An Introduction to MCMC methods and Bayesian Statistics

What will we cover in this first session?

- What is Bayesian Statistics? (as opposed to classical or frequentist statistics)
- What is MCMC estimation?
- MCMC algorithms and Gibbs Sampling
- MCMC diagnostics
- MCMC Model comparisons

WHAT IS BAYESIAN STATISTICS?

Why do we need to know about Bayesian statistics?

- The rest of this workshop is primarily about MCMC methods which are a family of estimation methods used for fitting realistically complex models.
- MCMC methods are generally used on Bayesian models which have subtle differences to more standard models.
- As most statistical courses are still taught using classical or frequentist methods we need to describe the differences before going on to consider MCMC methods.

Bayes Theorem

Bayesian statistics named after Rev. Thomas Bayes (1702-1761)

Bayes Theorem for probability events A and B

$$p(A \mid B) = \frac{p(B \mid A)\,p(A)}{p(B)}$$

Or for a set of mutually exclusive and exhaustive events (i.e. $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$):

$$p(A_i \mid B) = \frac{p(B \mid A_i)\,p(A_i)}{\sum_j p(B \mid A_j) P(A_j)}$$

Example – coin tossing

Let A be the event of 2 Heads in three tosses of a fair coin. B be the event of 1st coin is a Head.

Three coins have 8 equally probable patterns
{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}

A = {HHT,HTH,THH} →p(A)=3/8

B = {HHH,HTH,HTH,HTT} →p(B)=1/2

A|B = {HHT,HTH}|{HHH,HTH,HTH,HTT} →p(A|B)=1/2

B|A = {HHT,HTH}|{HHT,HTH,THH} →p(B|A)=2/3

P(A|B) = P(B|A)P(A)/P(B) = (2/3*3/8)/(1/2) = 1/2

Example 2 – Diagnostic testing

A new HIV test is claimed to have “95% sensitivity and 98% specificity”

In a population with an HIV prevalence of 1/1000, what is the chance that a patient testing positive actually has HIV?

Let A be the event patient is truly positive, A’ be the event that they are truly negative

Let B be the event that they test positive

Diagnostic Testing continued:

We want p(A|B)

"95% sensitivity" means that p(B|A) = 0.95

"98% specificity" means that p(B|A') = 0.02

So from Bayes Theorem:

$$p(A \mid B) = \frac{p(B \mid A)\,p(A)}{p(B \mid A)\,p(A) + p(B \mid A')\,p(A')}$$

$$= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} = 0.045$$

Thus over 95% of those testing positive will, in fact, not have HIV.

Being Bayesian!

So the vital issue in this example is *how should this test result change our prior belief that the patient is HIV positive?*

The disease prevalence (p=0.001) can be thought of as a '*prior*' probability.

Observing a positive result causes us to modify this probability to p=0.045 which is our '*posterior*' probability that the patient is HIV positive.

This use of Bayes theorem applied to *observables* is uncontroversial however its use in general statistical analyses where *parameters* are unknown quantities is more controversial.

Bayesian Inference

In Bayesian inference there is a fundamental distinction between

- Observable quantities *x,* i.e. the data
- Unknown quantities θ

θ can be statistical parameters, missing data, latent variables…

- Parameters are treated as random variables

In the Bayesian framework we make probability statements about model parameters

In the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data.

Prior distributions

As with all statistical analyses we start by positing a model which specifies $p(x|\theta)$

This is the **likelihood** which relates all variables into a '**full probability model**'

However from a Bayesian point of view :

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data
- Therefore we specify a **prior distribution** $p(\theta)$

Note this is like the prevalence in the example

Posterior Distributions

Also *x* is known so should be conditioned on and here we use Bayes theorem to obtain the conditional distribution for unobserved quantities given the data which is known as the **posterior distribution**.

$$p(\theta \mid x) = \frac{p(\theta)\,p(x \mid \theta)}{\int p(\theta)\,p(x \mid \theta)d\theta} \propto p(\theta)\,p(x \mid \theta)$$

The prior distribution expresses our uncertainty about θ **before** seeing the data.

The posterior distribution expresses our uncertainty about θ **after** seeing the data.

Examples of Bayesian Inference using the Normal distribution

Known variance, unknown mean

It is easier to consider first a model with 1 unknown parameter.

Suppose we have a sample of Normal data:

$$x_i \sim N(\mu, \sigma^2), i = 1, \ldots, n.$$

Let us assume we know the variance, σ^2 and we assume a prior distribution for the mean, μ based on our prior beliefs:

$$\mu \sim N(\mu_0, \sigma_0^2)$$

Now we wish to construct the posterior distribution $p(\mu|x)$.

Posterior for Normal distribution mean

So we have

$$p(\mu) = (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mu - \mu_0)^2 / \sigma_0^2)$$

$$p(x_i \mid \mu) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(x_i - \mu)^2 / \sigma^2)$$

and hence

$$p(\mu \mid x) = p(\mu)\, p(x \mid \mu)$$

$$= (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mu - \mu_0)^2 / \sigma_0^2) \times$$

$$\prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(x_i - \mu)^2 / \sigma^2)$$

$$\propto \exp(-\tfrac{1}{2}\mu^2(1/\sigma_0^2 + n/\sigma^2) + \mu(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2) + cons)$$

Posterior for Normal distribution mean (continued)

For a Normal distribution with response *y* with mean θ and variance ϕ we have

$$f(y) = (2\pi\phi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(y-\theta)^2/\phi\}$$

$$\propto \exp\{-\tfrac{1}{2} y^2 \phi^{-1} + y\theta/\phi + cons\}$$

We can equate this to our posterior as follows:

$$\propto \exp(-\tfrac{1}{2}\mu^2(1/\sigma_0^2 + n/\sigma^2) + \mu(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2) + cons)$$

$$\rightarrow \phi = (1/\sigma_0^2 + n/\sigma^2)^{-1} \text{ and } \theta = \phi(\mu_0/\sigma_0^2 + \sum_i x_i/\sigma^2)$$

Precisions and means

In Bayesian statistics the precision = 1/variance is often more important than the variance.

For the Normal model we have

$$1/\phi = (1/\sigma_0^2 + n/\sigma^2) \text{ and } \theta = \phi(\mu_0/\sigma_0^2 + \bar{x}/(\sigma^2/n))$$

In other words the posterior precision = sum of prior precision and data precision, and the posterior mean is a (precision weighted) average of the prior mean and data mean.

