Foundations of Bayesian Methods and Software for Data Analysis

presented by
Bradley P. Carlin

Division of Biostatistics, School of Public Health, University of Minnesota

TARDIS 2016
University of North Texas, Denton TX
September 9, 2016

Course Outline

- Morning Session I (9:00 - 10:20 am)
  - Introduction: Motivation and overview, potential advantages, key differences with traditional methods (Ch 1, C&L and BCLM texts)
  - Bayesian inference: prior determination, point and interval estimation, hypothesis testing (Ch 2 C&L)

  (Break)
- Morning Session II (10:40 am - 12:00 pm)
  - Bayesian computation: Markov chain Monte Carlo (MCMC) methods, Gibbs sampling, Metropolis algorithm, extensions (Ch 3 C&L)
  - Model criticism and selection: Bayesian robustness, model assessment, and model selection via Bayes factors, predictive approaches, and penalized likelihood methods including DIC (Ch 4 C&L)

  (Lunch)

Course Outline (cont'd)

- Afternoon Session (1:00 - 2:30 pm)
  - MCMC software options: `WinBUGS` and its variants: `R2WinBUGS`, `BRugs`, extensions
  - Computer Lab Session 1: Experimentation with `R` and `WinBUGS` for elementary models (conjugate priors; simple failure rate)

  (Break)
- Afternoon Session II (2:50 - 4:00 pm)
  - Computer Lab Session 2: Experimentation with `WinBUGS` for more advanced models (linear, nonlinear, and logistic regression; random effects; meta-analysis; missing data; etc.)
  - Floor discussion; Q&A; Wrap-up

Textbooks for this course

- Strongly Recommended
  - ("C&L"): *Bayesian Methods for Data Analysis*, 3rd ed., by B.P. Carlin and T.A. Louis, Boca Raton, FL: Chapman and Hall/CRC Press, 2009.
- Recommended:
  - Your favorite math stat and linear models books

Textbooks for this course

- Other books of interest:
  - *Bayesian Data Analysis*, 3rd ed., by A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, Boca Raton, FL: Chapman and Hall/CRC Press, 2013.
  - *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, 2nd ed., by John Kruschke, New York: Academic Press, 2014.
  - *The BUGS Book: A Practical Introduction to Bayesian Analysis*, by D. Lunn, C. Jackson, N. Best, A. Thomas, and D.J. Spiegelhalter, Boca Raton: Chapman and Hall/CRC Press, 2012.
  - *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed., by S. Banerjee, B. Carlin, and A.E. Gelfand, Boca Raton, FL: Chapman and Hall/CRC Press, 2014.
  - *Bayesian Adaptive Methods for Clinical Trials* by S.M. Berry, B.P. Carlin, J.J. Lee, and P. Müller, Boca Raton, FL: Chapman and Hall/CRC Press, 2010.

Ch 1: Overview

- Biostatisticians in the drug and medical device industries 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

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.

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

Some preliminary Q&A

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

$$\text{posterior information} \;\geq\; \text{prior information} \;\geq\; 0\,,$$

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 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	$\star$	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for $\star$, 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,

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

Let X = # Study B patients with no AE ("successes").

- If the prior is $\theta \sim Beta(a = 110, b = 7)$ (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 θ 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 2010 Guidance Document, `www.fda.gov/cdrh/osb/guidance/1601.html`
  - see also the new Bayesian clinical trials textbook by Berry, Carlin, Lee, and Müller (CRC Press, 2010)!

Bayesians have a problem with p-values

- $p =$ Pr(results as surprising as you got *or more so*).
  The “or more so" part gets us in trouble with:
  - *The Likelihood Principle:* When making decisions, only the observed data can play a role.

  This can lead to bad decisions (esp. false positives)
- Are p-values at least more objective, because they are not influenced by any prior distribution?
  - No, because they are influenced crucially by the design of the experiment, which determines the reference space of events for the calculation.
- Purely practical problems also plague p-values:
  - Ex: Unforeseen events: First 5 patients develop a rash, and the trial is stopped by clinicians.
    $\Longrightarrow$ this aspect of design wasn’t anticipated, so strictly speaking, the p-value is not computable!

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 $p(\boldsymbol{\theta})$, and use Bayes' Theorem (1763):

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\mathbf{x}|\theta)p(\boldsymbol{\theta}) \\ \text{(“posterior} &\propto \text{likelihood} \times \text{prior”)} \end{aligned}$$

- 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)!

Bayesian Advantages in Inference

- Ability to formally incorporate prior information
- Probabilities of parameters; answers are more easily interpretable (e.g., confidence intervals)
- All analyses follow directly from the posterior; no separate theories of estimation, testing, multiple comparisons, etc. are needed
- Role of randomization: minimizes the possibility of selection bias, balances treatment groups over covariates... but does not serve as the basis of inference (which is model-based, not design-based)
- Inferences are conditional on the actual data
- Bayes procedures possess many optimality properties (e.g. consistent, impose parsimony in model choice, define the class of optimal frequentist procedures, ...)

Ch 2: Basics of Bayesian Inference

- Start with the discrete finite case: Suppose we have some event of interest A and a collection of other events $B_j,\ j = 1, \ldots, J$ that are mutually exclusive and exhaustive (that is, exactly one of them must occur).
- Given the event probabilities $P(B_j)$ and the conditional probabilities $P(A|B_j)$, Bayes' Rule states

$$\begin{aligned} P(B_j|A) &= \frac{P(A, B_j)}{P(A)} = \frac{P(A, B_j)}{\sum_{j=1}^{J} P(A, B_j)} \\ &= \frac{P(A|B_j)P(B_j)}{\sum_{j=1}^{J} P(A|B_j)P(B_j)}, \end{aligned}$$

where $P(A, B_j) = P(A \cap B_j)$ indicates the *joint* event where both A and B_j occur.

Example in Discrete Finite Case

Example: Ultrasound tests for determining a baby's gender.
When reading preliminary ultrasound results, errors are not "symmetric" in the following sense: girls are virtually always correctly identified as girls, but boys are sometimes misidentified as girls.

Suppose a leading radiologist states that

$$P(test+|G) = 1 \text{ and } P(test+|B) = .25 ,$$

where "test +" denotes that the ultrasound test predicts the child is a girl. Thus, we have a 25% false positive rate for girl, but no false negatives.

Question: Suppose a particular woman's test comes back positive for girl. Assuming 48% of babies are girls, what is the probability she is actually carrying a girl?

Example in Discrete Finite Case (cont'd)

Solution: Let "boy" and "girl" provide the $J = 2$ mutually exclusive and exhaustive cases B_j, and let A being the event of a positive test.

Then by Bayes' Rule we have

$$\begin{aligned} P(G \mid test+) &= \frac{P(test+|G)P(G)}{P(test+|G)P(G) + P(test+|B)P(B)} \\ &= \frac{(1)(.48)}{(1)(.48) + (.25)(.52)} = .787 \ , \end{aligned}$$

or only a 78.7% chance the baby is, in fact, a girl.

Bayes in the Continuous Case

- Now start with a likelihood (or model) $f(\mathbf{y}|\boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ given the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, where the parameters are continuous (meaning they can take an infinite number of possible values)
- Add a prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of hyperparameters.
- The posterior distribution for $\boldsymbol{\theta}$ is given by

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

We refer to this continuous version of Bayes' Rule as *Bayes' Theorem*.

Bayes in the Continuous Case (cont'd)

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

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

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

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

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

Illustration of Bayes' Theorem

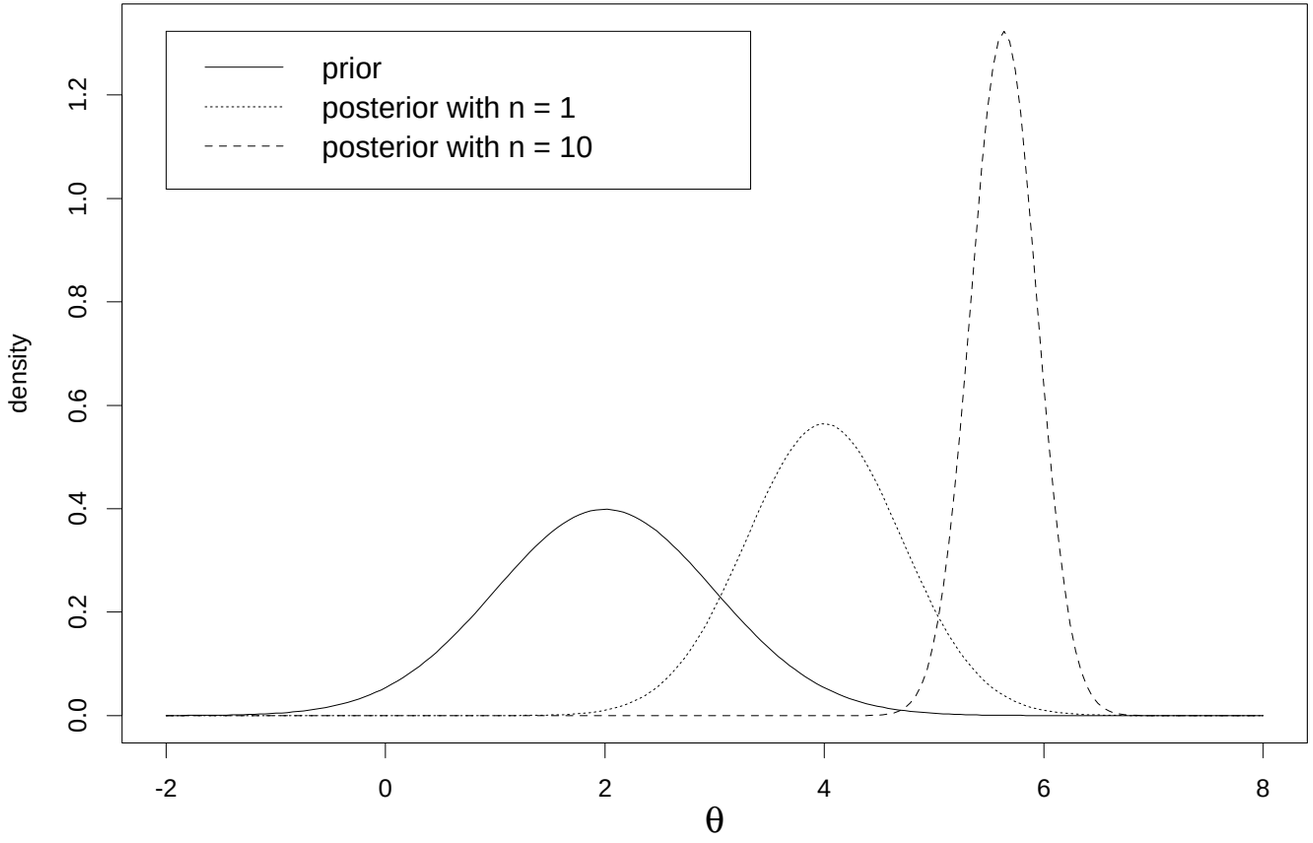
- Suppose $f(y|\theta) = N(y|\theta, \sigma^2)$, $\theta \in \Re$ and $\sigma > 0$ known
- If we take $\pi(\theta|\boldsymbol{\lambda}) = N(\theta|\mu, \tau^2)$ where $\boldsymbol{\lambda} = (\mu, \tau)'$ is fixed and known, then it is easy to show that

$$p(\theta|y) = N\left(\theta \,\Big|\, \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y\,,\; \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) .$$

- Note that
  - The posterior mean $E(\theta|y)$ is a weighted average of the prior mean μ and the data value y, with weights depending on our relative uncertainty
  - the posterior precision (reciprocal of the variance) is equal to $1/\sigma^2 + 1/\tau^2$, which is the sum of the likelihood and prior precisions.
- R and BUGS code for this: first two entries at `http://www.biostat.umn.edu/~brad/data.html`

Illustration (continued)

As a concrete example, let $\mu = 2, \tau = 1, \bar{y} = 6$, and $\sigma = 1$:



- When $n = 1$, prior and likelihood receive equal weight
- When $n = 10$, the data dominate the prior
- The posterior variance goes to zero as $n \to \infty$

Addendum: Notes on prior distributions

- The prior here is *conjugate:* it leads to a posterior distribution for θ that is available in closed form, and is a member of the same distributional family as the prior.
- Note that setting $\tau^2 = \infty$ corresponds to an arbitrarily vague (or noninformative) prior. The posterior is then

$$p\left(\theta|y\right) = N\left(\theta|\overline{y}, \sigma^2/n\right) ,$$

  the same as the likelihood! The limit of the conjugate (normal) prior here is a uniform (or "flat") prior, and thus the posterior is the renormalized likelihood.
- The flat prior is appealing but improper here, since $\sum_\theta p(\theta) = +\infty$. However, the posterior is still well defined, and so improper priors are often used!

Quick preview: Hierarchical modeling

- The hyperprior for $\boldsymbol{\eta}$ might itself depend on a collection of unknown parameters $\boldsymbol{\lambda}$, resulting in a generalization of our three-stage model to one having a third-stage prior $h(\boldsymbol{\eta}|\boldsymbol{\lambda})$ and a fourth-stage hyperprior $g(\boldsymbol{\lambda})$...
- This enterprise of specifying a model over several levels is called hierarchical modeling, which is often helpful when the data are nested:
- **Example:** Test scores Y_{ijk} for student k in classroom j of school i:

$$\begin{aligned} Y_{ijk}|\theta_{ij} &\sim N(\theta_{ij}, \sigma^2) \\ \theta_{ij}|\mu_i &\sim N(\mu_i, \tau^2) \\ \mu_i|\lambda &\sim N(\lambda, \kappa^2) \end{aligned}$$

Adding $p(\lambda)$ and possibly $p(\sigma^2, \tau^2, \kappa^2)$ completes the specification!

Prediction

- Returning to two-level models, we often write

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ ,$$

since the likelihood may be multiplied by any constant (or any function of $\mathbf{y}$ alone) without altering $p(\boldsymbol{\theta}|\mathbf{y})$.

- If y_{n+1} is a future observation, independent of $\mathbf{y}$ given $\boldsymbol{\theta}$, then the predictive distribution for y_{n+1} is

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

thanks to the conditional independence of y_{n+1} and $\mathbf{y}$.

- The naive frequentist would use $f(y_{n+1}|\widehat{\boldsymbol{\theta}})$ here, which is correct only for large n (i.e., when $p(\boldsymbol{\theta}|\mathbf{y})$ is a point mass at $\widehat{\boldsymbol{\theta}}$).

Prior Distributions

- Suppose we require a prior distribution for

  $\theta =$ true proportion of U.S. men who are HIV-positive.

- We cannot appeal to the usual long-term frequency notion of probability – it is not possible to even imagine "running the HIV epidemic over again" and reobserving θ. Here θ is random only because it is unknown to us.
- Bayesian analysis is predicated on such a belief in subjective probability and its quantification in a prior distribution $p(\theta)$. But:
  - How to create such a prior?
  - Are "objective" choices available?

