

Estimation and Inference Via Bayesian Simulation: A Practical Introduction

Herbert M. Blalock Lecture Series
on Advanced Topics in Social Research

Simon Jackman
Department of Political Science
Stanford University

July 30 - August 3, 2001

Contents

Click on the highlighted text to jump ahead:

[Inference for a Proportion](#)

[Meta-analysis via hierarchical modeling](#)

What is “Bayesian Simulation”?

2

What is “Bayesian Simulation”?

2

Shorthand for the following suite of algorithms for estimation and inference:

What is “Bayesian Simulation”?

2

Shorthand for the following suite of algorithms for estimation and inference:

- Markov chain Monte Carlo (MCMC)

What is “Bayesian Simulation”?

2

Shorthand for the following suite of algorithms for estimation and inference:

- Markov chain Monte Carlo (MCMC)
- Gibbs sampler

What is “Bayesian Simulation”?

Shorthand for the following suite of algorithms for estimation and inference:

- Markov chain Monte Carlo (MCMC)
- Gibbs sampler
- **Metropolis-Hastings**

What is “Bayesian Simulation”?

2

Shorthand for the following suite of algorithms for estimation and inference:

- Markov chain Monte Carlo (MCMC)
- Gibbs sampler
- Metropolis-Hastings
- **data augmentation**

What is “Bayesian Simulation”?

Shorthand for the following suite of algorithms for estimation and inference:

- Markov chain Monte Carlo (MCMC)
- Gibbs sampler
- Metropolis-Hastings
- data augmentation
- **multiple imputation**

What is “Bayesian Simulation”?

3

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:

What is “Bayesian Simulation”?

3

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. **hierarchical models**

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. hierarchical models
 2. data sets with missing data

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. hierarchical models
 2. data sets with missing data
 3. **item-response models (measurement models)**

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. hierarchical models
 2. data sets with missing data
 3. item-response models (measurement models)
 4. **using the t distribution to model heavily-tailed data**

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. hierarchical models
 2. data sets with missing data
 3. item-response models (measurement models)
 4. using the t distribution to model heavily-tailed data
 5. **mixture models**

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. hierarchical models
 2. data sets with missing data
 3. item-response models (measurement models)
 4. using the t distribution to model heavily-tailed data
 5. mixture models
 6. **models with dynamics in latent variables** (transitional models for discrete longitudinal data)

What is “Bayesian Simulation”?

- Any model that can be estimated by maximum likelihood can be estimated by Bayesian simulation
- But Bayesian simulation lets us work with models and data set long thought to be “too hard”:
 1. hierarchical models
 2. data sets with missing data
 3. item-response models (measurement models)
 4. using the t distribution to model heavily-tailed data
 5. mixture models
 6. models with dynamics in latent variables (transitional models for discrete longitudinal data)
 7. **combinations of these types of models**

What is Bayesian Simulation?

4

Bayesian simulation is

especially given continuing advances in computing power

What is Bayesian Simulation?

4

Bayesian simulation is

easy

especially given continuing advances in computing power

The Monte Carlo Principle

- Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .

First anticipated in the first half of the 20th century
Metropolis and Ulam (1949)

The Monte Carlo Principle

- Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .
- Moreover, the **precision** with which we learn about x is limited only by the number of samples from $f(x)$ we can generate, store, and summarize.

First anticipated in the first half of the 20th century
Metropolis and Ulam (1949)

The Monte Carlo Principle

- Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .
- Moreover, the **precision** with which we learn about x is limited only by the number of samples from $f(x)$ we can generate, store, and summarize.
- **Modern computing power lets us exploit this principle as never before possible.** First anticipated in the first half of the 20th century Metropolis and Ulam (1949)

Example 1: Use your RNG instead of limit theorems!

6

Go to Splus demo.

Example 1: Use your RNG instead of limit theorems!

6

Statistical theory tells us that if

$$x \sim N(0, 1)$$

Go to Splus demo.

Example 1: Use your RNG instead of limit theorems!

Statistical theory tells us that if

$$x \sim N(0, 1)$$

then

$$\Pr(x > 1.96) =$$

Go to Splus demo.

Example 1: Use your RNG instead of limit theorems!

Statistical theory tells us that if

$$x \sim N(0, 1)$$

then

$$\Pr(x > 1.96) = .025$$

Go to Splus demo.

Example 2: Use your RNG instead of limit theorems!

7

But what about the following example?

$$x \sim N(0, 1)$$

- Use simulation methods to directly inspect the small-sample distribution of z .

Example 2: Use your RNG instead of limit theorems!

7

But what about the following example?

$$x \sim N(0, 1)$$

$$y \sim N(0, 1)$$

- Use simulation methods to directly inspect the small-sample distribution of z .

Example 2: Use your RNG instead of limit theorems!

7

But what about the following example?

$$x \sim N(0, 1)$$

$$y \sim N(0, 1)$$

$$z = \frac{x}{\sqrt{y}}$$

- Use simulation methods to directly inspect the small-sample distribution of z .

Example 2: Use your RNG instead of limit theorems!

7

But what about the following example?

$$x \sim N(0, 1)$$

$$y \sim N(0, 1)$$

$$z = \frac{x}{\sqrt{y}}$$

- The distribution of z is...?
- Use simulation methods to directly inspect the small-sample distribution of z .

Example 2: Use your RNG instead of limit theorems!

7

But what about the following example?

$$x \sim N(0, 1)$$

$$y \sim N(0, 1)$$

$$z = \frac{x}{\sqrt{y}}$$

- The distribution of z is...?
- Can derive an *asymptotically-valid* approximation using a technique known as the delta method.
- Use simulation methods to directly inspect the small-sample distribution of z .

Example 2: Use your RNG instead of limit theorems!

But what about the following example?

$$x \sim N(0, 1)$$

$$y \sim N(0, 1)$$

$$z = \frac{x}{\sqrt{y}}$$

- The distribution of z is...?
- Can derive an *asymptotically-valid* approximation using a technique known as the delta method. But how does this work in a small or “finite” sample?
- Use simulation methods to directly inspect the small-sample distribution of z .

A Quick Review of Maximum Likelihood

The likelihood function plays a critical role in Bayesian simulation, and so we briefly review it here.

Doing Maximum Likelihood

- Specify a parametric probability model for the data:

assuming independence across observations

In this case,

$$L(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}) \propto f(\mathbf{y} | \boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^n \phi\left(\frac{y_i - \mu}{\sigma}\right)$$

where ϕ is the standard Normal probability density function.

Doing Maximum Likelihood

- Specify a parametric probability model for the data: e.g.,

$$y_i \sim N(\mu, \sigma^2) \quad i = 1, \dots, n$$

assuming independence across observations

In this case,

$$L(\mu, \sigma^2 | \mathbf{y}) \propto f(\mathbf{y} | \mu, \sigma^2) = \prod_{i=1}^n \phi\left(\frac{y_i - \mu}{\sigma}\right)$$

where ϕ is the standard Normal probability density function.

Doing Maximum Likelihood

- Specify a parametric probability model for the data: e.g.,

$$y_i \sim N(\mu, \sigma^2) \quad i = 1, \dots, n$$

- **Definition:** Any function proportional to the joint probability density function for the data is known as the **likelihood function**. In this case, assuming independence across observations

$$L(\mu, \sigma^2 | \mathbf{y}) \propto f(\mathbf{y} | \mu, \sigma^2) = \prod_{i=1}^n \phi\left(\frac{y_i - \mu}{\sigma}\right)$$

where ϕ is the standard Normal probability density function.

Doing Maximum Likelihood

- Find those values of μ and σ^2 that maximize the likelihood function.

Or a “mode-finder” such as the EM algorithm.

Doing Maximum Likelihood

- Find those values of μ and σ^2 that maximize the likelihood function. These are the MLEs.

Or a “mode-finder” such as the EM algorithm.

Doing Maximum Likelihood

- Find those values of μ and σ^2 that maximize the likelihood function. These are the MLEs.
- Get the MLEs via some kind of hill climbing algorithm: e.g., Newton-Raphson, quasi-Newton, EM algorithm, finding that point in the parameter space where the likelihood function is at its maximum. Or a “mode-finder” such as the EM algorithm.

Inference for MLEs

- Precision of the MLEs is based on the shape of the likelihood function in the neighborhood of the MLEs.

- Precision of the MLEs is based on the shape of the likelihood function in the neighborhood of the MLEs.
- How far can I move in the parameter space, away from the MLEs, before I encounter a statistically significant fall in likelihood?

- Precision of the MLEs is based on the shape of the likelihood function in the neighborhood of the MLEs.
- How far can I move in the parameter space, away from the MLEs, before I encounter a statistically significant fall in likelihood?
- How wide is the set of parameter values that are equally likely?

- Precision of the MLEs is based on the shape of the likelihood function in the neighborhood of the MLEs.
- How far can I move in the parameter space, away from the MLEs, before I encounter a statistically significant fall in likelihood?
- How wide is the set of parameter values that are equally likely?
- Does zero lie in this set of equally likely parameter values?

- Almost all applications of MLE characterize the shape of the likelihood function with the matrix of second derivatives (the Hessian matrix) evaluated at the MLEs, \mathbf{H} .

- Almost all applications of MLE characterize the shape of the likelihood function with the matrix of second derivatives (the Hessian matrix) evaluated at the MLEs, \mathbf{H} .
- $\widehat{\text{var}}(\hat{\theta}_{\text{MLE}}) = -\mathbf{H}^{-1}$.

