A General Framework for Updating Belief Distributions

P.G. Bissiri, C.C. Holmes & S.G. Walker *

Abstract

We propose a general framework for Bayesian inference that does not require the specification of a complete probability model, or likelihood, for the data. As data sets become larger and systems under investigation more complex it is increasingly challenging for Bayesian analysts to attempt to model the true data generating mechanism. Moreover, when the object of interest is a low dimensional statistic, such as a mean or median, it is cumbersome to have to achieve this via a complete model for the whole data distribution. If Bayesian analysis is to keep pace with modern applications it will need to forsake the notion that it is either possible or desirable to model the complete data distribution. Our proposed framework uses loss-functions to connect information in the data to statistics 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 given, yet provides coherent subjective inference in much more general settings. We demonstrate our approach in important application areas for which Bayesian inference is problematic including variable selection in survival analysis models and inference on a set of quantiles of a sampling distribution. Connections to other inference frameworks are highlighted.

Keywords: Bayesian updating; Decision theory; Generalized estimating equations; Information; Loss function; Maximum entropy; Self–information loss function.

*Pier Giovanni Bissiri is Research Associate, Universita degli Studi di Milano-Bicocca, Italy, (email: pier.bissiri@unimib.it); Chris Holmes is Professor of Statistics, Department of Statistics, University of Oxford, Oxford, U. K. (email: c.holmes@stats.ox.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). Holmes is supported by the Oxford-Man Institute, Oxford, and the Medical Research Council, UK.

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. So even if the task is inference for a low-dimensional statistic of the population, the analyst is required to model the complete data distribution and, moreover, assume that the model is "true". We propose a coherent procedure for general Bayesian inference that does not require complete knowledge of $f_0(x)$ and which connects information in the data to the value of an unknown object or parameter of interest via the use of loss functions. By "coherent" we mean that all relevant information is contained in the posterior probability distribution. For ease of exposition we shall use the terminology "parameter of interest" and "statistic of interest" interchangeably. We show how the approach leads to conventional Bayesian updating when the true likelihood is known but allows for rational updating of beliefs in much more general settings.

The central tenet of our paper is this: as applications get more complex Bayesian analysts will increasingly be forced to forsake the notion that they can precisely model all aspects of the data. Settling for a misspecified model undermines the traditional Bayesian approach leading to interpretability problems along with the reliability of the posterior distribution. If the analyst acknowledges this then they should seek an alternative coherent way to proceed. We aim to contribute to this task.

1.1 The idea. Let θ denote a parameter or statistic of interest, for example the mean or median of a population $F_0(x)$, and let x denote a set of observables x from $F_0(x)$, with F_0 unknown. We are interested in a formal, optimal, way to update prior beliefs $\pi(\theta)$ to posterior beliefs $\pi(\theta|x)$ given x.

Bayesian inference proceeds through knowledge of a complete, true, model for $f_0(x)$. This is often parameterised via a sampling distribution $f(x;\beta)$ and a prior $\pi(\beta)$, such that,

$$f_0(x) = \int_\beta f(x;\beta)\pi(d\beta)\,,$$

and following de Finetti we know that all exchangeable distributions can be modelled in such form, see for example Bernardo and Smith (1994). Then inference for the statistic of interest, θ, can occur via,

$$\pi(\theta|x) = \int_\beta g[f(\cdot;\beta)]\pi(d\beta|x)$$

where $g[\cdot]$ defines the statistic; for example, if θ is the mean then $g[f(\cdot;\beta)] = \int x f(x;\beta)\,\mathrm{d}x$; or if θ denotes the median then $g[f(\cdot;\beta)] = F_\beta^{-1}(0.5)$. 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, x may contain a vast number of data points and β might be high-dimensional. Taken together, this makes the Bayesian approach somewhat cumbersome.

We are interested in the rational updating of beliefs, $\pi(\theta) \rightarrow \pi(\theta|x)$, under more realistic and manageable assumptions. To do so we relax the assumption that $f_0(x)$ is known and 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 will 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 state space of θ. We shall construct a loss function to select an optimal posterior measure $\widehat{\nu}(\theta)$ given a prior $\pi(\theta)$ and data x. To achieve this we construct a loss-function $L(\nu; \pi, x)$ on the space of probability measures on θ, and then present

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

as the optimal "honest" 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 data x is 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

$$L(\nu; \pi, x) = h_1(\nu, x) + h_2(\nu, \pi), \tag{1}$$

where h_1 and h_2 are themselves loss functions 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.

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. We can deal immediately with the loss function $h_2(\nu, \pi)$. Somewhat remarkably, as proved later, for coherent inference h_2 must be the Kullback–Leibler divergence, Kullback and Leibler (1951), and given by

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

Regarding h_1, since $\widehat{\nu}(\theta)$ is a probability measure representing beliefs about θ, the only choice is to take the loss-to-data $h_1(\nu, x)$ as the *expected* loss (see von Neumann and Morgenstern, 1944) of $l(\theta, x)$; that is

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

with the particular types of the loss-function $l(\theta, x)$ to be discussed later. In general there the form of $l(\theta, x)$ will be problem specific as discussed in Section 3.

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

$$L(\nu; \pi, x) = \int_{\Theta} l(\theta, x)\, \nu(\mathrm{d}\theta) + d_{KL}(\nu, \pi). \tag{2}$$

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

$$\begin{aligned} \widehat{\nu}(\theta) &= \arg\min_{\nu} L(\nu; \pi, x) \\ &= \frac{\exp\{-l(\theta, x)\}\pi(\theta)}{\int_{\Theta} \exp\{-l(\theta, x)\}\pi(\mathrm{d}\theta)}. \end{aligned} \tag{3}$$

This has the form of a Bayesian update but where the complete log-likelihood, $\log f(x; \beta)$, is replaced by a loss function $l(x, \theta)$ targeting the parameter of interest. As is usual in decision problems involving the use of loss function, it is incumbent on the decision maker to ensure solutions exist. So $l(\theta, x)$ needs to be constructed so that $0 < \int_{\Theta} \exp\{-l(\theta, x)\}\pi(\mathrm{d}\theta) < +\infty$.

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

1.2 Connections with other work. There is a vast amount of literature on procedures for robustly estimating a parameter of interest by minimizing the cumulative loss

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

See, for example, Hüber (2009), where we note that the primary aim is not modeling the data but rather estimating a statistic. This is an advantage when a probability model for the data is too hard to formulate. We are presenting a Bayesian extension of this idea. Since we are interested in a belief distribution for θ given data, and have further information provided by π, we claim the appropriate Bayesian version is given by (2).

Some of the ideas presented in the paper have been considered in a less general setting by Zhang (2006a, 2006b) 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 (3), is described in Section IV of his paper. This is our procedure when data is regarded as stochastic. Zhang then concentrates on the properties of the Gibbs posterior.

Further theoretical work is done in Zhang (2006b). In Jiang and Tanner (2008) a Gibbs posterior is 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.

We build on the work of Zhang (2006a, 2006b) and Jiang and Tanner (2008) in a number of important directions. The first is that we develop an approach for inference and statistical applications rather than studying the theoretical properties of the posterior under misspecification. We provide a principled approach to scale the relative information in the data to information in the prior; that is left as an arbitrary free parameter in Zhang (2006a, 2006b) and Jiang and Tanner (2008). We show that in order to remain coherent, the modeller *must* adopt the Kullback-Leibler divergence as the loss between prior π and ν. Finally, we demonstrate how to incorporate non-stochastic information into the cumulative loss function, which provides a definition of a conditional probability in the presence of non-stochastic information.

Another similar construct to $L(\nu; \pi, x)$ is provided by Zellner (1988), who presents what is essentially a loss function for the posterior distribution 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's paper. 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.

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 considers 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) describe procedures for adjusting likelihood functions when the model is misspecified. More recently, Doucet and Shepherd (2012), and Muller (2012) consider 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). Ribatet et al (2009) consider pseudo-Bayesian approaches with composite likelihoods.

Several authors have considered issues with Bayesian updating with proxy models, $f(x; \theta)$, for example, Key et al. (1999), when (x_i) is known not to arise from $f(x; \theta)$ for any value of θ. That is, 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) describe one such idea using a cross–validation approach. While this may be a 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. That is, assume the observations are coming from $f(x;\theta)$ even though it is known 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; that is, 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) point out, how does one construct $\pi_j(\theta_j)$ and p_j when one knows none of the postulated models are correct. So the Bayesian update breaks down in that nothing has any interpretation.

A recent popular idea is to use Bayesian nonparametrics. See Ghosh and Ramamoorthi (2003), and Hjort et al. (2010) for reviews. The idea here is making the choice of modeling density $f(x)$ so large by constructing a prior directly on a space of density functions, and written as $\pi(\mathrm{d}f)$, which has such a large support that it is reasonable to assume $f_0(x)$ lies in the support. A well known model is the infinite mixture model, whereby $\pi(\mathrm{d}f)$ is generating random density functions of the type

$$f(x) = \int_{z\in Z} K(x|z)\,\mathrm{d}P(z),$$

where K is a density for each z, often the normal density and z denotes the mean and variance, and P is a random distribution function, usually of the type

$$P = \sum_{l=1}^{\infty} w_l\,\delta_{z_l},$$

and the prior is assigned to $(w_l, z_l)_{l=1}^{\infty}$. Here the (w_l) are weights and sum to unity. The Dirichlet process, Ferguson (1973), is widely used in such contexts; see Lo (1984) and Escobar (1988) for the origins of the model and sampling based algorithms for estimating the model. While this methodology has made rapid developments in recent years, including the development of sampling algorithms, for complex data structures there are still issues about just how large the supports are and indeed how complicated inference can be and how to construct priors which capture reasonable beliefs about dependencies in the data. Moreover this still requires the specification of complete beliefs on $f_0(x)$ even when the objective is inference for a summary statistic of the data distribution.

Finally, we note that it is informative to view the selection of $\widehat{\nu}$, i.e.

$$\widehat{\nu}(\theta) \quad = \quad \arg\min_{\nu} \{h_1(\nu;x) + h_2(\nu,\pi)\} \tag{4}$$

as trading off fidelity to the data and fidelity to the prior. This highlights connections with penalised likelihood and regularized regression, see for example Hastie et al (2009). But whereas in penalised likelihood the objective is to select a single parameter estimate $\widehat{\theta}$, the general Bayesian approach (4) selects a probability distribution $\widehat{\nu}(\theta)$.

The layout of the remainder of the paper is as follows. In Section 2 we discuss how (3) arises as the unique minimiser of expected loss. In Section 3 we discuss forms for the loss-to-data functions and calibration. Section 4 then considers general forms of data, such as partial information and non-stochastic information. Section 5 provides some numerical illustrations including inference based on the Cox proportional hazards model and inference about the median of a distribution function. Section 6 concludes with a discussion on a number of points.

2. Information in the prior. Here we discuss the choice of the Kullback–Leibler divergence as being appropriate for quantifying the loss-to-prior $h_2(\nu, \pi)$ in (1). With n independent pieces of information $x = (x_1, \ldots, x_n)$ we take the cumulative loss as

$$L(\nu; \pi, x) = \sum_{i=1}^{n} h_1(\nu, x_i) + h_2(\nu, \pi), \tag{5}$$

where h_1 will be taken in the integral form, i.e. the average or expected loss:

$$h_1(\nu, x_i) = \int_{\Theta} l(\theta, x_i)\ \nu(\mathrm{d}\theta).$$

Now, adhering to the "likelihood principle" (see Bernardo and Smith 1994), for any $0 < m < n$, all the information contained in $(x_1, \ldots, x_m)$ is to be found in $\widehat{\nu}_m$, where $\widehat{\nu}_m$ minimizes

$$L(\nu; \pi, x_1, \ldots, x_m) = \sum_{i=1}^{m} h_1(\nu, x_i) + h_2(\nu, \pi).$$

and hence it follows that,

$$L(\nu; \pi, x) = \sum_{i=m+1}^{n} h_1(\nu, x_i) + h_2(\nu, \widehat{\nu}_m),$$

where $\widehat{\nu}_m$ now serves as the prior for future information $(x_{m+1}, \ldots, x_n)$. For coherence, the solution from L for all cases of m must be the same. To derive the form of h_2 we start with the family of g–divergences, that is

$$h_2(\nu, \pi) = d_g(\nu, \pi) = \int g(\mathrm{d}\pi/\mathrm{d}\nu)\, \mathrm{d}\nu \tag{6}$$

where g is a convex function from $(0, \infty)$ to the real line and $g(1) = 0$. See Ali and Silvey (1966). For this coherence to be in force, it is necessary that the discrepancy h_2 is the Kullback-Leibler divergence. To be more precise, the following theorem can be stated:

Theorem. *Let the loss $L(\nu; \pi, [x_1, x_2])$ be defined by* (5) *and* (6)*. Moreover, let $\widehat{\nu}_{(\pi,x_1,x_2)}$ be the probability measure that minimizes the loss*

$$L(\nu; \pi, [x_1, x_2])$$

among the probability measures on Θ that are absolutely continuous with respect to π. Similarly, let $\widehat{\nu}_{(\pi,x_1)}$ and $\widehat{\nu}_{(\widehat{\nu}_{(\pi,x_1)},x_2)}$ be the probability measures minimizing the loss $L(\nu; \pi, x_1)$ and $L(\nu; \widehat{\nu}_{(\pi,x_1)}, x_2)$, respectively. Assume that

$$\widehat{\nu}_{(\widehat{\nu}_{(\pi,x_1)},x_2)} = \widehat{\nu}_{(\pi,[x_1,x_2])} \tag{7}$$

for every probability measure π on Θ and for every choice of the loss functions $h_1(\nu, x_1)$ and $h_2(\nu, x_2)$ such that $\widehat{\nu}_{(\pi,[x_1,x_2])}$, $\widehat{\nu}_{(\pi,x_1)}$, $\widehat{\nu}_{(\widehat{\nu}_{(x_1)},x_2)}$, are all properly defined. Then h_2 is the Kullback–Leibler divergence.

