

Consistency in nonparametrics: a revisit and some recent developments

*Consistenza in ambito non parametrico:
rivisitazione e sviluppi recenti*

Gabriella Salinetti

Dipartimento di Statistica
Probabilità e Statistiche Applicate
Università di Roma “La Sapienza”
gabriella.salinetti@uniroma1.it

Luca Tardella

Dipartimento di Statistica
Probabilità e Statistiche Applicate
Università di Roma “La Sapienza”
luca.tardella@uniroma1.it

Riassunto: All'interno della vasta letteratura sulla consistenza nell'inferenza non parametrica il lavoro rivisita alcune analisi di consistenza per le stime di massima verosimiglianza e per le procedure bayesiane con l'obiettivo di mettere in luce quali aspetti del problema possono determinarne un comportamento parallelo talvolta rilevato ma non ancora completamente sviluppato in ambito non parametrico.

Keywords: Consistency, Nonparametric Inference, Maximum Likelihood, Bayesian Inference, Glivenko-Cantelli Class, Uniformly consistent test, Hypo-convergence.

1. Introduction and problem setting

In the usual context of a parametric model for independent and identically distributed observations consistency of inferential procedure represent a fundamental property which ought to be preliminarily ascertained. In this paper we aim at reviewing some old and recent results concerning conditions which guarantee consistency in a context where the dimension of the parameter space is infinite, commonly referred to as nonparametric setting. More specifically we aim at focussing on inferential problems where the parameter space is a set of density functions and we will be considering as inferential techniques only maximum likelihood estimation and Bayesian procedures. In illustrating the main ideas underlying some of the recent approach towards determining general conditions for consistency we try to bring to the light the presence of a common ground for conditions ensuring consistency in both approaches.

Let $\mathcal{M} = \{P_\theta; \theta \in \Theta\}$ a family of distributions on the observation space $\mathcal{X}$. Here $\mathcal{X}$ is a finite dimensional euclidean space and $\mathcal{A}_\mathcal{X}$ its Borel σ -field. It is understood that the mapping $\theta \mapsto P_\theta$ is 1-1, so that we can directly refer to the space Θ endowed with a metric d with respect to which (Θ, d) will be separable and complete. We assume that $\mathcal{M}$ is dominated by a σ -finite measure μ and for any P_θ we denote with $f(\cdot, \theta)$ the corresponding density with respect to μ . Furthermore we assume that there exists $\theta_0 \in \Theta$ such that $X_1, \dots, X_n, \dots$ are independent and identically distributed according to P_{θ_0} . In dealing with consistency issues it is convenient to refer to the space $(\Omega, \mathcal{A}_\Omega, \nu)$ underlying the whole process, where $\Omega = \mathcal{X}^\infty$, $\mathcal{A}_\Omega$ and $\nu = P_{\theta_0}^\infty$.

In the context of a parameter space Θ endowed with metric d , a sequence of estimators $T_n^\omega = T_n(X_1(\omega), \dots, X_n(\omega))$ is said to be d -consistent in θ_0 when, almost surely with

respect to ν (heretoeafter denoted a.s.[ν]), it is $d(T_n^\omega, \theta_0) \rightarrow 0$ that is

$$T_n^\omega \rightarrow \theta_0. \quad (1)$$

We recall that in many density estimation problems a privileged distance is the Hellinger distance defined as follows:

$$d_H^2(f, g) = \frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 \mu(dx) = 1 - \int_{\mathcal{X}} \sqrt{f(x)g(x)} \mu(dx).$$

The likelihood function corresponding to the first n observations $(X_1(\omega), \dots, X_n(\omega))$ is given by

$$\theta \mapsto L_n(\omega, \theta) = \prod_{i=1}^n f(X_i(\omega), \theta)$$

and the maximum likelihood estimate is the maximizing argument of $L_n(\omega, \theta)$ when θ varies in Θ , that is

$$\hat{\theta}_n^\omega \in \arg \max_{\theta \in \Theta} L_n(\omega, \theta).$$

One can define

$$H_n(\omega, \theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i(\omega), \theta)}{f(X_i(\omega), \theta_0)} = \int \log \frac{f(x, \theta)}{f(x, \theta_0)} P_n(\omega, dx)$$

where $P_n(\omega, \cdot)$ is the empirical measure associated to the first n observations and rephrase the maximum likelihood estimate as follows

$$\hat{\theta}_n^\omega \in \arg \max_{\theta \in \Theta} H_n(\theta, \omega). \quad (2)$$

