Test (2008) 17: 585–605
DOI 10.1007/s11749-007-0057-7

ORIGINAL PAPER

Compatible priors for Bayesian model comparison with an application to the Hardy–Weinberg equilibrium model

Guido Consonni · Eduardo Gutiérrez-Peña · Piero Veronese

Received: 22 July 2005 / Accepted: 2 March 2007 /
Published online: 12 May 2007

Abstract Suppose we entertain Bayesian inference under a collection of models. This requires assigning a corresponding collection of prior distributions, one for each model's parameter space. In this paper we address the issue of relating priors across models, and provide both a conceptual and a pragmatic justification for this task. Specifically, we consider the notion of "compatible" priors across models, and discuss and compare several strategies to construct such distributions. To explicate the issues involved, we refer to a specific problem, namely, testing the Hardy–Weinberg Equilibrium model, for which we provide a detailed analysis using Bayes factors.

Keywords Bayes factor · Conjugate prior · Dirichlet distribution · Jeffreys conditioning · Kullback–Leibler projection · Nested model · Predictive distribution · Trinomial distribution

Mathematics Subject Classification (2000) 62F15 · 62P10 · 92D25

G.C.'s research was partially supported by MIUR, Rome, (PRIN 2003138887) and the University of Pavia. P.V.'s research was partially supported by MIUR, Rome, (PRIN 2003138887) and by L. Bocconi University. The research of E.G. was partially supported by CONACyT (Grant 32256-E) and SNI, Mexico. The authors also benefited from partial support through Project No. 4 of the Bilateral Executive Programme for Technological and Scientific Co-operation Between Italy and Mexico 2003–2005.

G. Consonni
Università di Pavia, Pavia, Italy

E. Gutiérrez-Peña (✉)
IIMAS, Universidad Nacional Autónoma de México, Apartado Postal 20-726,
01000 Mexico D.F., Mexico
e-mail: eduardo@stats.iimas.unam.mx

P. Veronese
Università L. Bocconi, Milano, Italy

1 Introduction

Suppose we wish to make inference, using a Bayesian approach, under a collection of models for the same observable. If the models are not nested, a separate prior distribution on the parameter space of each model is typically required.

On the other hand, when models are nested within a unique encompassing model $\mathcal{M}$, it appears natural to perform inference using the prior assigned on the parameter $\theta \in \Theta$ under $\mathcal{M}$, since all models under investigation are obtained through a suitable restriction of Θ. This idea has proved to be especially fruitful in the framework of model choice/comparison. For instance, Goutis and Robert (1998) and Bernardo and Rueda (2002) use the expectation, relative to the posterior distribution of θ, of a measure of divergence between a model and a submodel in order to assess the validity of model simplification. Specifically, the latter paper uses a decision-theoretic approach to model choice, based on the concept of intrinsic discrepancy, and the corresponding reference prior, which only depends on the structure of the model.

When model comparison is performed through the Bayes factor, a specific prior under each submodel is still required. If each prior is derived from that on θ under $\mathcal{M}$, we achieve some "compatibility" of prior distributions across models (thus alleviating the sensitivity of the Bayes factor to prior specification), and we reduce the burden of the elicitation procedure, which can be especially heavy when the collection of models is large, see McCulloch and Rossi (1992), Dawid and Lauritzen (2001), and Roverato and Consonni (2004). A related course of action, based on an objective approach, is pursued in Casella and Moreno (2006).

Despite these efforts, the notion of compatible priors is still elusive, and it is difficult, when confronted with a practical problem, to offer firm guidance on how to proceed. This paper is a step in this direction. Specifically, we consider the problem of testing the Hardy–Weinberg equilibrium model of population genetics and offer a careful discussion of prior specifications together with the resulting Bayes factors.

The structure of the paper is as follows. In the next section we review and discuss several strategies for the construction of compatible prior distributions, focusing, in particular, on priors obtained via Kullback–Leibler (KL) projections. In Sect. 3 we describe the Hardy–Weinberg equilibrium model of population genetics. Section 4 discusses prior specifications for testing the Hardy–Weinberg equilibrium model with an application to a data set previously analysed in the literature. Section 5 contains a simulation study comparing the Bayes factors for the various compatible priors. The results show that the KL-projection priors perform well relative to other choices of compatible priors discussed in this paper. Finally, in Sect. 6 we present some concluding remarks.

2 Strategies to construct compatible priors

As already mentioned in the introduction, the issue of compatibility of prior distributions across models has been relatively neglected. We present here a brief account of the available strategies; see also Dawid and Lauritzen (2001).

Consider a model $\mathcal{M}$, and a submodel $\mathcal{M}_0$ thereof, for a (vector-valued) observation X. In a parametric setting this means that if $\mathcal{M} = \{p(\cdot|\theta) : \theta \in \Theta\}$, where

$p(\cdot|\theta)$ is a density for X, and $\Theta \subseteq \mathbb{R}^d$, then $\mathcal{M}_0 = \{p(\cdot|\theta) : \theta \in \tilde{\Theta}_0\}$, with $\tilde{\Theta}_0 \subset \Theta$. Note that $\tilde{\Theta}_0$ lives in $\mathbb{R}^d$ although its dimension is typically lower than d. Henceforth we shall assume $\dim(\tilde{\Theta}_0) = d_0 < d$, so that $\tilde{\Theta}_0$ is isomorphic to a space Θ_0 in $\mathbb{R}^{d_0}$. Accordingly, we shall write $\mathcal{M}_0 = \{p_0(\cdot|\theta_0) : \theta_0 \in \Theta_0\}$ with $\Theta_0 \subseteq \mathbb{R}^{d_0}$.

As an illustration consider the following simple example.

Example 1 Take $\mathcal{M}$ to be a bivariate normal model with mean $\mu = (\mu_1, \mu_2) \in \Theta = \mathbb{R}^2$ and known covariance matrix $\Sigma = (\sigma_{ij})$, and assume that, under $\mathcal{M}_0$, $\mu_1 = \mu_2 \equiv \mu_0$. Then $\tilde{\Theta}_0 = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_1 = \mu_2\}$, while $\Theta_0 = \{\mu_0 : \mu_0 \in \mathbb{R}\}$.

Let π_θ denote a prior density over Θ and consider the problem of assigning a "compatible" prior π_{θ_0} on Θ_0. We now discuss some possible strategies.

2.1 Kullback–Leibler projection prior

Quite often in statistical modelling the parameter θ and the corresponding space Θ have a concrete meaning, and θ is not merely a label indexing a distribution in $\mathcal{M}$. This is the case in Example 1 above, wherein Θ is the mean space under $\mathcal{M}$, and similarly for Θ_0 under $\mathcal{M}_0$. A way to relate $\mathcal{M}$ and $\mathcal{M}_0$ is through a projection map $\tau : \Theta \mapsto \Theta_0$. Given a prior law π_θ on Θ, the prior induced on $\tau(\theta)$ represents a natural choice for a compatible prior on Θ_0, which we name the τ-projection prior and denote by $\pi^\tau_{\theta_0}$.

Notice that this procedure is not well defined when π_θ is improper. This happens because the dimension of Θ_0 is lower than that of Θ, and in order to compute $\pi^\tau_{\theta_0}$ an integration with respect to π_θ is required: the latter, however, diverges, and thus no meaningful result can be derived. This feature can be regarded as a difficulty when an objective Bayesian analysis is looked for, since so-called objective priors are typically improper. One possible approach is the following: if the objective prior π_θ can be interpreted as the limit of a sequence of proper priors $\{\pi_{\theta,h}(\cdot), h = 1, 2, \ldots\}$, a natural suggestion is to compute the τ-projection prior as

$$\pi^\tau_{\theta_0}(\cdot) = \lim_{h\to\infty} \pi^\tau_{\theta_0,h}(\cdot),$$

where the limit is defined appropriately, for instance, in the Kullback–Leibler sense or in terms of intrinsic convergence (see Definition 2 of Bernardo and Rueda 2002). We will not explore this idea further in this paper.

The next issue we address is the choice of the map τ. The following example shows some of the difficulties involved.

Example 1 (ctd.) Suppose we reparametrise $\mathcal{M}$ by writing $\mu_1 = \mu_0$ and $\mu_2 = \mu_0 + c$, $c \in \mathbb{R}$, so that $\mathcal{M}_0$ is identified through $c = 0$ and parametrised by μ_0. One could then think of obtaining π_{θ_0} as the marginal of μ_1 under π_θ. On the other hand, one might have reparametrised $\mathcal{M}$ as $\mu_2 = \tilde{\mu}_0$ and $\mu_1 = \tilde{\mu}_0 + \tilde{c}$ so that now $\mathcal{M}_0$ is identified through $\tilde{c} = 0$ and parametrised by $\tilde{\mu}_0$, which leads to setting π_{θ_0} as the marginal of μ_2 under π_θ. On the other hand, π_θ could assign quite different priors to μ_1 and μ_2, thus leading to two distinct compatible priors on μ_0 and $\tilde{\mu}_0$. However, μ_0 and $\tilde{\mu}_0$ both represent the mean under $\mathcal{M}_0$ and, therefore, should share the same prior.

