An Objective Bayesian Criterion to Determine Model Prior Probabilities

CRISTIANO VILLA and STEPHEN WALKER
School of Mathematics, Statistics and Actuarial Science, University of Kent
Division of Statistics and Scientific Computation, University of Texas at Austin

Abstract

We discuss the problem of selecting among alternative parametric models within the Bayesian framework. For model selection problems which involve non-nested models, the common objective choice of a prior on the model space is the uniform distribution. The same applies to situations where the models are nested. It is our contention that assigning equal prior probability to each model is over simplistic. Consequently, we introduce a novel approach to objectively determine model prior probabilities conditionally on the choice of priors for the parameters of the models. The idea is based on the notion of the *worth* of having each model within the selection process. At the heart of the procedure is the measure of this *worth* using the Kullback–Leibler divergence between densities from different models.

Some key words: Bayesian model selection, Kullback–Leibler divergence, objective Bayes, self-information loss

1 Introduction and Background

This paper introduces a novel approach to objectively determine model prior probabilities for model selection problems. We focus on the case where the model is the pair $M = \{f(x|\theta), \pi(\theta)\}$, where $f(x|\theta)$ is the probability distribution, characterised by the parameter θ (possibly, a vector of parameters), and $\pi(\theta)$ is the prior distribution representing beliefs on the model parameter. We assume these have been specified, so our objectivity is for the model priors only. Furthermore, our approach can deal with either non-nested or nested models.

Suppose we want to compare k models for observed data $x = (x_1, \ldots, x_n)$. The models are denoted by

$$M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}, \qquad j = 1, \ldots, k.$$

The usual way to perform the comparison is to compute pairwise Bayes factors between models in the model space $\{M_1, \ldots, M_k\}$; for example, Berger & Pericchi (2001), Robert (2001) and Pericchi (2005) and the references therein. In general, the Bayes factor between model M_j and model M_i is given by

$$B_{ji} = \frac{m_j(x)}{m_i(x)} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j)\, d\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i)\, d\theta_i}, \qquad i \neq j \in \{1, \ldots, k\},$$

where $m_j(x)$ and $m_i(x)$ are the marginal densities of x under, respectively, model M_j and model M_i. We see that the Bayes factor B_{ji} is a weighted likelihood ratio (for the observed data) of M_j over M_i, where the weights are represented by the prior probabilities $\pi_j(\theta_j)$ and $\pi_i(\theta_i)$. Then, given model prior probabilities, $P(M_j)$, $j = 1, \ldots, k$, the posterior mass for each element in the model space, given the data x, is given by

$$P(M_j|x) = \left[\sum_{j=1}^{k} \frac{P(M_j)}{P(M_i)} B_{ji}\right]^{-1},$$

which can be used, for example, to select the model with the highest posterior probability. Alternatively, the weights can be used in a model averaging procedure. Berger & Pericchi (2001) and Chipman *et al.* (2001), for example.

There are other methods which allow for model selection; such as intrinsic Bayes factors (Casella & Moreno, 2006) and fractional Bayes factors (Carvalho & Scott, 2009; O'Hagan, 1995). A review of these approaches and others can be found, for example, in Pericchi (2005). Other noteworthy publications include Berger & Pericchi (2001), Chipman *et al.* (2001), Pérez & Berger (2002), Stracham & van Dijk (2003) and Bayarri *et al.* (2012), which defines a set of criteria that an objective prior for model parameter should satisfy.

As the scope of the paper is limited to model priors, we will not discuss further methods for model choice or for parameter-specific priors, referring to the specific literature on the subjects.

When it comes to defining priors for models, the literature is somewhat sparse. It appears that the common objective choice for a prior on the space of models is the uniform; i.e. $P(M_j) = 1/k$, for $j = 1, \ldots, k$. As in, for example, Berger & Pericchi (2001) or Robert (2001). In other words, the objective approach assigns equal importance to each model in the set of all the possible models. Whilst the uniform prior is regarded as the common objective choice for non-nested models, for nested model selection problems, such as regression models or graphical models, Scott & Berger (2010) have proposed a different model prior. The aim of this prior is to correct for multiplicity in variable selection problems, and it assigns the highest probability to the extreme models (i.e. the null model and the full model).

On the other hand, we propose a novel objective method to assign prior mass to each model on the basis of the *worth* that it has, with respect to the other models in the model space. In fact, we believe that the assignment of equal importance to each model, as a result of objective thinking, is too simplistic, and that models do not necessarily have the same *worth* in relation to each other. This approach is a generalisation of the idea proposed to define objective priors on discrete parameter spaces (Villa & Walker, 2014a,b).

The outline of the paper is as follows. In Section 2 we discuss our idea and introduce the notation that is used throughout the paper. Section 3 is dedicated to non-nested models. In particular, we illustrate the case of selecting between two models for discrete data with one parameter, and selecting between two and three multiparameter continuous models. In Section 4 we study model selection when there are nested models in the scenario, including an illustration where there are, simultaneously, nested and non-nested models. Section 5 contains a brief discussion.

2 The idea

We introduce our idea to assign prior mass to models by means of the following illustration.

Let us consider three trivial models $M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}$, for $j = 1, 2, 3$. Here, each model represents a single density. We also assume that models M_1 and M_2 (and so, densities f_1 and f_2) are very similar, and that the third model M_3 (density f_3) is significantly different from the other two. We do not question the rational behind this scenario set up, we just assume that there is one.

By analysing this scenario in the light of the utility of each model we note that the *worth* of models M_1 or M_2 is less than the one of model M_3. In fact, should we lose either M_1 or M_2, we would still have the remaining one to "represent" that position in the set of all possible models. On the other hand, M_3 would be more valuable, as its removal from the set of choices would lead to bad inference if it turned out to be the true model.

Having identified this approach to assign the mass to each model on the basis of *worth*, we see it takes into consideration the "position" of each model with respect to the others. The quantification of the *worth* comes from a result in Berk (1966) which says that, if the model is misspecified, the posterior distribution asymptotically tends to accumulate at the nearest model in terms of the Kullback–Leibler divergence (Kullback & Leibler, 1951). Therefore, if we were to remove model M_j from the set of possible models, and it is the true one, the loss we would incur is given by the Kullback–Leibler divergence from it to the nearest of $\{f_i\}$, $i \neq j$. And because the nearest model is determined by the choice of the possible models only, the objectivity of the approach emerges. Thus, by defining the Kullback–Leibler divergence between M_j and M_i by $D_{KL}(f_j \| f_i) = \int f_j \log(f_j/f_i)$, the loss associated with model M_j would be

$$l(M_j) = l_j = -\min_{j \neq i} D_{KL}(f_j \| f_i). \tag{1}$$

That is, the larger the value of $\min_{j \neq i} D_{KL}(f_j \| f_i)$ the greater the utility (or, equivalently, the smaller the loss) of keeping the model.

If we consider the mass to be put on each model $P(M_j)$, this can be linked to the *worth* of the model via the *self-information* loss function. The *self-information* loss function (also known as the *log-loss* function in machine learning) measures the performance of a probability statement with respect to an outcome. Thus, for every probability assignment $P = \{P(A), A \in \Omega\}$, the *self-information* loss function is defined as

$$l(P, A) = -\log P(A).$$

More details and properties of this particular loss function can be found, for example, in Merhav & Feder (1998). Therefore, for each model M_j we have a measure of the information loss related to its *worth*, given by (1), and related to the *self-information*, given by $-\log P(M_j)$. We then equate the two losses, yielding

$$-\log P(M_j) = -\min_{j \neq i} D_{KL}(f_j \| f_i),$$

equivalently

$$P(M_j) \propto \exp\left\{\min_{j \neq i} D_{KL}(f_j \| f_i)\right\}. \tag{2}$$

In other words, the mass that we assign to each model is proportional to the exponential of the Kullback–Leibler divergence between the model and the nearest one in the set of options.

