View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Kent Academic Repository

Bayesian priors from loss matching

Philip J. Brown & Stephen G. Walker *

Abstract

This paper is concerned with the construction of prior probability measures for parametric families of densities where the framework is such that only beliefs or knowledge about a single observable data point is required. We pay particular attention to the parameter which minimizes a measure of divergence to the distribution providing the data. The prior distribution reflects this attention and we discuss the application of the Bayes rule from this perspective. Our framework is fundamentally nonparametric and we are able to interpret prior distributions on the parameter space using ideas of matching loss functions, one of which is coming from the data model and the other from the prior.

Keywords: Conjugate prior; Dirichlet process; Kullback–Leibler divergence; Loss function; Model choice; $\mathcal{M}$-open; Prior distribution; Self–information loss.

1. Introduction. A key component of the statistical approach to the modeling of independent and identically distributed or exchangeable outcomes is the choice of a parametric family of densities, denoted by $f(x|\theta)$, with $\theta \in \Theta$ and Θ the parameter space. For the Bayesian, the additional task is to construct a prior distribution $\pi(\theta)$ on Θ. This task has been a well debated subject in the literature (see the Appendix for a systematic review) and in particular how one should proceed when information or knowledge about θ is scant or non–existent. Priors under such a scenario have been termed non–informative, vague, objective, reference, default. This issue, according to Bernardo and Smith (1994), "is far more complex than the apparent intuitive immediacy of these words and phrases would suggest". According to Hartigan

*Philip J. Brown is Professor of Statistics, School of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, U. K. (email: pjb8@kent.ac.uk); Stephen G. Walker is Professor of Statistics, School of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, U. K. (email: S.G.Walker@kent.ac.uk)

(1998), "The selection and justification of these priors is an important part of the Bayesian approach".

The subjective Bayesian approach insists that π is a proper density and exploits information about θ to construct it. There are a number of considerations here. The obvious one is that there may be no qualitative information to use. Another, and more poignant one is that if, as is typical, $f(x|\theta)$ is selected as an approximation or for pragmatic reasons, then assigning a subjective probability to θ must be problematic. For what is θ? We are asked to express uncertainty about θ using the language of probability, but before we can do this we need to define θ. We need to know what we are expressing uncertainty about. These problems have been widely reported in the literature; see, for example, Goldstein (1981).

To elaborate here. If a subjective prior has been assigned then presumably quantities such as $\mathrm{P}(\theta \in A)$ for suitable $A \subseteq \Theta$ must mean something. But this can only be the case if θ means something. We would need then to complete the statement "θ is the parameter value which ...". The only evident and worthwhile identification of θ here is the parameter value which minimizes a distance, or divergence, (such as the Kullback–Leibler divergence) to the true sampling density function. Therefore, in the case of the Kullback–Leibler divergence we would be interested in the θ which minimizes

$$l(\theta) = -\int \log f(x|\theta)\, \mathrm{d}F_0(x)$$

where we use F_0 to denote the distribution of the observations. If F_0 is associated with a $f(x|\theta_0)$ for some true $\theta_0 \in \Theta$ then this θ_0 becomes the true parameter value. Outside of this scenario we believe it is essential to acknowledge what interest should focus on, rather than express beliefs via probabilities about non–existent "true" parameter values. In this context, we believe a prior should be targeted at θ^* which is the parameter value minimizing $l(\theta)$. Hence $\mathrm{P}(\theta^* \in A)$ means something as this θ^* is a real parameter value.

The likelihood approach, and in particular the maximum likelihood estimator, implicitly acknowledge the appropriateness of thinking about θ^*. Using the data $(X_1, \ldots, X_n)$ and the empirical distribution function to substitute for F_0, (which is something sensible to do), we have the estimated $l(\theta)$ as the simple empirical average

$$l_n(\theta) = -n^{-1} \sum_{i=1}^{n} \log f(X_i|\theta).$$

Minimizing this yields the maximum likelihood estimator $\widehat{\theta}$.

In both Bayes and classical approaches, explicitly acknowledging that it is θ^* which is of interest, means that the subsequent statistics: confidence intervals, testing, asymptotics, consistency (in the case

of classical methods) and the application of Bayes theorem itself for Bayesian analysis, now lack justification.

For Bayesian inference assigning a subjective prior to θ with interest in θ^* is inconsistent with an application of Bayes theorem and there seems little theoretical room for manoeuvre here. These concerns can then motivate an objective prior, not solely when prior information is scant. And for this there seems no reason to insist on a proper prior distribution. However, an obvious choice of objective prior based on ignorance is problematic. According to Bernardo and Smith (1994), "There is no objective prior that represents ignorance."

To deal with the problem of assigning a prior $\pi(\theta)$, which reflects beliefs about θ^* while wishing to use something equivalent to a Bayes theorem, our approach is not the traditional subjective idea in that we do not construct $\pi(\theta)$ based on the expression of beliefs directly. Rather, we require the Bayesian to express beliefs about the distribution of the data directly, and this first guess will be denoted by M_0. We also require the specification of a parameter, $c > 0$, which reflects the degree of belief in the choice of M_0. These, i.e. (c, M_0), are necessary specifications in most Bayesian nonparametric prior models. With these we then provide a means, using loss functions, by which to build a prior distribution for $\pi(\theta)$.

We note here that a selection of (c, M_0) already defines a particular Bayesian (nonparametric) prior. It is the Dirichlet process prior (Ferguson, 1973). Details of this prior will be outlined in Section 1.2. The use of this nonparametric prior in the present paper is essentially the same as used in Gutiérrez-Peña and Walker (2005), and this will also be described in Section 1.2.

We see our approach as intermediate between the traditional ideas for subjective and objective priors. Indeed, the subjective choice of (c, M_0) can be made with no reference to any parametric model $f(\cdot|\theta)$. On the other hand, once (c, M_0) have been specified, and then the model $f(\cdot|\theta)$ selected, we provide an objective criterion for the choice of $\pi(\theta)$. As far as we are aware, this is a new idea.

We will discuss our idea for prior construction using exchangeable distributed observations, say $(x_1, x_2, \ldots)$. Then, according to de Finetti (1937), and later also studied by Hewitt and Savage (1951), there exists a probability measure on the space of relevant distribution functions such that the sequence can be generated by first generating a random distribution function from the probability measure and then taking the (x_i) to be independent and identically distributed from this randomly generated distribution. That is, the density function for an n sequence is given by

$$m(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} f(x_i)\, \pi(\mathrm{d}f),$$

where π is the probability measure generating the random distribution functions, with $f(x)$ denoting the corresponding density function.

This proceeds by observing part of the sequence, say $(x_1, \ldots, x_n)$, and use this to learn about the density function which renders the sequence independent and identically distributed. The vehicle for doing this is the posterior distribution $\pi(\mathrm{d}f|x_1, \ldots, x_n)$ and one would anticipate, that as n gets large, the sequence of posterior distributions accumulate about this density.

However, for this to happen, the density generating the sequence must be in the support of the prior. By this we mean that the densities that π can generate are going to be able to land arbitrarily close to the density generating the sequence, with respect to some suitable distance, which for the sake of concreteness and for reasons which will become clear later on, we shall take to be the Kullback–Leibler divergence.

Traditional approaches to the construction of $\pi(\mathrm{d}f)$ relied heavily on parametric models. So it is easy to generate random densities by generating a random parameter $\theta \in \Theta$ and then have $f(x|\theta)$ as the random density. Hence, in this way, a $\pi(\mathrm{d}f)$ has been constructed, and with this we can represent the prior by the pair $\{f(x|\theta), \pi(\theta)\}$. It is important to note that the prior is not just $\pi(\theta)$; it is the pair. Probably the more important part of the prior is $f(x|\theta)$, since it determines the support of the prior $\pi(\mathrm{d}f)$. For some reason, which is not so clear, the prior has become known to be exclusively as $\pi(\theta)$ in the Bayesian literature.

