Bayesian Model Specification: Quantifying the Price of Model Uncertainty

David Draper (joint work with
Milovan Krnjajić and **Thanasis Kottas**)

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

draper@ams.ucsc.edu

www.ams.ucsc.edu/∼draper

ISBA 2008
Hamilton Island, Australia

22 July 2008

Modeling

We use **models** constantly in our **inferential** and **predictive** work.

Most of the time, one or more **features** of such models are arrived at after a **search** (typically **guided by the data**) among possible modeling choices.

Often we deal with the **uncertainty in the modeling process** implied by this search by **ignoring** it: we find the **best model** (in some sense) and carry out our inferences and predictions **conditional on this model**, as if it were "correct."

Sometimes this produces **well-calibrated** answers (for instance, when there's **little ambiguity** about sensible modeling choices); sometimes it **doesn't**.

When we follow through to see if our inferential or predictive statements were right **about as often as we asserted they would be**, we find most frequently that **lack of calibration** is in the direction of **insufficient conservatism** — in other words, we had **more uncertainty** than we were willing to admit.

(**Cross-validation** can help, but not necessarily as it's **usually practiced**; more later.)

The Price of Model Uncertainty

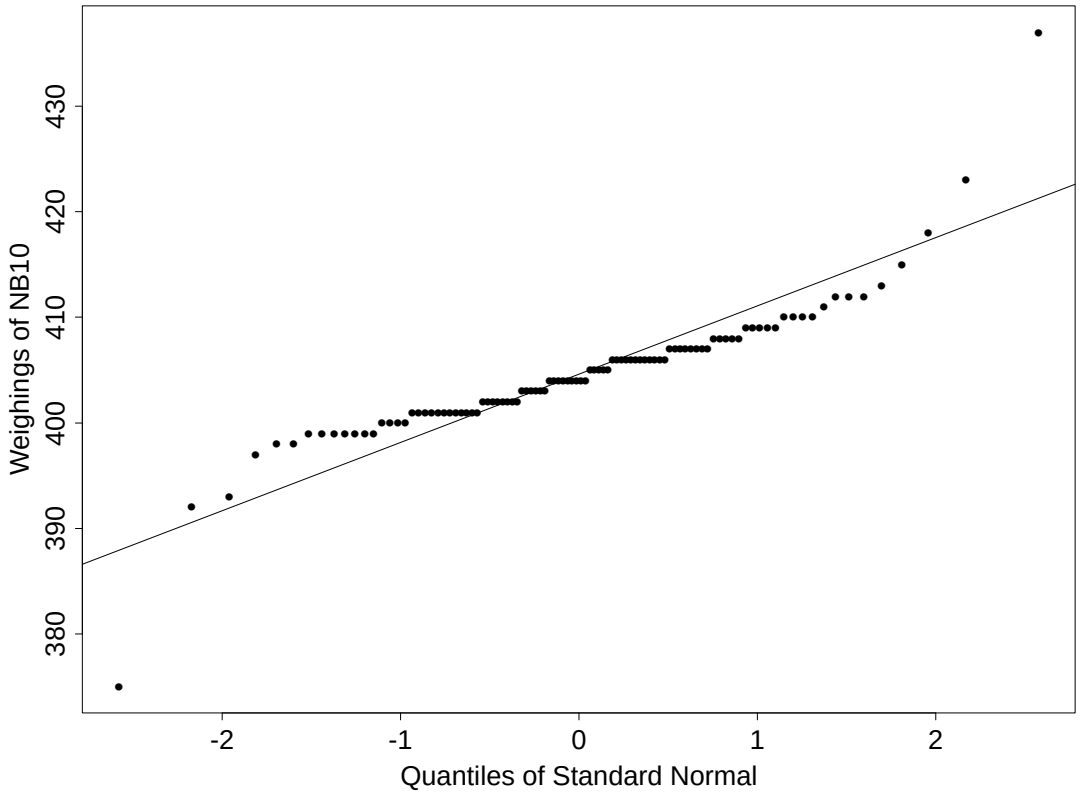
One possible explanation for this **insufficient conservatism** is the **underpropagation of modeling uncertainty** just mentioned.

One would ordinarily expect that **acknowledging greater modeling uncertainty** — for example, by making a **standard modeling choice** a **special case** of a larger family of models (**model expansion**) and enlisting the help of the data to **identify a better model** in this larger family, rather than just assuming the standard model is "correct" — would lead to an **increase in inferential or predictive uncertainty**.

Interestingly, this is **often, but not always, true**: the **price** of model uncertainty is usually **positive** (i.e., uncertainty about the **quantity of main interest** is **larger** when model uncertainty is properly acknowledged, which is like having **less data** than we think we have, which is like having to **pay more** to get the accuracy we want) but **can sometimes be negative**.

I'll start with an **example** of this **relatively rare behavior**, and then I'll describe one approach to **quantifying the price of model uncertainty** rather generally.

NB10



This is a **normal quantile-quantile plot** of $n = 100$ weighings of a checkweight called **NB10**, made by workers at the U.S. National Bureau of Standards in 1962–63 under conditions as close to **IID** as possible (the units are **micrograms** below the nominal weight of **10g**).

How much does NB10 **weigh**?

Gaussian Versus t Likelihood

The plot above shows that it's plausible in answering this question to assume a **symmetric location-scale model**, but the **form** of the **error distribution** is **less clear**.

The standard choice is **Gaussian**; with μ as true weight and little or no prior information, **posterior distribution** for μ based on an assumption of Gaussian errors is **close to normal** with mean **404.6** and standard deviation **0.65**.

But the **solid line** in the plot is the **target shape** for the plot if the data were in fact **Gaussian**, and there is **clear evidence for heavier tails**.

If one were to instead adopt (say) a t_k **model** for the errors and treat the **degrees of freedom** k as **unknown** — which corresponds to an **increase in model uncertainty** — it turns out that the posterior SD **drops** to **0.46**, a **29% decrease**.

Is this a **fluke** or an example of a **general phenomenon**?

Consider $Y_i, i = 1, \ldots, n$, IID (given (μ, σ)) from a **symmetric location-scale family** $Y_i = \mu + \sigma e_i$, where the e_i are assumed to have **two finite moments** and **support** $(-\infty, \infty)$.

Fisher Information Is Of Course Key

Suppose as above that the exact form of the density $f(y|\mu, \sigma)$ of the Y_i is **not known** a priori, and interest focuses on the **effect**, on uncertainty assessments about μ, implied by this **model uncertainty** about f.

Without loss of generality take $E(e_i) = 0$ and $V(e_i) = 1$ so that μ and $\sigma^2 = V(Y_i)$ have the **same meaning** in all models to be compared.