We can now take the basis of the idea to proper models, that is in a full model selection scenario. Let us assume that we have to select only between two models

$$M_1 = \{f_1(x|\theta_1), \pi_1(\theta_1)\} \quad \text{and} \quad M_2 = \{f_2(x|\theta_2), \pi_2(\theta_2)\},$$

where we assume that the prior of the parameter $\theta_1 \in \Theta_1$, $\pi_1(\theta_1)$, and the prior on the parameter $\theta_2 \in \Theta_2$, $\pi(\theta_2)$, are known and proper. Following the criterion illustrated above, the prior mass on M_1, $P(M_1)$, is determined on the basis of what is lost if model M_1 is removed, and it is the true one. To elaborate, by applying Berk's result, if model M_1 is removed and θ_1 is the true parameter value, the posterior asymptotically accumulates on the density in M_2 which minimises the Kullback–Leibler divergence from $f_1(\cdot|\theta_1)$. So the utility is given by $\inf_{\theta_2} D_{KL}(f_1(\cdot|\theta_1) \| f_2(\cdot|\theta_2))$; but since θ_1 is unknown, we evaluate the expected utility as

$$\int_{\Theta_1} \inf_{\theta_2} D_{KL}(f_1(\cdot|\theta_1)\|f_2(\cdot|\theta_2))\pi_1(\theta_1)\, d\theta_1.$$

In other words, we associate a *worth* to the whole model M_1 which is the expectation of the *worth* weighted by the prior we chose to put on the parameter, that is $\pi_1(\theta_1)$. Thus, by considering (2), $P(M_1)$ is proportional to the exponential of the expected minimum loss between the models, and so

$$P(M_1) \propto \exp\left\{ \int_{\Theta_1} \inf_{\theta_2} D_{KL}\Big(f_1(x|\theta_1)\|f_2(x|\theta_2)\Big)\pi_1(\theta_1)\, d\theta_1 \right\}. \tag{3}$$

Similarly, the mass associated to M_2, $P(M_2)$, is proportional to the exponential of the expected minimum loss between model M_2 and model M_1, given by

$$P(M_2) \propto \exp\left\{ \int_{\Theta_2} \inf_{\theta_1} D_{KL}\Big(f_2(x|\theta_2)\|f_1(x|\theta_1)\Big)\pi_2(\theta_2)\, d\theta_2 \right\}. \tag{4}$$

An important and fundamental aspect of our approach is that the prior probability assigned to a model, $P(M_j)$, depends on the prior assigned to the parameter of the model, $\pi_j(\theta_j)$. Section 4 gives a rational for this aspect in a clear setting: in particular, the so-called *Jeffreys-Lindley paradox* (Lindley, 1957).

The most general scenario is represented by a model space of k elements, where each model is specified by a vector of parameters of finite dimension. Let us consider a model selection problem with model space $\{M_1, \ldots, M_k\}$, with $M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}$, $j = 1, \ldots, k$. A compact notation for the prior mass for model M_j is then given by

$$P(M_j) \propto \exp\left[\mathbb{E}_{\pi_j}\left\{ \inf_{\theta_m, m\neq j} D_{KL}\Big(f_j(x|\theta_j)\|f_m(x|\theta_m)\Big) \right\} \right], \qquad \text{for } j = 1, \ldots, k,$$

where the expectation is taken with respect to the prior assigned to the parameters of model f_j, that is $\pi_j(\theta_j)$. In other words, the prior assigned to model M_j can be seen as if it is obtained by measuring the divergence between $f_j(x|\theta_j)$ and any other model, and selecting the smaller one.

In the following sections we discuss some illustrations for the non-nested and the nested model selection case. To simplify the notation, unless otherwise specified, the numbering of the various models (including the probability and prior distribution that form them) starts afresh in each illustration.

3 Non-nested models

In this section we examine the objective approach that we are proposing in the paper and, in particular, to scenarios where the elements of the model space are non-nested. In the first illustra-

tion we compare two discrete models; a Poisson and a geometric probability mass function. Then, we consider a model selection problem with two multiparameter continuous densities: Weibull and log-normal. Finally, in the third illustration, we extend the latter problem to a three model selection problem by adding a gamma density.

3.1 Poisson and Geometric

Let us assume that we have observed a set of observations x from a phenomenon we know to have support $\mathcal{X} = \{0, 1, 2, \ldots\}$. The two models are given by

$$M_1 = \left\{ f_1(x|\theta) = \theta^x e^{-\theta}/x!\,, \pi_1(\theta) \right\} \quad \text{and} \quad M_2 = \left\{ f_2(x|\phi) = \phi(1-\phi)^x, \pi_2(\phi) \right\},$$

that is, M_1 is a Poisson distribution with rate parameter $\theta \in (0, +\infty)$, and M_2 is a geometric distribution with probability of success $\phi \in (0, 1)$.

Following the objective approach we have outlined in Section 2, we first consider the mass to be assigned to model M_1. This mass depends on what we lose if we remove model M_1, which in the two model case implies choosing M_2, and it is the true one. By applying (3) we have

$$P(M_1) \propto \exp\left\{ \int \inf_{\phi} D_{KL}\Big(f_1(x|\theta) \| f_2(x|\phi) \Big) \pi_1(\theta)\, d\theta \right\}. \tag{5}$$

To determine the mass in (5), we first find the Kullback–Leibler divergence between a Poisson distribution with parameter θ and a geometric distribution with parameter ϕ. As shown by Theorem 1 in the supporting information, this is given by

$$\begin{aligned} D_{KL}(f_1(x|\theta) \| f_2(x|\phi)) &= \sum_{x=0}^{\infty} \left[\frac{\theta^x}{x!} e^{-\theta} \log\left\{ \frac{e^{-\theta}\theta^x/x!}{\phi(1-\phi)^x} \right\} \right] \\ &= \theta \log\theta - \sum_{x=0}^{\infty} \left(\log x! \, \frac{\theta^x}{x!} e^{-\theta} \right) - \theta - \log\phi - \theta\log(1-\phi). \end{aligned} \tag{6}$$

The divergence (6) is minimised, with respect to ϕ, by $\phi = 1/(1+\theta)$. By replacing this result into (6), we obtain the minimum Kullback–Leibler divergence between a Poisson and a geometric distributions,

$$\inf_{\phi} D_{KL}(f_1(x|\theta) \| f_2(x|\phi)) = -\theta + \theta\log(1+\theta) + \log(1+\theta) - \sum_{x=0}^{\infty} \left(\log x! \, \frac{\theta^x}{x!} e^{-\theta} \right).$$

For this illustration, we have considered a gamma prior on the parameter θ, with shape and scale parameter both equal to one; that is, $\pi_1(\theta) \sim Ga(1,1) = \exp(-\theta)$. Therefore

$$\begin{aligned} P(M_1) &\propto \exp\left\{\int \inf_{\phi} D_{KL}\Big(f_1(x|\theta)\|f_2(x|\phi)\Big)e^{-\theta}d\theta\right\} \\ &= \exp(0.09) \\ &= 1.09. \end{aligned} \tag{7}$$

The result in (7) is obviously affected by the choice of the prior. In particular, we note that if the variance of $\pi_1(\theta)$ increases, corresponding on an increase of uncertainty about the true value of the parameter, the mass assigned to model M_1 increases. For example, if we chose the prior to be $\pi_1(\theta) \sim Ga(10, 1)$ (corresponding to a variance of 10), the corresponding mass on M_1 would be $P(M_1) \propto 2.16$. Similarly, if the variance decreases, therefore the uncertainty about the parameter is more limited, the approach will assign a lower mass. For example, for $\pi_1(\theta) \sim Ga(1, 5)$ (variance equal to 0.04), we have $P(M_1) \propto 1.01$. Intuitively, if we have a relatively high uncertainty about the true value of the parameter, the loss (in expectation) we would incur in choosing the wrong model would be relatively large. Hence, the model assumes more importance in the overall scenario. Vice versa, if our prior knowledge about the true value of the parameter is relatively precise (i.e. low uncertainty), the loss of information in choosing the wrong model would be (in expectation) relatively low.

With a similar procedure, by applying (4) we obtain the mass for model M_2. In fact, the Kullback–Leibler divergence between a geometric distribution and a Poisson distribution is given by

$$\begin{aligned} D_{KL}(f_2(x|\phi)\|f_1(x|\theta)) &= \sum_{x=0}^{\infty}\left[\phi(1-\phi)^x \log\left\{\frac{\phi(1-\phi)^x}{e^{-\theta}\theta^x/x!}\right\}\right] \\ &= \log\phi + \frac{1-\phi}{\phi}\log(1-\phi) - \frac{1-\phi}{\phi}\log\theta + \theta \\ &\quad + \sum_{x=0}^{\infty}\Big\{\phi(1-\phi)^x \log x!\Big\}, \end{aligned} \tag{8}$$

which is minimised by $\theta = (1-\phi)/\phi$ (refer to Theorem 1 in the supporting information). We replace this result in (8), and obtain

$$\inf_{\theta} D_{KL}(f_2(x|\phi)\|f_1(x|\theta)) = \log\phi + \frac{1-\phi}{\phi}\log\phi + \frac{1-\phi}{\phi} + \sum_{x=0}^{\infty}\Big\{\phi(1-\phi)^x \log x!\Big\}.$$

The prior for parameter ϕ has been selected to be a beta distribution with both shape parameter values equal to two. That is, $\pi_2(\phi) \sim Be(2, 2) \propto \phi(1-\phi)$. Thus, the mass to be put on model M_2

is determined to be

$$\begin{aligned} P(M_2) &\propto \exp\left\{\int \inf_{\theta} D_{KL}\Big(f_2(x|\phi)\|f_1(x|\theta)\Big)\phi(1-\phi)\,d\phi\right\} \\ &= \exp(0.47) \\ &= 1.60. \end{aligned} \tag{9}$$

Also in the computation of $P(M_2)$ we have noted, as expected, that the prior mass assigned to the model depends on the variance of the prior distribution for ϕ. In particular, similarly to the computation of $P(M_1)$, the larger the variance the more the mass, and vice versa.

