On Bayesian Models and Consistency

STEPHEN WALKER* AND PAUL DAMIEN**

*Department of Mathematical Sciences, University of Bath

**University of Michigan Business School

ABSTRACT. We discuss the relevance of consistency to the Bayesian. Should consistency be dismissed as irrelevant or thought about seriously when constructing prior distributions? Strong opinions have been held on this matter, but it is probably fair to say it is a largely neglected area. Pioneers, such as de Finetti, Savage, Lindley, have had very little to say on the matter. The aim of this paper is to give specific reasons why Bayesians should be concerned with consistency and exactly what type of consistency they should be concerned with, if the goal is to make rational and good decisions. We also discuss the notion of a true model and define useful connections with consistency.

KEYWORDS: Decision theory, exchangeability, expected utility rule, true model.

1. Introduction

While it is acknowledged that consistency is crucially important in Classical statistics, for example, providing justification for many estimators and procedures, no such consensus exists in the domain of Bayesian statistics. This is, we would suggest, due to an understanding that Bayesian decision theory/inference are the logical consequences of a set of axioms of rationality and a subjective interpretation of probability (see, for example, Bernardo and Smith, 1994). Hence, Bayesian procedures are logical consequences of an axiomatic system and so, apparently, asymptotic studies of such procedures is not required. Large sample studies have generally focused on the mathematical aspects (Freedman, 1963; Schwartz, 1965; Berk, 1966, 1970) with no practical implications for Bayesian inference being made as a consequence

of the results. That is, in a traditional Bayesian analysis, consistency is not typically, if ever, considered.

Procedural components required to be specified in order to carry out a Bayesian analysis of observed data are undefined, it is only their existence which is guaranteed. The prior distribution is thought to be the most problematic to specify in practice; the existence of which is guaranteed by the Representation Theorem of de Finetti (1938), assuming exchangeable observations; see, also, Hewitt and Savage (1955).

Since we do only have an existence theorem, it might seem appropriate to think hard about the specification of the prior and in particular how it should be constructed. This, perhaps, could then lead to procedural principles that might further solidify the Bayesian paradigm.

The question which this paper is interested in is the following: Should Bayesians construct priors so that inference is *consistent* in some sense? Clearly it is not incorrect or wrong to do this. But is it sensible? Does it matter?

A first point to make is that consistency (in whatever form it takes) is not an accident. It has to be mathematically designed. This imposes conditions on a prior which might be difficult to establish in specific cases; see, for example, Schwartz (1965); Barron et al. (1999); Ghosal et al. (1999). Another point to make is whether a procedure which is the logical consequence of a desire for rational behaviour implies it is a *good* procedure. Clearly, an existence theorem for the prior means precisely and just that – anything goes. The prior could be a single point mass on some unusual density, which is clearly *not good*. Making use of relevant prior information would (hopefully) eliminate this impractical prior from consideration, but the point is made: the theory does not automatically lead to a good procedure. More effort is required.

In this paper, we endeavour to give reasons why a particular type of consistency is important and to discuss the consequences of this. In particular, we concentrate on what we believe to be a practical perspective and hence consider the notion of consistency with respect to Bayesian decision theory.

Due to the fact that inference is based entirely on models, we start by looking at Bayesian models and in particular *true* Bayesian models. This is done in Section 2, where the notion of a true Bayesian model is discussed. In Section 3, we propose a general definition of a true Bayesian model in terms of consistency, claiming this a useful definition and in Section 3.1 consider

the ramifications of this definition. Section 4 then considers which type of consistency is required by considering Bayesian decision theory, acknowledging the fact that data collection leads to decisions having to be made. Section 5 extends on work presented in Section 4, by considering infinite rather than finite decision spaces.

2. True models

A survey of the Bayesian literature will involve meeting the term "true model" with high frequency. This is particularly apparent in the Bayesian model selection and model averaging literature. For example, in Raftery et al. (1996) it is stated; "A typical approach to data analysis is to carry out a model selection exercise leading to a single 'best' model and to then make inference as if the selected model were the true model". Then later on in the paper; "... $\mathrm{pr}(M_k)$ is the prior probability that M_k is the true model."

