Bayesian Modeling, Inference and Prediction

5: Bayesian Model Specification

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu

www.ams.ucsc.edu/∼draper

San Francisco Chapter, American Statistical Association

9 October 2010

What is a Bayesian Model?

Definition: A **Bayesian model** is a **mathematical framework** (embodying **assumptions** and **judgments**) for **quantifying uncertainty about unknown quantities** by relating them to **known quantities**.

Desirable for the **assumptions** and **judgments** in the model to arise as directly as possible from **contextual information** in the problem under study.

The most satisfying approach to **achieving this goal** appears to be that of de Finetti (1930): a **Bayesian model** is a **joint predictive distribution**

$$p(y) = p(y_1, \ldots, y_n) \tag{1}$$

for as-yet-unobserved **observables** $y = (y_1, \ldots, y_n)$.

Example 1: Data = **health outcomes** for all patients at one hospital with heart attack admission diagnosis.

Simplest possible: $y_i = 1$ if patient i **dies within 30 days of admission**, 0 otherwise.

Exchangeability

de Finetti (1930): **in absence of any other information**, my predictive uncertainty about y_i is **exchangeable**.

Representation theorem for binary data: if I'm willing to regard $(y_1, \ldots, y_n)$ as part of an **infinitely exchangeable sequence** (meaning that I judge all **finite subsets** exchangeable; this is like **thinking** of the y_i as having been **randomly sampled** from the **population** $(y_1, y_2, \ldots)$), then to be **coherent** my joint predictive distribution $p(y_1, \ldots, y_n)$ must have the simple **hierarchical** form

$$\begin{aligned} \theta &\sim p(\theta) \\ (y_i|\theta) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \end{aligned} \tag{2}$$

where $\theta = P(y_i = 1) =$ **limiting value of mean of** y_i in infinite sequence.

Mathematically $p(\theta)$ is **mixing distribution** in

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} p(y_i|\theta)\, p(\theta)\, d\theta \ . \tag{3}$$

Model = Exchangeability + Prior

Statistically, $p(\theta)$ provides opportunity to quantify **prior information** about θ and combine with information in y.

Thus, in simplest situation, **Bayesian model specification** = choice of **scientifically appropriate prior distribution** $p(\theta)$.

Example 2 (elaborating Example 1): Now I want to predict real-valued **sickness-at-admission score** instead of mortality (still no **covariates**).

Uncertainty about y_i still **exchangeable**; de Finetti's (1937) **representation theorem** for real-valued data: if $(y_1, \ldots, y_n)$ part of **infinitely exchangeable sequence**, all **coherent** joint predictive distributions $p(y_1, \ldots, y_n)$ must have hierarchical form

$$\begin{aligned} F &\sim p(F) \\ (y_i|F) &\overset{\text{IID}}{\sim} F, \end{aligned} \tag{4}$$

where F = **limiting empirical cumulative distribution function** (CDF) of infinite sequence $(y_1, y_2, \ldots)$.

Bayesian Nonparametrics

Thus here Bayesian model specification = choosing **scientifically appropriate mixing (prior) distribution** $p(F)$ for F.

However, F is **infinite-dimensional parameter**; putting probability distribution on $\mathcal{D} = \{\text{all possible CDFs}\}$ is harder.

Specifying distributions on **function spaces** is task of **Bayesian nonparametric** (BNP) modeling (e.g., Dey et al. 1998).

Example 3 (elaborating Example 2): In practice, in addition to **outcomes** y_i, **covariates** x_{ij} will typically be available.

For instance (Hendriksen et al. 1984), 572 elderly people **randomized**, 287 to **control** (C) group (standard care) and 285 to **treatment** (T) group (standard care plus **in-home geriatric assessment** (IHGA): **preventive medicine** in which each person's medical/social needs assessed, acted upon individually).

One important **outcome** was **number of hospitalizations** (in two years):
y_i^T, y_j^C = numbers of hospitalizations for **treatment** person i, **control** person j, respectively.

Conditional Exchangeability

Suppose **treatment/control** (T/C) status is **only available covariate**.

Unconditional judgment of exchangeability across all 572 outcomes **no longer automatically scientifically appropriate**.

Instead **design of experiment** compels (at least initially) judgment of **conditional exchangeability given T/C status** (e.g., de Finetti 1938, Draper et al. 1993), as in

$$\begin{array}{ccc} (F_T, F_C) & \sim & p(F_T, F_C) \\ (y_i^T | F_T, F_C) & \overset{\text{IID}}{\sim} & F_T \;\Big|\; (y_j^C | F_T, F_C) \overset{\text{IID}}{\sim} F_C \end{array} \tag{5}$$

This framework, in which (a) **covariates** specify **conditional exchangeability judgments**, (b) de Finetti's **representation theorem** reduces model specification task to placing appropriate prior distributions on CDFs, covers much of field of **statistical inference/prediction**.

Data-Analytic Model Specification

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 **same effect** as C) versus **conditional** exchangeability ($F_T \neq F_C$; T and C effects **differ**)).

Basic problem of Bayesian **model choice**: Given future observables $y = (y_1, \ldots, y_n)$, I'm **uncertain** about y (**first-order**), but I'm also **uncertain about how to specify my uncertainty** about y (**second-order**); I want to cope with **both of these kinds of uncertainty** in a **well-calibrated** manner.

Standard (**data-analytic**) approach to model specification involves initial choice, for **structure** of model, of **standard parametric family**, followed by **modification** of initial choice—once data begin to arrive—if data suggest **deficiencies** in original specification.

This approach (e.g., Draper 1995) is **incoherent** (unless I pay an **appropriate price** for **shopping around** for the model).

Cromwell's Rule

The **data-analytic** approach uses data both to specify **prior distribution on structure space** and to **update** using **data-determined prior** (result will typically be **uncalibrated** (too narrow) predictive distributions for future data).

Dilemma is example of **Cromwell's Rule** (if $p(\theta) = 0$ then $p(\theta|y) = 0$ for all y): initial model choice placed **0 prior probability** on **large regions of model space**; formally all such regions **must also have 0 posterior probability** even if data indicate **different prior on model space** would have been better.

Two possible solutions:

- **BNP** (which solves the problem by **"not putting zero probability on anything"**), and
- **3CV** (a modification of the usual **cross-validation** approach, which solves the problem by **paying an appropriate price for model exploration**).

Two Solutions: BNP and 3CV

• If use prior on F that places **non-zero probability on all Kullback-Leibler neighborhoods of all densities** (Walker et al. 2003; e.g., Pólya trees, Dirichlet process mixture priors, when chosen well), then BNP **directly avoids** Cromwell's Rule dilemma, at least for large n: as $n \to \infty$ posterior on F will **shrug off** any incorrect details of prior specification, will **fully adapt** to actual data-generating F (**NB** this assumes correct exchangeability judgments).

• **Three-way cross-validation** (3CV; Draper and Krnjajić 2007): taking usual cross-validation idea one step further,

(1) Partition data at random into *three* (non-overlapping and exhaustive) subsets S_i.

(2) Fit tentative {likelihood + prior} to S_1. **Expand** initial model in all feasible ways suggested by data exploration using S_1. **Iterate** until you're happy.

3CV (continued)

(3) Use 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 likelihood as necessary, **retune prior** as necessary, to get good calibration.
Iterate until you're happy.

(4) Announce **final model** (fit to $S_1 \cup S_2$) from (3), and report **predictive calibration** of this model on data points in S_3 as indication of how well it would perform with new data.

With **large** n probably only need to do this **once**; with **small** and **moderate** n probably best to **repeat** (1–4) several times and **combine** results in some appropriate way (e.g., **model averaging**).

How large should the S_i be? (**Preliminary answer** below.)

Model Selection as Decision Problem

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

(1) Is M_1 **better than** M_2? (this tells me **when it's OK to discard a model in my search**)

(2) Is M_1 **good enough**? (this tells me **when it's OK to stop searching**)

It would seem self-evident that **to specify a model you have to say to what purpose the model will be put**, for how else can you answer these two questions?

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

To take **two examples**,

(Case 1) If you're going to choose which of several ways to behave in future, then model has to be **good enough to reliably aid in choosing best behavior** (e.g., Fouskakis and Draper 2005); or

Choosing Utility Function

(Case 2) If you wish to make scientific summary of what's known, then—remembering that hallmark of good science is good prediction—the model has to be **good enough to make sufficiently accurate predictions of observable outcomes** (in which dimensions along which accuracy is to be monitored are driven by what's **scientifically relevant**).

How can a **utility function** driven by predictive accuracy be specified in a **reasonably general way** to answer **model specification question (1)** above? (Is M_1 **better than** M_2?)

Need **scoring rule** that measures **discrepancy** between observation y^* and predictive distribution $p(\cdot|y, M_i)$ for y^* 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

$$\boxed{\log p(y^*|y)}.$$

Log Score as Utility

On **calibration** grounds it would **seem** to be a mistake to **use data twice** in measuring this sort of thing (once to make predictions, again with same data to see how good they are; but see below).

Out-of-sample predictive validation (e.g., Geisser and Eddy 1979, Gelfand et al. 1992) addresses this apparent concern directly: e.g., successively remove each observation y_j one at a time, construct predictive distribution for y_j based on y_{-j} (data vector with y_j removed), see where y_j falls in this distribution.

This motivates **cross-validated** version of **log scoring rule** (e.g., Gelfand and Dey 1994; Bernardo and Smith 1994): with n data values y_j, when choosing among k models $M_i, i \in I$, find that model M_i which maximizes

$$LS_{CV}(M_i|y) = \frac{1}{n}\sum_{j=1}^{n} \log p(y_j|M_i, y_{-j}). \tag{6}$$

Approximating LS_{CV}

It has been argued that this can be given direct **decision-theoretic justification**: with utility function for model i

$$U(M_i|y) = \log p(y^*|M_i, y), \tag{7}$$

where y^* is **future data value**, expectation in MEU is over **uncertainty about** y^*; Gelfand et al. (1992) and Bernardo and Smith (1994) claim that this expectation can be accurately **estimated** (assuming exchangeability) by LS_{CV} (I'll revisit this claim below).

With **large data sets**, in situations in which **predictive distribution** has to be **estimated by MCMC**, direct calculation of LS_{CV} is **computationally expensive**; need **fast approximation** to it.

To see how this might be obtained, examine log score in **simplest possible model** M_0: for $i = 1, \ldots, n$,

$$\mu \sim N\left(\mu_0, \sigma_\mu^2\right), \quad (Y_i|\mu) \overset{\text{IID}}{\sim} N(\mu, \sigma^2), \quad \sigma^2 \text{ known}; \tag{8}$$

Approximating LS_{CV} (continued)

take **highly diffuse prior** on μ so that **posterior** for μ is approximately

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

where $y = (y_1, \ldots, y_n)$.

Then **predictive distribution** for next observation is approximately

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

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

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

where as before y_{-j} is y with observation j **set aside**.

But by **same reasoning**

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

Approximating LS_{CV} (continued)

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

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

for some **constants** c_1 and c_2 with $c_2 > 0$. Now it's **interesting fact** (related to behavior of **jackknife**), which you can prove by **induction**, that

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

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

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

Deviance Information Criterion (DIC)

i.e., in this model log score is **almost perfectly negatively correlated with sample variance**.

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

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

for some $c_3 > 0$, encouraging suspicion that **log score should be strongly related to deviance**.

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

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

DIC (continued)

where $\bar{\theta}$ is posterior mean of θ; they suggest that models with **low** DIC value are to be **preferred** over those with higher value.

When p_D is **difficult to read directly from 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:

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

i.e., difference between **posterior mean of deviance** and **deviance evaluated at posterior mean** of parameters (`WinBUGS` release 1.4.1 will **estimate** these quantities).