Elicited Priors

- Histogram approach: Assign probability masses to the "possible" values in such a way that their sum is 1, and their relative contributions reflect the experimenter's prior beliefs as closely as possible.
  - BUT: Awkward for continuous or unbounded θ.
- Matching a functional form: Assume that the prior belongs to a parametric distributional family $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, choosing $\boldsymbol{\eta}$ so that the result matches the elicitee's true prior beliefs as nearly as possible.
  - This approach limits the effort required of the elicitee, and also overcomes the finite support problem inherent in the histogram approach...
  - BUT: it may not be possible for the elicitee to "shoehorn" his or her prior beliefs into any of the standard parametric forms.

Conjugate Priors

- Defined as one that leads to a posterior distribution belonging to the same distributional family as the prior.
- Conjugate priors were historically prized for their computational convenience, but the emergence of modern computing methods and software (e.g., `WinBUGS`) has greatly reduced our need for them.
- Still, they remain popular, due both to historical precedent and a desire to make our modern computing methods as fast as possible: in high-dimensional problems, priors that are conditionally conjugate are often available (and helpful).
- a finite mixture of conjugate priors may be sufficiently flexible (allowing multimodality, heavier tails, etc.) while still enabling simplified posterior calculations.

Noninformative Prior

– is one that does not favor one θ value over another

- Examples:
  - $\Theta = \{\theta_1, \ldots, \theta_n\} \Rightarrow p(\theta_i) = 1/n,\ i = 1, \ldots, n$
  - $\Theta = [a, b],\ -\infty < a < b < \infty$
    $\Rightarrow p(\theta) = 1/(b-a),\ a < \theta < b$
  - $\Theta = (-\infty, \infty) \Rightarrow p(\theta) = c,\ \text{any } c > 0$

    This is an improper prior (does not integrate to 1), but its use can still be legitimate if $\sum_\theta f(\mathbf{x}|\theta) = K < \infty$, since then

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) \cdot c}{\sum_\theta f(\mathbf{x}|\theta) \cdot c} = \frac{f(\mathbf{x}|\theta)}{K},$$

    so the posterior is just the renormalized likelihood!

Bayesian Inference: Point Estimation

- Easy! Simply choose an appropriate distributional summary: posterior mean, median, or mode.
- Mode is often easiest to compute (no integration), but is often least representative of “middle”, especially for one-tailed distributions.
- Mean has the opposite property, tending to "chase" heavy tails (just like the sample mean $\bar{X}$)
- Median is probably the best compromise overall, though can be awkward to compute, since it is the solution θ^{median} to

$$\sum_{\theta=-\infty}^{\theta^{median}} p(\theta|x) = \frac{1}{2} \; .$$

Example: The General Linear Model

- Let $\mathbf{Y}$ be an $n \times 1$ data vector, X an $n \times p$ matrix of covariates, and adopt the likelihood and prior structure,

$$\mathbf{Y}|\boldsymbol{\beta} \sim N_n\left(X\boldsymbol{\beta}, \Sigma\right) \text{ and } \boldsymbol{\beta} \sim N_p\left(A\boldsymbol{\alpha}, V\right)$$

- Then the posterior distribution of $\boldsymbol{\beta}|\mathbf{Y}$ is

$$\boldsymbol{\beta}|Y \sim N\left(D\mathbf{d}, D\right) \text{ , where}$$

$D^{-1} = X^T\Sigma^{-1}X + V^{-1}$ and $\mathbf{d} = X^T\Sigma^{-1}\mathbf{Y} + V^{-1}A\boldsymbol{\alpha}$.

- $V^{-1} = 0$ delivers a "flat" prior; if $\Sigma = \sigma^2 I_p$, we get

$$\boldsymbol{\beta}|Y \sim N\left(\hat{\boldsymbol{\beta}}\,,\, \sigma^2(X'X)^{-1}\right) \text{ , where}$$

$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \Longleftrightarrow$ usual likelihood approach!

Bayesian Inference: Interval Estimation

- The Bayesian analogue of a frequentist CI is referred to as a credible set: a $100 \times (1-\alpha)$% credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$1 - \alpha \leq P(C|\mathbf{y}) = \sum_{\boldsymbol{\theta} \in C} p(\boldsymbol{\theta}|\mathbf{y}) \ .$$

- In continuous settings, we can obtain coverage exactly $1-\alpha$ at minimum size via the highest posterior density (HPD) credible set,

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \ p(\boldsymbol{\theta}|\mathbf{y}) \geq k(\alpha)\} \ ,$$

where $k(\alpha)$ is the largest constant such that

$$P(C|\mathbf{y}) \geq 1 - \alpha \ .$$

Interval Estimation (cont'd)

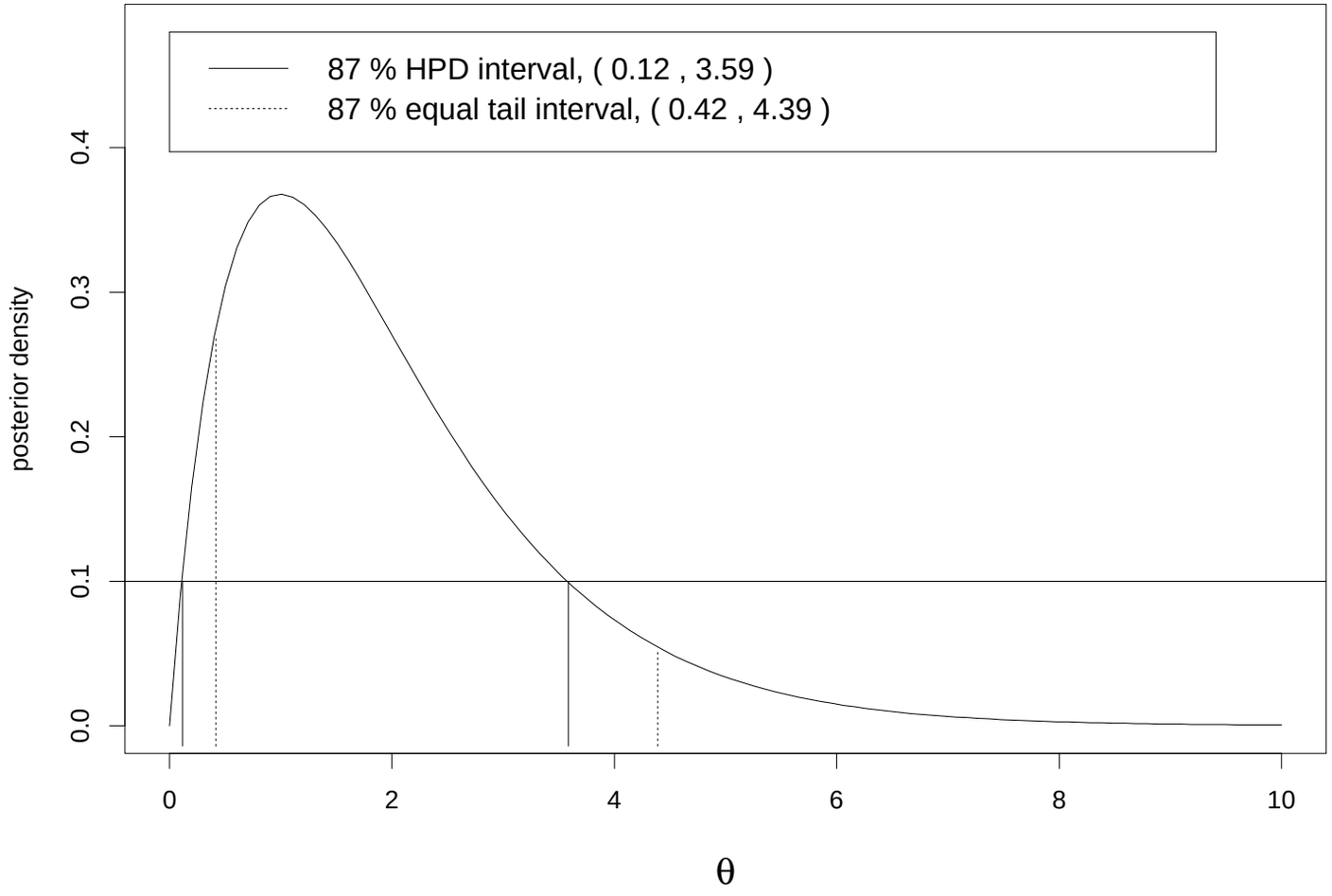
- Simpler alternative: the equal-tail set, which takes the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$.
- Specifically, consider q_L and q_U, the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$:

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

  Then clearly $P(q_L < \theta < q_U|\mathbf{y}) = 1-\alpha$; our confidence that θ 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 θ.
- This interval is relatively easy to compute, and enjoys a direct interpretation ("The probability that θ lies in (q_L, q_U) is $(1-\alpha)$") that the frequentist interval does not.

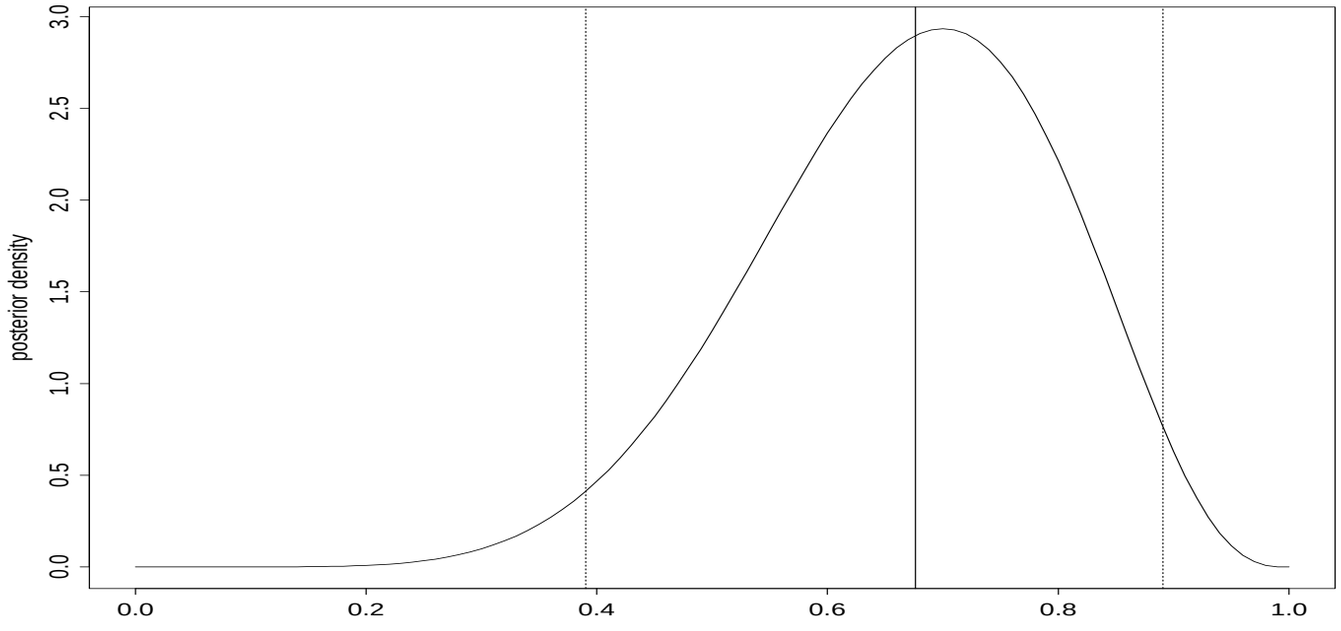
Interval Estimation: Example

Using a $Gamma(2, 1)$ posterior distribution and $k(\alpha) = 0.1$:



Equal tail interval is a bit wider, but easier to compute (just two gamma quantiles), and also transformation invariant.

Ex: $Y \sim Bin(10, \theta)$**,** $\theta \sim U(0, 1)$**,** $y_{obs} = 7$



Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A
- Several *troubles* with this approach:
  - hypotheses must be nested
  - p-value can only offer evidence against the null
  - p-value is not the "probability that H_0 is true" (but is often erroneously interpreted this way)
  - As a result of the dependence on "more extreme" $T(\mathbf{Y})$ values, two experiments with different designs but identical likelihoods could result in different p-values, violating the Likelihood Principle!

Bayesian hypothesis testing (cont'd)

- Bayesian approach: Select the model with the largest posterior probability, $P(M_i|\mathbf{y}) = p(\mathbf{y}|M_i)p(M_i)/p(\mathbf{y})$,

$$\text{where} \quad p(\mathbf{y}|M_i) = \sum_{\boldsymbol{\theta}_i} f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i) \ .$$

- For two models, the quantity commonly used to summarize these results is the Bayes factor,

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

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

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

Bayesian hypothesis testing (cont'd)

When the BF is not well-defined, several alternatives:

- Modify the definition of BF: partial Bayes factor, fractional Bayes factor (text, p.54)
- Switch to the conditional predictive distribution,

$$f(y_i|\mathbf{y}_{(i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(i)})} = \sum_{\boldsymbol{\theta}} f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) \ ,$$

  which will be proper if $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is. Assess model fit via plots or a suitable summary (say, $\prod_{i=1}^{n} f(y_i|\mathbf{y}_{(i)})$).
- Penalized likelihood criteria: the Akaike information criterion (AIC), Bayesian information criterion (BIC), or Deviance information criterion (DIC).
- IOU on all this – Chapter 4!

Example: Consumer preference data

- Suppose 16 taste testers compare two types of ground beef patty (one stored in a deep freeze, the other in a less expensive freezer). The food chain is interested in whether storage in the higher-quality freezer translates into a "substantial improvement in taste."
- Experiment: In a test kitchen, the patties are defrosted and prepared by a single chef/statistician, who randomizes the order in which the patties are served in double-blind fashion.
- Result: 13 of the 16 testers state a preference for the more expensive patty.