Another way of looking at the problems involved in the example above is through the notion of invariance. Let $\eta = (\mu_0, c)$ and $\gamma = (\tilde{\mu}_0, \tilde{c})$. Under the former parametrisation the projection map is $\tau_\eta(\eta) = \mu_0$, while under the latter it is $\tau_\gamma(\gamma) = \tilde{\mu}_0$. Since $\gamma = g(\eta) = (\mu_0 + c, -c)$, so that $g^{-1}(\gamma) = (\tilde{\mu}_0 + \tilde{c}, -\tilde{c})$, invariance considerations would require that $\tau_\gamma(\gamma) = \tau_\eta(g^{-1}(\gamma))$, which is violated in this case, since $\tau_\eta(g^{-1}(\gamma)) = \tilde{\mu}_0 + \tilde{c}$.

In view of the remarks above, a natural suggestion is to take $\tau(\theta)$ as the Kullback–Leibler (KL) projection of θ onto Θ_0, i.e.,

$$\tau_\theta^{\mathrm{KL}}(\theta) = \arg\min_{\theta_0 \in \Theta_0} \mathrm{KL}\big(p_0(\cdot|\theta_0)|p(\cdot|\theta)\big),$$

where

$$\mathrm{KL}(q|p) = E^p\left(\log \frac{p(X)}{q(X)}\right),$$

denotes the KL divergence between the densities p and q relative to a common dominating measure.

A very important feature of the KL-projection is invariance to reparametrisation. This means that if $\gamma = g(\theta)$ is a reparametrisation under $\mathcal{M}$, then $\tau_\gamma^{\mathrm{KL}}(\gamma) = \tau_\theta^{\mathrm{KL}}(g^{-1}(\gamma))$. This guarantees that the procedure generates a prior $\pi_{\theta_0}^{\tau}$ that does not depend on the specific parametrisation that is chosen. Henceforth, we shall omit the subscript θ and set $\theta_0^{\perp} = \tau^{\mathrm{KL}}(\theta)$.

Notice that $\mathrm{KL}(q|p)$ is not symmetric, and this feature may be unsuitable in some contexts. One way to overcome this difficulty, as suggested by a referee, is to define a symmetric version such as the intrinsic discrepancy between p and q, i.e.,

$$\delta(p,q) = \min\big\{\mathrm{KL}(q|p), \mathrm{KL}(p|q)\big\}, \tag{1}$$

see Bernardo and Rueda (2002).

One further advantage of $\delta(p,q)$ is that it is typically finite, even if the support of q is strictly contained in that of p, in which case $\mathrm{KL}(q|p)$ is infinite.

Despite the attractive properties of δ, we prefer to use $\mathrm{KL}(q|p)$ in the sequel for the following reasons: (i) in our approach p denotes the encompassing model while q represents a possible model simplification of p. Notice that the validity of p is not questioned, and thus it represents a "benchmark" relative to which all other models are evaluated. From this point of view taking expectations with respect to p, as in the directed divergence $\mathrm{KL}(q|p)$, appears a sensible procedure. Moreover, it is precisely in this context that $\mathrm{KL}(q|p)$ can be interpreted as the loss in expected utility incurred when choosing the density q to report inferences instead of the true density p if a logarithmic scoring rule is used; (ii) for regular nested models (wherein the support is independent of the parameter), p and q have the same support, so that $\mathrm{KL}(q|p)$ is well defined; (iii) the use of $\delta(p,q)$, instead of $\mathrm{KL}(q|p)$, adds complexity from an analytical viewpoint. Specifically, if we let

$$\theta_0^* = \arg\min_{\theta_0 \in \Theta_0} \mathrm{KL}\big(p(\cdot|\theta)|p_0(\cdot|\theta_0)\big)$$

then the δ-projection of θ onto Θ_0 is given by

$$\begin{aligned}\theta_0^\delta &= \arg\min_{\theta_0\in\Theta_0} \delta\big(p(\cdot|\theta), p_0(\cdot|\theta_0)\big)\\ &= \min\big[\mathrm{KL}\big(p_0\big(\cdot|\theta_0^\perp\big)|p(\cdot|\theta)\big), \mathrm{KL}\big(p(\cdot|\theta)|\, p_0\big(\cdot|\theta_0^*\big)\big)\big].\end{aligned} \tag{2}$$

In general we shall have

$$\theta_0^\delta = \begin{cases}\theta_0^\perp, & \text{if } \theta\in A,\\ \theta_0^*, & \text{if } \theta\in\bar{A},\end{cases}$$

for some $A\subseteq\Theta$.

The derivation of the δ-projection prior requires the calculation of the induced prior on θ_0^δ, which can be very hard or not possible analytically, since it implies knowledge of both $\theta_0^\perp$ and θ_0^*, together with the structure of the set A. Of course, the δ-projection prior can be obtained by means of simulation, provided one can sample θ-values from the given prior distribution π_θ. Subsequent analysis, however, can only be performed using simulated samples.

When the submodel $\mathcal{M}_0$ admits a sufficient statistic $T(X)=T$, the map τ^{KL} is given by

$$\tau^{\mathrm{KL}}(\theta) = \arg\max_{\theta_0\in\Theta_0} E_\theta\big(\log\big(p_0^T\big(T(X)|\theta_0\big)\big)\big), \tag{3}$$

where $p_0^T(\cdot|\theta_0)$ is the density of T under $\mathcal{M}_0$ and E_θ denotes expectation with respect to $p(\cdot|\theta)$. Equation (3) holds since $p_0(x|\theta_0) = p_0^T(T(x)|\theta_0)\,h(x)$, because of the factorisation theorem, so that

$$\mathrm{KL}\big(p_0(\cdot|\theta_0)|\, p(\cdot|\theta)\big) = E_\theta\left(\log\frac{p^T(T(X)|\theta)}{p_0^T(T(X)|\theta_0)}\right) + E_\theta\left(\log\frac{p(X|\theta)}{p^T(T(X)|\theta)h(X)}\right).$$

Clearly, only the first term involves θ_0 through the denominator, whence the result.

The next proposition, which holds for two arbitrary models (not necessarily nested), provides an explicit way of finding the KL-projection when one of the models belongs to an exponential family.

Proposition 1 *Consider two models $\mathcal{M}_i = \{p_i(\cdot|\theta_i),\ \theta_i\in\Theta_i\}$, $i=0,1$. Assume that $\mathcal{M}_0$ is an exponential family with natural parameter $\xi\in\Xi\subseteq\mathbb{R}^{d_0}$ and density, with respect to a suitable measure, given by*

$$p_0(x|\theta_0) = \exp\big\{\xi(\theta_0)^T T(x) - M_0\big(\xi(\theta_0)\big)\big\}, \tag{4}$$

and that $E_{\theta_1}^1(T(X))$ is finite, where $E_{\theta_i}^i$ denotes expectation with respect to the model $\mathcal{M}_i$.

Consider the KL-divergence $\mathrm{KL}(p_0(\cdot|\theta_0)\,|\,p_1(\cdot|\theta_1))$ and the corresponding KL-projection of θ_1 onto Θ_0, denoted by $\theta_0^\perp$. Then $\theta_0^\perp$ satisfies

$$E_{\theta_0^\perp}^0\big(T(X)\big) = E_{\theta_1}^1\big(T(X)\big). \tag{5}$$

Proof By (3), minimising $\mathrm{KL}(p_0(\cdot|\theta_0)\,|\,p_1(\cdot|\theta))$ with respect to θ_0 is equivalent to maximising

$$E^1_{\theta_1}\big(\log p_0(X|\theta_0)\big) = \xi(\theta_0)^T E^1_{\theta_1}\big(T(X)\big) - M_0\big(\xi(\theta_0)\big).$$

Differentiating both sides with respect to θ_0 and equating to zero, we get the equation

$$\left.\frac{\partial M_0(\xi)}{\partial \xi}\right|_{\xi=\xi(\theta_0)} = E^1_{\theta_1}\big(T(X)\big),$$

which is equivalent to

$$E^0_{\theta_0}\big(T(X)\big) = E^1_{\theta_1}\big(T(X)\big),$$

using standard exponential family theory. The solution $\theta_0^{\perp}$ of the above equation with respect to the unknown θ_0 is indeed a maximum because of the log-concavity of exponential family likelihoods. □

Example 1 (ctd.) Since $\mathcal{M}_0$ is an exponential family with sufficient statistic $T(X_1, X_2) = (X_1(\sigma_{22} - \sigma_{12}) + X_2(\sigma_{11} - \sigma_{12}))/(\sigma_{11}\sigma_{22} - \sigma_{12}^2)$, one can apply Proposition 1 with $\mathcal{M}_1 = \mathcal{M}$. From (5) one obtains

$$\frac{\mu_0^{\perp}(\sigma_{11} + \sigma_{22} - 2\sigma_{12})}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} = \frac{\mu_1(\sigma_{22} - \sigma_{12}) + \mu_2(\sigma_{11} - \sigma_{12})}{\sigma_{11}\sigma_{22} - \sigma_{12}^2},$$

whence

$$\mu_0^{\perp} = \frac{\mu_1(\sigma_{22} - \sigma_{12}) + \mu_2(\sigma_{11} - \sigma_{12})}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}$$

is the KL-projection.

