Published in the Proceedings of the AMSA Conference, September 2011, Novosibirsk, Russia

Bayesian model specification: some problems related to model choice and calibration

Milovan Krnjajić
National University of Ireland, Galway
David Draper
University of California at Santa Cruz
e-mail: milovan.krnjajic@nuigalway.ie

Abstract

In the development of Bayesian model specification for inference and prediction we focus on the conditional distributions $p(\theta|\mathcal{B})$ and $p(D|\theta,\mathcal{B})$, with data D and background assumptions $\mathcal{B}$, and consider *calibration* (an assessment of how often we get the right answers) as an important integral step of the model development. We compare several predictive model-choice criteria and present related calibration results. In particular, we have implemented a simulation study to compare predictive model-choice criteria LS_{CV}, a *log-score* based on cross-validation, LS_{FS}, a full-sample log score, with deviance information criterion, DIC. We show that for several classes of models DIC and LS_{CV} are (strongly) negatively correlated; that LS_{FS} has better small-sample model discrimination performance than either DIC, or LS_{CV}; we further demonstrate that when validating the model-choice results, a standard use of *posterior predictive tail-area* for hypothesis testing can be poorly calibrated and present a method for its proper calibration.

Keywords: log-score, deviance information criterion, posterior predictive tail areas, hypothesis testing.

Introduction

Bayesian approach to modeling comprises *inference, prediction* and *decision-making* and considers three main objects: θ, a model parameter vector; D, an information (data) source about θ; and $\mathcal{B}$, a set of propositions summarizing background assumptions about of θ and D, for example, that $\theta > 0$ if θ represents the mean remission time for a specified set of patients with a given disease; or that the data set arose as the result of a randomized controlled trial with the specified design). From the results of Cox (1946) and Ramsey (1926) each of these three basic Bayesian statistical activities is governed conceptually by a single equation and requires a series of specification tasks:

- (inference) $p(\theta|D,\mathcal{B}) = c\,p(\theta|\mathcal{B})\,p(D|\theta,\mathcal{B})$, where $c > 0$, and $p(\theta|D,\mathcal{B})$ *posterior distribution*, quantifies the information about θ, both internal and external to D;
- (prediction) $p(D^*|D,\mathcal{B}) = \int_\Theta p(D^*|\theta,\mathcal{B})\,p(\theta|D,\mathcal{B})\,d\theta$, where D^* is future data;
- (decision) The optimal action is given by $a^* = \text{argmax}_{a\in\mathcal{A}}\, E_{(\theta|D,\mathcal{B})}U(a,\theta)$.

In problems of realistic complexity it is uncertain how to specify $p(D|\theta,\mathcal{B})$. In our view, a leading principle governing this specification should be *calibration*, which consists of checking how often one obtains the right answers. For example, a statement such as "$p(a < \theta < b|D,\mathcal{B}) = 0.9$" should be verifiably correct about 90% of the time. To address the uncertainty in specifying a model, $p(D|\theta,\mathcal{B})$, we search for an ensemble, $\mathcal{M}$, of such specifications in a well calibrated manner, carefully avoiding a double use of data (to specify priors on model space and again to update this prior when carrying out inference and prediction). In the paper

we present calibration results related to the following basic questions in Bayesian model specification, "Q_1: Is model M_j better than M_k?" and "Q_2: Is model M_j good enough?" These questions are not complete without a clear reference to the purpose of the models. However, once the purpose is made explicit, the inferential task transforms into a decision problem, best solved by maximizing expectation of utility (MEU) specific for the model's purpose.

A standard way to answer question Q_1 is to use *Bayes factors* and related criteria, for example (for reasons of space we comment on this very briefly). A well known problem with Bayes factors is a possibly extreme sensitivity to the way diffuse priors are specified on the model parameters, (see e.g. Bernardo and Smith (1994)). The consequence of this instability is that the evidence in favor of one model over the other may be made arbitrarily large, based on a range of plausible parameter values, even regardless of the data set. Motivated in part by this well known problem, we focus on stable model-choice criteria based on the *posterior predictive* distribution (of the future data, D^*, given the observed sample, D), which has a sound basis as a utility for model comparison and is entirely stable relative to the specification of diffuse priors: $p(D^*|D, M_j, \mathcal{B}) = E_{(\eta_j|D,M_j,\mathcal{B})} p(D^*|\eta_j, M_j, \mathcal{B})$.

