Case Studies in Bayesian Data Science

3: Bayesian Non-Parametric Case Studies

David Draper

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

draper@ucsc.edu

Short Course (Day 5)
University of Reading (UK)

27 Nov 2015

users.soe.ucsc.edu/∼draper/Reading-2015-Day-5.html

Pólya Tree Case Study

A key issue in the consolidation process of the nuclear fuel cycle is the **safe disposal of radioactive waste**.

At present, deep geological disposal based on a multibarrier concept is considered the most promising option (visualize a **deep underground chamber** within which radioactive materials such as spent fuel rods are entombed in layers of concrete and other barriers;
e.g,, PSAC User Group, 1989).

The **safety** of this concept ultimately relies on the safety of the mechanical, chemical and physical barriers offered by the geological formation itself.

In spite of recent worldwide efforts, the physico-chemical behavior of such a disposal system over geological time scales (hundreds or thousands of years) is **far from known with certainty** (e.g., Sinclair, 1996).

<u>**Goal**</u>: Predicting outcomes, including **radioactive dose** for people on the earth's surface, as a function of factors like time, how far the disposal chamber is underground, ...

Uncertainty

Radioactive dose is estimated by **computer simulation models** such as AEA's `MASCOT`, which numerically solve complex systems of partial differential equations.

The output of such models is deterministic given fixed **scenario** and **parametric** inputs, but these are uncertain. **Structural** and **predictive** uncertainty are also part of a full uncertainty audit (Fig. 1; Draper, 1997).

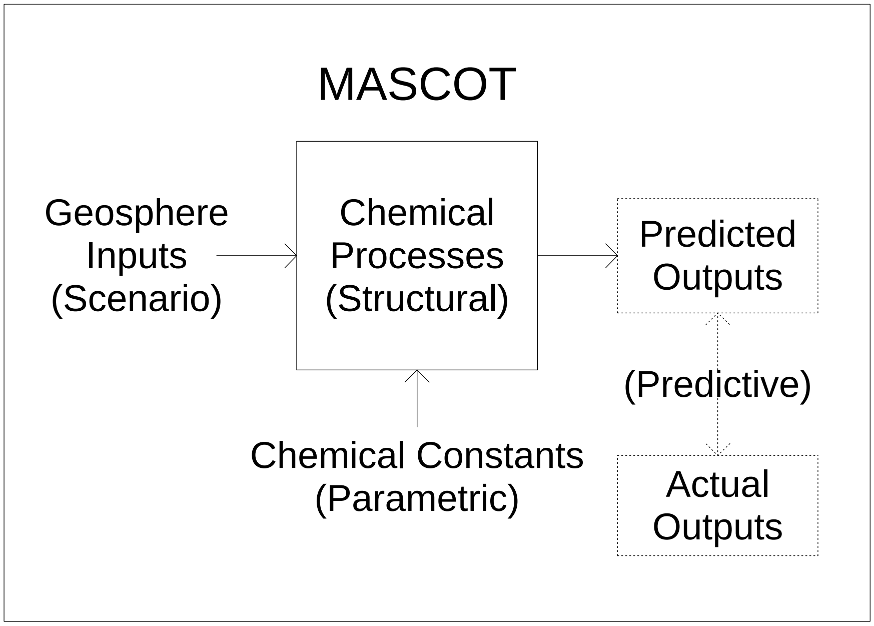

MASCOT
Geosphere Inputs (Scenario)
Chemical Processes (Structural)
Predicted Outputs
(Predictive)
Chemical Constants (Parametric)
Actual Outputs

Figure 1. *Illustrating the four potential sources of uncertainty in stochastic modeling of radioactive dose with programs like* `MASCOT`.

Parametric uncertainty is typically quantified with **probability distributions** across all the model inputs: the program is run N times, with different stochastically generated inputs each time, obtaining N dose estimates at each of T time points.

Focus on the Mean

Regulatory bodies insist on summarizing the dose distribution f at a given time point by its $\boxed{\textbf{mean } \theta = \int y f(y)\, dy}$ (even though this may be a very unstably estimated quantity; Sinclair and Robinson, 1994).

Technical challenge: f is typically **extremely (positively) skewed**, with many zeros and a few comparatively huge values, and the number of Monte Carlo repetitions N is constrained by time and money (often $\leq 10,000$, sometimes $\leq$ 500–1000).

With relatively small N, the concern is that **you haven't seen all of the right tail yet**.

Problem statement in contract proposal:

To develop an improved understanding of the issue of convergence of probabilistic safety assessment (PSA) calculations, together with specific algorithms that could underlie improved analysis of statistical errors associated with estimating **mean values** *or other statistical performance measures, in the context of risk assessments for long-term safety studies in radioactive waste disposal.*

The Problem (continued)

Recent elaboration by Jim Sinclair:

Given a set of observations,
what can I say *about the true mean?*

Is the internal evidence of my sample distribution sufficient for me to quote a **best estimate** *and some* **interval limits***?*

Is the evidence such that I can say **I shouldn't even be estimating a mean unless I get many more samples***?*

What kind of **external information** *about the distribution, such as knowledge of its general shape, or something like an upper bound, could improve my ability to predict the mean?*

Time permitting, can similar questions be answered about **other statistics***, such as various percentiles of the distribution?*

Can the extent to which, say, the **99th percentile** *is more robustly predictable than the mean be quantified?*

An Example of the Data

Consider $N = 10,000$ dose values from `MASCOT` at $t = 100$ years, based on a scenario permitting relatively large doses of **Strontium 90 (Sr–90)** with relatively low probability. The outcome examined is **total dose** from three nuclides including Sr–90.

9864 (98.6%) of the 10,000 values are **0**; 134 of the other 136 (1.36%) range smoothly from 1.059e–14 to 8.552e–01; the two largest values are 3.866 and **189.3** (!). The sample mean is 0.01964. (The true mean at 100 years, obtained from another program, AEA's `ESCORT`, is **9.382e–4** (21 times smaller); the sample mean omitting the largest observation is 7.138e–4.)

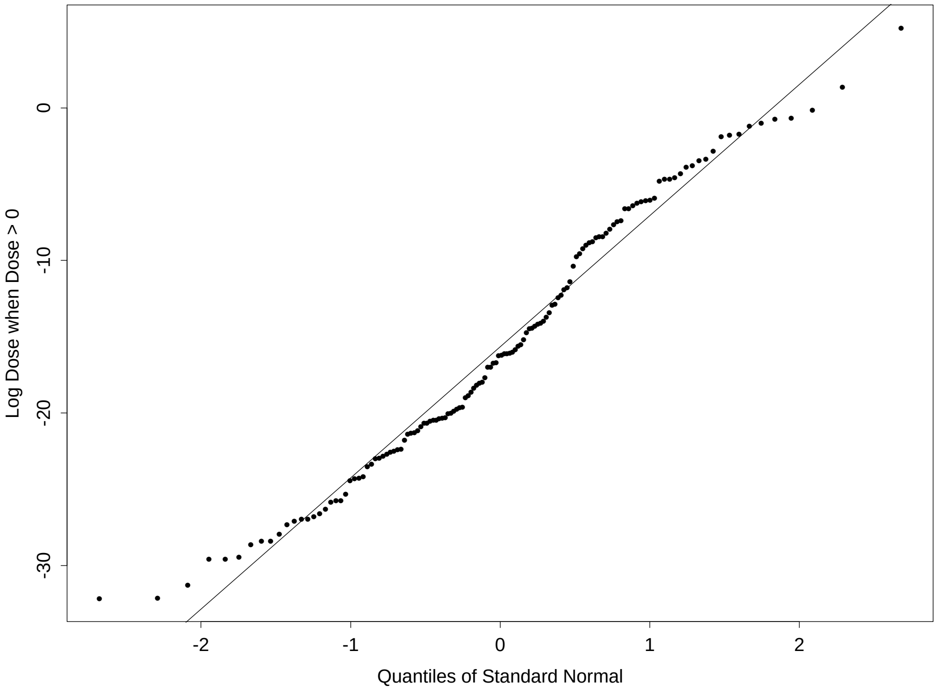


Figure 2. *A normal quantile-quantile plot of the positive log dose values (the line shows ideal behavior if Gaussian).*

Method 1: Naive Frequentist Nonparametrics

This distribution fairly closely follows a **two-part** or **mixture** model, in which each observation is 0 with probability p and **lognormal** with probability $(1-p)$ (Fig. 2).

Method 1: Central Limit Theorem (CLT). We are trying to estimate θ, the population mean. Why not use $\widehat{\theta}_1 =$ the sample mean?

Dose values $D_i, i = 1, \dots, N$;

$$\text{Point estimate} \quad \widehat{\theta}_1 = \bar{D} = \frac{1}{N}\sum_{i=1}^{N} D_i. \qquad (1)$$

The standard (frequentist) **interval estimate** is based on the hope—with such a large N—that the distribution of $\widehat{\theta}_1$ in repeated sampling from the population density f is close to normal (by the **Central Limit Theorem**):

$$\text{95\% Interval estimate} \quad \widehat{\theta}_1 \pm 1.96 \frac{s_D}{\sqrt{N}}, \qquad (2)$$

where s_D is the sample standard deviation $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(D_i - \bar{D})^2}$.

This is a **nonparametric** method, because no assumptions about f are used (except that its variance is finite).

Method 1: CLT (continued)

Here $\widehat{\theta}_1 = \bar{D} = 0.01964$, $s_D = 1.893$, and the 95% interval estimate is $0.01964 \pm 1.96 \frac{1.893}{\sqrt{10000}} = (-0.01746, 0.05675)$, which does include the true value 9.382e–4 but has made itself look silly in doing so by going **negative**. ("Guttman" multiplier 2.68 [Woo, 1989] just makes this problem worse.)

Moreover the largest observation occurred at iteration number 6132, and many of the CLT intervals based on observations 1–k for $k < 6132$ fail to cover: only **63%** of the 100 "95%" intervals based on observations 1–100, 1–200, ... include the true value.

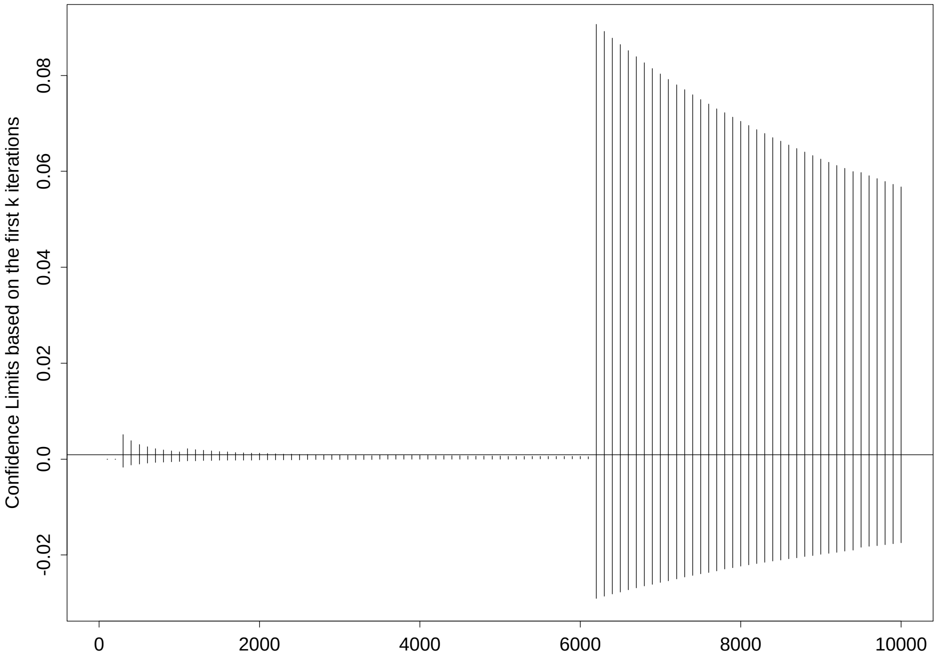


Figure 3. *Upper and lower 95% Central Limit Theorem intervals, with the true mean superimposed, based on observations 1–100; 1–200; ..., 1–10,000. 73 of these 100 intervals go negative.*

Method 1: Simulation Results

I regarded the 10,000 dose values at 100 years as a **population to be sampled with replacement**. I repeatedly ($S = 5,000$ times) took samples of size N from this population, with N varying from 100 to 1,000,000, and computed the **actual coverage** of nominal "95%" intervals from the CLT method, with results as shown in Table 1.

Table 1. *Actual coverage of nominal 95% intervals based on the Central Limit Theorem, as a function of sample size N (simulation standard errors in parenthesis; results for the 2.68 multiplier were only slightly better).*

Sample Size (N)	Coverage (1.96 Multiplier)	% Left Endpoint Negative	Average Length
100	.0296 (2.41e–3)	74.7	.0914 (1.14e–2)
1,000	.0976 (4.20e–3)	99.3	.0757 (3.17e–3)
5,000	.391 (6.90e–3)	81.4	.0656 (1.16e–3)
10,000	.622 (6.86e–3)	76.5	.0565 (6.53e–4)
25,000	.878 (4.63e–3)	67.9	.0434 (2.55e–4)
50,000	.909 (4.07e–3)	25.9	.0324 (1.13e–4)
100,000	.930 (3.61e–3)	0.8	.0231 (5.32e–5)
500,000	.950 (3.08e–3)	0.0	.0105 (2.12e–5)
1,000,000	.945 (3.22e–3)	0.0	.00742 (5.28e–6)

For $N \leq 10,000$ mistakes were always from the interval lying **entirely to the left** of the true mean. 24.1% of the data sets with $N = 100$ consisted of **all zeros**, but this dropped to 0% for $N \geq 500$.

