

Bayesian model specification

David Draper[†] and Milovan Krnjajić
University of California, Santa Cruz, USA

Summary. A standard (data-analytic) approach to statistical model specification, practiced with equal vigor in both Bayesian and non-Bayesian approaches to model-building, involves the initial choice, for the structure of the model, of one or another of a variety of standard parametric families, followed by modification of this initial choice—once data begin to arrive—if the data suggest deficiencies in the original specification. In this paper (a) we argue that this approach is formally incoherent, because it amounts to using the data both to specify the prior distribution on structure space and to update using this data-determined prior; (b) we identify two approaches to avoiding (at least in principle, and with a fair amount of data) the incoherence in (a): (1) Bayesian semi-parametric modeling and (2) three-way out-of-sample predictive validation; (c) we provide details on implementing (2); (d) we argue that to make progress in coherent Bayesian model specification in complicated problems You (the modeler) have to either implicitly or explicitly choose a utility structure which defines, for You, when the model currently being examined is “good enough”; (e) we argue that it is best to make this choice explicitly on the basis of real-world considerations regarding the use to which the model will be put; and (f) we contrast model selection methods based on the log score and deviance information criteria (DIC) as two examples of (e) with utilities governed by predictive accuracy.

Keywords: Bayesian model specification, DIC, model selection as a decision problem, predictive log scoring rule, three-way out-of-sample predictive validation

1 Introduction: what is a Bayesian model?

This paper is about methods for comparing, criticizing, and specifying Bayesian statistical models. The question in the title of this section has a bewildering array of possible answers (e.g., in September 2005 the search string *Bayesian model* generated more than 3.9 million hits at a leading web search engine, and yielded about 8,400 published articles since 1975 using the searching capabilities of a leading electronic article data base). We regard a *Bayesian model* as a mathematical framework for quantifying uncertainty about unknown quantities by relating them to known quantities. The model will typically embody a variety of assumptions \mathcal{A} and judgments \mathcal{J} , and it is desirable for \mathcal{A} and \mathcal{J} to arise as directly as possible from the contextual information in the problem we study.

The most appealing approach to achieving this goal appears to be that of de Finetti (1970), who regarded a Bayesian model as a *joint predictive distribution* $p(y)$ for observables $y = (y_1, \dots, y_n)$ which have not yet been observed. The following examples illustrate the model specification process from this point of view.

[†]*Address for correspondence:* 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, web www.ams.ucsc.edu/~draper).

1.1 Example 1: Binary outcomes, no predictors

Consider observing health outcomes over a specified time window for all patients at one hospital with an admission diagnosis such as heart attack. In this first example we focus on the simplest possible observables: y_i is 1 if patient i dies within 30 days of admission and 0 otherwise ($i = 1, \dots, n$), and no predictor variables are available. As de Finetti (1930) noted, in the absence of any other information our predictive uncertainty about the y_i is *exchangeable*, in the usual sense that $p(y)$ is invariant under permutation of the labels on the patients. Continuing to follow de Finetti, if we are willing to regard (y_1, \dots, y_n) as part of an infinitely exchangeable sequence of binary outcomes (meaning that we regard our uncertainty about all finite subsets of this sequence as exchangeable), then any *coherent* predictive distribution $p(y)$ can be given the simple hierarchical representation $(y_i|\theta) \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta), \theta \sim p(\theta)$ for some density $p(\theta)$ on $(0, 1)$, where θ has a dual interpretation as (a) the limiting value of the mean of the y_i in the infinite sequence (which must exist by exchangeability) and (b) the marginal probability that any of the patients in the sequence will die within 30 days of admission (which must be the same for all patients by exchangeability). Mathematically $p(\theta)$ is just a mixing distribution in the expression

$$p(y_1, \dots, y_n) = \int_0^1 \prod_{i=1}^n p(y_i|\theta) p(\theta) d\theta, \quad (1)$$

which is equivalent to the previously stated simple hierarchical model; statistically, of course, $p(\theta)$ provides an opportunity for us to quantify any (prior) information (external to the current set of observables) about θ and combine this with the information in y . Thus, in this simplest situation, Bayesian model specification is equivalent to choosing a scientifically appropriate prior distribution $p(\theta)$; the rest of the model specification details in (1) arise directly from exchangeability, which in turn is driven by the real-world context.

1.2 Example 2: Continuous outcomes, no predictors.

In a small elaboration of Example 1, consider predicting a real-valued observable y_i for each patient, such as a score measuring sickness on admission to hospital, in a setting in which there are still no predictor variables. Our uncertainty about the y_i is still exchangeable; de Finetti's (1937) representation theorem for real-valued data now gives that, if we regard (y_1, \dots, y_n) as part of an infinitely exchangeable sequence, then any coherent joint predictive distribution $p(y)$ can be expressed in the hierarchical form $(y_i|F) \stackrel{\text{iid}}{\sim} F, F \sim Q(F)$, where F is the limiting empirical cumulative distribution function (CDF) of the infinite sequence (y_1, y_2, \dots) , which again must exist by exchangeability. The corresponding equivalent mixture form for the predictive distribution is

$$p(y_1, \dots, y_n) = \int_{\mathcal{D}} \prod_{i=1}^n p(y_i|F) dQ(F), \quad (2)$$

where \mathcal{D} is the set of all possible CDFs (note that (1) is a special case of (2)). As in the previous example, Bayesian model specification amounts to choosing a scientifically appropriate

prior distribution, $Q(F)$. However, here the unknown distribution F is in effect an infinite-dimensional parameter, requiring us to put a scientifically relevant probability distribution on \mathcal{D} . Specifying distributions on function spaces is the task of Bayesian nonparametric (BNP) modeling (e.g., Dey et al. 1998), which is considered in detail, e.g., in Krnjajić, Draper and Kottas (2005).

1.3 Example 3: Integer-valued outcomes with a covariate

In practice, of course, in addition to outcomes y_i , covariates x_{ij} will typically be available. For instance, in a case study (Hendriksen et al. 1984) to which we will return several times, 572 elderly people were randomized, 287 to a control (C) group (which received standard care) and 285 to a treatment (T) group (which received 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 (T/C) status is the only available covariate. Then the assumption of *unconditional* exchangeability across all 572 outcomes is no longer automatically scientifically appropriate. Instead the design of the experiment implies (at least initially) *partial* or *conditional exchangeability* (e.g., de Finetti 1938; Draper et al. 1993) given T/C status, and this leads by a simple generalization of the representation theorem in Example 1.2 to the Bayesian nonparametric model

$$\begin{aligned} (F_T, F_C) &\sim p(F_T, F_C) \\ (y_i^T | F_T, F_C) &\stackrel{\text{iid}}{\sim} F_T \quad \text{and} \quad (y_j^C | F_T, F_C) \stackrel{\text{iid}}{\sim} F_C. \end{aligned} \tag{3}$$

Note that even in this rather general nonparametric framework it will be necessary to have a good tool for discriminating between the quality of two models (here: unconditional exchangeability ($F_T = F_C$; T has the same effect as C) versus conditional exchangeability ($F_T \neq F_C$; the T and C effects differ)).

This framework—in which (a) covariates specify conditional exchangeability assumptions in a manner driven completely by the problem context and (b) one version or another of de Finetti’s representation theorems reduces the model specification task to placing appropriate prior distributions on CDFs—seems to cover much of the field of statistical inference and prediction with minor extensions (e.g., multiple discrete covariates can be cross-tabulated, and continuous covariates can be discretized; when the number of cells in the resulting layout becomes too large, assumptions about how those cells are related to each other become necessary). However, placing real-world-relevant prior distributions on CDFs is not straightforward, and the statistics profession does not have much experience with this process yet; in the meantime, in parallel with efforts to accumulate such experience, a great deal of parametric modeling will occur, and tools for specifying such models will often be employed. We review both parametric and nonparametric Bayesian model specification below and offer some new methodological details.

The plan of the paper is as follows. In Section 2 we examine a data-analytic approach to model specification which is employed frequently in both frequentist and Bayesian approaches

to statistical work. Section 3 demonstrates that model choice is really a decision problem which should be approached via maximization of expected utility, with a utility structure that is sensitive to the real-world context. In Section 4 (a) we examine the log score LS , a generic utility structure for model choice appropriate in situations where predictive accuracy is key; (b) we establish a connection between a cross-validation version LS_{CV} of the log score idea and the deviance information criterion DIC ; and (c) we discuss a full-sample version LS_{FS} of the log score approach and demonstrate its small-sample superiority over LS_{CV} and DIC for model discrimination in fixed- versus random-effects Poisson modeling. Section 5 explores connections between LS and Bayes factors. In Section 6 we examine the question “Could the data have arisen from model M ?” and illustrate the use of an algorithm for answering this question in a well-calibrated way, and Section 7 offers some conclusions.

2 Data-analytic model specification

The basic problem of statistical model-building can be stated as follows: In modeling our uncertainty about future observables $y = (y_1, \dots, y_n)$, we recognize that we are uncertain about y (this might be termed *first-order uncertainty*), but we also acknowledge that we are uncertain about how to specify our uncertainty about y (which might be called *second-order uncertainty*). A fundamental problem in Bayesian modeling is how to cope with both of these levels of uncertainty in a manner that is both coherent and *well-calibrated*. These criteria are of course not the same: we want to be coherent in our implementation of Bayes (otherwise there are internal inconsistencies in our probability assessments), but coherence by itself is not enough to guarantee that our Bayesian answer is a good answer to a real-world question (we are always free in the coherent Bayesian paradigm to insert extremely strong prior information that is, after the fact, seen to be out of step with the world, and if we do so our Bayesian solution will be poor indeed). This forces us to be guided, not only by coherence, but also by calibration: as scientific collaborators we want to be free to use Bayesian methods, but if we want to get invited back to collaborate again (and again) we had better pay attention to how often we get the right answer (e.g., meteorologists who consistently get it wrong about when it will rain will quickly be ignored, or fired, or both), and this is a fundamentally calibrative activity. The *objective Bayes* movement (e.g., Berger 2006) has points of contact with this view; also see Rubin (1984).

