ISBA 2000, Proceedings, pp. 000–000

Compatible Prior Distributions

A. PHILIP DAWID and STEFFEN L. LAURITZEN
University College London, UK and
Aalborg University, Denmark

SUMMARY

We investigate two approaches to constructing compatible prior laws over alternative models: 'projection' and 'conditioning'. Each of these is shown to require additional inputs. We suggest that these can be chosen in a natural way in each case, leading to 'Kullback-Leibler projection' and 'Jeffreys conditioning'. We recommend the former for the case of coexisting models, and the latter for competing models.

Keywords: BAYES FACTOR; COMPATIBLE PRIORS; CONDITIONING; JEFFREYS CONDITIONING; KULLBACK-LEIBLER PROJECTION; MARGINALISATION; MODEL AVERAGING; MODEL SELECTION; PROJECTION.

1. INTRODUCTION

Suppose we wish to compare two models, $\mathcal{M}$ and $\mathcal{M}_0$, for the same observable X. Let $\mathcal{M}$ be parametrised by θ, with model densities $f(x \mid \theta)$, and $\mathcal{M}_0$ by ϕ, with model densities $f_0(x \mid \phi)$. Within each model, we have a prior density, $\pi(\theta)$ and $\pi_0(\phi)$ respectively, representing uncertainty about its parameter conditional on the model. If we observe $X = x$, the impact of this on model uncertainty can be isolated in the corresponding $Bayes\ factor$:

$$B(\mathcal{M} : \mathcal{M}_0) := \frac{\int_{\mathcal{M}} f(x \mid \theta)\, \pi(\theta)\, d\theta}{\int_{\mathcal{M}_0} f_0(x \mid \phi)\, \pi_0(\phi)\, d\phi}.$$

In general there is no compelling reason to relate the prior densities over different models, even when one is a submodel of another. However, the task of assessing a prior density for a model parameter is difficult enough when we only have a single model to deal with, and we might hope that, having once conducted this exercise, we can use the resulting distribution to assist in assigning an appropriate prior over a different model. Furthermore, it is known that Bayes factors can be quite sensitive to the specific choice of priors. It would lend some degree of objectivity if the priors over the different models were chosen to be, in some sense still to be made precise, as similar as possible — we shall then call them $compatible$. In this case we can argue that the Bayes factor is truly responding to the data, rather than merely reflecting prior prejudices.

Our purpose in this paper is to investigate possible explications of this informal concept of 'compatibility'. For simplicity, we restrict attention to the case that $\mathcal{M}_0$ is a lower dimensional submodel of $\mathcal{M}$, and examine two methods that have commonly been used

in this case: 'projection' and 'conditioning'. We point out that neither method is uniquely defined, each depending on additional inputs. We suggest that there are natural choices for these, leading to 'Kullback-Leibler projection' and 'Jeffreys conditioning'. And we make tentative recommendations as to when each of these methods might be appropriate.

2. EXAMPLES

We focus attention on two simple examples, each involving a random sample of n observations on a pair of outcome variables. In the first example the outcome variables have a joint Gaussian distribution, while in the second they are binary. Our notation for multivariate and matrix-variate distributions follows Dawid (1981).

Example 1 [Two Gaussian variables] Our data arise as n independent copies of a pair of continuous variables $X = (X_0, X_1)$. Under model $\mathcal{M}$, X is assumed to follow a bivariate Gaussian distribution: $X \sim \mathcal{N}_2(0, \Sigma)$, with arbitrary dispersion matrix

$$\Sigma = \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{pmatrix}.$$

The standard conjugate prior for Σ in this model is the inverse Wishart distribution: $\Sigma \sim \mathcal{IW}(\delta; \Phi)$. That is, the concentration matrix $K := \Sigma^{-1}$ is assumed to have the Wishart distribution, $K \sim \mathcal{W}_2(\delta + 1; \Phi^{-1})$, with density, with respect to Lebesgue measure on the set of positive definite matrices:

$$f(K) \propto \{\det(K)\}^{\frac{1}{2}\delta - 1} \exp\{-\operatorname{tr}(\Phi K)/2\}. \tag{1}$$

We take as $\mathcal{M}_0$ the model under which X_0 and X_1 are independent, each still being normal with zero mean. This model may be parametrised by (τ_0, τ_1), where $\tau_i := \operatorname{var}(X_i)$. It can also be described as the submodel of $\mathcal{M}$ under which Σ satisfies the restriction $\sigma_{01} = 0$, or, equivalently, K satisfies $k_{01} = 0$.

