Inference after checking multiple Bayesian models for data conflict

May 21, 2014

David R. Bickel

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa; 451 Smyth Road; Ottawa, Ontario, K1H 8M5

Abstract

Two major approaches have developed within Bayesian statistics to address uncertainty in the prior distribution and in the overall model more generally. First, methods of model checking, including those assessing prior-data conflict, determine whether the prior and the rest of the model are adequate for purposes of inference and estimation or other decision-making. The main drawback of this approach is that it provides little guidance for inference in the event that the model is found to be inadequate, that is, in conflict with the data. Second, the robust Bayes approach determines the sensitivity of inferences and decisions to the prior distribution and other model assumptions. This approach includes rules for making decisions on the basis of a set of posterior distributions corresponding to the set of reasonable model assumptions. Drawbacks of the second approach include the inability to criticize the set of models and the lack of guidance for specifying such a set.

Those two approaches to model uncertainty are combined in order to overcome the limitations of each approach. The first approach checks each model within a large class of models to assess which models are in conflict with the data and which models are adequate for purposes of data analysis. The resulting set of adequate models is then used for inference according to decision rules developed for the robust Bayes approach and for imprecise probability more generally. This proposed framework is illustrated by the application of a class of hierarchical models to a simple data set.

Keywords: Bayesian model selection; empirical Bayes; imprecise probability; local false discovery rate; model assessment; model checking; possibility theory; posterior predictive checks; prior-data conflict; robust Bayesian decision theory; strength of statistical evidence

1 Introduction

In Bayesian statistics, a model, including some prior distribution, is considered inadequate for inference purposes if it is assessed to conflict with the observed data. Otherwise, the model is considered adequate, and inference proceeds according to Bayes's theorem. If a loss function is specified, the Bayes rule requires the minimization of the posterior expected loss to determine the result of a hypothesis test, parameter estimation, or other action taken on the basis of the model passing the check for data conflict (Ando, 2010, pp. 6-7).

While this approach guards against excessive reliance on a single model, it faces threats from two fronts, one corresponding to passing the model check and the other to failing it. The fact that a model passes a reasonable check is insufficient to justify its use in inference and decision-making. That is seen from the fact that such assessments are conservative, each only checking for certain deviations from the data. A lack of evidence for a model's conflict with data does not warrant the sole use of that model. In fact, it is usually the case that many models would pass the same check for adequate agreement with the data. In general, there is no single best explanation of the data (Psillos, 2002, §4.2).

Worse, when a model fails the assessment, there is no general prescription for how to proceed with data analysis without relying on the model deemed inadequate (Lele and Dennis, 2009). The usual advice is to return to the model-formulation stage, perhaps with a more careful elicitation of an expert's prior opinion. There is little algorithmic guidance at this point other than to to modify or relax assumptions made by the original model. Caution is needed here since moving the model in the direction of the data introduces some bias due to again using the same data for going from the prior distribution under the new model to the posterior distribution. Even without consciously using the data twice, a mechanical iteration of the modeling-assessment routine until a model is found that passes the check eventually leads to some model in better agreement with the data than previous models. Cross validation and other methods that split the data into training and testing sets attenuate such reuse of the data but often at the expense of the ability to draw useful conclusions from a subset of the original data. Moreover, such methods lack the theoretical foundations that make Bayesian statistics appealing.

While it is not always practical to attempt to eliminate all double-use of the data, a more principled approach can keep it under control while at the same time avoiding undue reliance on the first model that passes the check. This is achieved in two stages: Divide a broad class of possible models into a set of adequate models, those passing the assessment of data agreement, and a set of inadequate models, those failing the check. Perform inference, estimation, and other actions on the basis of the set of adequate models by using one of the algorithms that generalizes the Bayes rule.

This paper develops that solution as follows. Section 2 will consider sets of models that are sufficiently close to the data according to the chosen criterion of model checking. In noted respects, this approach differs from a previous approach leading to sets of models generated by a non-Bayesian model selection criterion (Burnham and Anderson, 2002, pp. 169-171). For concreteness without many of the interpretation problems that posterior predictive checks share with frequentist p-values, Bayes factors will be used to assess model adequacy. Since this is done in the absence of the prior distributions needed to convert them to posterior odds, it is outside the scope of traditional Bayesian model selection and Bayesian model averaging. To make the method clear, examples of sets of adequate models under various assumptions are provided in Section 3.

Data analysis may stop with reporting the set of posterior probabilities or probability distributions corresponding to the set of adequate models. Alternatively, as described in Section 4 data analysis may proceed to decisions such as hypothesis tests and point estimates using methods originally developed for the use of sets of models that do not depend on the data. For example, the robust Bayes literature has methods of determining optimal actions on the basis of sets of prior distributions in some neighborhood of the initial prior distribution (e.g., Betrò and Ruggeri, 1992). Likewise, the economics, operations research, and imprecise probability literatures offer decision rules for sets of models corresponding to partially ordered preferences (e.g., Giron and Rios, 1980; Gajdos et al., 2004) and differences the prices of buying and selling a gamble (e.g., Walley, 1991; Troffaes, 2007). Since the sets of models used do not depend on data, with the notable exception of confidence sets (Remark 1), the sets are in principle beyond criticism in the light of new information. Thus, applying those methods to the results of Bayesian model assessment extends their domain and may broaden their appeal to the mainstream statistics community.

Finally, Section 5 applies the proposed approach to a well studied data set, and the remarks in Section 6 conclude the paper with generalizations of the proposed two-stage approach to methods of model assessment differing from the featured use of the Bayes factor.

2 Defining sets of adequate models

Suppose the observation x is a member of some set $\mathcal{X}$ of possible observations. Let $\mathcal{M}$ denote a set of models such that for every $M \in \mathcal{M}$, there is a density function (Radon-Nikodym derivative) π_M on a parameter set Θ_M and $\mathcal{F}_M = \{f_M(\bullet|\theta) : \theta \in \Theta_M\}$, a family of probability density functions on $\mathcal{X}$, such that $(\pi_\bullet, \mathcal{F}_\bullet) : \mathcal{M} \to \{(\pi_M, \mathcal{F}_M) : M \in \mathcal{M}\}$ is invertible. For any model $M \in \mathcal{M}$, the density function π_M is called the *prior distribution*, the density function $f_M(\bullet|\theta)$ is called the *sampling distribution* given some $\theta \in \Theta$, and the density $f_M(x|\theta)$, considered as a function of θ, is called the *likelihood function.*

Every model $M \in \mathcal{M}$, there is an alternative model $\neg M$ corresponding to a prior distribution $\pi_{\neg M}$ such that $\pi_{\neg M} \neq \pi_M$ and to a family $\mathcal{F}_{\neg M} = \mathcal{F}_M$ of sampling distributions. The alternative model would be used for inference given only the information that model M is false. The corresponding integrated likelihoods are

$$f_M(x) = \int f_M(x|\theta)\,\pi_M(\theta)\,d\theta \tag{1}$$

$$f_{\neg M}(x) = \int f_M(x|\theta)\,\pi_{\neg M}(\theta)\,d\theta. \tag{2}$$

Thus, the *Bayes factor* in favor of model M over its alternative model is

$$B(M;x) = \frac{f_M(x)}{f_{\neg M}(x)} = \frac{\int f_M(x|\theta)\,\pi_M(\theta)\,d\theta}{\int f_M(x|\theta)\,\pi_{\neg M}(\theta)\,d\theta},$$

with the numerator and denominator only differing by the prior density function with respect to which the integration is performed.

