Journal of Statistical Planning and Inference 166 (2015) 67–77

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

About the non-asymptotic behaviour of Bayes estimators

Lucien Birgé *

Sorbonne Universités, UPMC Université Paris 06, U.M.R. 7599, L.P.M.A. F-75005, Paris, France

ARTICLE INFO

Article history:
Received 11 February 2014
Accepted 16 July 2014
Available online 30 October 2014

Keywords:
Bayes estimation
Concentration of posterior distributions
Risk of Bayes estimators

ABSTRACT

This paper investigates the *nonasymptotic* properties of Bayes procedures for estimating an unknown distribution from n i.i.d. observations. We assume that the prior is supported by a model $(\mathscr{S}, h)$ (where h denotes the Hellinger distance) with suitable metric properties involving the number of small balls that are needed to cover larger ones. We also require that the prior put enough probability on small balls.

We consider two different situations. The simplest case is the one of a parametric model containing the target density for which we show that the posterior concentrates around the true distribution at rate $1/\sqrt{n}$. In the general situation, we relax the parametric assumption and take into account a possible misspecification of the model. Provided that the Kullback–Leibler Information between the true distribution and the model $\mathscr{S}$ is finite, we establish risk bounds for the Bayes estimators.

1. Introduction

The purpose of this paper is to derive in a simple way some non-asymptotic results about posterior distributions and Bayes estimators from a frequentist viewpoint, therefore offering a complementary point of view to the classical results by Ghosal et al. (2000) – see also the related papers: Ghosal et al. (2003) and van der Vaart (2003). It can also be considered as a new and extended presentation of Le Cam (1973, 1982). In any case, it has been strongly influenced by these three papers.

We shall work here within the following framework: we have at disposal a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of size n, the X_i being measurable mappings from $(\Omega, \mathscr{A})$ to $(\mathscr{X}, \mathcal{X})$ with a common unknown distribution P. This distribution is an element of the metric space $(\mathscr{P}, h)$ of all probability measures on $(\mathscr{X}, \mathcal{X})$ endowed with the Hellinger distance h given by

$$h^2(R, T) = \frac{1}{2}\int \left(\sqrt{\frac{dR}{d\lambda}} - \sqrt{\frac{dT}{d\lambda}}\right)^2 d\lambda,$$

where λ is an arbitrary positive measure which dominates both R and T. We then introduce a model for P, i.e. a dominated family $\mathscr{S} = \{P_t \mid t \in S\} \subset \mathscr{P}$ of probabilities on $\mathscr{X}$ with densities $f_t = dP_t/d\mu$ with respect to some reference measure μ on $\mathscr{X}$. We assume that the mapping $t \mapsto P_t$ is one-to-one which allows us to systematically identify S and $\mathscr{S}$, thus considering S as a metric space with distance $h - h(t, u) = h(P_t, P_u)$ – and the corresponding Borel σ-algebra. We then introduce a prior distribution ν on S, turning the parameter t into a random variable $\boldsymbol{t}$. The prior ν and the sample $\boldsymbol{X}$ give rise to a posterior distribution $\overline{\nu} = \overline{\nu}(\cdot|\boldsymbol{X})$ and, given a loss function $w \circ h$ on $S \times S$, to a corresponding Bayes estimator $\widetilde{s}$ defined by

$$\widetilde{s}(\cdot|\boldsymbol{X}) = \operatorname*{argmin}_{u \in S} \mathbb{E}\left[w\left(h(u, \boldsymbol{t})\right)|\boldsymbol{X}\right] = \int_S w\left(h(u, \boldsymbol{t})\right) d\overline{\nu}(\boldsymbol{t}|\boldsymbol{X}), \tag{1.1}$$

* Correspondence to: UMR 7599 "Probabilités et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Paris 06, 4 Place Jussieu, F-75252 Paris Cedex 05, France.
E-mail address: lucien.birge@upmc.fr.

http://dx.doi.org/10.1016/j.jspi.2014.07.009
0378-3758/

where *argmin* refers to any minimizer in case it is not unique. In the sequel we shall write $\mathbb{E}_s[f(\boldsymbol{X})]$ to indicate that the X_i are i.i.d. with distribution P_s and $\mathbb{P}_s$ for the corresponding probability on Ω that gives $\boldsymbol{X}$ the distribution $P_s^{\otimes n}$.

Our purpose here will be twofold. When $P = P_s$ truly belongs to $\mathscr{S}$ and the metric structure of $(\mathscr{S}, h)$ is similar to that of a compact subset of some Euclidean space, we shall study the concentration rate of the posterior distribution $\overline{\nu}(\cdot|\boldsymbol{X})$ of $\boldsymbol{t}$ around P_s. When P does not belong to $\mathscr{S}$ or the metric structure of $\mathscr{S}$ does not follow the previous requirements, we shall study the performance of the Bayes estimator(s) $P_{\widehat{s}}$ of P defined via the loss function $w \circ h$ for suitable functions w. The main feature of our approach is its non-asymptotic viewpoint, explicit deviation bounds being provided for fixed n.

