ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)

A general framework for updating belief distributions

P. G. Bissiri,

University of Milano-Bicocca, Italy

C. C. Holmes

University of Oxford, UK

and S. G. Walker

University of Texas at Austin, USA

[Received December 2013. Final revision November 2015]

Summary. We propose a framework for general Bayesian inference. We argue that a valid update of a prior belief distribution to a posterior can be made for parameters which are connected to observations through a loss function rather than the traditional likelihood function, which is recovered as a special case. Modern application areas make it increasingly challenging for Bayesians to attempt to model the true data-generating mechanism. For instance, when the object of interest is low dimensional, such as a mean or median, it is cumbersome to have to achieve this via a complete model for the whole data distribution. More importantly, there are settings where the parameter of interest does not directly index a family of density functions and thus the Bayesian approach to learning about such parameters is currently regarded as problematic. Our framework uses loss functions to connect information in the data to functionals of interest. The updating of beliefs then follows from a decision theoretic approach involving cumulative loss functions. Importantly, the procedure coincides with Bayesian updating when a true likelihood is known yet provides coherent subjective inference in much more general settings. Connections to other inference frameworks are highlighted.

Keywords: Decision theory; General Bayesian updating; Generalized estimating equations; Gibbs posteriors; Information; Loss function; Maximum entropy; Provably approximately correct Bayes methods; Self-information loss function

1. Introduction

Data sets are increasing in size and modelling environments are becoming more complex. This presents opportunities for Bayesian statistics but also major challenges, perhaps the greatest of which is the requirement to define the true sampling distribution, or likelihood, for the data generator $f_0(x)$, regardless of the study objective. Even if the task is inference for a low dimensional parameter, Bayesian analysis is required to model the complete data distribution and, moreover, to assume that the model is 'true'.

In this paper we present a coherent procedure for general Bayesian inference which is based on the updating of a prior belief distribution to a posterior when the parameter of interest is connected to observations via a loss function. Briefly here, and in the simplest scenario, suppose that interest is in the θ minimizing the expected loss

Address for correspondence: C. C. Holmes, Department of Statistics, University of Oxford, 24–29 St Giles, Oxford, OX1 3LB, UK.
E-mail: c.holmes@stats.ox.ac.uk

1369–7412/16/79000

$$L(\theta) = \int l(\theta, x)\, \mathrm{d}F_0(x), \tag{1}$$

for some loss function $l(\theta, x)$, e.g. $l(\theta, x) = |\theta - x|$ for estimating a median, where $F_0(x)$ is the unknown distribution function from which independent and identically distributed observations arise. If $\pi(\theta)$ represents prior beliefs about this θ, and x is observed from F_0, then we argue that a valid and coherent update of $\pi(\cdot)$ is to the posterior $\pi(\cdot|x)$, where

$$\pi(\theta|x) \propto \exp\{-l(\theta, x)\}\, \pi(\theta). \tag{2}$$

The argument for this is given later in the paper and to some extent relies on the idea that an update of beliefs *must* exist. For we have a well-defined parameter of interest θ, an initial belief distribution about the location of the parameter, $\pi(\theta)$, and gain further independent information about θ via x coming from $F_0(x)$. To update, it is clear for some function ψ that we must have

$$\pi(\theta|x) = \psi\,\{l(\theta, x), \pi(\theta)\}.$$

That the form for ψ is expression (2) is detailed later and a coherence property plays a key role:

$$\psi[l(\theta, x_2), \psi\{l(\theta, x_1), \pi(\theta)\}] \equiv \psi\{l(\theta, x_1) + l(\theta, x_2), \pi(\theta)\}. \tag{3}$$

This ensures that we end up with $\pi(\theta|x_1, x_2)$ as the same object whether we update with (x_1, x_2) together or $\{(x_1), (x_2)\}$ one after the other.

A special case is when it is known that $F_0(x) = F(x; \theta_0)$ for some parametric family of distributions $F(\cdot; \theta)$, with corresponding density function $f(\cdot; \theta)$, and $l(\theta, x) = -\log\{f(x; \theta)\}$. For minimizing $L(\theta)$ here yields θ_0 and the update (2) is the usual Bayesian update. It is important to note that the general Bayesian update using loss functions should not to be seen as an approximation to anything; rather, it is targeting the parameter of interest, employing the necessary loss function with a valid coherent update of beliefs.

Classical inference based on the likelihood function can be regarded as using the 'negative log-likelihood function' as a loss function; for example, in the case of independent and identically distributed observations, we can regard

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

as a loss function connecting data (x_i) with a parameter θ indexing the family of density functions $f(x|\theta)$. And, in this setting, we do not even need to assume the correctness of the model; we are merely expressing interest in the parameter θ_0 minimizing

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

which is the parameter minimizing the Kullback–Leibler divergence between the unknown $f_0(\cdot)$ and the family $f(\cdot; \theta)$.

1.1. The idea

Here we provide further elaboration on the outline of the idea given previously. Let θ denote a parameter or functional of interest, e.g. the mean or median of a population $F_0(x)$, and let x denote an observation from $F_0(x)$, with F_0 unknown. We are interested in a formal way to update prior beliefs $\pi(\theta)$ to posterior beliefs $\pi(\theta|x)$ given x.

Bayesian inference proceeds through knowledge of a complete and true model for $f_0(x)$. This is often parameterized via a sampling distribution $f(x;\theta)$ and a prior $\pi(\theta)$, and defines the marginal likelihood

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

Then (see for example Bernardo and Smith (1994)), inference for θ can occur via Bayes theorem

$$\pi(\theta|x) = f(x;\theta)\,\pi(\theta)/m(x).$$

However, the statement 'inference for θ' is meaningless unless the true parametric family $f(\cdot;\theta)$ is known. In this case, following the Savage axioms (Savage, 1954), the Bayesian update can be shown to be the rational way to proceed. However, $f_0(x)$ may be unknown and, even if $f(\cdot;\theta)$ is correct, θ might be ultrahigh dimensional mainly made up of nuisance parameters relative to a low dimensional subset of the parameters of interest. Taken altogether, these points can make the Bayesian approach cumbersome.

We are interested in the rational updating of beliefs under more general and less stringent conditions. To do so we make use of loss functions to connect information in data to parameters of interest. Informally for now, we write such loss functions as $l(\theta, x)$, and we shall discuss specific types later in the paper. We shall consider the reporting of subjective beliefs $\pi(\theta|x)$ as an action made under uncertainty and use decision theory to guide the optimal action. See, for example, Hirshleifer and Riley (1992).

To outline the theory, let ν denote a probability measure on the space of θ. We shall construct a loss function to select an optimal posterior distribution $\hat{\nu}(\theta)$ given a prior $\pi(\theta)$ and data x. (We use $\hat{\nu}$ to denote optimality rather than an approximation or estimate.) To achieve this we construct a loss function $L(\nu;\pi,x)$ on the space of probability measures on θ-space, and then present

$$\hat{\nu} = \arg\min_{\nu} L(\nu;\pi,x)$$

as the representation of beliefs about the unknown value of θ given the prior information, represented via the belief distribution π, and data x. As it is widely assumed that data x are an independent piece of information to that which gave rise to the prior, it is appropriate to consider an additive, or cumulative, loss function of the form

$$L(\nu;\pi,x) = h_1(\nu,x) + h_2(\nu,\pi), \tag{4}$$

where h_1 and h_2 are themselves loss functions on probability measures, representing fidelity to data and fidelity to prior respectively. See, for example, Berger (1993) for more about ideas on uses of loss functions within decision theory.

The question is whether we can claim a probability measure selected as the solution to a decision problem, i.e. minimizing a loss function, can be viewed as representing beliefs about a parameter. To answer this, given the aim (1), we would clearly prefer probability measure ν_1 to ν_2 as representing beliefs if

$$\int\int l(\theta,x)\,\mathrm{d}F_0(x)\,\nu_1(\mathrm{d}\theta) \leqslant \int\int l(\theta,x)\,\mathrm{d}F_0(x)\,\nu_2(\mathrm{d}\theta). \tag{5}$$

Indeed, it would be incoherent to select ν_2 rather than ν_1 when condition (5) holds. Thus the answer is affirmative. Though we are not minimizing or comparing condition (5), since we do not have F_0, we can substitute the expression

$$L_0(\nu; F_0) = \int \int l(\theta, x)\, \mathrm{d}F_0(x)\, \nu(\mathrm{d}\theta) \tag{6}$$

with the Bayesian finite sample expression of the form (4). We now discuss the choices of h_1 and h_2 which give equation (4) as a Bayesian finite sample version of equation (6).

Under this approach the analyst needs to specify h_1 and h_2 in such a way that they proceed in an optimal, rational and coherent manner. Somewhat remarkably, as proved in the on-line supplementary material, for coherent inference (3), h_2 must be the Kullback–Leibler divergence (Kullback and Leibler, 1951), and given by

$$h_2(\nu, \pi) = d_{\mathrm{KL}}(\nu, \pi) = \int \nu(\mathrm{d}\theta)\, \log\{\nu(\mathrm{d}\theta)/\pi(\mathrm{d}\theta)\},$$

where, with a slight abuse of notation, we also use $\pi(\mathrm{d}\theta)$ to denote the probability measure version of π, i.e. $\pi(\mathrm{d}\theta) = \pi(\theta)\, \mathrm{d}\theta$.

Regarding h_1, since $\nu(\theta)$ is a probability measure representing beliefs about θ, the only choice here is to take the loss to data $h_1(\nu, x)$ as the *expected* loss (see von Neumann and Morgenstern (1944)) of $l(\theta, x)$, i.e.

$$h_1(\nu, x) = \int l(\theta, x)\, \nu(\mathrm{d}\theta),$$

with the particular types of the loss function on the parameter of interest $l(\theta, x)$ to be discussed later.

Substituting in h_1 and h_2, the cumulative loss function is then given by

$$L(\nu; \pi, x) = \int l(\theta, x)\, \nu(\mathrm{d}\theta) + d_{\mathrm{KL}}(\nu, \pi). \tag{7}$$

This then, i.e. equation (7), is our finite sample version of equation (6), and note that equation (7) becomes, under mild regularity conditions, equation (6) as $n \to \infty$. The solution to equation (7) provides the $\hat{\nu}$ which the statistician believes best minimizes equation (6). This is, according to our approach, done by using the empirical distribution function as a substitute for F_0 and using a penalty term which prevents the answer from being too far from the prior in a Kullback–Leibler sense; the Kullback–Leibler appearing here for the necessary coherence property of the answer. Of interest, as discussed later on, is the provably approximately correct (PAC) Bayes solution to the problem (Langford, 2005) that finds an approximation which minimizes an upper bound for equation (6); see Section 3.

Surprisingly, but quite easy to show, the minimizer of $L(\nu; \pi, x)$ is given by

$$\begin{aligned} \hat{\nu}(\theta) &= \arg\min_{\nu} L(\nu; \pi, x) \\ &= \frac{\exp\{-l(\theta, x)\}\, \pi(\theta)}{\int \exp\{-l(\theta, x)\}\, \pi(\mathrm{d}\theta)}. \end{aligned} \tag{8}$$

This can be seen by observing that

$$\int l(\theta, x)\, \nu(\mathrm{d}\theta) + d_{\mathrm{KL}}(\nu, \pi) = \int \nu(\mathrm{d}\theta)\, \log\left[\frac{\nu(\theta)}{\exp\{-l(\theta, x)\}\, \pi(\theta)}\right].$$

So equation (8) has the form of a Bayesian update using exponentiated negative loss in place of the likelihood function. We stress again that equation (8) is not an approximation, or pseudo-posterior, but rather a valid coherent representation of subjective uncertainty in the minimizer of expression (1). As is usual in decision problems involving the use of loss functions, it is

incumbent on the decision maker to ensure that solutions exist. So $l(\theta, x)$ needs to be constructed such that

$$0 < \int \exp\{-l(\theta, x)\}\, \pi(\mathrm{d}\theta) < \infty.$$

Whereas the Bayesian approach requires the construction of a probability model for all possible outcomes conditionally on all unknown states of nature, the general Bayesian approach requires the construction of loss functions given the outcomes for only the parameter of interest. This allows the decision maker to concentrate on modelling only those quantities that are important to the task at hand.