Note that $\arg \max_{\theta \in \Theta} H_n(\theta, \omega)$ need not to be made of a single point and it might be empty. However, in order to keep out of the presentation formal detailed considerations about the existence of $\hat{\theta}_n$ and the measurability of events related to maximizing arguments we will invoke *permissible* conditions to use a terminology introduced by Pollard (1984); In other words we assume that conditions are met so that a.s.[ν] $\arg \max H_n \neq \emptyset$ for all n and all the measurability issues can be given a positive answer.

In the Bayesian approach inference on θ is based on the posterior distribution which can be determined in our setting by the Bayes formula

$$B \mapsto \Pi_n(\omega, B) = \frac{\int_B \prod_{i=1}^n f(X_i(\omega), \theta) \Pi(d\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i(\omega), \theta) \Pi(d\theta)}$$

or equivalently

$$\Pi_n(\omega, B) = \frac{\int_B \exp \{n H_n(\omega, \theta)\} \Pi(d\theta)}{\int_{\Theta} \exp \{n H_n(\omega, \theta)\} \Pi(d\theta)} \quad (3)$$

for any $B \in \mathcal{B}_{\Theta}$, where $\mathcal{B}_{\Theta}$ is the Borel σ -field of subsets of Θ . The posterior distributions is then viewed as a random probability measure, defined on Ω , valued in the space of all probability distributions on $(\Theta, \mathcal{B}_{\Theta})$.

In the Bayesian approach one notion of consistency is referred to the asymptotic behaviour of the posterior distribution i.e. when a.s.[ν] the sequence of posterior distributions Π_n converges weakly to the distribution concentrated at P_0 . Formally

The sequence of posterior distributions Π_n is said to be consistent at θ_0 if there exists $N \in \mathcal{A}$ with $\nu(N) = 0$ such that for all $\omega \in \Omega \setminus N$ and for any choice of an open neighborhood $B(\theta_0)$ of θ_0 ,

$$\Pi_n(\omega, B(\theta_0)) = \frac{\int_{B(\theta_0)} \prod_{i=1}^n f(X_i(\omega), \theta) \Pi(d\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i(\omega), \theta) \Pi(d\theta)} \rightarrow 1. \quad (4)$$

Remark It is not difficult to prove that under mild conditions consistency in the above sense implies consistency of the Bayesian estimator $T_n^\omega = E[\theta | X_1(\omega), \dots, X_n(\omega)]$ in the former classical sense of (1).

Apparently from (2) and (3) both maximum likelihood and Bayesian consistency depends on the asymptotic behaviour of H_n . In this sense we immediately observe that under general conditions the strong law of large numbers implies for any $\theta \in \Theta$, the a.s.[ν] convergence of $H_n(\omega, \theta)$ to $H_0(\theta)$, where

$$H_0(\theta) = \int \log \frac{f(x, \theta)}{f(x, \theta_0)} P_{\theta_0}(dx) = -K(\theta, \theta_0)$$

and $K(\theta, \theta_0)$ is the well-known Kullback-Leibler divergence of $f(\cdot, \theta)$ from $f(\cdot, \theta_0)$. We also anticipate at this point that this type of convergence is not enough to obtain consistency.

Hypo-convergence instead is a suitable type of convergence which will allow to obtain useful consistency results. We limit ourselves to give its definition and briefly mention the essential features which will be used in the next sections.