Example: Consumer preference data

- Likelihood: Let

$$\theta = \text{prob. consumers prefer more expensive patty}$$

$$Y_i = \begin{cases} 1 & \text{if tester } i \text{ prefers more expensive patty} \\ 0 & \text{otherwise} \end{cases}$$

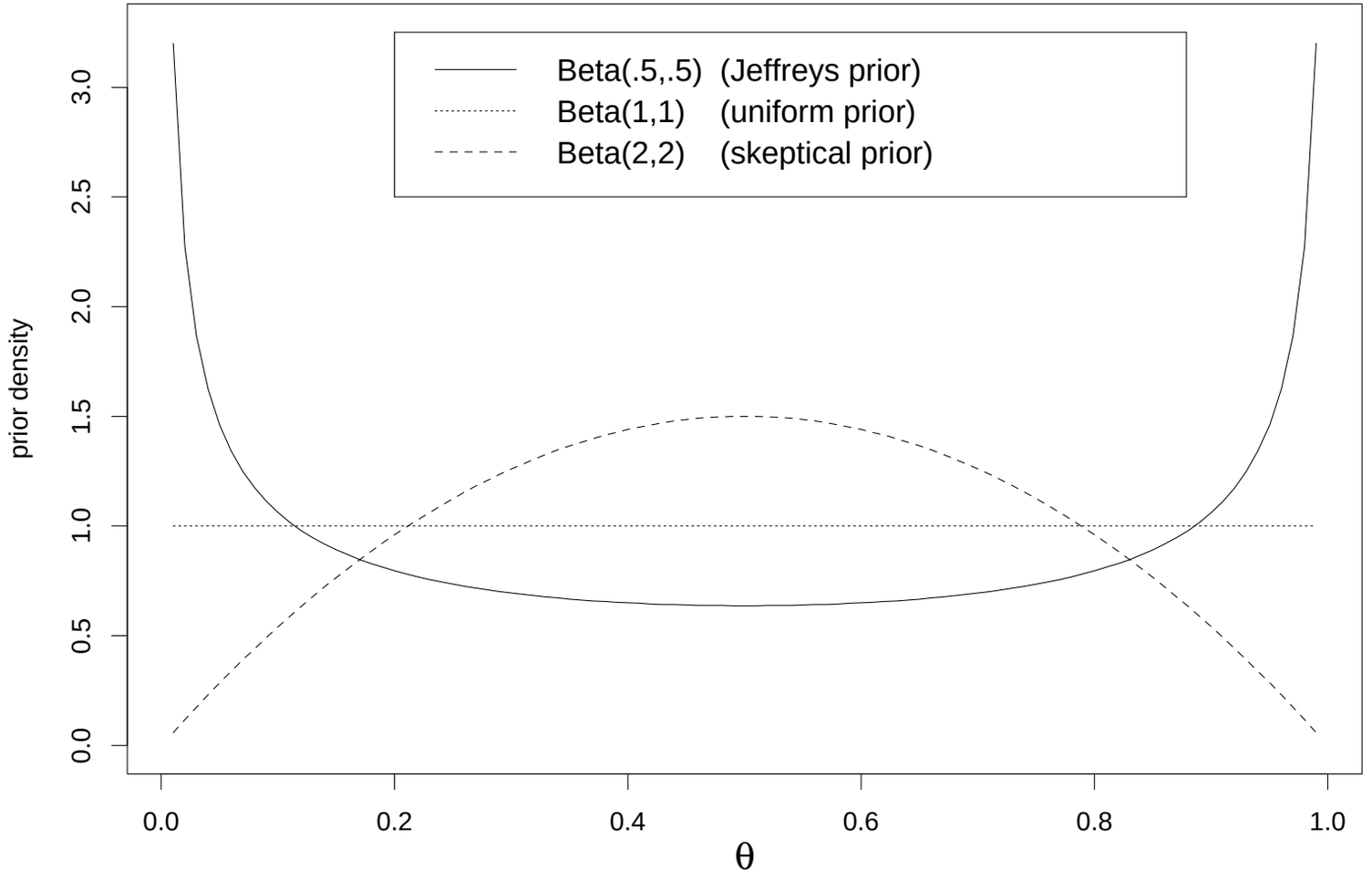
- Assuming independent testers and constant θ, then if $X = \sum_{i=1}^{16} Y_i$, we have $X|\theta \sim Binomial(16, \theta)$,

$$f(x|\theta) = \binom{16}{x} \theta^x (1-\theta)^{16-x} .$$

- The beta distribution offers a conjugate family, since

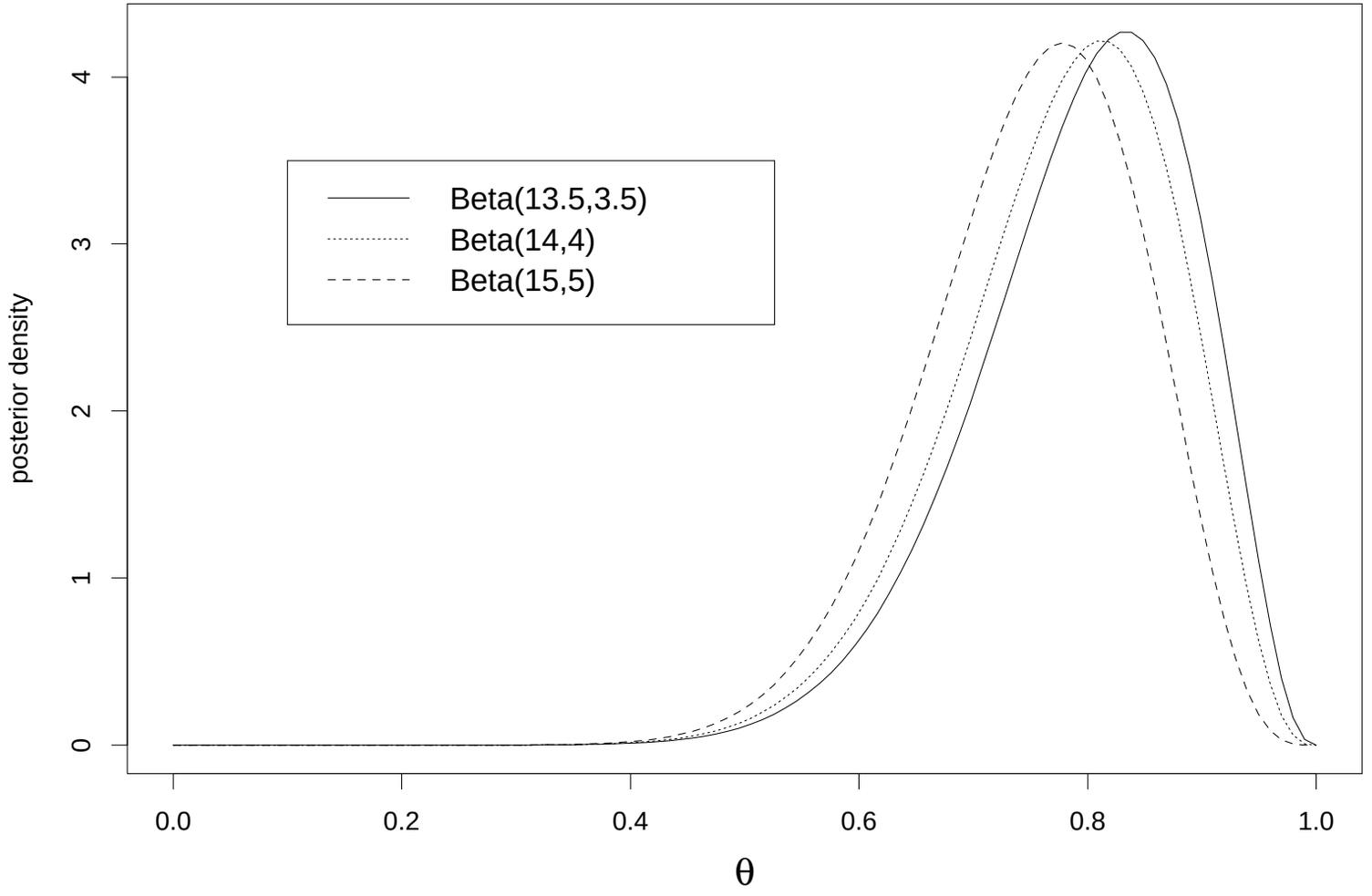
$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} .$$

Three "minimally informative" priors



The posterior is then $Beta(x + \alpha, 16 - x + \beta)$...

Three corresponding posteriors



- Note ordering of posteriors; consistent with priors.
- All three produce 95% equal-tail credible intervals that exclude 0.5 $\Rightarrow$ there is an improvement in taste.

Posterior summaries

Prior distribution	Posterior quantile .025	.500	.975	$P(\theta > .6\|x)$
$Beta(.5, .5)$	0.579	0.806	0.944	0.964
$Beta(1, 1)$	0.566	0.788	0.932	0.954
$Beta(2, 2)$	0.544	0.758	0.909	0.930

- Suppose we define "*substantial* improvement in taste" as $\theta \geq 0.6$. Then under the uniform prior, the Bayes factor in favor of $M_1 : \theta \geq 0.6$ over $M_2 : \theta < 0.6$ is

$$BF = \frac{0.954/0.046}{0.4/0.6} = 31.1 \ ,$$

or fairly strong evidence (adjusted odds about 30:1) in favor of a substantial improvement in taste.

Bayesian computation

- prehistory (1763 – 1960): Conjugate priors
- 1960's: Numerical quadrature – Newton-Cotes methods, Gaussian quadrature, etc.
- 1970's: Expectation-Maximization ("EM") algorithm – iterative mode-finder
- 1980's: Asymptotic methods – Laplace's method, saddlepoint approximations
- 1980's: Noniterative Monte Carlo methods – Direct posterior sampling and indirect methods (importance sampling, rejection, etc.)
- 1990's: Markov chain Monte Carlo (MCMC) – Gibbs sampler, Metropolis-Hastings algorithm

$\Rightarrow$ MCMC methods broadly applicable, but require care in parametrization and convergence diagnosis!

Asymptotic methods

- When n is large, $f(\mathbf{x}|\boldsymbol{\theta})$ will be quite peaked relative to $p(\boldsymbol{\theta})$, and so $p(\boldsymbol{\theta}|\mathbf{x})$ will be approximately normal.
- "Bayesian Central Limit Theorem": Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_i(x_i|\boldsymbol{\theta})$, and that the prior $p(\boldsymbol{\theta})$ and the likelihood $f(\mathbf{x}|\boldsymbol{\theta})$ are positive and twice differentiable near $\widehat{\boldsymbol{\theta}}^p$, the posterior mode of $\boldsymbol{\theta}$. Then for large n

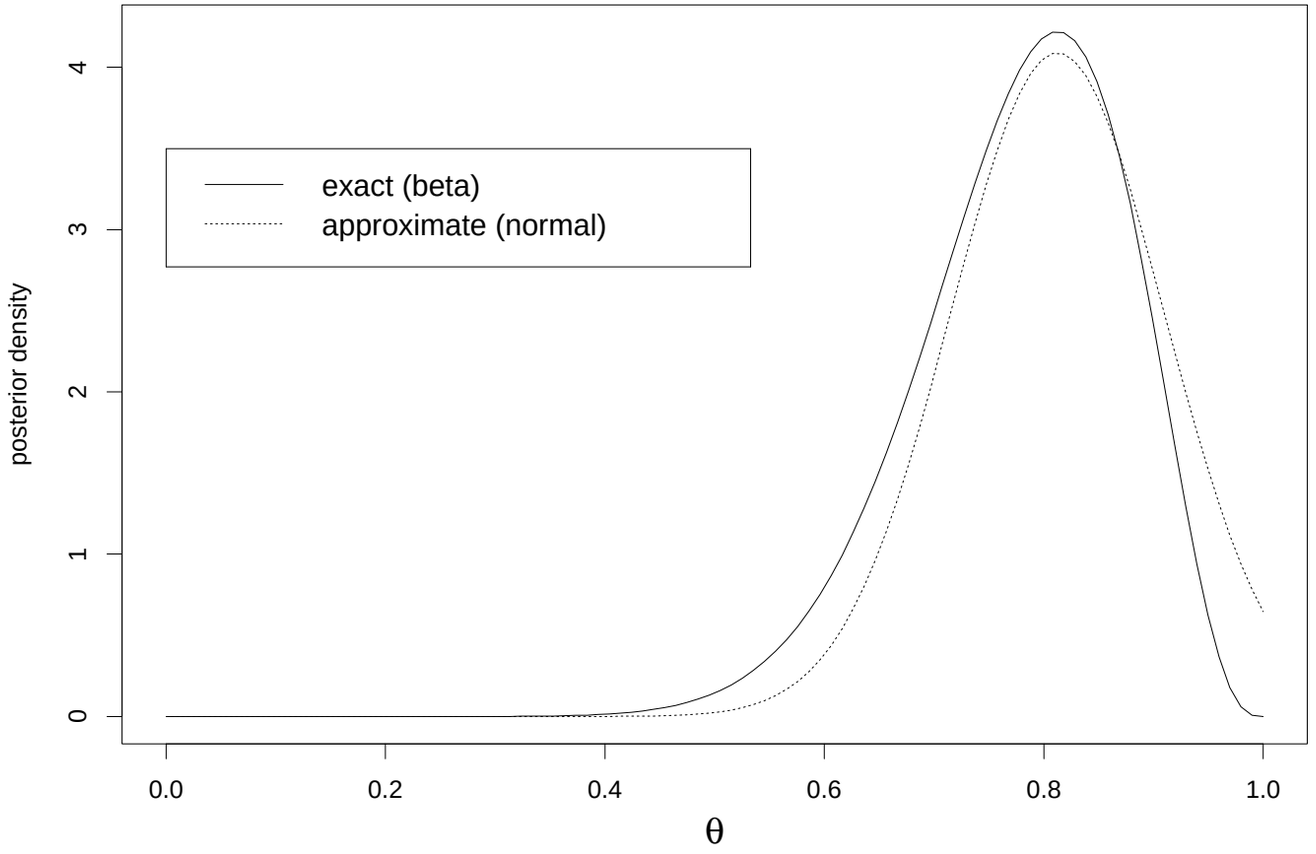
$$p(\boldsymbol{\theta}|\mathbf{x}) \dot{\sim} N(\widehat{\boldsymbol{\theta}}^p, [I^p(\mathbf{x})]^{-1}) \ ,$$

where $[I^p(\mathbf{x})]^{-1}$ is the "generalized" observed Fisher information matrix for $\boldsymbol{\theta}$, i.e., minus the inverse Hessian of the log posterior evaluated at the mode,

$$I^p_{ij}(\mathbf{x}) = - \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log \left(f(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \right) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^p} \ .$$

Example 3.1: Hamburger patties again

Comparison of this normal approximation to the exact posterior, a $Beta(14, 4)$ distribution (recall $n = 16$):



Similar modes, but very different tail behavior: 95% credible sets are (.57, .93) for exact, but (.62, 1.0) for normal approximation.

Higher order approximations

- The Bayesian CLT is a first order approximation, since

$$E(g(\boldsymbol{\theta})) = g(\hat{\boldsymbol{\theta}})\left[1 + O\left(1/n\right)\right] \ .$$

- Second order approximations (i.e., to order $O(1/n^2)$) again requiring only mode and Hessian calculations are available via Laplace's Method (C&L, Sec. 3.2.2).
- Advantages of Asymptotic Methods:
  - deterministic, noniterative algorithm
  - substitutes differentiation for integration
  - facilitates studies of Bayesian robustness
- Disadvantages of Asymptotic Methods:
  - requires well-parametrized, unimodal posterior
  - $\boldsymbol{\theta}$ must be of at most moderate dimension
  - n must be large, *but is beyond our control*

Gibbs sampling

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

$$\begin{aligned}
\text{Draw } \theta_1^{(1)} &\sim p_1(\theta_1|\theta_2^{(0)}, \ldots, \theta_K^{(0)}), \\
\text{Draw } \theta_2^{(1)} &\sim p_2(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_K^{(0)}), \\
&\vdots \\
\text{Draw } \theta_K^{(1)} &\sim p_K(\theta_K|\theta_1^{(1)}, \ldots, \theta_{K-1}^{(1)}),
\end{aligned}$$

- Under mild conditions,

$$(\theta_1^{(t)}, \ldots, \theta_K^{(t)}) \xrightarrow{d} (\theta_1, \cdots, \theta_K) \sim p \text{ as } t \to \infty\ .$$

Gibbs sampling (cont'd)

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

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

- 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 preconvergence period, since these samples will be discarded anyway

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

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

  where now the j subscript indicates chain number.

