Rev.R.Acad.Cienc.Exact.Fis.Nat. (Esp)
Vol. 93, N.º 3, pp 295-302, 1999
Monográfico: Métodos Bayesianos en las Ciencias

MODEL-FREE OBJECTIVE BAYESIAN PREDICTION

(Bayesian statistics/kernel density estimation/information theory/prediction/predictive distributions/reference analysis/scoring rules)

JOSÉ M. BERNARDO*

* Departament d'Estadística i I. O., Universitat de València. Facultad de Matemàtiques, 46100-Burjassot, València, Spain. jose.m.bernardo@uv.es. URL: www.uv.es/~bernardo/.

ABSTRACT

Probabilistc prediction of the value of a given observable quantity given a random sample of past observations of that quantity is a frequent problem in the sciences, but a problem which has *not* a commonly agreed solution. In this paper, *Bayesian* statistical methods and *information theory* are used to propose a new procedure which is *model-free,* in that no assumption is required about an underlying statistical model, and it is *objective,* in that a reference non-subjective prior distribution is used. The proposed method may be seen as a Bayesian analogue to conventional *kernel density estimation,* but one with an appropriate *predictive* behaviour not previously available. The procedure is illustrated with the analysis of some published astronomical data.

RESUMEN

Predicción Bayesiana objetiva con modelos probabilísticos desconocidos

En la investigación científica se plantea frecuentemente el problema de especificar una distribución de probabilidad que, a la vista de una muestra aleatoria de observaciones experimentales de una magnitud, permita predecir el valor de una observación *futura* de la misma magnitud; este problema de *predicción probabilística* carece sin embargo de una solución generalmente aceptada. En este trabajo se utilizan los métodos estadísticos Bayesianos y la teoría de la información para proponer una solución al problema descrito que no requiere suponer conocido un modelo paramétrico que describa el comportamiento de las observaciones, y que proporciona resultados objetivos en el sentido de que utiliza una distribución inicial de referencia que, por definición, no es subjetiva. El método propuesto puede ser descrito como un análogo Bayesiano al procedimiento convencional de estimación de densidades mediante núcleos (*kernel density estimation*), lo que proporciona un comportamiento predictivo adecuado del que no se dispone en la metodología convencional. El procedimiento es ejemplificado mediante el análisis de un conjunto conocido de datos relativos a la velocidad con la que las galaxias se desplazan por el universo.

[1] Research partially funded with grant PB97-1403 of the DGICYT, Madrid, Spain.

1. THE PREDICTION PROBLEM

Let $\boldsymbol{x} = \{x_1, ..., x_n\}$ be a set of n real-valued observations of some *observable* real-valued quantity x, and consider a situation where one is interested in a (necessarily probabilistic) *prediction* of a future observation of the same quantity. Let us suppone that the observed values $\{x_1, ..., x_n\}$ may be assumed to be a subset of an *exchangeable* sequence, so that the *order* in which these observations have been obtained is assumed to contain no relevant information on the behaviour of the x's. Note that, in particular, this includes *all* cases in which $\boldsymbol{x}$ may be assumed to be a random sample from some underlying probability model.

It then follows from the general representation theorem (see *e.g.,* Bernardo and Smith, 1994, Ch. 4 and references therein) that there exists some probability model $m(x_i \mid \theta)$, labelled by some parameter $\theta \in \Theta$, such that the joint probability density of $\boldsymbol{x}$ may be written as

$$p(\boldsymbol{x}) = p(x_1, ..., x_n) = \int_\Theta \prod_{i=1}^{n} m(x_i \mid \theta)\, p(\theta)\, d\theta \qquad (1)$$

Consequently, $\boldsymbol{x}$ may always be regarded as a random sample from *some,* typically unknown, probability model $m(x_i \mid \theta)$, indexed by some unknown (possibly multidimensional) paramenter $\theta \in \Theta$, defined as the limit as $n \to \infty$ of some function of $\boldsymbol{x}$, for which a prior distribution $p(\theta)$ *necessarily* exists. Note that this result is an *existence theorem* in probability theory and, hence, it is

not subject to any of the polemics often associated to the use of Bayesian statistics in the sciences with a subjetive prior specification.

An immediate corollary of the representation theorem is that *all* the information about the value of future observation x contained in the observed data $\boldsymbol{x}$ is encapsulated in its (posterior) *predictive* distribution

$$p(x \mid \boldsymbol{x}) = p(x \mid x_1, \ldots, x_n) = \int_\Theta m(x \mid \theta)\, p(\theta \mid \boldsymbol{x})\, d\theta, \quad (2)$$

where, by Bayes' theorem the *posterior* distribution $p(\theta \mid \boldsymbol{x})$ of the unknown parameter θ is of the form

$$p(\theta \mid \boldsymbol{x}) = p(\theta \mid x_1, \ldots, x_n) \propto p(\theta) \prod_{i=1}^{n} m(x_i \mid \theta). \quad (3)$$

For any exchangeable data set $\boldsymbol{x}$, the posterior predictive distribution $p(x \mid \boldsymbol{x})$ given by (2) is *the* solution to the problem posed: it precisely describes *all* available information about a future observation x. If a point estimate $\hat{x}$ is desired, the mode, the median or the mean of $p(x \mid x_1, \ldots, x_n)$ could be used; confidence regions $R(\alpha)$ with posterior probability $1 - \alpha$ may he obtained as solutions of the equation $\int_{R(\alpha)} p(x \mid \boldsymbol{x})\, dx = 1 - \alpha$. Those are however only *partial* (if very useful) descriptions of the available information about a future values of x; the *complete* solution is simply and elegantly encapsulated in $p(x \mid \boldsymbol{x})$. Moreover, any other from of solution will *necessarily* violate the basic rules of probability theory; unfortunately, this includes most conventional proposals, such as those obtained by plug-in estimates of the form $m(x \mid \hat{\theta})$, for some estimate $\hat{\theta}$ of θ. Naturally, the problem is to find a suitable model $m(x \mid \theta)$, and to specify the prior distribution, $p(\theta)$, for its associated parameter θ.