Large sample properties

As $n \to \infty$

Posterior precision $1/\phi = (1/\sigma_0^2 + n/\sigma^2) \to n/\sigma^2$

So posterior variance $\to \sigma^2/n$

Posterior mean $\theta = \phi(\mu_0/\sigma_0^2 + \bar{x}/(\sigma^2/n)) \to \bar{x}$

And so posterior distribution

$p(\mu \mid x) \to N(\bar{x}, \sigma^2/n)$

Compared to $p(\bar{x} \mid \mu) = N(\mu, \sigma^2/n)$

in the frequentist setting

Girls Heights Example

10 girls aged 18 had both their heights and weights measured.

Their heights (in cm) were as follows:

169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3

We will assume the population variance is known to be 50.

Two individuals gave the following prior distributions for the mean height:

Individual 1 $p_1(\mu) \sim N(165, 2^2)$

Individual 2 $p_2(\mu) \sim N(170, 3^2)$

Constructing posterior 1

To construct the posterior we use the formulae we have just calculated

From the prior, $\mu_0 = 165, \sigma_0^2 = 4$

From the data, $\bar{x} = 165.52, \sigma^2 = 50, n = 10$

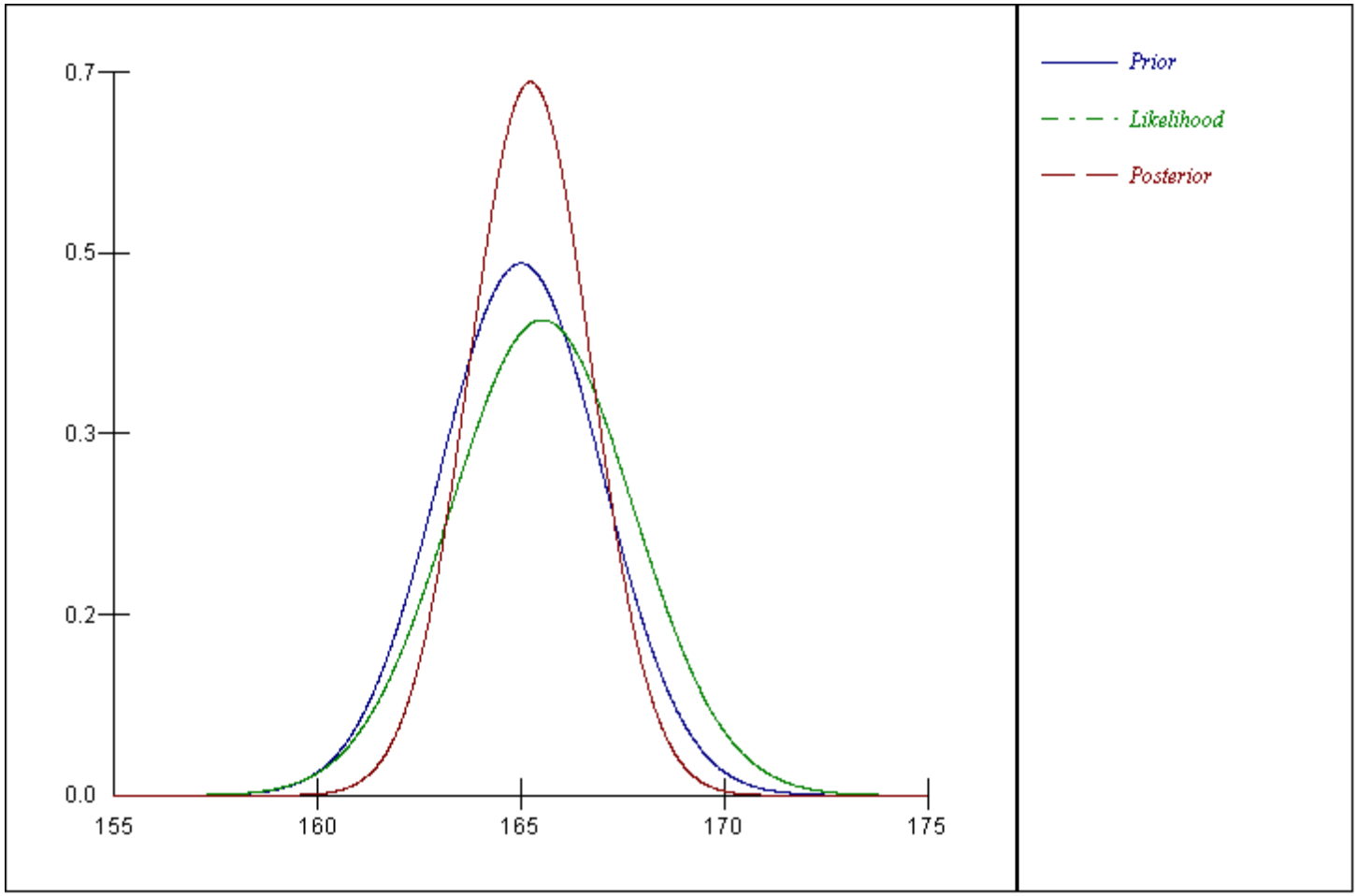
The posterior is therefore

$$p(\mu \mid x) \sim N(\theta_1, \phi_1)$$

$$\text{where } \phi_1 = (\tfrac{1}{4} + \tfrac{10}{50})^{-1} = 2.222,$$

$$\theta_1 = \phi_1(\tfrac{165}{4} + \tfrac{1655.2}{50}) = 165.23.$$

Prior and posterior comparison



Constructing posterior 2

- Again to construct the posterior we use the earlier formulae we have just calculated
- From the prior, $\mu_0 = 170, \sigma_0^2 = 9$
- From the data, $\bar{x} = 165.52, \sigma^2 = 50, n = 10$
- The posterior is therefore