Note that $\mu_0^{\perp}$ is a weighted average of the two means and reduces to $(\mu_1 + \mu_2)/2$, when $\Sigma = \sigma^2 I_2$, where I_2 is the identity matrix of order 2.

Given $\theta \sim \pi_\theta$, we define the KL-projection prior on Θ_0 as the prior on $\theta_0^{\perp}$ induced from π_θ and denote it by $\pi_{\theta_0}^{\mathrm{KL}}$. This prior was considered by McCulloch and Rossi (1992). Goutis and Robert (1998) and Dupuis and Robert (2003) use the notion of a KL-projection to perform Bayesian hypothesis testing (or model selection), although they do not resort to KL-projection priors.

Consider again Example 1 and suppose that, under $\mathcal{M}$, $\mu \sim N_2(m, V)$. Then $\mu_0^{\perp}$ is also normally distributed with the appropriate mean and variance. Typically, however, the KL-projection, let alone its distribution, is not analytically available. For this reason McCulloch and Rossi (1992) resorted to MCMC computations.

In order to obtain analytic expressions, one can approximate the KL-projection prior. In particular, if $\mathcal{M}_0$ is a Natural Exponential Family (NEF; see Kotz et al. 2000, Chap. 54) a natural class wherein to search for such an approximation is represented by a conjugate family for θ_0. One can then try to approximate $\pi_{\theta_0}^{\mathrm{KL}}$ via a specific prior in such a family, which we choose to be the prior that minimises the Kullback–Leibler divergence from $\pi_{\theta_0}^{\mathrm{KL}}$; a justification for this procedure will be provided in

Sect. 2.3. Accordingly, we denote the resulting prior by $\pi_{\theta_0}^{\text{KLCA}}$, where KLCA stands for Kullback–Leibler Conjugate Approximation. It is expedient to employ the natural parametrisation ξ for model $\mathcal{M}_0$, introduced in (4), and the corresponding standard conjugate family $\mathcal{C}_\xi(s'_0, n'_0)$ whose densities, with respect to the Lebesgue measure, are given by

$$\pi_\xi^C(\xi) = \exp\{\xi^T s'_0 - n'_0 M_0(\xi)\} h(s'_0, n'_0), \tag{6}$$

with $(s'_0, n'_0) \in \mathcal{H}$, where $\mathcal{H}$ is the interior of the set $\{(s'_0, n'_0) \in \mathbb{R}^{d_0+1} : (h(s'_0, n'_0))^{-1} < \infty\}$. Thus, we have the following.

Theorem 1 *Consider two models $\mathcal{M}$ and $\mathcal{M}_0$, parametrised by θ and θ_0, respectively, with $\mathcal{M}_0$ a submodel of $\mathcal{M}$. Assume that $\mathcal{M}_0$ is a NEF with density given in (4) and natural parameter ξ. Let π_ξ^{KL} be the prior induced on ξ by $\pi_{\theta_0}^{\text{KL}}$ and assume that $\pi_\xi^C \in \mathcal{C}_\xi(s'_0, n'_0)$. Then $\text{KL}(\pi_\xi^C \mid \pi_\xi^{\text{KL}})$ is minimised when (s'_0, n'_0) is a solution of*

$$\begin{cases} E^C(\xi) = E^{\text{KL}}(\xi), \\ E^C(M_0(\xi)) = E^{\text{KL}}(M_0(\xi)), \end{cases} \tag{7}$$

where E^C and E^{KL} denote expectations w.r.t. π_ξ^C and π_ξ^{KL}, respectively.

The value for (s'_0, n'_0) obtained in Theorem 1 identifies a unique distribution in $\mathcal{C}_\xi(s'_0, n'_0)$, labelled π_ξ^{KLCA}. Finally, $\pi_{\theta_0}^{\text{KLCA}}$ is obtained from π_ξ^{KLCA} by transformation.

Remark 1 Since the prior π_ξ^{KL} is induced from π_θ, we can compute $E^{\text{KL}}(g(\xi))$ as $E(g(\xi(\tau^{\text{KL}}(\theta))))$, where E denotes expectation w.r.t. π_θ, so that the explicit form of π_ξ^{KL} is not required.

Proof The standard conjugate prior (6) is itself an exponential family with natural parameter (s'_0, n'_0) and minimal sufficient statistic $(\xi, M_0(\xi))$. The result then follows from Proposition 1. □

Example 2 Let $X = (X_1, X_2)$ and assume that, under $\mathcal{M}$, the X_i, $i = 1, 2$, given $\mu = (\mu_1, \mu_2)$ are independent Poisson with mean μ_i, while $\mathcal{M}_0$ requires $\mu_1 = \mu_2 \equiv \mu_0$. In this case the sufficient statistic T under $\mathcal{M}_0$ is given by $X_1 + X_2$, so that using Proposition 1 one obtains $\mu_0^\perp = (\mu_1 + \mu_2)/2$.

If one wants to use the intrinsic discrepancy, instead of KL, one should first obtain the KL-projection relative to $\text{KL}(p(\cdot|\mu_1, \mu_2) \mid p_0(\cdot|\mu_0))$ which is given by $\mu_0^* = \sqrt{\mu_1 \mu_2}$, i.e., the geometric mean of μ_1 and μ_2. One can verify that $\text{KL}(p_0(\cdot|\mu_0^\perp)|p(\cdot|\mu_1, \mu_2)) \leq \text{KL}(p(\cdot|\mu_1, \mu_2)|p_0(\cdot|\mu_0^*))$, for all (μ_1, μ_2), whence

$$\mu_0^\delta = \mu_0^\perp,$$

so that in this case the use of the discrepancy δ leads exactly to the same projection as $\text{KL}(p_0|p)$.

If the μ_is are independent Gamma(α_i, β_i) with expectation α_i/β_i, the KL-projection prior on μ_0 is no longer analytically available. However, we can approximate it using Theorem 1. From (4) we have $\xi = \log(\mu_0)$ and $M_0(\xi) = 2\exp(\xi)$. Thus, the standard conjugate prior (6) is

$$\pi_\xi^C(\xi) \propto \exp\bigl\{\xi s_0' - n_0' 2\exp(\xi)\bigr\}. \tag{8}$$

Equations (7) with respect to the unknowns s_0' and n_0' become

$$\begin{aligned} &E^C(\xi) = E^{\mathrm{KL}}(\xi) \equiv E\bigl(\log\bigl((\mu_1+\mu_2)/2\bigr)\bigr), \\ &E^C\bigl(2\exp(\xi)\bigr) = E^{\mathrm{KL}}\bigl(2\exp(\xi)\bigr) \equiv E(\mu_1+\mu_2) = \frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}. \end{aligned} \tag{9}$$

Although it may be possible to find a numerical solution of (9) for given α_i, β_i, $i = 1, 2$, a simple analytic approximation may be derived in a more general way as described in the following proposition.

Proposition 2 *Consider the setting of Theorem 1, with the addition that $\mathcal{M}_0$ is a real NEF. Then an approximation to the system of equations (7) is given by*

$$\begin{cases} E^C(\mu_{0T}) = E^{\mathrm{KL}}(\mu_{0T}), \\ \mathrm{Var}^C(\mu_{0T}) = \mathrm{Var}^{\mathrm{KL}}(\mu_{0T}), \end{cases} \tag{10}$$

where $\mu_{0T} = \partial M_0(\xi)/\partial\xi$ denotes the mean parameter of $T(X)$ under $\mathcal{M}_0$.

With some abuse of notation we will still denote by π_ξ^{KLCA} the prior in $\mathcal{C}_\xi(s_0', n_0')$ with s_0' and n_0' satisfying (10).

Remark 2 The system of (10) is typically easier to solve than (7). Indeed, for regular NEFs, $E^C(\mu_{0T}) = s_0'/n_0'$. Moreover, $\mathrm{Var}^C(\mu_{0T})$ is also explicitly available when the variance function of $p_0(\cdot|\theta_0)$ is quadratic; see, for example, Morris (1982).