- Almost all applications of MLE characterize the shape of the likelihood function with the matrix of second derivatives (the Hessian matrix) evaluated at the MLEs, \mathbf{H} .
- $\widehat{\text{var}}(\hat{\theta}_{\text{MLE}}) = -\mathbf{H}^{-1}$.
- $\hat{\theta} \stackrel{\text{asy}}{\sim} N\left(\hat{\theta}_{\text{MLE}}, \widehat{\text{var}}\left(\hat{\theta}_{\text{MLE}}\right)\right)$

- Almost all applications of MLE characterize the shape of the likelihood function with the matrix of second derivatives (the Hessian matrix) evaluated at the MLEs, \mathbf{H} .
- $\widehat{\text{var}}(\hat{\theta}_{\text{MLE}}) = -\mathbf{H}^{-1}$.
- $\hat{\theta} \overset{\text{asy}}{\sim} N\left(\hat{\theta}_{\text{MLE}}, \widehat{\text{var}}\left(\hat{\theta}_{\text{MLE}}\right)\right)$
- Probability statements about $\hat{\theta}$ follow (e.g., confidence intervals).

- (1) Use of the Normal distribution as a characterization of uncertainty in $\hat{\theta}$; and (2) the use of the Hessian matrix to characterize the shape of the likelihood function are valid *asymptotically*,
- No need to rely on asymptotically-valid approximations with Bayesian simulation.

- (1) Use of the Normal distribution as a characterization of uncertainty in $\hat{\theta}$; and (2) the use of the Hessian matrix to characterize the shape of the likelihood function are valid *asymptotically*, but in finite samples are *merely approximations*.
- No need to rely on asymptotically-valid approximations with Bayesian simulation.

- (1) Use of the Normal distribution as a characterization of uncertainty in $\hat{\theta}$; and (2) the use of the Hessian matrix to characterize the shape of the likelihood function are valid *asymptotically*, but in finite samples are *merely approximations*.
- Most of the time, probably a good approximation.
- No need to rely on asymptotically-valid approximations with Bayesian simulation.

- (1) Use of the Normal distribution as a characterization of uncertainty in $\hat{\theta}$; and (2) the use of the Hessian matrix to characterize the shape of the likelihood function are valid *asymptotically*, but in finite samples are *merely approximations*.
- Most of the time, probably a good approximation. But we often need someone to do analytics and/or Monte-Carlo work to examine the finite-sample properties of particular estimators in particular situations.
- No need to rely on asymptotically-valid approximations with Bayesian simulation.

How Can the Maximum Likelihood Paradigm Let Us Down?

14

At the frontiers of quantitative social science:

(Jackman 2000, 377)

How Can the Maximum Likelihood Paradigm Let Us Down?

At the frontiers of quantitative social science:

Substantively interesting statistical models can give rise to complex likelihood functions, having either lots of parameters or a computationally intractable functional form, or both....

(Jackman 2000, 377)

How Can the Maximum Likelihood Paradigm Let Us Down?

At the frontiers of quantitative social science:

Substantively interesting statistical models can give rise to complex likelihood functions, having either lots of parameters or a computationally intractable functional form, or both....
Maximization algorithms may reach terminal solutions extremely slowly or not at all....

(Jackman 2000, 377)

How Can the Maximum Likelihood Paradigm Let Us Down?

At the frontiers of quantitative social science:

Substantively interesting statistical models can give rise to complex likelihood functions, having either lots of parameters or a computationally intractable functional form, or both.... Maximization algorithms may reach terminal solutions extremely slowly or not at all.... **In other cases the likelihood will be known *a priori* not to have a unique maximum....**

(Jackman 2000, 377)

How Can the Maximum Likelihood Paradigm Let Us Down?

At the frontiers of quantitative social science:

Substantively interesting statistical models can give rise to complex likelihood functions, having either lots of parameters or a computationally intractable functional form, or both.... Maximization algorithms may reach terminal solutions extremely slowly or not at all.... In other cases the likelihood will be known *a priori* not to have a unique maximum.... **In yet another class of cases, the researcher may want to estimate not just parameters, but the values of missing data points as well, complicating the optimization problem substantially..** (Jackman 2000, 377)

Bayesian simulation lets us do more, and better

As King, Tomz and Wittenberg (2000, 353) put it

...there is a simulation-based alternative to nearly every analytical method of computing quantities of interest and conducting statistical tests, but the reverse is not true. Thus, simulation can provide accurate answers even when no analytical solutions exist.

Relies on two principles:

Or, a big/hard problem can be broken down into a interrelated series of smaller/easier problems.

Together these principles constitute MCMC: Markov-chain Monte Carlo.

Relies on two principles:

1. **The Monte Carlo Principle:** Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .

Or, a big/hard problem can be broken down into a interrelated series of smaller/easier problems.

Together these principles constitute MCMC: Markov-chain Monte Carlo.

Relies on two principles:

1. **The Monte Carlo Principle:** Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .

Moreover, the **precision** with which we learn about x is limited only by the number of samples from $f(x)$ we can generate, store, and summarize.

Or, a big/hard problem can be broken down into a interrelated series of smaller/easier problems.

Together these principles constitute MCMC: Markov-chain Monte Carlo.

Relies on two principles:

1. **The Monte Carlo Principle:** Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .

Moreover, the **precision** with which we learn about x is limited only by the number of samples from $f(x)$ we can generate, store, and summarize.

Modern computing power lets us exploit this principle as never before possible.

Or, a big/hard problem can be broken down into a interrelated series of smaller/easier problems.

Together these principles constitute MCMC: Markov-chain Monte Carlo.

Relies on two principles:

1. **The Monte Carlo Principle:** Anything we want to know about a random variable x can be learned by sampling many times from $f(x)$, the pdf of x .

Moreover, the **precision** with which we learn about x is limited only by the number of samples from $f(x)$ we can generate, store, and summarize.

Modern computing power lets us exploit this principle as never before possible.

2. **High-dimensional joint densities are completely characterized by lower-dimensional conditional densities.** Or, a big/hard problem can be broken down into a interrelated series of smaller/easier problems.

Together these principles constitute MCMC: Markov-chain Monte Carlo.

What is Bayesian about Bayesian simulation?

- These two ideas have been long known by Bayesians.

(learn about y , conditional on what we know about x)

- Iterate these two steps

What is Bayesian about Bayesian simulation?

- These two ideas have been long known by Bayesians.
- In particular, the idea that a joint distribution is completely characterized by conditional densities
- To learn about $f(x, y)$:
 1. $g(x|y)$
(learn about y , conditional on what we know about x)
- Iterate these two steps

What is Bayesian about Bayesian simulation?

- These two ideas have been long known by Bayesians.
- In particular, the idea that a joint distribution is completely characterized by conditional densities
- To learn about $f(x, y)$:
 1. $g(x|y)$ (learn about x , conditional on what we know about y)
(learn about y , conditional on what we know about x)
- Iterate these two steps

What is Bayesian about Bayesian simulation?

- These two ideas have been long known by Bayesians.
- In particular, the idea that a joint distribution is completely characterized by conditional densities
- To learn about $f(x, y)$:
 1. $g(x|y)$ (learn about x , conditional on what we know about y)
 2. $g(y|x)$ (learn about y , conditional on what we know about x)
- Iterate these two steps

What is Bayesian about Bayesian simulation?

18

What is Bayesian about Bayesian simulation?

- Bayesian statistics: “what should I believe about the parameters having seen the data?”

What is Bayesian about Bayesian simulation?

- Bayesian statistics: “what should I believe about the parameters having seen the data?”
- A *subjective* notion of probability,

What is Bayesian about Bayesian simulation?

- Bayesian statistics: “what should I believe about the parameters having seen the data?”
- A *subjective* notion of probability, sometimes controversial, though less so now than 25 years ago

What is Bayesian about Bayesian simulation?

- Bayesian statistics: “what should I believe about the parameters having seen the data?”
- A *subjective* notion of probability, sometimes controversial, though less so now than 25 years ago
- The **posterior density** is the formalization of these beliefs,

What is Bayesian about Bayesian simulation?

- Bayesian statistics: “what should I believe about the parameters having seen the data?”
- A *subjective* notion of probability, sometimes controversial, though less so now than 25 years ago
- The **posterior density** is the formalization of these beliefs, written as $\pi(\boldsymbol{\theta}|\text{data})$.

The Bayesian Mantra

The Bayesian Mantra

a posterior is proportional to a prior times a likelihood

The Bayesian Mantra

a posterior is proportional to a prior times a likelihood

The Bayesian Mantra

a posterior is proportional to a prior times a likelihood

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}$$

The Bayesian Mantra

a posterior is proportional to a prior times a likelihood

POSTERIOR \propto PRIOR \times LIKELIHOOD

$$\pi(\boldsymbol{\theta}|\text{data}) \propto \pi(\boldsymbol{\theta}) f(\text{data}|\boldsymbol{\theta})$$

What is Bayesian about Bayesian simulation?

20

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}$$

a “trivial” answer (our prior).

Perhaps

What is Bayesian about Bayesian simulation?

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}$$

- As priors become vague, $\pi(\boldsymbol{\theta}) \rightarrow c$ (a constant), and the posterior is proportional to the likelihood.

a “trivial” answer (our prior).

Perhaps

What is Bayesian about Bayesian simulation?

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}$$

- As priors become vague, $\pi(\boldsymbol{\theta}) \rightarrow c$ (a constant), and the posterior is proportional to the likelihood.
- When priors are vague or “diffuse”, Bayesian approaches yield the same answers as likelihood approaches.

a “trivial” answer (our prior).

Perhaps

What is Bayesian about Bayesian simulation?