A frequently-employed *data-analytic* approach to model-building involves an initial choice, for the structure of the model, of a standard parametric family, followed by modification of the initial choice—once data begin to arrive—if the data suggest deficiencies in the original specification; indeed, a search is typically conducted, based on the data, for the apparently “best” model M^* . This approach (e.g., Draper 1995) is formally incoherent if no attempt is made to pay an appropriate price for having chosen the structure of the model in a data-driven fashion: in effect it uses the data both to specify the prior distribution on structure space and to update that prior using the same data. The result will typically be uncalibrated predictive distributions for future data, and the lack of calibration will typically manifest itself as a bias in favor of predictive intervals that are too narrow to accommodate the full uncertainty which the future holds.

This dilemma, of how to approach the problem of both first- and second-order uncertainty,

is an example of what Lindley (1985) termed *Cromwell's Rule*: if the initial model choice places zero prior probability on large regions of model space then, formally, all such regions must also have zero posterior probability even if the data indicate that a different prior on model space would have been better.

We are aware of only two possible solutions to the dilemma posed by Cromwell's Rule in Bayesian model specification: (a) Bayesian nonparametric modeling and (b) a modified data-analytic approach that might be termed *three-way cross-validation* (3CV).

- As noted by, e.g., Walker et al. (2004), if we use a prior that places non-zero probability on all Kullback-Leibler neighborhoods of all CDFs F (both *Pólya trees* (e.g., Walker et al. 1999) and *Dirichlet process mixture priors* (e.g., Dey et al. 1998) succeed in this goal if specified properly), then Bayesian nonparametric modeling directly avoids the Cromwell's Rule dilemma, at least for large n : as $n \rightarrow \infty$ the posterior on F will discard any incorrect details of prior specification and will fully adapt to the actual data-generating F (this line of reasoning of course assumes correct exchangeability judgments). When well specified, BNP priors on CDFs thus solve the problem by, in effect, not placing zero prior probability on any scientifically relevant subsets of the space of all possible models.
- Three-way cross-validation solves the problem in a different way, by exploring the data for the best model but then paying an appropriate price for the exploration. 3CV takes the usual cross-validation idea one step further, as follows:
 - (1) Partition the data at random into *three* (non-overlapping and exhaustive) subsets S_i .
 - (2) Fit a tentative model ($\{\text{likelihood} + \text{prior}\}$) to S_1 . Expand the initial model in all feasible ways suggested by data exploration using S_1 . Iterate until the model fit is satisfactory (methods for assessing the fit will be examined later in this chapter).
 - (3) Use the final model (fit to S_1) from (2) to create predictive distributions for all data points in S_2 . Compare actual outcomes with these distributions, checking for predictive calibration. Go back to (2), change the likelihood as necessary, re-tune the priors as necessary, and so on, to get good calibration. Iterate until the predictive distributions accurately capture the data in both S_1 and S_2 .
 - (4) Announce the final model (fit to $S_1 \cup S_2$) from (3), and report predictive calibration of this model on the data points in S_3 as an indication of how well it would perform with new data.

With large n it is only necessary to do steps (1–4) once; with small and moderate n it is best to repeat the above steps several times and use Bayesian model averaging (e.g., Draper 1995) to combine the results.

Both of these approaches lead to uncertainty bands that are typically (and appropriately) somewhat wider than those obtained by the data-analytic M^* approach described above:

- (a) If a parametric model can be found that fits the data equally well using the M^* approach, BNP modeling will typically produce predictive and inferential intervals that are somewhat wider than those from the parametric modeling, but the parametric intervals are

narrower than they should be because no price was paid, in finding the “best” parametric model, for the model search.

- (b) By explicitly holding out subset S_3 as a proxy for future data in the 3CV approach, the final model fit to $S_1 \cup S_2$ will yield somewhat wider intervals than if it were fit to the entire data set. We have no definitive results yet for the optimal fractions of data to assign to the subsets S_i (this is a subject of on-going study); we conjecture that $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is a reasonable (not far from optimal) choice. 3CV will be further explored and illustrated elsewhere (Draper and Krnjajić 2005).

3 Model selection as a decision problem

Given a method like 3CV which permits us to explore model space without forfeiting calibration, two kinds of model specification questions (in both parametric and nonparametric Bayesian modeling) arise:

- (1) Is model M_1 better than M_2 ? (this tells us when it is reasonable to discard a model in our search), and
- (2) Is M_1 good enough? (this tells us when it is reasonable to stop searching).

To bring these two questions into sufficiently sharp focus to begin answering them, the terms “better than” and “good enough” must be made more precise. The following principle, which seems to us to be essentially self-evident, is crucial in this effort.

Model Selection Principle (MSP). It is not possible to choose a model well without contemplating the purpose to which it will be put; for how else will you know if the model under scrutiny is “good enough”? (Good enough for what?)

Specifying this purpose demands a *decision-theoretic* basis for model choice (e.g., Draper 1996; Key et al. 1998).

It is useful to distinguish two cases:

- (1) If we are going to choose which of several ways to behave in the future, then the model has to be good enough to reliably aid in choosing the best behavior; or
- (2) If instead we simply wish to make a scientific summary of what’s known, then—remembering that a hallmark of good science is good prediction—the model has to be good enough to make sufficiently accurate predictions of observable outcomes (in which the dimensions along which accuracy is to be monitored are driven by what is scientifically relevant).

As an example of case (1), Draper and Fouskakis (2000, 2005) (also see Fouskakis and Draper 2002) give a case study of decision-theoretic model choice in action. The problem they addressed was to construct a scale measuring sickness at admission to hospital for elderly pneumonia patients, in an environment in which costs were constrained. The main issue of

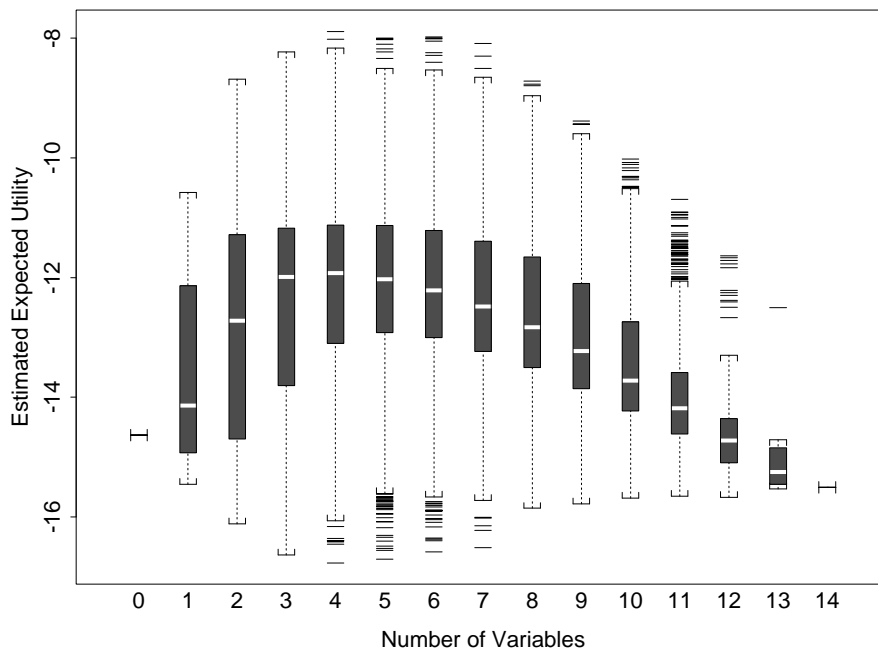


Figure 1: *Variable selection in a generalized linear model regression setting, from Draper and Fouskakis (2000). Estimated expected utility as a function of the number of predictor variables, in a problem involving construction of a cost-effective scale to measure sickness at hospital admission of elderly pneumonia patients.*

model specification was the usual problem of variable selection, but standard (“benefit-only”) methods that pay attention only to how well variables predict the outcome of interest are sub-optimal in this setting; instead a cost-benefit trade-off is needed, in which the final model should only contain variables that predict well enough given how much they cost to collect. Draper and Fouskakis solved this model specification problem with Bayesian decision theory, by formulating a utility function with two components—one quantifying data collection costs associated with the construction of a given sickness scale, and the other rewarding and penalizing the scale’s predictive successes and failures—and maximizing expected utility (MEU). Figure 1 gives an example of their results, in a setting with 14 predictors (chosen via standard benefit-only methods from the 85 available variables) and $2^{14} = 16,384$ possible models. It is evident from the figure that the best models only have 4–6 sickness indicators and that these models are far better at solving the real-world problem than the standard benefit-only solution (which includes all 14 variables).

In case (2) the main goal instead is simply a summary of scientific knowledge, which suggests (as noted above) a utility function that rewards predictive accuracy. In specifying such a utility function in a reasonably general way—to answer model specification question (1) above, “Is M_1 better than M_2 ?”—we need a *scoring rule* that measures the discrepancy between an observation y^* and the predictive distribution $p(\cdot|y, M_i)$ for that observation under model M_i given data y . As noted (e.g.) by Good (1950) and O’Hagan and Forster (2004), the optimal

(*impartial, symmetric, proper*) scoring rules are linear functions of $\log p(y^*|y)$.