This is not our main point, but it does mean that the literature on Bayesian prior construction has narrowly focused on the $\pi(\theta)$. This has led to issues between interpretation of θ and whether an objective or subjective construction is needed. See recent discussions of Goldstein (2006) and Berger (2006).

We will adopt the common stance with Bayesian analysis which is that $f(x|\theta)$ has been selected for pragmatic reasons (e.g. Box, 1980). In particular, there is no θ conditional on which the (x_i) are independent and identically distributed with density $f(x|\theta)$. We are in what Bernardo and Smith (1994) refer to as the $\mathcal{M}$–open view. It is then true to say that the usual update provided by Bayes' theorem is not valid and this is what led Key et al. (1999) to propose a cross–validation updating rule. Many other concerns along these lines have also been detailed in the literature.

Additionally, within this $\mathcal{M}$–open view, it is difficult to see how a $\pi(\theta)$ can really be developed. What is being targeted, what is uncertainty being expressed about? This sentiment is central to de Finetti's subjectivist approach to inference. The idea of separating out a part of the distribution for which the complementary part is unknown but trials are independent given it, he emphasizes cannot be 'stripped of its, so to speak, "metaphysical" character'. With respect to statements of

independence conditional on some *unknown* parameter he states, 'From our point of view these statements are devoid of sense, and no one has given them a justification which seems satisfactory, even in relation to a different point of view.' (de Finetti, 1964, chapter V, p141). Central to his approach is the supremacy of observables over unobserved parameters; these parameters being mere constructs with no operational meaning in his formulation except that which can be deduced via the representation theorem of exchangeability on sequences of trials.

Taking this paradigm as the cue, it is our aim to undertake Bayesian inference by only expressing beliefs about the observables. We aim to develop the theory with only a $M_0(\mathrm{d}x)$ and a $c > 0$ being specified. Here the $M_0(\mathrm{d}x)$ is a prior guess as to the distribution generating each x_i; and c is a measure of the degree of belief in the choice of $M_0(\mathrm{d}x)$.

The essence of our ideas is how to map

$$[(c, M_0), f(\cdot|\theta)] \Rightarrow \pi_\Theta(\theta).$$

We will show that the procedure we develop is coherent and coincides with an application of Bayes theorem; so that once $(X_1, \ldots, X_n)$ has been observed, and we have updated (c, M_0) to (c_n, M_n) (using the nonparametric model (see Section 1.2)) then

$$[(c_n, M_n), f(\cdot|\theta)] \Rightarrow \pi_\Theta(\theta|X_1, \ldots, X_n)$$

where

$$\pi_\Theta(\theta|X_1, \ldots, X_n) \propto \pi_\Theta(\theta) \times \prod_{i=1}^{n} f(X_i|\theta).$$

In the next sub–section 1.1 we look at a key component of our approach: the loss function, and in sub–section 1.2 we review the Dirichlet process and how we use in the current context.

1.1 Loss functions. Loss functions are a key to the paper. In the most broad of definitions a loss function denotes the loss (incurred to an individual) when outcome u arises. The loss is measured as $l(u)$. Or there could be two outcomes u and v for which we write the loss as $l(u, v)$. These two become distinct situations when one of the outcomes arises out of the control of the individual and the other is an action determined by the individual. But this is not essential and the loss is the loss however the outcomes arise.

A common situation is when one of the outcomes is selected as an action and the other is an as yet unknown outcome. The choice of action by the individual can then be made, apparently rationally, by selecting a belief distribution for the unknown outcome, let this be v and the belief distribution be $Q(v)$, and the appropriate loss function

is given by the expected loss:

$$l(u) = \int l(u, v)\, \mathrm{d}Q(v).$$

See for example Hirshleifer and Riley (1992).

Loss functions are therefore a means by which to connect outcomes and choices, or just outcomes. We are interested in connecting θ with a number of outcomes and/or choices. These are with X, F_0 and π_Θ. For ease of notation we will refer to π_Θ as simply π.

Our loss function connecting θ and F_0 is standard; it is based on the logarithmic score function which is fundamentally linked to the Kullback–Leibler divergence (Kullback and Leibler, 1951):

$$l(\theta, F_0) = -\int \log f(x|\theta)\, \mathrm{d}F_0(x).$$

Since F_0 is unknown, the rational approach is to construct a probability distribution for it and to replace F_0 by the expectation of this probability distribution. For us this would be M_0 and therefore the loss function for θ would be

$$l(\theta) = -\int \log f(x|\theta)\, dM_0(x).$$

This can also be seen as an expected loss for the loss function

$$l(\theta, X) = -\log f(X|\theta).$$

For if X is observed then one can evaluate the loss for θ using this loss function as we discussed earlier. Based on a sample of size n the cumulative loss would be $l_n(\theta)$ and minimizing this yields the maximum likelihood estimator.

Looking at this loss function the other way round also is also perfectly natural. For if we know θ then the x minimizing $-\log f(x|\theta)$ is the x maximizing $f(x|\theta)$ and so $\widehat{x}$ is the mode of $f(\cdot|\theta)$.

Hence we have dealt with the loss functions connecting θ with M_0 and X. Finally we look at loss functions connecting θ and π. We must note that however we select π the ultimate aim is to use the posterior as a representation of the information we have in the form of a belief probability on Θ. Thus implicitly we are thinking about θ^*. Hence to us $\pi(\theta)$ is a belief distribution about this particular value of θ. Thus $l(\theta, \pi)$ should be equivalent in structure to the loss for $l(\theta, X)$ but with the roles switched. Hence,

$$l(\theta, \pi) = -\log \pi(\theta).$$

Here now π plays the role of θ, and θ the role of X, in $l(\theta, X) = -\log f(X|\theta)$. This can be seen more clearly if we index π with a

parameter ϕ and write

$$l(\theta, \pi) = l(\phi, \theta) = -\log \pi(\theta|\phi).$$

But as we have mentioned earlier, loss functions connecting outcomes make sense whichever one regards as the outcome of choice or of uncertainty.

In summary, the building blocks for loss functions considered in this paper are derived from the two following standard or benchmark loss functions:

1. If x is modeled via $f(x|\theta)$ and x is a possible outcome, then the logarithmic loss function connecting θ and x is given by

   $$l(\theta, x) = -\log f(x|\theta).$$

   If M_0 represents the current belief about the distribution of x, then the expected loss is

   $$l(\theta) = -\int \log f(x|\theta)\, M_0(\mathrm{d}x).$$

   Effectively then, this loss is measuring the Kullback–Leibler divergence between $M_0(\cdot)$ and $f(\cdot;\theta)$.

2. If $\pi(\theta)$ expresses beliefs about a θ, then the *self–information* loss function connecting θ and π is given by

   $$l(\theta, \pi) = -\log \pi(\theta).$$

   See, for example, Merhav (1998) for details about this loss function and use.

Whenever we write a loss function it must be noted that these are defined, at least for us, up to scalar and additive constants; so in reality, even though we write, for example, $l(\theta, \pi) = -\log \pi(\theta)$, for some constants α and β, not depending on θ, we could have $l(\theta, \pi) = a - \beta \log \pi(\theta)$.

We obtain loss functions for the pieces of information available. A loss function for the information that $f(\cdot|\theta)$ has been chosen to act as the model will be referred to as $l_M(\theta)$, and developed in Section 2.1, whereas a loss function for the information provided by (c, M_0) shall be referred to as $l_N(\theta)$ and developed in Section 2.2.

For more on loss functions in statistical decision theory, see Berger (1993).

1.2 Bayesian nonparametric prior. Another key component of our idea is the use of a nonparametric prior which is acting as the "true"

model for the observations. This is the foundation for the work in Gutiérrez–Peña and Walker (2005). The point is that one can make low dimensional decisions or inference within a large or nonparametric framework. The nonparametric model used was the Dirichlet process (Ferguson, 1973) and we briefly describe it here.

