THE UNIVERSITY OF IOWA

Introduction to Bayesian Psychometric Modeling

PSQF 7375; Spring 2019

Lecture 1

18 January, 2019

Today's Class

- An introduction to the course
  - Bayesian Overview
  - MCMC Overview
  - Psychometric Model Overview

AN INTRODUCTION TO BAYESIAN STATISTICS

Bayesian Statistics: The Basics

THE UNIVERSITY OF IOWA

- Bayesian statistical analysis refers to the use of models where some or all of the parameters are treated as **random components**
  - Each parameter comes from some type of distribution

- The likelihood function of the data is then augmented with an additional term that represents the likelihood of the **prior distribution** for each parameter
  - Think of this as saying each parameter has a certain likelihood – the height of the prior distribution

- The final estimates are then considered summaries of the **posterior distribution** of the parameter, conditional on the data
  - We use these estimates to make inferences, just like we do when using the non-Bayesian approaches (e.g., maximum likelihood/least squares)

Bayesian Statistics: Why It Is Used

THE UNIVERSITY OF IOWA

- Bayesian methods get used because the ***relative*** accessibility of one method of estimation (MCMC – to be discussed shortly)

- There are at least four main reasons why people use MCMC:

1. Missing data
   - Multiple imputation: MCMC is used to estimate model parameters then "impute" data
   - More complicated models for certain types of missing data

2. Lack of software capable of handling large sized analyses
   - Have a zero-inflated negative binomial with 21 multivariate outcomes per 18 time points?

3. New models/generalizations of models not available in software
   - Have a new model?
   - Need a certain link function not in software?

4. Membership in the "cult" of Bayes
   - Cult members believe philosophical differences exist between numbers from Bayesian analysis and other types of estimators

Bayesian Statistics: Perceptions and Issues

THE UNIVERSITY OF IOWA

- The use of Bayesian statistics has been controversial
  - The use of certain prior distributions can produce results that are biased or reflect subjective judgment rather than objective science
- Most MCMC estimation methods are **computationally intensive** with long compute times relative to other estimators
  - Until very recently, very few methods available for those who aren't into programming in Fortran, C, or C++
- Understanding of what Bayesian methods are and how they work is limited outside the field of mathematical statistics
  - Especially the case in the educational and social sciences
- Over the past 20 years, Bayesian methods have become widespread – making new models estimable and becoming standard in some social science fields (quantitative psychology and educational measurement)

As an Example…

THE UNIVERSITY OF IOWA

Statistical Modelling

These new articles for Statistical Modelling are available online

View online

OnlineFirst Alert

Article

Bayesian residual analysis for spatially correlated data

Viviana GR Lobo, Thaís CO Fonseca
Statistical Modelling
Jan 13, 2019 | OnlineFirst

A Bayesian approach of analysing semi-continuous longitudinal data with monotone missingness

Jayabrata Biswas, Kiranmoy Das
Statistical Modelling
Jan 10, 2019 | OnlineFirst

Bayesian semiparametric latent variable model with DP prior for joint analysis: Implementation with nimble

Zhihua Ma, Guanghui Chen
Statistical Modelling
Jan 8, 2019 | OnlineFirst

Exploring and modelling team performances of the Kaggle European Soccer database

Maurizio Carpita, Enrico Ciavolino, Paola Pasca
Statistical Modelling
Jan 10, 2019 | OnlineFirst

A PRIMER ON HOW BAYESIAN METHODS WORK

How Bayesian Statistics Work

THE UNIVERSITY OF IOWA

- The term Bayesian refers to Thomas Bayes (1701-1761)
  - Formulated Bayes' Theorem

- Bayesian methods rely on Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  - $P(A)$ is the **prior distribution (pdf) of A → WHY THINGS ARE BAYESIAN**
  - $P(B)$ is the **marginal distribution (pdf) of B**
  - $P(B|A)$ is the **conditional distribution (pdf) of B, given A**
  - $P(A|B)$ is the **posterior distribution (pdf) of A, given B**

- Bayes' Theorem Example…

Imagine a patient takes a test for a rare disease (present 1% of the population) that has a 95% accuracy rate…what is the probability the patient actually has the disease?

Bayes' Theorem Example

Imagine a patient takes a test for a rare disease (present 1% of the population) that has a 95% accuracy rate…what is the probability the patient actually has the disease?

- D = the case where the person actually has the disease
- ND = the case where the person does not have the disease
- + = the test for the disease is positive

The question is asking for: P(D|+)

From Bayes' Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)}$$

What we know:

$$P(D) = .01$$
$$P(+|D) = .95$$

Back to Distributions

- We don't know $P(+)$ directly from the problem, but we can figure it out if we recall how distributions work:

- $P(+)$ is a marginal distribution
- $P(+|D)$ is a conditional distribution

- We can get to the marginal by summing across the conditional:

$$P(+) = P(+|D)P(D) + P(+|ND)P(ND)$$
$$= .95 * .01 + .05 * .99 = .059$$

- So, to figure out the answer, if a person tests positive for the disease, the **posterior probability** they actually have the disease is:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{.01 * .99}{.059} = .17$$

A (Perhaps) More Relevant Example

THE UNIVERSITY OF IOWA

- The old-fashioned Bayes' Theorem example I've found to be difficult to generalize to your actual data, so…

- Imagine you administer an IQ test to a sample of 50 people
  - y_p = person p's IQ test score

- To put this into a linear-models context, the empty model for Y:

$$y_p = \beta_0 + e_p$$

where $e_p \sim N(0, \sigma_e^2)$

- From this empty model, we know that:
  - β_0 is the mean of the Y (the mean IQ)
  - σ_e^2 is the sample variance of Y
  - The conditional distribution of Y is:

$$f\left(y_p \middle| \beta_0, \sigma_e^2\right) \sim N(\beta_0, \sigma_e^2)$$

Non-Bayesian Analysis

THE UNIVERSITY OF IOWA

- Often, least squares (which is equivalent to REML) is used to estimate this model
  - We could also use ML…

- For ML, we maximized the joint likelihood of the sample with respect to the two unknown parameters β_0 and σ_e^2

$$L(\beta_0, \sigma_e^2) = \prod_{p=1}^{N} f\left(y_p | \beta_0, \sigma_e^2\right) = \prod_{p=1}^{N} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_p - \beta_0\right)^2}{2\sigma_e^2}\right)$$

- Here, using gls(), I found:

$$\beta_0 = 102.769$$
$$\sigma_e^2 = 239.490$$

- Also, I found:

$$LogL = -207.91$$

Setting up a Bayesian Model

THE UNIVERSITY OF IOWA

- The (fully) Bayesian model would treat each parameter as a random instance from some **prior distribution**

- Let's say you know that this version of the IQ test is supposed to have a mean of 100 and a standard deviation of 15
  - So β_0 should be 100 and σ_e^2 should be 225

- Going a step further, let's say you have seen results for administrations of this test that led you to believe that the mean came from a normal distribution with a SD of 2.13
  - This indicates the prior distribution for the **mean**…or

$$f(\beta_0) \sim N(100, 2.13^2)$$

- Let's also say that you don't really have an idea as for the distribution of the variance, but you have seen it range from 200 to 400, so we can come up with a prior distribution for the **variance** of:

$$f(\sigma_e^2) \sim U(200, 400)$$

- Here the prior is a uniform distribution meaning all values from 200 to 400 are equally likely

More on the Bayesian Approach

- The Bayesian approach is now to seek to find the **posterior distribution** of the parameters given the data:

$$f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- We can again use Bayes' Theorem (but for continuous parameters):

$$f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right) = \frac{f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0, \sigma_e^2)}{f\left(\boldsymbol{y}_p\right)} = \frac{f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2)}{f\left(\boldsymbol{y}_p\right)}$$

- Because $f\left(\boldsymbol{y}_p\right)$ essentially is a constant (which involves integrating across β_0 and σ_e^2 to find its value), this term is often referred to as:

$$f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right) \propto f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2)$$

- The symbol $\propto$ is read as "is proportional to" – meaning it is the same as when multiplied by a constant
  - So it is the same for all values of β_0 and σ_e^2

Unpacking the Posterior Distribution

- $f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right)$ is the **conditional distribution** of the data given the parameters – which comes from our linear model distribution

$$f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) = \prod_{p=1}^{N} f\left(y_p \middle| \beta_0, \sigma_e^2\right) = \prod_{p=1}^{N} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_p - \beta_0\right)^2}{2\sigma_e^2}\right)$$

- $f(\beta_0)$ is the **prior distribution** of β_0, which we decided would be $N(100, 2.13^2)$, giving the height of any β_0:

$$f(\beta_0) = \frac{1}{\sqrt{2\pi\sigma_{\beta_0}^2}} \exp\left(-\frac{\left(\beta_0 - \mu_{\beta_0}\right)^2}{2\sigma_{\beta_0}^2}\right)$$

$$= \frac{1}{\sqrt{2\pi * 2.13^2}} \exp\left(-\frac{(\beta_0 - 100)^2}{2 * 2.13^2}\right)$$

Unpacking the Posterior Distribution

THE UNIVERSITY OF IOWA

- $f(\sigma_e^2)$ is the **prior distribution** of σ_e^2, which we decided would be $\mathrm{U}(200,400)$, giving the density function result of any value of σ_e^2 as:

$$f(\sigma_e^2) = \frac{1}{b_{\sigma_e^2} - a_{\sigma_e^2}} = \frac{1}{400 - 200} = \frac{1}{200} = .005$$

- Some useful terminology:
  - The parameters of the model (for the data) get prior distributions
  - The prior distributions each have parameters – these parameters are called **hyper-parameters**
  - The hyper-parameters are not estimated in our example, but could be – giving us a case where we would call our priors **empirical priors**
    - **AKA random intercept variance**

Up Next: Estimation

THE UNIVERSITY OF IOWA

- Although MCMC is commonly thought of as the only method for Bayesian estimation, there are several other forms

- The form analogous to ML (where the value of the parameters that maximize the likelihood or log-likelihood) is called **Maximum (or Modal) a Posteriori estimation (MAP)**
  - The term modal comes from the maximum point coming at the peak (the mode) of the posterior distribution

- In practice, this functions similar to ML, instead of maximizing the joint likelihood of the data, we now have to worry about the prior:

$$f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right) = \frac{f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2)}{f\left(\boldsymbol{y}_p\right)} \propto f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2)$$

- Because it is often more easy to work with, the log is often used:

$$\log\left(f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right)\right) \propto \log f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) + \log f(\beta_0) + \log f(\sigma_e^2)$$

Grid Searching for the MAP Estimate of $\boldsymbol{\beta_0}$

THE UNIVERSITY OF IOWA

- To demonstrate, let's imagine we know $\sigma_e^2 = 239.490$
  - Later we won't know this…when we use MCMC

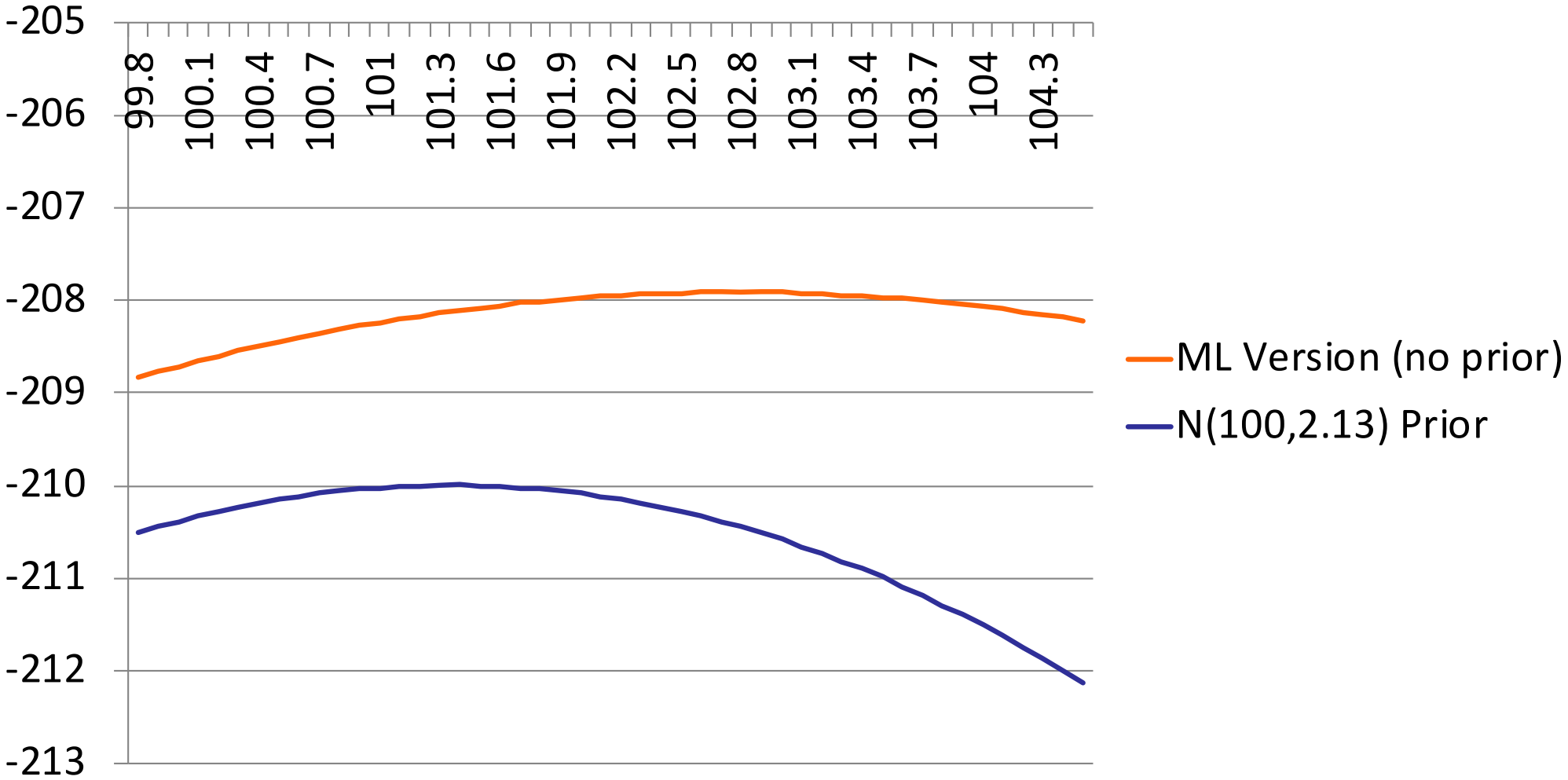
- We will use Excel to search over a grid of possible values for β_0

