Calibration results for Bayesian model specification

David Draper* and Milovan Krnjajić†

9 Apr 2013

Abstract: When the goal is inference about an unknown θ and prediction of future data D^* on the basis of data D and background assumptions/judgments $\mathcal{B}$, the process of Bayesian model specification involves two ingredients: the conditional probability distributions $p(\theta|\mathcal{B})$ and $p(D|\theta,\mathcal{B})$. Here we focus on specifying $p(D|\theta,\mathcal{B})$, and we argue that *calibration* considerations — paying attention to how often You get the right answer — should be an integral part of this specification process. After contrasting Bayes-factor-based and predictive model-choice criteria, we present some calibration results, in fixed- and random-effects Poisson models, relevant to addressing two of the basic questions that arise in Bayesian model specification: (Q_1) Is model M_j better than $M_{j'}$? and (Q_2) Is model M_{j^*} good enough? In particular, we show that LS_{FS}, a *full-sample log score* predictive model-choice criterion, has better small-sample model discrimination performance than either DIC or a cross-validation-style log-scoring criterion, in the simulation setting we consider; we examine the large-sample behavior of LS_{FS}; and we (a) demonstrate that the popular *posterior predictive tail-area* method for answering a question related to Q_2 can be poorly calibrated, and (b) document the success of a method for calibrating it.

Keywords: Asymptotics, Bayesian and frequentist probability paradigms, Bayes factors, BIC, cross-validation log score, DIC, fixed- and random-effects Poisson modeling, foundations of statistics, full-sample log score, logical consistency, posterior predictive tail areas, relevance of point-null hypothesis testing.

1 Introduction

1.1 The role of calibration in Bayesian modeling

We begin at the beginning, to fix notation and ideas. In the Bayesian modeling paradigm for inference, prediction and decision-making, there are three fundamental ingredients: θ, something unknown or only partially known to You (a generic person wishing to reason sensibly in the face of uncertainty; Good (1950)); D, an information (data) source that You judge to be relevant to decreasing Your uncertainty about θ; and $\mathcal{B}$, a set of propositions (true-false statements) summarizing Your background assumptions and judgments about relevant aspects of θ (e.g., that $\theta > 0$ if θ represents the mean remission time for a specified set of patients with a given disease) and

*David Draper is with the Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064 USA; email draper@ams.ucsc.edu

†Milovan Krnjajić is is a Stokes Lecturer of Statistics in the School of Mathematics, Statistics and Applied Mathematics at the National University of Ireland, Galway; email milovan.krnjajic@nuigalway.ie

D (e.g., that the data set arose as the result of a randomized controlled trial with the following design: ...).

In this framework the physicist R. T. Cox (1946, 1961) (also see Jaynes (2003)) began with a conditional plausibility operator $p(A|B)$ acting on propositions A and B and developed from this a full probability calculus in which, for example, propositions such as $\theta \leq t$ for real-valued θ bring the usual machinery of cumulative distribution functions and densities to bear on the process of uncertainty quantification. Cox proved that — under a reasonable set of axioms involving internal logical consistency and representation of degrees of plausibility by real numbers — Your uncertainty quantification, if You wish to be rational, should be based (for inference and prediction) on the conditional probability distributions $p(\theta|\mathcal{B})$ and $p(D|\theta, \mathcal{B})$; for decision-making, it had previously been shown by Ramsey (1926) that the only other relevant ingredients are a set $\mathcal{A}$ of possible actions a and a *utility function* $U(a, \theta^*)$, expressing (in real-valued terms) the gain that would result if You chose action a and the unknown θ actually took the value θ^*. In this formulation $p(\theta|\mathcal{B})$ — usually called Your *prior distribution*, although temporal considerations need not arise in specifying it — quantifies all of Your information about θ *external* to D, and $p(D|\theta, \mathcal{B})$ — typically referred to as Your *sampling distribution*, even though it is not necessary in this approach to consider other data sets that might have been observed but were not — quantifies Your predictive uncertainty about D given θ, before D has arrived. Note that Cox's use of the phrase *logical consistency* has nothing to do with the repeated-sampling idea of *asymptotic consistency* (see Section 5 for asymptotic considerations).

It then follows, from Cox's and Ramsey's results in this framework, that the three basic statistical activities of inference, prediction and decision-making are each governed by a single equation (familiar in Bayesian work), as follows:

- (inference) $p(\theta|D, \mathcal{B}) = c\, p(\theta|\mathcal{B})\, p(D|\theta, \mathcal{B})$, in which c is a positive normalizing constant and $p(\theta|D, \mathcal{B})$ — usually called Your *posterior distribution*, although again temporal considerations are not central to the formulation — quantifies the totality of Your information about θ, both internal and external to D;

- (prediction) $p(D^*|D, \mathcal{B}) = \int_\Theta p(D^*|\theta, D, \mathcal{B})\, p(\theta|D, \mathcal{B})\, d\theta$, where D^* is future data and Θ is the space over which You acknowledge Your uncertainty about θ. Often D provides no additional information about D^* if θ is known, in which case this equation simplifies to $p(D^*|D, \mathcal{B}) = \int_\Theta p(D^*|\theta, \mathcal{B})\, p(\theta|D, \mathcal{B})\, d\theta$ and involves only previously-defined quantities; and

- (decision) The optimal action a^* is given by

$$a^* = (a^*|D, \mathcal{B}) = argmax_{a \in \mathcal{A}} E_{(\theta|D,\mathcal{B})} U(a, \theta) = argmax_{a \in \mathcal{A}} \int_\Theta U(a, \theta)\, p(\theta|D, \mathcal{B})\, d\theta \ ;$$

  in other words, a^* is the action that maximizes expected utility, where the expectation is over Your posterior uncertainty about θ given D.

<u>*Remark.*</u> De Finetti (1974) obtained the same foundational results from different premises, involving betting odds as the primitive concept for probabilities and *coherence* (a desire not to offer bets that would guarantee You a monetary loss) instead of Cox's logical consistency axioms. De Finetti's approach may be seen as an extension of the original probability notion of Bayes (1763): in Bayes's view, Your assessment of $p(A|\mathcal{B})$, for a true-false proposition A, is equal to some value π if You judge Yourself indifferent between (a) receiving $\pi \cdot m$ monetary units for sure (for

some small $m > 0$) and (b) betting with someone in such a way that You will get m monetary units if A turns out to be true and nothing if not. We prefer (cf. Jaynes (2003)) to regard probability as a measure of the weight of evidence in favor of the truth of A — an extension of Boolean true-false logic to settings in which You are uncertain about A's truth — without an appeal to betting odds. ♠

It is clear that, in this framework, in a given application everything comes down to a series of specification tasks: You have to specify $p(\theta|\mathcal{B})$ and $p(D|\theta, \mathcal{B})$ for inference and prediction, and additionally $\mathcal{A}$ and $U(a, \theta)$ for decision-making. It is natural to refer to this set of tasks as *Bayesian model specification*. We focus here on the specification of the prior predictive (sampling) distribution $p(D|\theta, \mathcal{B})$, not because it is the only important ingredient but simply for lack of space; see, e.g., O'Hagan and Forster (2004) and Parmigiani and Inoue (2009) for useful suggestions on specifying $p(\theta|\mathcal{B})$ and $\{\mathcal{A}, U(a, \theta)\}$, respectively.

An important question remains: what *principles* should govern the process of specifying $p(D|\theta, \mathcal{B})$? An appeal to logical consistency is not helpful; any (proper) choice will yield logically consistent results as long as the three equations above are employed. In uncomplicated settings, $p(D|\theta, \mathcal{B})$ may arise directly from the background of the data-gathering process (an example is Your judgment of *exchangeability* (De Finetti (1930); e.g., Draper (2008)) of binary observables, leading — via the simplest of De Finetti's representation theorems — essentially uniquely to a Bernoulli sampling distribution), but in problems of realistic complexity You will typically be uncertain about how to specify $p(D|\theta, \mathcal{B})$. In our view, a central principle guiding the specification of $p(D|\theta, \mathcal{B})$ should be *calibration*: a desire on Your part for uncertainty quantifications by You such as $p(u < \theta < v|D, \mathcal{B}) = 0.9$ (for real-valued θ, and taking 0.9 for illustration) and $p(u < f(D^*) < v|D, \mathcal{B}) = 0.9$ (for a real-valued function $f(D^*)$ of future data D^*) to be verifiably correct approximately 90% of the time. Taking as an axiom that You want to help positively advance the course of science, we base this view on the following observations:

- It appears to be a fundamental scientific question, independent of the meaning of probability, to wonder how often Your methods for uncovering truth are getting the right answer, and
- There is nothing in the Bayesian paradigm to prevent You from making one or both of the following mistakes — (a) choosing $p(D|\theta, \mathcal{B})$ unwisely; (b) inserting strong (prior) information about θ external to D that turns out after the fact to have been out of step with reality — and, if You repeatedly insist on doing so, (i) it would seem likely that Your colleagues will stop inviting You into their projects as a statistical collaborator and (ii) this runs counter to Your axiomatic desire to aid in the scientific enterprise.

Remark. An interest in calibration may appear to bring a non-Bayesian (repeated-sampling, frequentist) element into the overall story; in response to this possible perception, we would make two comments:

- The Bayesian probability paradigm has been explored since the early work of Bayes (1763) and Laplace (1774) in the 18th century; the frequentist probability paradigm began with Venn (1866) and was formalized by von Mises (1928) and Kolmogorov (1933). Intelligent people have therefore been arguing about the merits of these two approaches for about 150 years, like chess grandmasters working out the weaknesses of each other's opening strategies in tournament games, and both strategies are still employed in tournament play after all that intellectual conflict; we take from this a kind of empirical theorem that there must be elements

of merit in both approaches. The frequentist paradigm has the strength that it draws Your attention naturally to calibration issues, but this comes with many weaknesses (e.g., anything that cannot be cast uniquely in terms of a repeatable collective is off-limits for frequentist uncertainty quantification); the Bayesian approach appears to us to be the most flexible framework so far developed for successfully quantifying all kinds of uncertainty, whether arising from repeatable phenomena or not, but it has no built-in concept of calibration. Throughout much of the 20th century it was tacitly assumed that You had to choose one of the two paradigms and defend it fiercely against attacks from people who chose the other one, but this is a mis-framing of the problem: given that each paradigm has pluses and minuses, it appears to us instead that Your job is to try to develop a fusion of the two approaches that emphasizes the strengths, and de-emphasizes the weaknesses, of the fusion. For us this involves reasoning in a Bayesian way when formulating our inferences, predictions and decisions (this promotes internal consistency), and then paying attention calibratively to how often we get the right answer (this promotes external consistency). See, e.g., Box (1980), Rubin (1984), Draper (1996) and Little (2006) for additional thoughts on the value of a Bayesian-frequentist fusion.

- Calibration can also be given an entirely Bayesian justification via decision theory. Taking a broader perspective over Your career, not just within any single attempt to solve an inferential/predictive problem in collaboration with other investigators, Your desire (noted above) to avoid the loss of collaborative opportunities, arising from getting the wrong answer too often, and to take part positively in the progress of science can be quantified in a utility function that trades off a bonus for being well-calibrated against the length (or volume) of Your inferential and predictive intervals (or regions), and in this context calibration-monitoring emerges as a natural and inevitable Bayesian activity. We formalize these considerations elsewhere (Draper and von Brzeski (2010)). ♠

Remark. The calibration of predictive statements of the form $p(u < f(D^*) < v|D, \mathcal{B}) = 0.9$ is relatively straightforward to verify once the new data D^* have arrived; and if You judge that Your uncertainty about the past D and the future D^* is *conditionally exchangeable* (De Finetti (1938); e.g., Draper et al. (1993)) given θ, You can still (while waiting for the future to unfold) undertake verification exercises in which the available data set D is partitioned exchangeably into modeling and validation subsets D_M and D_V (respectively), and predictive distributions for the data in D_V based on the data in D_M can be compared with the actual validation data values (this provides the beginning of a Bayesian justification for *cross-validation*). The calibration of inferential statements such as $p(u < \theta < v|D, \mathcal{B}) = 0.9$ may also readily be checked, by creating a simulation world (similar to the structure of the problem under study) in which known values of θ drive the data-generating mechanism and inferential intervals (or regions) for θ may be compared calibratively with the known truth (see Section 6 for more on this idea). There are points of contact between our position on calibration in Bayesian modeling and both (a) the *prequential* approach to statistical analysis (Dawid (1984, 1985, 1991, 1997)) and (b) so-called *objective Bayesian* methodology (e.g., Berger (2006, 2009), Bayarri (2009)); in our view (Draper (2006)), since all inferential, predictive and decision-theoretic work in statistics is inherently subjective (i.e., based on assumptions and judgments), a better name for "objective Bayes" would be *calibrated Bayes*. ♠

As noted above, in practice You will typically be uncertain about how to specify the ingredients $p(\theta|\mathcal{B})$ and $p(D|\theta, \mathcal{B})$ necessary for inference and prediction, and this implies a willingness by You to consider an ensemble $\mathcal{M}$ of such specifications. In the important special case in which θ is a

vector of real parameters (of finite length), a Bayesian inferential and predictive model M_j is a choice $\{p(\eta_j|M_j, \mathcal{B}), p(D|\eta_j, M_j, \mathcal{B})\}$, in which $\eta_j = (\theta, \lambda_j)$ collects together both the unknowns θ common to all models in $\mathcal{M}$ and the unknowns λ_j specific to M_j. When λ_j is also a vector of real parameters (of finite length), M_j is typically referred to as a *parametric* Bayesian model.

The process of Bayesian model specification typically involves a search among scientifically and statistically plausible models, generally with a first goal of identifying a reasonable ensemble $\mathcal{M} = (M_1, \ldots, M_m)$ of models that are worthy of consideration. (We strongly agree with Box (1980), who notes — as we do — that statisticians are often in a collaborative role with subject-matter experts and who emphasizes that the interplay between statistical and substantive considerations in identifying the models to be considered in the ensemble is vital.) After $\mathcal{M}$ has been specified, often the second goal is either (i) to perform *Bayesian model averaging* (Leamer (1978); e.g., Draper (1995), Hoeting et al. (1999)) as a way to capture Your uncertainty over the possibilities in $\mathcal{M}$ or (ii) to choose a single model M_{j^*} among those in $\mathcal{M}$, if its posterior probability $p(M_{j^*}|D, \mathcal{B})$ is sharply dominant over the other $p(M_j|D, \mathcal{B})$ values across the models in $\mathcal{M}$.

In our view, consistent with the calibration principle described above, this search to specify $\mathcal{M}$ should be performed in a well-calibrated manner. It will often not be sufficient for good calibration to conduct the search using all of the data in D and then to use all of D again to draw inferential or predictive conclusions conditional on the results of the search; this amounts to using D to specify Your prior distribution on the space of all models and then using D again to update that prior to draw inferential and predictive conclusions, and this double-use of the data will be seen on closer examination to be logically inconsistent. For us, the right price to be paid for the model search can be quantified via a version of Bayesian cross-validation (Draper and Krnjajić (2010)) involving partitioning D into three (rather than two) subsets, in a manner somewhat related to a method used in machine learning (e.g., Hastie et al. (2001)); for reasons of space we again do not pursue this issue here. The point is that, in our view, it *is* possible to solve the problem of Bayesian model specification in a well-calibrated manner, even when this involves a data-driven search, as long as the right price is paid for the search.