The prior Dirichlet process is characterized by parameters (c, M_0) and generates random distribution functions F such that $\mathrm{E}(F) = M_0$ and

$$\mathrm{Var}(F(A)) = \frac{M_0(A)M_0(A^c)}{c+1},$$

where A^c is the complementary set of A. The posterior given $(x_i)_{i=1}^n$ is also a Dirichlet process with updated parameters $(c+n, M_n)$, where

$$M_n(\mathrm{d}x) = \frac{c\,M_0(\mathrm{d}x) + n\,P_n(\mathrm{d}x)}{c+n},$$

and P_n is the empirical distribution of the observations. The sequence of posterior distributions is always consistent in the sense that for any suitable set A it is that $\mathrm{E}[F(A)|X_1, \ldots, X_n] \to F_0(A)$ a.s. and

$$\mathrm{Var}[F(A)|X_1, \ldots, X_n] \to 0 \quad \text{a.s.}$$

Hence, the Dirichlet process prior can be thought of as a "true" model.

An issue with the Dirichlet process is that it only generates discrete distribution functions. Hence, if the overall target is density estimation, for example, then this is a problem. However, if the overall aim is one of decision making via the use of utility or loss functions, then this discreteness is irrelevant because we take expectations of quantities with respect to M_n.

Suppose we wish to select an action $a \in \mathcal{A}$, the best action being known if the distribution generating the (x_i) is known and a loss function is in place measuring the loss in taking action a when F is the true distribution function. Call this loss function $l(a, F)$. With beliefs about F being represented by a posterior distribution $\Pi_n(\mathrm{d}F)$, then the best action is to select the action a which minimizes

$$l(a) = \int l(a, F)\,\Pi_n(\mathrm{d}F).$$

Thus, in particular, if $l(a, F) = \int l(a, x)\,\mathrm{d}F(x)$, where $l(a, x)$ is a loss function directly connecting the action with observable x, and we use the Dirichlet process model, then

$$l(a) = \int l(a, x)\,\mathrm{d}M_n(x).$$

The basis of the work in Gutiérrez–Peña and Walker (2005) is that action a is a statistical decision, such as parameter estimation or model selection.

In fact, the point of the paper by Gutiérrez–Peña and Walker (2005) was to discuss the incoherence of Bayesian model selection as it currently stands. The incoherence is due, ironically, to the lack of a prior. There is no prior for Bayesian model selection problems, i.e. there is no single assessment of prior uncertainty. Gutiérrez–Peña and Walker (2005) resolve this issue by using a Dirichlet process prior as the prior and then deal with issues such as Bayesian model selection using decision theory with respect to the posterior Dirichlet process. In this way, and undertaken in this framework, the incoherence disappears. For example, if k indexes a number of possible models with θ_k denoting the parameter for model $f_k(\cdot|\theta_k)$, then the best model can be selected based on the loss function

$$l(k) = \min_{\theta_k \in \Theta_k} \left\{ - \int \log f_k(x|\theta_k)\, M_n(\mathrm{d}x) + \gamma(p_k) \right\}$$

where p_k is the dimension of model k and $\gamma(p)$ a function which penalizes high dimensional models. The idea is that the best model is the one with a parameter which takes the family of densities closest to F_0 with respect to the Kullback–Leibler divergence. And in the traditional approach, F_0 is replaced by the current best guess which is clearly M_n.

Here we discuss more explicitly the role of c. We are not using the Dirichlet process to model the data as a final goal; it is being used to make decisions and as such its use appears solely in the form $\int l(a,x)\,\mathrm{d}M_n(x)$. The role of c is ambiguous in general, but its use in making decisions is less so. For $c = 0$, the most controversial choice simply yields M_n as the empirical distribution function P_n. This is hardly any cause for concern. This is natural as we can use c literally as a prior sample size by recalling that M_n is a weighted mixture of the empirical and M_0, with the weighting determined by c. Hence, we see no conceptual or practical problems with the choice of c. See also Walker and Mallick (1997).

The paper is laid out as follows: In Section 2 we detail our idea for prior construction based on the notion of matching loss functions. Section 3, 4 and 5 then illustrate the approach for a range of models including regression and hierarchical. Section 6 concludes with a discussion. In the Appendix we provide a brief review of many popular and current approaches to the construction of prior distributions.

2 Priors from loss matching. The idea for matching of loss functions is quite straightforward. Since π is to encapsulate all the information from $[(c, M_0), f_\Theta]$, then the requirement must be that, up to

additive and scalar constants, the loss functions must match, i.e.

$$l(\theta, \pi) = l_N(\theta) + l_M(\theta) = l(\theta, (c, M_0)|f_\Theta) + l(\theta, f_\Theta). \tag{1}$$

We see this as a highly appropriate means by which to deduce π from $[(c, M_0), f_\Theta]$.

This section breaks into a number of sub-sections. In Section 2.1 we provide the loss function $l_M(\theta)$, in Section 2.2 the loss function $l_N(\theta)$, and in Section 2.3 we consider the loss function derived from the choice of prior $\pi(\theta)$. Section 2.4 then puts all these together to derive a choice for $\pi(\theta)$ and Section 2.5 relates our results to the existing literature.

2.1 Loss from choice of model. The first point, which is presumably well known, is that it is imperative to construct any prior on Θ with reference to the family $f(x|\theta)$, and not just to Θ. We start by considering the utility of the density $f(\cdot|\theta)$ for a particular θ as it sits in the family of densities. This may be motivated first by discretizing the densities so $f_j(\cdot) = f(\cdot|\theta_j)$ for a set of discrete $(\theta_j) \in \Theta$.

To illustrate this, we consider a simple and extreme case, when three densities have been chosen only, say (f_1, f_2, f_3) to model the density generating the outcomes. These have been chosen in such a way that f_1 and f_2 are barely indistinguishable from each other, yet f_3 is far from these two. One can imagine that there is knowledge that has been used to construct a model in this way.

However, something concrete is needed and we believe this is to be found using notions of utility functions (equivalently loss functions but it is convenient to come at this from utility functions to start with). We would assess the utility of f_3 to be greater than that of either f_2 or f_1. It would be more serious to lose f_3 from the model than it would be to lose either f_2 or f_1. And this utility can only be made concrete by considering how close each f_j is to its neighbours.

If we were to remove f_j from the model, and it was the true density, then we would lose the Kullback–Leibler distance from f_j to its nearest density. This is because the Bayesian model would eventually put all the mass on the density closest to f_j with respect to the Kullback–Leibler divergence (see Berk, 1966), hence this is the loss. Thus, the utility of f_j would be of the type

$$u(f_j) = \inf_{k \neq j} D(f_j, f_k),$$

where $D(f, g) = \int f \log(f/g)$ is the Kullback–Leibler divergence between f and g.

When we have a continuum of densities, and indexed by a parameter θ, then for each $\epsilon > 0$ we would consider the utility function for θ of the type

$$u_\epsilon(\theta) = D(f_\theta, f_{\theta+\epsilon}).$$

This would arise by assuming we have a model with discrete θ and the nearest neighbour is at $\theta + \epsilon$. We would need this to have a limiting form as $\epsilon \to 0$ to get something out in the continuum. Hence, we would need to take

$$u_\epsilon(\theta) = \epsilon^{-2}\, D(f_\theta, f_{\theta+\epsilon})$$

and if θ is p–dimensional then the limit as $\epsilon \to 0$, (Blyth, 1994), is given by

$$u(\theta) = \sum_{1 \le j,k \le p} I_{jk}(\theta),$$

where

$$I_{jk}(\theta) = \mathrm{E}\left(\frac{\partial}{\partial \theta_j} \log f(x|\theta)\ \frac{\partial}{\partial \theta_k} \log f(x|\theta) \right).$$

Hence, reinterpreting this as a loss function, and putting on the log–scale which is where we will be operating, we would use

$$l_M(\theta) = -\log \left(\sum_{1 \le j,k \le p} I_{jk}(\theta) \right). \tag{2}$$

Of course if $p = 1$ then we have

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right]$$

and so $l_M(\theta) = -\log I(\theta)$.

