Practical Bayesian Asymptotics

STEPHEN WALKER AND PAUL DAMIEN

University of Bath, UK, and University of Michigan Business School, USA

ABSTRACT. We discuss Bayesian consistency from a decision theoretic perspective. It turns out that if we are concerned with making consistent decisions then we only require the posterior distributions to be weakly consistent. We provide full motivation for our views.

KEYWORDS: Bayesian decision theory, consistency.

1. Introduction

In this paper, we discuss the relevance and importance of Bayesian consistency in the case involving *exchangeable* observations. There has recently been considerable activity in this area particularly in a nonparametric context (Barron et al, 1999; Ghosal et al., 1999). We wish to make our discussion as broad as possible, i.e. not restricted to finite dimensional parametric models, and hence we consider prior distributions defined on an arbitrary set of distributions, say $\mathcal{F}$, with topology induced by a metric d, and we assume the distributions in $\mathcal{F}$ have densities with respect to the Lebesgue measure.

Let Π be a prior distribution on $\mathcal{F}$, $y^n = (y_1, \ldots, y_n)$ a random sample from F_0 and Π^n the posterior distribution on $\mathcal{F}$ given y^n. We can differentiate between two types of consistency. Let d_W be a metric, such as the Prokhorov metric, which induces the weak topology.

DEFINITION 1. If $\Pi(A|y^n) \to 1$ a.s. as $n \to \infty$ for all weak neighbourhoods A of F_0, i.e. $A = \{F : d_W(F_0, F) \leq \varepsilon\}$ for all $\varepsilon > 0$, then Π is said to be weakly consistent at F_0.

Now let d_S be a metric which defines strong neighbourhoods of F_0; e.g. the Hellinger metric,

$$d_S(F_0, F) = \left\{ \int \left(\sqrt{\mathrm{d}F/\mathrm{d}y} - \sqrt{\mathrm{d}F_0/\mathrm{d}y} \right)^2 \mathrm{d}y \right\}^{1/2}.$$

DEFINITION 2. If $\Pi(A|y^n) \to 1$ a.s. as $n \to \infty$ for all strong neighbourhoods A of F_0, i.e. $A = \{F : d_S(F_0, F) \leq \varepsilon\}$ for all $\varepsilon > 0$, then Π is said to be strongly consistent at F_0.

Schwartz (1965) established the following result: if Π puts positive mass on all Kullback-Leibler neighbourhoods of F_0, which we now refer to as condition (A), then Π is weakly consistent at F_0.

The condition of Schwartz (1965) is not a necessary condition; see Wasserman (1998) who describes a counter-example first presented by Ghosal et al. (1997).

Additional sufficient, but not necessary, conditions on (A) for strong (Hellinger) consistency are provided by Barron et al. (1999) and Ghosal et al. (1999). A good review is provided by Wasserman (1998). The need for strong consistency is based on the fact that weak neighbourhoods of F_0 include distributions F with densities f that are far from f_0 with respect to a strong metric, such as the Hellinger metric. An example highlighting this phenomenon is provided by Barron et al. (1999).

While density estimation is an important problem in statistics (consider, for example, the huge amount of literature related to kernel density estimation), data collection usually involves a different endpoint – a decision. We will be discussing Bayesian consistency from this perspective. Before presenting our views on this, we collect together the elements of Bayesian decision theory.

2. Bayesian decision theory

We take our notation from Hirshleifer and Riley (1992). The elements of a Bayesian decision problem are as follows :

(1) a finite set of actions indexed by x and without loss of generality we assume $x \in \{1, \ldots, K\}$ for some integer K;

(2) a set of states of nature, which we take to be $\mathcal{F}$ equiped with the weak topology [1];

(3) a consequence function $c(x, F)$ showing outcomes under all combinations of actions and states of nature. We assume that $c(x, F)$ is uniformly continuous in F for all x;

[1] We assume that the relevant unknown state of nature is the distribution giving rise to the data. This gives us a general framework to work with. Certainly, knowing the true distribution will solve all decision problems connected to the data.

(4) a preference scaling function $v(c)$ measuring the desirability of the different consequences. We assume $v(\cdot)$ is uniformly continuous;

(5) a probability distribution on $\mathcal{F}$ representing beliefs in the true state of nature. In a Bayesian context this probability is the prior Π in the no sample problem and is Π^n once the data y^n has been observed.

The *expected-utility rule* then asserts that the best decision is to take the action x which maximises

$$U_n(x) = \int v\{c(x,F)\}\,\Pi^n(\mathrm{d}F).$$

This expected-utility rule is applicable if and only if the $v(\cdot)$ function has been determined in a particular way which leads to $v(c)$ being bounded, specifically $0 \leq v(c) \leq 1$. That is, the $v(c)$ are interpreted as probabilities and the $U_n(x)$ are then derived via the law of total probability. See Hirshleifer and Riley (1992) for further details. It is not our intention to discuss the expected-utility rule in this paper. Our aim, motivated by discussion in Section 3, is to provide sufficient conditions for which the rule is consistent. Since we assume $c(x,F)$ to be uniformly continuous in F for all x and $v(\cdot)$ is bounded and uniformly continuous, it follows that $v_x(F) = v\{c(x,F)\}$ is bounded and uniformly continuous in F for all x.

