Xn} and F_{1n} are empirical d.f.s, and hence the couple $(F_{Xn}(x), F_{1n}(y))$ is asymptotically normal for every fixed x, y . It can be also shown that the stochastic process $(\sqrt{n}((F_{Xn}(x) - F_{X0}(x)), (F_{1n}(y) - F_{10}(y)); x, y \in \mathbb{R}))$ converges weakly to a Gaussian process $(G(x, y); x, y \in \mathbb{R})$, with $E[G(x, y)] = 0$ and an appropriate covariance function (cfr. Gill and Johansen (1990)). As shown in Gill and Johansen (1990), the map $\theta(F_{X0}, F_{10}) = (\Lambda_{T0}(t); 0 \leq t \leq \tau)$ is Hadamard-differentiable at (F_{X0}, F_{10}) , for every positive τ for which $F_{X0}^-(\tau) < 1$. Hence, the sequence stochastic processes $(\sqrt{n}(\Lambda_n(t) - \Lambda_{0T}(t); 0 \leq t \leq \tau))$ converges weakly to a Gaussian process.

As seen in Ex. 5, the cumulative risk is equal to $\Lambda_{T0}(t)$ is equal to $\int_0^t (1 - F_{T0}^-)^{-1} dF_{T0}$. The map $\psi(F_{T0}) = \Lambda_{T0}$ is invertible, and its inverse can be written as the *product integral* of $-\Lambda_{T0}$: $1 - F_{T0}(t) = \phi(\Lambda_{T0}) = \prod_{0 < u \leq t} (1 - \Lambda_{T0}(u)) \exp(-\Lambda_{T0}^c(t))$, Λ_{T0}^c being the continuous part of Λ_{T0}^c . In this way, we obtain the *Kaplan-Meyer estimator* of $1 - F_{T0}$:

$$1 - F_{Tn}^K(t) = \phi(\Lambda_n) = \prod_{i: X_i \leq t} \frac{\#(j : X_j \geq X_i) - \Delta_i}{\#(j : X_j \geq X_i)}.$$

Under appropriate regularity conditions, the map $\phi(\cdot)$ is Hadamard-differentiable at Λ_{T0} , and this implies the asymptotic normality of the Kaplan-Meyer estimator. $\square$

Example 17 (Ex. 7 contd.) Consider again Ex. 7. As already seen, the population adjustment coefficient R_0 is the smallest root of the equation $G_0(u) = 1$. Now, an integration by parts shows that for every positive u the relationships

$$G_{C0}(u) = u \int_{-\infty}^{+\infty} e^{ux} (1 - F_{C0}(x)) dx, \quad G_{W0}(u) = u \int_{-\infty}^{+\infty} e^{ux} (1 - F_{W0}(x)) dx$$

hold. For $u > 0$, let $h_u(x) = \exp\{\max(0, ux)\}$, denote by D_u the set of functions $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $h_u(x)f(x)$ is in $D[-\infty, +\infty]$ (cfr. Section 4), and define the norm $\|f\|_u = \sup_x |h_u(x)f(x)|$. If $f \in D_u$, then define $\varphi(f)(u) = u \int e^{ux} f(x) dx$, $u > 0$. The function $\psi(f_1, f_2)(u) = (\varphi(f_1)(u))(\varphi(f_2)(u))$, $f_1, f_2 \in D_u$, maps the set D_u into the set $C_\epsilon^1[0, u_1]$ of real functions continuous on $[0, u_1]$ and differentiable (with continuous derivative) on $[\epsilon, u_1]$, with $0 < \epsilon < u_1 < u$. Consider next the set A of all functions $q \in C_\epsilon^1[0, u_1]$ such that $q(0) = 1$, q is convex and differentiable in $[0, \epsilon]$, with $q'(0) < 0$. The set A is equipped by the norm $\|q\| = \sup_x |q(x)|$. For $q \in A$ define the functional $\psi(q) = \sup\{x \geq 0 : q(x) = 1\}$. From now on, we will assume that $G_0(u)$ is finite for every $u \in [0, \beta]$, $\beta > 2R_0$. Clearly, the population adjustment coefficient R_0 is equal to $\psi(\phi(1 - F_{C0}, 1 - F_{W0}))$. As a by-product of results in Capitanio and Conti (2003), the function ψ is Hadamard-differentiable at $(1 - F_{C0}, 1 - F_{W0})$, tangentially to $D_u \times D_u$ for every $2R_0 < u \leq \gamma$, with $2R_0 < \gamma < \beta$. Its Hadamard derivative is:

$$\phi'_{F_{C0}, F_{W0}}(f_1, f_2) = \varphi(1 - F_{C0})(R_0) \varphi(f_2)(R_0) + \varphi(1 - F_{W0})(R_0) \varphi(f_1)(R_0).$$