In virtue of this Theorem, which is proved in Appendix A, for coherence it is required to take

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

the Kullback–Leibler divergence. So, in the case of $m = 0$, we have

$$L(\nu; \pi, x) = \sum_{i=1}^{n} h_1(\nu, x_i) + d_{KL}(\nu, \pi),$$

where π is the initial choice of probability measure representing beliefs about θ in the absence of x.

The solution to this minimization problem is easy to find and is given by

$$\nu(\mathrm{d}\theta) = \frac{\exp\{-\sum_{i=1}^{n} l_i(\theta, x_i)\} \; \pi(\mathrm{d}\theta)}{\int_\Theta \exp\{-\sum_{i=1}^{n} l_i(\theta, x_i)\} \; \pi(\mathrm{d}\theta)},$$

and this is the solution since one can see that

$$\begin{aligned} \textstyle\int_\Theta l(\theta, x) \, \nu(\mathrm{d}\theta) \quad & + \textstyle\int_\Theta \nu(\mathrm{d}\theta) \; \log\{\nu(\theta)/\pi(\theta)\} \\ & = \textstyle\int_\Theta \nu(\mathrm{d}\theta) \; \log\{\nu(\theta)/[\exp(-l(\theta, x)) \, \pi(\theta)]\}. \end{aligned}$$

3. Information in the data. In this section we will consider the form of h_1 in (1) that connects information in the data to the value of the unknown

θ. We shall consider three broad situations, first when the analyst really believes they know the complete family of distributions from which (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, the so called M–open perspective. Finally, when the statistic θ does not index a complete sampling distribution or proxy model for x.

3.1 M–closed and self-information loss. When the analyst knows the family from which (x_i) arose then the Bayesian approach to learning is fully justified, well known and widely used as a statistical approach to inference; the book of Bernardo and Smith (1994) is comprehensive. Here we recall the essence of it: A parameter of a density function $f(x;\theta)$, $\theta \in \Theta$, is unknown and beliefs about it are encapsulated via a prior distribution $\pi(\theta)$. Once (conditionally) independent samples $(x_1, \ldots, x_n)$ are observed from the density function $f(x;\theta)$, the prior is updated to the posterior distribution via Bayes' Theorem; given by

$$\pi(\mathrm{d}\theta|x_1, \ldots, x_n) = \frac{l_n(\theta)\,\pi(\mathrm{d}\theta)}{\int_\Theta l_n(\theta)\,\pi(\mathrm{d}\theta)},$$

where $l_n(\theta) = \prod_{i=1}^n f(x_i;\theta)$ is the likelihood function. The posterior then represents revised beliefs taking into account both the prior distribution and the observations. Mathematically, it is an application of Bayes' Theorem via the standard definition of conditional probability.

So the Bayesian update works and is applicable in the case when the (x_i) come from the density $f(x;\theta)$ for some $\theta \in \Theta$. In Bernardo and Smith (1994) this is referred to as the M–closed view. 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 "honest" loss function in this case is the self–information, or logarithmic loss function, given by

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

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 to ensure that the analyst remains honest in expressing subjective beliefs, under which we recover the Bayesian updating rule. However, there are different ideas behind our derivation of this rule, with different assumptions being made. Most crucially, we need the (x_i) to provide independent pieces of information to maintain the credibility of the cumulative loss function.

3.2 M–open and the use of proxy models. As has been mentioned by many authors, for example, Key et al. (1999), issues with the Bayesian rule arise

when $f(x;\theta)$ is known not to be the family of densities from which the (x_i) come. Equivalently, 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). In many situations, the correct sampling density, $f_0(x)$, is unknown or unavailable or too complex to work with. There are a number of ways to attempt to resolve this issue from a Bayesian perspective.

One idea is to use methods based on approximations and Key et al. (1999) describe one such idea using a cross–validation approach. While this may be a suitable idea which can work in practice 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. That is, assume the observations are coming from $f(x;\theta)$ even though it is known 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; that is, 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) point out, how does one construct $\pi_j(\theta_j)$ and p_j when one knows none of the postulated models are correct. So the Bayesian update breaks down in that nothing has any interpretation.

We show in Appendix B that it is possible to learn about this θ_0 since an infinite collection of (x_i) yields $f_0(\cdot)$ via the empirical distribution function and so θ_0 will be found with sufficient samples. Then we would wish the sequence of $\nu(\mathrm{d}\theta)$ to accumulate about θ_0. So what is the appropriate loss $l(\theta,x)$ in the case where we're trying to learn about the value of θ_0? The loss function $l(\theta,x) = -\log f(x;\theta)$ is still the right choice. For the standardized cumulative loss based on a sequence of observations is given by

$$-n^{-1}\log f(x_i;\theta) \to \int -\log f(x;\theta)\,\mathrm{d}F_0(x) \quad \text{a.s.}$$

which is minimized by θ_0.

When an approximate model $f(x;\theta)$ has been supposed, it is often prudent to consider a number of models, say $f_j(x;\theta_j)$ for $j = 1,\ldots,M$, as we have mentioned previously. We can deal with this in a simple way. So let $\theta = (\theta_1,\ldots,\theta_M)$ and let $\pi(\theta)$ be the prior distribution for θ on $\Theta = \cup_{j=1}^{M}\Theta_j$. This would be constructed by considering beliefs about which θ_j from $f_j(\cdot;\theta_j)$ takes this family closest to $f_0(\cdot)$. The model $f(x;\theta)$ would then be given by

$$f(x;\theta) = \sum_{j=1}^{M} p_j\, f_j(x;\theta_j)$$

and the (p_j) would now be the probabilities describing beliefs about which model provides the closest density to $f_0(\cdot)$. Hence, unlike the Bayesian ap-

proach to model selection in the M–open case, all the quantities to be specified have clear interpretation. We can recover the Bayesian update when we take, for each $i \in (1, \ldots, n)$,

$$l(\theta, x_i) = -\log f(x_i; \theta).$$

So while the Bayesian approach has some issues to deal with whether the M–open or M–closed view hold, for us it is irrelevant. If one adopts θ_0 as the parameter value taking the family closest to $f_0(\cdot)$ then one does not need to worry if in M–open or M–closed, 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 it may be that one simply does not know if one is in the M–open or M–closed view (though strictly speaking this puts you in M–open) and then one needs a framework in which the same approach is adopted and justified regardless of which view is held. We have provided such a framework.

3.3 M–free. Often the analyst might not wish to express a full probability model for the data, either as it's too cumbersome or too problematic. A motivating example is inference for the median of a population of iid data. However, the analyst knows the object or statistic θ that they wish to express beliefs about. It is incumbent on them to choose a specification for $l(\theta, x)$ that provides greatest information on the unknown value. The literature on this is in the area of *Robust Statistics* and loss functions can be found in the literature pertaining to M-estimation and estimating equations. See, for example, Hüber (2009). We refer to this setting as M–free to highlight the model free aspect of inference.

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

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

with solution provided by

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

For example, one possible application would be the Generalized Estimating Equations, see Liang and Zeger (1986). For the grouped observations $(x_{i1}, \ldots, x_{in_i})$,

$$\rho(x_i, \theta) = \tfrac{1}{2}(x_i - \mu_i(\beta))' V_i(\phi, \alpha)^{-1}(x_i - \mu_i(\beta))$$

where $\theta = (\beta, \phi, \alpha)$ and for some link function g, $g(\mu_{ij}(\beta)) = x'_{ij}\beta$, and for some correlation matrix $R_i(\alpha)$ and diagonal matrix A_i, with j entry given by $a(\mu_{ij}(\beta))$, with a a specified variance function, $V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$, with ϕ a scale parameter. There is by now an abundance of literature on M-estimation, estimating equations and generalized estimating equations. Our point is that all such equations can be viewed as loss functions connecting independent units with parameters of interest. Hence, all fit within our framework and we would extend the loss function to include the prior π and we obtain an explicit expression for $\nu(\mathrm{d}\theta)$. In cases when the parameter estimation is done via iterative methods, which is typically the case, Markov chain Monte Carlo methods would substitute for our sampling strategies from $\nu(\mathrm{d}\theta)$.

In essence, this is the practical innovations of the framework we are proposing. We are claiming that any loss function of the type

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

can be extended to the Bayesian type updating mechanism. 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 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 $\int \rho(x;\theta)\,\mathrm{d}F_0(x)$.

3.4 M-free calibration of relative losses. In the M–closed and M–open settings the use of the self-information loss $l(\theta, x) = -\log f(x;\theta)$ results in a fully specified form for (3). However in the M–free setting 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 $l(\nu, x)$ relative to $l(\nu, \pi)$, in that we are free to multiply either by an arbitrary factor. So equivalently we are interested in the 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 the classical literature on estimation using such loss functions since there is no combining with different styles of loss functions. However the calibration of different types of loss function is

not a unique problem. It arises in many applied contexts; possibly the most well known be in health economics where losses pertaining to costs need to be balanced against losses pertaining to health benefits. There are a number of ideas for the choice of w and we discuss them here.

3.4.1 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 & 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. On the other hand, 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; that is $\nu(\mathrm{d}\theta) \approx \delta_{\widehat{\theta}}(d\theta)$, where $\widehat{\theta}$ maximizes $\prod_{i=1}^{n} f(x_i;\theta)$.

Alternative ideas for setting w include a data dependent assignment based on cross-validation and a random assignment once the parameter has appeared in the Gibbs posterior. That is, one considers

$$\widehat{\nu}(\theta|x) = \int \widehat{\nu}(\theta|w,x)\, \pi_w(\mathrm{d}w)$$

for some probability measure $\pi_w(\mathrm{d}w)$.

3.4.2 Unit information loss. Here we discuss a subjective assignment but a more orientated and direct allocation. The subjective choice is based on a prior evaluation of the expected value of $l(\theta, x)$.

To aid in the calibration of the loss functions and the selection of w we can consider the following. Write the loss function with an additional term $\log \pi(\widehat{\theta})$, which is a constant, and where $\widehat{\theta}$ maximizes $\pi(\theta)$, so that the cumulative loss becomes

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

In order to calibrate the information in the data relative to the prior we now assume that both loss functions, $l(\theta, x)$ and $\log\{\pi(\widehat{\theta})/\pi(\theta)\}$ are non-negative, and we standardise $l(\theta, x)$ such that $\min_\theta l(\theta, x) = 0$ for any x. If this is not the case then we replace $l(\theta, x)$ by $l(\theta, x) - l(\theta_x, x)$ where now θ_x minimizes $l(\theta, x)$. Hence, we can regard

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

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

$$w\, l(\theta, x) \quad \text{and} \quad \log\{\pi(\widehat{\theta})/\pi(\theta)\}.$$

These are two loss functions for θ and to adhere with the notion that at the outset before we have data, there is a single piece of information, we can calibrate the two losses by making the expected losses to match. That is, whether someone takes a θ and is penalized by the loss $\log\{\pi(\widehat{\theta})/\pi(\theta)\}$, or takes a (θ, x) and is penalized by the loss $wl(\theta, x)$, 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 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

$$w\mathrm{E}\left(l(\theta, x)\right) = \mathrm{E}\left(\log\{\pi(\widehat{\theta})/\pi(\theta)\}\right).$$

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

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

One empirical choice is then given by

$$w = \frac{\int \log\{\pi(\widehat{\theta})/\pi(\theta)\}\, \pi(\mathrm{d}\theta)}{\int\int l(\theta, x)\, \pi(\mathrm{d}\theta)\, \mathrm{d}F_n(x)}.$$

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

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

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 $1/(2\sigma^2)$ 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. One further suggestion is

to replace the prior evaluation of the expected datum-loss with the observed unit information loss given x,

$$\int\int l(\theta, x)\, m(dx, d\theta) \approx \frac{1}{n-p} \sum_i l(\hat{\theta}_x, x_i) \tag{8}$$

where $\hat{\theta}_x = \arg\min_\theta \left[\sum_i l(\theta, x_i)\right]$ is the data-loss estimate of θ and p is the dimension of θ. For instance, in the above example this leads to,

$$w = \frac{1}{2\hat{\sigma}^2}$$

where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$.