This idea for extracting information from a choice of model deserves close inspection and is currently being investigated by a PhD student of the second author.

2.2 Loss from choice of (c, M_0). Following Gutiérrez–Peña and Walker (2005) and Section 1.1, we assess the loss at (θ, x) to be given by the logarithmic loss function:

$$l(\theta, x) = -\log f(x|\theta).$$

This is standard and even forms the basis for classical estimation via maximum likelihood, since based on a sample of size n, the cumulative loss would be

$$l(\theta, x_1, \ldots, x_n) = -\sum_{i=1}^{n} \log f(x_i|\theta)$$

and minimizing this yields the maximum likelihood estimator.

Hence, if M_0 represents beliefs about the distribution of x, then the expected loss is precisely

$$l_N(\theta) = -\int \log f(x|\theta)\, M_0(\mathrm{d}x). \tag{3}$$

We will see the role of c when we put all the loss functions together in Section 2.4.

2.3 Loss from choice of prior distribution. We now think about a loss function for (θ, π), where the idea is that π is now being used to encapsulate the information in $[(c, M_0), f_\Theta]$ in a probability density function on Θ. If $\pi(\theta)$ is to represent beliefs about which value of θ is best, for us to be able to make statements such as $\mathrm{P}(\theta \in A) = \pi(A)$, we need to have $\pi(\theta)$ as the belief distribution function for θ. Hence, if we construct a loss for θ based on the choice of $\pi(\theta)$, we would need to use the "honest" loss function $-\log \pi(\theta)$, see Bernardo (1979a). His Theorem 2 shows this to be the unique local proper scoring rule. See also Section 1.1.

2.4 Matching the loss functions. We have relevant information $I = [(c, M_0), f_\Theta]$ and we have a loss function connecting this information for each value of θ, this is, up to additive and scalar constants,

$$l_M(\theta) = l(\theta, f_\Theta) \quad \text{and} \quad l_N(\theta) = l(\theta, (c, M_0)|f_\Theta).$$

The loss functions for the model f_Θ and $[(c, M_0)|f_\Theta]$ are cumulative and provide a loss for each θ based on these pieces of information, i.e. putting arbitrary scalars into (**??**),

$$l_I(\theta) = \beta l_N(\theta) + \gamma l_M(\theta),$$

where $\beta > 0$ and $\gamma > 0$ are as yet undefined constants, and l_M and l_N are given in equations (**??**) and (**??**), respectively. We will work out appropriate values for them.

On the other hand, the loss for the choice of $\pi(\theta)$, which is supposed to encapsulate the information I in a probability form, is given by $l_\pi(\theta) = -\log \pi(\theta)$.

We have two loss functions for the same θ, and hence they must match. Therefore, for some α,

$$-\log \pi(\theta) = \alpha - \beta \int \log f(x|\theta)\, M_0(\mathrm{d}x) - \gamma \log \left(\sum_{1 \le j,k \le p} I_{jk}(\theta) \right).$$

Hence, our choice of prior is

$$\pi(\theta) \propto J(\theta)^\gamma \exp \left\{ \beta \int \log f(x|\theta)\, M_0(\mathrm{d}x) \right\},$$

where we have written

$$J(\theta) = \sum_{1 \le j,k \le p} I_{jk}(\theta).$$

In order to understand the role of c and β here, we need to move forward, applying Bayes theorem at overarching Dirichlet level, and look at the posterior distribution. This is given by

$$\pi(\theta|x_1,\ldots,x_n) \propto J(\theta)^{\gamma} \exp\left\{(\beta+n)\int \log f(x|\theta)\, M_n(\mathrm{d}x)\right\},$$

where

$$M_n(\mathrm{d}x) = \frac{\beta M_0(\mathrm{d}x) + n\, P_n(\mathrm{d}x)}{\beta+n}.$$

Within the embedded framework we cannot utilize Bayes theorem directly, so to see how this can be derived from the matching of loss functions and the Bayes theorem applied to the Dirichlet process, it is worth returning to the discussion about Bayesian nonparametric inference using the Dirichlet process prior. If we are interested in the loss

$$l(\theta, F) = -\int \log f(x|\theta)\, F(\mathrm{d}x)$$

and model F with a Dirichlet process prior with parameters (c, M_0), then the posterior expected loss is given by

$$l(\theta) = -\int \log f(x|\theta)\, M_n(\mathrm{d}x),$$

where now the $M_n(\mathrm{d}x)$ is as given before except with $\beta = c$.

Using the match of loss functions at this point, we obtain, for some α_n and $\beta_n > 0$,

$$-\log \pi(\theta|x_1,\ldots,x_n) = \alpha_n - \beta_n \int \log f(x|\theta)\, M_n(\mathrm{d}x) - \gamma \log J(\theta).$$

Consequently, we obtain coherence for the procedure, and based on a necessary cumulative loss function, only if we take $\beta_n = \beta + n$ and $\beta = c$.

For the choice of γ, we do not need this to change with n as the choice of model does not alter with the sample size. This parameter is difficult to assess, yet if we consider $p = 1$, then $J(\theta)$ is the Fisher information and hence the choice of $\gamma = \frac{1}{2}$ yields the Jeffreys prior (Jeffreys, 1946), which is a standard choice of objective prior from many perspectives, and which includes an invariance property. We will use this value in the examples which follow in Section 3.

Hence, the prior, based on choices $I = [(c, M_0), f_\Theta]$ is given by

$$\pi(\theta) \propto J(\theta)^{\frac{1}{2}} \exp\left\{c \int \log f(x|\theta)\, M_0(\mathrm{d}x)\right\}.$$

It is to be noted the model is conjugate. As data accumulate the c changes to $c+n$ and the M_0 changes to

$$M_n = \frac{cM_0 + nP_n}{c+n}$$

which are the updates for the Dirichlet process prior. The coherence of the procedure based on the matching of loss functions follows since it coincides with an application of Bayes theorem: The $l_N(\theta)$ is changing from

$$-c \int \log f(x|\theta)\, M_0(\mathrm{d}x)$$

to

$$-(c+n) \int \log f(x|\theta)\, M_n(\mathrm{d}x)$$

which is a consequence of revised expected loss to concur with current updated beliefs in light of the data $(x_1, \ldots, x_n)$. All that is required is an ability to express beliefs about the distribution of the initial observable.

Note that we can write the prior as

$$\pi(\theta) \propto J(\theta)^{\gamma} \, \exp\left\{-c\, D(M_0(\cdot), f(\cdot|\theta))\right\}$$

and so the prior puts more weight to those θ which make the Kullback–Leibler divergence between $f(\cdot|\theta)$ and M_0 small (and only this when $\gamma = 0$). This is what is achieved by the data, since one can write the likelihood function as, with a mild abuse of notation,

$$\prod_{i=1}^{n} f(x_i|\theta) \propto \exp\left\{-n\, D(P_n(\cdot), f(\cdot|\theta))\right\}.$$

It makes sense then to have this aspect a part of the prior distribution. Our aim is to have it as the key part of the prior.

We are content to apply Bayes theorem for the Dirichlet process model. It is a true model, as discussed earlier, and hence there does exist an F, a distribution function conditional on which the data are independent and identically distributed. We may be more circumspect about applying Bayes theorem to the model $f(\cdot|\theta)$ as it is not the case that there is a θ for which the observations are independent and identically distributed from $f(\cdot|\theta)$. However, through the application of Bayes theorem for the Dirichlet process and the coherent application of loss matching we can derive the Bayes rule for the parametric family.