The following is taken from Key et al. (1999); "... the M-closed view, corresponds to believing that one of the models ... is 'true', but which one is unknown." Section 4 of Key et al. (1999) is entitled : "Choosing among models when none of them is true". The discussion concluding Key et al. (1999) brings into question the existence of a true model. M.J.Bayarri states "In a sense, an honest Bayesian can only agree with the main theme of this paper (i.e. Key et al., 1999), namely that (i) no model is true ...". This remark is essentially attributable to Savage who said "no model is true, some models are useful." See also Box (1980) who writes "No statistical model can safely be assumed adequate".

It is with these latter sentiments that we wish to disagree. None of the participants in the "true model" discussions define exactly what they mean by such a phrase. In this paper, we consider the case when f_0 is a density and $y^n = (y_1, \ldots, y_n)$ are available as a random sample from f_0, the first n observations of a possibly "infinite" sequence $y_1, y_2, \ldots$; Key et al. (1999) refer to this as the *exchangeable* case. It is quite likely that the above commentators were referring to more complex data structures than the exchangeable case, though it was never mentioned explicitly in their comments. Nevertheless, discussing the concept of a true model is valid in the exchangeable case and is obviously the correct place to start since it is the most minimal assumption that can be made to develop Bayesian theory.

We examine when a model is or is not the true model with respect to f_0. There seems to be no definition of exactly what a true Bayesian model is. It might be assumed that a precise definition would be as follows :

DEFINITION 1. If $M = \{f(\cdot\,;\theta),\ \pi(\theta)\}$ and $f_0(\cdot) \equiv f(\cdot\,;\theta_0)$ and θ_0 is in the support of the prior then M is a true Bayesian model.[1]

We use the phrase 'is a true Bayesian model' since there are, for example, many $\pi(\theta)$ which will have θ_0 in its support.

The main point to mention here is that if the true model is parametric (i.e. finite dimensional) then Doob (1949) proves posterior consistency holds, in the sense that the posterior distribution accumulates about θ_0 as the sample size tends to infinity, for almost all $\theta_0 - \pi$.

We would argue that without the result of Doob (1949), Definition 2 would be hard to substantiate as defining a true model, since without consistency it implies nothing favourable about using the model in preference to any other model.

Perhaps the feeling that no true model exists, as indicated by Bayarri, Savage, Box etc., stems from the problems associated with a parametric model. Of course, how can it really be known that a θ_0 exists? It cannot. To be sure that the true density does lie in the support of a prior we need to consider bigger, or nonparametric models.

Consequently, a general definition of a true model should also allow for θ being an infinite dimensional parameter, even a density or a distribution. In these nonparametric cases, the support of the prior depends on the topology or metric used to define distances between densities or distributions. Suppose the metric is $d(\theta, \theta')$.

DEFINITION 2. The support of a prior π is defined to be

$$S_\pi = \{\theta : \pi(N_\varepsilon(\theta)) > 0 \text{ for all } \varepsilon > 0\},$$

where $N_\varepsilon(\theta) = \{\theta' : d(\theta, \theta') \leq \varepsilon\}$.

For example, if θ is a density with respect to the Lebesgue measure, then a

[1] This version of M suggests a Bayesian model has two components, a parametric family with prior. We prefer, and this will be apparent later on, to think of M as a probability on a subset of all densities, the subset being a parametric family. This then extends readily to the idea of a nonparametric prior probability on the set of all densities.

possible choice of metric is the Hellinger distance:

$$d(\theta, \theta') = \left\{ \int \left(\sqrt{\theta} - \sqrt{\theta'} \right)^2 \right\}^{1/2}.$$

Another possibility is a metric which metrizes weak convergence, such as the Prokhorov metric. See, for example, Billingsley (1968).
Thus, a nonparametric version of Definition 1 would be as follows:

DEFINITION 1 (revised). If $M = \{\pi(\mathrm{d}\theta)\}$ and θ_0 lies in the support of the prior with respect to the metric d, then M is a true Bayesian model.

Now π is a probability measure on densities or distributions, rather than a density as in the parametric case. When θ is a density the likelihood function is defined to be

$$\mathrm{L}_n(\theta) = \prod_{i=1}^{n} \theta(y_i)$$

and the posterior distribution is characterised via

$$\pi(A|y^n) \propto \int_A \mathrm{L}_n(\theta)\, \pi(\mathrm{d}\theta).$$

This supports the notion that a Bayesian model is indeed just a probability on sets of densities rather than a likelihood and prior combination.