3.4.3 Hierarchical loss. Another way to proceed is to extend the loss function to include w as an unknown parameter. Standard ideas here would suggest 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 hyper-prior 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

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

and given that w^ξ can be absorbed in to the prior π it is perfectly reasonable to assess ξ subjectively. That is, it seems unreasonable to accept that π can be chosen subjectively but that ξ can not.

3.4.4 Operational Characteristics. The idea here is to set w so that the posterior quantiles match up 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 $C_\alpha(w, x_1, ..., x_n)$ is the $100(1-\alpha)\%$ level confidence interval for θ, then we would select the w such that the posterior distribution of θ, with parameter w, is such that

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

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

4. General forms of information. In this Section we discuss more general forms of information to condition on, rather than a complete stochastic data sample x from unknown $F_0(x)$. In particular we provide a definition of conditional probability when non–stochastic information is available, and updating using partial-information in a data set.

4.1 Conditional probability distributions and non-stochastic data. The theory of conditional probability distributions is a well-established mathematical theory that provides a procedure to update 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; that is, events to which a probability is assigned. In other words, such information needs to be already included into the probability model. In this section, we shall show how the 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 if an expert declares "θ is close to 0". This type of information has been discussed by a number of authors 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 the paper by Diaconis and Zabell (1982) and in particular refer the reader to example in Section 1.1 of their paper.

If we denote by I a piece of information for which no probability model for each θ is possible. In other words it is not and can not be determined to be stochastic in any way. However, a loss function $l(\theta, I)$ can be assigned. Our theory does not preclude such a loss function based on such a piece of information. The answer $\widehat{\nu}(\theta)$ based on I and π can then be considered as a means of defining a conditional probability distribution in the presence of non-stochastic information. This section develops this argument.

Before proceeding, we introduce the notation for this section, being different to put the discussion in a more broader context than simply a Bayesian statistical style updating. Let Y be a random variable on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, which will be the outcome of interest, and valued into a measurable space $(\mathbb{Y}, \mathscr{Y})$ with probability distribution P_Y. Hence, P_Y represents initial belief about the outcome concerning Y. Now, assume that that the outcome of another random variable, say X, is known. So, let X be a random variable from $(\Omega, \mathscr{F}, \mathbb{P})$ into $(\mathbb{X}, \mathscr{X})$ with probability distribution P_X and the additional information I about Y will be assumed to be an outcome of X. Then it is possible to update the unconditional distribution of Y to the

probability distribution of Y given X.

In probability theory, a conditional distribution of Y given X is a map p from $\mathscr{Y} \times \mathbb{X}$ into $\mathbb{R}$ such that:

- for each x in $\mathbb{X}$, $p(\cdot, x)$ is a probability measure on $\mathscr{Y}$,
- for each B in $\mathscr{Y}$, $p(B, X(\omega))$ is a version of the conditional probability $\mathbb{P}(Y \in B \mid X(\omega))$, i.e. for each A in $\mathscr{X}$ and each B in $\mathscr{Y}$,

$$\mathbb{P}\{X \in A,\, Y \in B\} = \int_A p(B, x)\, \mathrm{d}P_X(x), \tag{9}$$

  where P_X denotes the probability distribution of X.

The conditional distribution is known to be essentially unique, i.e. unique only up to almost sure equality. This is a consequence of X being stochastic. In fact, as (Feller, 1971, p. 160) points out, if, for instance, the distribution of X is concentrated on a subset $\mathbb{X}_0$ of $\mathbb{X}$, no natural definition of $p(B, x)$ is possible for x outside $\mathbb{X}_0$. Nevertheless, in individual cases, there usually exists a natural choice dictated by regularity requirements.

Moreover, it is well known that conditional distributions do not always exist unless some conditions are satisfied by the spaces $(\mathbb{X}, \mathscr{X})$ and $(\mathbb{Y}, \mathscr{Y})$. For more information about conditional probability distributions, see, for instance, Feller (1971) or Billingsley (1995).

Here, we will consider the case in which there are two σ-finite measures μ_1 and μ_2 on $\mathscr{F}$ such that the probability distribution of $(X,\, Y)$ is absolutely continuous with respect to $\mu_1 \times \mu_2$. Denote its density by f. This is a general framework which includes most applications and enables us to find easily an expression for the conditional distributions. Generally, $\mathbb{X}$ and $\mathbb{Y}$ are subsets of $\mathbb{R}^k$, for some k, and μ_1 and μ_2 are the corresponding Lebesgue measures.

If f is the density of the probability distribution of $(X,\, Y)$ with respect to $\mu_1 \times \mu_2$, then one can take

$$p(B, x) = \frac{\int_B f(x, y)\ \mu_2(\mathrm{d}y)}{\int_{\mathbb{Y}} f(x, y)\ \mu_2(\mathrm{d}y)}, \tag{10}$$

for every B in $\mathbb{Y}$ and every x in $\mathbb{X}$ such that

$$0 \,<\, f_X(x) := \int_{\mathbb{Y}} f(x, y)\ \mu_2(\mathrm{d}y) \,<\, \infty. \tag{11}$$

Note that $p(\cdot, x)$ is absolutely continuous w.r.t. μ_2 and its density is

$$f_{Y|X}(y|x) := f(x, y)/f_X(x), \tag{12}$$

for every x in $\mathbb{X}$ satisfying (11). The density (12), which is called the conditional density of Y given X, is what is used in most application to find an expression for the conditional distribution. Therefore, (10) deserves to

be considered as the "practical definition" of conditional probability distribution. Indeed, it is the natural version of the conditional distribution of Y given X whenever a joint density f exists for X and Y.

Clearly, this approach relies on the joint distribution of X and Y and therefore is not available when X is replaced by some non–stochastic information I. Moreover, even if I coincides with an outcome of the random variable X, to define the conditional distribution of Y given X, it is required to know all the possible alternatives of I, that is, all the outcomes of X. It is also required to assess the joint distribution of X and Y or the conditional distribution of X given Y. This is quite easy if, for instance, I is known to be an outcome of some well-defined random experiment. In many situations, one has seen the outcome X and in order to establish an update of the distribution of Y, one needs to retrospectively ponder and imagine a joint probability model.

This difficulty arises in different puzzles such as, for instance, Freund's puzzle of the two aces, introduced by Freund (1965). For other puzzles about conditional probabilities, see, for instance, Gardener (1959).

These puzzles have been widely used to discuss the concept of conditional probability. Hutchison (1999, 2008) emphasizes that the updating process needs to take into account the circumstances under which the truth of I was conveyed. Also, Bar–Hillel and Falk (1982) claim that to know how the knowledge was obtained is "a crucial ingredient to select the appropriate model". These scholars present different views about the concept of conditionalization, but all agree on the fact that there would not be a problem if it was known how the information I became available, and therefore one could build a model including I.

The concept of conditional probability distributions is certainly appropriate as a procedure to update probabilities on the basis of any new information that was already included in the probability model. But it can be difficult to construct a model that considers all possible relevant information that in the future could become available. Therefore, the problem arises when one obtains some new and possibly unexpected information and wants to use it to update a probability distribution. Indeed, it does not seem appropriate to assess the probability of something which has been already observed.

First, we shall now show that if instead I is the outcome of a random variable X and there is a joint density f for (X, Y), then one can recover as particular case the conditional distribution of Y given X.

If there is a joint density f for (X, Y), then the conditional distribution (10) of Y given X arises as the solution of a decision theoretic problem related to a loss function of the form (2). For every x in $\mathbb{X}$ satisfying (11),

such a loss function is:

$$-\int_S \log(f(x,y)/f_Y(y))\,\mu_2(\mathrm{d}y) \;+\; d_{KL}(\nu, P_Y), \tag{13}$$

where

$$f_Y(y) := \int_{\mathbb{X}} f(x,y)\,\mu_1(\mathrm{d}x),$$

S is the set of all y in S such that $0 < f_Y(y) < \infty$, P_Y is the probability distribution of Y, ν is a probability measure on $\mathbb{Y}$ absolutely continuous w.r.t. P_Y. The loss (13) is of the form (2) with

$$\begin{aligned} l(\theta, I) = h(y,x) : &= -\mathbb{I}_S(y)\log(f(x,y)/f_Y(y)) \\ &= -\mathbb{I}_S(y)\log f_{X|Y}(x|y), \end{aligned} \tag{14}$$

where $\mathbb{I}_S(y)$ is equal to 1 or 0 depending on whether y belongs to S or not. For every x in $\mathbb{X}$ satisfying (11), the conditional distribution $p(\cdot, x)$ given by (10) minimizes the loss (13).

If the random variable X is replaced by some non–stochastic information I, then the self–information loss (14) cannot be defined, but one can still resort to a loss function of the form (2), by choosing a different loss $l(\theta, I)$. So, the approach introduced in Section 2 provides a general definition of conditional distributions based on non-stochastic information.

4.2 Partial information. We now consider a partial information problem. Here the parameter of interest is θ yet the information I collected is more informative; it is possible to identify $I_\Theta \subset I$ which provides all the information about θ. One is therefore interested in constructing the loss function $l(\theta, I_\Theta)$. A particular example to be looked at in detail is the proportional hazards model. If the model is that the hazard function is $g_0(t)\,\exp(g(\theta, z_i))$ for individual i, where z_i is the covariate value for individual i, g_0 is the baseline hazard function and $g(\theta, z)$ is the regression function, then the information about θ is provided by individual i failing from the set of possible failures $S_i = \{j : t_j \geq t_i\}$, where t_i is the time of failure of individual i, and these are assumed to be different for each individual. The assumption for us to use the partial information is that information of the failure times provide information about θ only through the sets $\{S_i\}$. Hence, there are $k \leq n$ pieces of information, where k is the number of individuals whose failure time is known, and it is usual to denote this by setting $\delta_i = 1$.

Using the partial self–information or logarithmic loss function, we have

$$\begin{aligned} l(\theta, I_\Theta) \;&= -\textstyle\sum_{\delta_i=1} \log \mathrm{P}(i|S_i, z) \\ &= -\textstyle\sum_{\delta_i=1} \left\{ h(\theta, z_i) - \log\left(\sum_{j\in S_i} \exp(h(\theta, z_j))\right)\right\} \end{aligned}$$

and so the solution to the decision problem is given by

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

This is a new approach and not taken on by Bayesians due to the lack of motivation. In Appendix C we consider other stylized models.

5. Illustrations. In this section we discuss the application of our approach to two important inferential problems. The first is an analysis of variation in survival times of colon cancer patients incorporating genetic information as potential predictors. The second is for joint inference on a set of quantiles. In both 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 we are implementing. Yet the updates we employ do allow us to learn about the specified parameters of interest. All of the models used to generate results are available as open source code in R.

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 eitiology and underlying disease heterogeneity. To this aim collaborators at the Wellcome Trust Centre for Human Genetics, University of Oxford, obtained survival times on 918 cancer patients with germline genotype data at 100,000's of markers genome-wide. For demonstration purposes we only consider one chromosome's worth of data containing 15,608 genotype measurements. The data table X then has $n = 918$ rows and $p = 15,608$ columns, where $(X)_{ij} \in \{0, 1, 2\}$ denotes the genotype of the i'th individual at the j'th 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 hazard (PH) model, 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 into the partial-likelihood for estimating the parameters of interest $\boldsymbol{\beta}$.

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

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

where R_i denotes the risk set, those individuals not censored or at time t_i, and in this way obtain a conditional distribution $\pi(\boldsymbol{\beta}|\mathbf{x})$.