For **large** n and **little or no prior information** about μ, **uncertainty assessments** for μ will be based on the **Fisher information** for location.

The posterior distribution for μ is then **approximately normal** with mean given by the **maximum likelihood estimator** (MLE) $\hat{\mu}$ and variance $\hat{I}^{-1}(f)\sigma^2/n$, where

$$I(f) = \int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)} dx. \tag{1}$$

Here $\hat{I}$ is **observed information** for a single observation and f is the (mean 0, SD 1) density of the **normalized errors** e_i.

The Gaussian Is Conservative

It's a fact (Kagan, Linnik & Rao, 1973; Draper, 2008 (calculus-of-variations proof that seems to be new)) that in this situation the **off-the-shelf Gaussian choice** for f is **conservative**:

Theorem: Under the above assumptions, **$I(f)$ is minimized by the standard Gaussian distribution**.

This means that if one were to **place the Gaussian in a larger family** of densities f_β in which it's a **special case** ($\beta = 0$, say), and compare **two modeling strategies**,

- **Strategy 1:** I assert that the Y_i are **Gaussian**, which corresponds to placing all my prior mass on $\beta = 0$; or
- **Strategy 2:** I express **little prior knowledge** of β and await the **data's information** about **plausible values** for it,

the second strategy would admit **greater model uncertainty** than the first, and yet — at least for large n — would lead to **smaller** uncertainty assessments for μ.

t and Generalized Power-Exponential

The t_k family mentioned above (with $k > 2$; take $\beta = \frac{1}{k}$ to place the Gaussian at 0) is **one instance** of this model; it's been studied in **location-scale problems** by Lange, Little & Taylor (1989).

Another example, this time including distributions with tails **both heavier and lighter than those of the Gaussian**, is the **generalized power-exponential distributions** examined by Box & Tiao (1962),

$$f(x|\beta) = c \ \exp\left\{-\frac{1}{2}|x|^{2/(1+\beta)}\right\}, \tag{2}$$

where c is a normalizing constant.

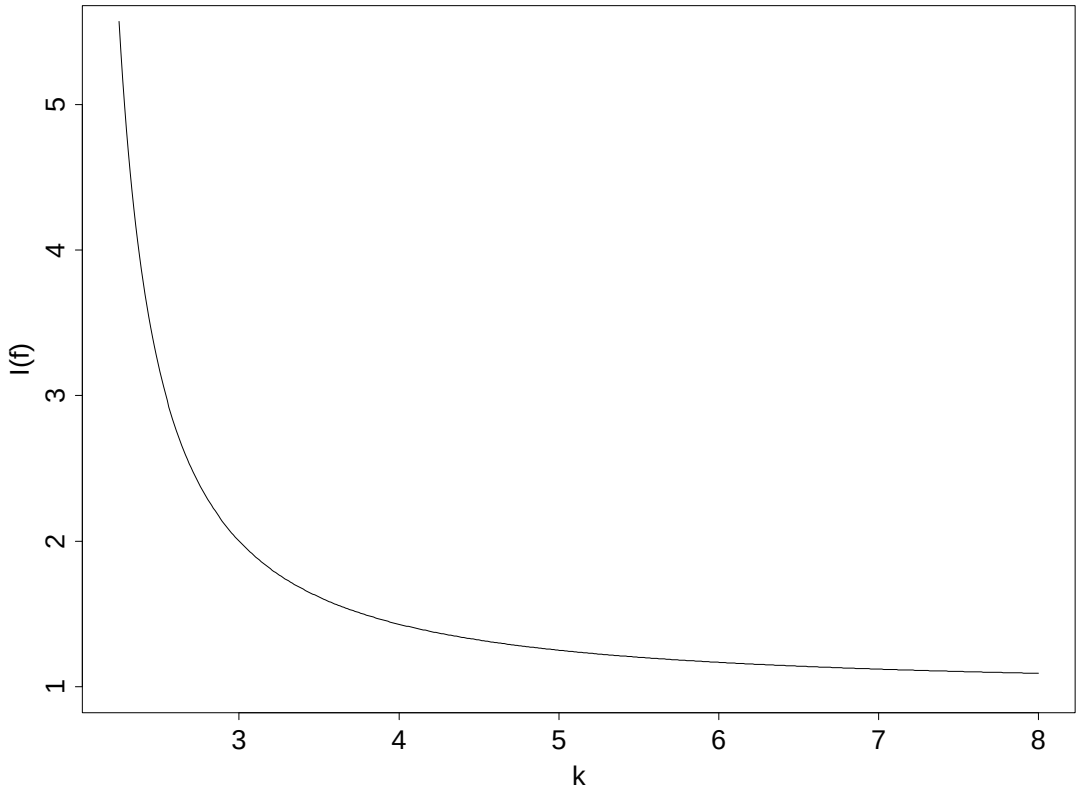
Lange, Little & Taylor note, in both of these models, that in large samples **scale** σ and **shape** β may be **confounded** (although insisting that $V(Y_i) = \sigma^2$ in all models permits direct examination of the effect of β on the posterior variance of μ given the data), but that (σ, β) will be **uncorrelated** with location μ, so that (at least with large n) **one pays no price in Strategy 2, in uncertainty about μ, for one's uncertainty about β.**

Fisher Information For the t_k Family

In the t_k family with $k > 2$, $I(f)$ has the **simple expression**

$$I(t_k^*) = \frac{k(k+1)}{(k+3)(k-2)}, \tag{3}$$

where t_k^* is the **scaled** t-distribution with k **degrees of freedom** and variance 1 (cf. Taylor, 1992).



NB10 Is Not a Fluke

k	2.1	2.5	3.0	4.0	5.0	6.0	7.0	8.0	10.0	20.0
c_k^*	0.28	0.56	0.71	0.84	0.89	0.93	0.95	0.96	0.97	0.99

The table above gives some values of the **multiplier** $c_k^* = I^{-\frac{1}{2}}(t_k^*)$ in the expression

$$\left(\begin{array}{c} \text{posterior SD for } \mu \\ \text{assuming } t_k \end{array} \right) = c_k^* \left(\begin{array}{c} \text{posterior SD for } \mu \\ \text{assuming normality} \end{array} \right) \tag{4}$$