There are special issues to consider when θ is a random distribution function which do not admit densities with respect to any dominating measure. Such a case in point is the Dirichlet process – random distributions with random mass allocated to random points – formally introduced by Ferguson (1973). Then alternative procedures need to be developed to obtain the posterior distributions.

Definition 1 is now actually incomplete as a definition since it says nothing about d. It is our intention to convince the reader that Definition 1 is inappropriate as a definition of a true Bayesian model, even if d has been specified, and to propose and support an alternative definition. Whereas Definition 1 may well be appropriate for finite dimensional cases, it does not sit well in nonparametric cases. The key is that Definition 1 implies consistency in the parametric finite dimensional cases and hence the notion of a true model. However, Definition 1 does not imply consistency in infinite dimensional cases, even if, for example, d is a strong metric.

Our aim is to provide a definition of a true Bayesian model, whether we decide to put all the mass on all densities (and call ourself a nonparametric Bayesian) or restrict the mass to parametric families of densities (and call ourself a parametric Bayesian). The consideration of nonparametric priors is warranted due to the recent explosion in their use. See, for example, Dey et al. (1998) and Walker et al. (1999).

3. Defining a true Bayesian model

Suppose that $M = \{\pi(\mathrm{d}\theta)\}$ and $\pi(A_\varepsilon|y^n) \to 1$ with probability 1 as $n \to \infty$ for all $\varepsilon > 0$, where $\pi(\cdot\,|y^n)$ is the posterior based on a sample of size n and $A_\varepsilon = \{\theta : d(\theta_0, \theta) \leq \varepsilon\}$. If this holds for M then it is clear that M is a true Bayesian model with respect to the metric d. It would be hard to argue that M was not a true model if it possessed the property of consistency. Hence we are saying a consistent model is a true model.

Here problems of using Definition 1 for defining a true Bayesian model are encountered. A model which satisfies the conditions of Definition 1 is not necessarily consistent. Counter-examples exist. One such counter-example is presented in Barron et al. (1999). Their prior is nonparametric and although the true density $f_0 \equiv \mathrm{Un}(0,1)$ lies within the (Kullback-Leibler) support of the prior, the posterior is demonstrably not consistent, with repsect to the Hellinger distance. It is however weakly consistent. But even weak consistency can cause problems. Diaconis and Freedman (1986) present an example which illustrates that even if a model has the true distribution in the weak support, weak consistency is not guaranteed.

The upshot is that Definition 1 includes models which, not being consistent in any sense, would be difficult to justify as true. The *true* in true model should mean just that; the true f_0 or F_0 is available, which is not the case unless the model is consistent with respect to a suitable metric. Without the property of consistency, there is nothing to distinguish; i.e., to set apart, a model which satisfies the conditions of Definition 1 and is not consistent with any other model. That is, the property of having the "truth" in the support counts for nothing unless it is acting as (part of) the sufficient condition for consistency. Hence, it is argued that a true model must be consistent and we have already indicated that a consistent model is a true model, leading to :

DEFINITION 3. A Bayesian model $M = \{\pi(\mathrm{d}\theta)\}$ is said to be true with

respect to the metric d, if and only if

$$\pi(A_\varepsilon|y^n) \to 1 \text{ with probability 1 for all } \varepsilon > 0,$$

where $A_\varepsilon = \{\theta : d(\theta_0, \theta) \leq \varepsilon\}$.

In the next sub-section we discuss the ramifications of using Definition 3 as defining a true model, in the exchangeable case.

3.1 Consequences of Definition 3

The fundamental assumption adopted here is that a true model is more desirable to work with than any other model, so long as all other important aspects are approximately equal. [2]

We mainly concentrate on the consequences of Definition 3 for model selection. The first point to make is that with careful construction of a (non-parametric) model it is possible to ensure the prior is consistent, i.e. the posteriors are consistent with respect to either the Hellinger or Prokhorov metrics. Sufficient conditions have been given by both Barron et al. (1999) and Ghosal et al. (1999) for the former and Schwartz (1965) for the latter. Additionally, Barron et al. (1999) and Ghosal et al. (1999) work out these sufficient conditions for specific priors; including infinite-dimensional exponential family, Pólya trees, Dirichlet mixture models, and histograms. With the sufficient conditions, these models then meet the requirements for being called a true Bayesian model.