For a detailed description of Bayesian prediction, including the use of dynamic models, see the excellent review paper by West (1998), and references therein.

In some scientific contexts, there are good reasons to select a particular model $m(x \mid \theta)$; this may be suggested, for instance, by an underlying physical theory, by invariance considerations, or by judicious application of some limit theorem. If this is the case, the problem reduces to specifying an appropriate, non-subjective, model based, 'reference' prior distribution $\pi(\theta)$ which would let the data 'speak for themselves'. The prediction problem would then be immediately solved by the corresponding reference posterior predictive distribution

$$\pi(x \mid \boldsymbol{x}) = \pi(x \mid x_1, \ldots, x_n) = \int_\Theta m(x \mid \theta)\, \pi(\theta \mid x_1, \ldots, x_n)\, d\theta,$$
$$\pi(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta) \prod_{i=1}^{n} m(x_i \mid \theta). \quad (4)$$

In the long quest for these 'baseline' non-subjective distributions, a number of requirements have emerged which may reasonably be regarded as their necessary properties. These include invariance, consistent marginalization, good frequency properties, general applicability and limiting admissibility. The *reference analysis* algorithm, introduced by Bernardo (1979b) and further developed by Berger and Bernardo (1989, 1992) is, to the best of our knowledge, the only available method to derive non-subjective distributions which satisfy all these desiderata. For a discussion of the many polemic issues in this topic, see Bernardo (1997). For an introduction to reference analysis, see Bernardo and Smith (1994, Ch. 5), or Bernardo and Ramón (1998).

In many situations however, it is very difficult to specify the probability model $m(x \mid \theta)$ with a reasonable degree of confidence. An *exact* Bayesian approach then requires to specify a very large class of models $m(x \mid \theta)$, where $\theta \in \Theta$ is often infinitely dimensional, one of whose members hopefully provides a good approximation to the underlying probability mechanism, *and* a prior $p(\theta)$ which describes available information on this structure; popular choices are mixture models with Dirichlet priors (see *e.g.,* West, 1992; Escobar and West, 1995, Roeder and Wasserman, 1997, and references therein). However, subjetive prior specification within this framework is very difficut —an often polemic—, and the reference priors for those models are typically *very* difficult to derive.

A possible alternative, which will be described in this paper, is to consider an *approximate,* data-based 'model' which may be used as a *proxy* to the actual, unknown underlying model. The more successful techniques to achieve such a type of aproximation are known under the general heading of *kernel density estimation.* Those are considered in the next section.

2. KERNEL DENSITY ESTIMATION

2.1. Conventional Approach

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from some unknown underlying model $m(x \mid \theta)$. Conventional kernel density estimation consists on assuming that an appropriate proxy for the required predictive density is provided by

$$\hat{p}(x \mid \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} q(x \mid x_i, \hat{\sigma}), \quad (5)$$

where the *kernel* $q(\cdot \mid \mu, \sigma)$ is some location-scale probability model

$$q(\cdot \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right), \; f(t) > 0, \; \int_{\mathbb{R}} f(t)\, dt = 1 \quad (6)$$

and $\hat{\sigma} = \hat{\sigma}(\boldsymbol{x})$ is an estimate of the unknown parameter σ (see *e.g.,* Silverman, 1986).

A large proportion of the literature on kernel density estimation deals with the appropriate selection of the kernel function and the corresponding estimate $\hat{\sigma}$ of its 'window' σ. The more pupular choice seems to be a normal kernel, $q(x \mid \mu, \sigma) = \mathrm{N}(x \mid \mu, \sigma)$, with the so-called normal reference rule, given by

$$\hat{\sigma} = (4/3)^{1/5} \tilde{s}\, n^{-1/5} \approx 1.06\, \tilde{s}\, n^{-1/5},\ (n-1)\,\tilde{s}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad (7)$$

as its corresponding estimate (see Scott, 1992, p. 131, and references therein).

This is a plug-in estimate solution and, therefore, it is bound to violate basic probability theory principles. Indeed the use of (5) is found to be both inconsistent under marginalization, and incompatible with Bayes theorem (West, 1991).

2.2. A Bayesian Approach

As described in Section 1, if data $\boldsymbol{x} = \{x_1, ..., x_n\}$ are assumed to be a subset of some exchangeable sequence, then they may be considered as a random sample from some unknown underlying model. Note that the exchangeability assumption is *not* unduly restrictive; for instance, the underlying model may well be a mixture model, thus allowing to model outlying observations.