Good (1979) defined the *weight of evidence* in favor of model M and against model $\neg M$ by

$$w(M) = \log B(M;x),$$

where the logarithm is of base 2 in the present paper, yielding $w(M)$ as the number of bits of information in favor of M and against $\neg M$. Consequently, the weight of evidence in favor of $\neg M$ and against M is $-w(M)$. Negative values of $w(M)$ are interpreted according to Table 1. For example, if $w(M) \leq -1$ or $w(M) \leq -3$, then there is at least "weak" or at least "strong" evidence, respectively, against model M and in favor of model $\neg M$.

Negligible	**Weak**	**Moderate**	**Strong**	**Very strong**	**Overwhelming**
$w(M) \leq 0$	$w(M) \leq -1$	$w(M) \leq -2$	$w(M) \leq -3$	$w(M) \leq -5$	$w(M) \leq -7$

Table 1: Grades of evidence against a model M and in favor of its alternative $\neg M$. Such grades correspond to intervals of $w(M)$, the weight of evidence in bits, an adaptation (Royall, 1997; Bickel, 2011) of the base-10 grades of Jeffreys (1948).

For any $a \in \mathbb{R}$, the set of *a-adequate* models is defined as

$$\mathcal{M}(a) = \{M \in \mathcal{M} : w(M) > a\}. \tag{3}$$

It follows that, for any $a < 0$, the a-adequate set contains all models against which there are no more than $|a|$ bits of evidence.

If $\vartheta_M \sim \pi_M$ is defined for the same measurable space $(\Theta, \mathfrak{H})$ for all $M \in \mathcal{M}$, the posterior probabilities that the parameter lies in $\mathcal{H}$ for some $\mathcal{H} \in \mathfrak{H}$ given a model adequacy of a constitute the set

$$\mathcal{P}_a(\mathcal{H}|x) = \left\{ \int_{\mathcal{H}} \pi_M(\theta|x)\, d\theta : M \in \mathcal{M}(a) \right\}, \tag{4}$$

called the set of *a-adequate posterior probabilities* of $\mathcal{H}$. A set rather than a real number, $\mathcal{P}_a(\mathcal{H}|x)$ is an *imprecise probability* (Walley, 1991). Some of the properties of imprecise probability that are useful in decision making are explained in Section 4.

3 Example sets of adequate models

Example 1 uses a simple and well-known class of models to illustrate the basic concepts introduced in Section 2. More general classes of models can be approached using the methods of Example 2, which are also simple but which do not require explicit specification of prior distributions. The concept of sets of adequate models is applied to recently adopted multiple testing procedures in Example 3. Lastly, Example 4 combines aspects of the previous two examples, providing the framework for the case study of Section 5.

Example 1. Let $\mathcal{M} = \mathbb{R}$, and, for all $M \in \mathbb{R}$, let $\mathcal{F}_M = \{\phi_{\mu,1} : \mu \in \mathbb{R}\}$, the set of all normal density functions of unit variance, let π_M be the Dirac-delta function on $\mathbb{R}$ with its mass at M, and let $\pi_{\neg M} = \phi_{M,\sigma_0^2}$ denote the normal density function of zero mean and variance $\sigma_0^2 = 10^6$. Thus, the observation x is modeled as coming from some normal distribution of unit variance and some unknown mean, with each possible value of that mean corresponding to a model and to an alternative model in which the mean has a diffuse prior centered at the same mean. Consequently, the Bayes factor favoring the model that says x was drawn from a distribution of density $\phi_{M,1}$ is

$$B(M;x) = \frac{\phi_{M,1}(x)}{\int \phi_{\mu,1}(x)\, \phi_{M,\sigma_0^2}(\mu)\, d\mu}.$$

According to Table 1, the set of models against which there is not at least strong evidence is the set of -3-adequate models,

$$\mathcal{M}(-3) = \{M \in \mathbb{R} : w(M) > -3\} = \left\{M \in \mathbb{R} : \phi_{M,1}(x) > 2^{-3} \int \phi_{\mu,1}(x)\, \phi_{M,\sigma_0^2}(\mu)\, d\mu\right\}.$$

▲

Example 2. Under various classes of weak conditions (Kass and Wasserman, 1995), the weight of evidence may be approximated by

$$w_{\text{BIC}}(M) = \log \exp\left(-\frac{s(M) - s(\neg M)}{2}\right) = \log \frac{n^{-\frac{\nu_M}{2}} f_M\left(x|\widehat{\theta}_M\right)}{n^{-\frac{\nu_{\neg M}}{2}} f_{\neg M}\left(x|\widehat{\theta}_{\neg M}\right)}, \tag{5}$$

where n is the number of independent observations constituting x, $\widehat{\theta}_M$ is the maximum likelihood estimate $\arg\sup_{\theta\in\Theta_M} f_M\left(x|\theta\right)$, ν_M is the number of free parameters in model M, and $s\left(M\right)$ is the Bayesian information criterion (BIC) $-2\ln f_M\left(x|\widehat{\theta}_M\right)+\nu_M\ln n$, and similarly for model $\neg M$ (Carlin and Louis, 2009, p. 53). If the approximation were exact, the set of a-adequate models would be

$$\begin{aligned}\mathcal{M}_{\mathrm{BIC}}\left(a\right) &= \left\{M\in\mathcal{M}:w_{\mathrm{BIC}}\left(M\right)>a\right\}\\ &= \left\{M\in\mathcal{M}:f_M\left(x|\widehat{\theta}_M\right)>2^a n^{\frac{\nu_M-\nu_{\neg M}}{2}}f_{\neg M}\left(x|\widehat{\theta}_{\neg M}\right)\right\}.\end{aligned}\tag{6}$$

The set $\mathcal{M}_{\mathrm{BIC}}\left(a\right)$ is similar to $\mathcal{N}_{\mathrm{BIC}}\left(a\right)=\left\{M\in\mathcal{M}:s\left(M\right)<b+\underline{s}\right\}$, where $b=-2a\ln 2$ and $\underline{s}=\inf_{M\in\mathcal{M}}s\left(M\right)$, except that the sets differ in whether they are relative to each alternative model $\neg M$ or to $\underline{s}$, the infinum over all models. In many cases of $\mathcal{M}_{\mathrm{BIC}}\left(a\right)$, the value of $f_{\neg M}\left(x|\widehat{\theta}_{\neg M}\right)$ is the same for all $M\in\mathcal{M}$ (see Remark 2), while $\mathcal{N}_{\mathrm{BIC}}\left(a\right)$ is analogous to $\mathcal{N}_{\mathrm{AIC}}\left(a\right)=\left\{M\in\mathcal{M}:t\left(M\right)<b+\underline{t}\right\}$, the set of models with values of $t\left(M\right)$, the Akaike information criterion (AIC), less than the sum of $\underline{t}=\inf_{M\in\mathcal{M}}t\left(M\right)$ and b (Burnham and Anderson, 2002, pp. 169-171). Whether using $\neg M$ for each $M\in\mathcal{M}$ or an extremum over all $M\in\mathcal{M}$, the set of adequate models can alternatively be based on each of many other model selection criteria (Remark 1), but the former use has calibration advantages (Remark 2).