- A density estimate $\hat{p}(\theta_i|\mathbf{y})$ may be obtained by smoothing the histogram of the $\{\theta_{i,j}^{(t)}\}$, or as

$$\begin{aligned} \hat{p}(\theta_i|\mathbf{y}) &= \frac{1}{m(T-t_0)} \sum_{j=1}^{m} \sum_{t=t_0+1}^{T} p(\theta_i|\theta_{k\neq i,j}^{(t)}, \mathbf{y}) \\ &\approx \int p(\theta_i|\theta_{k\neq i}, \mathbf{y}) p(\theta_{k\neq i}|\mathbf{y}) d\theta_{k\neq i} \end{aligned}$$

Example 3.6 (2.7 revisited)

- Consider the model

$$Y_i|\theta_i \overset{ind}{\sim} Poisson(\theta_i s_i),\ \theta_i \overset{ind}{\sim} G(\alpha,\beta),$$
$$\beta \sim IG(c,d),\ i=1,\ldots,k,$$

where α, c, d, and the s_i are known. Thus

$$\begin{aligned} f(y_i|\theta_i) &= \frac{e^{-(\theta_i s_i)}(\theta_i s_i)^{y_i}}{y_i!},\ y_i \geq 0,\ \theta_i > 0, \\ g(\theta_i|\beta) &= \frac{\theta_i^{\alpha-1} e^{-\theta_i/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\ \alpha > 0,\ \beta > 0, \\ h(\beta) &= \frac{e^{-1/(\beta d)}}{\Gamma(c) d^c \beta^{c+1}},\ c > 0,\ d > 0. \end{aligned}$$

Note: g is conjugate for f, and h is conjugate for g

Example 3.6 (2.7 revisited)

- To implement the Gibbs sampler, we require the full conditional distributions of β and the θ_i.
- By Bayes' Rule, each of these is proportional to the complete Bayesian model specification,

$$\left[\prod_{i=1}^{k} f(y_i|\theta_i)g(\theta_i|\beta)\right] h(\beta)$$

- Thus we can find full conditional distributions by dropping irrelevant terms from this expression, and normalizing!

Good news: BUGS will do all this math for you! :)

Example 3.6 (2.7 revisited)

- BUGS can sample the $\{\theta_i^{(t)}\}$ and $\beta^{(t)}$ directly
- If α were also unknown, harder for BUGS since

$$p(\alpha|\{\theta_i\},\beta,\mathbf{y}) \propto \left[\prod_{i=1}^{k} g(\theta_i|\alpha,\beta)\right] h(\alpha)$$

is not proportional to any standard family. So resort to:

- adaptive rejection sampling (ARS): provided $p(\alpha|\{\theta_i\},\beta,\mathbf{y})$ is log-concave, or
- Metropolis-Hastings sampling – IOU for now!

Note: This is the order the `WinBUGS` software uses when deriving full conditionals!

This is the standard "hybrid approach": Use Gibbs overall, with "substeps" for awkward full conditionals

Example 7.2: Rat data

- Consider the longitudinal data model

$$Y_{ij} \overset{ind}{\sim} N\left(\alpha_i + \beta_i x_{ij}\ ,\ \sigma^2\right)\ ,$$

where Y_{ij} is the weight of the i^{th} rat at measurement point j, while x_{ij} denotes its age in days, for $i = 1, \ldots, k = 30$, and $j = 1, \ldots, n_i = 5$ for all i (see text p.337 for actual data).

- Adopt the random effects model

$$\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}\ ,\ \Sigma\right),\ i = 1, \ldots, k\ ,$$

which is conjugate with the likelihood (see general normal linear model in Section 4.1.1).

Example 7.2: Rat data

- Priors: Conjugate forms are again available, namely

$$\begin{aligned} \sigma^2 &\sim IG(a,\, b)\ , \\ \boldsymbol{\theta}_0 &\sim N(\boldsymbol{\eta},\, C)\ , \quad \text{and} \\ \Sigma^{-1} &\sim W\left((\rho R)^{-1},\, \rho\right)\ , \end{aligned} \tag{1}$$

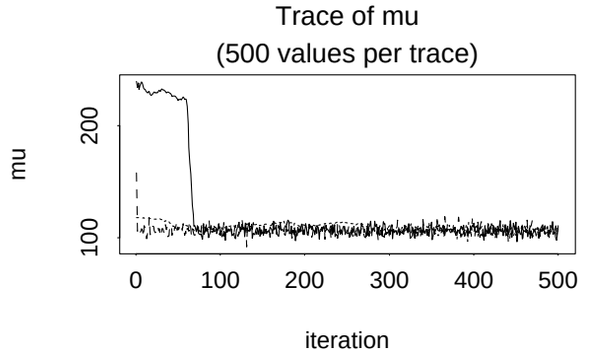
  where W denotes the Wishart (multivariate gamma) distribution; see Appendix A.2.2.

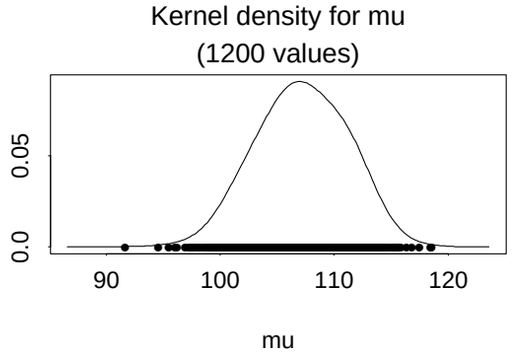
- We assume the hyperparameters ($a, b, \boldsymbol{\eta}, C, \rho$, and R) are all known, so there are 30(2) + 3 + 3 = 66 unknown parameters in the model.

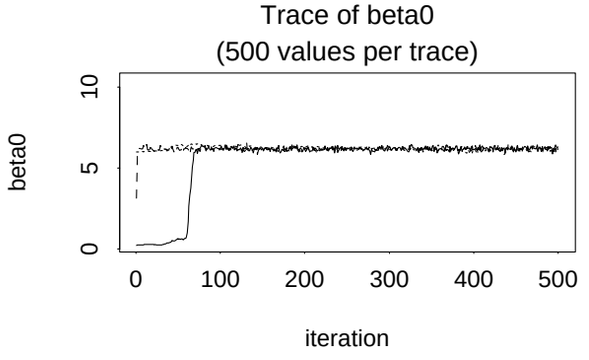
  Yet the Gibbs sampler is relatively straightforward to implement here, thanks to the conjugacy at each stage in the hierarchy.

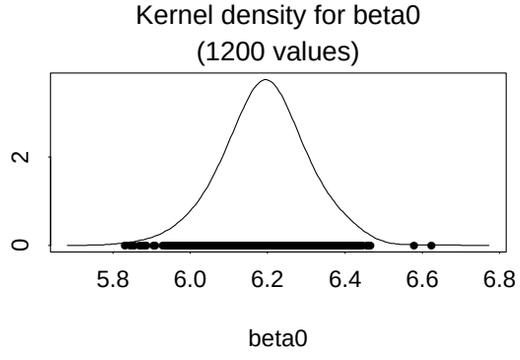
Example 7.2: Rat data

- Using vague hyperpriors, run 3 initially overdispersed parallel sampling chains for 500 iterations each:









- The output from all three chains over iterations 101–500 is used in the posterior kernel density estimates (col 2)
- The average rat weighs about 106 grams at birth, and gains about 6.2 grams per day.

Metropolis algorithm

- What happens if the full conditional $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$ is not available in closed form? Typically, $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$ will be available up to proportionality constant, since it is proportional to the portion of the Bayesian model (likelihood times prior) that involves θ_i.
- Suppose the true joint posterior for $\boldsymbol{\theta}$ has unnormalized density $p(\boldsymbol{\theta})$.
- Choose a candidate density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ that is a valid density function for every possible value of the conditioning variable $\boldsymbol{\theta}^{(t-1)}$, and satisfies

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*) \ ,$$

  i.e., q is symmetric in its arguments.

Metropolis algorithm (cont'd)

Given a starting value $\boldsymbol{\theta}^{(0)}$ at iteration $t = 0$, the algorithm proceeds as follows:

- **Metropolis Algorithm:** For ($t \in 1 : T$), repeat:
  1. Draw $\boldsymbol{\theta}^*$ from $q(\cdot|\boldsymbol{\theta}^{(t-1)})$
  2. Compute the ratio
     $$r = p(\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^{(t-1)}) = \exp[\log p(\boldsymbol{\theta}^*) - \log p(\boldsymbol{\theta}^{(t-1)})]$$
  3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$;

     If $r < 1$, set $\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^* \text{ with probability } r \\ \boldsymbol{\theta}^{(t-1)} \text{ with probability } 1 - r \end{cases}$.

- Then a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.
- Note: When used as a substep in a larger (e.g., Gibbs) algorithm, we often use $T = 1$ (convergence still OK).

Metropolis algorithm (cont'd)

- How to choose the candidate density? The usual approach (after $\boldsymbol{\theta}$ has been transformed to have support $\Re^k$, if necessary) is to set

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}, \widetilde{\Sigma}) \ .$$

  In one dimension, MCMC "folklore" suggests choosing $\widetilde{\Sigma}$ to provide an observed acceptance ratio near 50%.

- Hastings (1970) showed we can drop the requirement that q be symmetric, provided we use

$$r = \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}$$

  – useful for asymmetric target densities!
  – this form called the Metropolis-Hastings algorithm

Convergence assessment

When it is safe to stop and summarize MCMC output?

- We would like to ensure that $\int |\hat{p}_t(\boldsymbol{\theta}) - p(\boldsymbol{\theta})| d\boldsymbol{\theta} < \epsilon$, but all we can hope to see is $\int |\hat{p}_t(\boldsymbol{\theta}) - \hat{p}_{t+k}(\boldsymbol{\theta})| d\boldsymbol{\theta}$!
- Controversy: Does the eventual mixing of "initially overdispersed" parallel sampling chains provide worthwhile information on convergence?
  - While one can never "prove" convergence of a MCMC algorithm using only a finite realization from the chain, poor mixing of parallel chains can help discover extreme forms of nonconvergence
- Still, it's tricky: a slowly converging sampler may be indistinguishable from one that will never converge (e.g., due to nonidentifiability)!

Convergence diagnostics

Various summaries of MCMC output, such as

- sample autocorrelations in one or more chains:
  - close to 0 indicates near-independence, and so chain should more quickly traverse the entire parameter space :)
  - close to 1 indicates the sampler is “stuck” :(
- Gelman/Rubin shrink factor,

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN}\frac{B}{W}\right)\frac{df}{df-2}} \overset{N\to\infty}{\longrightarrow} 1\ ,$$

where B/N is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of an approximating t density to the posterior.

Convergence diagnosis strategy

- Run a few (3 to 5) parallel chains, with starting points believed to be overdispersed
  - say, covering ± 3 prior standard deviations from the prior mean
- Overlay the resulting sample traces for a representative subset of the parameters
  - say, most of the fixed effects, some of the variance components, and a few well-chosen random effects)
- Annotate each plot with lag 1 sample autocorrelations and perhaps Gelman and Rubin diagnostics
- Investigate bivariate plots and crosscorrelations among parameters suspected of being confounded, just as one might do regarding collinearity in linear regression.

Variance estimation

How good is our MCMC estimate once we get it?

- Suppose a single long chain of (post-convergence) MCMC samples $\{\lambda^{(t)}\}_{t=1}^{N}$. Let

$$\hat{E}(\lambda|\mathbf{y}) = \hat{\lambda}_N = \frac{1}{N}\sum_{t=1}^{N}\lambda^{(t)} \ .$$

- Then by the CLT, under iid sampling we could take

$$\widehat{Var}_{iid}(\hat{\lambda}_N) = s_\lambda^2/N = \frac{1}{N(N-1)}\sum_{t=1}^{N}(\lambda^{(t)} - \hat{\lambda}_N)^2 \ .$$

But this is likely an underestimate due to positive autocorrelation in the MCMC samples.

Variance estimation (cont'd)

- To avoid wasteful parallel sampling or "thinning," compute the *effective sample size*,

$$ESS = N/\kappa(\lambda) \ ,$$

where $\kappa(\lambda) = 1 + 2\sum_{k=1}^{\infty} \rho_k(\lambda)$ is the *autocorrelation time*, and we cut off the sum when $\rho_k(\lambda) < \epsilon$

- Then

$$\widehat{Var}_{ESS}(\hat{\lambda}_N) = s_\lambda^2 / ESS(\lambda)$$

Note: $\kappa(\lambda) \geq 1$, so $ESS(\lambda) \leq N$, and so we have that $\widehat{Var}_{ESS}(\hat{\lambda}_N) \geq \widehat{Var}_{iid}(\hat{\lambda}_N)$, in concert with intuition.

Variance estimation (cont'd)

- Another alternative: Batching: Divide the run into m successive batches of length k with batch means $b_1, \ldots, b_m$. Then $\hat{\lambda}_N = \bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$, and

$$\widehat{Var}_{batch}(\hat{\lambda}_N) = \frac{1}{m(m-1)}\sum_{i=1}^{m}(b_i - \hat{\lambda}_N)^2 \ ,$$

  provided that k is large enough so that the correlation between batches is negligible.

- For any $\widehat{V}$ used to approximate $Var(\hat{\lambda}_N)$, a 95% CI for $E(\lambda|\mathbf{y})$ is then given by

$$\hat{\lambda}_N \pm z_{.025}\sqrt{\widehat{V}} \ .$$

Overrelaxation

- Basic idea: Try to speed MCMC convergence by inducing negative autocorrelation within the chains
- Neal (1998): Generate $\{\theta_{i,k}\}_{k=1}^{K}$ independently from the full conditional $p(\theta_i|\theta_{j\neq i}, \mathbf{y})$. Ordering these along with the old value, we have

$$\theta_{i,0} \leq \theta_{i,1} \leq \cdots \leq \theta_{i,r} \equiv \theta_i^{(t-1)} \leq \cdots \leq \theta_{i,K} \ ,$$

so that r is the index of the old value. Then take

$$\theta_i^{(t)} = \theta_{i,K-r} \ .$$

- Note that $K = 1$ produces Gibbs sampling, while large K produces progressively more overrelaxation.
- Generation of the K random variables can be avoided if the full conditional cdf and inverse cdf are available.

Model criticism and selection

- Three related issues to consider:
  - Robustness: Are any model assumptions having an undue impact on the results? (text, Sec. 4.2)
  - Assessment: Does the model provide adequate fit to the data? (text, Sec. 4.3)
  - Selection: Which model (or models) should we choose for final presentation? (text, Secs. 4.4–4.6)

- Consider each in turn...

Sensitivity analysis

Make modifications to an assumption and recompute the posterior; any impact on interpretations or decisions?

- No: The data are strongly informative with respect to this assumption (robustness)
- Yes: Document the sensitivity, think more carefully about it, and perhaps collect more data.
- Examples of assumptions to modify: increasing/ decreasing a prior mean by one prior s.d.; doubling/ halving a prior s.d.; case deletion.
- Importance sampling and asymptotic methods can greatly reduce computational overhead, even if these methods were *not* used in analysis of original model.

  $\Rightarrow$ Run and diagnose convergence for "base" model; use approximate method for robustness study

Prior partitioning

– a "backwards" approach to robustness!

- What if the range of plausible assumptions is unimaginably broad, as in the summary of a government-sponsored clinical trial?
- Potential solution: Determine the set of prior inputs that are consistent with a given conclusion, given the data observed so far. The consumer may then compare this prior class to his/her own personal prior beliefs.
- Thus we are partitioning the prior class based on possible outcomes.
- *Example:* Find set of all prior means μ such that

$$P(\theta \geq 0|\mathbf{y}) > .025$$

(for otherwise, we will decide $\theta < 0$).