A sequence of functions $\{H_n : \Theta \mapsto \bar{\mathbb{R}}; n \in \mathbb{N}\}$ hypo-converge to $H_0 : \Theta \mapsto \bar{\mathbb{R}}$, denotes $H_n \xrightarrow{\text{hypo}} H_0$, if for any $\theta \in \Theta$

$$\begin{aligned} \exists \theta_n \rightarrow \theta : \quad H_0(\theta) &\leq \liminf H_n(\theta_n) \\ \forall \theta_m \rightarrow \theta, \forall \{n_m\} \subset \{n\} \quad H_0(\theta) &\geq \limsup H_{n_m}(\theta_m). \end{aligned} \quad (5)$$

It is immediate to verify that in presence of hypo-convergence, from (5) the following holds true

$$\liminf_{n \rightarrow \infty} \sup_{\theta \in \Theta} H_n \geq \sup_{\theta \in \Theta} H_0; \quad (6)$$

furthermore, one can show that

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Theta} H_n \leq \sup_{\theta \in \Theta} H_0$$

if and only if for any $\epsilon > 0$, there exists a compact subset $C \subset \Theta$ such that for any n it is

$$\epsilon\text{-arg max}_{\theta \in \Theta} H_n \cap C \neq \emptyset \quad (7)$$

where $\epsilon\text{-arg max}_{\theta \in \Theta} H_n = \{\theta \in \Theta : H_n(\theta) > \sup_{\theta \in \Theta} H_n(\theta) - \epsilon\}$. In addition if H_0 admits a unique maximum, θ_0 , then hypo-convergence and (7) imply that for any choice of $\theta_n \in \arg \max H_n$, θ_0 it is the unique cluster point for the sequence $\{\theta_n; n \in \mathbb{N}\}$. However, this is not enough to ensure that for any choice of $\theta_n \in \arg \max H_n$, one has $\theta_n \rightarrow \theta_0$.

2. Maximum likelihood consistency

Sufficient conditions for maximum likelihood consistency have been investigated in a large number of articles even though not always in the general context of infinite-dimensional parameter space. Among the fundamental contributions we like to mention Wald (1949), LeCam (1953), Kiefer and Wolfowitz (1956), Bahadur (1967), Huber (1967) e Perlman (1972) for their ability to highlight with examples and counterexamples crucial points in obtaining consistency. In all these works it is clear that obtaining consistency calls for determining an appropriate way to control the asymptotic behaviour of $H_n(\omega, \theta)$ in a stronger sense than what is guaranteed by the law of large numbers; just simple paradigmatic counterexamples, see for instance Le Cam (1990), show that this type of convergence is not sufficient to guarantee consistency even in a finite-dimensional parameter space.

One way to single out an appropriate asymptotic behaviour can be suggested observing that θ_0 is in fact solution for the following maximization problem

$$H_0(\theta_0) = \sup_{\theta \in \Theta} H_0(\theta) = - \inf_{\theta \in \Theta} K(\theta, \theta_0) = -K(\theta_0, \theta_0) = 0,$$

that is

$$\{\theta_0\} = \arg \max_{\theta \in \Theta} H_0(\theta).$$

Since

$$\hat{\theta}_n^\omega \in \arg \max_{\theta \in \Theta} H_n(\theta, \omega)$$

it is then clear that maximum likelihood consistency require the analysis of the convergence of stochastic suprema.

In light of the brief considerations remarked in the introduction it seems natural to tackle consistency issues for maximum likelihood with the aid of the almost sure hypo-convergence i.e.,

$$H_n(\omega, \cdot) \xrightarrow{hypo} H_0(\cdot) \quad \text{a.s.}[\nu].$$

This can be regarded as a minimal context in order to study the convergence of stochastic maximization problems (values and arguments).

In fact, this kind of general approach revealed itself particularly appropriate to study consistency of a large class of estimators obtained as solutions of stochastic optimization problems and has been already adopted in Salinetti (1990) within a finite-dimensional framework. One can keep the same approach when considering infinite-dimensional parameter space and as long as the extension is allowed one can get similar conditions for nonparametric consistency. Along this direction we recall the recent contributions of Hess (1996) e Dong and Wets (2000).

