

# Introduction to Hierarchical Bayes Methods for Data Analysis

ASA Biopharmaceutical Section Webinar  
October 13, 2009, noon – 2 pm Eastern time

presented by

**Bradley P. Carlin**

brad@biostat.umn.edu

Division of Biostatistics  
School of Public Health  
University of Minnesota

## Overview

- Statisticians are increasingly faced with data that are:
  - highly multivariate, with many important predictors and response variables
  - temporally correlated (longitudinal, survival studies)
  - costly and difficult to obtain, but often with historical data on previous but similar drugs or devices
- Recently, the FDA Center for Devices has encouraged hierarchical Bayesian statistical approaches –
  - Methods are not terribly novel: Bayes (1763)!
  - But their practical application has only become feasible in the last decade or so due to advances in computing via Markov chain Monte Carlo (MCMC) methods and related WinBUGS software

## **Role of Bayes in biometric settings**

- Safety/efficacy studies: Historical data and/or information from published literature can be used to reduce sample size, reducing time and expense. Unlimited looks at accumulating data are also permitted (due to different framework for testing).
- Equivalence studies: Bayes allows one to make direct statements about the probability that one drug is equivalent to another, rather than merely “failing to reject” the hypothesis of no difference.
- Meta-analysis: Bayes facilitates combining disparate but similar studies of a common drug or device.
- Hierarchical models: Realistic models can be fit to complicated, multilevel data (e.g., multiple observations per patient, or multiple patients per clinical site), accounting for all sources of uncertainty.

## **Bayesian design of experiments**

- In traditional sample size formulae, one often plugs in a “best guess” or “smallest clinically significant difference” for  $\theta \Rightarrow$  “Everyone is a Bayesian at the design stage.”
- In practice, frequentist and Bayesian outlooks arise:
  - Applicants may have a more Bayesian outlook:
    - to take advantage of historical data or expert opinion (and possibly stop the trial sooner), or
    - to “peek” at the accumulating data without affecting their ability to analyze it later
  - Regulatory agencies may appreciate this, but also retain many elements of frequentist thinking:
    - to ensure that in the long run they will only rarely approve a useless or harmful product

---

Applicants must thus design their trials accordingly!

## Some preliminary Q&A

- *What is the philosophical difference between classical (“frequentist”) and Bayesian statistics?*
  - To a frequentist, unknown model parameters are fixed and unknown, and only estimable by replications of data from some experiment.
  - A Bayesian thinks of parameters as random, and thus having distributions (just like the data). We can thus think about unknowns for which no reliable frequentist experiment exists, e.g.

$\theta$  = proportion of US men with  
untreated atrial fibrillation

## Some preliminary Q&A

- *How does it work?*
  - A Bayesian writes down a prior guess for  $\theta$ ,  $p(\theta)$ , then combines this with the information that the data  $X$  provide to obtain the posterior distribution of  $\theta$ ,  $p(\theta|X)$ . All statistical inferences (point and interval estimates, hypothesis tests) then follow as appropriate summaries of the posterior.
  - Note that

posterior information  $\geq$  prior information  $\geq 0$ ,

with the second “ $\geq$ ” replaced by “=” only if the prior is noninformative (which is often uniform, or “flat”).

## Some preliminary Q&A

- *Is the classical approach “wrong”?*
  - While a “hardcore” Bayesian might say so, it is probably more accurate to think of classical methods as merely “limited in scope”!
  - The Bayesian approach expands the class of models we can fit to our data, enabling us to handle:
    - any outcome (binary, count, continuous, censored)
    - repeated measures / hierarchical structure
    - complex correlations (longitudinal, spatial, or cluster sample) / multivariate data
    - unbalanced or missing data
    - and many other settings that are awkward or infeasible from a classical point of view.
  - The approach also eases the interpretation of and learning from those models once fit.

## Simple example of Bayesian thinking

- From *Business Week*, online edition, July 31, 2001:

“Economists might note, to take a simple example, that American turkey consumption tends to increase in November. A Bayesian would clarify this by observing that Thanksgiving occurs in this month.”
- Data: plot of turkey consumption by month
- Prior:
  - location of Thanksgiving in the calendar
  - knowledge of Americans’ Thanksgiving eating habits
- Posterior: Understanding of the pattern in the data!