5.1.1 Single marker association. As is standard practice, e.g. Balding (2006), we initially investigate the evidence of genetic association by testing each of the 15,608 markers in turn using a univariate model with loss,

$$l(\beta_j, \mathbf{x_j}) = \sum_{i=1}^{n} \log \left(\frac{\exp(x_{ij}\beta_j)}{\sum_{l \in R_i} \exp(x_{lj}\beta_j)} \right),$$

for each of the $j = 1, \ldots 15,608$ genetic makers. An advantage of our approach is the incorporation of prior information into the analysis. In most modern genome-wide genetic association studies we expect *a priori* that the coefficient values of predictive markers will be small, as otherwise we would have detected association of the marker using historic linkage based methods with lower resolution but higher power. Hence, we have additional information on the coefficient values. For unknown markers truly associated with survival we assume,

$$\beta_j \sim N(0, v_j)$$

and set $v_j = 0.5$ for our study, reflecting beliefs that associated coefficients will be modest. For each marker we now include an indicator variable, $\gamma_j \in 0, 1$ that specifies whether there is any association at the corresponding marker or not. This defines a hierarchical prior with,

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

and our prior $\pi(\delta_j)$ reflects beliefs about whether the corresponding β_j will be zero or not. For now we shall simply assume $\pi(\delta_j = 1) = 0.5$, although we note it is straightforward to incorporate genetic prior information here.

In this way we can use our framework to calculate a posterior measure $\pi(\delta_j, \beta_j|\mathbf{x}, \mathbf{y})$ for each marker. Interest lies in the evidence for a non-zero effect, i.e., in the marginal,

$$\pi(\delta_j|\mathbf{x}, \mathbf{y}) = \int_{\beta_j} \pi(\beta_j, \delta_j|\mathbf{x}, \mathbf{y}) \mathrm{d}\beta_j.$$

In particular we can define the general Bayes Factor of association at the j th marker as,

$$BF_j = \frac{\int_{\beta_j} \exp\left[-l(\beta_j|\mathbf{x}_j)\right] \pi(\beta_j|\delta_j = 1)\mathrm{d}\beta_j}{\exp\left[-l(\beta_j = 0|\mathbf{x}_j)\right]}$$

The one-dimensional integral in the numerator is simple to evaluate using quadrature or Monte Carlo methods. However, with a large sample size and over 15,000 integrals to calculate it is convenient to adopt a Laplace approximation to the integral, namely,

$$\int_{\beta_j} \exp\left[-l(\beta_j|\mathbf{x}_j)\right] \pi(\beta_j|\delta_j = 1)\mathrm{d}\beta_j \approx |\hat{\Sigma}_j|^{1/2} \exp[-l(\tilde{\beta}_j|\mathbf{x})]\pi(\tilde{\beta}_j|\delta_j = 1)$$

where $\tilde{\beta}_j$ is the MAP estimator, mode of the posterior $\pi(\beta_j|\delta_j, \mathbf{x}, \mathbf{y})$, and $\hat{\Sigma}_j$ is an estimate of the Hessian at the mode. Both the MAP estimate and the Hessian can be calculated efficiently under our loss and normal prior for β_j. We calculated the general Bayes Factors for each marker and in Fig (1) we plot the log Bayes Factors over the chromosome. While there is considerable variation we observe strong evidence of association around marker 10,000. To test if the Laplace approximation is accurate we selected 500 markers at random and ran a Monte Carlo importance sampler with $N(\tilde{\beta}_j, \tilde{\Sigma}_j^{-1})$, and 500 samples. Fig (2) indicates that the Laplace approximation appears accurate. This is not so surprising given we have 918 observations and a single parameter.

It is interesting to compare the evidence of association provided by the Bayes Factor Fig (1) in comparison to that obtained using a conventional Cox PH partial-likelihood based test. In Fig (3) we plot the log Bayes Factors versus $-\log_{10}$ p-values obtained from a likelihood ratio test. We can see general agreement especially at the markers with strongest association as one would expect for a large sample size. Interestingly there appears to be greater dispersion at markers of weaker association. In Fig (4) we highlight the region of weaker association and colour the points by the standard error of the maximum likelihood estimate. We can see a tendency for markers with less information, greater standard error, to get attenuated towards a logBF of 0 under the general Bayesian approach. This is further highlighted in Fig (5) where we plot the standard error against log 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 10,000, in Fig (6) we see the portion of the graph from Fig (1) containing 800 makers around the marker of strongest association. Due to high colinearity 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. In order to investigate this we developed multiple marker methods.

R code to calculate Bayes Factors for single marker association using Laplace and Monte Carlo Importance Sampling is available.

5.1.2 Multiple marker variable selection. With the aim of determining if there are multiple markers underlying the signal of association in Fig (6) we consider a model using potentially all 800 makers 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 AIC or BIC, that adjusts fit to the data by the number of covariates in the model. Inference proceeds using an optimization algorithm, such as forward or stepwise selection, to find a model that minimises the cost. More recently, penalized-likelihood methods have proved popular (Tibshirani, 1997; Fan and Li, 2002) where the partial-likelihood is maximised subject to some constraint on the norm of the regression coefficients defined by some appropriate sparsity inducing metric.

Despite the enormous impact of Cox PH models and the importance of variable selection, the Bayesian literature in this area is very limited. This is because of the lack of a theoretical foundation to treat $h_0(t)$ as a nuisance parameter, leading to either ad hoc methods or the full specification of a joint probability model. For instance, Faraggi and Simon (1998) and Volinsky et al. (1997) adopt pseudo-Bayesian approaches. The paper of Volinsky et al. (1997) take the BIC as an approximation to the marginal likelihood and they use 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, while the methods are important and well motivated, they are ultimately ad hoc. Moreover, prior information on $\pi(\boldsymbol{\beta})$ does not enter into the calculation of the BIC, meaning that an important aspect of the Bayesian approach is lost.

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

Here we use the partial-information relevant to the regression coefficients $\boldsymbol{\beta}$ via the cumulative loss function,

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

where R_i denotes the risk set, those individuals not censored or at time t_i. As

in Section 6.1.1 we assume proper priors, $\pi(\boldsymbol{\beta})$ on the regression coefficient,

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

where $\delta_j \in \{0, 1\}$ is an indicator variable on covariate relevance with,

$$\pi(\delta_j) = \mathrm{Bn}(a_j)$$

where $\mathrm{Bn}(\cdot)$ denotes the Bernoulli distribution but 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) 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 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,

$$q(\boldsymbol{\beta}'|\boldsymbol{\delta}') = \mathrm{N}(\tilde{\boldsymbol{\beta}}_{\delta'}, \tilde{\boldsymbol{V}}_{\delta'})$$

where $\{\tilde{\boldsymbol{\beta}}_{\delta'}, \tilde{\boldsymbol{V}}_{\delta'}\}$ are the MAP and approximate Information Matrix obtained from the combination of log-partial-loss and normal prior. The joint proposal is then accepted with probability,

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

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 (7) 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 suggest overwhelming evidence for a single marker in the region of index 10200 but also weaker evidence of independent signal in a couple of other regions. R code to perform the reversible jump MCMC multiple variable sampling for the Cox PH partial-likelihood with normal priors is available on request.

5.2 Joint inference for quantiles and the Bayesian Boxplot. We discuss this illustration for three reasons. The first is that there is a unique loss function for learning about a set of quantiles, countering the notion that loss functions are arbitrary, and second there is no traditional Bayesian version for

updating a set of quantiles which can coincide with our approach. Finally, we show how boxplots, one of the most widely used exploratory graphical tool, can be enhanced by taking into account uncertainty in the plot due to a finite sample size.

Let us start with the median solely. The unique loss function for learning about the median of a distribution function is given by $l(\theta, x) = w|\theta - x|$ for some $w > 0$. Hence, the posterior distribution is given by

$$\pi(\theta|x_1, \ldots, x_n) \propto \exp\left\{-w\sum_{i=1}^{n}|x_i - \theta|\right\}\pi(\theta).$$

One might be tempted to argue that this is merely a Bayesian update using the Laplace distribution and hence falls within the Bayesian paradigm. This is correct but it would put the Bayesian in an awkward quandary if she knew, for example, the observations were coming from a normal distribution.

In fact we are, as we have stated previously, not assigning a probability model for x. To make this distinction more explicit let us consider the situation where we want to learn about the three quartiles $(\theta_1, \theta_2, \theta_3)$ jointly, where θ_1 is the lower quartile, θ_2 the median, and θ_3 the upper quartile. The prior will be denoted by $\pi(\theta_1, \theta_2, \theta_3)$ which would obviously include the constraint $\theta_1 < \theta_2 < \theta_3$. The loss function $l(\theta, x)$ in this case, treating the learning of the quartiles with equal importance, is given by

$$\begin{aligned} l(\theta, x) \quad &= w\left\{0.25(\theta_1 - x)_+ + 0.75(x - \theta_1)_+ + \right. \\ &\left. +0.5|\theta_2 - x| + 0.75(\theta_3 - x)_+ + 0.25(x - \theta_3)_+\right\} \end{aligned}$$

for some $w > 0$. Then the posterior distribution is given by

$$\pi(\theta|x_1, \ldots, x_n) \propto \pi(\theta)\exp\left\{\sum_{i=1}^{n} l(\theta, x_i)\right\}.$$

This can not be obtained by any Bayesian model that has currently been proposed. It is certainly therefore not classifiable as a Bayesian update.

We can illustrate the utility of this by considering a boxplot. In Fig (8) we show a boxplot of data taken from the example used in MATLAB help file for the function `boxplot.m`, in the statistics toolbox. The plot illustrates the distribution of miles per gallon (MPG) from records of a selection of cars taken in the 1970s, broken down by manufacturing country. The data set is available as `carbig.mat` in MATLAB, we have omitted the 'England' group which contains only 1 observation.

The boxplot is one of the most important and widely used graphical tool applied to summarise the distribution of data and highlight potential differences in the distributions across groups, but there is traditionally no uncertainty displayed in the summary statistics of the distributions used in the

boxplot. In fact, for this data there are only 13 observations for "French" cars while there are 249 observations for the "USA", yet the conventional boxplot fails to inform on this.

We placed a prior on the median, upper and lower quartiles defined by the blue boxes in Fig (8) and account for the uncertainty by inferring the posterior distribution on these unknowns. Let θ_1 denote the lower quartile, θ_2 the median and θ_3 the upper quartile. We adopted a normal, fairly vague, prior,

$$\theta_1 \sim N(10, 100); \quad \theta_2 \sim N(20, 100); \quad \theta_3 \sim N(30, 100),$$

with the constraint $\theta_1 < \theta_2 < \theta_3$. We adopt the "observed unit information loss" in the setting of w, see Section 3.4.2,

$$\hat{w} = \frac{\int \log\{\pi(\widehat{\theta})/\pi(\theta)\}\, \pi(\mathrm{d}\theta)}{\frac{1}{n-p}\sum_i l(\hat{\theta}_x, x_i)}$$

where we estimate $\int \log\{\pi(\widehat{\theta})/\pi(\theta)\}\, \pi(\mathrm{d}\theta)$ via Monte Carlo and use a Nelder-Mead optimiser for $\hat{\theta}_x$.

We then implemented a Metropolis-Hastings MCMC algorithm to sample from the posterior $\pi(\theta_1, \theta_2, \theta_3|x)$, for each of the 6 groups of cars shown in Fig (8), using 100,000 samples with a 50,000 sample burn-in.

In Fig (9) we show our "Bayesian boxplot" which includes the original boxes (empirical estimates) overlaid with 95% credible intervals for $(\theta_1, \theta_2, \theta_3)$. Credible intervals are shown as extended dotted lines from the empirical estimates with a small diamond denoting the edge of the interval. In comparison with Fig (8) we see that Fig (9) contains much more information. For example, we see that while in Fig (8) the median MPG of Italian and Swedish cars look different, in fact the 95% credible intervals overlap in Fig (9). In addition we see that there is considerable overlap in the distribution of medians between Sweden and the USA; and in general, comparison of medians or distributions in the conventional boxplot are obscured and confounded by sample size.

The MCMC samples approximately from $\pi(\theta_1, \theta_2, \theta_3|x)$ for France and USA are shown in Figs (10), (11). The data set for France contains 13 observations and hence there is much greater uncertainty in the posterior marginals. Moreover, looking at the joint densities of (θ_1, θ_2) and (θ_2, θ_3) we can see the constraints imposed by the prior. In contrast, due to the higher sample size the posterior samples for the USA are tighter and hence exhibit less dependence. An interesting extension would be to include hierarchical priors on the quartiles whereby one could borrow strength across groups.

6. Discussion. We have provided a basis for general learning and the updating of information 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 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 currently in use.

The scope of our findings provides extensive generalizations to the Bayes updating rule. For the Bayesian, when it is problematic to construct a probability model with all the implications about assigning probability one to events, can be compared to the ease of introducing a loss function which has no further implications. A probability model needs to assert a sample space with alternatives and assign probabilities to all outcomes. On the other hand, a loss function can be constructed after the information has been received and determined solely for the known information without need to consider which alternative information could have been received. Yet, surprisingly, both approaches can coincide which suggests the Bayesian support theory is more than is really needed.