Some notations. To begin with, let us fix some notations to be used throughout the paper. In the metric space (S, h), we denote by $\mathcal{B}(t, r)$ the closed Hellinger ball with center $t \in S$ and radius r while the ball with center y and radius r in the Euclidean space $\mathbb{R}^d$ will be denoted $\mathcal{B}_d(y, r)$. The set of positive integers $\mathbb{N} \setminus \{0\}$ is denoted by $\mathbb{N}^*$, the cardinality of the set N by $|N|$ and we write $a \vee b$ for $\max\{a, b\}$. The distance between two sets A and B is $h(A, B) = \inf_{t\in A,\, u\in B} h(t, u)$ and if $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n$ we write $f_t(\boldsymbol{x})$ instead of $\prod_{i=1}^n f_t(x_i)$.

For any measurable subset B of S such that $\nu(B) > 0$, we define the density g_B with respect to $\mu^{\otimes n}$ and the probability P_B on $\mathscr{X}^n$ by

$$g_B(\boldsymbol{x}) = \frac{1}{\nu(B)} \int_B f_t(\boldsymbol{x})\, d\nu(t) \quad \text{and} \quad P_B = g_B \cdot \mu^{\otimes n}. \tag{1.2}$$

We denote by $\mathbb{P}_B$ the probability on Ω that gives $\boldsymbol{X}$ the distribution P_B.

2. A toy example

Let us first consider, in order to motivate our approach, the very particular situation of a finite or countable parameter set S containing the true density s to estimate. Besides, we shall assume that $\nu(t) > 0$ for all $t \in S$. The posterior probability is given in this case by

$$\overline{\nu}(B|\boldsymbol{X}) = \frac{\sum_{t\in B} \nu(t) f_t(\boldsymbol{X})}{\sum_{t\in S} \nu(t) f_t(\boldsymbol{X})} = \left(1 + \frac{\sum_{t\in B^c} \nu(t) f_t(\boldsymbol{X})}{\sum_{t\in B} \nu(t) f_t(\boldsymbol{X})}\right)^{-1} \geq 1 - \frac{\sum_{t\in B^c} \nu(t) f_t(\boldsymbol{X})}{\sum_{t\in B} \nu(t) f_t(\boldsymbol{X})}$$

for all $B \subset S$ and it follows that

$$\overline{\nu}(B|\boldsymbol{X}) \geq 1 - \frac{\sum_{t\in B^c} \nu(t) f_t(\boldsymbol{X})}{\nu(s) f_s(\boldsymbol{X})} \quad \text{for all } B \ni s. \tag{2.1}$$

In order to evaluate the concentration of the posterior distribution $\overline{\nu}(.|\boldsymbol{X})$ around s, we focus on those sets B_k which are Hellinger balls centered at s with radius $k/\sqrt{n}$, $k \in \mathbb{N}^*$. Bounding $\overline{\nu}(B_k|\boldsymbol{X})$ from below requires to bound from above ratios of the form $f_t(\boldsymbol{X})/f_s(\boldsymbol{X})$ when the X_i are distributed according to P_s, which implies that $f_s(\boldsymbol{X}) > 0$ a.s. This control derives from Lemma 7 in Birgé (2006) which implies the following inequality:

Lemma 1. *Given n i.i.d. random variables $X_1 \ldots X_n$ with distribution P and another distribution Q, then* $\log\left((dQ/dP)(X_i)\right) \in [-\infty, +\infty)$ *a.s. (with the convention* $\log 0 = -\infty$*) and, for all* $y \in \mathbb{R}$,

$$\mathbb{P}\left[\sum_{i=1}^n \log\left(\frac{dQ}{dP}(X_i)\right) \geq y\right] \leq \exp\left[-\frac{y}{2}\right] \rho^n(P, Q) \quad \textit{with } \rho(P, Q) = \int \sqrt{\frac{dP}{d\lambda}\frac{dQ}{d\lambda}}\, d\lambda.$$

We recall here that $\rho(P, Q)$ is called the *Hellinger affinity* between P and Q, the definition being independent of the choice of the dominating measure λ, and that it satisfies

$$\rho(P, Q) = 1 - h^2(P, Q) \quad \text{and} \quad \rho\left(P^{\otimes n}, Q^{\otimes n}\right) = \rho^n(P, Q) \leq \exp\left[-nh^2(P, Q)\right], \tag{2.2}$$

hence

$$h^2\left(P^{\otimes n}, Q^{\otimes n}\right) = 1 - \rho^n(P, Q) = 1 - \left(1 - h^2(P, Q)\right)^n \leq nh^2(P, Q). \tag{2.3}$$

We therefore derive from Lemma 1 and (2.2) that, for $\delta > 0$,

$$\mathbb{P}_s\left[f_t(\boldsymbol{X}) \geq \delta\nu(s) f_s(\boldsymbol{X})\right] \leq [\delta\nu(s)]^{-1/2} \rho^n(P_s, P_t) \leq [\delta\nu(s)]^{-1/2} \exp\left[-nh^2(P_s, P_t)\right].$$

Setting, for $k \in \mathbb{N}^*$,

$$\Gamma_k = \left\{\boldsymbol{x} \;\middle|\; \sup_{t\in B_k^c} f_t(\boldsymbol{x}) \geq \delta\nu(s) f_s(\boldsymbol{x})\right\}$$