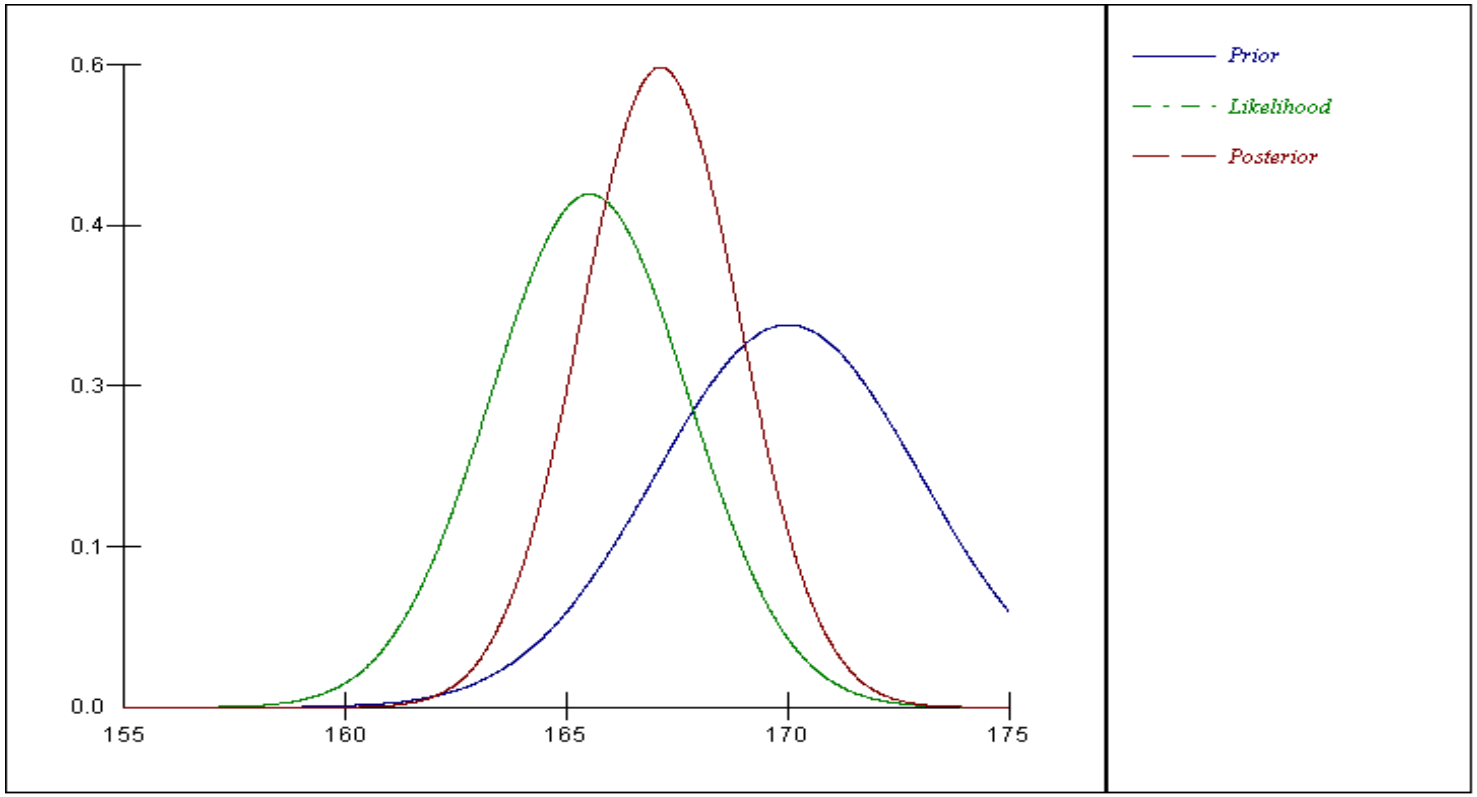
$$p(\mu \mid x) \sim N(\theta_2, \phi_2)$$

$$\text{where } \phi_2 = (\tfrac{1}{9} + \tfrac{10}{50})^{-1} = 3.214,$$

$$\theta_2 = \phi_2(\tfrac{170}{9} + \tfrac{1655.2}{50}) = 167.12.$$

Prior 2 comparison

Note this prior is not as close to the data as prior 1 and hence posterior is somewhere between prior and likelihood.



Other conjugate examples

When the posterior is in the same family as the prior we have *conjugacy*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	Mean	Normal	Normal
Normal	Precision	Gamma	Gamma
Binomial	Probability	Beta	Beta
Poisson	Mean	Gamma	Gamma

In all cases

The posterior mean is a compromise between the prior mean and the MLE

The posterior s.d. is less than both the prior s.d. and the s.e. (MLE)

'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule' (Senn)

As $n \to \infty$

The posterior mean $\to$ the MLE

The posterior s.d. $\to$ the s.e. (MLE)

The posterior does not depend on the prior.

Non-informative priors

We often do not have any prior information, although true Bayesian's would argue we always have some prior information!

We would hope to have good agreement between the frequentist approach and the Bayesian approach with a non-informative prior.

Diffuse or flat priors are often better terms to use as no prior is strictly non-informative!

For our example of an unknown mean, candidate priors are a Uniform distribution over a large range or a Normal distribution with a huge variance.

Point and Interval Estimation

In Bayesian inference the outcome of interest for a parameter is its full posterior distribution however we may be interested in summaries of this distribution.

A simple point estimate would be the mean of the posterior. (although the median and mode are alternatives.)

Interval estimates are also easy to obtain from the posterior distribution and are given several names, for example credible intervals, Bayesian confidence intervals and Highest density regions (HDR). All of these refer to the same quantity.

Credible Intervals

If we consider the heights example with our first prior then our posterior is

$$P(\mu|x) \sim N(165.23, 2.222),$$

and a 95% credible interval for μ is

$$165.23 \pm 1.96 \times \text{sqrt}(2.222) = (162.31, 168.15).$$

Similarly prior 2 results in a 95% credible interval for μ is

$$(163.61, 170.63).$$

Note that credible intervals can be interpreted in the more natural way that there is a probability of 0.95 that the interval contains μ rather than the frequentist conclusion that 95% of such intervals contain μ.

MCMC METHODS

How does one fit models in a Bayesian framework?

In the first section we illustrated a use of conjugate priors to evaluate a posterior distribution for a model with one unknown parameter.

Let us now consider a simple linear regression:

$$weight_i = \beta_0 + \beta_1 height_i + e_i$$

$$e_i \sim N(0, \sigma^2)$$

With conjugate priors:

$$\beta_0 \sim N(0, m_0), \beta_1 \sim N(0, m_1),$$

$$\sigma^2 \sim \Gamma^{-1}(\varepsilon, \varepsilon)$$

where $m_0 = m_1 = 10^6, \varepsilon = 10^{-3}$

So our goal now is to make inferences on the joint posterior distribution:

$$p(\beta_0, \beta_1, \sigma^2 \mid y)$$

MCMC Methods

Goal: To sample from joint posterior distribution:

$$p(\beta_0, \beta_1, \sigma^2 \mid y)$$

Problem: For complex models this involves multidimensional integration

Solution: It may be possible to sample from conditional posterior distributions,

$$p(\beta_0 \mid y, \beta_1, \sigma^2),\ p(\beta_1 \mid y, \beta_0, \sigma^2),\ p(\sigma^2 \mid y, \beta_0, \beta_1)$$

It can be shown that after *convergence* such a sampling approach generates dependent samples from the joint posterior distribution.

