Test (2007) 16: 188–197
DOI 10.1007/s11749-006-0008-8

ORIGINAL PAPER

Bayesian parametric inference in a nonparametric framework

Stephen G. Walker · Eduardo Gutiérrez-Peña

Received: 24 June 2004 / Accepted: 20 February 2005 /
Published online: 27 February 2007

Abstract This paper considers the problem of reporting a "posterior distribution" using a parametric family of distributions while working in a nonparametric framework. This "posterior" is obtained as the solution to a decision problem and can be found via a well-known optimization algorithm.

Keywords Decision theory · Expected utility · Nonparametric prior · Parametric predictive density

Mathematics Subject Classification (2000) 62C10 · 62G07

1 Introduction

There is a standard and well known route for the Bayesian to construct a parametric posterior distribution. The idea is to formulate a prior distribution on the parameter space Θ, which connects up with the parametric family of densities, say $f(x;\theta)$, with $\theta \in \Theta$. Here Θ is a finite dimensional parameter space. The prior distribution on Θ, say $\pi(\theta)$, combines with the data $X^n = \{X_1, \ldots, X_n\}$ to give the posterior distribution

$$\pi\left(\theta \mid X^n\right) = \frac{\prod_{i=1}^n f(X_i;\theta)\pi(\theta)}{\int \prod_{i=1}^n f(X_i;\theta)\pi(\theta)\,d\theta}.$$

S.G. Walker
Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK

E. Gutiérrez-Peña (✉)
Departamento de Probabilidad y Estadística, IIMAS, Universidad Nacional Autónoma
de México, Apartado Postal 20-726,
01000 Mexico D.F., Mexico
e-mail: eduardo@sigma.iimas.unam.mx

However, it is typically the case that any chosen parametric model is wrong, in the sense that there is no $\theta_0 \in \Theta$, with $\pi(\theta_0) > 0$, such that, for example, the $\{X_i\}$ are independent and identically distributed from $f(x; \theta_0)$. Under these circumstances, honest specification of a prior on θ is, to say the least, problematic. We would argue that, if the posited parametric model is thought to be wrong, it does not make much sense to try to specify a prior on θ.

In the traditional Bayesian approach to statistics, having acknowledged that the parametric model may be wrong, it seems prudent to compare different models using Bayes factors or other Bayesian criteria such as those discussed in Bernardo and Smith (1994, Chap. 6). But this ignores the fact that the use of the parametric model involves the declaration that

$$\Pr(f \in \Omega) = 1,$$

where

$$\Omega = \left\{ f : f(\cdot) \equiv f(\cdot; \theta) \text{ for some } \theta \in \Theta \right\}.$$

Here, for a set of densities B,

$$\Pr(f \in B) \equiv \Pi(B) = \int_{\{\theta : f(\cdot;\theta) \in B\}} \pi(\theta)\, d\theta.$$

The prior specification of $\Pr(f \in \Omega) = 1$, which carries through to the posterior, is clearly at odds with the experimenter who readily acknowledges the model should be checked once the data has been observed in order to verify that the model and the data are compatible. Authors such as Linsdey (1999) and Draper (1999) see this as a serious problem for Bayesians and we argue it is nothing short of irrational behavior on the part of the statistician. Draper (1999) also discusses the practical implications of model switching once $\Pr(f \in \Omega) = 1$ is questioned, suggesting poor calibration is the most likely scenario. Equivalently, poor statistics. Knowing this, Draper (1995) proposed model expansion (leading to model averaging) as a means to circumvent such problems.

Draper (1995) equates *good calibration* to *honest uncertainty assessments*. Being calibrated here can be loosely interpreted as not being surprised by the data, no matter how many are observed. Thus, a forecaster "is well calibrated if, for example, of those events to which he assigns a probability of 30 percent, the long run proportion that actually occurs turns out to be 30 percent" (Dawid 1982). If we condition on a single parametric model, poor calibration of the resulting inference is likely since model uncertainty is not being taken into account.