Example 3. A typical problem of testing N null hypotheses involves reduced data consisting of N test statistics $X=\left(X_1,\ldots,X_N\right)$ of observed values $x=\left(x_1,\ldots,x_N\right)\in\mathbb{R}^N$. Consider the parameter $\theta\in\Theta=\left\{0,1\right\}$ and the hyperparameters $p\left(0\right)\in\mathfrak{P}$, $p\left(1\right)=1-p\left(0\right)$, and $\gamma\in\Gamma$ for some sets $\mathfrak{P}\subseteq\left[0,1\right]$ and $\Gamma\subseteq\mathbb{R}$ and with $M=\left(p\left(0\right),\gamma\right)$ and $\mathcal{M}=\mathfrak{P}\times\Gamma$. Let g_0 and, for any $\gamma\in\Gamma$, g_γ denote probability density functions on $\mathbb{R}$ corresponding to $\theta=0$

and $\theta = 1$, respectively. The probability density at x_i is

$$f_{p(0),\gamma}\left(x_i\right) = p\left(0\right) g_0\left(x_i\right) + p\left(1\right) g_\gamma\left(x_i\right), \tag{7}$$

for all $i = 1, \dots, N$, in agreement with equation (1). Equation (7) specifies the *two-component mixture*, which has been used extensively in empirical Bayes applications to high-dimensional data sets ($N \gg 1$), especially those of genetics and genomics (Allison et al., 2002; Genovese and Wasserman, 2002; McLachlan et al., 2006; Strimmer, 2008; Efron, 2010); see Bickel (2013) for additional references. The model as stated here is parametric, but it can be made nonparametric by instead specifying some large class Γ of probability distributions. Assuming the independence of the test statistics, the MLE based on equation (7) is

$$\left(\widehat{p}\left(0\right), \widehat{\gamma}\right) = \arg \sup_{\left(p(0),\gamma\right) \in \mathfrak{P} \times \Gamma} \prod_{i=1}^{N} f_{p(0),\gamma}\left(x_i\right). \tag{8}$$

For $i = 1, \dots, N$, the random variable ϑ_i is drawn from $\text{Bern}_{p(1)}$, the Bernoulli distribution with success probability $p\left(1\right)$. The statement that $\vartheta_i = 0$ typically is a null hypothesis such as the hypothesis of no effect on the expression of the ith gene or the hypothesis of no association between a trait and the ith genetic variant. According to Bayes's theorem, the posterior probability that the ith of the N null hypotheses is true is the conditional probability that $\vartheta_i = 0$ given $X_i = x_i$:

$$\psi_{p(0),\gamma}\left(x_i\right) = \frac{p\left(0\right) g_0\left(x_i\right)}{f_{p(0),\gamma}\left(x_i\right)}, \tag{9}$$

which, for historical reasons, is called the *local false discovery rate* (LFDR) (Efron et al., 2001; Efron and Tibshirani, 2002).

Let $\mathcal{M}(a) \subseteq \mathfrak{P} \times \Gamma$ denote the set of a-adequate models. For any $i = 1, \ldots, N$, the set of *a-adequate LFDRs* for the ith hypothesis is the corresponding set of posterior probabilities that $\vartheta_i = 0$:

$$\begin{aligned} \Psi_i(a) &= \left\{ \psi_{p(0),\gamma}(x_i) : \gamma \in \Gamma, (p(0), \gamma) \in \mathcal{M}(a) \right\} \qquad (10) \\ &= \left\{ \psi_{p(0),\gamma}(x_i) : \gamma \in \Gamma, w(p(0), \gamma) > a \right\}, \end{aligned}$$

where $w(p(0), \gamma)$ is the $w(M)$ in equation (3) for each $M = (p(0), \gamma)$. If $w(p(0), \gamma)$ is continuous as a function of $p(0)$ and γ, then the set of a-adequate LFDRs for the ith hypothesis is an interval between $\underline{\psi}_i(a) = \inf \Psi_i(a)$ and $\overline{\psi}_i(a) = \sup \Psi_i(a)$.

Example 4. If the BIC approximation of Example 2 holds for the $n = N$ independent observations of Example 3, then equations (6) and (10) give

$$\mathcal{M}_{\mathrm{BIC}}(a) = \left\{ (p(0), \gamma) \in \mathfrak{P} \times \Gamma : \prod_{i=1}^{N} f_{p(0),\gamma}(x_i) > 2^a N^{\frac{0-\nu_{\neg M}}{2}} \prod_{i=1}^{N} f_{\widehat{p}(0),\widehat{\gamma}}(x_i) \right\} \qquad (11)$$

$$\Psi_{\mathrm{BIC},i}(a) = \left\{ \psi_{p(0),\gamma}(x_i) : (p(0), \gamma) \in \mathcal{M}_{\mathrm{BIC}}(a) \right\} \qquad (12)$$

for all $i = 1, \ldots, N$ and $a \in \mathbb{R}$, where $\nu_{\neg M} \in \{1, 2\}$ is the number of degrees of freedom in the maximization of the likelihood function on $\mathfrak{P} \times \Gamma$. A simple application of this example is worked out in detail in Section 5. This example can be studied in terms of possibility theory (Remark 2). ▲

4 Decisions based on a set of adequate models

The theory of Section 2 incorporates uncertainty in prior distributions, thereby overcoming the commonly voiced objections against the Bayesian practice of specifying a single prior distribution in a situation permitting many other prior distributions and against empirical Bayes methods that substitute point estimates of the prior as if the prior were known.

Decision-theoretic approaches can be cautious or incautious given a feasible set of prior distributions or, more generally, models (Hurwicz, 1951; Jaffray, 1989a,b; Bickel, 2012a). Various decision rules that map imprecise probabilities to actions have been studied, including five of the rules compared by Troffaes (2007).

The rest of this section illustrates some of the available rules for making decisions on the basis of $\mathcal{M}(a)$, the set of a-adequate models for a specified $a \in \mathbb{R}$. Since each rule is the second stage of the first-stage rule that determines $\mathcal{M}(a)$ for some value of a, the overall decision procedure corresponds to lexicographic preferences (see Bickel, 2012b, Remark 3).

4.1 Caution toward ambiguity

Let $\mathcal{D}$ denote a set of possible decisions. Jaffray (1989b) and others studied the criterion of Hurwicz (1951), which prescribes decision

$$\delta(a,\kappa) = \arg\inf_{\delta \in \mathcal{D}} \left(\kappa \sup_{M \in \mathcal{M}(a)} E_M(\ell(\vartheta,\delta)) + (1-\kappa) \inf_{M \in \mathcal{M}(a)} E_M(\ell(\vartheta,\delta)) \right), \tag{13}$$

where $E_M(\ell(\vartheta,\delta)) = \int \ell(\theta,\delta)\,\pi_M(\theta)\,d\theta$, and $\kappa \in [0,1]$ is the parameter encoding the attitude of the agent toward ambiguity in $\mathcal{M}(a)$. Weichselberger (2001) and Augustin (2002) refer to this κ as *caution*, with the result that the most cautious attitude ($\kappa = 1$) corresponds to the conditional Γ-minimax strategy (Gilboa and Schmeidler, 1989; Betrò and Ruggeri, 1992).