From an algorithmic point of view, the process of specifying $\mathcal{M}$ will typically have four steps — (a) start, (b) propose a move from M_j to $M_{j'}$, (c) decide whether to make the move in (b), and (d) stop — with considerable iteration of the (b–c) steps. In this paper we present some calibration results that are helpful in answering two basic questions arising in steps (c) and (d):

- Q_1: Is model M_j better than $M_{j'}$?
- Q_2: Is model M_{j^*} good enough?

For us these questions, although they appear both reasonable and fundamental, are not yet well formed: Is model M_j better than $M_{j'}$, *for what purpose*? Is model M_{j^*} good enough, *for what purpose*? It is easy to imagine a situation in which M_j is better than $M_{j'}$ at helping to attain real-world goal G_1, and yet the two models are essentially equivalent in achieving goal G_2; similarly, M_{j^*} might be good enough for G_3 but wholly inadequate to reach G_4. To us, making clear the purpose to which the modeling will be put transforms model specification into a decision problem, which should (as noted above) be solved by maximizing expected utility (MEU; see, e.g., Fouskakis and Draper (2008) for an example, involving variable selection in generalized linear models, in which the model-specification problem is solved decision-theoretically). Others who share this decision-analytic view of the modeling process include Bernardo and Smith (1994) and Key et al. (1999).

Thus it is difficult to provide general-purpose methodology for model search and specification of the model ensemble $\mathcal{M}$, above and beyond that inherent in the MEU framework itself. Despite the fact that model specification is really a decision problem, many methods for at least indirectly addressing Q_1 (without explicit identification of a problem-specific utility function) have been proposed, including (Section 1.2) *Bayes factors* (which define the MEU solution to the M_{j^*} model-selection problem with a utility function — in which You have to pretend that one of the models in $\mathcal{M}$ is the actual data-generating mechanism M_{DG} and You reward Yourself with $c > 0$ utiles if Your chosen M_{j^*} is M_{DG} and 0 otherwise — that may be rather far from quantifying Your actual goals in a specific application) and the *Deviance Information Criterion* (*DIC*: Spiegelhalter et al. (2002)), which — when it works well (see Section 3.2 below) — is essentially a Bayesian generalization of *AIC* (Akaike (1974)) to parametric models in which it is not straightforward to identify the effective dimensionality of η_j. We return to *DIC* and (briefly) to *AIC* in Sections 2, 3 and 6.

1.2 Bayes factors

It will be helpful in what follows to further examine the strengths and weaknesses of Bayes factors. Applying Bayes's Theorem in odds form to the comparison of two models M_j and $M_{j'}$ yields the familiar expression

$$\frac{p(M_j|D,\mathcal{B})}{p(M_{j'}|D,\mathcal{B})} = \left[\frac{p(M_j|\mathcal{B})}{p(M_{j'}|\mathcal{B})}\right] \cdot \left[\frac{p(D|M_j,\mathcal{B})}{p(D|M_{j'},\mathcal{B})}\right] , \tag{1}$$

in which $\frac{p(D|M_j,\mathcal{B})}{p(D|M_{j'},\mathcal{B})}$ is the Bayes factor in favor of M_j over $M_{j'}$. When (as in Section 1.1) M_j is specified by the parametric representation $\{p(\eta_j|M_j,\mathcal{B}), p(D|\eta_j,M_j,\mathcal{B})\}$ with $\eta_j = (\theta,\lambda_j)$, the numerator of the Bayes factor, generally referred to as the *marginal* or *integrated likelihood*, is

$$p(D|M_j,\mathcal{B}) = \int_{H_j} p(D|\eta_j,M_j,\mathcal{B})\, p(\eta_j|M_j,\mathcal{B})\, d\eta_j = E_{(\eta_j|M_j,\mathcal{B})}\, p(D|\eta_j,M_j,\mathcal{B}) , \tag{2}$$

in which H_j is the parameter space for M_j; in other words, (2) is the expectation of the sampling distribution $p(D|\eta_j,M_j,\mathcal{B})$ under M_j (evaluated at the observed data set D) with respect to the *prior distribution* specified by M_j. If strong information about η_j and $\eta_{j'}$ external to D is available, then the prior distributions $p(\eta_j|M_j,\mathcal{B})$ and $p(\eta_{j'}|M_{j'},\mathcal{B})$ will be relatively stable with respect to small variations in how they are specified, and (2) — together with its analogue $p(D|M_{j'},\mathcal{B})$ for $M_{j'}$ — will lead to a stable Bayes factor that may serve as a useful basis for model comparison (if You are satisfied that the utility structure underlying Bayes factors is an adequate approximation to the real-world situation in the problem at hand). However, if — as is frequently the case — the information about η_j and $\eta_{j'}$ external to D is relatively weak (*diffuse*), then the expectation in (2) can be highly unstable as a function of small variations in how the diffuseness is specified.

To take a simple example that illustrates the problem with non-negative integer-valued data $D = y = (y_1,\ldots,y_n)$, consider two models: M_1 specifies a (conditionally-IID) Geometric(θ_1) sampling distribution with a Beta(α_1,β_1) prior on θ_1; M_2 is based on a (conditionally-IID) Poisson(θ_2) sampling distribution with a Gamma(α_2,β_2) prior on θ_2. The Bayes factor in favor of M_1 over M_2 (cf. Bernardo and Smith (1994)) is

$$\frac{\Gamma(\alpha_1+\beta_1)\Gamma(n+\alpha_1)\Gamma(s+\beta_1)\Gamma(\alpha_2)(n+\beta_2)^{s+\alpha_2}\,(\prod_{i=1}^n y_i!)}{\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(n+s+\alpha_1+\beta_1)\Gamma(s+\alpha_2)\beta_2^{\alpha_2}}, \tag{3}$$

where $s = \sum_{i=1}^{n} y_i$. Common choices for diffuse priors would include taking $(\alpha_1, \beta_1) = (1, 1)$ and $(\alpha_2, \beta_2) = (\epsilon, \epsilon)$ for some $\epsilon > 0$. The Bayes factor then reduces to

$$\frac{\Gamma(n+1)\Gamma(n\bar{y}+1)\Gamma(\epsilon)(n+\epsilon)^{n\bar{y}+\epsilon}\left(\prod_{i=1}^{n} y_i!\right)}{\Gamma(n+n\bar{y}+2)\Gamma(n\bar{y}+\epsilon)\epsilon^{\epsilon}}. \tag{4}$$

Expression (4) goes to $+\infty$ as $\epsilon \downarrow 0$; in other words, the evidence in favor of the Geometric model over the Poisson can be made as large as You might wish, *no matter what the data set is*, as a function of a quantity near 0 that scientifically has no basis for unique specification. (There is nothing special about the diffuse priors used here; the same sharp sensitivity to a prior specification with little scientific grounding appears, e.g., with a Uniform$(0, c)$ prior for θ_2 as a function of the nearly arbitrary c, and You can create equally bizarre behavior in the opposite direction by letting α_1 and β_1 approach 0 in the Beta prior for θ_1.)

Of course, this phenomenon is well known, and many attempts have been made to circumvent it, including {partial, intrinsic, fractional} Bayes factors, well-calibrated priors, conventional priors, intrinsic priors, and expected posterior priors (see, e.g., Pericchi (2005)). Many of these approaches attempt, either implicitly or explicitly, to stabilize (2) in the presence of diffuse prior information by calculating the expectation with respect to a *posterior* distribution that may be less subject to prior sensitivity. For example, *intrinsic Bayes factors* (Berger and Pericchi (1996)) partition the data set D into two subsets D_A and D_B, update the diffuse priors in models M_j and $M_{j'}$ — using the *training sample* D_A — to posterior distributions on the parameters, compute *partial* Bayes factors using these posterior distributions and the data in D_B, and then average the resulting partial Bayes factors across the (many) ways in which the partition could be performed; this requires defining criteria for what constitutes a "good" training sample and exploring which of a variety of averaging methods is "best." The most blatant attempt along these lines is *posterior Bayes factors* (PBF: Aitkin (1991)), in which the idea is to favor model M_j over $M_{j'}$ if $\bar{L}_j^{PBF} > \bar{L}_{j'}^{PBF}$, where

$$\bar{L}_j^{PBF} = \int_{H_j} p(D|\eta_j, M_j, \mathcal{B})\, p(\eta_j|D, M_j, \mathcal{B})\, d\eta_j = E_{(\eta_j|D,M_j,\mathcal{B})}\, p(D|\eta_j, M_j, \mathcal{B})\,; \tag{5}$$

in other words, $\bar{L}_j^{PBF}$ is the expectation of the sampling distribution $p(D|\eta_j, M_j, \mathcal{B})$ under M_j (evaluated at the observed data set D) with respect to the *posterior* distribution $p(\eta_j|D, M_j, \mathcal{B})$. With an even moderately informative data set D this solves the problem of sensitivity to a diffuse prior, but it creates a serious new problem of its own: by vigorously using the data twice (to update the sampling distribution to a posterior, and then to average the sampling distribution over this posterior) it lacks a correct probabilistic basis and is therefore logically inconsistent.

Another approach to avoiding the problem of instability of Bayes factors in the presence of diffuse prior distributions on the parameters is provided by the *Bayesian information criterion* (*BIC*; Schwarz (1978)). Letting k_j be the dimension of the parameter vector $\eta_j = (\theta, \lambda_j)$ in model M_j, a Laplace approximation to the integrated likelihood $p(D|M_j, \mathcal{B})$ on the log scale yields

$$\begin{aligned} \log p(D|M_j, \mathcal{B}) &= \log p(D|\hat{\eta}_j, M_j, \mathcal{B}) + \log p(\hat{\eta}_j|M_j, \mathcal{B}) \\ &\quad + \frac{k_j}{2}\log 2\pi - \frac{1}{2}\log|\hat{I}_j| + O\left(\frac{1}{n}\right), \end{aligned} \tag{6}$$

in which $\hat{\eta}_j$ is the maximum likelihood estimate of η_j under model M_j, $\hat{I}_j$ is the observed information matrix for that model and n is the sample size in the data set D. Using a less precise Taylor

expansion than that underlying (6), Schwarz proposed the cruder approximation

$$\log p(D|M_j, \mathcal{B}) = \log p(D|\hat{\eta}_j, M_j, \mathcal{B}) - \frac{k_j}{2} \log n + O(1). \tag{7}$$

By equating (6) and (7) You can solve to see what implied prior distribution BIC is using, from the point of view of the more accurate Laplace approximation; with all of the components of η_j living continuously on the real line, this yields the *unit-information prior*

$$(\eta_j|M_j, \mathcal{B}) \sim N_{k_j}(\hat{\eta}_j, n\hat{I}_j^{-1}), \tag{8}$$

a rather gently data-dependent diffuse prior distribution that is equivalent (with large n) to one additional observation with the same likelihood summaries as the data set D under M_j (if some components of η_j live on restricted ranges within $\Re$, a similar calculation may be made after transforming those components to take values on all of $\Re$). Thus BIC can be seen as the basis of an approximate Bayes factor with a diffuse prior that is not sensitive to how the diffuseness is specified; in effect BIC protects You from Your worst possible excesses in trying to be "noninformative" about the parameters.

1.3 Predictive model comparison

All of the work described in Section 1.2 seems to be unresponsive to the following point: there is an approach to Bayesian model comparison that both (a) has an arguably sounder basis in utility than Bayes factors and (b) entirely avoids instability with respect to diffuse prior specification. Of all the quantities arising in Bayesian modeling, the one that simultaneously (i) has the most to do with model comparison and (ii) is the most stable in the presence of diffuse prior distributions is the predictive distribution, for future data D^* given present data D, under model M_j:

$$\begin{aligned} p(D^*|D, M_j, \mathcal{B}) &= \int_{H_j} p(D^*|\eta_j, M_j, \mathcal{B})\, p(\eta_j|D, M_j, \mathcal{B})\, d\theta \\ &= E_{(\eta_j|D,M_j,\mathcal{B})} p(D^*|\eta_j, M_j, \mathcal{B})\,. \end{aligned} \tag{9}$$

Note the similarity with $\bar{L}_j^{PBF}$, which however incorrectly computes the expectation of $p(D|\eta_j, M_j, \mathcal{B})$, rather than $p(D^*|\eta_j, M_j, \mathcal{B})$, with respect to $p(\eta_j|D, M_j, \mathcal{B})$.

In our view, the closest You may be able to come to a useful generic utility structure driving model comparison is (a) to recognize the basic scientific fact that *good (bad) models make good (bad) predictions* and (b) to therefore reward a given model M_j in a manner driven by the quality of its predictions. To compare a predictive distribution with the actual data value it is trying to predict, You need a *scoring rule*; it has been shown (e.g., O'Hagan and Forster (2004)) that all of the optimal scoring rules, for comparing a single (real) data value y^* with its predictive distribution $p(\cdot|D, M_j, \mathcal{B})$ under a model M_j, are linear functions of the logarithm of the height of $p(\cdot|D, M_j, \mathcal{B})$ at y^*. This motivates the *log scoring* criterion (e.g, Gelfand and Dey (1994), Gelfand and Ghosh (1998)); for instance, in a one-sample setting in which the data set D consists of a vector $y = (y_1, \ldots, y_n)$ of real values of length n, a jackknife-style cross-validated version of this idea would be based on

$$LS_{CV}(M_j|y, \mathcal{B}) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|y_{-i}, M_j, \mathcal{B}), \tag{10}$$

where y_{-i} is y with observation i set aside; this is also referred to as the *conditional predictive ordinate* (CPO) approach (e.g., Geisser (1980), Pettit (1990)). We examine the decision-theoretic motivation of this and a related log-score criterion in Section 3.1.

It may appear that LS_{CV} is computationally expensive in settings in which the predictive distribution $p(y_i|y_{-i}, M_j, \mathcal{B})$ is not available in closed form and therefore has to be estimated via MCMC, because a naive implementation of (10) would appear to require n separate MCMC runs (one for each omitted observation). However, Gelfand (1996) has shown — in the common situation in which (under model M_j) y_i and y_{-i} are conditionally independent given η_j — that $p(y_i|y_{-i}, M_j, \mathcal{B})$ has the alternative representation

$$p(y_i|y_{-i}, M_j, \mathcal{B}) = \left\{ \int \left[p(y_i|M_j, \eta_j, \mathcal{B})\right]^{-1} p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j \right\}^{-1}, \tag{11}$$

and this permits estimation of LS_{CV} with a single MCMC run via

$$\widehat{LS}_{CV}(M_j|y, \mathcal{B}) = -\frac{1}{n}\sum_{i=1}^{n} \log \overline{p^{-1}(y_i|\eta_j, M_j, \mathcal{B})}\,; \tag{12}$$

here $\overline{p^{-1}(y_i|\eta_j, M_j, \mathcal{B})}$ is the posterior mean of the reciprocal of the sampling distribution for y_i under model M_j.