Gibbs Sampling

When we can sample directly from the conditional posterior distributions then such an algorithm is known as Gibbs Sampling.

This proceeds as follows for the linear regression example:

Firstly give all unknown parameters starting values,

$$\beta_0(0), \beta_1(0), \sigma^2(0).$$

Next loop through the following steps:

Gibbs Sampling ctd.

Sample from

$p(\beta_0 \mid y, \beta_1(0), \sigma^2(0))$ to generate $\beta_0(1)$ and then from

$p(\beta_1 \mid y, \beta_0(1), \sigma^2(0))$ to generate $\beta_1(1)$ and then from

$p(\sigma^2 \mid y, \beta_0(1), \beta_1(1))$ to generate $\sigma^2(1)$.

These steps are then repeated with the generated values from this loop replacing the starting values. The chain of values produced by this procedure is known as a Markov chain, and it is hoped that this chain converges to its equilibrium distribution which is the joint posterior distribution.

Calculating the conditional distributions

In order for the algorithm to work we need to sample from the conditional posterior distributions.

If these distributions have standard forms then it is easy to draw random samples from them.

Mathematically we write down the full posterior and assume all parameters are constants apart from the parameter of interest.

We then try to match the resulting formulae to a standard distribution.

The next 4 slides will probably be skipped but are in the talk for reference purposes

Note the new STAT-JR software gives these derivations!

Matching distributional forms

If a parameter θ follows a Normal(μ,σ^2) distribution then we can write

$$p(\theta) \propto \exp(a\theta^2 + b\theta + const)$$

$$\text{where } a = -\frac{1}{2\sigma^2} \text{ and } b = \frac{\mu}{\sigma^2}$$

Similarly if θ follows a Gamma(α,β) distribution then we can write

$$p(\theta) \propto \theta^a \exp(b\theta)$$

$$\text{where } a = \alpha - 1 \text{ and } b = -\beta$$

Step 1: β_0

$$p(\beta_0 \mid y, \beta_1, \sigma^2) \propto p(\beta_0)\, p(y \mid \beta_0, \beta_1, \sigma^2)$$

$$\propto \frac{1}{\sqrt{m_0}} \exp\left(-\frac{\beta_0^2}{2m_0}\right) \prod_i \left[\frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - x_i\beta_1)^2\right) \right]$$

$$\propto \exp\left[\left(-\tfrac{1}{2}\left(\frac{1}{m_0} + \frac{N}{\sigma^2}\right)\right)\beta_0^2 + \frac{1}{\sigma^2}\sum_i (y_i - x_i\beta_1)\beta_0 + const \right]$$

Matching powers gives

$$\frac{1}{\sigma^2_{\beta_0}} = \frac{1}{m_0} + \frac{N}{\sigma^2} \rightarrow \sigma^2_{\beta_0} = \left[\frac{1}{m_0} + \frac{N}{\sigma^2}\right]^{-1}$$

$$\mu_{\beta_0} = \sigma^2_{\beta_0} b = \left[\frac{1}{m_0} + \frac{N}{\sigma^2}\right]^{-1} \frac{1}{\sigma^2}\sum_i (y_i - x_i\beta_1)$$

$$\text{as } m_0 \rightarrow \infty, \beta_0 \sim N\left(\frac{1}{N}\sum_i (y_i - x_i\beta_1), \frac{\sigma^2}{N}\right)$$

Step 2: β_1

$$p(\beta_1 \mid y, \beta_0, \sigma^2) \propto p(\beta_1)\, p(y \mid \beta_0, \beta_1, \sigma^2)$$

$$\propto \frac{1}{\sqrt{m_1}} \exp\left(-\frac{\beta_1^2}{2m_1}\right) \prod_i \left[\frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - x_i\beta_1)^2\right) \right]$$

$$\propto \exp\left[\left(-\tfrac{1}{2}\left(\frac{1}{m_1} + \frac{\sum_i x_i^2}{\sigma^2}\right)\right)\beta_1^2 + \frac{1}{\sigma^2}\sum_i (y_i - \beta_0) x_i \beta_1 + const \right]$$

Matching powers gives

$$\frac{1}{\sigma^2_{\beta_1}} = \frac{1}{m_1} + \frac{\sum_i x_i^2}{\sigma^2} \rightarrow \sigma^2_{\beta_1} = \left[\frac{1}{m_1} + \frac{\sum_i x_i^2}{\sigma^2} \right]^{-1}$$

$$\mu_{\beta_1} = \sigma^2_{\beta_1} b = \left[\frac{1}{m_1} + \frac{\sum_i x_i^2}{\sigma^2} \right]^{-1} \frac{1}{\sigma^2} \sum_i (x_i (y_i - \beta_0))$$

$$\text{as } m_1 \rightarrow \infty, \beta_1 \sim N\left(\frac{\sum_i x_i y_i - \beta_0 \sum_i x_i}{\sum_i x_i^2}, \frac{\sigma^2}{\sum_i x_i^2} \right)$$

Step 3: $1/\sigma^2$

$$p(1/\sigma^2 \mid y, \beta_0, \beta_1) \propto p(1/\sigma^2)\, p(y \mid \beta_0, \beta_1, \sigma^2)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\varepsilon-1} \exp\left[-\frac{\varepsilon}{\sigma^2}\right] \prod_i \left[\frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - x_i\beta_1)^2\right)\right]$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}+\varepsilon-1} \exp\left[-\frac{1}{\sigma^2}\left(\varepsilon + \tfrac{1}{2}\sum_i (y_i - \beta_0 - x_i\beta_1)^2\right)\right]$$

Matching terms gives

$$p(1/\sigma^2 \mid y, \beta_0, \beta_1) \sim \Gamma(a, b)$$

where $a = \varepsilon + \frac{N}{2}, b = \varepsilon + \frac{1}{2}\sum_i e_i^2$

Algorithm Summary

Repeat the following three steps

1. Generate β_0 from its Normal conditional distribution.
2. Generate β_1 from its Normal conditional distribution.
3. Generate $1/\sigma^2$ from its Gamma conditional distribution

Convergence and burn-in

Two questions that immediately spring to mind are:

1. We start from arbitrary starting values so when can we safely say that our samples are from the correct distribution?
2. After this point how long should we run the chain for and store values?