Results in (7) and (9) can be normalised. The resulting prior distribution for this model selection problem (i.e. given the chosen models and the prior distributions of the respective parameters), is $P_N(M_1) = 0.41$ and $P_N(M_2) = 0.59$. It is not possible to perform a comparison between the variances of the two prior distributions, $\pi_1(\theta)$ and $\pi_2(\phi)$. However, it is plausible to assume that there is always the possibility to chose them in a way that the prior masses on the models are equal. In fact, if we consider as prior distribution for θ a gamma with shape parameter 5 and rate parameter 1, and as prior for ϕ a beta with both parameters equal to two, we obtain $P(M_1) \propto 1.59$ and $P(M_2) \propto 1.60$. Normalising, we have the uniform prior of the models given by $P_N(M_1) = 0.50$ and $P_N(M_2) = 0.50$. Under these circumstances, we can assume that the level of uncertainty about θ and ϕ is virtually the same.

It is also interesting to examine what happens when the uncertainty about the parameter value of one model is much larger than the uncertainty on the parameter of the other model. For example, let us keep the prior on ϕ fixed, that is $\pi_2(\phi) \sim Be(2, 2)$, and set $\pi_1(\theta) \sim Ga(20, 1/2)$. In this case, the variance of $\pi_1(\theta)$ is equal to 80, which is a much larger value than the case where $\pi_1(\theta) \sim Ga(1, 1)$. Thus, we have that $P(M_1) \propto \exp(1.43) = 4.17$. Normalising, $P_N(M_1) = 0.72$ and $P_N(M_2) = 0.28$.

3.2 Weibull and Log-normal

In this illustration we consider a scenario where the quantity of interest x has a continuous support $\mathcal{X} = (0, +\infty)$. We also show how the approach can be applied to models with dimension of the parameter space larger than one. We consider model M_1 to be a Weibull density with scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$. Model M_2 is a log-normal density with location parameter $\mu \in \mathbb{R}$ (in the log-scale), and shape parameter σ^2. These distributions are often considered as option to model data, for example, in survival analysis studies (Klein & Moeschberger, 1997). Note that we will consider the parametrisation expressed with the precision parameter $\tau = 1/\sigma^2 > 0$. Therefore

$$M_1 = \left\{ f_1(x|\lambda,\kappa) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\}, \pi_1(\lambda,\kappa)\right\},$$
$$M_2 = \left\{ f_2(x|\mu,\tau) = \frac{1}{x}\left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}\tau(\log x - \mu)^2\right\}, \pi_2(\mu,\tau)\right\}.$$

On the basis of our approach, the prior mass to be assigned to model M_1 and model M_2 is determined, respectively, by

$$P(M_1) \propto \exp\left\{\int\int \inf_{\mu,\tau} D_{KL}\Big(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau)\Big)\pi_1(\lambda,\kappa)\, d\kappa d\lambda\right\}, \tag{10}$$

and

$$P(M_2) \propto \exp\left\{\int\int \inf_{\lambda,\kappa} D_{KL}\Big(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa)\Big)\pi_2(\mu,\tau)\, d\mu d\tau\right\}. \tag{11}$$

Recall that the expression in the exponential in (10) represents the expected loss we would incur in choosing model M_2 when model M_1 is the true one. Similarly, the expression at the exponential in (11) represents the expected loss should we chose model M_1 when M_2 is the true model.

To compute the mass for model M_1, we first obtain the Kullback–Leibler divergence between a Weibull density and a log-normal density, as shown in Theorem 2 in the supporting information.

$$\begin{aligned} D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau)) &= \int_0^\infty f_1(x|\lambda,\kappa)\log\left\{\frac{f_1(x|\lambda,\kappa)}{f_2(x|\mu,\tau)}\right\}dx \\ &= \log\kappa + \kappa\,\mathbb{E}_1(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}_1(x^\kappa) - \frac{1}{2}\log\tau + \frac{1}{2}\log(2\pi) \\ &\quad + \frac{1}{2}\tau\,\mathbb{E}_1(\log^2 x) - \tau\mu\,\mathbb{E}_1(\log x) + \frac{1}{2}\tau\mu^2, \end{aligned} \tag{12}$$

where the expectations are with respect to $f_1(x|\lambda,\kappa)$, with $\mathbb{E}_1(\log x) = \log\lambda - \gamma/\kappa$ ($\gamma \approx 0.5772$ is the Euler's constant), $\mathbb{E}_1(x^\kappa) = \lambda^\kappa$, and $\mathbb{E}_1(\log^2 x) = \pi^2/(6\kappa^2) + (\log\lambda - \gamma/\kappa)^2$ ($\pi^2/(6\kappa^2)$ is the variance of the logarithm of x, that is $Var(\log x) = \pi^2/(6\kappa^2)$). The infimum of the divergence in (12), with respect to parameters μ and τ, is attained at $\mu = \mathbb{E}_1(\log x) = \log\lambda - \gamma/\kappa$ and $\tau = 1/Var(\log x) = 6\kappa^2/\pi^2$. Recalling that, if random variable x is log-normally distributed with parameters μ and τ, then random variable $y = \log x$ has a normal distribution with mean μ and precision τ, we see that the minimum divergence between a Weibull and a log-normal is attained when, in the log scale, both densities have the same mean and variance. And this is a sensible result. Thus, by replacing the expressions of the expectations of the functions of x into equation (12), we have

$$\begin{aligned}\inf_{\mu,\tau} D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau)) &= \log\kappa + \kappa\,\mathbb{E}_1(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}_1(x^\kappa) \\ &\quad + \frac{1}{2}\log\{Var(\log x)\} + \frac{1}{2}\log(2\pi) + \frac{1}{2} \\ &= \frac{1}{2}\log(2\pi) + \log\pi - \gamma - \frac{1}{2}\log 6 - \frac{1}{2}.\end{aligned}$$

We note that the minimum divergence, with respect to μ and τ, from a Weibull density to a log-normal density, does not depend on the values of parameters λ and κ, and it has value 0.09. An important aspect of this result is that the mass to be assigned to model M_1 does not depend on the choice of the priors for λ and κ. By applying (10), the prior mass for the Weibull density is $P(M_1) \propto \exp(0.09) = 1.09$.

With an analogous approach, we compute the value of $P(M_2)$. The Kullback–Leibler divergence from a log-normal density with parameters μ and τ, and a Weibull density with parameters λ and κ (refer to Theorem 2 in the supporting information) is given by

$$\begin{aligned}D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa)) &= \int_0^\infty f_2(x|\mu,\tau)\log\left\{\frac{f_2(x|\mu,\tau)}{f_1(x|\lambda,\kappa)}\right\}dx \\ &= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\,\mathbb{E}_2(\log^2 x) + \tau\mu\,\mathbb{E}_2(\log x) - \frac{1}{2}\tau\mu^2 \\ &\quad - \log\kappa - \kappa\,\mathbb{E}_2(\log x) + \kappa\log\lambda + \frac{1}{\lambda^\kappa}\mathbb{E}_2(x^\kappa), \qquad (13)\end{aligned}$$

where in this case the expectations are with respect to the log-normal density. In particular, $\mathbb{E}_2(\log x) = \mu$, $\mathbb{E}_2(x^\kappa) = \exp\left\{\kappa^2/(2\tau) + \mu\kappa\right\}$ and $\mathbb{E}_2(\log^2 x) = 1/\tau + \mu^2$. The divergence in (13) has infimum for $\lambda = \exp\{1/(2\sqrt{\tau}) + \mu\}$ and $\kappa = \sqrt{\tau}$, giving

$$\begin{aligned}\inf_{\lambda,\kappa} D_{KL}(f_2(x|\mu,\tau)\|f_1(x|\lambda,\kappa)) &= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\left(\frac{1}{\tau} + \mu^2\right) + \tau\mu^2 - \frac{1}{2}\tau\mu^2 - \frac{1}{2}\log\tau \\ &\quad - \sqrt{\tau}\mu + \sqrt{\tau}\left(\frac{1}{2\sqrt{\tau}} + \mu\right) + 1 \\ &= 1 - \frac{1}{2}\log(2\pi).\end{aligned}$$

Again, we note that the minimum divergence between the models is a constant, and its value is of 0.08. As such, the choice of $\pi_2(\mu,\tau)$ does not have impact on the prior mass that, in accordance to our approach, is assigned to model M_2. We then compute this mass as $P(M_2) \propto \exp(0.08) = 1.08$.