Consequently, understanding consistency, the notion of M-closed and M-open views, as discussed by Bernardo and Smith (1994), lose importance. That is, it is possible to force the M-closed view with a single model. If there are two models, one consistent, the other not, all else being equal in the prior information sense, we would obviously select to work solely with the consistent model. There is no good reason to use the non-consistent model. Hence model selection will only be necessary when all the models under consideration have not, for whatever reason, been constructed to be consistent. Obviously, the comments and consequences relating to model selection are equally valid for model averaging (Draper, 1995). That is, knowing a true model eliminates the need for model averaging.

[2] Two priors, say π_1 and π_2, can be considered equal in a prior information sense if, for example, $\mathrm{E}_{\pi_1}(A) = \mathrm{E}_{\pi_2}(A)$ and $\mathrm{var}_{\pi_1}(A) = \mathrm{var}_{\pi_2}(A)$ for all A.

Having said all of this, constructing and making inference from a true model may be inconvenient and inappropriate. For example, a simpler (proxy) model might be needed to explain the statistial procedure to a non-expert or for ease of inference. In this case a number of simple models can be considered and the 'best' one used. However, the benchmark model is a true model and consequently the best model should be chosen which is closest in some sense to a true model. Walker and Gutiérrez-Peña (1999) implicitly advocate choosing the model which has the closest predictive density in the sense of Kullback Leibler divergence to the predictive density of a true model.

Statements highlighted at the outset of the paper, i.e. no model is true etc. are incorrect. A consistent model, attainable (at least) in the exchangeable case, is a true model. A question from Key et al. (1999) goes: "But when does it actually make sense to speak of a 'true' model and hence to adopt the M-closed perspective?" The answer, we believe, is when a consistent model has been constructed.

In summary, we can achieve true (consistent) models. On the assumption of exchangeability, de Finetti's celebrated Representation Theorem implies the existence of a prior π and a procedure for updating/predicting based on observations. That is,

$$\mathrm{pr}\{y_{n+1} \in B | y^n\} = E_{y^n}\left\{\int_B f(y)\,\mathrm{d}y\right\} = \int\left\{\int_B f(y)\,\mathrm{d}y\right\}\pi(\mathrm{d}f|y^n).$$

This, however, does not imply it is a "good" procedure in any sense. Consistency is one way to define a good procedure in the sense implied above. It extends on the existence theorem to provide a criterion for the selection of π. The key is this: if an arbitrary parametric model is chosen and a prior constructed, it is no great effort to incorporate certain defining characteristics from this parametric model into a nonparametric one, with the merit that the nonparametric model, with the regularity conditions for consistency, can then justifiably be called 'true'.
The question left for us to consider is which type of consistency is required.

4. Which kind of consistency?

Clearly the type of consistency depends on the problem at hand which will be determined by the reason the data was collected in the first place. Density

estimation is an important problem in statistics and if this is the aim then priors which lead to strong consistency will be needed. We would argue that data collection is usually motivated by a decision problem and we discuss the type of consistency required from this perspective. We wish to make this discussion as broad as possible, i.e. not restricted to parametric models, and hence we consider prior distributions defined on an arbitrary set of distributions, say $\mathcal{F}$, with topology induced by the 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, as has already been pointed out, distinguish between two types of consistency. Let d_W be a metric, such as the Prokhorov metric, which metrizes weak convergence, i.e. F_k converges weakly to F iff $d_W(F_k, F) \to 0$. In the following notation, we replace θ with the distribution function F.

DEFINITION 4. 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 any $\varepsilon > 0$, 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 5. 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 any $\varepsilon > 0$, 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 in addition to condition (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 apparent need for strong consistency is based on the fact that

weak neighbourhoods of F_0 include distributions F with densities 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). It should be noted that the (current) extra conditions for strong consistency are quite strict and could clash with prior information, i.e. it is difficult to both insist on strong consistency and incorporate realistic prior information. An example of this is Pólya trees.

Next, we review the elements of Bayesian decision theory.

4.1 Bayesian decision theory

Taking notation from Hirshleifer and Riley (1992), the elements of a decision problem are as follows :

(1) a finite set of actions indexed by x and for practical purposes we assume $x \in \{1, \ldots, X\}$ for some integer X. While most theory is associated with finite decision spaces (Raiffa, 1970; Lindley, 1985), a relaxation of this assumption to a non-finite decision space will be discussed in Section 5;

(2) a set of states of nature, which we take to be $\mathcal{F}$ equiped with the weak topology; [3]

(3) a consequence function $c(x, F)$ showing outcomes under all combinations of actions and states of nature.