Proof From well known properties of exponential families, the function $\mu_{0T} = M_0'(\xi)$ is invertible, so we can write $\xi = \xi(\mu_{0T})$. The first equation in (7) becomes $E^C(\xi(\mu_{0T})) = E^{\mathrm{KL}}(\xi(\mu_{0T}))$ and, using a first order Taylor series approximation about the corresponding expectations of μ_{0T}, we obtain $\xi(E^C(\mu_{0T})) = \xi(E^{\mathrm{KL}}(\mu_{0T}))$, i.e., $E^C(\mu_{0T}) = E^{\mathrm{KL}}(\mu_{0T})$. Using a second order approximation for the expectations involved in the second equation in (7), we obtain

$$\begin{aligned} &E^C\bigl(M_0\bigl(\xi(\mu_{0T})\bigr)\bigr) \\ &\quad = M_0\bigl(\xi\bigl(E^C(\mu_{0T})\bigr)\bigr) + \frac{1}{2}\frac{\partial^2 M_0(\xi(\mu_{0T}))}{\partial\mu_{0T}^2}\bigg|_{\mu_{0T}=E^C(\mu_{0T})} \mathrm{Var}^C(\mu_{0T}) \end{aligned}$$

and

$$\begin{aligned} &E^{\mathrm{KL}}\bigl(M_0\bigl(\xi(\mu_{0T})\bigr)\bigr) \\ &\quad = M_0\bigl(\xi\bigl(E^{\mathrm{KL}}(\mu_{0T})\bigr)\bigr) + \frac{1}{2}\frac{\partial^2 M_0(\xi(\mu_{0T}))}{\partial\mu_{0T}^2}\bigg|_{\mu_{0T}=E^{\mathrm{KL}}(\mu_{0T})} \mathrm{Var}^{\mathrm{KL}}(\mu_{0T}). \end{aligned}$$

Equating the last two expressions and using the result obtained from the first equation, the proof is completed. (Note that if a second order approximation were also used for the first equation the same result would follow.) □

Example 2 (ctd.) Since the prior (8) induces a Gamma(s'_0, n'_0) distribution on $\mu_{0T} = \mu_1 + \mu_2 = 2\mu_0$, we have $E^C(\mu_{0T}) = s'_0/n'_0$ and $\mathrm{Var}^C(\mu_{0T}) = s'_0/n'^2_0$. Thus, using Proposition 2, we obtain

$$\begin{cases} \frac{s'_0}{n'_0} = E^{\mathrm{KL}}(\mu_{0T}) = E(\mu_1 + \mu_2) = \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right), \\ \frac{s'_0}{n'^2_0} = \mathrm{Var}^{\mathrm{KL}}(\mu_{0T}) = \mathrm{Var}(\mu_1 + \mu_2) = \left(\frac{\alpha_1}{\beta_1^2} + \frac{\alpha_2}{\beta_2^2}\right), \end{cases}$$

which leads to the solution

$$s'_0 = \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right)^2 \Big/ \left(\frac{\alpha_1}{\beta_1^2} + \frac{\alpha_2}{\beta_2^2}\right)$$

and

$$n'_0 = \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right) \Big/ \left(\frac{\alpha_1}{\beta_1^2} + \frac{\alpha_2}{\beta_2^2}\right).$$

Note that if $\beta_1 = \beta_2 = \beta$ then $\pi^{\mathrm{KL}}_{\mu_{0T}}$ is Gamma$(\alpha_1 + \alpha_2, 2\beta)$ which belongs to the standard conjugate family. In this case, even if we use the approximate solution given by (10) to compute $\pi^{\mathrm{KLCA}}_{\mu_{0T}}$, we obtain $\pi^{\mathrm{KL}}_{\mu_{0T}} = \pi^{\mathrm{KLCA}}_{\mu_{0T}}$ since s'_0 and n'_0 become $\alpha_1 + \alpha_2$ and β, respectively.

2.2 Conditioning priors

Another strategy to build compatible priors is via conditioning. Let $\mathcal{M}_0$ be a submodel of $\mathcal{M}$ identified by $\tilde{\Theta}_0 = \{\theta \in \Theta : t(\theta) = t_0\}$, where $t(\cdot)$ is (a vector-valued) function on Θ and t_0 a suitable constant. For a given prior π_θ on Θ, a natural way to obtain a compatible prior on $\tilde{\Theta}_0$ is to condition on $t(\theta) = t_0$; the corresponding prior on Θ_0 is denoted by $\pi^{\mathrm{UC}}_{\theta_0}$, where UC stands for Usual Conditioning. Unfortunately, the UC procedure may lead to different answers depending on the choice of the constraint function $t(\cdot)$; this, of course, is an instance of the Borel–Kolmogorov paradox.

Example 1 (ctd.) Since, under $\mathcal{M}_0$, one has $\mu_1 = \mu_2 \equiv \mu_0$, two possible choices for the constraint function $t(\cdot)$ are $t_1(\mu_1, \mu_2) = \mu_1 - \mu_2$ and $t_2(\mu_1, \mu_2) = \mu_1/\mu_2$. It may be checked, however, that the conditional distribution of μ_1 given $t_1(\mu_1, \mu_2) = 0$ is different from that of μ_1 given $t_2(\mu_1, \mu_2) = 1$.

To overcome the ambiguity associated with the UC-approach, which will be further exemplified in Sect. 3, Dawid and Lauritzen (2001) introduced a modification of the UC procedure named Jeffreys conditioning. The resulting expression for the prior on $\tilde{\Theta}_0$ is given by

$$\pi^{\mathrm{JC}}_0(\theta) \propto \pi(\theta)\frac{j_0(\theta)}{j(\theta)}, \quad \theta \in \tilde{\Theta}_0, \tag{11}$$

where $j(\theta) = |H(\theta)|^{1/2}$ and $|H(\theta)|$ is the determinant of the Fisher information matrix for θ under $\mathcal{M}$, so that $j(\theta)$ is the Jeffreys prior for θ, and analogously for $j_0(\theta)$ under the model $\mathcal{M}_0$. As usual, we can re-express π_0^{JC} in terms of $\theta_0 \in \Theta_0$ and, accordingly, we shall employ the notation $\pi_{\theta_0}^{\mathrm{JC}}$.

A useful feature of Jeffreys conditioning is that it is invariant under model reparametrisation. A potential difficulty with Jeffreys conditioning is that the resulting prior $\pi_{\theta_0}^{\mathrm{JC}}$ may be improper even though π_θ is proper.

Finally, we remark that Jeffreys conditioning is a special case of a more general procedure, named reference conditioning, developed in Roverato and Consonni (2004) for the analysis of models having a causal structure.

2.3 A decision theoretical approach

Another approach to the choice of a compatible prior is to state this as a decision problem (see, for example, Bernardo and Smith 1994, Chap. 6). Specifically, consider a statistical decision problem with the following elements:

Decision space

$$\mathcal{D} = \{\pi_0 : \pi_0 \text{ is a p.d.f. on } \Theta_0\},$$

where p.d.f. stands for probability density function. Typically, $\mathcal{D}$ is a parametric family so that $\pi_0(\cdot) \equiv \pi_0(\cdot|\omega_0)$, with $\omega_0 \in \Omega_0$, where Ω_0 is the set of hyperparameters. For example, in the setting described in Theorem 1, $\mathcal{D}$ is equal to the class of priors on Θ_0 induced from the standard conjugate family on ξ, $\mathcal{C}_\xi(s_0', n_0')$, so that $\omega_0 = (s_0', n_0')$. In such cases we can identify $\mathcal{D}$ with Ω_0.

States of the world

$$\mathcal{P} = \big\{p(\cdot|\theta) : \theta \in \Theta\big\}.$$

This is the (parametric) family of densities corresponding to the model $\mathcal{M}$. As in the previous case, we can identify $\mathcal{P}$ with Θ.

Prior

$$\pi_\theta, \text{ a p.d.f. on } \Theta,$$

describing prior beliefs about the value of θ.

Utility function

$$U(\omega_0, \theta) = \int p^S(s|\theta) \log m_0^S(s|\omega_0)\eta(ds),$$

where $S = S(X)$ is a statistic, η is an appropriate carrier measure, and

$$m_0^S(s|\omega_0) = \int p_0^S(s|\theta_0)\pi_0(\theta_0|\omega_0)\, d\theta_0$$

denotes the (prior) predictive distribution of S under $\mathcal{M}_0$.