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}$$

- As priors become vague, $\pi(\boldsymbol{\theta}) \rightarrow c$ (a constant), and the posterior is proportional to the likelihood.
- When priors are vague or “diffuse”, Bayesian approaches yield the same answers as likelihood approaches.
- But Bayesian methods give us answers when likelihood can’t! Perhaps a “trivial” answer (our prior).

What is Bayesian about Bayesian simulation?

- Bayesian simulation means we recover the posterior or likelihood *exactly*, without any recourse to asymptotically-valid approximations about the shape of the likelihood.

This is of tremendous help when working with models that are barely identified or even unidentified.

What is Bayesian about Bayesian simulation?

- Bayesian simulation means we recover the posterior or likelihood *exactly*, without any recourse to asymptotically-valid approximations about the shape of the likelihood.
- It is always possible to use “informative” priors. This is of tremendous help when working with models that are barely identified or even unidentified.

What is Bayesian about Bayesian simulation?

22

What is Bayesian about Bayesian simulation?

22

- Bayesian estimation and inference consists of *computing* and *communicating* interesting features of the posterior density

What is Bayesian about Bayesian simulation?

- Bayesian estimation and inference consists of *computing* and *communicating* interesting features of the posterior density (e.g., how much of the posterior density lies above/below zero?)

Example: inference about a log-odds ratio

- The relationship between smoking and lung cancer (Johnson and Albert 1999, 35; Dorn 1954):

	Cancer	Control
Smokers	83	72
Nonsmokers	3	14

Example: inference about a log-odds ratio

- The relationship between smoking and lung cancer (Johnson and Albert 1999, 35; Dorn 1954):

	Cancer	Control
Smokers	83	72
Nonsmokers	3	14

- **Scientific question:** is there a *significant difference* between the smoking habits in the two groups?

Example: inference about a log-odds ratio

- p_L : population proportion of lung-cancer patients who smoke
- p_C : population proportion of control who smoke

Example: inference about a log-odds ratio

- p_L : population proportion of lung-cancer patients who smoke
- p_C : population proportion of control who smoke
- Scientific question of interest could be answered by examining the posterior distribution of $p_L - p_C$.

Example: inference about a log-odds ratio

- p_L : population proportion of lung-cancer patients who smoke
- p_C : population proportion of control who smoke
- Scientific question of interest could be answered by examining the posterior distribution of $p_L - p_C$.
- But this quantity has a highly skewed distribution;

Example: inference about a log-odds ratio

- p_L : population proportion of lung-cancer patients who smoke
- p_C : population proportion of control who smoke
- Scientific question of interest could be answered by examining the posterior distribution of $p_L - p_C$.
- But this quantity has a highly skewed distribution; asymptotic approximations to either (a) sampling distribution of MLE; (b) posterior density are not precise (Johnson and Albert 1999, 35).

Example: inference about a log-odds ratio

- Transform the problem, and study the *log-odds ratio*

$$\lambda = \log \left(\frac{p_L / (1 - p_L)}{p_C / (1 - p_C)} \right)$$

If $p_L > p_C$ then $\lambda > 0$.

- In Bayesian context, we want to compute $\Pr(\lambda > 0 | \text{data})$.
- log-odds ratio has a more symmetric looking distribution -- approximately normal in even a moderate sample

Example: inference about a log-odds ratio

- Transform the problem, and study the *log-odds ratio*

$$\lambda = \log \left(\frac{p_L / (1 - p_L)}{p_C / (1 - p_C)} \right)$$

- $\lambda = 0$ when two proportions are the same. If $p_L > p_C$ then $\lambda > 0$.
- In Bayesian context, we want to compute $\Pr(\lambda > 0 | \text{data})$.
- log-odds ratio has a more symmetric looking distribution -- approximately normal in even a moderate sample

Example: inference about a log-odds ratio

- **Likelihood:** product of two binomial likelihoods, one for the smokers, one for the control group.

- Note that the quantity of scientific interest

$$\lambda = \log \left(\frac{p_L / (1 - p_L)}{p_C / (1 - p_C)} \right)$$

is a nonlinear function of the model parameters.

Example: inference about a log-odds ratio

- **Likelihood:** product of two binomial likelihoods, one for the smokers, one for the control group. n.b., exploiting independence across smokers and control built into the design:

- Note that the quantity of scientific interest

$$\lambda = \log \left(\frac{p_L / (1 - p_L)}{p_C / (1 - p_C)} \right)$$

is a nonlinear function of the model parameters.

Example: inference about a log-odds ratio

- **Likelihood:** product of two binomial likelihoods, one for the smokers, one for the control group. n.b., exploiting independence across smokers and control built into the design:

$$L(p_L, p_C) = p_L^{83}(1 - p_L)^3 p_C^{72}(1 - p_C)^{14}, \quad 0 < p_L, p_C < 1$$

- Note that the quantity of scientific interest

$$\lambda = \log \left(\frac{p_L/(1 - p_L)}{p_C/(1 - p_C)} \right)$$

is a nonlinear function of the model parameters.

Example: inference about a log-odds ratio

- Use simulation to recover the exact distribution of λ .
- Exploit the fact that the components of λ , p_L and p_C have (a) well-known distributions that (b) are independent.

nor does the log of the odds-ratio of two Beta variates.

Example: inference about a log-odds ratio

- Use simulation to recover the exact distribution of λ .
- Exploit the fact that the components of λ , p_L and p_C have (a) well-known distributions that (b) are independent.
- p_L has a Beta distribution, specifically Beta(83,3)
- p_C has Beta(72,14)
- Beta distributions are highly flexible distributions for quantities on the unit interval (i.e., proportions).

nor does the log of the odds-ratio of two Beta variates.

Example: inference about a log-odds ratio

- Use simulation to recover the exact distribution of λ .
- Exploit the fact that the components of λ , p_L and p_C have (a) well-known distributions that (b) are independent.
- p_L has a Beta distribution, specifically Beta(83,3)
- p_C has Beta(72,14)
- Beta distributions are highly flexible distributions for quantities on the unit interval (i.e., proportions). Uniform distribution is a special case. Splus-DEMO.

nor does the log of the odds-ratio of two Beta variates.

Example: inference about a log-odds ratio

- Use simulation to recover the exact distribution of λ .
- Exploit the fact that the components of λ , p_L and p_C have (a) well-known distributions that (b) are independent.
- p_L has a Beta distribution, specifically Beta(83,3)
- p_C has Beta(72,14)
- Beta distributions are highly flexible distributions for quantities on the unit interval (i.e., proportions). Uniform distribution is a special case. Splus-DEMO.
- But neither the sum nor the difference of Beta variates has a standard form, nor does the log of the odds-ratio of two Beta variates.

Example: inference about a log-odds ratio

Repeat the following many times:

1. Sample $p_L^{(t)}$ from a Beta(83,3) density

3. Compute

$$\lambda^{(t)} = \log \left(\frac{p_L^{(t)} / (1 - p_L^{(t)})}{p_C^{(t)} / (1 - p_C^{(t)})} \right)$$

Example: inference about a log-odds ratio

Repeat the following many times:

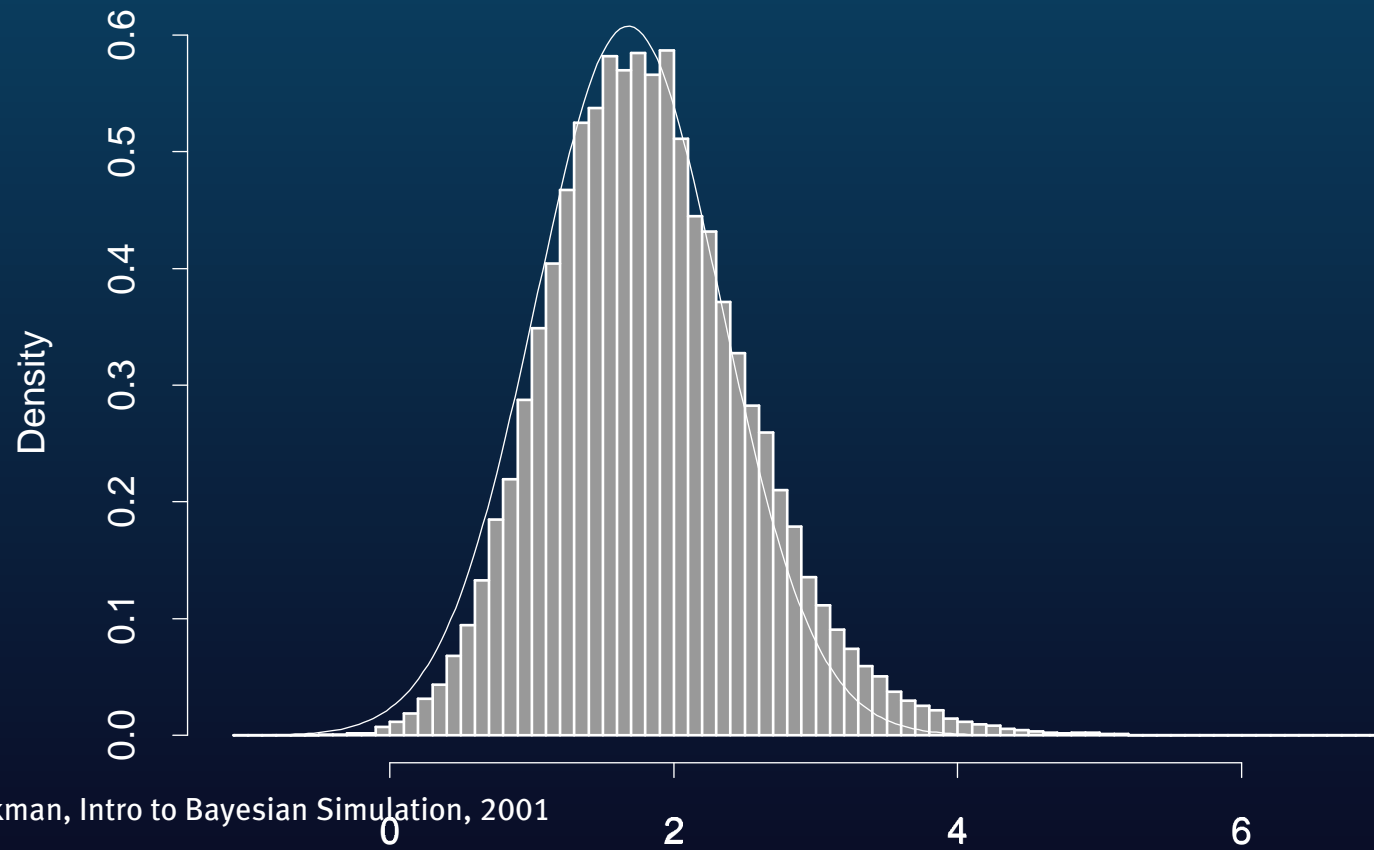
1. Sample $p_L^{(t)}$ from a Beta(83,3) density
2. Sample $p_C^{(t)}$ from a Beta(72,14) density
3. Compute