A key condition for obtaining maximum likelihood consistency but which will play a fundamental role even for posterior consistency is the following:

there exists a subset $N \subset \Omega$ with $\nu(N) = 0$ such that for any $\omega \in \Omega \setminus N$

$$(C) \quad \forall B(\theta_0), \exists \epsilon > 0 : \sup_{\theta \in B(\theta_0)^c} H_n(\omega, \theta) < -\epsilon,$$

for n sufficiently large

where $B(\theta_0)^c$ stands for the complement of $B(\theta_0)$, open neighborhood of θ_0 .

One can easily get the following result (see also Perone Pacifico *et al.* (2000))

Proposition 1 *Suppose condition (C) holds: then, ν -almost surely, for any sequence $\hat{\theta}_n^\omega$ con $\hat{\theta}_n^\omega \in \arg \max_{\theta \in \Theta} H_n(\omega, \theta)$, one get*

$$\hat{\theta}_n^\omega \rightarrow \theta_0. \quad (8)$$

Furthermore, if ν -almost surely, $H_n(\omega, \cdot) \xrightarrow{hypo} H_0(\cdot)$, then

$$\sup_{\theta \in \Theta} H_n(\omega, \theta) \rightarrow \sup_{\theta \in \Theta} H_0(\theta) = H_0(\theta_0) = 0. \quad (9)$$

Proof Since $H_n(\omega, \theta_0) = 0$, condition (C) implies $\arg \max_{\theta \in \Theta} H_n(\omega, \theta) \subset B(\theta_0)$ for n sufficiently large. The same argument can be repeated for any $B(\theta_0)$ so that (8) is proved. Similarly one can show that $\sup_{\theta \in \Theta} H_n(\omega, \theta) = \sup_{\theta \in B(\theta_0)} H_n(\omega, \theta)$ hence $\limsup_{n \rightarrow \infty} \sup_{\theta \in \Theta} H_n(\omega, \theta) \leq \sup_{\theta \in \Theta} H_0(\theta) - \epsilon$; hence, using (6), one gets the convergence in (9). $\diamond\diamond$

Remark When almost sure hypo-convergence holds true and (Θ, d) is a compact metric space then it is not difficult to verify that condition (C) is met and that one always gets d -consistency. One can weaken the compactness requirement, with local compactness as long as supplementary convexity property of the parameter space together with the functions $d(\cdot, \theta)$ and $H_0(\cdot)$ are satisfied.

Within all the literature devoted to studying maximum likelihood consistency in a nonparametric setting one of the fundamental approaches that we would like to pause is given by those results which exploit the use of Glivenko-Cantelly-class-type results, that is by using uniform strong law of large numbers. In fact, since the final aim is to get convergence of argmax, one can focus on those classes of density functions $f(\cdot, \theta)$ which ensures the ν -almost sure convergence of $H_n(\omega, \cdot)$ to $H_0(\cdot)$ or some variants from that. In these perspective one can look at results obtained in van de Geer (1993) recently extended by Dudley (1998).

We will concentrate on van de Geer (1993) as a fundamental article representative of that approach. There, the metric for Θ is the Hellinger distance d_H and the starting point which can be easily derived is the following inequality:

$$\begin{aligned} d_H(\hat{\theta}_n^\omega, \theta_0) &\leq \int_{x: f(x, \theta_0) > 0} \left[\sqrt{\frac{f(x, \hat{\theta}_n^\omega)}{f(x, \theta_0)}} - 1 \right] P_n(\omega, dx) + \\ &\quad - \int_{x: f(x, \theta_0) > 0} \left[\sqrt{\frac{f(x, \hat{\theta}_n^\omega)}{f(x, \theta_0)}} - 1 \right] P_{\theta_0}(\omega, dx) \\ &= \int_{\mathcal{X}} g(x, \hat{\theta}_n^\omega) P_n(\omega, dx) - \int_{\mathcal{X}} g(x, \hat{\theta}_n^\omega) P_{\theta_0}(dx) \end{aligned}$$

where

$$g(x, \theta) = \begin{cases} \sqrt{\frac{f(x, \theta)}{f(x, \theta_0)}} - 1 & x : f(x, \theta) > 0 \\ 0 & \text{otherwise} \end{cases}$$