Example 5. Consider the 0-1 loss function $\ell : \{0,1\} \times \{0,1\} \to \{0, \ell_{\text{I}}, \ell_{\text{II}}\}$ defined such that

$$\ell\left(\theta,\delta\right) = \begin{cases} 0 & \text{if } \delta = \theta \in \{0,1\} \\ \ell_{\text{I}} & \text{if } \delta = 1, \theta = 0 \\ \ell_{\text{II}} & \text{if } \delta = 0, \theta = 1 \end{cases}$$

for $\ell_{\text{I}}, \ell_{\text{II}} > 0$. In hypothesis testing terminology, θ is the hypothesis indicator equal to 0 if the null hypothesis is true or otherwise equal to 1, $\delta = 1$ is the decision to reject the null hypothesis, and ℓ_{I} (resp. ℓ_{II}) is the loss due to making a Type I error (resp. Type II error). Let $p_M\left(\bullet\right)$ denote the probability mass function such that $p_M\left(\theta\right) = \int_{\{\theta\}} \pi_M\left(\theta_0\right) d\theta_0$ for $\theta = 0, 1$, where π_M is a Dirac-delta function. Then

$$\begin{aligned} E_M\left(\ell\left(\vartheta,\delta\right)\right) &= \ell\left(0,\delta\right) p_M\left(0\right) + \ell\left(1,\delta\right) p_M\left(1\right) \\ &= \begin{cases} \ell_{\text{I}} p_M\left(0\right) & \text{if } \delta = 1 \\ \ell_{\text{II}} p_M\left(1\right) & \text{if } \delta = 0, \end{cases} \end{aligned} \tag{14}$$

and equation (13) gives

$$\delta\left(a,\kappa\right) = \arg \inf_{\delta \in \{0,1\}} \begin{cases} \ell_{\text{I}}\left(\kappa \overline{p}\left(0;a\right) + \left(1-\kappa\right) \underline{p}\left(0;a\right)\right) & \text{if } \delta = 1 \\ \ell_{\text{II}}\left(\kappa \overline{p}\left(1;a\right) + \left(1-\kappa\right) \underline{p}\left(1;a\right)\right) & \text{if } \delta = 0, \end{cases} \tag{15}$$

where $\underline{p}\left(\theta;a\right) = \inf_{M \in \mathcal{M}(a)} p_M\left(\theta\right)$ and $\overline{p}\left(\theta;a\right) = \sup_{M \in \mathcal{M}(a)} p_M\left(\theta\right)$ for $\theta = 0, 1$. In the case of maximal caution ($\kappa = 1$),

$$\begin{aligned}\delta\left(a,1\right) &= \arg\inf_{\delta\in\{0,1\}}\begin{cases}\ell_{\mathrm{I}}\overline{p}\left(0;a\right) & \text{if } \delta = 1\\ \ell_{\mathrm{II}}\overline{p}\left(1;a\right) & \text{if } \delta = 0\end{cases}\\ &= \begin{cases}1 & \text{if } \overline{p}\left(1;a\right)/\overline{p}\left(0;a\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}}\\ 0 & \text{if } \overline{p}\left(1;a\right)/\overline{p}\left(0;a\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}}.\end{cases}\end{aligned} \tag{16}$$

Similarly, the opposite extreme of no caution ($\kappa = 0$) yields

$$\delta\left(a,0\right) = \begin{cases}1 & \text{if } \underline{p}\left(1;a\right)/\underline{p}\left(0;a\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}}\\ 0 & \text{if } \underline{p}\left(1;a\right)/\underline{p}\left(0;a\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}}.\end{cases} \tag{17}$$

4.2 E-admissibility

The set of *E-admissible* (Troffaes, 2007) decisions is

$$\mathcal{D}\left(a\right) = \left\{\arg\inf_{\delta\in\mathcal{D}} E_M\left(\ell\left(\vartheta,\delta\right)\right) : M \in \mathcal{M}\left(a\right)\right\},$$

the set of all Bayes rules compatible with one or more of the a-adequate set of models.

Example 6. Consider the 0-1 loss function of Example 5. By equation (14),

$$\begin{aligned}\arg\inf_{\delta\in\{0,1\}} E_M\left(\ell\left(\vartheta,\delta\right)\right) &= \begin{cases}1 & \text{if } p_M\left(1\right)/p_M\left(0\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}}\\ 0 & \text{if } p_M\left(1\right)/p_M\left(0\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}}\end{cases}\\ &= \begin{cases}1 & \text{if } p_M\left(0\right) < \left(1+\ell_{\mathrm{I}}/\ell_{\mathrm{II}}\right)^{-1}\\ 0 & \text{if } p_M\left(0\right) > \left(1+\ell_{\mathrm{I}}/\ell_{\mathrm{II}}\right)^{-1}\end{cases}\end{aligned}$$

for any $M\in\mathcal{M}\left(a\right)$. Thus,

$$\mathcal{D}\left(a\right) = \begin{cases}\{1\} & \text{if } \exists_{M\in\mathcal{M}(a)} p_M\left(0\right) < \beta, \nexists_{M\in\mathcal{M}(a)} p_M\left(0\right) > \beta\\ \{0\} & \text{if } \nexists_{M\in\mathcal{M}(a)} p_M\left(0\right) < \beta, \exists_{M\in\mathcal{M}(a)} p_M\left(0\right) > \beta\\ \{0,1\} & \text{if } \exists_{M\in\mathcal{M}(a)} p_M\left(0\right) < \beta, \exists_{M\in\mathcal{M}(a)} p_M\left(0\right) > \beta,\end{cases} \tag{18}$$

where $\beta = \left(1+\ell_{\mathrm{I}}/\ell_{\mathrm{II}}\right)^{-1}$. ▲

4.3 E-admissibility with caution

Using $\mathcal{D}\left(a\right)$ in place of $\mathcal{D}$ in equation (13), that is,

$$\Delta\left(a,\kappa\right) = \arg\inf_{\delta\in\mathcal{D}(a)}\left(\kappa\sup_{M\in\mathcal{M}(a)} E_M\left(\ell\left(\vartheta,\delta\right)\right) + \left(1-\kappa\right)\inf_{M\in\mathcal{M}(a)} E_M\left(\ell\left(\vartheta,\delta\right)\right)\right), \tag{19}$$

can avoid certain pathologies that can otherwise occur with the Hurwicz criterion (Troffaes, 2007; Seidenfeld, 2004).