Model assessment

Many of the tools mentioned in Chapter 2 are now easy to compute via Monte Carlo methods!

- Example: Find the cross-validation residual

$$r_i = y_i - E(y_i|\mathbf{y}_{(i)}) \ ,$$

where $\mathbf{y}_{(i)}$ denotes the vector of all the data except the $_i{}^{th}$ value, i.e.

$$\mathbf{y}_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)'$$

Model assessment

- Using MC draws $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$, we have

$$\begin{aligned} E(y_i|\mathbf{y}_{(i)}) &= \int\int y_i f(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) dy_i d\boldsymbol{\theta} \\ &= \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \\ &\approx \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &\approx \frac{1}{G} \sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)}) \ . \end{aligned}$$

  - Approximation should be adequate unless the dataset is small and y_i is an extreme outlier
  - Same $\boldsymbol{\theta}^{(g)}$'s may be used for each $i = 1, \ldots, n$.

Bayes factors

the most basic Bayesian model choice tool!

- Given models M_1 and M_2, computable as

$$BF = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \; .$$

- Sadly, unlike posteriors and predictives, *marginal* distributions are not easily estimated via MCMC! So...

◊ Direct methods: Since $p(\mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, we could draw $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta})$ and compute

$$\hat{p}(\mathbf{y}) = \frac{1}{G} \sum_{g=1}^{G} f(\mathbf{y} \mid \boldsymbol{\theta}^{(g)}) \; .$$

Easy, but terribly inefficient.

Bayes factors

- Better: Draw $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$ and compute the *harmonic mean* estimate

$$\hat{p}(\mathbf{y}) = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{1}{f(\mathbf{y} \mid \boldsymbol{\theta}^{(g)})} \right]^{-1},$$

But this is terribly unstable (division by 0)!

- Better yet: try

$$\hat{p}(\mathbf{y}) = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{h(\boldsymbol{\theta}^{(g)})}{f(\mathbf{y}|\boldsymbol{\theta}^{(g)})\, p(\boldsymbol{\theta}^{(g)})} \right]^{-1},$$

where $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$ and $h(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta}|\mathbf{y})$.
(If h equals the prior, we get the harmonic mean again.)

Predictive Model Selection

Less formal approaches, useful when Bayes factor is unavailable or inappropriate (e.g., when using improper priors). These include:

- Cross-validatory checks, such as $\sum_i \log f(y_i^{obs}|\mathbf{y}_{(i)})$ or $\sum_i [y_i - E(y_i|\mathbf{y}_{(i)})]^2$.
- Expected predicted "model discrepancy,"

$$E[d(\mathbf{y}_{new}, \mathbf{y}_{obs})|\mathbf{y}_{obs}, M_i] \ ,$$

where $d(\mathbf{y}_{new}, \mathbf{y}_{obs})$ is an appropriate discrepancy function, e.g.,

$$d(\mathbf{y}_{new}, \mathbf{y}_{obs}) = (\mathbf{y}_{new} - \mathbf{y}_{obs})^T (\mathbf{y}_{new} - \mathbf{y}_{obs}) \ .$$

Choose the model that minimizes discrepancy!

Predictive Model Selection

- Likelihood criteria: think of $\ell \equiv \log L(\boldsymbol{\theta})$ as a parametric function of interest, and compute

$$\hat{\ell} \equiv E[\log L(\boldsymbol{\theta})|\mathbf{y}] \approx \frac{1}{G}\sum_{g=1}^{G} \log L(\boldsymbol{\theta}^{(g)})$$

  as an overall measure of model fit.

- Penalized likelihood criteria: Subtract a "penalty" from the likelihood score, in order to avoid flooding unhelpful predictors into the model. Most common example: the Bayesian Information (Schwarz) Criterion,

$$\widehat{BIC} = 2\hat{\ell} - p \log n$$

  where p is the number of parameters in the model, and n is the number of datapoints.

Extension to Hierarchical Models

- Penalized likelihood criteria (BIC, AIC) trade off "fit" against "complexity"
- But what is the "complexity" of a hierarchical model?
- Example: One-way ANOVA model

$$Y_i|\theta_i \overset{ind}{\sim} N(\theta_i, 1/\tau_i) \text{ and } \theta_i \overset{iid}{\sim} N(\mu, 1/\lambda),\ i = 1, \ldots, p$$

Suppose μ, λ, and the τ_i are known. How many parameters are in this model?

- If $\lambda = \infty$, all $\theta_i = \mu$ and there are 0 free parameters
- If $\lambda = 0$, the θ_i are unconstrained and there are p free parameters

- In practice, $0 < \lambda < \infty$ so the "effective number of parameters" is somewhere in between! How to define?....

Hierarchical model complexity

- Proposal: use the effective number of parameters,

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

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

is the deviance score, computed from the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ and a standardizing function $h(\mathbf{y})$.

- Example: For the one-way ANOVA model,

$$p_D = \sum_{i=1}^{p} \frac{\tau_i}{\tau_i + \lambda} \ ,$$

  - Clearly $0 \leq p_D \leq p$ as desired
  - If we place a hyperprior on λ, the effective model size p_D will depend on the dataset!

Model selection via DIC

- Given the p_D measure of model complexity, suppose we now summarize fit of a model by

$$\bar{D} = E_{\theta|\mathbf{y}}[D] \ ,$$

- Compare models via the Deviance Information Criterion,

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

  a generalization of the Akaike Information Criterion (AIC), since $AIC \approx \bar{D} + p$ for nonhierarchical models.

- Smaller values of DIC indicate preferred models.
- While p_D has a scale (effective model size), DIC does not, so only differences in DIC across models matter.

Issues in using DIC

- p_D and DIC are very broadly applicable provided $p(\mathbf{y}|\boldsymbol{\theta})$ is available in closed form
- Both building blocks of DIC and p_D, $E_{\theta|y}[D]$ and $D(E_{\theta|y}[\boldsymbol{\theta}])$, are easily estimated via MCMC methods
- ...and in fact are directly available within `WinBUGS`!

- p_D and DIC may not be invariant to reparametrization
- p_D can be negative for non-log-concave likelihoods, or when there is strong prior-data conflict
- p_D and DIC will depend on our "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 η"

Bayesian Software Options

- BUGS: WinBUGS, OpenBUGS, R2WinBUGS, BRugs, rbugs
  `http://www.openbugs.info/w/`
- JAGS: JAGS, rjags, R2jags, runjags
  `http://mcmc-jags.sourceforge.net/`
  `http://cran.r-project.org/web/packages/rjags`
- R: mcmc (general purpose), JMBayes (Joint Modeling)
  `cran.r-project.org/web/packages/JMbayes`
- SAS: PROC MCMC
  `support.sas.com/rnd/app/da/Bayesian/MCMC.html`
- Other MCMC-based: Stan and RStan, WBDev, PyMC
  `http://www.mc-stan.org/`
- Other non-MCMC-based: INLA (Integrated Nested Laplace Approx)
  `http://www.r-inla.org/`

Example using R: Heart Valves Study

- Goal: Show that the thrombogenicity rate (TR) is less than two times the objective performance criterion
- Data: From both the current study and a previous study on a similar product (St. Jude mechanical valve).
- Model: Let T be the total number of patient-years of followup, and θ be the TR per year. We assume the number of thrombogenicity events $Y \sim Poisson(\theta T)$:

$$f(y|\theta) = \frac{e^{-\theta T}(\theta T)^y}{y!} \ .$$

- Prior: Assume a $Gamma(\alpha, \beta)$ prior for θ:

$$p(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \ , \theta > 0 \ .$$

Heart Valves Study

- The gamma prior is conjugate with the likelihood, so the posterior emerges in closed form:

$$\begin{aligned} p(\theta|y) &\propto \theta^{y+\alpha-1} e^{-\theta(T+1/\beta)} \\ &\propto Gamma(y+\alpha, (T+1/\beta)^{-1}) \ . \end{aligned}$$

The study objective is met if

$$P(\theta < 2 \times OPC \mid y) \geq 0.95 \ ,$$

where $OPC = \theta_0 = 0.038$.

- Prior selection: Our gamma prior has mean $M = \alpha\beta$ and variance $V = \alpha\beta^2$. This means that if we specify M and V, we can solve for α and β as

$$\alpha = M^2/V \text{ and } \beta = V/M \ .$$

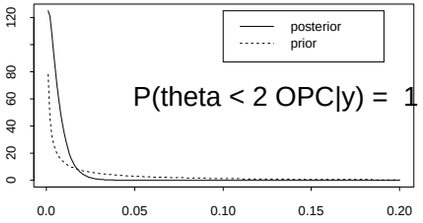
Heart Valves Study

- A few possibilities for prior parameters:
  - Suppose we set $M = \theta_0 = 0.038$ and $\sqrt{V} = 2\theta_0$ (so that 0 is two standard deviations below the mean). Then $\alpha = 0.25$ and $\beta = 0.152$, a rather vague prior.
  - Suppose we set $M = 98/5891 = .0166$, the overall value from the St. Jude studies, and $\sqrt{V} = M$ (so 0 is one sd below the mean). Then $\alpha = 1$ and $\beta = 0.0166$, a moderate (exponential) prior.
  - Suppose we set $M = 98/5891 = .0166$ again, but set $\sqrt{V} = M/2$. This is a rather informative prior.
- We also consider event counts that are lower (1), about the same (3), and much higher (20) than for St. Jude.
- The study objective is not met with the "bad" data – *unless* the posterior is "rescued" by the informative prior (lower right corner, next page).

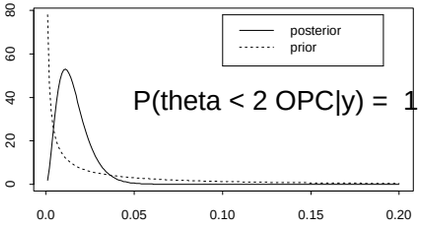
Heart Valves Study

Priors and posteriors, Heart Valves ADVANTAGE study, Poisson-gamma model
for various prior (M, sd) and data (y) values; T = 200 , 2 OPC = 0.076

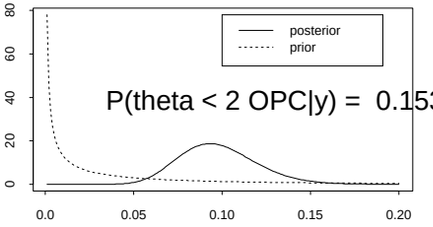
vague prior



M. sd = 0.038 0.076 : Y = 1

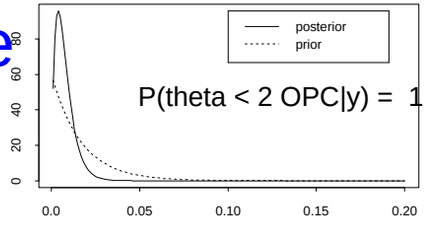


M. sd = 0.038 0.076 : Y = 3

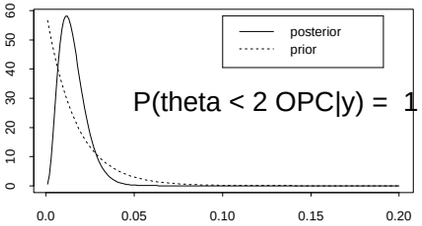


M. sd = 0.038 0.076 : Y = 20

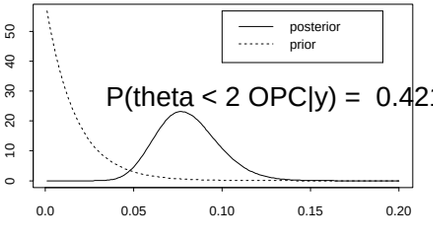
moderate prior



M. sd = 0.017 0.017 : Y = 1

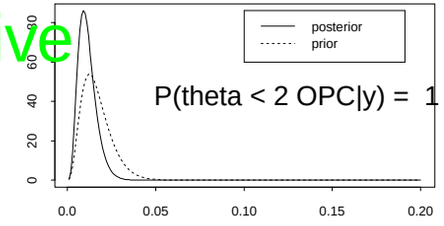


M. sd = 0.017 0.017 : Y = 3

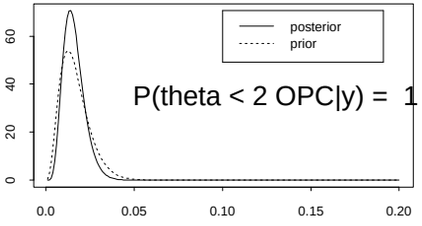


M. sd = 0.017 0.017 : Y = 20

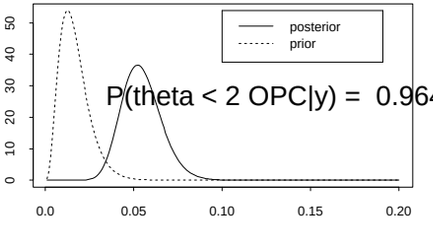
informative prior



M. sd = 0.017 0.008 : Y = 1



M. sd = 0.017 0.008 : Y = 3



M. sd = 0.017 0.008 : Y = 20

- S code to create this plot is available in www.biostat.umn.edu/∼brad/hv.S
  – try it yourself in S-plus or R (http://cran.r-project.org)

Alternate hierarchical models

- One might be uncomfortable with our implicit assumption that the TR is the same in both studies. To handle this, extend to a hierarchical model:

$$Y_i \sim Poisson(\theta_i T_i),\ i = 1, 2,$$

where $i = 1$ for St. Jude, and $i = 2$ for the new study.

- Borrow strength between studies by assuming

$$\theta_i \overset{iid}{\sim} Gamma(\alpha, \beta),$$

i.e., the two TR's are exchangeable, but not identical.

- We now place a third stage prior on α and β, say

$$\alpha \sim Exp(a) \text{ and } \beta \sim IG(c, d).$$

- Fit in WinBUGS using the pump example as a guide!

BUGS Example 1: Poisson Failure Rates

Example 2.7 revisited again!

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

$i = 1, \ldots, k$, where μ, 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 θ_i for the i^{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 α and β.