2.5 Relation to the literature. Looking through the literature, we find this idea is in the same spirit as an idea appearing in Barron (1998). Effectively, Barron (1998) is matching two loss functions for θ. An

asymptotic expression for the Kullback–Leibler divergence (Kullback and Leibler, 1951) between

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

and the marginal joint density

$$p(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} f(x_i|\theta)\, \pi(\mathrm{d}\theta)$$

is given by

$$D\big(f(x_1, \ldots, x_n|\theta), p(x_1, \ldots, x_n)\big) = K_n + \log\{|I(\theta)|^{1/2}/\pi(\theta)\} + o(1),$$

where K_n does not depend on θ and $I(\theta)$ is the Fisher information matrix.

Re–arranging this, we see we have

$$-\log \pi(\theta) = D\big(f(x_1, \ldots, x_n|\theta), p(x_1, \ldots, x_n)\big) - \tfrac{1}{2}\log|I(\theta)|,$$

which is very similar to our matching of loss functions. However, we have something different in the multipara–meter case, replacing $I(\theta)$, and (effectively) instead of our subjective $D\big(M_0(\cdot), f(\cdot|\theta)\big)$, Barron has $a(\theta) = D\big(f(x_1, \ldots, x_n|\theta), p(x_1, \ldots, x_n)\big)$ for which an objective choice is sought.

If it is possible to assess what $a(\theta)$ should be, then this would result in the choice of prior as

$$\pi(\theta) \propto \sqrt{|I(\theta)|}\, e^{-a(\theta)}.$$

However, as noted by Sweeting (1998) in the discussion of Barron (1998), it seems a non–trivial task to find a suitable or well motivated choice for $a(\theta)$.

Nevertheless, the idea we present for constructing prior distributions is closely related to the idea of Barron (1998). Effectively it is the same if one views Barron's idea is indeed matching loss functions.

Prudent observers will note we are breaking a supposed prior construction rule which is that

$$m(x) = \int f(x|\theta)\, \pi(\theta)\, \mathrm{d}\theta,$$

where $m(x)$ is the density function corresponding to the distribution $M_0(x)$. But this rule is only valid if it is thought that for some θ the x is coming from $f(x|\theta)$. It is a law of total probability statement

and required to hold for the Bayesian update via Bayes Theorem. But we do not rely on this for the update as we can do it via the match of loss functions. While we believe $m(x)$ to be the best choice for the initial distribution of x, once we have constructed $\pi(\theta)$ via the matching of losses, there is no reason whatsoever to now believe that $\int f(x|\theta)\,\pi(\mathrm{d}\theta)$ is also this initial belief since we do not connect x and θ through a probability model $f(x|\theta)$, but rather through a loss function $-\log f(x|\theta)$.

We will discuss further aspects in Section 6. In the next section we will consider the application for independent and identically distributed observations. In Section 4 we consider the application to regression models and in Section 5 to hierarchical models.

3. Illustrations. For a variety of models we now consider the priors constructed through the matching of loss functions approach.

3.1 Normal model. Here we consider the case when $\theta = (\mu, \lambda)$ and

$$f(x|\theta) \propto \lambda^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\lambda(x-\mu)^2\right\}.$$

If the choice of $m(x)$ is Normal with mean ν and variance σ^2, then

$$\int_{\mathbb{R}} \log f(x|\theta)\, m(x)\, \mathrm{d}x = K + \tfrac{1}{2}\log\lambda - \tfrac{1}{2}\lambda(\mu^2+\sigma^2),$$

where K does not depend on θ.

It is also easy to verify that $J(\theta) = (\lambda + \frac{1}{2}\lambda^{-2})$, thus we have

$$\pi(\mu,\lambda) \propto [\lambda + \tfrac{1}{2}\lambda^{-2}]^{\frac{1}{2}}\lambda^{c/2} \exp\left\{-\tfrac{1}{2}c\lambda(\mu^2+\nu^2+\sigma^2-2\mu\nu)\right\}.$$

The conjugacy works on the parameters (c, ν, s^2), where $s^2 = \nu^2 + \sigma^2$. Then it is easy to see that $c \to c+n$,

$$\nu \to \frac{c\nu + n\bar{x}}{c+n}$$

and

$$s^2 \to \frac{cs^2 + \sum_i x_i^2}{c+n}$$

where $\bar{x}$ is the sample mean.

3.2 Bernoulli model. Here we have

$$\log f(x|\theta) = x\log\theta + (1-x)\log(1-\theta)$$

with $x \in \{0,1\}$ and $0 < \theta < 1$. If $m(1) = p$ then it is easy to see that

$$\sum_{x\in\{0,1\}} [x\log(1-\theta) + (1-x)\log(1-\theta)]m(x) = p\log\theta + (1-p)\log(1-\theta).$$

Given that $J(\theta) = \theta^{-1}(1-\theta)^{-1}$, we would have

$$\pi(\theta) \propto \theta^{cp-\frac{1}{2}}(1-\theta)^{c(1-p)-\frac{1}{2}}.$$

The conjugacy here operates on c and p and $c \to c+n$ and

$$p \to \frac{cp + n\bar{x}}{c+n}.$$

3.3 Poisson model. Here we have $\log f(x|\theta) = K + x\log\theta - \theta$, where K does not depend on θ, with $x \in \{0, 1, 2, \ldots\}$ and $\theta > 0$. So

$$\sum_x \log f(x|\theta)\, m(x) = K + \mu\log\theta - \theta$$

where μ is the prior guess at the mean of x. So, given that $J(\theta) = \theta^{-1}$ we have

$$\pi(\theta) \propto \theta^{c\mu-\frac{1}{2}}\ \exp(-c\theta).$$

The pattern for conjugacy is now clear and so $c \to c+n$ and

$$\mu \to \frac{c\mu + n\bar{x}}{c+n}.$$

3.4 Gamma model. Here we have

$$f(x|\theta) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-xb},$$

where $\theta = (a, b)$ and $a, b > 0$. Then

$$\log f(x|\theta) = a\log b - \log\Gamma(a) + (a-1)\log x - xb$$

and so

$$\int_X \log f(x|\theta)\, m(x)\,\mathrm{d}x = a\log b - \log\Gamma(a) + (a-1)\xi - \mu b$$

where ξ is the guess at the expected value of $\log x$ and μ the guess at the expected value of x. Some calculations give

$$J(a, b) = \psi(a) - 2/b + a/b^2,$$

where $\psi(a)$ is the tri–gamma function. This $J(a, b)$ is positive since

$$J(a, b) = \psi(a) - \frac{1}{a} + \left(\frac{1}{\sqrt{a}} - \frac{\sqrt{a}}{b}\right)^2$$

and $\psi(a) > 1/a$. So

$$\pi(a,b) \propto J(a,b)^{\frac{1}{2}} \, \frac{b^{ca}}{\Gamma(a)^c} \xi^{ca} e^{-c\mu b}.$$

3.5 Exponential family. Here we consider the exponential family; so

$$f(x|\theta) = c(x) \exp\{x\theta - b(\theta)\}.$$

If we consider $m(x)$ as the prior guess for the density of x, and $\xi = \int x m(x)\,\mathrm{d}x$, then, given that

$$J(\theta) = b''(\theta),$$

we would take the prior as

$$\pi(\theta) \propto \sqrt{b''(\theta)}\, \exp\{c\xi\theta - cb(\theta)\}.$$

Conjugacy here is that $c \to c + n$ and

$$\xi \to \frac{c\xi + n\bar{x}}{c+n}.$$

4. Regression models. In this section we extend our idea of loss function matching to regression models. We will write the models as $f(y|x,\theta)$ where now y is the dependent variable and x the independent variable and θ the parameter of interest to which a prior is to be assigned. We will consider the priors for θ under which the x are generated stochastically with known density function $m(x)$.