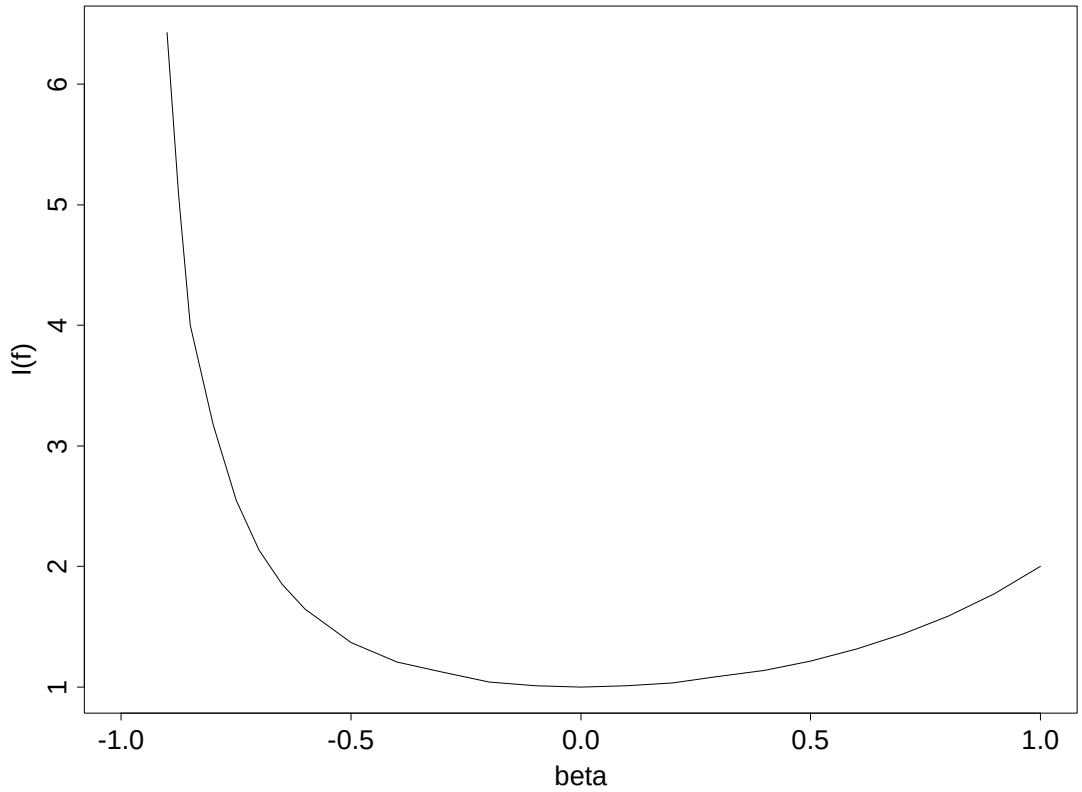
for **selected values** of k.

The table indicates that **noticeable decreases** in **inferential uncertainty** from that implied by the Gaussian will only occur in the t_k model in datasets for which the posterior distribution for k concentrates **most of its mass** on $k \leq 5$ **or so**.

For the **NB10 data** the MLE of k is **3.0** (with a standard error of 0.86), and the observed **decrease in inferential uncertainty** when moving from Strategy 1 to 2 agrees closely with the relevant value from the table: the ratio of posterior SDs is $\frac{0.46}{0.65} = \mathbf{0.71}$.

The Power-Exponential Family

In the **power-exponential family** no closed-form expression for $I(f)$ is available, but **numerical comparisons** may still be made.



In the limit as $\beta \to -1$ one obtains the **uniform** distribution, and $\beta = +1$ is the **double-exponential** distribution.

Power-Exponential Results

β	-0.9	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0
c^*_β	0.39	0.56	0.78	0.91	0.98	1.00	0.98	0.94	0.87	0.79	0.71

The table above gives values of the **multiplier** c^*_β analogous to c^*_k in the previous table.

In the limit as $\beta \to -1$, $I(f)$ becomes **infinite**, reflecting the fact that inferences about a **location parameter** with **uniform errors** are an **order of magnitude** (in powers of n on the variance scale) **more accurate** than with $\beta > -1$.

These are **large-sample results**; with small n (e.g., Box & Tiao, 1962: **Darwin's data** on the heights of self- and cross-fertilized plants ($n = 15$)) the **uncertainty** added through **estimating** β can make members of the power-exponential or t families other than the Gaussian have **(slightly) larger posterior variance** for μ than that induced by the Gaussian.

However, with $Y = (Y_1, \ldots, Y_n)$,

$$V(\mu|Y) = V_{(\beta|Y)}[E(\mu|Y,\beta)] + E_{(\beta|Y)}[V(\mu|Y,\beta)]. \tag{5}$$

Asymptotics

$$V(\mu|Y) = V_{(\beta|Y)}[E(\mu|Y,\beta)] + E_{(\beta|Y)}[V(\mu|Y,\beta)].$$

The first term captures **variability** in the **conditional expectations** of μ as a function of β; the second is a **summary** of the **conditional variance** of μ given β.

As n increases both terms on the right-hand side **go to 0** like $\frac{c}{n}$, but **simulations** reveal that the numerator constant c for the **first** term is **far smaller** than that for the **second** term, which may be **approximated** by $V(\mu|Y,\beta = \hat{\beta}_{\text{MLE}}) \leq V(\mu|Y,\beta = 0)$; thus for **large** n

$$\begin{aligned} V(\mu|Y) &= V_{(\beta|Y)}[E(\mu|Y,\beta)] + E_{(\beta|Y)}[V(\mu|Y,\beta)] \\ &\doteq E_{(\beta|Y)}[V(\mu|Y,\beta)] \\ &\doteq V(\mu|Y,\beta = \hat{\beta}_{\text{MLE}}) \leq V(\mu|Y,\beta = 0). \end{aligned} \tag{6}$$

The **conclusions** here also apply to **more general location-scale problems**, including **regression** (see, e.g., Lange, Little & Taylor, 1989).

The Default Choice (Unusually) Is Conservative

All of this may be **summarized** with the statement that

For whatever reason — historical accident or otherwise — from a model uncertainty point of view the default Gaussian choice for the underlying error distribution in large–n location-scale problems is conservative.

(This result is related to the **maximum-entropy** property of the Gaussian — see, e.g., Rao, 1973 — although it's **not straightforward** to connect **entropy** and **Fisher information** for location **algebraically**.)

This conclusion has a **frequentist parallel** in the case of **long-tailed data** arising from **robustness considerations**: the point of robust estimators is to **downweight** outliers, and when it's appropriate to do so the result will be **sharper inferential statements**.

It's perhaps **less frequently** noted that the **same effect** occurs with **light-tailed data**, and for a **different reason**: moving from the **Gaussian** to the **uniform** involves **crossing over** from **inferential uncertainty assessments** of the form $V(\mu|Y) = O(n^{-1})$ to $O(n^{-2})$.

Conditional Inherent Accuracy

One **natural reaction** to the results above is to attribute the phenomenon to **goodness-of-fit**.

For example, it might seem **intuitively plausible** to point out that **moving** from the **Gaussian** to the t_3 model for the NB10 data amounts to **switching** from a model that **doesn't fit well** to one that **does**, and one may expect to enjoy a decrease in inferential uncertainty as a result.