More generally, we can use loss functions currently employed in a classical context for robust estimation; for example, generalized estimating equations. We can also deal appropriately with partial information where it is only a part of some observed information 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.

6.1 Optimal Decisions. Let us now recap the story from a slightly different perspective when observations are independent and identically distributed from $F_0(x)$ and action $a \in A$ is to be made. The decision maker is happy to make an action if the minimizer, θ_0, of $\int l(\theta, x)\, dF_0(x)$ is known, for some loss function $l(\theta, x)$. This action is based on the utility function $u(a, \theta)$ and hence the action would be the one maximizing $u(a, \theta_0)$.

With θ_0 not being known, as F_0 is not known, a prior distribution $\pi(\theta)$

is constructed expressing beliefs about the location of θ_0. Then, with data $(x_i)_{i=1}^n$, the loss function picking out the appropriate probability measure $\nu(\theta)$, with which to provide an action a through the maximization of expected utility, i.e. $U(a) = \int_\Theta u(a,\theta)\,\nu(\mathrm{d}\theta)$, itself minimizes the loss function

$$L(\nu) = \sum_{i=1}^{n} \int_\Theta l(\theta, x_i)\,\nu(\mathrm{d}\theta) + d_{KL}(\nu, \pi).$$

In this way it is seen that the sequence of $\nu(\theta)$ should accumulate about θ_0.

To us, now, there seems to be no reason whatsoever why $l(\theta, x)$ should be exclusively based on a probability distribution. For example, if we want the median then $l(\theta, x) = |\theta - x|$; if we want the mean then $l(\theta, x) = (\theta - x)^2$; whereas if we want the θ taking us closest in Kullback–Leibler divergence to f_0, then $l(\theta, x) = -\log f(x;\theta)$.

6.2 Conclusion. We acknowledge we have presented a general framework which at first sight might appear to sanction "anything goes". This is wrong. We have replaced a subjective probability model with an objective loss function, since the parameter of interest is typically defined by the statistical problem. In this case, the loss function connecting the information to the parameter is unique. See, for example, Section 5.2, in the case of the parameter of interest being the median. On the other hand, there is no unique probability distribution to use to first model the data and then use this to estimate the median.

When the interest is in a parameter indexing a family of densities and the parameter to target is the one which makes this family closest to the true model, then the unique loss function in this case is the self–information loss, which yields the Bayesian update.

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

References.

Ali, S.M. and Silvey, S.D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B* **28**, 131–142.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**(10), 781–791.

Bar-Hillel, M. and Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition*, 11:109–122, 1982.

Barron, A., Schervish, M.J and Wasserman, L. (1999) The consistency of posterior distributions in nonparametric problems *Annals of Statistics* **27**, 536–561.

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

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

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

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

Billingsley, P. (1995). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. ISBN 0-471-00710-2. A Wiley-Interscience Publication.

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

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Datta, G. S., and Sweeting, T. J. (2005). Probability matching priors. *Handbook of statistics*, **25**, 91-114.

De Blasi, P. and Walker, S.G. (2012). Bayesian asymptotics with misspecified models. *Statistica Sinica* **23**, 169-187.

Dempster, A.P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B* **30**, 205–247.

Diaconis, P. and Zabell, S.L. (1982). Updating subjective probability. *Journal of the American Statistical Association* **77**, 822–830.

Doucet, A., and Shephard, N. (2012). Robust inference on parameters via particle filters and sandwich covariance matrices. *University of Oxford, Department of Economics.* No. 606.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincaré* **7**, 1–68.

Escobar, M.D. (1988). *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means.* Unpublished PhD dissertation, Department of Statistics, Yale University.

Fan, J and Li, R. (2002). Variable Selection for Cox's proportional Hazards Model and Frailty Model. *Ann. Statist.* **30**, 1, 74-99.

Faraggi, D. and Simon R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54**, 1475-1485.

Feller, W. (1971). An Introduction to Probability Theory and its Applications. Vol. II. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York-London-Sydney, second edition, 1971.

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

Freund., J. E. (1965) Puzzle or paradox? *Am. Stat.*, 19 (4): 29–44, 1965.

Gardner, M. (1959). The Scientific American Book of Mathematical Puzzles and Diversions. Simon and Schuster, New York, 1959.

Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics* Berlin: Springer–Verlag.

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

Goldstein, M., and Wooff, D. (2007). *Bayes Linear Statistics, Theory & Methods.* **716**. Wiley.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *Elements of Statistical Learning.* Springer.

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

Hjort, N.L., Holmes, C.C., Müller, P. and Walker, S.G. (2010). *Bayesian Nonparametrics.* Cambridge University Press.

Hoff, P. and Wakefield, J.C. (2013). Bayesian sandwich posteriors for pseudo-true parameters. To appear in *Journal of Statistical Planning and Inference*

Hüber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73-101.

Hüber, P. (2009). Robust Statistics (2nd ed.). Hoboken, NJ: John Wiley & Sons Inc.

Hutchison, K. (1999). What are conditional probabilities conditional upon? *Brit. J. Phi. Sci.*, 50:665–695, 1999.

Hutchison,K. (2008). Resolving some puzzles of conditional probability. *Adv. Sci. Lett.*, 1:212–221, 2008.

Ibrahim, J.G. and Chen, M.H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46–60.

Ibrahim, J.G., Chen, M.H. and MacEachern, S.N. (1999). Bayesian variable selection for proportional hazards models. *The Canadian Journal of Statistics*. **27**, 701-171.

Jiang, W. and Tanner, M.A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36**, 2207-2231

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.

Kleijn, B.J.K. and van der Vaart, A.W. (2006). Misspecification in infinite dimensional Bayesian statistics. *Annals of Statistics* **34**, 837–877.

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

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates I. Density estimates. *Annals of Statistics* **12**, 351–357.

Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory* **44**, 2124–2147.

Muller, U. (2012). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. Department of Economics, Princeton University.

Ribatet, M., Cooley, D. and Davison, A. C. (2009). Bayesian inference from composite likelihoods, with an application to spatial extremes. *arXiv preprint* :0911.5357.

Royall, R., and Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B* **65**, 391–404.

Savage, L.J. (1954). *The Foundations of Statistics*. New York. Wiley.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Tibshirani, R. J. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385-395.

Volinsky, C.T., Madigan, D., Raftery, A. E. and Kronmal, A. (1997). Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke. *Journal of the Royal Statistical Society, Series C*. **46**, 4, 433–448.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton University Press.

Walker, S.G. and Hjort, N.L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society, Series B*

Walker, S.G. (2004). New approaches to Bayesian consistency. *Annals of Statistics* **32**, 2028–2043.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician* **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. Theory* **52**, 1307-1321.

Appendix A: Proof of Theorem in Section 2.1. This result is proven by Bissiri and Walker (2012) in their Theorem 2. Here, a shorter proof is given by assuming the differentiability of g.

Assume that Θ contains at least two distinct points, say θ_1 and θ_2. Otherwise, π is degenerate and the thesis is trivially satisfied. To prove this theorem, it is sufficient to consider the case $n = 2$ and a very specific choice for π, taking $\pi = p_0\delta_{\theta_1} + (1 - p_0)\delta_{\theta_2}$, where $0 < p_0 < 1$. Any probability measure ν absolutely continuous with respect to π has to be equal to $p\delta_{\theta_1} + (1 - p)\delta_{\theta_2}$, for some $0 \leq p \leq 1$. Therefore, in this specific situation, the loss $L(\nu; I, \pi)$ becomes:

$$l(p, p_0, h_I) := p\, h_I(\theta_1) \,+\, (1-p)\, h_I(y_1) \\ +\, p_0\, g\left(\frac{p}{p_0}\right) \,+\, (1-p_0)\, g\left(\frac{1-p}{1-p_0}\right),$$

where $h_I(\theta_i) = h(\theta_i, I_1)+h(\theta_i, I_2)$ for $I = (I_1, I_2)$ and $h_I(\theta_i) = h_1(\theta_i, I_j)$ for $I = I_j$, $i, j = 1, 2$. Denote by p_1 the probability $\pi_{I_1}(\{\theta_1\})$, i.e. the minimum point of $l(p, p_1, h_{(I_1,I_2)})$ as a function of p, and by p_2 the probability $\pi_{(I_1,I_2))}(\{\theta_1\})$. By hypotheses, p_2 is the unique minimum point of both loss functions $l(p, p_1, h_{I_2})$ and $l(p, p_0, h_{(I_1,I_2)})$. Again by hypothesis, we shall consider only those functions h_{I_1} and h_{I_2} such that each one of the functions $l(p, p_0, h_{I_1})$, $l(p, p_1, h_{I_2})$, and $l(p, p_0, h_{(I_1,I_2)})$, as a function of p, has a unique minimum point, which is p_1 for the first one and p_2 for the second and third one. The values p_1 and p_2 have to be strictly bigger than zero and strictly smaller than one: this was proved by Bissiri and Walker (2012) in their Lemma 2. Hence, p_1 has to be a stationary point of $l(p, p_0, h_{I_1})$ and p_2 of both the functions $l(p, p_1, h_{I_2})$ and $l(p, p_0, h_{(I_1,I_2)})$. Therefore,

$$g'\left(\frac{p_1}{p_0}\right) \,-\, g'\left(\frac{1-p_1}{1-p_0}\right) \,=\, h_{I_1}(y_1) \,-\, h_{I_1}(\theta_1), \tag{16}$$

$$g'\left(\frac{p_2}{p_0}\right) \,-\, g'\left(\frac{1-p_2}{1-p_0}\right) \,=\, h_{(I_1,I_2)}(y_1) \,-\, h_{(I_1,I_2)}(\theta_1), \tag{17}$$

$$g'\left(\frac{p_2}{p_1}\right) \,-\, g'\left(\frac{1-p_2}{1-p_1}\right) \,=\, h_{I_2}(y_1) \,-\, h_{I_2}(\theta_1). \tag{18}$$

Recall that $h_{(I_1,I_2)} = h_{I_2} + h_{I_1}$. Therefore, summing up term by term (16) and (18), and considering (17), one obtains:

$$g'\left(\frac{p_2}{p_0}\right) \,-\, g'\left(\frac{1-p_2}{1-p_0}\right) \\ =\, g'\left(\frac{p_1}{p_0}\right) \,-\, g'\left(\frac{1-p_1}{1-p_0}\right) \,+\, g'\left(\frac{p_2}{p_1}\right) \,-\, g'\left(\frac{1-p_2}{1-p_1}\right). \tag{19}$$

Recall that by hypothesis (16)–(18) need to hold for every two functions h_{I_1} and h_{I_2} arbitrarily chosen with the only requirement that p_1 and

p_2 uniquely exist. Hence, (19) needs to hold for every (p_0, p_1, p_2) in $(0,1)^3$. By substituting $t = p_0$, $x = p_1/p_0$ and $y = p_2/p_1$, (19) becomes

$$\begin{aligned} g'(xy) \,-\, g'\left(\frac{1-txy}{1-t}\right) \\ = g'(x) \,-\, g'\left(\frac{1-tx}{1-t}\right) \,+\, g'(y) \,-\, g'\left(\frac{1-txy}{1-tx}\right), \end{aligned} \tag{20}$$

which holds for every $0 < t < 1$, and every $x, y > 0$ such that $x < 1/t$ and $y < 1/(xt)$. Being g convex and differentiable, its derivative g' is continuous. Therefore, letting t go to zero, (20) implies that

$$g'(xy) = g'(x) \,+\, g'(y) \,-\, g'(1) \tag{21}$$

holds true for every $x, y > 0$. Define the function $\varphi(\cdot) = g'(\cdot) - g'(1)$. This function is continuous, being g' such, and by (21), $\varphi(xy) = \varphi(x) \,+\, \varphi(y)$ holds for every $x, y > 0$. Hence, $\varphi(\cdot)$ is $k \ln(\cdot)$ for some k, and therefore

$$g'(x) \;=\; k\, \ln(x) \,+\, g'(1), \tag{22}$$

where $k \;=\; (g'(2) \,-\, g'(1))/\ln(2)$. Being g convex, g' is not decreasing and therefore $k \geq 0$. If $k = 0$, then g' is constant, which is impossible, otherwise, for any h_I, p_1 satisfying (16) either would not exist or would not be unique. Therefore, k must be positive. Being $g(1) = 0$ by assumption, (22) implies that $g(x) \;=\; k\, x \ln(x) \,+\, (g'(1) - k)(x-1)$. Hence,

$$h_2(\nu_1, \nu_2) = k \int \ln\left(\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}\right) \mathrm{d}\nu_1$$

holds true for some $k > 0$ and for every couple of measures (ν_1, ν_2) on Θ such that ν_1 is absolutely continuous with respect to ν_2.