In **model** M_0, p_D is of course 1, and $\bar{\theta} = \bar{y}$, so

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

and conclusion is that

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

for $c_2 > 0$, i.e., (in this simple example) **choosing model by maximizing LS_{CV} and by minimizing DIC are approximately equivalent behaviors**.

Milovan and I have **explored the scope** of (20); in several **simple models** M we find for $c_2 > 0$ that

$$-DIC(M|y) \doteq c_1 + c_2 LS_{CV}(M|y), \tag{21}$$

i.e., across repeated data sets generated from given model, even with small n DIC and LS_{CV} can be **fairly strongly negatively correlated**.

Above argument **generalizes** to **any situation** in which **predictive distribution** is **approximately Gaussian** (e.g., **Poisson**(λ) likelihood with **large** λ, **Beta**(α, β) likelihood with **large** $(\alpha + \beta)$, etc.).

Example 3 continued. With **one-sample count data** (like number of hospitalizations in the T and C portions of IHGA data), people often choose between **fixed-** and **random-effects Poisson model formulations**: for $i = 1, \ldots, n$, and, e.g., with **diffuse priors**,

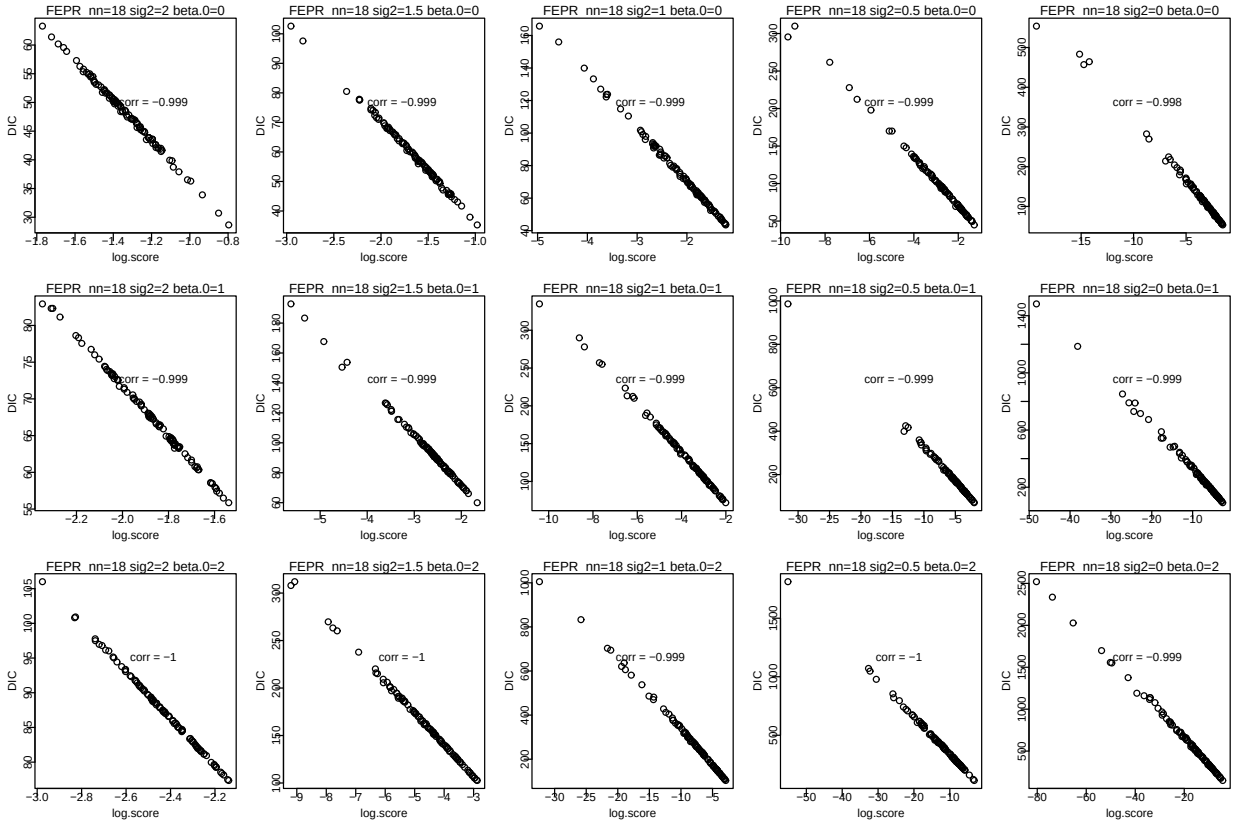
$$M_1: \left\{ \begin{array}{ccc} \lambda & \sim & p(\lambda) \\ (y_i|\lambda) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) \end{array} \right\} \quad \text{versus} \tag{22}$$

$$M_2: \left\{ \begin{array}{ccc} (\beta_0, \sigma^2) & \sim & p(\beta_0, \sigma^2) \\ (y_i|\lambda_i) & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = & \beta_0 + e_i \\ e_i & \overset{\text{IID}}{\sim} & N(0, \sigma^2) \end{array} \right\} \tag{23}$$

We conducted **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

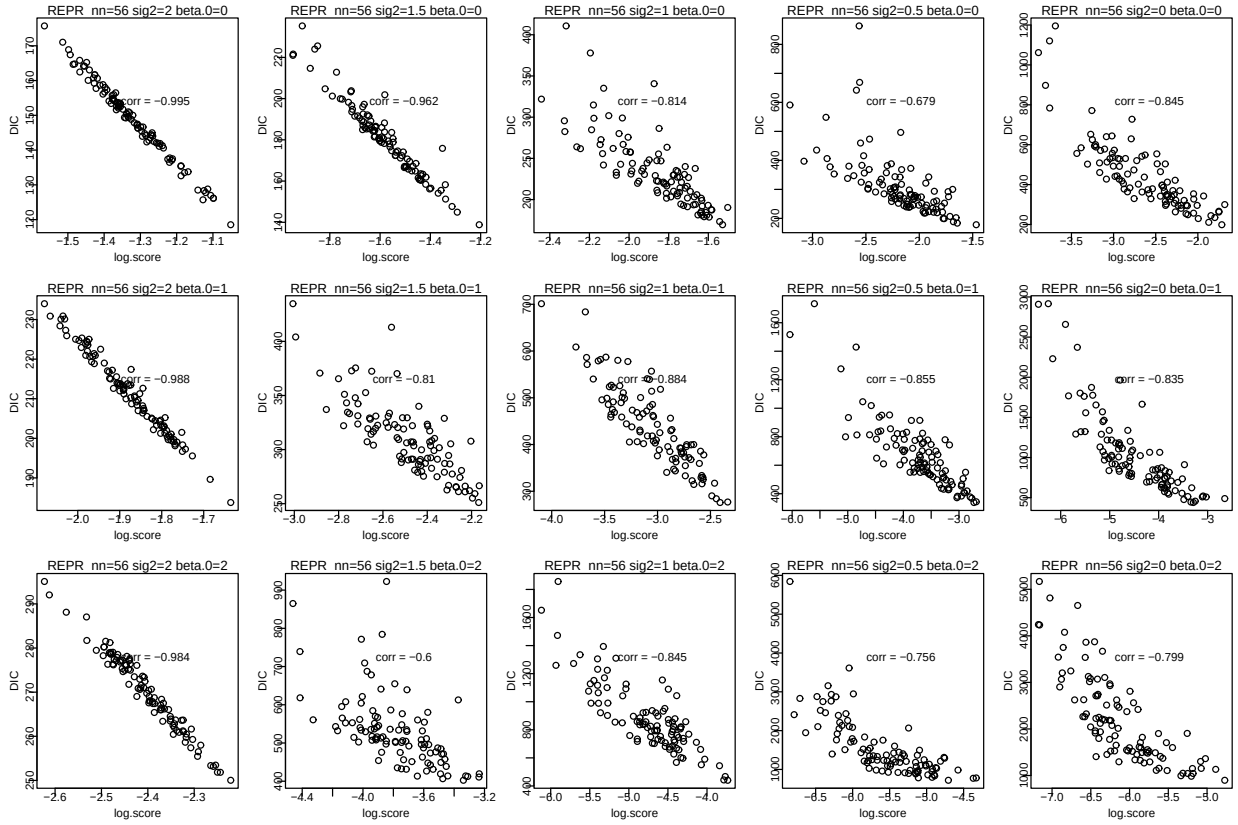
$LS_{CV} \leftrightarrow DIC$? (continued)

(**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**.



When **assumed model** is M_1 (**fixed-effects** Poisson), LS_{CV} and DIC are **almost perfectly negatively correlated**.

$LS_{CV} \leftrightarrow DIC$? (continued)



When **assumed model** is M_2 (**random-effects** Poisson), LS_{CV} and DIC are **less strongly negatively correlated** (DIC can **misbehave** with **mixture models**; see below), but **correlation increases with** n.

Example 3

As example of **correspondence between** LS_{CV} **and** DIC in real problem, IHGA data were as follows:

Distribution of number of hospitalizations in IHGA study over 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

Evidently IHGA **lowered mean hospitalization rate** (for these elderly Danish people, at least) by $(0.944 - 0.768) = \mathbf{0.176}$, which is about $100\left(\frac{0.768-0.944}{0.944}\right) = \mathbf{19\%}$ reduction from control level, a difference that's **large in clinical terms**.

Four **possible models** for these data (not all of them good):

- **Two-independent-sample Gaussian** (diffuse priors);
- **One-sample Poisson** (diffuse prior), pretending treatment and control λs are equal;

Example 3 (continued)

- **Two-independent-sample Poisson** (diffuse priors), which is equivalent to **fixed-effects Poisson regression** (FEPR); and
- **Random-effects Poisson regression** (REPR), because C and T **variance-to-mean ratios** (VTMRs) are 1.63 and 1.32, respectively:

$$\begin{aligned} (y_i \,|\lambda_i\,) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_i + e_i \\ e_i &\overset{\text{IID}}{\sim} N\big(0, \sigma_e^2\big) \\ \big(\beta_0, \beta_1, \sigma_e^2\big) &\sim \text{diffuse ,} \end{aligned} \tag{24}$$

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

DIC and LS_{CV} **results** on these four models:

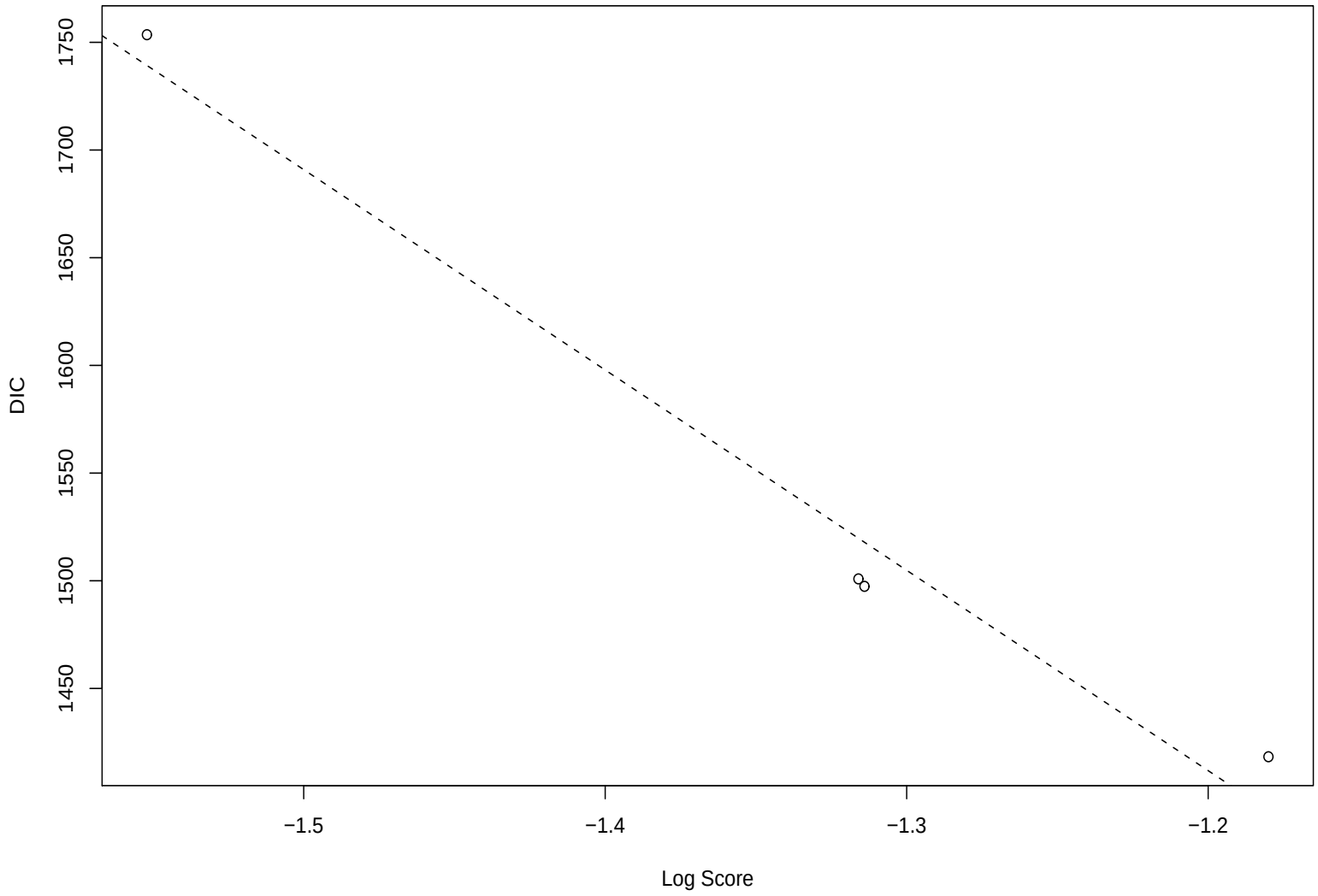
Example 3 (continued)

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	1132.0	143.2	1418.3	−1.180
	1274.7	1131.3	143.5	1418.2	
	1274.4	1130.2	144.2	1418.6	

(3 REPR rows were based on **different monitoring runs**, all of length 10,000, to give idea of Monte Carlo noise level.)

As $\sigma_e \to 0$ in **REPR** model, you get **FEPR** model, with $p_D = 2$ parameters; as $\sigma_e \to \infty$, in effect **all subjects in study have their own** λ and p_D would be 572; in between at $\sigma_e \doteq 0.675$ (posterior mean), `WinBUGS` estimates that there are about **143 effective parameters in REPR model**, but its deviance $D(\bar{\theta})$ is so much lower that it **wins** DIC **contest** hands down.

Example 3 (continued)



Correlation between LS_{CV} and DIC across these four models is **–0.98**.

But DIC Can Misbehave

$y = (0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6)$ is a data set generated from the **negative binomial** distribution with parameters $(p, r) = (0.82, 10.8)$ (in `WinBUGS` notation); y has `mean 2.35` and `VTMR 1.22`.

Using **standard diffuse priors** for p and r as in the `BUGS` examples manuals, the **effective number of parameters** p_D for the negative binomial model (which **fits the data quite well**) is estimated at **–66.2**.

The basic problem here 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** helps but can still lead to **poor estimates** of p_D.

It's evident that DIC can sometimes provide an accurate and **fast (indirect) approximation** to LS_{CV}; what about a **fast** **direct** **approximation**?

Fast (Direct) Approximation to LS_{CV}

An obvious thing to try is the following **full-sample** version of LS: 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 (cf. Laud and Ibrahim 1995)

$$LS_{FS}(M_i|y) = \frac{1}{n}\sum_{j=1}^{n} \log p^*(y_j|y, M_i). \tag{25}$$

The **naive** approach to calculating LS_{CV}, when MCMC is needed to compute the predictive distributions, requires n MCMC runs, **one for each omitted observation**; by contrast LS_{FS} needs only a **single** MCMC run, making its computational speed (a) n **times faster** than naive implementations of LS_{CV} and (b) **equivalent** to that of DIC.

- The **log score approach** works equally well with **parametric** and **nonparametric** Bayesian models; DIC is **only defined for parametric models**.

Asymptotic Properties of LS_{FS}

• When **parametric** model M_i is fit via **MCMC** the **predictive ordinate** $p(y^*|y, M_i)$ in LS_{FS} is easy to approximate: with m identically distributed (not necessarily independent) MCMC **monitoring** draws θ_k from $p(\theta|y, M_i)$,

$$\begin{aligned} p^*(y^*|y, M_i) &= \int p(y^*|\theta, M_i)\, p(\theta|y, M_i) d\theta \\ &= E_{(\theta|y,M_i)}\left[p(y^*|\theta, M_i)\right] \\ &\doteq \frac{1}{m}\sum_{k=1}^{m} p(y^*|\theta_k, M_i). \end{aligned} \tag{26}$$

Recall the claim that LS_{CV} **approximates expectation of logarithmic utility**:

$$E\left[U(M_i|y)\right] \approx LS_{CV} = \frac{1}{n}\sum_{j=1}^{n} \log p(y_j|M_i, y_{-j}) \tag{27}$$

Berger et al. (2005) recently proved that **difference** between LHS and RHS of (27) **does not vanish** for large n but is instead $O_p(\sqrt{n})$.

Asymptotic Properties of LS_{FS} (continued)

(However **unpleasant**, this fact does not automatically invalidate use of LS_{CV} as estimated expected utility, since when comparing two models we effectively look at the **difference** between two LS_{CV} values, and the discrepancy should largely **cancel out**.)

We have proved in the same setting as Berger et al. (2005) that LS_{FS} is **free from this deficiency**: the difference between $E[U(M_i|y)]$ and $LS_{FS} = \frac{1}{n}\sum_{j=1}^{n} \log p^*(y_j|y, M_i)$ is $O_p(1)$.

Q: Does this **asymptotic superiority** of LS_{FS} over LS_{CV} translate into **better small-sample performance**?

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?

LS_{CV}, LS_{FS} and DIC Model Discrimination

Example: Recall that in **earlier simulation study**, for $i = 1, \ldots, n$, and with **diffuse priors**, we considered

$$M_1: \left\{ \begin{array}{ccc} \lambda & \sim & p(\lambda) \\ (y_i|\lambda) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) \end{array} \right\} \quad \text{versus}$$

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

As **extension** of previous simulation study, we generated data from M_2 and computed LS_{CV}, LS_{FS}, and DIC for models M_1 and M_2 in **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 **100** simulation replications in each cell, and monitored **percentages of correct model choice** (here M_2 is always correct).

Model Discrimination (continued)

Examples of **results** for (e.g.) LS_{CV}:

$n = 32$

% Correct Decision

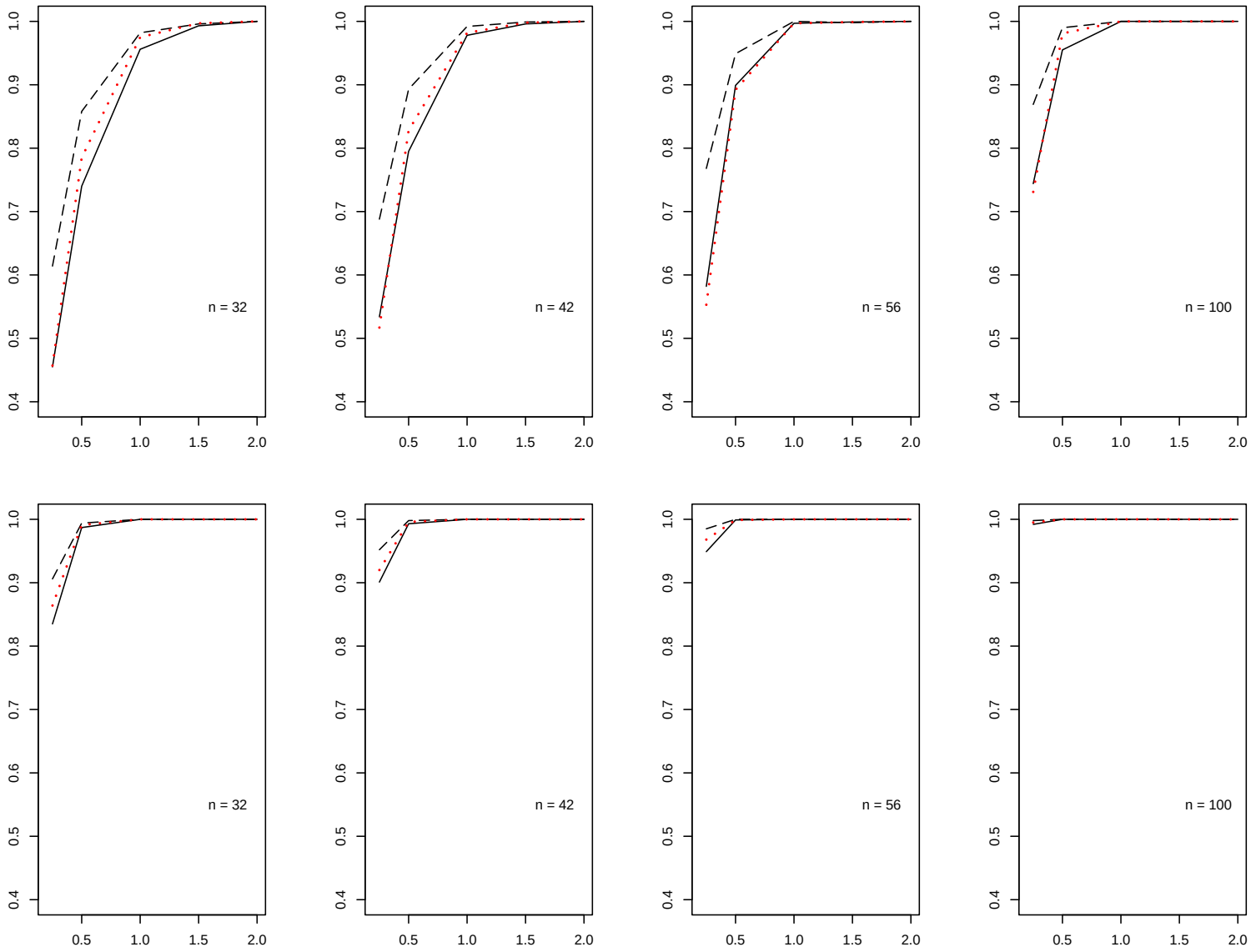
	β_0	
σ^2	0	1
0.10	31	47
0.25	49	85
0.50	76	95
1.00	97	100
1.50	98	100
2.00	100	100

Mean Absolute Difference in LS_{CV}

	β_0	
σ^2	0	1
0.10	0.001	0.002
0.25	0.002	0.013
0.50	0.017	0.221
1.00	0.237	4.07
1.50	1.44	17.4
2.00	12.8	63.9

Even with n 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 $\beta_0 = 0$.

Model Discrimination (continued)



LS_{CV} (**solid** lines), LS_{FS} (**long dotted** lines), and DIC (**short dotted** lines).

Bayes Factors

Remarkably, not only is LS_{FS} **much quicker computationally** than LS_{CV}, it's also **more accurate** with small and moderate sample sizes at **identifying the correct model** than LS_{CV} or DIC.