1.2. On equation (3) implying equation (2)

We shall now go into the details of how equation (3) and some other natural assumptions imply equation (2). We are asking for the unique ψ which provides the update for all Θ, i.e. is Θ invariant? This is a reasonable requirement since how we update should not depend on Θ. In fact we show that equation (3) uniquely implies equation (2) for $|\Theta| = 3$, i.e. the cardinality of Θ is 3, and hence the update is the only unique update that applies for all Θ. So consider the following assumptions.

Assumption 1. Condition (3) holds true.

Assumption 2. For any set $A \subset \Theta$,

$$\frac{\psi\{l(\theta, x), \pi(\theta)\}}{\displaystyle\int_A \psi\{l(\theta, x), \pi(\theta)\}\, \mathrm{d}\theta} = \psi\{l(\theta, x), \pi_A(\theta)\}, \tag{9}$$

where π_A is π restricted and normalized to A, i.e. $\pi_A(\theta) = \pi(\theta)\, \mathbf{1}(\theta \in A) / \int_A \pi(\theta)\, \mathrm{d}\theta$. This condition says that whether we update the prior restricted to the set A, or update the prior and then restrict to the set A, we obtain the same update.

Assumption 3. Lower evidence (larger loss) for a state should yield smaller posterior probabilities under the same prior. So, if for some $A \subset \Theta$, $l(\theta, x) > l(\theta, y)$ for $\theta \in A \subset \Theta$ and $l(\theta, x) = l(\theta, y)$ for $\theta \in A^{\mathrm{c}}$, then

$$\int_A \psi\{l(\theta, x), \pi(\theta)\}\, \mathrm{d}\theta < \int_A \psi\{l(\theta, y), \pi(\theta)\}\, \mathrm{d}\theta.$$

Assumption 4. If $l(\theta, x) \equiv$ constant, then $\psi\{l(\theta, x), \pi(\theta)\} = \pi(\theta)$, i.e., if the observation provides no information about θ, since the loss function is a constant, then the posterior is the same as the prior.

Assumption 5. If $\tilde{l}(\theta, x) = l(\theta, x) + c$ for some constant c, then

$$\psi\{\tilde{l}(\theta, x), \pi(\theta)\} = \psi\{l(\theta, x), \pi(\theta)\}.$$

Theorem 1. If assumptions 1–5 hold, then for $|\Theta| = 3$ equation (3) uniquely implies equation (2).

The proof is given in Appendix A. It is quite straightfoward to extend the uniqueness argument to all countably infinite Θ, which would replace the uniqueness argument for all Θ. However, we would need more work to extend separate uniqueness to general Θ.

Clearly Sections 1.1 and 1.2 are different derivations of the same result, i.e. the support of update (2).

1.3. Connections with related work

There is a large literature on procedures for robustly estimating a parameter of interest by minimizing the cumulative loss

$$L(\theta; x) = \sum_{i=1}^{n} l(\theta, x_i). \tag{10}$$

This is clearly the finite sample version of

$$L(\theta) = \int l(\theta, x)\, \mathrm{d}F_0(x).$$

Our claim is that equation (7) is the general Bayesian version of equation (10), where interest is on probability measures on θ-space rather than single states θ.

Hüber (2009) provided examples of equation (10), where we note that the primary aim is not modelling the data but rather estimating a parameter. This is an advantage when a probability model for the data is too difficult to formulate. We are presenting a general Bayesian extension of this idea. Since we are interested in a belief distribution for θ given data, and we have further information provided by π, we claim that the appropriate Bayesian version is given by equation (8).

Some of the ideas that are presented in the paper have been considered by Zhang (2006a, b) and Jiang and Tanner (2008). In Zhang (2006a) an estimation procedure, named information risk minimization, also known as a Gibbs posterior, which has the same form as equation (8), is described in section IV of his paper. Zhang then concentrated on the properties of the Gibbs posterior. Further theoretical work was done in Zhang (2006b).

In Jiang and Tanner (2008) a Gibbs posterior was studied in comparison with a true Bayesian posterior where the model is assumed to be misspecified. The claim is that posterior performance of a Bayesian model can be unreliable when misspecified, whereas a Gibbs posterior which targets points of interest can have better performance. The comparison involves variable selection for high dimensional classification problems involving a logit model.

Here we show that solutions of the form (8) are the only coherent, decision theoretic representation of posterior beliefs under model misspecification. We also provide a principled approach to scale the relative information in the data to information in the prior (see Section 3); that was left as an arbitrary free parameter in Zhang (2006a, b) and Jiang and Tanner (2008).

Bissiri and Walker (2010) used equation (7) with Bernoulli observations and found sufficient conditions on $l(\theta, x)$ for the sequence of posteriors, based on equation (8), to be consistent. This result for consistency was extended to more general independent and identically distributed observations in Bissiri and Walker (2012a). In Bissiri and Walker (2012b), it was shown starting from the class of g-divergences (Ali and Silvey, 1966), for a coherent sequence of updates, that we need the Kullback–Leibler divergence as the loss between prior π and ν. In the on-line supplementary material, we present an updated proof that is simplified and more intuitive to that appearing before now.

A similar construct to $L(\nu; \pi, x)$ was provided by Zellner (1988), who presented what is essentially a loss function for the posterior distribution by using ideas of information processing from prior to posterior. The motivation is different and relies on notions of information present in log-probabilities and log-likelihoods, which may not be compatible as noted by J. M. Bernardo

in the discussion of Zellner (1988). Furthermore, our derivation of the loss function allows a broader interpretation of the elements, which does not require the existence of a probability distribution for the observation; see Section 4.

Concerns that the specification of a complete model for the data-generating distribution is unachievable date back to de Finetti (1937) and the notion of 'prevision'. In his work de Finetti considered conditional expectation as the fundamental primitive, or statistic, of interest on which prior beliefs are expressed and updated. Recently other researchers have further developed this approach under the field of Bayesian linear statistics; see Goldstein and Wooff (2007).

There has been increasing awareness of the restrictive assumptions that formal Bayesian analysis entails. Royall and Tsou (2003) described procedures for adjusting likelihood functions when the model is misspecified. More recently, Doucet and Shephard (2012) and Müller (2012) considered formal approaches to pseudo-Bayesian methods using sandwich estimators to update subjective beliefs, motivated by robustness to model misspecification; see also Hoff and Wakefield (2013). Cooley *et al.* (2009) considered pseudo-Bayesian approaches with composite likelihoods. More generally there is increasing recognition that formal Bayesian analysis can be restrictive for example through computational issues, such as arise in the area of approximate Bayesian computation (see, for example, Marin *et al.* (2012)).

Several researchers have considered issues with Bayesian updating by using proxy models $f(x;\theta)$ (for example, see Key *et al.* (1999)), when (x_i) are known not to arise from $f(x;\theta)$ for any value of θ, i.e. there is no θ conditional on which x is from $f(x;\theta)$. This is referred to as the M-open case in Bernardo and Smith (1994). One suggested solution is to use methods based on approximations and Key *et al.* (1999) described one such idea using a cross-validation approach. Although this may be pragmatic it does have some shortcomings. Most serious is that there is little back-up theory and this has repercussions in that the update suffers from a lack of coherence.

Another approach is to ignore the problem, i.e. to assume that the observations are coming from $f(x;\theta)$ even though it is known that they are not. According to Goldstein (1981), 'there is no obvious meaning for Bayesian analysis in this case'. The disaster of making horribly wrong inference can be protected to some extent by model selection, i.e. postulating a number of models for $f_0(x)$, say $f_j(x;\theta_j)$, with corresponding priors $\pi_j(\theta_j)$, and model probabilities p_j, for $j=1,\ldots,M$. But, as Key *et al.* (1999) pointed out, how do we construct $\pi_j(\theta_j)$ and p_j when we know that none of the postulated models are correct? So the Bayesian update breaks down in that nothing has any interpretation.

Finally, and we acknowledge the contribution of the reviewers for pointing this out, we discuss connections with PAC Bayes methods; see, Shawe-Taylor and Williamson (1997), Langford (2005), Alquier (2008) and McAllester (1998). PAC Bayes is an interesting emerging field in machine learning concerned with techniques for bounding the generalization error (empirical risk) of a Bayesian model. The motivation behind PAC Bayes methods is to find an upper bound for the empirical risk of a probability measure ν on a model $L(\nu;F_0)$ in equation (6), which is termed generalization error in the PAC Bayes literature. Given observation x and prior π, the upper bound will be written as $U(\nu;x,\pi)$, i.e. for all ν

$$L(\nu;F_0)\leqslant U(\nu;x,\pi).$$

See Catoni (2003) where the form of U is provided. Then it can be shown that an upper bound $U(\nu;x,\pi)$ is provided by equation (8). The PAC Bayes approach is complementary to our work. The motivation and construction are very different. We are interested in a framework for the rational updating of beliefs, rather than seeking bounds on the empirical risk of a probability measure on models. The minimizer of an upper bound is interesting but does not justify using $\hat{\nu}$ as an update of a belief distribution for Bayesian style inference, and hence whether $\hat{\nu}$ forms

a coherent sequence of belief distributions is not discussed in the PAC Bayes formulation of U; the requirement of coherence is central to Bayesian style learning. Moreover the scaling of the loss to data h_1 to the loss to prior h_2 enters as a constant in the margin of the error bound in PAC Bayes methods, whereas here it has explicit meaning in the relative weight of information provided by the two sources, prior and data (see Section 3).

The general Bayesian approach also coincides with the prediction-motivated approach of Cesa-Bianchi and Lugosi (2006) and is known as aggregation with exponential weight, which does not rely on stochastic information; see also our Section 4.1.

This said, there are clear synergies and the operational characteristics of PAC Bayes methods are similar; they must be since we gather the same answer. However, the motivation and consequences are different. Moreover, as we shall see later, the derivation here provides insights into the necessary calibration of loss functions h_1 and h_2.

1.4. Layout of the paper

The layout of the remainder of the paper is as follows. In Section 2 we discuss types of loss function. When the self-information loss function is used then the update is the traditional Bayes update. With other loss functions there is a calibration issue between the two styles of loss function used, i.e. the loss to the data and the loss to the prior. This calibration problem is discussed and potential solutions provided in various ways in Section 3. In Section 4 we discuss forms of information other than the usual data arising from some unknown distribution function. This includes non-stochastic information and also partial information. Section 5 provides some numerical illustrations including inference based on partial information and a clustering problem. Section 6 concludes with a discussion on various points.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Types of loss function

In this section we shall consider the form of h_1 in equation (4) that connects information in the data to the value of the unknown θ. We shall consider three broad situations: first, when the analyst believes that they know the complete family of distributions from which the (x_i) arose, the so called M-closed scenario; second, when $f_0(x)$ is unknown but where a complete likelihood $f(x;\theta)$ is being used as a proxy model; finally, when there is no sampling distribution or proxy model for x and the parameter of interest is connected to x via a loss function $l(\theta, x)$.

2.1. M-closed case and self-information loss

When the analyst knows the family from which (x_i) arose, the so-called M-closed view, then the Bayesian approach to learning is fully justified, well known and widely used as a statistical approach to inference; Bernardo and Smith (1994) is comprehensive. To see how Bayes arises in our framework, we would need to construct a loss function for $l(\theta, x)$ with the knowledge that x came from $f(x;\theta)$. It is well known that the appropriate and sole loss function in this case is the self-information, or logarithmic loss function, given by

$$l(\theta, x) = -\log\{f(x;\theta)\}.$$

Indeed, the cumulative loss version of this is the log-likelihood function. See Bernardo (1979) and Merhav and Feder (1998) for more on the self-information loss function. This amounts to

the use of proper scoring rules when the parametric family $f(x;\theta)$ is known, and under which our approach coincides with the Bayesian updating rule.