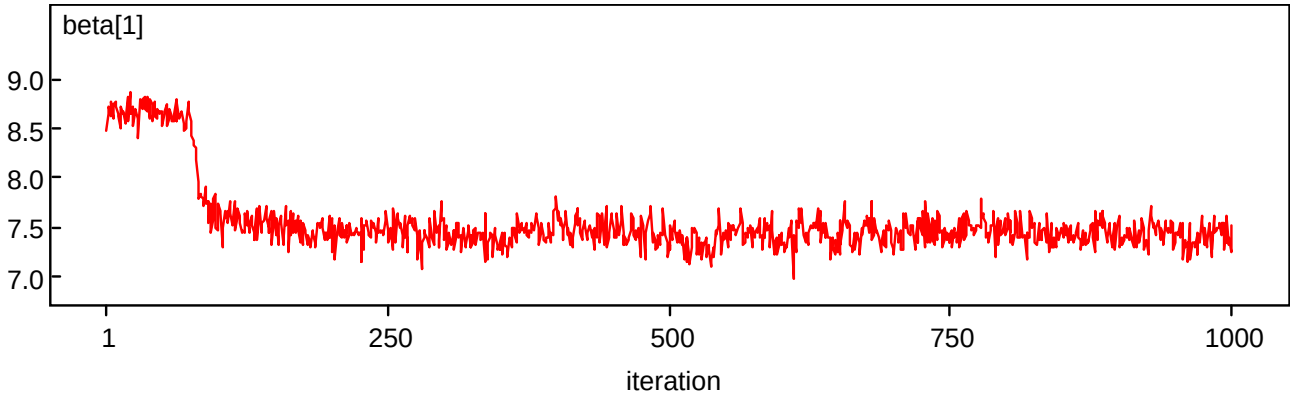
MCMC DIAGNOSTICS

Checking Convergence

This is the researchers responsibility!

Convergence is to a target **distribution** (the required posterior), not to a single value as in ML methods.

Once convergence has been reached, samples should look like a random scatter about a stable mean value.



Convergence occurs here at around 100 iterations.

How many iterations after convergence?

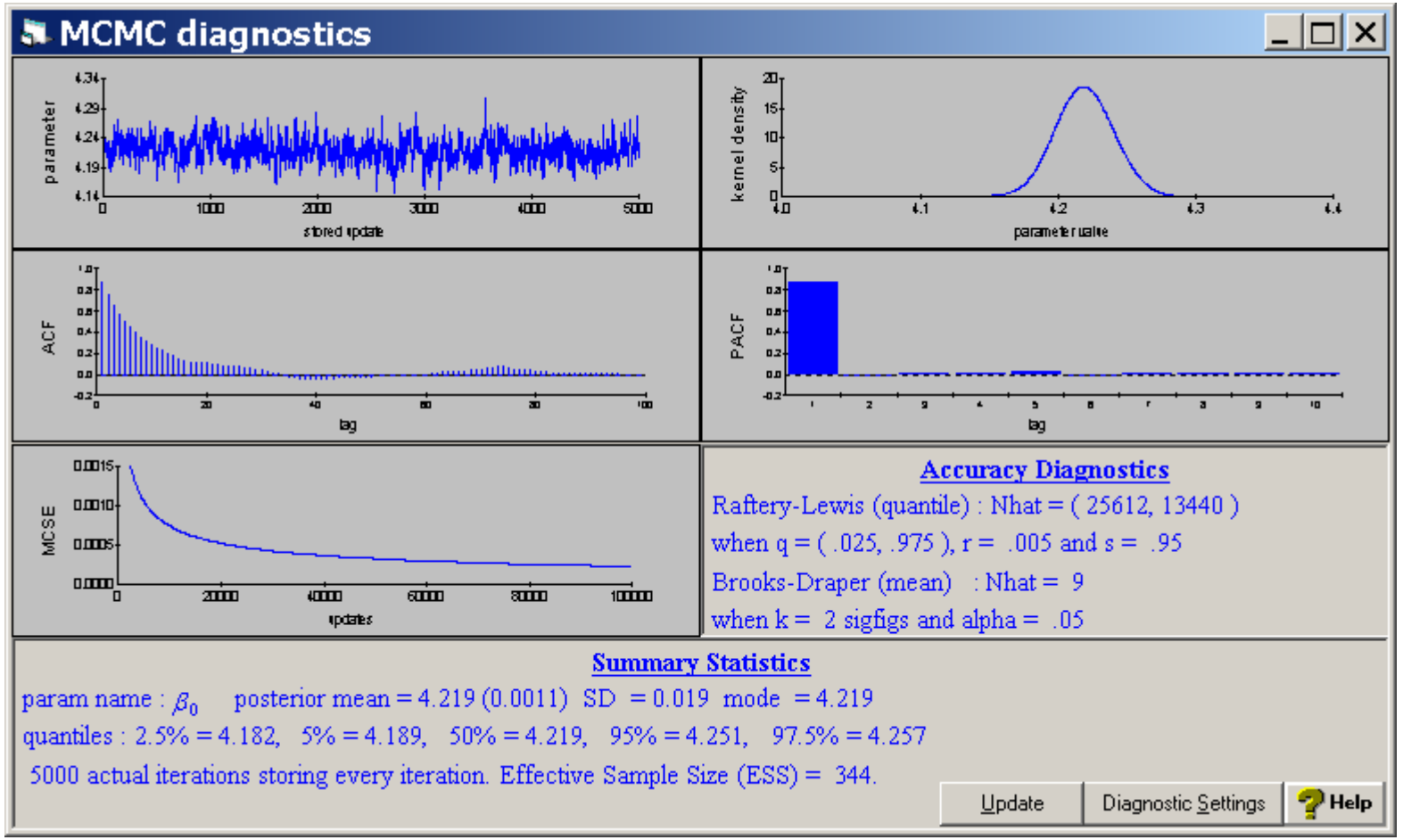
After convergence, further iterations are needed to obtain samples for posterior inference.

More iterations = more accurate posterior estimates.

MCMC chains are dependent samples and so the dependence or autocorrelation in the chain will influence how many iterations we need.