Example 7. Consider the 0-1 loss function of Example 5. In the case of maximal caution

$(\kappa = 1)$, equations (18), (16), and (19) yield

$$\begin{aligned}
\Delta\left(a,1\right) &= \arg\inf_{\delta\in\mathcal{D}(a)}\begin{cases}\ell_{\mathrm{I}}\overline{p}\left(0;a\right) & \text{if } \delta = 1\\ \ell_{\mathrm{II}}\overline{p}\left(1;a\right) & \text{if } \delta = 0\end{cases}\\
&= \begin{cases}1 & \text{if } \overline{p}\left(1;a\right)/\overline{p}\left(0;a\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}} \text{ or } \exists_{M\in\mathcal{M}(a)}p_{M}\left(0\right) < \beta, \nexists_{M\in\mathcal{M}(a)}p_{M}\left(0\right) > \beta\\ 0 & \text{if } \overline{p}\left(1;a\right)/\overline{p}\left(0;a\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}} \text{ or } \nexists_{M\in\mathcal{M}(a)}p_{M}\left(0\right) < \beta, \exists_{M\in\mathcal{M}(a)}p_{M}\left(0\right) > \beta\end{cases}\\
&= \begin{cases}1 & \text{if } \left(1-\underline{p}\left(0;a\right)\right)/\overline{p}\left(0;a\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}} \text{ or } \underline{p}\left(0;a\right) < \beta, \overline{p}\left(0;a\right) \leq \beta\\ 0 & \text{if } \left(1-\underline{p}\left(0;a\right)\right)/\overline{p}\left(0;a\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}} \text{ or } \underline{p}\left(0;a\right) \geq \beta, \overline{p}\left(0;a\right) > \beta\end{cases}\\
&= \begin{cases}1 & \text{if } \left(1-\underline{p}\left(0;a\right)\right)/\overline{p}\left(0;a\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}}\\ 0 & \text{if } \left(1-\underline{p}\left(0;a\right)\right)/\overline{p}\left(0;a\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}}.\end{cases} \qquad (20)
\end{aligned}$$

This applies directly to each hypothesis of the N-hypothesis framework of Example 3 if the loss function is additive, that is, if the total loss is $\sum_{i=1}^{N}\Delta\left(a,1\right)$, where $\Delta_{i}\left(a,1\right)$ is the optimal decision for ϑ_{i} according to equation 19. With the substitutions $\underline{p}\left(0;a\right) = \underline{\psi}_{i}\left(a\right)$ and $\overline{p}\left(0;a\right) = \overline{\psi}_{i}\left(a\right)$, equation (20) gives the optimal decision for whether the ith null hypothesis is rejected $(\Delta_{i}\left(a,1\right) = 1)$ or not $(\Delta_{i}\left(a,1\right) = 0)$:

$$\Delta_{i}\left(a,1\right) = \begin{cases}1 & \text{if } \left(1-\underline{\psi}_{i}\left(a\right)\right)/\overline{\psi}_{i}\left(a\right) > \ell_{\mathrm{I}}/\ell_{\mathrm{II}}\\ 0 & \text{if } \left(1-\underline{\psi}_{i}\left(a\right)\right)/\overline{\psi}_{i}\left(a\right) < \ell_{\mathrm{I}}/\ell_{\mathrm{II}}.\end{cases} \qquad (21)$$

Thus, if $\overline{\psi}_{i}\left(a\right) = 1$,

$$\Delta_{i}\left(a,1\right) = \begin{cases}1 & \text{if } \underline{\psi}_{i}\left(a\right) < 1-\ell_{\mathrm{I}}/\ell_{\mathrm{II}}\\ 0 & \text{if } \underline{\psi}_{i}\left(a\right) > 1-\ell_{\mathrm{I}}/\ell_{\mathrm{II}}.\end{cases}$$

Hence, $\overline{\psi}_i(a) = 1$ with maximal caution ($\kappa = 1$) prevents rejecting the ith null hypothesis unless $\ell_{\mathrm{I}} < \ell_{\mathrm{II}}$. This case is important since $\overline{\psi}_i(a) = 1$ for all $\gamma \in \Gamma$ and all $i = 1, \ldots, N$ if 1 is an a-adequate value of $p(0)$, that is, if there is a $\gamma \in \Gamma$ such that $(1, \gamma) \in \mathcal{M}(a)$ or, equivalently, $f_{1,\gamma}(x) \geq a f_{\widehat{p}(0),\widehat{\gamma}}(x)$.

Analogously, equation (17), with $\underline{p}(1; a) = 1 - \overline{\psi}_i(a)$ and $\underline{p}(0; a) = \underline{\psi}_i(a)$, gives the the case of minimal caution ($\kappa = 0$). Again assuming $\overline{\psi}_i(a) = 1$, the odds in equation (17) is necessarily 0, that is, $\underline{p}(1; a) / \underline{p}(0; a) = 0$, preventing rejecting the ith null hypothesis regardless of the value of $\ell_{\mathrm{I}}/\ell_{\mathrm{II}}$.

Still assuming $\overline{\psi}_i(a) = 1$, intermediate caution ($0 < \kappa < 1$) instead brings the decision

$$\begin{aligned} \Delta_i(a, \kappa) &= \inf_{\delta \in \mathcal{D}(a)} \begin{cases} \ell_{\mathrm{I}} \left(\kappa \times 1 + (1 - \kappa) \underline{p}(0; a) \right) & \text{if } \delta = 1 \\ \ell_{\mathrm{II}} \left(\kappa \left(1 - \underline{p}(0; a) \right) + (1 - \kappa) \times 0 \right) & \text{if } \delta = 0, \end{cases} \\ &= \begin{cases} 1 & \text{if } \underline{\psi}_i(a) < \psi^\star(\kappa) \\ 0 & \text{if } \underline{\psi}_i(a) > \psi^\star(\kappa), \end{cases} \end{aligned}$$

where $\psi^\star(\kappa) = \left(1 + \kappa^{-1} \left(\ell_{\mathrm{II}}/\ell_{\mathrm{I}} - 1 \right)^{-1} \right)^{-1}$, as seen from equation (15) with $1 - \underline{p}(1; a) = \overline{p}(0; a) = \overline{\psi}_i(a) = 1$. Note that $\psi^\star(\kappa) > 0$ only if $1 - \kappa^{-1} \left(1 - \ell_{\mathrm{II}}/\ell_{\mathrm{I}} \right)^{-1} > 0$, which would imply that $\left(1 - \ell_{\mathrm{II}}/\ell_{\mathrm{I}} \right)^{-1} < \kappa$, but $\ell_{\mathrm{I}} \geq \ell_{\mathrm{II}}$ entails $\left(1 - \ell_{\mathrm{II}}/\ell_{\mathrm{I}} \right)^{-1} > 1$ and thus leads to the contradiction $\left(1 - \ell_{\mathrm{II}}/\ell_{\mathrm{I}} \right)^{-1} > \kappa$. Thus, $\ell_{\mathrm{I}} \geq \ell_{\mathrm{II}}$ requires that $\psi^\star(\kappa) \leq 0$, which means $\Delta_i(a, \kappa) = 0$, even for the smallest values of $\underline{\psi}_i(a)$. In conclusion, except in the unusual situation that a Type I error is less harmful than a Type II error ($\ell_{\mathrm{I}} < \ell_{\mathrm{II}}$), the case of $\overline{\psi}_i(a) = 1$ prevents rejecting the ith null hypothesis regardless of the value of κ, with the consequence that no null hypothesis can be rejected if 1 is the first component any a-adequate model $(p(0), \gamma) \in \mathcal{M}(a)$. ▲

4.4 Combining the adequate posterior distributions

Another approach to making decisions on the basis of the a-adequate models is to combine their posterior distributions into a single distribution for Bayes rule, the minimization of posterior expected loss. A *distribution-combination method* is a function that transforms each set $\mathcal{P}$ of distributions to a combined distribution $\pi^{\mathcal{P}}$ on the same domain.