Appendix B: Asymptotics under M–open. Here we discuss the asymptotic properties of the general Bayesian learning model. The difference to typical asymptotic studies is that we need to understand what happens when the proxy model chosen is "wrong", in a sense to be made precise. We will do this for the parametric model; $f(x;\theta)$, $\theta \in \Theta$, and the idea is that we want the posterior distribution to accumulate about θ_0; the parameter which minimizes the Kullback–Leibler divergence between the family and the true density function $f_0(x)$; i.e. θ_0 minimizes

$$D(f_0(\cdot), f(\cdot;\theta)) = \int_X f_0(x) \,\log\{f_0(x)/f(x;\theta)\}\,\mathrm{d}x,$$

and we will let

$$\delta = \int f_0(x) \,\log\{f_0(x)/f(x;\theta_0)\}\,\mathrm{d}x.$$

Early work in this direction has been done by Berk (1966) and more recently by Bunke and Milhaud (1998), Kleijn and van der Vaart (2006) and De Blasi and Walker (2012).

For our idea, two assumptions in order to achieve this almost sure accumulation are:

1. The likelihood ratio satisfies

$$n^{-1}\sum_{i=1}^{n}\log\{f(x_i;\widehat{\theta})/f(x_i;\theta_0)\}\to 0 \text{ a.s}$$

   where $\widehat{\theta}$ is the maximum likelihood estimator; that is, $\widehat{\theta}$ maximizes $\prod_{i=1}^{n} f(x_i|\theta)$. We of course assume that this exists in the first place.

2. The best parameter θ_0 is in the support of the prior, so

$$\pi(\theta : 0 < D(f_0(x), f(x;\theta)) < \delta + \eta) > 0$$

   for all $\eta > 0$.

The first is that the maximum likelihood estimator converges to the best parameter θ_0. The topic is dealt with by White (1982) and gives conditions under which $\widehat{\theta}\to\theta_0$ a.s., and the additional assumptions under which condition 1. is satisfied.

Condition 2. is clearly a support condition, so that the prior actually does put mass in a suitable neighborhood of θ_0.

It is sufficient to consider the following problem. Take out a neighborhood N about θ_0 so that now the parameter closest to f_0 has a Kullback–Leibler distance $\delta^* > \delta$, and label the parameter as θ_0^*. We will now show that

$$I_{n1}/I_{n2}\to+\infty \text{ a.s}$$

where

$$I_{n1}=\int_N\left\{\prod_{i=1}^{n}f(x_i)/f_0(x_i)\right\}\pi_N(\mathrm{d}\theta)$$

and

$$I_{n2}=\int_{\Theta-N}\left\{\prod_{i=1}^{n}f(x_i)/f_0(x_i)\right\}\pi_{\Theta-N}(\mathrm{d}\theta)$$

where, for example, π_N is π restricted to the set N. Now, using assumption 2., and following ideas in Barron, Schervish and Wasserman (1999), it can be shown that

$$I_{n1}\geq e^{-nc} \text{ a.s}$$

for all large n, for any $c > \delta$. Also, based on assumption 1.,

$$I_{n2}\leq\prod_{i=1}^{n}\widehat{f^*}(x_i)/f_0(x_i)$$

where $\widehat{f^*}$ is the maximum likelihood restricted to $\Theta - N$. A similar technique is used in Walker and Hjort (2001). Using the result of White (1982), we have that

$$\limsup_n n^{-1} \log I_{n2} \leq -\delta^* \text{ a.s.}$$

Putting this together we see that we have

$$\liminf_n n^{-1} \log I_{n1}/I_{n2} \geq -c + \delta^* > 0 \text{ a.s.}$$

and hence the desired result, since we can choose $\delta < c < \delta^*$.

More recently, De Blasi and Walker (2012) have extended the consistency result of Walker (2004) to the misspecified model case. In Walker (2004) the support condition along with

$$\sum_j \pi(A_{j,\epsilon})^{1/2} < +\infty$$

for all $\epsilon > 0$, where the $(A_{j,\epsilon})$ form a Hellinger partition of the space of densities with balls of size ϵ is sufficient for consistency. For the misspecified case, for accumulation about f_0, the condition becomes

$$\sum_j \pi(A_{j,\epsilon(\alpha)})^{\alpha} < +\infty$$

for all $\alpha > 0$, where, e.g., $\epsilon(\alpha) = (\alpha^2/2)^{2\alpha}$. This is an important result. If the target is θ_0 then we need to be sure we can find it given an arbitrary large amount of information.

Appendix C: Stylized inference problems. The form of the problem is as follows. We have independent stochastic pieces of information I_i. We identify a θ of interest to aid us in the decision process from which we will construct a utility $u(a, \theta)$. Equally, if θ_0 were known we would be happy to select the action a maximizing $u(a, \theta)$. The information (I_i) provides further knowledge about θ_0 through an appropriate loss function $l(\theta, I_i)$.

Let us return to the case when we observe (x_i) independent and identically distributed from some density $f_0(x)$ and $f(x; \theta)$ is the chosen family to model this. In our framework we do not need to concern ourselves whether this family contains $f_0(x)$ or not, provided we use $l(\theta, x) = -\log f(x; \theta)$ under both scenarios.

We then obtain the standard Bayesian updating rule, but now the prior $\pi(\mathrm{d}\theta)$ and posterior $\nu(\mathrm{d}\theta)$ represent our best beliefs about which θ gets us closest to $f_0(x)$. All other aspects of inference can be done with this interpretation of θ and ν. So, for an action $a \in A$, we would maximize $U(a)$ defined previously in terms of ν.

There is also a difference to be highlighted with prediction. It is clear that as far as prediction is concerned,

$$p(x) = \int_{\Theta} f(x;\theta)\,\nu(\mathrm{d}\theta)$$

is the estimate of the density closest to $f_0(x)$ and, unlike the Bayesian interpretation, one knows that the next x is certainly not coming from this density. We can formally obtain $p(x)$ as the estimate for the density closest to $f_0(x)$ by using the utility function

$$u(p,\theta) = -\int_X (p(x) - f(x;\theta))^2\,\mathrm{d}x.$$

Let us move on to more complicated data structures.

C1 Regression model. We will first consider a standard regression model; so we consider the case when x_i come from the density $f_0(x|z)$ and we use the model $f(\cdot|z_i,\theta)$, where the (z_i) are covariate information and the (x_i) are independent observations. So $I_i = (x_i, z_i)$. We recover the Bayesian approach when we take $l(\theta,(x_i,z_i)) = -\log f(x_i|z_i,\theta)$ and as before the usual interpretation of θ is the parameter which takes us closest to $f_0(x_i|z_i)$. If the (z_i) are independent and identically distributed with probability measure μ then we can define θ_0 to minimize

$$\int_Z d_{KL}(f(\cdot|z,\theta), f_0(\cdot|z))\mu(\mathrm{d}z).$$

An infinite collection of the (x_i, z_i) will give us $f_0(x|z)$ and so θ_0 is defined asymptotically. An equivalent idea would be to define θ_0 as the θ minimizing

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} d_{KL}(f(\cdot|z_i,\theta), f_0(\cdot|z_i)).$$

In both of these cases, the loss function is suitable for learning about this θ_0; in the sense that the asymptotic minimizer of

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

is, under mild regularity conditions, precisely θ_0. Alternatively, we could, if the (z_i) are non–stochastic, define θ_0 as minimizing

$$\sup_{z\in Z} d_{KL}(f(\cdot|z,\theta), f_0(\cdot|z)).$$

In this case it would be necessary to construct a loss function $l(\theta,(x,z))$ which asymptotically yielded this θ_0.

C2 Hierarchical model. A random effects hierarchical model will be similar to the above described regression model; yet here we would have the $f(x|z,\theta)$ in a particular form given by

$$f(x_i|z_i,\theta) = \int_B f(x_i|z_i,\beta_i,\theta)\, f(\beta_i|z_i,\theta)\, \mathrm{d}\beta_i.$$

We retain $I_i = (x_i, z_i)$. One determines here that there is a θ to be learnt about which involves an unobserved set of (β_i). We can define θ_0 as in the regression case; that is the θ which minimizes

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} d_{KL}(f(\cdot|z_i,\theta), f_0(\cdot|z_i)).$$

In this case it would be quite challenging to find an alternative to the $l(\theta,(x_i,z_i)) = -\log f(x_i|z_i,\theta)$ loss function. For inference one can solve for $\nu(\mathrm{d}\ \theta)$ and then set up the joint measure

$$\nu(\mathrm{d}\theta,\beta_1,\ldots,\beta_n) = \left\{\prod_{i=1}^{n} f(x_i|z_i,\beta_i,\theta)\, f(\beta_i|z_i,\theta)\right\} \pi(\mathrm{d}\theta)$$

to allow inference via Markov chain Monte Carlo methods, for example.

C3 Time series model. Now let us consider a time series setting whereby it is deemed that x_i depends on $(x_{i-1},\ldots,x_{i-p})$; that is a p–autoregressive model. In this case $I_i = (x_i, x_{i-1},\ldots,x_{i-p})$ and if we model the observations through $f(x_i|x_{i-1},\ldots,x_{i-p},\theta)$ then the Bayesian update arises by taking $l(\theta,I_i) = -\log f(x_i|x_{i-1},\ldots,x_{i-p},\theta)$. In this case the target θ_0 will be the parameter minimizing

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} d_{KL}(f(\cdot|x_{i-1},\ldots,x_{i-p},\theta), f_0(\cdot|x_{i-1},\ldots,x_{i-p})).$$

Of course this assumes that the order is known to be p and typically this will be unknown. Writing the true model as $f_0(x_i|x_{i-1},\ldots,x_1)$ we can construct a general model as

$$f(x_i|x_{i-1},\ldots,x_1,\theta) = \sum_{p=1}^{\infty} w_p\, f_p(x_i|x_{i-1},\ldots,x_{i-p},\theta_p)$$

where w_p is the probability that the correct order is p and $\theta = (\theta_1,\theta_2,\ldots)$ so $\pi(\theta) = \prod_{p=1}^{\infty} \pi_p(\theta_p)$ and $\pi_p(\theta_p)$ represents the beliefs about which θ_p takes $f_p(\cdot|x_{i-1},\ldots,x_{i-p},\theta_p)$ closest to $f_0(\cdot|x_{i-1},\ldots,x_{i-p})$, conditional on the truth of p being the correct order. The Bayesian update now arises by taking $l(\theta,(x_i,x_{i-1},\ldots,x_1)) = -\log f(x_i|x_{i-1},\ldots,x_i,\theta)$.

C4 Grouped data model. Here we consider the case when we have repeated observations on independent units; so $I_i = (x_{i1}, \ldots, x_{in_i}, z_i)$, where the z_i are unit specific covariates. If it assumed that the $(x_{ij})_j$ are conditionally independent given an unobserved parameter β_i then we would have a model of the type

$$f(x_i|z_i, \theta) = \int \prod_{j=1}^{n_i} f(x_{ij}|z_i, \beta_i, \theta)\, f(\beta_i|z_i, \theta)\, \mathrm{d}\beta_i$$

and we recover the Bayesian update when we take

$$l(\theta, (x_i, z_i)) = -\log f(x_i|z_i, \theta)$$

and the interpretation for the prior $\pi(\theta)$ is again to do with beliefs about where the θ taking this model closest to the true model is to be located.

This section shows that it is possible to undertake Bayesian inference with models in the M–open view by taking the logarithmic loss functions associated with these models. The interpretation of θ is different however. We construct prior distributions and learn about the best parameter θ_0 which takes us closest to the true model. It is assumed the data can give the true model completely and therefore there is access to θ_0.