By normalising, we have that $P_N(M_1) = 0.50$ and $P_N(M_2) = 0.50$, which is uniform and, in

this case, traces back to the common objective approach to assign equal prior probability to two models.

The result deriving from Theorem 2 is easy to derive and it is discussed, for example, in Dumonceaux *et al.* (1973) and Dumonceaux & Antle (1972). In essence, if we consider two models with location and scale parameters, say M_a and M_b, the minimum Kullback–Leibler divergence between M_a and M_b, $D_{KL}(M_a\|M_b)$, does not depend on the parameters of the model M_b. Vice versa, $D_{KL}(M_b\|M_a)$ has an infimum which does not depend on the parameters of model M_a. In the light of our approach, this means that the choice of the prior distribution for the parameters has no influence on the value of prior mass assigned to each model. Furthermore, for the Weibull and log-normal models, the Kullback–Leibler divergences are very similar, resulting in a prior mass that is basically uniform.

3.3 Weibull, Log-normal and Gamma

The approach we propose, as discussed in Section 2, can be applied to model spaces with a number of elements as large as necessary. To illustrate this, we consider the case where, in addition to the two models introduced in Section 3.2, we add a third one. In particular, a gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. This distribution as well, is considered as an option to model survival analysis data (Klein & Moeschberger, 1997). The model space is then formed by the following three models

$$
\begin{aligned}
M_1 &= \left\{ f_1(x|\lambda,\kappa) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\}, \pi_1(\lambda,\kappa)\right\},\\
M_2 &= \left\{ f_2(x|\mu,\tau) = \frac{1}{x}\left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}\tau(\log x-\mu)^2\right\}, \pi_2(\mu,\tau)\right\},\\
M_3 &= \left\{ f_3(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x), \pi_3(\alpha,\beta)\right\}.
\end{aligned}
$$

Given that our approach assigns mass on a model on the basis of what it is lost if the model is removed from the model space and it is the true model, and that this loss is measured by the expected Kullback–Leibler divergence between the model and the nearest one, we have to identify, for each model M_j, $j = 1, 2, 3$, the model M_i, $j \neq i$ that is nearer.

Let us first consider the Weibull model M_1. The $\log P(M_1)$ is proportional to the minimum value between

$$
\left\{ \mathbb{E}_{\pi_1}\left[\inf_{\mu,\tau} D_{KL}(f_1(x|\lambda,\kappa)\|f_2(x|\mu,\tau))\right],\ \mathbb{E}_{\pi_1}\left[\inf_{\alpha,\beta} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))\right]\right\}, \tag{14}
$$

where the expectations are taken with respect to the prior $\pi_1(\lambda,\kappa)$. From Section 3.2, we know

that the value of the first element in (14) is 0.09, as the minimum divergence from a Weibull density to a log-normal density does not depend on $\pi_1(\lambda, \kappa)$. To compute the expected minimum divergence from model M_2 to model M_3, we proceed as seen in Section 3.2. First, we determine the Kullback–Leibler divergence from M_1 to M_3, as shown in Theorem 3 in the supporting information, which gives

$$\begin{aligned} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta)) &= \int_0^\infty f_1(x|\lambda,\kappa)\log\left\{\frac{f_1(x|\lambda,\kappa)}{f_3(x|\alpha,\beta)}\right\}dx \\ &= \log\kappa + \kappa\,\mathbb{E}_1(\log x) - \kappa\log\lambda - \frac{1}{\lambda^\kappa}\mathbb{E}_1(x^\kappa) - \alpha\log\beta + \log\Gamma(\alpha) \\ &\quad -\alpha\,\mathbb{E}_1(\log x) + \beta\,\mathbb{E}_1(x). \end{aligned} \tag{15}$$

Where, again, the expectations are taken with respect to model $f_1(x|\lambda, \kappa)$. The infimum of (15), with respect to the parameter α and β of the gamma density, is found by solving the following system of equations

$$\begin{cases} \mathbb{E}_1(\log x) = \Psi(\alpha) - \log\beta \\ \mathbb{E}_1(x) = \alpha/\beta, \end{cases} \tag{16}$$

from which we see that the two densities are nearer when they have equal expectation for x and $\log x$ (refer to Theorem 3 in the supporting information). In fact, if a random variable has a gamma distribution with shape parameter α and rate parameter β, its expectation is α/β and the expectation of its logarithm is $\Psi(\alpha) - \log\beta$; where $\Psi(\alpha) = d\left\{\log\Gamma(\alpha)\right\}/d\alpha$ is the digamma function. System (16) is solved with numerical methods, and the minimum divergence between a Weibull and a gamma has the form

$$\begin{aligned} \inf_{\alpha,\beta} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta)) &= \log\kappa - \gamma - 1 - \alpha\log\beta + \log\Gamma(\alpha) - \alpha\log\lambda + \alpha\frac{\gamma}{\kappa} \\ &\quad +\beta\lambda\Gamma\left(1+\frac{1}{\kappa}\right), \end{aligned}$$

where we have considered that, if x has a Weibull distribution with parameters λ and κ, then $\mathbb{E}_1(x) = \Gamma(1/\kappa)\lambda/\kappa$. In this illustration we assume that the parameters of the Weibull are independent. Therefore, the prior $\pi_1(\lambda, \kappa)$ is the product of the marginal prior assigned on each parameter, which have been chosen to be identical and, in particular, gamma distributed with shape parameter equal to 25 and rate parameter equal to 1. That is, distributions with relatively large variance. With this prior, we have obtained $\mathbb{E}_{\pi_1}\left\{\inf_{\alpha,\beta} D_{KL}(f_1(x|\lambda,\kappa)\|f_3(x|\alpha,\beta))\right\} = 0.05$. Thus, given that this result gives a smaller expected divergence in comparison to the one mea-

sure to the log-normal (as computed in Section 3.2), the mass to be assigned to model M_1 is $P(M_1) \propto \exp(0.05) = 1.06$.

It is legitimate to wonder if it is possible, by selecting a different prior $\pi_1(\lambda, \kappa)$, to define a Weibull density which is nearer to the log-normal than to the gamma. For example, if we chose the gamma distributions for λ and κ with the rate parameter equal to 2, the expected minimum divergence would have value 0.14, and the prior mass for M_1 would be based on the expected minimum divergence with respect to the log-normal density.

To determine the prior probability for model M_2, we need to identify the minimum between

$$\left\{ \mathbb{E}_{\pi_2} \left[\inf_{\lambda,\kappa} D_{KL}(f_2(x|\mu,\tau) \| f_1(x|\lambda,\kappa)) \right], \; \mathbb{E}_{\pi_2} \left[\inf_{\alpha,\beta} D_{KL}(f_2(x|\mu,\tau) \| f_3(x|\alpha,\beta)) \right] \right\}.$$

Where, in this case, the expectations are taken with respect to the prior $\pi_2(\mu, \tau)$. In Section 3.2 we have shown that the first term does not depend on the parameters of the log-normal ad has value 0.08. The Kullback–Leibler divergence between M_2 and M_3 is (refer to Theorem 4 in the supporting information)

$$\begin{aligned} D_{KL}(f_2(x|\mu,\tau) \| f_3(x|\alpha,\beta)) &= \int_0^\infty f_2(x|\mu,\tau) \log \left\{ \frac{f_2(x|\mu,\tau)}{f_3(x|\alpha,\beta)} \right\} dx \\ &= \frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{1}{2}\tau\, \mathbb{E}_2(\log^2 x) + \tau\mu\, \mathbb{E}_2(\log x) - \frac{1}{2}\tau\mu^2 \\ &\quad -\alpha \log\beta + \log\Gamma(\alpha) - \alpha\, \mathbb{E}_2(\log x) + \beta\, \mathbb{E}_2(x). \end{aligned} \tag{17}$$

In this case, the expectations are with respect to model $f_2(x|\mu, \tau)$. The minimum of (17), with respect to α and β, is attained when simultaneously $\mathbb{E}_2(x) = \alpha/\beta$ and $\mathbb{E}_2(\log x) = \Psi(\alpha) - \log\beta$. That is, when the two densities have equal mean and equal expectation of the logarithm of x. Note that this result is analogous to the one obtained when we have determined the minimum divergence from M_1 to M_3. To compute the expected minimum Kullback–Leibler divergence between the log-normal density and the gamma density, for coherence, we have again assumed the parameters as independent. Therefore, $\pi_2(\mu, \tau)$ is given by the product of the two marginals. For the location parameter we have set a log-normal prior with location parameter zero and precision 1/10, and for the parameter τ, the prior is a gamma with parameters 25 and 1. We have obtained $\mathbb{E}\left\{\inf_{\alpha,\beta} D_{KL}(f_2(x|\mu,\tau) \| f_3(x|\alpha,\beta))\right\} = 0.06$. With this prior, the mass for model M_2 is determined on the basis of its distance to the gamma density, and it is $P(M_2) \propto \exp(0.06) = 1.06$.