Failure of the CLT

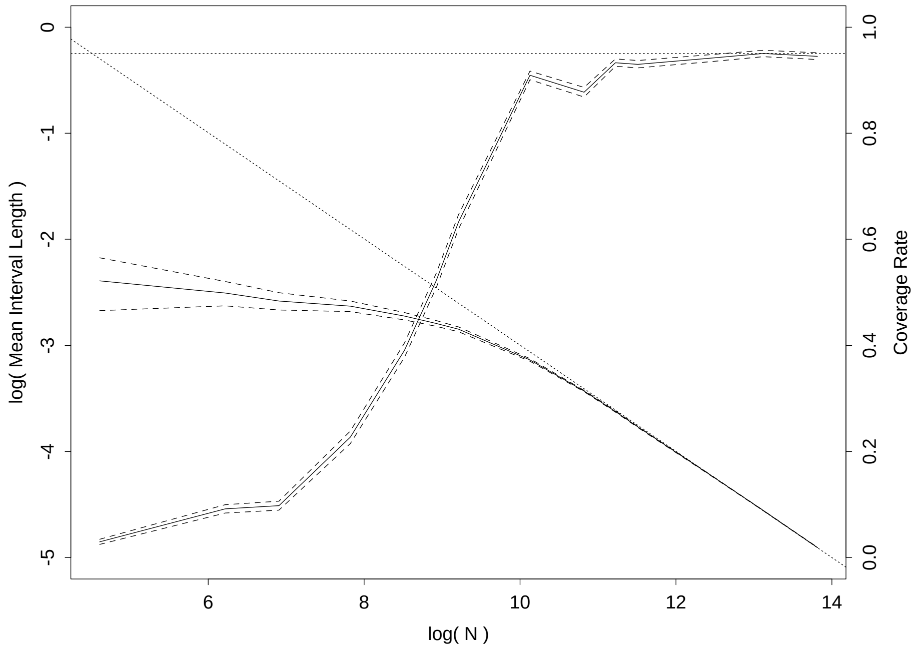


Figure 4. *Log mean interval length and coverage rate against* $\log(N)$ *for the CLT intervals.*

$\log$(interval length) vs. $\log(N)$ should be **linear** if the sample SD is doing its job properly: length $= 3.92 \cdot s_N/\sqrt{N}$, so for N large enough that $s_N \doteq \sigma = 1.893$, $\log(\text{length}) = 2.004 - 0.5 \log(N)$; but the actual curve (Fig. 4) **does not approximate this line** until $N > \exp(10) \doteq 22,000$.

Not coincidentally, that is just about exactly where the CLT starts producing **decent performance**: coverage rate against $\log(N)$ shows an ogive shape that does not exceed (say) 0.9 until N is also roughly 22,000 or more.

Failure of the CLT

The reason this method performs so poorly is that even with (say) 7,500 observations going into each average, the distribution of the sample mean is far from Gaussian (Fig. 4.1), because of the **extreme skewness** of the population.

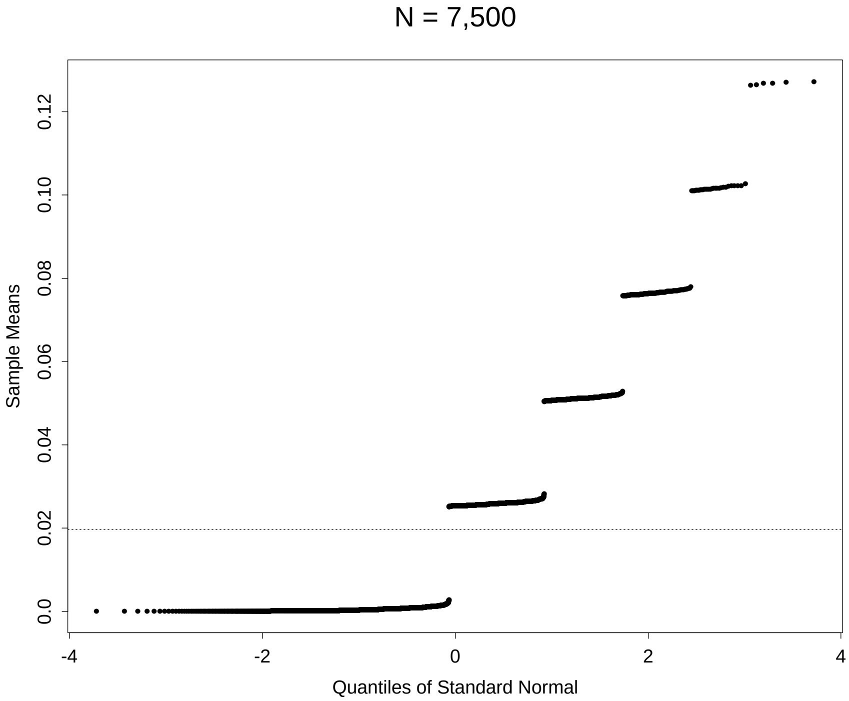


Figure 4.1. *Normal quantile-quantile plot of the 5,000 simulated means in Table 1, each based on* $N = 7,500$ *observations.*

CLT Nightmare

In fact with this population you don't even begin to get a really decent normal approximation to the mean until $N \geq$ **100,000**:

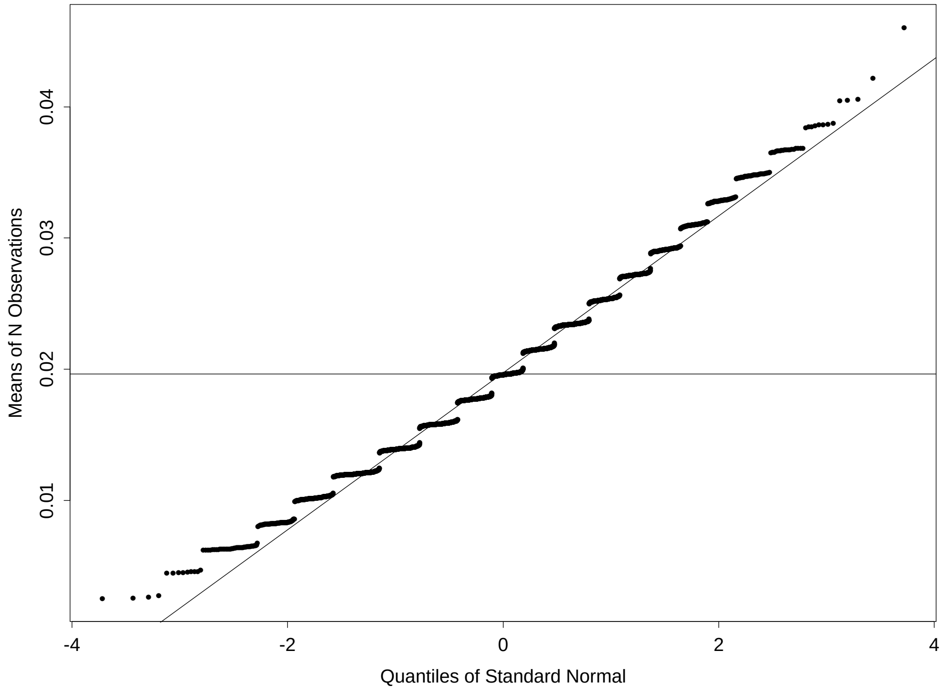


Figure 5. *Normal quantile-quantile plot of 5,000 simulated means, each based on* $N = 100,000$ *observations.*

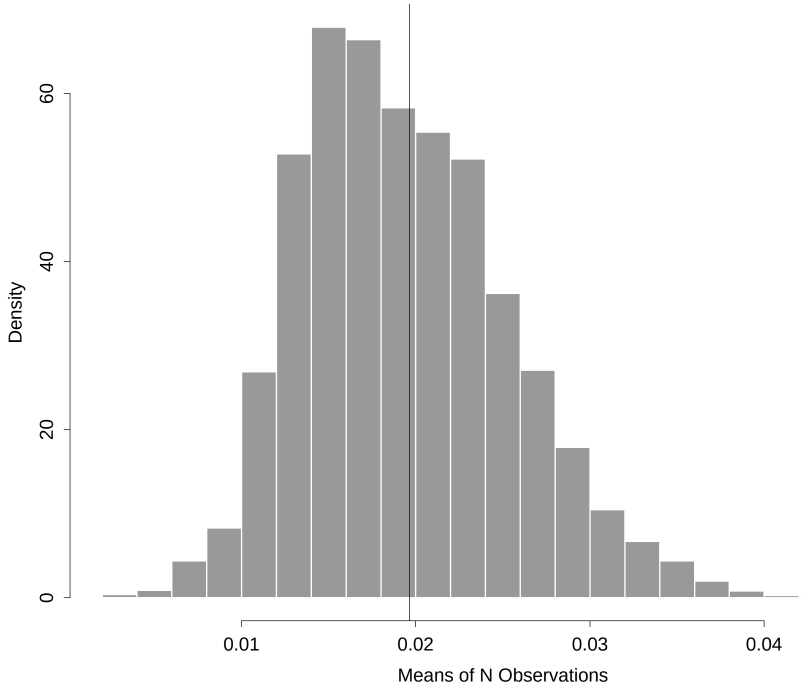


Figure 6. *Histogram of the 5,000 simulated means in Fig. 5.*

Method 2: Less Naive Frequentist Nonparametrics

The best frequentist nonparametric confidence interval technology to date is the **BC$_a$** method, based on the **bootstrap** (Efron and Tibshirani, 1993).

The bootstrap idea in this context is to repeatedly ($B = 1,000$ times, say) choose a sample of size N **with replacement** from $D_1, \ldots, D_N$, calculate the means of each of these samples, and use the distribution $\mathcal{D}$ of these B means as the basis for confidence intervals.

The **percentile** method is literal: to produce a 95% interval, choose the $\alpha_1 = 2.5\%$ and $\alpha_2 = 97.5\%$ points of $\mathcal{D}$.

This method works OK with large samples from reasonably "standard" data sets (not too far from Gaussian), but can produce **poor coverage** for small N and with (e.g.) highly skewed data.

The BC$_a$ method improves on the percentile method by choosing different α_1 and α_2 values which yield **better (closer to nominal) coverage**.

The Bootstrap

This method also makes **no use** of information about f apart from assuming its variance is finite.

With the $N = 10,000$ dose values at 100 years the BC$_a$ method yields the nominal 95% interval (3.64e–4, 0.134), which also includes the true mean; moreover, BC$_a$ intervals are **incapable of going negative**.

However, in the analogue of Fig. 3 for the bootstrap (Fig. 4), it is still true that only **85%** of the nominal 95% intervals include the truth.

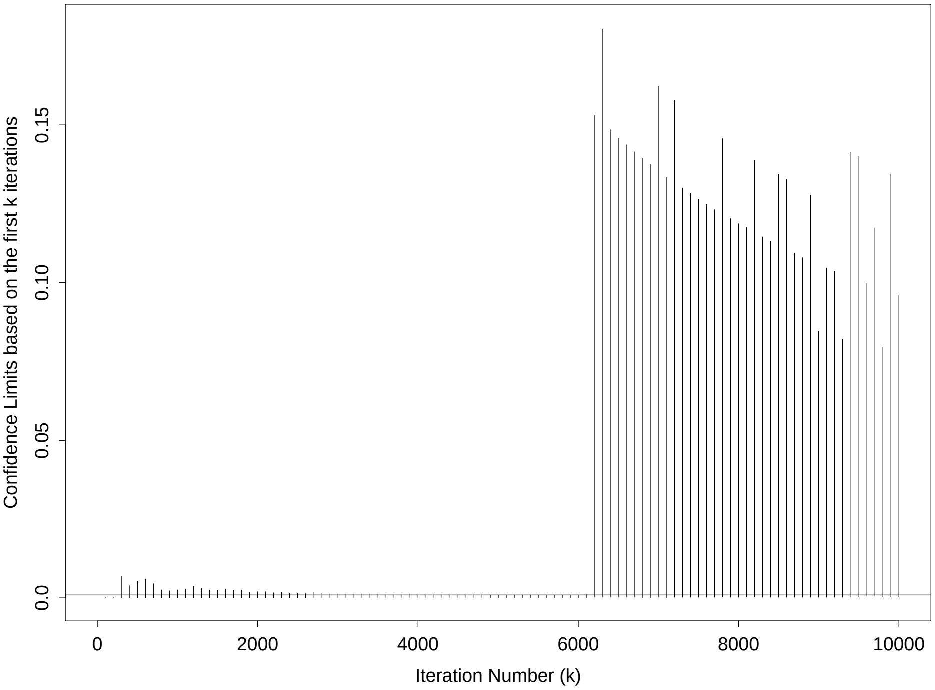


Figure 7. *Upper and lower 95% bootstrap intervals, with the true mean superimposed, based on observations 1–100; 1–200; ..., 1–10,000. None of these intervals go negative.*

Bootstrap Results

Performance in the analogue of Table 1 is a bit better, but still **pretty bad**:

Table 2. *Actual coverage of nominal 95% intervals based on the bootstrap, as a function of sample size N (simulation standard errors in parenthesis; number of simulation repetitions $S = 1000$ except where otherwise indicated).*

Sample Size (N)	Actual Coverage	Mean Length
100	.0373 (2.68e–3)	.0695 (9.89e–3)
500	.100 (4.24e–3)	.0727 (4.42e–3)
1,000	.0938 (4.12e–3)	.0689 (2.99e–3)
5,000	.394 (6.91e–3)	.0655 (1.16e–3)
10,000	.632 (6.82e–3)	.0582 (6.65e–4)

(**No point** in continuing the simulations: very similar to Table 1.)