1.4 Outline

The plan of the paper is as follows: Sections 2 and 3 are devoted to aspects of some answers to question Q_1, and Section 4 addresses an issue arising from Q_2. In Section 2 we explore similarities and differences between DIC and LS_{CV} in Gaussian and Poisson models. Section 3 examines a *full-sample* version LS_{FS} of the log-score idea and presents results on the small-sample abilities of DIC, LS_{CV} and LS_{FS} to discriminate between fixed- and random-effects Poisson data-generating mechanisms. In Section 4 we show that *posterior predictive tail areas* (Gelman et al. (1996)), a popular method for answering a question related to Q_2 — namely, could the data have arisen from model M_{j^*}? — can be poorly calibrated, and we document the success of an approach to calibrating it. Section 5 addresses some asymptotic considerations, and in Section 6 we conclude the paper with a brief discussion.

2 DIC and LS_{CV}

Consider M_0, one of the simplest possible parametric models for continuous outcomes:

$$M_0\text{: } \left\{ \begin{array}{ccc} (\mu|\mathcal{B}) & \sim & N\left(\mu_0, \sigma_\mu^2\right) \\ (Y_i|\mu, \mathcal{B}) & \overset{\text{IID}}{\sim} & N(\mu, \sigma^2) \end{array} \right\},$$

with $(\sigma^2, \mu_0, \sigma_\mu^2)$ known. To see how DIC and LS_{CV} are related in this simple setting, take a highly diffuse (large σ_μ^2) prior on μ so that the posterior for μ is approximately

$$(\mu|y, \mathcal{B}) = (\mu|\bar{y}, \mathcal{B}) \,\dot{\sim}\, N\left(\bar{y}, \frac{\sigma^2}{n}\right), \tag{13}$$

where $\bar{y}$ is the sample mean of $y = (y_1, \ldots, y_n)$. The predictive distribution for the next observation is then approximately

$$(y_{n+1}|y, \mathcal{B}) = (y_{n+1}|\bar{y}, \mathcal{B}) \dot{\sim} N\left[\bar{y}, \sigma^2\left(1 + \frac{1}{n}\right)\right], \tag{14}$$

and LS_{CV}, ignoring linear scaling constants, is

$$LS_{CV}(M_0|y, \mathcal{B}) = \sum_{i=1}^{n} \ln p(y_i|y_{-i}, \mathcal{B}). \tag{15}$$

But by the same reasoning

$$p(y_i|y_{-i}, \mathcal{B}) \doteq N\left(\bar{y}_{-i}, \sigma_n^2\right), \tag{16}$$

where $\bar{y}_{-i}$ is the sample mean with observation i omitted and $\sigma_n^2 = \sigma^2\left(1 + \frac{1}{n-1}\right)$, so that

$$\begin{aligned} \ln p(y_i|y_{-i}, \mathcal{B}) &\doteq c - \frac{1}{2\sigma_n^2}(y_i - \bar{y}_{-i})^2 \quad \text{and} \\ LS_{CV}(M_0|y, \mathcal{B}) &\doteq c_1 - c_2 \sum_{i=1}^{n}(y_i - \bar{y}_{-i})^2 \end{aligned} \tag{17}$$

for some constants c_1 and c_2 with $c_2 > 0$. Now it is an interesting fact (related to the behavior of the jackknife), which can be proved by induction, that

$$\sum_{i=1}^{n}(y_i - \bar{y}_{-i})^2 = c\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{18}$$

for some $c > 0$, so finally for $c_2 > 0$ the result is that

$$LS_{CV}(M_0|y, \mathcal{B}) \doteq c_1 - c_2 \sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{19}$$

i.e., in M_0 with a diffuse prior the log score is almost perfectly negatively correlated with the sample variance.

In this model the *deviance* (minus twice the log likelihood) is

$$\begin{aligned} D(\mu) = D(\mu|\mathcal{B}) &= -2\ln l(\mu|y, \mathcal{B}) = c_0 - 2\ln p(y|\mu, \mathcal{B}) \\ &= c_0 + c_3 \sum_{i=1}^{n}(y_i - \mu)^2 \end{aligned} \tag{20}$$

for some $c_3 > 0$. Given a parametric model $p(y|\theta)$, Spiegelhalter et al. (2002) define the *deviance information criterion* (DIC) (by analogy with other information criteria) to be an estimate $D(\bar{\theta}|\mathcal{B})$ of the model lack of fit (as measured by the deviance) plus a penalty for complexity equal to twice the effective number of parameters p_D of the model:

$$DIC(M|y, \mathcal{B}) = D(\bar{\theta}) + 2\,\hat{p}_D, \tag{21}$$

where $\bar{\theta}$ is the posterior mean of θ; they suggest that models with low DIC values are to be preferred over those with higher values. When p_D is difficult to read directly from the model (e.g.,

in complex hierarchical settings, especially those with random effects), they motivate the following estimate, which is easy to compute from standard MCMC output:

$$\hat{p}_D = \overline{D(\theta)} - D(\bar{\theta}), \tag{22}$$

where $\overline{D(\theta)}$ is the posterior mean of the deviance and $D(\bar{\theta})$ is the deviance evaluated at the posterior mean of θ. In model M_0, p_D is of course 1, and with a diffuse prior $\bar{\theta} \doteq \bar{y}$, so

$$DIC(M_0|y, \mathcal{B}) \doteq c_0 + c_3 \sum_{j=1}^{n} (y_j - \bar{y})^2 + 2 \tag{23}$$

and the conclusion is that

$$-DIC(M_0|y, \mathcal{B}) \doteq c_1 + c_2 LS_{CV}(M_0|y, \mathcal{B}) \tag{24}$$

for $c_2 > 0$. In other words, in this simple setting, choosing a model by maximizing LS_{CV} and by minimizing DIC are approximately equivalent behaviors. This connection was hinted at in the discussion of Spiegelhalter et al. (2002) but was never made explicit. It is evident that this argument readily generalizes to any situation in which the predictive distribution is approximately Gaussian (e.g., Poisson(λ) likelihoods with large λ, Beta(α, β) likelihoods with large $(\alpha + \beta)$, and so on).

As a second example of the relationship between LS_{CV} and DIC, consider a single sample of counts of the number of occurrences of some (typically rather rare) event in a given time interval. With data of this form, modelers often choose between fixed- and random-effects Poisson model structures: for $i = 1, \ldots, n$, and, e.g., with diffuse priors, one implementation of this comparison involves choosing between

$$M_1\text{: } \left\{ \begin{array}{ccc} (\lambda|\mathcal{B}) & \sim & p(\lambda|\mathcal{B}) \\ (y_i|\lambda, \mathcal{B}) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) \end{array} \right\} \quad \text{and} \tag{25}$$

$$M_2\text{: } \left\{ \begin{array}{ccc} (\beta_0, \sigma^2|\mathcal{B}) & \sim & p(\beta_0, \sigma^2|\mathcal{B}) \\ (y_i|\lambda_i, \mathcal{B}) & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = & \beta_0 + e_i \\ (e_i|\sigma^2, \mathcal{B}) & \overset{\text{IID}}{\sim} & N(0, \sigma^2) \end{array} \right\}. \tag{26}$$

M_1 is of course a special case of M_2 with $\left(\sigma^2 = 0, \lambda = e^{\beta_0}\right)$; the likelihood in M_2 is a Lognormal mixture of Poissons (this is often similar to fitting a Negative Binomial distribution, which is a Gamma mixture of Poissons).

It is not entirely straightforward to express all of $\{LS_{CV}$ (in M_1 and M_2), DIC (in $M_2)\}$ algebraically in this setting, so we conducted a partial-factorial simulation study with factors $\{n = 18, 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0, 2.0\}$, and $\{\sigma^2 = 0.0, 0.5, 1.0, 1.5, 2.0\}$ in which $\{$(data-generating mechanism, assumed model)$\} = \{(M_1, M_1), (M_1, M_2), (M_2, M_1), (M_2, M_2)\}$; in each cell of this grid we used 100 simulation replications. Figures 1 and 2 summarize some of the results of this simulation (see Krnjajić (2005) for additional details). The first of these two Figures demonstrates that when both the data-generating model and the assumed model were M_1 (the fixed-effects Poisson), LS_{CV} and DIC are almost perfectly negatively correlated; the second Figure shows by contrast that when the data-generating and assumed models were M_2 (the random-effects Poisson), LS_{CV} and DIC

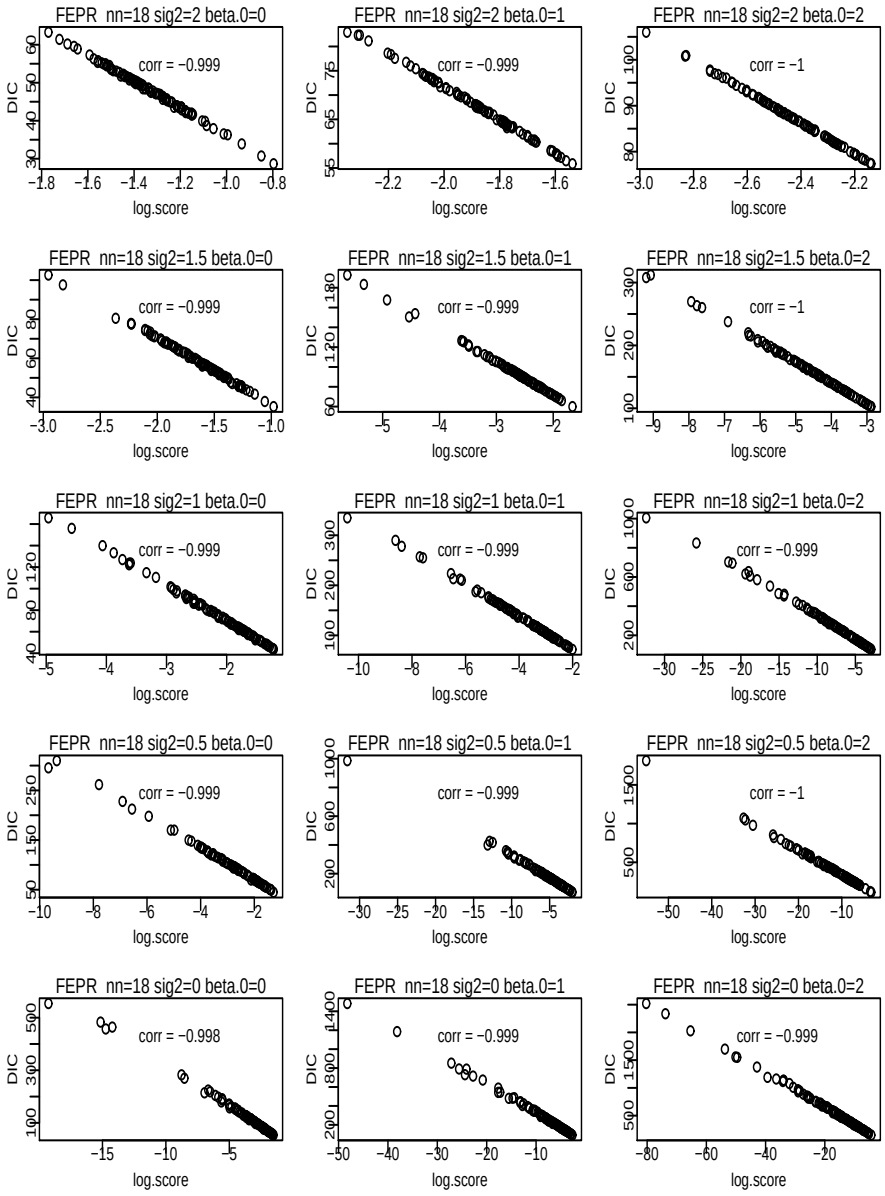

Figure 1: *DIC versus* LS_{CV} *with* $n = 18$*; the data-generating and assumed models were both* M_1 *(fixed-effects Poisson).*

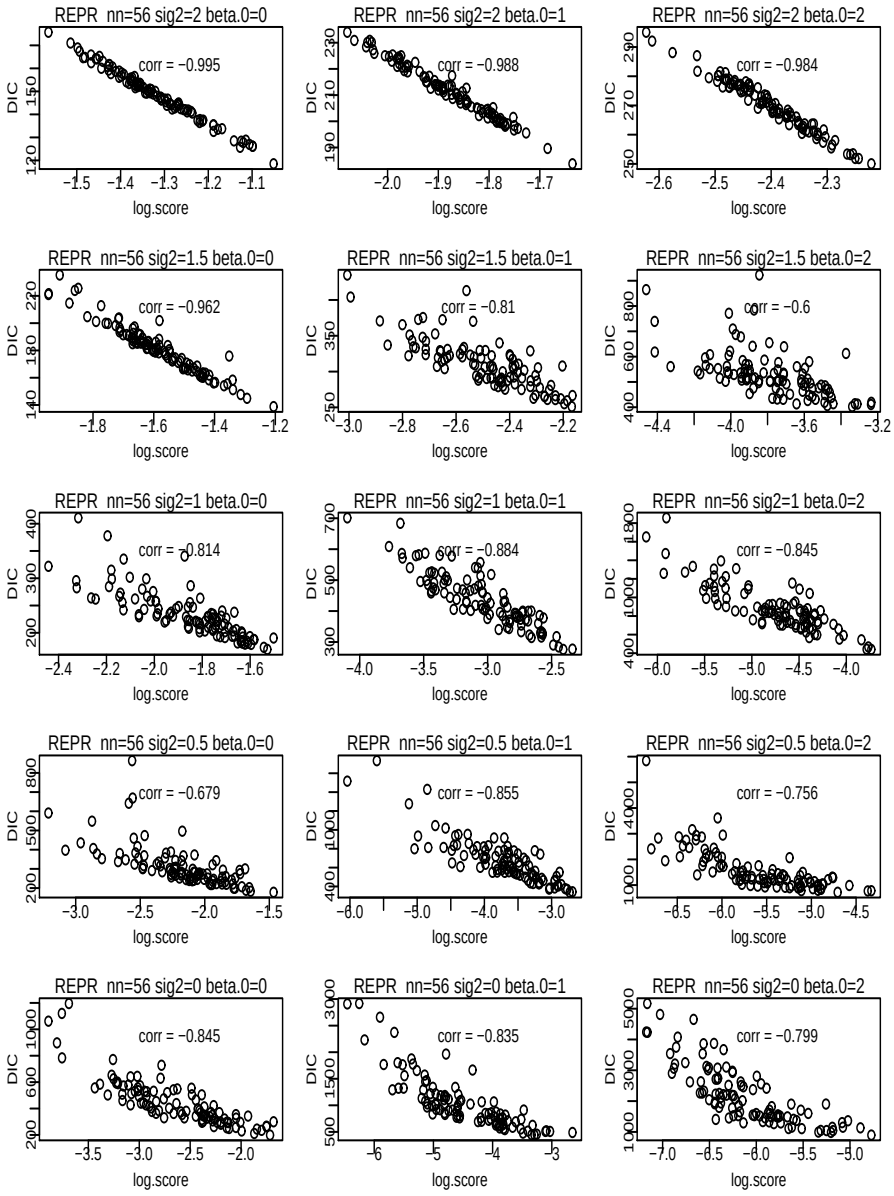

Figure 2: *DIC versus* LS_{CV} *with* $n = 56$*; the data-generating and assumed models were both* M_2 *(random-effects Poisson).*

Table 1: *Distribution of number of hospitalizations in the IHGA study over a two-year period.*

	Number of Hospitalizations										
Group	0	1	2	3	4	5	6	7	n	Mean	SD
Control	138	77	46	12	8	4	0	2	287	0.944	1.24
Treatment	147	83	37	13	3	1	1	0	285	0.768	1.01

are less strongly negatively correlated, although (not shown in the Figure) the correlation increases with n.