A sufficient condition for obtaining consistency of the maximum likelihood estimator can be derived by requiring that the class of functions $\mathcal{G} = \{g(\cdot, \theta), \theta \in \Theta\}$ has the Glivenko-Cantelli property, that is, ν -almost surely

$$(U) \quad \sup_{g \in \mathcal{G}} \left| \int_{\mathcal{X}} g(x, \theta) P_n(\omega, dx) - \int_{\mathcal{X}} g(x, \theta) P_{\theta_0}(dx) \right| \rightarrow 0.$$

Hence to arrive at consistency one can use sufficient conditions which ensure that $\mathcal{G}$ is a Glivenko-Cantelli class. These results can be found in many books about the asymptotic behaviour of empirical processes (for instance van der Vaart and Wellner (1996)) and are mainly related to the limiting behaviour of entropy numbers.

It is relevant to point out at this point that ν -almost sure uniform convergence in (U) implies the former condition (C). In fact, (U) implies that for any $\omega \in \Omega \setminus N$ with $\nu(N) = 0$, and for any choice of $\delta > 0$ one has eventually and uniformly for $\theta \in \Theta$

$$\int_{\mathcal{X}} g(x, \theta) P_n(\omega, dx) \leq \int_{\mathcal{X}} g(x, \theta) P_{\theta_0}(dx) + \delta = -d_H^2(\theta, \theta_0) + \delta \leq \delta;$$

since $H_n(\omega, \theta) \leq 2 \int_{\mathcal{X}} g(x, \theta) P_n(\omega, dx)$ this is enough to show that (C) is true since

$$\begin{aligned} \sup_{\theta \in B(\theta_0)^c} H_n(\omega, \theta) &\leq 2 \sup_{\theta \in B(\theta_0)^c} \int_{\mathcal{X}} g(x, \theta) P_n(\omega, dx) \\ &\leq 2 \sup_{\theta \in B(\theta_0)^c} \int_{\mathcal{X}} g(x, \theta) P_{\theta_0}(dx) + \delta \\ &= \delta - \inf_{\theta \in B(\theta_0)^c} d_H^2(\theta, \theta_0) \end{aligned}$$

and considering that outside of an Hellinger neighborhood of θ_0 , $d_H^2(\theta, \theta_0) \geq \eta > 0$ hence (C) is obtained by choosing $\delta < \eta^2$ and $\epsilon = \eta^2 - \delta$.

3. Posterior consistency

The interest for studying the consistency property of posterior distributions as defined in (4) in a non parametric setting has been recording an intensification relatively recently. Actually a different notion of Bayesian consistency had been formalized and proved in an old result due to Doob (1949). That result, often referred to as Doob's theorem, claims under mild measurability conditions, generally verified, that there exists a subset Θ' with prior probability $\Pi(\Theta') = 1$ such that for all $\theta' \in \Theta'$ the sequence of posterior distributions tends to degenerate to the distribution with a single point mass at θ' . However the fact that Θ' might exclude some elements of the whole Θ prevent that property from being a sufficient consistency warrantee for those who deem relevant to study the inferential properties of statistical procedure uniquely in the probability space $(\mathcal{X}^\infty, \mathcal{A}_{\mathcal{X}^\infty}, P_{\theta_0}^\infty)$ where, since θ_0 is not known can be potentially *any* element of Θ .

Some additional arguments have been explained in Diaconis and Freedman (1986) in order to convince about the usefulness of the study of posterior consistency as in (4). Those arguments, appealing to the notion of *merging* of posterior distributions corresponding to two possibly different prior input, together with revealing that inconsistency pathologies were not so rare or cumbersome, provided a decisive boost to the investigation of posterior consistency in the sense of (4).