$$\lambda^{(t)} = \log \left(\frac{p_L^{(t)} / (1 - p_L^{(t)})}{p_C^{(t)} / (1 - p_C^{(t)})} \right)$$

Posterior Density for λ

29

	$E(\lambda)$	$sd(\lambda)$	$\Pr(\lambda > 0)$
Direct simulation:	1.82	.70	.999
Normal approximation:	1.68	.66	.995



What happens if we (arbitrarily) assume we were working with a much smaller sample?

Posterior Density for λ

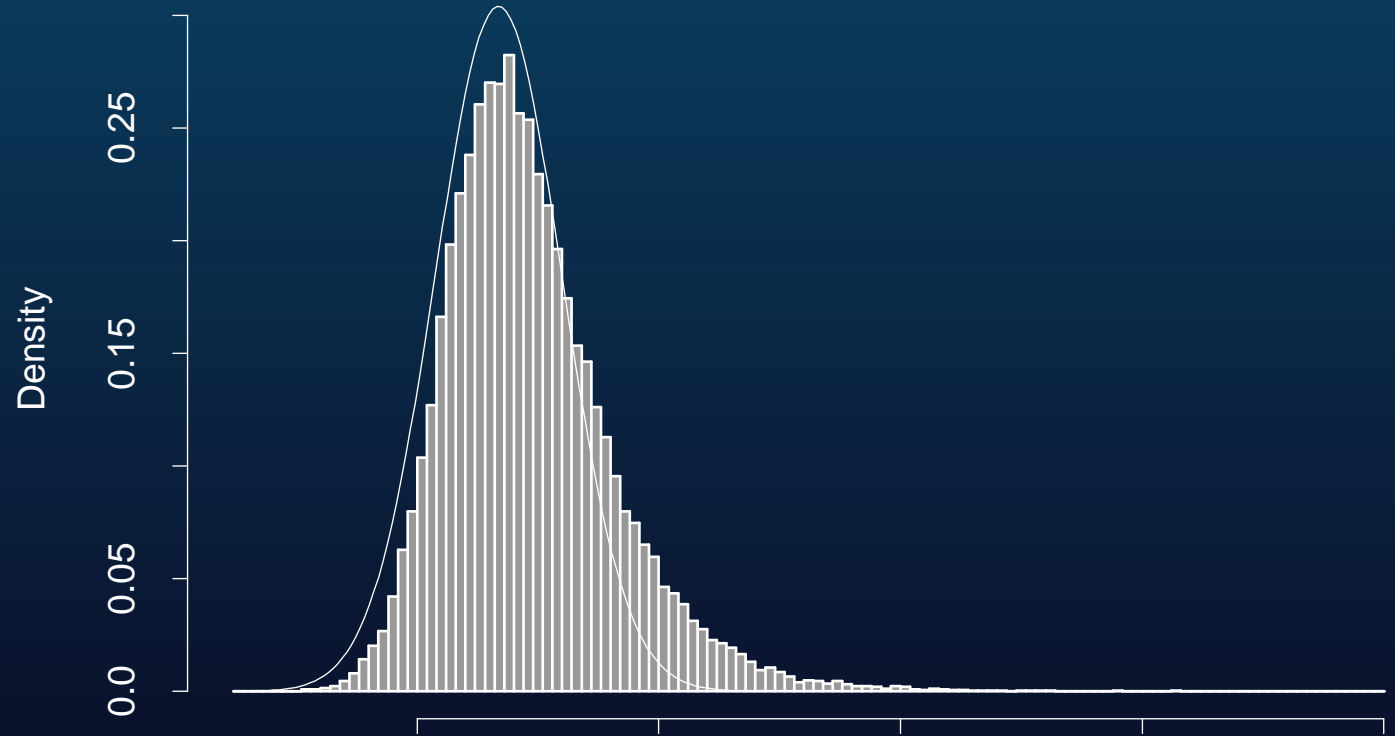
What happens if we (arbitrarily) assume we were working with a much smaller sample? How good is the normal approximation now?

	Cancer	Control
Smokers	83/4	72/4
Nonsmokers	3/4	14/4

Posterior Density for λ , 1/4 of original n

32

	$E(\lambda)$	$sd(\lambda)$	$Pr(\lambda > 0)$
Direct simulation:	2.34	1.72	.948
Normal approximation:	1.68	1.31	.900



Example: inference about a log-odds ratio

Can also use simulation methods to compute the difference

$$\delta = p_L - p_C$$

and ask questions such as

$$\Pr(\delta > 0 | \text{data})$$

Full Sample: .998

Quarter Sample: .950

- Bayesian Inference Using Gibbs Sampling
- Lets re-do the lung cancer example using WinBUGS

All in S-PLUS.

Computing Posterior Densities with The Gibbs Sampler 37

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

Computing Posterior Densities with The Gibbs Sampler 37

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

The high-dimensional posterior density $\pi(\boldsymbol{\theta}|\text{data})$ can be computed via the following scheme

Computing Posterior Densities with The Gibbs Sampler 37

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

The high-dimensional posterior density $\pi(\boldsymbol{\theta}|\text{data})$ can be computed via the following scheme (the Gibbs sampler):

Computing Posterior Densities with The Gibbs Sampler

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

The high-dimensional posterior density $\pi(\boldsymbol{\theta}|\text{data})$ can be computed via the following scheme (the Gibbs sampler): we start iteration t with the output of iteration $t - 1$: $\boldsymbol{\theta}^{(t-1)} = (\boldsymbol{\theta}_1^{(t-1)}, \boldsymbol{\theta}_2^{(t-1)})'$

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

The high-dimensional posterior density $\pi(\boldsymbol{\theta}|\text{data})$ can be computed via the following scheme (the Gibbs sampler): we start iteration t with the output of iteration $t - 1$: $\boldsymbol{\theta}^{(t-1)} = (\boldsymbol{\theta}_1^{(t-1)}, \boldsymbol{\theta}_2^{(t-1)})'$

1. sample $\boldsymbol{\theta}_1^{(t)}$ from $g_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t-1)}, \text{data})$

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

The high-dimensional posterior density $\pi(\boldsymbol{\theta}|\text{data})$ can be computed via the following scheme (the Gibbs sampler): we start iteration t with the output of iteration $t - 1$: $\boldsymbol{\theta}^{(t-1)} = (\boldsymbol{\theta}_1^{(t-1)}, \boldsymbol{\theta}_2^{(t-1)})'$

1. sample $\boldsymbol{\theta}_1^{(t)}$ from $g_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t-1)}, \text{data})$
2. sample $\boldsymbol{\theta}_2^{(t)}$ from $g_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t)}, \text{data})$

Consider a high-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$.

The high-dimensional posterior density $\pi(\boldsymbol{\theta}|\text{data})$ can be computed via the following scheme (the Gibbs sampler): we start iteration t with the output of iteration $t - 1$: $\boldsymbol{\theta}^{(t-1)} = (\boldsymbol{\theta}_1^{(t-1)}, \boldsymbol{\theta}_2^{(t-1)})'$

1. sample $\boldsymbol{\theta}_1^{(t)}$ from $g_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t-1)}, \text{data})$
2. sample $\boldsymbol{\theta}_2^{(t)}$ from $g_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t)}, \text{data})$

This yields $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)})$.

The Gibbs Sampler

- Initialize the sampler with starting values $\theta^{(0)}$.

The Gibbs Sampler

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally:

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally: The sampler moves away from the initial values,

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally: The sampler moves away from the initial values, providing a “random tour” of the (high-dimensional) parameter space,

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally: The sampler moves away from the initial values, providing a “random tour” of the (high-dimensional) parameter space, visiting locations in the parameter space with frequencies proportional to the posterior density.