As a further example of the correspondence between LS_{CV} and DIC, consider the following case study, in which fixed- and random-effects Poisson modeling — of the type examined above — arises naturally. In a controlled experiment (Hendriksen et al. (1984)) to assess the value of a gerontological intervention, 572 elderly Danish people were randomized, 287 to a control (C) group (receiving standard health care) and 285 to a treatment (T) group (receiving standard care plus *in-home geriatric assessment* (IHGA), a kind of preventive medicine in which each person's medical and social needs were assessed and acted upon individually). A major outcome of interest in this experiment was the number of hospitalizations experienced by the subjects during the two-year life of the study. Let y_i^T and y_j^C be the numbers of hospitalizations for treatment person i and control person j, respectively, and suppose (as was true of the published results of the study) that treatment/control status is the only available covariate.

Table 1 presents the data values. Evidently IHGA lowered the mean hospitalization rate (for these elderly Danish people, at least) by $(0.944 - 0.768) \doteq 0.176$, which is approximately a $100\left(\frac{0.768-0.944}{0.944}\right)\% = 19\%$ reduction from the control level, a difference that is large in clinical terms; as usual, the next question is whether this difference is large in statistical terms, and a model is needed to answer this second question.

Four possible models for these data (not all of them good) are as follows:

- A two-independent-sample Gaussian model with diffuse priors (based on the usual advice that in repeated sampling the two-independent-samples z or t procedures are robust to non-Normality, at least as far as false-positive validity is concerned);
- A one-sample Poisson model with a diffuse prior, which in effect assumes that the treatment and control λs are equal;
- A two-independent-sample Poisson model with diffuse priors, which is equivalent to a *fixed-effects Poisson regression* (FEPR) model; and
- a *random-effects Poisson regression* (REPR) model (which may be preferable to the FEPR model because the C and T variance-to-mean ratios (VTMRs) are 1.63 and 1.32, respectively, and the FEPR model assumes that these ratios are 1):

$$\begin{aligned} (y_i|\lambda_i, \mathcal{B}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_i + e_i \\ (e_i|\sigma_e^2, \mathcal{B}) &\overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right) \\ \left(\beta_0, \beta_1, \sigma_e^2|\mathcal{B}\right) &\sim \text{diffuse} , \end{aligned} \tag{27}$$

Table 2: *DIC and LS_{CV} results for four models applied to the IHGA example.*

Model	$\overline{D(\theta)}$	$D(\bar{\theta})$	$\hat{p}_D$	DIC	LS_{CV}
1 (Gaussian)	1749.6	1745.6	3.99	1753.5	−1.552
2 (Poisson, common λ)	1499.9	1498.8	1.02	1500.9	−1.316
3 (FEPR, different λs)	1495.4	1493.4	1.98	1497.4	−1.314
4 (REPR)	1275.7 1274.7 1274.4	1132.0 1131.3 1130.2	143.2 143.5 144.2	1418.3 1418.2 1418.6	−1.180

where $x_i = 1$ is a binary indicator for T/C status.

The DIC and LS_{CV} results on these four models are given in Table 2 (the three REPR rows were based on different monitoring runs, all of length 10,000, to give an idea of the size of the Monte Carlo noise level in the components of DIC). As $\sigma_e \to 0$ in the REPR model, the result is the FEPR model, with $p_D = 2$ parameters; as $\sigma_e \to \infty$, in effect all subjects in the study have their own λs and p_D would be 572; in between at $\sigma_e \doteq 0.675$ (the posterior mean), DIC estimates that there are about 144 effective parameters in the REPR model, but its deviance $D(\bar{\theta})$ is so much lower that it wins the DIC contest handily. The correlation between LS_{CV} and DIC across these four models turned out to be −0.98, providing another example of a situation where the two approaches lead to similar model-choice behaviors (this is due to the rather large samples in both the T and C groups in the experiment). As noted in Krnjajić et al. (2008), the REPR model fits the data well (with one caveat, addressed below in Section 3.2), and leads to the inferential conclusion that $p(\gamma < 0|D, \mathcal{B}) \doteq 0.97$, where γ is the mean difference $(T - C)$ in hospitalizations per two years in the population of all elderly people similar to the participants in the experiment.

3 Small-sample model discrimination

3.1 Full-sample log scores

In addition to LS_{CV}, our interest was drawn (on the basis of simulation results presented below) to another version of the log-score idea in which no cross-validation is employed: in the one-sample situation, for instance, You can compute a single predictive distribution $p(\cdot|y, M_j)$ for a future data value with each model M_j under consideration, based on the entire data set y (without omitting any observations), and define (cf. Laud and Ibrahim (1995)) the *full-sample log score*

$$LS_{FS}(M_j|y, \mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y, M_j, \mathcal{B}). \tag{28}$$

Remark. This appears to use the data twice, but (a) all LS_{FS} is actually doing is evaluating

the posterior predictive distribution for the *next* data value at the observed data, and (b) when n is even moderate in size, any effect this may induce is small. ♠

Remark. Revisiting the example in Section 1.2 that compared Geometric (M_1) and Poisson (M_1) models for one-sample count data, the LS_{FS} values for the two models are

$$LS_{FS}(M_1|y,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\Gamma(\alpha_1+\beta_1+n+s)\,\Gamma(\beta_1+s+y_i)\,(\alpha_1+n)}{\Gamma(\beta_1+s)\,\Gamma(y_i+\alpha_1+n+\beta_1+s+1)}\right] \tag{29}$$

and

$$\begin{aligned} LS_{FS}(M_2|y,\mathcal{B}) &= \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\Gamma(\alpha_2+s+y_i)}{\Gamma(\alpha_2+s)\Gamma(y_i+1)}\left(\frac{\beta_2+n}{\beta_2+n+1}\right)^{\alpha_2+s}\cdot\right. \\ &\qquad \left.\left(\frac{1}{\beta_2+n+1}\right)^{y_i}\right]\,, \end{aligned} \tag{30}$$

where $s = \sum_{i=1}^{n} y_i$. It is evident that both of these expressions are entirely stable as any or all of $\{\alpha_1,\beta_1,\alpha_2,\beta_2\} \downarrow 0$; thus — in this example, which is typical of results in both parametric and nonparametric Bayesian modeling — LS_{FS} has none of the difficulty with diffuse prior distributions exhibited by Bayes factors. ♠

Remark. The utility justification of LS_{FS} (cf. Bernardo and Smith (1994)) is as follows: with the unknown θ as a future data value y^*, the action space $\mathcal{A}$ as the models M_j in $\mathcal{M}$, and $U(a,\theta) = U(M_j,y^*) = \log p(y^*|y,M_j,\mathcal{B})$, the expectation in $E_{(\theta|D,\mathcal{B})}U(a,\theta)$ is over Your uncertainty about how y^* will turn out in the future, and (if y^* and the components of y are exchangeable given M_j, as will typically be the case) LS_{FS} is a direct Monte Carlo approximation to $E_{(\theta|D,\mathcal{B})}U(a,\theta)$.

As Mukhopadhyay et al. (2005) point out, another way to arrive at this same conclusion is as follows: positing the existence of a "true" data-generating mechanism M_{DG}, model M_j is better than $M_{j'}$ if the "distance" from M_j to M_{DG} is smaller than the "distance" from $M_{j'}$ to M_{DG}; evaluating the models on the basis of their predictive distributions for future data, and using Kullback-Leibler (relative entropy) divergence as the "distance" measure, You would prefer M_j if

$$\begin{aligned} &\int p(y^*|y,M_{DG},\mathcal{B}) \log\left[\frac{p(y^*|y,M_{DG},\mathcal{B})}{p(y^*|y,M_j,\mathcal{B})}\right] dy^* < \\ &\qquad \int p(y^*|y,M_{DG},\mathcal{B}) \log\left[\frac{p(y^*|y,M_{DG},\mathcal{B})}{p(y^*|y,M_{j'},\mathcal{B})}\right] dy^*\,, \end{aligned} \tag{31}$$

which is equivalent to choosing M_j if

$$\begin{aligned} &\int \log p(y^*|y,M_j,\mathcal{B})\, p(y^*|y,M_{DG},\mathcal{B})\, dy^* = E_{(y^*|y,M_{DG},\mathcal{B})} \log p(y^*|y,M_j,\mathcal{B}) > \\ &\int \log p(y^*|y,M_{j'},\mathcal{B})\, p(y^*|y,M_{DG},\mathcal{B})\, dy^* = E_{(y^*|y,M_{DG},\mathcal{B})} \log p(y^*|y,M_{j'},\mathcal{B})\,. \end{aligned} \tag{32}$$

But if You know M_{DG}, there is no new information in y, so the expectations in (32) are actually over the sampling distribution $p(y^*|M_{DG},\mathcal{B})$ for y^* under M_{DG}, and a direct Monte Carlo approximation to this distribution (based on the observed $y = (y_1,\ldots,y_n)$) yields the log-score criterion: prefer

M_j over $M_{j'}$ if

$$\hat{E}_{(y^*|M_{DG},\mathcal{B})} \log p(y^*|y, M_j, \mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y, M_j, \mathcal{B}) = LS_{FS}(M_j|y, \mathcal{B}) >$$
$$\hat{E}_{(y^*|M_{DG},\mathcal{B})} \log p(y^*|y, M_{j'}, \mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y, M_{j'}, \mathcal{B}) = LS_{FS}(M_{j'}|y, \mathcal{B}). \quad (33)$$

Mukhopadhyay et al. (2005) have shown that asymptotically the difference between {the actual expected utility $E_{(y^*|D,\mathcal{B})}U(M_j, y^*)$ called for by decision-theoretic model choice} and {its Monte Carlo approximation $LS_{CV}(M_j|D, \mathcal{B})$ based on a line of reasoning like the one above} is $O_p(\sqrt{n})$ (and the same argument would evidently apply to LS_{FS}), but this fact does not automatically imply poor model-discrimination behavior for either LS_{CV} or LS_{FS} with a fixed sample size, because model comparison with, e.g., LS_{FS} involves calculating $LS_{FS}(M_j|y, \mathcal{B}) - LS_{FS}(M_{j'}|y, \mathcal{B})$ and the asymptotic bias documented by Mukhopadhyay et al. (2005) would be expected to (largely or entirely) cancel out in this comparison (this is an application of the old idea that a biased scale can nevertheless be valuable in evaluating the *difference* in weight between two objects). ♠

Remark. When model M_j is fit via MCMC, the predictive ordinate $p(y^*|y, M_j, \mathcal{B})$ in LS_{FS} is easy to approximate: with m identically distributed (not necessarily independent) MCMC monitoring draws η_{jk}^* from $p(\eta_j|y, M_j, \mathcal{B})$,

$$\begin{aligned} p(y^*|y, M_j, \mathcal{B}) &= \int p(y^*|\eta_j, M_j, \mathcal{B})\, p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j \\ &= E_{(\eta_j|y,M_j,\mathcal{B})}\left[p(y^*|\eta_j, M_j, \mathcal{B})\right] \\ &\doteq \frac{1}{m}\sum_{k=1}^{m} p(y^*|\eta_{jk}^*, M_j, \mathcal{B})\,. \quad ♠ \end{aligned} \quad (34)$$

Remark. Along with a discussion of what LS_{FS} is, it is perhaps useful to point out several things that it is *not*.

- It may seem at first glance (e.g., O'Hagan and Forster (2004)) that the behavioral rule based on posterior Bayes factors is the same as the rule based on LS_{FS}, which favors model M_j over $M_{j'}$ if
$$n\, LS_{FS}(M_j|y, \mathcal{B}) > n\, LS_{FS}(M_{j'}|y, \mathcal{B}). \quad (35)$$
But not so: for example, in the common situation in which the data set D consists of observations y_i that are conditionally IID from $p(y_i|\eta_j, M_j, \mathcal{B})$ under M_j,
$$nLS_{FS}(M_j|y, \mathcal{B}) = \log \prod_{i=1}^{n} \left[\int p(y_i|\eta_j, M_j, \mathcal{B})\, p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j\right], \quad (36)$$
and this is not the same as
$$\log \int \left[\prod_{i=1}^{n} p(y_i|\eta_j, M_j, \mathcal{B})\right] p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j = \bar{L}_j^{PBF} \quad (37)$$
because the product and integral operators do not commute.

- Also, comparing models based on the posterior expectation of the log likelihood (or, equivalently, the log sampling distribution; this is related to one of the two additive terms in DIC, namely the one that penalizes lack of fit) has sometimes been suggested, and this is not the same as LS_{FS} either: by Jensen's inequality

$$\begin{aligned} nLS_{FS}(M_j|y,\mathcal{B}) &= \sum_{i=1}^{n} \log p(y_i|y,M_j,\mathcal{B}) \\ &= \sum_{i=1}^{n} \log \int p(y_i|\eta_j,M_j,\mathcal{B})\, p(\eta_j|y,M_j,\mathcal{B})\, d\eta_j \\ &= \sum_{i=1}^{n} \log E_{(\eta_j|y,M_j,\mathcal{B})}\, p(y_i|\eta_j,M_j,\mathcal{B}) \\ &> \sum_{i=1}^{n} E_{(\eta_j|y,M_j,\mathcal{B})} \log p(y_i|\eta_j,M_j,\mathcal{B}) \qquad (38) \\ &= E_{(\eta_j|y,M_j,\mathcal{B})} \sum_{i=1}^{n} \log p(y_i|\eta_j,M_j,\mathcal{B}) \\ &= E_{(\eta_j|y,M_j,\mathcal{B})} \log \prod_{i=1}^{n} p(y_i|\eta_j,M_j,\mathcal{B}) \\ &= E_{(\eta_j|y,M_j,\mathcal{B})} \log p(y|\eta_j,M_j,\mathcal{B}). \quad \spadesuit \end{aligned}$$

Including LS_{FS} in the mix gives rise to the three behavioral rules we examine in the rest of this Section: {maximize LS_{CV}, maximize LS_{FS}, minimize DIC}. With (e.g.) two models to choose between, how accurately do these behavioral rules discriminate between M_1 and M_2?

As an extension of the previous simulation study, we generated data from the random-effects Poisson model M_2 (equation (26)) and computed LS_{CV}, LS_{FS}, and DIC for models M_1 (the fixed-effects Poisson model (25)) and M_2 in the full-factorial grid $\{n = 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0\}$, $\sigma^2 = 0.1, 0.25, 0.5, 1.0, 1.5, 2.0\}$, with 1000 simulation replications in each cell (the simulation was performed on a cluster of 100 Linux-based CPUs), and we monitored the percentages of correct choice for each model specification method (in this simulation M_2 is always correct).

Table 3 gives examples of the results of this simulation, using LS_{CV} for illustration. Even with a sample size of only 32, LS_{CV} makes the right model choice more than 90% of the time when $\sigma^2 > 0.5$ for $\beta_0 = 1$ and when $\sigma^2 > 1.0$ for $\beta_0 = 0$ (these are parameter ranges that lead to large enough amounts of extra-Poisson variability that random-effects models would be contemplated). The right part of the table shows that even rather small differences in LS_{CV} can separate correct and incorrect model choice, which encourages the question "How do You know when a difference on the log score scale is big?" (we return to this point in Section 4). The graphs in Figure 3 compare model discrimination curves for LS_{CV}, LS_{FS}, and DIC; increasing σ^2 makes it easier for all three methods to conclude that random effects, to describe the Poisson over-dispersion, are needed. Interestingly, in this simulation environment LS_{FS} was more accurate, with small samples of data, at identifying the correct model than LS_{CV} or DIC; for this reason, we focus on LS_{FS} in what follows.