We note that, by increasing the uncertainty around the parameters, this mass increases as well. For example, by setting the rate parameter of the prior for τ to 1/4, we would have an expected minimum divergence of 0.09. In this case, the prior probability for the log-normal would be based on the distance with respect to the Weibull.

For the prior probability of model M_3, we need to compare

$$\left\{ \mathbb{E}_{\pi_3} \left[\inf_{\lambda,\kappa} D_{KL}(f_3(x|\alpha,\beta) \| f_1(x|\lambda,\kappa)) \right], \; \mathbb{E}_{\pi_3} \left[\inf_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta) \| f_2(x|\mu,\tau)) \right] \right\}.$$

Obviously, the expectation are now taken with respect to the prior for the parameters of model $f_3(x|\alpha,\beta)$. First, we see that the divergence from model M_3 to model M_1 is given by

$$\begin{aligned} D_{KL}(f_3(x|\alpha,\beta) \| f_1(x|\lambda,\kappa)) &= \int_0^\infty f_3(x|\alpha,\beta) \log \left\{ \frac{f_3(x|\alpha,\beta)}{f_1(x|\lambda,\kappa)} \right\} dx \\ &= \alpha \log \beta - \log \Gamma(\alpha) + \alpha \, \mathbb{E}_3(\log x) - \beta \, \mathbb{E}_3(x) - \log \kappa - \kappa \, \mathbb{E}_3(\log x) \\ &\quad + \kappa \log \lambda + \frac{1}{\lambda^\kappa} \mathbb{E}_3(x^\kappa), \end{aligned} \tag{18}$$

where the expectations are with respect to model $f_3(x|\alpha,\beta)$, with $\mathbb{E}_3(x) = \alpha/\beta$, $\mathbb{E}_3(\log x) = \Psi(\alpha) - \log \beta$ and $\mathbb{E}_3(x^\kappa) = \beta^{-\kappa}\Gamma(\kappa+\alpha)/\Gamma(\alpha)$ (refer to Theorem 2 in the supporting information). The infimum of (18), with respect to λ and κ, is found by solving

$$\begin{cases} \mathbb{E}_3(x^\kappa) = \lambda^\kappa \\ \Psi(\kappa+\alpha) - 1/\kappa = \Psi(\alpha). \end{cases} \tag{19}$$

Solving system (19), with numerical methods, the minimum Kullback–Leibler divergence between the gamma density and the Weibull density has the following expression

$$\begin{aligned} \inf_{\lambda,\kappa} D_{KL}(f_3(x|\alpha,\beta) \| f_1(x|\lambda,\kappa)) &= -\log \Gamma(\alpha) + \alpha\Psi(\alpha) - \alpha - \log \kappa - \kappa\Psi(\alpha) + \kappa \log \beta \\ &\quad + \kappa \log \lambda + 1. \end{aligned}$$

Assuming α and β independent, prior $\pi_3(\alpha,\beta)$ can be set as the product of two gamma distributions. For coherence with previous decisions, we have chosen both gamma with shape parameter equal to 25, in order to have relatively high variance, thus relatively high uncertainty about the parameter values. The expected minimum divergence is $\mathbb{E}_{\pi_3}\{\inf_{\lambda,\kappa} D_{KL}(f_3(x|\alpha,\beta) \| f_1(x|\lambda,\kappa))\} = 0.02$.

To assess $\mathbb{E}_{\pi_3}\{\inf_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta) \| f_2(x|\mu,\tau))\}$, we consider the Kullback–Leibler divergence between the two models (refer to Theorem 4 in the supporting information)

$$
\begin{aligned}
D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau)) &= \int_0^\infty f_3(x|\alpha,\beta)\log\left\{\frac{f_3(x|\alpha,\beta)}{f_2(x|\mu,\tau)}\right\}dx \\
&= \alpha\log\beta-\log\Gamma(\alpha)+\alpha\,\mathbb{E}_3(\log x)-\beta\,\mathbb{E}_3(x)-\frac{1}{2}\log\tau+\frac{1}{2}\log(2\pi) \\
&\quad +\frac{1}{2}\tau\,\mathbb{E}_3(\log^2 x)-\tau\mu\,\mathbb{E}_3(\log x)+\frac{1}{2}\tau\mu^2. \qquad (20)
\end{aligned}
$$

The divergence in (20) is minimised with respect to μ and τ when, simultaneously, $\mu=\mathbb{E}_3(\log x)=\Psi(\alpha)-\log\beta$ and $\tau=1/Var(\log x)=1/\Psi'(\alpha)$, with $\Psi'(\alpha)=d\left\{\Psi(\alpha)\right\}/d\alpha$ the trigamma function. We note that the two models are at their nearest distance when expectation and variance (of the logarithm) are equal. The expression of this minimum divergence is

$$
\begin{aligned}
\inf_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau)) &= -\log\Gamma(\alpha)+\alpha\Psi(\alpha)-\alpha+\frac{1}{2}\log\Psi'(\alpha) \\
&\quad +\frac{1}{2}\log 2\pi+\frac{1}{2}.
\end{aligned}
$$

We used the same prior we have used to compute the expected minimum divergence between M_3 and M_1. The result is $\mathbb{E}_{\pi_3}\left\{\inf_{\mu,\tau} D_{KL}(f_3(x|\alpha,\beta)\|f_2(x|\mu,\tau))\right\}=0.06$. Therefore, the prior mass for M_3 is based on the "distance" from the gamma to the Weibull, and has value $P(M_3)\propto\exp(0.02)=1.02$. We note that, in this case, the expected minimum divergence depends only on the value of the shape parameter.

TABLE 1 HERE

Table 1 summarises the expected minimum divergences among models M_1, M_2 and M_3 and, as previously computed, the appropriate prior mass. The normalised prior for this particular model selection problem, and given the selected priors for the parameter of the models, are $P_N(M_1)=0.34$, $P_N(M_2)=0.33$ and $P_N(M_3)=0.33$. Even though is not possible to make a direct comparison among the level of uncertainty that we have expressed for the parameters of each model (via the appropriate prior distributions), we note that, by keeping variances relatively large, the prior mass is basically uniform.

4 Nested Models

Let us now consider the case where models are nested. The simplest scenario is when we have only two models, and where we can identify an inner (or simple) model and an outer (or complex)

model. Logic dictates that, if we remove the smaller model when it is the true, there would be no loss, for the inner is a special case of the outer (unless some additional loss is placed on the larger model simply because it is more complex). Therefore, the prior mass to be assigned to the inner model will be proportional to one (associated with a loss of zero). The mass for the outer model will be determined with a procedure analogous to the one we have repeatedly examined in Section 3; that is, by computing the expected minimum Kullback–Leibler divergence with respect to the inner model. Consider the following example.

***Example* 1.** Let us assume that we want to select between a standard normal density and a normal density with the same precision but with the mean that is allowed to be different from zero. That is

$$M_1 = \left\{ f(x|0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \right\} \text{ and } M_2 = \left\{ f(x|\mu,1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2\right], \pi(\mu) \right\}.$$

The general expression of the Kullback–Leibler divergence between two normal densities with different means and precisions, say $f(x|\mu_1,\tau_1) = N(\mu_1,\tau_1)$ and $f(x|\mu_2,\tau_2) = N(\mu_2,\tau_2)$, is given by

$$\begin{aligned} D_{KL}(f(x|\mu_1,\tau_1)\|f(x|\mu_2,\tau_2)) &= \int_{-\infty}^{\infty} f(x|\mu_1,\tau_1) \log\left\{\frac{f(x|\mu_1,\tau_1)}{f(x|\mu_2,\tau_2)}\right\} dx \\ &= \frac{\tau_2}{2}(\mu_1-\mu_2)^2 + \frac{1}{2}\left(\frac{\tau_2}{\tau_1} - 1 - \log\frac{\tau_2}{\tau_1}\right). \end{aligned} \tag{21}$$

To assign a mass to M_1, we have to find the infimum of $D_{KL}(f(x|0,1)\|f(x|\mu,1))$ which, considering (21), is attained for $\mu = 0$, resulting in a divergence equal to zero. As such, $P(M_1) \propto 1$. For M_2, we note that $D_{KL}(f(x|\mu,1)\|f(x|0,1)) = \mu^2/2$, which is also the minimum, given μ. Therefore, the minimum expected divergence, with respect to the prior $\pi(\mu)$, is given by

$$\begin{aligned} \int D_{KL}\Big(f(x|\mu,1)\|f(x|0,1)\Big)\pi(\mu)\, d\mu &= \int \frac{\mu^2}{2}\pi(\mu)\, d\mu \\ &\propto \mathbb{E}(\mu^2). \end{aligned}$$

Thus, taking $\mathbb{E}(\mu) = 0$, we have $\mathbb{E}(\mu^2) = Var(\mu)$. Hence, $P(M_2) \propto \exp\{Var(\mu)\}$. First, we note that the mass associated with the simpler model is proportional to one. Second, the mass on the more complex model is related to the variance of the prior distribution for μ. If $Var(\mu) = 0$ (i.e. we put a point mass at $\mu = 0$), we have that $P(M_1) = P(M_2) = 1/2$, as it should be. On the other hand, if $Var(\mu) \to \infty$, $P(M_2)$ increases, as we believe more and more that model M_1 is wrong.