Several distribution-combination methods have been proposed (Genest and Zidek, 1986). Many of them were originally designed to combine the opinions of multiple experts, with each opinion represented by a probability distribution parameter space (Cooke, 1991, Ch. 11). For example, minimizing a weighted sum of divergences from the distributions being combined yields a linear combination of the distributions (Toda, 1956; Kracík, 2011). Any linearly combined marginal distribution is the same whether marginalization or combination is carried out first (McConway, 1981). Alternatively, a weighted multiplicative combination of the distributions is invariant to the order of combination and Bayesian updating (Berger, 1985, §4.11.1).

A common approach to combining non-subjective distributions maximizes the entropy over the plausible set of distributions to be combined relative to some benchmark distribution or other base measure (Grünwald and Dawid, 2004; Bickel, 2012a). Other information-theoretic methods (Ryabko, 1979; Gallager, 1979; Davisson and Leon-Garcia, 1980), also used as methods of distribution combination (Bickel, 2012b), do not require a base measure.

For use with a distribution-combination method $\pi^{\bullet}$, the set of *a-adequate posterior distributions* is defined as $\mathcal{P}_{a|x} = \{\pi_M(\bullet|x) : M \in \mathcal{M}(a)\}$, in analogy with equation (4). Thus, the Bayes rule for $\pi^{\mathcal{P}_{a|x}}$, the probability density function that combines the a-adequate posterior distributions, delivers $\arg\inf_{\delta\in\mathcal{D}} \int \ell(\theta,\delta)\,\pi^{\mathcal{P}_{a|x}}(\theta)\,d\theta$ as the optimal decision.

In the case that the base measure ϖ in an entropy-type method (Grünwald and Dawid,

2004) is a posterior distribution (Bickel, 2012a), $\pi^{\mathcal{P}_{a|x}} = \varpi$ if $\varpi \in \mathcal{P}_{a|x}$; otherwise, $\pi^{\mathcal{P}_{a|x}}$ is the distribution in $\mathcal{P}_{a|x}$ that is most similar to ϖ in an information-theoretic sense. This thereby implements not only the Bayesian practice of using a single posterior ϖ for inference given its passing a model check but also prescribes the posterior to use for inference in the event that ϖ fails the model check. In short, use the initial posterior distribution if it is assessed to be a-adequate or the a-adequate posterior as close as possible to the initial posterior if not. This overcomes an objection Lele and Dennis (2009) raised against the Bayesian approach but still relies heavily on the specification of a single model (§1), the one leading to ϖ.

5 Case study for a set of adequate models

This section uses a multi-dimensional data set to illustrate the framework proposed in Sections 2 and 4 for the special case of hypothesis testing. Specifically, educational data illustrate the application of sets of adequate models to the two-component mixture (Example 3). Fig. 1 displays the standardized mean SAT exam score difference between students participating in a training program and students not participating in the program for each of $N = 8$ exam sites, as reported by Rubin (1981).

Suppose x_i, the square of the ith test statistic, is χ^2 with 1 degree of freedom if the ith null hypothesis is true (meaning there is no effect of training at site i) but is noncentral χ^2 with 1 degree of freedom and noncentrality parameter γ if the ith alternative hypothesis is true (meaning there is an effect of training at site i). Assume a per-site prior probability of 50% that the training affects the SAT exam score. In the notation of Example 3, the null and alternative hypotheses correspond to $\vartheta_i = 0$ and $\vartheta_i = 1$, respectively; $\mathfrak{P} = [50\%, 50\%]$, $\Gamma =]0, \infty[$, and g_0 (resp. g_γ) are the central (resp. noncentral) χ^2 density functions. Further, suppose the BIC approximation holds for the 8 independent observations, as in Example 4. Consequently, equations (8), (11), and (12) reduce to

$$\widehat{\gamma} = \arg\sup_{\gamma>0} \prod_{i=1}^{8} f_{50\%,\gamma}\left(x_i\right)$$

$$\Gamma_{\text{BIC}}\left(a\right) = \left\{\gamma > 0 : \prod_{i=1}^{8} f_{50\%,\gamma}\left(x_i\right) > 2^a 8^{-\frac{1}{2}} \prod_{i=1}^{8} f_{50\%,\widehat{\gamma}}\left(x_i\right)\right\}$$

$$\mathcal{M}_{\text{BIC}}\left(a\right) = \left[50\%, 50\%\right] \times \Gamma_{\text{BIC}}\left(a\right)$$

$$\Psi_{\text{BIC},i}\left(a\right) = \left\{\psi_{50\%,\gamma}\left(x_i\right) : \gamma \in \Gamma_{\text{BIC}}\left(a\right)\right\}$$

for all $i = 1, \ldots, 8$ and $a \in \mathbb{R}$.

While Fig. 2 indicates that γ, the noncentrality parameter value of any sites with a training effect, could be as high as 7.2, Figs. 3 and 4 reveal the inadequacy of models assigning an LFDR value lower than 25% to any site. Thus, no site has more than 75% posterior probability of a training effect. In fact, models in which all sites have a 50% LFDR are 1-adequate.

Based on the values in the captions of Figs. 3 and 4, applying the maximally cautious ($\kappa = 1$) decision rule of Example 7 and the standard of strong evidence ($a = -3$) uses $\left(1 - \underline{\psi}_1\left(-3\right)\right)/\overline{\psi}_1\left(-3\right) = 3/2$ and $\left(1 - \underline{\psi}_2\left(-3\right)\right)/\overline{\psi}_2\left(-3\right) = 5/9$ in equation (21). The null hypothesis that the SAT exam score is not affected by the training at Site A (resp. Site B) is only rejected if $\ell_{\text{I}}/\ell_{\text{II}} < 3/2$ (resp. $\ell_{\text{I}}/\ell_{\text{II}} < 5/9$). In the case of equal losses for Type I and Type II errors ($\ell_{\text{I}} = \ell_{\text{II}}$), the conclusion is that the training affected the score at Site A but not at Site B.

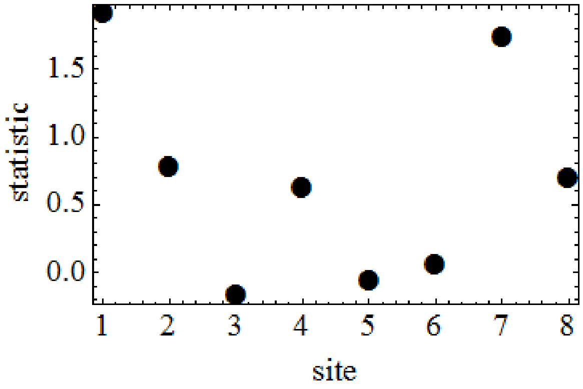

Figure 1: Standardized effect of training on exam scores.