In the spirit of the above 3CV discussion, on calibration grounds it would seem to be a mistake to use the data twice in assessing predictive accuracy (once to make predictions, and again with the same data to see how good they are). We will see later in this paper that this is not necessarily true, but for now we begin our examination of the log scoring idea by revisiting the out-of-sample predictive validation method of Geisser and Eddy (1979) and Gelfand et al. (1992): successively remove each observation y_j one at a time, construct the predictive distribution for y_j based on y_{-j} (the data vector with y_j removed) and see where y_j falls in this distribution. This motivates a cross-validation variant of the log-scoring rule (e.g., Good 1950; Gelfand and Dey 1994; Bernardo and Smith 1994): with n data values y_j , when choosing among k models $M_i, i = 1, \dots, k$, find that model M_i which maximizes

$$LS_{CV}(M_i|y) = \frac{1}{n} \sum_{j=1}^n \log p(y_j|M_i, y_{-j}). \quad (4)$$

(Item (1) in the Appendix gives details on how to calculate $p(y_j|M_i, y_{-j})$ via MCMC.)

It has been argued that this can be given a direct decision-theoretic justification: defining the utility function for model i as

$$U(M_i|y) = \log p(y^*|M_i, y), \quad (5)$$

where y^* is a future data value, the expectation in MEU is over our uncertainty about y^* . Bernardo and Smith (1994) claim that this expectation can be closely approximated (assuming exchangeability) by (4):

$$E[U(M_i|y)] \approx \frac{1}{n} \sum_{j=1}^n \log p(y_j|M_i, y_{-j}). \quad (6)$$

We shall revisit this claim below.

It can also be revealing when the predictive distributions are approximately Gaussian to compute predictive z -scores, for observation j under model i :

$$z_{ij} = \frac{y_j - E(y_j|M_i, y_{-j})}{\sqrt{V(y_j|M_i, y_{-j})}}. \quad (7)$$

For good predictive calibration the $\{z_{ij}, j = 1, \dots, n\}$ should have mean 0 and standard deviation (SD) 1 for each i ; we often find instead that the SD is larger than 1, signifying that the predictive uncertainty bands are not wide enough.

4 Log-score and the Deviance Information Criterion

With large data sets, in situations in which the predictive distribution has to be estimated by MCMC, direct calculation of LS_{CV} is computationally expensive, since it requires $O(n)$ MCMC runs for a sample of size n (with discrete or count data the number of MCMC runs may be smaller than n , i.e., equal to the number of unique points in the sample). In this section we look for a computationally efficient alternative to LS_{CV} and explore the relation between two variants of LS and a recent popular method for Bayesian model choice, the *deviance information criterion* (DIC), proposed by Spiegelhalter et al. (2002).

4.1 LS_{CV} and DIC

To see how a fast approximation to LS_{CV} might be obtained, it is useful to examine how the log score works in a simple model, e.g., M_0 : for $i = 1, \dots, n$,

$$\begin{aligned} \mu &\sim N(\mu_0, \sigma_\mu^2) \\ (Y_i|\mu) &\stackrel{\text{iid}}{\sim} N(\mu, \sigma^2) \end{aligned} \quad (8)$$

with σ known; take a highly diffuse prior on μ so that the posterior for μ is approximately

$$(\mu|y) = (\mu|\bar{y}) \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right), \quad (9)$$

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

$$(y_{n+1}|y) = (y_{n+1}|\bar{y}) \sim N\left[\bar{y}, \sigma^2 \left(1 + \frac{1}{n}\right)\right], \quad (10)$$

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

$$LS_{CV}(M_0|y) = \sum_{j=1}^n \ln p(y_j|y_{-j}), \quad (11)$$

where as before y_{-j} is y with observation j set aside. But by the same reasoning

$$p(y_j|y_{-j}) \doteq N(\bar{y}_{-j}, \sigma_n^2), \quad (12)$$

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

$$\begin{aligned} \ln p(y_j|y_{-j}) &\doteq c - \frac{1}{2\sigma_n^2}(y_j - \bar{y}_{-j})^2 \quad \text{and} \\ LS_{CV}(M_0|y) &\doteq c_1 - c_2 \sum_{j=1}^n (y_j - \bar{y}_{-j})^2 \end{aligned} \quad (13)$$

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

$$\sum_{j=1}^n (y_j - \bar{y}_{-j})^2 = c \sum_{j=1}^n (y_j - \bar{y})^2 \quad (14)$$

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

$$LS_{CV}(M_0|y) \doteq c_1 - c_2 \sum_{j=1}^n (y_j - \bar{y})^2, \quad (15)$$

i.e., in M_0 the log score is almost perfectly negatively correlated with the sample variance. But in this model the *deviance* (minus twice the log likelihood) is

$$\begin{aligned} D(\mu) &= -2 \ln l(\mu|y) = c_0 - 2 \ln p(y|\mu) \\ &= c_0 + c_3 \sum_{j=1}^n (y_j - \mu)^2 \end{aligned} \quad (16)$$

for some $c_3 > 0$, encouraging the suspicion that LS_{CV} should be strongly related to the deviance.

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 $D(\bar{\theta})$ 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:

$$DIC(M|y) = D(\bar{\theta}) + 2 \hat{p}_D, \quad (17)$$

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

$$\hat{p}_D = \overline{D(\theta)} - D(\bar{\theta}), \quad (18)$$

i.e., \hat{p}_D is the difference between {the posterior mean of the deviance} and {the deviance evaluated at the posterior mean of the parameters} (the popular freeware package WinBUGS release 1.4 will estimate these quantities). In model M_0 , p_D is of course 1, and $\bar{\theta} \doteq \bar{y}$, so

$$DIC(M_0|y) \doteq c_0 + c_3 \sum_{j=1}^n (y_j - \bar{y})^2 + 2 \quad (19)$$

and the conclusion is that

$$-DIC(M_0|y) \doteq c_1 + c_2 LS_{CV}(M_0|y) \quad (20)$$

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 count data, e.g., the number of hospitalizations in each of the T and C portions of the IHGA data (Example 3 in Section 1.3). With data of this type modelers often choose between fixed- and random-effects Poisson model formulations: for $i = 1, \dots, n$, and, e.g., with diffuse priors,

$$M_1: \left\{ \begin{array}{c} \lambda \\ (y_i|\lambda) \end{array} \begin{array}{c} \sim \\ \stackrel{\text{iid}}{\sim} \end{array} \begin{array}{c} p(\lambda) \\ \text{Poisson}(\lambda) \end{array} \right\} \quad \text{versus} \quad (21)$$

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

Group	Number of Hospitalizations								n	Mean	SD
	0	1	2	3	4	5	6	7			
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

$$M_2: \left\{ \begin{array}{ll} (\beta_0, \sigma^2) & \sim p(\beta_0, \sigma^2) \\ (y_i | \lambda_i) & \stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = \beta_0 + e_i \\ (e_i | \sigma^2) & \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \end{array} \right\} \quad (22)$$

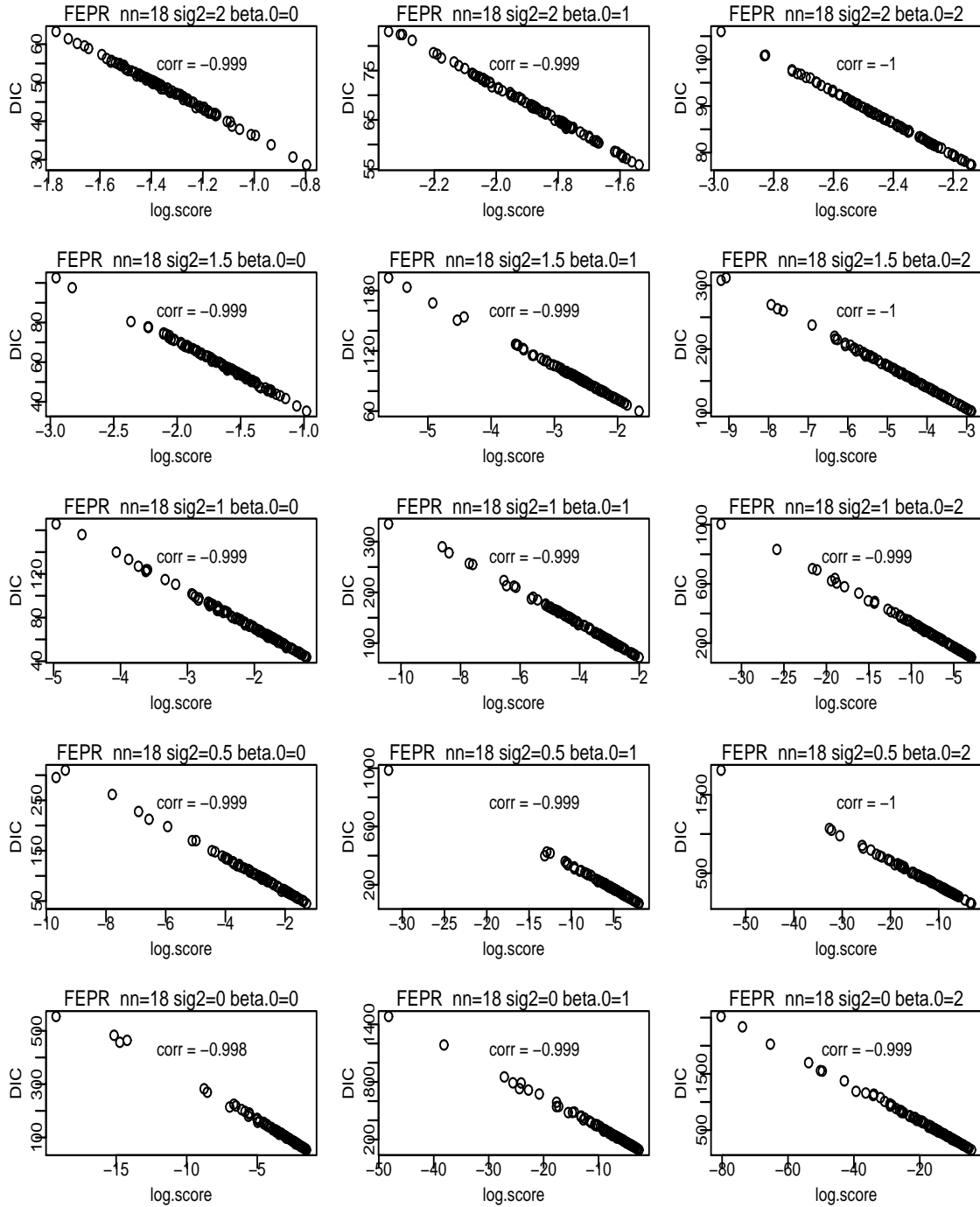
M_1 is of course a special case of M_2 with $(\sigma^2 = 0, \lambda = e^{\beta_0})$; the likelihood in M_2 is a Lognormal mixture of Poissons (this is often similar to fitting a Negative Binomial distribution, which is a Gamma mixture of Poissons).