To summarize, in **computational efficiency**

$$LS_{CV} < DIC \doteq LS_{FS} \tag{28}$$

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

$$LS_{CV} \doteq DIC < LS_{FS} \tag{29}$$

Much has been written about use of **Bayes factors** for **model choice** (e.g., Jeffreys 1939, Kass and Raftery 1995; excellent recent book by O'Hagan and Forster (2004) devotes almost **40 pages** to this topic).

Why not use **probability scale** to choose between M_1 and M_2?

Bayes Factors (continued)

$$\left[\frac{p(M_1|y)}{p(M_2|y)}\right] = \left[\frac{p(M_1)}{p(M_2)}\right] \cdot \left[\frac{p(y|M_1)}{p(y|M_2)}\right] \tag{30}$$

$$\begin{pmatrix} \text{posterior} \\ \text{odds} \end{pmatrix} = \begin{pmatrix} \text{prior} \\ \text{odds} \end{pmatrix} \cdot \begin{pmatrix} \text{Bayes} \\ \text{factor} \end{pmatrix}$$

Kass and Raftery (1995) **note** that

$$\begin{aligned} \log\left[\frac{p(y|M_1)}{p(y|M_2)}\right] &= \log p(y|M_1) - \log p(y|M_2) \\ &= LS^*(M_1|y) - LS^*(M_2|y), \end{aligned} \tag{31}$$

where

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

Bayes Factors (continued)

Thus **log Bayes factor** equals **difference** between models in **something that looks like a log score**, i.e., aren't LS_{CV} and LS_{FS} equivalent to choosing M_i whenever the Bayes factor in favor of M_i **exceeds 1**?

No; crucially, LS^* is defined via **sequential** prediction of y_2 from y_1, y_3 from (y_1, y_2), etc., 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 particular **form** in which diffuseness is **specified**, but this defect is **entirely absent** from LS_{CV} and LS_{FS} (and from other properly-defined **utility-based model choice criteria**).

Example: Integer-valued data $y = (y_1, \ldots, y_n)$;

$M_1 =$ **Geometric**(θ_1) likelihood with **Beta**(α_1, β_1) prior on θ_1;

$M_2 =$ **Poisson**(θ_2) likelihood with **Gamma**(α_2, β_2) prior on θ_2.

Bayes Factors (continued)

Bayes factor in favor of M_1 over M_2 is

$$\frac{\Gamma(\alpha_1+\beta_1)\Gamma(n+\alpha_1)\Gamma(n\bar{y}+\beta_1)\Gamma(\alpha_2)(n+\beta_2)^{n\bar{y}+\alpha_2}\left(\prod_{i=1}^n y_i!\right)}{\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(n+n\bar{y}+\alpha_1+\beta_1)\Gamma(n\bar{y}+\alpha_2)\beta_2^{\alpha_2}}.$$

Diffuse priors: take $(\alpha_1,\beta_1)=(1,1)$ and $(\alpha_2,\beta_2)=(\epsilon,\epsilon)$ for some $\epsilon>0$.

Bayes factor reduces to

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

This goes to $+\infty$ as $\epsilon \downarrow 0$, i.e., you can make the evidence in **favor** of the **Geometric model** over the **Poisson** as **large** as you want, **no matter what the data says**, as a function of a quantity near 0 that **scientifically** you have **no basis** to specify.

Bayes Factors (continued)

By **contrast**, e.g.,

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

and

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

(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.

What LS_{FS} Is Not

(Various **attempts** have been made to **fix** this defect of Bayes factors, e.g., {partial, intrinsic, fractional} Bayes factors, well calibrated priors, conventional priors, intrinsic priors, 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., Berger et al. (2005) re-interpret LS_{CV} as the "Gelfand-Dey (1994) **predictive Bayes factor**" BF^{GD}; connections like these are the subject of **ongoing investigation**.)

---

(1) **Likelihood** part of (parametric) model

M_j: $(y_i|\theta_j, M_j) \overset{\text{IID}}{\sim} p(y_i|\theta_j, M_j)(j = 1, 2)$, with **prior** $p(\theta_j|M_j)$ for model M_j.

Ordinary Bayes factor involves comparing quantities of the form

$$\begin{aligned} p(y|M_j) &= \int \left[\prod_{i=1}^{n} p(y_i|\theta_j, M_j)\right] p(\theta_j|M_j)\, d\theta_j, \\ &= E_{(\theta_j|M_j)} L(\theta_j|y, M_j), \end{aligned} \tag{32}$$

What LS_{FS} Is Not (continued)

i.e., 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 badly with diffuse priors**).

Aitkin (1991; **posterior Bayes factors**): compute 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

$$\log \bar{L}_j^A = \log \int \left[\prod_{i=1}^{n} p(y_i|\theta_j, M_j) \right] p(\theta_j|y, M_j)\, d\theta_j. \tag{33}$$

This **solves** the problem of sensitivity to a diffuse prior but **creates new problems of its own**, e.g., it's **incoherent**.

It may **seem** at first glance (e.g., O'Hagan and Forster (2004)) that **PBF is the same thing as** LS_{FS}: favor model M_1 if

$$n\, LS_{FS}(M_1|y) > n\, LS_{FS}(M_2|y). \tag{34}$$

What LS_{FS} Is Not (continued)

But **not so**:

$$nLS_{FS}(M_j|y) = \log \prod_{i=1}^{n} \left[\int p(y_i|\theta_j, M_j)\, p(\theta_j|y, M_j)\, d\theta_j \right], \tag{35}$$

and this is **not the same** because the **integral** and **product** operators **do not commute**.

Also, some people like to compare models based on the **posterior expectation of the log likelihood** (this is **one of the ingredients** in DIC), and this is **not the same** as LS_{FS} either: by **Jensen's inequality**

$$\begin{aligned} nLS_{FS}(M_j|y) &= \sum_{i=1}^{n} \log p(y_i|y, M_j) \\ &= \sum_{i=1}^{n} \log \int p(y_i|\theta_j, M_j)\, p(\theta_j|y, M_j)\, d\theta_j \\ &= \sum_{i=1}^{n} \log E_{(\theta_j|y,M_j)} L(\theta_j|y_i, M_j) \end{aligned}$$

When is a Model Good Enough?

$$
\begin{aligned}
&> \sum_{i=1}^{n} E_{(\theta_j|y,M_j)} \log L(\theta_j|y_i, M_j) \\
&= E_{(\theta_j|y,M_j)} \sum_{i=1}^{n} \log L(\theta_j|y_i, M_j) \qquad (36) \\
&= E_{(\theta_j|y,M_j)} \log \prod_{i=1}^{n} L(\theta_j|y_i, M_j) \\
&= E_{(\theta_j|y,M_j)} \log L(\theta_j|y, M_j).
\end{aligned}
$$

LS_{FS} **method** described here (**not** LS* method) can **stably** and **reliably** help in choosing between M_1 and M_2; but suppose 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, i.e., what about model specification question (2): Is M_1 **good enough**?

Calibrating LS_{FS} Scale

As mentioned above, a **full judgment of adequacy** requires **real-world input** (to what purpose will the model be put?), but you can answer a somewhat related question—**could the data have arisen from a given model**?—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 (Draper and Krnjajić 2007).

This is related to the **posterior predictive model-checking** method of Gelman, Meng and Stern (1996); however, this sort of thing cannot be done **naively**, or result will be **poor calibration**—indeed, Robins et al. (2000) demonstrated that the Gelman et al. procedure may be (sharply) **conservative**.

Using **modification** of idea in Robins et al., we have developed method for **accurately calibrating the log score scale**.

Inputs to our procedure: (1) A **data set** (e.g., with regression structure); (2) A **model** (can be parametric, non-parametric, or semi-parametric).

Calibrating LS_{FS} Scale (continued)

Simple example: data set $y = (1, 2, 2, 3, 3, 3, 4, 6, 7, 11)$, $n = 10$.

Given **model** $(*)$

$$\begin{aligned} (\lambda) &\sim \text{diffuse} \\ (y_i|\lambda) &\overset{\text{IID}}{\sim} \text{Poisson}(\lambda) \end{aligned} \tag{37}$$

Step 1:

Calculate LS_{FS} for this data set; say get $LS_{FS} = -1.1$; call this **actual log score** (ALS).

Obtain posterior for λ given y based on this data set; call this **actual posterior**.

Calibrating LS_{FS} Scale (continued)

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 ]

}
```

Output of this loop is a vector of log scores; call this **V.LS**.

Calibrating LS_{FS} Scale (continued)

Locate ALS in distribution of LS_{FS} values by computing percentage of LS_{FS} values in V.LS that are $\leq$ ALS; call this percentage **unadjusted actual tail area** (say this is 0.22).

So far this is just Gelman et al. with LS_{FS} as the **discrepancy function**.

We know from our own simulations and the literature (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**.

We've **modified** and implemented one of the ways suggested by Robins et al., and we've shown that it does indeed work even in rather small-sample situations, although our approach to implementing the basic idea can be **computationally intensive**.

Calibrating LS_{FS} Scale (continued)

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**.

Calibrating LS_{FS} Scale (continued)

Compute the percentage of tail areas in V.P that are $\leq$ the unadjusted actual tail area; this is the **adjusted actual tail area**.

Step 3 in this procedure **solves the calibration problem** by applying the old idea that if $X \sim F_X$ then $F_X(X) \sim U(0,1)$.

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.

This claim can be verified by building a **big loop** around steps 1–3 as follows:

Calibrating LS_{FS} Scale (continued)

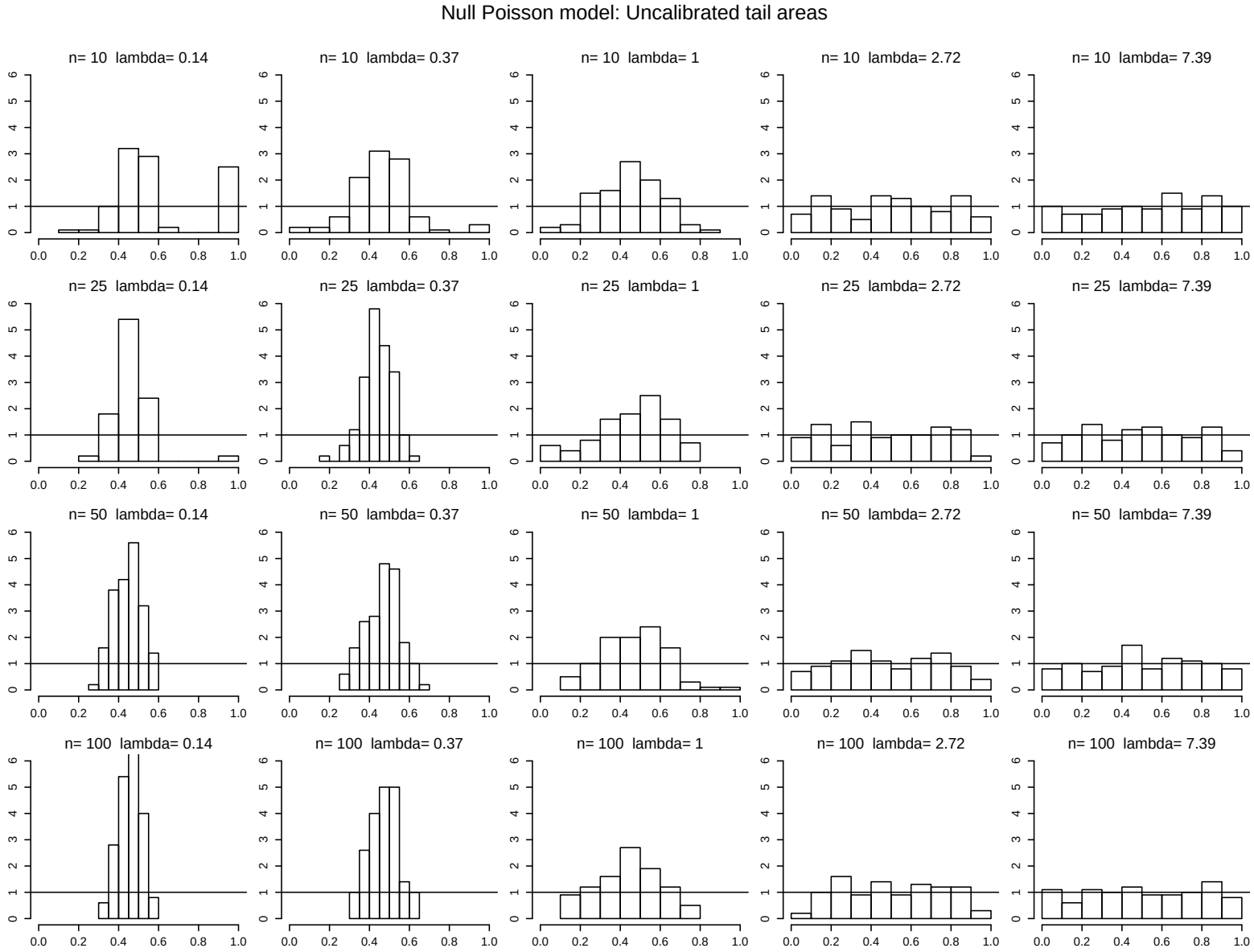
```
  Choose a lambda value of interest; call it lambda.sim

  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**.