2.2. M-open case and the use of proxy models

Issues with the Bayesian rule arise when the form of $f(x;\theta)$ is not known; for example, see Key *et al.* (1999). Equivalently, there is no θ conditional on which x is from $f(x;\theta)$; more bluntly, there is no connection between any x and any θ via $f(x;\theta)$. This is referred to as the M-open case in Bernardo and Smith (1994). In many situations, the correct sampling density $f_0(x)$ is unknown or unavailable or too complex to work with.

Under a general Bayesian approach we may proceed by considering θ_0, the value of θ that minimizes the Kullback–Leibler divergence between a proxy model $f(x;\theta)$ and the true density function $f_0(x)$, i.e. θ_0 minimizes

$$d_{\mathrm{KL}}\{f_0(\cdot), f(\cdot;\theta)\} = \int f_0(x)\log\{f_0(x)/f(x;\theta)\}\,\mathrm{d}x.$$

Then prior beliefs $\pi(\theta)$ will be expressed on this unknown value. It is possible to learn about this θ_0 since an infinite collection of (x_i) yields θ_0. Then we would wish the sequence of $\nu(\theta)$ to accumulate about θ_0. The appropriate loss function in this case is still $l(\theta, x) = -\log\{f(x;\theta)\}$. The standardized cumulative loss based on a sequence of observations $(x_i)_{i=1}^n$ is given by $-n^{-1}\sum_{i=1}^n \log\{f(x_i;\theta)\} \to -\int \log\{f(x;\theta)\}\,\mathrm{d}F_0(x)$ almost surely for all θ, which is minimized by θ_0.

So although the Bayesian approach has foundational issues to deal with whether the M-open or M-closed view holds, for the approach here it is irrelevant. If we adopt θ_0 as the parameter value taking the family closest to $f_0(\cdot)$ then we do not need to worry if we are in the M-open or M-closed scenario, since if $f(\cdot;\theta)$ is the true family then obviously θ_0 reverts to the true parameter value. This point is crucial, since for the Bayesian being in the M-open or M-closed state forces us to adopt different inference approaches; see Bernardo and Smith (1994). Moreover our approach supports the use of the relevant partial information in the data for updating beliefs on the parameter of interest, an example of which is shown in Section 5.1. This can be especially important when the data are high dimensional. Such updates have no formal justification from a Bayesian perspective.

2.3. Parameter minimizing a loss function

In the most general scenario the parameter of interest minimizes a loss function of the type (1). In the classical literature, this type of estimation problem is in the area of *robust statistics* and specific loss functions can be found in the literature, pertaining to M-estimation and estimating equations. See, for example, Hüber (2009).

An important class of loss functions is provided by the M-estimators for a location parameter; Hüber (1964). So, rather than using the loss function $-\log\{f(x_i;\theta)\}$, a $\rho(x_i;\theta)$ is used in an attempt to obtain robust estimation, rather than the traditional maximum likelihood estimator, which can be suspect if the model is incorrect. This idea has been generalized to the class of estimating equations, whereby the estimate of θ is obtained by minimizing

$$\sum_{i=1}^{n} \rho(x_i;\theta).$$

Our approach, which mirrors this classical robust procedure, would use the loss function

$$L(\nu; x_1, \ldots, x_n, \pi) = \int \sum_{i=1}^{n} \rho(x_i;\theta)\,\nu(\mathrm{d}\theta) + d_{\mathrm{KL}}(\nu, \pi)$$

with solution provided by

$$\hat{\nu}(\mathrm{d}\theta) \propto \exp\left\{-\sum_{i=1}^{n} \rho(x_i;\theta)\right\} \pi(\mathrm{d}\theta).$$

The θ_0 of interest is implicitly assumed to be the limit of the sequence of minimizers of the cumulative losses. This would be the minimizer of $\int \rho(x;\theta)\,\mathrm{d}F_0(x)$ and hence the prior beliefs are being expressed about this unknown value. Then the loss function $l(\theta, x) = \rho(x;\theta)$ is ensuring that the updates are indeed 'moving towards' θ_0. To complete the picture, it would have been that the decision maker would be happy to make a decision given the minimizer of $\int \rho(x;\theta)\,\mathrm{d}F_0(x)$.

3. Calibration of relative losses

This section deals with the important aspect of specifying the relative information in the data to the information in the prior in general settings. In the M-closed and M-open, including partial information, settings the use of the self-information loss $l(\theta, x) = -\log\{f(x;\theta)\}$ results in a fully specified form for equation (8). However, in the setting of Section 2.3 there is an issue about the scale of the loss function h_1 which is a consequence of the apparent arbitrariness in the weight of $h_1(\nu, x)$ relative to $h_2(\nu, \pi)$, in that we are free to multiply either by an arbitrary factor. So, equivalently, we are interested in a loss function $w\,l(\theta, x)$ for some $w > 0$. The question is how to select w, noting that w controls the relative weight of loss to data to loss to prior.

Of course, such an issue does not arise in general in the classical literature on *parameter* estimation since there is typically no combining with different styles of loss function. A notable exception is the class of regularized regressions, such as the lasso, where one minimizes

$$L(\beta) = w \sum_{i=1}^{n} l(\beta, y_i, x_i) + |\beta|.$$

Note the substantial difference in that this loss is for a parameter, whereas the losses that we consider are for a measure.

The calibration of different types of loss is not a unique problem to us or to the lasso. It arises in many applied contexts; possibly the most well known are in health economics where losses pertaining to costs need to be balanced against losses pertaining to health benefits.

The most common ideas for assigning w in the Gibbs posteriors and PAC Bayes literature typically involve cross-validation and subjective choices. As mentioned above, in PAC Bayes methods the weighting w is a constant that enters the margin of the error bound. Here we discuss some ideas in the context of a general Bayesian update intended to help the analyst. We do not claim to be exhaustive in the approaches, or to be prescriptive in advocating one approach over another. Our intention is to provide tools and suggestions for elicitation of the relative loss to data to loss to prior.

3.1. Annealing

In the literature on Gibbs posteriors, the weighting parameter is labelled as a 'temperature' and selected subjectively. There are clear connections here with the use of 'power priors' (Ibrahim and Chen, 2000) where

$$\nu(\mathrm{d}\theta) \propto \prod_{i=1}^{n} f(x_i;\theta)^w \pi(\mathrm{d}\theta).$$

Such an idea has also been discussed in Walker and Hjort (2001). It is evident what w achieves; if $0 < w < 1$ then the loss to prior is given more prominence than in the Bayesian update and

the data will be less influential. In the extreme case when $w=0$ we retain the prior throughout. In contrast, when $w>1$ the loss $-\log\{f(x;\theta)\}$ is given more prominence than in the Bayesian update and in the extreme case when w is very large the ν is accumulating about the maximum likelihood estimator for the model, i.e.

$$\nu(\mathrm{d}\theta) \approx \delta_{\hat{\theta}}(\mathrm{d}\theta),$$

where $\hat{\theta}$ maximizes $\Pi_{i=1}^{n} f(x_i;\theta)$.

3.2. Unit information loss

Here we discuss a procedure for default subjective assignment based on a prior evaluation of the expected value of $l(\theta,x)$. The idea originates from work in the specification of reference priors and 'objective Bayes' methods; see for example Kass and Wasserman (1996).

To begin it helps to ensure that both losses are non-negative for all θ. Hence we write the prior loss function with an additional term $\log\{\pi(\hat{\theta})\}$, which is a constant, and where $\hat{\theta}$ maximizes $\pi(\theta)$, so that the cumulative loss becomes

$$L(\nu;x,\pi) = \int [w\,l(\theta,x) + \log\{\pi(\hat{\theta})/\pi(\theta)\}]\nu(\mathrm{d}\theta) + \int \nu(\mathrm{d}\theta)\log\{\nu(\theta)\}.$$

and we would additionally standardize $l(\theta,x)$ such that $\min_\theta l(\theta,x)=0$ for any x. Hence, we can regard

$$L(\theta;x,\pi) = w\,l(\theta,x) + \log\{\pi(\hat{\theta})/\pi(\theta)\}$$

as a loss function for θ with information provided by x and π. So, assuming that $l(\theta,x)>0$, we want to calibrate the two loss functions given by

$$w\,l(\theta,x)$$

and

$$\log\{\pi(\hat{\theta})/\pi(\theta)\}.$$

These are two loss functions for θ and to adhere with the notion that, before we have any data, there is a single piece of information, we can calibrate the two losses by making the joint expected losses, taken over θ and x, to match, i.e. whether someone takes a θ and is penalized by the loss

$$\log\{\pi(\hat{\theta})/\pi(\theta)\},$$

or takes a (θ,x) and is penalized by the loss $w\,l(\theta,x)$, at the outset, the expected losses should match. They are confronted by two choices of loss with one piece of information and thus the losses can be calibrated by ensuring that their expected losses coincide. The connection between expected information and expected loss can be found in Bernardo (1979).

Thus w can be set by ensuring that

$$w\,E_{\theta,x}\{l(\theta,x)\} = E_\theta[\log\{\pi(\hat{\theta})/\pi(\theta)\}].$$

Here E is with respect to a joint belief distribution in x and θ; say $m(x,\theta)$, the marginal for θ of which is $\pi(\theta)$. So

$$w = \frac{\int \log\{\pi(\hat{\theta})/\pi(\theta)\}\,\pi(\mathrm{d}\theta)}{\int\int l(\theta,x)\,m(\mathrm{d}\theta,\mathrm{d}x)}. \tag{11}$$

Let us consider an example, where $l(\theta, x) = (\theta - x)^2$ with $\pi(\theta) = N(\theta|0, \tau^2)$ with $m(x|\theta)$ being any density with mean θ and variance σ^2. Then we can evaluate

$$\int \log\{\pi(\hat{\theta})/\pi(\theta)\}\,\pi(\mathrm{d}\theta) = \tfrac{1}{2}$$

and

$$\int\int (\theta - x)^2\, m(\mathrm{d}x, \mathrm{d}\theta) = \sigma^2,$$

so $w = \frac{1}{2}\sigma^{-2}$. Hence, this calibration idea yields the 'correct' value of $\frac{1}{2}\sigma^{-2}$ in this case. This construction requires the user specification of a joint density $m(\mathrm{d}x, \mathrm{d}\theta)$ which in some circumstances may prove difficult. Here we propose an empirical expression for this.

Now $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ should predict x_i and the best θ-value to achieve this would minimize

$$\sum_{j \neq i} l(\theta, x_j).$$

Denote the minimizer as $\hat{\theta}_{-i}$. Then we would empirically estimate the denominator of equation (11) by

$$E\{l(\theta, x)\} = n^{-1} \sum_{i=1}^{n} l(\hat{\theta}_{-i}, x_i).$$

To see this more easily, we relate it to a standard Bayesian cross-validation idea. So assume that we wish to estimate

$$E[\log\{f(x|\theta)\}] \tag{12}$$

empirically. Given x_{-i} we would predict x_i by using the plug-in density $f(x_i|\hat{\theta}_{-i})$, where $\hat{\theta}_{-i}$ maximizes $\Pi_{j \neq i} f(x_j|\theta)$. Hence, we would estimate expectation (12) via

$$n^{-1} \sum_{i=1}^{n} \log\{f(x_i|\hat{\theta}_{-i})\}$$

based on the idea that $(\hat{\theta}_{-i}, x_i)$ represent an empirical sample from $m(\theta, x)$.

If we illustrate this on a toy example, for which $l(\theta, x) = (x - \theta)^2$ and $\pi(\theta)$ is normal with zero mean and variance $1/\tau$, then it is easy to show that

$$w = \frac{1}{2} \frac{n}{\sum_{i=1}^{n} (x_i - \bar{x}_{-i})^2}$$

which asymptotically becomes $\frac{1}{2}\sigma^{-2}$, with σ^2 the variance of the data.