If the conditions of Fubini's theorem hold, the corresponding expected utility is given by

$$\begin{aligned}\bar{U}(\omega_0) &= \int U(\omega_0,\theta)\pi_\theta(\theta)\,d\theta = \int\left\{\int p^S(s|\theta)\log m_0^S(s|\omega_0)\eta(ds)\right\}\pi_\theta(\theta)\,d\theta \\ &= \int\left\{\int p^S(s|\theta)\pi_\theta(\theta)\,d\theta\right\}\log m_0^S(s|\omega_0)\eta(ds) \\ &= \int m^S(s)\log m_0^S(s|\omega_0)\eta(ds), \end{aligned} \tag{12}$$

where $m^S(s)$ denotes the predictive distribution of S under $\mathcal{M}$.

The rationale behind this specific choice of utility function is that predictive distributions can be directly compared across models (unlike prior distributions, especially in the case of nested models).

The solution to the decision problem is the prior $\pi_0(\cdot|\omega_0)$ corresponding to the value of ω_0, which maximises the expected utility (12). Note that maximising $\bar{U}(\omega_0)$ with respect to ω_0 is equivalent to minimising the Kullback–Leibler divergence between $m^S(\cdot)$ and $m_0^S(\cdot|\omega_0)$. We emphasise that this utility function gives a precise meaning to the notion of "compatibility" between priors: two priors are most compatible when the corresponding predictive distributions of the chosen statistic under $\mathcal{M}$ and $\mathcal{M}_0$ are closest to each other, and the Kullback–Leibler criterion is a natural measure of the divergence between two distributions. For a related view see Ibrahim (1997) and the references therein.

Unfortunately, the computation of $\bar{U}(\omega_0)$ is difficult in general and so an approximation will be typically required. However, we can provide a simple asymptotic approximation which further motivates the result of Theorem 1.

Take $S=\hat{\theta}$, where $\hat{\theta}$ is the maximum likelihood estimator of θ under $\mathcal{M}$ based on a sample of size n. The predictive distribution of $\hat{\theta}$ is $m^{\hat{\theta}}(s)=\int p^{\hat{\theta}}(s|\theta)\pi_\theta(\theta)\,d\theta$. Under suitable regularity conditions and for each given θ, as $n\to\infty$, the sampling distribution of $\hat{\theta}$ degenerates at the point θ, so that $m^{\hat{\theta}}(\theta)\approx\pi_\theta(\theta)$. Similarly, given $\mathcal{M}_0$ and under suitable regularity conditions, $\hat{\theta}\to\theta_0^\perp$ as $n\to\infty$; see, for example, Wald (1949). Hence, asymptotically $m_0^{\hat{\theta}}(\theta_0^\perp|\omega_0)\approx\pi_0(\theta_0^\perp|\omega_0)$. Recalling from Sect. 2.1 that $\theta_0^\perp=\tau^{\mathrm{KL}}(\theta)$, with the latter defined in (3), the expected utility (12) can then be approximated by

$$\bar{U}(\omega_0)\approx\int\pi_\theta(\theta)\log\pi_0\big(\tau^{\mathrm{KL}}(\theta)|\omega_0\big)\,d\theta. \tag{13}$$

Maximising the right-hand side of this expression with respect to ω_0 is equivalent to minimising the Kullback–Leibler divergence between the KL-projection prior on Θ_0, $\pi_{\theta_0}^{\mathrm{KL}}$, and the prior $\pi_0(\cdot|\omega_0)$. To see this, note that the right-hand side of (13) can be written as

$$E^{\pi_\theta}\big\{\log\pi_0\big(\tau^{\mathrm{KL}}(\theta)|\omega_0\big)\big\},$$

which is equivalent to $E^{\pi_{\theta_0}^{\mathrm{KL}}}\{\log\pi_0(\theta_0|\omega_0)\}$.

The previous argument shows that, in the setting of Theorem 1, the Kullback–Leibler Conjugate Approximation, $\pi_{\theta_0}^{\text{KLCA}}$, is also an approximate solution to the decision problem described in this section. Specifically, $\pi_{\theta_0}^{\text{KLCA}}$ gives rise to a predictive distribution for $\hat{\theta}$, under $\mathcal{M}_0$, which is approximately closest in KL-divergence to the corresponding predictive distribution for $\hat{\theta}$ under $\mathcal{M}$.

3 The Hardy–Weinberg model

The Hardy–Weinberg (HW) model of equilibrium has been of interest to population geneticists in a variety of contexts, most notably evolutionary theory and forensic science. Lindley (1988) discusses Bayesian testing for HW-equilibrium and, in particular, computes, for four different data sets, the Bayes factor in favour of the HW-model versus a general (disequilibrium) model.

We follow Lindley to provide a brief introduction to the genetic problem and set some notation. At a single locus with two alleles, a diploid individual can be one of three possible genotypes, namely: AA, Aa, aa (aA being indistinguishable from Aa). Let p_1, p_2, p_3 with $p_i \geq 0$ and $p_3 = 1 - p_1 - p_2$ be the genotype frequencies in the population. Alternatively, p_i may be thought of as the probability that an individual, randomly chosen from the population, be of genotype i.

Consider a random sample of n individuals from the population. Conditionally on (p_1, p_2), let X_1 and X_2 represent genotype 1 and 2 counts whose joint sampling distribution is trinomial with index n and probabilities (p_1, p_2); we name this the "General Model" (GM), which corresponds to $\mathcal{M}$ in the notation of the previous section. Note that GM is a two-dimensional NEF with canonical parameter $\theta = (\theta_1, \theta_2)$, where $\theta_i = \log\{p_i/(1 - p_1 - p_2)\}$, and canonical statistic (X_1, X_2). The population is said to be in HW-equilibrium if

$$p_1 = p^2, \qquad p_2 = 2p(1-p), \qquad p_3 = (1-p)^2, \tag{14}$$

for some $0 < p < 1$. Note that (14) can be equivalently stated as saying that $p_2 = 2\sqrt{p_1}(1 - \sqrt{p_1})$. The trinomial model under assumption (14) is a curved exponential family and we name it the HW-model. It corresponds to $\mathcal{M}_0$ in the notation of the previous section. It can be verified that the HW-model is itself a one-dimensional NEF, actually Binomial$(p, 2n)$, with canonical parameter $\xi = \log\{p/(1-p)\}$ and canonical statistic $T(X_1, X_2) = 2X_1 + X_2$.

It is often convenient to reparametrise GM so as to show more explicitly the departure from the HW-model by means of disequilibrium parameters. There are several ways to do this, as described in Shoemaker et al. (1998), to which we refer for further details and references. We present here three such reparametrisations.

The first, due to Hernández and Weir (1989), writes the trinomial probabilities as

$$p_1 = p^2 + D, \qquad p_2 = 2p(1-p) - 2D, \qquad p_3 = (1-p)^2 + D, \tag{15}$$

where D represents a disequilibrium parameter and is subject to the following constraints

$$\max\{-p^2, -(1-p)^2\} \leq D \leq p(1-p). \tag{16}$$

Clearly, $D = 0$ corresponds to equilibrium.

The second parametrisation (Weir 1996) uses the inbreeding coefficient within populations, here denoted by f. It is given by

$$\begin{aligned} &p_1 = p^2 + p(1-p)f, \qquad p_2 = 2p(1-p)(1-f), \\ &p_3 = (1-p)^2 + p(1-p)f. \end{aligned} \tag{17}$$

The constraints on f are

$$\max\bigl\{-p/(1-p), -(1-p)/p\bigr\} \le f \le 1, \tag{18}$$

and $f = 0$ corresponds to HW-equilibrium.

Lindley (1988) suggested the following reparametrisation

$$\alpha = \frac{1}{2}\log\frac{4p_1p_3}{p_2^2}, \qquad \beta = \frac{1}{2}\log\frac{p_1}{p_3}. \tag{19}$$

Note that if HW-equilibrium (14) obtains then $\alpha = 0$, and $\beta = \log\{p/(1-p)\}$; conversely, if $\alpha = 0$ then $p_2 = 2\sqrt{p_1}(1-\sqrt{p_1})$ which is equivalent to (14), whence, setting $p_1 = p^2$ one has $\beta = \log\{p/(1-p)\} = \xi$. In other words $\alpha = 0$ identifies the HW-model. An important advantage of the (α, β) parametrisation is that the two parameters are variation independent, as opposed to the awkward dependence between p and D, exhibited in (16), or between p and f, as shown in (18).

4 Compatible priors for testing Hardy–Weinberg equilibrium

4.1 Priors under the general model

We assume that, under GM, (p_1, p_2) is distributed according to a Dirichlet prior, with hyperparameters $m_i > 0$, written $\mathrm{Di}(m_1, m_2, m_3)$. We define $M = m_1 + m_2 + m_3$ to be the "precision" of the Dirichlet family. The Dirichlet family is (standard) conjugate for the general model and allows a closed-form expression for the marginal distribution of the data, which is especially useful in order to compute the Bayes factor. Moreover, it covers a wide range of possible prior specifications as we now detail.