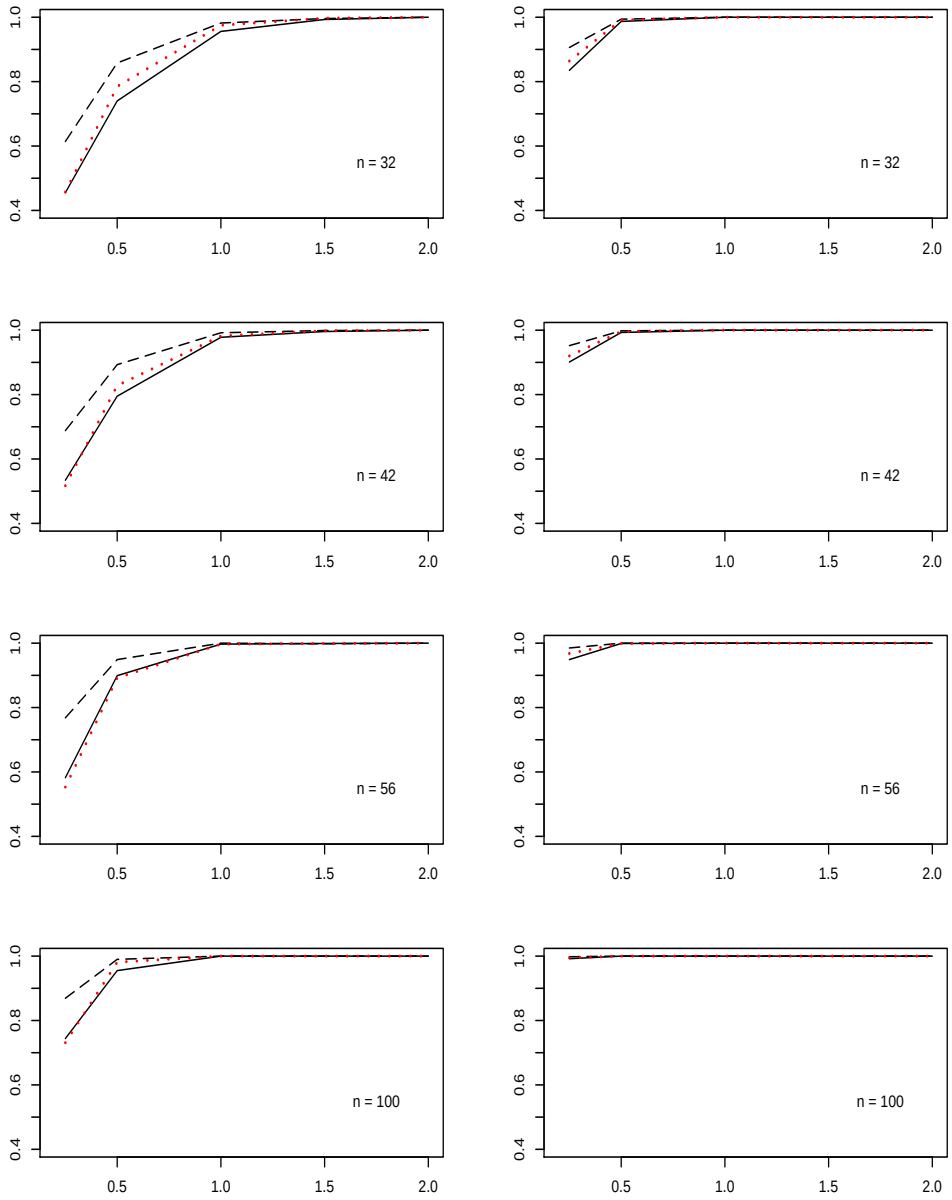

Figure 3: *Model discrimination curves for LS_{CV} (solid lines), LS_{FS} (long dotted lines), and DIC (short dotted lines) (column 1: $\beta_0 = 0$; column 2: $\beta_0 = 1$; rows are indexed by sample size n; in all plots the horizontal scale is σ^2 and the vertical scale is the proportion of correct model choices).*

Table 3: *Percentages of correct model choice and mean absolute difference in LS_{CV} between M_1 and M_2 when the right model is M_2, for $n = 32$.*

$n = 32$					
% Correct Decision			Mean Absolute Difference in LS_{CV}		
	β_0			β_0	
σ^2	0	1	σ^2	0	1
0.10	31	47	0.10	0.001	0.002
0.25	49	85	0.25	0.002	0.013
0.50	76	95	0.50	0.017	0.221
1.00	97	100	1.00	0.237	4.07
1.50	98	100	1.50	1.44	17.4
2.00	100	100	2.00	12.8	63.9

3.2 Comparison of LS_{FS} and DIC

We noted earlier that DIC can be thought of as a useful generalization of AIC to settings in which it is difficult to estimate the model complexity simply by, e.g., reading a count of the number of parameters directly from the model. However, it is worth emphasizing the point made by Spiegelhalter et al. (2002) (and the discussants of that paper) that DIC can be quite sensitive to parameterization. For example, $y = (0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6)$ is a data set with $n = 17$ observations generated with parameters $(\theta, r) = (0.82, 10.8)$ from the Negative Binomial distribution, in the parameterization under which the marginal sampling distribution is

$$p(y_i|\theta, r) = \frac{\Gamma(y_i + r)}{\Gamma(y_i + 1)\,\Gamma(r)}\theta^r(1-\theta)^{y_i} \ ; \tag{39}$$

y has mean 2.35 and variance-to-mean ratio 1.22. Using a Uniform$(0, 1)$ prior for θ and a popular (if possibly ill-advised) prior for r ($\Gamma(\epsilon, \epsilon)$ with $\epsilon = 0.001$), the effective number of parameters p_D for the negative binomial model (which fits the data quite well) is estimated to be -66.2, when of course the right answer is $+2.0$. The basic problem, as usual with DIC, is that the MCMC estimate of p_D can be quite poor if the marginal posteriors for one or more parameters (using the parameterization that defines the deviance) are far from Normal. Reparameterization can help — here, for example, working with Uniform$(-c, c)$ priors on logit(θ) and $\log(r)$, with c chosen large enough in each case not to truncate the likelihood function, yields $\hat{p}_D = 1.1$ (which is, however, still too low by 45%) — but may nevertheless lead in other problems to regrettable estimates of p_D. The log-score approach to model choice does not suffer from any such instability as a function of parameterization.

Remark. While the subject of complexity penalty is on the table, so to speak, it may be natural to wonder why such a penalty does not appear explicitly in LS_{FS}. In fact, a penalty for excess model complexity is implicitly built into LS_{FS}: models with unnecessary parameters will yield predictive distributions with larger variances (and therefore smaller predictive density values at the observed data) than models in which such unnecessary parameters are removed. ♠

Remark. Note also that the log-scoring approach works equally well with both parametric and nonparametric Bayesian models. As an example of the use of log scores in the comparison of such

Table 4: *LS_{FS} values for three models applied to the control and treatment samples in the IHGA case study (from Krnjajić et al. (2008)).*

	Model		
Sample	(26)	(40)	(41)
Treatment	−1.199	−1.198	−1.205
Control	−1.343	−1.342	−1.336

models, Table 4 (based on results given in Krnjajić et al. (2008): KKD) presents LS_{FS} values for three models applied separately to the treatment and control samples of IHGA data in Table 1: the random-effects Poisson model (26), a Dirichlet-process (DP) mixture model

$$\begin{aligned} (y_i|\theta_i) &\overset{\text{indep}}{\sim} \text{Poisson}(\theta_i) \\ (\theta_i|G) &\overset{\text{IID}}{\sim} G \\ (G|\alpha,\mu,\sigma^2) &\sim DP[\alpha\, G_0(\,\cdot\,|\mu,\sigma^2)] \\ (\alpha,\mu,\sigma^2) &\sim p(\alpha)\, p(\mu)\, p(\sigma^2) \end{aligned} \tag{40}$$

in which the DP was applied to the latent variables θ_i and was centered on model (26) (here $G_0(\,\cdot\,|\mu,\sigma^2) = N(\,\cdot\,;\mu,\sigma^2)$; KKD employed a Gamma prior for α, a Normal prior for μ, and an inverse-Gamma prior for σ^2), and a DP model

$$\begin{aligned} (y_i|F) &\overset{\text{IID}}{\sim} F \\ (F|\alpha,\theta) &\sim DP[\alpha\, F_0(\,\cdot\,|\theta)] \\ (\alpha,\theta) &\sim p(\alpha)\, p(\theta) \end{aligned} \tag{41}$$

applied directly to the observed counts (centered on the Poisson distribution $F_0(\,\cdot\,|\theta) = \text{Poisson}[\,\cdot\,;\exp(\theta)]$; KKD used a Gamma prior for α and a Gaussian prior for θ). There is some evidence that the Gaussian distributional assumption for the latent variables e_i in (26), which is conventional rather than arising directly from the science of the problem, is questionable. This sort of comparison cannot be made with DIC and would be difficult or impossible to achieve in a sound manner with Bayes factors (see Carota (2006) for a clear analysis of some of the problems that can arise when using Bayes factors to compare Bayesian parametric and nonparametric models). ♠

4 Calibrating posterior predictive tail areas

Section 3 demonstrates that full-sample log scores can stably and reliably help in choosing between two or more models (without loss of generality, consider just M_1 and M_2); but suppose that M_1 has a (substantially) higher LS_{FS} value than M_2. This doesn't say that M_1 is adequate; it just says that M_1 is better than M_2, which still leaves open model specification question Q_2 near the beginning of the paper: Is M_1 good enough?

As discussed in Section 1.1, in our view a full judgment of adequacy requires real-world input ("To what purpose will the model be put?"), so it does not seem possible to propose generic

methodology to answer Q_2 (apart from MEU, with a utility function that is appropriately tailored to the problem at hand), but the somewhat related question

- $Q_{2'}$: Could the data have arisen from model M_j?

can be answered in a general way by simulating from M_j many times, developing a distribution of (e.g.) LS_{FS} values, and seeing how unusual the actual data set's log score is in this distribution.

This is related to the *posterior predictive model-checking* method of Gelman et al. (1996). However, this sort of thing needs to be done carefully (Draper (1996)), or the result will be poor calibration; indeed, Bayarri and Berger (2000) and Robins et al. (2000) have demonstrated that the Gelman et al. procedure may be (sharply) conservative. Using a modification of an idea suggested by Robins et al., we have developed a method for accurately calibrating the log score scale.

The inputs to our procedure are: (1) a data set (e.g., with regression structure), and (2) a model (which can be parametric or non-parametric). To take a simple example to fix ideas, consider a one-sample data set of counts and suppose the goal is to judge whether this data set could have arisen from the model (call it $(*)$)

$$\begin{aligned} (\lambda|\mathcal{B}) &\sim \text{diffuse} \\ (y_i|\lambda,\mathcal{B}) &\overset{\text{IID}}{\sim} \text{Poisson}(\lambda) \end{aligned} \tag{42}$$

Step 1: Calculate LS_{FS} for this data set; call this the *actual log score* (ALS). Obtain the posterior for λ given y based on this data set; call this the *actual posterior*.

Step 2:

```
for ( i in 1:m1 ) {

  Make a lambda draw from the actual posterior; call it
    lambda[ i ].

  Generate a data set of size n from the second line of model (*)
    above, using lambda = lambda[ i ].

  Compute the log score for this generated data set; call it
    LS[ i ].

}
```

The output of this loop is a vector of log scores; call this *V.LS*. Locate the ALS in the distribution of LS_{FS} values by computing the percentage of LS_{FS} values in V.LS that are no greater than ALS; call this percentage the *unadjusted actual tail area* (suppose, e.g., that this comes out 0.22).

So far this is just Gelman et al. with LS_{FS} as the *discrepancy function*. We know from our own simulations (summarized below) and the literature (Bayarri and Berger (2000), Robins et al. (2000)) that this tail area (a P-value for a composite null hypothesis, e.g., Poisson(λ) with λ unspecified) is conservative, i.e., with the 0.22 example above an adjusted version of it that is well calibrated would be smaller (and might be much smaller, e.g., 0.02). We have modified and implemented one of the ways suggested by Robins et al. for improving calibration, and we have shown that it does indeed work even in rather small-sample situations, although implementing the basic idea can be computationally intensive.

Step 3:

```
for ( j in 1:m2 ){

  Make a lambda draw from the actual posterior; call it lambda*.

  Generate a data set of size n from the second line of model (*)
    above, using lambda = lambda*; call this the simulated
    data set.

  Repeat Steps 1 and 2 above on this simulated data set.

}
```

The result will be a vector of unadjusted tail areas; call this *V.P.* Compute the percentage of tail areas in V.P that are no greater than the unadjusted actual tail area; this is the *adjusted actual tail area.*

The claim is that the 3–step procedure above is well-calibrated, i.e., if the sampling part of model $(*)$ really did generate the observed data, the distribution of adjusted actual tail areas obtained in this way would be uniform, apart from simulation noise. Step 3 in this procedure solves the calibration problem by applying the old idea that if $X \sim F_X$ then $F_X(X) \sim U(0,1)$.

Our claim of calibration can be verified by building a further loop around steps 1–3 as follows:

```
Choose a lambda value of interest; call it lambda.sim .

for ( k in 1:m3 ) {

  Generate a data set of size n from the second line of model (*)
    above, using lambda = lambda.sim; call this the validation
    data set.

  Repeat Steps 1-3 on the validation data set.

}
```

The result will be a vector of *adjusted tail areas*; call this *V.Ta.* We have verified (via simulation, which was again performed on a cluster of 100 Linux-based CPUs) in several simple (and some less simple) situations that the values in V.Ta are close to $U(0,1)$ in distribution.