3. Consistent decisions

Bayesian decision theory can be demonstrated to be the logical consequence of a set of axioms of rationality and the subjective interpretation of probability. These axioms and the derivation of the consequences leading to Bayesian decision theory are presented in great detail in Bernardo and Smith (1994). However, the sound foundations for Bayesian decision theory do not necessarily imply that the resulting practice is good. The theory implies the existence of a prior Π on the states of nature, by acknowledging dealing with uncertainty via probability, but says nothing about what Π is or what it should do.

If the Bayesian intellectual process is to be credible, the corresponding Bayesian machinery must work. By the latter we mean that as more data accumulate the quality of the decisions made must improve. Common sense dictates this; if not why collect any data at all? That is, as the sample size

increases to infinity, we must require that eventually the correct decision is made. It does not look good in principle if, even with an infinite sample. equivalently an infinite amount of information, we can not make the correct decision. This position is not incompatible with the subjectivist Bayesian viewpoint (to which we adhere). We are merely imposing further (sensible) conditions on the prior above and beyond those required by standard Bayesian theory. We differ from researchers insisting on posterior consistency *per se* since our position is that data is collected to make decisions and we need to ensure our decisions are consistent. Hence, we are interested in the appropriate mathematical conditions under which this is achieved.

The correct decision is made if F_0 is known and we can evaluate $U_0(x) = v\{c(x, F_0)\}$ for each x and hence select the action x which maximises $U_0(x)$. For the correct decision to be made eventually, we require

$$\mathrm{pr}_{y^n}(x_n \neq x_0 \text{ i.o.}) = 0$$

where x_0 is the correct action, i.e. $U_0(x_0) \geq U_0(x)$ for all x. While the theory does not require us to make consistent decisions, it is a useful condition to impose so that a good and theoretically sound decision making procedure emerges.

A sufficient condition for the desired result to hold is that

$$\max_x\{U_n(x)\} \to \max_x\{U_0(x)\} \text{ a.s.}$$

which, since x indexes a finite set, holds if

$$U_n(x) \to U_0(x) \text{ a.s.}$$

for all x. Now

$$U_n(x) = \int v_x(F)\,\Pi^n(\mathrm{d}F)$$

and if Π satisfies condition (A), that is Π^n converges weakly to Π_0, the distribution with point mass at F_0, then the Portmanteau theorem (see, for example, Billingsley, 1968, Theorem 2.1), gives the desired convergence result for $U_n(x)$. This also requires that $v_x(F)$ be bounded and uniformly continuous, which we have assumed.

Hence, provided a prior Π satisfies condition (A), a decision maker following Bayesian decision theory is *guaranteed* making the correct decision as more samples, i.e. as more information, is accumulated.

4. Discussion

Bayesian consistency is extremely important. However, we would also argue that since decision making (equivalently inference) is at the heart of statistical practice, it is sufficient that posteriors are weakly consistent rather than insisting they are strongly consistent.

Our recommendation to decision makers in the exchangeable case is to construct priors which do put positive mass on all Kullback-Leibler neighbourhoods of F_0. This is currently the simplest sufficient condition which gives weak consistency. Barron et al. (1999) and Ghosal et al. (1999) illustrate how to achieve this for a number of nonparametric priors, including Dirichlet mixtures, Pólya-trees. infinite dimensional exponential family, among others.

Unfortunately, unless it is known that F_0 does belong to some finite dimensional parametric family, it is not possible to ensure the Kullback-Leibler property using a prior restricted to some parametric family. Hence, we are also advocating the use of nonparametric priors in the decision making process; indeed it appears to be the only recourse. Using and interpreting such priors is now becoming routine (Walker et al., 1999).

References

BARRON, A., SCHERVISH, M.J. and WASSERMAN, L. (1999). The consistency of distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.

BERNARDO, J.M. AND SMITH, A.F.M. (1994). *Bayesian Theory.* Wiley & Sons.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley & Sons.

GHOSAL, S., GHOSH, J.K. and RAMAMOORTHI, R.V. (1997). Consistency issues in Bayesian nonparametrics. Unpublished.

GHOSAL, S., GHOSH, J.K. and RAMAMOORTHI, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–158.

HIRSHLEIFER, J. AND RILEY, J.G. (1992). *The Analysis of Uncertainty and Information.* Cambridge University Press.

SCHWARTZ, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete.* **4**, 10–26.

WALKER, S.G., DAMIEN P., LAUD, P.W. AND SMITH, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society Series B* **61**, 485–527.

WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.), 293–304. *Lecture Notes in Statistics*, Springer.