It is interesting to note in the above that if it is thought that the appropriate choice for $\pi(\theta)$ is flat, which is possible if the θ-space is bounded, then clearly we have $\log\{\pi(\hat{\theta})/\pi(\theta)\} = 0$. Thus, to be coherent, we would equally believe that $\int l(\theta, x)\, m(\mathrm{d}x|\theta)$ does not depend on θ, where $m(\cdot|\theta)$ is a belief distribution for x given θ. This is a condition which would be difficult to justify, as it would then be also for the uniform prior for θ. If one is used, then we only recommend that the value of w is not assigned in the above way.

3.3. Hierarchical loss

Another way to proceed is to extend the loss function to include w as an unknown parameter. Standard ideas here would suggest that we take

$$L(\theta, w; x, \pi) = w\, l(\theta, x) + \xi\, l(w) - \log\{\pi(\theta, w)\}$$

for some $\xi > 0$. We would appear to be making no progress since we now have a ξ to assign. However, this is akin to the hierarchical Bayesian model where uncertainty is propagated via hyperprior distributions to robustify the ultimate prior choice at some level. Hence, the allocation of a ξ would not be as crucial as the assignment of a w.

For example, as w is a scale parameter on loss to data, taking $l(w) = \log(w)$ the solution is given by

$$\hat{\nu}(\theta, w|x, \pi) \propto w^{\xi} \exp\{-w\, l(\theta, x)\}\, \pi(\theta, w)$$

and given that w^{ξ} can be absorbed in the prior π it is reasonable to assess ξ subjectively, i.e. it seems unreasonable to accept that π can be chosen subjectively but that ξ cannot.

3.4. Operational characteristics and subjective calibration

The idea here is to set w so that the posterior quantiles are calibrated at some level of error to frequentist confidence intervals based on the estimation of θ via minimizing the loss

$$\sum_{i=1}^{n} l(\theta, x_i).$$

So, if $C_\alpha(w, x_1, \ldots, x_n)$ is the $100(1-\alpha)\%$ level confidence interval for θ, then we could select the w such that the posterior distribution of θ, with parameter w, is such that

$$P\{\theta \in C_\alpha(w, x_1, \ldots, x_n)|x_1, \ldots, x_n\} = 1 - \alpha.$$

See, for example, Datta and Sweeting (2005) for references to probability matching priors and posteriors, and Cooley *et al.* (2009) for ideas in pseudo-Bayesian approaches with composite likelihoods.

More generally we can consider the subjective setting of w where knowledge of the frequentist sampling statistic of $\Sigma_{i=1}^{n} l(\theta, x_i)$ can assist. To begin note that w is explicitly related to the Bayes factor quantifying the posterior-to-prior odds,

$$\log\left\{ \frac{\pi(\theta|x)}{\pi(\theta'|x)} \Big/ \frac{\pi(\theta)}{\pi(\theta')} \right\} = -w\{l(\theta, x) - l(\theta', x)\}$$

where $w\{l(\theta, x) - l(\theta', x)\}$ measures the update in beliefs in favour of θ from θ' on observing x. Clearly the larger the difference $l(\theta, x) - l(\theta', x)$ is the greater the relative evidence in favour of θ, with w determining the scale for unit change. It is interesting to note that, should the Bayes factor be known for any three points $\{\theta, \theta', x\}$ in the joint parameter sample space, $\Omega_{\theta^2} \times \Omega_{X^n}$, then w would be fixed. The idea here is that the analyst is free to contemplate any specific values $\{\theta, \theta', x\}$ for which the distribution of the statistic $S = l(\theta, x) - l(\theta', x)$ may be known, and to use this knowledge in turn to help to elicit a Bayes factor and therefore setting w. A concrete example will help.

Suppose that θ_0 denotes the unknown mean of a population with prior $N(0, v)$ and loss function $l(\theta, x) = \Sigma_{i=1}^{n} (\theta - x_i)^2$. Consider the design points $\{\theta = \bar{x}, \theta' = 0, x\}$ so that the statistic S is then

$$S = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} (\bar{x} - x_i)^2,$$

the difference in the sum of squares to the sum of squares around the mean, with log-Bayes-factor

$$\log\left\{\frac{\pi(\theta|x)}{\pi(\theta'|x)}\Big/\frac{\pi(\theta)}{\pi(\theta')}\right\}=-wS.$$

The analyst is free to contemplate any value of n and any $x=\{x_1,\ldots,x_n\}$ to help in the elicitation. Let n be chosen large and contemplate x such that the $(1-\alpha)\%$ confidence interval for the unknown mean touches $\theta'=0$. In this case, for large n, we know that $S=F^{-1}_{1,n-1}(1-\alpha)$, where F denotes the F-distribution. If the analyst is prepared to say how their prior beliefs would be updated on observing x in knowledge of this symmetric confidence interval for θ_0 then the w can be set via

$$w=-\log(\text{Bayes factor})/S.$$

We give a concrete illustration of this approach in Section 5.

3.5. Conjugate loss prior

If prior beliefs about θ can be expressed in the form

$$\pi(\theta)\propto\exp\{-\lambda\, l(\theta,\mu)\}$$

for given parameters (λ,μ), then the posterior has a conjugate-type property, i.e.

$$\pi(\theta|x)\propto\exp\{-w\, l(\theta,x)-\lambda\, l(\theta,\mu)\}.$$

Thus the prior has interpretation of prior observation μ with precision λ. Thus μ and λ would be standard objects for a Bayesian to specify. If the prior can then be established as the equivalent of m observations, then we obtain w via $w/\lambda=1/m$.

If the prior is thought not to be able to be specified in such a way, then a good approximation to any prior can be found with choices of $(M,(\mu_j),(\lambda_j))$ such that

$$\pi(\theta)\propto\exp\left\{-\sum_{j=1}^{M}\lambda_j\, l(\theta,\mu_j)\right\}.$$

If we now write

$$\pi(\theta|x)\propto\exp\left\{-w\, l(\theta,x)-\Lambda\sum_{j=1}^{M}(\lambda_j/\Lambda)\, l(\theta,\mu_j)\right\},$$

where $\Lambda=\Sigma_{1\leqslant j\leqslant M}\lambda_j$, then we see that now $w/\Lambda=1/m$.

Thus there is an apparent new concept here in that the experimenter is required to think about how much information, in the form of the number of prior observations, is available. However, this is not completely new, since in some conjugate problems there are parameters which do have the interpretation of a prior sample size; the exponential family, for example.

4. General forms of information

In this section we discuss more general forms of information x rather than assume that it arises from some unknown $F_0(x)$. The argument is that provided that $l(\theta,x)$ has been specified then an update of a belief distribution about θ is available. Clearly this does not rely on any assumption about where x came from or indeed how it became known.

In particular, we provide a definition of conditional probability when non-stochastic information is available. This allows for updating or refinement of prior beliefs to be applied in much more general settings than Bayesian models, which require a stochastic x.

4.1. Conditional probability distributions and non-stochastic data

The theory of conditional probability distributions is a well-established mathematical theory which provides a procedure to update prior probabilities taking into account new information. Such a procedure is available *only* if the information which is used to update the probability concerns stochastic events, i.e. events to which a probability is preassigned. In other words, such information needs to be already included in the probability model. In this section, we shall show how the updating approach can be used to define conditional probability distributions based on non-stochastic information.

Information about θ may arrive in the form of non-stochastic data, such as if an expert declares that

$$I = \text{`}\theta \text{ is close to 0'}. \tag{13}$$

This type of information has been discussed by various researchers and is known to be problematic for the Bayesian especially when such information arises after or during the arrival of stochastic observations (x_i). We cite Diaconis and Zabell (1982) and in particular refer the reader to the example in section 1.1 of their paper.

We denote by I a piece of information for which no probability model for each θ is assigned; in other words I is not and cannot be considered stochastic in any way. So we cannot represent equation (13) by using a probability model whereby we could reasonably assume $I \sim F_0(\cdot)$ in any meaningful sense.

Although a probability model cannot connect equation (13) and θ, they can be connected via a loss function without much difficulty. For example, $l(\theta, I) = w\theta^2$ for some $w > 0$ could be deemed appropriate. Note here that we use I to denote information now, replacing the stochastic x. The update $\hat{\nu}(\theta)$ based on I and π can then be considered as a means of defining an operational conditional probability distribution in the presence of non-stochastic information, given by

$$\hat{\nu}(\theta|I) = \frac{\exp\{-w\, l(\theta, I)\}\, \pi(\theta)}{\displaystyle\int \exp\{-w\, l(\theta, I)\}\, \pi(\mathrm{d}\theta)}.$$

So, the general Bayesian approach provides a general definition of conditional distributions based on non-stochastic information, which may also be useful in the construction of priors from multiple information sources.

For literature on paradoxes related to forcing non-stochastic events into a probability model with a determination of all the alternatives to I we refer the reader to Freund (1965), Gardener (1959), Bar-Hillel and Falk (1982) and Hutchison (1999, 2008).

4.2. Partial information

As noted in Section 2, although the parameter of interest is θ, the information I that is collected may be more informative, i.e. there is within I information which does not assist with the learning about θ, for which it is possible to identify $I_\theta \subset I$ which provides *all* the information about θ. We are therefore interested in constructing the loss function $l(\theta, I_\Theta)$, leading to

$$\hat{\nu}(\mathrm{d}\theta) \propto \exp\{-l(\theta, I_\Theta)\}\, \pi(\mathrm{d}\theta). \tag{14}$$

The partial likelihood, or partial self-information loss, that is used in proportional hazards models is one such example. Whereas Bayesian practitioners may have adopted such a procedure in the past it would be regarded as lacking motivation. However, our point is that expression (14) represents a valid update of beliefs. We illustrate this approach in Section 5.

5. Illustrations

In this section we discuss the application of our approach to important inferential problems. The first problem is one from survival analysis where we have a well-motivated proxy likelihood based on partial information, and hence it is natural to use $w = 1$ in this setting. The second example is from model-free clustering where we have a general loss function so that calibration of w is important. A third example, which is to be found in the on-line supplementary material, is for joint inference on a set of quantiles. In all cases we claim that the choice of loss function is well founded (and unique) and that there is no traditional Bayesian interpretation of the updates that we are implementing. Yet the updates that we employ do allow us to learn about the specified parameters of interest. All of the models that are used to generate results are available as open-source code in R or MATLAB.

5.1. Colon cancer genetic survival analysis

Colon cancer is a major worldwide disease with increasing prevalence particularly within western societies. Exploring the genetic contribution to variation in survival times following incidence of the cancer may shed light into the disease aetiology and underlying disease heterogeneity. For this collaborators at the Wellcome Trust Centre for Human Genetics, University of Oxford, obtained survival times on 918 cancer patients with germline genotype data at hundreds of thousands of markers genomewide. For demonstration we consider only one chromosome previously identified as holding a potential association signal containing 15608 genotype measurements. The data table X then has $n = 918$ rows and $p = 15608$ columns, where $(X)_{ij} \in \{0, 1, 2\}$ denotes the genotype of the ith individual at the jth marker. Alongside this we have the corresponding $n \times 2$ response table of survival times Y with a column of event times, $y_{i1} \in \mathbb{R}^+$ and a column of indicator variables $y_{i2} \in \{0, 1\}$, denoting whether the event is observed or right censored at y_{i1}.

To explore association between genetic variation and time to event we employ a loss function derived under proportional hazards, treating the loss to the baseline hazard as a nuisance parameter. This is based on the Cox proportional hazards model, which has been one of the most widely used methods in survival analysis since its introduction in Cox (1972). In this log-linear model the hazard rate at time t for an individual with covariate $\mathbf{x} = \{x_1, \ldots, x_p\}$ is defined as

$$h(t|\mathbf{x}) = h_0(t) \exp\left(\sum_{j=1}^{p} x_j \beta_j \right)$$

where $h_0(t)$ is a baseline hazard function. In the seminal work of Cox (1972), $h_0(t)$ is treated as a nuisance parameter (or process) that does not enter the partial likelihood for estimating the parameters of interest $\boldsymbol{\beta}$.