We will assume that for some k, with $0 < k < n$, the underlying model may be *approximated* by a kernel-type mixture based on a subset of size k of the observed data. Intuitively, we are assuming that the probabilistic behaviour of the exchangeable sequence from which the data have been sampled may approximately be described by mixtures with k components, where the value of k has yet to be specified. Formally,

Kernel approximation assumption. *Let $\boldsymbol{x}_k = \{x_1, ..., x_k\}$ be a subset of size k of some exchangeable sequence. It is assumed that there is a location-scale kernel $q(\cdot \mid \mu, \sigma)$ indexed by positive parameter σ, which may depend on $\boldsymbol{x}_k$, such that, for any other element x in the sequence,*

$$p(x \mid \sigma) \approx \frac{1}{k} \sum_{j=1}^{k} q(x \mid x_j, \sigma). \quad (8)$$

Under the kernel assumption, an approximate expression for the required posterior predictive density $p(x \mid \boldsymbol{x}_n)$ may be obtained. Indeed, it follows from (8) that for any partition of the observed data $\boldsymbol{x}_n = \{x_1, ..., x_n\}$ of the form $\boldsymbol{x}_n = \{\boldsymbol{x}_k, \boldsymbol{y}_m\}$, where $\boldsymbol{x}_k$ is a size k subset of $\boldsymbol{x}_n$, and $\boldsymbol{y}_m$ consists of those observations in $\boldsymbol{x}_n$ which are not in $\boldsymbol{x}_k$, with $m = n - k$ and $0 < k < n$, one may obtain a reasonable *approximation* to $p(\boldsymbol{y}_m \mid \sigma)$, namely

$$p(\boldsymbol{y}_m \mid \sigma) = \prod_{i=1}^{m} p(y_i \mid \sigma) \approx \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} q(y_i \mid x_j, \sigma) \right\}. \quad (9)$$

Thus, for any other element x in the exchangeable sequence,

$$\begin{aligned} p(x \mid \boldsymbol{x}_k, \boldsymbol{y}_m) &= \int_0^\infty p(x \mid \sigma)\, p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m)\, d\sigma \\ &\approx \int_0^\infty \frac{1}{k} \sum_{j=1}^{k} q(x \mid x_j, \sigma)\, p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m)\, d\sigma, \quad (10) \\ &= \frac{1}{k} \sum_{j=1}^{k} \int_0^\infty q(x \mid x_j, \sigma)\, p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m)\, d\sigma \end{aligned}$$

which is the average of k *integrated* kernels with respect to the posterior distribution of σ,

$$\begin{aligned} p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m) &\propto p(\sigma)\, p(\boldsymbol{y}_m \mid \boldsymbol{x}_k, \sigma) \approx \\ &\approx p(\sigma) \prod_{j=1}^{m} \left\{ \sum_{j=1}^{k} q(y_i \mid x_j, \sigma) \right\}. \end{aligned} \quad (11)$$

Since this is true for all partitions of this type, an estimate of the desired posterior predictive distribution may be obtained as

$$p(x \mid k, \boldsymbol{x}_n) = \frac{1}{n_p} \sum_{l=1}^{n_p} p(x \mid \boldsymbol{x}_k^{(l)}, \boldsymbol{y}_m^{(l)}), \quad (12)$$

where n_p is an arbitrary number of random partitions of the form $\boldsymbol{x}_n = \{\boldsymbol{x}_k, \boldsymbol{y}_m\}$. It is suggested that n_p should be of the same order that the sample size n; in the examples quoted in this paper, the number of simulations n_p has been chosen to be equal to the corresponding sample size. Note that the solution explicity depends on the number k of components in the mixtures which are judged necessary for an accurate description the behaviour of the data; we postpone to Section 4 our discussion of the choice of k.

The proposed solution conditions on one part of the data $\boldsymbol{x}_k$, to build the model, and on the rest of the data, $\boldsymbol{y}_m$, to learn about its parameter σ. This is intended as a workable *approximation* to an exact Bayesian appropach which would requiere a probability model on the unknown sampling distribution *and* a prior over its paramenters what, as mentioned before, may be extremely difficult to implement from a non-subjetive viewpoint.

2.3. Choice of the Kernel Function

The procedure described could be implemented for any choice for the kernel density. However there are several arguments which suggest the use of *normal* kernels:

i) Published literature on both kernel density estimation and Bayesian mixture models suggests that normal mixtures are typically able to provide good approximations to predictive densities (see *e.g.,* Diaconis and Ylvisaker, 1985).

ii) A 'maximum entropy' argument may be used to argue that normal kernels are the 'less demanding' of all possible location-scale kernels on the real line. Indeed, (see *e.g.,* Bernardo and Smith, 1994, Sec. 3.4 and references therein) if x is a real-valued location quantity defined on $(-c, c)$, then the positive, invariant, logarithmic divergence between a density $p(x)$ and the uniform density on $(-c, c)$, $\pi(x) = (2c)^{-1}$,

$$\delta\{p(\cdot), c\} = \int_{-c}^{c} p(x) \log \frac{p(x)}{\pi(x)}\, dx = \\ = \log 2c - \int_{-c}^{c} p(x) \log p(x)\, dx, \qquad (13)$$

measures the amount of information about x contained in $p(x)$. If $p(x)$ has both finite mean μ and finite variance σ^2 for all c, then a simple calculus of variations argument may be used to prove that, as $c \to \infty$, $\delta\{p(\cdot), c\}$ is minimized if, and only if $p(x) = \mathrm{N}(x \mid \mu, \sigma)$, so that normal kernels may be described as those containing the minimum amount of information among all possible location-scale kernels on the real line. Thus, normal kernels suggest themselves as a 'default' option for kernel estimation.