- In each, we will use $\log f\left(\boldsymbol{y}_p \middle| \beta_0\right) + \log f\left(\beta_0\right)$

- As a comparison, we will also search over the ML log likelihood function $\log f\left(\boldsymbol{y}_p \middle| \beta_0\right)$

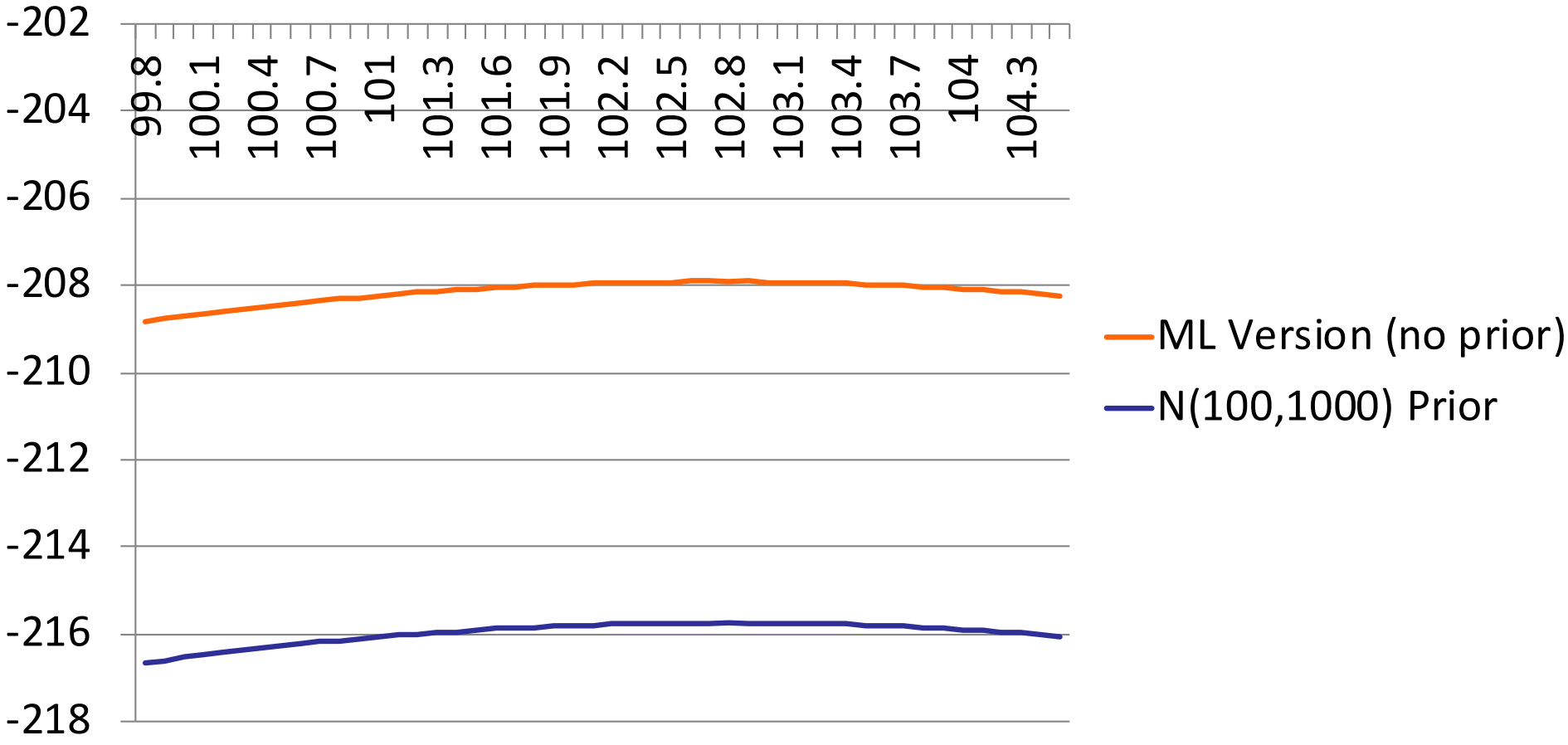
ML v. Prior for β_0 of N(100, 2.13²)

- Maximum for ML: 102.8
- Maximum for Bayes: 101.4
  (estimate is closer to mean of prior)



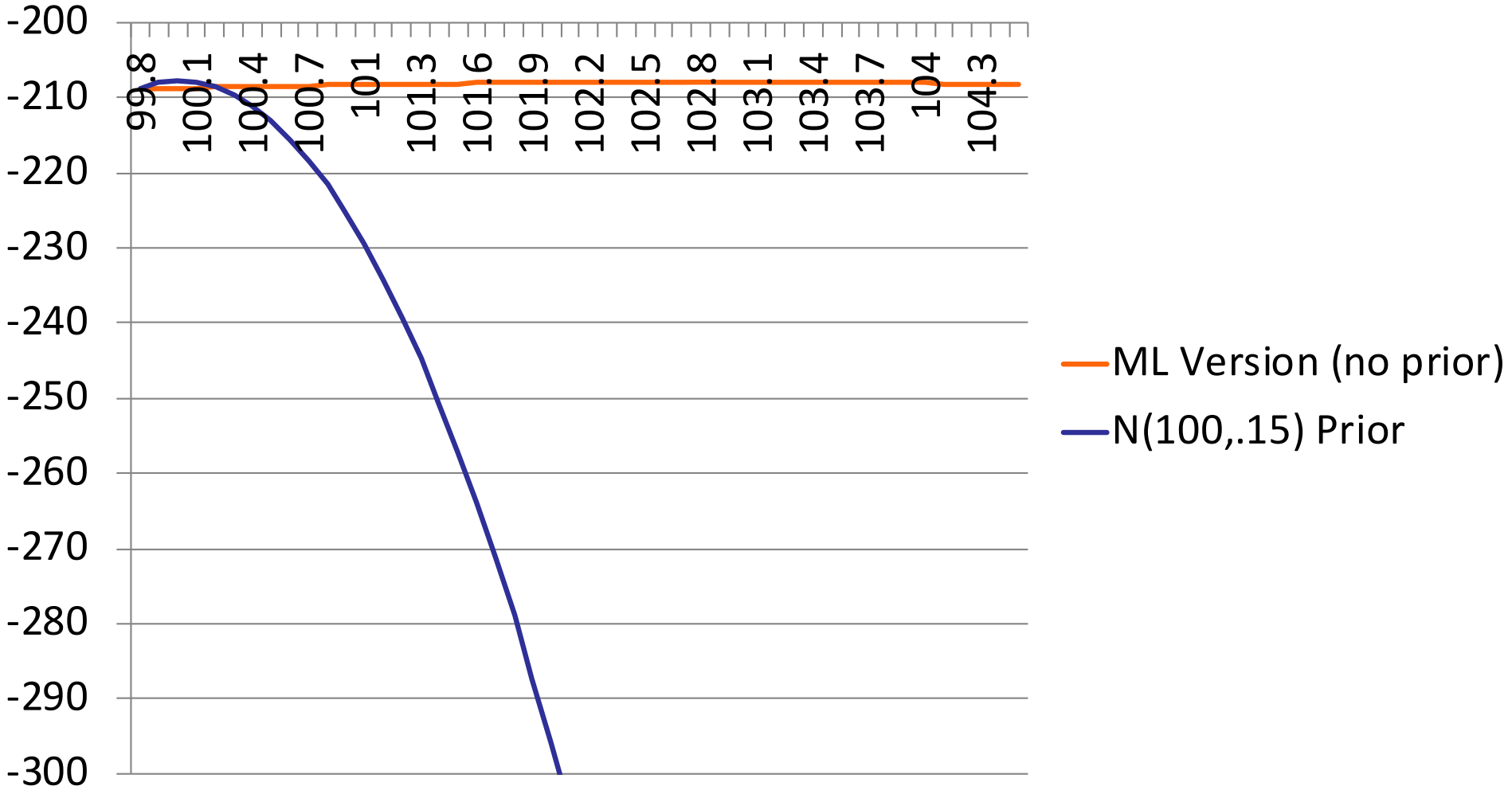
ML vs. Prior for β_0 of N(100, 1000²)

- Maximum for ML: 102.8
- Maximum for Bayes: 102.8



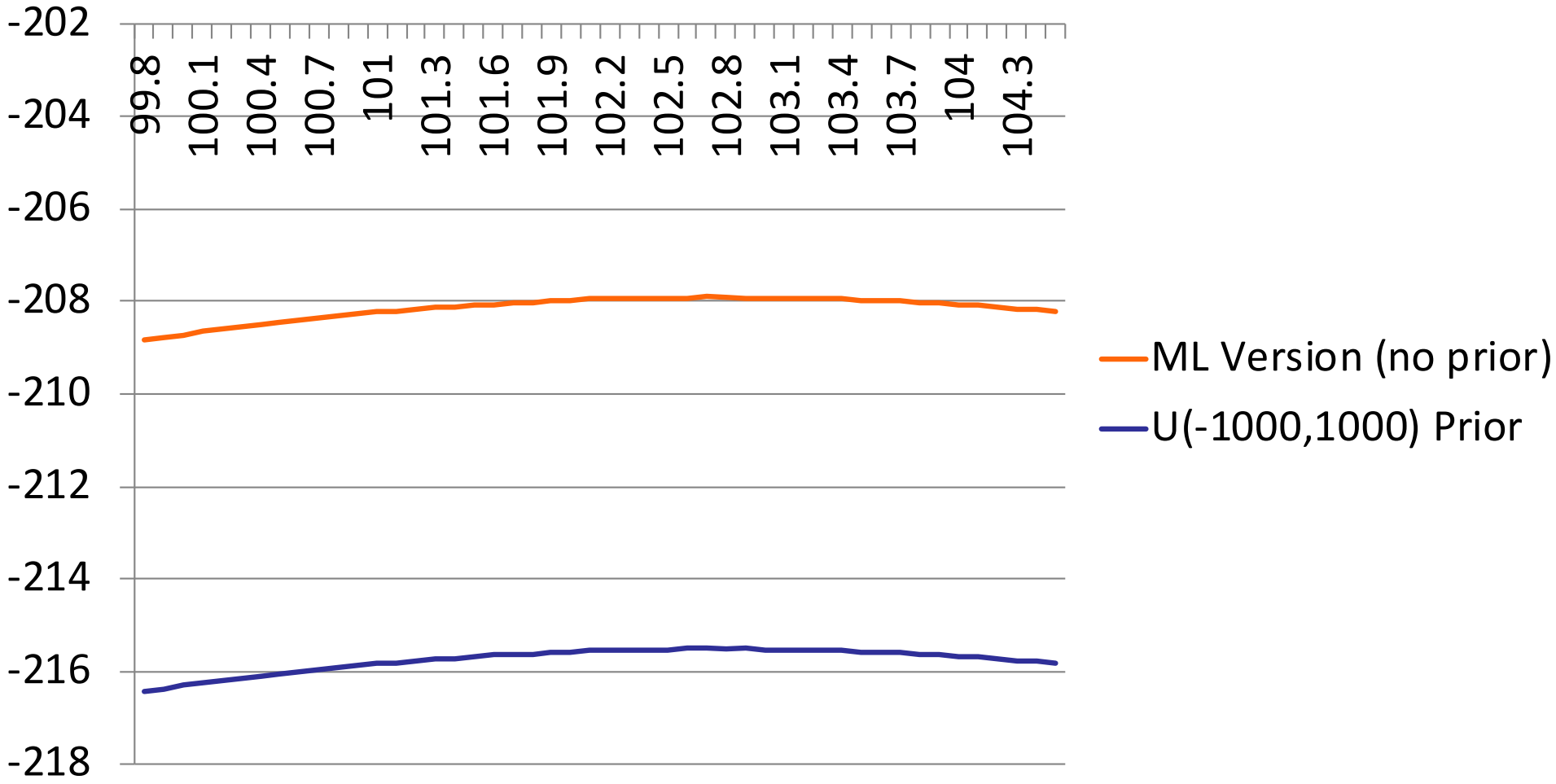
ML vs. Prior for β_0 of N(100, 0.15²)

- Maximum for ML: 102.8
- Maximum for Bayes: 100



ML vs. Prior for β_0 of U(-1000,1000)

- Maximum for ML: 102.8
- Maximum for Bayes: 102.8



Summarizing Bayesian So Far

THE UNIVERSITY OF IOWA

- Bayesian → parameters have prior distributions

- Estimation in Bayesian → MAP estimation is much like estimation in ML, only instead of likelihood of data, now have to add in likelihood for prior of all parameters
  - But…MAP estimation may be difficult as figuring out derivatives for gradient function (for Newton-Raphson) are not always easy
  - Where they are easy: **Conjugate** priors → prior distributions that are the same as the posterior distribution (think multilevel with normal outcomes)

- Priors can be **informative** (highly peaked) or **uninformative** (not peaked)
  - Some uninformative priors will give MAP estimates that are equal to ML

- Up next: estimation by brute force: Markov Chain Monte Carlo

MARKOV CHAIN MONTE CARLO ESTIMATION: THE BASICS

How Estimation Works (More or Less)

THE UNIVERSITY OF IOWA

- Most estimation routines do one of three things:

1. **Minimize Something:** Typically found with names that have "least" in the title. Forms of least squares include "Generalized", "Ordinary", "Weighted", "Diagonally Weighted", "WLSMV", and "Iteratively Reweighted." Typically the estimator of last resort…

2. **Maximize Something:** Typically found with names that have "maximum" in the title. Forms include "Maximum likelihood", "ML", "Residual Maximum Likelihood" (REML), "Robust ML". Historically the gold standard of estimators.