We argue that the quality of model prediction is also a solid basis for a useful generic utility in model comparison and hence we focus on working with posterior predictive distributions. In order to compare a predictive distribution with the actual data point, y^*, we use two *log-score* criteria, LS_{CV}, based on cross validation and defined as $n\, LS_{CV}(M_j|y, \mathcal{B}) = \sum_{i=1}^{n} \log p(y_i|y_{-i}, M_j, \mathcal{B})$, and LS_{FS}, the *full-sample log-score* defined as $n\, LS_{FS}(M_j|y, \mathcal{B}) = \sum_{i=1}^{n} \log p(y_i|y, M_j, \mathcal{B})$, and which uses all data in the sample only once (see, for example, Gelfand and Dey (1994) and Laud and Ibrahim (1995)). Considering how to address the question Q_1, we contrast deviance information criterion, DIC with the *log-score* rules, LS_{CV} and LS_{FS}.

The plan of the paper is as follows: In Sections 1 and 2 we present aspects of some answers to question Q_1 whereas Section 3 addresses a calibration issue arising from Q_2. Specifically, in Section 1, we consider how to obtain answers to Q_1, explore similarities and differences between DIC and LS_{CV} in Gaussian and Poisson models, and show results on the small-sample performance of DIC, LS_{CV} and LS_{FS} in discriminating between nested models. In Section 3 we show that the *posterior predictive tail areas* (Gelman et al. (1996)), a standard method for answering "could model M_j have generated the data?" (a question related to Q_2) can be poorly calibrated, and we document an approach to calibrating the answer.

1 LS_{CV} and DIC

In order to show the relationship between LS_{CV} and DIC as model-comparison criteria let us consider a simple parametric model, M_0, for continuous outcomes, where: $(y_i|\mu, \mathcal{B}) \overset{\text{IID}}{\sim} N(\mu, s^2)$ and $(\mu|\mathcal{B}) \sim N\big(a, b^2\big)$.

With (s^2, a, b^2) known and a diffuse prior on μ (large b^2), the posterior for μ is: $p(\mu|y, \mathcal{B}) \doteq N\big(\bar{y}, s^2/n\big)$, where $\bar{y}$ is the sample mean of $y = (y_1, \ldots, y_n)$. The predictive distribution for the next observation is then $p(y_{n+1}|y, \mathcal{B}) \doteq N\big(\bar{y}, s^2/(1 + 1/n)\big)$. Similarly, $p(y_i|y_{-i}, \mathcal{B}) \doteq N\big(\bar{y}_{-i}, s_n^2\big)$, where $\bar{y}_{-i}$ is the sample mean with observation i omitted and $s_n^2 = s^2\,(1 + 1/(n-1))$, so that $LS_{CV}(M_0|y, \mathcal{B}) \doteq c_1 - c_2 \sum_{i=1}^{n} (y_i - \bar{y}_{-i})^2$ for some constants c_1 and c_2 with $c_2 > 0$.

With a bit of algebra it can be shown that $LS_{CV}(M_0|y, \mathcal{B}) \doteq c_1 - c_2 \sum_{i=1}^{n} (y_i - \bar{y})^2$, $(c_2 > 0)$, meaning that for M_0 with a diffuse prior the LS_{CV} is almost perfectly *negatively correlated* with the sample variance.

In model M_0 the *deviance* is $D(\mu) = -2 \ln l(\mu|y, \mathcal{B}) = c_0 + c_3 \sum_{i=1}^{n} (y_i - \mu)^2$ for some $c_3 > 0$. Given a parametric model $p(y|\theta)$, Spiegelhalter et al. (2002) define the *deviance information criterion* (DIC) as: $DIC(M|y, \mathcal{B}) = D(\bar{\theta}) + 2\,\hat{p}_D$, where p_D is the effective number of model parameters, and $\bar{\theta}$ is the posterior mean of θ, so 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 with random effects), it can be estimated from standard MCMC output as $\hat{p}_D = \overline{D(\theta)} - D(\bar{\theta})$, 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.

Model M_0 has just one parameter ($p_D = 1$), a diffuse prior for which implies $\bar{\theta} \doteq \bar{y}$, so that we get $DIC(M_0|y, \mathcal{B}) \doteq c_0 + c_3 \sum_{j=1}^{n} (y_j - \bar{y})^2 + 2$ concluding that

$$-DIC(M_0|y, \mathcal{B}) \doteq c_1 + c_2 LS_{CV}(M_0|y, \mathcal{B}) \tag{1}$$

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 argument readily generalizes to any situation in which the predictive distribution is approximately Gaussian.