In contrast, a Bayesian approach to the Cox model necessarily involves the baseline hazard function. There is a limiting argument for the use of the partial likelihood but this is rarely, if at all, used. Most common is the finite partitioning of the time axis and using a piecewise constant baseline hazard function. Though typically regarded as a nuisance parameter, the Bayesian must

specify a full probability model for it. See Ibrahim *et al*. (2001), chapter 3, for details, where they noted that the proportional hazards model is obtained under a limiting improper prior on the baseline, but it is not known what effect this has on marginal quantities of interest such as marginal model choice probabilities.

Using a general Bayes construction we can consider only the order of events as partial information relevant to the regression coefficients $\boldsymbol{\beta}$, via the cumulative loss function,

$$l(\boldsymbol{\beta}, \mathbf{x}) = \sum_{i=1}^{n} \log \left\{ \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j \right)}{\sum_{l \in R_i} \exp\left(\sum_{j=1}^{p} x_{lj}\beta_j \right)} \right\}, \tag{15}$$

where R_i denotes the risk set, i.e. those individuals alive or not censored at time y_{i1}, and in this way obtain a conditional update. We assume that $\beta_j \sim N(0, v_j)$ and set $v_j = 0.5$ for our study, reflecting beliefs that associated coefficients will be modest, and we note that one advantage of our approach is that subjective prior information can be integrated into the analysis.

Initially we consider each marker in turn for evidence of effects, i.e. $\beta_j \neq 0$, within a univariate regression and we can calculate the general Bayes factor of association at the jth marker, assuming equal prior probability in there being an effect or not, as,

$$\mathrm{gBF}^{(j)} = \frac{\int_{\beta_j} \exp\{-l(\beta_j|\mathbf{x}_j)\}\pi(\beta_j)\,\mathrm{d}\beta_j}{\exp\{-l(\beta_j = 0|\mathbf{x}_j)\}}$$

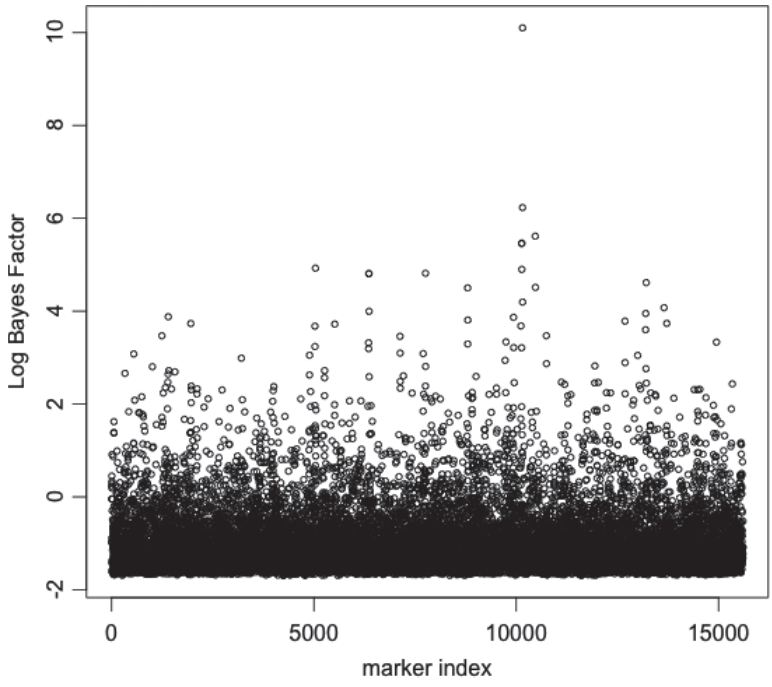
which involves a one-dimensional integral that we calculate via importance sampling.

We calculated the general Bayes factors for each marker and in Fig. 1(a) we plot the log-general-Bayes factors over the chromosome. Although there is considerable variation we observe strong evidence of association around marker 10000. It is interesting to compare the evidence of association that is provided by the Bayes factor Fig. 1(a) with that obtained by using a conventional Cox proportional hazards partial-likelihood-based test. In Fig. 1(b) we plot the log-general-Bayes factors against $-\log_{10}(p\text{-values})$ obtained from a conventional likelihood ratio test at each marker. We can see general agreement especially at the markers with strongest association as we would expect for a large sample size. Interestingly there appears to be greater dispersion at markers of weaker association as highlighted in Fig. 2 where we plot the standard error against log-general-Bayes factors. Markers with high standard error relate to genotypes of rarer alleles and the attenuation reflects a greater degree of uncertainty for association at these markers that contain less information.

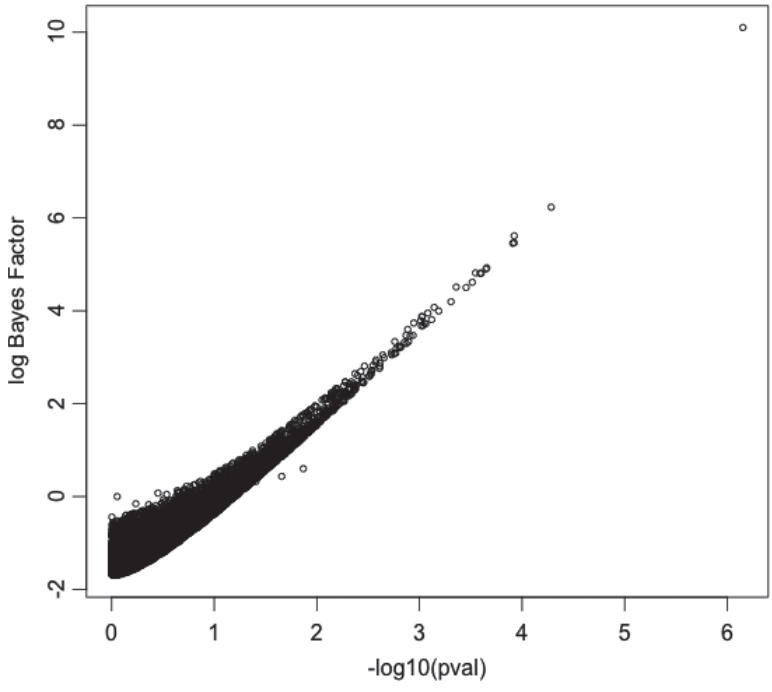
Returning to the 'hit region' showing strongest association around marker 10000, owing to high collinearity between markers it is not clear whether the signal of association arises from a single effect correlated with others, or from multiple independent association signals. To investigate this we developed multiple-marker methods.

We consider a model using potentially all 800 markers in the region and phrase the problem as a variable selection task under a partial likelihood (loss), in which the user suspects that some of the $p = 800$ recorded covariates (15) may not be relevant to variation in survival times.

In the non-Bayesian paradigm, variable selection can proceed by defining a cost function, such as the Akaike information criterion AIC or Bayesian information criterion BIC, that adjusts the fit to the data by the number of covariates in the model. Inference proceeds by using an optimization algorithm, such as forward or stepwise selection, to find a model that minimizes



(a)



(b)

Fig. 1. (a) log-Bayes-factor *versus* marker index and (b) log-Bayes-factor *versus* $\log_{10}(p\text{-value})$ of association

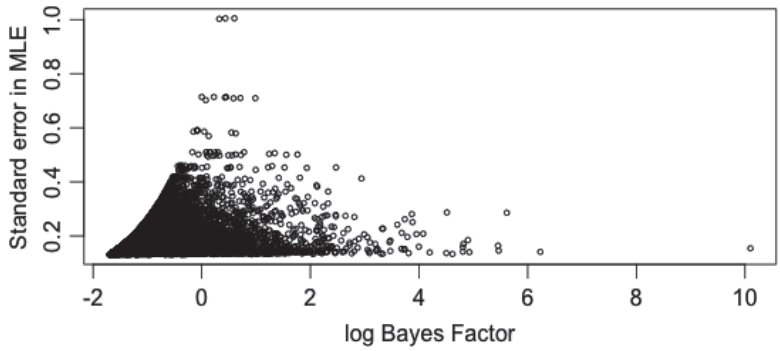


Fig. 2. Standard error in maximum likelihood estimate *versus* log-Bayes-factor

the cost. More recently, penalized likelihood methods have proved popular (Tibshirani, 1997; Fan and Li, 2002).

Despite the enormous influence of Cox proportional hazards models and the importance of variable selection, the Bayesian literature in this area is limited. This is because of the lack of a theoretical foundation to treat $h_0(t)$ as a nuisance parameter, leading to either approximate methods or the full specification of a joint probability model (Faraggi and Simon, 1998; Volinsky *et al*., 1997). Volinsky *et al*. (1997) took BIC as an approximation to the marginal likelihood and they used a branch-and-bound algorithm to find a set of models with differing sets of covariates with high BIC-scores. The difficulty here is that, although the methods are important and well motivated, they are ultimately *ad hoc*. Moreover, prior information on $\pi(\boldsymbol{\beta})$ does not enter the calculation of BIC, meaning that an important aspect of the Bayesian approach is lost.

In contrast, Ibrahim *et al*. (1999) considered variable selection within a full joint model using a prior specification of a gamma process for the baseline hazard (see also Ibrahim *et al*. (2001)). This provides a formal Bayesian solution but inference is then conditional on, and sensitive to, the specification of the prior on $h_0(t)$, which is something that the partial likelihood model explicitly avoids.

Here we use the partial information that is relevant to the regression coefficients $\boldsymbol{\beta}$ via the cumulative loss function (15). We assume proper priors $\pi(\boldsymbol{\beta})$ on the regression coefficient,

$$\pi(\beta_j) = \begin{cases} 0 & \text{if } \delta_j = 0, \\ N(0, v_j) & \text{otherwise,} \end{cases}$$

where $\delta_j \in \{0, 1\}$ is an indicator variable on covariate relevance with $\pi(\delta_j) = \text{Bin}(a_j)$ and we now treat $\{\delta_1, \ldots, \delta_{800}\}$ as a vector in a joint model. In this way the posterior $\pi(\boldsymbol{\delta}|\mathbf{x})$ quantifies beliefs about which variables are important to the regression. We use Markov chain Monte Carlo (MCMC) sampling to draw samples approximately from $\pi(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{x})$ from which the marginal distribution on $\boldsymbol{\delta}$ can be examined. In particular we make use of an efficient joint updating proposal, $q(\boldsymbol{\delta}', \boldsymbol{\beta}'|\boldsymbol{\delta})$, within the MCMC algorithm as $q(\boldsymbol{\delta}', \boldsymbol{\beta}'|\boldsymbol{\delta}) = q(\boldsymbol{\delta}'|\boldsymbol{\delta})\, q(\boldsymbol{\beta}'|\boldsymbol{\delta}')$ where $q(\boldsymbol{\delta}'|\boldsymbol{\delta})$ proposes a local move to add, remove or swap one variable per MCMC iteration in or out of the current model indexed by $\boldsymbol{\delta}$, and $q(\boldsymbol{\beta}'|\boldsymbol{\delta}')$ is a joint independence Metropolis update proposal, $q(\boldsymbol{\beta}'|\boldsymbol{\delta}') = N(\tilde{\boldsymbol{\beta}}_{\delta'}, \tilde{\mathbf{V}}_{\delta'})$ where $\{\tilde{\boldsymbol{\beta}}_{\delta'}, \tilde{\mathbf{V}}_{\delta'}\}$ are the maximum *a posteriori* and approximate information matrix obtained from the combination of log-partial-loss and normal prior. The joint proposal is then accepted with probability

$$\alpha = \min\left[1, \frac{\exp\{-l(\boldsymbol{\beta}'|\mathbf{x})\}\, \pi(\boldsymbol{\beta}'|\boldsymbol{\delta}')\, \pi(\boldsymbol{\delta}')\, q(\boldsymbol{\beta}, \boldsymbol{\delta}|\boldsymbol{\delta}')}{\exp\{-l(\boldsymbol{\beta}|\mathbf{x})\}\, \pi(\boldsymbol{\beta}|\boldsymbol{\delta})\, \pi(\boldsymbol{\delta})\, q(\boldsymbol{\beta}', \boldsymbol{\delta}'|\boldsymbol{\delta})}\right].$$