(4) a preference scaling function $v(c)$ measuring the desirability of the consequence c;

(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.

We assume that $v\{c(x, F)\}$ is uniformly continuous in F for each x. This makes sense since small changes in F should result in small changes to our elementary utility $v(\cdot)$.

[3] 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 associated with the data.

The Von Neumann–Morgenstern (1947) *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 $v(c)$ has a probabilistic interpretation. Hirshleifer and Riley (1992) say:

> It turns out that the expected-utility rule is applicable *if and only if the $v(c)$ function has been determined in a particular way that has been termed the assignment of "cardinal" utilities to consequences.*

They go on to say:

> To formally justify the joint use of a cardinal preference scaling function and the expected-utility rule, for dealing with choices among risky prospects, involves a somewhat higher order of technical difficulty. What follows ... how the required type of preference scaling function can be developed. ... The essential point is that the $v(c)$ measure obtained via the reference lottery technique is in the form of a *probability*, so that the expected-utility rule becomes equivalent to the standard formula for compounding probabilities.

See Hirshleifer and Riley (1992) for further details. There are differing opinions on the point of a bounded elementary utility function. De Groot (1970) states:

> In many axiomatic developments of the theory of utility, assumptions are made which are stronger in certain respects than those which have been made here [De Groot, 1970]. These strengthened assumptions [Von Neumann-Morgenstern, 1947] make it possible to conclude that the utility function must be a bounded function.

It therefore depends on which set of axioms; i.e., the strength of them, the experimenter is willing to adhere to. We would adhere to those of Von Neumann

and Morgenstern (1947), summarised in Hirshleifer and Riley (1992), which lead to bounded utilities since with unbounded utilities, i.e. unbounded $v(\cdot)$, it is not guaranteed that $U_n(x)$ even exists. Berger (1985) effectively works with bounded utilities; he only considers lower bounded loss functions, i.e. $l(x, F) > -K > -\infty$ and we can consider $v_x(F) = K - l(x, F)$.

It is not our intention to discuss the expected utility rule further in this paper. Our aim, motivated by discussion in Section 4.2, is to provide sufficient conditions for which the rule is consistent. Note that, since we assume $c(x, F)$ to be uniformly continuous in F for all x and $v(\cdot)$ is bounded and uniformly continuous, $v_x(F) = v\{c(x, F)\}$ is bounded and uniformly continuous in F for all x.

4.2 Consistent decisions

Although we did not cover the ground; i.e., present the axioms and derivations, in the previous subsection, the point is that Bayesian decision theory is the logical consequence of axioms of rationality and the subjective interpretation of probability. The goal is to maximise expected utility. However, good foundations do not necessarily lead to good procedures. As more data accumulate, it is essential that the quality of decisions improve. The theory does not touch on this area. However, it is common sense that with large samples, correct decisions (defined later) must be made. How could it ever be argued that collecting more information (i.e. data) is worthwhile unless such an asymptotic property exists.

To highlight the above point, let us suppose there are two actions, one of which must be taken. If we do not have consistency then the rule will (randomly) point to one action or the other *ad infinitum* as the sample size increases, or, alternatively, eventually sticks to the incorrect decision. In this latter case, a decision maker is going to reject an infinite and free no sampling cost sample, which is clearly ridiculous; that is, at least if he/she guesses at the start (i.e. $n = 0$), he/she has a 50% chance of getting it right. An experimenter of any type should never be in a position where it is preferable to reject an infinite and free sample rather than accepting it. Although we accept that all decisions *will* be made with finite samples, the methodology used to perform this task should be such that it would be desirable to accept an infinite and free sample.

This ties up with the notion of *perfect information*, and is connected to

the idea of deciding on how many samples to collect; i.e., the size of n (when sampling costs money). The following is a quote from Raiffa and Schlaifer (1961):

> Let us imagine an ideal experiment e_∞ with a known cost c_∞ which is capable of yielding exact or perfect information concerning the true state of F and let us suppose that the decision maker wishes to choose between e_∞ and the null experiment (no sampling).

This decision problem only makes sense if we do indeed have consistency. The discussion for the number of samples to collect must be done under the fundamental assumption that an infinite sample does indeed lead to perfect information, i.e. the correct decision. Otherwise, there appears no reason for sampling at all.