We will as usual take $l_\pi(\theta) = -\log \pi(\theta)$ and for the model part we will take logarithmic loss function; so if

$$l(y,x,\theta) = -\log f(y|x,\theta),$$

then the expected Kullback–Leibler loss is with respect to the distribution assigned to (y,x), with density $m(y,x)$, so

$$l_N(\theta) = -\int\int \log f(y|x,\theta)\, m(y,x)\,\mathrm{d}y\,\mathrm{d}x.$$

On the other hand, we would have

$$l_M(\theta) = -\log \int \sum_{1\le j,k\le p} I_{jk}(\theta,x)\, m(x)\,\mathrm{d}x,$$

with an obvious interpretation of $I_{jk}(\theta,x)$. Therefore, we have

$$\pi(\theta) \propto J(\theta)^{1/2} \exp\left(c\int\int \log f(y|x,\theta)\, m(y,x)\,\mathrm{d}y\,\mathrm{d}x\right),$$

where

$$J(\theta) = \exp\{-l_M(\theta)\}.$$

As before, this is a conjugate prior, since the posterior density for θ is given by

$$\pi(\theta|(x_1, y_1), \ldots, (x_n, y_n)) \propto J(\theta)^{1/2} \exp\left(c_n \int\int \log f(y|x, \theta)\, \mathrm{d}M_n(y, x)\, \mathrm{d}y\, \mathrm{d}x\right),$$

where $c_n = c + n$ and

$$M_n(y, x) = \frac{cM(y, x) + nP_n(y, x)}{c + n}.$$

4.1 Normal regression model. Here we consider a normal example, so $\theta = (\alpha, \beta, \lambda)$, where

$$f(y|x, \theta) = \mathrm{N}(y|\alpha + \beta x, \lambda),$$

a normal distribution with mean $\alpha + \beta x$, x being a real scalar, and variance λ^{-1}. To obtain the prior we need to find the expectation of

$$\tfrac{1}{2}\log\lambda - \tfrac{1}{2}\lambda(y - \alpha - \beta x)^2,$$

with respect to $m(y, x)$, which is given by

$$\tfrac{1}{2}\log\lambda - \tfrac{1}{2}\lambda\left\{(\mu_y - \alpha - \beta\mu_x)^2 + \beta\sigma_x^2 - 2\beta\rho_{xy} + \sigma_y^2\right\},$$

where μ_x is $\mathrm{E}(x)$ and σ_x^2 is $\mathrm{Var}(x)$; μ_y is the prior choice for $\mathrm{E}(y)$, σ_y^2 is the prior choice for $\mathrm{Var}(y)$ and ρ_{xy} is the prior choice for the $\mathrm{Cov}(x, y)$. Therefore,

$$\pi(\theta) \propto J(\alpha, \beta, \lambda)^{1/2}\lambda^{c/2} \exp\left[-c\tfrac{1}{2}\lambda\left\{(\mu_y - \alpha - \beta\mu_x)^2 + \beta\sigma_x^2 - 2\beta\rho_{xy} + \sigma_y^2\right\}\right],$$

where

$$J(\alpha, \beta, \lambda) = 2\lambda^{-2} + \lambda + \lambda\int x^2 m(x)\, \mathrm{d}x.$$

4.2 Bernoulli regression model. Now we consider a Bernoulli model for y, whereby for some θ we have

$$\Pr(y = 1|x, \theta) = \frac{e^{\theta x}}{1 + e^{\theta x}}.$$

Therefore,

$$\begin{aligned}
&\textstyle\sum_{y\in\{0,1\}} \int \log f(y|x, \theta)\, m(y, x)\, \mathrm{d}x \\
&= \int \left\{\theta x - \log(1 + e^{\theta x}) m(1|x)\, m(x)\right\} \mathrm{d}x + \int -\log(1 + e^{\theta x})[1 - m(1|x)]\, m(x)\, \mathrm{d}x \\
&= \int \theta\, x\, m(1|x)\, m(x)\, \mathrm{d}x - \int \log(1 + e^{\theta x})\, m(x)\, \mathrm{d}x.
\end{aligned}$$

Hence,

$$\pi(\theta) \propto J(\theta)^{1/2} \exp\left\{c\theta\xi - c\int \log(1+e^{\theta x})\, m(x)\, \mathrm{d}x\right\},$$

where ξ is the prior choice for $\int x\, m(1|x)\, m(x)\, \mathrm{d}x$, and

$$J(\theta) = \int \frac{x^2 e^{x\theta}}{1+e^{x\theta}}\, m(x)\, \mathrm{d}x.$$

5. Hierarchical models. The form of model we consider here is in hierarchical form given by

$$y_i|x_i \sim f_1(y_i|x_i,\theta)$$

$$x_i \sim f_2(x_i|\phi)$$

where (θ,ϕ) are the parameters and (x_i) are unobserved random effects. This model, if we integrated out the (x_i), would not yield independent (y_i) and hence the unsuitability of cumulative loss in this case. Moreover the integration is often intractable. Hence, we can and must consider a conditional loss function.

It is easy to see that $l(\theta;x,y) = -\log f_1(y|x,\theta)$ and therefore if $m(y,x)$ represents the a priori guess for the joint density of (y,x), then we would take, for some $c_1 > 0$,

$$\pi(\theta) \propto J_1(\theta)^{1/2} \exp\left\{c_1 \int\int \log f_1(y|x,\theta)\, m(y,x)\, \mathrm{d}y\, \mathrm{d}x\right\},$$

where $J_1(\theta)$ is defined as

$$J_1(\theta) = \int \sum_{1\le j,k\le p_1} I_{1jk}(\theta,x)\, m(x)\, \mathrm{d}x$$

with $I_{1jk}(\theta,x)$ being the elements of the Fisher information matrix based on $f_1(y|x,\theta)$, and $m(x) = \int m(y,x)\, \mathrm{d}y$.

The second level of the hierarchy can be dealt in the usual way, so for some $c_2 > 0$, we would have

$$\pi(\phi) \propto J_2(\phi)^{1/2} \exp\left\{c_2 \int \log f_2(x|\phi)\, m(x)\, \mathrm{d}x\right\},$$

where

$$J_2(\phi) = \sum_{1\le j,k\le p_2} I_{2jk}(\phi),$$

with $I_{2jk}(\phi)$ being the elements of the Fisher information matrix based on $f_2(x|\phi)$.

These are conjugate type prior distributions, in that we can easily obtain $\pi(\theta|(x_i, y_i)_{i=1}^n)$ and $\pi(\phi|x_1, \ldots, x_n)$. The updates are, respectively, $c_1 + n$ and

$$\frac{c_1 \mathrm{d}M(y,x) + n\mathrm{d}P_n(y,x)}{c_1 + n},$$

and $c_2 + n$ and

$$\frac{c_2 \mathrm{d}M(x) + n\mathrm{d}P_n(x)}{c_2 + n}.$$

Although the (x_i) are not observed, within a Gibbs sampler algorithm, these are key conditional distributions. To complete the sampler, one would need to evaluate and sample, for each i, $f(x_i|y_1, \ldots, y_n, \theta, \phi)$.

6. Discussion. We have shown how the idea of matching loss functions, based on the self–information loss and Kullback–Leibler information loss functions, lead to the probability belief distribution being represented as

$$-\log \pi_n(\theta) = K_n + (c+n) \int \log f(x|\theta)\, M_n(\mathrm{d}x) - \gamma \log J(\theta),$$

which holds for all $n \geq 0$. This is the Bayesian rule applied to the model $f(\cdot|\theta)$ for which the direct application of Bayes theorem can be seen as problematic. The Bayes theorem hiding in the background has been applied to the Dirichlet process, this we are happy to apply due to the true model status we have allocated to it. It is a consistent model for all sampling distributions. The loss matching idea we can view as a justification for the Bayes rule in the parametric case. So we have no need for the formal Bayes theorem for the parametric model and all the restrictive assumptions that need to be made to effect it. While $f(x|\theta)$ is a probability density function, it enters the learning mechanism only through its role as providing a loss function $-\log f(x|\theta)$. The precise rôle is to measure its loss with respect to beliefs about the distribution of the observables, using the Kullback–Leibler divergence.