iii) If restrictions in the range of possible x values, to say an interval (a, b), are relevant, then one may work with the unrestricted transformed data $z_i = \log\{(x_i - a)/(b - x_i)\}$, use normal kernels to obtain $p(z \mid k, \boldsymbol{z})$, and transform back to the original metric to derive the required predictive density

$$p(x \mid k, \boldsymbol{x}) = p(z \mid k, \boldsymbol{z}) \frac{b - a}{(x - a)(b - x)}, \\ z = \log\{(x - a)/(b - x)\}. \qquad (14)$$

In the rest of this paper, we will restrict attention to normal kernels so that, with the notation established above, $q(y \mid \mu, \sigma) = \mathrm{N}(y \mid \mu, \sigma)$. We will find more convenient to work in terms of the variance $\phi = \sigma^2$, so that we will use kernels of the form

$$q(y \mid \mu, \phi) = \frac{\phi^{-1/2}}{\sqrt{2\pi}} \exp\left[-\frac{(y - \mu)^2}{2\phi}\right]. \qquad (15)$$

The relevant mixture model will be therefore $p(y \mid \boldsymbol{x}, \phi) = k^{-1} \Sigma_j q(y \mid x_j, \phi)$, where the x_j's are known constants and $\phi > 0$ is an unknown parameter.

To implement our proposal, there are two problems which remain to be solved. First, an appropriate *reference* prior $\pi(\phi)$ with respect to the model $p(y \mid \boldsymbol{x}, \phi)$ has to be chosen; then, a *computable* expression for the corresponding posterior density for $\pi(\phi \mid \boldsymbol{y}_m)$ given a random sample $\boldsymbol{y}_m = \{y_1, ..., y_m\}$ of m observations from $p(y \mid \boldsymbol{x}, \phi)$ has to be found. In words, we have to provide a reference analysis of the mixture model $p(y \mid \boldsymbol{x}, \phi)$. This is done in the next section.

3. REFERENCE ANALYSIS OF A MIXTURE OF NORMAL KERNELS

3.1. Mixture of Normal Models with Known Locations

For a given *known* vector $\boldsymbol{x} = \{x_1, ..., x_k\} \in \mathbb{R}^k$ and unknown $\phi > 0$, consider the mixture of k normal densities centered at each of the x_j's, with common variance ϕ, that is

$$p(y \mid \boldsymbol{x}, \phi) = \frac{1}{k} \sum_{j=1}^{k} q(y \mid x_j, \phi) = \\ = \frac{1}{k} \sum_{j=1}^{k} \left\{\frac{\phi^{-1/2}}{\sqrt{2\pi}} \exp\left[-\frac{(y - x_j)^2}{2\phi}\right]\right\}, \quad y \in \mathbb{R}. \qquad (16)$$

This is a probability model with a single unknown parameter $\phi > 0$, whose first two moments are immediately found to be

$$\mathrm{E}[y \mid \boldsymbol{x}, \phi] = \bar{x}, \quad \bar{x} = \frac{1}{k} \sum_{j=1}^{k} x_j, \\ \mathrm{Var}[y \mid \boldsymbol{x}, \phi] = s^2 + \phi, \quad s^2 = \frac{1}{k} \sum_{j=1}^{k} (x_j - \bar{x})^2 \qquad (17)$$

The likelihood function which corresponds to a sample $\boldsymbol{y}_m = \{y_1, ..., y_m\}$ of size m is

$$L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m) = \prod_{i=1}^{m} \left\{\frac{1}{k} \sum_{j=1}^{n} q(y_i \mid x_j, \phi)\right\} \propto \\ \propto \prod_{i=1}^{m} \left\{\sum_{j=1}^{k} \frac{\phi^{-1/2}}{\sqrt{2\pi}} \exp\left[-\frac{d_{ij}}{2\phi}\right]\right\}, \qquad (18)$$

where $d_{ij} = (y_i - x_j)^2$. Clearly, $L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m)$ is a computationally formidable quantity for large k and m values; it is known, however that, by definition, the reference prior only depends on the *asymptotic* behaviour of the likelihood function.

3.2. Asymptotic Behaviour of the Likelihood Function

The probability density of an inverted gamma distribution with parameters α and β is given by

$$\text{Ig}(\phi \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \phi^{-(\alpha+1)} \exp\left[-\frac{\beta}{\phi}\right], \quad \alpha > 0, \quad \beta > 0;$$

therefore, the likelihood function (18) may be reexpressed as

$$\begin{aligned} L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m) &= \prod_{i=1}^{m} \left\{ \frac{1}{k} \sum_{j=1}^{n} q(y_i \mid x_j, \phi) \right\} \propto \\ &\propto \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} \frac{\phi^{-1/2}}{\sqrt{d_{ij}}} \text{Ig}\left(\phi \left| \frac{1}{2}, \frac{d_{ij}}{2}\right.\right) \right\} \\ &= \phi^m \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} w_{ij} \text{Ig}\left(\phi \left| \frac{1}{2}, \frac{d_{ij}}{2}\right.\right) \right\}, \\ w_{ij} &= \frac{d_{ij}^{-1/2}}{\sum_{j=1}^{k_j} x_{jk}\, d_{ij}^{-1/2}}; \end{aligned} \tag{19}$$

thus, the likelihood function is proportional to the product of m mixtures of k inverted gamma densities $\text{Ig}(\phi \mid a, b_{ij})$ with $a = 1/2$, $b_{ij} = d_{ij}/2$, and weights inversely proportional to $\sqrt{d_{ij}}$.