It might be suggested that a parametric model is simply a conditional model; that is, conditional on the assumption that it is a "good" model, inferences being reported conditionally on this fact. This then begs the question as to what the prior actually is. In the context of model comparison, the conditional idea implicitly assumes there are a number of possible models, say $M_1, \ldots, M_k$, with conditional prior weights, $\pi_1, \ldots, \pi_k$. One approach would then be to use posterior odds or Bayes factors for comparing models. But then none of the priors can be an actual prior. None of them acknowledge the uncertainty inherent in considering other models, so none can actually reflect true beliefs. When one model is selected after seeing the data, all the original uncertainty has been suppressed artificially and hence the experimenter is

clearly underestimating the uncertainty. Moreover, for a posterior to represent posterior beliefs the prior must represent prior beliefs, and none of them do in this case. We believe that the only rational approach in this conditional setting is for the Bayesian to use the prior model

$$\sum_{j=1}^{k} \pi_k M_k.$$

This model would then not be checked post data observation. Note, however, that such a prior is far from natural and in most cases its support will not contain the true density f.

Here we introduce what we believe should be at the heart of the solution: To put probability one on a set of densities for which the experimenter will undertake no checks no matter what data arrive. This is implicit in the approach of Draper (1995). In most cases this will require a prior distribution which puts mass on all densities. This can be achieved with the use of a nonparametric prior. If all densities are included in the prior then the data can now offer no surprises and there is no check of the assignment of probability one to be made. Of course, if a particular parametric model is thought to be "correct" (that is, a θ_0 does exist) then this is to be used, and no check will then be made. This procedure avoids the contradictory (incoherent, irrational) behavior of both assigning probability one to a model and a willingness to check the model once the data has been observed.

A referee has pointed out that our approach does not protect us from the very criticism we apply to the parametric Bayesian, since we are after all assuming that the data are independent and identically distributed conditional on the unknown density f. We would argue, however, that inference and prediction "always involve an assumption of conditional exchangeability of known and unknown quantities at some level of conditioning" (Draper 1995, Sect. 7). We would also argue that, while (conditional) independence is a more fundamental assumption, in many instances it can be justified by a good experimental design or a careful data collection process. Moreover, unlike the parametric model assumption, the independence assumption is typically not checked when assumed to be true. Finally, this is also a problem for the parametric Bayesian (besides the problems associated with a poor choice of model).

The stance of this paper is Bayesian nonparametric, and for us, we would report all summaries in that context. However, we do acknowledge that many statisticians prefer to work in a parametric framework. In this paper, then, we show how it is possible to obtain a parametric "posterior distribution" which avoids the irrational behavior discussed above and takes into account the uncertainty implicit in the choice of the parametric model(s).

So consider the parametric model represented by $f(\cdot;\theta)$ with $\theta \in \Theta$. The task would be to select a probability distribution on Θ which was somehow derived from the actual (nonparametric) posterior distribution

$$\Pi(df \mid X_1, \ldots, X_n) \propto \prod_{i=1}^{n} f(X_i)\Pi(df),$$

where Π denotes our nonparametric prior distribution. Therefore, we are looking for a probability distribution $\mu(\theta)$. We then suggest that the solution to a decision prob-

lem, to be described later, can as well be used as a parametric posterior distribution on Θ which will allow the Bayesian to undertake standard tasks such as model selection without behaving irrationally.

As pointed out above, we are looking at this problem from the nonparametric perspective, and not from a parametric one. Consequently, the only prior we acknowledge is Π and we do not ourselves see $\mu(\theta)$ as a posterior distribution but, as we shall see, as a solution to a well defined decision problem. However, we are proposing that $\mu(\theta)$ can be used by a parametric statistician as though it were a posterior distribution. Within our framework, it is then not irrational or incoherent to undertake model selection procedures using $\{f(\cdot;\theta), \mu(\theta)\}$.

The outline of the paper is as follows. In the next section we state the elements of the decision problem and discuss how it can be solved in practice. In Sect. 3 we consider the special case where the nonparametric predictive distribution is given by the empirical distribution function, that is, the Bayesian bootstrap. This simple case, however, can be easily extended to provide a solution for more general, informative nonparametric priors. This is illustrated in Sect. 4 with an example. In Sect. 5 we discuss asymptotic properties and Sect. 6 contains some concluding remarks.