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally: The sampler moves away from the initial values, providing a “random tour” of the (high-dimensional) parameter space, visiting locations in the parameter space with frequencies proportional to the posterior density.
- Even more formally:

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally: The sampler moves away from the initial values, providing a “random tour” of the (high-dimensional) parameter space, visiting locations in the parameter space with frequencies proportional to the posterior density.
- Even more formally: The output of the Gibbs sampler forms a **Markov chain** on the parameter space for $\boldsymbol{\theta}$,

- Initialize the sampler with starting values $\boldsymbol{\theta}^{(0)}$.
- Let the sampler run, generating $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$.
- Under a very wide set of conditions, as $t \rightarrow \infty$, each Gibbs sample $\boldsymbol{\theta}^{(t)}$ can be regarded as samples from the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.
- More formally: The sampler moves away from the initial values, providing a “random tour” of the (high-dimensional) parameter space, visiting locations in the parameter space with frequencies proportional to the posterior density.
- Even more formally: The output of the Gibbs sampler forms a **Markov chain** on the parameter space for $\boldsymbol{\theta}$, with transition probabilities such that the “equilibrium”, “limiting”, or “stationary” distribution of the Gibbs sampler is the posterior density $\pi(\boldsymbol{\theta}|\text{data})$.

Inference with the Gibbs Sampler

Extremely simple.

Extremely simple.

- Store output of the Gibbs sampler

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals).

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals). Compute any function of the Gibbs sampler output

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals). Compute any function of the Gibbs sampler output

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals). Compute any function of the Gibbs sampler output
- These estimates get “better” with more Gibbs samples;

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals). Compute any function of the Gibbs sampler output
- These estimates get “better” with more Gibbs samples; i.e., how fast and how big is your computer?

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals). Compute any function of the Gibbs sampler output
- These estimates get “better” with more Gibbs samples; i.e., how fast and how big is your computer?
- Modern computing power makes this approach to estimation and inference possible.

Extremely simple.

- Store output of the Gibbs sampler
- Compute any summary statistic you like (mean, median, confidence intervals). Compute any function of the Gibbs sampler output
- These estimates get “better” with more Gibbs samples; i.e., how fast and how big is your computer?
- Modern computing power makes this approach to estimation and inference possible. First anticipated in the first half of the 20th century **Metropolis and Ulam (1949)**

Example: Probit model for binary data

40

Binary response models can be written as a regression where the *latent dependent variable*, y^* , is observed only in terms of sign.

Treat y_i^* as a set of “nuisance parameters”, or missing data.

Example: Probit model for binary data

40

Binary response models can be written as a regression where the *latent dependent variable*, y^* , is observed only in terms of sign.

Probit:

$$y_i^* \sim N(\mu_i, \sigma^2)$$

Treat y_i^* as a set of “nuisance parameters”, or missing data.

Example: Probit model for binary data

Binary response models can be written as a regression where the *latent dependent variable*, y^* , is observed only in terms of sign.

Probit:

$$y_i^* \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

Treat y_i^* as a set of “nuisance parameters”, or missing data.

Example: Probit model for binary data

Binary response models can be written as a regression where the *latent dependent variable*, y^* , is observed only in terms of sign.

Probit:

$$y_i^* \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

$$y_i = \begin{cases} 0 & \iff y_i^* < 0 \\ 1 & \iff y_i^* \geq 0 \end{cases}$$

Treat y_i^* as a set of “nuisance parameters”, or missing data.

Example: Probit model for binary data

40

Binary response models can be written as a regression where the *latent dependent variable*, y^* , is observed only in terms of sign.

Probit:

$$y_i^* \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

$$y_i = \begin{cases} 0 & \iff y_i^* < 0 \\ 1 & \iff y_i^* \geq 0 \end{cases}$$

$$\sigma^2 = 1$$

Treat y_i^* as a set of “nuisance parameters”, or missing data.

Example: Probit model for binary data

Binary response models can be written as a regression where the *latent dependent variable*, y^* , is observed only in terms of sign.

Probit:

$$y_i^* \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

$$y_i = \begin{cases} 0 & \iff y_i^* < 0 \\ 1 & \iff y_i^* \geq 0 \end{cases}$$

$$\sigma^2 = 1$$

If we had y_i^* , we could estimate $\boldsymbol{\beta}$ with regression methods. Treat y_i^* as a set of “nuisance parameters”, or missing data.

1. Sample y^* , conditional on data and β :

$$y_i^* | (\mathbf{x}_i, y_i, \beta) \sim N(\mathbf{x}_i \beta, 1)$$

Recall $\sigma^2 = 1$.

1. Sample y^* , conditional on data and β :

$$y_i^* | (\mathbf{x}_i, y_i, \beta) \sim N(\mathbf{x}_i \beta, 1) I(a_i, b_i)$$

Recall $\sigma^2 = 1$.

1. Sample y^* , conditional on data and β :

$$y_i^* | (\mathbf{x}_i, y_i, \beta) \sim N(\mathbf{x}_i \beta, 1) I(a_i, b_i)$$
$$(a_i, b_i) = \begin{cases} (-\infty, 0) & \iff y_i = 0 \\ (0, \infty) & \iff y_i = 1 \end{cases}$$

where $I(\cdot, \cdot)$ is an indicator function.

2. Sample β , conditional on data and y^* :

Recall $\sigma^2 = 1$.

1. Sample y^* , conditional on data and β :

$$y_i^* | (\mathbf{x}_i, y_i, \beta) \sim N(\mathbf{x}_i \beta, 1) I(a_i, b_i)$$
$$(a_i, b_i) = \begin{cases} (-\infty, 0) & \iff y_i = 0 \\ (0, \infty) & \iff y_i = 1 \end{cases}$$

where $I(\cdot, \cdot)$ is an indicator function.

2. Sample β , conditional on data and y^* :

$$\beta | y^*, \mathbf{X}, \mathbf{y} \sim N(\tilde{\beta}, \tilde{\mathbf{B}})$$
$$\tilde{\beta} = (\mathbf{B}_{\text{prior}}^{-1} + \mathbf{X}'\mathbf{X})^{-1} (\mathbf{B}_{\text{prior}}^{-1} \beta_{\text{prior}} + \mathbf{X}'\mathbf{y}^*)$$
$$\tilde{\mathbf{B}} = (\mathbf{B}_{\text{prior}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

Recall $\sigma^2 = 1$.

1. Sample y^* , conditional on data and β :

$$y_i^* | (\mathbf{x}_i, y_i, \beta) \sim N(\mathbf{x}_i \beta, 1) I(a_i, b_i)$$

$$(a_i, b_i) = \begin{cases} (-\infty, 0) & \iff y_i = 0 \\ (0, \infty) & \iff y_i = 1 \end{cases}$$

where $I(\cdot, \cdot)$ is an indicator function.

2. Sample β , conditional on data and y^* :

$$\beta | y^*, \mathbf{X}, \mathbf{y} \sim N(\tilde{\beta}, \tilde{\mathbf{B}})$$

$$\tilde{\beta} = (\mathbf{B}_{\text{prior}}^{-1} + \mathbf{X}'\mathbf{X})^{-1} (\mathbf{B}_{\text{prior}}^{-1} \beta_{\text{prior}} + \mathbf{X}'\mathbf{y}^*)$$

$$\tilde{\mathbf{B}} = (\mathbf{B}_{\text{prior}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

i.e., run a “Bayesian regression” -- note that with an uninformative prior $\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$ and $\tilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}$ (i.e., the posterior moments are given by simply running a regression of \mathbf{y}^* on \mathbf{X}). Recall $\sigma^2 = 1$.

Example: voter turnout

42

See **Turnout** example in WinBUGS.

Example: measuring political information with item-response models

- See write-up in *Political Analysis* piece.
- See French Test example in WinBUGS.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items,

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has,

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has, which party controls both which house of Congress,

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has, which party controls both which house of Congress, and so on.
- Implement this in France.

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has, which party controls both which house of Congress, and so on.
- Implement this in France. Problem of *item calibration*.
- **Scientific questions: Were our items too hard or too easy?**

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has, which party controls both which house of Congress, and so on.
- Implement this in France. Problem of *item calibration*.
- **Scientific questions:** Were our items too hard or too easy? **Do some items tap political information more so than others?**

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has, which party controls both which house of Congress, and so on.
- Implement this in France. Problem of *item calibration*.
- **Scientific questions:** Were our items too hard or too easy? Do some items tap political information more so than others? **What are the properties of any resulting scale measure?**

each respondent administered 10 items, 12 items administered in all.

Example: measuring political information with item-response models

- “Objective” or “factual” political information items, pioneered by American National Election studies.
- E.g., what job William Rehnquist has, which party controls both which house of Congress, and so on.
- Implement this in France. Problem of *item calibration*.
- **Scientific questions:** Were our items too hard or too easy? Do some items tap political information more so than others? What are the properties of any resulting scale measure?
- Pre-tests of 26 and 25 interviews in April 2000; each respondent administered 10 items, 12 items administered in all.