It is also a matter of inspiring confidence. A procedure for making decisions must inspire confidence in those paying for experiments, collecting the data and then handing it over to the expert decision maker (i.e. the statistician). If the decision maker with his/her procedure can not guarantee good decisions with large samples then this is not going to inspire confidence. It is not good enough to say procedure A will be implemented for small samples and procedure B for large samples. If procedure B is good for large samples; i.e., leads to correct decisions, why is it not good for small samples? If A is good for small samples, why is it not good for large samples? And what precisely is a small/large sample? As we have mentioned before, we can adequately incorporate both prior information and consistency into a single prior which sorts out both large and small sample desirabilities of a prior (or procedure).

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 $U_0(x)$. As the sample size increases to infinity, at some point we require the correct decision to be made. For this we need

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

where x_0 is the correct action, i.e. $U_0(x_0) \geq U_0(x)$ for all x and x_n maximises $U_n(x)$. A sufficient condition for this 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), i.e., π^n converges weakly to π_0 a.s. – the probability with point mass at F_0 – then from the Portmanteau theorem (see, for example, Billingsley, 1968, Theorem 2.1), the desired convergence result for $U_n(x)$ holds. Hence, provided a prior satisfies condition (A), a decision maker is guaranteed making the correct decision as more samples; i.e., as more information, is accumulated.

We would argue, therefore, that Bayesian consistency is important. We do need decisions to be correct as the sample size goes to infinity. Therefore, it is sufficient that models are weakly consistent rather than insisting that they are strongly consistent. This means that provided a prior puts positive mass on all Kullback Leibler neighbourhoods of the true distribution, a Bayesian ends up with the right decision. To ensure the Kullback–Leibler property, it may well be necessary to use a large or nonparametric prior, since obviously the true distribution is unknown.

In summary, we propose that priors are constructed to ensure the Kullback-Leibler property. It can be done, and this ensures consistent decisions are made. To guarantee the Kullback–Leibler property it may be required to use nonparametric priors. Indeed, this is necessary to avoid the problem of a decision maker who knows that decisions made with a free and infinite sample will either be random or wrong and hence is in the undesirable situation of having to turn down such a sample.

4.3 Objective prior and rates of convergence

A particular type of utility arises if we think of the decision problem having y_{n+1} as the unknown state of nature, having witnessed y^n. Then $\mathrm{pr}(y_{n+1} \in A|y^n) = F^n(A)$, where $F^n = \int F\, \pi^n(\mathrm{d}F)$. Then we would have

$$U_n(x) = \int v_x(y)\, \mathrm{d}F^n(y) = \int \left\{ \int v_x(y)\, \mathrm{d}F(y) \right\} \pi^n(\mathrm{d}F)$$

and hence previous theory applies with $v_x(F) = \int v_x(y)\, \mathrm{d}F(y)$. A so-called *objective* prior may well be warranted in many contexts and one is available

using a Dirichlet process prior for F (Ferguson, 1973) with diffuse base measure; effectively this is equivalent to using the Baysian bootstrap (Rubin, 1981). Then

$$U_n(x) = n^{-1} \sum_{i=1}^{n} v_x(y_i)$$

which clearly only depends on the data and the choices of $v_x(\cdot)$. [4]

In this case we can establish rates of convergence of U_n to U_0 (dropping subscript x). We make use of a well known result of Hoeffding (1963) which gives

$$\text{pr}\{|Z_1 + \cdots + Z_n| \geq \eta\} \leq 2\exp\left\{\frac{-2\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\},$$

where the $\{Z_i\}$ are independent, $\text{E}Z_i = 0$ and $a_i \leq Z_i \leq b_i$ for each i. In our situation we have $Z_i = v(y_i) - \int v(y)\,\text{d}F_0(y)$ so that $-1 \leq Z_i \leq +1$ and so

$$\text{pr}\{|U_n - U_0| \geq \varepsilon\} \leq 2\exp\{-\tfrac{1}{2}\varepsilon^2 n\}.$$

The Borel-Cantelli theorem confirms (what we know) the a.s. convergence of U_n to U_0. The result is useful from a practical perspective. If we require $\text{pr}\{|U_n - U_0| \geq \varepsilon\} \leq \delta$ then we require a sample of size n to be at least $-2\log(\delta/2)/\varepsilon^2$.