3. **Use Simulation to Sample from Something:** more recent advances in simulation use resampling techniques. Names include "Bayesian Markov Chain Monte Carlo", "Gibbs Sampling", "Metropolis Hastings", "Metropolis Algorithm", and "Monte Carlo". Used for complex models where ML is not available or for methods where prior values are needed.

How MCMC Estimation Works

- MCMC estimation works by taking samples from the posterior distribution of the data given the parameters:

$$f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right) = \frac{f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2)}{f\left(\boldsymbol{y}_p\right)}$$

  - How is that possible? We don't know $f(\boldsymbol{y}_p)$…but…we'll see…

- After enough values are drawn, a rough shape of the distribution can be formed
  - From that shape we can take summaries and make them our parameters (i.e., mean)

- How the sampling mechanism happens comes from several different algorithms that you will hear about, the most popular being:
  - **Gibbs Sampling**: used when $f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right)$ is known
    - Parameter values are drawn and kept throughout the chain
  - **Metropolis-Hastings (within Gibbs):** used when $f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right)$ is unknown
    - Parameter values are proposed, then either kept or rejected
    - SAS PROC MCMC uses the latter
    - TRIVIA NOTE: The Metropolis algorithm comes from Chemistry (in 1950)
  - **Hybrid MC:** Newer versions (1980s; implemented in Stan)

- In some fields (Physics in particular), MCMC estimation is referred to as Monte Carlo estimation

MCMC Estimation with MHG

- The Metropolis-Hastings algorithm works a bit differently than Gibbs sampling:

1. Each parameter (here β_0 and σ_e^2) is given an initial value

2. In order, a new value is proposed for each model parameter from some distribution:

$$\beta_0^* \sim Q(\beta_0^* | \beta_0); \sigma_e^{2^*} \sim Q\left(\sigma_e^{2^*} \middle| \sigma_e^2\right)$$

3. The proposed value is then accepted as the current value with probability $\max(r_{MHG}, 1)$:

$$r_{MHG} = \frac{f\left(\boldsymbol{y}_p \middle| \beta_0^*, \sigma_e^{2^*}\right) f(\beta_0^*) f\left(\sigma_e^{2^*}\right) Q\left(\beta_0 \middle| \beta_0^*\right) Q\left(\sigma_e^2 \middle| \sigma_e^{2^*}\right)}{f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2) Q(\beta_0^* | \beta_0) Q\left(\sigma_e^{2^*} \middle| \sigma_e^2\right)}$$

4. The process continues for a pre-specified number of iterations (more is better)

Notes About MHG

- The constant in the denominator of the posterior distribution:

$$f\left(\beta_0, \sigma_e^2 \middle| \boldsymbol{y}_p\right) = \frac{f\left(\boldsymbol{y}_p \middle| \beta_0, \sigma_e^2\right) f(\beta_0) f(\sigma_e^2)}{f\left(\boldsymbol{y}_p\right)}$$

…cancels when the ratio is formed

- The proposal distributions $Q(\beta_0^* | \beta_0)$ and $Q\left(\sigma_e^{2^*} \middle| \sigma_e^2\right)$ can literally be any statistical distribution
  - The trick is picking ones that make the chain "converge" quickly
  - Want to find values that lead to moderate number of accepted parameters
  - SAS PROC MCMC/JAGS don't make you pick these

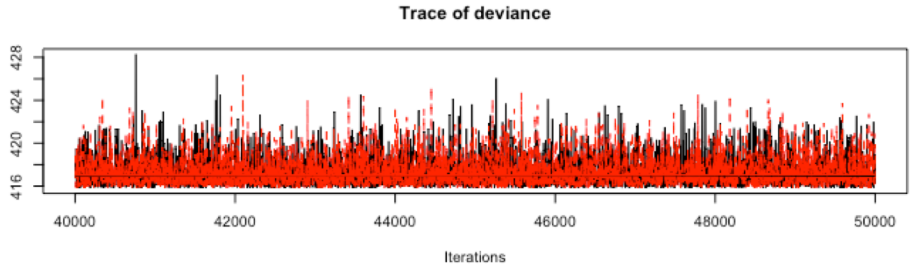
- Given a long enough chain, the final values of the chain will come from the posterior distribution
  - From that you can get your parameter estimates

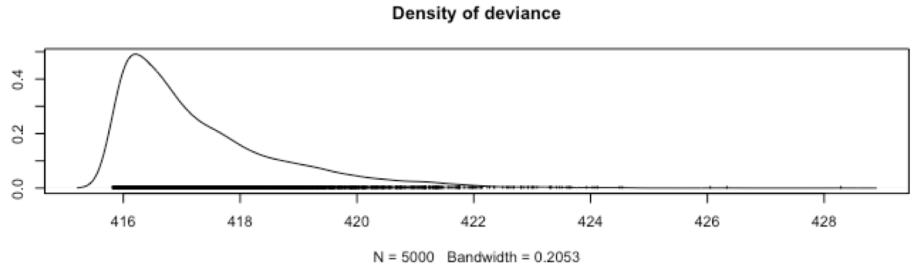
Introducing JAGS…

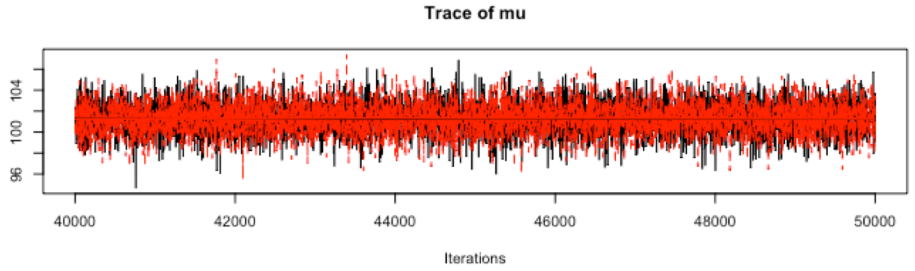
THE UNIVERSITY OF IOWA

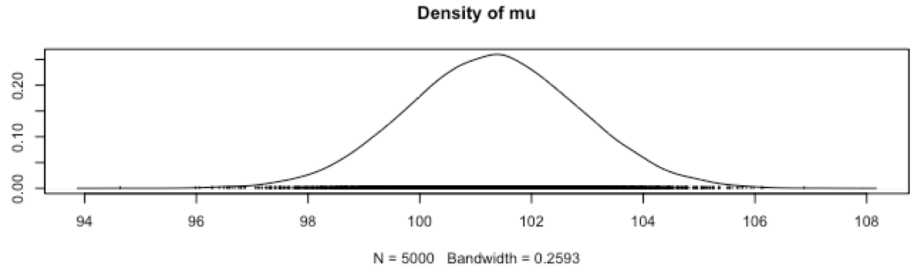
```
# estimation with Bayesian

model01Bayes = function(){

  # likelihood
  for (i in 1:n){
    y[i] ~ dnorm(mu, tau)
  }

  #priors
  mu ~ dnorm(100, 1/2.13^2)
  tau ~ dunif(1/400, 1/200)
  sigma2 = 1/tau
}

data = list(y = dataIQ$y, n = nrow(dataIQ))

jags.param = c("mu", "tau", "sigma2")

fit <- jags.parallel(data=data,
                     parameters.to.save=jags.param,
                     n.iter=50000, n.chains=2,n.thin=2,n.burnin=40000,
                     model.file=model01Bayes)
```

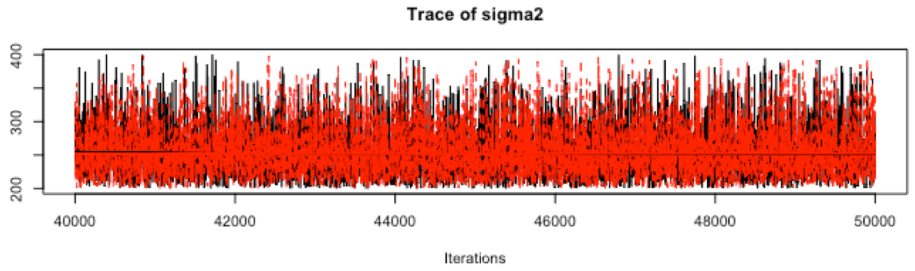
Examining the Chain and Posteriors

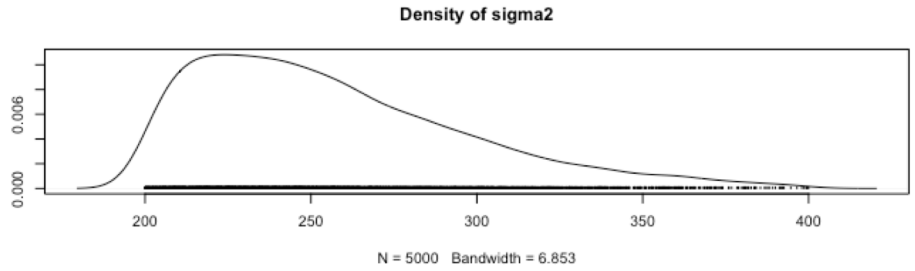


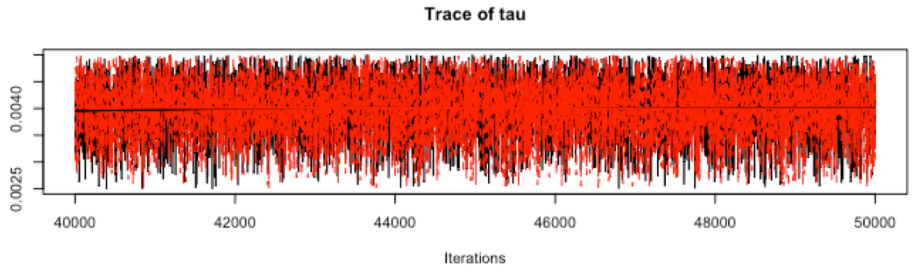


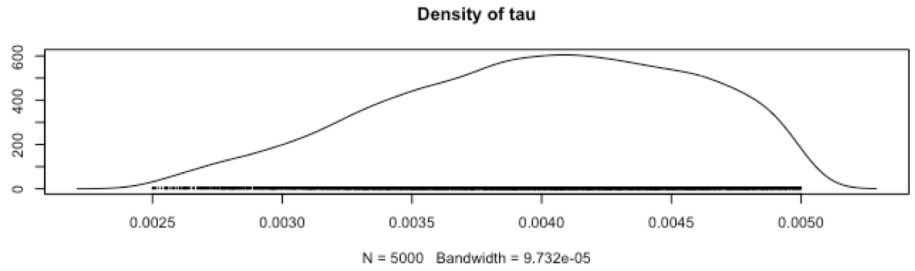












Practical Specifics in MCMC Estimation

THE UNIVERSITY OF IOWA

- A **burn-in** period is used where a chain is run for a set number of iterations before the sampled parameter values are used in the posterior distribution

- Because of the rejection/acceptance process, any two iterations are likely to have a high correlation (called **autocorrelation**) → posterior chains use a **thinning interval** to take every Xth sample to reduce the autocorrelation
  - A high autocorrelation may indicate the standard error of the posterior distribution will be smaller than it should be

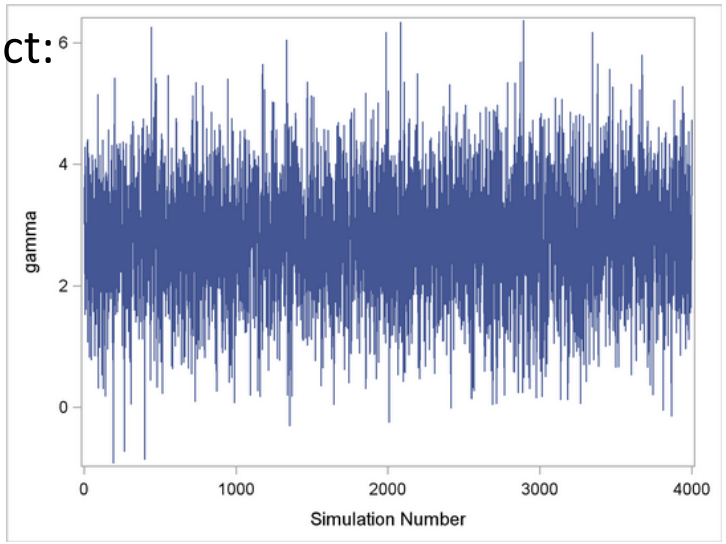
- The **chain length** (and sometimes number of chains) must also be long enough so the rejection/acceptance process can reasonably approximate the posterior distribution

- How does one what values to pick for these? Output diagnostics
  - Trial. And. Error.