Test Items

45

	Item	Correct Response	Percent Correct	n
1	Michèle Alliot Marie est la présidente du RPR	True	77%	26

Test Items

45

	Item	Correct Response	Percent Correct	n
1	Michèle Alliot Marie est la présidente du RPR	True	77%	26
2	La Finlande fait partie de l'Union européenne	True	47%	51

Test Items

45

	Item	Correct Response	Percent Correct	n
1	Michèle Alliot Marie est la présidente du RPR	True	77%	26
2	La Finlande fait partie de l'Union européenne	True	47%	51
3	Jean Pierre Chevènement appartient au Parti socialiste	False	35%	26

Test Items

45

	Item	Correct Response	Percent Correct	n
1	Michèle Alliot Marie est la présidente du RPR	True	77%	26
2	La Finlande fait partie de l'Union européenne	True	47%	51
3	Jean Pierre Chevènement appartient au Parti socialiste	False	35%	26
4	Le premier ministre a le droit de dissoudre l'Assemblée nationale	False	59%	51

Test Items

45

	Item	Correct Response	Percent Correct	n
1	Michèle Alliot Marie est la présidente du RPR	True	77%	26
2	La Finlande fait partie de l'Union européenne	True	47%	51
3	Jean Pierre Chevènement appartient au Parti socialiste	False	35%	26
4	Le premier ministre a le droit de dissoudre l'Assemblée nationale	False	59%	51
5	Il y a des ministres communistes dans le gouvernement de Lionel Jospin	True	78%	51

Test Items

45

	Item	Correct Response	Percent Correct	n
1	Michèle Alliot Marie est la présidente du RPR	True	77%	26
2	La Finlande fait partie de l'Union européenne	True	47%	51
3	Jean Pierre Chevènement appartient au Parti socialiste	False	35%	26
4	Le premier ministre a le droit de dissoudre l'Assemblée nationale	False	59%	51
5	Il y a des ministres communistes dans le gouvernement de Lionel Jospin	True	78%	51
6	Le Président de la République est élu pour un mandat de 5 ans	False	86%	51

Test Items

	Item	Correct Response	Percent Correct	n
7	Le Sénat a le pouvoir de renverser le gouvernement	False	55%	51

Test Items

	Item	Correct Response	Percent Correct	n
7	Le Sénat a le pouvoir de renverser le gouvernement	False	55%	51
8	Les députés sont élus au scrutin proportionnel	False	18%	51

Test Items

	Item	Correct Response	Percent Correct	n
7	Le Sénat a le pouvoir de renverser le gouvernement	False	55%	51
8	Les députés sont élus au scrutin proportionnel	False	18%	51
9	Les étrangers qui résident en France depuis 5 ans ont le droit de voter à l'élection présidentielle	False	57%	51

Test Items

46

	Item	Correct Response	Percent Correct	n
7	Le Sénat a le pouvoir de renverser le gouvernement	False	55%	51
8	Les députés sont élus au scrutin proportionnel	False	18%	51
9	Les étrangers qui résident en France depuis 5 ans ont le droit de voter à l'élection présidentielle	False	57%	51
10	L'Etat aide financièrement les partis politique	True	71%	51

Test Items

46

	Item	Correct Response	Percent Correct	n
7	Le Sénat a le pouvoir de renverser le gouvernement	False	55%	51
8	Les députés sont élus au scrutin proportionnel	False	18%	51
9	Les étrangers qui résident en France depuis 5 ans ont le droit de voter à l'élection présidentielle	False	57%	51
10	L'Etat aide financièrement les partis politique	True	71%	51
11	Laurent Fabius appartient au Parti socialiste	True	76%	25

Test Items

	Item	Correct Response	Percent Correct	n
7	Le Sénat a le pouvoir de renverser le gouvernement	False	55%	51
8	Les députés sont élus au scrutin proportionnel	False	18%	51
9	Les étrangers qui résident en France depuis 5 ans ont le droit de voter à l'élection présidentielle	False	57%	51
10	L'Etat aide financièrement les partis politiques	True	71%	51
11	Laurent Fabius appartient au Parti socialiste	True	76%	25
12	Jorg Haider est le leader du parti libéral autrichien	False	20%	25

Two-Parameter Item-Response Model

47

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

Two-Parameter Item-Response Model

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- y_{ij} is the i -th respondent's answer to the j -th political information item (1 if correct, 0 if incorrect, with no answers considered an incorrect response)

Two-Parameter Item-Response Model

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- y_{ij} is the i -th respondent's answer to the j -th political information item (1 if correct, 0 if incorrect, with no answers considered an incorrect response)
- x_i is i -th respondent's latent level of political information
- β_{j1} is an unknown parameter, tapping the *item discrimination* of the j -th item, the extent to which the probability of a correct answer responds to levels of political information

Two-Parameter Item-Response Model

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- y_{ij} is the i -th respondent's answer to the j -th political information item (1 if correct, 0 if incorrect, with no answers considered an incorrect response)
- x_i is i -th respondent's latent level of political information
- β_{j1} is an unknown parameter, tapping the *item discrimination* of the j -th item, the extent to which the probability of a correct answer responds to levels of political information
- β_{j2} is an unknown *item difficulty* parameter, tapping the probability of a correct answer irrespective of levels of political information

- $F(\cdot)$ maps from the real line to the unit probability interval; use probit, ⁴⁸
 $F \equiv \Phi$

Likelihood function for 2P Item-Response Model

49

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\cdot 1}, \boldsymbol{\beta}_{\cdot 2}, \mathbf{x})'$.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\cdot 1}, \boldsymbol{\beta}_{\cdot 2}, \mathbf{x})'$.
- Lots of parameters: 51 latent traits and 24 item parameters, for a total of 75 parameters.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\cdot 1}, \boldsymbol{\beta}_{\cdot 2}, \mathbf{x})'$.
- Lots of parameters: 51 latent traits and 24 item parameters, for a total of 75 parameters.
- Formidable computing problem with bigger data sets:

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\cdot 1}, \boldsymbol{\beta}_{\cdot 2}, \mathbf{x})'$.
- Lots of parameters: 51 latent traits and 24 item parameters, for a total of 75 parameters.
- Formidable computing problem with bigger data sets: e.g., 1,000 subjects taking tests with 50 items, we have 1,100 parameters to estimate (1,000 x_i parameters, and $2 \times 50 = 100$ item parameters).

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

(or vice-versa).

- Bayesian simulation lets us do better.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Likelihood folks typically use EM to find joint MLEs;

(or vice-versa).

- Bayesian simulation lets us do better.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Likelihood folks typically use EM to find joint MLEs;
- Inference via *marginalization*:

(or vice-versa).

- Bayesian simulation lets us do better.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Likelihood folks typically use EM to find joint MLEs;
- Inference via *marginalization*: hold latent traits \mathbf{x} fixed, so as to do perform inference for item-parameters $\boldsymbol{\beta}$;

(or vice-versa).

- Bayesian simulation lets us do better.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Likelihood folks typically use EM to find joint MLEs;
- Inference via *marginalization*: hold latent traits \mathbf{x} fixed, so as to do perform inference for item-parameters $\boldsymbol{\beta}$;
- Marginalization gives rise to a much smaller matrix inversion problem (e.g., 24 item parameters vs 75 total parameters).

(or vice-versa).

- Bayesian simulation lets us do better.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Likelihood folks typically use EM to find joint MLEs;
- Inference via *marginalization*: hold latent traits \mathbf{x} fixed, so as to do perform inference for item-parameters $\boldsymbol{\beta}$;
- Marginalization gives rise to a much smaller matrix inversion problem (e.g., 24 item parameters vs 75 total parameters).
- But standard errors produced by this method will generally be too small;
(or vice-versa).
- Bayesian simulation lets us do better.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

- Likelihood folks typically use EM to find joint MLEs;
- Inference via *marginalization*: hold latent traits \mathbf{x} fixed, so as to do perform inference for item-parameters $\boldsymbol{\beta}$;
- Marginalization gives rise to a much smaller matrix inversion problem (e.g., 24 item parameters vs 75 total parameters).
- But standard errors produced by this method will generally be too small; uncertainty in latent traits not allowed to propagate into uncertainty in item parameters (or vice-versa).
- Bayesian simulation lets us do better.

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- Likelihood folks estimate subject to an additivity constraint: $\sum_{i=1}^n x_i = 0$.

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- Recall we're estimating both x_i and β_j
- Likelihood folks estimate subject to an additivity constraint: $\sum_{i=1}^n x_i = 0$.

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- Recall we're estimating both x_i and β_j
- Unidentified!
- Likelihood folks estimate subject to an additivity constraint: $\sum_{i=1}^n x_i = 0$.

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- Recall we're estimating both x_i and β_j
- Unidentified! *Scale invariance* problem.
- Likelihood folks estimate subject to an additivity constraint: $\sum_{i=1}^n x_i = 0$.

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- Recall we're estimating both x_i and β_j
- Unidentified! *Scale invariance* problem. Can arbitrarily re-scale x_i with offsetting re-scaling of β_{j1} .
- Likelihood folks estimate subject to an additivity constraint: $\sum_{i=1}^n x_i = 0$.

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- With Bayes, unidentified parameters can always be identified via a proper prior.
- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- With Bayes, unidentified parameters can always be identified via a proper prior. How so?
- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- With Bayes, unidentified parameters can always be identified via a proper prior. How so?

$$\pi(\theta|\text{data}) \propto \pi(\theta) f(\text{data}|\theta)$$

- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- With Bayes, unidentified parameters can always be identified via a proper prior. How so?

$$\pi(\theta|\text{data}) \propto \pi(\theta) f(\text{data}|\theta)$$

$$\log \pi(\theta|\text{data}) = \log \pi(\theta) + \log f(\text{data}|\theta) + \log K$$

- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- With Bayes, unidentified parameters can always be identified via a proper prior. How so?

$$\pi(\theta|\text{data}) \propto \pi(\theta) f(\text{data}|\theta)$$

$$\log \pi(\theta|\text{data}) = \log \pi(\theta) + \log f(\text{data}|\theta) + \log K$$

$$H(\log \text{posterior}) = H(\log \text{prior}) + H(\log \text{likelihood})$$

- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

$$p_{ij} \equiv \Pr[y_{ij} = \text{“Correct”}] = F(\beta_{j1}x_i - \beta_{j2})$$

- With Bayes, unidentified parameters can always be identified via a proper prior. How so?

$$\pi(\theta|\text{data}) \propto \pi(\theta) f(\text{data}|\theta)$$

$$\log \pi(\theta|\text{data}) = \log \pi(\theta) + \log f(\text{data}|\theta) + \log K$$

$$H(\log \text{posterior}) = H(\log \text{prior}) + H(\log \text{likelihood})$$

- i.e., non-pd + pd = pd
- Prior for x_i : $x_i \sim N(0, 1)$.
- Solve problem of *rotational invariance* with bounded priors on (at least some) β_{j1} .