But the **fit** of a model M and what may be termed the **conditional inherent accuracy** (for a given unknown quantity like μ) given M are **two different things**.

This may perhaps be **seen most directly** by **running the NB10 experience in reverse**: if the data had followed a **Gaussian** model and one had **begun** by instead assuming t_3 errors, **embedding** the t_3 model in the **larger** t_k framework would reveal that the **Gaussian** fit better, and yet the move from t_3 to Gaussian would involve **at most a negligible decline in inferential uncertainty**.

General Quantification of Model Uncertainty

The **Bayesian formulation** of **model uncertainty** is **clarifying**: with y as an **unknown** quantity of interest, x as what's **known**, and M as a **model** relating y to x,

$$p(y|x) = \int p(y|x, M)\, p(M|x)\, dM. \tag{7}$$

The **second term** — $p(M|x)$ — in the product in this integral captures **goodness-of-fit**, the **first** — $p(y|x, M)$ — represents **conditional inherent accuracy**; a retrospectively **well-calibrated uncertainty assessment** for y relies on **both**, and the two terms play **different roles** in such an assessment.

How can one **quantify the price of model uncertainty** generally?

One idea involves a **comparison** of **how much data Bayesian parametric** and **nonparametric models** need to achieve the **same inferential accuracy** about the **main quantity of interest**.

Modeling of Count Data

Example: **Count data — parametric random-effects Poisson** (PREP) modeling versus **Bayesian nonparametric** (BNP) modeling with a **Dirichlet process prior**.

The **first** thing to try **parametrically** with count data is usually a **fixed-effects Poisson model**; in the one-sample case, for simplicity, and parameterizing with θ as the **log** of the usual mean parameter λ (for $i = 1, \ldots, n$),

$$\begin{array}{rcl} (y_i|\theta) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta)] \\ (\theta|\mu,\sigma^2) & \overset{\text{iid}}{\sim} & N(\mu,\sigma^2) \\ (\mu,\sigma^2) & \sim & p(\mu,\sigma^2). \end{array} \tag{8}$$

This uses a **Lognormal** prior for λ rather than the more usual **conjugate Gamma** choice; the two families are **similar**, and the Lognormal **generalizes more readily**, as follows.

In practice there will often be **heterogeneity** (**extra-Poisson variability**), manifesting as a **variance-to-mean ratio** greater than **1**.

Parametric Random-Effects Poisson Model

The next thing to try **parametrically** would then be a **random-effects Poisson model** (PREP):

$$\begin{array}{rcl} (y_i|\theta_i) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta_i)] \\ (\theta_i|G) & \overset{\text{iid}}{\sim} & G \\ G & \equiv & \text{N}(\mu, \sigma^2) \\ (\mu, \sigma^2) & \sim & p(\mu, \sigma^2), \end{array} \tag{9}$$

assuming a parametric CDF G (the **Gaussian**) for the **latent variables** or **random effects** θ_i.

But the **mixing distribution** in the **population** to which it's appropriate to **generalize** may be **multimodal** or **skewed**, which a **single Gaussian can't capture**; if so, this PREP model can **fail to be valid**.

Moreover, this would usually be **diagnosed** with something like **density trace** of **posterior means** of θ_i, looking for need to use **mixture of Gaussians** instead of single one, but choosing G to be Gaussian will tend to make diagnostics **support Gaussian model** even when it's not right.

Dirichlet Process Mixture Model

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

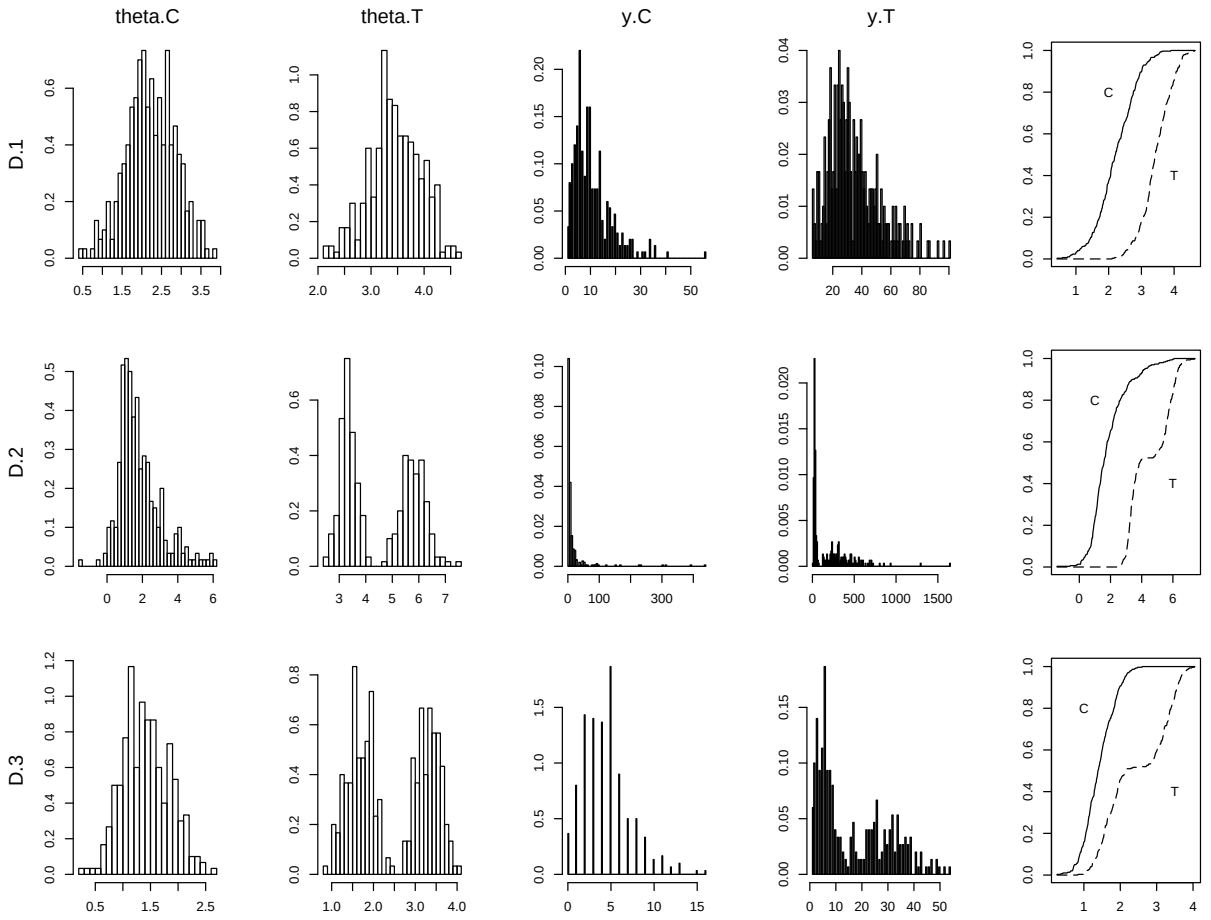
Specifying a prior for an **unknown distribution** requires a **stochastic process** with realizations (sample paths) that are CDFs.