All that is required to start this process is a prior choice for the $M_0(x)$ and a measure of precision with this, i.e. some scalar $c > 0$. Bayesians should not have any qualms about specifying such objects, or should question whether they really are Bayesians at all. And these are precisely the quantities needed in any Bayesian nonparametric application. It is also interesting to note that there are no objective priors that have been presented in the Bayesian nonparametric literature. So a (c, M_0) must be specified in such models.

6.1 Model selection. Another issue is the model selection problem. Suppose for each k there is a model to be considered based on $f_k(x|\theta_k)$, with $\theta_k \in \Theta_k$. The ambition for each parameter θ_k would be

to learn about the value θ_{k0} which takes the family $f_k(\cdot|\theta_k)$ closest in the Kullback–Leibler sense to the true density function. A measure of this discrepancy is precisely what we are using as the loss function

$$L(\theta_k) = -\int \log f_k(x|\theta_k)\, M_n(\mathrm{d}x),$$

where M_n is the current estimate for the true distribution function. Hence, if $\widehat{\theta}_k$ minimizes $L(\theta_k)$ then we would select the k which minimizes $L(k) = L(\widehat{\theta}_k)$. To us this seems a natural decision rule. For large n, $\widehat{\theta}_k$ would be approximately the maximum likelihood estimator in Θ_k since M_n is approximately the empirical distribution function.

6.2 Predictive density. An interesting point of discussion is what does the predictive

$$f_n(x) = \int f(x|\theta)\, \pi_n(\mathrm{d}\theta)$$

represent. Given the wrong model is being used, and known to be wrong, but used for pragmatic purposes, nevertheless we have constructed $\pi_n(\theta)$ to represent beliefs about which θ is getting us closest in the Kullback–Leibler sense to the correct predictive $M_n(\mathrm{d}x)$. Hence, to us, $f_n(x)$ is merely the expected density averaged over the family $f(x|\theta)$ using $\pi_n(\theta)$. It therefore has no special interpretation other than an estimate of the sampling density.

6.3 Objective prior. With the choice of $c = 0$, we have no M_0 entering the model at all. So now effectively the prior becomes

$$\pi(\theta) \propto J(\theta)^{\frac{1}{2}}$$

which we could claim as an objective prior. As far as we are aware, this has not been previously proposed as an objective prior, though it does coincide with the Jeffreys prior in the case $p = 1$.

Yet its claim for an objective prior are good. It extracts the information inherent in the fact that the model $f(\cdot|\theta)$ has been chosen to model the data, and this idea appears new. We are not trying to be minimally informative or ignorant, but using the available information in that the model has been chosen

For $p > 1$ we are not able to consider invariance under one-to-one re–parameterizations. Datta and Ghosh (1996) investigate the invariance, or lack of, for various noninformative or objective priors under re-parameterizations. They establish the invariance under certain re–parameterizations for particular priors and lack of invariance for others. The message would appear to be that while invariance is a desirable property it is by no means an overriding one.

6.4 Asymptotics. We can write the posterior distribution as

$$\pi(\theta|x_1,\ldots,x_n) \propto J(\theta)^{\frac{1}{2}} \exp\left\{-(c+n)\, D(M_n(\cdot), f(\cdot|\theta))\right\}.$$

Here we provide heuristic discussions of the asymptotics; models that do not behave as indicated below will be unusual ones. Now M_n converges a.s. to F_0 and so the posterior will accumulate about θ^* and one may recall that this is the parameter which minimizes the Kullback–Leibler divergence between F_0 and $f(\cdot|\theta)$. See also Berk (1966) and Bunke and Milhaud (1998). If this is where the posterior will end up then it is natural to think about this result when constructing the prior and a prior of the form

$$\pi(\theta) \propto J(\theta)^{\frac{1}{2}} \exp\left\{-c\, D(M_0(\cdot), f(\cdot|\theta))\right\}$$

is doing precisely this as M_0 represents the initial belief about the value of F_0. That is the further $f(\cdot|\theta)$ is away from M_0 the lower the weight attributed to that value of θ.

6.5 Summary. If it is acknowledged that $f(x|\theta)$ and F_0 are different then one is interested in the θ which maximizes

$$U(\theta) = \int \log f(x|\theta)\, \mathrm{d}F_0(x).$$

There are two unknowns, the minimizing θ value and F_0. If we assign a Dirichlet process prior for F, with parameters (c, M_0), then we need a compatible prior for θ, $\pi(\theta)$, which accounts for all the interpretations associated with the desire to find the maximizer of $U(\theta)$. This we believe has the key component

$$\pi(\theta) \propto \exp\left\{c \int \log f(x|\theta)\, M_0(\mathrm{d}x)\right\},$$

which clearly assigns greater density value to those θ making the prior estimate of $U(\theta)$ large. The procedure updates naturally, using the likelihood and Dirichlet process update, to

$$\pi(\theta|P_n) \propto \exp\left\{(c+n) \int \log f(x|\theta)\, M_n(\mathrm{d}x)\right\},$$

where M_n is now the best guess for F_0 once the data have been seen.

Appendix. Since we advocate a particular choice of prior, here we review the main current ideas for prior construction. We start with subjective or personal probabilities.

A.1 Subjective prior. Perhaps the dominant theory in recent years has been that which acknowledges explicitly and transparently the rôle of subjectivity in statistical inference. The seminal paper of Ramsey (1926, 1964) developed the skeleton of a theory relying on *coherent* betting behaviour and the existence of a canonical 'ethically neutral' event. Independently, de Finetti (1937, 1964), coming from an actuarial background in Italy, showed how the notion of symmetry of beliefs or *exchangeability* gives rise to the duality of model and prior $\{f(x|\theta), \pi(\theta)\}$ so that for de Finetti probability is assigned to observables, and unknown parameters merely emerge as constructs from his theorem on exchangeability. Again, coherence, or the non existence of bets such that an individual is sure to win is crucial to the interrogation of beliefs, and has been developed by a number of authors since. See Lindley (1972) for an insightful review, and French (1982), for the axiomization of subjective probability,

These pure subjectivist views do not countenance using data in hand for mixing with the assignment process. Many of the alternative methods of assigning prior probabilities for $\pi(\theta)$, when $f(x|\theta)$ is assumed given and 'true', do take some cognizance of how the sample arose, e.g. through the Fisher Information involving averaging over the sample space. In fact all the objective priors listed next contradict the pure subjectivist credo in this respect.

We now look at the so–called objective ideas. A recent review is given in Kass and Wasserman (1996) where details on most of the priors discussed below are given in more depth and for priors which we have not mentioned, such as the maximum entropy prior.

A.2 Uniform prior. One idea is the uniform prior, $\pi(\theta) \propto 1$. This is clearly improper except in the case when Θ is bounded and has finite Lebesgue measure; i.e. $\int_\Theta d\theta < +\infty$.

A.3 Jeffreys' prior. A commonly used objective prior is the Jeffreys' prior (1946), whereby $\pi(\theta) \propto \sqrt{|I(\theta)|}$, where $I(\theta)$ is the Fisher information matrix, given by

$$I_{ij}(\theta) = -\int_X f(x|\theta)\,\mathrm{d}x\,\frac{\partial^2}{\partial\theta_i\partial\theta_j}(\log f(x|\theta))$$

where $\theta = (\theta_1, \ldots, \theta_p)$ and $I_{ij}(\theta)$ is the ijth element of the matrix. The motivation for this is the invariance to transformation property. If we denote $J(\theta) = \sqrt{|I(\theta)|}$ and we consider the transform $\phi = \phi(\theta)$, then it is not difficult to show that $\pi(\phi) = \pi(\theta)|\partial\theta/\partial\phi| = \sqrt{|I(\phi)|} = J(\phi)$.