We immediately note that an unescapable condition which must be assumed for obtaining posterior consistency is that the prior probability assigned to any neighborhood of θ_0 must be positive that is, for any $B(\theta_0)$,

$$\Pi(B(\theta_0)) > 0.$$

In fact, if $\Pi(B(\theta_0)) = 0$ for some $B(\theta_0)$, then the numerator of (4) would be null for all n . When Θ is infinite-dimensional the above requirement is usually strengthened by imposing that the prior probability assigned to any Kullback-Leibler neighborhood of θ_0 , namely to subsets of the following kind $N_\delta(\theta_0) = \{\theta \in \Theta : K(\theta, \theta_0) < \delta\}$. We will denote $\theta_0 \in \mathcal{S}_K(\Pi)$, when the prior distribution Π is such that

$$(K) \quad \Pi(N_\delta(\theta_0)) > 0 \quad \forall \delta > 0.$$

The role of this hypothesis is that it allows to lower-bound the asymptotic behaviour of the numerator of (3).

In this section we provide an essential review, certainly not covering all the results related to posterior consistency in nonparametrics.

In a sense the pioneering result about posterior consistency as defined in (4) has been established in Schwartz (1965) where it is exploited the existence of a sequence of uniformly consistent test for the following hypotheses: $H_0 : \theta = \theta_0$ against $H_1 : \theta \in B(\theta_0)^c$. In case some mild measurability conditions are met the result can be stated as follows

[Schwartz (1965)] *Let (Θ, d) a complete and separable metric space. If $\theta_0 \in \mathcal{S}_K(\Pi)$ and for any open neighborhood $B(\theta_0)$ there exists a sequence of test functions $0 \leq \phi(x_1, \dots, x_n) \leq 1$ so that*

$$(S) \quad \int_{\mathcal{X}} \phi_n(x_1, \dots, x_n) \prod_{i=1}^n f(x_i, \theta_0) dx_1 \dots dx_n < \inf_{\theta \in U^c} \int_{\mathcal{X}} \phi_n(x_1, \dots, x_n) \prod_{i=1}^n f(x_i, \theta) dx_1 \dots dx_n$$

then the sequence of posterior distributions is d -consistent in θ_0 .

In fact this result can be very powerful for it allows, with particular choices of d to get broad consistency results. Just to mention an example if Θ represent a subset of the space, $\mathcal{F}_\lambda$, of distributions which are absolutely continuous with respect to the Lebesgue measure, and d_w is a distance compatible with the topology of weak convergence on Θ it is known (Ghosal *et al.*, 1999) that condition (S) always holds. While this result is very pleasing for its minimal requirement about $\theta_0 \in \mathcal{S}_K(\Pi)$ on the other hand it produces a consistency result with respect to a metric d_w certainly not particularly appropriate if one considers that $\mathcal{F}_\lambda$ represent a subset of density function for which a more natural distance would be the L^1 norm or, which is equivalent, the Hellinger distance. However the possibility of ensuring the existence of uniformly consistent tests is at risk and cannot hold when Θ is endowed with the Hellinger distance (LeCam (1973), Barron (1989)). For

this reason, investigating d_H -consistency for the posteriors has been addressed through controlling the asymptotic behaviour of the likelihood within an appropriately defined sequence of subsets (sieves), here denoted with V_n , in such a way that, whatever is the likelihood behaviour outside V_n , this cannot matter substantially as long as the prior probability $\Pi(V_n^c)$ vanishes at exponential rate. All the recent results obtained in Barron *et al.* (1999), Ghosal *et al.* (1999) and Petrone and Wasserman (2002)) obey to this principle. To give a more precise idea of this way of guaranteeing posterior consistency we only report the result in Ghosal *et al.* (1999) where the way of appropriately defining the sieves calls for the notion of metric entropy numbers in L^1 assigned to a subset $\mathcal{F}$ of L^1 . We denote $J(\delta, \mathcal{F})$ the logarithm of the minimum number of functions $f_1, \dots, f_k$ in $\mathcal{F}_\lambda$ such that $\mathcal{F} \subset \cup_{i=1}^k \{f : \|f - f_i\|_{L^1} < \delta\}$

[Ghosal *et al.* (1999)] Let $\theta_0 \in \mathcal{S}_K(\Pi)$; if for any $\epsilon > 0$ there exist $\delta < \epsilon$, $c_1, c_2 > 0$, $\beta < \epsilon/2$ and a sequence of subsets V_n such that eventually

- (i) $\Pi(V_n^c) < c_1 e^{-nc_2}$
- (ii) $J(\delta, V_n) < n\beta$

then the sequence of posterior distributions is d_H -consistent in θ_0 .