The question is: Which prior distribution for (τ_0, τ_1) in $\mathcal{M}_0$ 'corresponds naturally' to the inverse Wishart distribution $\mathcal{IW}(\delta; \Phi)$ specified for Σ in $\mathcal{M}$? As we shall see, this question can be answered in a variety of ways. □

Example 2 [Two binary variables] Consider now n independent copies of a pair of binary variables I and J. Under model $\mathcal{M}$, the associated probabilities

$$\theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix},$$

where $\theta_{ij} := P(I = i, J = j)$, are arbitrary, subject only to

$$\theta_{ij} > 0 \text{ and } \sum \theta_{ij} = 1. \tag{2}$$

The standard conjugate prior for θ in this model is the Dirichlet distribution $\mathcal{D}(\alpha)$, where

$$\alpha = \begin{pmatrix} \alpha_{00} & \alpha_{01} \\ \alpha_{10} & \alpha_{11} \end{pmatrix}.$$

Then the prior density of θ (with respect to Lebesgue measure on the simplex (2)) is

$$f(\theta) \propto \prod_{ij} \theta_{ij}^{\alpha_{ij}-1}.$$

We again take as $\mathcal{M}_0$ the model in which I and J are independent. We can parametrise this by $\phi := P(I = 0)$ and $\psi := P(J = 0)$. We can also describe $\mathcal{M}_0$ as the submodel of $\mathcal{M}$ for which θ satisfies the restriction: for all i and j, $\theta_{ij} = \theta_{i+}\,\theta_{+j}$, where $\theta_{i+} := \sum_j \theta_{ij}$ and $\theta_{+j} := \sum_i \theta_{ij}$.

Again the question is: Which prior distribution for (ϕ, ψ) in $\mathcal{M}_0$ 'corresponds naturally' to the Dirichlet distribution $\mathcal{D}(\alpha)$ specified for θ in $\mathcal{M}$? And again, this question can be answered in a variety of ways. □

3. PROJECTION

A statistical model parameter can be regarded, either as an abstract label, identifying one out of many possible distributions in the model; or, in some cases at least, more concretely as a direct measure of some aspect of that distribution: for example, its mean, or correlation. If we take the latter approach then, since $\mathcal{M}_0$ is a submodel of $\mathcal{M}$, we might continue to describe any distribution in $\mathcal{M}_0$ by the same parameter θ we used for $\mathcal{M}$. In fact, because $\mathcal{M}$ has lower dimension, only a subvector θ_0 of θ will typically be required. We can then use the original prior distribution for θ to induce a marginal distribution for θ_0, thus supplying a 'compatible' prior for $\mathcal{M}_0$.

The following example demonstrates that this seemingly 'natural' construction is in fact non-unique, being dependent on exactly how we choose to interpret our parametrisation.

Example 3 In Example 1, there are two equally 'obvious' ways of identifying the parameters of model $\mathcal{M}_0$ with (a subset of) those of model $\mathcal{M}$. The first sets

$$\tau_i = \sigma_{ii}, \quad i = 0, 1$$

while the second takes

$$\tau_i = 1/k_{ii}, \quad i = 0, 1.$$

That is, in the first case we identify the variances in the two models, whereas in the second case we identify the concentrations.

In the first case, using the induced joint distribution for $(\sigma_{00}, \sigma_{11})$ under the $\mathcal{IW}(\delta; \Phi)$ prior for Σ leads to

$$\tau_i \sim \phi_{ii}/\chi^2_\delta$$

whereas in the second case we obtain

$$\tau_0 \sim \phi_{00.1}/\chi^2_{\delta+1}, \quad \tau_1 \sim \phi_{11.0}/\chi^2_{\delta+1},$$

where

$$\phi_{00.1} := \phi_{00} - \phi^2_{01}/\phi_{11}, \quad \phi_{11.0} := \phi_{11} - \phi^2_{01}/\phi_{00}.$$