We conducted a partial-factorial simulation study with factors $\{n = 18, 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0, 2.0\}$, $\{\sigma^2 = 0.0, 0.5, 1.0, 1.5, 2.0\}$ 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 2 and 3 summarize some of the results of this simulation. The first of these two Figures demonstrates that when the assumed model is M_1 (the fixed-effects Poisson), LS_{CV} and DIC are almost perfectly negatively correlated (we have a mathematical explanation of this which will be presented elsewhere); the second Figure shows by contrast that when the assumed model is M_2 (the random-effects Poisson), LS_{CV} and DIC are less strongly negatively correlated, but (not shown in the Figure) the correlation increases with n .

As a further example of the correspondence between LS_{CV} and DIC , the full IHGA data are given in Table 1. 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's 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);
- A one-sample Poisson model with a diffuse prior, which in effect assumes that the treatment and control λ s 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,

Figure 2: DIC versus $\log\text{-score}$; true model is M_1 .

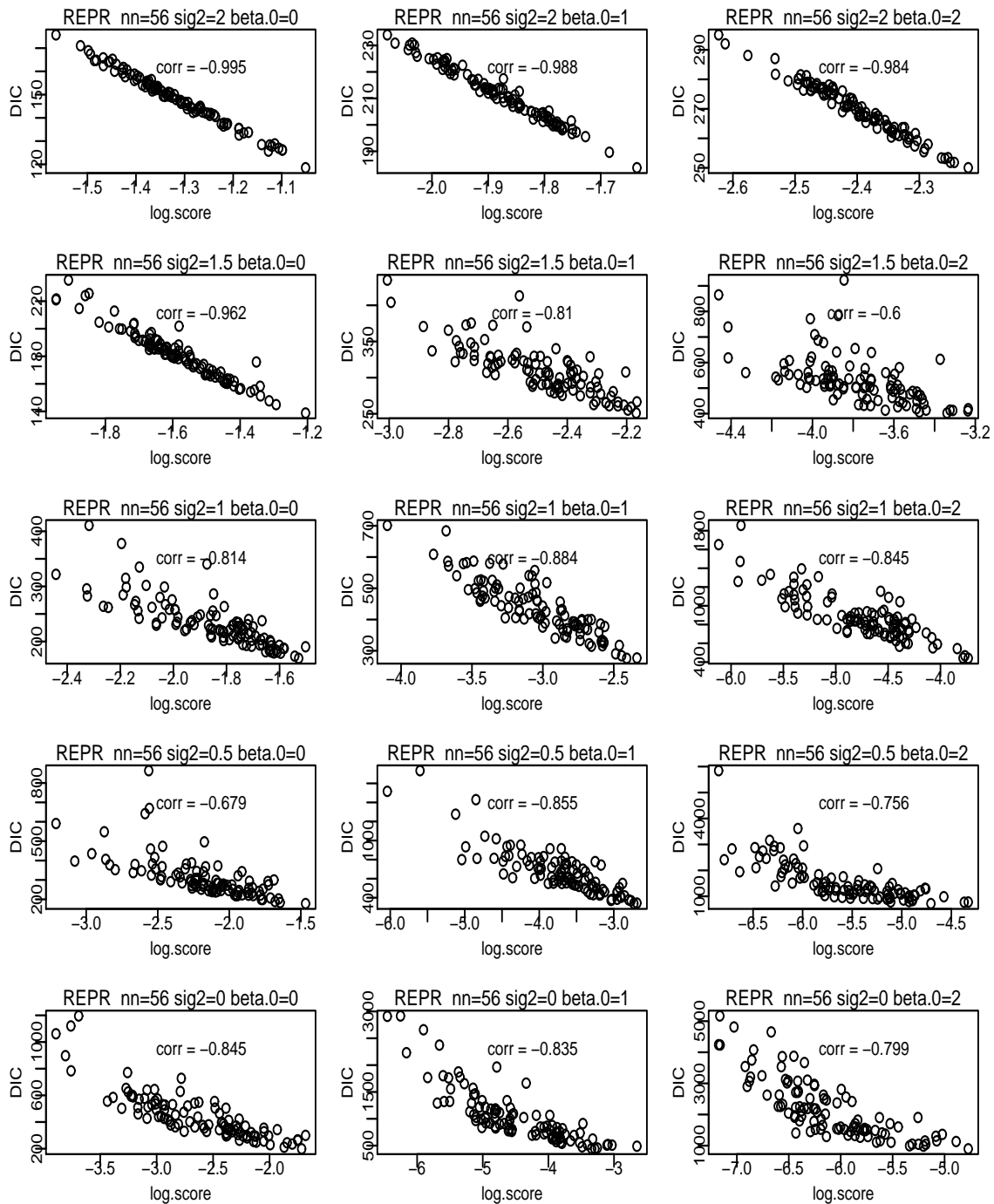


Figure 3: DIC versus $\log\text{-score}$; true model is M_2 .

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

Model	$\overline{D(\bar{\theta})}$	$D(\bar{\theta})$	\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 λ s)	1495.4	1493.4	1.98	1497.4	-1.314
4 (REPR)	1275.7	1132.0	143.2	1418.3	
	1274.7	1131.3	143.5	1418.2	-1.180
	1274.4	1130.2	144.2	1418.6	

respectively, and the FEPR model assumes that these ratios are 1):

$$\begin{aligned}
(y_i | \lambda_i) &\stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \beta_0 + \beta_1 x_i + e_i \\
(e_i | \sigma_e^2) &\stackrel{\text{iid}}{\sim} N(0, \sigma_e^2) \\
(\beta_0, \beta_1, \sigma_e^2) &\sim \text{diffuse} ,
\end{aligned} \tag{23}$$

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 $\sigma_e \rightarrow 0$ in the REPR model, the result is the FEPR model, with $p_D = 2$ parameters; as $\sigma_e \rightarrow \infty$, in effect all subjects in the study have their own λ and p_D would be 572; in between at $\sigma_e \doteq 0.675$ (the posterior mean), DIC estimates that there are about 143 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).

The conclusion we draw from the results presented so far is that, while DIC does not have a direct utility-based decision-theoretic basis, it may in some cases provide a computationally quick approximation to LS_{CV} (since DIC requires only one MCMC run rather than the $O(n)$ runs needed for direct implementations of LS_{CV}). However, it is worth emphasizing the point made by Spiegelhalter et al. (2002) 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

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

y has mean 2.35 and VTMR 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 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 $\text{Uniform}(-c, c)$ priors on $\text{logit}(\theta)$ and $\log(r)$, with c chosen large enough in each case not to truncate the likelihood function, yields $\hat{p}_D = 1.1$ —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.

4.2 Full-sample log score

Evidently, while DIC can sometimes provide an accurate and fast (indirect) approximation to LS_{CV} , it would be useful to have a fast direct approximation. An obvious thing to try is the following *full-sample* version of LS (cf. Laud and Ibrahim 1995): in the one-sample situation, for instance, compute a single predictive distribution $p^*(\cdot|y, M_i)$ for a future data value with each model M_i under consideration, based on the entire data set y (without omitting any observations), and define

$$LS_{FS}(M_i|y) = \frac{1}{n} \sum_{j=1}^n \log p^*(y_j|y, M_i). \quad (24)$$

As noted earlier, the naive approach to calculating LS_{CV} , when MCMC is needed to compute the predictive distributions, requires $O(n)$ MCMC runs, one for each omitted observation; by contrast LS_{FS} needs only a single MCMC run, making its computational speed (a) $O(n)$ times faster than naive implementations of LS_{CV} and (b) equivalent to that of DIC . Note also that the log score approach works equally well with both parametric and nonparametric Bayesian models (this remark applies to both LS_{CV} and LS_{FS}), whereas DIC is only defined for parametric models.