## Bayes means revision of estimates

Humans tend to be Bayesian in the sense that most revise their opinions about uncertain quantities as more data accumulate. For example:

- Suppose you are about to make your first submission to a particular academic journal
- You assess your chances of your paper being accepted (you have an opinion, but the “true” probability is unknown)
- You submit your article and it is accepted!
- Question: What is your revised opinion regarding the acceptance probability for papers like yours?
- If you said anything other than “1”, you are a Bayesian!

## Bayes can account for structure

County-level breast cancer rates per 10,000 women:

79	87	83	80	78
90	89	92	99	95
96	100	*	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for \*, what estimate would you use?
- Is 200 reasonable?
- Probably not: all the other rates are around 100
- Perhaps use the average of the “neighboring” values (again, near 100)

## Accounting for structure (cont'd)

- Now assume that data become available for county  $\star$ : 100 women at risk, 2 cancer cases. Thus

$$rate = \frac{2}{100} \times 10,000 = 200$$

Would you use this value as the estimate?

- Probably not: The sample size is very small, so this estimate will be unreliable. How about a compromise between 200 and the rates in the neighboring counties?
- Now repeat this thought experiment if the county  $\star$  data were 20/1000, 200/10000, ...
- Bayes and empirical Bayes methods can incorporate the structure in the data, weight the data and prior information appropriately, and allow the data to dominate as the sample size becomes large.

## Has Bayes Paid Real Dividends?

Yes! Here is an example of a dramatic savings in sample size from my work:

- Consider Safety Study B, in which we must show freedom from severe drug-related adverse events (AEs) at 3 months will have a 95% lower confidence bound at least 85%.
- Problem: Using traditional statistical methods, we obtain an estimated sample size of over 100 – *too large!*
- But: We have access to the following (1-month) data from Safety Study A:

	No AE	AE	total
count	110	7	117
(%)	(94)	(6)	

## Bayes Pays Real Dividends

- Since we expect similar results in two studies, use Study A data for the prior  $\Rightarrow$  reduced sample size!
- Model: Suppose  $N$  patients in Study B, and for each,

$$\theta = \text{Pr}(\text{patient does not experience the AE})$$

Let  $X = \#$  Study B patients with no AE (“successes”).

- If the prior is  $\theta \sim \text{Beta}(a = 100, b = 7)$  (the target prior), Bayes delivers equal weighting of Studies A and B.
- The company wound up opting for 50% downweighting of the Study A data (in order to obtain suitable Type I error behavior). This still delivered 79% power to ensure a  $\theta$  lower confidence bound of at least 87% with just  $N=50$  new Study B patients!

## Bayes Pays Real Dividends

Other dividends Bayes can offer:

- Time and Money: Bayesian approaches are natural for adaptive trials, where more promising treatments are emphasized as the trial is running, and for seamless Phase I-II or Phase II-III trials, reducing a compound’s “travel time” from development to FDA approval.
- Ethical: By reducing sample size, Bayesian trials expose fewer patients to the inferior treatment (regardless of which this turns out to be).
- These dividends are already being realized at FDA! CDRH has been an aggressive promoter of Bayesian methods, especially via the 2006 Guidance Document, [www.fda.gov/cdrh/osb/guidance/1601.html](http://www.fda.gov/cdrh/osb/guidance/1601.html) c.f. the forthcoming Bayesian clinical trials textbook by Berry, Berry, Carlin, Lee, and Müller (CRC Press, 2010)!

## Motivating Theoretical Example

- From Berger and Berry (1988, *Amer. Scientist*): Consider a clinical trial to study the effectiveness of Vitamin C in treating the common cold.
- Observations are matched pairs of subjects (twins?), half randomized (in “double blind” fashion) to vitamin C, half to placebo. We count how many pairs had C giving superior relief after 48 hours.