Mistakes for small N were again always from the interval lying **entirely to the left** of the true mean. These intervals cover slightly more often than the CLT intervals, and are slightly narrower.

However the coverage is still abysmal, and the BC$_a$ method is **slow**: it took 25 minutes of CPU time to do the calculations leading to Fig. 5, and **51 hours** of CPU time to produce Table 2, on an otherwise unburdened 333Mhz `DECalpha` workstation.

Method 3: Parametric Bayesian

The data analysis on p. 6 above suggests the following **mixture model**.

At a particular time t, let D_i be the observed dose on simulation run $i = 1, \ldots, N$. Then

$$D_i = \left\{ \begin{array}{ccc} 0 & \text{with probability} & p \\ LN(\mu, \sigma^2) & \text{with probability} & (1 - p) \end{array} \right\}, \quad (3)$$

where $LN(\mu, \sigma^2)$ denotes the lognormal distribution with mean μ and standard deviation σ **on the log scale**.

In this model the true population mean θ is given **theoretically** (Johnson and Kotz, 1970) by

$$\theta = (1 - p) e^{\mu + \frac{1}{2}\sigma^2}. \quad (4)$$

In a Bayesian formulation **prior distributions** are needed for the parameters p, μ, and σ (or σ^2, or the **precision** $\tau = \frac{1}{\sigma^2}$).

I have so far used **diffuse** priors that are relatively **flat** in the regions of high likelihood for the parameters.

With $N = 10,000$ observations this is reasonable; with $N \leq 1,000$ (say) the priors would probably need to be **more informative**.

MCMC

One simple initial idea for fitting the mixture model (3): **Gibbs sampling** via $\boxed{\texttt{BUGS}}$ (Spiegelhalter et al., 1997).

Unfortunately BUGS cannot fit model (3) above, but it can fit a **functionally equivalent** model:

$$D_i = \left\{ \begin{array}{lll} LN(\mu_1, \sigma_1^2) & \text{with probability} & p_1 \\ LN(\mu_2, \sigma_2^2) & \text{with probability} & p_2 = (1 - p_1) \end{array} \right\}, \tag{5}$$

where the **zeros** in the data set are replaced by values of the form (ϵ ± tiny lognormal noise) to correspond to the first component of the mixture.

In this model the **underlying mean** of the distribution of the D_i is **theoretically**

$$\theta = p_1 \exp\left(\mu_1 + \frac{1}{2}\sigma_1^2\right) + p_2 \exp\left(\mu_2 + \frac{1}{2}\sigma_2^2\right). \tag{6}$$

With the $N = 10,000$ observations of dose at $t = 100$ years examined on p. 6, I used initial values that permitted a **short burn-in** (100 iterations):

```
list( mu = c(-45.03454, NA ), eta = 29.31891,
  p = c( 0.0136, NA ),tau = c( 0.2522539, 0.01351527 ) )
```

An Example BUGS Program

```
model d100.2;

const

  N = 10000, mu.mu1.p = 0.0, tau.mu1.p = 1.0E-6,
    mu.eta.p = 0.0, tau.eta.p = 1.0E-6,
    epsilon = 0.001, kappa = 0.437;

var

  d[N], mu[2], eta, alpha[2], p[2], tau[2], T[N],
    theta;

data d in "d100.gibbs.dat";
inits in "d100-2.in";

{

  mu[1] ~ dnorm( mu.mu1.p, tau.mu1.p );
  mu[2] <- mu[1] + eta;
  eta ~ dnorm( mu.eta.p, tau.eta.p ) I( 0, );
  alpha[1] <- 1;
  alpha[2] <- 1;
  p[] ~ ddirch( alpha[] );
  tau[1] ~ dgamma( epsilon, epsilon );
  tau[2] ~ dgamma( epsilon, epsilon );

  for ( i in 1:N ) {

    T[i] ~ dcat( p[] );
    d[i] ~ dlnorm( mu[ T[i] ], tau[ T[i] ] );

  }

  theta <- p[1] * exp( mu[1] + 0.5 / tau[1] ) +
    p[2] * exp( mu[2] + 0.5 * kappa / tau[2] );

}
```

Extreme Behavior of Lognormal Model

With values for the parameters similar to those in the dose data at 100 years (with 0's changed to lognormal draws centered at a very small positive value) ($\mu_1 = -45.0, \sigma_1 = 1.99, \mu_2 = -15.7, \sigma_2 = 8.60, p = 0.986$), formula (6) gives a shock: θ comes out **2.46e+7**!

To look at this from another angle, I repeatedly (10,000 times) sampled **10,000** draws from model (5) and calculated the mean of these draws.

The smallest of the 10,000 means was 5.33e–6, and their median was 0.0787; but their mean was **1.36e+4**, and the biggest was **9.88e+7**!

The problem is that the lognormal distribution is extremely sensitive to assumptions about the **rate at which the tails fall off toward 0** in the normal distribution.

With a mean of –15.7 and an SD of 8.60 on the log scale, the median observation on the dose scale would be 1.52e–7, but **one time in 10,000** you would get a value like $\exp(-15.7 + 3.72 \cdot 8.60)$ = **1.21e+7** (which contributes .0001· 1.21e+7 = 1.21e+4 to the mean), one time in 100,000 you would get something like $\exp(-15.7 + 4.26 \cdot 8.60)$ = 1.32e+9 (which adds another 1.32e+4), and so on.

Parametric Bayesian Results

One possible fix is to use a **truncated** lognormal model: in the second component, $\log(D_i) = \mu_2 + \sigma_2 e_i$, with $e_i \sim N(0, 1)$ truncated at $-A$ and A (or just at A). Then for this component of the mixture $V[\log(D_i)] = \kappa\,\sigma_2^2$ with

$$\kappa = 1 - \frac{2\Phi^{-1}(1-\gamma)\,\phi[\Phi^{-1}(1-\gamma)]}{1-2\gamma}, \qquad (7)$$

where $\gamma = \Phi(-A)$.

Table 2.5. *Rough estimates in the truncated lognormal mixture model as a function of the number of points k set aside in each tail.*

k	$\widehat{\mu}_2$	$\widehat{\sigma}_2$	$\widehat{\theta}$
0	-15.68	8.601	24,086,800.
1	-15.71	8.352	585,834.
2	-15.71	8.157	50,485.
3	-15.71	7.988	6,739.
4	-15.72	7.843	1,176.
5	-15.73	7.692	248.8
6	-15.74	7.539	60.89
8	-15.77	7.250	5.086
10	-15.79	6.952	0.5939
12	-15.82	6.695	0.08936
14	-15.83	6.434	0.01645

To bring θ in line with the true mean, $k = 14$ (about $\gamma =$ **10%** in each tail) corresponds to $\kappa \doteq 0.43$.

Parametric Bayesian Results

Figs. 8–11 and Table 3 present **exploratory results** with this model on the modified 100-year dose data, using a burn-in of 500 and a monitoring run of 5000 draws (this took 4.5 hours of CPU time at 333Mhz).

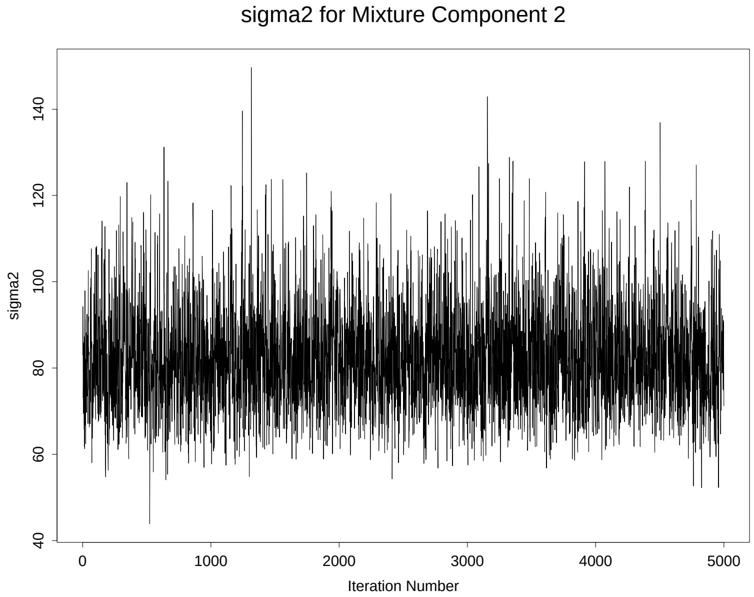


Figure 8. *Time series trace for σ_2^2, showing a bit of positive serial correlation (0.26) but not enough to be worrisome (the traces for the other parameters are similar).*

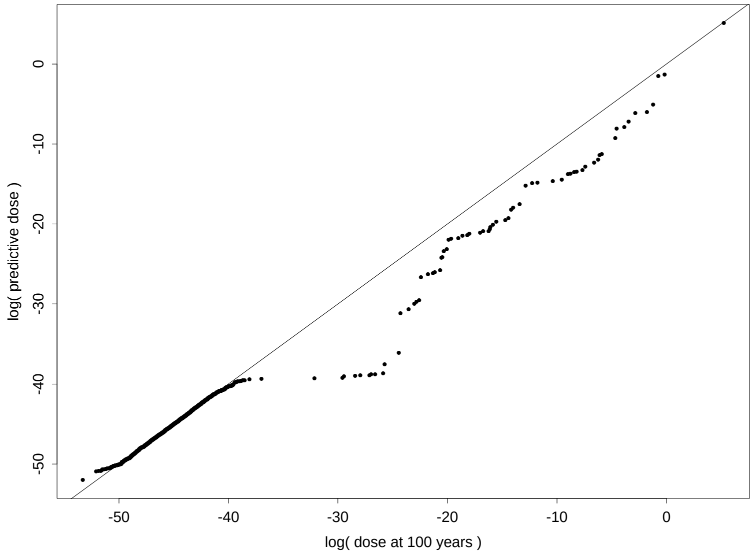


Figure 8.1. *qqplot of* log(*predictive dose*) *against* log(*actual dose*) *at 100 years.*

Bayesian Results (continued)

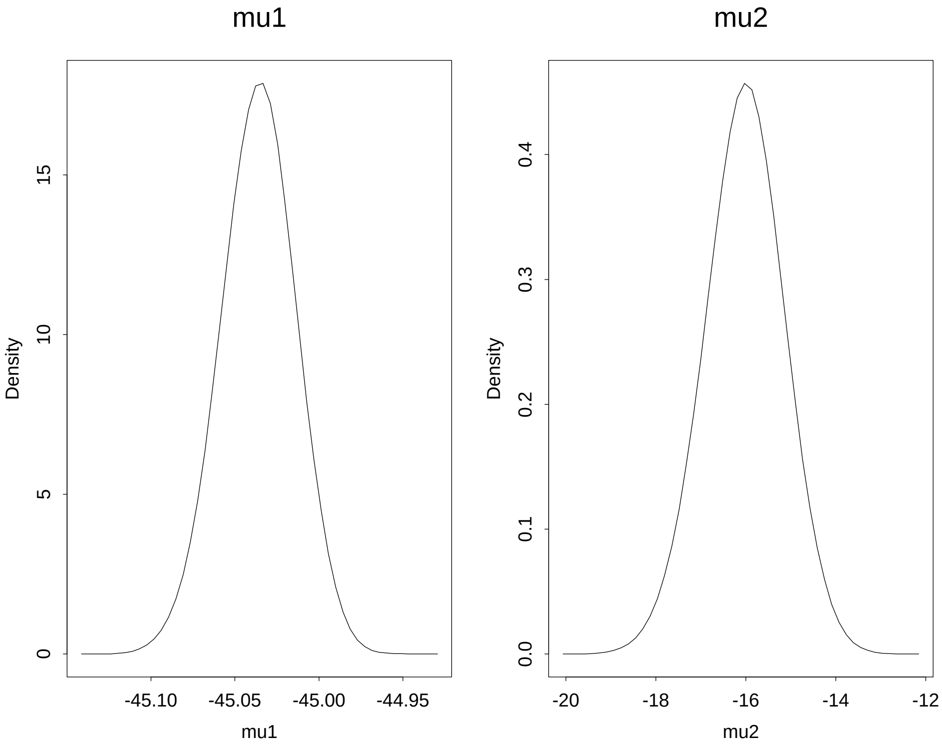


Figure 9. *Density traces of the marginal posterior distributions of μ_1 and μ_2, both of which are not far from normal.*