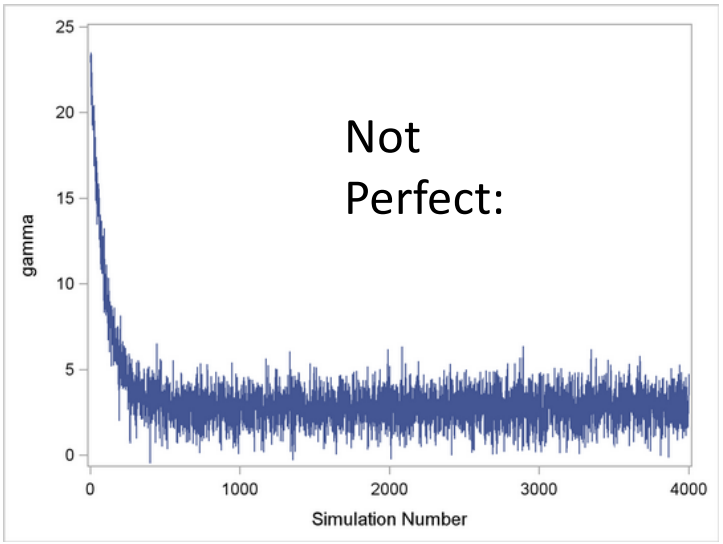
Best Output Diagnostics: the Eye Ball Test

THE UNIVERSITY OF IOWA

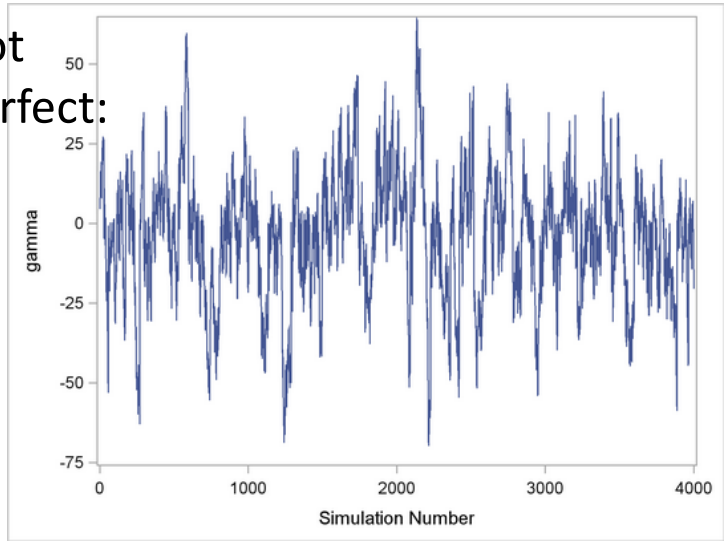
Perfect:



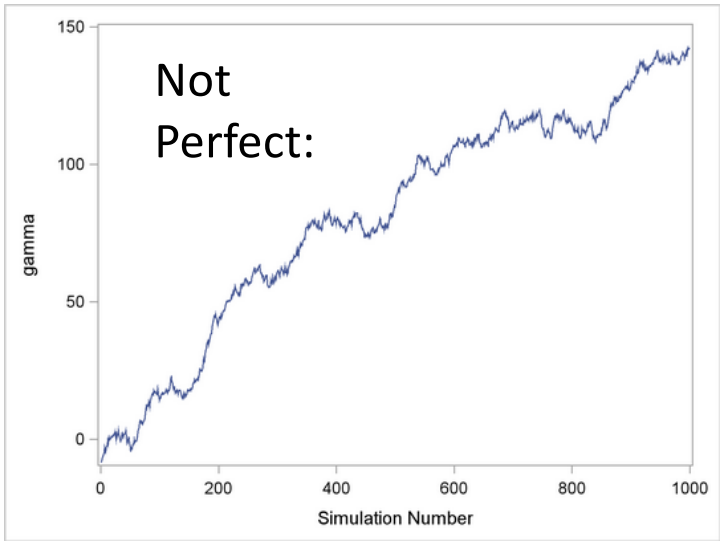
Not Perfect:



Not Perfect:



Not Perfect:



Output Statistics and Diagnostics

THE UNIVERSITY OF IOWA

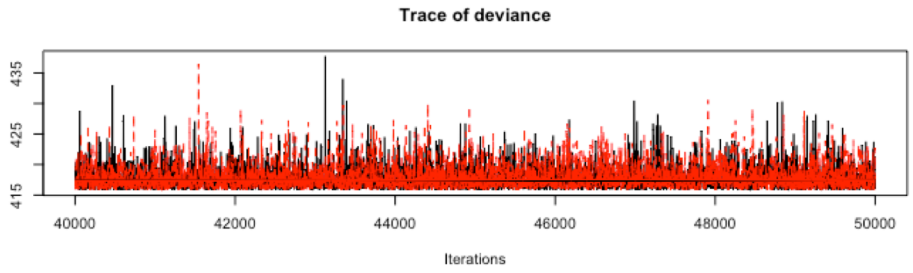
```
> fit
Inference for Bugs model at "model01Bayes", fit using jags,
 2 chains, each with 50000 iterations (first 40000 discarded), n.thin = 2
 n.sims = 10000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
mu        101.312   1.546  98.280 100.279 101.316 102.347 104.349 1.001 10000
sigma2    256.627  40.791 202.918 224.724 248.240 280.247 358.019 1.001 10000
tau         0.004   0.001   0.003   0.004   0.004   0.004   0.005 1.001 10000
deviance  417.317   1.404 415.862 416.284 416.869 417.922 421.059 1.001 10000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 1.0 and DIC = 418.3
DIC is an estimate of expected predictive error (lower deviance is better).
```

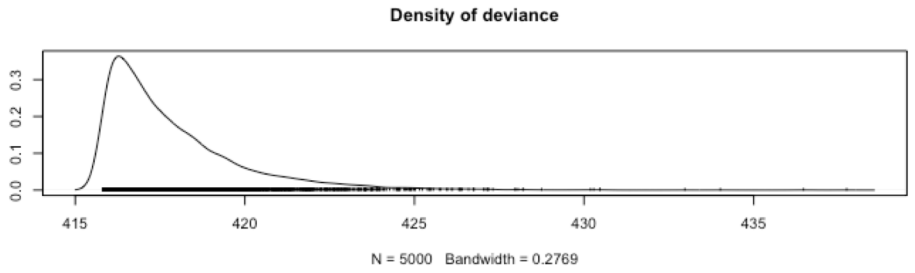
Changing the Prior

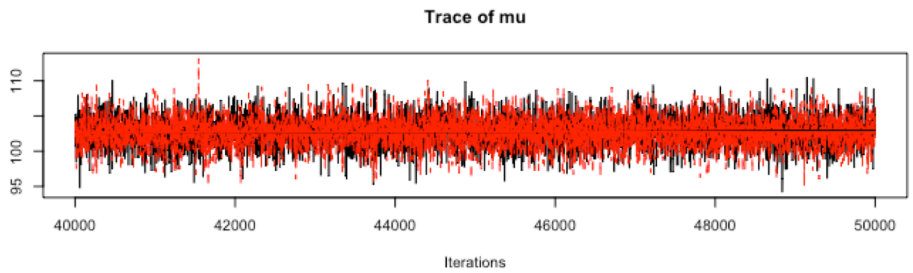
- To demonstrate different priors affect the analysis, we will now try a few prior distributions for our parameters
- Prior: $\beta_0 \sim U(-10000,10000); \sigma_e^2 \sim U(0,5000)$

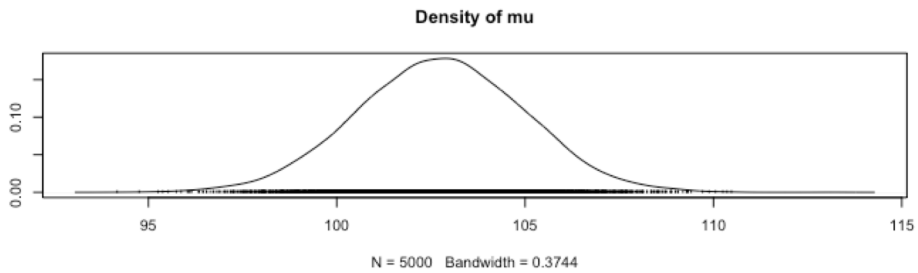
```
> fit2
Inference for Bugs model at "model02Bayes", fit using jags,
 2 chains, each with 50000 iterations (first 40000 discarded), n.thin = 2
 n.sims = 10000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
mu        102.750   2.229  98.385 101.239 102.758 104.251 107.100 1.001 10000
sigma2    244.899  51.326 164.789 208.259 238.353 273.458 362.843 1.001 10000
tau         0.004   0.001   0.003   0.004   0.004   0.005   0.006 1.001 10000
deviance  417.856   2.028 415.869 416.415 417.241 418.624 423.285 1.001  3200

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 2.1 and DIC = 419.9
DIC is an estimate of expected predictive error (lower deviance is better).
```

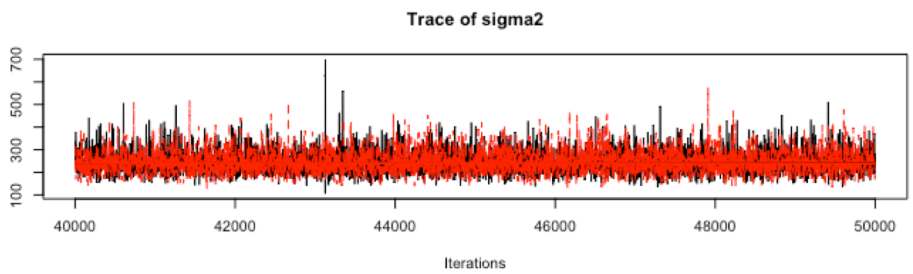
Chain Plots

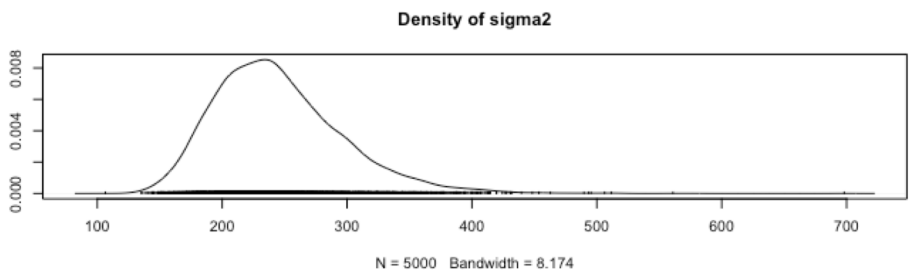

Trace of deviance
435
425
415
40000
42000
44000
46000
48000
50000
Iterations

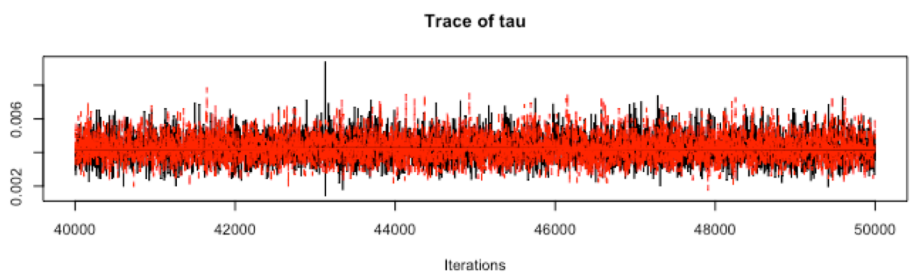

Density of deviance
0.3
0.2
0.1
0.0
415
420
425
430
435
N = 5000 Bandwidth = 0.2769

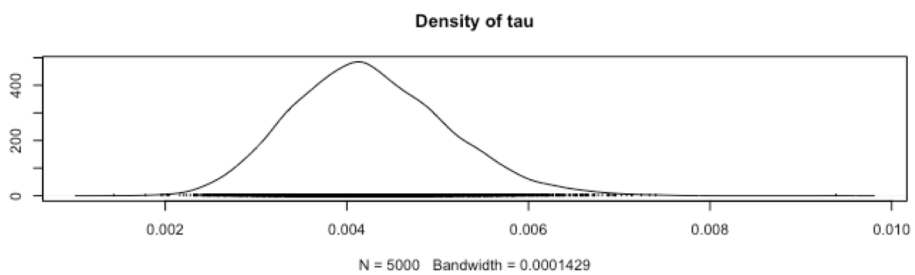

Trace of mu
110
100
95
40000
42000
44000
46000
48000
50000
Iterations


Density of mu
0.10
0.00
95
100
105
110
115
N = 5000 Bandwidth = 0.3744


Trace of sigma2
700
500
300
100
40000
42000
44000
46000
48000
50000
Iterations


Density of sigma2
0.008
0.004
0.000
100
200
300
400
500
600
700
N = 5000 Bandwidth = 8.174