## Two Designs

- Design #1: Sample  $n = 17$  pairs, and test

$$H_0 : P(C \text{ better}) = \frac{1}{2} \text{ vs. } H_A : P(C \text{ better}) \neq \frac{1}{2}$$

Suppose we observe  $x = 13$  preferences for C. Then

$$\text{p-value} = P(X \geq 13 \text{ or } X \leq 4) = .049$$

So if  $\alpha = .05$ , stop and reject  $H_0$ .

- Design #2: Sample  $n_1 = 17$  pairs. Then:
  - if  $x_1 \geq 13$  or  $x_1 \leq 4$ , stop.
  - otherwise, sample an additional  $n_2 = 27$  pairs.Reject  $H_0$  if  $X_1 + X_2 \geq 29$  or  $X_1 + X_2 \leq 15$ .

## Two Designs (cont'd)

- We choose this second stage since under  $H_0$ ,  
 $P(X_1 + X_2 \geq 29 \text{ or } X_1 + X_2 \leq 15) = .049$   
– the same as Stage 1!
- Suppose we again observe  $X_1 = 13$ . Now:  
$$\begin{aligned} \text{p-value} &= P(X_1 \geq 13 \text{ or } X_1 \leq 4) \\ &\quad + P(X_1 + X_2 \geq 29 \text{ and } 4 < X_1 < 13) \\ &\quad + P(X_1 + X_2 \leq 15 \text{ and } 4 < X_1 < 13) \\ &= .085 \leftarrow \text{no longer significant at } \alpha = .05! \end{aligned}$$
- Yet the observed data was exactly the same; all we did was contemplate a second stage (no effect on data), and it changed our answer!

## Additional Q & A

- Q: What if we kept adding stages?
  - A: p-value  $\rightarrow 1$ , even though  $x_1$  still 13!
- Q: So are p-values really “objective evidence”?
  - A: No, since extra info (like design) critical!
- Q: What about unforeseen events?  
Example: First 5 patients develop an allergic reaction to the treatment – trial is stopped by clinicians.  
Can a frequentist analyze these data?
  - A: No: This aspect of design wasn't anticipated, so p-values not computable!
- Q: Can a Bayesian?
  - A: (obviously) Yes – as we shall see.....

## Conditional (Bayesian) Perspective

- Always condition on data which has actually occurred; the long-run performance of a procedure is of (at most) secondary interest. Fix a prior distribution  $p(\theta)$ , and use Bayes' Theorem (1763):

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta)$$

(“posterior  $\propto$  likelihood  $\times$  prior”)

- Indeed, it often turns out that using the Bayesian formalism with relatively vague priors produces procedures which perform well using traditional *frequentist* criteria (e.g., low mean squared error over repeated sampling)!
  - several examples in Carlin and Louis (2009, Chapter 5)!

## Basics of Bayesian Inference

- As usual, we start with a likelihood (or model)  $f(\mathbf{y}|\theta)$  for the observed data  $\mathbf{y} = (y_1, \dots, y_n)$  given the unknown parameters  $\theta = (\theta_1, \dots, \theta_K)$
- Add a prior distribution  $\pi(\theta|\lambda)$ , where  $\lambda$  is a vector of hyperparameters.
- The posterior distribution for  $\theta$  is given by

$$\begin{aligned} p(\theta|\mathbf{y}, \lambda) &= \frac{p(\mathbf{y}, \theta|\lambda)}{p(\mathbf{y}|\lambda)} = \frac{p(\mathbf{y}, \theta|\lambda)}{\int p(\mathbf{y}, \theta|\lambda) d\theta} \\ &= \frac{f(\mathbf{y}|\theta)\pi(\theta|\lambda)}{\int f(\mathbf{y}|\theta)\pi(\theta|\lambda) d\theta} = \frac{f(\mathbf{y}|\theta)\pi(\theta|\lambda)}{m(\mathbf{y}|\lambda)}. \end{aligned}$$

We refer to this formula as *Bayes' Theorem*.