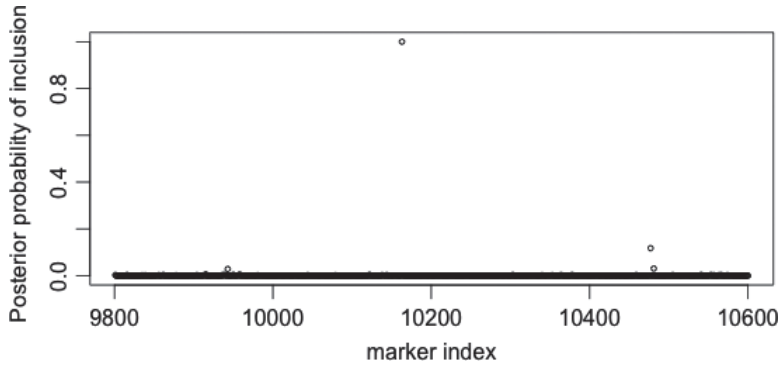


Fig. 3. Posterior marginal inclusion probability from the multiple-marker model using 800 markers around the peak of association

We ran our MCMC algorithm for 100 000 iterations with prior parameter settings, $\{v_j = 0.5, a_j = 1/800\}$, for all $j = 1, \ldots, p$, equivalent to a prior assumption of a single associated marker. In Fig. 3 we show the marginal inclusion probability, after discarding 10 000 samples as a burn-in. The algorithm showed an overall acceptance rate of 8% for proposed moves. The model suggests overwhelming evidence for a single marker in the region of index 10 200 but also weaker evidence of independent signal in a couple of other regions.

5.2. *Bayesian model-free clustering*

Cluster analysis is one of the most widely used and important areas of applied statistics (Hastie *et al.*, 2009). In cluster analysis, a primary objective is to identify self-similar groups within data such that observations within a group are deemed more closely related to one another than observations between groups, K-means clustering being arguably the most popular clustering method in use today.

The clustering problem is interesting from a formal Bayesian perspective as it raises several challenges. The object of interest is the cluster partition mapping S, which allocates observations to clusters. However, the partition S as it stands is not a generative model (a sampling distribution for observables). To implement clustering the Bayesian analyst is forced to define a sampling distribution for observations within a cluster, $f(x|C_j)$, where C_j denotes parameters that are associated with the jth cluster, with an associated prior probability of cluster membership p_j. This leads to the well-known marginal mixture representation

$$f(x|C) = \sum_{j=1}^{K} p_j f_j(x|C_j),$$

the canonical example being with Gaussian mixture components, $f(x|C_j) = N(\mu_j, \Sigma_j)$, which necessitates a further layer of hierarchical priors $\pi(\mu_j, \Sigma_j)$. Cluster membership can be sensitive to the choice of sampling distribution and hierarchical prior, both of which are nuisance to the task, and computation is complicated by the well-known label switching problem (Jasra *et al.*, 2005).

Non-Bayesian model-free segmentation methods have a distinct advantage in allowing the analyst to concentrate on the object of interest, namely the clustering S, typically defined through the specification of a pairwise dissimilarity score between observations $d(x_i, x_j)$. An optimization algorithm is then used to find the optimal partition $\hat{S}$ which minimizes the score over pairs within

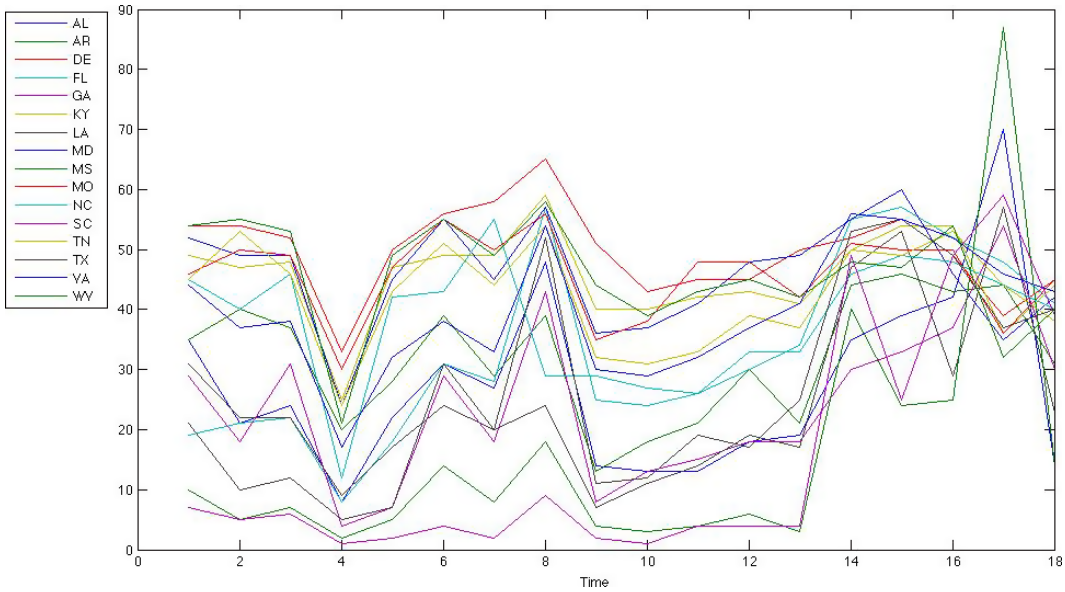


Fig. 4. Voting of southern states, illustrating the percentage of the Republican vote for Presidential elections every 4 years beginning in 1900: AL, Alabama; AR, Arkansas; DE, Delaware; FL, Florida; GA, Georgia; KY, Kentucky; LA, Louisiana; MD, Maryland; MS, Mississippi; MO, Missouri; NC, North Carolina; SC, South Carolina; TN, Tennessee; TX, Texas; VA, Virginia; WV, West Virginia

and/or between clusters. However, quantifying uncertainty in $\hat{S}$, even assuming that the global minima can be found, is far from trivial as we typically have only a single realization of dependent multivariate data x, although see Seldin and Tishby (2001), who used PAC Bayesian ideas to consider uncertainty in a predictive regression model when clustering the covariates.

We define a prior distribution directly on the partition, $\pi(S)$, and a loss function $l(S, x_1, \ldots, x_n)$ and we use general Bayesian updating. To illustrate this we consider uncertainty analysis of a classic data set considered in Hartigan (1972), illustrated in Fig. 4, in his highly influential paper that introduced biclustering. Biclustering refers to the simultaneous clustering of observations and covariates (rows and columns) of a data matrix and has proved extremely useful in modern application areas, particularly in genomics (Cheng and Church, 2000; Tanay *et al.*, 2002; Heard *et al.*, 2005).

Hartigan (1972) considered the percentage Republican Presidential vote of 16 southern states in the USA over 18 elections covering the years 1900–1968. Hartigan treated the time series as independent covariates in his co-clustering approach. Here, for simple illustration, we maintain the time series ordering, so that the co-clustering is akin to clustering multiple-change-point time series with common but unknown change points. We assume that the cluster memberships are constant over time, but the time series change at specific break points. Our loss function is defined as in Hartigan (1972) using a sum-of-squares decomposition,

$$l(S, x_1, \ldots, x_n) = w \sum_{C_k \in S} \sum_{ij \in C_k} (x_{ij} - \bar{x}_{C_k})^2,$$

where i denotes state and j denotes time, C_k denotes the kth grouping of states over a particular time period and $\bar{x}_{C_k}$ denotes the mean over all $(i, j) \in C_k$. The posterior distribution is therefore

$$P(S|x) \propto \pi(S) \exp\{-l(S, x)\}.$$

The setting of the loss parameter w is a crucial part of the model specification. Following the procedures that were discussed in Section 3, it is difficult to consider a conjugate specification or a unit information prior on the discrete structures. We instead propose to use a frequentist calibration approach in the following manner. Recall that under a flat prior on S we can set w via a subjective assessment of the posterior ratio at a reference point,

$$\frac{P(S|x)}{P(S'|x)} = \exp[-w\{l(S,x) - l(S',x)\}],$$

and where we can solve for w if all the other elements are given. In elicitation of w we propose to make use of classical results from analysis of variance. We take as our first reference point the null partition using a single global cluster, so that the loss $l(S,x) = \Sigma_i \Sigma_j (x_{ij} - \bar{x})^2$ is simply the sum of squares around the mean. Then consider a randomized data partition $\{x, S'\}$ that allocates the data uniformly at random to k clusters. Under this scheme we expect that

$$\frac{\{l(S,x) - l(S',x)\}/k - 1}{l(S,x)/(n-k)} \sim F_{k-1,n-k}$$

where F denotes the F-distribution. We can then use the F-distribution to help in the calibration. For example, if we consider a point in the tails of F, such that $f^*_\alpha = F^{-1}_{k-1,n-k}(\alpha)$ with $\alpha \in (0,1)$, and specify

$$l(S',x) = \frac{l(S,x)}{1 + f^*(k-1)/(n-k)}$$

then $l(S',x)$ represents the value of loss such that a randomized allocation has probability $1-\alpha$ of producing a smaller loss. Equivalently, with probability α a random allocation would lead to a reduction in loss as high as $1 + f^*_\alpha(k-1)/(n-k)$ relative to the single cluster. When α is large we can be confident that a partition achieving a loss of $l(S',x)$ represents a significant clustering. The analyst can then calibrate w in the following way.

(a) Define a reference value for $R = P(S|x)/P(S'|x)$ under a uniform prior, setting R small, say $R = 0.001$, relative to the global cluster S.

Table 1. Average loss of partitions across MCMC samples (and log-posterior probabilities in parentheses)†

Number of state clusters k_s	*Average loss $\times 10^4$ for the following numbers of change points in time k_t (groups $= k_t + 1$)*		
	$k_t = 0$	$k_t = 1$	$k_t = 2$
1	7.98 (−14.49)	6.82 (−14.34)	6.72 (−14.73)
2	5.36 (−13.69)	5.13 (−13.65)	3.19 (−13.58)
3	5.09 (−13.64)	3.92 (−13.38)	2.36 (*−13.28*)
4	4.99 (−13.91)	3.32 (−13.50)	2.02 (−13.41)

†The average loss is $T^{-1} \Sigma_{i=1}^{T} l(S_i, x)$ with $S_i \sim \pi(S|x, k_s, k_t)$, where k_s denotes the number of clusters of states and k_t denotes the number of time series change points. Log-posterior-probabilities are shown in parentheses using a Poisson(3) and Poisson(2) prior on the number of groups and number of time clusters $k_t + 1$. The maximum posterior clustering is shown in italics.