Trace of tau
0.006
0.002
40000
42000
44000
46000
48000
50000
Iterations


Density of tau
400
200
0
0.002
0.004
0.006
0.008
0.010
N = 5000 Bandwidth = 0.0001429

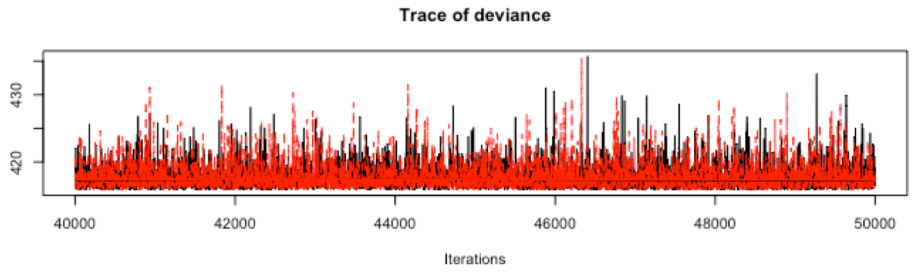
Changing the Prior

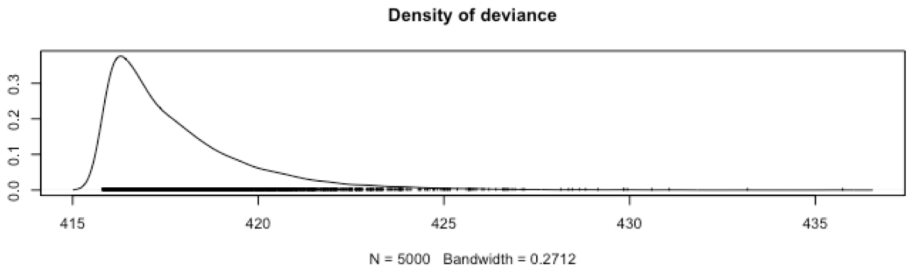
THE UNIVERSITY OF IOWA

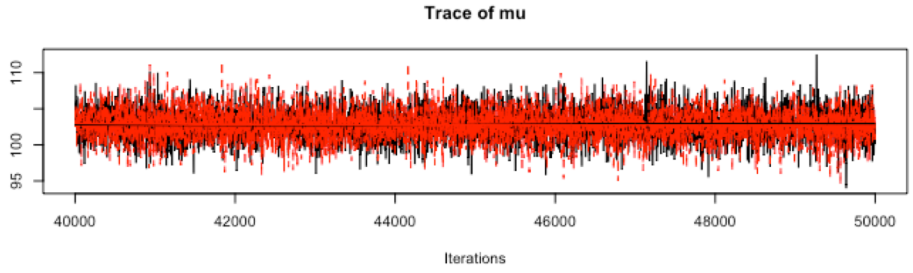
- Prior: $\beta_0 \sim N(0,100{,}000)$;
- $\sigma_e^{-2} \sim gamma(r = .01, \lambda = .01)$

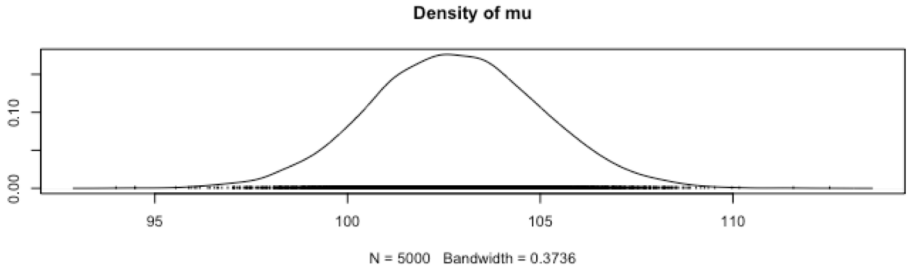
```
> fit3
Inference for Bugs model at "model03Bayes", fit using jags,
 2 chains, each with 50000 iterations (first 40000 discarded), n.thin = 2
 n.sims = 10000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
mu        102.784   2.224  98.426 101.274 102.758 104.254 107.175 1.001  7300
sigma2    253.996  52.970 171.522 216.053 247.279 284.606 375.192 1.001  9500
tau         0.004   0.001   0.003   0.004   0.004   0.005   0.006 1.001  9500
deviance  417.812   1.994 415.872 416.409 417.203 418.572 423.105 1.003  1700

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 2.0 and DIC = 419.8
DIC is an estimate of expected predictive error (lower deviance is better).
```

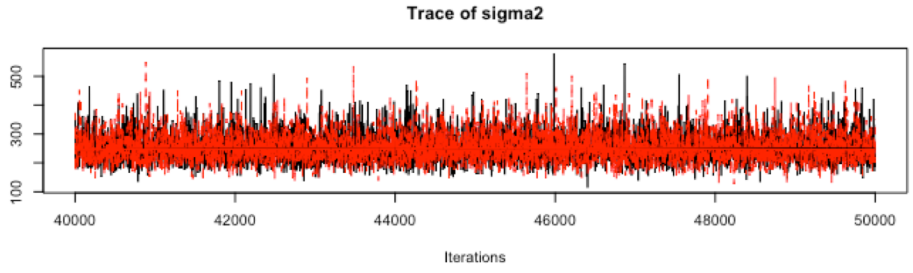
Chain Plots

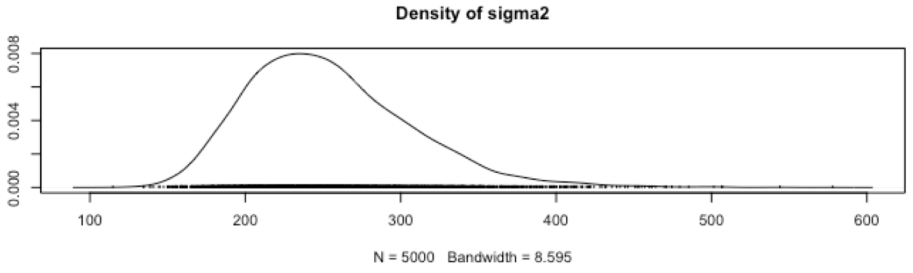

Trace of deviance
420
430
40000
42000
44000
46000
48000
50000
Iterations

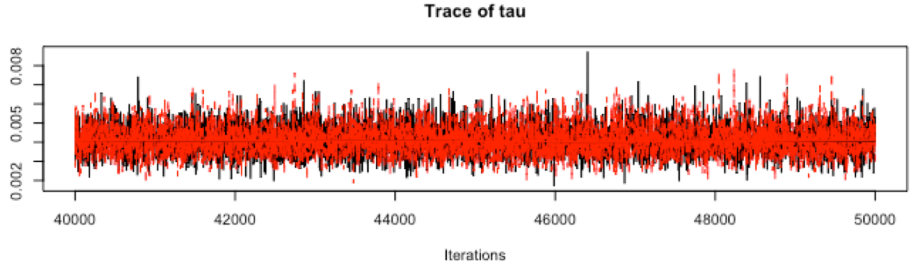

Density of deviance
0.0
0.1
0.2
0.3
415
420
425
430
435
N = 5000 Bandwidth = 0.2712

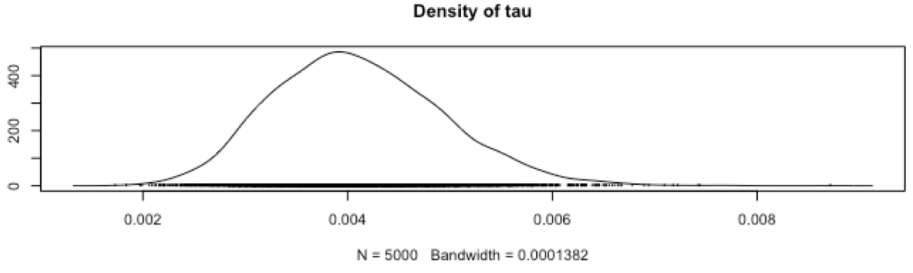

Trace of mu
95
100
110
40000
42000
44000
46000
48000
50000
Iterations


Density of mu
0.00
0.10
95
100
105
110
N = 5000 Bandwidth = 0.3736


Trace of sigma2
100
300
500
40000
42000
44000
46000
48000
50000
Iterations


Density of sigma2
0.000
0.004
0.008
100
200
300
400
500
600
N = 5000 Bandwidth = 8.595


Trace of tau
0.002
0.005
0.008
40000
42000
44000
46000
48000
50000
Iterations


Density of tau
0
200
400
0.002
0.004
0.006
0.008
N = 5000 Bandwidth = 0.0001382

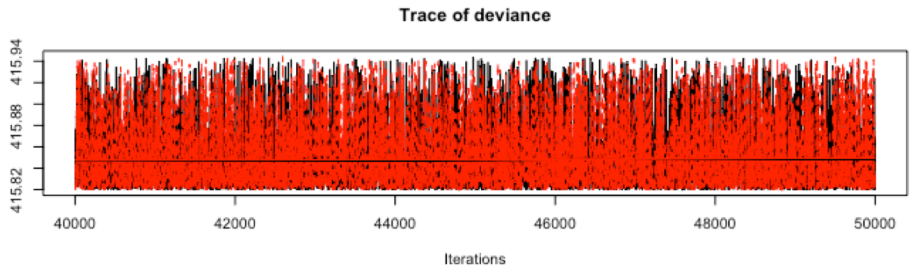
What About an Informative Prior?

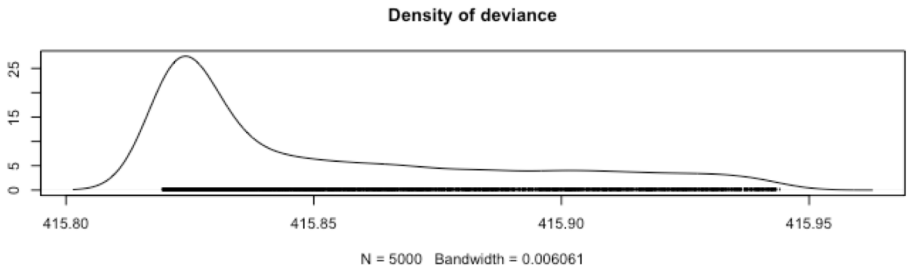
THE UNIVERSITY OF IOWA

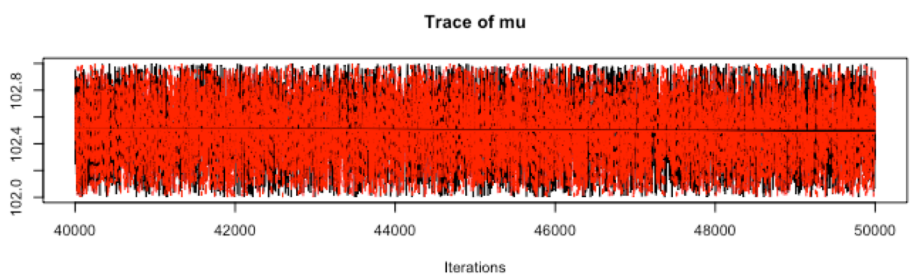
- Prior: $\beta_0 \sim U(102,103); \sigma_e^2 \sim U(238,242)$

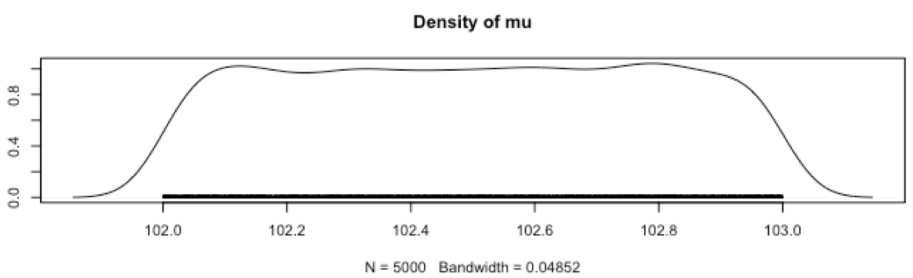
```
> fit4
Inference for Bugs model at "model04Bayes", fit using jags,
 2 chains, each with 50000 iterations (first 40000 discarded), n.thin = 2
 n.sims = 10000 iterations saved
         mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
mu       102.500   0.289 102.026 102.250 102.502 102.752 102.975 1.001  8200
sigma2   239.992   1.155 238.104 238.979 240.011 240.993 241.890 1.001 10000
tau        0.004   0.000   0.004   0.004   0.004   0.004   0.004 1.001 10000
deviance 415.853   0.036 415.820 415.823 415.835 415.876 415.935 1.000     1

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 0.0 and DIC = 415.9
DIC is an estimate of expected predictive error (lower deviance is better).
```