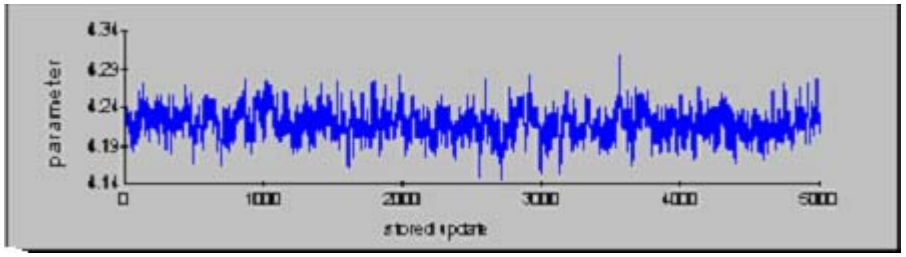
Accuracy of the posterior estimates can be assessed by the Monte Carlo standard error (MCSE) for each parameter.

MCMC diagnostics in MLwiN - Example



We will describe each pane separately – Note Stat-JR has similar six way plots!

Trace plot

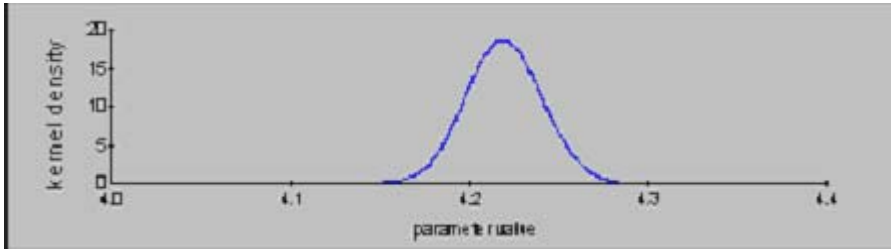


This graph plots the generated values of the parameter against the iteration number.

A crude test of mixing is the ‘blue finger’ test.

This chain doesn’t mix that well but could be worse!

Kernel Density plot



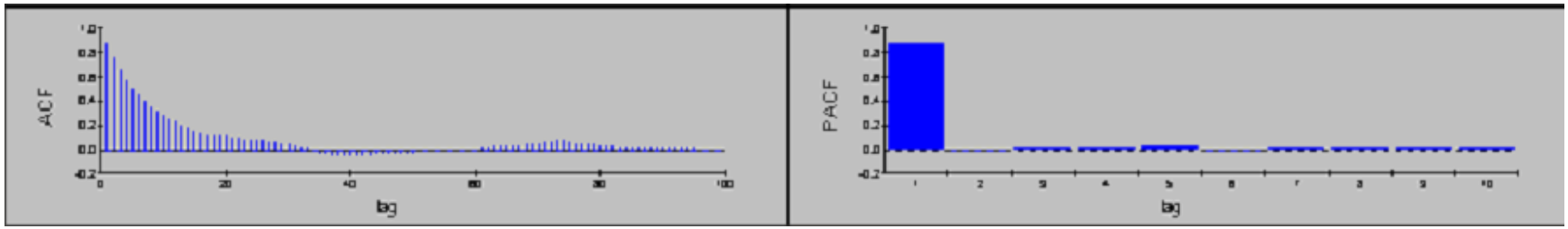
This plot is like a smoothed histogram.

Instead of counting the estimates into bins of particular widths like a histogram, the effect of each iteration is spread around the estimate via a Kernel function e.g. a normal distribution.

This means that at each point we get the sum of the Kernel function parts for each iteration.

The Kernel density plot has a smoothness parameter that can be modified.

Time series diagnostics



Here we have the Auto correlation function (ACF) and partial autocorrelation function (PACF) plots.

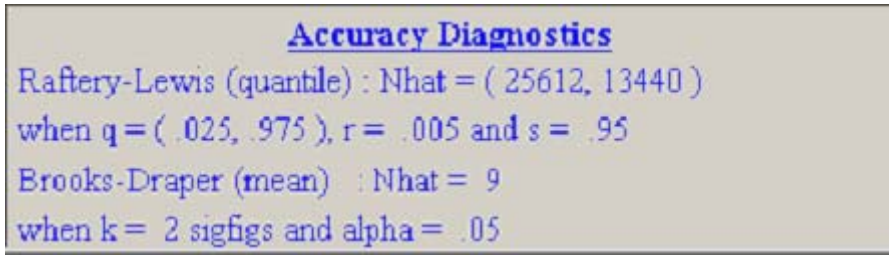
The ACF measures how correlated the values in the chain are with their close neighbours. The lag is the distance between the two chains to be compared.

An independent chain will have approximately zero autocorrelation at each lag.

A Markov chain should have a power relationship in the lags i.e. if ACF(1) = ρ then ACF(2) = ρ^2 etc. This is known as an AR(1) process.

The PACF measures discrepancies from such a process and so should normally have values 0 after lag 1.

Accuracy Diagnostics



MLwiN has 2 accuracy diagnostics:

- Raftery-Lewis works on quantiles of distribution (Not given in Stat-JR).
- Brooks-Draper works on quoting the mean to *n* significant figures. It's formulae uses the estimate, it's s.d. and the ACF and it can often give very small or very large values! Available from SummaryStats in Stat-JR.

Summary Statistics

Summary Statistics

param name : β_0 posterior mean = 4.219 (0.0011) SD = 0.019 mode = 4.219
quantiles : 2.5% = 4.182, 5% = 4.189, 50% = 4.219, 95% = 4.251, 97.5% = 4.257
5000 actual iterations storing every iteration. Effective Sample Size (ESS) = 344

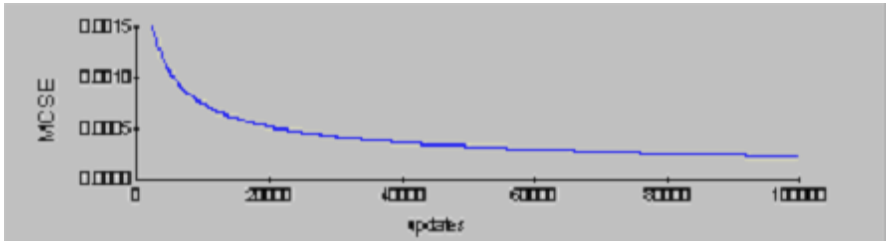
Three estimates of location are given:

- Mean – from the chain.
- Mode – from the kernel plot.
- Median (50% quantile) – by sorting the chain and finding the middle value.

The SD is calculated from the chain and the other quantiles are used to give (possibly) non-symmetric interval estimates.

The MCSE and ESS will be discussed next.