In either case τ_0 and τ_1 are not independent, their joint distribution being complicated. We note in passing that the compatible hyper inverse Wishart priors suggested in Dawid and Lauritzen (1993) would have the same marginal distributions as in the first case, but in addition τ_0 and τ_1 would be independent. □

A more abstract way of thinking about the above approach is that, explicitly or implicitly, we have to specify a way of associating, with each distribution P in $\mathcal{M}$, a corresponding distribution P_0 in $\mathcal{M}_0$. Thus in the first method suggested in Example 3, P_0 is the (unique) distribution in $\mathcal{M}_0$ having the same values for var(X_0) and var(X_1) as does P; in the second method, P_0 has the same value for var$(X_1 \mid X_0)$ and var$(X_0 \mid X_1)$. Such a specification can be described by an appropriate mapping $r : \mathcal{M} \to \mathcal{M}_0$, such that $P_0 = r(P)$; we can think of r as a generalised 'projection' function. Given r, the distribution of $r(\tilde{P})$, when $\tilde{P}$ varies according to its assigned prior law in $\mathcal{M}$, will supply the 'compatible' prior law for $\tilde{P}_0$ over $\mathcal{M}_0$. Here we are using the term "law" to indicate a "distribution for a distribution". However, as there is an almost limitless choice for the function r, the above 'projection method' does not, without additional input, yield a unique answer, as Example 3 makes clear.

3.1. *Kullback–Leibler projection*

One approach to defining a projection map r is in terms of a 'discrepancy function' $D(P, Q)$ between distributions P and Q for X. We could then define the 'minimum discrepancy' projection map onto $\mathcal{M}_0$ by:

$$r(P) := \arg \min_{Q \in \mathcal{M}_0} D(P, Q),$$

provided a mimimiser exists and is uniquely defined. A popular discrepancy function is the *Kullback–Leibler divergence*,

$$KL(P, Q) := E_P\{\log p(X)/q(X)\},$$

where $p(\cdot)$ denotes the density of P, *etc.*, leading to the projection

$$r_{KL}(P) := \arg \min_{Q \in \mathcal{M}_0} KL(P, Q).$$

This KL-projection approach was used by McCulloch and Rossi (1992), who proposed Monte Carlo methods for calculating the associated Bayes factors. Among other advantages, KL-projection does not depend on the parametrisations of $\mathcal{M}$ and $\mathcal{M}_0$, is invariant under one-to-one transformations of the observable X, and scales by sample size for observables modelled as independent and identically distributed under both P and Q, so that, for this common case, the associated projection map does not depend on sample size.

Example 4 In the multivariate Gaussian case, the KL divergence between $P_\Sigma = \mathcal{N}(0, \Sigma)$ and $P_\Phi = \mathcal{N}(0, \Phi)$ is

$$KL(P_\Sigma, P_\Phi) = \frac{1}{2}\left\{\mathrm{tr}(\Sigma\Phi^{-1} - I) - \log\det(\Sigma\Phi^{-1})\right\},$$

and for Φ restricted to being diagonal this is minimised when $\phi_{ii} = \sigma_{ii}$. That is, using KL-projection onto the model of complete independence identifies variances rather than concentrations, thus resolving the ambiguity seen in Example 3. □

Example 5 In Example 2, the KL-projection method relates the parameters (ϕ, ψ) of $\mathcal{M}_0$ to those of $\mathcal{M}$ through

$$\phi = \theta_{0+}, \quad \psi = \theta_{+0}.$$

Used with the Dirichlet prior $\mathcal{D}(\alpha)$ for $\mathcal{M}_0$, this leads to

$$\phi \sim \beta(\alpha_{0+}, \alpha_{1+}), \quad \psi \sim \beta(\alpha_{+0}, \alpha_{+1}).$$

Once again, however, ϕ and ψ are not independent in this induced distribution, in contrast to the hyper Markov specifications of Dawid and Lauritzen (1993), who suggest these marginal distributions for ϕ and ψ under model $\mathcal{M}_0$, but in addition require their independence. □

4. CONDITIONING

We now consider a different general approach to constructing compatible priors.