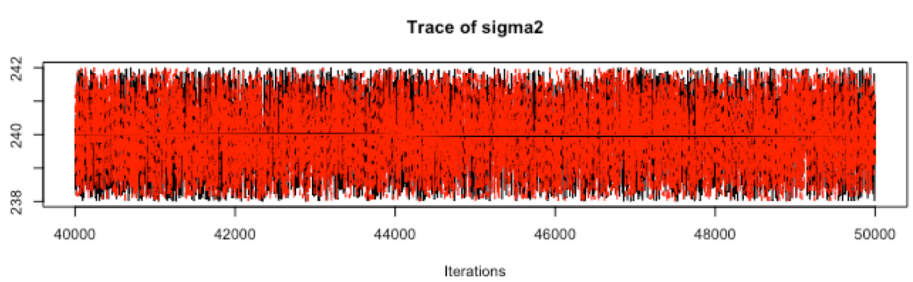
Chain Plots

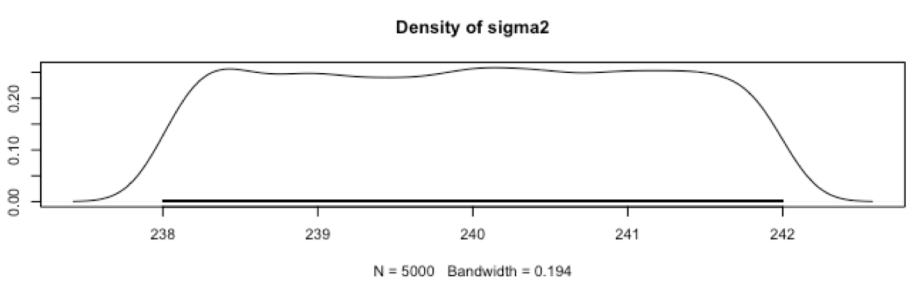

Trace of deviance
415.82
415.88
415.94
40000
42000
44000
46000
48000
50000
Iterations

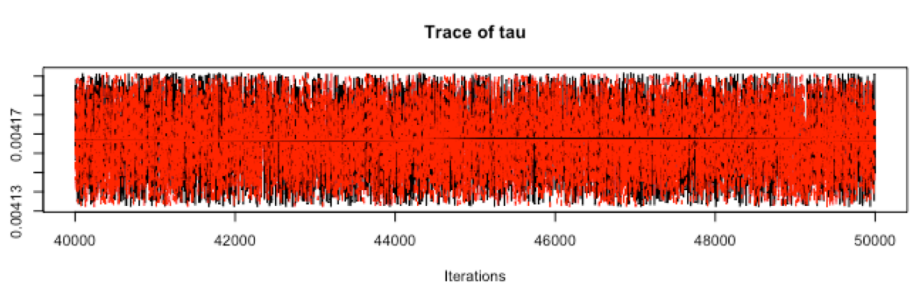

Density of deviance
0
5
15
25
415.80
415.85
415.90
415.95
N = 5000 Bandwidth = 0.006061

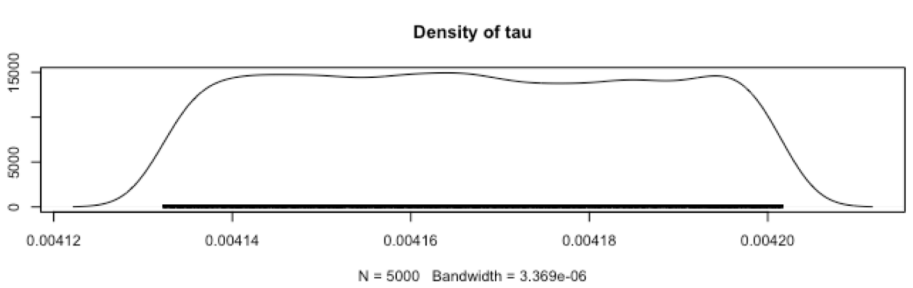

Trace of mu
102.0
102.4
102.8
40000
42000
44000
46000
48000
50000
Iterations


Density of mu
0.0
0.4
0.8
102.0
102.2
102.4
102.6
102.8
103.0
N = 5000 Bandwidth = 0.04852


Trace of sigma2
238
240
242
40000
42000
44000
46000
48000
50000
Iterations


Density of sigma2
0.00
0.10
0.20
238
239
240
241
242
N = 5000 Bandwidth = 0.194


Trace of tau
0.00413
0.00417
40000
42000
44000
46000
48000
50000
Iterations


Density of tau
0
5000
15000
0.00412
0.00414
0.00416
0.00418
0.00420
N = 5000 Bandwidth = 3.369e-06

MCMC in R

THE UNIVERSITY OF IOWA

- R itself does not have an MCMC engine native to the language – but there are many free versions available outside of R

- For instance, if you wanted to estimate a path model with MCMC you can:
  - Install the blavaan package (Bayesian lavaan)
  - Run the path analysis with MCMC

PSYCHOMETRIC MODELS OVERVIEW

Measurement Models

THE UNIVERSITY OF IOWA

- Measurement models can be divided into two families of models based on **response format alone:**
  - Continuous responses - **Confirmatory Factor Models**
  - Categorical responses - **Item Response Models**

- Both of these families fall under a larger framework: **Generalized Linear Latent and Mixed Models**
  - Provide measurement models for other types of responses

- Other relevant families:
  - **Structural Equation Models** - provides estimates of correlations amongst latent variables in measurement models
  - **Path Analysis** - simultaneous regression amongst multiple observed variables

Confirmatory Factor Analysis Models

THE UNIVERSITY OF IOWA

- Main idea of CFA: Build a measurement model for response variables that measure the same trait
  - **CFA = Linear regression model** predicting each continuous observed outcome variable (item, subscale) from a latent trait predictor variable(s)

$$Y_{si} = \mu_i + \lambda_{i1} F_{s1} + e_{si}$$

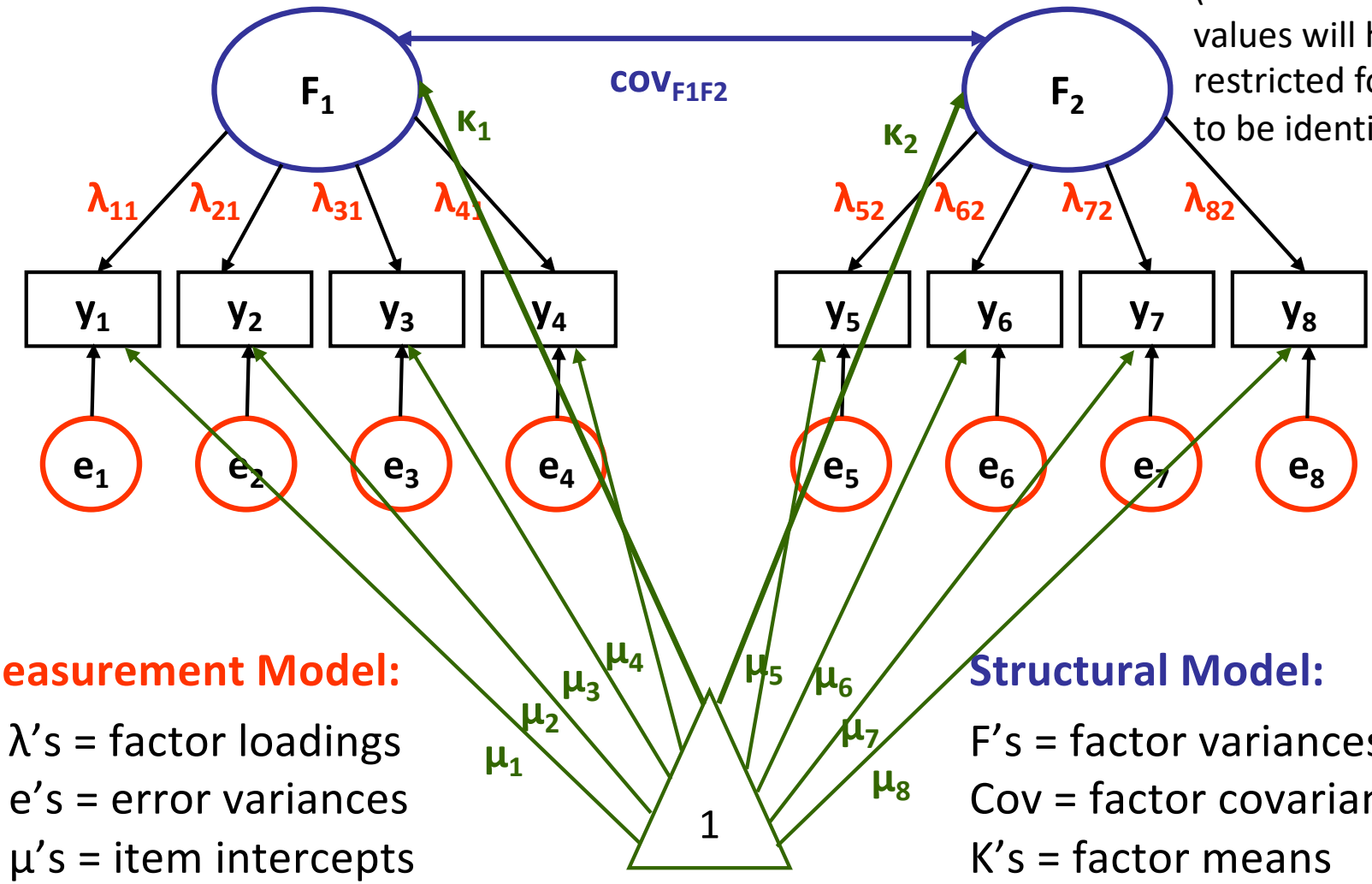
- i - item; s - subject; μ_i is the item intercept; λ_{i1} is the item slope (factor loading for factor 1); e_{is} is the error for the item and subject; Y_{is} is the item response (***assumed continuous***)

- Differs from exploratory factor analysis:
  - Number and content of factors is decided a priori
  - Alternative models are comparable and testable

- Uses of confirmatory factor analysis models:
  - Analyze relationships among subscales that have normal, continuous distributions
  - Provide comparability across persons, items, and occasions

CFA Model WITH Factor Means and Item Intercepts

THE UNIVERSITY OF IOWA



(But some of these values will have to be restricted for the model to be identified.)

Measurement Model:

λ's = factor loadings
e's = error variances
μ's = item intercepts

Structural Model:

F's = factor variances
Cov = factor covariances
K's = factor means

Factor Analysis (Y Observed; F latent)

THE UNIVERSITY OF IOWA

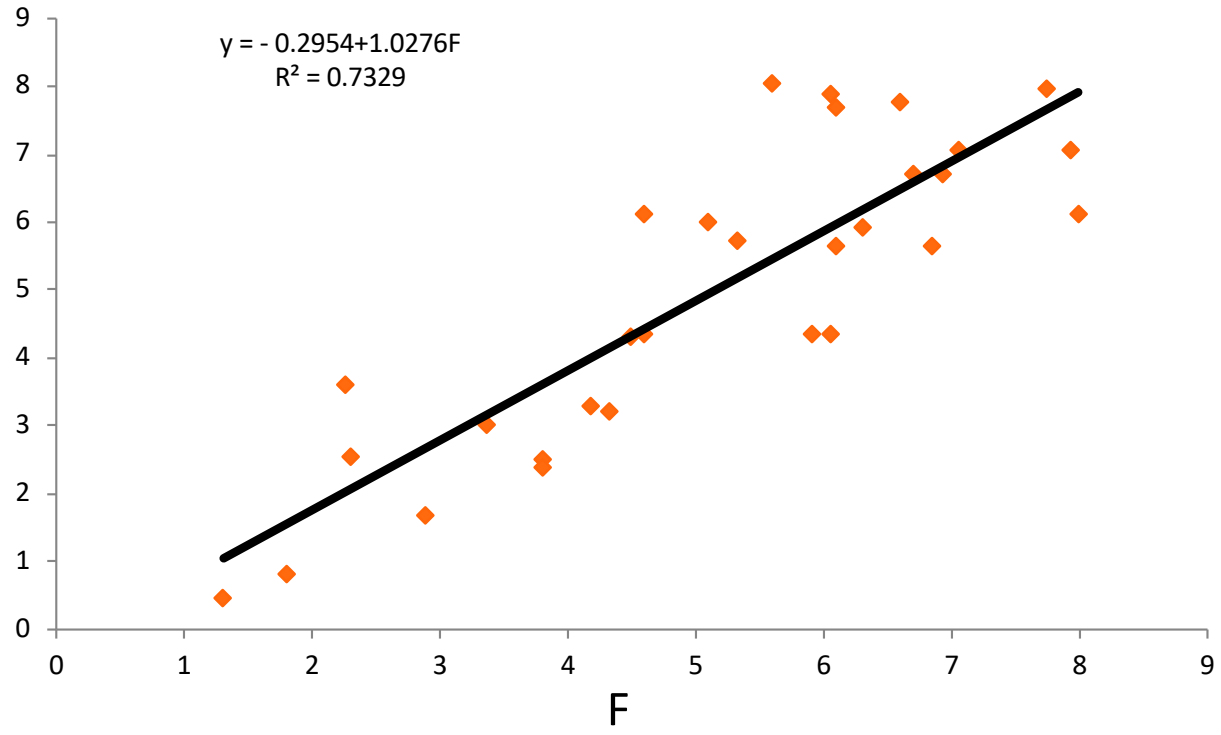
- The prediction of Y is done using a linear regression:

$$Y_{si} = \mu_i + \lambda_{i1} F_{s1} + e_{si}$$

μ is the intercept (where the line crosses the Y axis)

λ is the slope (the increase in Y for a one unit increase in F)

e is the error (or residual), with estimated error variance ψ_i^2



Item Response Models

THE UNIVERSITY OF IOWA

- Main idea of IRT: Build a measurement model for response variables that measure the same trait…but the responses are discrete
  - **Binary Item IRT = Logistic regression model** predicting each discrete observed outcome variable (item, subscale) from a latent trait predictor variable(s)

$$logit(Y_{si} = 1) = \mu_i + \lambda_{i1} F_{s1}$$
$$= a_i(\theta_s - b_i)$$

- i - item; s - subject; μ_i is the item intercept; λ_{i1} is the item slope (factor loading for factor 1); e_{is} is the error for the item and subject; Y_{is} is the item response (***assumed continuous***)