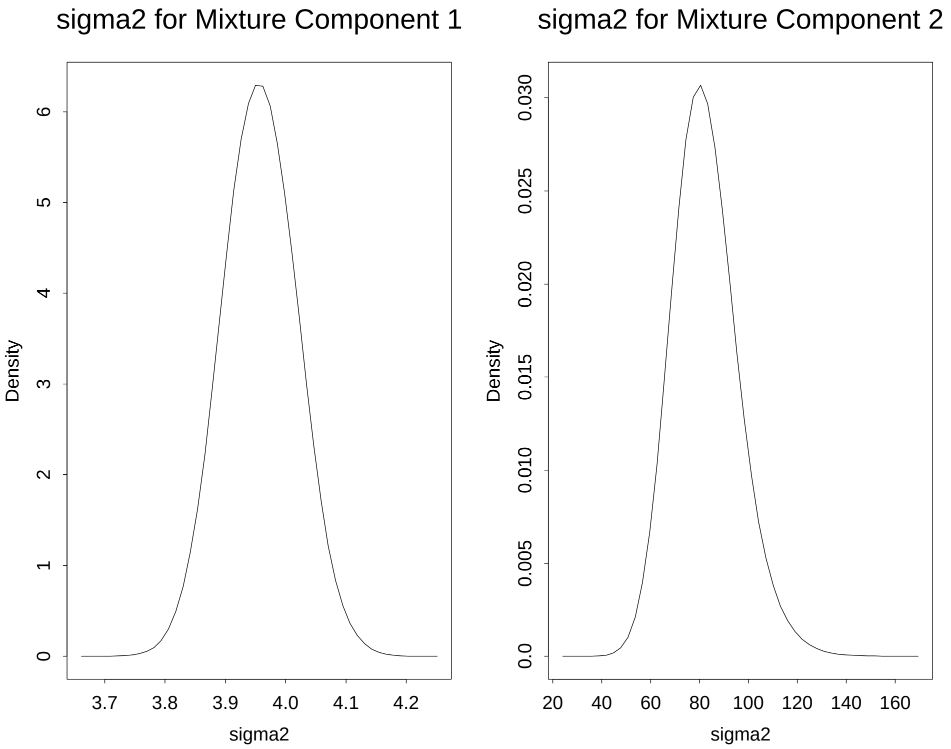


Figure 10. *Density traces of the marginal posterior distributions of σ_1^2 and σ_2^2, both of which exhibit the sort of skewness you would expect for variances.*

Bayesian Results (continued)

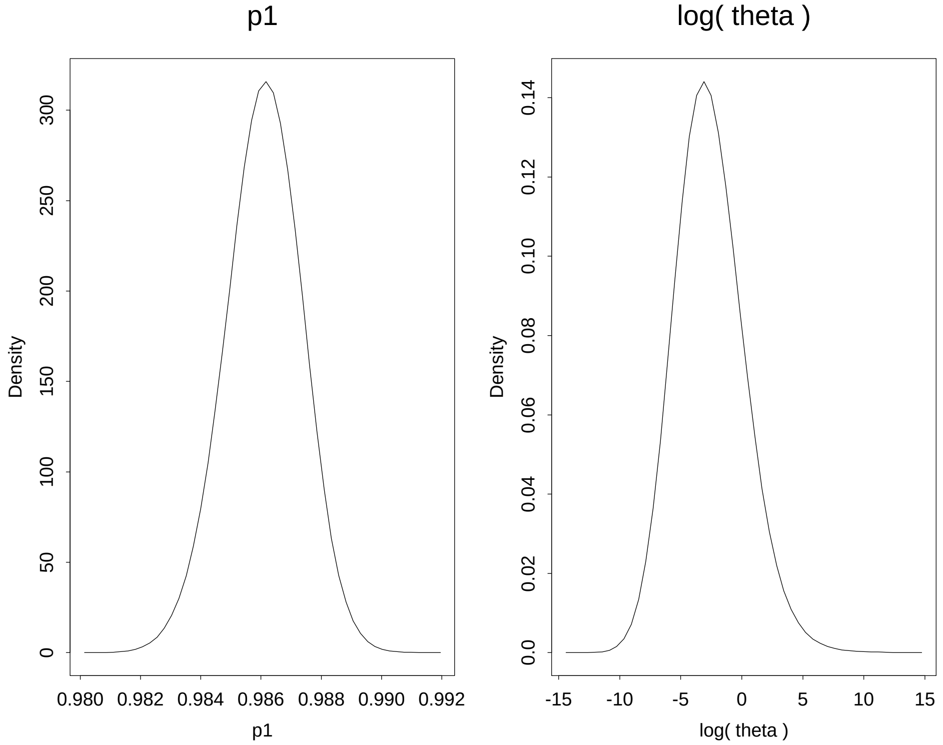


Figure 11. *Density traces of the marginal posterior distributions of* p_1 *and* $\log(\theta)$*.* **NB** θ*'s distribution is even heavier-tailed than lognormal.*

Table 3. *Numerical summaries of the posterior distributions for the parameters of model (5) with the 100-year dose data (*$\kappa = 0.43$*).*

	Posterior		Maximum Likelihood	
Variable	Mean	SD*	Estimate	SE**
p_1	0.986	1.16e–3	0.986	1.16e–3
μ_1	–45.0	0.0203	-45.0	0.0199
σ_1^2	3.96	0.0563	3.96	0.0563
p_2	0.0139	1.16e–3	0.0139	1.16e–3
μ_2	–16.0	0.812	-15.7	0.738
σ_2^2	82.5	11.9	74.0	10.0
θ	28.0	959 (!)	0.076	0.205

* SD = standard deviation ** SE = standard error

NB median(θ) = 0.058, q_{85} = 1.0, max = 47728 (!), 95% central interval = (9.35e-4, 19.2); 95% maximum likelihood CI = (3.92e-4, 14.7)

Method 4: Bayesian Nonparametrics

Method 3 coverage properties. I repeatedly (100 times) drew samples of size 10,000 with replacement from the modified 100-year dose data and used `BUGS` to construct 95% interval estimates of the true mean 0.0196 (based on the 2.5% and 97.5% points of the simulated posterior distributions).

With $\kappa = 0.43$, actual coverage was **89%** (with simulation SE 3.1%), but the intervals were extremely long (median length 238 (!), mean length 409 (!)).

Tentative conclusion: Intervals probably still wider than necessary for decent coverage. How well will this method work with small N (500, say)? Still have to actually implement truncated lognormal idea instead of κ approximation. (Work in progress.)

$\boxed{\textbf{Method 4: Bayesian Nonparametrics}}$. A sample of size $N = 1000$ from the 100-year dose data would only be expected to have about **14** values from the non-zero part of the distribution.

Clearly with such samples it would be necessary to **teach** the interval-generating process about the right tail, above and beyond what it can learn directly from the data (the lack of such a way to learn is why the bootstrap fails).

Bayesian Nonparametrics (continued)

One approach: the parametric Bayesian Method 3. But the lognormal is only **approximately "correct"** at $t = 100$ years, and the approximation may well be even more vague at other times and for other scenarios.

It would be good to be able to build a model that is **centered** at the lognormal, but which can adapt to other distributions when the data suggest this is necessary.

Continuing Part 3, a fairly recently developed modeling approach based on **Pólya trees** (Lavine, 1992, 1994; Walker et al., 1998), first studied by Ferguson (1974), is promising.

Consider just the $n = 136$ non-zero dose values Y_i in the 100-year data. One way to write the **parametric Bayesian lognormal model** is

$$\begin{array}{rcl} \log(Y_i) & = & \mu + \sigma \, e_i \\ (\mu, \sigma^2) & \sim & p(\mu, \sigma^2) \\ e_i & \stackrel{\text{IID}}{\sim} & N(0, 1), \end{array} \qquad (8)$$

for some prior distribution $p(\mu, \sigma^2)$ on μ and σ^2.

The Pólya trees idea is to replace the last line of (7), which expresses certainty about the distribution of the e_i, with a **distribution on the set of possible distributions** F for the e_i.

Pólya Trees

The new model is

$$\begin{array}{rcl} \log(Y_i) & = & \mu + \sigma\, e_i \\ (\mu, \sigma^2) & \sim & p(\mu, \sigma^2) \\ (e_i | F) & \overset{\text{IID}}{\sim} & F \quad \text{(mean 0, SD 1)} \\ F & \sim & PT\,(\Pi, \mathcal{A}_c)\,. \end{array} \tag{9}$$

Here (a) $\Pi = \{B_\epsilon\}$ is a **binary tree partition** of the real line, where ϵ is a binary sequence which locates the set B_ϵ in the tree.

You get to choose these sets B_ϵ in a way that **centers the Pólya tree on any distribution you want**, in this case the standard normal.

This is done by choosing the cutpoints on the line, which define the partitions, **based on the quantiles of** $N(0,1)$:

Level	Sets	Cutpoint(s)
1	(B_0, B_1)	$\Phi^{-1}(\frac{1}{2}) = 0$
2	$(B_{00}, B_{01},$ $B_{10}, B_{11})$	$\Phi^{-1}(\frac{1}{4}) = -0.674, \Phi^{-1}(\frac{1}{2}) = 0,$ $\Phi^{-1}(\frac{3}{4}) = +0.674$
$\vdots$	$\vdots$	$\vdots$

(Φ is the $N(0,1)$ CDF.) In practice this process has to stop somewhere; I use a tree **defined down to level** $M = 8$, which is like working with **random histograms**, each with $2^8 = 256$ bins.

Pólya Trees (continued)

And (b) Walker et al. (1998):

A helpful image is that of a **particle cascading through the partitions** B_ϵ. It starts [on the real line] and moves into B_0 with probability C_0 or into B_1 with probability $C_1 = 1 - C_0$. In general, on entering B_ϵ the particle could either move into $B_{\epsilon 0}$ or into $B_{\epsilon 1}$. Let it move into the former with probability $C_{\epsilon 0}$ or into the latter with probability $C_{\epsilon 1} = 1 - C_{\epsilon 0}$. For Pólya trees, these probabilities are random, **beta** variables, $(C_{\epsilon 0}, C_{\epsilon 1}) \sim \text{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ with non-negative $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$. If we denote the collection of α's by $\mathcal{A}$, a particular Pólya tree distribution is completely defined by Π and $\mathcal{A}$.

To make a Pólya tree distribution choose a continuous distribution with probability 1, the α's have to **grow quickly** as the level m of the tree increases. Following Walker et al. (1998) I take

$$\alpha_\epsilon = c\, m^2 \text{ whenever } \epsilon \text{ defines a set at level } m, \tag{10}$$

and this defines $\mathcal{A}_c$.

$c > 0$ is a kind of **tuning constant**: with small c the posterior distribution for the CDF of the e_i will be based almost completely on $\widehat{F}_n$, the empirical CDF (the "data distribution") for the e_i, whereas with large c the posterior will be based almost completely on the prior centering distribution, in this case $N(0, 1)$.

Prior to Posterior Updating

Prior to posterior updating is easy with Pólya trees: if

$$\begin{array}{rcl} F & \sim & PT(\Pi, \mathcal{A}) \\ (Y_i | F) & \overset{\text{IID}}{\sim} & F \end{array} \qquad (11)$$

and (say) Y_1 is observed, then the posterior $p(F|Y_1)$ for F given Y_1 is **also a Pólya tree** with

$$(\alpha_\epsilon | Y_1) = \left\{ \begin{array}{cc} \alpha_\epsilon + 1 & \text{if } Y_1 \in B_\epsilon \\ \alpha_\epsilon & \text{otherwise} \end{array} \right\}. \qquad (12)$$

In other words the updating follows a **Pólya urn scheme** (e.g., Feller, 1968): at each level of the tree, if Y_1 falls into a particular partition set B_ϵ, then 1 is added to the α for that set.

Inference for μ and σ^2

With $Y = (Y_1, \ldots, Y_n)$ as the vector of non-zero dose values, as usual Bayes' Theorem gives

$$p(\mu, \sigma^2 | Y) \propto p(\mu, \sigma^2) \, l(\mu, \sigma^2 | Y). \qquad (13)$$

Here I use the **conjugate** prior for μ and σ^2,

$$\begin{aligned} \sigma^2 & \sim \chi^{-2}(\nu_p, \sigma_p^2) \\ (\mu | \sigma^2) & \sim N\left(\mu_p, \frac{\sigma^2}{\kappa_p}\right) \end{aligned} \qquad (14)$$

(χ^{-2} denotes the distribution of the reciprocal of a χ^2 variate), and $l(\mu, \sigma^2 | Y)$ is the **likelihood** function (the sampling distribution for Y given μ and σ^2, re-interpreted as a function of μ and σ^2 for fixed Y).

(13) is hard to use to draw inferences about μ and σ^2 for two reasons: (a) there is the usual difficulty in extracting **marginal** information about μ or σ^2 singly, and (b) $l(\mu, \sigma^2 | Y)$ is not directly evaluable, and depends in a complicated way on something that is directly computable, the **conditional** likelihood $l(\mu, \sigma^2 | Y, F)$.

Figs. 12–14 plot the conditional likelihood, which has a remarkable, almost **fractal**, character in this nonparametric setting.