First of all we remark that (p_1, p_2) and (α, β), introduced in (19), are conjugate parametrisations, as defined in Gutiérrez-Peña and Smith (1995). This means that the Dirichlet family on (p_1, p_2) is equivalent to the standard conjugate family on (α, β), in the sense that the latter can be obtained via a change of variable technique.

Lindley (1988) argues in favour of a data-dependent choice for the hyperparameters (m_1, m_2, m_3), namely

$$m_1 = M\hat{p}^2, \qquad m_2 = M2\hat{p}(1-\hat{p}), \qquad m_3 = M(1-\hat{p})^2, \tag{20}$$

where $\hat{p}$ is the maximum likelihood estimate of p under the HW-model, i.e., $\hat{p} = (2x_1 + x_2)/(2n)$. The choice of (m_1, m_2, m_3), described in (20), is such that $E(p_i) = m_i/M$ obey the HW-equilibrium, while the corresponding prior on (α, β) has a mode at $\alpha = 0$, which characterises the HW-model; accordingly, we name it

"HW-Dirichlet", since this prior always favours the HW-model. Moreover, there is no disagreement between prior expectation and the ML-estimate under the HW-model: in Lindley's opinion the advantage of (20) is that "any effects observed are not confounded with the difference in the allele-proportion between observed and expected".

Now let $\hat{p}_i = x_i/n$ be the MLE of p_i under GM. Since $M\hat{p}_i$ is unbiased for m_i under the predictive distribution of $\hat{p}_i$, an alternative data-dependent choice, based on an empirical Bayes argument, for the hyperparameters of the Dirichlet family is

$$m_1 = M\hat{p}_1, \qquad m_2 = M\hat{p}_2, \qquad m_3 = M\hat{p}_3,$$

whose expectations are in agreement with the GM. We name this prior "GM-Dirichlet".

A further, data-independent Dirichlet prior we consider is the symmetric Dirichlet with $m_i = M/3$. The special case $M = 3/2$ (so that $m_i = 1/2$) corresponds to Jeffreys prior. Note that this prior cannot favour the HW-model, since its expectation structure does not satisfy the corresponding constraint.

4.2 Compatible priors

Given the Dirichlet prior $(p_1, p_2) \sim \mathrm{Di}(m_1, m_2, m_3)$ under GM, one can compute the corresponding compatible prior according to Jeffreys conditioning or KL-projection.

For completeness, we also mention the possibility of applying usual conditioning (UC), although, as discussed above, this is not particularly advisable given the lack of invariance. For example, if one reparametrises the problem in terms of (α, β) and then conditions on $\alpha = 0$, one obtains $p \sim \mathrm{Beta}(2m_1 + m_2, 2m_3 + m_2)$, as in Lindley (1988); we label this prior π_p^{UC}. On the other hand, if one had started with the alternative parametrisation (15), the conditional distribution of p given $D = 0$ would have been $\mathrm{Beta}(2m_1 + m_2 - 2, 2m_3 + m_2 - 2)$, provided $2m_1 + m_2 > 2$ and $2m_3 + m_2 > 2$; we label this distribution π_p^{UCD}, where UCD stands for Usual Conditioning given $D = 0$.

4.2.1 KLCA-prior

In order to identify the KL-prior, recall that $T = (2X_1 + X_2)$ is a sufficient statistic for the HW-model with a Binomial$(p, 2n)$ distribution. From Proposition 1, the KL-projection is the solution with respect to the unknown p of the following equation

$$E_p^{\mathrm{HW}}\big(T(X_1, X_2)\big) = E_{p_1,p_2}^{\mathrm{GM}}\big(T(X_1, X_2)\big),$$

i.e., $2np = n(2p_1 + p_2)$, whence $p^{\perp} = p_1 + p_2/2$.

Since π_p^{KL} cannot be obtained in a closed form, we apply Proposition 2. Equating the expectations and variances of p under π_p^{KL} and π_p^{C}, the latter being a Beta(a, b), and solving for a and b we obtain

$$a = \frac{1}{2}\frac{(2m_1 + m_2)(m_2^2 + 2m_1m_2 + m_2 + 2m_2m_3 + 4m_1m_3)}{(m_1m_2 + 4m_1m_3 + m_2m_3)},$$

$$b = \frac{1}{2}\frac{(8m_1m_3^2 + 8m_1m_2m_3 + 2m_1m_2^2 + m_2^2 + 2m_2m_3 + 4m_2m_3^2 + m_2^3 + 4m_2^2m_3)}{(m_1m_2 + 4m_1m_3 + m_2m_3)},$$

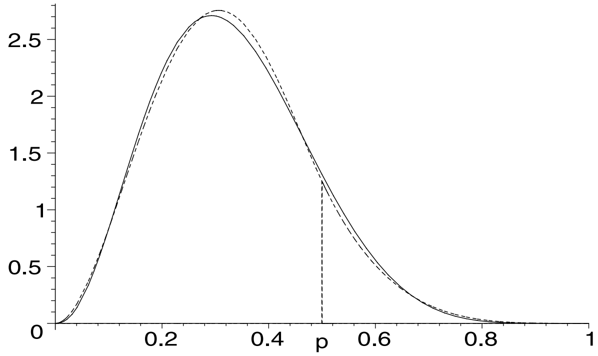

Fig. 1 Prior π_p^{KL} (*dotted line*) and its approximation π_p^{KLCA} (*solid line*) for $m_1 = 1$, $m_2 = 2$, $m_3 = 3$

which reduces to $a = b = (3M + 1)/4$ when $m_1 = m_2 = m_3 = M/3$.

Figure 1 shows the actual prior π_p^{KL} along with its approximation π_p^{KLCA} for a specific choice of the hyperparameters (m_1, m_2, m_3). We have experimented with several choices of (m_1, m_2, m_3) and always found an excellent fit of the approximation.

4.2.2 JC-prior

In order to compute the JC-prior we need the following items. First of all, $j(p_1, p_2)$ is given by the square-root of the determinant of the Fisher information matrix for the trinomial model. This is easily computed as $j(p_1, p_2) \propto \{p_1 p_2(1 - p_1 - p_2)\}^{-1/2}$. Since the distribution of $2X_1 + X_2$ under the HW model is Binomial$(p, 2n)$, one can derive $j_0(p) \propto \{p(1 - p)\}^{-1/2}$. We can finally obtain the JC-prior on p as

$$\begin{aligned}\pi_p^{\text{JC}}(p) &\propto \pi_{p_1 p_2}\big(p_1(p), p_2(p)\big) \times \frac{j_0(p)}{j(p_1(p), p_2(p))}\\ &\propto p^{2m_1+m_2-2}(1 - p)^{2m_3+m_2-2},\end{aligned} \tag{21}$$

namely a Beta$(2m_1 + m_2 - 1, 2m_3 + m_2 - 1)$. Clearly, π_p^{JC} is proper, provided $2m_1 + m_2 > 1$ and $2m_3 + m_2 > 1$.

Recalling that the "precision" of a Beta distribution with parameters (a, b) is $(a + b)$, it follows that the precision under the UC-prior is $2M$, while that under the JC-prior is $2(M - 1)$. The precision under the KLCA-prior is more complex but, assuming $m_1 = m_2 = m_3$, it becomes $\frac{3}{2}M + \frac{1}{2}$, which is smaller than either of the previous two when $M > 5$. Furthermore, it is interesting to compare the expected values of p under each of the three priors which are $E^{\text{UC}} = E^{\text{KLCA}} = (2m_1 + m_2)/(2M)$ and $E^{\text{JC}} = (2m_1 + m_2 - 1)/(2M)$ (provided π_p^{JC} is proper).

4.3 Bayes factors

From the previous subsections we conclude that, when the prior under GM is Di(m_1, m_2, m_3), all compatible priors under HW are of type Beta(a, b). As a consequence, for trinomial data (x_1, x_2, x_3), it can be shown that the Bayes factor in favour of the HW-model takes the form

$$B_{01} = \frac{2^{x_2}\Gamma(M + n)\Gamma(a + 2x_1 + x_2)\Gamma(b + 2n - 2x_1 - x_2)\Gamma(a + b)\Gamma(m_1)\Gamma(m_2)\Gamma(m_3)}{\Gamma(M)\Gamma(a)\Gamma(b)\Gamma(a + b + 2n)\Gamma(m_1 + x_1)\Gamma(m_2 + x_2)\Gamma(m_3 + n - x_1 - x_2)}.$$

In order to compare the behaviour of B_{01} under different compatible priors, we consider four data sets, each having sample size $n = 100$, previously analysed by Emigh (1980) and Lindley (1988): $\{31, 38, 31\}$, $\{6, 22, 72\}$, $\{2, 6, 92\}$ and $\{1, 8, 91\}$. For the first three sets, the classical Haldane's "exact" test (Haldane 1954) rejects the null hypothesis of HW-equilibrium with significance level below 3.4%, whereas for the last data set the HW-model is not rejected, its p-value being around 20%.