- Differs from confirmatory factor analysis:
  - Items are discrete (although CFA has item factor analysis)
  - Traits are normal

Latent Class Models

THE UNIVERSITY OF IOWA

- Main idea of LCA models: Build a measurement model for response variables that measure the same trait…but the responses are discrete and people belong to a finite set of groups that were not observed (latent classes)
  - **Binary Item LCA = Logistic ANOVA model** predicting each discrete observed outcome variable (item, subscale) from a latent trait predictor variable(s)

$$logit(Y_{si} = 1) = \mu_{ic} = \mu_i + \lambda_{ic} d_c$$

- i - item; s - subject; μ_i is the item intercept; λ_{i1} is the item slope (factor loading for factor 1); e_{is} is the error for the item and subject; Y_{is} is the item response (***assumed continuous***)

- Differs from confirmatory factor analysis and IRT:
  - Traits are discrete-–representing a "nominal" level of measurement
  - Model identification is very different (enables connections to AI)
  - Often used in exploratory analyses

- Note: many parameterizations exist (very flexible models)

DCMs are Confirmatory LCA Models

THE UNIVERSITY OF IOWA

- DCMs are confirmatory LCA models
  - Most defined for a set of *A* dichotomous attributes (**α**)
    - Attributes are either possessed (α = 1) or not (α = 0)
    - DCM attributes can have more than two levels
  - DCMs are LCA models with 2^A latent classes
    - Each possible combination of attribute possession
    - i.e., a test measuring 3 dichotomous attributes has 8 latent classes

- LCA measurement model parameters
  - Items measure only some attributes (so-called Q-matrix indicator)
  - Equated for classes with equivalent status of measured attributes

Building the LCDM

THE UNIVERSITY OF IOWA

- To demonstrate the LCDM, consider the item 2+3-1=?
  - Measures addition (attribute 1: α_{r1}) and subtraction (attribute 2: α_{r2})
- Only attributes defined by the Q-matrix are modeled for an item
- The LCDM provides the logit of a correct response as a function of the latent attributes mastered by a respondent:

$$Logit(Y_{ri} = 1|\boldsymbol{\alpha}_r) = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{r1} + \lambda_{i,1,(2)}\alpha_{r2} + \lambda_{i,2,(1,2)}\alpha_{r1}\alpha_{r2}$$

Bayesian Inference Networks (BayesNets)

THE UNIVERSITY OF IOWA

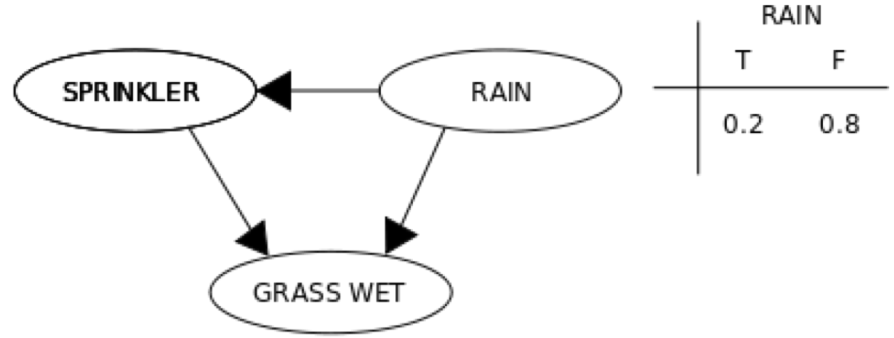
- Introduce a new vocabulary for psychometric things
- Nodes: categorical latent variables
  - Analogs to latent factors in factor analysis or item response theory
- Nodes can be Parents or Children
  - Parents: Not predicted by anything (we would call this an Exogenous variable)
  - Children: Predicted by parents (we would call this an Endogenous variable)
- Edges: conditional dependencies between:
  - Nodes
  - Nodes and items
- Often represented with a Directed Acyclic Graph or DAG

Woefully Short Primer on Bayesian Networks

THE UNIVERSITY OF IOWA

- BINs describe multivariate data using conditional probabilities

Image source: Wikipedia (yeah, that Wikipedia)



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

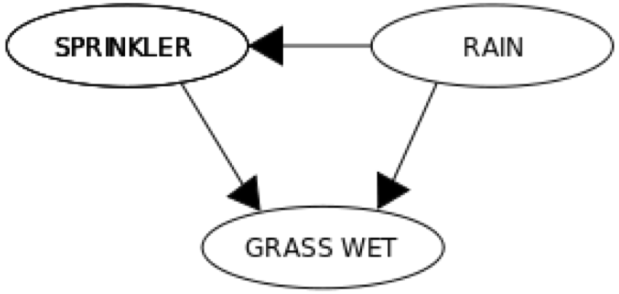
- In the image, three variables observed:
  - Did it rain?
  - Were the sprinklers on?
  - Was the grass wet?

- The BIN includes the set of parameters leading to the probabilities in the tables

Woefully Short Primer on Bayesian Networks

THE UNIVERSITY OF IOWA

SPRINKLER		
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Joint distribution of Rain, Sprinkler, and Grass Wet given by:

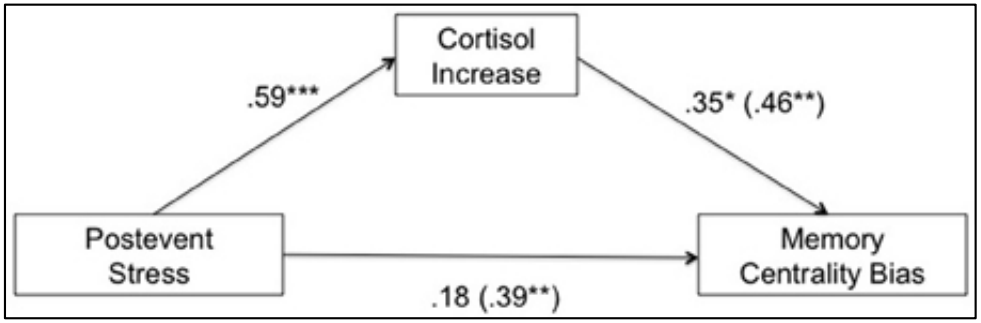
$$= P(Grass, Sprinkler, Rain)$$

$$P(GrassWet = T|Rain, Sprinkler)P(Sprinkler|Rain)P(Rain)$$

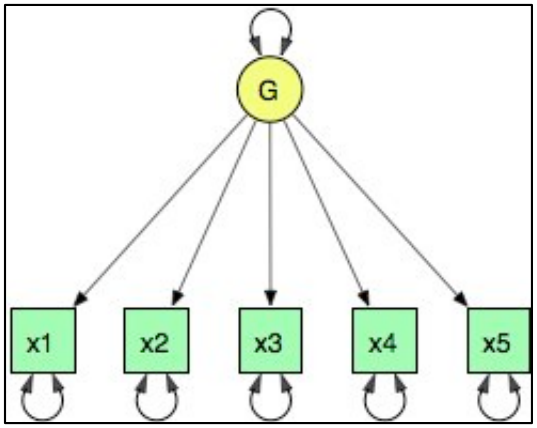
- Conditional/Marginal distribution of each variable: Bernoulli
- This example has all observed variables, but latent variables can also be defined
  - Hidden/unobserved nodes

Worlds Colliding….
Psychometric Models are BINs

- Here are some BINs that may be more familiar in the social sciences…



Conditional/Marginal distribution of each variable: Normal
Nodes: Observed variables (or more specifically, X, Y, and M)



Conditional/Marginal distribution of each variable: Normal
Nodes: 5 Observed variables (X1 – X5)
1 unobserved variable (G)

More BIN Terminology

THE UNIVERSITY OF IOWA

- Network Learning/Training = Estimation of model parameters
  - Often done with Bayesian/MCMC where priors are placed on nearly all parameters
- Estimation typically done using cross-validation
  - Estimation on one/several samples of data
  - Prediction done with left-out samples of data
- From Psychometrics: Model fit…not evaluated in same way
  - BIN model fit based on:
    - Prediction of left-out samples
    - Posterior predictive checks
    - Entropy (for categorical hidden nodes)
      - This is like saying your CFA model fits because your Omega reliability coefficient is high

COMMONALITIES ACROSS MODELS

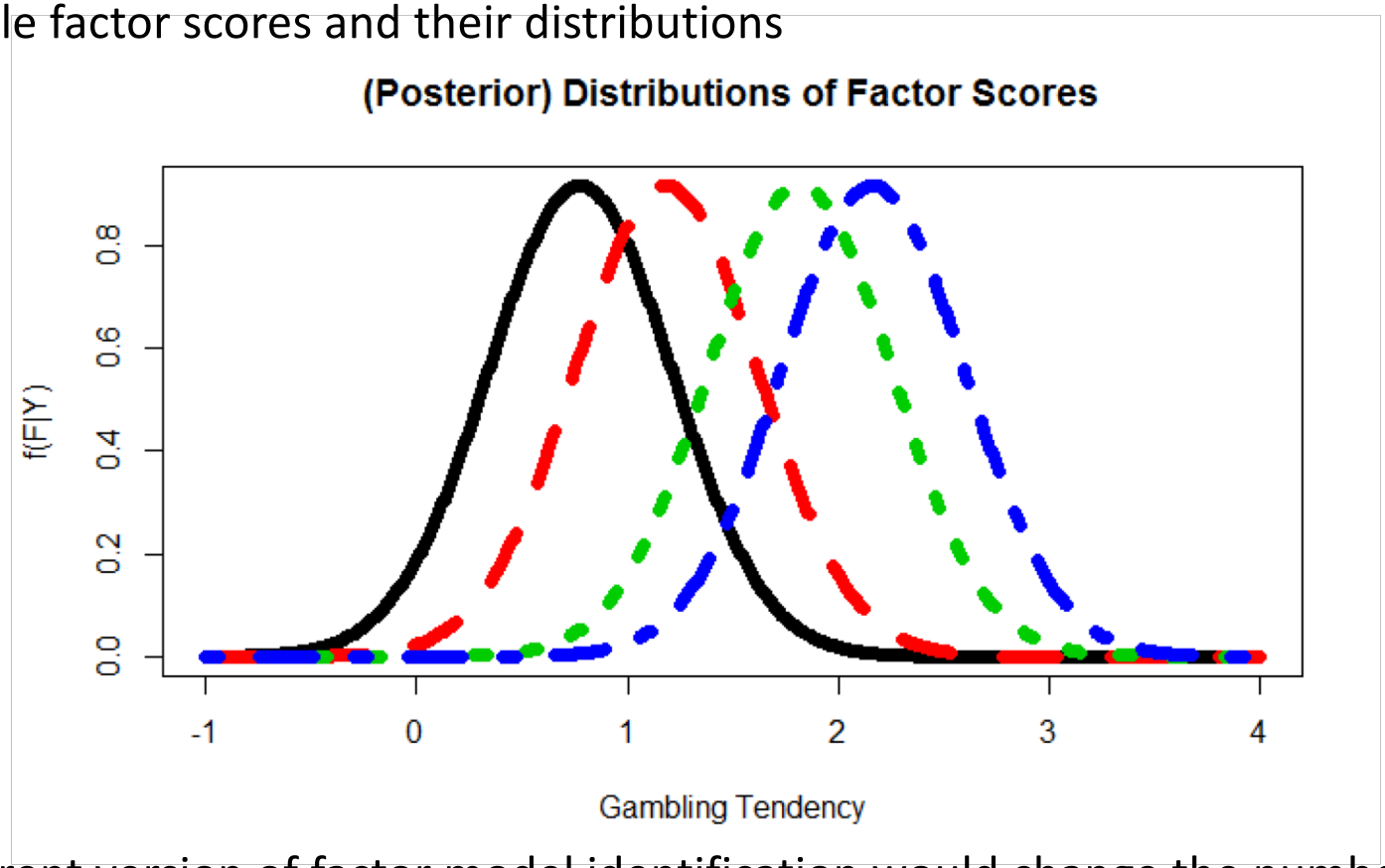
Where (Modern) Test Scores Come From

THE UNIVERSITY OF IOWA

- Factor scores (by other names) are used in many domains
  - Item response theory (CFA w/categorical items): GRE scores are factor scores
- A factor score is the estimate of a subject's unobserved latent trait
- Because this latent variable is not measured directly, it acts like it is missing data: you really cannot know with certainty its true value
- It is difficult to pin down what the missing data value (factor score value) should be precisely
  - Each factor score has a posterior distribution of possible values
  - Often, the mean of the posterior distribution is the "factor score"
    - In CFA, the mean is the most likely value
  - Depending on the test, there may be a lot of error (variability) in the distribution
- Therefore, the use of factor scores must reflect that the score is not known and is represented by a distribution

Draw Templin, Draw!