## Basics of Bayesian Inference

- Since  $\lambda$  will usually not be known, a second stage (hyperprior) distribution  $h(\lambda)$  will be required, so that

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda}) d\boldsymbol{\lambda}}{\int \int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda}) d\boldsymbol{\theta}d\boldsymbol{\lambda}} .$$

- Alternatively, we might replace  $\lambda$  in  $p(\boldsymbol{\theta}|\mathbf{y}, \lambda)$  by an estimate  $\hat{\lambda}$ ; this is called empirical Bayes analysis
- For prediction of a future value  $y_{n+1}$ , we would use the predictive distribution,

$$p(y_{n+1}|\mathbf{y}) = \int p(y_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} ,$$

which is nothing but the posterior of  $y_{n+1}$ .

## Gibbs sampling

- **Gibbs Sampler:** Suppose the joint distribution of  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$  is uniquely determined by the full conditional distributions,  $\{p_i(\theta_i|\theta_{j \neq i}), i = 1, \dots, K\}$ .
- Given an arbitrary set of starting values  $\{\theta_1^{(0)}, \dots, \theta_K^{(0)}\}$ ,

$$\text{Draw } \theta_1^{(1)} \sim p_1(\theta_1|\theta_2^{(0)}, \dots, \theta_K^{(0)}),$$

$$\text{Draw } \theta_2^{(1)} \sim p_2(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_K^{(0)}),$$

⋮

$$\text{Draw } \theta_K^{(1)} \sim p_K(\theta_K|\theta_1^{(1)}, \dots, \theta_{K-1}^{(1)}),$$

- Under mild conditions,

$$(\theta_1^{(t)}, \dots, \theta_K^{(t)}) \xrightarrow{d} (\theta_1, \dots, \theta_K) \sim p \text{ as } t \rightarrow \infty .$$

## Gibbs sampling (cont'd)

- For  $t$  sufficiently large (say, bigger than  $t_0$ ),  $\{\theta^{(t)}\}_{t=t_0+1}^T$  is a (correlated) sample from the true posterior.
- Can use a sample mean to estimate the posterior mean,

$$\hat{E}(\theta_i|\mathbf{y}) = \frac{1}{T - t_0} \sum_{t=t_0+1}^T \theta_i^{(t)} .$$

- The time from  $t = 0$  to  $t = t_0$  is commonly known as the *burn-in* period; one can safely adapt (change) an MCMC algorithm during this pre-convergence period, since these samples will be discarded anyway
- Most popular software package for this: WinBUGS
  - Uses R-like syntax to specify models
  - freely available from <http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>

## Gibbs sampling (cont'd)

- In practice, we may actually run  $m$  *parallel* Gibbs sampling chains, instead of only 1, for some modest  $m$  (say,  $m = 5$ ). Discarding the burn-in period, we obtain

$$\hat{E}(\theta_i|\mathbf{y}) = \frac{1}{m(T - t_0)} \sum_{j=1}^m \sum_{t=t_0+1}^T \theta_{i,j}^{(t)} ,$$

where now the  $j$  subscript indicates chain number.

- If the full conditional  $p(\theta_i|\theta_{j \neq i}, \mathbf{y})$  is not available in closed form, it will typically still be available up to proportionality constant. So WinBUGS uses:
  - adaptive rejection sampling (log-concave densities)
  - slice sampling (bounded domains)
  - Metropolis sampling (all other cases)

## Bayesian estimation

- **Point estimation:** Choose an appropriate measure of centrality: the posterior mean, median, or mode.
- **Interval estimation:** Consider  $q_L$  and  $q_U$ , the  $\alpha/2$ - and  $(1 - \alpha/2)$ -quantiles of  $p(\theta|\mathbf{y})$ :

$$\int_{-\infty}^{q_L} p(\theta|\mathbf{y})d\theta = \alpha/2 \quad \text{and} \quad \int_{q_U}^{\infty} p(\theta|\mathbf{y})d\theta = \alpha/2 .$$

Then clearly  $P(q_L < \theta < q_U|\mathbf{y}) = 1 - \alpha$ ; our confidence that  $\theta$  lies in  $(q_L, q_U)$  is  $100 \times (1 - \alpha)\%$ . Thus this interval is a  $100 \times (1 - \alpha)\%$  credible set (“Bayesian CI”) for  $\theta$ .