In particular, the larger the variance (i.e. the more uncertainty about the parameter we have), the larger the mass associated to the larger model. Furthermore, our approach allows us to "explain" the *Jeffreys-Lindley paradox*, by assigning a model prior that depends on the model, namely $\{f(x|\mu, 1), \pi(\mu)\}$. The paradox arises when the posterior of the null model, M_1 in our example above, converges to one when $Var(\mu)$ tends to infinity; therefore, in terms of hypothesis testing, the point null-hypothesis will always be accepted. According to our approach, the prior on model M_2 is proportional to the variance of μ, which allows to avoid the "paradoxical" inconvenient of a posterior on M_1 that tends to one when the above variance tends to infinity. Robert (1993), although through a different approach, derives a solution that makes the prior on the models function of the variance of μ. For other discussions on the paradox refer to, for example, Shafer (1982), Bernardo (1999), and Dellaportas *et al.* (2012).

To generalise, let us assume that we have to select between the following two nested models,

$$M_1 = \{f(\cdot|\theta_1), \pi_1(\theta_1)\} \quad \text{and} \quad M_2 = \{f(\cdot|\theta_1, \theta_2), \pi_1(\theta_1)\pi_2(\theta_2|\theta_1)\}\,,$$

with $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, and where the prior distributions for the parameters are known. The fact that model M_1 is nested into model M_2, implies that $D_{KL}(f(\cdot|\theta_1)\|f(\cdot|\theta_1, \theta_2))$ is minimised, with respect to the pair (θ_1, θ_2), when θ_2 degenerates to a fixed value. As such, $P(M_1) \propto 1$.

The prior mass to be put on model M_2, following our approach, will be found in the following way. First, if we assume that $D_{KL}(f(\cdot|\theta_1, \theta_2)\|f(\cdot|\theta))$ attains its minimum at $\theta = \theta_1$, we note that it is not necessary to identify the minimum Kullback–Leibler divergence from model M_2 to model M_1, as parameter θ_1 would have the same value for both models. Thus, the mass to assign to M_2 is given by

$$P(M_2) \propto \exp\left\{\int\int D_{KL}(f(\cdot|\theta_1, \theta_2)\|f(\cdot|\theta_1))\pi_2(\theta_2|\theta_1)\pi_1(\theta_1)\, d\theta_2 d\theta_1\right\}.$$

This result can be further generalised if we consider a set of models nested one into each other. In this case, it becomes obvious that the mass assigned to each model, except for the largest one (i.e. the most complex), will be proportional to one. Furthermore, the only mass that has to be actually computed is the one to be assigned to the largest model.

From this result, we note that when a model is nested into another one, say M_1 is nested in M_2, the prior mass on the simpler model will never be larger than the prior mass on the more complex model. That is, $P(M_2) \geq P(M_1)$. The complex model expresses a more detailed representation of the phenomenon than the simple model. Therefore, in general, it has to be $P(M_2) > P(M_1)$, and we would have $P(M_2) = P(M_1) = 1/2$ if and only if M_1 and M_2 are the same model.

Let us now see an example on how the general approach is applied to the selection of two nested models of the same family. This example is a generalisation the previous one.

***Example* 2.** Let us assume that we are interested in selecting between a normal density with mean μ and precision one, and a normal density with the same mean parameter and precision τ. The models are

$$M_1 = \left\{ f(x|\mu,1) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x-\mu)^2 \right\}, \pi_1(\mu) \right\},$$
$$M_2 = \left\{ f(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{ -\frac{\tau}{2}(x-\mu)^2 \right\}, \pi_2(\mu,\tau) \right\}.$$

Applying (21), we have that $D_{KL}(f(x|\mu,1)\|f(x|\mu,\tau)) = \tau(\mu-\theta)^2/2 + (\tau - 1 - \log\tau)/2$, where the mean in M_2 has been rewritten as θ in order to distinguish it from the mean of model M_1. By differentiating with respect to θ and τ, we find that the minimum is attained when $\theta = \mu$ and $\tau = 1$. And, as expected, the value of the divergence at these points is zero. Thus, $P(M_1) \propto 1$. The prior mass for M_2 is based on the divergence $D_{KL}(f(x|\mu,\tau)\|f(x|\mu,1)) = \tau(\theta-\mu)^2/2 + (1/\tau - 1 + \log\tau)/2$. We can see that this is minimised, with respect to μ, when the two means are equal, and the value is $\inf_\mu D_{KL}(f(x|\mu,\tau)\|f(x|\mu,1)) = (1/\tau - 1 + \log\tau)$. Therefore, we have

$$P(M_2) \propto \exp\left\{ \int\int \frac{1}{2}\left(\frac{1}{\tau} - 1 + \log\tau \right) \pi_2(\mu,\tau)\, d\mu d\tau \right\}. \tag{22}$$

Similarly, as seen in Example 1, we note that the further π_2 is from a point mass at one, the larger $P(M_2)$ becomes. This is shown by the fact that $(1/\tau + \log\tau)$ in (22) is minimised at $\tau = 1$. And this expresses the idea that the more uncertain we are about the simpler model being the true one, the more mass we assign to the more complex model. As an illustration, we consider the prior for τ to be a gamma distribution with shape parameter 5 and rate parameter 1. We then obtain that $P(M_2) \propto \exp(0.38) = 1.46$. With this result, the normalised prior mass is $P_N(M_1) = 0.41$ and $P_N(M_2) = 0.59$. It is of course possible, by changing the prior π_2, to obtain a different prior mass for M_1 and M_2.

Again, we note from Example 2 that, when we consider nested models, the *worth* of the larger model is, at least, as large as the *worth* of the inner model.

4.1 Normal and Student's t

The first illustration for nested models not belonging to the same family of distributions, considers a normal density and a Student's t density. The normal and the Student's t can be found, for example, as alternative models in financial applications, such as modelling logarithmic financial returns (Fabozzi *et al.*, 2010). We then consider model M_1 to be a normal distribution with mean μ and precision τ, and model M_2 to be a Student's t distribution with location parameter θ, precision parameter λ and parameter ν representing the number of degrees of freedom. That is

$$M_1 = \left\{ f_1(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{ -\frac{\tau}{2}(x-\mu)^2 \right\}, \pi_1(\mu,\tau) \right\},$$
$$M_2 = \left\{ f_2(x|\theta,\lambda,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\lambda}{\nu\pi}\right)^{1/2} \left\{ 1 + \frac{\lambda}{\nu}(x-\theta)^2 \right\}^{-\frac{\nu+1}{2}}, \pi_2(\theta,\lambda,\nu) \right\}.$$

The Student's t distribution converges to a normal distribution when the number of degrees of freedom tends to infinity (for example, Chu (1956)). As such, the two models can be considered nested models which differ from the number of degrees of freedom only (for example, Casellas *et al.* (2008)). Therefore, as discussed above, we have that the infimum of the Kullback–Leibler divergence between M_1 and M_2 is zero, resulting in a prior mass on the normal model $P(M_1) \propto 1$.