Figures 4–7 summarize some of our results (see Krnjajić (2005) for additional findings) and illustrate uncalibrated and calibrated tail areas from one-sample Poisson and Gaussian models (we used $m_1 = m_2 = m_3 = 1{,}000$). Figures 4 and 6 present histograms of the unadjusted actual tail area distributions, which are in many cases far from the target (uniform) distribution; figures 5 and 7 give uniform quantile plots of the adjusted tail area distributions. Consider, for example, the case $(n = 100, \lambda = 0.14)$ in the fourth row and first column of Figure 4: if the Gelman et al. tail area came out 0.35 in this situation, it would be natural to conclude that the data could very well have come from the Poisson model, but this part of Figure 4 demonstrates clearly that in fact an uncalibrated tail area of 0.35 with $(n = 100, \lambda = 0.14)$ is highly unusual under the Poisson model. Our procedure solves the calibration problem by asking "How often would You get 0.35 or less for an uncalibrated tail area in this situation?", and it is evident from Figure 4 that the answer

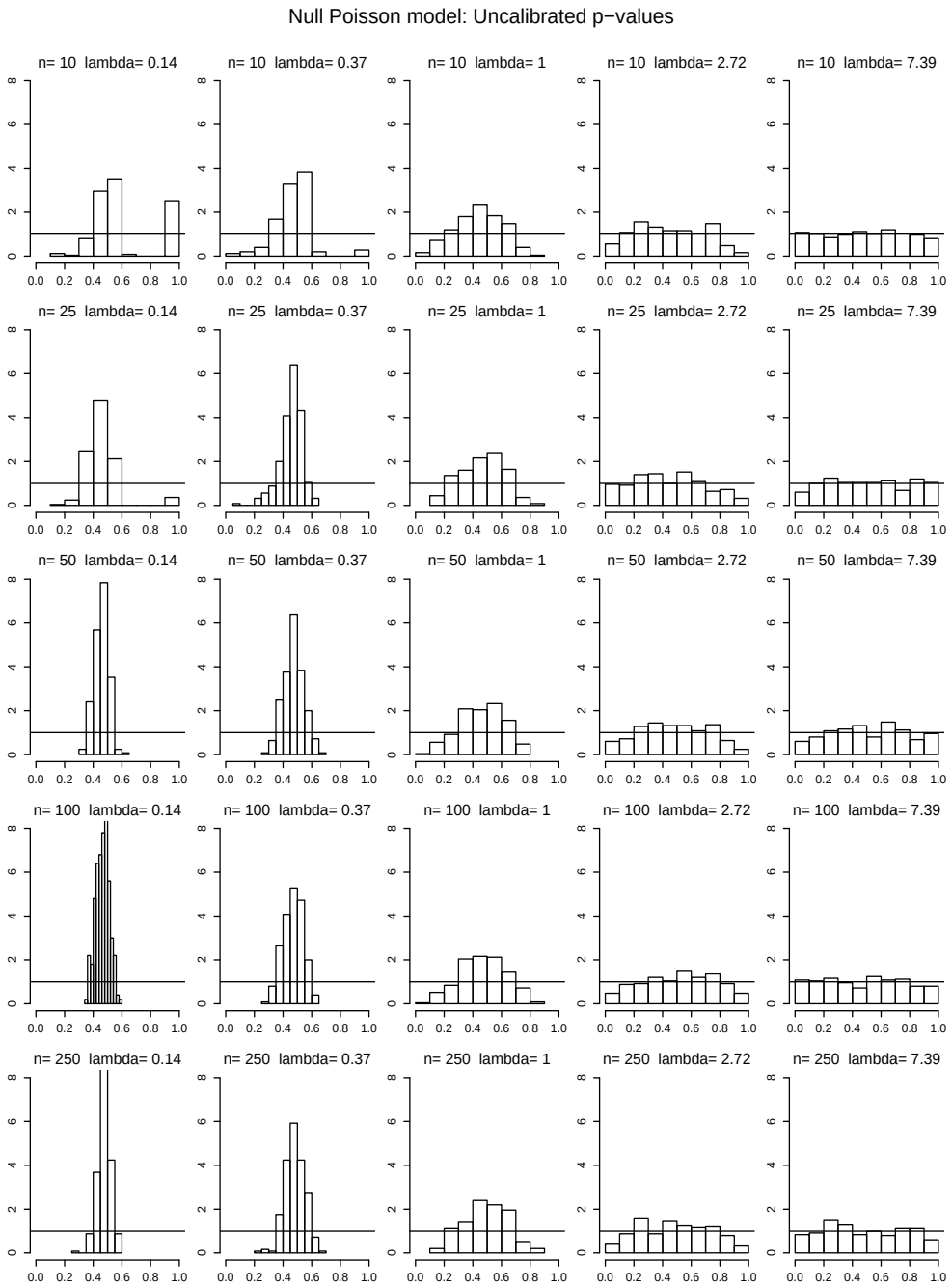

Figure 4: *Poisson model: uncalibrated tail-area values.*

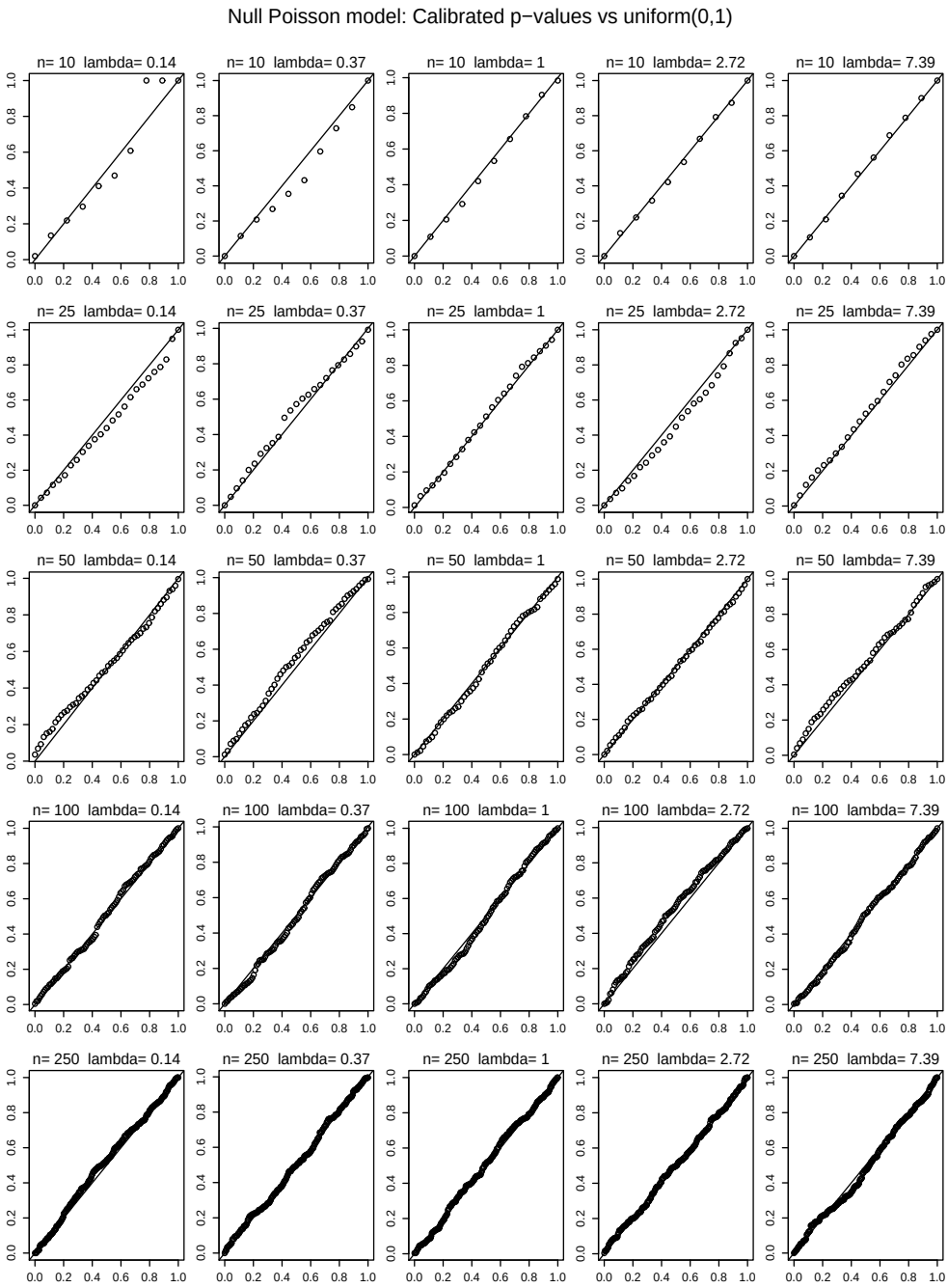

Figure 5: *Poisson model: calibrated tail-area values.*

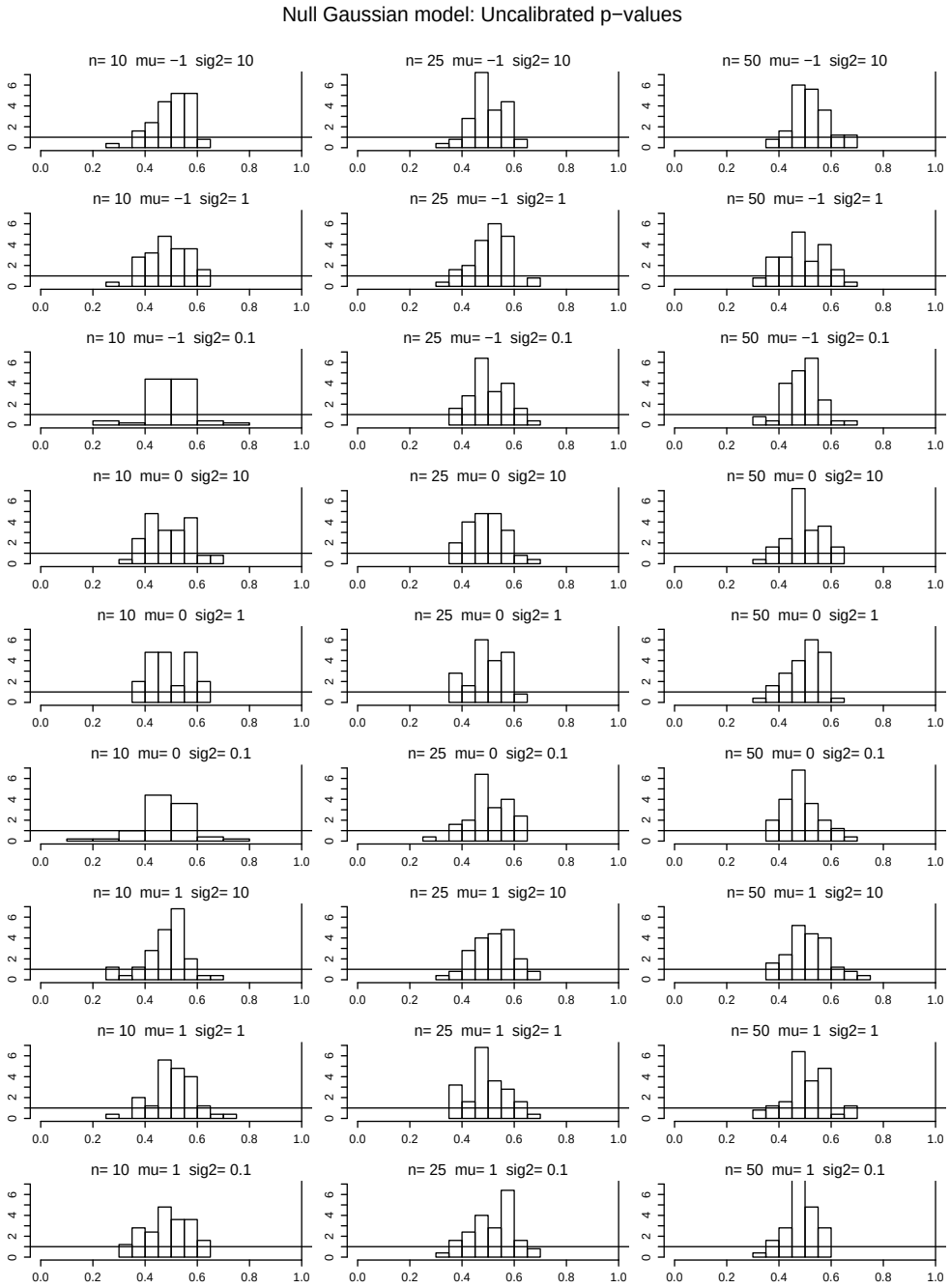

Figure 6: *Gaussian model: uncalibrated tail-area values.*

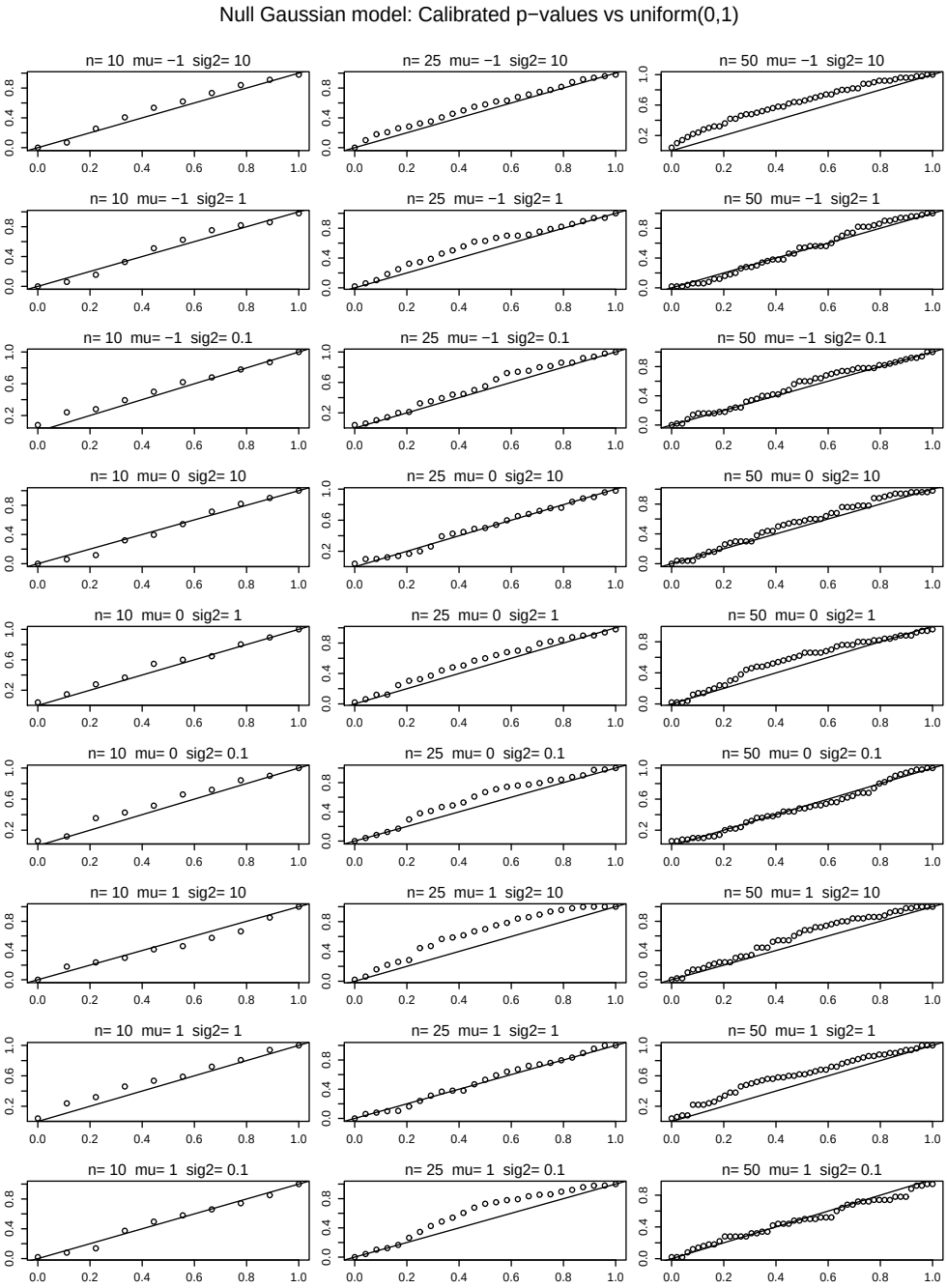

Figure 7: *Gaussian model: calibrated tail-area values.*

is not very often (in fact, only about 0.035 of the time, i.e., in this case the calibrated version of the uncalibrated Gelman et al. tail area is 10 times smaller). Figure 4 shows that the calibration of the Gelman et al. unadjusted approach improves in the one-sample Poisson setting, even for small n, as λ increases, but Figure 6 demonstrates that in the Gaussian model with both μ and σ^2 unknown, the Gelman et al. unadjusted approach is poorly calibrated across the entire subset $\{-1 \leq \mu \leq +1\} \times \{0.1 \leq \sigma^2 \leq 10\}$ of parameter space we examined, and things actually seem to get worse as n increases. Our adjusted results, by contrast (Figures 5 and 7), are nearly perfectly calibrated for all parameter values and sample sizes examined.