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

Poisson **DP mixture model** (**PDPMM**: this talk's **BNP model**), for $i = 1, \ldots, n$:

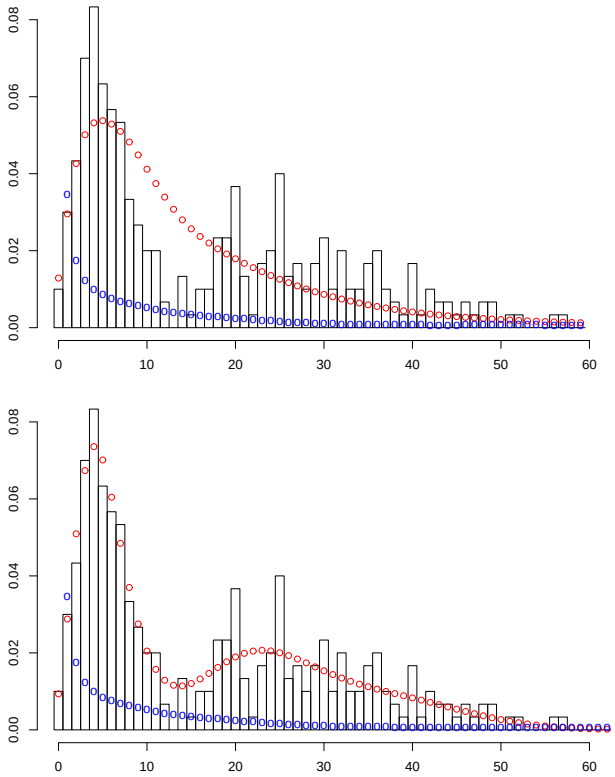
$$\begin{array}{rcl} (y_i \mid \theta_i) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta_i)] \\ (\theta_i \mid G) & \overset{\text{iid}}{\sim} & G \\ G & \sim & \text{DP}(\alpha G_0), \\ G_0 & \equiv & N(\mu, \sigma^2) \\ (\alpha, \mu, \sigma^2) & \sim & p(\alpha, \mu, \sigma^2). \end{array} \tag{10}$$

Simulation: Random-Effects and Data Sets



Simulation data sets for control (C) and treatment (T) groups in more interesting **two-sample RCT** case ($n = 300$ 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).

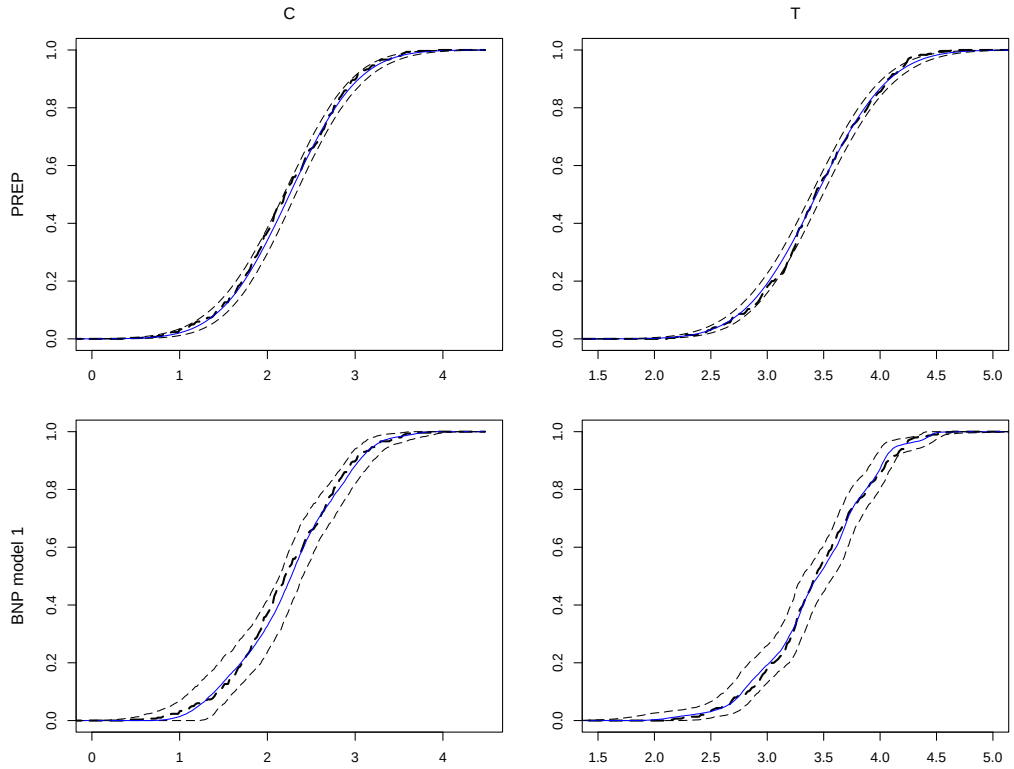
Predictive: PREP Versus BNP Model 1



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

The PREP model **can't 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**.