Recall the claim by Bernardo and Smith (1994), discussed earlier, that LS_{CV} approximates the expectation of logarithmic predictive utility:

$$E[U(M_i|y)] \approx LS_{CV} = \frac{1}{n} \sum_{j=1}^n \log p(y_j|M_i, y_{-j}) \quad (25)$$

Mukhopadhyay et al. (2005) recently proved that the difference between the left- and right-hand sides of (25) does not vanish for large n but is instead $O_p(\sqrt{n})$. (However unpleasant, this fact does not automatically invalidate the use of LS_{CV} as an approximate expected utility, since when comparing two models we effectively look at the difference between two LS_{CV} values, and the bias in using LS_{CV} as an approximation to $E[U(M_i|y)]$ should largely cancel out.) We have proved, under mild regularity conditions similar to those specified by Mukhopadhyay et al. (2005), that LS_{FS} is free from this deficiency: the difference between $LS_{FS} = \frac{1}{n} \sum_{j=1}^n \log p^*(y_j|y, M_i)$ and $E[U(M_i|y)]$ and is $o_p(1)$ (this proof will be presented elsewhere).

It is natural to wonder if this asymptotic superiority of LS_{FS} over LS_{CV} translates into better small-sample performance; this is the subject of the next section.

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}		
σ^2	β_0		σ^2	β_0	
	0	1		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

4.3 Log-score model discrimination

We now have three behavioral rules: 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 (22)) and computed LS_{CV} , LS_{FS} , and DIC for models M_1 (the fixed-effects Poisson model (21)) 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 which 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 6). The graphs in Figure 4 compare Bayesian decision-theoretic power curves for LS_{CV} , LS_{FS} , and DIC . Remarkably, not only is LS_{FS} much quicker computationally than LS_{CV} , in our simulation environment it was also more accurate with small samples of data at identifying the correct model than LS_{CV} or DIC .

To summarize our comparative findings, in computational efficiency

$$LS_{CV} < DIC \doteq LS_{FS}, \quad (26)$$

and in fixed- and random-effects Poisson modeling the results in model discrimination power are

$$LS_{CV} \doteq DIC < LS_{FS}. \quad (27)$$

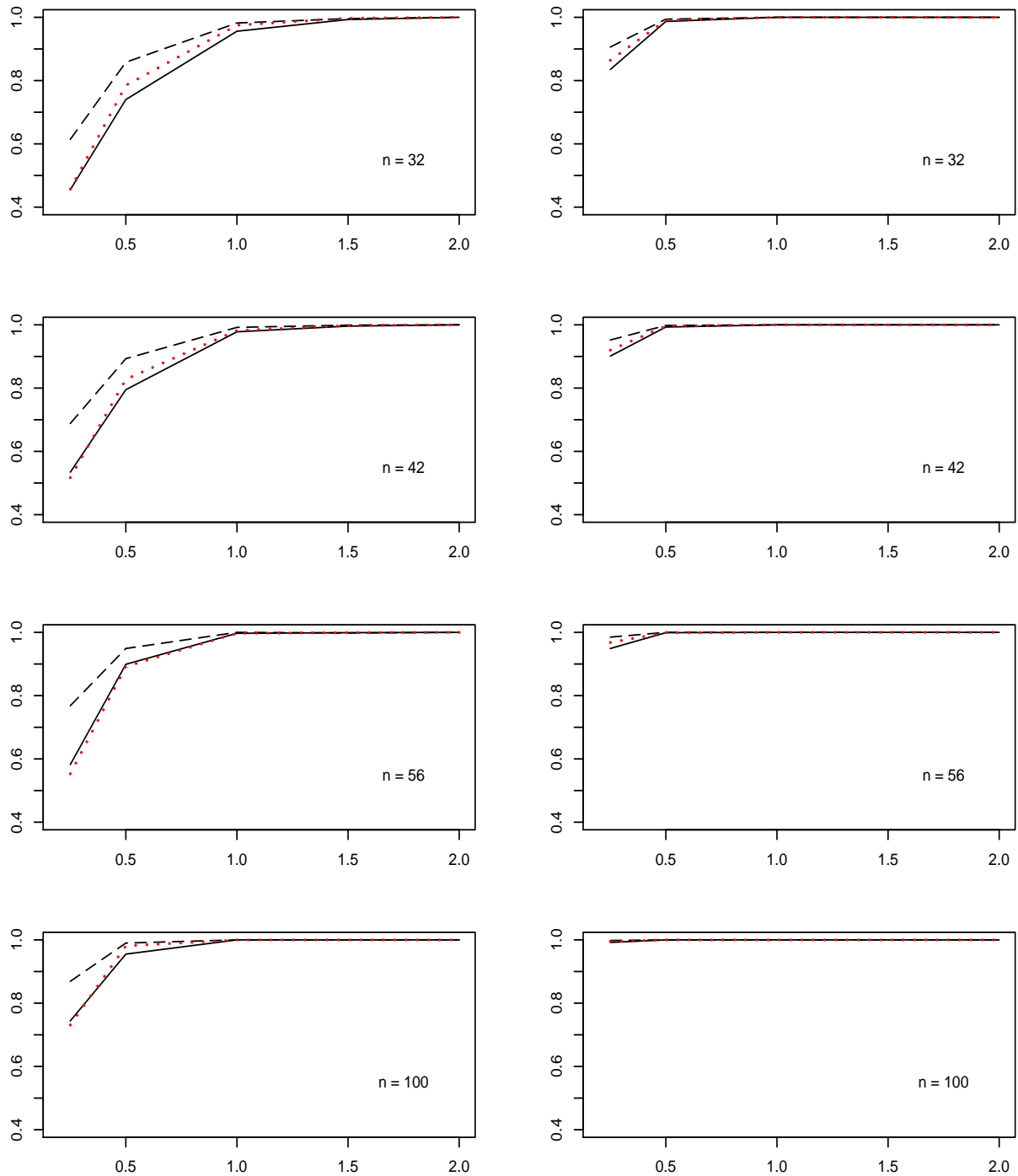


Figure 4: Bayesian decision-theoretic power curves for LS_{CV} (solid lines), LS_{FS} (long dotted lines), and DIC (short dotted lines) (column 1: $\beta_0 = 0$; column 2: $\beta_0 = 1$; rows indexed by sample size n ; horizontal scale in all plots is σ^2).

5 Connections with Bayes factors

An extensively explored alternative to log-score-based predictive model choice is *Bayes factors* (e.g., Jeffreys 1939; Berger and Sellke 1987; Berger and Pericchi 1996, 2001; Kass and Raftery 1995; Pericchi 2004; O'Hagan and Forster 2004), which implicitly perform model comparison on the probability scale:

$$\begin{aligned} \left[\frac{p(M_1|y)}{p(M_2|y)} \right] &= \left[\frac{p(M_1)}{p(M_2)} \right] \cdot \left[\frac{p(y|M_1)}{p(y|M_2)} \right] \\ \left(\begin{array}{c} \text{posterior} \\ \text{odds} \end{array} \right) &= \left(\begin{array}{c} \text{prior} \\ \text{odds} \end{array} \right) \cdot \left(\begin{array}{c} \text{Bayes} \\ \text{factor} \end{array} \right). \end{aligned} \quad (28)$$

On the surface, in fact, there appears to be a connection between the two approaches: Kass and Raftery (1995) note that

$$\begin{aligned} \log \left[\frac{p(y|M_1)}{p(y|M_2)} \right] &= \log p(y|M_1) - \log p(y|M_2) \\ &= LS^*(M_1|y) - LS^*(M_2|y), \end{aligned} \quad (29)$$

where

$$\begin{aligned} LS^*(M_i|y) &\equiv \log p(y|M_i) \\ &= \log [p(y_1|M_i) p(y_2|y_1, M_i) \cdots p(y_n|y_1, \dots, y_{n-1}, M_i)] \\ &= \log p(y_1|M) + \sum_{j=2}^n \log p(y_j|y_1, \dots, y_{j-1}, M_i). \end{aligned} \quad (30)$$

Thus the log Bayes factor equals the difference between the models in something that looks like a log score, which gives rise to the question “Isn't the rule {choose the model with the biggest LS_{CV} or LS_{FS} } equivalent to choosing M_i whenever the Bayes factor in favor of M_i exceeds 1?”

Looking more closely at (30), the answer is no: crucially, LS^* is defined via sequential prediction of y_2 from y_1 , y_3 from (y_1, y_2) , and so on, whereas LS_{CV} and LS_{FS} are based on averaging over all possible out-of-sample predictions. This distinction really matters: as is well known, with diffuse priors Bayes factors are hideously sensitive to the particular form in which the diffuseness is specified, but this defect is entirely absent from LS_{CV} and LS_{FS} .

As an example, with non-negative integer-valued data $y = (y_1, \dots, y_n)$, consider two models:

- $M_1 = \text{Geometric}(\theta_1)$ likelihood with $\text{Beta}(\alpha_1, \beta_1)$ prior on θ_1 ;
- $M_2 = \text{Poisson}(\theta_2)$ likelihood with $\text{Gamma}(\alpha_2, \beta_2)$ prior on θ_2 .