Remark. In drawing inferential conclusions, we do not support the practice of hypothesis/significance testing in general, and tail-area calculations (P-values) in particular, because (a) this approach is far less informative than posterior distributions, and the interval estimates they imply, on the scale of the data and (b) if the two hypotheses being compared in hypothesis testing correspond to different behavioral choices, the problem is really decision-theoretic rather than inferential and should be approached via MEU. However, we find it difficult to avoid using something like a tail area to calibrate model discrimination methods such as LS_{FS} in answering $\{Q_{2'}$: could the data have arisen from model M_{j^*}?$\}$. Other ways exist for judging how unusual $LS_{FS,actual}$ is in the calibrated distribution of LS_{FS} values — an obvious alternative is the ratio of the maximum height of the calibration density to its height at $LS_{FS,actual}$ — but all of them have an element of ad-hockery about where to draw the line. ♠

5 Asymptotic considerations

It has sometimes been suggested that *asymptotic consistency* is an important desideratum for both Bayesian and non-Bayesian model specification methods (note that this has nothing to do with *logical* consistency in the sense of Cox's formulation in Section 1.1). According to the main way this line of reasoning has played out in the literature to date, You are comparing two models M_1 and M_2, and You consider a data-generating mechanism M_{DG} under which one of the two models is correct (M_2, say); then if You generate data sets D_n with M_{DG}, using larger and larger sample sizes n, people who believe in the normative value of asymptotic consistency suggest that You should prefer a model-choice method that chooses M_2 more and more emphatically as $n \to \infty$ — with the model choice set $\mathcal{M} = \{M_1, M_2\}$ *remaining fixed* — over a model-choice method that does not have this property.

An appealing feature of this desideratum is that it appears to be founded on a desire to be well-calibrated; however, there are two caveats to consider:

- Since You will generally only have a data set D of finite sample size n_{actual}, what counts for You is how the model choice criteria You are comparing work in Your problem setting with Your n_{actual} value; asymptotic calculations that shed no light on this issue are of little value to You in actually solving Your problem.

- Also, as noted by Spiegelhalter et al. (2002), it may well be (emphatically) irrelevant to obtain asymptotic results that let $n \to \infty$ while holding the model-comparison set $\mathcal{M}$ fixed; in actual statistical practice, as n increases Your appetite for more complex models — to do a better job of approximating reality — will also increase, and it is not at all clear that lessons learned under the $\mathcal{M}$-fixed scenario have any normative value when the complexities of the models being compared are also on the rise.

To explore the relevance of asymptotic consistency in Bayesian model specification, consider the following simple example presented by Mukhopadhyay et al. (2005) (hereafter MGB): they compare M_1 — under which the data values $y = (y_1, \ldots, y_n)$ are modeled as $(y_i|\mathcal{B}) \overset{\text{IID}}{\sim} N(0,1)$ — with M_2, under which $(y_i|\theta,\mathcal{B}) \overset{\text{IID}}{\sim} N(\theta,1)$ for $\theta \in \Re$ with the improper prior $p(\theta|\mathcal{B}) = 1$ (the same results would be obtained with a proper and highly diffuse prior; note that, in frequentist hypothesis-testing language, this formulation amounts to testing the point-null hypothesis H_0: $\theta = 0$ against the compositive alternative H_A: $\theta \neq 0$ in the Gaussian sampling-distribution setting). It is straightforward to show in this situation that (a) as $|\theta| \to \infty$ with n fixed, the probability that LS_{FS} selects M_2 goes to 1, but (b) MGB show that this also occurs with LS_{CV} as $n \to \infty$ for $\theta = 0$ (and the same would be true of LS_{FS}); for MGB this is a serious criticism of the log-score approach. More generally it can be shown that $\{AIC, DIC, LS_{FS}\}$ may be inconsistent in the $n \to \infty$ sense in situations — e.g., the $\mathcal{M}$-closed setting (for finite $|\mathcal{M}|$) in which the data-generating M_{DG} is assumed to be in $\mathcal{M}$ and model M_j has k_j parameters, with k_j remaining fixed as $n \to \infty$ — in which {some Bayes factors, BIC} are asymptotically consistent. In response to this observation, (a) we re-emphasize the point made above, about the lack of realism of settings in which $\mathcal{M}$ remains fixed while n grows, and (b) we examine what appears to us to be the artificiality of the MGB example, as follows.

The prior MGB use in their model M_2 treats θ as a continuous quantity on $\Re$, which is appropriate scientifically in many applied settings, but the specification $\theta = 0$ in their model M_1 is logically inconsistent with the continuous treatment of θ on $\Re$. To fix ideas in seeing how to make this example more realistic, consider assessing the performance of a drug, for lowering systolic blood pressure (SBP) in hypertensive patients, in a phase–II clinical trial, and suppose (as did MGB) that a Gaussian sampling distribution for the outcome variable is reasonable (possibly after transformation). The two most frequent designs in settings of this type are:

- (quantifying improvement) Here You want to estimate the mean decline in blood pressure under this drug, and it would be natural to choose a repeated-measures (pre-post) experiment, in which SBP values are obtained for each patient, both before and after taking the drug for a sufficiently long period of time for its effect to become apparent. Let θ stand for the mean difference $(SBP_{before} - SBP_{after})$ in the population of patients to which it is appropriate to generalize from the patients in Your trial. There is nothing special about $\theta = 0$ in this setting, and in fact You *know* scientifically that θ is not exactly 0 (because the outcome variable in this experiment is conceptually continuous); what matters here is whether $\theta > \Delta$, where Δ is a *practical significance improvement threshold* below which the drug is not worth advancing into phase III (for example, any drug that did not lower SBP for severely hypertensive patients — those whose pre-drug values average 160 mmHg or more — by at least 15 mmHg would not deserve further attention). Thus, in the spirit of what MGB were attempting to examine, what counts in this situation is not a comparison of the models M_1 and M_2 above but a choice between

$$M_{1'}\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } \theta \leq \Delta \\ (y_i|\theta,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta,\sigma^2) \end{array} \right\} \tag{43}$$

  and

$$M_{2'}\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } \theta > \Delta \\ (y_i|\theta,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta,\sigma^2) \end{array} \right\}, \tag{44}$$

  in which, for simplicity, we follow MGB and take σ^2 to be known. As discussed in Section 1.1, an optimal real-world choice between $M_{1'}$ and $M_{2'}$ in this case would be based on a utility

function that quantified the costs and benefits of {taking the drug forward to phase III when it was correct to do so, taking it forward when it should have been abandoned, not taking it forward when it should have been, not taking it forward when it was correct not to do so}, but here we examine the performance of LS_{FS} in comparing $M_{1'}$ and $M_{2'}$, so that our asymptotic consistency results may be compared with those of MGB. Note that a natural inferential competitor to LS_{FS} in this case is simply to compute $\pi' = p(\theta > \Delta | y, \mathcal{B})$ and favor $M_{2'}$ if $\pi' > 0.5$.

- (establishing bio-equivalence) In this case there is a previous hypertension drug B (call the new drug A) and You are wondering if the mean effects of the two drugs are close enough to regard them as bio-equivalent. A good design here would again have a repeated-measures character, in which each patient's SBP is measured four times: before and after taking drug A, and before and after taking drug B (allowing enough time to elapse between taking the two drugs for the effects of the first drug to disappear). Let θ stand for the mean difference

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \tag{45}$$

in the population of patients to which it is appropriate to generalize from the patients in Your trial. Again in this setting there is nothing special about $\theta = 0$, and as before You *know* scientifically that θ is not exactly 0; what matters here is whether $|\theta| \leq c_{BE}$, where $c_{BE} > 0$ is a *practical significance bio-equivalence threshold* (e.g., 5 mmHg). Here again what counts is not a choice between MGB's models M_1 and M_2 but a comparison of

$$M_{1''}: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq c_{BE} \\ (y_i|\theta, \mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{46}$$

$$M_{2''}: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > c_{BE} \\ (y_i|\theta, \mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{47}$$

in which σ^2 is again taken to be known. As before, a careful real-world choice between $M_{1''}$ and $M_{2''}$ in this case would be based on a utility function that quantified the costs and benefits of {claiming the two drugs were bio-equivalent when they were, concluding that they were bio-equivalent when they were not, deciding that they were not bio-equivalent when they were, judging that they were not bio-equivalent when they were not}, but here we again examine the asymptotic consistency of LS_{FS} for comparison purposes. As above, a natural competitor to LS_{FS} here is simply to compute $\pi'' = p(|\theta| > c_{BE} | y, \mathcal{B})$ and choose $M_{2''}$ if $\pi'' > 0.5$.

The posterior predictive distributions $p(y^* | y, M_j, \mathcal{B})$ on which LS_{FS} is based are truncated-Normal mixtures of Normal sampling distributions under all of $\{M_{1'}, M_{2'}, M_{1''}, M_{2''}\}$, and it is straightforward to show that (a) the rate at which LS_{FS} correctly selects $M_{1'}$ over $M_{2'}$ goes to 1 as $n \to \infty$ for all data-generating $\theta_{DG} < \Delta$, (b) the rate at which LS_{FS} correctly selects $M_{2'}$ over $M_{1'}$ goes to 1 as $n \to \infty$ for all data-generating $\theta_{DG} > \Delta$, (c) the rate at which LS_{FS} correctly selects $M_{1''}$ over $M_{2''}$ goes to 1 as $n \to \infty$ for all data-generating $|\theta_{DG}| < c_{BE}$, and (d) the rate at which LS_{FS} correctly selects $M_{2''}$ over $M_{1''}$ goes to 1 as $n \to \infty$ for all data-generating $|\theta_{DG}| > c_{BE}$ (it does not matter what happens for $\theta_{DG} = \Delta$ in comparing $M_{1'}$ and $M_{2'}$ in the quantifying-improvement case, because $\theta_{DG} = \Delta$ is a zero-probability proposition under both $M_{1'}$ and $M_{2'}$, and similarly for $\theta_{DG} = c_{BE}$ in the bio-equivalence case). Thus we would argue that MGB's apparent asymptotic

inconsistency of LS_{CV} (and therefore also of LS_{FS}) in choosing between their M_1 and M_2 when $\theta_{DG} = 0$ was an artifact of their scientifically curious attempt to test a point-null hypothesis in a setting in which their modeling choices showed that their uncertainty about θ was continuous.

Table 5 documents the small-sample model discrimination performance of LS_{FS} in comparing $\{M_{1'}$ with $M_{2'}\}$ and $\{M_{1''}$ with $M_{2''}\}$, and contrasts this with the $\pi' > 0.5$ and $\pi'' > 0.5$ rules based directly on the posterior distribution for θ. We used equations (28) and (34) to evaluate LS_{FS}, with $m = 10{,}000$ Monte Carlo draws from $p(\theta|y, M_j, \mathcal{B})$ and with at least 4,000 simulation replications in each row of the table (the maximum Monte Carlo standard error in the proportion estimates was 0.008). The following conclusions may be drawn from the results in Table 5.

- $|LS_{FS}|$ increases nearly linearly with n in all scenarios examined.
- In the improvement-quantification case,
  - The results in the scenarios $(\Delta, \theta_{DG}, \sigma) = (10, 11, 10)$ and $(10, 9, 10)$ are symmetric up to Monte Carlo noise, as they must be.
  - It is remarkable that the LS_{FS} and π' rules, which use the data vector y in such different ways, yield identical model-discrimination results to almost three significant figures.
  - As a kind of stress test, the scenario $(\Delta, \theta_{DG}, \sigma) = (10, 10.1, 10)$ presented a difficult model-discrimination task for both LS_{FS} and π', because Δ and θ_{DG} were so close in relation to the standard deviation σ; at least $n = 1{,}280$ observations were required to achieve correct discrimination rates of 64% or more.
- With the bio-equivalence setup,
  - When θ_{DG} was chosen to make $M_{2''}$ true, the model-discrimination abilities of the LS_{FS} and π'' methods were again identical to nearly three significant figures.
  - In the not-very-difficult scenario $(c_{BE}, \theta_{DG}, \sigma) = (0.5, 0, 1)$ in which $M_{1''}$ was true, the π'' approach had a small-sample edge over LS_{FS} that (of course) decreased to 0 with increasing n.
  - As another stress test, the scenario $(c_{BE}, \theta_{DG}, \sigma) = (0.1, 0, 1)$ was much more challenging for both methods, because the region of practical equivalence was so small and the choice of θ_{DG} made $M_{1''}$ true. With $n \geq 160$, the central region $\bar{y} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}$, in which most of the posterior mass resides (for small α), was narrow enough for both approaches to identify the correct model at least 72% of the time, but with smaller samples sizes LS_{FS} typically had the better performance, with the π'' method choosing the wrong model 100% of the time for $n < 46$.

Remark. It is possible to object that the clinical trial setting explored here is insufficiently general, but — in our experience — when scientists come to us as consultants and ask for our help to test $H_0 : \theta = 0$, (a) this is because that is the frequentist problem formulation they have been trained to adopt and (b) some exploratory questioning reveals that what they really want to do is to assess the relative plausibility of $\theta \leq c$ versus $\theta > c$ or $|\theta| \leq c$ versus $|\theta| > c$ (for a problem-specific practical-significance threshold c), and these are just the improvement-assessment and bio-equivalence problems examined above. In our more than 40 combined years of consulting experience, we have never encountered an applied setting (call such a problem (∗∗)) in

Table 5: *Small-sample model discrimination performance of LS_{FS} and the π' or π'' rules.*

Improvement Quantification

				Simulation Mean of $LS_{FS}(M_j\|y,\mathcal{B})$		Rate at Which $M_{2'}$ Was Chosen By		Data-Generating
Δ	θ_{DG}	σ	n	$j=1'$	$j=2'$	LS_{FS}	π'	Model
10	11	10	10	−37.59	−37.15	0.631	0.630	$M_{2'}$
			20	−74.85	−74.25	0.681	0.679	
			40	−149.38	−148.57	0.720	0.720	
			80	−298.86	−297.61	0.821	0.821	
10	9	10	10	−37.20	−37.59	0.383	0.383	$M_{1'}$
			20	−74.21	−74.80	0.325	0.326	
			40	−148.55	−149.40	0.257	0.258	
			80	−297.40	−298.63	0.190	0.189	
10	10.1	10	10	−37.35	−37.30	0.513	0.514	$M_{2'}$
			20	−74.61	−74.53	0.527	0.525	
			40	−148.94	−148.82	0.535	0.535	
			80	−297.95	−298.83	0.548	0.549	
			160	−595.72	−595.56	0.558	0.557	
			320	−1190.4	−1190.2	0.571	0.573	
			640	−2382.3	−2382.0	0.602	0.602	
			1280	−4764.4	−4763.9	0.636	0.637	

Bio-Equivalence

				Simulation Mean of $LS_{FS}(M_j\|y,\mathcal{B})$		Rate at Which $M_{2''}$ Was Chosen By		Data-Generating
c_{BE}	θ_{DG}	σ	n	$j=1''$	$j=2''$	LS_{FS}	π''	Model
0.5	0	1	10	−13.81	−14.23	0.157	0.107	$M_{1''}$
			20	−27.91	−29.00	0.056	0.027	
			40	−56.31	−59.20	0.026	0.011	
0.5	1	1	10	−16.19	−13.85	0.943	0.943	$M_{2''}$
			20	−31.78	−28.03	0.989	0.989	
			40	−62.43	−56.28	1.000	1.000	
0.1	0	1	10	−14.16	−13.78	0.721	1.000	$M_{1''}$
			20	−28.25	−27.89	0.712	1.000	
			45	−63.79	−63.51	0.705	1.000	
			46	−65.24	−64.94	0.703	0.925	
			50	−70.83	−70.54	0.684	0.775	
			80	−113.56	−113.35	0.598	0.451	
			160	−227.34	−227.38	0.276	0.204	
			320	−453.80	−454.33	0.095	0.058	

which simultaneously (i) it was appropriate scientifically to model uncertainty about an observable outcome with a continuous sampling distribution indexed by a real-valued parameter θ and (ii) there was a real scientific need to distinguish between two theories, one of which held that θ was precisely equal to some value θ_0 and the other of which held that θ was free to vary continuously in $\Re$.