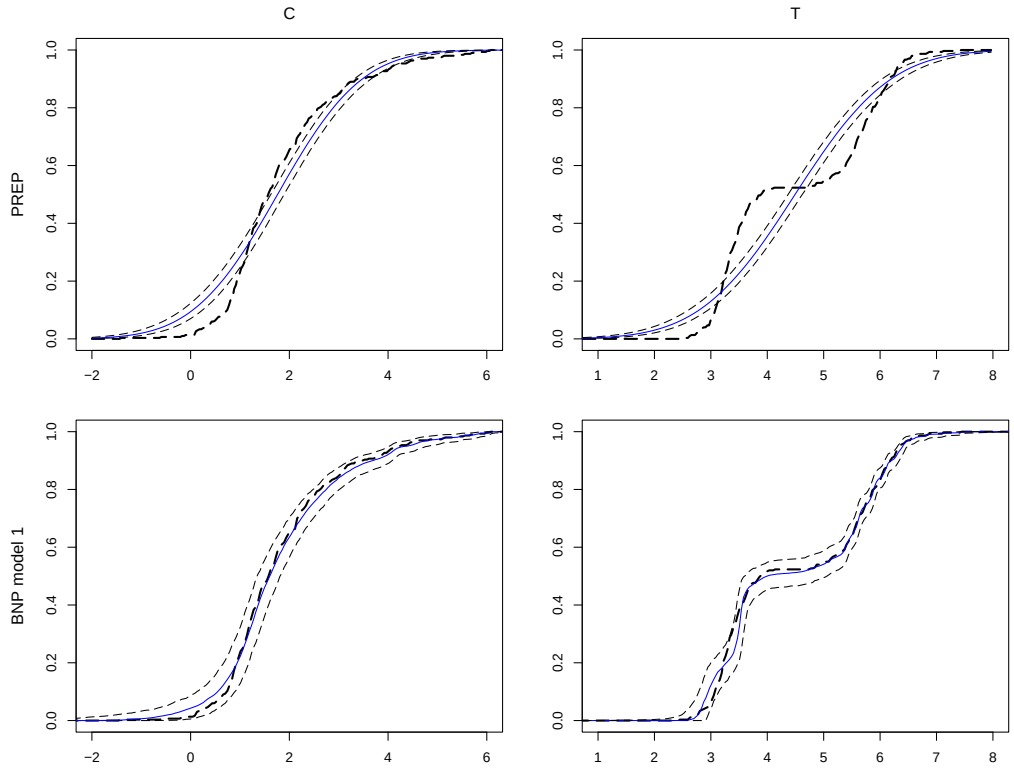
Normal Random Effects: PREP vs. BNP



Normal random effects (data set D_1): Posterior MCMC estimates of the **random effects distributions** G for PREP model (first row) and BNP model (second row); first column C, second column T.

When PREP is **correct** it (naturally) yields **narrower uncertainty bands**, but **direct comparison not fair** because PREP model typically arrived at via **data-analytic search on entire data set**.

Skewed and Bimodal Random Effects, PREP vs. BNP

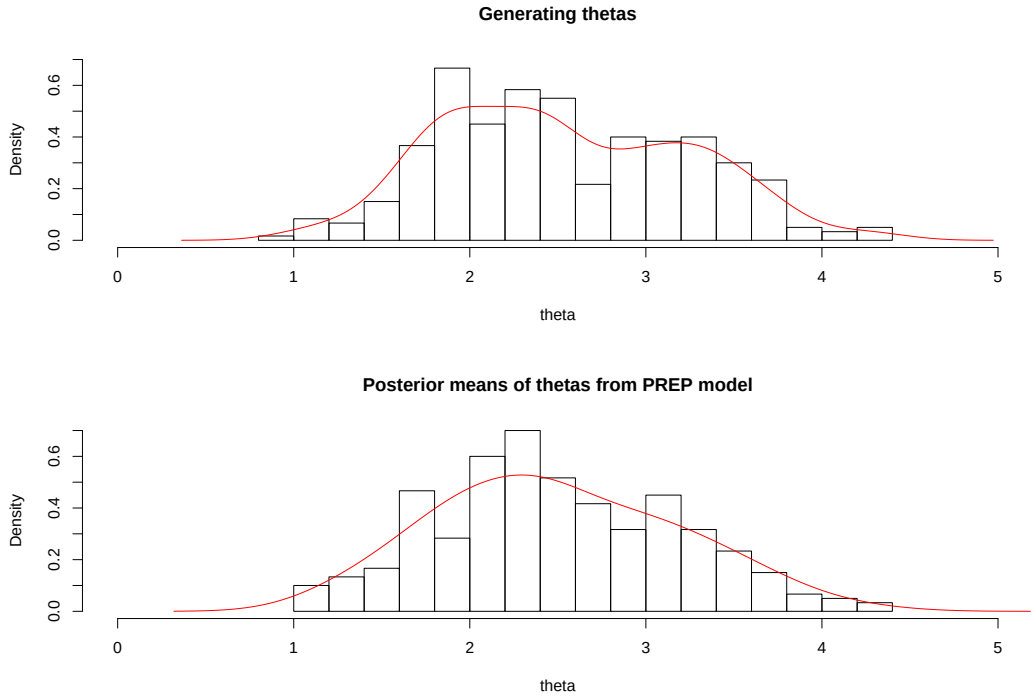


Skewed and **bimodal** random effects (data set D_2): Posterior MCMC estimates of **random effects distributions** G for PREP model (first row) and BNP model (second row); first column C, second column T.

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

A Parametric Pitfall

Warning: the **Gaussian assumption** on the **latent variables scale** in the **PREP model** can **make the model look plausible when it's not**:



Top panel: **bimodal mixture of Gaussians** as true latent-variable distribution of $\theta_i = \log(\lambda_i)$; bottom panel: **posterior means** of θ_i values from **PREP** model ($n = 300$ observations).

Diagnostic checking of PREP model would make it look appropriate when it's not; by contrast **BNP correctly picks up the bimodality**.

What Price Data-Analytic Model Specification?

One way to pay the right price for conducting a data-analytic search to arrive at a final parametric model — **three-way cross-validation** (3CV; Draper and Krnjajić, 2008): taking usual cross-validation idea one step further,

(1) Partition data at random into *three* (non-overlapping and exhaustive) subsets S_i, of size n_i (respectively).

(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.

(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.

Quantifying the Price of Model Uncertainty

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**).

Note that I'm advocating **holding back** n_3 observations in S_3 that are **not to be used** in summarizing inferential uncertainty about the main quantities of interest but are instead used to estimate **calibration** of the **entire data-analytic modeling process**.

How should the n_i be **specified**?

General idea for **quantifying the price of model uncertainty**:

(a) Bayesian **parametric models** are just **BNP models** with **stronger prior information** (example: **PREP** model takes $G \equiv N(\mu, \sigma^2)$ while **DP mixture model** takes $G \sim DP(\alpha G_0), G_0 \equiv N(\mu, \sigma^2)$), and **stronger prior information** often leads to **narrower uncertainty bands**; on this line of reasoning a **BNP model** would require **more data** (sample size n_{BNP}) to achieve the **same accuracy** as the **best-fitting parametric model** (sample size $n = n_{BP} < n_{BNP}$).