The Bayes factor in favor of M_1 over M_2 is (Bernardo and Smith 1994)

$$\frac{\Gamma(\alpha_1 + \beta_1) \Gamma(n + \alpha_1) \Gamma(s + \beta_1) \Gamma(\alpha_2) (n + \beta_2)^{s + \alpha_2} (\prod_{i=1}^n y_i!)}{\Gamma(\alpha_1) \Gamma(\beta_1) \Gamma(n + s + \alpha_1 + \beta_1) \Gamma(s + \alpha_2) \beta_2^{\alpha_2}}, \quad (31)$$

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

$$\frac{\Gamma(n+1)\Gamma(n\bar{y}+1)\Gamma(\epsilon)(n+\epsilon)^{n\bar{y}+\epsilon}(\prod_{i=1}^n y_i!)}{\Gamma(n+n\bar{y}+2)\Gamma(n\bar{y}+\epsilon)\epsilon^\epsilon}. \quad (32)$$

This 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 desired 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, e.g., the same fierce sensitivity to a prior specification with little scientific grounding appears with a Uniform(0, c) prior for θ_2 as a function of the nearly arbitrary c .) By contrast, e.g.,

$$\begin{aligned} LS_{CV}(M_1|y) &= \log \left[\frac{(\Gamma\alpha_1 + n - 1)\Gamma(\beta_1 + s)}{\Gamma(\alpha_1 + n + \beta_1 + s)} \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \log \left[\frac{\Gamma(\alpha_1 + n - 1 + \beta_1 + s_i)}{\Gamma(\beta_1 + s_i)} \right] \end{aligned} \quad (33)$$

and

$$\begin{aligned} LS_{CV}(M_2|y) &= \frac{1}{n} \sum_{i=1}^n \log \left[\frac{\Gamma(\alpha_2 + s)}{\Gamma(y_i + 1)\Gamma(\alpha_2 + s_i)} \right] \\ &\quad \cdot \left(\frac{\beta_2 + n}{\beta_2 + n + 1} \right)^{\alpha_2 + s_i} \left(\frac{1}{\beta_2 + n + 1} \right)^{y_i} \end{aligned} \quad (34)$$

where $s_i = \sum_{j \neq i} y_j$ (with similar expressions for LS_{FS}); both of these quantities are entirely stable as a function of (α_1, β_1) and (α_2, β_2) near zero.

Various attempts have been made to fix this fundamental defect of Bayes factors, e.g., {partial, intrinsic, fractional} Bayes factors, well calibrated priors, conventional priors, intrinsic priors, and expected posterior priors (e.g., Pericchi 2004); all of these methods appear to require an appeal to ad-hockery which is not required by the log score approach. Some bridges can be built between LS and BF , e.g., Mukhopadhyay et al. (2005) re-interpret LS_{CV} as the ‘‘Gelfand-Dey (1994) predictive Bayes factor’’ BF^{GD} ; connections such as these are the subject of on-going investigation. One thing that can clearly be said: despite an assertion to the contrary in O’Hagan and Forster (2004), LS_{FS} is not the same as Aitkin’s (1991) *posterior Bayes factor* in disguise. (A sketch of the proof is given as item (2) in the Appendix.)

6 When is a model good enough?

We have demonstrated that the LS_{FS} method described above 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} 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 (2) in Section 3: Is M_1 good enough?

As mentioned in Section 3, under the Model Selection Principle 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 question (2), but a somewhat related question—“Could the data have arisen from a given model?”—can be answered in a general way by simulating from that model 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 cannot be done naively, or the result will be poor calibration—indeed, 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); (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 $(*)$)

$$\begin{aligned} (\lambda) &\sim \text{diffuse} \\ (y_i|\lambda) &\stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda) \end{aligned} \tag{35}$$

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 (Robins et al. 2000) that this tail area (a p -value for a composite null hypothesis, e.g., $\text{Poisson}(\lambda)$ 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., and we have shown that it does indeed work even in rather small-sample situations, although our approach to 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, 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

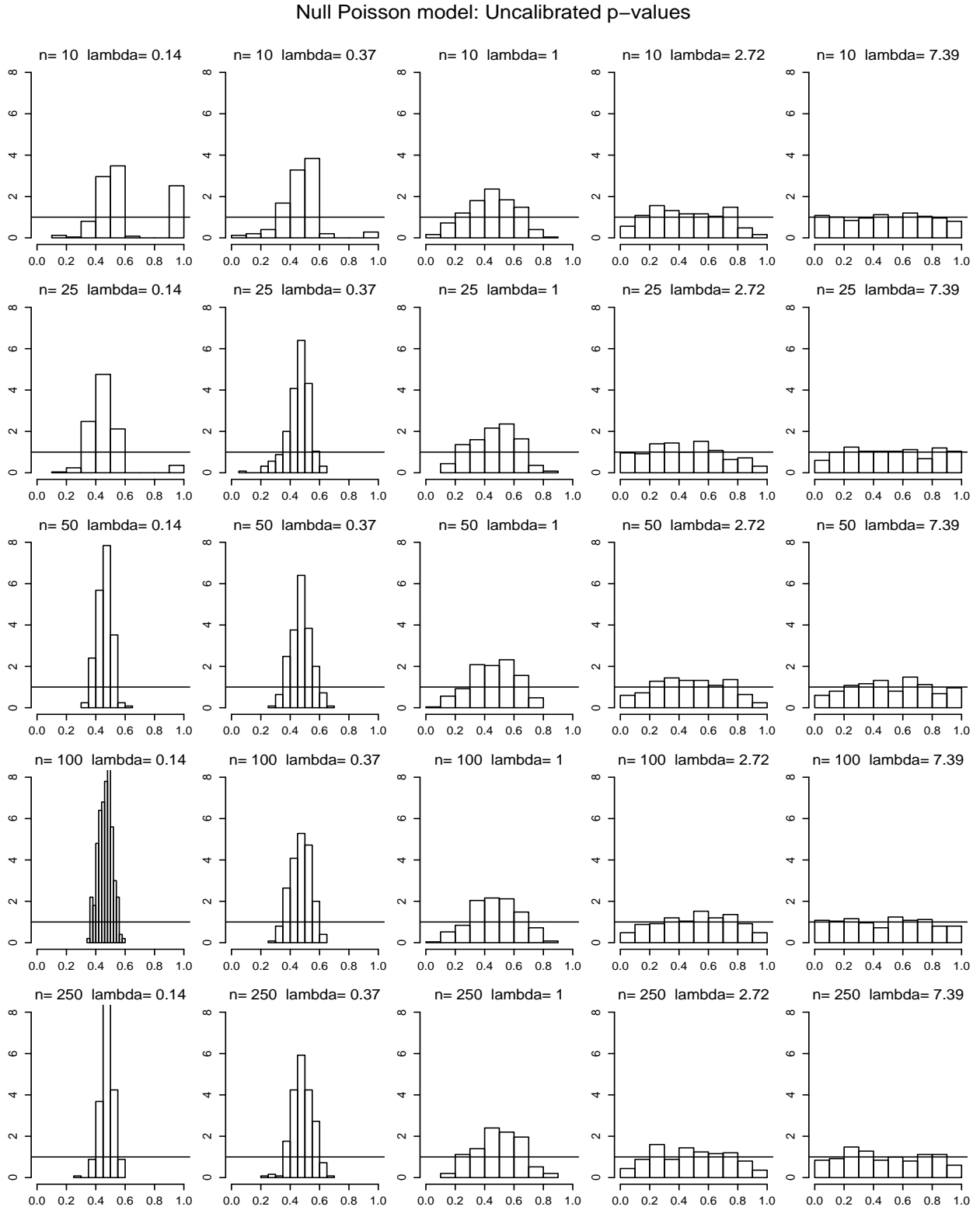
  repeat steps 1–3 on the validation data set