Pump Example

- Recall that the full conditional distributions for the θ_i and β are available in closed form (gamma and inverse gamma, respectively), but that no conjugate prior for α exists.
- However, the full conditional for α,

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

can be shown to be log-concave in α. 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)
```

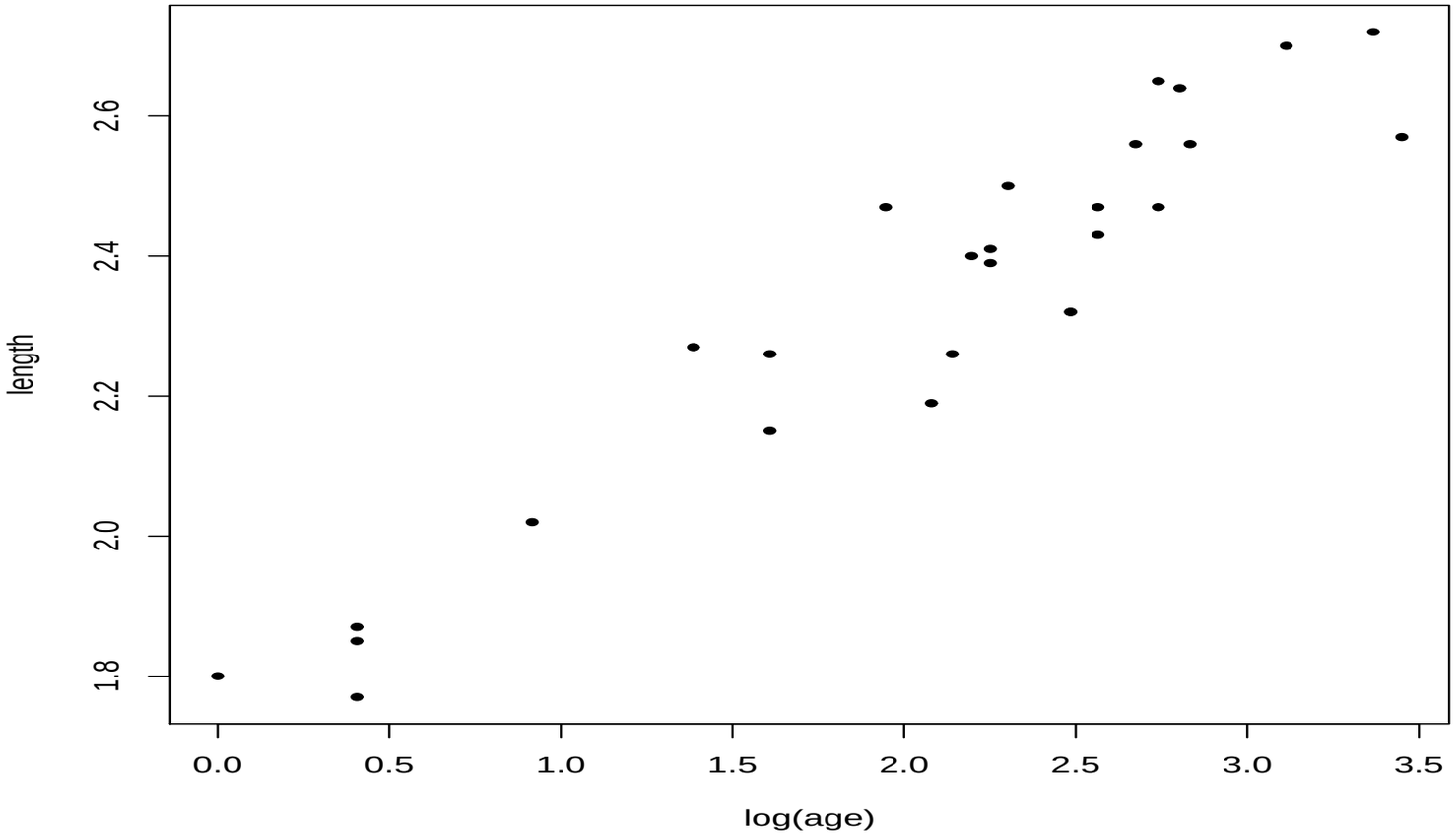
Pump Example Results

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 θ_5 and θ_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!

BUGS Example 2: Linear Regression



- For $n = 27$ captured samples of the sirenian species *dugong* (sea cow), relate an animal's length in meters, Y_i, to its age in years, x_i.
- To avoid a nonlinear model for now, transform x_i to the log scale; plot of Y versus $\log(x)$ looks fairly linear!

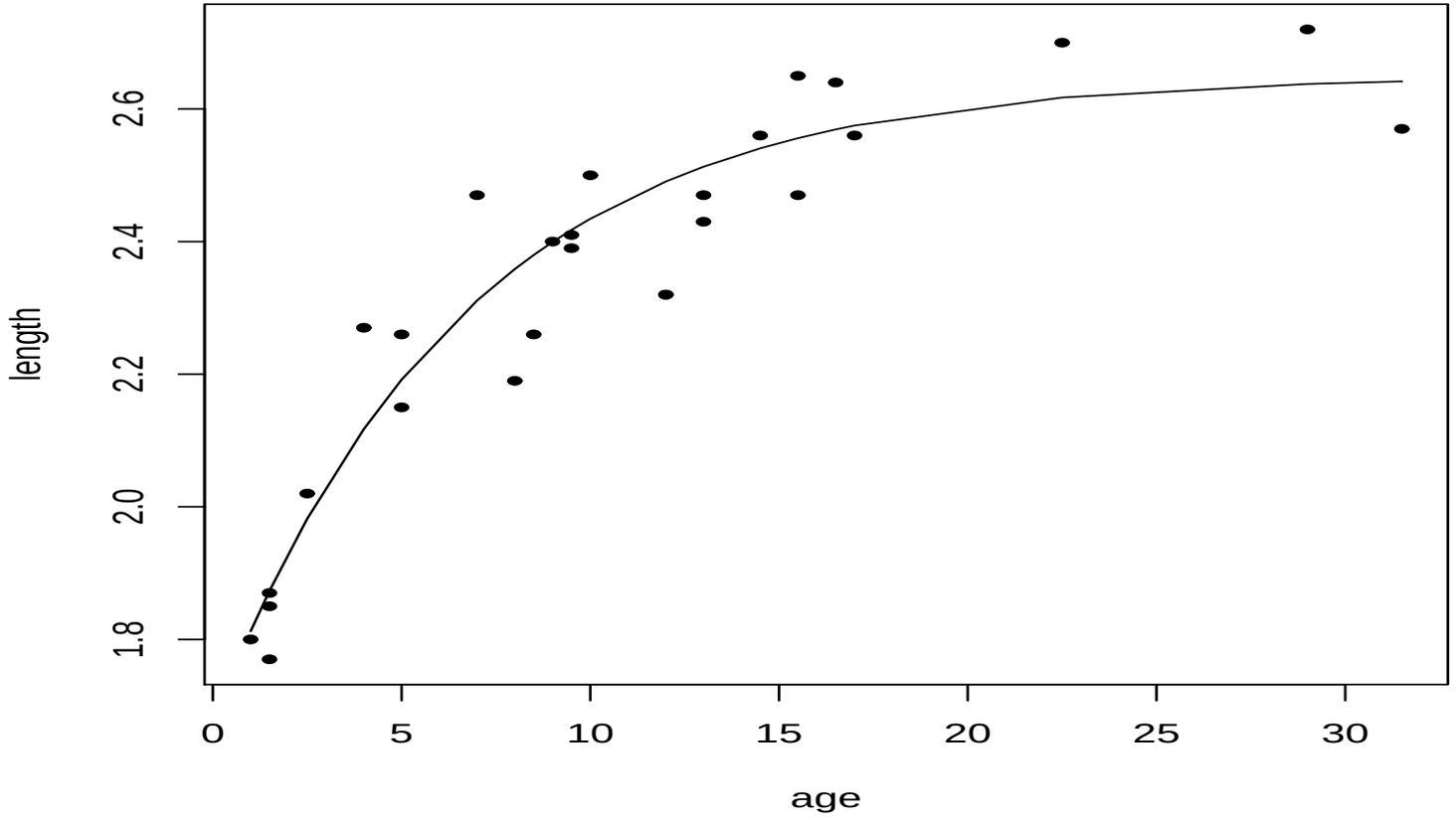
Simple linear regression in WinBUGS

$$Y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i, \ i = 1, \ldots, n$$

where $\epsilon_i \overset{iid}{\sim} N(0, \tau)$ and $\tau = 1/\sigma^2$, the precision in the data.

- Prior distributions:
  - flat for β_0, β_1
  - vague gamma on τ (say, Gamma(0.1, 0.1), which has mean 1 and variance 10) is traditional
- posterior correlation is reduced by centering the $\log(x_i)$ around their own mean
- Andrew Gelman suggests placing a uniform prior on σ, bounding the prior away from 0 and $\infty \Longrightarrow U(.01, 100)$?
- Code: www.biostat.umn.edu/∼brad/data/dugongs_BUGS.txt

BUGS Example 3: Nonlinear Regression



- Model the untransformed dugong data as

$$Y_i = \alpha - \beta\gamma^{x_i} + \epsilon_i,\ i = 1, \ldots, n\ ,$$

where $\alpha > 0$, $\beta > 0$, $0 \leq \gamma \leq 1$, and as usual $\epsilon_i \overset{iid}{\sim} N(0, \tau)$ for $\tau \equiv 1/\sigma^2 > 0$.

Nonlinear regression in WinBUGS

- In this model,
  - α corresponds to the average length of a fully grown dugong ($x \to \infty$)
  - $(\alpha - \beta)$ is the length of a dugong at birth ($x = 0$)
  - γ determines the growth rate: lower values produce an initially steep growth curve while higher values lead to gradual, almost linear growth.
- Prior distributions: flat for α and β, $U(.01, 100)$ for σ, and $U(0.5, 1.0)$ for γ (harder to estimate)
- Code: www.biostat.umn.edu/$\sim$brad/data/dugongsNL_BUGS.txt
- Obtain posterior density estimates and autocorrelation plots for α, β, γ, and σ, and investigate the bivariate posterior of (α, γ) using the Correlation tool on the Inference menu!

BUGS Example 4: Logistic Regression

- Consider a binary version of the dugong data,

$$Z_i = \begin{cases} 1 & \text{if } Y_i > 2.4 \text{ (i.e., the dugong is “full-grown”)} \\ 0 & \text{otherwise} \end{cases}$$

- A logistic model for $p_i = P(Z_i = 1)$ is then

$$logit(p_i) = \log[p_i/(1-p_i)] = \beta_0 + \beta_1 log(x_i) \ .$$

- Two other commonly used link functions are the probit,

$$probit(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 log(x_i) \ ,$$

and the complementary log-log (cloglog),

$$cloglog(p_i) = \log[-\log(1-p_i)] = \beta_0 + \beta_1 log(x_i) \ .$$

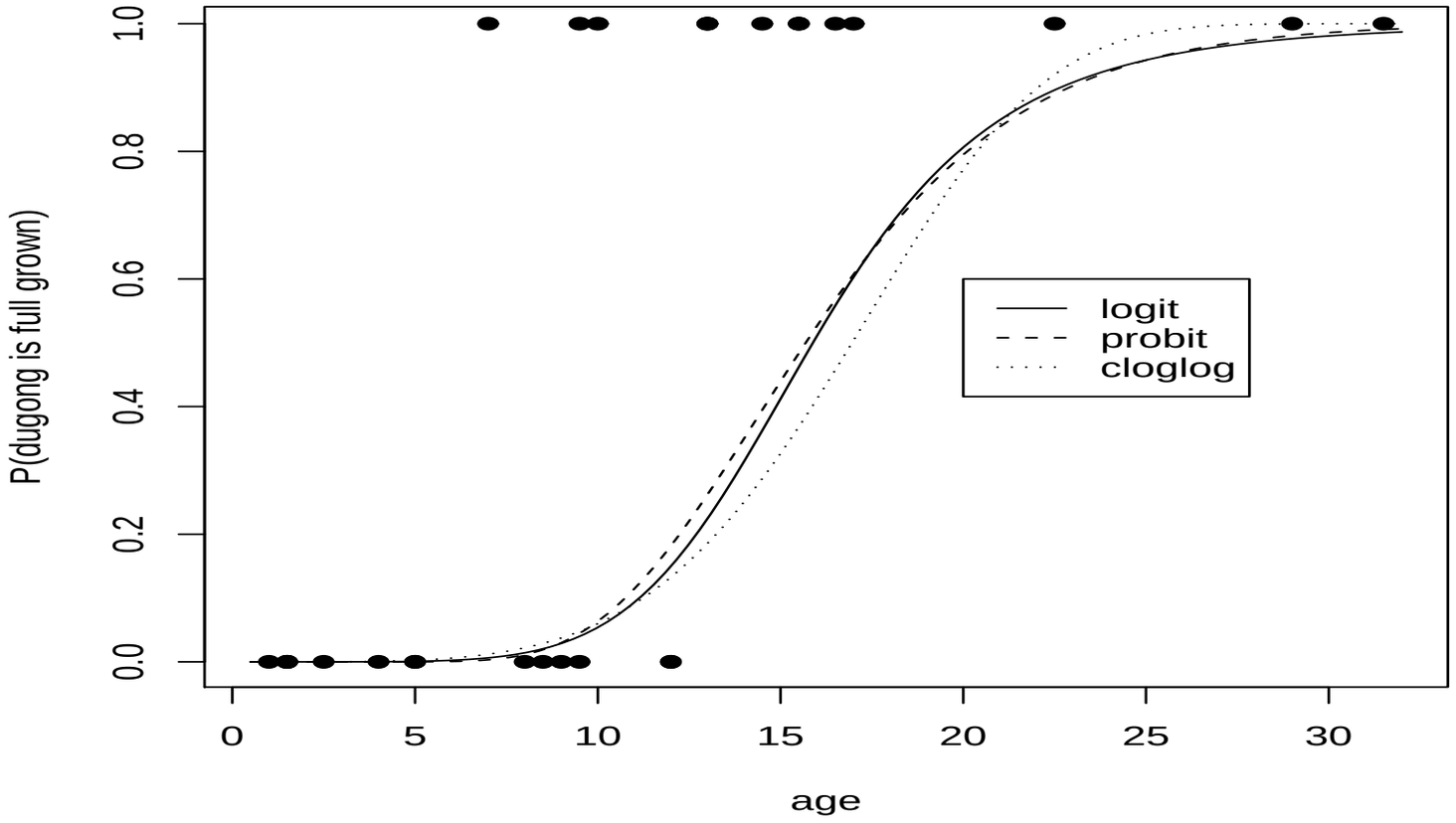
Binary regression in WinBUGS

- Code: www.biostat.umn.edu/∼brad/data/dugongsBin_BUGS.txt
- Code uses flat priors for β_0 and β_1, and the phi function, instead of the less stable probit function.
- DIC scores for the three models:

model	$\overline{D}$	p_D	DIC
logit	19.62	1.85	21.47
probit	19.30	1.87	21.17
cloglog	18.77	1.84	20.61

  In fact, these scores can be obtained from a single run; see the "trick version" at the bottom of the BUGS file!
- Use the Comparison tool to compare the posteriors of β_1 across models, and the Correlation tool to check the bivariate posteriors of (β_0, β_1) across models.

Fitted binary regression models



- The logit and probit fits appear very similar, but the cloglog fitted curve is slightly different
- You can also compare p_i posterior boxplots (induced by the link function and the β_0 and β_1 posteriors) using the Comparison tool.

BUGS Example 5: Hierarchical Models

- Extend the usual two-stage (likelihood plus prior) Bayesian structure to a hierarchy of L levels, where the joint distribution of the data and the parameters is