Conditional Likelihood Plots

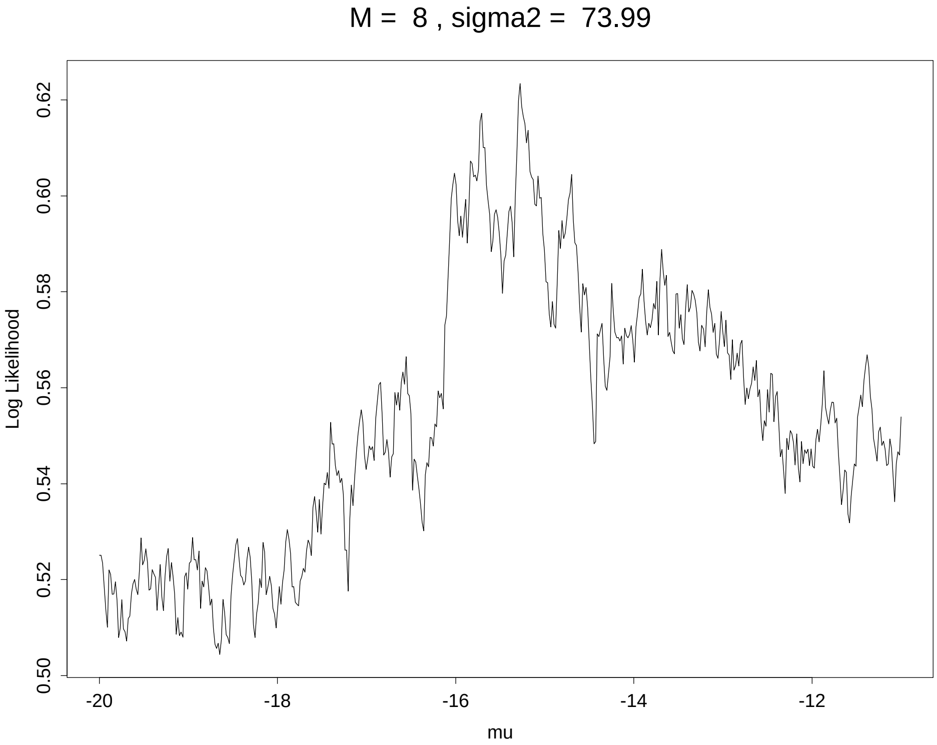


Figure 12. *The conditional log likelihood function for μ given a particular estimate of F based on the 100 year data and with $\sigma^2 = 73.99$. The log likelihood in the parametric version of this model is much smoother.*

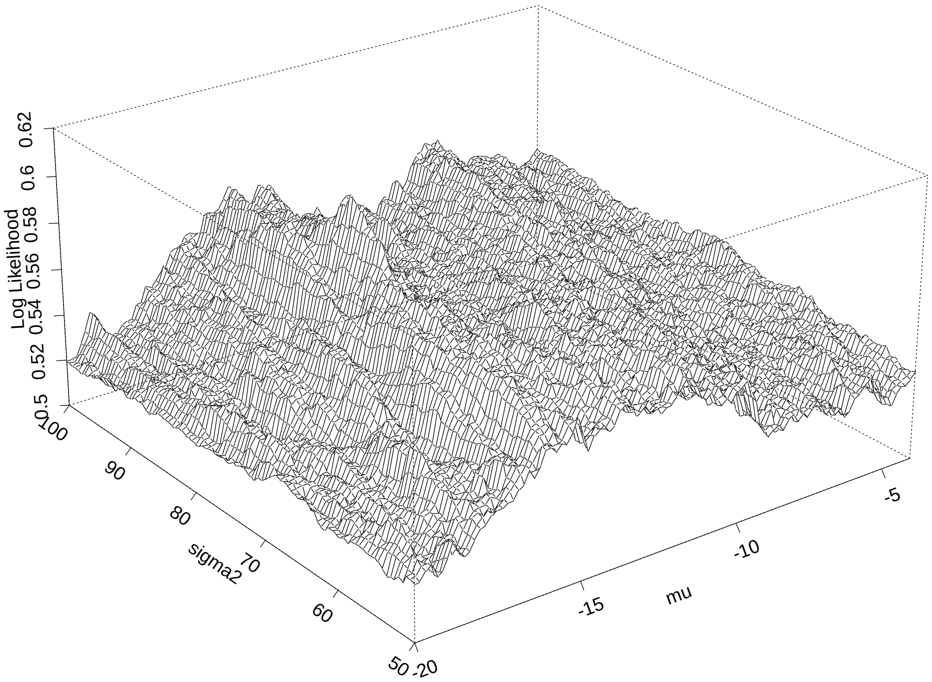


Figure 13. *The (joint) conditional log likelihood function for μ and σ^2 with the 100 year data. The global maximum is barely visible near $(\mu, \sigma^2) = (-15, 80)$.*

MCMC Again

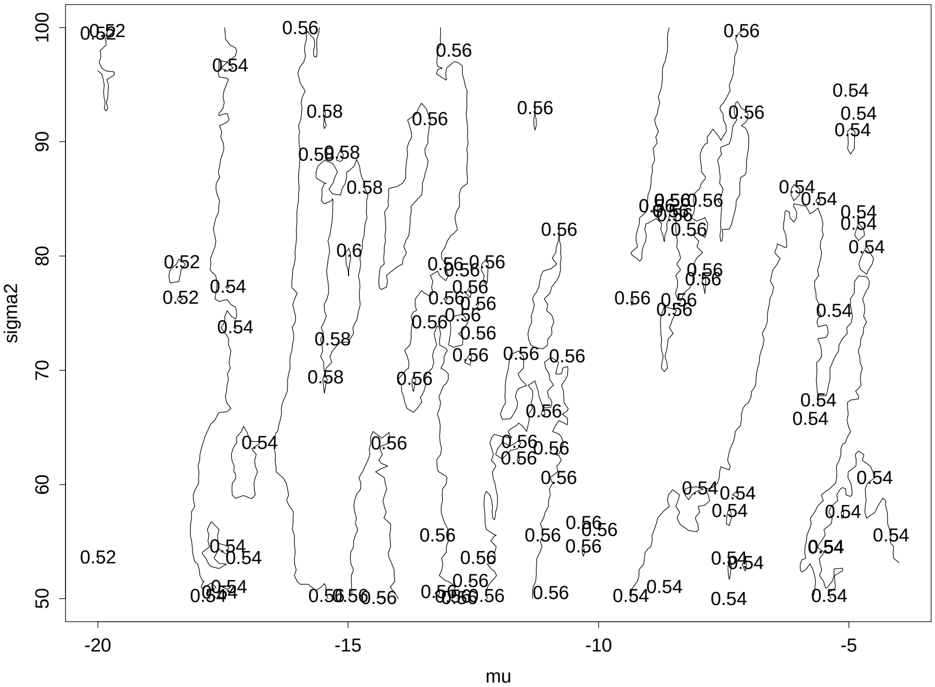


Figure 14. *A contour plot of the same (joint) conditional log likelihood function for μ and σ^2 as in Fig. 13. Now the global max is easier to spot, as the only place with a conditional log likelihood of 0.6.*

To overcome the computational problems I again use MCMC. Walker et al. (1998) sketch a **Metropolis within Gibbs** algorithm for a model like (8), except they pretend that σ^2 is known. I have extended this algorithm to the more realistic case of unknown σ^2.

(Sample F from its full conditional given (μ, σ^2) (easy: just Pólya updating), and use a random-walk Metropolis to sample $[\mu, \log(\sigma^2)]$, given F, with a bivariate normal proposal distribution. On each MCMC sweep renormalize F to have mean 0 and SD 1.)

Inference for θ

It still remains to relate all this to $\theta_Y = E(Y)$ (and to incorporate the **mixture** aspect of model (5)).

With $W = \log(Y)$, each iteration of the MCMC obtains an estimate of the CDF F_W (in the form of a **histogram** estimate of the density f_W with $2^M = 256$ bins). But

$$\theta_Y = E(Y) = E(e^W) = \int_{-\infty}^{\infty} e^w f_W(w)\, dw, \tag{15}$$

so θ_Y can be estimated from each MCMC iteration by

$$\hat{\theta}_Y = \sum_{j=1}^{2^M} e^{w_j^*} \hat{p}_j(w_j^*), \tag{16}$$

where the $\hat{p}$'s are the **current bin proportions** and the w_j^* are the bin centers.

As an example of the results, Figs. 15–18 present posterior summaries based on 5,000 monitoring iterations, arising from the following modeling inputs: (Pólya tree prior) $M = 8, c = 1$ (**NB** and the centering distribution was the standard normal **Winsorized** to ±2.42 (roughly the 0.008 point of the distribution), to damp down the tail); (prior on μ_2 and σ_2^2) $\nu_p = \kappa_p = 50, \mu_p = -15.7$ and $\sigma_p^2 = 74.0$ (the sample mean and variance of the $\log(y_i)$); and Metropolis proposal distribution bivariate normal with covariance matrix

$$\Sigma = K \begin{pmatrix} 0.54 & 0.000 \\ 0.00 & 0.015 \end{pmatrix}, \tag{17}$$

(the matrix values are based on Fisher information), with $K = 0.5$ (Metropolis acceptance rate 76%).

Results

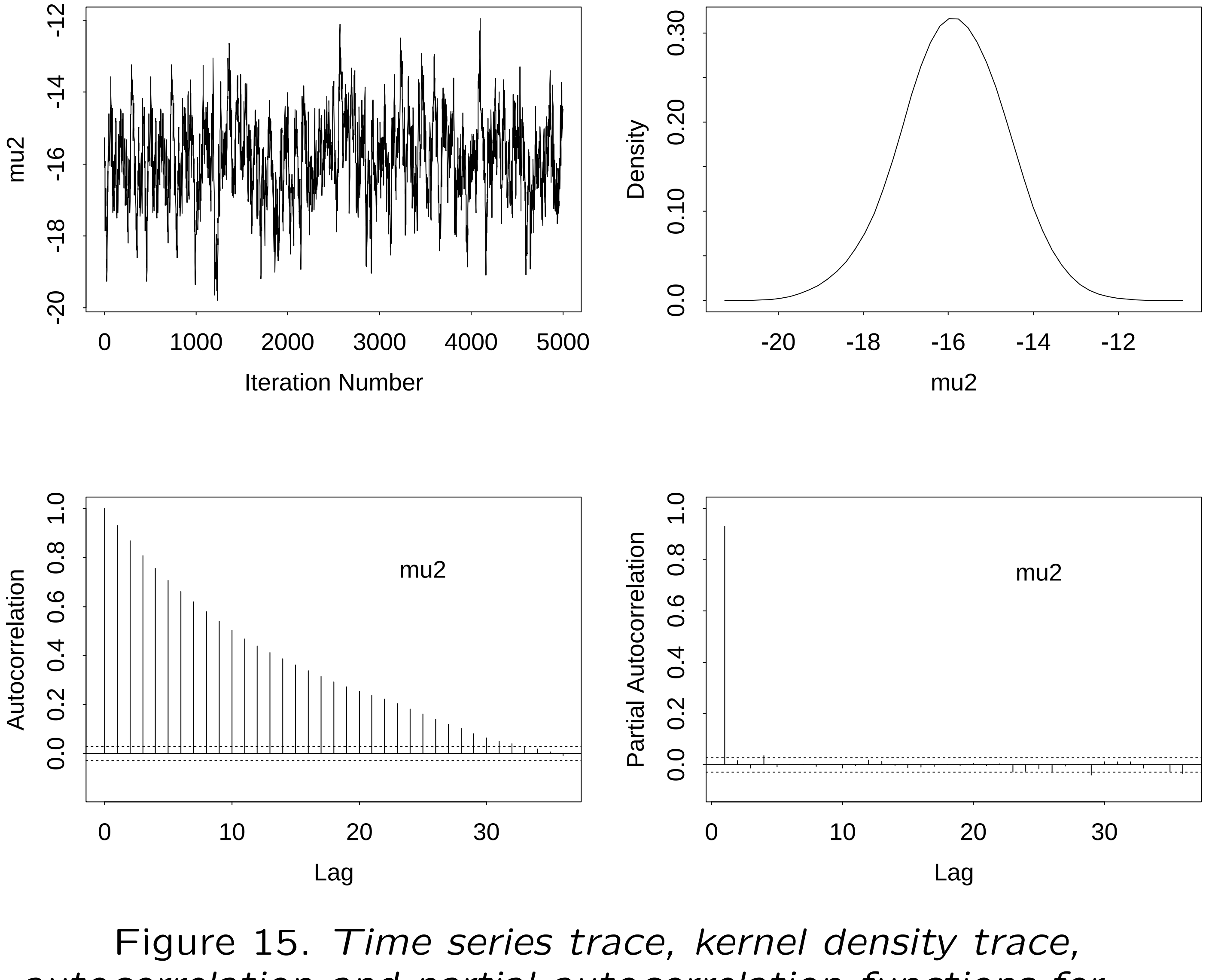


Figure 15. *Time series trace, kernel density trace, autocorrelation and partial autocorrelation functions for* μ_2.