As a second example of the relationship between LS_{CV} and DIC we consider two models for count data a fixed-effects Poisson (FEP), model M_1 where $(y_i|\lambda, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$ and $(\lambda|\mathcal{B}) \sim p(\lambda|\mathcal{B})$, and random-effects Poisson (REP), model M_2:

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

where $i = 1, ..., n$. 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.

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. Here we summarize only a small part of the results of this simulation (see Krnjajić (2005) for additional details).

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 (graph not shown); By contrast, the Figure 1 shows that when the data-generating and assumed models were M_2 (the random-effects Poisson), LS_{CV} and DIC are less strongly negatively correlated, although the correlation increases with n (graph not shown).

2 Model-comparison criteria and small data samples

In addition to LS_{CV}, which requires n model fitting exercises, our interest was drawn to another version of the log-score idea in which no cross-validation is employed. Instead, in the one-sample situation, for instance, it suffices to compute only a single predictive distribution $p(\cdot\,|y, M_j)$ for future data, for each model M_j under consideration and based on the entire data set y. Thus, we define the *full-sample log score* $n\,LS_{FS}(M_j|y, \mathcal{B}) = \sum_{i=1}^{n} \log p(y_i|y, M_j, \mathcal{B})$ (cf. Laud and Ibrahim (1995)). *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

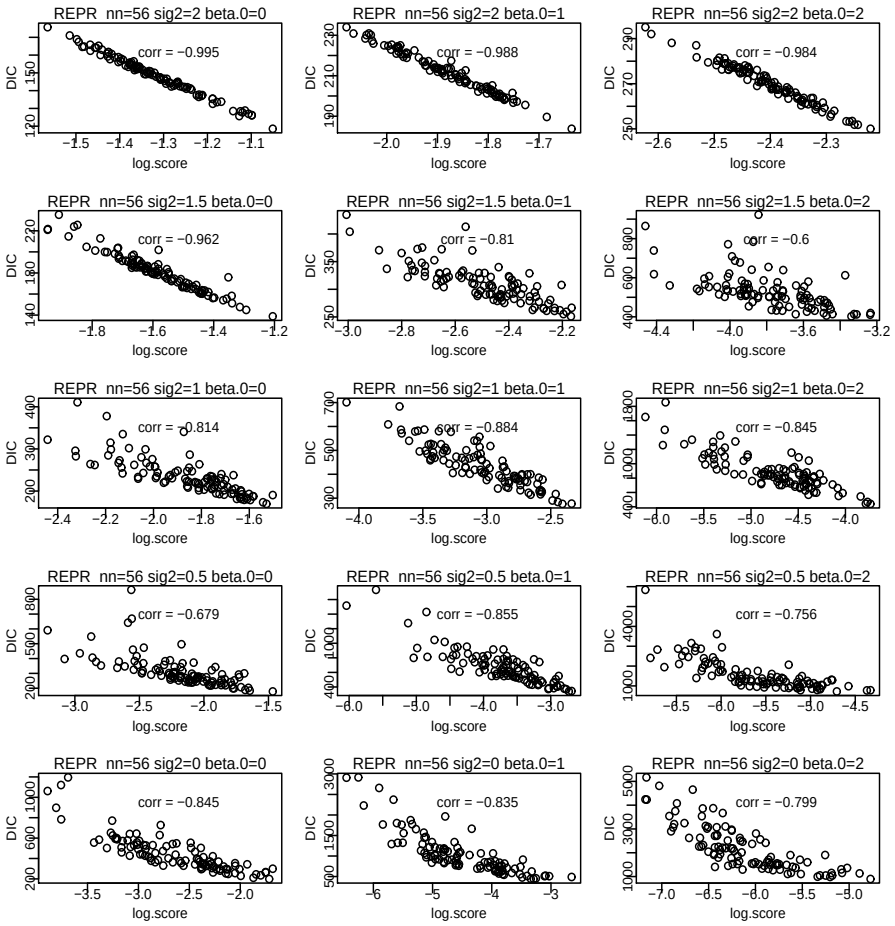

Figure 1: *DIC versus LS_{CV} with $n = 56$; the data-generating and assumed models were both M_2 (random-effects Poisson).*

small. The calculation of LS_{FS}, as opposed to Bayes factors, is entirely stable and does not have any difficulties related to the way diffuse priors may be specified.

Here we examine three model-choice rules: {maximize LS_{CV}, maximize LS_{FS}, minimize DIC}. and consider two models M_1 and M_2 to choose between. Our objective is to find out how accurately do these 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 (2)) and computed LS_{CV}, LS_{FS}, and DIC for models M_1 (the fixed-effects Poisson, FEP) and M_2 (the random-effects Poisson, REP) 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, and we monitored the percentages of correct choice for each model specification method (in this simulation M_2 is always correct).