5. Infinite decision spaces

In this section we assume that the decision space, say Ω, is not necessarily finite, and we let m be a metric on Ω. Of course, with an infinite set Ω, consistent decisions are no longer guaranteed, using the theory developed in Section 4. First, we keep to (Ω, m) being compact. We will now go through the mathematics in detail (dropping a.s. from the following). We will need an *equicontinuity* condition for $\{U_n(x)\}$.

DEFINITION 6. $\{U_n(x)\}$ is equicontinuous at x if for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $m(x, x') < \delta$ implies $|U_n(x) - U_n(x')| < \varepsilon$ for all n.

To ensure this, we require a condition on $v_x(F)$ of the type

$$|v_x(F) - v_{x'}(F)| \leq K\, m(x, x') \text{ for all } F$$

[4] All priors are subjective; some are more data-dependent than others. At times, authors have gone through considerable pain to construct families of objective priors, usually in the parametric framework. Little work has been done on objective nonparametric priors.

for some $K < \infty$. Then it is easy to see that

$$|U_n(x) - U_n(x')| \leq K\, m(x, x')$$

and hence we have equicontinuity for $\{U_n(x)\}$.

From our conditions on Π, we know that we have the pointwise convergence of $U_n \to U_0$. A well known theorem from Analysis is useful here:

THEOREM 1. If $U_n \to U_0$ pointwise for all $x \in \Omega$, Ω is compact and $\{U_n\}$ is equicontinuous at each $x \in \Omega$, then $U_n \to U_0$ uniformly, i.e. $\sup_{x\in\Omega}|U_n(x) - U_0(x)| \to 0$.

Consequently, condition (A) implies pointwise convergence, the compactness and equicontinuity provide uniform convergence.

THEOREM 2. Uniform convergence implies consistent decisions.

PROOF. Suppose there exists a subsequence n' such that $|x_{n'} - x_0| > \varepsilon$ where x_n maximises U_n and x_0 maximises U_0. Since $\{x_{n'}\}$ lie in a compact space there exists a sub-subsequence n'' and an α such that $|x_{n''} - \alpha| \to 0$, i.e. $m(x_{n''}, \alpha) \to 0$. Equicontinuity implies

$$|U_{n''}(x_{n''}) - U_{n''}(\alpha)| \to 0$$

and since

$$|U_{n''}(\alpha) - U_0(\alpha)| \to 0$$

we must have

$$|U_{n''}(x_{n''}) - U_0(\alpha)| \to 0.$$

Since $\alpha \neq x_0$, for all large n'', $U_{n''}(x_{n''}) < U_0(x_0) - \eta$ for some $\eta > 0$. However, we know that $U_{n''}(x_0) \to U_0(x_0)$ and hence we encounter a contradiction, indicating that $x_n \to x_0$. That is, $\sup_{x\in\Omega}U_n(x) \to \sup_{x\in\Omega}U_0(x)$ and hence we are again making consistent decisions.

Let us recall the type of decision and "objective" prior considered in Section 4.3. We now wish to consider

$$U_n(x) = n^{-1}\sum_{i=1}^{n} v_x(y_i).$$

For equicontinuity we require $|v_x(y) - v_{x'}(y)| < Km(x, x')$ for all y. For consistency in a non-compact space, we require a uniform strong law of large numbers; i.e.,

$$\sup_{x\in\Omega}|U_n(x) - U_0(x)| \to 0 \text{ a.s.}$$

Following Pollard (1984, Def. 23) let us define $N_1(\varepsilon, F^n, \mathcal{V})$, where F^n is the empirical distribution function and $\mathcal{V}$ is the set of functions $\{v_x : x \in \Omega\}$, as the smallest integer $m(n, \varepsilon)$ for which there exists functions $\{g_1, \ldots, g_{m(n,\varepsilon)}\}$ such that

$$\min_j \int |v - g_j| \mathrm{d}F^n \leq \varepsilon$$

for each $v \in \mathcal{V}$. Then Pollard (1984, Theorem 24) provides the required result if $\log m(n, \varepsilon) = o_p(n)$ for each $\varepsilon > 0$. Rates of convergence are also discussed in Pollard (1984).