The functional $\psi(q)$ is Hadamard-differentiable at $G_0 = \phi(1 - F_{C0}, 1 - F_{W0})$, tangentially to the set $\phi'_{G_0}(D_u \times D_u)$, with Hadamard derivative $\psi'_{G_0}(g) = -g(\psi(G_0))/G'_0(\psi(G_0)) = -g(R_0)/G'_0(R_0)$. By applying the chain rule, it is seen that $\psi(\phi(\cdot))$ is Hadamard-differentiable at $(1 - F_{C0}, 1 - F_{W0})$, tangentially to $D_u \times D_u$, with Hadamard derivative:

$$(\psi(\phi))'_{F_{C0}, F_{W0}}(f_1, f_2) = -\frac{\varphi(1 - F_{C0})(R_0) \varphi(f_2)(R_0) + \varphi(1 - F_{W0})(R_0) \varphi(f_1)(R_0)}{G'_0(R_0)}.$$

Finally, the sequence of stochastic process $(G_{1n}(x), G_{2n}(y); x, y \in \mathbb{R}), n \geq 1$, with $G_{1n}(x) = \sqrt{n}(F_{Cn}(x) - F_{C0}(x))$, $G_{2n}(y) = \sqrt{n}(F_{Wn}(y) - F_{W0}(y))$, converges weakly to a two-parameter Gaussian process $(G(x, y) = (G_1(x), G_2(y)); x, y \in \mathbb{R})$, where $G_1(\cdot), G_2(\cdot)$ are independent Gaussian processes with $G_1(x) = B_1(F_{C0}(x))$, $G_2(y) = B_2(F_{W0}(y))$, B_1, B_2 being two independent Brownian bridges. Taking into account that G_1, G_2 take a.s. values in D_u , we conclude that the limiting distribution $\sqrt{n}(R_n - R_0)$ coincides with the law of $-(\varphi(1 - F_{C0})(R_0) \varphi(G_2)(R_0) + \varphi(1 - F_{W0})(R_0) \varphi(G_1)(R_0) / G'_0(R_0))$, that turns out to be normal with mean zero and variance $(G_{C0}(2R_0)G_{W0}(R_0)^2 + G_{W0}(2R_0)G_{C0}(R_0)^2 - 2G_{C0}(R_0)^2G_{W0}(R_0)^2) / G'_0(R_0)^2$. $\square$

7. Approximating empirical processes via bootstrap

The bootstrap method provides a simple approximation of the probability distribution of estimators (and test-statistics, as well). Assume that $X_1, \dots, X_n$ are *i.i.d.* with common probability distribution P_0 , and let $\theta_0 = \theta(P_0)$ be a parameter of interest. Denote by $\hat{\theta}_n = \hat{\theta}(P_n)$ an estimator of θ_0 , where P_n is the empirical measure introduced in Section 2. Consider next n r.v.s $\tilde{X}_1, \dots, \tilde{X}_n$ *i.i.d.* conditionally on $\mathbf{X}_n = (X_1, \dots, X_n)$, with $Pr(\tilde{X}_i \in B | \mathbf{X}_n) = P_n(B)$. The *bootstrap* probability measure $\tilde{P}_n$ is defined as:

$$\tilde{P}_n(B) = \frac{\#(\tilde{x}_i \in B)}{n}$$

for every subset B of $\mathcal{X}$. When n is large, the bootstrap measure $\tilde{P}_n(B)$ is close to the empirical measure $P_n(B)$, by the strong law of large numbers. On the other hand, when the sample size is large $P_n(B)$ is close to $P_0(B)$. From this simple remark, we conclude that, for every fixed B , $\tilde{P}_n(B)$ is close to $P_0(B)$ as the sample size n is large. As a consequence, when the sample size is large the probability distribution of $\hat{\theta}_n$ should be close to the conditional probability distribution of $\tilde{\theta}_n = \theta(\tilde{P}_n)$, given $\mathbf{X}_n$. Of course, the advantage of this procedure is that, conditionally on $\mathbf{X}_n$, the distribution of $\tilde{P}_n$ is *known*, and hence also the conditional distribution of $\tilde{\theta}_n$ is known. The precise meaning of these sentences heavily rests on the techniques developed so far. Using the same notation as in Section 2, consider a real-valued function g lying in some class $\mathcal{G}$, and let $\tilde{P}_n g = \int g d\tilde{P}_n$. Consider next the empirical process indexed by $\mathcal{G}$, $(G_n(g) = \sqrt{n}(P_n g - P_0 g); g \in \mathcal{G})$, and the corresponding “bootstrapped” process $(\tilde{G}_n(g) = \sqrt{n}(\tilde{P}_n g - P_n g); g \in \mathcal{G})$. The basic result in this direction is due to Giné and Zinn (1990), and tells us that, as n goes to infinity, the conditional law of $\tilde{G}_n(\cdot)$, given $\mathbf{X}_n$, possesses the same limit as the unconditional law of the empirical process $G_n(\cdot)$. More precisely, assume that $\mathcal{G}$ is a Donsker class, so that $G_n(\cdot)$ converges weakly to a Gaussian process $G(\cdot)$ (on $l^\infty(\mathcal{G})$) as n increases. Denote by $\tilde{E}$ the expectation w.r.t. the probability law of the bootstrap sample, conditionally on $\mathbf{X}_n$, and consider the set $BL(l^\infty(\mathcal{G}))$ of all uniformly Lipschitz functions $h : l^\infty(\mathcal{G}) \rightarrow [-1, 1]$. The following result holds.

Theorem 5 (Donsker theorem for bootstrapped processes) *Let $\mathcal{G}$ be a Donsker class with finite envelope function $\bar{g}$ with $P_0 \bar{g}^2$, so that $G_n(\cdot)$ converges weakly to $G(\cdot)$. Then:*

$$\sup_{h \in BL(l^\infty(\mathcal{G}))} \left| \tilde{E}[h(\tilde{G}_n) | \mathbf{X}_n] - E[h(G)] \right| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \square \quad (23)$$

In view of (13), statement (23) tells us that, for a set of data sequence having probability one, the bootstrap process $\tilde{G}_n(\cdot)$ converges weakly to $G(\cdot)$. In other words, the *conditional* law of $\tilde{G}_n(\cdot)$ possesses the same limit as the *unconditional* law of $G_n(\cdot)$.

Using Theorem 5, it is possible to prove a version of the functional delta method for bootstrapped processes. Denote by $\tilde{F}_n(x) = \sum I_{(\tilde{X}_i \leq x)}/n$ the bootstrap d.f., and consider a parameter $\theta = \theta(F_0)$, and estimators $\hat{\theta} = \theta(F_n)$, $\tilde{\theta} = \theta(\tilde{F}_n)$. Suppose further that the functional $\theta(\cdot)$ is Hadamard differentiable at F_0 , with Hadamard derivative $\theta'_{F_0}(\cdot)$. We already know that $\sqrt{n}(F_n - F_0)$ converges weakly to a Gaussian process $G(\cdot) = B(F_0(\cdot))$, $B(\cdot)$ being a Brownian bridge. As seen in Theorem 3, $\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}(\theta(F_n) - \theta(F_0))$, as n goes to infinity, converges in distribution to $\theta'_{F_0}(G(\cdot)) = \theta'_{F_0}(B(F_0(\cdot)))$.

Theorem 6 *If the functional $\theta(\cdot)$ takes values in $\mathbb{R}^k$, $k \geq 1$, then*

$$\sup_{h \in BL(\mathbb{R}^k)} \left| E[h(\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)) | \mathbf{X}_n] - E[h(\theta'_{F_0}(G))] \right| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty. \square \quad (24)$$

Taking into account relationship (13), from Theorems 3, 6 we conclude that the conditional distribution of $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$, given $\mathbf{X}_n$, tends to the same limit as the unconditional distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Note that in (24) we have convergence in probability instead of almost sure. This means that the result now obtained is weaker than those of Theorem 5. More general version of Theorem 6 are in the paper by Giné and Zinn (1990).

8. Asymptotic efficiency in nonparametric models

A very important point in estimation theory is the choice among different estimators of the same parameter. In parametric models, apart from a few special cases, this problem can be effectively approached only when the sample size is large. The reason for this fact is simple. To study the efficiency of estimators, it is first necessary to establish a tight upper bound for their mean squared errors, and this is possible only when the sample size goes to infinity. For this reason, from now on we will consider the *asymptotic efficiency* of estimator. Asymptotic efficiency for *parametric* models is well established since the beginning of the Seventies; an excellent account is in Lehmann and Casella (1998). The first step towards a theory of asymptotic efficiency for nonparametric models is in Levit (1978). Many results are in Gill (1989). An excellent account, with special emphasis on methods for constructing estimators, is in Bickel *et al.* (1993).