- Though not necessarily narrowest, this equal tail interval is easy to compute.
- Unlike frequentist CIs, interpretation of Bayesian CIs is direct: “The probability that  $\theta$  lies in  $(q_L, q_U)$  is  $(1 - \alpha)$ .”

## Bayesian hypothesis testing

- Classical approach bases accept/reject decision on  
 $p\text{-value} = P\{T(\mathbf{Y}) \text{ more “extreme” than } T(\mathbf{y}_{obs})|\theta, H_0\}$  ,  
where “extremeness” is in the direction of  $H_A$
- Bayesian approach: for two models, a commonly used summary historically is the Bayes factor,

$$BF = \frac{P(M_1|\mathbf{y})/P(M_2|\mathbf{y})}{P(M_1)/P(M_2)} = \frac{p(\mathbf{y} | M_1)}{p(\mathbf{y} | M_2)} ,$$

i.e., the likelihood ratio if both hypotheses are simple

- Problem: If  $\pi_i(\theta_i)$  is improper, then  $p(\mathbf{y}|M_i)$  necessarily is as well  $\implies BF$  is not well-defined!...

## Bayesian hypothesis testing via DIC

- A generalization of the Akaike Information Criterion (AIC) to the case of hierarchical models based on the posterior distribution of the deviance statistic,

$$D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y}|\boldsymbol{\theta}) + 2 \log h(\mathbf{y}) ,$$

where  $f(\mathbf{y}|\boldsymbol{\theta})$  is the likelihood and  $h(\mathbf{y})$  is any standardizing function of the data alone

- Summarize the fit of a model by the posterior expectation of the deviance,  $\bar{D} = E_{\theta|y}[D]$
- Summarize the complexity of a model by the effective number of parameters,

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\boldsymbol{\theta}]) = \bar{D} - D(\bar{\boldsymbol{\theta}}) .$$

## Bayesian hypothesis testing via DIC

- The *Deviance Information Criterion* (DIC) is then

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\boldsymbol{\theta}}) ,$$

with smaller values indicating preferred models.

- Both building blocks of DIC and  $p_D$ , that is,  $E_{\theta|y}[D]$  and  $D(E_{\theta|y}[\boldsymbol{\theta}])$ , are easily estimated via MCMC methods, and in fact are automatic within WinBUGS.
- While  $p_D$  has a scale (effective model size), DIC does not, so only differences in DIC across models matter.
- DIC can be sensitive to parametrization and “focus” (i.e., what is considered to be part of the likelihood)
  - $f(\mathbf{y}|\boldsymbol{\theta})$ : “focused on  $\boldsymbol{\theta}$ ”
  - $p(\mathbf{y}|\eta) = \int f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\eta)d\boldsymbol{\theta}$ : “focused on  $\eta$ ”
- Like AIC, DIC tends to select “bigger” models

## Pump Example

Carlin and Louis (2009) Examples 2.7, 3.6  
c.f. WinBUGS Examples Vol. I

$$Y_i | \theta_i \overset{\text{ind}}{\sim} \text{Poisson}(\theta_i t_i),$$
$$\theta_i \overset{\text{ind}}{\sim} G(\alpha, \beta),$$
$$\alpha \sim \text{Exp}(\mu), \beta \sim \text{IG}(c, d),$$

$i = 1, \dots, k$ , where  $\mu, c, d$ , and the  $t_i$  are known, and *Exp* denotes the exponential distribution.

- We apply this model to a dataset giving the numbers of pump failures,  $Y_i$ , observed in  $t_i$  thousands of hours for  $k = 10$  different systems of a certain nuclear power plant.
- The observations are listed in increasing order of raw failure rate  $r_i = Y_i/t_i$ , the classical point estimate of the true failure rate  $\theta_i$  for the  $i^{\text{th}}$  system.