Monte Carlo Standard Error



The Monte Carlo Standard Error (MCSE) is an indication of how much error is in the estimate due to the fact that MCMC is used.

As the number of iterations increases the MCSE$\rightarrow$0.

For an independent sampler it equals the SD/$\sqrt{n}$.

However it is adjusted due to the autocorrelation in the chain.

The graph above gives estimates for the MCSE for longer runs.

Effective Sample Size

This quantity gives an estimate of the equivalent number of independent iterations that the chain represents.

This is related to the ACF and the MCSE.

Its formula is: n / κ where $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k).$

For this parameter our 5,000 actual iterations are equivalent to only 344 independent iterations!

Inference using posterior samples from MCMC runs

A powerful feature of MCMC and the Bayesian approach is that all inference is based on the joint posterior distribution.

We can therefore address a wide range of substantive questions by appropriate summaries of the posterior.

Typically report either the mean or median of the posterior samples for each parameter of interest as a point estimate

2.5% and 97.5% percentiles of the posterior sample for each parameter give a 95% posterior credible interval (interval within which the parameter lies with probability 0.95)

Derived Quantities

Once we have a sample from the posterior we can answer lots of questions simply by investigating this sample.

Examples:

What is the probability that $\theta>0$?

What is the probability that $\theta_1> \theta_2$?

What is a 95% interval for $\theta_1/(\theta_1+ \theta_2)$?

MODEL COMPARISON

Model Comparison in MCMC

In frequentist statistics there are many options including:

- Likelihood ratio (deviance) tests
- Wald Tests
- Information Criterion – e.g. AIC/BIC

Here we look at a criterion that can be used with MCMC and which for a linear regression model is equivalent to the AIC – the Deviance information criterion (DIC).

DIC

A natural way to compare models is to use a criterion based on a trade-off between the fit of the data to the model and the corresponding complexity of the model.

DIC does this in a Bayesian way.

DIC = 'goodness of fit' + 'complexity'.

Fit is measured by deviance $D(\theta) = -2\log L(data \mid \theta)$

Complexity is measured by an estimate of the 'effective number of parameters' defined as

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

i.e. Posterior mean deviance minus the deviance evaluated at the posterior mean of the parameters.

DIC (continued)

The DIC is then defined analagously to AIC as

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

Models with smaller DIC are better supported by the data.

- DIC is available in Stat-JR in the ModelResults object.
- DIC can be monitored in other packages such as MLwiN under the Model/MCMC menu and WinBUGS from (Inference/DIC menu).

Deviance Information Criterion

- Diagnostic for model comparison
- Goodness of fit criterion that is penalized for model complexity
- Generalization of the Akaike Information Criterion (AIC; where df is known)
- Used for comparing non-nested models (eg same number but different variables)
- Valuable in MLwiN for testing improved goodness of fit of non-linear model (eg Logit) because Likelihood (and hence Deviance is incorrect)
- Estimated by MCMC sampling; on output get

Bayesian Deviance Information Criterion (DIC)

Dbar	D(thetabar)	pD	DIC
9763.54	9760.51	3.02	9766.56

Dbar: the average deviance from the complete set of iterations

D(thetaBar): the deviance at the expected value of the unknown parameters

pD: the **Estimated** degrees of freedom consumed in the fit, ie Dbar-D(thetaBar)

DIC: Fit + Complexity; Dbar + pD

NB lower values = better parsimonious model

- Somewhat controversial!

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**: 583-640.

Some guidance on DIC

- any decrease in DIC suggests a better model
- But stochastic nature of MCMC; so, with small difference in DIC you should confirm if this is a real difference by checking the results with different seeds and/or starting values.
- More experience with AIC, and common rules of thumb……

- A model with a Δ value within 1-2 of the best model has substantial support in the data, and should be considered along with the best model.
- A Δ value within 4-7 of the best model has considerably less support.
- A Δ value > 10 indicates that the worse model has virtually no support and can be omitted from further consideration.

Example: House price dataset
Sequence of Models (in practical)

- Model 0 Null one-level model; two parameters a fixed mean and a variance
- Model 1b: Null model but districts as 50 fixed effects (49 dummies & constant); separate estimation of each intercept
- Model 1: Null two-level random intercepts model: differential intercepts coming from a distribution
- Model 2b: with size-5 in model, but districts as 50 fixed effects; separate estimation of each intercept
- Model 2: Same basic model but with districts as random effects; differential intercepts coming from a distribution
- Model 3b: Differential district slopes and intercepts as 100 fixed effects; separate estimation of each intercept and slope for each district.
- Model 3: Random slopes model; quadratic variance function at level 2; differential intercepts and slopes coming from a joint multivariate distribution

DIC for Sequence of Models

Model	Nominal DF	Estimated DF	DIC
0: single level	2	2.00	10728.31
1b: 50 district effects	51	50.95	10504.88
1: 2 level Random intercepts	51	44.43	10498.74
2b 1b+ Size	52	51.93	9874.45
2 : 1 + Size	52	45.27	9868.08
3b: 2b + 49 slopes	101	101.16	9843.51
3: + Random Slopes	101	65.24	9807.47

Model 1b: constant + 49 differential intercepts + level 1 variance = 51 effective parameters.

Model 1: districts as random effects; only 44 parameters as these effects are estimated as coming from an overall distribution. The nominal 51 parameters are shrunk due to sharing a prior distribution and therefore do not contribute whole parameters to the parameter count

Model 3: random intercepts & slopes: Lowest DIC; most parsimonious parameterization: most strongly favoured by this approach

SUMMARY & COMPARISON WITH FREQUENTIST APPROACH

Markov chain Monte Carlo (MCMC)

- MCMC methods are Bayesian estimation techniques which can be used to estimate multilevel models
- MCMC works by drawing a random sample of values for each parameter from its probability distribution
- The mean and standard deviation of each random sample gives the point estimate and standard error for that parameter

Estimating a model using MCMC estimation

- We start by specifying the **model** and our **prior** knowledge for each parameter (nearly always no knowledge!)
- Next we specify **initial values** for the model parameters (nearly always the IGLS estimates)
- We then run the MCMC algorithm and obtain the **parameter chains**
- We discard the initial **burn-in** iterations when the chains are settling down (converging to their **posterior** distributions)
- Summary statistics for the remaining **monitoring iterations** are then calculated:
  - Point estimates and standard errors are given by the means and standard deviations of the chains

IGLS vs. MCMC (1)