The submodel $\mathcal{M}_0$ can typically be derived from $\mathcal{M}$ by imposing a constraint on its parameter θ: say $\eta = \eta_0$, for an appropriate parameter $\eta := \eta(\theta)$. It might then appear 'natural' to derive the compatible prior distribution in $\mathcal{M}_0$ by simply conditioning that for θ in $\mathcal{M}$ on $\eta(\theta) = \eta_0$. But, once again, this proposal turns out to be subject to ambiguity: there is typically a variety of choices for the function η, each leading to a different answer. This phenomenon is sometimes termed the *Borel-Kolmogorov paradox*.

Example 6 Consider Example 1 from the point of view of conditioning. The constraint $X \perp\!\!\!\perp Y$ can be expressed in any of the following ways, among others:

(i). $k_{01} = 0$

(ii). $\sigma_{01} = 0$

(iii). $\beta_{1.0} = 0$, where $\beta_{1.0} := \sigma_{01}/\sigma_{00}$ is the *regression coefficient* of X_1 on X_0

(iv). $\rho = 0$, where $\rho := \sigma_{01}/\{\sigma_{00}\,\sigma_{11}\}^{\frac{1}{2}}$ is the *coefficient of correlation* between X_0 and X_1.

We shall see that we obtain different answers, depending on which of these constraints we condition on.

(i). The joint prior density $\pi(k_{00}, k_{01}, k_{11})$ is given by (1). To condition on $k_{01} = 0$, we substitute this value and renormalise. We obtain:

$$\pi_0(k_{00}, k_{11}) \propto k_{00}^{\frac{1}{2}\delta - 1} \exp\left(-\phi_{00}\, k_{00}/2\right) \times k_{11}^{\frac{1}{2}\delta - 1} \exp\left(-\phi_{11}\, k_{11}/2\right),$$

i.e., in terms of $\tau_0 = k_{00}^{-1}$, $\tau_1 = k_{11}^{-1}$,

$$\tau_0 \sim \phi_{00}/\chi^2_\delta, \quad \tau_1 \sim \phi_{11}/\chi^2_\delta, \quad \text{independently.}$$

(ii). Making a change of variables in (1), we find that the joint prior density $\pi(\sigma_{00}, \sigma_{01}, \sigma_{11})$ is proportional to

$$(\det \Sigma)^{-\frac{1}{2}\delta-2} \exp\left\{-\operatorname{tr}\left(\Phi\Sigma^{-1}\right)/2\right\}. \tag{3}$$

Restricting this to $\sigma_{01} = 0$ yields the conditional joint density:

$$\pi_0(\sigma_{00}, \sigma_{11}) \propto \sigma_{00}^{-\frac{1}{2}\delta-2} \exp\left(-\phi_{00}\,\sigma_{00}^{-1}/2\right) \times \sigma_{11}^{-\frac{1}{2}\delta-2} \exp\left(-\phi_{11}\,\sigma_{11}^{-1}/2\right),$$

i.e., in terms of $\tau_0 = \sigma_{00}$, $\tau_1 = \sigma_{11}$,

$$\tau_0 \sim \phi_{00}/\chi^2_{\delta+2}, \quad \tau_1 \sim \phi_{11}/\chi^2_{\delta+2}, \quad \text{independently.}$$

(iii). To condition on $\beta_{1.0} = 0$, we note (Dawid, 1988, Lemma 2) that, under the specified $\mathcal{IW}(\delta; \Phi)$ prior distribution for Σ, we have $\sigma_{00} \perp\!\!\!\perp (k_{11}, \beta_{1.0})$, with joint distribution given by:

$$\begin{aligned} \sigma_{00} &\sim \phi_{00}/\chi^2_{\delta} \\ k_{11} &\sim \phi_{11.0}^{-1}\,\chi^2_{\delta+1} \\ \beta_{1.0}\,|\,k_{11} &\sim \mathcal{N}\left\{b_{1.0}, (\phi_{00}\,k_{11})^{-1}\right\}, \end{aligned}$$

where $b_{1.0} := \phi_{01}/\phi_{00}$. Consequently, conditioning on $\beta_{1.0} = 0$, we still have $\sigma_{00} \sim \phi_{00}/\chi^2_{\delta}$, independently of k_{11}; while the conditional distribution of k_{11} is readily found to be $\phi_{11}^{-1}\,\chi^2_{\delta+2}$. Thus, in terms of $\tau_0 = \sigma_{00}, \tau_1 = k_{11}^{-1}$, we have:

$$\tau_0 \sim \phi_{00}/\chi^2_{\delta}, \quad \tau_1 \sim \phi_{11}/\chi^2_{\delta+2}, \quad \text{independently.}$$