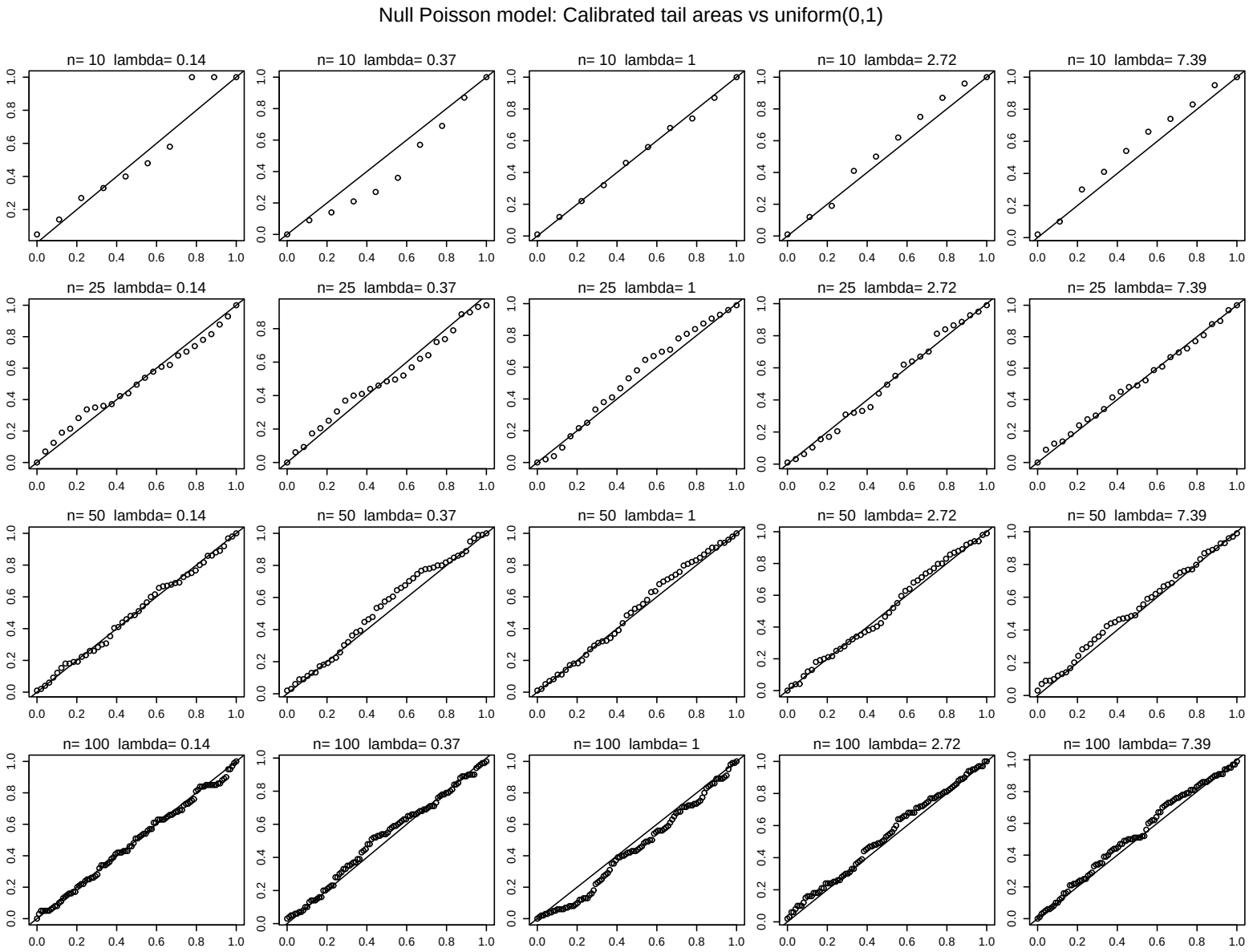
We have **verified** (via simulation) in several simple (and some less simple) situations that the values in V.Pa are close to $U(0, 1)$ in distribution.

Two **examples**—Poisson(λ) and Gaussian(μ, σ^2):

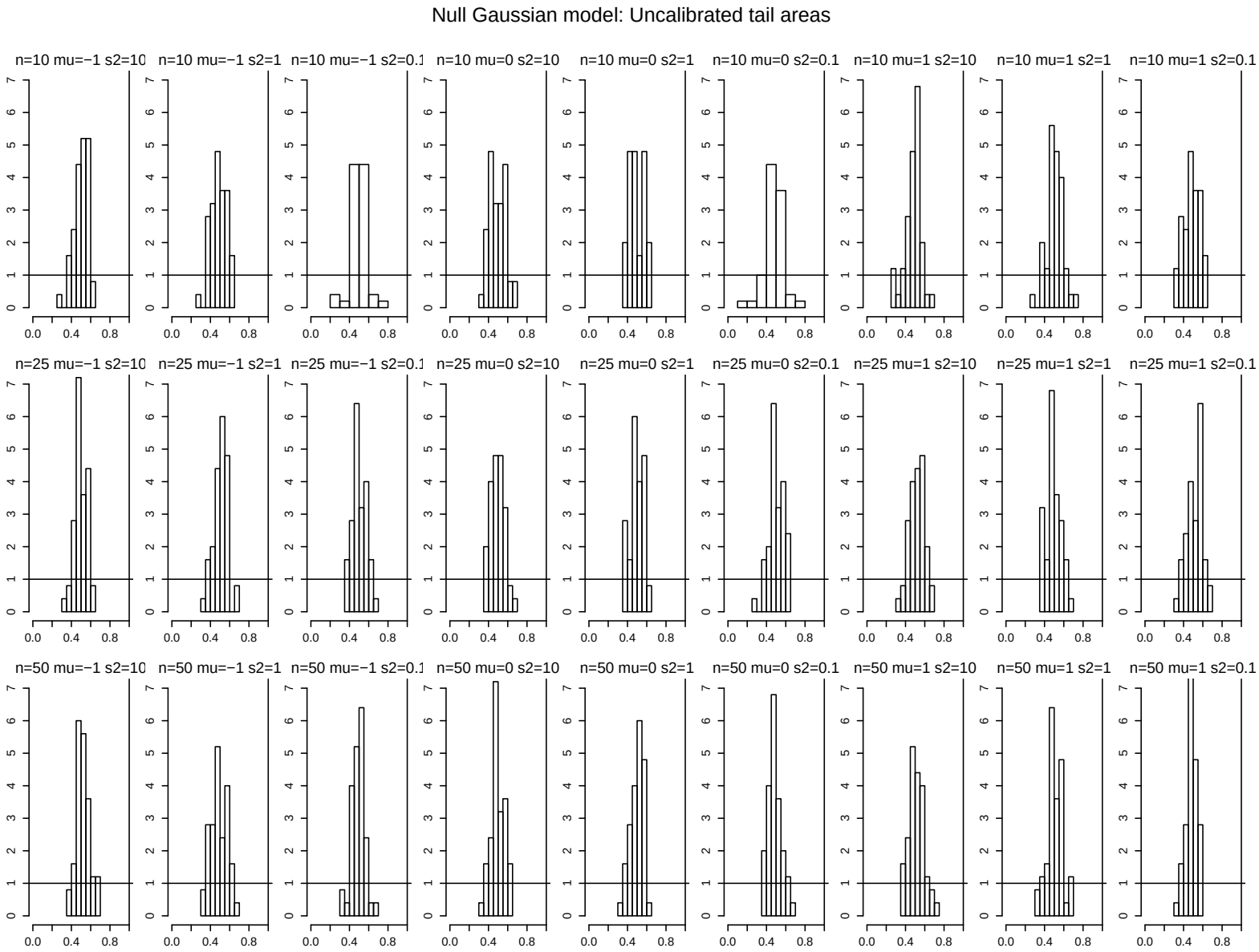
Calibrating LS_{FS} Scale (continued)



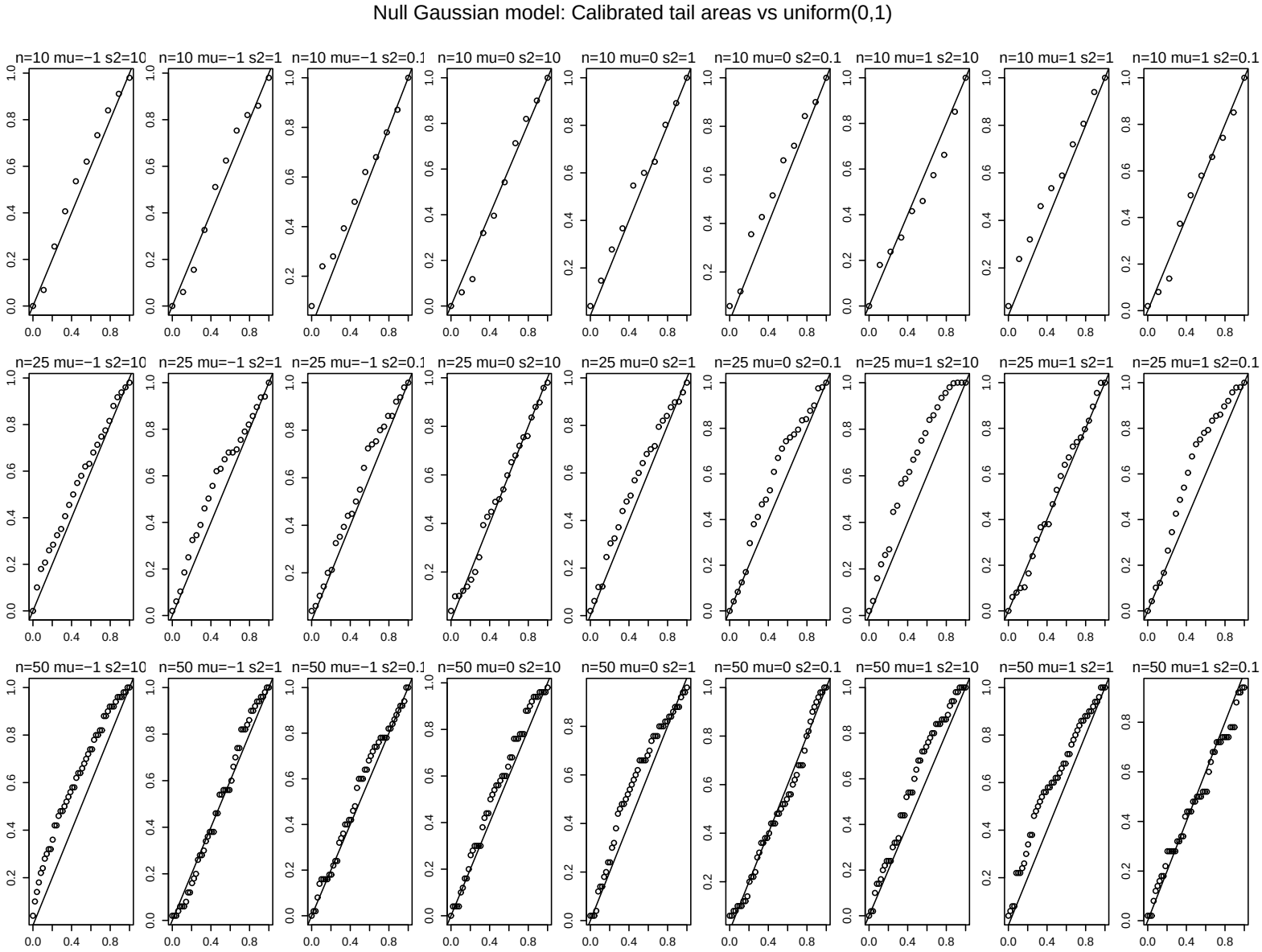
Calibrating LS_{FS} Scale (continued)