Results (continued)

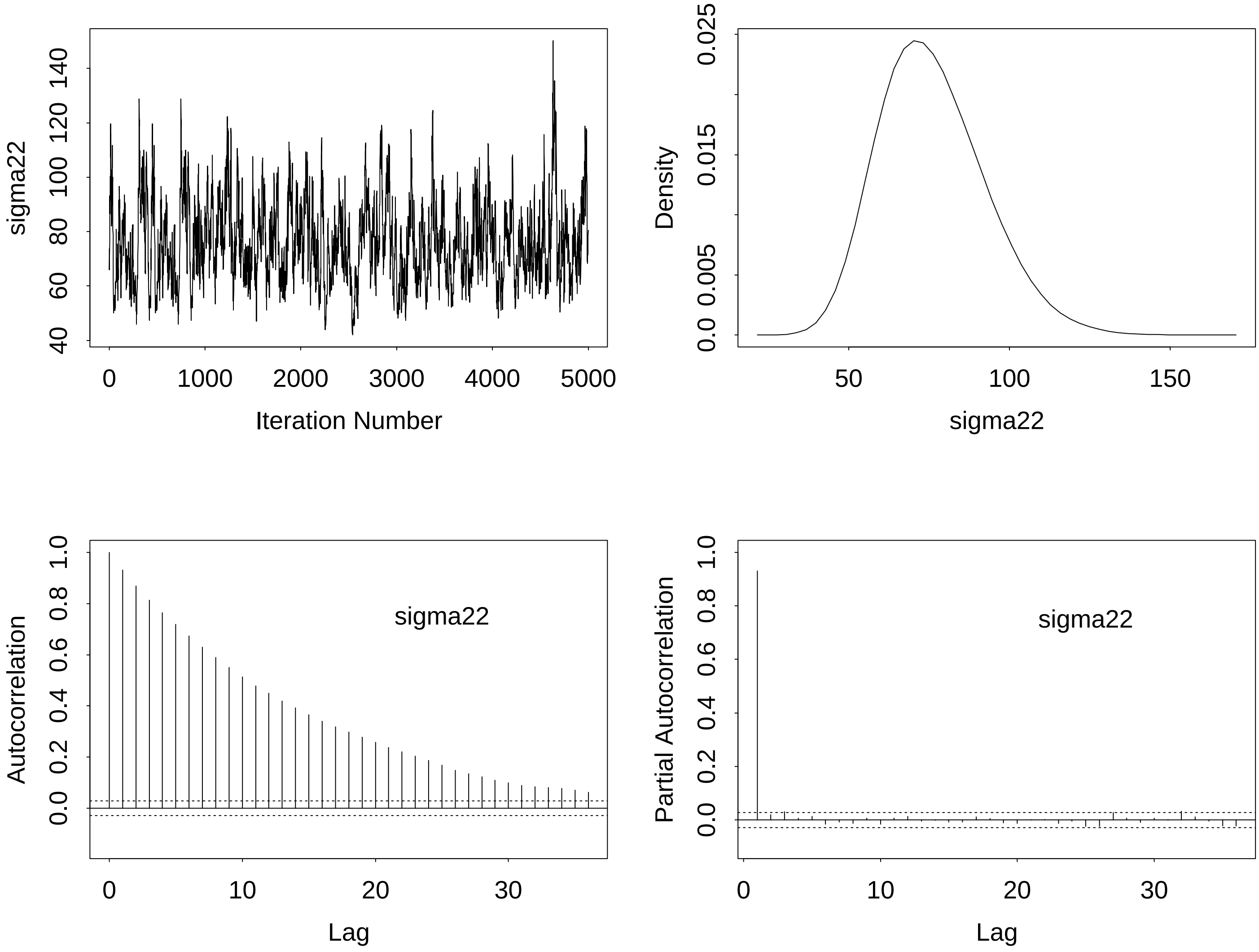


Figure 16. *Time series trace, kernel density trace, autocorrelation and partial autocorrelation functions for σ_2^2.*

Results (continued)

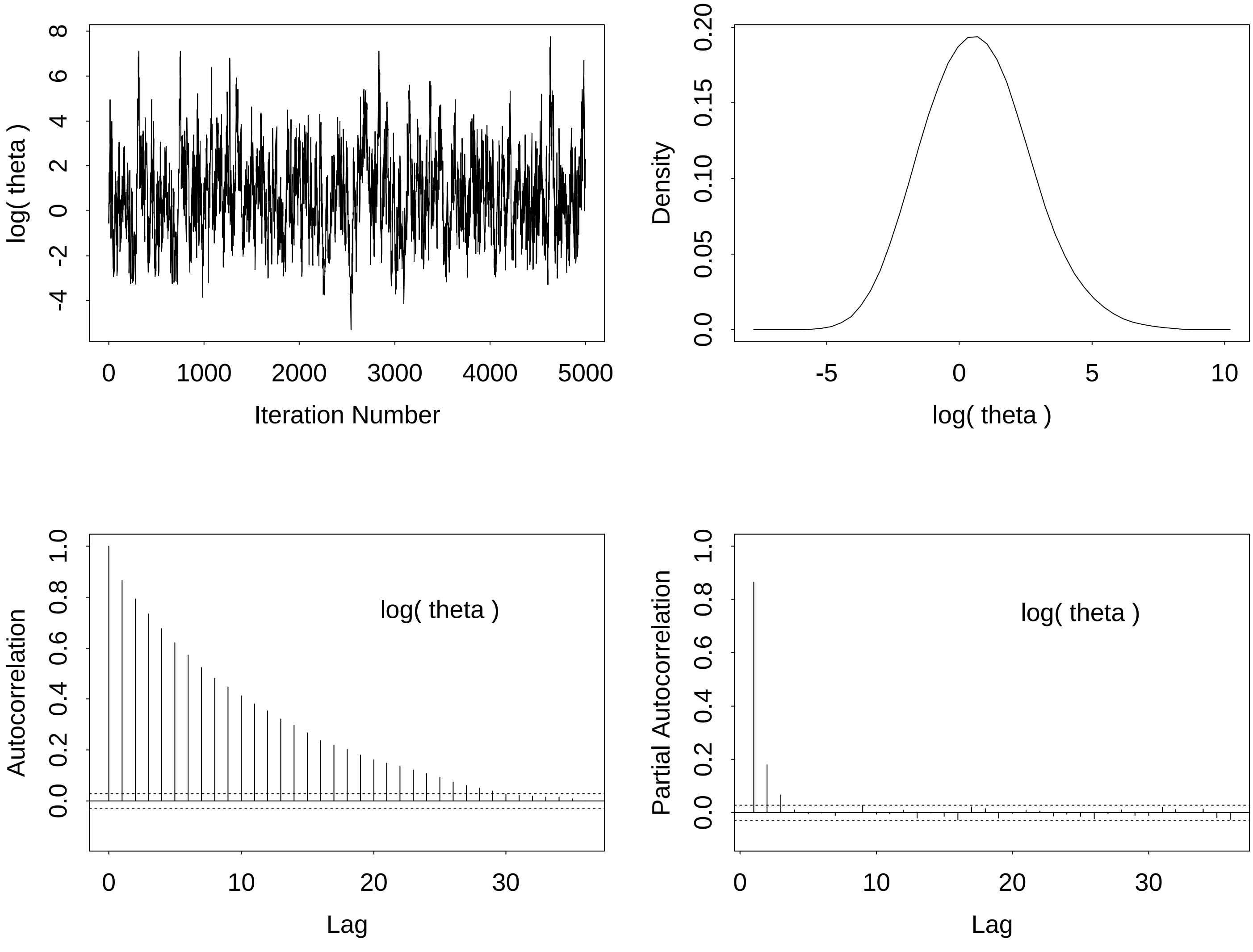


Figure 17. *Time series trace, kernel density trace, autocorrelation and partial autocorrelation functions for* $\log(\theta)$.

Results (continued)

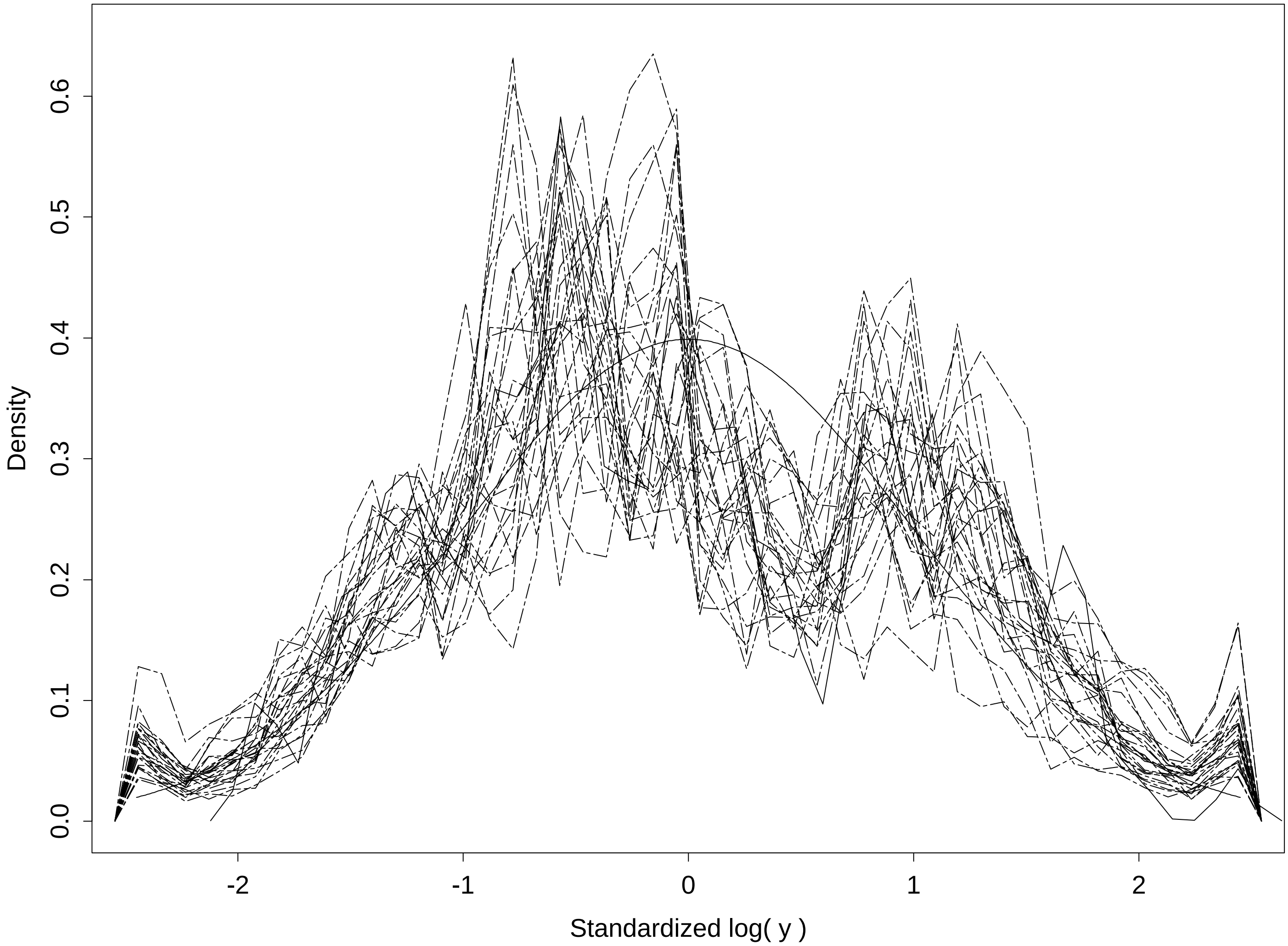


Figure 18. *Prior (standard normal) for CDF F of standardized* log(y) *and density trace of sample of y values (bold lines), with 25 density traces of MCMC draws from the posterior of F (dotted lines). Note the compromise effected between the prior and the sample with $c = 1$.*

Results (continued)

Preliminary results are given in Table 4 for the 100 year data, using $c = 1$ as a kind of **compromise** tuning constant. The sample mean of the e^y values in the data was 1.44.

Table 4. *Parametric versus nonparametric results with the nonzero part of the 100 year data.*

	Parametric Posterior		Nonparametric Posterior	
Parameter	Mean	SD	Mean	SD
μ_2	–16.0	0.812	-15.9	1.17
σ_2^2	82.5	11.9	76.1	15.1
θ_Y	28.0	959	14.7	78.7

(median of θ_Y = 1.80,
95% central interval = (0.0660, 102.8))

Note that the nonparametric approach results in larger posterior SDs for μ_2 and σ_2^2 (correctly acknowledging greater uncertainty) but a smaller posterior SD for θ (correctly damping down the extreme lognormal tail).

It appears from a small preliminary simulation that by varying c from 0.1 to 10 it is possible to obtain **actual coverage close to 95%** for nominal 95% intervals with this approach without unnecessarily wide intervals (work in progress).

DP Mixture Model Case Study

• (joint work with **Thanasis Kottas** and **Milovan Krnjajić**)

• 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 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; Part 1) 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):

$$\begin{array}{ccc} (y_i|\theta_i) & \overset{\text{ind}}{\sim} & \text{Poisson}(\exp(\theta_i)) \\ (\theta_i|G) & \overset{\text{iid}}{\sim} & G \\ G & \equiv & \text{N}(\mu, \sigma^2) \end{array} \qquad (18)$$

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

• 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).

Dirichlet Process Mixture Model

• 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**:

$$\begin{array}{ccc} (y_i \mid \theta_i) & \overset{ind}{\sim} & \text{Poisson}(\exp(\theta_i)) \\ (\theta_i \mid G) & \overset{iid}{\sim} & G \\ G & \sim & \text{DP}(\alpha G_0), \quad G_0 \equiv G_0(\cdot;\psi), \end{array} \tag{19}$$

where $i = 1, ..., n$ (we refer to (19) as **BNP model 1**).

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

$$(y_i \mid G) \overset{iid}{\sim} f(\cdot; G) = \int \text{Poisson}(y_i; \exp(\theta))\text{d}G(\theta), \; G \sim \text{DP}(\alpha G_0), \tag{20}$$

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

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

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:

$$\begin{array}{ccc} Y_{1i} \mid \theta_i & \overset{ind}{\sim} & \text{Poisson}(\exp(\theta_i)), i = 1, n_1 \\ Y_{2k} \mid \theta_{1,n_1+k}, \theta_{2k} & \overset{ind}{\sim} & \text{Poisson}(\exp(\max(\theta_{1,n_1+k}, \theta_{2k}))), k = 1, n_2 \\ \theta_{1i} \mid H_1 & \overset{iid}{\sim} & H_1, i = 1, n_1 + n_2 \\ \theta_{2k} \mid H_2 & \overset{iid}{\sim} & H_2, k = 1, n_2 \\ H_r \mid \alpha_r, \mu_r, \sigma_r^2 & \sim & DP(\alpha_r H_{r0}) \end{array} \tag{21}$$

where the **base distributions** of Dirichlet processes, H_{10} and H_{20}, are again **Normal** with parametric priors on hyperparameters. We refer to (21) as BNP model 2.

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

Posterior Predictive Distributions

- 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**.