Table 1 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

Table 1: *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 $n = 32$.*

% 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

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 begs the question "When a difference on the log score scale is big enough?" (we return to this point in Section 3). Based on model discrimination results for LS_{CV}, LS_{FS}, and DIC we created a series of performance graphs (not shown) and observed (expectedly) that increasing σ^2 makes it easier for all three methods to conclude that random effects model (the nesting model) is needed to accommodate the Poisson over-dispersion. 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.

3 Calibration of posterior predictive tail areas

Section 2 shows that full-sample log scores can stably and reliably decide between two models by choosing one with higher LS_{FS} (or LS_{CV}) value. However, this still leaves open model specification question Q_2: Is M_1 good enough?

In our view, a full judgment of adequacy requires real-world input about the purpose of the model, so it does not seem possible to propose generic methodology to answer Q_2. Instead, the somewhat related question "Q_2': Could model M_j have generated the data?" 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 kind of simulation needs to be done carefully (Draper (1996)), or the result will be poor calibration; indeed, Berger (2000) and Robins et al. (2000) have demonstrated that the procedure in (Gelman et al. (1996)) 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: a data set and a model (which may be parametric or non-parametric). For simplicity, consider a one-sample data set, D, of counts and suppose the goal is to quantify whether this data set could have come from the model $(y_i|\lambda, \mathcal{B}) \stackrel{\text{IID}}{\sim} \text{Poisson}(\lambda)$, and $(\lambda|\mathcal{B}) \sim$ diffuse (call it model $(*)$). Now, consider the following procedure:

Step 1: Calculate LS_{FS} for data set D and call it the *actual log score* (ALS). Obtain

the posterior for λ given y based on data set D; call this the *actual posterior*. **Step 2:**

```
for ( i in 1:m1 ) {
  Let lambda[ i ] be a draw from the actual posterior.
  Sample n data points from model (*) above, using lambda = lambda[ i ].
  Compute the full-sample log-score, LS.FS[ i ], for this data set.
}
```