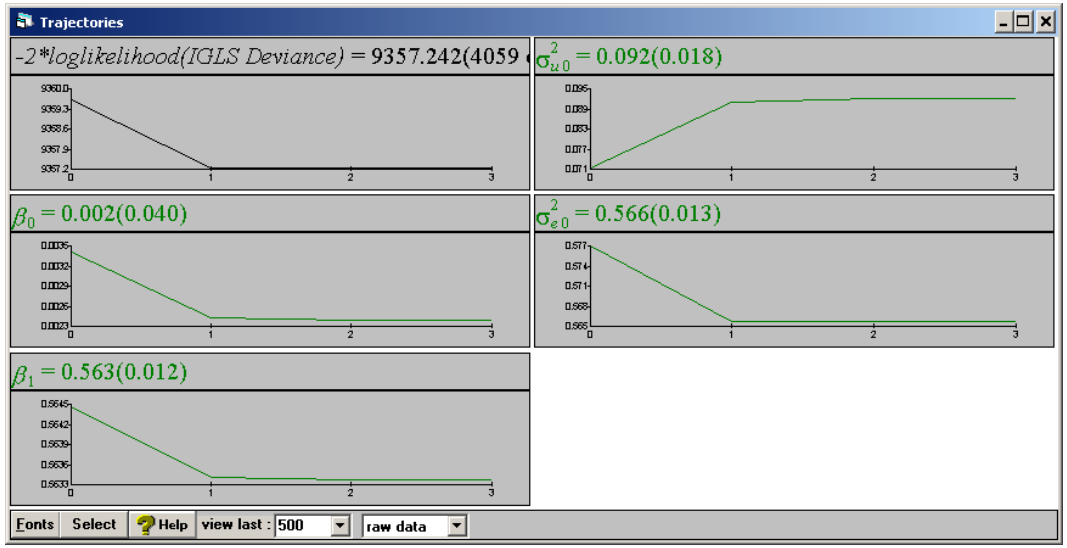
IGLS	MCMC
Fast	Slow
Uses MQL/PQL approximations to fit discrete response models, which can sometimes produce biased estimates	Produces unbiased estimates
Cannot incorporate prior information	Can incorporate prior information

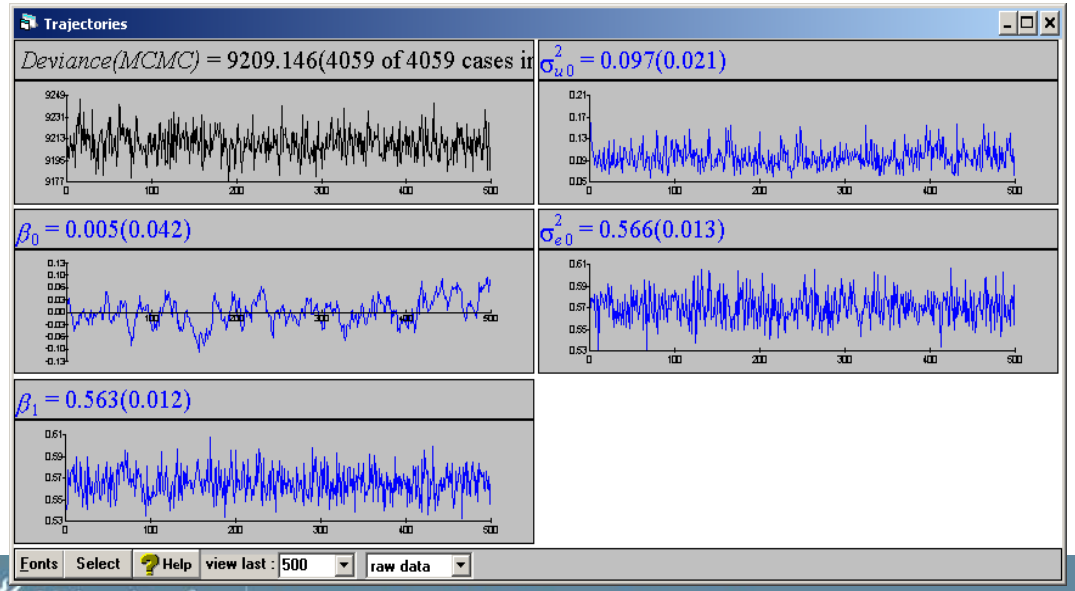
- Note that in practice we often do not incorporate prior information
- We want to protect our inferences from being influenced by our prior beliefs
  - True Bayesians have a very different take

IGLS vs. MCMC (2)

IGLS	MCMC
Confidence intervals based on normality are unreasonable for variance parameters	Normality not assumed
Hard to calculate confidence intervals for functions of parameters	Easy to calculate confidence intervals for arbitrarily complex functions of parameters
Difficult to extend to new models	Easy to extend
Model convergence is judged for you	You have to judge model convergence for yourself

IGLS vs. MCMC (3)





IGLS algorithm converges deterministically to a point

Convergence is therefore judge for you

MCMC algorithm converges stochastically to the equilibrium probability distribution

You have to judge convergence for yourself

Priors

- Our prior knowledge for each parameter is summarised by a probability distribution referred to as the **prior distribution**
  - Typically, we specify that we have no prior knowledge as we like the 'data to speak for it self'
  - We therefore specify **vague**, **diffuse** or **uninformative priors**

$$\beta_1 \sim N(0,10000) \approx U(-\infty,\infty)$$

MCMC samplers

- At the t^{th} iteration we want to sample from the posterior distribution of each parameter in turn
  - If we can write down an analytical expression for the posterior distribution then we can use **Gibbs sampling**
    - Computationally efficient algorithm
    - Continuous response models
  - If we can't write down an analytical expression for the posterior then we use **Metropolis-Hastings** sampling
    - Discrete response models (see later)

Deviance information criterion (DIC) for model comparison

- **DIC** can be viewed as an **AIC** or **BIC** statistic for MCMC
- DIC balances goodness of fit and model complexity (i.e. deviance and number of parameters)
- Want to maximise fit and minimise complexity
  - Lower deviance and fewer parameters
- So “better” models have smaller DIC
- Note that the DIC does not have universal approval!

When to consider using MCMC Estimation

- **Discrete response data** (categorical, counts). PQL often gives quick and accurate estimates but a good idea to check against MCMC, especially if you have small clusters – see later
- If you want to obtain accurate **confidence intervals** for level 2 variances
- Some **complex models** only estimated using MCMC (e.g. multilevel factor analysis)
- Some models can be fitted more easily using MCMC (e.g. cross-classified and multiple membership models)