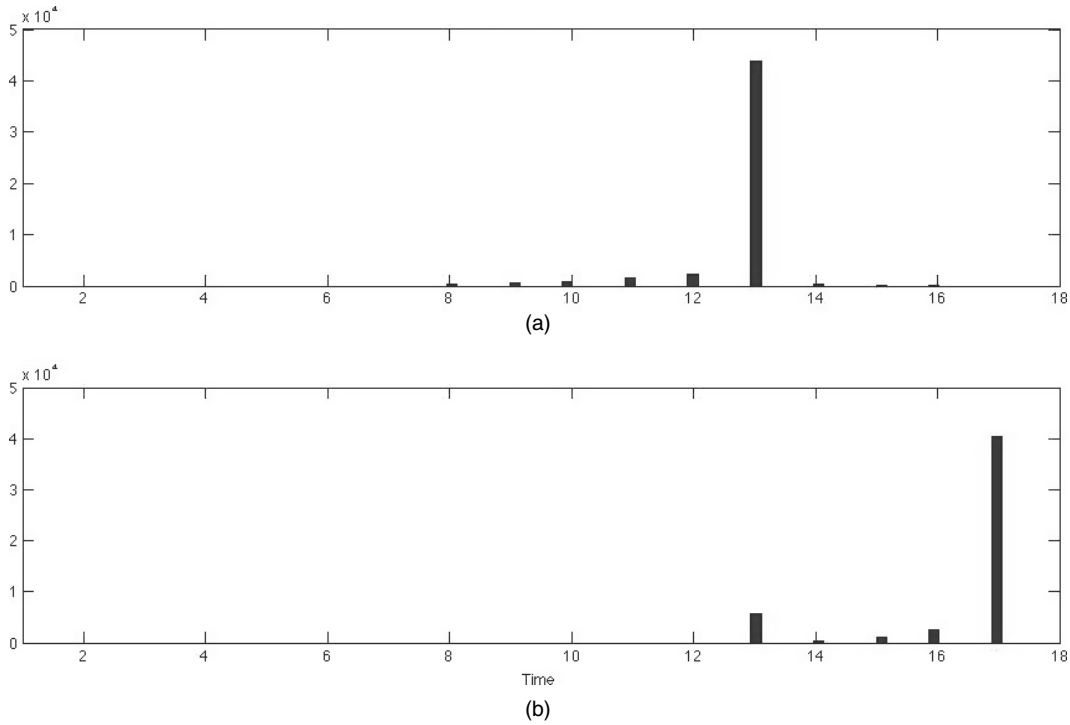


Fig. 5. Time change point locations for the two-change-point, $k_t = 2$, model and $k_s = 3$ groups: (a) change point 1; (b) change point 2

(b) Define a tail area value α such that, should a partition S' achieve a relative reduction in loss of $1 + f^*_\alpha(k-1)/(n-k)$, then you would assign relative posterior beliefs of R.
(c) Set $w = -\log(R)\, l(S,x)\, f^*_\alpha(k-1)/(n-k)$.

For the election data we found that w is quite stable to the calibration choice of $\{\alpha, R\}$; for example with $k = 3$ we find $w = 0.0036$ for $\{\alpha = 0.99, R = 0.01\}$ and $w = 0.0012$ for $\{\alpha = 0.999, R = 0.01\}$. We choose $w = 0.0012$ and ran an MCMC algorithm for 100 000 iterations using a burn-in of 50 000. The iteration numbers were chosen after experimentation to deliver stable results over multiple runs. The MCMC algorithm was rerun for differing numbers of partitions of states and differing number of time series change points. Table 1 presents the results of the average loss achieved over each run alongside the estimate of the posterior probability for each configuration shown in parentheses by using a Poisson(3) prior on the number of groups and a Poisson(2) prior on the number of *time groupings*, which is the number of change points plus 1. Note that the first column in Table 1 equates to standard clustering of the states with zero change points in time, whereas the first row represents a multivariate change point model. Table 1 suggests strong evidence for clustering in both time and across states. The maximum posterior probability favours the model with three groups of states and three time groupings.

We investigated uncertainty in the partitions and in the cluster allocation of the maximal posterior model. To illustrate this we plot in Fig. 5 the distribution of the location of time series change points for the $\{k_s = 3, k_t = 2\}$ model. We can see strong evidence that the change points occur late in the series, which is visually supported by the data in Fig. 4. The pairwise co-clustering probabilities of this model are shown in Fig. 6, where each element represents

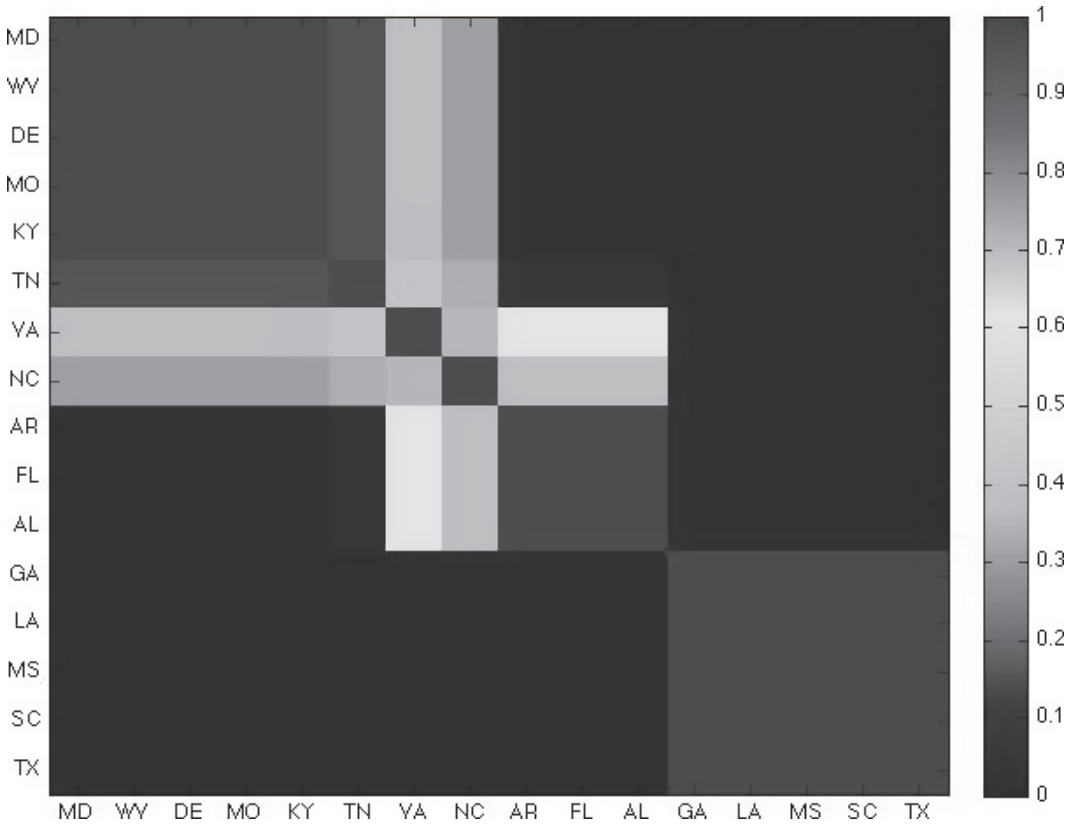


Fig. 6. Pairwise co-clustering probabilities across three groups and two time change points: AL, Alabama; AR, Arkansas; DE, Delaware; FL, Florida; GA, Georgia; KY, Kentucky; LA, Louisiana; MD, Maryland; MS, Mississippi; MO, Missouri; NC, North Carolina; SC, South Carolina; TN, Tennessee; TX, Texas; VA, Virginia; WV, West Virginia

the pairwise probability events $\Sigma_S P(S|x)\mathbf{1}[\mathrm{CI}(x_i) = \mathrm{CI}(x_j)|S]$, where $\mathrm{CI}(x_i)$ is the cluster index for the ith state. The cluster blocks show strong concordance with the single co-cluster that was reported by Hartigan; see Fig. 6(a) in Hartigan (1972). However, our method highlights considerable uncertainty in the pairing of Virginia and North Carolina, which is something that we can quantify by using our general Bayesian approach.

6. Discussion

We have provided a basis for general Bayesian learning and the updating of information by using belief probability distributions. Loss functions constructed on spaces of probability measures allow for coherent updating. Specifically, information is connected to the parameter of interest via a loss function and this is the fundamental concept, replacing the restrictive connection based on probability models. We can recover precisely the traditional updating rules such as the Bayes rule when we select the self-information loss function, when it is appropriate to do so.

The assumptions that we make are minimal: that information can be connected to unknown parameters via loss functions and that individuals then act rationally by minimizing their expected loss. If information is assumed to come from some probability model then we can accommodate this within our framework by appealing to the self-information loss function equivalent

to the negative log-likelihood and so we can argue that loss functions are sufficient for learning mechanisms that are currently in use.

More generally, we can use loss functions that are currently employed in a classical context for robust estimation, e.g. generalized estimating equations. We can also deal with partial information where it is only a part of some observed information that is useful or relevant for learning about the decision-making process based on a particular relevant parameter of interest.

We have developed a rigorous approach to updating beliefs where we are required only to think about which is the best parameter from a chosen model needed to make a decision rather than have to think about a non-existent true model parameter which coincides with the true data-generating mechanism.

We believe that it is fundamental to identify parameters of interest through loss functions. The alternative route through a probability model is, we argue, highly restrictive and leads to narrow types of Bayesian updating. The necessary supporting theory for us is minimal (the construction and minimization of loss functions), whereas for the use of probability models it is more intricate and restrictive.

Acknowledgements

The authors are grateful for the detailed comments of two reviewers and an Associate Editor on revisions of the paper.

Holmes is supported by the Oxford-Man Institute, Oxford, the Engineering and Physical Sciences Research Council programme grant i-like EP/K014463/1, and the Medical Research Council, UK, grant MC_UP_A390_1107. Walker is partially supported by National Science Foundation grant DMS 1506879.

Appendix A: Proof of theorem 1

In the two-point case, say $\Theta = \{0, 1\}$, the prior π is determined by a real number z in the unit interval, and the loss $l(\theta, x)$ takes only two values, say l_0 and l_1. By condition 5, we can replace the pair (l_1, l_0) with $(l_1 - l_0, 0)$. Therefore, the posterior is a function of $l = l_1 - l_0$ and z, say $\bar{\psi}(l, z)$. To proceed, it is convenient to think in terms of odds rather than probabilities. So, if z is the prior for 0, we can consider $t = z/(1-z)$ and $z = t/(1+t)$. We can do the same with the posterior. The posterior odds are a function of the loss difference l and the prior odds t, which we denote by $\phi(l, t)$.

Dealing with the odds, equation (3) becomes

$$\phi(l + h, t) = \phi\{l, \phi(h, t)\}, \tag{16}$$

where h is a replication of l with a possibly different x. Moreover, with a constant loss, the posterior is equal to the prior (condition 4), i.e.

$$\phi(0, t) = t, \tag{17}$$

for every $t > 0$.

At this stage, we consider a prior with more than two mass points, say three and given by $\{1, 2, 3\}$; the prior is given by $\{z_1, z_2, 1 - z_1 - z_2\}$, with loss functions $\{l_1(x), l_2(x), l_3(x)\}$. The loss can be 0 at one point without loss of generality (condition 5), so it takes values $(l_1, l_2, 0)$. Let us consider the updating rule ϕ for priors with just two mass points. We can use this to update the conditional probability of 1 given $\{1, 3\}$, i.e. $z_1/(z_1 + z_3)$.

For this, we update the prior with masses $z_1/(z_1 + z_3)$ and $z_3/(z_1 + z_3)$ considering just the loss values $(l_1, 0)$, i.e. disregarding the point 2 with its loss l_2. In other words, we aim at the right-hand side of equation (9).