Table 1 shows the Bayes factors for each of the four data sets, each choice of Dirichlet prior under the general model (HW, GM, symmetric, and Jeffreys), and each compatible prior (UC, JC, KLCA), for various values of M. Note that the combination HW-Dirichlet/UC corresponds to Table 3 of Lindley (1988).

Consider first the comparative behaviour of the priors under the general model.

In the case of the HW-Dirichlet prior, it is apparent that the values of B_{01} are higher than the corresponding values for the other priors under the general model, and this is consistent with the fact that the HW-Dirichlet supports the HW-model. Moreover, for the first three data sets B_{01} initially decreases, as M grows, and then increases; for the last data set, the decrease of B_{01} is monotone. In any case, its limiting value is 1; see the discussion in Lindley (1988). Under the GM- and symmetric Dirichlet prior, B_{01} decreases monotonically for all data sets as M increases; its limiting value is, however, data-dependent.

From the above comments, we conclude that only moderate values of M should be used in the analysis.

To fix the value of M we may use an argument based on imaginary observations, see Spiegelhalter and Smith (1980). Here is the idea expressed in a more general context. Given a model $\mathcal{M}$ and a submodel $\mathcal{M}_0$, consider a *minimal* imaginary training sample that provides *maximal* support (irrespective of the prior) to $\mathcal{M}_0$. Then it is reasonable to require that the Bayes factor should be around one, since the evidence in favour of $\mathcal{M}_0$ is based on a limited sample size.

We implement the above argument as follows. Since under the GM there are two unknown parameters p_1 and p_2, we consider an imaginary minimal training sample of two observations (y_1, y_2). Next we identify those realisations that provide maximal support to the HW-model by computing the Likelihood Ratio Test (LRT) statistic

$$\Lambda(y_1, y_2) = \frac{\mathcal{L}_0(\hat{p})}{\mathcal{L}_1(\hat{p}_1, \hat{p}_2)},$$

where $\mathcal{L}_0$ (respectively, $\mathcal{L}_1$) is the likelihood under the HW-model (respectively, the GM-model); $\hat{p} = \frac{2y_1+y_2}{2n}$ and $\hat{p}_i = \frac{y_i}{n}$, $i = 1, 2$, with $n = 2$. One can verify that $\Lambda(y_1, y_2)$ is maximised at $(y_1, y_2) = (0, 0)$ or at $(y_1, y_2) = (2, 0)$, with $\Lambda(y_1, y_2)$ taking the value one. Consider for concreteness the symmetric prior under the GM and the KLCA prior under the HW. Then the Bayes factor is the same under the above two realisations and is equal to

$$B_{01} = 9\frac{M+1}{M+3}\frac{\Gamma(a^{\mathrm{KL}}+4)}{\Gamma(2a^{\mathrm{KL}}+4)}\frac{\Gamma(2a^{\mathrm{KL}})}{\Gamma(a^{\mathrm{KL}})}, \tag{22}$$

where $a^{\mathrm{KL}} = \frac{3M+1}{4}$. Setting $B_{01} = 1$ in (22) and solving for M leads to $M = 0.24$; on the other hand, setting $M = 1$ into (22) gives $B_{01} = 0.9$, which is essentially indistinguishable from 1 for our purposes. If, instead of KLCA, we use the JC-procedure,

Table 1 Bayes factors in favour of HW for three compatible priors

	UC	JC	KLCA	UC	JC	KLCA
Data set	$\{31, 38, 31\}$			$\{6, 22, 72\}$		
	($\hat{p} = 0.50$)			($\hat{p} = 0.17$)		
	Jeffreys prior					
	0.617	0.310	0.587	0.804	0.710	0.820
M	HW-Dirichlet					
1	0.997	–	1.267	4.556	–	5.940
5	0.349	0.312	0.366	1.208	0.620	1.269
20	0.241	0.236	0.243	0.600	0.545	0.607
500	0.691	0.691	0.691	0.794	0.794	0.794
M	GM-Dirichlet					
1	0.857	–	0.915	2.720	–	2.968
5	0.277	0.248	0.258	0.733	0.376	0.686
20	0.137	0.134	0.126	0.324	0.295	0.300
500	0.060	0.060	0.059	0.136	0.136	0.134
M	Symmetric Dirichlet					
1	0.852	–	0.852	1.172	–	1.172
5	0.281	0.251	0.251	0.251	0.375	0.375
20	0.147	0.144	0.131	0.030	0.042	0.162
500	0.089	0.089	0.087	6.4e–6	6.4e–6	4.2e–5
Data set	$\{2, 6, 92\}$			$\{1, 8, 91\}$		
	($\hat{p} = 0.05$)			($\hat{p} = 0.05$)		
	Jeffreys prior					
	0.096	0.244	0.111	1.083	2.748	1.256
M	HW-Dirichlet					
1	7.739	–	10.645	65.956	–	90.718
5	1.753	–	1.857	14.008	–	14.841
20	0.661	0.376	0.668	4.303	2.447	4.352
500	0.678	0.676	0.679	1.136	1.132	1.136
M	GM-Dirichlet					
1	1.482	–	1.435	19.534	–	23.000
5	0.335	–	0.286	4.323	–	4.152
20	0.123	0.070	0.107	1.483	0.844	1.399
500	0.046	0.046	0.045	0.511	0.510	0.505
M	Symmetric Dirichlet					
1	0.164	–	0.164	1.718	–	1.718
5	0.011	0.039	0.039	0.164	0.584	0.584
20	7.4e–5	1.7e–4	0.005	0.001	0.003	0.080
500	7e–11	7e–11	2.4e–9	4e–10	4e–10	1.4e–8

the value $M = 1$ is not acceptable, since it leads to an improper prior under the HW-model; however, if $M = 2.24$ then $B_{01} = 1$, and indeed $M = 2$ gives $B_{01} = 1.1$, so that in this case a value of $M = 2$ seems appropriate. We conclude that a value of M between 1 and 2 represents a reasonable choice for a weakly informative prior.

In order to better appreciate the relative merits of the priors involved under the general model, it is expedient to use a scale for the value of the Bayes factor *against* the HW-model, B_{10}, according to a suggestion of Jeffreys, as described in Kass and Raftery (1995, p. 777). Specifically, we consider four classes of evidence *against* the HW-model based on $\log_{10} B_{10} = -\log_{10} B_{01}$, namely: $(-\infty, 1/2]$ "no evidence or not worth more than a bare mention"; $(1/2, 1]$ "substantial"; $(1, 2]$ "strong; $(2, \infty)$ "decisive". In practice a value of B_{01} less than 0.31 leads to a rejection of $\mathcal{M}_0$.

It is worth focusing on the third and fourth data sets, since they are seemingly similar (and, in fact, give rise to the same MLE for p under the HW-model), but lead to different conclusions according to the frequentist approach as described above. The latter conclusion is confirmed if one adopts Jeffreys prior (corresponding to $M = 1.5$) and any of the compatible priors.

Lindley (1988) used a reference prior on (α, β), see (19), and plotted the corresponding posterior distribution of α for the data set $\{1, 8, 91\}$. Recall that the HW-model is identified by $\alpha = 0$. Since the reference prior is improper, we can evaluate the evidence in favour of the HW-model (i.e., $\alpha = 0$) by means of an HPD (High Posterior Density) interval for α at levels 95% and 99%. In both cases $\mathcal{M}_0$ is rejected for the third data set and is accepted for the last data set. For the first two data sets $\mathcal{M}_0$ is rejected at level 95% while it is accepted at level 99%. There is thus perfect agreement between the conclusion based on the 99% HPD reference interval for α and that obtained under Jeffreys prior together with any compatible prior, according to the Jeffreys scale.

From the previous discussion of the four data sets analysed by Lindley it appears that there is no appreciable difference between the various compatible priors under investigation.

5 Simulation study

In order to provide a more detailed comparison of the Bayes factors using different compatible priors, we have performed both an asymptotic analysis based on Laplace approximation and a simulation study. As far as the former is concerned, the reader is referred to the full technical report (Consonni et al. 2005, Sect. 4.4), from which it appears that, at least asymptotically, the KLCA-prior is an appealing choice. Here we consider a simulation study with two scenarios. The first, labelled "equilibrium", assumes that the HW-model is true: specifically, we generated $1,000$ trinomial data sets for each value of the parameter p in the set $\{0.05, 0.17, 0.50\}$ and each sample size n in the set $\{10, 50, 100, 500, 1000\}$. The second scenario, labelled "disequilibrium", assumes that the GM holds under a variety of disequilibrium situations. Specifically, we used the parametrisation discussed in (17), and generated $1,000$ trinomial data sets for each value of the inbreeding coefficient f in the set $\{-0.15, 0.5, 0.95\}$ and $p = 0.17$. The values of the sample size n were chosen as in the equilibrium scenario.