2 Formal decision problem

The elements to be specified for setting up and solving a decision problem are well known (see, for example, Bernardo and Smith 1994). For us the unknown state of nature is the density function f generating the data $X_1, \ldots, X_n$. We assume that this density is such that $f \in \mathcal{F}$, the set of all densities with respect to the Lebesgue measure. The elements of our decision problem are as follows:

(1) A set of *decisions*; $\{\mu \in \mathcal{G}\}$, where $\mathcal{G}$ is the set of probability distributions on Θ.
(2) A set of *states of nature*; $\{f \in \mathcal{F}\}$.
(3) A *utility function* $U(\mu, f)$ evaluating the desirability of μ when f is the true density function.
(4) A *probability distribution* on the space of density functions representing beliefs about the true state of nature. In a Bayesian context, this probability is the prior Π in the no-sample problem and is $\Pi(\cdot|X^n)$ once the data $X^n = x^n$ have been observed.

The problem we focus on is one-step ahead prediction. It is apparent that it is also possible to construct a utility function for the problem of multiple predictions into the future. The importance of the (one-step ahead) predictive density is that it serves as the Bayes estimate of the unknown density given all the information available at any given point in time; see, for example, Haussler and Opper (1997).

Several authors, including Good (1952) and Bernardo (1979), advocate the logarithmic score as a utility function when the decision space consists of density functions. Other loss functions are possible, but we will settle with this one. Consequently, we consider

$$U(\mu, f) = \int \log\{p(x;\mu)\} f(x)\, dx,$$

where

$$p(x;\mu) = \int f(x;\theta)\mu(\theta)\,d\theta$$

is the "predictive" density for the model $f(\cdot;\theta)$ with probability distribution $\mu(\theta)$. Note that the pair $\{f(\cdot;\theta), \mu(\theta)\}$ is not the standard Bayesian parametric model since we do not regard μ as an actual prior; recall that, for us, the only real prior is Π. The distribution μ is merely a probability distribution on the parametric space for the family of densities $f(\cdot;\theta)$. Thus we regard $\{f(\cdot;\theta), \mu(\theta)\}$ as nothing more than a *working model* that the parametric statistician may want to use in order to make inferences about θ, despite the fact that the actual Bayesian model is nonparametric.

The solution to the decision problem is given by maximizing the expected utility

$$U_n(\mu) = \int \log\big\{p(x;\mu)\big\} f_n(x)\,dx, \tag{2.1}$$

where f_n is the nonparametric predictive density function. That is,

$$f_n(x) = \int f(x)\Pi\big(df \mid X^n\big).$$

More precisely, f_n is the predictive density associated with the prior which is sufficiently large to ensure that the experimenter will not be interested in checking, no matter what data arrive.

Suppose now that we have two working models, $M_1 = \{f^{(1)}(\cdot,\theta), \mu^{(1)}(\theta)\}$ and $M_2 = \{f^{(2)}(\cdot,\theta), \mu^{(2)}(\theta)\}$, and that we want to select one of them in order to report parametric inferences. Then we would obviously select M_1 over M_2 whenever $U_n(\widehat{\mu}^{(1)}) > U_n(\widehat{\mu}^{(2)})$, where $\widehat{\mu}^{(j)}$ is the maximizer of (2.1) under model j, $j = 1, 2$. In this way, a parametric statistician would be able to undertake model selection without incurring the contradiction alluded to in Sect. 1.

We now consider an interesting special case where the predictive distribution turns out to be the empirical distribution function.

3 The Bayesian bootstrap

Here we consider the Bayesian bootstrap (Rubin 1981) for constructing the nonparametric predictive distribution. In fact, in this case, the predictive distribution is the empirical distribution function and hence

$$U_n(\mu) = \frac{1}{n}\sum_{i=1}^{n} \log\big\{p(x_i;\mu)\big\}.$$

It is easily seen that the optimal μ is obtained via

$$\max_{\mu\in\mathcal{G}} \prod_{i=1}^{n} \int f(x_i;\theta)\mu(\theta)\,d\theta. \tag{3.1}$$