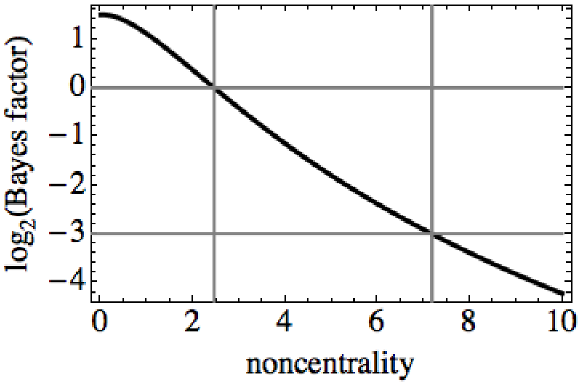

Figure 2: The logarithm of the Bayes factor versus the value of the noncentrality parameter. The set of noncentrality parameter values against which there is no evidence or no strong evidence is $\Gamma_{\mathrm{BIC}}(0) =]0, 2.5]$ or $\Gamma_{\mathrm{BIC}}(-3) =]0, 7.2]$, respectively, according to Table 1.

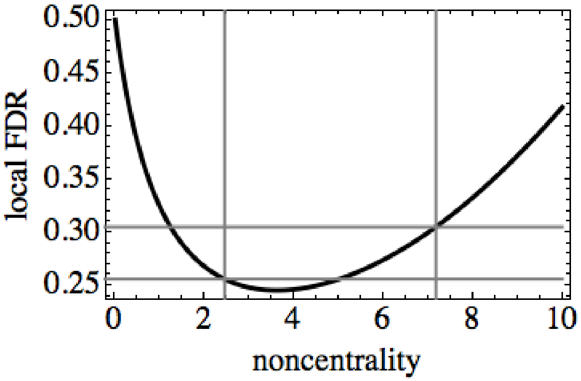

Figure 3: The logarithm of the local false discovery rate (LFDR) versus the value of the noncentrality parameter for Site A. The set of LFDR values against which there is no evidence or no strong evidence is $\left[\underline{\psi}_1(0), \overline{\psi}_1(0)\right[= [26\%, 50\%[$ or $\left[\underline{\psi}_1(-3), \overline{\psi}_1(-3)\right[= [25\%, 50\%[$, respectively, according to Fig. 2 and Table 1.

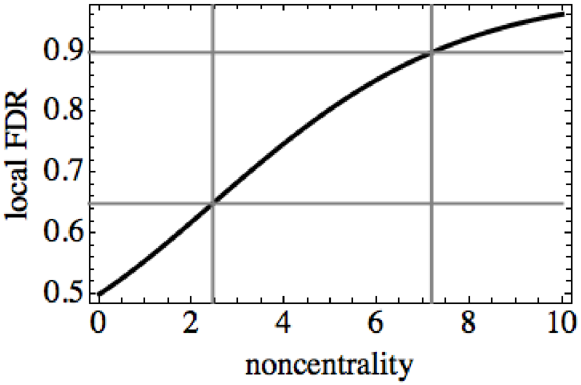

Figure 4: The logarithm of the local false discovery rate (LFDR) versus the value of the noncentrality parameter for Site B. The set of LFDR values against which there is no evidence or no strong evidence is $\left]\underline{\psi}_2(0), \overline{\psi}_2(0)\right] =]50\%, 65\%]$ or $\left]\underline{\psi}_2(-3), \overline{\psi}_2(-3)\right] =]50\%, 90\%]$, respectively, according to Fig. 2 and Table 1.

6 Remarks

Remark 1. The methodology of this paper can be generalized to any method of eliminating models that do not adequately agree with the observed data. For example, the weight of evidence can be replaced with various measures of statistical support (Bickel, 2013) or other minimum description lengths (Grünwald, 2007), and the BIC in equation (5) can be replaced by the AIC (Burnham and Anderson, 2002), by the deviance information criterion (Spiegelhalter et al., 2002), or by another penalized likelihood (Sin and White, 1996).

Eliminating models with p-values less than a significance threshold is isomorphic to considering a confidence set as the set of adequate models to which decision theory may be applied, a practice equivalent to that of Kyburg and Teng (2001, §9.8) and Gajdos et al. (2004, Example 3). In fact, Box (1980) argued that p-values have a valid role in testing Bayesian models, including their prior distributions. Specifically, he proposed prior predictive statistics for Bayesian model checking. Rubin (1984), Meng (1994), Bayarri and Berger (2000), and others developed variations of the p-value for Bayesian model assessment. See Bernardo and Smith (1994, pp. 409-417), Bayarri and Berger (2004, §4.3), Little (2006), and Gelman and Shalizi (2013, §4) for reviews. Kelly and Smith (2011, §3.5.7) described a software implementation of a p-value as a function of the preposterior distribution. A closely related approach assesses agreement between the prior distribution and the data via an invariant p-value based on the observed probability density of the data as the test statistic (Evans and Jang, 2010). Once models that fail the predictive checks or that otherwise have sufficiently low p-values are eliminated from consideration, the methods of Section 4 may be applied to the remaining set of adequate models.

Remark 2. If, as in Example 4, there is a constant $f_{\neg\star}(x)$ such that $f_{\neg\star}(x) = f_{\neg M}(x)$ for all $M \in \mathcal{M}$, then equation (3) yields

$$\mathcal{M}(a) = \left\{ M \in \mathcal{M} : \text{poss}(M) > \left(f_{\neg\star}(x) / \overline{f(x)} \right) 2^a \right\}, \tag{22}$$

where poss : $\mathcal{M} \to [0,1]$ is defined by $\text{poss}(M) = f_M(x) / \overline{f(x)}$ and $\overline{f(x)} = \sup_{M \in \mathcal{M}} f_M(x)$ for all $M \in \mathcal{M}$. Since poss has a range of $[0,1]$ and since $\sup_{M \in \mathcal{M}} \text{poss}(M) = 1$, the function poss is a *possibility profile* defining the *possibility measure* Poss, a function given by $\text{Poss}(\mathcal{M}_0) = \sup_{M \in \mathcal{M}_0} \text{poss}(M)$ for any measurable subset $\mathcal{M}_0 \subseteq \mathcal{M}$ (Wang, 2008). In this context, Bickel (2012c) called $\text{Poss}(\mathcal{M}_0) / \text{Poss}(\mathcal{M}_1)$ the "strength of statistical evidence" for the hypothesis that $\theta \in \mathcal{M}_0$ relative to the hypothesis that $\theta \in \mathcal{M}_1$, with $\mathcal{M}_0$ and $\mathcal{M}_1$ in the same σ-field of subsets of $\mathcal{M}$. Due to the calibrating factor of $f_{\neg\star}(x) / \overline{f(x)}$ in equation (22), $\mathcal{M}(a)$ differs from

$$\mathcal{N}(a) = \{ M \in \mathcal{M} : \text{poss}(M) > 2^a \}$$

for the same value of a in that $\mathcal{M}(a)$ avoids pathologies and biases associated with the profile likelihood that result from dependence on $\overline{f(x)}$ (e.g., Berger et al., 1999; Walley and Moral, 1999; Bickel, 2013). Nevertheless, the use of $\mathcal{N}(a)$ with decision rules such as those of Section 4 represents a novel operational interpretation of the possibility measure, differing from treating it as an upper probability measure, the approach of Walley (1997) and of de Cooman and Walley (2002). The case of $\mathcal{N}(a)$ that appears in Example 2 as $\mathcal{N}_{\text{BIC}}(a)$ puts the BIC in place of the AIC in $\mathcal{N}_{\text{AIC}}(a)$ of the same example.