(iv). Using $k_{01} = -\rho\,(k_{00}\,k_{11})^{\frac{1}{2}}$ with (1) to evaluate $\pi(k_{00}, k_{11}, \rho)$, we now find that, conditional on $\rho = 0$,

$$\tau_0 \sim \phi_{00}/\chi^2_{\delta+1}, \quad \tau_1 \sim \phi_{11}/\chi^2_{\delta+1}, \quad \text{independently.}$$

□

Example 7 Similar instances of the Borel-Kolmogorov paradox arise in Example 2. Introducing $\kappa := \theta_{0\,|\,0}, \quad \lambda := \theta_{0\,|\,1}$, the joint distribution of (I, J) can be parametrised by (ψ, κ, λ) where $\psi = P(J = 0) = \theta_{+0}$ as earlier. Note that, under the constraint of independence between I and J imposed by $\mathcal{M}_0$, we have $\kappa = \lambda = \phi = P(I = 0)$. We now define:

$$\begin{aligned} \eta_1 &:= \kappa - \lambda \\ \eta_2 &:= \frac{\kappa}{\lambda} \\ \eta_3 &:= \frac{\kappa\,(1-\lambda)}{(1-\kappa)\,\lambda} \equiv \frac{\theta_{00}\,\theta_{11}}{\theta_{10}\,\theta_{01}}. \end{aligned}$$

These parameters are familiar quantities in epidemiology: η_1 is the *excess risk* (for having $I = 0$, due to having $J = 0$), η_2 is the *risk ratio*, and η_3 is the *odds ratio*. The logarithm of η_3 is the *interaction* between I and J.

The independence constraint of $\mathcal{M}_0$ can be expressed in any of the forms:

(i). $\eta_1 = 0$

(ii). $\eta_2 = 1$

(iii). $\eta_3 = 1$.

Again we start with the prior distribution $\theta \sim \mathcal{D}(\alpha)$ in $\mathcal{M}$, and condition on the chosen independence constraint. We find that, conditionally on any of (i), (ii) or (iii), ϕ and ψ are independent, with $\psi \sim \beta(\alpha_{+0}, \alpha_{+1})$; but ϕ has different distributions in the three cases:

(i). $\phi \sim \beta(\alpha_{0+} - 1, \alpha_{1+} - 1)$,

(ii). $\phi \sim \beta(\alpha_{0+}, \alpha_{1+} - 1)$,

(iii). $\phi \sim \beta(\alpha_{0+}, \alpha_{1+})$.

□

4.1. *Jeffreys conditioning*

The fact that the method of conditioning depends on the particular parameter-function η used to define the submodel $\mathcal{M}_0$ suggests the need for an alternative formulation of the method.

One possible generalisation proceeds by choosing *reference measures*, ν and ν_0, on $\mathcal{M}$ and $\mathcal{M}_0$ respectively. Given a prior law Π in $\mathcal{M}$, we can form its density function with respect to ν: $\pi(\cdot) := d\Pi/d\nu$. This is a scalar function, with a value for each $P \in \mathcal{M}$ (we suppose that there is sufficient smoothness in the problem that we are able to define this function everywhere, rather than merely almost everywhere). We can then construct a law Π_0 in $\mathcal{M}_0$ by requiring that its density with respect to ν_0, $\pi_0(\cdot) := d\Pi_0/d\nu_0$, which is defined for $P \in \mathcal{M}_0$, should be proportional to π there. That is, using the given measures ν and ν_0 as the basis for our construction of densities, we are merely restricting the density on $\mathcal{M}$ to the submodel $\mathcal{M}_0$, and then renormalising.