We remark that in the above result, depending of course from the chosen distance, consistency is ensured basically through additional conditions imposed on Π .

A deeper look at the results of this kind Ghosal *et al.* (1999) highlights that the basic role of V_n is to limit from above the functionals $H_n(\omega, \cdot)$ on $B(\theta_0)^c \cap V_n$. This fact drives back to conditions such as (C). In fact one can show (see also Perone Pacifico *et al.* (2000))

Proposition 2 If $\theta_0 \in \mathcal{S}_K(\Pi)$ and condition (C) holds then the sequence of posterior distributions is d -consistent.

Proof For any $\delta > 0$ let $N_\delta(\theta_0)$ the Kullback-Leibler neighborhood of θ_0 ; starting from the strong law of large number one can get (see for instance Barron *et al.* (1999)) that ν -almost surely

$$\int_{\Theta} \exp\{nH_n(\omega, \theta)\} \Pi(d\theta) < e^{-n\delta} \Pi(N_\delta(\theta_0)).$$

From (C), one has that ν -almost surely that for any $B(\theta_0)$ there is $\epsilon > 0$ such that $\sup_{\theta \in B(\theta_0)^c} H_n(\omega, \theta) < -\epsilon$ for n sufficiently large. Choosing $\delta < \epsilon$ one obtain:

$$\begin{aligned} \Pi_n(\omega, B(\theta_0)^c) &= \frac{\int_{B(\theta_0)^c} \exp\{nH_n(\omega, \theta)\} \Pi(d\theta)}{\int_{\Theta} \exp\{nH_n(\omega, \theta)\} \Pi(d\theta)} \\ &\leq e^{-n(\epsilon-\delta)} \frac{\Pi(B(\theta_0)^c)}{\Pi(N_\delta(\theta_0))} \end{aligned}$$

hence $\Pi_n(\omega, B(\theta_0)^c) \rightarrow 0$. ◇◇

4. Concluding remarks

The attempt of this work is to review some of the main approaches to get consistency in a nonparametric setting in order to provide useful hints on what conditions allow for a parallel behaviour of maximum likelihood and posterior consistency. This search for a joint consistency has been somewhat investigated, but never fully developed in a nonparametric setting. Besides the obvious point that both consistencies depends on the asymptotic behaviour of the likelihood, we noticed that the core of the problem can be reduced, as one can naturally expect, to its behaviour on the “tails”; condition (C) renders explicit and formal this intuitive point and it has been proved to establish consistency for both bayesian and maximum likelihood. A joint behaviour in terms of consistency can be also extended to the case of maximum likelihood approach based on *sieves* and that posterior consistency based on appropriately designed *sieves* as those in Barron *et al.* (1999), Ghosal *et al.* (1999) and Petrone and Wasserman (2002)). Along a similar line of extension we would like to mention a recent result in Walker and Hjort (2001). This work contains a general consistency result actually referred to a sequence of distributions $\tilde{\Pi}_{1/2}^n(\cdot)$ which are *not* posterior distributions, but clearly represent a formal analog since they are defined through the following ratio of integrals

$$\tilde{\Pi}_{1/2}^n(B) = \frac{\int_B (\prod_{i=1}^n f(X_i(\omega), \theta))^{\frac{1}{2}} \Pi(d\theta)}{\int_{\Theta} (\prod_{i=1}^n f(X_i(\omega), \theta))^{\frac{1}{2}} \Pi(d\theta)}.$$

The authors show that with the only hypothesis $\theta_0 \in \mathcal{S}_K(\Pi)$, the sequence of distributions $\tilde{\Pi}_{1/2}^n(B)$ are d_H -consistent. This result, interesting *per se*, hints at studying consistency properties of sequence of distributions that can be similar to the true posterior distribution, pointing out that in order to keep control of the likelihood growth a “little” penalization can make a difference. This line of research interestingly recall an analogue direction within the classical study of asymptotic properties of penalized or approximate maximum likelihood. In fact, the authors hints at the possibility of generalizing their result by using a likelihood function raised to a power $\alpha \in (0, 1)$ possibly close to 1. Formal definition and derivation of such results have been appeared in Scricciolo (2001). Even in this case we find an interesting connection suggesting a parallel study and ascertainment of maximum likelihood based on penalization and posterior consistency.