For the sake of simplicity, consider a parameter real $\theta(P)$, $P \in \mathcal{P}$, and denote by P_0 and $\theta_0 = \theta(P_0)$ the true population distribution and the true parameter, respectively. For every one-dimensional submodel $\mathcal{P}_0 = \{P_t; 0 < t < \tau\} \subseteq \mathcal{P}$ containing the true P_0 (obtained as t goes to 0), define the score function g at $t = 0$ as a (measurable) function $s : \mathcal{X} \rightarrow \mathbb{R}$ such that

$$\lim_{t \rightarrow 0} \int \left\{ \frac{1}{t} (dP_t^{1/2} - dP_0^{1/2}) - \frac{1}{2} s(x) dP_0^{1/2} \right\}^2 = 0 \quad (25)$$

In several cases the score function takes a simple form. Suppose for instance that the probability distributions P_t possess a density p_t w.r.t. the Lebesgue measure, so that (25) reduces to $\int \{(p_t^{1/2} - p_0^{1/2})/t - s(x)p_0(x)\}^2 dx$. Then, $s(x)$ reduces to $\partial \log p_t / \partial t|_{t=0}$,

which is the “usual” score function. The set $\dot{\mathcal{P}}_0$ of score functions obtained as $\mathcal{P}_0$ ranges over a class of submodels, is a *tangent set* of the model $\mathcal{P}$ at P_0 . Denoting by $L^2(P_0)$ the set of all functions g for which $P_0 g^2$ is finite, the relationship $\dot{\mathcal{P}}_0 \subseteq L^2(P_0)$ holds.

To define the information for estimating $\theta(P_0)$, we consider only submodels $\mathcal{P}_0$ such that $\theta(P_t)$ is Hadamard-differentiable at $t = 0$, tangentially to $\dot{\mathcal{P}}_0$. According to (20), there exists a linear, continuous function $\dot{\theta}_0: L^2(P_0) \rightarrow \mathbb{R}^k$ such that $\lim_{t \rightarrow 0} (\theta(P_t) - \theta(P_0))/t = \dot{\theta}_0(s)$ for every s in $\dot{\mathcal{P}}_0$. Using the Riesz representation theorem, it is possible to write $\dot{\theta}_0(s) = \int s(x) \tilde{\theta}_0(x) dP_0$ for an appropriate function $\tilde{\theta}_0(\cdot)$, which is unique in the closure $\overline{\text{span}}(\dot{\mathcal{P}}_0)$ of all linear combinations of elements of $\dot{\mathcal{P}}_0$. Consider now a submodel $\mathcal{P}_{0,s}$ with score function s . With the same notation as in Section 2, the optimal asymptotic variance to estimate $\theta(P_0)$, under $\mathcal{P}_{0,s}$, is equal to (Lehmann and Casella (1998))

$$(d\theta(P_t)/dt|_{t=0})^2 / P_0 s^2 = \dot{\theta}_0(s)^2 / P_0 s^2. \quad (26)$$

The supremum of (26), taken w.r.t. all score functions s in $\overline{\text{span}}(\dot{\mathcal{P}}_0)$, defines the *optimal asymptotic variance for the nonparametric model $\mathcal{P}$* . By Schwarz inequality, it is not difficult to see that the optimal asymptotic variance for $\mathcal{P}$ is equal to $P_0 \tilde{\theta}_0^2$, and plays a role similar to that of the optimal asymptotic variance in parametric models. In particular, under appropriate restrictions of the estimators considered (necessary to avoid superefficiency; see Lehmann and Casella (1998), Bickel *et al.* (1993)), the asymptotic variance of every (regular) sequence of estimators $(\hat{\theta}_n; n \geq 1)$ is bounded from below by $P_0 \tilde{\theta}_0^2$. The sequence $(\hat{\theta}_n; n \geq 1)$ is *asymptotically efficient* if $\sqrt{n}(\hat{\theta} - \theta_0)$ tends in distribution, as n increases, to a normal r.v. with mean zero and variance $P_0 \tilde{\theta}_0^2$. The notion of asymptotic efficiency can be easily extended to vector parameters, and is particularly important for semiparametric models. For these (and many other) topics, the reader is deferred to Bickel *et al.* (1993), van der Vaart (1996), Robins and Ritov (1997).