Lindsay (1983) discusses the existence and uniqueness of the (nonparametric) maximum likelihood estimator of a mixing distribution, and provides an algorithm based on the vertex direction method (VDM) to carry out such a maximization. He also shows that the maximizer of (3.1) is a discrete distribution with support on at most n points. Mallet (1986) proposes a related method for estimating the distribution of the parameters of a random effects model. Both authors discuss the connection between these procedures and those occurring in the theory of optimal design of experiments; see, for example, Fedorov (1972) and Silvey (1980). Schumitzky (1991) develops an alternative procedure based on the EM algorithm. He then extends the basic algorithm to cope with the case where a continuous solution is desired. However, in this latter case the solution is typically not a smooth density and tends to follow the discrete solution too closely. Böhning (1995) reviews several other algorithms. Also relevant here is the work of Magder and Zeger (1996), who propose a method of estimating the mixing distribution using maximum likelihood over the class of arbitrary mixtures of normals subject to the constraint that the component variances be bounded below by a value h. The nonparametric maximum likelihood estimate can then be obtained as a limiting case as $h \to 0$. This procedure can also be extended to estimate multivariate mixing distributions.

4 An example

For illustrative purposes, and to show how we can find μ in general, we consider as nonparametric prior the mixture of Dirichlet Process (MDP) model. This is a well known and a widely used nonparametric prior. See, for example, Escobar and West (1995). The parametric model will be normal with unknown mean parameter θ and known variance 1, written as $\mathrm{N}(\theta, 1)$. The nonparametric model is given in hierarchical form as

$$\begin{aligned} X_i|\theta_i &\sim \mathrm{N}(\theta_i, 1), \\ \theta_i|F &\sim F, \\ F &\sim \mathcal{D}(c, G). \end{aligned}$$

Here $\mathcal{D}(c, G)$ is a Dirichlet process prior (Ferguson 1973). We will not go into too much detail here except to say that the parameters $c > 0$ and G, a distribution function, represent scale and location, respectively. We will, for the sake of illustration, take $c = 1$ and $G = \mathrm{N}(0, 10^2)$. It is well known that F can be integrated out of the model leading to the marginal distribution of $(\theta_1, \ldots, \theta_n)$ being given by

$$p(\theta_1, \ldots, \theta_n) \propto g(\theta_1) \prod_{i=2}^{n} \left\{ g(\theta_i) + \sum_{j<i} \delta_{\theta_j}(\theta_i) \right\},$$

where δ_θ is the probability mass function with mass 1 at θ. Here g is the density function corresponding to G.

Inference is achieved via sampling from full conditionals $p(\theta_i|\theta_{-i})$ in a Gibbs sampler. This is easy to do and one can sample from the predictive density $f_n(x)$ by

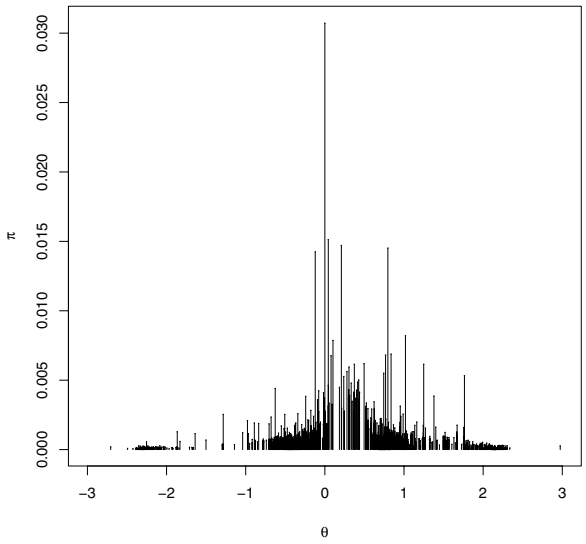

Fig. 1 Optimal distribution, $\widehat{\mu}$: $n = 10$

sampling θ_{n+1} from $p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n)$ and then taking X_{n+1} from $f(x; \theta_{n+1})$. We will denote a sample of size M from f_n as $\{Z_{n1}, \ldots, Z_{nM}\}$.

Consequently, we have an approximation to the expected utility function as

$$\frac{1}{M} \sum_{j=1}^{M} \log p(z_{nj}; \mu).$$

Therefore, we are now interested in finding the μ which maximizes

$$\prod_{j=1}^{M} p(z_{nj}; \mu), \tag{4.1}$$

which can be done using one of the algorithms mentioned in the previous section. As pointed out there, the solution will be a discrete distribution with support on at most M points. Note, however, that in this case we can control the "smoothness" of the solution simply by increasing the value of M, the size of the Monte Carlo sample.