Exploit the fact that probit models are latent linear regression models; i.e., partially observed dependent variable \mathbf{y}^* .

1. sample $\mathbf{Y}^{*(t)}$ from $\pi(\mathbf{Y}^* | \mathbf{x}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y})$
2. sample $\mathbf{x}^{(t)}$ from $\pi(\mathbf{x} | \mathbf{Y}^{*(t)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y})$
3. sample $\boldsymbol{\beta}^{(t)}$ from $\pi(\boldsymbol{\beta} | \mathbf{Y}^{*(t)}, \mathbf{x}^{(t)}, \mathbf{Y})$

where t indexes iterations of the Gibbs sampler.

- What is a *hierarchical* model?

or time.

- What is a *hierarchical* model? Very broad term, actually refers to lots of set-ups.
- **Multi-level** models are a classic case:

or time.

- What is a *hierarchical* model? Very broad term, actually refers to lots of set-ups.
- **Multi-level** models are a classic case:

$$y_{ij} \sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma^2)$$

or time.

- What is a *hierarchical* model? Very broad term, actually refers to lots of set-ups.
- **Multi-level** models are a classic case:

$$y_{ij} \sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma^2)$$
$$\boldsymbol{\beta}_j \sim N(\mathbf{z}_j\boldsymbol{\gamma}, \omega^2)$$

or time.

- What is a *hierarchical* model? Very broad term, actually refers to lots of set-ups.
- **Multi-level** models are a classic case:

$$y_{ij} \sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma^2)$$

$$\boldsymbol{\beta}_j \sim N(\mathbf{z}_j\boldsymbol{\gamma}, \omega^2)$$

$$\boldsymbol{\gamma} \sim N(\mathbf{g}_0, \mathbf{G}_0)$$

or time.

- What is a *hierarchical* model? Very broad term, actually refers to lots of set-ups.
- **Multi-level** models are a classic case:

$$y_{ij} \sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma^2)$$

$$\boldsymbol{\beta}_j \sim N(\mathbf{z}_j\boldsymbol{\gamma}, \omega^2)$$

$$\boldsymbol{\gamma} \sim N(\mathbf{g}_0, \mathbf{G}_0)$$

where, say, $i = 1, \dots, n_j$, indexes survey respondents or time.

- What is a *hierarchical* model? Very broad term, actually refers to lots of set-ups.
- **Multi-level** models are a classic case:

$$y_{ij} \sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma^2)$$

$$\boldsymbol{\beta}_j \sim N(\mathbf{z}_j\boldsymbol{\gamma}, \omega^2)$$

$$\boldsymbol{\gamma} \sim N(\mathbf{g}_0, \mathbf{G}_0)$$

where, say, $i = 1, \dots, n_j$, indexes survey respondents and $j = 1, \dots, m$ indexes geographic units or time.

- Estimation and inference for hierarchical models not trivial

- Switching to a Bayesian setting and using MCMC enables characterization of joint posterior of all model parameters, without any shortcuts/compromises (*a la* MLE).

- Estimation and inference for hierarchical models not trivial
- Nested stochastic structure complicates matters

- Switching to a Bayesian setting and using MCMC enables characterization of joint posterior of all model parameters, without any shortcuts/compromises (*a la* MLE).

- Estimation and inference for hierarchical models not trivial
- Nested stochastic structure complicates matters
- Issue for likelihood analyses is to ensure that uncertainty in 2nd level structure propagates into inferences about 1st level parameters
- Switching to a Bayesian setting and using MCMC enables characterization of joint posterior of all model parameters, without any shortcuts/compromises (*a la* MLE).

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

i.e., no cross-study heterogeneity

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

i.e., no cross-study heterogeneity

$$\begin{aligned}y_i &\sim N(\theta_i, V_i) \quad (\text{level 1}) \\ \theta_i &\stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \quad (\text{level 2}) \\ (\theta, \sigma^2) &\sim \pi(\theta, \sigma^2) \quad (\text{prior})\end{aligned}$$

i.e., no cross-study heterogeneity

$$\begin{aligned}y_i &\sim N(\theta_i, V_i) \quad (\text{level 1}) \\ \theta_i &\stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \quad (\text{level 2}) \\ (\theta, \sigma^2) &\sim \pi(\theta, \sigma^2) \quad (\text{prior})\end{aligned}$$

- The θ_i are sometimes called *random effects*

i.e., no cross-study heterogeneity

$$\begin{aligned}y_i &\sim N(\theta_i, V_i) \quad (\text{level 1}) \\ \theta_i &\stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \quad (\text{level 2}) \\ (\theta, \sigma^2) &\sim \pi(\theta, \sigma^2) \quad (\text{prior})\end{aligned}$$

- The θ_i are sometimes called *random effects*
- When $\sigma^2 = 0$ we get a *fixed effects* model:

i.e., no cross-study heterogeneity

$$\begin{aligned}y_i &\sim N(\theta_i, V_i) \quad (\text{level 1}) \\ \theta_i &\stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \quad (\text{level 2}) \\ (\theta, \sigma^2) &\sim \pi(\theta, \sigma^2) \quad (\text{prior})\end{aligned}$$

- The θ_i are sometimes called *random effects*
- When $\sigma^2 = 0$ we get a *fixed effects* model:

$$\theta_i = \theta \forall i = 1, \dots, n$$

i.e., no cross-study heterogeneity

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

(BLUPs -- best linear unbiased predictions).

Hierarchical Model

57

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

(BLUPs -- best linear unbiased predictions).

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

- Unknown parameters are θ (the mean of the distribution for the random effects, the θ_i) and σ^2 (the variance of this distribution).
- We don't estimate the θ_i directly; estimate the parameters of the distribution from which they come;
(BLUPs -- best linear unbiased predictions).

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

- Unknown parameters are θ (the mean of the distribution for the random effects, the θ_i) and σ^2 (the variance of this distribution).
- We don't estimate the θ_i directly; estimate the parameters of the distribution from which they come; then given these parameters and the observed data, we form expectations for the θ_i (BLUPs -- best linear unbiased predictions).

Hierarchical model

58

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

Hierarchical model

58

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

can be re-arranged to yield the estimates

$$\theta_i | y_i, \theta, \sigma^2 \sim N(\theta_i^*, V_i(1 - B_i))$$

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

can be re-arranged to yield the estimates

$$\theta_i | y_i, \theta, \sigma^2 \sim N(\theta_i^*, V_i(1 - B_i))$$

where

$$\theta_i^* = (1 - B_i)y_i + B_i\theta$$

$$B_i = V_i / (V_i + \sigma^2)$$

- Each study's conditional mean is a **weighted average** of the observed study mean and the overall mean. 59
- B_i are *shrinkage factors*.

- More study-specific precision, less shrinkage

- Each study's conditional mean is a **weighted average** of the observed study mean and the overall mean. 59
- B_i are *shrinkage factors*. The larger B_i , the more θ_i^* is shrunk back to the grand mean θ .
- More study-specific precision, less shrinkage

- Each study's conditional mean is a **weighted average** of the observed study mean and the overall mean. 59
- B_i are *shrinkage factors*. The larger B_i , the more θ_i^* is shrunk back to the grand mean θ .
- Less study-specific precision, more shrinkage
- More study-specific precision, less shrinkage

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}.$$

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}.$$

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

Likelihood function:

$$L(\theta, \sigma^2 | \mathbf{y}, \mathbf{V}) \propto \prod_{i=1}^n \frac{1}{\sqrt{V_i + \sigma^2}} \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \theta)^2}{V_i + \sigma^2} \right]$$

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}.$$

$$y_i \sim N(\theta_i, V_i) \text{ (level 1)}$$

$$\theta_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \text{ (level 2)}$$

Likelihood function:

$$L(\theta, \sigma^2 | \mathbf{y}, \mathbf{V}) \propto \prod_{i=1}^n \frac{1}{\sqrt{V_i + \sigma^2}} \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \theta)^2}{V_i + \sigma^2} \right]$$

After optimizing wrt θ and σ^2 :

$$\hat{\theta}_i = (1 - \hat{B}_i)y_i + \hat{B}_i\hat{\theta}, \hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}.$$

Asymptotic justification: as $n \rightarrow \infty$ (large number of units)

$$\hat{\theta} \sim N \left(\theta, \left[\sum_{i=1}^n \frac{1}{V_i + \hat{\sigma}^2} \right]^{-1} \right)$$
$$\hat{\theta}_i \sim N \left(\theta_i, V_i(1 - \hat{B}_i) \right),$$

- See **Browne and Draper (2000)**, who conclude that Bayesian simulation is probably the best way to estimate these models.

Asymptotic justification: as $n \rightarrow \infty$ (large number of units)

$$\hat{\theta} \sim N \left(\theta, \left[\sum_{i=1}^n \frac{1}{V_i + \hat{\sigma}^2} \right]^{-1} \right)$$
$$\hat{\theta}_i \sim N \left(\theta_i, V_i(1 - \hat{B}_i) \right),$$

- Draper notes that these expressions do not account fully for the uncertainty in σ^2 , and therefore underestimate the true sampling variances.
- See Browne and Draper (2000), who conclude that Bayesian simulation is probably the best way to estimate these models.