This statement applies to all types of problems on which we have consulted, including regression settings (cf. Draper (1999)): we *know* scientifically that a coefficient β_j measuring the effect of a predictor x_j on a continuous outcome y, after adjustment for the other predictor variables, is not precisely 0 — the scientific question is whether it is close enough to 0 *in both practical and statistical significance terms* for it to be sensible to exclude x_j — and, while people sometimes use *point-and-slab* priors (which put a point mass on 0 and spread the rest of the mass out continuously on $\Re$) on regression coefficients for computational convenience, we trust that they do so in the knowledge that such priors *never* express anyone's actual scientific uncertainty about a regression coefficient with a continuous outcome variable.

We are, of course, open to the possibility that (∗∗)-type problems exist, and we look forward to detailed descriptions of them. ♠

Remark. One possible defense of point-null hypothesis testing is as follows: "Of course we know that θ is not *exactly* 0 in the model $\{(y_i|\theta, \mathcal{B}) \stackrel{\text{IID}}{\sim} N(\theta, \sigma^2), p(\theta|\mathcal{B})$ diffuse, σ^2 known$\}$; the point of the test is to see if $\bar{y}$ is *close* to 0, with $1.96\frac{\sigma}{\sqrt{n}}$ as the yardstick." To this we would reply as follows:

- If the goal is inference about θ, we are in complete agreement with Box and Tiao (1973), who had no difficulty in writing a 588-page inferential text — which has long been recognized as a classic — without the need for the word "hypothesis" to appear in the index: the posterior distribution $(\theta|y, \mathcal{B}) \dot{\sim} N\left(\bar{y}, \frac{\sigma^2}{n}\right)$ summarizes the totality of Your information about θ, and gives rise to the 95% (*highest posterior density*: HPD) interval $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$; if You are wondering whether $\theta = 0$ has substantial support from the data, see if 0 is in the 95% HPD interval for θ.

- If an action needs to be taken based on whether θ is close to 0 or not, elicit the relevant utilities and use MEU, where the expectation is over uncertainty about θ as quantified via $(\theta|y, \mathcal{B}) \dot{\sim} N\left(\bar{y}, \frac{\sigma^2}{n}\right)$. ♠

Remark. It is also possible to note that in Sections 2 and 3.1 we compared fixed-effects (M_1) and random-effects (M_2) Poisson models, which amounts to choosing between $\sigma^2 = 0$ and $\sigma^2 > 0$ in the random-effects formulation M_2, and to object that (a) this does not cohere with the view expressed in this section about point-null hypothesis testing and (b) LS_{FS} will be inconsistent (in the MGB sense) for M_1, in that it tends to prefer M_2 even when You simulate with $\sigma^2 = 0$. To these objections we would reply that

- it is straightforward to re-formulate the results of Sections 2 and 3.1 by comparing $0 < \sigma^2 < c$ and $\sigma^2 > c$ in the random-effects model, for a small practical significance threshold $c > 0$, and when this is done (i) the results are similar to those presented earlier and (ii) in this formulation LS_{FS} is now asymptotically consistent, and

- even taking the MGB perspective, the behavior of LS_{FS} is actually desirable, because it is much worse to act as if $\sigma^2 = 0$ in random-effects models when it is not (this will result in a

potentially dramatic understatement of uncertainty) than to act as if $\sigma^2 > 0$ when (unknown to You) in fact $\sigma^2_{DG} = 0$. ♠

6 Discussion

We have argued in this paper that

- *calibration* — paying attention to how often You get the right answer — is a principle that (a) is important scientifically and (b) arises naturally in good Bayesian modeling;
- the question $\{Q_1$: Is model M_j better than $M_{j'}$?$\}$ is central to the process of well-calibrated Bayesian model specification; and
- this question is itself not well formed until You explicitly state the purpose to which the models will be put, which turns Bayesian model specification into a decision problem that should be addressed by maximizing expected utility (MEU), with a utility function that is sensitive to the purpose of the modeling exercise.

One may nevertheless observe empirically that, even though the above two points imply that MEU with a context-specific utility function is the only principled way to perform Bayesian model comparison, modelers have a powerful desire for more generic comparison tools that may serve as decent approximations to a principled (and therefore often resource-intensive) utility analysis. There appear to be three broad classes of generic tools of this type:

- Bayes factors — which are based on a 0–1 utility function in what Bernardo and Smith (1994) call the $\mathcal{M}$-*closed* view, in which You pretend that (a) there is a "true" data-generating model M_{DG} and (b) M_{DG} is in the set $\mathcal{M}$ of models You are considering — and their close cousin BIC;
- log-score criteria such as LS_{FS}, which are based on a predictive utility function; and
- methods that do not make any explicit appeal to utility at all, such as the information-theoretic AIC and DIC.

On the basis of the work presented here

- we regard DIC as a useful generalization of AIC and LS_{FS} as a further useful improvement upon DIC, with three advantages: LS_{FS} may well have better small-sample model discrimination behavior (as in the example simulated in Section 3.1); LS_{FS} is insensitive to model parameterization; and LS_{FS} can be used in Bayesian nonparametric as well as parametric settings;
- we contend that — if You wish to decide when to stop looking for better models by asking the question $\{Q_{2'}$: Could the data have arisen from model M_j?$\}$ — You should attempt to answer this question in a well-calibrated manner (see Section 4 for a method that achieves this with LS_{FS}); and

- until asymptotic calculations can be made, in problems of realistic difficulty, that closely mimic the process of (i) generating a set $\mathcal{M}$ of models that You are prepared to regard as a complete description of Your uncertainty — and doing so in a way that allows the complexity of the models to grow in a realistic manner with the amount of data — and (ii) examining the relative merits of these models, we believe that carefully-designed simulation studies will provide a more reliable guide to assessing calibration in practice than asymptotics that do not fully mimic the reality of practical model-building.

  In such simulation studies, the testbed should be constructed to be as similar as possible to the reality — in the actual problem under study — of generating $\mathcal{M}$ and comparing models within it, except that You know the truth in Your simulation world; then competitor methods for Bayesian model comparison such as {various flavors of Bayes factors, BIC} and LS_{FS} can be evaluated, in repeated sampling, on their ability to discover known truth (see Browne and Draper (2006) for examples of simulation environments along these lines, in variance-components and random-effects logistic regression models).

For reasons of space we intend to report elsewhere on comparisons between $\{BIC$ and other Bayes factors$\}$ and LS_{FS} in a variety of real-world-relevant modeling situations, and we hope that others will also undertake similar investigations, so that as a profession we can build up a body of comparative knowledge on which accurate and well-calibrated Bayesian model specification can rest.

Acknowledgments

We are grateful to colleagues in the Department of Applied Mathematics and Statistics at the University of California, Santa Cruz (UCSC), for comments and suggestions that improved this work, and to David Haussler and Jim Kent (in the Department of Biomolecular Engineering at UCSC) for access to the CPU cluster that made possible some of the simulation results we present here.

References

Aitkin, M. (1991). Posterior bayes factors (with discussion). *Journal of the Royal Statistical Society (Series B) 53*, 111–142.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Bayarri, M. J. and J. Berger (2000). p-values for composite null models (with discussion). *Journal of the American Statistical Association 95*, 1127–1170.

Bayarri, S. (2009). Objective bayesian testing and model selection. *Tutorial on Objective Bayesian Methodology: International Workshop on Objective Bayes Methodology, Philadelphia PA, 5 June 2009*.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Berger, J. (2006). The case for objective bayesian analysis (with discussion). *Bayesian Analysis 1*, 385–402.

Berger, J. (2009). Objective bayesian estimation. *Tutorial on Objective Bayesian Methodology: International Workshop on Objective Bayes Methodology, Philadelphia PA, 5 June 2009*.

Berger, J. and L. Pericchi (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association 91*, 109–122.

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

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society (Series A) 143*, 383–430.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Browne, W. J. and D. Draper (2006). Bayesian and likelihood methods for fitting multilevel models (with discussion). *Bayesian Analysis 1*, 473–550.

Carota, C. (2006). Some faults of the bayes factor in nonparametric model selection. *Statistical Methods and Applications 15*, 37–42.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics 14*, 1–13.

Cox, R. T. (1961). *The Algebra of Probable Inference*. Baltimore: Johns Hopkins University Press.

Dawid, A. P. (1984). Present position and potential developments: some personal views. statistical theory: the prequential approach (with discussion). *Journal of the Royal Statistical Society (Series A) 147*, 278–292.

Dawid, A. P. (1985). Calibration-based empirical probability (with discussion). *Annals of Statistics 13*, 1251–1285.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *Journal of the Royal Statistical Society (Series B) 53*, 79–109.

Dawid, A. P. (1997). Prequential analysis. In S. Kotz (Ed.), *Encyclopedia of Statistical Sciences*, Volume 1 (updated), New York, pp. 464–470. Wiley.

De Finetti, B. (1930). Funzione caratteristica di un fenomeno aleatorio. *Atti Reale Accademia Nazionale dei Lincei 4*, 86–133.

De Finetti, B. (1938). Sur la condition d'équivalence partielle. *Actualités Scientifiques et Industrielles 739*.

De Finetti, B. (1974). *Theory of Probability*. New York: Wiley (volumes 1 and 2).

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B) 57*, 45–97.

Draper, D. (1996). Utility, sensitivity analysis, and cross-validation in bayesian model checking. discussion of 'posterior predictive assessment of model fitness via realized discrepancies', by a. gelman, x.-l. meng, and h. stern. *Statistica Sinica 6*, 760–767.

Draper, D. (1999). Model uncertainty yes, discrete model averaging maybe. discussion of 'bayesian model averaging: a tutorial', by j. a. hoeting, d. madigan, a. e. raftery, and c. t. volinsky. *Statistical Science 14*, 405–409.

Draper, D. (2006). Coherence and calibration: comments on subjectivity and 'objectivity' in bayesian analysis. discussion of 'the case for objective bayesian analysis,' by j. berger and 'subjective bayesian analysis: principles and practice,' by m. goldstein. *Bayesian Analysis 1*, 423–428.

Draper, D. (2008). Exchangeability. In B. N. Petrov and F. Csaki (Eds.), *The New Palgrave Dictionary of Economics (second edition)*, Volume 3, Basingstoke, UK, pp. 99–103. Palgrave Macmillan.

Draper, D., J. Hodges, C. Mallows, and D. Pregibon (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A 156*, 9–37.

Draper, D. and M. Krnjajić (2010). 3cv: Well-calibrated bayesian model specification. *Technical Report, Department of Applied Mathematics and Statistics, University of California, Santa Cruz*.

Draper, D. and V. von Brzeski (2010). Bayesian decision theory and calibration. *Technical Report, Department of Applied Mathematics and Statistics, University of California, Santa Cruz*.

Fouskakis, D. and D. Draper (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. *Journal of the American Statistical Association 103*, 1367–1381.

Geisser, S. (1980). Discussion of "sampling and bayes' inference in scientific modelling and robustness", by g. e. p. box. *Journal of the Royal Statistical Society (Series A) 143*, 416–417.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, New York, pp. 145–161. Chapman & Hall.

Gelfand, A. E. and D. Dey (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society (Series B) 56*, 501–514.

Gelfand, A. E. and S. J. Ghosh (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika 85*, 1–11.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica 6*, 733–807.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Hendriksen, C., E. Lund, and E. Stromgard (1984). Consequences of assessment and intervention among elderly people: a three year randomised controlled trial. *British Medical Journal 289*, 1522–1524.

Hoeting, J. A., D. Madigan, A. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science 14*, 382–417.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: University Press.

Key, L., L. Pericchi, and A. Smith (1999). Bayesian model choice: what and why? (with discussion). In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics (Volume 6)*, Oxford, pp. 343–370. University Press.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Ergebnisse der Mathematik.

Krnjajić, M. (2005). Contributions to bayesian statistical analysis: model specification and nonparametric inference. *Ph.D. dissertation, Department of Applied Mathematics and Statistics, University of California, Santa Cruz*.

Krnjajić, M., A. Kottas, and D. Draper (2008). Parametric and nonparametric bayesian model specification: a case study involving models for count data. *Computational Statistics and Data Analysis 52*, 2110–2128.

Laplace, P. S. (1774). Mêmoire sur la probabilité des causes par les évènemens. *Mêmoires de mathématique et de physique presentés à l'Academie royale des sciences, par divers savans, & lûs dans ses assemblées 6*, 621–656.

Laud, P. and J. G. Ibrahim (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B 57*, 247–262.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference With Non-experimental Data*. New York: Wiley.

Little, R. (2006). Calibrated bayes: a bayes/frequentist roadmap. *The American Statistician 60*, 1151–1172.

Mukhopadhyay, N., J. K. Ghosh, and J. O. Berger (2005). Some bayesian predictive approaches to model selection. *Statistics and Probability Letters 73*, 369–379.

O'Hagan, A. and J. Forster (2004). *Kendall's Advanced Theory of Statistics, Second Edition, Volume 2B: Bayesian Inference*. London: Arnold.

Parmigiani, G. and L. Inoue (2009). *Decision Theory: Principles and Approaches*. New York: Wiley.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and bayes factors. In D. K. Dey and C. R. Rao (Eds.), *Handbook of Statistics (Volume 25): Bayesian Thinking, Modeling and Computation*, Elsevier, pp. 115–149. North Holland.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society (Series B) 52*, 175–184.

Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and Other Logical Essays*, London, pp. 156–198. Kegan Paul.

Robins, J. M., A. van der Vaart, and V. Ventura (2000). Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association 95*, 1143–1156.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics 12*, 1151–1172.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society (Series B) 64*, 583–639.

Venn, J. (1866). *The Logic of Chance*. London: Macmillan.

von Mises, R. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Vienna: Springer.