While the Jeffreys' prior is well known for its invariance property it is also the prior such that for small ϵ, puts equal mass on all Kullback–

Leibler balls of size ϵ. It also arises as an objective prior based on alternative criterion, such as in regular parametric families. It can be derived as the reference prior of Bernardo (1979b), see also Barron and Clarke (1994).

A.4 Kullback risk prior. This idea for constructing a prior distribution $\pi(\theta)$ is given in Barron (1998). An asymptotic expression for the Kullback–Leibler divergence between

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

and the marginal joint density

$$p(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} f(x_i|\theta)\, \pi(\mathrm{d}\theta)$$

is given by

$$D\big(f(x_1, \ldots, x_n|\theta), p(x_1, \ldots, x_n)\big) = K_n + \log\{|I(\theta)|^{1/2}/\pi(\theta)\} + o(1),$$

where K_n does not depend on θ. If it is possible to assess also that this risk is of the form $a(\theta) + C_n$ where $C_n \to 0$ then this would result in the choice of prior as

$$\pi(\theta) \propto \sqrt{|I(\theta)|}\, e^{-a(\theta)}.$$

A.5 Probability matching prior. The idea here is to find the prior $\pi(\theta)$ so that the posterior quantiles match up to some level of error, frequentist confidence intervals. So, if $z_\alpha(\pi, x_1, \ldots, x_n)$ is the α quantile of the posterior distribution of θ, with prior $\pi(\theta)$, then

$$\Pr(\theta \le z_\alpha(\pi, x_1, \ldots, x_n)|x_1, \ldots, x_n) = \alpha,$$

and the probability here refers to θ.

A probability matching prior would also ensure that

$$\Pr(\theta \le z_\alpha(\pi, x_1, \ldots, x_n)) = \alpha + O_p(n^{-1})$$

for all $0 < \alpha < 1$, where now the probability refers to $(x_1, \ldots, x_n)$ which are taken as independently and identically distributed from $f(x|\theta)$. See Datta and Sweeting (2005), for example.

A.6 Reference prior. Another idea is the reference prior of Bernardo (1979b). The idea is quite straightforward and involves the value of an

infinite amount of data if the prior $\pi(\theta)$ has been chosen. A minimally informative or reference prior will be the one which maximizes the information in the sample. While this might be a difficult task, with many possible options to define the value of an experiment, the chosen strategy is to maximize, asymptotically, the expected Kullback–Leibler divergence between the prior and posterior. That is, if

$$I_n(\pi) = \int_{X^n} p(x^n) \int_{\Theta} \pi(\theta|x^n) \log\{\pi(\theta|x^n)/\pi(\theta)\} \, \mathrm{d}\theta \, \mathrm{d}x^n,$$

where

$$p(x^n) = p(x_1, \ldots, x_n) = \int_{\Theta} \prod_{i=1}^{n} f(x_i|\theta) \, \pi(\theta) \, \mathrm{d}\theta.$$

then the reference prior π maximizes

$$\lim_{n\to\infty} I_n(\pi).$$

Due to the limit typically being infinite, some suitable adjustments need to be made. See Berger et al. (2009) for recent developments.

Other ideas include the maximum likelihood prior of Hartigan (1998). Here the problem is to find, if it exists, the prior for which the Bayes estimate is asymptotically negligibly different from the maximum likelihood estimator. Sweeting et al. (2006) provide asymptotic properties of priors derived via a posterior predictive entropy regret criterion. This is related to the prior predictive regret criterion described in Clarke and Barron (1990). We note the widespread use in these papers of the self–information loss function, known as the negative logarithmic score function. There is a recent contribution by Dicicio and Young (2010) where an objective Bayes methodology is considered for conditional frequentist inference about a canonical parameter in a multi-parameter exponential family. These authors derive a condition under which posterior Bayes quantiles match the conditional frequentist coverage to a higher-order approximation in terms of the sample size. Other default choices of prior appear in Fraser et al. (2010).

A recent article expanding on the objective point of view is given in Berger (2006).

Acknowledgments. The authors are grateful for the detailed comments of two referees and an Associate Editor which have led to a substantial rearrangement and development of the paper. The authors would also like to thank a co–Editor for his support of the paper.

References.

Barron, A.R. and Clarke, B. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference* **41**, 37–60.

Barron, A.R. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In (J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors), *Bayesian Statistics* Volume 6, Pages 27-52. Oxford University Press.

Berger, J.O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics.

Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385 – 402.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics* **37**, 905–938

Berk, R.H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37**, 51–58.

Bernardo, J.M. (1979a). Expected information as expected utility. *Annals of Statistics* **7**, 686–690.

Bernardo, J.M. (1979b). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B* **41**, 113–147 (with discussion).

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. Wiley.

Blyth, S. (1994). Local divergence and association. *Biometrika* **81**, 579–584.

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

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics* **37**, 905–938.

Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Annals of Statistics* **26**, 617–644.

Clarke, B. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **36**, 453–471.

Datta, G.S. and Ghosh, M. (1996). On the invariance of noninformative priors. *Annals of Statistics* **24**, 141–159.

Datta, G.S. and Sweeting, T.J. (2005). Probability matching priors. In *Handbook of Statistics* **25**, *Bayesian thinking: Modeling and computation* (D.K.Dey and C.R.Rao, eds.). Elsevier, 91–114.

Diciccio, T.J. and Young, G.A. (2010). Objective Bayes and conditional inference in exponential families. *Biometrika* **97**, 497–504.

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

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1–68.

de Finetti, B. (1964) Foresight: its logical laws, its subjective sources, *Studies in Subjective Probability*, H. E. Kyburg and H. E. Smokler, eds., Wiley, New York, 93-158. (Translation of *La prévision: ses lois logiques, ses sources subjectives*, *Ann. Inst. H. Poincaré*, **7** (1937), 1–68.

Fraser, D.A.S., Reid, N., Marras, G. and Yi, G.Y. (2010). Default priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society, Series B* **72**, 631–654.

French, S. (1982). On the axiomatization of subjective probabilities. *Theory and Decision* **14**, 19-33.

Goldstein, M. (1981). Revising previsions: A geometric interpretation. *Journal of the Royal Statistical Society, Series B* **43**, 105–130.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis* **1**, 403 – 420.

Gutiérrez–Peña, E. and Walker, S.G. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review* **73**, 309–330.

Hartigan, J.A. (1998). The maximum likelihood prior. *Annals of Statistics* **26**, 2083–2103.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* **80**, 470-501.

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

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences* **186**, 453-461.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343–1370.

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

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Lindley, D. V. (1972) Bayesian Statistics, a Review. Philidelphia, PA: *SIAM*.

Merhav, N. (1998). Universal prediction. *IEEE Trans. Inf. Theory* **44**, 2124–2147.

Ramsey, F. P. (1926) Truth and Probability. *The Foundations of Mathematics and other Logical Essays* (R. B. Braithwaite, ed) London, Kegan Paul (1931), 156-198. Reprinted in 1964 in *Studies in Subjective Probability* (H.E Kyburg and H. E. Smokler (Eds.) J Wiley, New York.

Sweeting, T.J. (1998). Discussion of "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems", by A.R.Barron. In J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, Bayesian Statistics, volume 6, pages 2752. Oxford University Press.

Sweeting, T. J., Datta, G. S. and Ghosh, M. (2006). Non–subjective priors via predictive relative entropy loss. *Annals of Statistics* **34**, 441–468.

Walker, S.G. and Mallick, B.K. (1997). A note on the scale parameter of the Dirichlet process. *Canadian Journal of Statistics* **25**, 473–479.