We generated three samples, of respective sizes 10, 100 and 1 000, from a $\mathrm{N}(\theta_0, 1)$ density with $\theta_0 = 0$. For each of these samples, we found $\widehat{\mu}$, the maximizer of (4.1), based on a Monte Carlo sample of size $M = 10\,000$ from the nonparametric predictive and using the simple VDM algorithm. Figures 1 to 3 show $\widehat{\mu}$ for each of these three cases. As would be expected, the resulting distributions tend to concentrate around the value $\theta_0 = 0$ as the sample size increases.

5 Asymptotics

We would like to know that if the data do indeed come from the parametric model selected to make summaries then the sequence of solutions to the decision problem

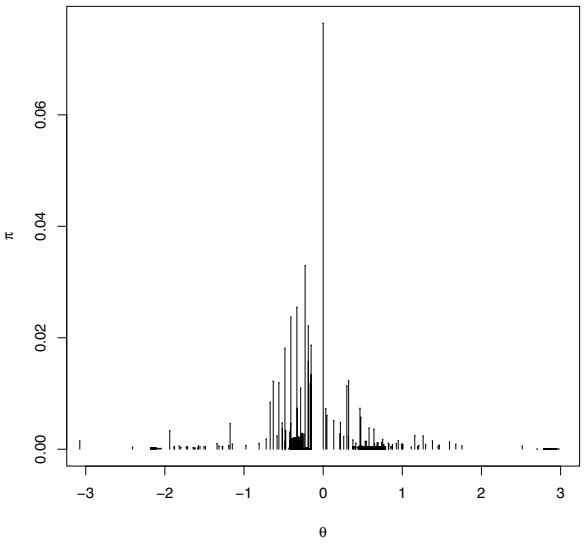

Fig. 2 Optimal distribution, $\widehat{\mu}$: $n = 100$

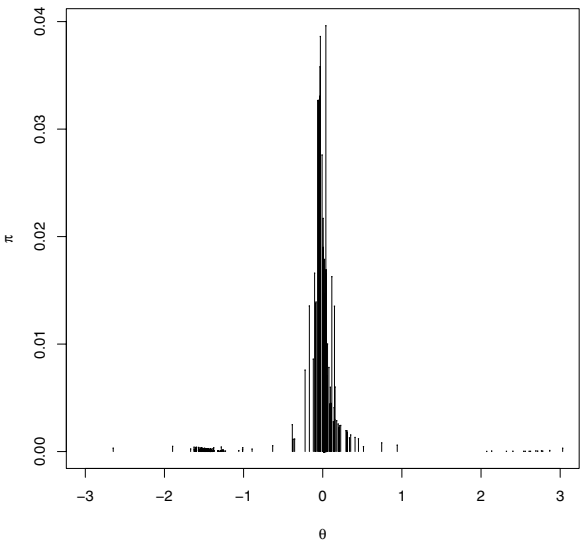

Fig. 3 Optimal distribution, $\widehat{\mu}$: $n = 1000$

$\mu_n(\theta)$ converges to the point mass at θ_0 with probability one. Let us denote $f(\cdot;\theta_0)$ by f_0 and $p(\cdot;\mu)$ by p_μ. In order to achieve this result we will assume the nonparametric predictive density is consistent in the sense that

$$d_K(f_n, f_0) \to 0$$

as $n \to \infty$ with probability one. Here $d_K(f,g) = \int f \log(f/g)$ is the Kullback–Leibler divergence between f and g. We will also assume that if $\tilde{\mu}_n$ is any sequence of probability distributions such that $p_{\tilde{\mu}_n} \to f_0$ with respect to the L_1 distance then $\tilde{\mu}_n \to \delta_{\theta_0}$ in the sense that $\tilde{\mu}_n(A) \to \mathbf{1}(\theta_0 \in A)$ $\forall$ sets A.