Quantifying the Price of Model Uncertainty (continued)

(b) This leads to the **recommendation**

$$n_3 = n\left(1 - \frac{n}{n_{BNP}}\right). \tag{11}$$

(c) Combining this with the usual **folklore cross-validation recommendation** that you should put **about twice as much data** in the **modeling** subset as in the **validation** subset yields

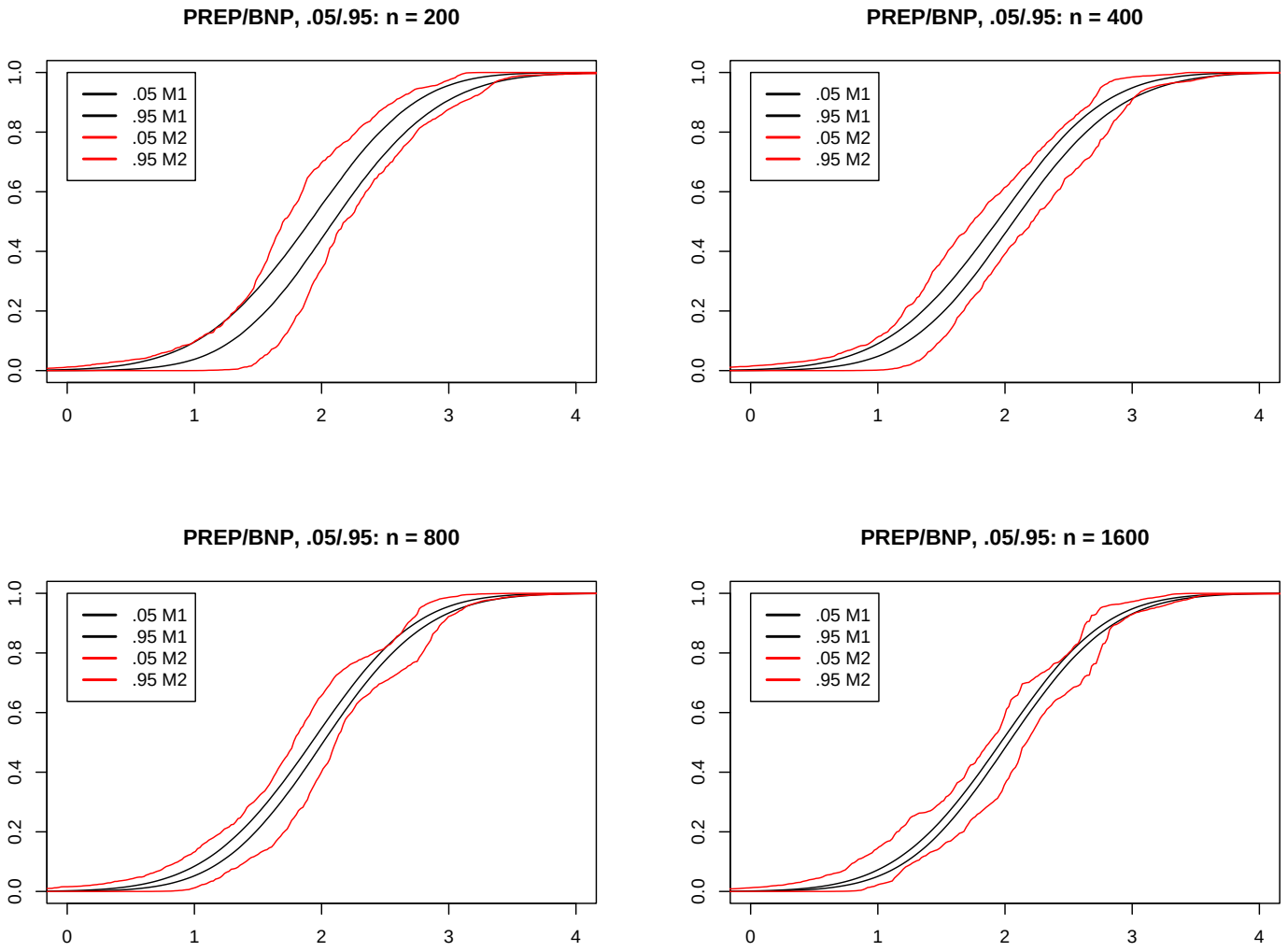
$$(n_1, n_2, n_3) = \left[\frac{2n^2}{3n_{BNP}}, \frac{n^2}{3n_{BNP}}, n\left(1 - \frac{n}{n_{BNP}}\right)\right]. \tag{12}$$

Example: With $n = \mathbf{1000}$ observations, if it takes about $n_{BNP} = \mathbf{1200}$ observations to achieve **BNP accuracy equivalent** to that of the **best parametric model** on the **main quantities of interest**, the subsets S_i should have about (**555, 278, 167**) observations in them.

Implementing this idea comes down to **estimating** n_{BNP}.

Milovan and I have been exploring this in the **PREP-PDPMM** context, and we've come up with some **surprising findings**.

Learning About G



Data-generating mechanism: **PREP** model; sample sizes **doubling** from **200** up to **1,600**; **black** **(red)** lines identify **5%** and **95%** points of posterior on G from fitting **PREP** **(PDPMM)** model.

Learning About G (continued)

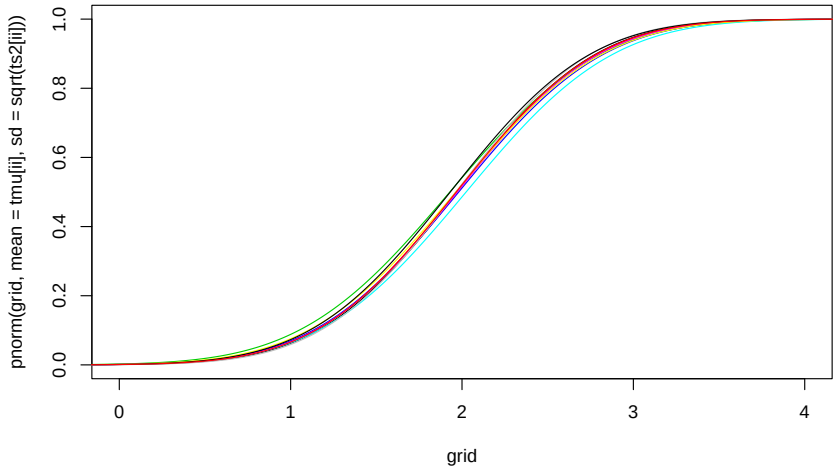
To **quantify the effect of sample size**, we computed the **areas** between the **0.05** and **0.95** pointwise quantiles of the CDF G along with the corresponding **maximum differences** between the two quantiles:

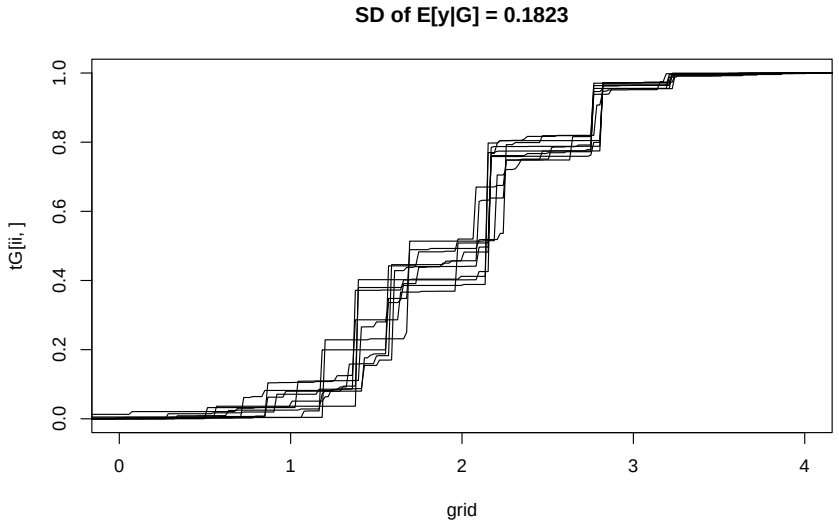
	PREP		PDPMM	
n	Area	Maximum Difference	Area	Maximum Difference
200	0.2256	0.11510	0.5556	0.3910
400	0.1608	0.07992	0.4788	0.2763
800	0.1122	0.05827	0.4195	0.2745
1600	0.0786	0.04002	0.3849	0.2445

You can see that the **PREP** model **learns** about G at a **substantially faster rate** than the **PDPMM** model; in fact, the **PREP learning rate** follows a **square root law** but the **PDPMM rate appears not to**.

Of course if the **data-generating mechanism** was **non-PREP** the **PREP model** would continue to **"learn" the wrong CDF** at a $\sqrt{n}$ rate, whereas the **PDPMM model** would (quite a bit more slowly) **learn the right** G.

Learning About $E(y|\text{data})$





However, **PDPMM** appears to learn about the **posterior mean on the data scale** at a **slightly faster rate** than **PREP**, at least for **small sample sizes**.

Learning About $E(y|\text{data})$ (continued)

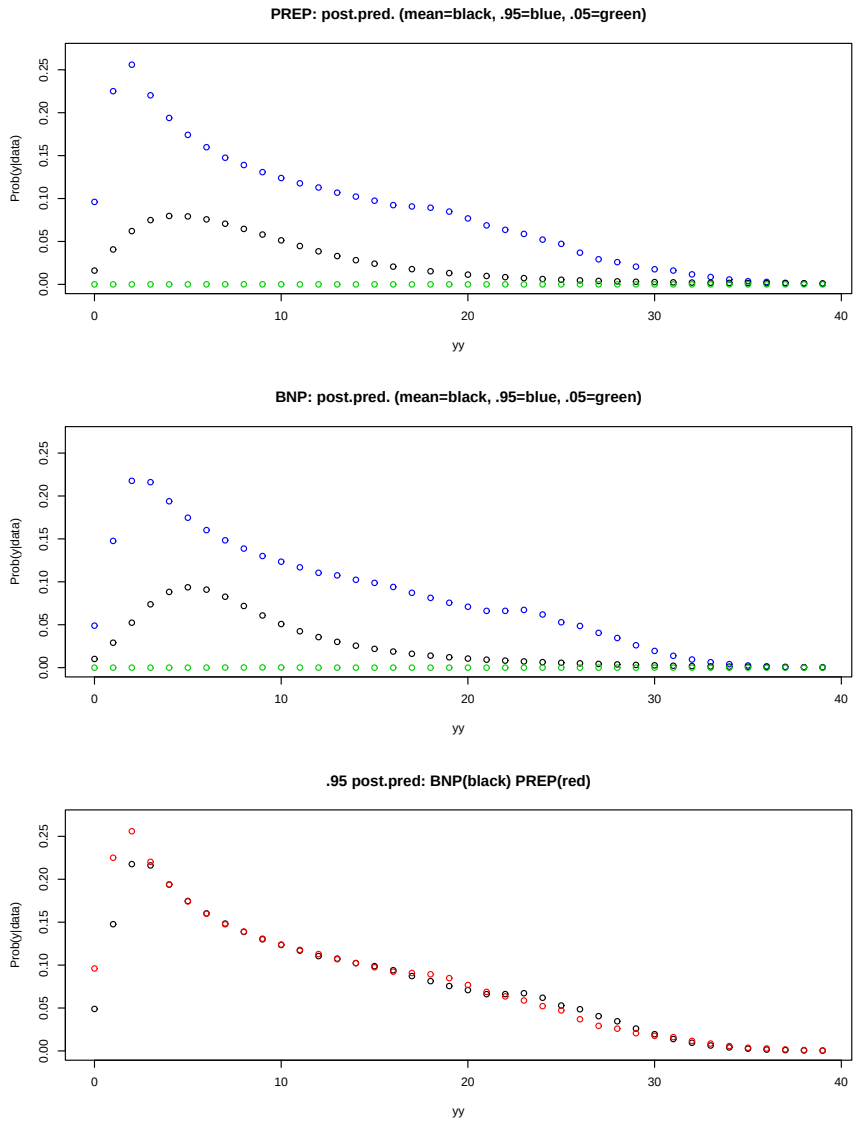
	90% Interval Width For $E(y\|\text{data})$	
n	PREP	PDPMM
200	1.793	1.679
400	1.251	1.179
800	0.800	0.795
1600	0.573	0.553

This is because the **standard MCMC estimate** of the **posterior mean on the data scale** is

$$u_j = \sum_{k=1}^{K} \exp(t_k) \left[G_j(t_k) - G_j(t_k^-) \right] \tag{13}$$

for each **MCMC iteration** j, where $\{t_1, \ldots, t_K\}$ is a **grid** of points at which $G_j(\cdot)$, the current MCMC iteration estimate of G, is evaluated; the **many flat spots** in G_j when the sample size is small can bring the **uncertainty assessment** for this estimate in **on the low side**.

Learning About Future Data



And on the **predictive scale**, results for the two models are **about the same**.

Comments

(1) The concept of **data equivalence** between **Bayesian parametric and nonparametric models** is **slippery**: the **answer** you get may well **depend on the scale on which you ask the question**.

(2) Some of what we've found may be **particular (peculiar?)** to the **DPMM approach** to implementing the idea of **placing a prior on CDF space**.

(3) Both the **NB10** t**–likelihood** model and the **PREP-PDPMM modeling of count data** provide examples of a fact that makes **Bayesian inference trickier** than it would otherwise be: **weaker prior information does not necessarily lead to weaker inferential conclusions**.

$$p(y|x) = \int p(y|x, M)\, p(M|x)\, dM.$$

$p_1(M)$ may be **weaker** (embody **more uncertainty**) than $p_2(M)$, and yet $p_1(M|x)$ may concentrate on models with **greater conditional inherent accuracy** $p(y|x, M)$ than those on which $p_2(M|x)$ concentrates, leading to **stronger inferential conclusions** from $p_1(y|x)$ than from $p_2(y|x)$.