$$f(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3)\cdots\pi_L(\boldsymbol{\theta}_L|\boldsymbol{\lambda}).$$

- L is often determined by the number of subscripts on the data. For example, suppose Y_{ijk} is the test score of child k in classroom j in school i in a certain city. Model:

$$\begin{aligned}
Y_{ijk}|\theta_{ij} &\overset{ind}{\sim} N(\theta_{ij}, \tau_\theta) \quad (\theta_{ij} \text{ is the classroom effect}) \\
\theta_{ij}|\eta_i &\overset{ind}{\sim} N(\eta_i, \tau_\eta) \quad (\eta_i \text{ is the school effect}) \\
\eta_i|\lambda &\overset{iid}{\sim} N(\lambda, \tau_\lambda) \quad (\lambda \text{ is the grand mean})
\end{aligned}$$

Priors for λ and the τ's now complete the specification!

Cross-Study (Meta-analysis) Data

- Data: estimated log relative hazards $Y_{ij} = \hat{\beta}_{ij}$ obtained by fitting separate Cox proportional hazards regressions to the data from each of $J = 18$ clinical units participating in $I = 6$ different AIDS studies.
- To these data we wish to fit the cross-study model,

$$Y_{ij} = a_i + b_j + s_{ij} + \epsilon_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J,$$

where a_i = study main effect

b_j = unit main effect

s_{ij} = study-unit interaction term, and

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_{ij}^2)$$

and the estimated standard errors from the Cox regressions are used as (known) values of the σ_{ij}.

Cross-Study (Meta-analysis) Data

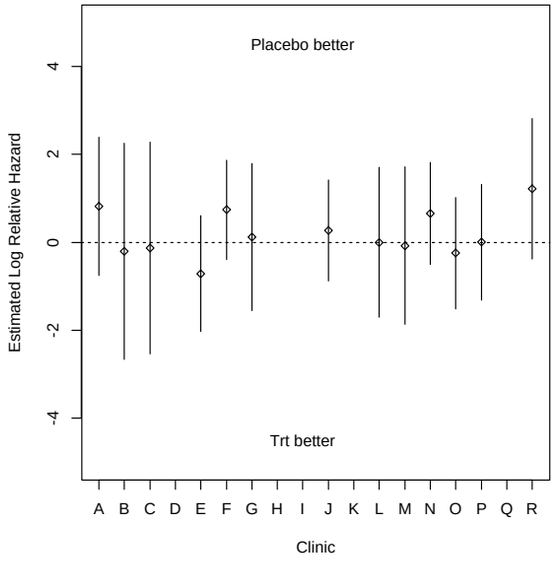
Estimated Unit-Specific Log Relative Hazards						
Unit	Toxo	ddI/ddC	NuCombo ZDV+ddI	NuCombo ZDV+ddC	Fungal	CMV
A	0.814	NA	-0.406	0.298	0.094	NA
B	-0.203	NA	NA	NA	NA	NA
C	-0.133	NA	0.218	-2.206	0.435	0.145
D	NA	NA	NA	NA	NA	NA
E	-0.715	-0.242	-0.544	-0.731	0.600	0.041
F	0.739	0.009	NA	NA	NA	0.222
G	0.118	0.807	-0.047	0.913	-0.091	0.099
H	NA	-0.511	0.233	0.131	NA	0.017
I	NA	1.939	0.218	-0.066	NA	0.355
J	0.271	1.079	-0.277	-0.232	0.752	0.203
K	NA	NA	0.792	1.264	-0.357	0.807
⋮	⋮	⋮	⋮	⋮	⋮	⋮
R	1.217	0.165	0.385	0.172	-0.022	0.203

Cross-Study (Meta-analysis) Data

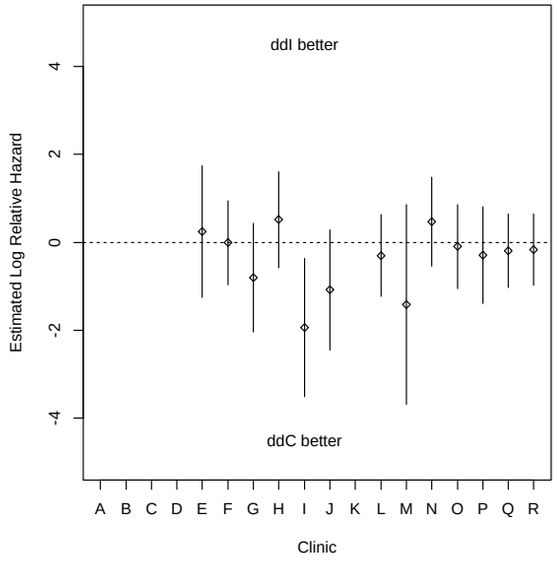
- Note that some values are missing ("NA") since
  - not all 18 units participated in all 6 studies
  - the Cox estimation procedure did not converge for some units that had few deaths
- *Goal:* To identify which clinics are opinion leaders (strongly agree with overall result across studies) and which are dissenters (strongly disagree).
- Here, overall results all favor the treatment (i.e. mostly negative Ys) except in Trial 1 (Toxo). Thus we multiply all the Y_{ij}'s by –1 for $i \neq 1$, so that larger Y_{ij} correspond in all cases to stronger agreement with the overall.
- Next slide shows a plot of the Y_{ij} values and associated approximate 95% CIs...

Cross-Study (Meta-analysis) Data

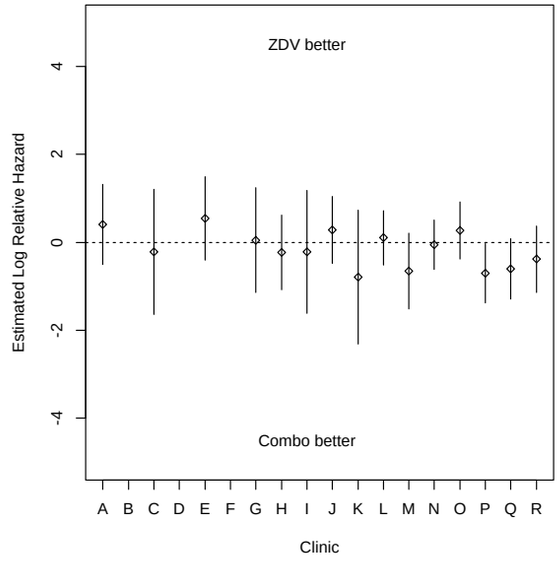
1: Toxo



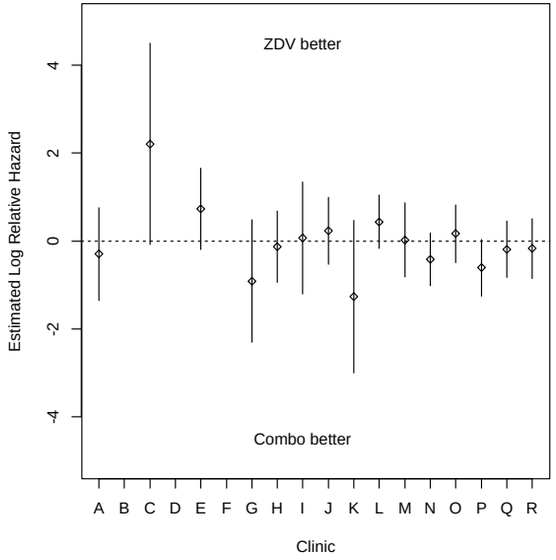
2: ddI/ddC



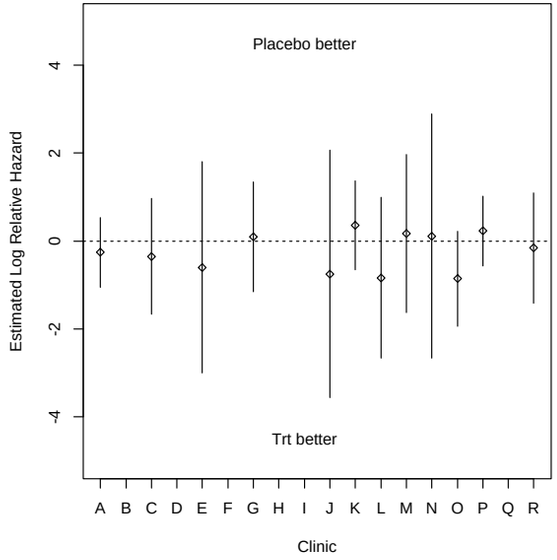
3: NuCombo-ddI



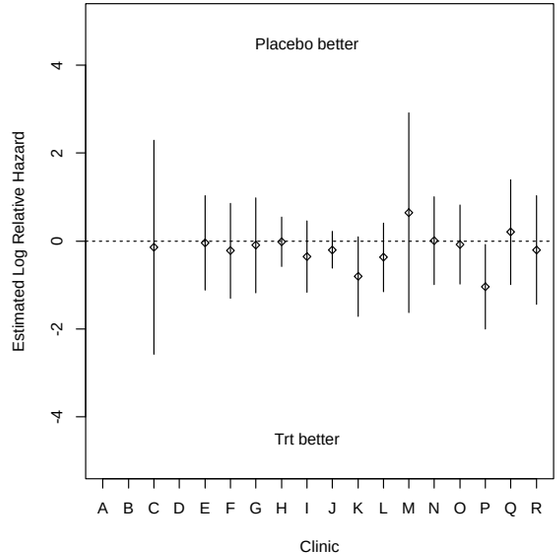
4: NuCombo-ddC



5: Fungal



6: CMV



Cross-Study (Meta-analysis) Data

- Second stage of our model:

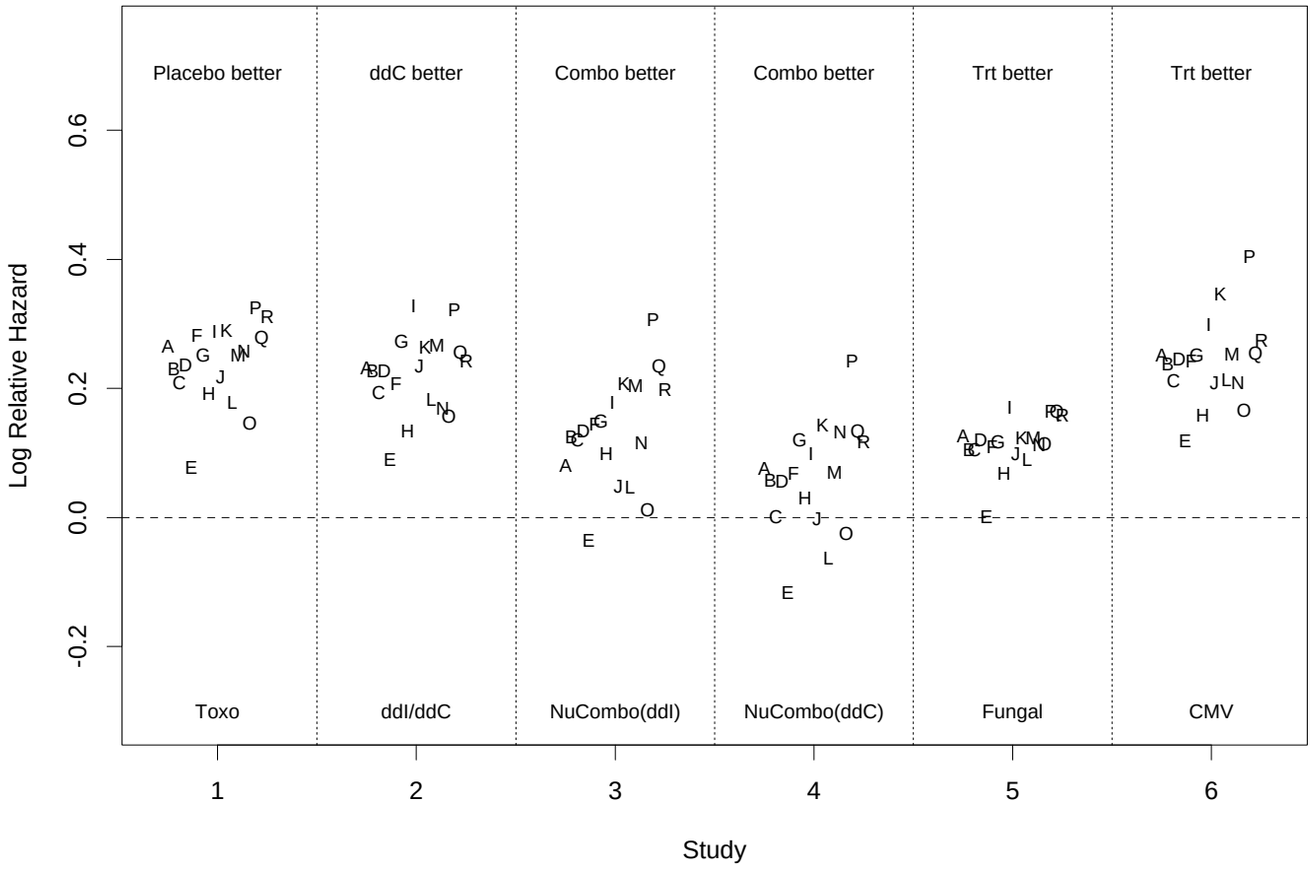
$$a_i \overset{iid}{\sim} N(0, 100^2), \quad b_j \overset{iid}{\sim} N(0, \sigma_b^2), \quad \text{and} \quad s_{ij} \overset{iid}{\sim} N(0, \sigma_s^2)$$

- Third stage of our model:

$$\sigma_b \sim Unif(0.01, 100) \quad \text{and} \quad \sigma_s \sim Unif(0.01, 100)$$

  That is, we
  - preclude borrowing of strength across studies, but
  - encourage borrowing of strength across units
- With $I + J + IJ$ parameters but fewer than IJ data points, some effects must be treated as random!
- Code: www.biostat.umn.edu/∼brad/data/crprot_BUGS.txt

Plot of θ_{ij} posterior means



- Unit P is an opinion leader; Unit E is a dissenter
- Substantial shrinkage towards 0 has occurred: mostly positive values; no estimated θ_{ij} greater than 0.6

Model Comparision via DIC

Since we lack replications for each study-unit (i-j) combination, the interactions s_{ij} in this model were only weakly identified, and the model might well be better off without them (or even without the unit effects b_j).

As such, compare a variety of reduced models:

```
  Y[i,j] ~ dnorm(theta[i,j],P[i,j])
#    theta[i,j] <- a[i]+b[j]+s[i,j]  # full model
#    theta[i,j] <- a[i] + b[j]       # drop interactions
#    theta[i,j] <- a[i] + s[i,j]     # no unit effect
#    theta[i,j] <- b[j] + s[i,j]     # no study effect
#    theta[i,j] <- a[1] + b[j]       # unit + intercept
#    theta[i,j] <- b[j]              # unit effect only
    theta[i,j] <- a[i]              # study effect only
```

Investigate p_D values for these models; are they consistent with posterior boxplots of the b_i and s_{ij}?

DIC results for Cross-Study Data:

model	$\overline{D}$	p_D	DIC
full model	122.0	12.8	134.8
drop interactions	123.4	9.7	133.1
no unit effect	123.8	10.0	133.8
no study effect	121.4	9.7	131.1
unit + intercept	120.3	4.6	124.9
unit effect only	122.9	6.2	129.1
study effect only	126.0	6.0	132.0

The DIC-best model is the one with only an intercept (a role played here by a_1) and the unit effects b_j.

These DIC differences are not much larger than their possible Monte Carlo errors, so almost any of these models could be justified here.

UGS Example 6: Nonlinear w/Random Effects

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

$$\begin{aligned} Y_{ij} &= \text{plasma concentration of the drug Cadralazine} \\ x_{ij} &= \text{time elapsed since dose given} \end{aligned}$$

where $i = 1, \ldots, 10$ indexes the patient, while $j = 1, \ldots, n_i$ 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 $\epsilon_{ij} \overset{ind}{\sim} N(0, \tau_i)$.

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

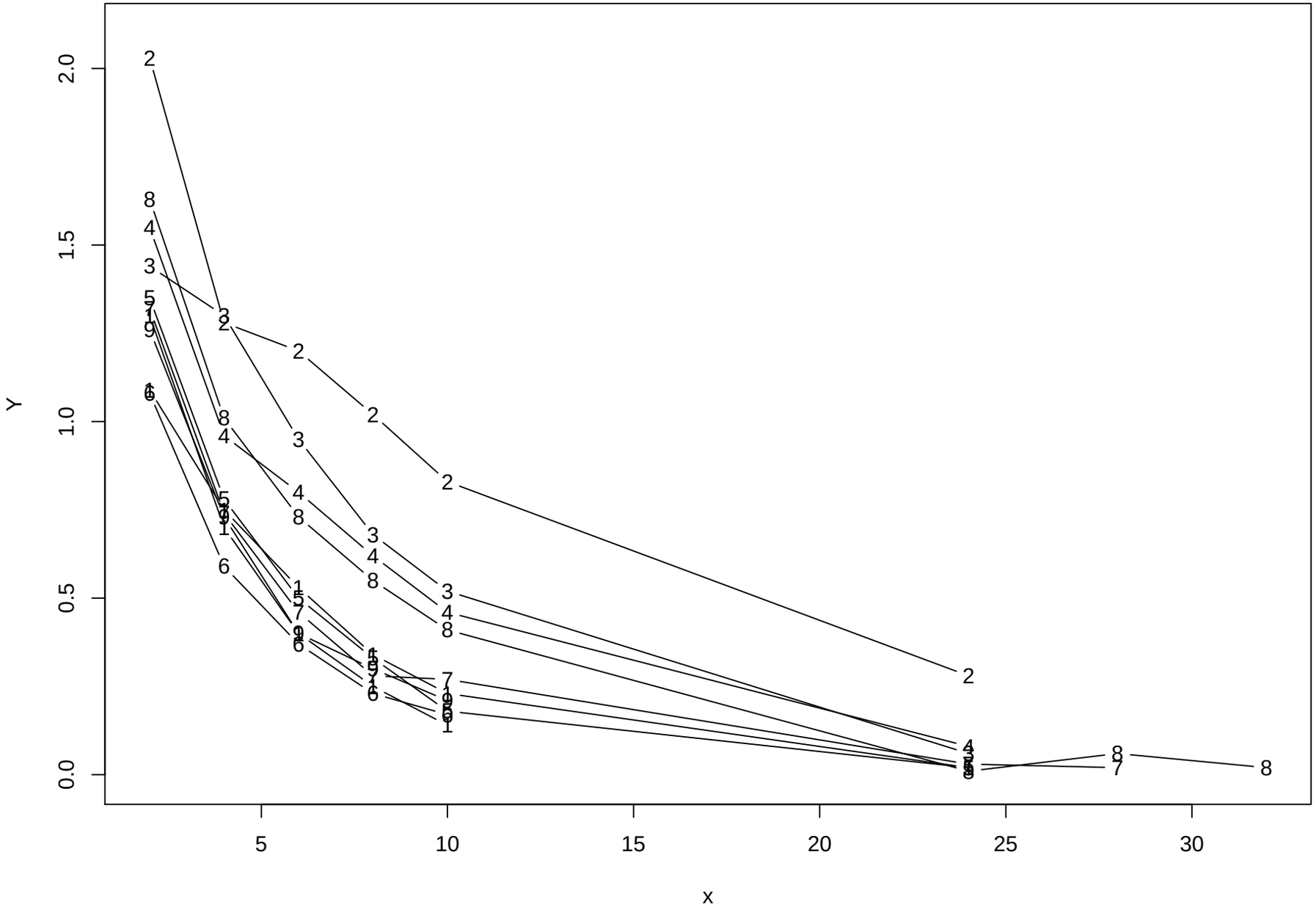
$$\begin{aligned} \log \eta_{ij}(x_{ij}) &= \log \left[30 \alpha_i^{-1} \exp(-\beta_i x_{ij}/\alpha_i) \right] \\ &= \log 30 - \log \alpha_i - \beta_i x_{ij}/\alpha_i \\ &= \log 30 - a_i - \exp(b_i - a_i) x_{ij} \ , \end{aligned}$$

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

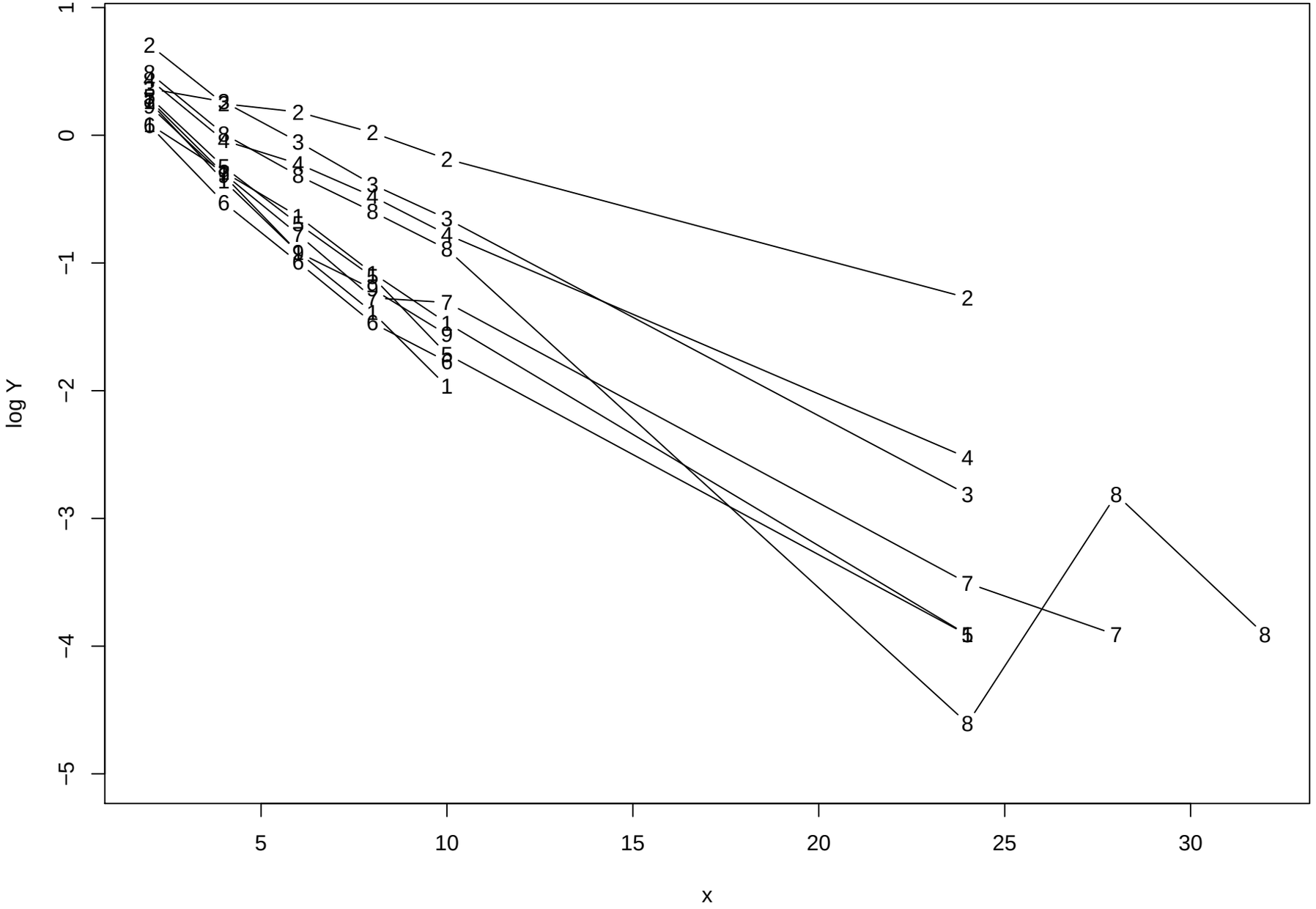
PK Data

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 $\boldsymbol{\theta}_i \equiv (a_i, b_i)'$,

$$\boldsymbol{\theta}_i \overset{iid}{\sim} N_2(\boldsymbol{\mu}, \Omega) \ , \text{ where } \boldsymbol{\mu} = (\mu_a, \mu_b) \ .$$

- Usual conjugate prior specification:

$$\begin{aligned} \boldsymbol{\mu} &\sim N_2(\boldsymbol{\lambda}, C) \\ \tau_i &\overset{iid}{\sim} G(\nu_0/2 \ , \ \nu_0\tau_0/2) \\ \Omega &\sim \text{Wishart}((\rho R)^{-1}, \rho) \end{aligned}$$

- Note that the $\boldsymbol{\theta}_i$ 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

PK Results (WinBUGS vs. Fortran)

	BUGS			Sargent et al. (2000)		
parameter	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
τ_1	362.4	260.4	0.313	380.8	268.8	0.220
τ_2	84.04	57.60	0.225	81.40	58.41	0.255
τ_7	18.87	12.07	0.260	15.82	11.12	0.237
τ_8	2.119	1.139	0.085	1.499	0.931	0.143
$Y_{2,8}$	0.1338	0.0339	0.288	0.1347	0.0264	–
$Y_{7,8}$	0.00891	0.00443	0.178	0.00884	0.00255	–

Homework: InterStim Example

Device that uses electrical stimulation of the brain to prevent urinary incontinences. For patients $i = 1, \ldots, 49$:

$$
\begin{aligned}
X_{1i} &= \text{number of incontinences per week at baseline} \\
X_{2i} &= \text{number of incontinences per week at 3 months} \\
X_{3i} &= \text{number of incontinences per week at 6 months} \\
X_{4i} &= \text{number of incontinences per week at 12 months}
\end{aligned}
$$

patient	X_{1i}	X_{2i}	X_{3i}	X_{4i}
1	60	0.7	0	16
2	8	0	0	0
...				
8	9	0.7	12	NA
9	3	0	0.7	0
...				
49	16	NA	NA	NA

InterStim Example

- Goal 1: Obtain full predictive inference for all missing X values (point and interval estimates)
- Goal 2: Obtain measure of percent improvement (relative to baseline) due to InterStim at 6 and 12 months
- Model: Let $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})'$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)'$. Clearly the X_{ij}'s are correlated, but an ordinary longitudinal model does not seem appropriate (we can't just use a linear model here). So instead, maintain the generality:

$$\begin{aligned} \mathbf{X}_i|\boldsymbol{\theta}, \Upsilon &\overset{iid}{\sim} N_4(\boldsymbol{\theta}, \Upsilon^{-1}) \\ \boldsymbol{\theta} &\sim N_4(\boldsymbol{\mu}, \Omega^{-1}) \\ \Upsilon &\sim Wishart_4(R, \rho) \end{aligned}$$

InterStim Example

- WinBUGS will generate all missing X's ("NA"s in the dataset) from their full conditional distributions as part of the Gibbs algorithm. Thus we will obtain samples from $p(X_{ij}|\mathbf{X}_{obs})$ for all missing X_{ij} (achieving Goal 1).
- Re: Goal 2 (percent improvement from baseline), let

$$\alpha = \frac{\theta_1 - \theta_3}{\theta_1} \text{ and } \beta = \frac{\theta_1 - \theta_4}{\theta_1}$$

  Then $p(\alpha|\mathbf{X}_{obs})$ and $p(\beta|\mathbf{X}_{obs})$ address this issue!

- Hyperparameters: all vague: $\Omega = Diag(10^{-6}, \ldots, 10^{-6})$, $\boldsymbol{\mu} = \mathbf{0}$, $\rho = 4$ (the smallest value for which the Wishart prior for Υ is proper) and $R = Diag(10^{-1}, \ldots, 10^{-1})$.
- Code and Data: www.biostat.umn.edu/∼brad/data/InterStim.txt

Basics of Model Building Recommendations

Brad Carlin and Harrison Quick

Data Type	Likelihood	Parameters	Prior
Continuous Data[1], $Y_i \in (-\infty, \infty)$			
No Covariates	$Y_i \sim N(\mu, \sigma^2)$	$\mu \in (-\infty, \infty)$ $\sigma^2 > 0$	$\mu \sim N(0, big)$ $\sigma^2 \sim IG(\epsilon, \epsilon)$
With Covariates — Identity Link	$Y_i \sim N(\mu_i, \sigma^2)$ $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$	$\beta_j \in (-\infty, \infty)$ $\sigma^2 > 0$	$\beta_j \sim N(0, big)$ $\sigma^2 \sim IG(\epsilon, \epsilon)$
Binary Data, $Y_i = \{0, 1\}$			
No Covariates	$Y_i \sim Bern(\theta)$	$\theta \in [0, 1]$	$\theta \sim Beta(a, b)$
With Covariates[2] — Logit Link	$Y_i \sim Bern(\theta_i)$ $\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$	$\beta_j \in (-\infty, \infty)$	$\beta_j \sim N(0, medium)$
Count Data[3][4], $Y_i = \{0, 1, \ldots, \}$			
No Covariates	$Y_i \sim \text{Pois}(\lambda)$	$\lambda > 0$	$\lambda \sim Gamma(a, b)$
With Covariates[5] — Log Link	$Y_i \sim \text{Pois}(\lambda_i)$ $\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}$	$\beta_j \in (-\infty, \infty)$	$\beta_j \sim N(0, medium)$
Time-to-Event Data, $t_i > 0$			
With Covariates, Cox PH	$h(t_i) = h_0(t_i)\exp[\mathbf{x}_i'\boldsymbol{\beta}]$ $h_0(t_i)$ unspecified	$\beta_j \in (-\infty, \infty)$	$\beta_j \sim N(0, medium)$
With Covariates, Weibull PH	$h(t_i) = h_0(t_i)\exp[\mathbf{x}_i'\boldsymbol{\beta}]$ $h_0(t_i) = pt_i^{p-1}$	$\beta_j \in (-\infty, \infty)$ $p > 0$	$\beta_j \sim N(0, medium)$ $p \sim Gamma(a, b)$

Table 1: Basic model recommendations for various data types. In the case of simple linear regression, $\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + x_i\beta_1$. For multiple linear regression, we could have $\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + x_{i1}\beta_1 + \ldots + x_{iJ}\beta_J$.

[1]For some types of data, it may be more appropriate to model a transformation of Y_i as being normally distributed; e.g., if Y_i is household income, we may want to model $\log Y_i \sim N(\mu, \sigma^2)$

[2]Because $P(Y_i = 1) \in [0, 1]$, we require the logit link function: $\text{logit}P(Y_i = 1) = \mathbf{x}_i'\boldsymbol{\beta} \in (-\infty, \infty)$

[3]While it is possible to model count data using a $\text{Bin}(n, \theta)$ distribution, the Poisson distribution is preferred when n is large and θ is small (e.g., the number of deaths due to heart disease in a state).

[4]Oftentimes (as presented in Harrison's slides), we want to account for the population size, n_i, in our Poisson model. This could be done by letting $Y_i \sim \text{Pois}(n_i\lambda)$ or (equivalently) $Y_i \sim \text{Pois}(E_i\lambda)$, as shown in the slides.

[5]Because $E(Y_i) > 0$, we require the log link: $\log E(Y_i) = \mathbf{x}_i'\boldsymbol{\beta} \in (-\infty, \infty)$