The logarithmic divergence of an inverted gamma density $\text{Ig}(\phi \mid \alpha, \beta)$ from a general density $p(\phi)$ is given by

$$\begin{aligned} \delta(\alpha, \beta) &= \int_0^{\infty} p(\phi) \log \frac{p(\phi)}{\text{Ig}(\phi \mid \alpha, \beta)}\, d\phi \\ &= c + \alpha \log \beta - \log \Gamma(\alpha) - (\alpha + 1)\text{E}[\log \phi] - \beta \text{E}[\phi^{-1}], \end{aligned} \tag{20}$$

where c is an irrelevant constant; this is minimized if, and only if,

$$\text{E}[\log \phi] = \log \beta - \psi(\alpha), \quad \text{E}[\phi^{-1}] = \alpha/\beta, \tag{21}$$

where $\psi(\cdot)$ is the digamma function. The right hand sides of (21) are, respectively, the expected values of $\{\log \phi\}$ and $\{\phi^{-1}\}$ when ϕ has an inverted gamma $\text{Ig}(\phi \mid \alpha, \beta)$ distribution; thus, according to the commonly accepted logarithmic divergence criterium, (Bernardo, 1987; West and Harrison, 1989, Ch. 12) to approximate the density of a positive random quantity ϕ by an inverted gamma distribution, one shoud match the expected values of both $\{\log \phi\}$ and $\{\phi^{-1}\}$.

Taking $p(\phi) = \Sigma_j\, p_j\, \text{Ig}(\phi \mid \frac{1}{2}, \beta_j)$, it follows, after some algebra, that the best approximation to this mixture of inverted gammas by a *single* inverted gamma $\text{Ig}(\phi \mid \alpha, \beta)$ is obtained by the solutionl to the non-linear equation system

$$\log \alpha - \psi(\alpha) = \log \tfrac{1}{2} - \psi(\tfrac{1}{2}) + \log \frac{\beta^{(0)}}{\beta^{(1)}}, \quad \beta = 2\, \alpha \beta^{(1)} \tag{22}$$

where

$$\beta^{(0)} = \exp[\Sigma_j\, p_j \log \beta_j], \quad \beta^{(1)} = (\Sigma_j\, p_j\, \beta_j^{-1})^{-1} \tag{23}$$

are, respectively, the weighted logarithmic and harmonic means of the β_j's.

An aproximate explicit solution to (22) may be obtained making use the Stirling approximation to the digamma function, namely, $\log t - \psi(t) \approx (2t)^{-1}$; this leads to

$$\{\alpha \approx \text{t}/2, \beta \approx t\, \beta^{(1)}\}, \quad t = \left(1 + \log \frac{\beta^{(0)}}{\beta^{(1)}}\right)^{-1}. \tag{24}$$

The use of (24) to approximate the mixtures of inverted gammas in (19) leads to

$$\begin{aligned} L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m) &\propto \phi^m \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} w_{ij} \text{Ig}\left(\phi \left| \frac{1}{2}, \frac{d_{ij}}{2}\right.\right) \right\} \approx \\ &\approx \phi^m \prod_{i=1}^{m} \{\text{Ig}(\phi \mid a_i, b_i)\} \propto \\ &\propto \phi^m\, \phi^{-\Sigma\{a_i+1\}} \exp(-\Sigma b_i/\phi) \propto \\ &\propto \phi^{-m\,\bar{a}} \exp(-m\, \bar{b}/\phi), \end{aligned} \tag{25}$$

where $\bar{a} = m^{-1}\Sigma_i a_i$ and $\bar{b} = m^{-1}\Sigma_i b_i$, with

$$\begin{aligned} a_i &= \frac{t_i}{2}, \quad b_i = \frac{t_i\, d_i^{(1)}}{2}, \quad t_i = \left(1 + \log \frac{d_i^{(0)}}{d_i^{(1)}}\right)^{-1}, \\ d_i^{(0)} &= \exp\left[\sum_{j=1}^{k} w_{ij} \log d_{ij}\right], \quad d_i^{(1)} = \left[\sum_{j=1}^{k} w_{ij}\, d_{ij}^{-1}\right], \\ w_{ij} &= \frac{d_{ij}^{-1/2}}{\sum_{j=1}^{k} d_{ij}^{-1/2}}, \end{aligned} \tag{26}$$

and where, as before, $d_{ij} = (y_i - x_j)^2$.

3.3. Reference Distributions for ϕ

The asymptotic approximation to the likekihood function derived above provides a *heuristic* argument to obtain the reference prior. Indeed, it follows from (25) that, for large sample sizes m, the posterior distribution of ϕ will be approximately proportional to $\phi^{-m\,\bar{a}} \exp(m\, \bar{b}/\phi)$, which has a maximum at $\hat{\phi} = \bar{b}/\bar{a}$, the approximate maximun likelihood estimate of ϕ. Taking logarithms and expanding around $\hat{\phi}$, one finds, after some algebra,

$$\log p(\phi \mid \boldsymbol{x}_k, \boldsymbol{y}_m) \approx c + \frac{m h(\hat{\phi})}{2} (\phi - \hat{\phi})^2, \quad h(\phi) = \bar{a}\, \phi^{-2} \tag{27}$$

where c is some irrelevant constant. Hence (Bernardo and Smith, 1994, p. 314) the required reference prior should be

$$\pi(\phi) \propto h(\phi)^{1/2} \propto \phi^{-1}, \tag{28}$$

as one could possibly expect for an scale-type parameter. A more detailed analysis of the asymptotics involved would be necessary for a formal proof.