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

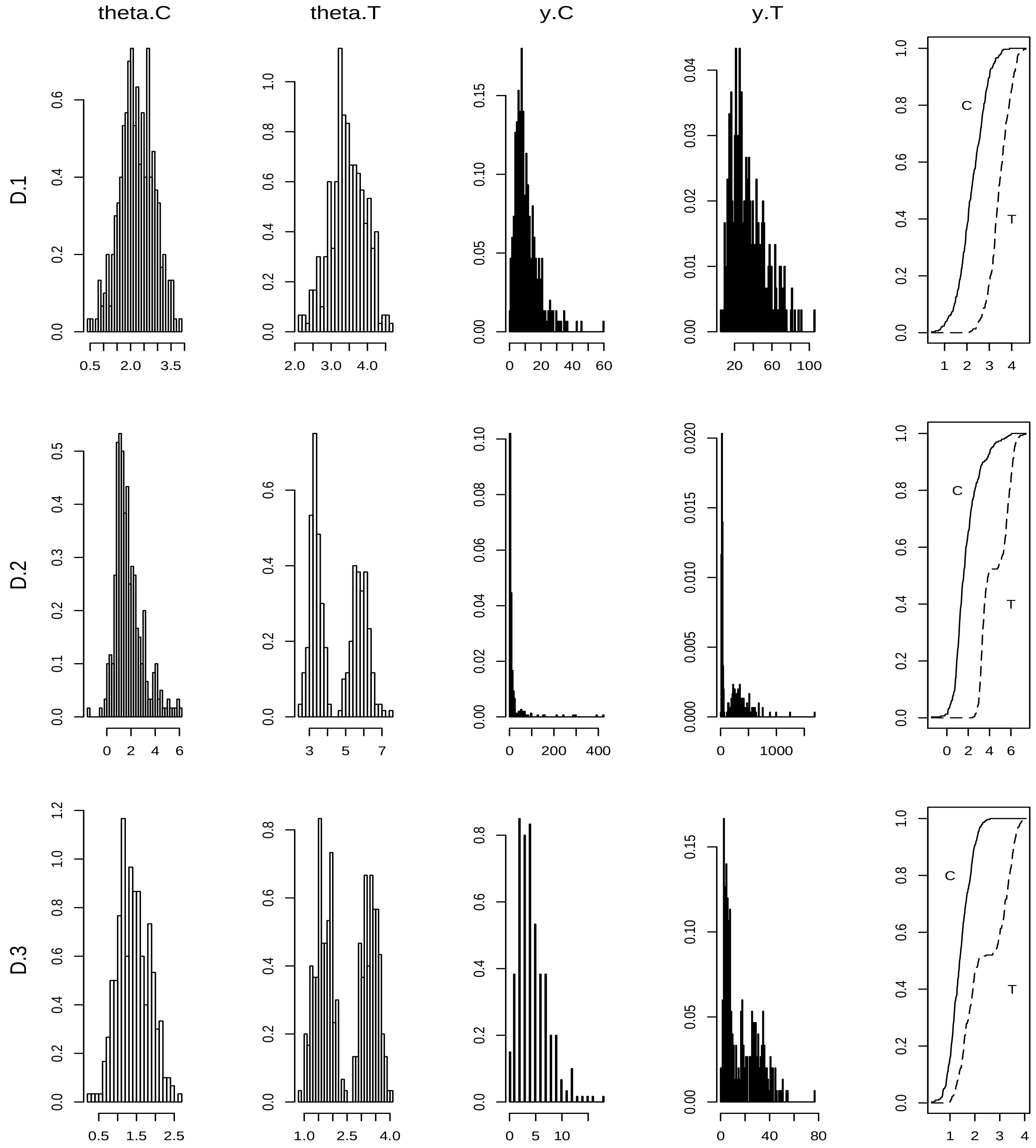
$$[Y^{\text{new}} \mid \text{data}] = \iint \text{Poisson}(Y^{\text{new}}; \exp(\theta^{\text{new}}))[\theta^{\text{new}} \mid \boldsymbol{\eta}][\boldsymbol{\eta} \mid \text{data}], \tag{22}$$

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

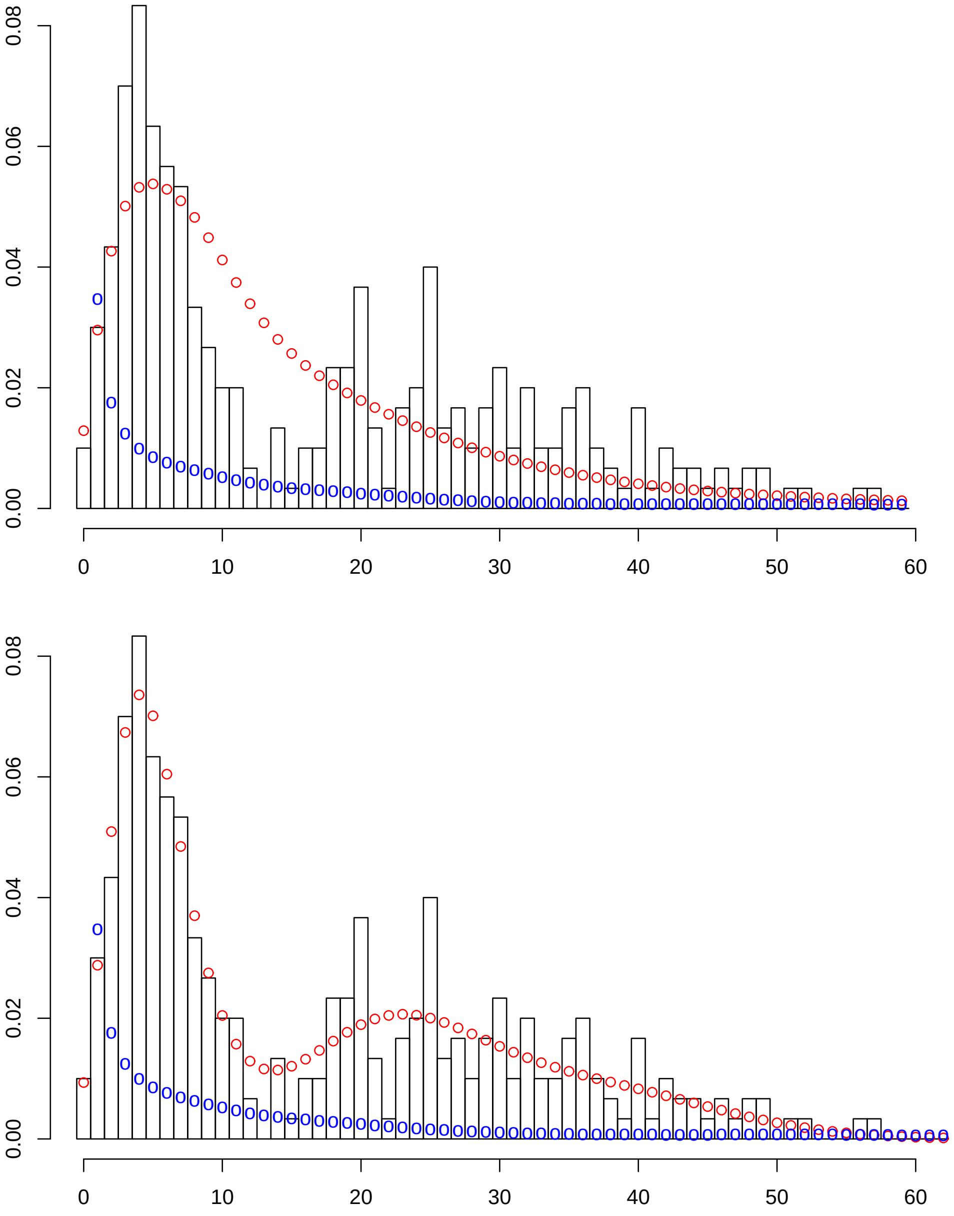
$$[\theta^{\text{new}} \mid \boldsymbol{\eta}] = \frac{\alpha}{\alpha + n} G_{r0}(\theta^{\text{new}} \mid \mu_r, \sigma^2) + \frac{1}{\alpha + n} \sum_{\ell=1}^{n} n_\ell \delta_{\theta_\ell}(\theta^{\text{new}}). \tag{23}$$

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



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**.

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),

$$[G|\text{data}] = \int [G|\theta, \alpha, \psi] \mathrm{d}[\theta, \alpha, \psi|\text{data}]. \qquad (24)$$

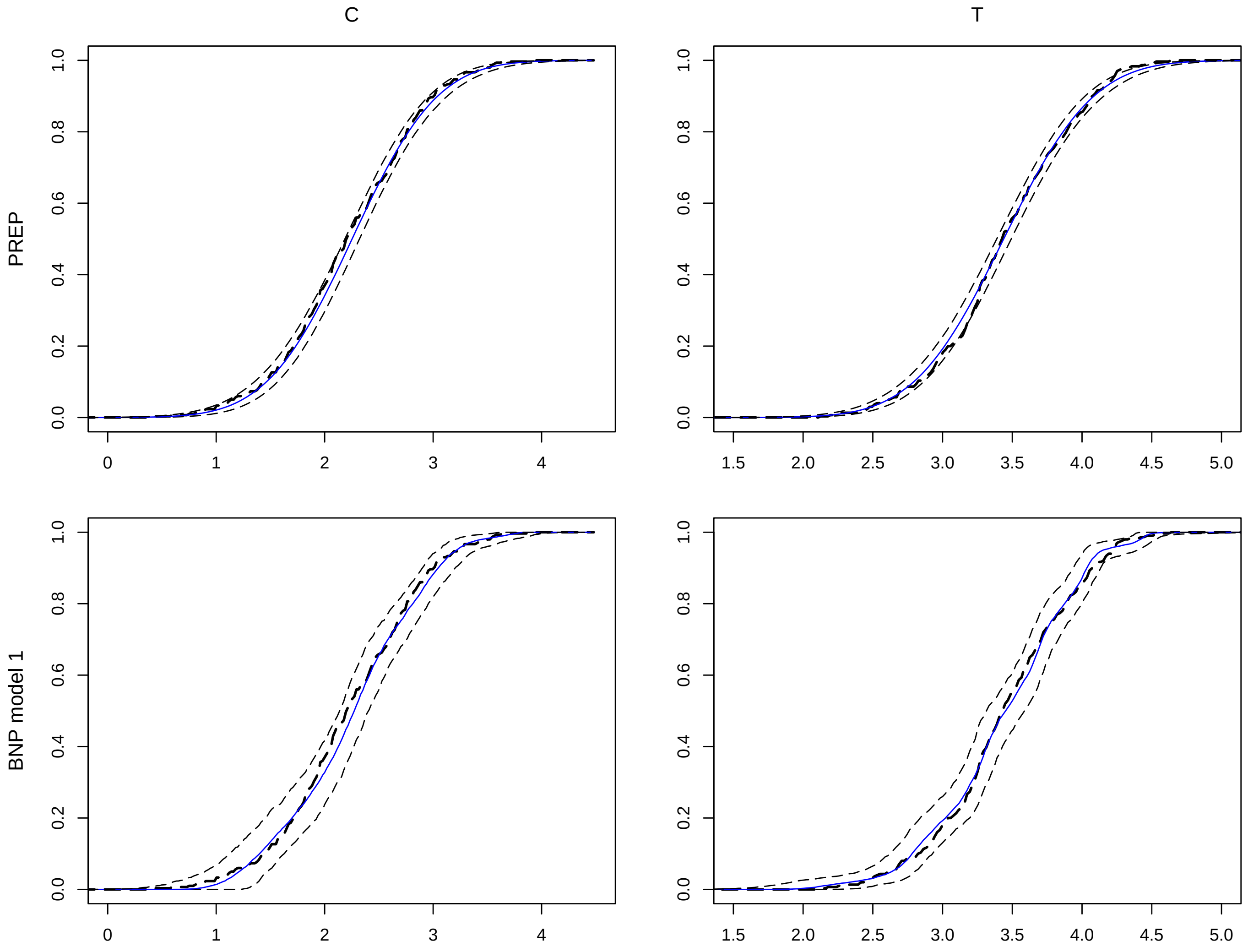
where $[G|\theta, \alpha, \psi]$ is **also a DP** with parameters $\alpha^{'} = \alpha + n$ and

$$G_0^{'}(\cdot|\psi) = \frac{\alpha}{\alpha+n} G_0(\cdot|\psi) + \frac{1}{\alpha+n} \sum\nolimits_{i=1}^{n} 1_{(-\infty,\theta_i]}(\cdot), \qquad (25)$$

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

- Using (24), (25) and the definition of DP we develop **computationally efficient** approach to obtaining **posterior sample paths** from $[G \mid \text{data}]$.