To determine the mass for M_2, following our approach, we consider that, as shown in Theorem 5 in the supporting information, we have

$$\begin{aligned} D_{KL}(f_2(x|\theta,\lambda,\nu) \| f_1(x|\mu,\tau)) &= \int_{-\infty}^{\infty} f_2(x|\theta,\lambda,\nu) \log\left\{ \frac{f_2(x|\theta,\lambda,\nu)}{f_1(x|\mu,\tau)} \right\} dx \\ &= \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{1}{2}\log\lambda - \frac{1}{2}\log\nu \\ &\quad -\frac{\nu+1}{2}\mathbb{E}_2\left\{ \log\left(1 + \frac{\lambda}{\nu}(x-\theta)^2\right) \right\} - \frac{1}{2}\log\tau + \frac{1}{2}\log 2 \\ &\quad +\frac{1}{2}\tau\,\mathbb{E}_2(x^2) - \tau\mu\,\mathbb{E}_2(x) + \frac{1}{2}\tau\mu^2. \end{aligned} \tag{23}$$

The divergence in (23) is minimised, with respect to μ and τ, when $\mu = \mathbb{E}_2(x) = \theta$ and $\tau = 1/Var(x) = \lambda$. That is, when the two distributions have location parameter and scale parameter of the same value. The minimum divergence is then

$$\begin{aligned}\inf_{\mu,\tau} D_{KL}(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau)) &= \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu+1}{2} \\ &\quad \mathbb{E}_2\left\{\log\left(1+\frac{\lambda}{\nu}(x-\theta)^2\right)\right\} + \frac{1}{2}\log\nu - \frac{1}{2}\log(\nu-2) + \frac{1}{2}.\end{aligned}$$

To compute the prior mass for M_2, we consider the following prior distribution $\pi_2(\theta,\lambda,\nu) = \pi_{2,1}(\nu)\pi_{2,2}(\lambda)\pi_{2,3}(\theta|\lambda)$. Where $\pi_{2,1}(\nu)$ is an exponential distribution (Geweke, 1993) with rate parameter equal to 1, $\pi_{2,2}(\lambda)$ is a gamma with shape parameter 25 and rate parameter 1, and $\pi_{2,3}(\theta|\lambda)$ is a normal distribution with mean zero and precision determined by the prior on λ. Thus

$$\begin{aligned}P(M_2) &\propto \exp\left\{\int\int\int \inf_{\mu,\tau} D_{KL}\Big(f_2(x|\theta,\lambda,\nu)\|f_1(x|\mu,\tau)\Big)\pi_2(\theta,\lambda,nu)\, d\theta\lambda d\nu\right\} \\ &= \exp(0.23) \\ &= 1.26,\end{aligned}$$

where the result has been obtained through numerical methods. By normalising, we have $P_N(M_1) = 0.44$ and $P_N(M_2) = 0.56$, which shows that more mass is given to the outer model. This is in line with the idea that, in relation to the other model, M_2 has more *worth*.

4.2 Nested and non-nested models

In this final illustration, we consider a realistic model selection problem where the model space has both nested and non-nested elements, and a total of four models. We do this by adding an exponential model to the selection scenario analysed in Section 3.3. That is, M_4 is an exponential density with rate parameter θ

$$M_4 = \left\{f_4(x|\theta) = \theta e^{-\theta x}, \pi_4(\theta)\right\}.$$

To identify the prior mass for model M_1, in addition to the results in Section 3.3, we need to consider the expected minimum Kullback–Leibler divergence with respect to the exponential density. This is given by (refer to Theorem 6 in the supporting information)

$$\begin{aligned} D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta)) &= \int_0^\infty f_1(x|\lambda,\kappa)\log\left\{\frac{f_1(x|\lambda,\kappa)}{f_4(x|\theta)}\right\}\,dx \\ &= \log\kappa+\kappa\,\mathbb{E}_1(\log x)-\mathbb{E}_1(\log x)-\kappa\log\lambda-\frac{1}{\lambda^\kappa}\mathbb{E}_1(x^\kappa)-\log\theta \\ &\quad +\theta\,\mathbb{E}_1(x), \end{aligned}$$

which is minimised for $\theta = 1/\mathbb{E}_1(x) = \lambda^{-1}\Gamma(1+1/\kappa)^{-1}$. As expected, the two densities have minimum distance when the respective first moments are equal. Then $\inf_\theta D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta)) = \log\kappa-\gamma+\gamma/\kappa+\log\Gamma(1+1/\kappa)$. We note that the minimum Kullback–Leibler divergence between the Weibull and the exponential densities does not depend on the scale parameter λ. To compute the expected minimum divergence, we have adopted the same prior distributions for the parameter of the Weibull we have used in Section 3.3. The result is $\mathbb{E}_{\pi_1}\{\inf_\theta D_{KL}(f_1(x|\lambda,\kappa)\|f_4(x|\theta))\} = 0.05$. Given that this is the smallest expected divergence for model M_1 (refer to Table 2), we have $P(M_1) \propto (0.05) = 1.06$.

With a similar process, we find the Kullback–Leibler divergence between model M_2 (log-normal) and the model M_4 (refer to Theorem 6 in the supporting information)

$$\begin{aligned} D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta)) &= \int_0^\infty f_2(x|\mu,\tau)\log\left\{\frac{f_2(x|\mu,\tau)}{f_4(x|\theta)}\right\}\,dx \\ &= -\mathbb{E}_2(\log x)+\frac{1}{2}\log\tau-\frac{1}{2}\log(2\pi)-\frac{1}{2}\tau\,\mathbb{E}_2(\log^2 x)+\tau\mu\,\mathbb{E}_2(\log x) \\ &\quad -\frac{1}{2}\tau\mu^2-\log\theta+\theta\,\mathbb{E}_2(x), \end{aligned}$$

which is minimised for $\theta = 1/\mathbb{E}_2(x) = 1/\exp\{\mu+1/(2\tau)\}$, as expected. The minimum divergence is $\inf_\theta D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta)) = \{\log\tau-\log(2\pi)+1+\tau\}/2$, which does not depend on the location parameter μ of the log-normal density. With the same priors for μ and τ considered in Section 3.3, we have obtained $\mathbb{E}_{\pi_2}\{\inf_\theta D_{KL}(f_2(x|\mu,\tau)\|f_4(x|\theta))\} = 0.05$. As the smallest expected divergence for model M_2 remains the one with respect to the gamma density (refer to Table 2), we have $P(M_2) \propto \exp(0.03) = 1.03$.

For the gamma model M_3, we have (refer to Theorem 6 in the supporting information)

$$\begin{aligned} D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta)) &= \int_0^\infty f_3(x|\alpha,\beta)\log\left\{\frac{f_3(x|\alpha,\beta)}{f_4(x|\theta)}\right\}\,dx \\ &= \alpha\log\beta-\log\Gamma(\alpha)+\alpha\,\mathbb{E}_3(\log x)-\mathbb{E}_3(\log x)-\beta\,\mathbb{E}_3(x)-\log\theta \\ &\quad +\theta\,\mathbb{E}_3(x), \end{aligned}$$

which is minimised by $\theta = 1/\mathbb{E}_3(x) = \beta/\alpha$. Therefore, we obtain the minimum divergence as $\inf_\theta D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta)) = -\log\Gamma(\alpha)+\alpha\Psi(\alpha)-\Psi(\alpha)-\alpha+\log\alpha+1$. The expected minimum divergence has been computed using the same priors for α and β defined in Section 3.3, obtaining $\mathbb{E}_{\pi_3}\{\inf_\theta D_{KL}(f_3(x|\alpha,\beta)\|f_4(x|\theta))\} = 0.05$. In this case as well, the divergence with respect to the exponential distribution does not constitute the minimum (refer to Table 2), so we have $P(M_3) \propto \exp(0.02) = 1.02$.

To compute the prior mass for model M_4, we note that, being the exponential nested into the Weibull and the gamma models (it is in fact a special case of these two densities), we obviously have $D_{KL}(f_4(x|\theta)\|f_1(x|\lambda,\kappa)) = 0$ and $D_{KL}(f_4(x|\theta)\|f_3(x|\alpha,\beta)) = 0$. Therefore, we can conclude that $P(M_4) \propto 1$. However, for completion, we have computed the expected minimum Kullback–Leibler divergence with respect to the log-normal density, and found $\mathbb{E}_{\pi_4}\{\inf_\theta D_{KL}(f_4(x|\theta)\|f_2(x|\mu,\tau))\} = 0.41$ (refer to Theorem 7 in the supporting information); where the expectation has been computed with respect to the priors for μ and τ defined in Section 3.3.

TABLE 2 HERE

Table 2 summarised the results for this particular selection problem. We note that all the normalised prior probabilities are close to 1/4. Given that we have kept the prior uncertainty about the parameter of the models at a relatively high level, the result is sensible. However, as we have already discussed in the previous illustrations, a change in the informational content within the prior distribution on the parameters will cause, in general, a different prior over the model space.