By Bayes' theorem $\pi(\phi \mid \boldsymbol{x}_k, \boldsymbol{y}_m) \propto \pi(\phi)\, L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m)$; thus, combining (28) and (25) we have an approximate

expression for the reference posterior distribution, immediately identified as an inverted gamma density, mamely

$$\pi(\phi \mid \boldsymbol{x}_k, \boldsymbol{y}_m) \propto \phi^{-1}\phi^{-m\,\bar{a}} \exp(-m\,\bar{b}/\phi) \propto \propto \mathrm{Ig}(\phi \mid m\bar{a}, m\bar{b}) \tag{29}$$

3.4. Approximate Reference Predictive Distribution

Introducing the approximation (29) in the procedure described by (10), and using the known fact that the mixture of normal distributions with inverted gamma distributed variances produces an Student t distribution, the required reference predictive distribution may be approximated by

$$\pi(x \mid \boldsymbol{x}_k, \boldsymbol{y}_m) = \frac{1}{k}\sum_{j=1}^{k}\int_0^\infty \mathrm{N}(x \mid x_j, \phi)\,\mathrm{Ig}(\phi \mid m\bar{a}, m\bar{b})\,d\phi = \frac{1}{k}\sum_{j=1}^{k}\mathrm{St}(x \mid x_j, \sqrt{d}, mt) \tag{30}$$

where

$$t = \frac{1}{m}\sum_{i=1}^{m} t_i, \qquad d = \frac{\Sigma_{i=1}^{m}\, t_i\, d_i^{(1)}}{\Sigma_{i=1}^{m}\, t_i}. \tag{31}$$

In words, for a given partition of $(\boldsymbol{x}_k, \boldsymbol{y}_m)$ of the data set $\boldsymbol{x}$, the desired reference predictive density may be approximated by a mixture or *Student* kernels centered at each of the x_i's, with a scale $\sqrt{d}$, the squared root of a weighted mean of weighted harmonic means of the squared distances $(y_i - x_j)^2$, which plays the same central role as that played by the 'window' in conventional kernel density estimation.

If n_p random partitions $\{(\boldsymbol{x}_k^{(l)}, \boldsymbol{y}_m^{(l)}),\ l = 1, \ldots, n_p\}$ of the same size k are performed, we can use (12) to obtain

$$\pi(x \mid k, \boldsymbol{x}) = \frac{1}{n_p}\sum_{l=1}^{n_p} p(x \mid \boldsymbol{x}_k^{(l)}, \boldsymbol{y}_m^{(l)}) = \frac{1}{n_p}\sum_{l=1}^{n_p}\frac{1}{k}\sum_{j=1}^{k}\mathrm{St}(x \mid x_j^{(l)}\sqrt{d^{(l)}}, mt^{(l)}) \tag{32}$$

We finally need a procedure to select k. This is developed in the next section.

4. PERFORMANCE

The choice of k is a particular case of the general problem of *model choice.* It has often been argued (see *e.g.,* Bernardo and Smith, 1994, Ch. 6 and reference therein) that model choice may usefully be treated as a *decision problem* where the utility function is a proper scoring rule evaluating the behaviour of the corresponding predictive distribution.

Moreover (Bernardo, 1979a; Bernardo and Smith, 1994, Sec. 3.4), it may be argued that the *logarithmic* scoring rule is the appropriate proper scoring rule to use in pure inference problems; it follows that the expected utility of using an approximate model $\hat{p}(x)$ to predict the value of an observable random quantity x with density $p(x)$ may reasonably be assumed to be of the form

$$u(\hat{p}) = a\int_X p(x)\,\log[\,\hat{p}(x)]\,dx + b, \tag{33}$$

where $a > 0$ and b are arbitrary constants. If the true distribution $p(x)$ is unknown but a random sample $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ of observations is available, then one may use the corresponding Monte Carlo approximation

$$\hat{u}(\hat{p}) \approx a\frac{1}{n}\sum_{j=1}^{n}\log[\hat{p}(x_j \mid \boldsymbol{x}_{n-1}(j))] + b, \tag{34}$$

where $\hat{p}(x_j \mid \boldsymbol{x}_{n-1}(j))$ is the predictive density of x_j based on the set all the *other* observations $\boldsymbol{x}_{n-1}(j) = \boldsymbol{x}_n - \{x_j\}$.

Equation (34) may be also seen as a cross-validation procedure, where the predictive value of the model $\hat{p}(\cdot)$ is judged by its average performance when predicting one observation based on all the others.