In the special case that ν is a probability measure, and ν_0 is obtained from ν by conditioning on a constraint $\eta = \eta_0$, this 'density restriction' method will give the same answer as simply conditioning Π on $\eta = \eta_0$. The arbitrariness in the form of the function η is now replaced by the arbitrariness in the choice of the measures ν and ν_0. However, this reinterpretation opens up new possibilities for resolving this ambiguity: by choosing measures ν and ν_0 that are in some way intrinsic to the models, and independent of specific ways of parametrising them.

One such intrinsic measure, for a general smooth model $\mathcal{M}$, is the *Jeffreys measure* $J = J_{\mathcal{M}}$, which, in any smooth parametrisation, has density $j = j_{\mathcal{M}}$ with respect to Lebesgue measure given by:

$$j(\theta) := \{\det I(\theta)\}^{1/2},$$

where $I(\theta)$ is the Fisher information matrix, having (r, s)-entry

$$\{I(\theta)\}_{rs} := -E\frac{\partial^2 \log f(X \mid \theta)}{\partial\theta_r\,\partial\theta_s}.$$

The Jeffreys measure is the Riemannian uniform measure when $\mathcal{M}$ is equipped with its Riemannian information geometry (Amari *et al.* , 1987; Kass and Vos, 1997) and is thus invariant under reparametrisation.

The Jeffreys measure is often used as a (typically improper) prior distribution representing ignorance about a parameter. Our usage here is entirely different: as a base measure for defining the density of a proper prior. We may term a density with respect to the Jeffreys measure *invariantised*. If π denotes the density of Π with respect to Lebesgue measure, the invariantised density ρ is given by:

$$\rho(\theta) = \pi(\theta)/j(\theta) = \pi(\theta)\{\det I(\theta)\}^{-1/2}.$$

When the method of density restriction is applied using the invariantised densities, formed with respect to the respective Jeffreys measures $\nu = J_{\mathcal{M}}$, $\nu_0 = J_{\mathcal{M}_0}$, we shall refer to the procedure as *Jeffreys conditioning*. Then the Jeffreys conditioning of π, with invariantised density $\rho_0 \propto \rho$, has density with respect to Lebesgue measure

$$\pi_0(\theta) \propto \rho(\theta) j_0(\theta) = \rho(\theta)\{\det I_0(\theta)\}^{1/2} = \pi(\theta) \left\{ \frac{\det I_0(\theta)}{\det I(\theta)} \right\}^{1/2}.$$

Example 8 In the bivariate Gaussian case of Example 1, we have, for Models $\mathcal{M}$ and $\mathcal{M}_0$ respectively,

$$\begin{aligned} j(\Sigma) &\propto (\det \Sigma)^{-\frac{3}{2}} \\ j_0(\tau_0, \tau_1) &\propto \frac{1}{\tau_0}\frac{1}{\tau_1}. \end{aligned}$$

The invariantised form of (3) is thus

$$\rho(\Sigma) \propto (\det \Sigma)^{-\frac{1}{2}(\delta+1)} \exp\left(-\Phi\Sigma^{-1}/2\right),$$

and restricting to $\mathcal{M}_0$ (and identifying $\tau_i = \sigma_{ii}$) yields

$$\rho_0(\tau_0, \tau_1) \propto \tau_0^{-\frac{1}{2}(\delta+1)} e^{-\frac{1}{2}\phi_{00}/\tau_0} \times \tau_1^{-\frac{1}{2}(\delta+1)} e^{-\frac{1}{2}\phi_{11}/\tau_1}.$$

The corresponding Lebesgue density is then

$$\pi_0(\tau_0, \tau_1) \propto \tau_0^{-\frac{1}{2}(\delta+1)-1} e^{-\frac{1}{2}\phi_{00}/\tau_0} \times \tau_1^{-\frac{1}{2}(\delta+1)-1} e^{-\frac{1}{2}\phi_{11}/\tau_1},$$

i.e. $\tau_0 \sim \phi_{00}/\chi^2_{\delta+1}$, $\tau_1 \sim \phi_{11}/\chi^2_{\delta+1}$, independently. In this case, Jeffreys conditioning yields the same answer as simple conditioning on the constraint $\rho = 0$. □

Example 9 Consider again the 2×2 table of Example 2. Using the notation of Example 7, the joint probability function of (I, J) is

$$p(i, j \mid \psi, \kappa, \lambda) = \psi^{1-j}(1-\psi)^{j}\, \kappa^{(1-i)(1-j)}(1-\kappa)^{i(1-j)}\, \lambda^{(1-i)j}(1-\lambda)^{ij}.$$