Another interesting consideration is that, by inspecting Table 2, we note that the expected Kullback–Leibler divergence between models M_1, M_2 and M_3 and model M_4 is constant and it is equal to 0.05. Recalling the results in this section, we have that the minimum divergences (in these three particular cases) depend only on the shape parameter of the models, respectively κ, τ and α. As we have used identical prior distributions for these parameters, the result obtained is sensible in the light of the prior information considered.

5 Discussion

The approach we propose aims to assign prior mass on models on the basis of their *worth*. We evaluate the *worth* by thinking about what is lost if we remove a model and it is correct. By doing this for all models, we obtain an objective value of each of their worth, which is then linked to the prior mass via the *self-information* loss function.

An important aspect of our approach is that the prior on a model depends on the model itself, in particular on the prior for the parameter of the model. This is evident in Example 1, where we compare the normal density $N(\mu, 1)$ to the normal density $N(0, 1)$. It is well known that assigning fixed probability to the two models (such as assigning 1/2 to both of them) may result in the so called Jeffreys-Lindley paradox. Our approach solves the paradox by assigning more mass to the model $N(\mu, 1)$ as the uncertainty on μ increases. Hence, the prior on a model must/should depend on the prior on the model parameter.

The proposed method to derive model prior masses can be applied to any selection problem. In fact, we have seen examples with discrete and continuous models; models with one parameter and models with more than one parameter. Particular results have been obtained in the presence of nested models: if model M_1 is contained in model M_2, then $P(M_2) \geq P(M_1)$. The result is not surprising, as the more complex model is (at least) as good as the simpler one, and there is no loss in removing the simpler model. Unless an additional loss is placed on model M_2 for model complexity/dimension.

Our approach can be applied to a certain specific category of variable selection problems. Suppose we have a variable of interest y, which outcome depends on potentially p covariates $(x_1, \ldots, x_p)$. In general, a variable selection problem would consider all the possible regression models where y is explained by *any* combination of the p covariates. To assign prior probabilities on this model space, our approach requires further considerations which are not discussed in this paper. However, we can consider the case where the optional models are formed by adding to the context model (i.e. the simplest model) one variable at a time, from x_1 to x_p, progressively. In this case, we would be in the presence of $p+1$ nested models, creating a scenario which is treatable with our approach.

To elaborate. Suppose that the normal linear model is defined to relate y to the potential covariates. That is

$$y \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

where $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression parameters, and σ^2 the constant error variance. Thus, the generic model can be represented as $M_j = \left\{\sum_{l=0}^{j} \beta_l x_{lm} + \varepsilon_m, \pi(\beta_0, \ldots, \beta_j)\right\}$, for $j = 0, \ldots, p$. On the basis of what has been discussed in Section 4, we assign prior mass proportional to one to each model from M_0 to M_{p-1}. Whilst the mass assigned to model M_p would be determined by the expected minimum Kullback–Leibler divergence from M_p to M_{p-1}. In this and similar contexts, we can add a penalty to the prior on the model space to take into account model complexity. The idea fits in our framework naturally, and results in the following

model prior

$$P(M_j) \propto \exp\left\{ \int_{\Theta_j} \left[\inf_{\theta_m, m \neq j} D_{KL}(f_j(\cdot|\theta_j) \| f_m(\cdot|\theta_m)) \pi_j(\theta_j)\, d\theta_j \right] - c \cdot |M_j| \right\}$$

where $c > 0$ and $|M_j|$ is the dimension of model M_j. However, we do not discuss this more in the current paper.

Acknowledgments

The authors would like to thank two reviewers and an Associate Editor for comments and criticism on earlier versions of the paper.

Supporting information Additional supporting information may be found in the online version of this article at the publisher's website.

References

Bayarri, M. J., Berger, J. O., Forte, A. & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40**, 1550–1577

Berger, J. O. & Pericchi, L. R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *IMS Lecture Notes - Monograph Series* **38**, 135–193

Berk, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. of Math. Statist.* **37**, 51–58

Bernardo, J. M. (1999). Nested hypothesis testing: the Bayesian reference criterion. *Bayesian Statist.* **6**, 101–130

Carvalho, C. & Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512

Casella, G. & Moreno, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.* **101**, 157–167

Casellas, J., Ibáñez-Escriche, N. & García-Cortéz, L. A. (2008). Bayes Factor Between Student t and Gaussian Mixed Models within an Animal Breeding Context. *Genet. Sel. Evol.* **40**, 395–413

Chipman, H., George, E. I. & McCulloch, R. E. (2001). The Practical Implementation of Bayesian Model Selection. *IMS Lecture Notes - Monograph Series* **38**, 65–116

Chu, J. T.1956 Errors in normal Approximation to the y, τ and Similar Types of Distributions. *Ann. of Math. Statist.* **27**, 780–789

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2012). Joint Specification of model space and parameter space prior distributions. *Statist. Sci.* **27**, 232–246

Dumonceaux, R. & Antle, C. E. (1973). Discrimination Between the log-normal and the Weibull Distributions. *Technometrics* **15**, 923–926

Dumonceaux, R., Antle, C. E. & Haas, G. (1973). Likelihood Ratio Test for Discrimination Between Two Models with Unknown Location and Scale Parameters. *Technometrics* **15**, 19–31

Fabozzi, F. J., Focardi, S. M., Höchstötter, M. & Rachev, S. T. (2010). *Probability and Statistics for Finance*. Wiley

Geweke, J. (1993). Bayesian Treatment of the Independent Student-t Linear Model. *J. App. Econometrics*. **8**, S19–S40

Johnson, N. L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*. Houghton Mifflin, Boston

Klein, J. P. & Moeschberger, M. L. (1997). *Survival Analysis*. Springer

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. of Math. Statist.* **22**, 79–86

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192

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

O'Hagan, A. (1995) Fractional Bayes Factors for Model Comparison. *J. R. Stat. Soc. B Stat. Methodol.* **57**, 99–138

Pérez, J. M. & Berger, J. O. (2002). Expected-Posterior Prior Distributions for Model Selection. *Biometrika* **89**, 491–511

Pericchi, L. R. (2005) Model selection and hypothesis testing based on objective probabilities and Bayes factors. In D. K. Dey & C. R. Rao, eds., *Handbook of Statistics, Bayesian Thinking and Computation.* Holland, 115–149

Robert, C. P. (1993). A Note on Jeffreys-Lindley Paradox. *Statist. Sinica* **3**, 601–608

Robert, C. P. (2001). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*. Springer

Scott, J. G. & Berger, J. O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Ann. Statist.* **38**, 2587–2619

Shafer, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* **77**, 325–334

Stracham, R. W. & van Dijk, H. K. (2003). Bayesian Model Selection with an Uninformative Prior. *Oxford Bulletin of Economics and Statistics* **65**, 863–876

Villa C. & Walker, S. G. (2014). Objective Prior for the Number of Degrees of Freedom of a t Distribution. *Bayesian Anal.* **9**, 197–220

Villa C. & Walker, S. G. (2014). An Objective Approach to Prior Mass Functions for Discrete Parameter Spaces. To appear in *J. Amer. Statist. Assoc.*

Cristiano Villa, School of Mathematics, Statistics and Actuarial Science, Cornwallis Building, University of Kent, Canterbury, Kent CT2 7NF, UK. email: cv88@kent.ac.uk

Table 1: Expected minimum Kullback-Leibler divergence (by column) among the models M_1 (Weibull), M_2 (log-normal) and M_3 (gamma). The divergences have been computed on the basis of the prior distributions on the parameters of the models as specified in Section 3.3. The mass is proportional to the exponential of the minimum divergence, and the last two rows show this mass for each model: non-normalised $P(M_j)$ and normalised $P_N(M_j)$, $j = 1, 2, 3$.

	M_1	M_2	M_3
M_1		0.08	0.06
M_2	0.09		0.02
M_3	0.05	0.03	
$P(M_j)$	1.06	1.03	1.02
$P_N(M_j)$	0.34	0.33	0.33

Table 2: Expected minimum Kullback–Leibler divergence (by column) among the models M_1 (Weibull), M_2 (log-normal), M_3 (gamma) and M_4 (exponential). The divergences are computed considering the priors for the parameter of the models as defined in Section 3.3 and Section 4.1. The prior mass is proportional to the exponential of the minimum divergence, and the last two rows report this mass for each model, non-normalised $P(M_j)$ and normalised $P_N(M_j)$, $j = 1, 2, 3, 4$.

	M_1	M_2	M_3	M_4
M_1		0.08	0.06	0.00
M_2	0.09		0.02	0.41
M_3	0.05	0.03		0.00
M_4	0.05	0.05	0.05	
$P(M_j)$	1.05	1.03	1.02	1.00
$P_N(M_j)$	0.26	0.25	0.25	0.24