The constants a and b in equations (33) and (34) may arbitrarily be chosen to define some easily understandable scale and origin. In the examples which follow, we use the values a and b defined by the equations

$$u\{\mathrm{N}(\cdot \mid 0, 1), 0\} = 1 \qquad u\{\mathrm{N}(\cdot \mid 0, 1), 3\} = 0, \tag{35}$$

leading to

$$a = 2/9 \approx 0.2222, \qquad b = 1 + \log(2\pi)/9 \approx 1.2042. \tag{36}$$

Thus, the utility of predicting the value of an observable quantity by a standard normal is set to be one if

Table 1. Mean and standard deviations of the predictive utilities of 20 reference predictive densities for partition sizes $k = 1, \ldots, 12$. The expected utility of the conventional kernel estimate is 0.709.

k	$\bar{u}$	s_u
1	0.623	0.007
2	0.701	0.011
3	0.742	0.009
4	0.761	0.010
5	0.765	0.005
6	0.764	0.008
7	0.767	0.006
8	0.766	0.006
9	0.762	0.005
10	0.753	0.007
11	0.739	0.006
12	0.698	0.008

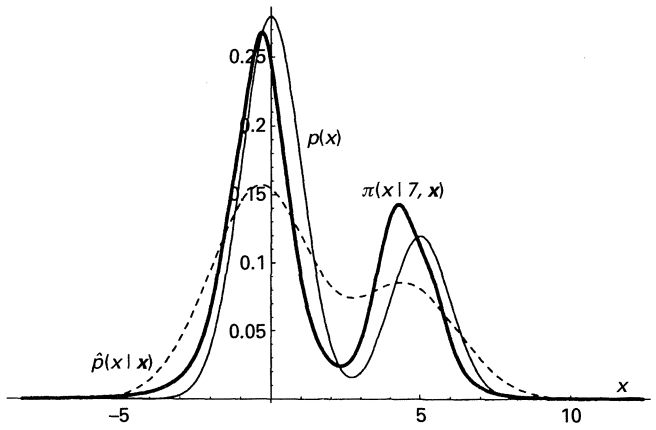

Figure 1. Analysis of 14 observations simulated from the mixture of two normals (thin line) $p(x) = 0.7\ \mathrm{N}(x \mid 0, 1) + 0.3\ \mathrm{N}(x \mid 5, 1)$. Conventional kernel estimate $\hat{p}(x \mid \boldsymbol{x})$ (dashed line), and optimal Bayes reference estimate (thick line) $\pi(x \mid \boldsymbol{x})$.

centered a its realized value, and zero if centered three standard deviations apart; consequently, a negative value would indicate a probabilistic prediction which associates to the actual observation a smaller density than the density of a standard normal at the point 3.

5. EXAMPLES

5.1. Simulated Data from a Mixture of Two Normals

In his interesting report on Bayesian prediction using mixtures of Dirichlet process models, West (1990) makes repeated used of the sample of 14 observations

$$\boldsymbol{x} = \{-1.39, -0.85, -0.54, -0.32, -0.31, -0.30, -0.19, -0.02, 0.54, 3.65, 4.21, 4.30, 4.98, 5.51\}$$

generated from the mixture of two normals

$$p(x) = 0.7\ \mathrm{N}(x \mid 0, 1) + 0.3\ \mathrm{N}(x \mid 5, 1).$$

We used (34), with the constants a and b set to the values provided by (36), to evaluate the behaviour of the reference predictive distribution $\pi(x \mid k, \boldsymbol{x})$ given by (32) for $k = 1, \ldots, 12$.

The procedure was repeated 20 times; Table 1 shows the mean and standard deviations of the estimated expected utilities. It may be appreciated that the expected utility is maximized with $k = 7$ leading to an expected utility 0.767. We also used (34) and (36) to evaluate the behaviour of the conventional kernel estimate provided by (5) and (7); this lead to an expected utility 0.709.

Figure 1 shows the density from which the data were actually generated, its conventional kernel estimate and a reference predictive density computed with the optimal partition size, $k = 7$. It is easily appreciated that the Bayesian solution provides a much better match to the true density.

5.2. Astronomical Data

Postman *et al.* (1986) describe a set of 82 measures of speed of galaxies, reproduced in Table 2, which have attracted considerable discussion over their underlying structure. We will now illustrate the proposed metodology with these data; for alternative Bayesian analysis see Roeder (1992), Escobar and West (1995), and Roeder and Wasserman (1997).

As with the simulated data above, we used (34) and (36), to evaluate the behaviour of the reference predictive

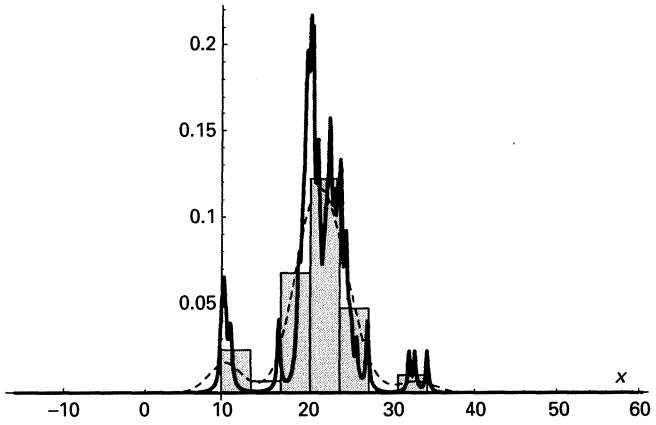

Figure 2. Speeds of Galaxies in the Corona Borealis Region $n = 82$. Conventional kernel estimate (dashed line) and Bayes optimal reference estimate ($k = 72$, continuous line).