## Pump Data

$i$	$Y_i$	$t_i$	$r_i$
1	5	94.320	.053
2	1	15.720	.064
3	5	62.880	.080
4	14	125.760	.111
5	3	5.240	.573
6	19	31.440	.604
7	1	1.048	.954
8	1	1.048	.954
9	4	2.096	1.910
10	22	10.480	2.099

Hyperparameters: We choose the values  $\mu = 1$ ,  $c = 0.1$ , and  $d = 1.0$ , resulting in reasonably vague hyperpriors for  $\alpha$  and  $\beta$ .

## Pump Example

- Recall that the full conditional distributions for the  $\theta_i$  and  $\beta$  are available in closed form (gamma and inverse gamma, respectively), but that no conjugate prior for  $\alpha$  exists.
- However, the full conditional for  $\alpha$ ,

$$\begin{aligned} p(\alpha|\beta, \{\theta_i\}, \mathbf{y}) &\propto \left[ \prod_{i=1}^k g(\theta_i|\alpha, \beta) \right] h(\alpha) \\ &\propto \left[ \prod_{i=1}^k \frac{\theta_i^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \right] e^{-\alpha/\mu} \end{aligned}$$

can be shown to be log-concave in  $\alpha$ . Thus WinBUGS uses adaptive rejection sampling for this parameter.

## WinBUGS code to fit this model

```
model {
  for (i in 1:k) {
    theta[i] ~ dgamma(alpha,beta)
    lambda[i] <- theta[i]*t[i]
    Y[i] ~ dpois(lambda[i])
  }
  alpha ~ dexp(1.0)
  beta ~ dgamma(0.1, 1.0)
}

DATA:
list(k = 10, Y = c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22),
     t = c(94.320, 15.72, 62.88, 125.76, 5.24, 31.44,
          1.048, 1.048, 2.096, 10.48))

INITS:
list(theta=c(1,1,1,1,1,1,1,1,1,1), alpha=1, beta=1)
```

## Pump Example Results

Results from running 1000 burn-in samples, followed by a “production” run of 10,000 samples (single chain):

node	mean	sd	MC error	2.5%	median	97.5%
alpha	0.7001	0.2699	0.004706	0.2851	0.6634	1.338
beta	0.929	0.5325	0.00978	0.1938	0.8315	2.205
theta[1]	0.0598	0.02542	2.68E-4	0.02128	0.05627	0.1195
theta[5]	0.6056	0.315	0.003087	0.1529	0.5529	1.359
theta[6]	0.6105	0.1393	0.0014	0.3668	0.5996	0.9096
theta[10]	1.993	0.4251	0.004915	1.264	1.958	2.916

- Note that while  $\theta_5$  and  $\theta_6$  have very similar posterior means, the latter posterior is much narrower (smaller sd).
- This is because, while the crude failure rates for the two pumps are similar, the latter is based on a far greater number of hours of observation ( $t_6 = 31.44$ , while  $t_5 = 5.24$ ). Hence we “know” more about pump 6!

## PK Example

Carlin and Louis (2009) Example 2.13, Chapter 4 Problem 2

- Wakefield et al. (1994) consider a dataset for which

$Y_{ij}$  = plasma concentration of the drug Cadralazine

$x_{ij}$  = time elapsed since dose given

where  $i = 1, \dots, 10$  indexes the patient, while

$j = 1, \dots, n_i$  indexes the observations,  $5 \leq n_i \leq 8$ .

- Attempt to fit the one-compartment nonlinear pharmacokinetic (PK) model,

$$\eta_{ij}(x_{ij}) = 30\alpha_i^{-1} \exp(-\beta_i x_{ij}/\alpha_i) .$$

where  $\eta_{ij}(x_{ij})$  is the mean plasma concentration at time  $x_{ij}$ .

## PK Example

- This model is best fit on the log scale, i.e.

$$Z_{ij} \equiv \log Y_{ij} = \log \eta_{ij}(x_{ij}) + \epsilon_{ij} ,$$

where  $\epsilon_{ij} \stackrel{ind}{\sim} N(0, \tau_i)$ .