We used a symmetric Dirichlet prior under the GM, which seems appropriate, at least as a benchmark, if one does not want to rely on data-dependent priors. Because of the remarks in Sect. 4.3, we considered only small to moderate values of M; specifically, we let M vary in the set $\{1.5, 5, 10, 20\}$. For each choice of the compatible prior UC, JC and KLCA, we computed the Bayes factor *against* the HW-model, B_{10}, so that we were able to use directly the four-class scale proposed by Jeffreys, as discussed in Sect. 4.3.

Due to space constraints, here we shall only summarise the main findings of our simulation study. Further details can be found in the technical report (Consonni et al. 2005), where the relative frequency distribution of $\log_{10} B_{10}$ across the four classes is presented, both under "equilibrium" (Table 2) and "disequilibrium" (Table 3), for $n = 10$ and $n = 100$.

Consider first the equilibrium scenario. The performance of KLCA is almost always superior relative to UC and JC in terms of correct classification. For example, when $p = 0.05$, $n = 100$, $M = 10$, the frequency of class 1, corresponding to "no evidence or not worth more than a bare mention" is only 20% under UC and 63% under JC, while it is 90% under KLCA. Moreover, KLCA is less sensitive to the choice of M, irrespective of the sample size; for example, when $M = 20$ (not reported in the table), the corresponding percentages are 1%, 2% and 63%.

Consider now the disequilibrium scenario. In particular, focus first on the case $f = -0.15$, corresponding to a situation which is actually "close" to equilibrium, not only for $p = 0.17$ but for all values of p; see Fig. 2a. For sample sizes up to $n = 100$, the modal class is by far the first one, formally corresponding to a wrong statement, although for $n \geq 500$ the performance markedly improves, with classes "strong" and "decisive" together accounting for at least 80% of the frequency. Summarising, while all compatible priors perform poorly for low to moderate n, this is not too worrisome, since the disequilibrium model under consideration is actually "close" to equilibrium; at any rate, the Bayes factors become sensible for large values of n. Overall, the performance of KLCA versus UC and JC is comparable for values of M up to 10.

We now consider the case corresponding to $f = 0.50$, which is substantially different from equilibrium; see Fig. 2b. For all priors the combined frequency of the third and fourth classes is at least 90%, when n exceeds 100. For lower values of the sample size this frequency decreases, and can be very low for $n = 10$. In the latter cases KLCA performs only slightly worse.

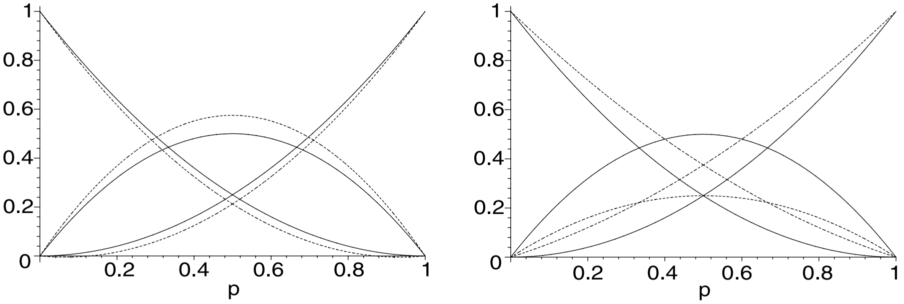

Fig. 2 Curves describing p_1, p_2 and p_3 under the equilibrium model (*solid lines*) and disequilibrium models (*dotted lines*) with: (**a**) $f = -0.15$ (*left*); (**b**) $f = 0.50$ (*right*)

We finally turn to the case $f = 0.95$, which represents an extreme form of disequilibrium. Only for $n = 10$ is the combined frequency of the third and fourth class not completely satisfactory: again KLCA's performance is comparable to that of the remaining priors.

For large n (e.g., 1000) the distribution is essentially degenerate at the correct class for all choices of compatible priors.

6 Discussion

In this paper we have discussed strategies to construct compatible priors for a collection of submodels. We have criticised the Usual Conditioning (UC) approach because of its lack of invariance to the choice of the constraint function, and we have discussed two alternative approaches, namely Jeffreys Conditioning (JC) and Kullback–Leibler (KL) projection, both of which are invariant to reparametrisation. A criticism of JC is that it is based on a procedure that may violate probability rules: specifically, it can lead to an improper compatible prior, even when the starting prior is proper.

We believe that the motivation behind the approach based on the KL-projection is appealing. Moreover, the approximation we have developed in this paper not only makes the method operationally applicable but also provides, at least asymptotically, an interesting interpretation from a decision theoretical point of view based on predictive distributions.

We tested the three procedures to construct compatible priors on a specific problem, namely the Hardy–Weinberg equilibrium model, using both real and synthetic data, for a variety of prior specifications under the encompassing model, including a weakly informative one. As far as the data sets analysed in Lindley (1988) are concerned, the KLCA and UC priors perform similarly, lending greater support to the HW-model than the JC-prior.

To further investigate the relative merits of the various compatible priors, we carried out a simulation study under a variety of situations involving different generating models, sample sizes and precisions of the prior under the general model. When the HW-model was the generating mechanism, the KLCA prior performed best in terms of model choice; on the other hand, it performed similarly to the other priors under a range of disequilibrium models, despite its tendency to favour the equilibrium model. From this perspective the approach based on the KL-projection seems promising and worth of further investigation in other applied domains.

Acknowledgements Part of this work was carried out while E.G. was visiting the Institute of Quantitative Methods, L. Bocconi University, and the Department of Economics and Quantitative Methods, University of Pavia, and G.C. together with P.V. were visiting IIMAS-UNAM. They are grateful to these institutions for their hospitality. The comments of a referee helped to improve the presentation of the paper.

References

Bernardo JM, Rueda R (2002) Bayesian hypothesis testing: a reference approach. Int Stat Rev 70:351–372
Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester
Casella G, Moreno E (2006) Objective Bayesian variable selection. J Am Stat Assoc 101:157–167

Consonni G, Gutiérrez-Peña E, Veronese P (2005) Compatible priors for Bayesian model comparison with an application to the Hardy–Weinberg equilibrium model. Technical Report, Dipartimento di Economia Politica e Metodi Quantitativi, University of Pavia. http://economia.unipv.it/~gconsonni/www/papers/CGV_TEST-TechRep.pdf

Dawid AP, Lauritzen SL (2001) Compatible prior distributions. In: George E (ed) Bayesian methods with applications to science, policy and official statistics. Monographs of official statistics Office for official publications of the European Communities, Luxembourg, pp 109–118. http://www.stat.cmu.edu/ISBA/index.html

Dupuis JA, Robert CP (2003) Variable selection in qualitative models via an entropic explanatory power. J Stat Plan Inference 111:77–94

Emigh TH (1980) A comparison of tests for Hardy–Weinberg equilibrium. Biometrics 36:627–642

Goutis C, Robert CP (1998) Model choice in generalised linear models: a Bayesian approach via Kullback–Leibler projections. Biometrika 85:29–37

Gutiérrez-Peña E, Smith AFM (1995) Conjugate parameterizations for natural exponential families. J Am Stat Assoc 90:1347–1356

Haldane JBS (1954) An exact test for randomness of mating. J Genet 52:631–635

Hernández JL, Weir BS (1989) A disequilibrium coefficient approach to Hardy–Weinberg testing. Biometrics 45:53–70

Ibrahim JG (1997) On properties of predictive priors in linear models. Am Stat 51:333–337

Kotz S, Balakrishnan N, Johnson NL (2000) Continuous multivariate distributions. Models and Applications, vol 1, 2nd edn. Wiley, New York

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Lindley DV (1988) Statistical inference concerning Hardy–Weinberg equilibrium. In: Bernardo JM, DeGroot MH, Lindley DV, Smith. AFM (eds) Bayesian statistics 3. University Press, Oxford, pp 307–326

McCulloch RE, Rossi PE (1992) Bayes factor for nonlinear hypotheses and likelihood distributions. Biometrika 79:663–676

Morris CN (1982) Natural exponential families with quadratic variance functions. Ann Stat 10:65–80

Roverato A, Consonni G (2004) Compatible prior distributions for DAG models. J Roy Stat Soc B 66:47–61

Shoemaker J, Painter I, Weir SB (1998) A Bayesian characterization of Hardy–Weinberg disequilibrium. Genetics 149:2079–2088

Spiegelhalter DJ, Smith AFM (1980) Bayes factor and choice criteria for linear models. J Roy Stat Soc B 42:215–220

Wald A (1949) Note on the consistency of the maximum likelihood estimate. Ann Math Stat 20:595–601

Weir BS (1996) Genetic data analysis. Sinuer, Sunderland