Let us summarise what we have in general terms. If (Ω, m) is compact and we construct $v(\cdot)$ and $c(\cdot, \cdot)$ so that $|v_x(F) - v_{x'}(F)| \leq K\, m(x, x')$ for all F then with the Kullback Leibler condition (A) on the prior distribution Π we make consistent decisions. Conditions for consistency are less strict if we have a finite decision space.

Suppose that Π does satisfy condition (A) and (Ω, m) is compact, then, for consistency, we need $|v_x(F) - v_{x'}(F)| \leq K\, m(x, x')$ for all F. Other subjective choices for v and c might not lead to this condition being satisfied. However, it would appear sensible to modify v and c appropriately, and to remain close in some sense to the subjective choice in order for claims about the decision making procedure to be made. Without these claims, there does not appear to be any good reason for the procedure. That it is *coherent* does not count for anything. Our actions may well stick together – but uniformly badly! It is only with consistency in place that (we believe) we can say anything positive about the decision–making procedure at all.

6. Discussion

Our conclusion is as follows: Bayesian theory suggests a *form* for the procedure by which data is studied – the choice of procedure effectively boils down to the choice of prior. This form of procedure includes both good and bad procedures. That the choice is left to subjective persuasions based on prior information alone will not provide a unique prior and will include some priors which are consistent (i.e. lead to posterior consistency). There is no good reason why the final choice should not be from this set of consistent priors. This intersection of consistent and subjective priors is the set of nonparametric priors, since consistency – as defined in this paper – is only guaranteed from such priors.

From a decision theoretic perspective and for a given utility function, we have argued the need to make correct decisions as sample sizes increase. This requires the use of nonparametric priors with the Kullback–Leibler property.

References

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

BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* 2nd edn. Springer-Verlag.

BERK, R.H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37**, 51–58.

BERK, R.H. (1970). Consistency a posteriori. *Ann. Math. Statist.* **41**, 894–906.

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

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

BOX, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A* **143**, 383–430.

DE FINETTI, B. (1938). Sur la condition d'equivalence partielle. VI Colloque Geneve. *Act. Sci. Ind.* **739**. Herman, Paris.

DE GROOT, M. (1970). *Optimal Statistical Decisions.* McGraw Hill Book Company.

DEY, D., MÜLLER, P. AND SINHA, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* springer, New York.

DIACONIS, P. AND FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14**, 1–67.

Doob, J.L. (1949). Application to the theory of martingales. In *Le Calcul de Probabilities et ses Applications*. Colloques Internationaux du Centre National de la Researche Scientifique, Paris 23 27.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B* **57**, 45 97.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Freedman, D. (1963). On the asymptotic behaviour of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403.

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.

Hewitt, E. and Savage, L.J. (1955). Symmetric measures on cartesian products. *Trans. Am. Math. Soc.* **80**, 470–501.

Hirshleifer, J. and Riley, J.G. (1992). *The Analysis of Uncertainty and Information*. Cambridge University Press.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1999). Bayesian model choice: what and why? In *Bayesian Statistics 6*, pp 343–370. J.M.Bernardo,J.O. Berger,A.P.Dawid and A.F.M.Smith (Eds.) Oxford University Press.

Lindley, D.V. (1985). *Making Decisions* (2nd edn). Wiley & Sons.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer–Verlag.

Raiffa, H. (1970). *Decision Analysis*. Addison–Wesley.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard, Boston.

RAFTERY, A.E., MADIGAN, D. AND VOLINSKY, C.T. (1996). Accounting for model uncerainty in survival analysis improves predictive performance. In *Bayesian Statistics 5*, pp 323 349. J.M.Bernardo,J.O.Berger,A.P.Dawid and A.F.M.Smith (Eds.) Oxford University Press.

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

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

SMITH, A.F.M. (1984). Bayesian Statistics: Present position and potential developments: some personal views. *J. Roy. Statist. Soc. A* **147**. 245–259 (with discussion).

VON NEUMANN, J. AND MORGENSTERN, O. (1947). *Theory of Games and Economic Behaviour* 2nd edn. Princeton University Press. Princeton N.J.

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.

WALKER, S.G. AND GUTIÉRREZ-PEÑA, E. (1999). Robustifying Bayesian procedures (with discussion). In *Bayesian Statistics 6*, pp 685 – 710. J.M.Bernardo,J.O.Berger,A.P.Dawid and A.F.M.Smith (Eds.) Oxford University Press.

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.