- The mean structure for the  $Z_{ij}$ 's thus emerges as

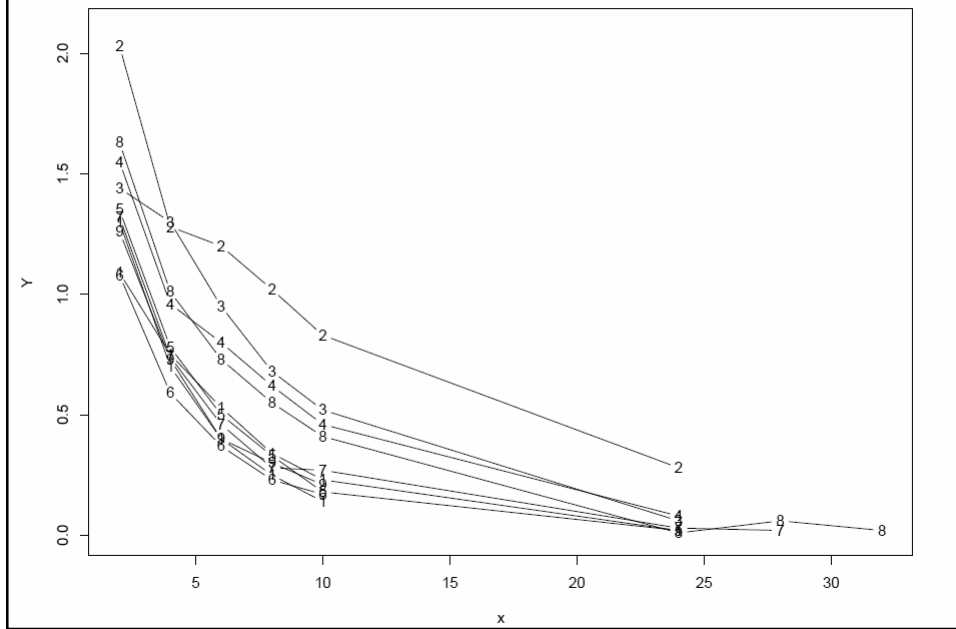
$$\begin{aligned} \log \eta_{ij}(x_{ij}) &= \log [30\alpha_i^{-1} \exp(-\beta_i x_{ij}/\alpha_i)] \\ &= \log 30 - \log \alpha_i - \beta_i x_{ij}/\alpha_i \\ &= \log 30 - a_i - \exp(b_i - a_i)x_{ij} , \end{aligned}$$

where  $a_i = \log \alpha_i$  and  $b_i = \log \beta_i$ .