References

- Alon N., Ben-David S., Cesa-Bianchi N. and Haussler D. (1997) Scale-sensitive dimensions, uniform convergence, and learnability, *Journal of the ACM*, 44, 615–631.
- Andersen E. (1957) On the collective theory of risk in the case of contagion between the claims, *Transactions of the XVth International Congress of Actuaries*.
- Bickel P., Klaassen C., Ritov Y. and Wellner J. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore.
- Billingsley P. (1999) *Convergence of Probability Measures (II Ed.)*, Wiley, New York.
- Birgé L. (1983) Approximation dans les espaces métriques et théorie de l'estimation, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 65, 181–237.
- Birgé L. and Massart P. (1993) Rates of convergence for minimum contrast estimators, *Probab. Theory Related Fields*, 97, 113–150.
- Birman M. and Solomjak M. (1967) Piece-wise polynomial approximations of functions in the classes w_p^α , *Math. USSR Sbornik*, 73, 295–317.
- Capitanio A. and Conti P. (2003) A Bayesian nonparametric approach to the estimation of the adjustment coefficient, with applications to insurance and telecommunications, *Sankhya A (to appear)*.
- Cifarelli D., Conti P. and Regazzini E. (1996) On the asymptotic distribution of a general measure of monotone dependence, *Ann. Statist.*, 24, 1386–1399.

- Donsker M. (1952) Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems, *Ann. Math. Statist.*, 23, 277–281.
- Doob J. (1949) Heuristic approach to the Kolmogorov-Smirnov theorems, *Ann. Math. Statist.*, 20, 393–403.
- Gill R. (1989) Non- and semi-parametric maximum likelihood estimators and the von mises method (part i), *Scan. J. Statist.*, 16, 97–128.
- Gill R. and Johansen S. (1990) A survey of product-integration with a view towards application in survival analysis, *Ann. Statist.*, 18, 1501–1555.
- Gill R. and van der Vaart A. (1993) Non- and semi-parametric maximum likelihood estimators and the von mises method (part ii), *Scan. J. Statist.*, 20, 271–288.
- Giné E. and Zinn J. (1990) Bootstrapping general empirical measures, *Ann. Probability*, 18, 851–869.
- Grandell J. (1991) *Aspects of Risk Theory*, Springer Verlag, New York.
- Györfi L., Kohler M., Krzyżak A. and Walk H. (2002) *A Distribution-Free Theory of Nonparametric Regression*, Springer Verlag, New York.
- Hausser D. (1992) Decision theoretic generalizations of the pac model for neural net and other learning applications, *Information and Computation*, 100, 78–150.
- Hoffmann-Jørgensen J. (1991) *Stochastic Processes in Polish Spaces*, Aarhus Universitet Matematisk Institut, Various Publication Series No. 39.
- Lehmann E. and Casella G. (1998) *Theory of Point Estimation (II Ed.)*, Springer Verlag, New York.
- Levit B. (1978) Infinite-dimensional informational lower bounds, *Theory Probab. Appl.*, 23, 388–394.
- Pfanzagl J. (1988) Consistency of the maximum likelihood estimator for certain nonparametric families, in particular: mixtures, *J. Statist. Plann. Inference*, 19, 137–158.
- Robins J. and Ritov Y. (1997) Towards a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models, *Statist. Med.*, 16, 285–319.
- Schweizer B. and Sklar A. (1983) *Probabilistic metric spaces*, North Holland, Amsterdam.
- Shao J. (1993) Differentiability of statistical functionals and consistency of the jackknife, *Ann. Statist.*, 21, 61–75.
- Shao J. and Tu D. (1995) *The Jackknife and Bootstrap*, Springer Verlag, New York.
- Talagrand M. (1996) The Glivenko-Cantelli problem, ten years later, *J. Theoret. Probab.*, 9, 371–384.
- van de Geer S. (1993) Hellinger-consistency of certain nonparametric maximum likelihood estimators, *Ann. Statist.*, 21, 14–44.
- van de Geer S. (2000) *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge.
- van de Geer S. and Wegkamp M. (1997) Consistency for the least squares estimators in nonparametric regression, *Ann. Statist.*, 24, 2513–2523.
- van der Vaart A. (1996) Efficient estimation in semiparametric models, *Ann. Statist.*, 24, 862–878.
- van der Vaart A. and Wellner J. (1996) *Weak Convergence and Empirical Processes*, Springer Verlag, New York.
- Vapnik V. (1998) *Statistical learning theory*, Wiley, New York.
- Vapnik V. and Chervonenkis A. (1971) On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab. Appl.*, 16, 264–280.