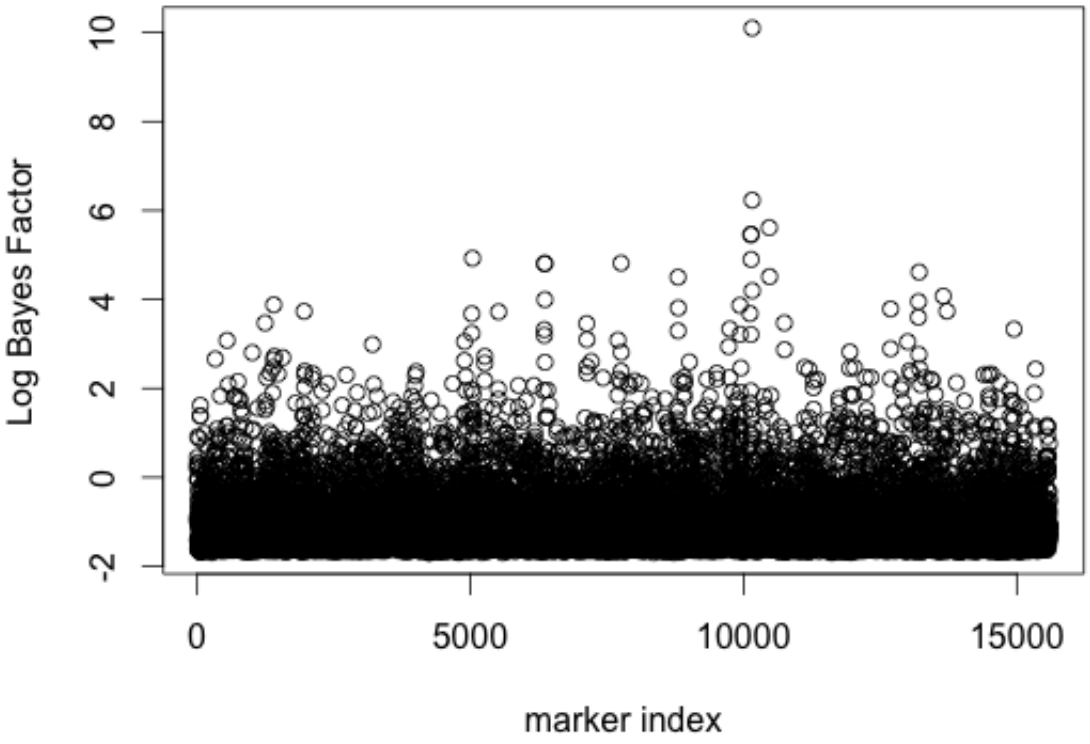

Figure 1: Log Bayes Factor (Laplace) vrs marker index along chromosome

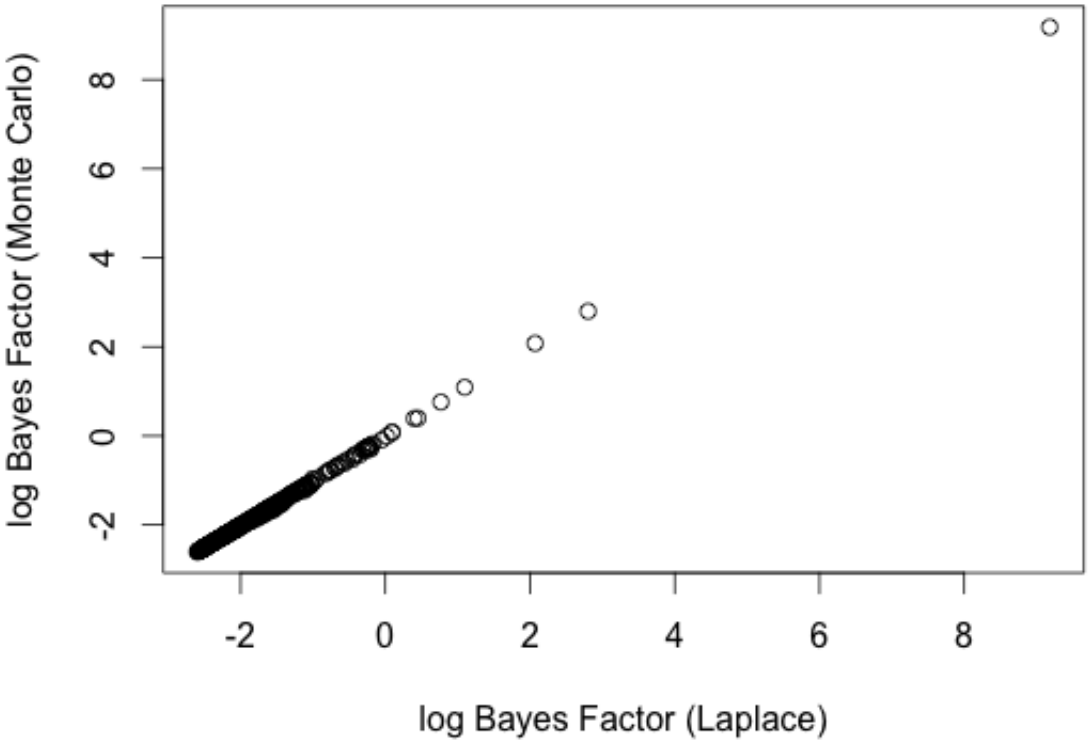

Figure 2: Log Bayes Factor using 500 Monte Carlo samples vrs Laplace approximation: at 500 random markers

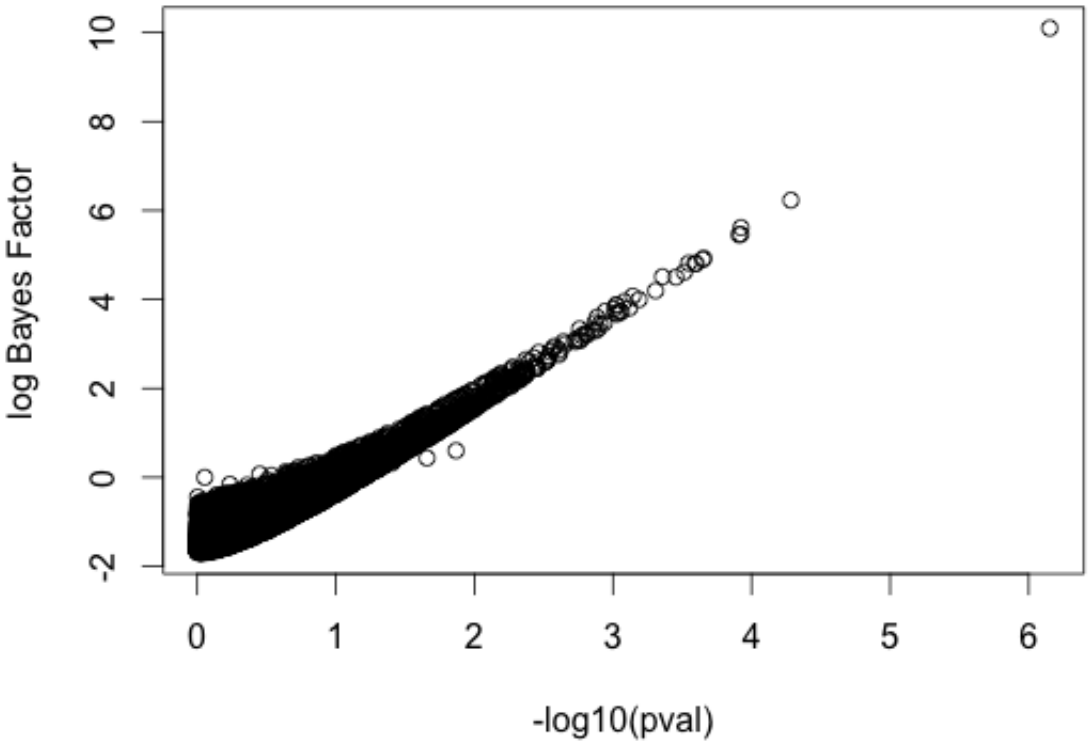

Figure 3: Log Bayes Factor vrs -log10 p-value of association

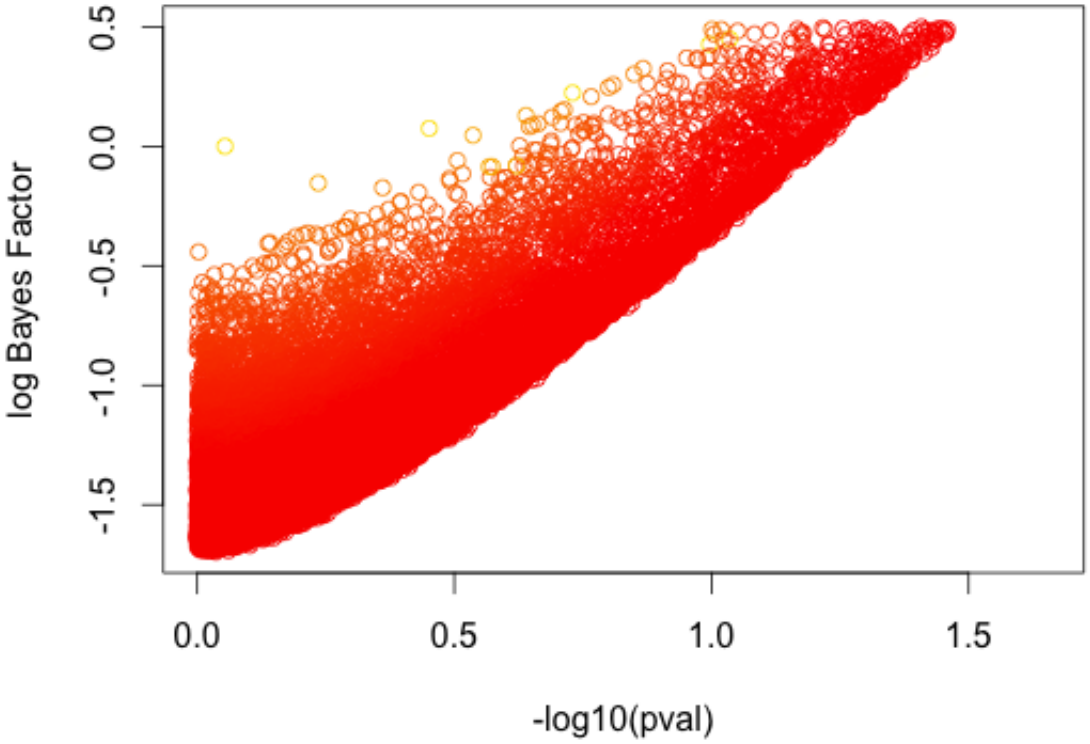

Figure 4: Log Bayes Factor vrs -log10 p-value of association coloured by standard error in MLE