Calibrating LS_{FS} Scale (continued)



Calibrating LS_{FS} Scale (continued)



BNP Modeling: An Example

• We describe **parametric** and **BNP** approaches to modeling **count data** and demonstrate advantages of BNP modeling using empirical, predictive, graphical and formal model comparisons (LS_{CV} and LS_{FS}).

• We examine models suitable for analyzing data in **control** (C) and **treatment** (T) setting as in the **IHGA case study** (Hendriksen et al. 1984) in which a number of elderly people were randomized in C group, receiving **standard** care, and T group, which also received **in-home geriatric assessment** (IHGA); the **outcome** of interest was **number of hospitalizations** during two years.

• **Parametric random-effects Poisson** (PREP) model is natural choice for C and T data sets (in parallel):

$$\begin{array}{ccc} (y_i|\theta_i) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta_i)] \\ (\theta_i|G) & \overset{\text{iid}}{\sim} & G \\ G & \equiv & \text{N}(\mu, \sigma^2) \end{array} \tag{38}$$

assuming a parametric CDF G for latent variables θ_i (random effects).

Dirichlet Process Mixture Model

• What if this **assumption** is **wrong**?

• Want to remove the **parametric assumption** on **distribution of random effects** by building a prior model on CDF G that may be centered on $N(\mu, \sigma^2)$, but permits **adaptation** (learning from data).

• Specifying prior for an **unknown distribution** requires a **stochastic process** with realizations (sample paths) that are CDFs.

• We use **Dirichlet process** (DP), in notation $G \sim DP(\alpha, G_0)$, where G_0 is the **center** or **base** distribution of the process and α a **precision** parameter (Ferguson 1973, Antoniak 1974).

• Poisson **DP mixture model**:

$$\begin{aligned} (y_i \mid \theta_i) &\overset{ind}{\sim} \text{Poisson}(\exp(\theta_i)) \\ (\theta_i \mid G) &\overset{iid}{\sim} G \\ G &\sim \text{DP}(\alpha G_0), \quad G_0 \equiv G_0(\cdot; \psi), \end{aligned} \tag{39}$$

where $i = 1, ..., n$ (we refer to (39) as **BNP model 1**).

Dirichlet Process Mixture Model (continued)

- **Equivalent formulation** of the Poisson DP mixture model:

$$(y_i \mid G) \overset{iid}{\sim} f(\cdot; G) = \int \text{Poisson}(y_i; \exp(\theta)) \mathrm{d}G(\theta), \; G \sim \text{DP}(\alpha G_0), \tag{40}$$

where $i = 1, \ldots, n$ and $G_0 = \text{N}(\mu, \sigma^2)$.

- MCMC implemented for a **marginalized** version of DP mixture. **Key idea:** G is integrated out over its prior distribution, (Antoniak 1974, Escobar and West 1995), resulting in $[\theta_1, ..., \theta_n \mid \alpha, \psi]$ that follows **Pólya urn** structure (Blackwell and MacQueen, 1973).

- **Specifically**, $[\theta_1, ..., \theta_n \mid \alpha, \psi]$ is

$$g_{r0}(\theta_{r1} \mid \mu_r, \sigma_r^2) \prod_{i=2}^{n_r} \left\{ \frac{\alpha_r}{\alpha_r + i - 1} g_{r0}(\theta_{ri} \mid \mu_r, \sigma_r^2) + \frac{1}{\alpha_r + i - 1} \sum_{\ell=1}^{i-1} \delta_{\theta_{r\ell}}(\theta_{ri}) \right\}.$$

DP Mixture Model with Stochastic Order

• There are cases when treatment **always has an effect**, only the **extent** of which is unknown. This can be expressed by introducing **stochastic order** for the random effects distributions: $G_1(\theta) \geq G_2(\theta), \theta \in R$, denoted by $G_1 \leq_{st} G_2$.

• Posterior **predictive** inference can be improved under this assumption if we incorporate stochastic order in the model. To that end we introduce a **prior** over the space $\mathcal{P} = \{(G_1, G_2) : G_1 \leq_{st} G_2\}$.

• A convenient way to **specify** such a prior is to work with subspace $\mathcal{P}'$ of $\mathcal{P}$, where $\mathcal{P}' = \{(G_1, G_2) : G_1 = H_1, G_2 = H_1 H_2\}$, with H_1 and H_2 d.f.-s on R, and then place **independent DP priors** on H_1 and H_2.

• Note: to obtain a **sample** θ from $G_2 = H_1 H_2$, **independently** draw θ_1 from H_1 and θ_2 from H_2, and then set $\theta = \max(\theta_1, \theta_2)$.

• Specifying **independent DP priors** on **mixing distributions** H_1 and H_2 we obtain the following model:

DPMM with Stochastic Order (continued)

$$
\begin{aligned}
Y_{1i} \mid \theta_i &\overset{ind}{\sim} \text{Poisson}(\exp(\theta_i)), i = 1, n_1 \\
Y_{2k} \mid \theta_{1,n_1+k}, \theta_{2k} &\overset{ind}{\sim} \text{Poisson}(\exp(\max(\theta_{1,n_1+k}, \theta_{2k}))), k = 1, n_2 \\
\theta_{1i} \mid H_1 &\overset{iid}{\sim} H_1, i = 1, n_1 + n_2 \\
\theta_{2k} \mid H_2 &\overset{iid}{\sim} H_2, k = 1, n_2 \\
H_r \mid \alpha_r, \mu_r, \sigma_r^2 &\sim DP(\alpha_r H_{r0})
\end{aligned}
\tag{41}
$$

where the **base distributions** of Dirichlet processes, H_{10} and H_{20}, are again **Normal** with parametric priors on hyperparameters. We refer to (41) as BNP model 2.

- We implement a **standard MCMC** with an extension for **stochastic order** (Gelfand and Kottas, 2002).

- To create a **level playing field** to compare quality of PREP and BNP models we compute **predictive distributions** for future data, based on predictive distribution for **latent variables** and posterior **parameter samples**.

Posterior Predictive Distributions

- For BNP model 1 the **posterior predictive** for a future Y^{new} is

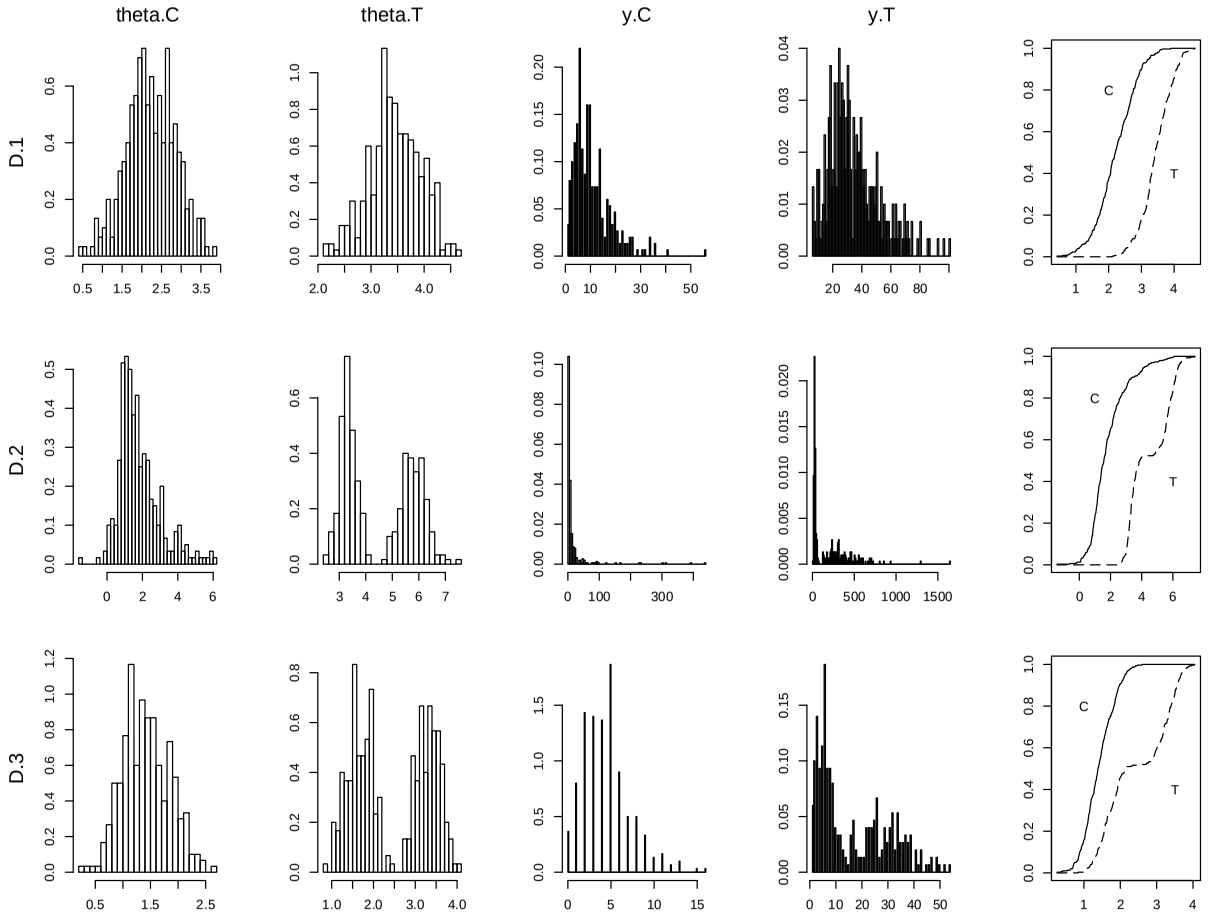
$$[Y^{\text{new}} \mid \text{data}] = \iint \text{Poisson}(Y^{\text{new}}; \exp(\theta^{\text{new}}))[\theta^{\text{new}} \mid \boldsymbol{\eta}][\boldsymbol{\eta} \mid \text{data}], \tag{42}$$

where θ^{new} is associated with Y^{new} and η collects **all model parameters** except θs (we use **bracket notation** of Gelfand and Smith (1990) to denote distribution function).