Example: meta-analysis of aspirin and heart attacks (Draper et al. 1992)

Data: 6 studies of effects of aspirin on survivorship after acute myocardial infarction

Study	Aspirin		Placebo	
	Patients	Mortality (%)	Patients	Mortality (%)
UK-1	615	7.97	624	10.74
CDPA	758	5.80	771	8.30
GAMS	317	8.52	309	10.36
UK-2	832	12.26	850	14.82
PARIS	810	10.49	406	12.81
AMIS	2267	10.85	2257	9.70
Total	5599	9.88	5217	10.73

n.b., the large AMIS study finds a negative effect, and tends to dominate any pooled analysis.

Example: meta-analysis of aspirin and heart attacks (Draper et al. 1992)

Data: 6 studies of effects of aspirin on survivorship after acute myocardial infarction (heart attack).

Study	Aspirin		Placebo	
	Patients	Mortality (%)	Patients	Mortality (%)
UK-1	615	7.97	624	10.74
CDPA	758	5.80	771	8.30
GAMS	317	8.52	309	10.36
UK-2	832	12.26	850	14.82
PARIS	810	10.49	406	12.81
AMIS	2267	10.85	2257	9.70
Total	5599	9.88	5217	10.73

n.b., the large AMIS study finds a negative effect, and tends to dominate any pooled analysis.

Example: meta-analysis of aspirin and heart attacks (Draper et al. 1992)

Data: 6 studies of effects of aspirin on survivorship after acute myocardial infarction (heart attack). Some patients given aspirin; some given a placebo.

Study	Aspirin		Placebo	
	Patients	Mortality (%)	Patients	Mortality (%)
UK-1	615	7.97	624	10.74
CDPA	758	5.80	771	8.30
GAMS	317	8.52	309	10.36
UK-2	832	12.26	850	14.82
PARIS	810	10.49	406	12.81
AMIS	2267	10.85	2257	9.70
Total	5599	9.88	5217	10.73

n.b., the large AMIS study finds a negative effect, and tends to dominate any pooled analysis.

Example: meta-analysis of aspirin and heart attacks

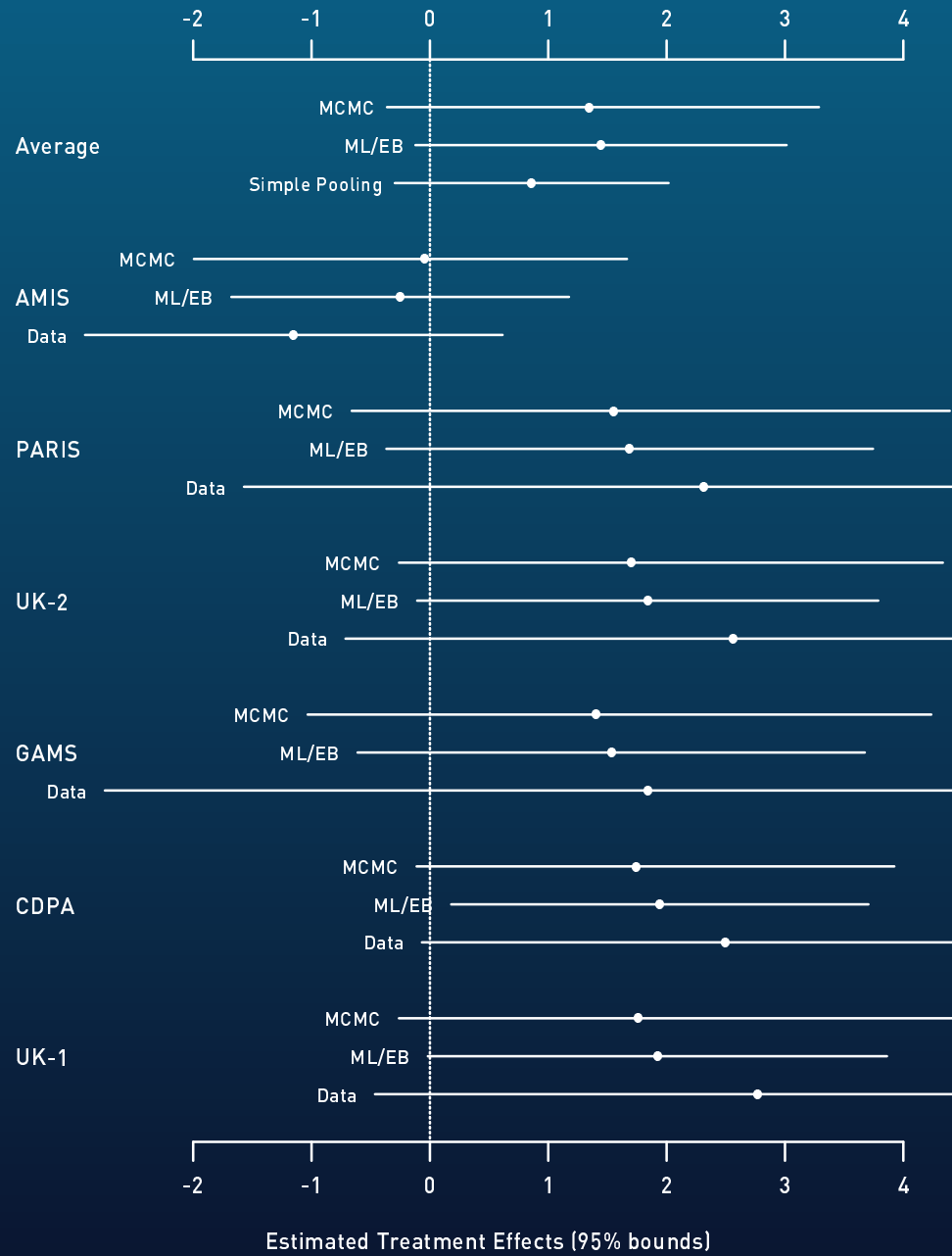
Data: each study yields a difference of means y_i and an accompanying standard error

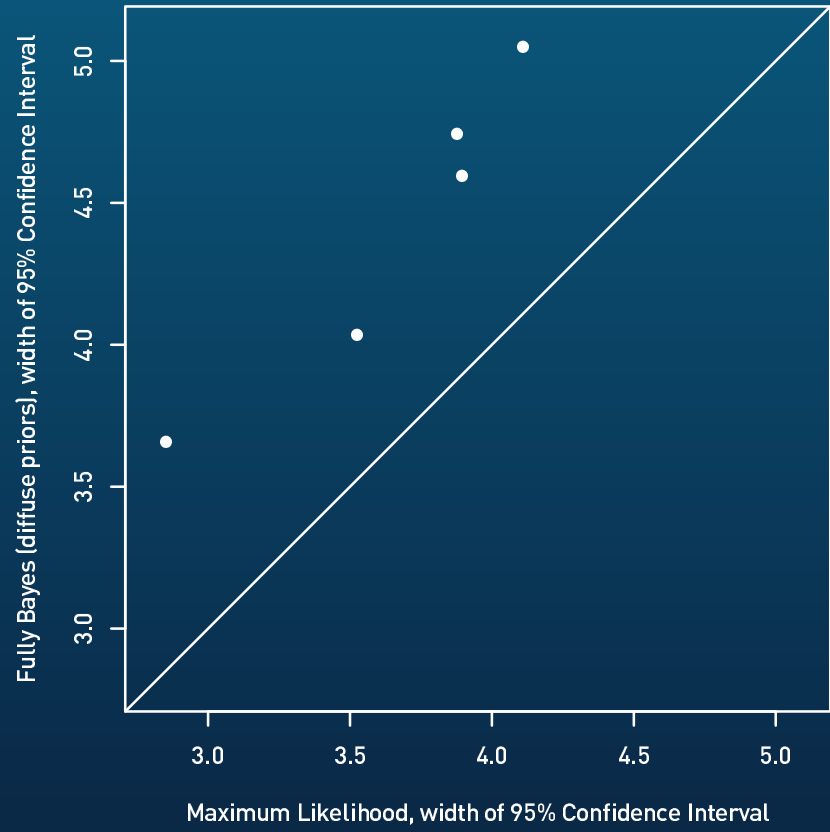
Study	y_i	se	Z_i	p_i
UK-1	2.77	1.65	1.68	.047
CDPA	2.50	1.31	1.91	.028
GAMS	1.84	2.34	0.79	.216
UK-2	2.56	1.67	1.54	.062
PARIS	2.31	1.98	1.17	.129
AMIS	-1.15	0.90	-1.27	.898
Total	0.86	0.59	1.47	.072

i.e., simple pooling finds a small effect, statistically indistinguishable from zero at conventional levels of significance.

Estimation and Inference for Hierarchical Model via Bayesian Simulation

- See WinBUGS job, aspirin.





Example: Hierarchical Model for Economic Growth

67

Corporatism example.

Extreme Missing Data

68

See my 1999 lecture notes, on my [MCMC pages](#), and the accompanying [Bimodal](#) WinBUGS ODC file on my web site.

References

Browne, William J. and David Draper. 2000. “Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models.” *Computational Statistics*. To appear.

Draper, D., Donald P. Gaver Jr, Prem K. Goel, Joel B. Greenhouse, Larry V. Hedges, Carl N. Morris, John R. Tucker and Christine M. Waternaux. 1992. *Combining Information: Statistical Issues and Opportunities for Research*. Number 1 in Contemporary Statistics. Alexandria, Virginia: American Statistical Association.

Jackman, Simon. 2000. “Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.” *American Journal of Political Science* 44:375--404.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. “Making the Most

of Statistical Analysis: Improving Interpretation and Presentation.”⁷⁰
American Journal of Political Science 44:341--355.

Metropolis, N. and S. Ulam. 1949. “The Monte Carlo Method.” *Journal of the American Statistical Association* 44:335--341.