References

- Bahadur R.R. (1967) Rates of convergence of estimates and test statistics, *The Annals of Mathematical Statistics*, 38, 303–324.
- Barron A., Schervish M.J. and Wasserman L. (1999) The consistency of posterior distributions in nonparametric problems, *Ann. Statist.*, 27, 2, 536–561.
- Barron A.R. (1989) Uniformly powerful goodness of fit tests, *Ann. Statist.*, 17, 1, 107–124.
- Diaconis P. and Freedman D. (1986) On the consistency of Bayes estimates, *Ann. Statist.*, 14, 1, 1–67, with a discussion and a rejoinder by the authors.
- Dong M.X. and Wets R.J.B. (2000) Estimating density functions: a constrained maximum likelihood approach, *J. Nonparametr. Statist.*, 12, 4, 549–595.

- Doob J.L. (1949) Application of the theory of martingales, in: *Le Calcul des Probabilités et ses Applications.*, Centre National de la Recherche Scientifique, Paris, 23–27.
- Dudley R.M. (1998) Consistency of M -estimators and one-sided bracketing, in: *High dimensional probability (Oberwolfach, 1996)*, Birkhäuser, Basel, 33–58.
- Ghosal S., Ghosh J.K. and Ramamoorthi R.V. (1999) Consistency issues in Bayesian nonparametrics, in: *Asymptotics, nonparametrics, and time series*, Dekker, New York, 639–667.
- Hess C. (1996) Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator, *Ann. Statist.*, 24, 3, 1298–1315.
- Huber P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions, in: *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, Vol. I: *Statistics*, Univ. California Press, Berkeley, Calif., 221–233.
- Kiefer J. and Wolfowitz J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Statist.*, 27, 887–906.
- Le Cam L. (1990) Maximum likelihood: An introduction, *International Statistical Review*, 58, 153–171.
- LeCam L. (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates, *Univ. California Publ. Statist.*, 1, 277–329.
- LeCam L. (1973) Convergence of estimates under dimensionality restrictions, *Ann. Statist.*, 1, 38–53.
- Perlman M.D. (1972) On the strong consistency of approximate maximum likelihood estimators, in: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971)*, Vol. I: *Theory of statistics*, Univ. California Press, Berkeley, Calif., 263–281.
- Perone Pacifico M., Salinetti G. and Tardella L. (2000) Consistenza degli stimatori di massima verosimiglianza e della probabilità a posteriori: un approccio unificante, Technical report, Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma "La Sapienza", serie A, #2-2000.
- Petrone S. and Wasserman L. (2002) Consistency of bernstein polynomial posteriors, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64, 1, 79–100.
- Pollard D. (1984) *Convergence of stochastic processes*, Springer-Verlag, New York.
- Salinetti G. (1990) On structural relationships between probability and statistics: Empirical processes and statistical functionals (Italian), in: *Proceedings of the Italian Statistical Society*, 75–96.
- Schwartz L. (1965) On Bayes procedures, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4, 10–26.
- Scricciolo C. (2001) *Consistency and Rate of Convergence of a Sequence of Posterior Distribution in Non-Parametric Problems*, Ph.D. thesis, Università di Padova.
- van de Geer S. (1993) Hellinger-consistency of certain nonparametric maximum likelihood estimators, *Ann. Statist.*, 21, 1, 14–44.
- van der Vaart A.W. and Wellner J.A. (1996) *Weak convergence and empirical processes*, Springer-Verlag, New York, with applications to statistics.
- Wald A. (1949) Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statistics*, 20, 595–601.
- Walker S. and Hjort N.L. (2001) On Bayesian consistency, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63, 4, 811–821.