Denote by $t_{1,3}$ the odds corresponding to the conditional probability of 1 given $\{1, 3\}$, i.e. $t_{1,3} = z_1/z_3$. We update $t_{1,3}$ on the basis of the loss values l_1 and 0, i.e. with $\phi(l_1, t_{1,3})$. Similarly, we can define $t_{1,2} = z_1/z_2$ and $t_{2,3} = z_2/z_3$. We update $t_{1,2}$ on the basis of l_1 and l_2, i.e. with $\phi(l_1 - l_2, t_{1,2})$ and $t_{2,3}$ with $\phi(l_2, t_{2,3})$. Clearly, $t_{1,3} = t_{1,2} t_{2,3}$, and this factorization of conditional odds must hold also after updating, i.e.

$$\phi(l_1, t_{1,3}) = \phi(l_1 - l_2, t_{1,2})\, \phi(l_2, t_{2,3}), \tag{18}$$

where $t_{1,3} = t_{1,2}t_{2,3}$. Formally, this identity is a consequence of condition 2, i.e. updating the conditional probability is the same as conditioning the updated probability. Since equation (18) must hold for every $t_{2,3}, t_{1,2} > 0$ and for every $l_1, l_2 \in \mathbb{R}$ we have that

$$\phi(l_1, ts) = \phi(l_1 - l_2, t)\, \phi(l_2, s),$$

for every $t, s > 0$ and $l_1, l_2 \in \mathbb{R}$. If in particular $l_2 = 0$, and say $l_1 = l$, recalling equation (17), we have that

$$\phi(l, ts) = \phi(l, t)s,$$

for every $t, s > 0$ and every real l. Letting $t = 1$, we find that

$$\phi(l, s) = \phi(l, 1)s, \tag{19}$$

for every $s > 0$ and every $l \in \mathbb{R}$. A combination of equations (16) and (19) yields $\phi(l + h, 1) = \phi\{l, \phi(h, 1)\} = \phi(l, 1)\, \phi(h, 1)$, for every $h, l \in \mathbb{R}$, which in turn, it is known, implies that $\phi(l, 1) = \exp(-wl)$ for some $w \in \mathbb{R}$, provided that $\phi(l, 1)$ is a monotone function of l owing to condition 3. Hence,

$$\phi(l, t) = \exp(-wl)t, \tag{20}$$

for every $l \in \mathbb{R}$ and every $t > 0$. In this way, we are basically done with the two-points case.

Let us now consider the three-point case: we want to update the prior with mass points at $\{1, 2, 3\}$ given by $\{z_1, z_2, 1 - (z_1 + z_2)\}$, where z_1 and z_2 are non-negative and their sum is less than or equal to 1, and it is convenient to set $z_3 := 1 - (z_1 + z_2)$. In terms of odds, the prior is given by the pair (t_1, t_2) (being $t_i = z_i/(1 - z_i)$, or equivalently $z_i = t_i/(1 + t_i) = 1 - 1/(1 + t_i)$, $i = 1, 2$) where $t_1, t_2 \geqslant 0$ and $t_1/(1 + t_1) + t_2/(1 + t_2) \leqslant 1$. Here, it is convenient to set $t_3 := [\{1 - t_1/(1 + t_1) + t_2/(1 + t_2)\}^{-1} - 1]^{-1}$. In this setting, we consider the loss values $(l_1, l_2, 0)$. Moreover, we consider an $\mathbb{R}^2$-valued function $\boldsymbol{\phi}(\mathbf{l}, \mathbf{t}) = (\phi_1(\mathbf{l}, \mathbf{t}), \phi_2(\mathbf{l}, \mathbf{t}))$, where $\mathbf{l} = (l_1, l_2)$ and $\mathbf{t} = (t_1, t_2)$. Now the question is how we could recover $\boldsymbol{\phi}$ from ϕ, where the latter gives the updating rule for the two-points case. Recall the notation that we have used for the conditional odds, i.e. $t_{i,j} = z_i/z_j$ are the odds corresponding to the conditional probability of i given $\{i, j\}$, for distinct $i, j = 1, 2, 3$. We can see that $t_1 = z_1/(z_2 + z_3) = (t_{2,1} + t_{3,1})^{-1}$. By condition 2, this identity will have to be satisfied also by the updated odds. Since we update $t_{2,1}$ with $\phi(l_2 - l_1, t_{2,1})$ and $t_{3,1}$ with $\phi(-l_1, t_{3,1})$, we must have

$$\phi_1(l_1, l_2; t_1, t_2) = \{\phi(l_2 - l_1, t_{2,1}) + \phi(-l_1, t_{3,1})\}^{-1},$$

which by equation (20) becomes

$$\phi_1(l_1, l_2; t_1, t_2) = \exp(-wl_1)\{\exp(-wl_2)t_{2,1} + t_{3,1}\}^{-1}. \tag{21}$$

As $t_{2,1} = z_2/z_1$ and $t_{3,1} = z_3/z_1$, equation (21) becomes

$$\phi_1(l_1, l_2; t_1, t_2) = \frac{\exp(-wl_1)z_1}{\exp(-wl_2)z_2 + z_3},$$

and the updated probability of 1 will be

$$1 - \{1 + \phi_1(l_1, l_2; t_1, t_2)\}^{-1} = \frac{\exp(-wl_1)z_1}{\exp(-wl_1)z_1 + \exp(-wl_2)z_2 + z_3}.$$

This is what we must obtain updating z_1 on the basis of the loss values l_1, l_2 and l_3 with $l_3 = 0$. Similarly, we can see that updating z_2 we must obtain

$$\frac{\exp(-wl_2)z_2}{\exp(-wl_1)z_1 + \exp(-wl_2)z_2 + z_3}.$$

In this way we have shown how to extend our coherent updating rule from the two-points case to the three-points case.

References

Ali, S. M. and Silvey, S. D. (1966) A general class of coefficients of divergence of one distribution from another. *J. R. Statist. Soc.* B, **28**, 131–142.

Alquier, P. (2008) PAC-Bayesian bounds for randomized empirical risk minimizers. *Math. Meth. Statist.*, **17**, 279–304.
Bar-Hillel, M. and Falk, R. (1982) Some teasers concerning conditional probabilities. *Cognition*, **11**, 109–122.
Berger, J. O. (1993) *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.
Bernardo, J. M. (1979) Expected information as expected utility. *Ann. Statist.*, **7**, 686–690.
Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.
Bissiri, P. G. and Walker, S. G. (2010) On Bayesian learning from Bernoulli observations. *J. Statist. Planng Inf.*, **140**, 3520–3530.
Bissiri, P. G. and Walker, S. G. (2012a) On Bayesian learning from loss functions. *J. Statist. Planng Inf.*, **142**, 3167–3173.
Bissiri, P. G. and Walker, S. G. (2012b) Converting information into probability measures via the Kullback–Leibler divergence. *Ann. Inst. Statist. Math.*, **64**, 1139–1160.
Catoni, O. (2003) A PAC Bayesian approach to adaptive classification. *Preprint 840.* Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, Paris.
Cesa-Bianchi, N. and Lugosi, G. (2006) *Prediction, Learning, and Games*. Cambridge: Cambridge University Press.
Cheng, Y. and Church, G. M. (2000) Biclustering of expression data. In *Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology* (eds P. Bourne, M. Gribskov, R. Altman, N. Jensen, D. Hope, T. Lengauer, J. Mitchell, E. Scheeff, C. Smith, S. Strande and H. Weissig), pp. 93–103. Menlo Park: American Association for Artificial Intelligence Press.
Cooley, D., Ribatet, M. and Davison, A. C. (2009) Bayesian inference from composite likelihoods, with an application to spatial extremes. *Preprint arXiv:0911.5357*. Department of Statistics, University of Colorado, Boulder.
Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.
Datta, G. S. and Sweeting, T. J. (2005) Probability matching priors. In *Handbook of Statistics* (eds D. Dey and C. R. Rao), vol. 25, pp. 91–114. Amsterdam: Elsevier.
Diaconis, P. and Zabell, S. L. (1982) Updating subjective probability. *J. Am. Statist. Ass.*, **77**, 822–830.
Doucet, A. and Shephard, N. (2012) Robust inference on parameters via particle filters and sandwich covariance matrices. *Report 606*. Department of Economics, University of Oxford, Oxford.
Fan, J and Li, R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, **30**, 74–99.
Faraggi, D. and Simon, R. (1998) Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475–1485.
de Finetti, B. (1937) La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1–68.
Freund, J. E. (1965) Puzzle or paradox? *Ann. Statist.*, **19**, 29–44.
Gardner, M. (1959) *The Scientific American Book of Mathematical Puzzles and Diversions*. New York: Simon and Schuster.
Goldstein, M. (1981) Revising previsions: a geometric interpretation. *J. R. Statist. Soc.* B, **43**, 105–130.
Goldstein, M. and Wooff, D. (2007) *Bayes Linear Statistics, Theory & Methods*. Chichester: Wiley.
Hartigan, J. A. (1972) Direct clustering of a data matrix. *J. Am. Statist. Ass.*, **67**, 123–129.
Hastie, T., Tibshirani, R. and Friedman, J. (2009) *Elements of Statistical Learning*. New York: Springer.
Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J. and Dimopoulos, G. (2005) Bayesian coclustering of Anopheles gene expression time series: study of immune defence response to multiple experimental challenges. *Proc. Natn. Acad. Sci. USA*, **102**, 16939–16944.
Hirshleifer, J. and Riley, J. G. (1992) *The Analytics of Uncertainty and Information*. Cambridge: Cambridge University Press.
Hoff, P. and Wakefield, J. C. (2013) Bayesian sandwich posteriors for pseudo-true parameters. *J. Statist. Planng Inf.*, **143**, 1638–1642.
Hüber, P. (1964) Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.
Hüber, P. (2009) *Robust Statistics*, 2nd edn. Hoboken: Wiley.
Hutchison, K. (1999) What are conditional probabilities conditional upon? *Br. J. Philos. Sci.*, **50**, 665–695.
Hutchison, K. (2008) Resolving some puzzles of conditional probability. *Adv. Sci. Lett.*, **1**, 212–221.
Ibrahim, J. G. and Chen, M. H. (2000) Power prior distributions for regression models. *Statist. Sci.*, **15**, 46–60.
Ibrahim, J. G., Chen, M. H. and MacEachern, S. N. (1999) Bayesian variable selection for proportional hazards models. *Can. J. Statist.*, **27**, 701–711.
Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001) *Bayesian Survival Analysis*. New York: Springer.
Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, **20**, 50–67.
Jiang, W. and Tanner, M. A. (2008) Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.*, **36**, 2207–2231.
Kass, R. E. and Wasserman, L. A. (1996) The selection of prior distributions by formal rules. *J. Am. Statist. Ass.*, **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* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 343–370. Oxford: Oxford University Press.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.
Langford, J. (2005) Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.*, **6**, 273–306.
Marin, J. M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statist. Comput.*, **22**, 1167–1180.
McAllester, D. (1998) Some PAC Bayes theorems. In *Proc. 11th A. Conf. Computational Learning Theory*, pp. 164–170. New York: Association for Computing Machinery.
Merhav, N. and Feder, M. (1998) Universal prediction. *IEEE Trans. Inform. Theor.*, **44**, 2124–2147.
Müller, U. (2012) Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. Department of Economics, Princeton University, Princeton.
von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.
Royall, R., and Tsou, T.-S. (2003) Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *J. R. Statist. Soc.* B, **65**, 391–404.
Savage, L. J. (1954) *The Foundations of Statistics*. New York: Wiley.
Seldin, Y. and Tishby, N. (2010) PAC Bayesian analysis of co-clustering and beyond. *J. Mach. Learn. Res.*, **11**, 3595–3646.
Shawe-Taylor, J. and Williamson, R. (1997) A PAC analysis of a Bayesian estimator. In *Proc. 10th A. Conf. Computational Learning Theory*, pp. 2–9. New York: Association for Computing Machinery.
Tanay, A., Sharan, R. and Shamir, R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, 136–144.
Tibshirani, R. J. (1997) The lasso method for variable selection in the Cox model. *Statist. Med.*, **16**, 385–395.
Volinsky, C. T., Madigan, D., Raftery, A. E. and Kronmal, R. A. (1997) Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Appl. Statist.*, **46**, 433–448.
Walker, S. and Hjort, N. L. (2001) On Bayesian consistency. *J. R. Statist. Soc.* B, **63**, 811–821.
Zellner, A. (1988) Optimal information processing and Bayes's theorem. *Am. Statistn*, **42**, 278–284.
Zhang, T. (2006a) From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Ann. Statist.*, **34**, 2180–2210.
Zhang, T. (2006b) Information theoretical upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theor.*, **52**, 1307–1321.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material'.