Example factor scores and their distributions

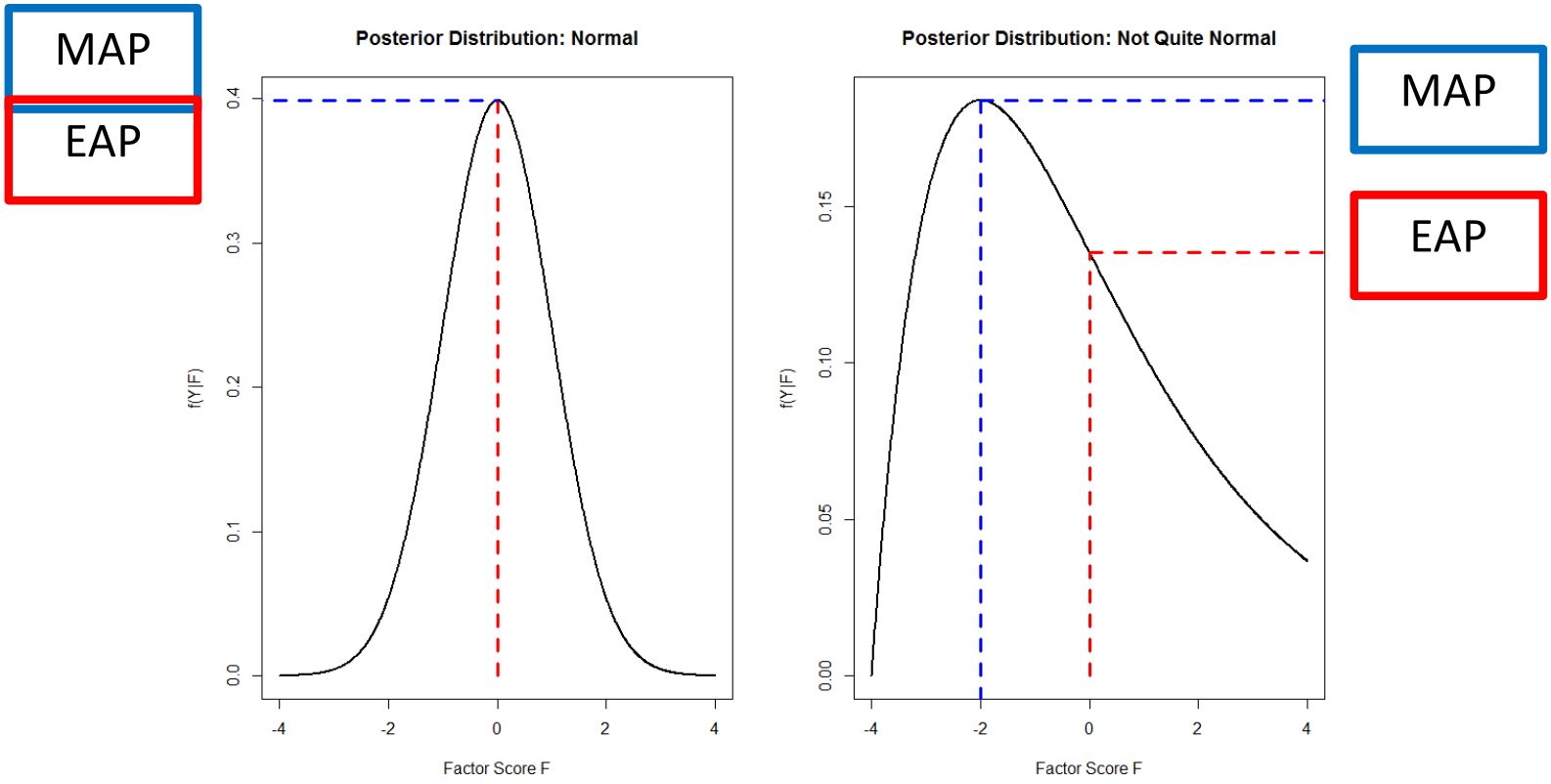


A different version of factor model identification would change the numbers on the X-axis, but the shapes and order of the distributions would not change

Factor scores provide a **weak ordering** of people (weak because of error)

How Distributions get Summarized into Scores

- There are two ways of providing a score from the factor score posterior distribution:
  - Expected a posteriori (EAP): the mean of the distribution
  - Maximum a posteriori (MAP): the most likely score from the distribution
- In CFA factor score distributions are normal (so EAP=MAP)



Additional Information on Factor Scores

THE UNIVERSITY OF IOWA

- For EAP factor scores:
  - $\hat{F}_p = E\left(f\left(F_p|\mathbf{Y}\right)\right)$
  - $SE\left(\hat{F}_p\right) = \sqrt{Var\left(f\left(F_p|\mathbf{Y}\right)\right)}$
- For MAP factor scores:
  - $\hat{F}_p = \arg\max_{F_p} f\left(F_p|\mathbf{Y}\right)$
  - $SE\left(\hat{F}_p\right) = \left[\frac{\partial^2}{\partial F_p^2} f\left(F_p|\mathbf{Y}\right)\Big|_{\hat{F}_p}\right]^{-\frac{1}{2}}$ (square root of Fisher's information)
- For CFA (Normal Data/Normal Factor) measurement models:
  - MAP = EAP
  - Variance is identical across all people, regardless of score
- For non-CFA measurement models:
  - MAP ≠ EAP (but does with infinite items)
  - Standard error is a function of the factor score

Scores Are Empirical Bayes Estimates

THE UNIVERSITY OF IOWA

- **For most (if not all) latent variable techniques**, the factor scores come from Empirical Bayes estimation—meaning there is a prior distribution present
  - Empirical = some or all of the parameters of the distribution of the latent variable are estimated (i.e., factor mean and variance)
  - Bayes = comes from the use of Bayes' Theorem

- Prior == Assumed factor distribution with mean/variance

- This is true for all CFA, IRT, mixed/multilevel/hierarchical models
  - And is true for models that don't have a label (e.g., Poisson Factor Analysis?)

Bayes' Theorem

- Bayes' Theorem states the conditional distribution of a variable A (soon to be our factor score) given values of a variable B (soon to be our data) is:

For Categorical A, replace integral with sum

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = \frac{f(B|A)f(A)}{\int_{a \in A} f(B|A=a)f(A=a)da}$$

- $f(A|B)$ is the **distribution** of A, conditional on B
  - We will come to know this as the posterior distribution of the factor score, conditional on the data observed or $f(\mathbf{F}|\mathbf{Y})$
- $f(B|A)$ is the **distribution** of B, conditional on A
  - We will come to know this as our measurement model or $f(\mathbf{Y}|\mathbf{F})$
- $f(A)$ is the marginal distribution of A
  - We will come to know this as the prior distribution of the factor or $f(\mathbf{F})$

Putting Together the Pieces of Empirical Bayes Factor Scores

THE UNIVERSITY OF IOWA

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = f(\mathbf{F}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{F})f(\mathbf{F})}{f(\mathbf{Y})}$$

- For $f(\mathbf{Y}|\mathbf{F})$, consider the measurement model (here CFA) for one item:

$$Y_{pi} = \mu_i + \lambda_i F_p + e_{pi}$$

Where: $e_{pi} \sim N(0, \psi_i^2)$

- Using expected values, we can show the distribution for this one item is:

$$f(Y_{pi}|F_p) \sim N(\mu_i + \lambda_i F_p, \psi_i^2)$$

- Therefore, for all I items, our conditional distribution is:

$$f(\mathbf{Y}|F_p) \sim N_I(\boldsymbol{\mu} + \boldsymbol{\Lambda} F_p, \boldsymbol{\Psi})$$

- With multiple factors, this becomes:

$$f(\mathbf{Y}|\mathbf{F}) \sim N_I(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{F}, \boldsymbol{\Psi})$$

Putting Together the Pieces of Empirical Bayes Factor Scores

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = f(\mathbf{F}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{F})f(\mathbf{F})}{f(\mathbf{Y})}$$

- For $f(\mathbf{F})$, consider the distribution assumed by the factor:
  - For one factor

$$f(F_p) \sim N(\mu_F, \sigma_F^2)$$

  - For multiple factors K

$$f(\mathbf{F}) \sim N_K(\boldsymbol{\mu}_F, \boldsymbol{\Phi})$$

- We must pick an identification method which determines if certain parameters of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Phi}$ are fixed or are estimated
  - Any method identification works, so we keep $\boldsymbol{\mu}_F$ and $\boldsymbol{\Phi}$ throughout

Putting Together the Pieces of Empirical Bayes Factor Scores

THE UNIVERSITY OF IOWA

$$f(A|B) = \frac{f(B|A)f(A)}{f(B)} = f(\mathbf{F}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{F})f(\mathbf{F})}{f(\mathbf{Y})}$$

- For $f(\mathbf{Y})$, we return to the model-implied mean vector and covariance matrix:

$$f(\mathbf{Y}) \sim N_I(\boldsymbol{\mu} + \boldsymbol{\Lambda}^T\boldsymbol{\mu}_F, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$$

A Quick Reminder About Types of Distributions

THE UNIVERSITY OF IOWA

- For two random variables x and z, a conditional distribution is written as: $f(z|x)$

- The conditional distribution is also equal to the joint distribution divided by the marginal distribution of the conditioning random variable

$$f(z|x) = \frac{f(z,x)}{f(x)}$$

- Therefore, the joint distribution can be found by the product of the conditional and marginal distributions:

$$f(z,x) = f(z|x)f(x)$$

- We can use this result in our analysis:

$$f(\mathbf{Y}|\mathbf{F})f(\mathbf{F}) = f(\mathbf{Y},\mathbf{F})$$

A Quick Reminder about Multivariate Normal Distributions

THE UNIVERSITY OF IOWA

- If $\mathbf{X}$ is distributed multivariate normally:

Conditional distributions of $\mathbf{X}$ are multivariate normal

- We can show that $f(\mathbf{Y}, \mathbf{F})$, the joint distribution of the data and the factors, is multivariate normal

- We can then use the result above (shown on the next slides) to show that our posterior distribution of the factor scores is also multivariate normal
  - This result **only** applies for measurement models assuming normally distributed data and normally distributed factors: CFA
  - For IRT (and other measurement models), this result will not hold—but this distribution is asymptotically normal as the number of items gets large

Conditional Distributions of MVN Variables are Multivariate Normal

- The conditional distribution of sets of variables from a MVN is also MVN

- If we were interested in the distribution of the first q variables, we partition three matrices:
  - The data: $[\mathbf{X}_{1:(N\ x\ q)} \mid \mathbf{X}_{2:(N\ x\ p-q)}]$
  - The mean vector: $\begin{bmatrix} \boldsymbol{\mu}_{1:(q\ x\ 1)} \\ \hline \boldsymbol{\mu}_{2:(p-q\ x\ 1)} \end{bmatrix}$
  - The covariance matrix: $\left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11:(q\ x\ q)} & \boldsymbol{\Sigma}_{12:(q\ x\ p-q)} \\ \hline \boldsymbol{\Sigma}_{21:(p-q\ x\ q)} & \boldsymbol{\Sigma}_{22:(p-q\ x\ p-q)} \end{array}\right]$

Conditional Distributions of MVN Variables

- The, $f(\mathbf{X}_1|\mathbf{X}_2)$, conditional distribution of $\mathbf{X}_1$ given the values of $\mathbf{X}_2 = \mathbf{x}_2$ is then:

$$\mathbf{X}_1|\mathbf{X}_2 \sim N_q(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$$

Where (using our partitioned matrices):

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2^T - \boldsymbol{\mu}_2)$$

And:

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Derive, Templin, Derive!

THE UNIVERSITY OF IOWA

- The joint distribution of all I items and K factor scores is

$$f(\mathbf{Y}, \mathbf{F}) = f\left(\begin{bmatrix}\mathbf{Y}\\ \mathbf{F}\end{bmatrix}\right)$$

$$= N_{I+K}\left(\begin{bmatrix}\boldsymbol{\mu} + \boldsymbol{\Lambda}^T\boldsymbol{\mu}_F\\ \boldsymbol{\mu}_F\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} & \boldsymbol{\Lambda}\boldsymbol{\Phi}\\ \boldsymbol{\Phi}\boldsymbol{\Lambda}^T & \boldsymbol{\Phi}\end{bmatrix}\right)$$

- Using the conditional distributions of MVNs result:

$f(\mathbf{F}_p|\mathbf{Y}_p)$ is MVN:

With mean: $\boldsymbol{\mu}_F + \boldsymbol{\Phi}\boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}(\mathbf{Y}_p^T - \boldsymbol{\mu})$

And Covariance: $\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}\,\boldsymbol{\Lambda}\boldsymbol{\Phi}$

#WTFTemplin

What All That Math Means for Factor Scores

THE UNIVERSITY OF IOWA

- **When using measurement models assuming normally distributed data and normally distributed factors (CFA):**
  - The posterior distribution of the factor scores is MVN
  - Therefore, the **most likely** factor score (MAP) and the **expected** factor score (EAP) is given by the mean from the previous slides
  - The factor score is a function of the model parameter estimates and the data

Factor Scores via Metropolis Hastings

- Place holder for R script…

WRAPPING UP

Wrapping Up

THE UNIVERSITY OF IOWA

- Today was an introduction to Bayesian statistics
  - Bayes = use of prior distributions on parameters
- We used two methods for estimation:
  - MAP estimation – far less common
  - MCMC estimation
    - Commonly, people will say Bayesian and mean MCMC – but Bayesian is just the addition of priors. MCMC is one way of estimating Bayesian models!
- MCMC is effective for most Bayesian models:
  - Model likelihood and prior likelihood are all that are needed
- MCMC is estimation by brute force:
  - Can be very slow, computationally intensive, and disk-space intensive