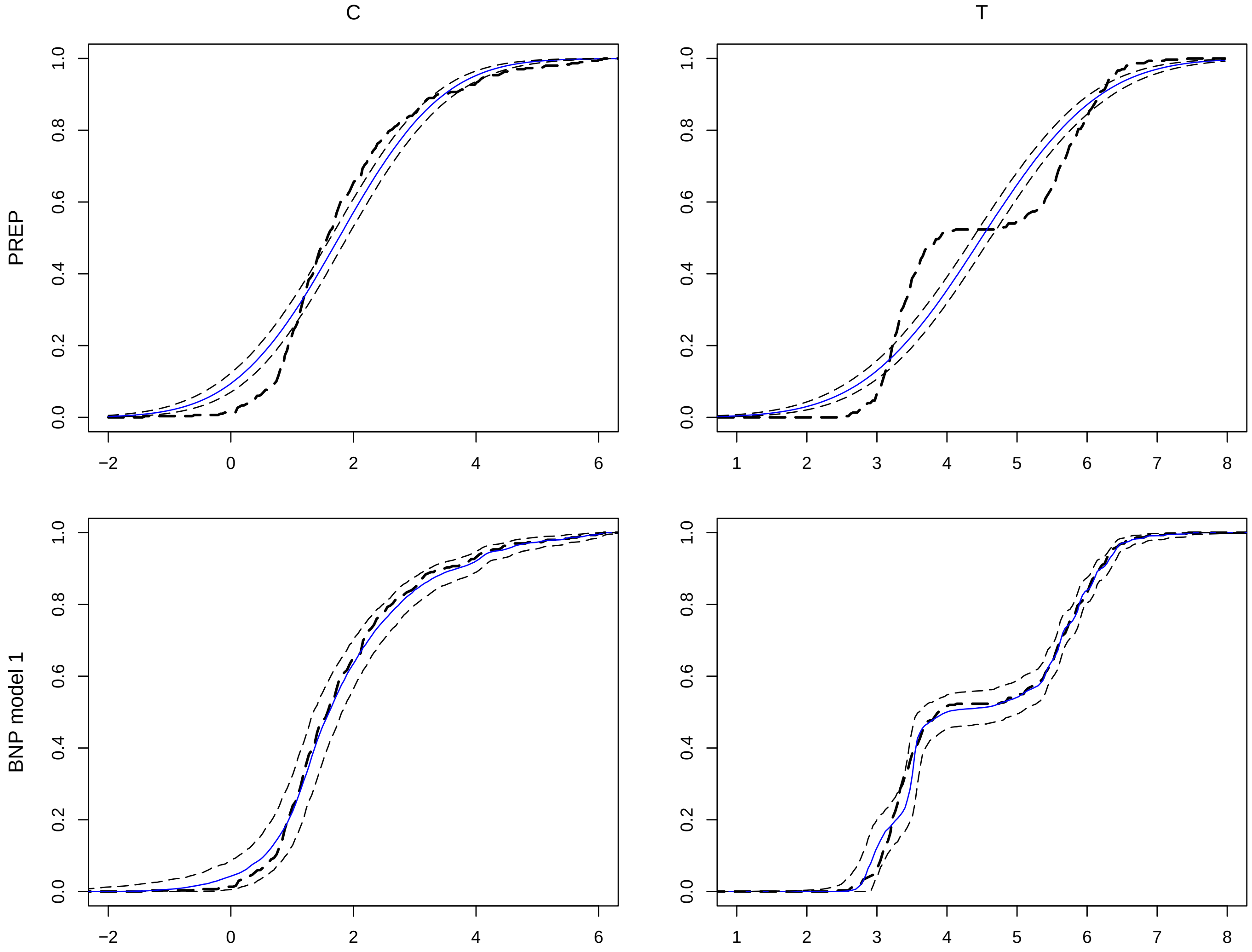
Normal Random Effects: PREP vs. BNP



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**.

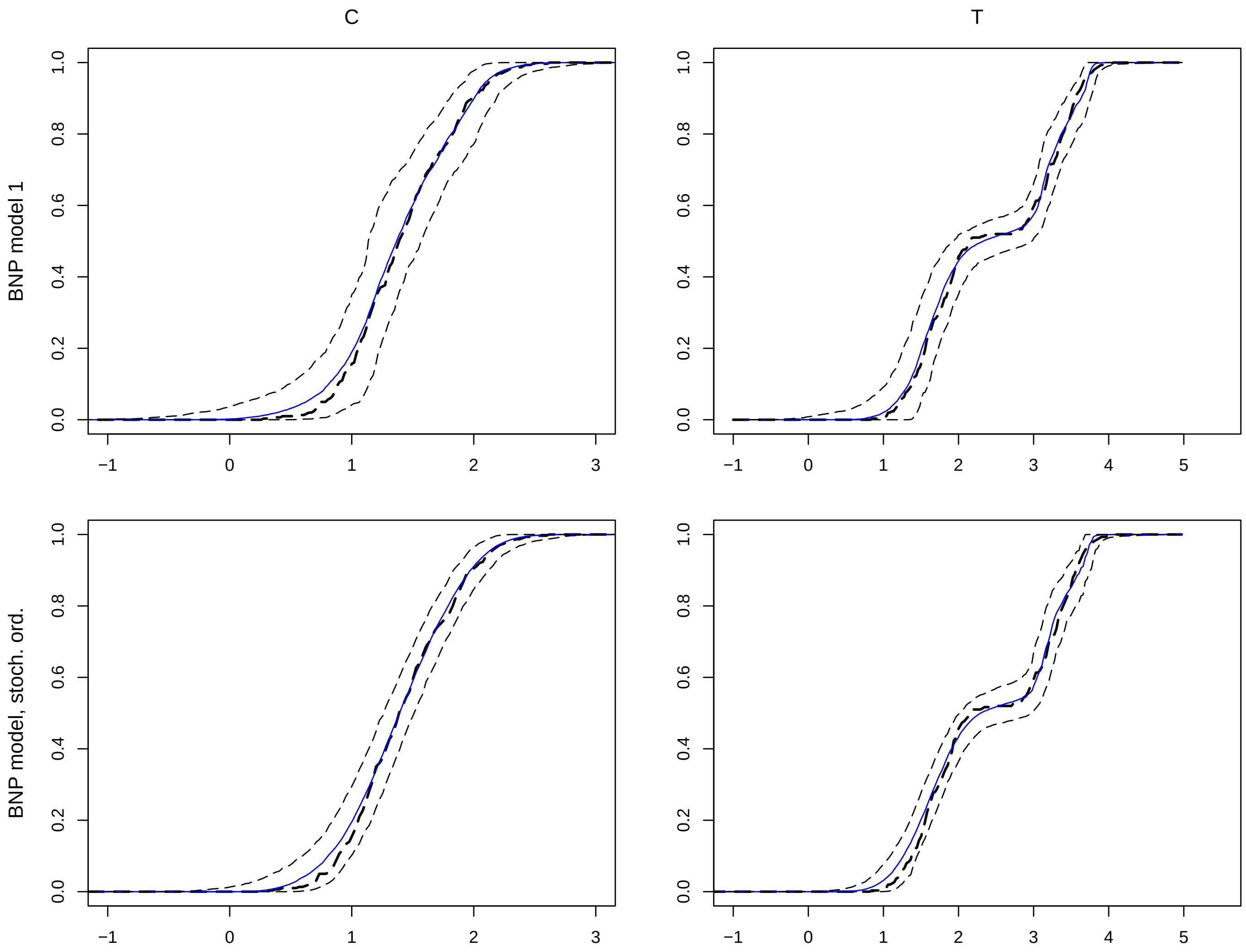
Skewed and Bimodal Random Effects, PREP vs. BNP



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.

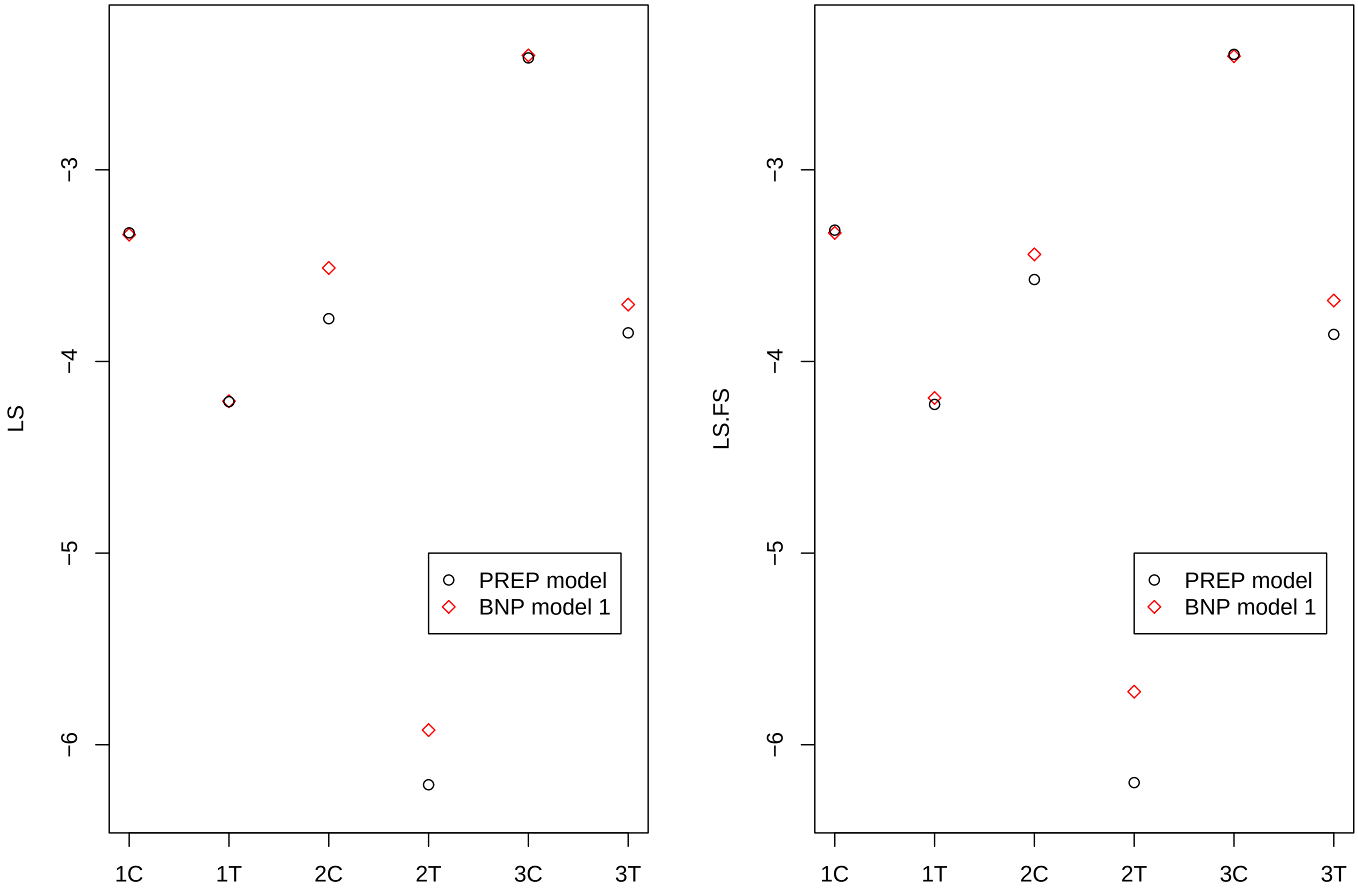
Bimodal Random Effects: BNP With and Without Stochastic Order



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 and LS_{FS}
For PREP and BNP Models



LS (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), LS and LS_{FS} for PREP and BNP **nearly coincide** (as they 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).

Conclusions

- The **BNP methods** we illustrate here allow the fitting of **random-effects models** without making **restrictive** (and **potentially incorrect**) parametric distributional assumptions about the random effects; these methods provide posterior inference for the **unknown random effects distribution** G and associated **functionals** of interest, as well as predictive distributions for **future data** (useful for model comparison).

- In Milovan's dissertation work, besides the BNP models shown here, we have also considered one more BNP model, with **bivariate base distribution** for DP to induce **dependence** between random effects C and T distributions.

- All BNP models exhibit **superior performance** compared to their parametric counterparts on **all data sets not generated from the parametric model** (e.g., with random effects drawn from **skewed** and **bimodal** distributions).

References

Draper D (1997). Model uncertainty in "stochastic" and "deterministic" systems. In *Proceedings of the 12th International Workshop on Statistical Modeling*, Minder C, Friedl H (eds.), Vienna: *Schriftenreihe der Österreichischen Statistichen Gesellschaft*, **5**, 43–59.

Draper D (1999). *Bayesian Hierarchical Modeling*. New York: Springer-Verlag, forthcoming.

Efron B, Tibshirani RJ (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Feller W (1968). *An Introduction to Probability Theory and Its Applications, Volume I*, Third Edition. New York: Wiley.

Ferguson TS (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615–629.

Gilks WR, Richardson S, Spiegelhalter DJ (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Johnson NL, Kotz S (1970). *Distributions in Statistics: Continuous Univariate Distributions, Volume 1*. Boston: Houghton-Mifflin.

Lavine M (1992). Some aspects of Pólya tree distributions for statistical modeling. *Annals of Statistics*, **20**, 1203–1221.

Lavine M (1994). More aspects of Pólya trees for statistical modeling. *Annals of Statistics*, **20**, 1161–1176.

PSAC (Probabilistic System Assessment Code) User Group (1989). *PSACOIN Level E Intercomparison*. Nuclear Energy Agency: Organization for Economic Co-operation and Development.

References (continued)

Sinclair J (1996). Convergence of risk estimates obtained from highly skewed distributions. AEA Technology briefing.

Sinclair J, Robinson P (1994). The unsolved problem of convergence of PSA. Presentation at the 15th meeting of the NEA Probabilistic System Assessment Group, 16–17 June 1994, Paris.

Spiegelhalter DJ, Thomas A, Best NG, Gilks WR (1997). *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.6*. Cambridge: Medical Research Council Biostatistics Unit.

Walker SG, Damien P, Laud PW, Smith AFM (1998). Bayesian nonparametric inference for random distributions and related functions. Technical report, Department of Mathematics, Imperial College, London.

Woo G (1989). Confidence bounds on risk assessments for underground nuclear waste repositories. *Terra Nova*, **1**, 79–83.