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 similar to Gelman et al. with LS_{FS} as the *discrepancy function*. We know from our own simulations (summarized below) and the literature such as 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 ){
  Let lambda* be a draw from the actual posterior.
  Generate a data set of size n from the 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: if the sampling part of model $(*)$ really did generate the observed data, the distribution of adjusted actual tail areas would be approximately uniform, since $X \sim F_X$ implies $F_X(X) \sim U(0,1)$. This 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 model (*) above,
     using lambda = lambda.sim; call this the validation data set.
  Repeat Steps 1-3 on the validation data set.
}
```

The result here is a vector, *V.TA*, of *adjusted tail areas*. We have verified (via simulation, performed on a cluster of 100 Linux-based CPUs) in several settings that the distribution of values in V.TA is (very) close to $U(0,1)$ indeed.

Figure 2 summarizes a set of histograms of the uncalibrated actual tail areas from one-sample Poisson model, indicating that in many cases the tail areas (p-values) are far from the target (uniform) distribution.

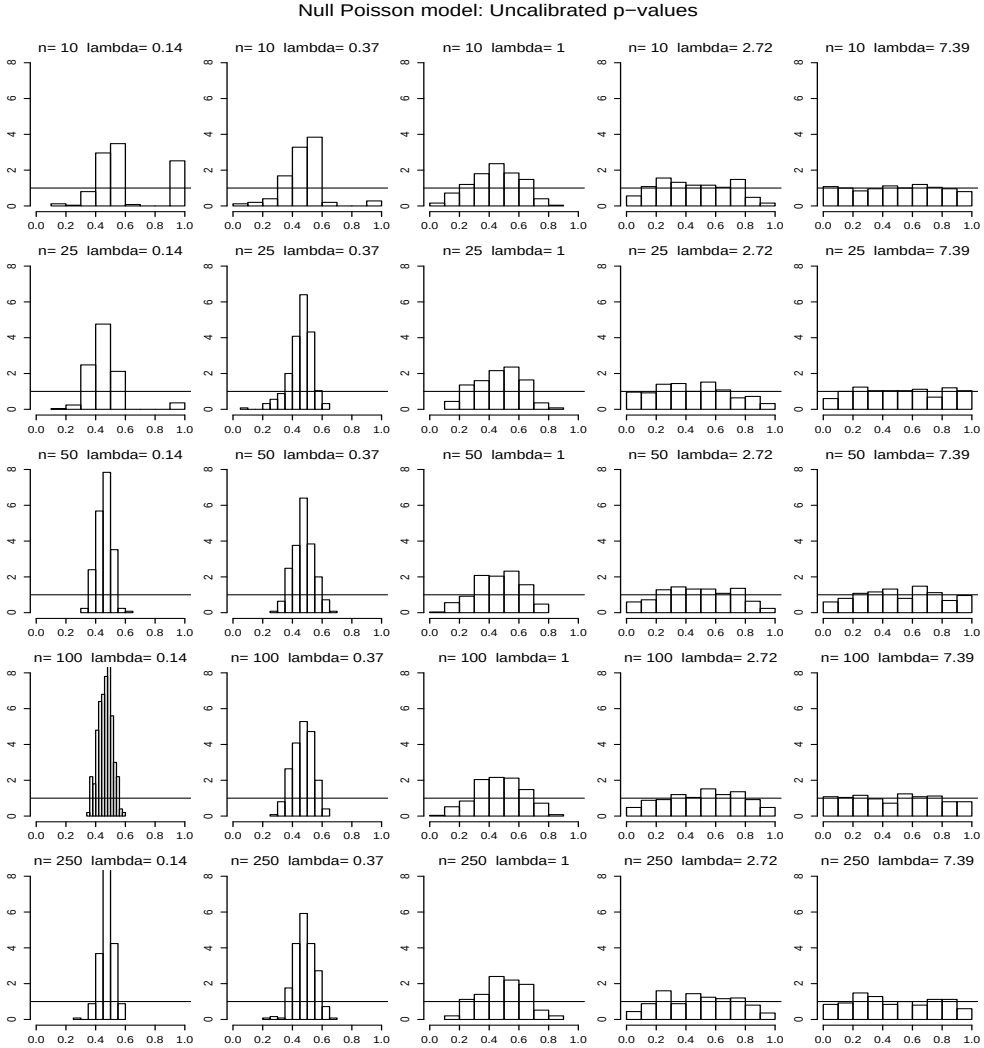

Figure 2: *Poisson model: uncalibrated tail-area values.*

Consider, for example, the case $(n = 100, \lambda = 0.14)$ in the fourth row and first column of the Figure 2: if the uncalibrated 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 2 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 one get 0.35 or less for an uncalibrated tail area in this situation?", and it is evident from Figure 2 that the answer 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 2 shows also that the calibration of the unadjusted approach improves in the one-sample Poisson setting, for increasing λ (even for small n), but in case of the Gaussian model with both μ and σ^2 unknown, the unadjusted approach remains 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 (not shown). However, the adjusted results, for the

Gaussian model, are nearly perfectly calibrated, having distributions close to $U(0,1)$ for all examind parameter values and sample sizes, (again, not shown). Note that for the reason of limited space here we could show only a small fraction of results and graphs.

Conclusions

We have argued that *calibration* (checking how often one obtains the right answer) is an important principle that that arises naturally in good Bayesian modeling; the question "Q_1: Is model M_j better than M_k?" is central to the process of well-calibrated Bayesian model specification; and it is not well formed unless the purpose of the model is considered. Once the purpose of the model is explicitly stated, the task of Bayesian model specification turns into a decision problem of maximizing expected utility (MEU), with a purpose-specific utility function (which may be computationally intensive).

LS_{FS} appears as a useful improvement upon DIC, with three advantages: LS_{FS} may well have better small-sample model discrimination behavior (as in the simulation of Section 3.1); LS_{FS} is insensitive to model parameterization; and LS_{FS} can be used both in Bayesian nonparametric and parametric settings; To decide when to stop looking for a better model, the question "Q_2': Could model M_j have generated the data?" can be answered in a well-calibrated manner, using LS_{FS} as a model choice criterion, as shown in the last section.

References

Berger, M. J. B. . J. (2000). "P-values for composite null models." *JASA*, 95: 1127–1170.

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

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

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.

Gelfand, A. E. and Dey, D. (1994). "Bayesian model choice: asymptotics and exact calculations." *JRSS (Series B)*, 56: 501–514.

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

Krnjajić, M. (2005). "Contributions to Bayesian statistical analysis: model specification and nonparametric inference." *Department of Applied Mathematics and Statistics, UCSC*.

Laud, P. and Ibrahim, J. (1995). "Predictive model selection." *JRSS/B*, 57: 247–262.

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

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). "Asymptotic distribution of P-values in composite null models." *JASA*, 95: 1143–1156.

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