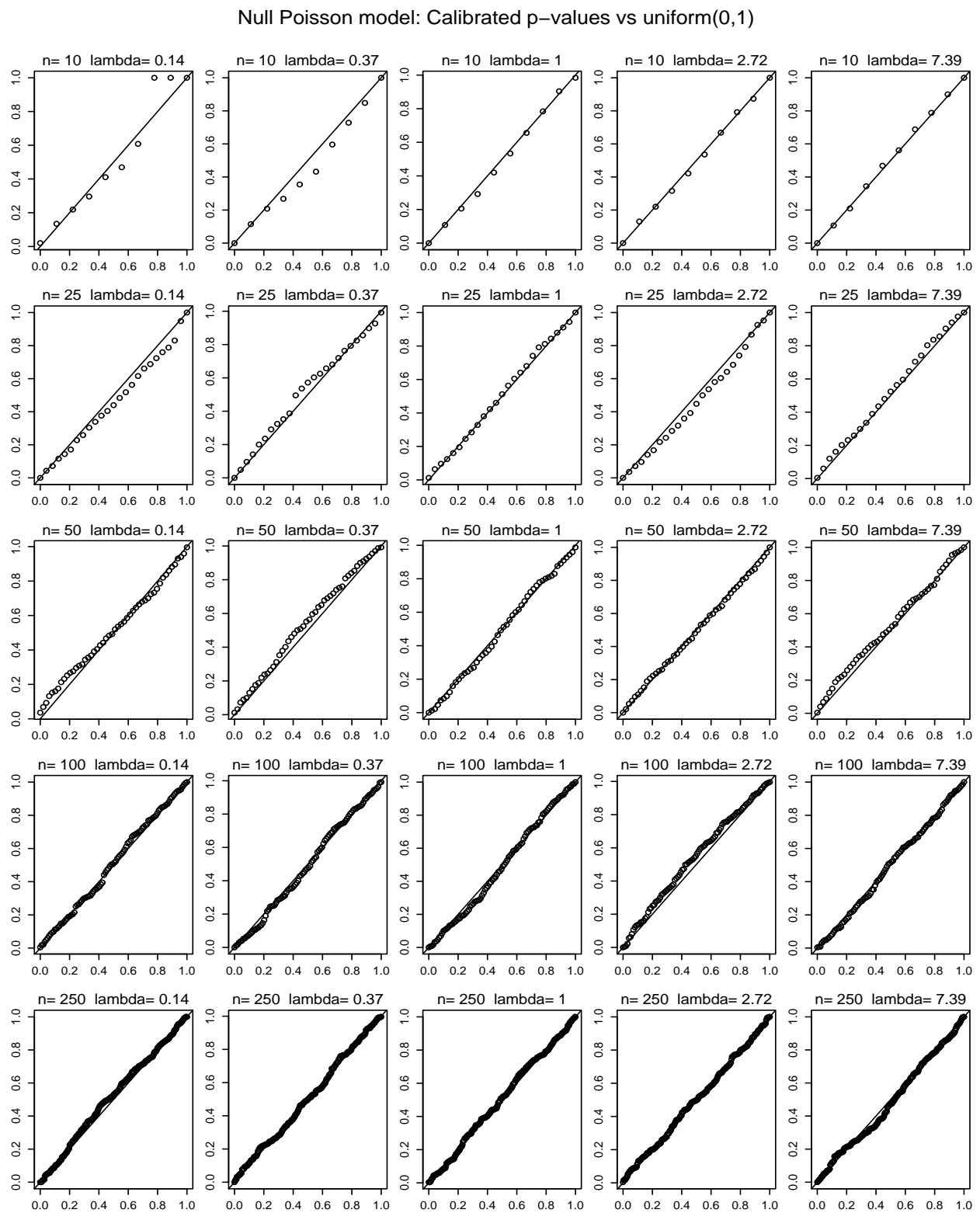
}

```

The result will be a vector of *adjusted p-values*; call this *V.Pa.* We have verified (via simulation) in several simple (and some less simple) situations that the values in *V.Pa.* are close to $U(0, 1)$ in distribution.

Figures 5–8 summarize our results and illustrate uncalibrated and calibrated *p-values* from one-sample Poisson and Gaussian models. Consider, for example, the case ($n = 100, \lambda = 0.14$) in the fourth row and first column of Figure 5: if the Gelman et al. *p-values* came out 0.35 in this situation, it would be natural to conclude that the data could very well have come from

Figure 5: *Poisson model, uncalibrated p-values.*

Figure 6: *Poisson model, calibrated p-values.*

Null Gaussian model: Uncalibrated p-values

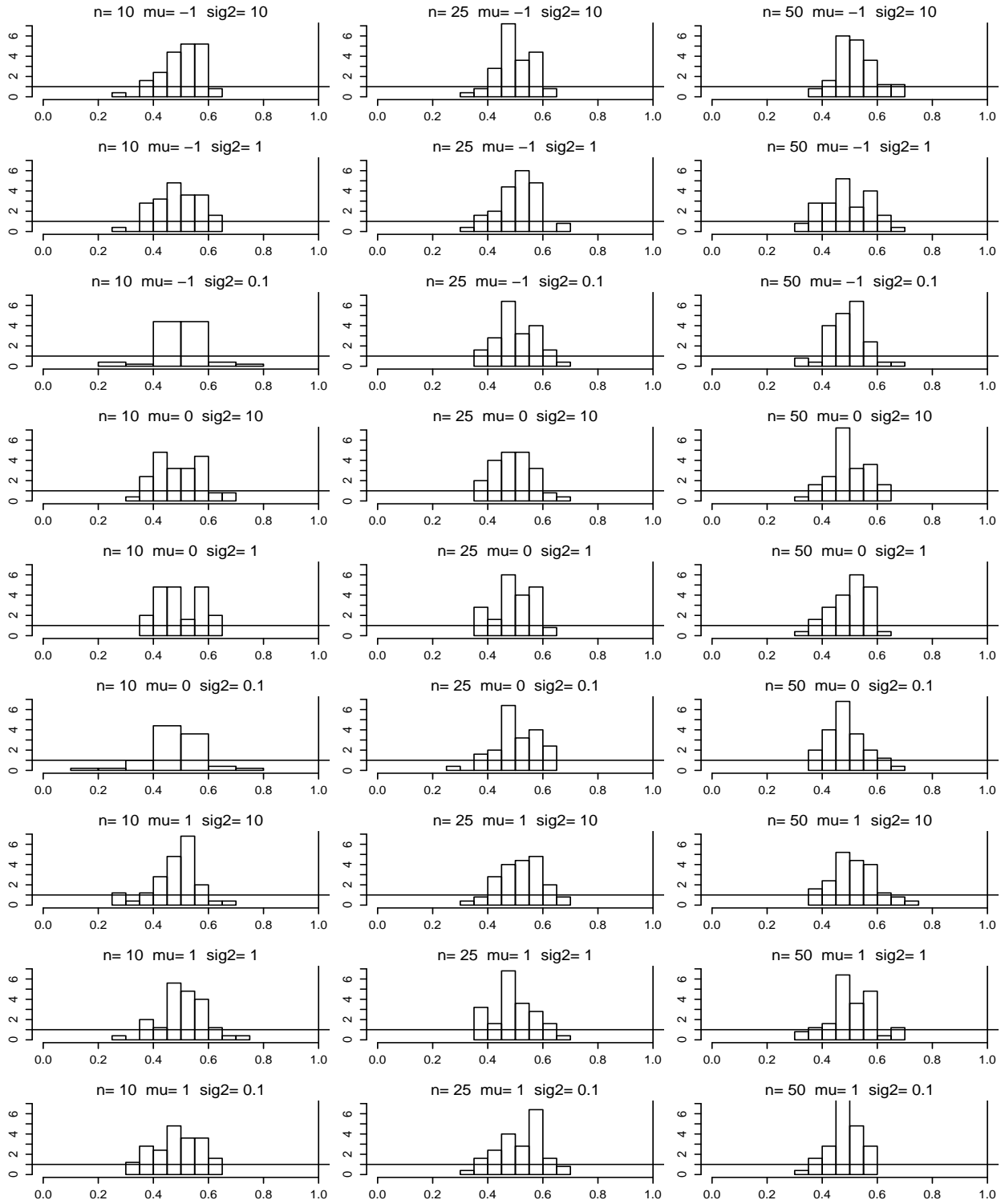


Figure 7: Gaussian model, uncalibrated p-values.

Null Gaussian model: Calibrated p-values vs uniform(0,1)

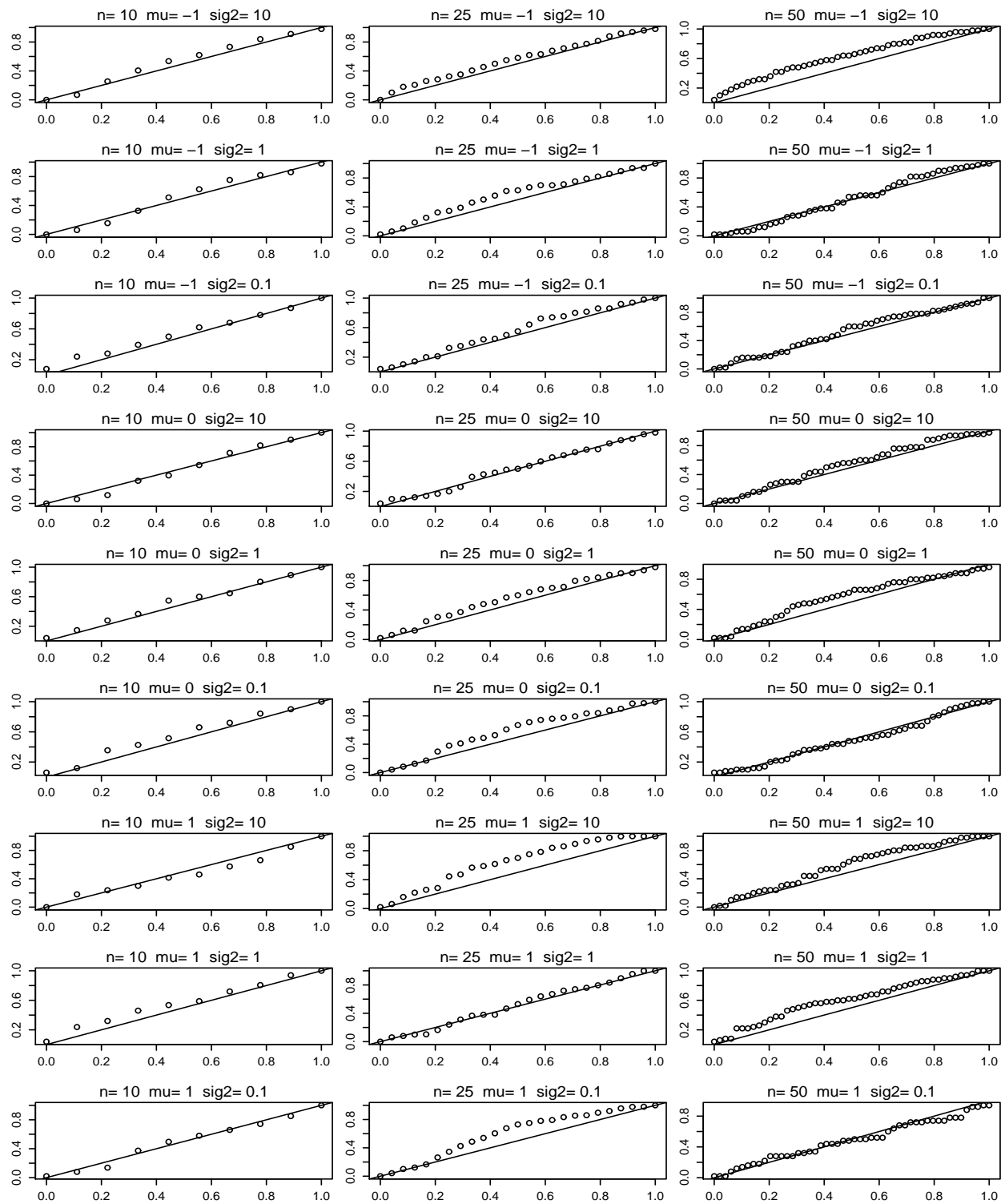


Figure 8: Gaussian model, calibrated p-values.

the Poisson model, but this part of Figure 5 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 5 that the answer is not very often (in fact, only about 0.035 of the time, i.e., the calibrated version of the uncalibrated Gelman et al. p -value is 10 times smaller). Figure 5 shows that the calibration of the Gelman et al. unadjusted approach improves, even for small n , as λ increases, but Figure 7 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 6 and 8), are nearly perfectly calibrated for all parameter values and sample sizes examined.

7 Conclusions

We draw the following conclusions from the results summarized above.

- In many problems Bayesian model specification = {exchangeability judgments plus non-parametric (BNP) modeling}.
- BNP is one way to avoid the dilemma posed by Cromwell’s Rule in Bayesian model specification; three-way cross-validation (3CV) is another. The goal in model choice is to pay attention both to coherence and to calibration.
- **Model Selection Principle (MSP):** It is not possible to choose a model well without contemplating the purpose to which it will be put; for how else will you know if the model under scrutiny is “good enough”? (Good enough for what?)
- The MSP implies that model choice is really a decision problem and should be approached via maximization of expected utility, with a utility structure that is sensitive to the real-world context.
- We believe that investigators should spend time figuring out an appropriate utility structure for model choice in each problem they tackle. For people in a hurry, when the goal is to make an accurate scientific summary of what’s known about something, the predictive log score has a sound generic utility basis and can yield stable and accurate model specification decisions.
- DIC can be thought of as a fast approximation to the leave-one-out predictive log score (LS_{CV}), but DIC can behave unstably as a function of parameterization (predictive log scores do not suffer from this defect).
- The full-sample log score (LS_{FS}) is n times faster than naive implementations of LS_{CV} , has better small-sample model discrimination power than either LS_{CV} or DIC , and has better asymptotic behavior than LS_{CV} .

- (Ordinary) Bayes factors are highly unstable when context suggests diffuse prior information; many methods for fixing this have been proposed, most of which seem to require an appeal to ad-hockery which is absent from the LS_{FS} approach.
- The basic Gelman et al. (1996) method of posterior predictive model checking is badly calibrated: when it yields a tail area of, e.g., 0.4, the calibrated equivalent may well be 0.04 or even 0.004.
- We have modified an approach suggested by Robins et al. (2000) to help answer the question “Could the data have arisen from model M ?” in a well-calibrated way.

Appendix

- (1) **Calculation of the height of a posterior predictive density via MCMC.** When parametric model M_i is fit via MCMC, the predictive ordinate $p(y_j|y_{-j}, M_i)$ in LS_{CV} or $p(y^*|y, M_i)$ in LS_{FS} is easy to approximate (cf. Gelfand and Mukhopadhyay 1995): with m identically distributed (not necessarily independent) MCMC monitoring draws θ_k from $p(\theta|y, M_i)$,

$$\begin{aligned} p(y^*|y, M_i) &= \int p(y^*|\theta, M_i)p(\theta|y, M_i)d\theta \\ &= E_{(\theta|y, M_i)} [p(y^*|\theta, M_i)] \\ &\doteq \frac{1}{m} \sum_{k=1}^m p(y^*|\theta_k, M_i). \end{aligned} \tag{36}$$

- (2) **LS_{FS} is different from the posterior Bayes factor.** Consider the likelihood part of a (parametric) model M_j : $(y_i|\theta_j, M_j) \stackrel{\text{IID}}{\sim} p(y_i|\theta_j, M_j)$ ($j = 1, 2$), with prior $p(\theta_j|M_j)$ for model M_j . The ordinary Bayes factor involves comparing quantities of the form

$$\begin{aligned} p(y|M_j) &= \int \left[\prod_{i=1}^n p(y_i|\theta_j, M_j) \right] p(\theta_j|M_j) d\theta_j, \\ &= E_{(\theta_j|M_j)} L(\theta_j|y, M_j), \end{aligned} \tag{37}$$

i.e., the Bayes factor involves comparing expectations of likelihoods with respect to the priors in the models under comparison (this is why ordinary Bayes factors behave so unstably with diffuse priors). Aitkin (1991; *posterior Bayes factors* (PBF)) suggested computing expectations instead with respect to the posteriors, i.e., PBF: favor model M_1 if $\log \bar{L}_1^A > \log \bar{L}_2^A$, where