Acknowledgments

This research was partially supported by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C.-K., Prolla, T. A., Weindruch, R., 2002. A mixture model approach for the analysis of microarray gene expression data. Computational Statistics and Data Analysis 38, 1–20.

Ando, T., 2010. Bayesian Model Selection and Statistical Modeling. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

Augustin, T., 2002. Expected utility within a generalized concept of probability - a comprehensive framework for decision making under ambiguity. Statistical Papers 43 (1), 5–22.

Bayarri, M., Berger, J., 2000. P values for composite null models. Journal of the American Statistical Association 95 (452), 1127–1142.

Bayarri, M. J., Berger, J. O., 2004. The interplay of Bayesian and frequentist analysis. Statistical Science 19, 58–80.

Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Berger, J. O., Liseo, B., Wolpert, R. L., 1999. Integrated likelihood methods for eliminating nuisance parameters. Statistical Science 14, 1–28.

Bernardo, J. M., Smith, A. F. M., 1994. Bayesian Theory. John Wiley and Sons, New York.

Betrò, B., Ruggeri, F., 1992. Conditional Γ-minimax actions under convex losses. Communications in Statistics - Theory and Methods 21, 1051–1066.

Bickel, D. R., 2011. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Canadian Journal of Statistics 39, 610–631.

Bickel, D. R., 2012a. Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes. Electron. J. Statist. 6, 686–709.

Bickel, D. R., 2012b. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. International Journal of Approximate Reasoning 53, 880–891.

Bickel, D. R., 2012c. The strength of statistical evidence for composite hypotheses: Inference to the best explanation. Statistica Sinica 22, 1147–1198.

Bickel, D. R., 2013. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. International Statistical Review 81, 188–206.

Box, G., 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society A 143, 383–430.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.

Carlin, B. P., Louis, T. A., 2009. Bayesian Methods for Data Analysis, Third Edition. Chapman & Hall/CRC, New York.

Cooke, R. M., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press.

Davisson, L., Leon-Garcia, a., 1980. A source matching approach to finding minimax codes. IEEE Transactions on Information Theory 26, 166–174.

de Cooman, G., Walley, P., 2002. A possibilistic hierarchical model for behaviour under uncertainty. Theory and Decision 52 (4), 327–374.

Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, Cambridge.

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70–86.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151–1160.

Evans, M., Jang, G. H., Feb. 2010. Invariant P-values for model checking. The Annals of Statistics 38 (1), 512–525.

Gajdos, T., Tallon, J. M., Vergnaud, J. C., SEP 2004 2004. Decision making with imprecise probabilistic information. Journal of Mathematical Economics 40, 647–681.

Gallager, R. G., 1979. Source coding with side information and universal coding. Technical Report LIDS-P-937, Laboratory for Information Decision Systems, MIT.

Gelman, A., Shalizi, C. R., 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical & Statistical Psychology 66 (1), 8–38.

Genest, C., Zidek, J. V., 1986. Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science 1, 114–135.

Genovese, C., Wasserman, L., 2002. Operating characteristics and extensions of the false discovery rate procedure. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64, 499–517.

Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18 (2), 141–153.

Giron, F. J., Rios, S., 1980. Quasi-Bayesian Behaviour: A more realistic approach to decision making? Trabajos de Estadistica Y de Investigacion Operativa 31 (1), 17–38.

Good, I. J., 1979. Studies in the history of probability and statistics. xxxvii a. m. turing's statistical work in world war ii. Biometrika 66 (2), 393–396.

Grünwald, P., Dawid, A. P., 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals of Statistics 32, 1367–1433.

Grünwald, P. D., 2007. The Minimum Description Length Principle. MIT Press, London.

Hurwicz, L., 1951. Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper 370.

Jaffray, J.-Y., 1989a. Généralisation du critère de l'utilité espérée aux choix dans l'incertain régulier. RAIRO: Recherche opérationnelle 23, 237–267.

Jaffray, J.-Y., 1989b. Linear utility theory for belief functions. Operations Research Letters 8, 107–112.

Jeffreys, H., 1948. Theory of Probability. Oxford University Press, London.

Kass, R. E., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. Journal of the American Statistical Association 90, 928–934.

Kelly, D., Smith, C., 2011. Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook. Springer Series in Reliability Engineering. Springer, New York.

Kracík, J., 2011. Combining marginal probability distributions via minimization of weighted sum of Kullback-Leibler divergences. International Journal of Approximate Reasoning 52, 659–671.

Kyburg, H. E., Teng, C. M., 2001. Uncertain Inference. Cambridge University Press, Cambridge.

Lele, S., Dennis, B., 2009. Bayesian methods for hierarchical models: Are ecologists making a faustian bargain? Ecological Applications 19, 581–584.

Little, R. J., 2006. Calibrated Bayes: a Bayes/frequentist roadmap. The American Statistician 60 (3), 213–223.

McConway, K. J., 1981. Marginalization and linear opinion pools. Journal of the American Statistical Association 76, 410–414.

McLachlan, G., Bean, R., Jones, L.-T., 2006. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics 22 (13), 1608–1615.

Meng, X.-L., 1994. Posterior predictive p-values. The Annals of Statistics 22, 1142–1160.

Psillos, S., 2002. Simply the best: A case for abduction. Computational Logic: From Logic Programming into the Future, 605–625.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. CRC Press, New York.

Rubin, D. B., 1981. Estimation in parallel randomized experiments. Journal of Educational Statistics 6, pp. 377–401.

Rubin, D. B., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann.Statist. 12, 1151–1172.

Ryabko, B., 1979. Encoding of a source with unknown but ordered probabilities. Prob. Pered. Inform. 15, 71–77.

Seidenfeld, T., 2004. A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibility. Synthese 140, 69–88.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207–225.

Spiegelhalter, D., Best, N., Carlin, B., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64, 583–616.

Strimmer, K., 2008. A unified approach to false discovery rate estimation. BMC Bioinformatics 9, art. 303.

Toda, M., 1956. Information-receiving behavior of man. Psychological Review 63, 204–212.

Troffaes, M. C. M., 2007. Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning 45 (1), 17–29.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.

Walley, P., 1997. Statistical inferences based on a second-order possibility distribution. International Journal of General Systems 26 (4), 337–383.

Walley, P., Moral, S., 1999. Upper probabilities based only on the likelihood function. Journal of the Royal Statistical Society.Series B: Statistical Methodology 61, 831–847.

Wang, Q., 2008. Probability distribution and entropy as a measure of uncertainty. Journal of Physics A: Mathematical and Theoretical 41, 065004.

Weichselberger, K., 2001. Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept. Physica-Verlag, Heidelberg.