Note that if μ puts all its mass on θ_0 then $p_\mu = f_0$. Now, since μ_n minimizes $d_K(f_n, p_\mu)$ and $d_K(f_n, f_0) \to 0$, it follows that

$$d_K(f_n, p_{\mu_n}) \to 0$$

almost surely. Note that $d_K(f_n, f_0) \to 0$ implies $d_1(f_n, f_0) \to 0$ and hence we have both $d_1(f_n, f_0) \to 0$ and $d_1(f_n, p_{\mu_n}) \to 0$ almost surely. Here $d_1(f, g) = \int |f - g|$ is the L_1 distance between f and g. Consequently,

$$d_1(p_{\mu_n}, f_0) \to 0$$

almost surely, and it follows that $\mu_n \to \delta_{\theta_0}$ almost surely.

Sufficient conditions under which $d_K(f_n, f_0) \to 0$ are currently not known. However, Ghosal et al. (1999) provide sufficient conditions under which $d_1(f_n, f_0) \to 0$, which in most cases will mean that $d_K(f_n, f_0) \to 0$.

6 Concluding remarks

We see the development of the solution to the decision problem described in Sect. 2 as a possible way of avoiding the internal contradiction and poor calibration associated with a fully parametric Bayesian analysis involving model comparison and selection.

It would be tempting to compare the solution $\mu_n(\cdot)$ with a posterior derived in a fully parametric way, that is, via a parametric prior. This, we believe, would be inappropriate because we do not advocate the use of such a parametric posterior distribution unless it is used outside of the context of model comparison. It is perhaps better to regard $\mu_n(\cdot)$ as some sort of *surrogate* "posterior" distribution which is optimal in the sense that it yields a predictive density that is as close as possible to the nonparametric predictive f_n in terms of Kullback–Leibler divergence.

We think of $\mu_n(\cdot)$ as nothing more than a probability measure on the space Θ which may be useful to parametric Bayesians wishing to undertake model selection procedures and/or inference coherently. The parametric posterior, on the other hand, permanently asserts that $\Pr(f \in \Omega \mid \text{data}) = 1$ and so leads to incoherent model selection procedures.

Acknowledgements The authors would like to thank two anonymous referees for their insightful comments which greatly improved the presentation of the paper. This work was carried out while the second author was on a sabbatical leave at the Department of Mathematical Sciences, University of Bath, UK. He is grateful to this institution for their hospitality. The first author is financed by an EPSRC Advanced Research Fellowship. Support from DGAPA-UNAM and SNI, Mexico, is also gratefully acknowledged by the second author.

References

Bernardo JM (1979) Expected information as expected utility. Ann Stat 7:686–690
Bernardo JM Smith AFM (1994) Bayesian theory. Wiley, Chichester
Böhning D (1995) A review of reliable maximum likelihood algorithms for semiparametric mixture model. J Stat Planning Inference 47:5–28
Dawid AP (1982) The well-calibrated Bayesian. J Am Stat Assoc 77:605–610

Draper D (1995) Assessment and propagation of model uncertainty (with discussion). J Roy Stat Soc Ser B 57:45–97
Draper D (1999) Discussion of the paper "Bayesian nonparametric inference for random distributions and related functions", by Walker et al. J Roy Stat Soc Ser B 61:510–513
Escobar MD West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:577–588
Fedorov VV (1972) Theory of optimal experiments. Academic, New York
Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230
Ghosal S Ghosh JK Ramamoorthi RV (1999) Posterior consistency of Dirichlet mixtures in density estimation. Ann Stat 27:143–158
Good I (1952) Rational decisions. J Roy Stat Soc Ser B 14:107–114
Haussler D Opper M (1997) Mutual information, metric entropy and cumulative relative entropy risk. Ann Stat 25:2451–2492
Lindsay BG (1983) The geometry of mixture likelihoods: a general theory. Ann Stat 11:86–94
Linsdey JK (1999) Some statistical heresies. J Roy Stat Soc Ser D Stat 48:1–40
Magder LS Zeger SL (1996) A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. J Am Stat Assoc 91:1141–1151
Mallet A (1986) A maximum likelihood estimation method for random coefficient regression models. Biometrika 73:645–656
Rubin DB (1981) The Bayesian bootstrap. Ann Stat 9:130–134
Schumitzky A (1991) Nonparametric EM algorithms for estimating prior distributions. Appl Math Comput 45:143–157
Silvey SO (1980) Optimal design. Champan & Hall, London