$$\log \bar{L}_j^A = \log \int \left[\prod_{i=1}^n p(y_i|\theta_j, M_j) \right] p(\theta_j|y, M_j) d\theta_j. \tag{38}$$

This solves the problem of sensitivity to a diffuse prior but creates new problems of its own, e.g., it is incoherent. It may seem at first glance (e.g., O’Hagan and Forster (2004) asserted this) that the PBF is the same thing as LS_{FS} : favor model M_1 if

$$n LS_{FS}(M_1|y) > n LS_{FS}(M_2|y). \tag{39}$$

But this is not so:

$$nLS_{FS}(M_j|y) = \log \prod_{i=1}^n \left[\int p(y_i|\theta_j, M_j) p(\theta_j|y, M_j) d\theta_j \right], \quad (40)$$

and this is not the same because the integral and product operators in (38) and (40) do not commute.

References

- Aitkin M (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society Series B*, **53**, 111-142.
- Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.
- Berger J (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, forthcoming.
- Berger J, Pericchi L (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of American Statistical Association*, **91**, 109-122.
- Berger J, Pericchi L (2001). Objective Bayesian model methods for model selection: Introduction and comparison (with discussion). In *IMS Lecture Notes-Monograph Series*, Lahiri P (editor), **38**, 135-207.
- Berger J, Sellke T (1987). Testing a point null hypothesis: The irreconcilability of p -values and evidence, *Journal of American Statistical Association*, **82**, 112-122.
- Dey D, Mueller P, Sinha D (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer Verlag (Lecture Notes in Statistics, Volume 133).
- de Finetti B (1930). Funzione caratteristica de un fenomeno aleatorio. *Mem. Acad. Naz. Lincei*, **4**, 86-133.
- de Finetti B (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1-68.
- de Finetti B (1938). Sur la condition d'équivalence partielle. *Actualités Scientifiques et Industrielles*, **739**.
- de Finetti B (1970). *Teoria della Probabilità*, Volumes 1 and 2. Turin: Einaudi.
- 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. *Statistica Sinica*, **6**, 760-767 (discussion of "Posterior predictive assessment of model fitness via realized discrepancies," by A Gelman, X-L Meng, and H Stern).
- Draper D, Fouskakis D (2000). A case study of stochastic optimization in health policy: problem formulation and preliminary results. *Journal of Global Optimization*, **18**, 399-416.
- Draper D, Fouskakis D (2005). Stochastic optimization methods for cost-effective quality assessment in health. Submitted.

- Draper D, Hodges J, Mallows C, Pregibon D (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society Series A*, **156**, 9–37.
- Draper D, Krnjajić M (2005). Three-way cross validation (3CV) for well-calibrated Bayesian model choice. In preparation.
- Fouskakis D, Draper D (2002). Stochastic optimization: a review. *International Statistical Review*, **70**, 315–349.
- Geisser S, Eddy WF (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.
- Gelfand AE, Dey DK (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B*, **56**, 501–514.
- Gelfand AE, Dey DK, Chang H (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors), Oxford: Oxford University Press, 147–167.
- Gelfand AE, Mukhopadhyay S (1995). On nonparametric Bayesian inference for the distribution of a random sample. *Canadian Journal of Statistics*, **23**, 411–420.
- Gelman A, Meng X-L, Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733–760.
- Good IJ (1950). *Probability and the Weighing of Evidence*. London: Griffin.
- Hendriksen C, Lund E, Stromgard E (1984). Consequences of assessment and intervention among elderly people: a three year randomised controlled trial. *British Medical Journal*, **289**, 1522–1524.
- Jeffreys H (1939). *Theory of Probability*. Oxford: Oxford University Press.
- Kass RE, Raftery AE (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Key JT, Pericchi LR, Smith AFM (1998). Bayesian model choice: what and why? (with discussion). In *Bayesian Statistics 6*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (editors). Oxford University Press, 343–370.
- Krnjajić M, Draper D, Kottas A (2005). Parametric and nonparametric Bayesian model specification: a case study. In preparation.
- Laud P, Ibrahim J (1995). Predictive Model Selection, *Journal of the Royal Statistical Society B*, **57**, 247–262.
- Lindley DV (1985). *Making Decisions*, second edition. New York: Wiley.
- Mukhopadhyay N, Ghosh J, Berger J (2005). Some Bayesian predictive approaches to model selection. Manuscript.
- O'Hagan A, Forster J (2004). *Bayesian Inference*, second edition. London: Arnold.

- Pericchi L (2004). Model selection and hypothesis testing based on objective probabilities and Bayes factors. Manuscript.
- Robins JM, van der Vaart A, Ventura V (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, **95**, 1143–1156.
- Rubin DB (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Annals of Statistics*, **12**, 1151–1172.
- Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B*, **64**, 583–616.
- Walker S, Damien P, Laud P, Smith AFM (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society Series B*, **61**, 485–527.
- Walker S, Damien P, Lenk P (2004). On priors with a Kullback-Leibler property. *Journal of the American Statistical Association*, **99**, 404–408.