The density j of the Jeffreys measure J in $\mathcal{M}$ is found to be

$$\begin{aligned} j(\psi, \kappa, \lambda) &= \left\{ \frac{1}{\psi(1-\psi)} \frac{\psi}{\kappa(1-\kappa)} \frac{(1-\psi)}{\lambda(1-\lambda)} \right\}^{1/2} \\ &= \{\kappa(1-\kappa)\,\lambda(1-\lambda)\}^{-1/2}, \end{aligned}$$

leading to the invariantised prior density ρ in $\mathcal{M}$:

$$\rho(\psi,\kappa,\lambda) \propto \psi^{\alpha_{+0}-1}(1-\psi)^{\alpha_{+1}-1}\,\kappa^{\alpha_{00}-1/2}(1-\kappa)^{\alpha_{10}-1/2}\,\lambda^{\alpha_{01}-1/2}(1-\lambda)^{\alpha_{11}-1/2}.$$

For $\mathcal{M}_0$ we similarly obtain

$$j_0(\psi,\phi) = \{\psi(1-\psi)\,\phi(1-\phi)\}^{-1/2}\,,$$

so that using Jeffreys conditioning we obtain for the prior density in $\mathcal{M}_0$ with respect to Lebesgue measure:

$$\begin{aligned}\pi_0(\psi,\phi) &= \rho(\psi,\phi,\phi)\,j_0(\psi,\phi)\\ &\propto \psi^{\alpha_{+0}-3/2}(1-\psi)^{\alpha_{+1}-3/2}\,\phi^{\alpha_{0+}-3/2}(1-\phi)^{\alpha_{1+}-3/2}\end{aligned}$$

i.e. $\psi \perp\!\!\!\perp \phi$, and

$$\psi \sim \beta(\alpha_{+0}-1/2,\alpha_{+1}-1/2), \quad \phi \sim \beta(\alpha_{0+}-1/2,\alpha_{1+}-1/2).$$

This is distinct from all the solutions found in Example 7. Note that the 'effective sample size' is reduced by 1 compared to conditioning with respect to the odds ratio η_3. We could regard Jeffreys conditioning as compensating for additional prior information arising from simplification of the model. □

5. CONCLUSIONS

There are many problems in which we might wish to entertain two (or more) distinct parametric models for our data. Within these, we can distinguish two somewhat different scenarios:

(i). We have *competing* models, only one of which can be true.

(ii). We have *coexisting* models, describing the same reality but at different levels of detail.

For example, $\mathcal{M}$ may describe a linear regression of weight on height and age, while $\mathcal{M}_0$ omits the regressor age. Under interpretation (i), $\mathcal{M}_0$ is valid only if, once height is known, age is of no further value in predicting height. Under interpretation (ii), we may feel that, even when further information on age might still be relevant, it could nevertheless be adequate for our purposes to use only height to predict weight. A Bayesian analysis of coexisting regression models may be found in Dawid (1988).

In case (i), since we are unsure about the true model, an appropriate Bayesian approach would be to assign probabilities to the truth of the various models (conditional on whatever data we have), and represent our opinion about reality as a mixture over the models, with these probabilities. This is the method of 'model averaging'. However, in case (ii) all models are true, and we merely wish to work with the most useful, *i.e.* it would be more appropriate to undertake 'model selection'.

Although further investigations are required, it is tempting to recommend the method of *Kullback-Leibler projection* to define model compatibility in the case of co-existing models, and *Jeffreys conditioning* in the case of competing models.

ACKNOWLEDGEMENTS

Steffen Lauritzen's work has received support from ESPRIT project 29105 (BaKE).

REFERENCES

Amari, S.-I., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., and Rao, C. R. (1987). *Differential Geometry in Statistical Inference*. Hayward, CA: IMS

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–74.

Dawid, A. P. (1988). The infinite regress and its conjugate analysis (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 95–110.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. New York: Wiley

McCulloch, R. E. and Rossi, P. E. (1992). Bayes factors for non-linear hypotheses and likelihood distributions. *Biometrika* **79**, 663–76.