- The posterior predictive for **latent variables**, induced by **Pólya urn** structure of DP, is

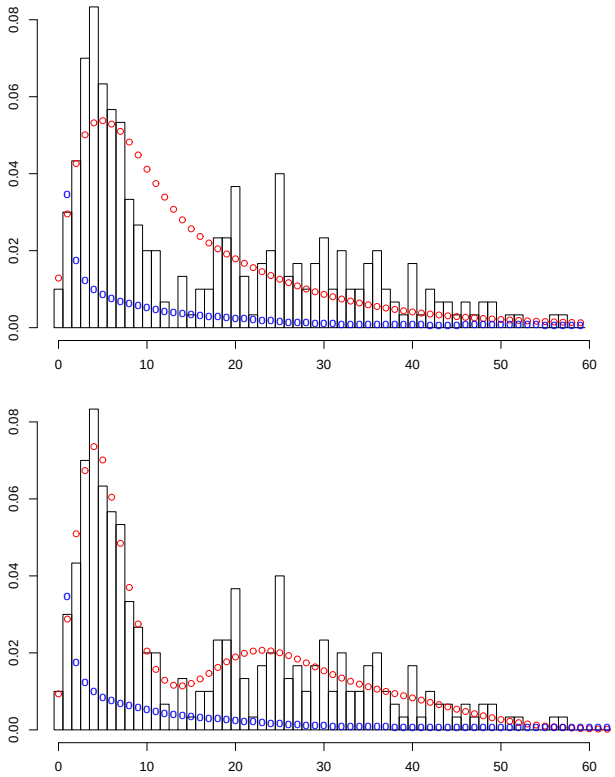
$$[\theta^{\text{new}} \mid \boldsymbol{\eta}] = \frac{\alpha}{\alpha + n} G_{r0}(\theta^{\text{new}} \mid \mu_r, \sigma^2) + \frac{1}{\alpha + n} \sum_{\ell=1}^{n} n_\ell \delta_{\theta_\ell}(\theta^{\text{new}}). \tag{43}$$

Simulation: Random-Effects and Data Sets



Simulation data sets for control (C) and treatment (T) ($n = 300$ observations in each), and distributions of **latent variables** (D_1: C and T both Gaussian; D_2: C skewed, T bimodal; D_3: C Gaussian, T bimodal, $C \leq_{st} T$).

Predictive: PREP Versus BNP Model 1



Prior (lower [blue] circles) and **posterior** (upper [red] circles) **predictive distributions** for PREP model (top) and BNP model 1 (bottom) for data set D_3 with **bimodal random effects**.

The PREP model **cannot adapt** to the bimodality (without **remodeling** as, e.g., a **mixture** of Gaussians on the latent scale), whereas the BNP modeling **smoothly adapts to the data-generating mechanism**.

Posterior Inference for G

• Perhaps more interestingly, using generic approach for inference about **random mixing distribution**, we can obtain $[G \mid \text{data}]$, based on which we can compute posterior of any **linear functional** of G, e.g. $[E(y|G)]$.

• With $G \sim DP(\alpha G_0)$, following Ferguson (1973) and Antoniak (1974),

$$[G|\text{data}] = \int [G|\theta, \alpha, \psi] \text{d}[\theta, \alpha, \psi|\text{data}]. \tag{44}$$

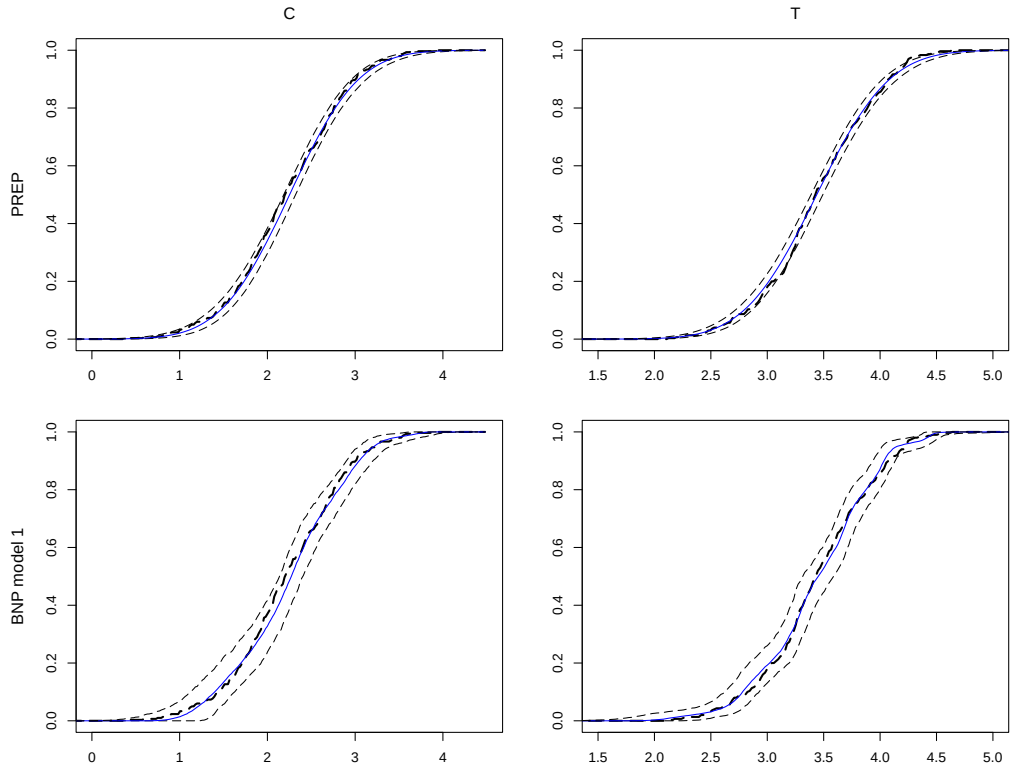
where $[G|\theta, \alpha, \psi]$ is **also a DP** with parameters $\alpha^{'} = \alpha + n$ and

$$G_0^{'}(\cdot|\psi) = \frac{\alpha}{\alpha + n} G_0(\cdot|\psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{(-\infty, \theta_i]}(\cdot), \tag{45}$$

where $\theta = (\theta_1, ..., \theta_n)$ and ψ collects parameters of G_0.

• Using (44), (45) and the definition of DP we develop **computationally efficient** approach to obtaining **posterior sample paths** from $[G \mid \text{data}]$.

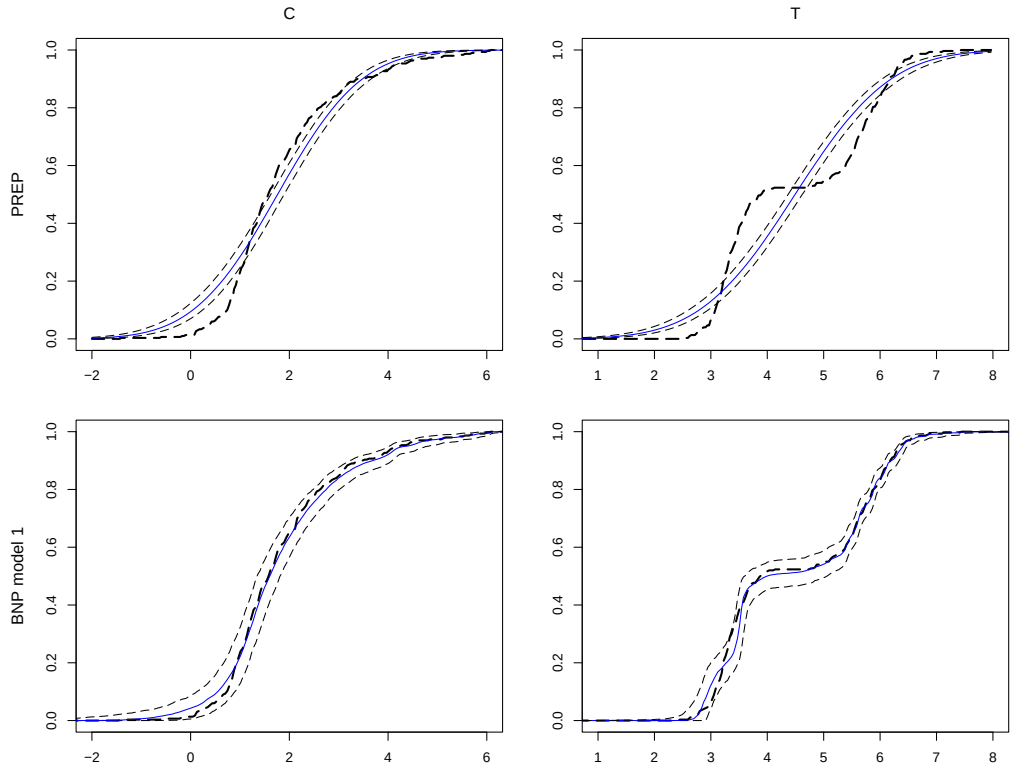
Normal Random Effects: PREP vs. BNP



Normal random effects (data set D_1): Posterior MCMC estimates of the **random effects distributions** for PREP model (first row) and BNP model 1 (second row).

When PREP is **correct** it (naturally) yields **narrower uncertainty bands** (but see below).

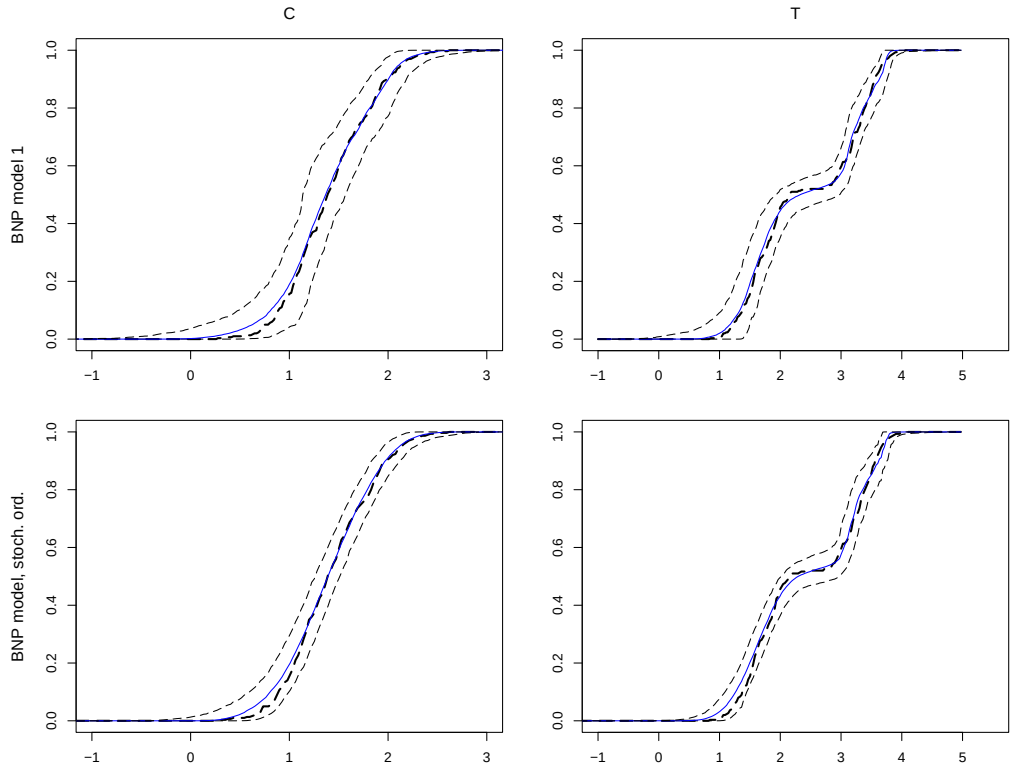
Skewed and Bimodal Random Effects, PREP vs. BNP



Skewed and **bimodal** random effects (data set D_2): Posterior MCMC estimates of **random effects distributions** for PREP model (first row) and BNP model 1 (second row).