Table 2. Ordered Speeds of Galaxies in the Corona Borealis Region ($\times 10^3$ m/seg)

9.172	9.350	9.483	9.558	9.775	10.227	10.406	16.084	16.170	18.419
18.552	18.600	18.920	19.052	19.070	19.330	19.343	19.349	19.440	19.473
19.529	19.541	19.547	19.663	19.846	19.856	19.863	19.914	19.918	19.973
19.989	20.166	20.175	20.179	20.196	20.215	20.221	20.415	20.629	20.795
20.821	20.846	20.875	20.986	21.137	21.492	21.701	21.814	21.921	21.960
22.185	22.209	22.242	22.249	22.314	22.374	22.495	22.746	22.747	22.888
22.914	23.206	23.241	23.263	23.484	23.538	23.542	23.666	23.706	23.711
24.129	24.285	24.289	24.366	24.717	24.990	25.633	26.960	26.995	32.065
32.789	34.279								

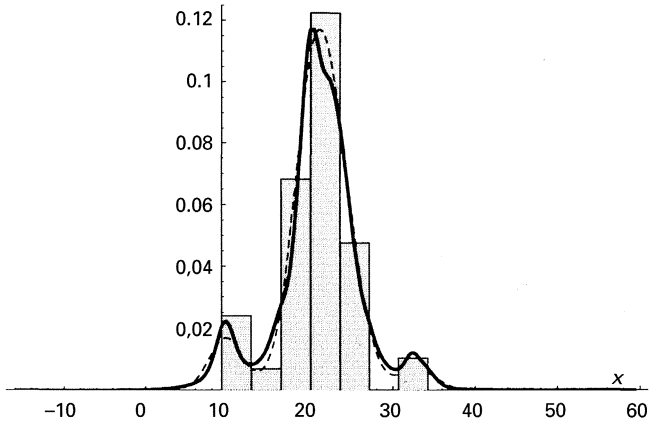

Figure 3. Speeds of Galaxies in the Corona Borealis Region $n = 82$. Conventional kernel estimate (dashed line) and Bayes reference estimate for $k = 25$ (continuous line).

distribution $\pi(x \mid k, \boldsymbol{x})$ given by (32) for $k = 1, ..., 80$. It was found that the best partition size corresponds to $k = 72$ leading to an expected utility 0.633. We also used (34) and (36) to evaluate the behaviour of the conventional kernel estimate provided by (5) and (7); this lead to an expected utility 0.604.

Over the background of a histogram of the data, Figure 2 shows its conventional kernel estimate and the reference predictive density computed with the optimal partition size, $k = 72$. It is easily appreciated that the proposed Bayesian solution suggests that, *to optimize predictive power,* the model has to be far more complex than the tri-modal solution given by conventional kernel estimation; speed galaxies appear to have *many* clusters, and those are duly reflected by the reference predictive distribution. Indeed, a trimodal solution, similar to that obtained by kernel estimation is obtained, for instance, with $k = 25$ (see Figure 3) but its expected utility is only 0.609 showing its smaller predictive power. If simplicity, rather than just predictive power, is to be taken into consideration, this may be done within the Bayesian framework by appropriately modifying the utility function.

It is important to note that the Bayesian solution *is* a predictive distribution, from which one is entitled to derive quantitative probabilistic predictions; since the reference predictive $\pi(x \mid k, \boldsymbol{x})$ is a mixture of Student densities this does not even require numerical integration, but may be done in terms of the Student distribution function. Thus, the probability that the speed of a galaxy is, say, larger than 35, is simply

$$\Pr[x > 35 \mid \boldsymbol{x}] \approx \int_{35}^{\infty} \pi(x \mid 72, \boldsymbol{x})\, dx = 0.0012.$$

This predictive interpretation, central to most scientific data analysis is *not* justifiable from a conventional kernel density estimation viewpoint.

REFERENCES

1. Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200-207.
2. Berger, J. O. and Bernardo J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35-60 (with discussion).
3. Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686-690.
4. Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113-147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.). Brookfield, VT: Edward Elgar, 1995, 229-263.
5. Bernardo, J. M. (1987). Approximations in statistics from a decision-theoretical viewpoint. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 53-60.
6. Bernardo, J. M. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference* **65**, 159-189 (with discussion).
7. Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 1-35.
8. Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.
9. Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Amsterdam: North-Holland, 133-156 (with discussion).
10. Escobar, M. D. and West, M. (1995). Bayesian density stimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.
11. Postman, M., Huchra, J. P., and Geller, M. J. (1986). Probes of large scale structures in the Corona Borealis region. *The Astronomical Journal* **92**, 1238-1247.
12. Roeder, K. (1992). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85**, 617-624.
13. Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894-902.
14. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.
15. Scott, D. W. (1992). *Multivariate Density Estimation.* New York: Wiley.
16. West, M. (1990). Bayesian kernel density estimation. *Tech. Rep.* **90-A02**, ISDS, Duke University.
17. West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika* **78**, 421-425.
18. West, M. (1992). Modelling with mixtures. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 503-524 (with discussion).
19. West, M. (1998). Bayesian Forecasting. *Encyclopedia of Statistical Sciences* (S. Kotz, C. B. Read and D. L. Banks, eds.). New York: Wiley, 50-60.
20. West, M. and Harrison (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer. Second edition in 1997.