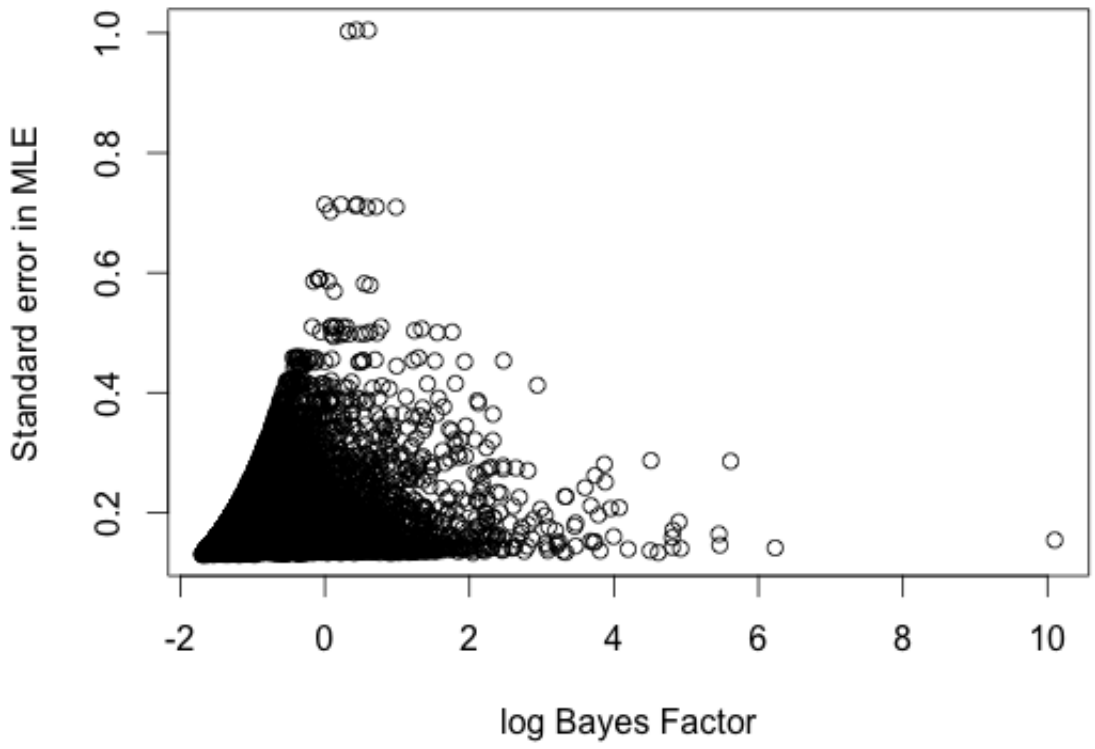

Figure 5: Standard Error in MLE vrs log Bayes Factor

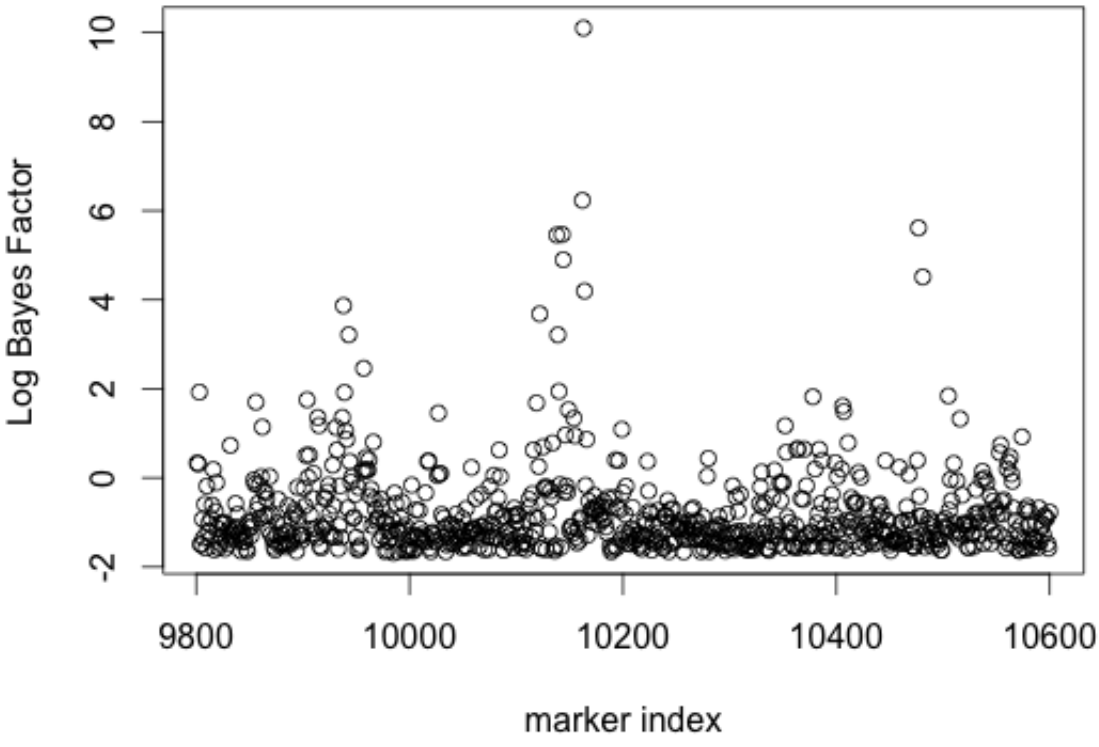

Figure 6: Log Bayes Factor vrs marker index in the “hit region”

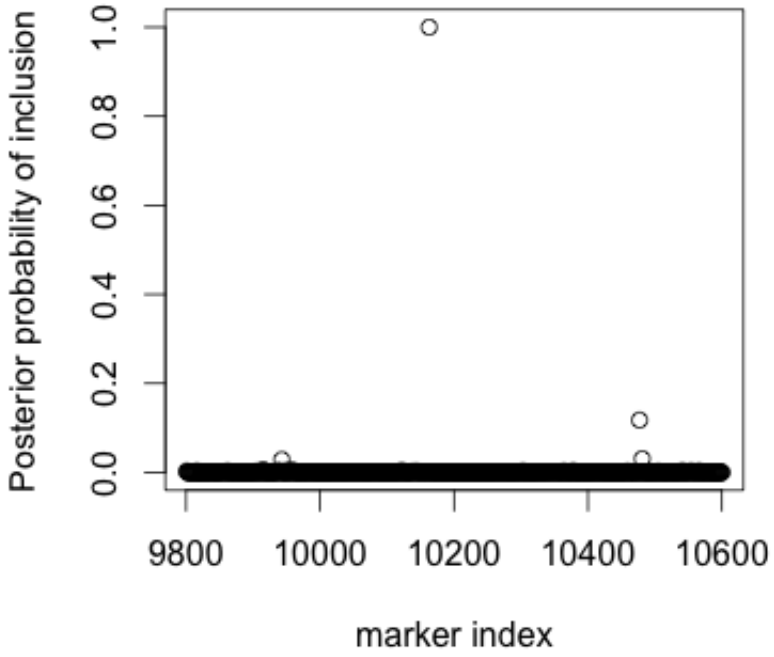

Figure 7: Posterior marginal inclusion probability from multiple marker model

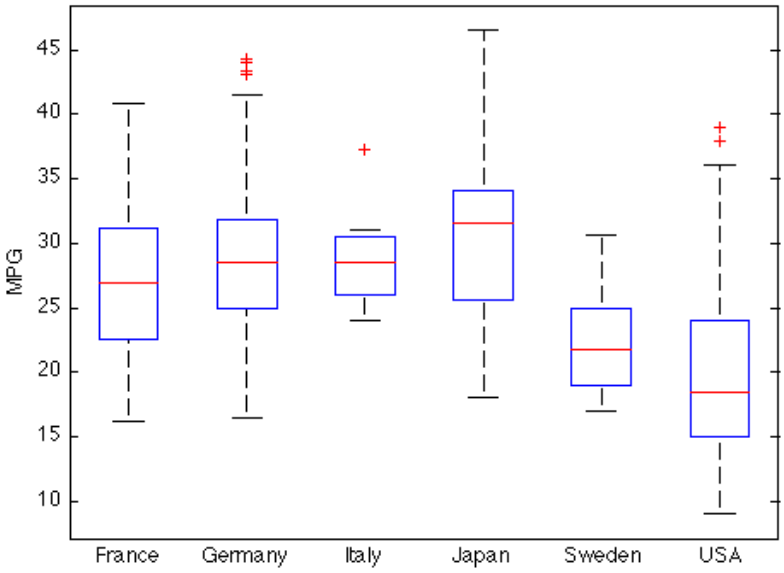

Figure 8: Boxplot of cars MPG data; taken from the MATLAB boxplot.m help file illustration.

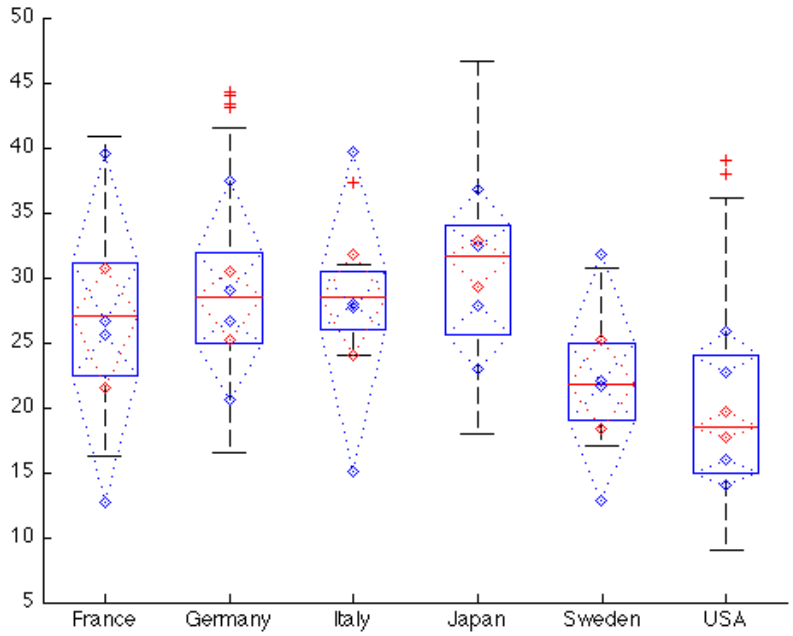

Figure 9: General Bayesian Boxplot of cars MPG data using Unit Information Loss

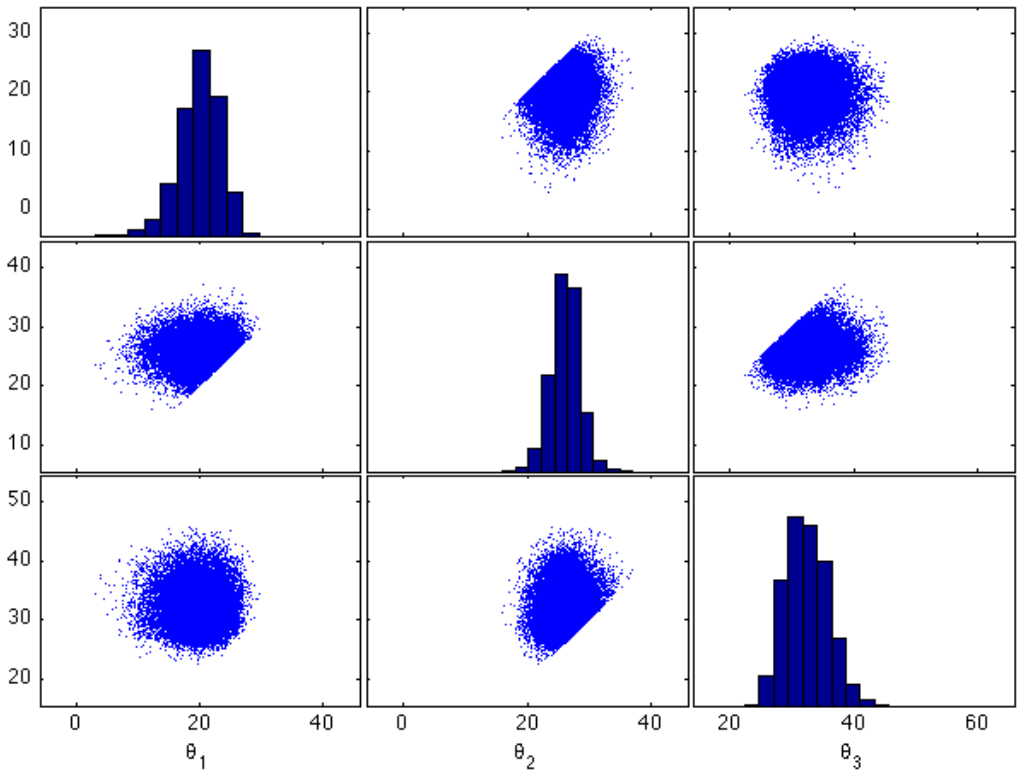

Figure 10: Posterior samples for quartiles of Franch cars MPG data using Unit Information Loss

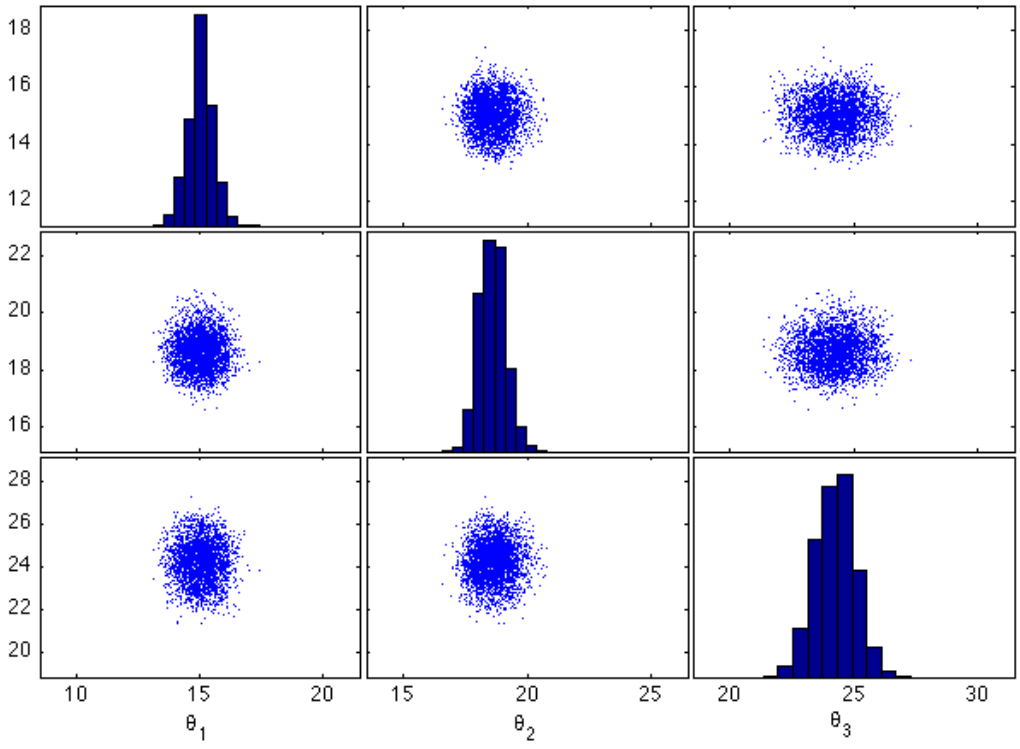

Figure 11: Posterior samples for quartiles of USA cars MPG data using Unit Information Loss