When PREP is **incorrect** it continues to yield **narrower uncertainty bands** that unfortunately **fail to include the truth**, whereas BNP model 1 **adapts successfully** to the data-generating mechanism.

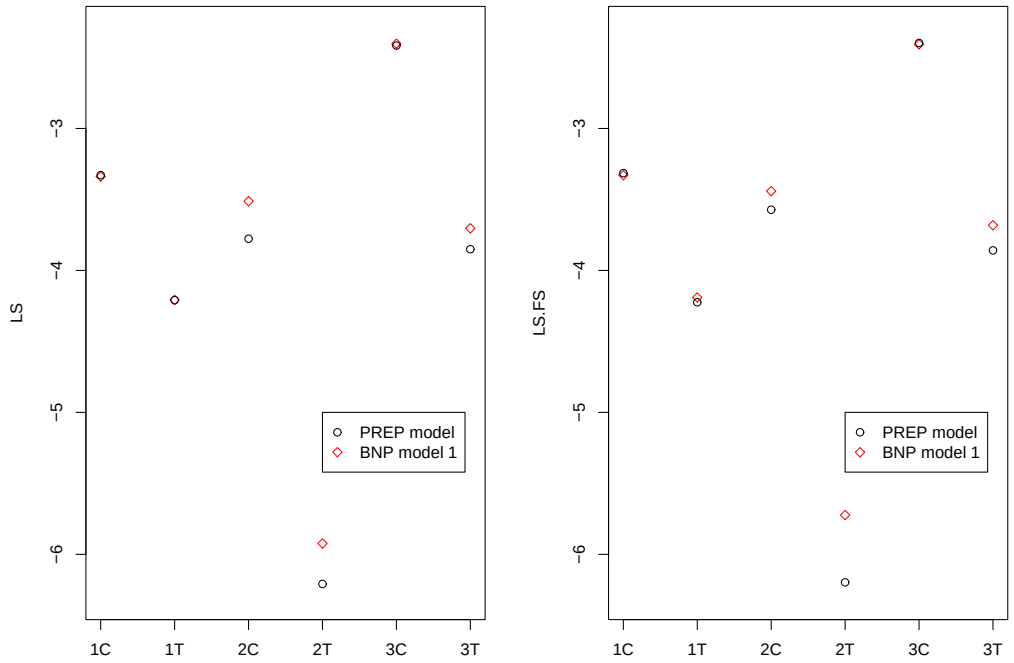
BNP With and Without Stochastic Order



Bimodal random effects in T (data set D_3): Posterior MCMC estimates of **random effects distributions** for BNP model 1 (first row) and BNP model with **stochastic order** (second row).

Extra assumption of **stochastic order**, when true, yields **narrower uncertainty bands** (as it should).

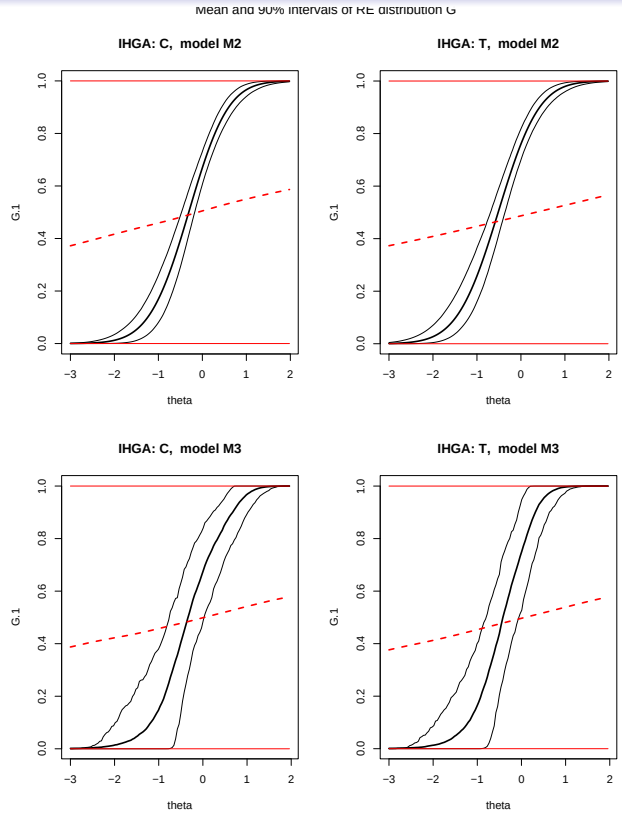
LS_{CV} and LS_{FS} For PREP and BNP Models



LS_{CV} (left panel) versus **full-sample log-score** LS_{FS} (right panel) for PREP and BNP models for all 3 data sets (C and T), $D_{1,C}, \ldots, D_{3,T}$.

When PREP is **correct** (1C, 1T, 3C), it has **small advantage** in LS_{CV} and LS_{FS} over BNP (as it should), but when PREP is **incorrect** (2C, 2T, 3T) both kinds of LS give a **clear preference for BNP model 1** (also as they should).

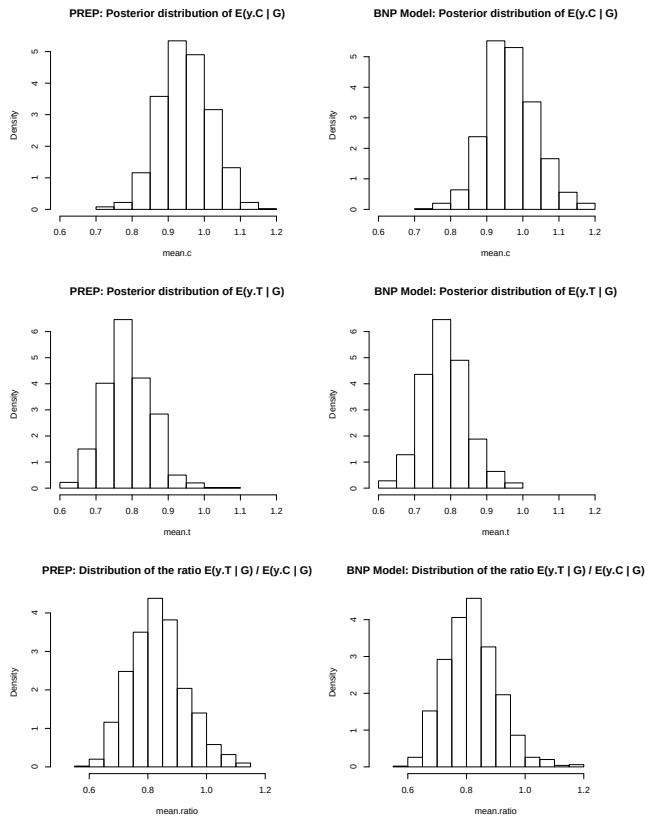
Results in IHGA Case Study



Results on **IHGA data** in case study: posterior mean and 90% intervals for **random-effects distribution** G (first column is C sample, second column is T; first row is PREP model, second row is BNP model 1).

Uncertainty bands are **wider** from BNP model 1, but **direct comparison not fair** because PREP model arrived at via **data-analytic search on entire data set**.

Results in IHGA Case Study (continued)



Results on **IHGA data** in case study: first row is posterior for C mean, second row is posterior for T mean, third row is posterior for **ratio of means**; first column is PREP model, second column is BNP model 1.

With **suitable amount of data** held out for **calibration check** in subset S_3 in 3CV (**about 25%**), BNP and 3CV achieve **comparable results**.

Conclusions

• Standard (**data-analytic**) (DA) approach to **model specification** "**shops around** for the **'right'** model," thereby often yielding **poorly calibrated (too narrow) predictive intervals** (a symptom of **incoherence**).

• I'm aware of **two principled solutions** to this problem:

– {**Exchangeability judgments** plus **Bayesian nonparametric (BNP)** modeling} (this solves the problem by **avoiding (some of) the shopping**: with BNP (and enough data), **no need to use DA** to specify many modeling details (**error distributions, response surfaces**); but will often still need to **compare** models with **different sets of exchangeability judgments**); and

– **3-way out-of-sample predictive cross-validation (3CV)**, a modification of DA in which the data are **partitioned** into **3** (rather than the usual 2) subsets S_1, S_2, S_3; a **DA search** is undertaken iteratively, **modeling** with S_1 and **predictively validating** with S_2; and S_3 is **not used in quoting final uncertainty assessments**, but is instead used to **evaluate predictive calibration of the entire modeling process** (this solves the problem by paying the **"right" price** for shopping around).

Conclusions (continued)

- **Two basic kinds of model choices** need to be made in **both BNP and 3CV**:

$\boxed{Q_1}$ Is M_1 **better than** M_2? $\boxed{Q_2}$ Is M_1 **good enough**?

- ($\boxed{Q_1}$ and $\boxed{Q_2}$) Model choice is really a **decision problem** and should be approached via **MEU**, with a utility structure that's **sensitive to the real-world context**.

- ($\boxed{Q_1}$ and $\boxed{Q_2}$) 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.

- ($\boxed{Q_1}$) 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**.

Conclusions (continued)

- ($\boxed{Q_1}$) The **full-sample log score** (LS_{FS}) is n times **faster** than naive implementations of LS_{CV}, has better **small-sample model discrimination accuracy** than either LS_{CV} or DIC, and **has better asymptotic behavior** than LS_{CV}.

- ($\boxed{Q_1}$) **Generic 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.

- ($\boxed{Q_2}$) The basic Gelman et al. (1996) method of **posterior predictive model-checking** is **badly calibrated**: when it gives you a tail area of, e.g., **0.4**, the calibrated equivalent may well be **0.04**.

- ($\boxed{Q_2}$) We have modified an **approach** suggested by Robins et al. (2000) to help answer the question "Could the data have arisen from M_1?" in a **well-calibrated** way.

Conclusions (continued)

- People often talk about **BNP modeling** as providing **"insurance"** against **mis-specified parametric models**:

(1) You can **simulate** from a known **("true") parametric model** M_1 and fit M_1 and BNP to the simulated data sets; both will be **valid** (both will **reconstruct** the **right answer averaging across simulation replications**) but the BNP **uncertainty bands** will typically be **wider**.

(2) You can also simulate from a **different parametric model** M_2 and fit M_1 and BNP to the simulated data sets; often now **only BNP will be valid**.

People refer to the **wider uncertainty bands** for BNP in (1) as the **"insurance premium"** you have to pay with BNP **to get the extra validity** of BNP in (2).

But this is **not a fair comparison**: the simulation results in (1) and (2) were all **conditional on a known "true" model**, and don't immediately apply to a **real-world setting** in which **you don't know what the "true" model is**.

Conclusions (continued)

When you **pay an appropriate price** for shopping around for the **"right" parametric model** (as in 3CV), the **discrepancy** between the parametric and BNP uncertainty bands **vanishes**.

- In preliminary results (with random-effects models in T versus C randomized trials), the **right amount of data** to allocate to subset S_3 to make this happen with moderate sample sizes is about **25%**, leading to a **recommended allocation of data** across (S_1, S_2, S_3) in the vicinity of **(50%, 25%, 25%)**.

In other words, with $n =$ **1,000** I should be prepared to **pay about 250 observations worth of information** in **quoting my final uncertainty assessments** (i.e., make these uncertainty assessments **about** $\sqrt{\frac{n}{0.75n}} \doteq 15\%$ **wider** than those based on the full data set), to **account in a well-calibrated manner** for my **search for a good model**.