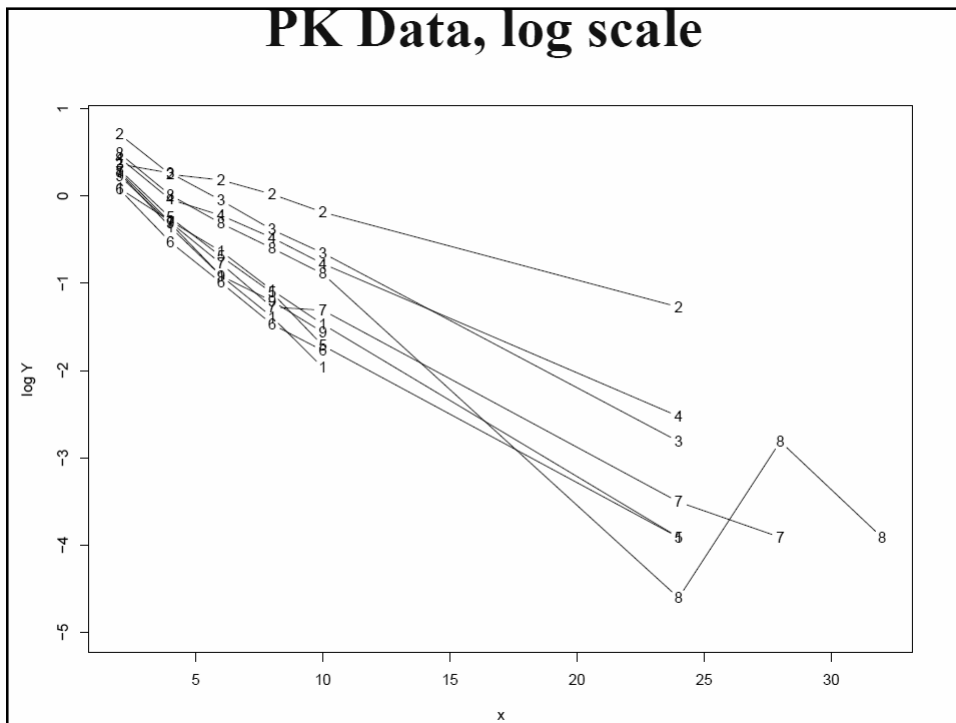
## PK Data

patient	no. of hours following drug administration, $x$							
	2	4	6	8	10	24	28	32
1	1.09	0.75	0.53	0.34	0.23	0.02	–	–
2	2.03	1.28	1.20	1.02	0.83	0.28	–	–
3	1.44	1.30	0.95	0.68	0.52	0.06	–	–
4	1.55	0.96	0.80	0.62	0.46	0.08	–	–
5	1.35	0.78	0.50	0.33	0.18	0.02	–	–
6	1.08	0.59	0.37	0.23	0.17	–	–	–
7	1.32	0.74	0.46	0.28	0.27	0.03	0.02	–
8	1.63	1.01	0.73	0.55	0.41	0.01	0.06	0.02
9	1.26	0.73	0.40	0.30	0.21	–	–	–
10	1.30	0.70	0.40	0.25	0.14	–	–	–

## PK Data, original scale



## PK Data, log scale



## PK Example

- For the subject-specific random effects  $\theta_i \equiv (a_i, b_i)'$ ,

$$\theta_i \stackrel{iid}{\sim} N_2(\boldsymbol{\mu}, \Omega), \text{ where } \boldsymbol{\mu} = (\mu_a, \mu_b).$$

- Usual conjugate prior specification:

$$\boldsymbol{\mu} \sim N_2(\boldsymbol{\lambda}, C)$$

$$\tau_i \stackrel{iid}{\sim} G(\nu_0/2, \nu_0\tau_0/2)$$

$$\Omega \sim \text{Wishart}((\rho R)^{-1}, \rho)$$

- Note that the  $\theta_i$  full conditional distributions are:

- not simple conjugate forms

- not guaranteed to be log-concave

Thus, the Metropolis capability of `winBUGS` is required:

[www.biostat.umn.edu/~brad/data/PKNL\\_BUGS.txt](http://www.biostat.umn.edu/~brad/data/PKNL_BUGS.txt)

## PK Results (WinBUGS vs. Fortran)

parameter	BUGS			Sargent et al. (2000)		
	mean	sd	lag 1 acf	mean	sd	lag 1 acf
$a_1$	2.956	0.0479	0.969	2.969	0.0460	0.947
$a_2$	2.692	0.0772	0.769	2.708	0.0910	0.808
$a_7$	2.970	0.1106	0.925	2.985	0.1360	0.938
$a_8$	2.828	0.1417	0.828	2.838	0.1863	0.934
$b_1$	1.259	0.0335	0.972	1.268	0.0322	0.951
$b_2$	0.234	0.0648	0.661	0.239	0.0798	0.832
$b_7$	1.157	0.0879	0.899	1.163	0.1055	0.925
$b_8$	0.936	0.1458	0.759	0.941	0.1838	0.932
$\tau_1$	362.4	260.4	0.313	380.8	268.8	0.220
$\tau_2$	84.04	57.60	0.225	81.40	58.41	0.255
$\tau_7$	18.87	12.07	0.260	15.82	11.12	0.237
$\tau_8$	2.119	1.139	0.085	1.499	0.931	0.143
$Y_{2,s}$	0.1338	0.0339	0.288	0.1347	0.0264	–
$Y_{7,s}$	0.00891	0.00443	0.178	0.00884	0.00255	–