Chapter 1: Introduction

Jun Yu

Bayes vs frequentist

Bayesian statistics is not simply another statistical tool to be added to ones analytical toolbox. It is a different way of thinking about, and doing, statistical analyses of all types.

Words of Caution

Economists should be aware that Bayesian methods constitute a radically different way of doing science. Bayesian statistics is not just another tool to be added into economists' repertoire of statistical methods. Instead, Bayesians categorically reject various tenets of statistics and the scientific method that are currently widely accepted in economics and other sciences. The Bayesian approach has split the statistics world into warring factions and it is fair to say that the Bayesian approach is growing rapidly in influence.

Very brief history of statistics

- Bayesian conceptual framework was developed by the Reverend Thomas Bayes (1702-1761), and published posthumously in 1764.
- Classical philosophy formalized in early 20th century (Karl Pearson, Ronald Fisher et al.) and quickly became dominant.
- Revival of Bayesian statistics in late 20th century due largely to computational advances (MCMC, WinBUGS software, etc.).

Classical/Frequentist Statistics

- Fixed-effects parameters are fixed and unknown.
- Probabilities are defined to be the long term average under repetition of the experiment.

1. If a balanced coin is tossed many times then, on average, it will be Heads half the time.

2. 95% confidence intervals are constructed so that

they will contain the parameter 95 times out of 100 under repetition of the experiment.

3. No probabilistic interpretation for the particular

experiment performed.

Classical/Frequentist Statistics

Pre-experiment interpretation:

- What is the probability of snow tomorrow?
- What is the probability for an increase in the stock price in next week?

Classical/Frequentist Statistics

Post-experiment interpretation:

- At SMU second year students in economics are used as replicates of the experiment that observes 100 values from a standard normal distribution, N(0,1). They each compute 95% confidence intervals for µ(=0).
- In a large class there will typically be at least one CI that lies entirely below zero, and another that lies entirely above.
- Suppose that I toss a balanced coin, but do not reveal the outcome. What is the probability of Heads?

Bayesian Statistics

- Requires specification of prior knowledge

  -- Experimental observation is used to update the prior knowledge to obtain posterior knowledge.

  -- Uses Bayes formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- In the next chapter:

$$\pi(\theta \mid y) = \frac{f(y, \theta)}{f(y)}$$

Bayesian Statistics

Post-experiment interpretation:

- At SMU second year students in economics are used as replicates of the experiment that observes 100 values from a standard normal distribution, N(0,1). They each compute 95% credible intervals for μ(=0).
- In a large class there will typically be at least one CI that lies entirely below zero, and another that lies entirely above.
- Suppose that I toss a balanced coin, but do not reveal the outcome. What is the probability of Heads?

Example: Future car sale

- **Objective**: Determine the future car sale, B, under various tariff rates and CoE prices

- Assume that we have some data to which an appropriate population dynamics model can be fitted.

Frequentist

- *Frequentist* – might be able to produce a confidence interval for B.

- How is this used to determine the consequences of various tariff and CoE rates?

Bayesian

- The posterior distribution for B incorporates prior information/uncertainty, information provided by the data.

- Moreover, it quantifies the probability of B taking certain values.

Bayesian vs Frequentist

Who’s right?

- Jury still divided after nearly a century.
- Bayesians have moral high ground.

  -- Frequentists rely on the conditionality principle (CP) to define “replication” of the experiment. However, CP implies the likelihood principle (LP). LP is consistent with Bayesian method.

- Frequentists claim objectivity.

  -- Specification of priors is seldom unequivocal and “non-informative” priors can only be rigorously defended in the simplest of cases. Frequentists view priors as subjective.

Classical: some good features

- Regarded as objective.
- Works well in many cases.
- Well developed methodology

1. Unbiasedness of estimators
2. Efficient estimators (e.g., minimum variance)
3. Very general large sample (asymptotic) properties based on maximum likelihood.
4. Model checking.

Classical: some bad features

- Technical problems re conditionality

  -- Numerous examples demonstrate the problems, notwithstanding that these examples are all rather contrived. Asymptotic theory not always available

- Not doing what we want

  1. The null hypothesis is always false, so why bother testing it?

  2. p-values do not quantify how well the data support the hypothesis.

Bayesian: some good features

- Provides probabilistic interpretations.
- Logically consistent method to update prior knowledge and uncertainty, using information contained in the data.
- Uses prior knowledge, which often exists.
- MCMC (Markov Chain Monte Carlo) provides very general purpose software for fitting complex models using off-the-shelf software.

Bayesian: some bad features

- Needs prior knowledge, which is seldom unequivocal.
- “Non-informative” priors are mythical in all but the simplest models.
- “Reference” priors may be complicated and improper.
- Posterior will depend strongly on prior if data poor.
- MCMC is dangerous.
- Techniques for model evaluation, diagnostics, sensitivity etc., are less well developed.

Bayesian vs Frequentist inference

Point Estimation

- Frequentists rely on the well developed theory of minimum variance unbiased estimators (MVUE).

  1. Sample mean and least squares estimators are MVUE for linear models with normal data.

  2. Maximum likelihood estimators are asymptotically MVUE.

- Bayesians typically use posterior mean, which is known as the “Bayes estimator”.

  -- Posterior mean is optimal under squared error loss.

Bayesian vs Frequentist inference

Hypothesis testing

- Frequentists attempt to reject a null hypothesis, H0, using a test which has small probability of falsely rejecting (though H0 is usually known to be false!).

  -- Methodology is based on computing most powerful tests – that is, tests with greatest probability of rejecting H0 when it is false.

- Bayesians specify prior probabilities on two (or more) hypotheses and obtain the posterior probabilities.

  -- Optimal choice is the hypothesis with highest posterior probability.

Bayesian vs Frequentist inference

Interval Estimation

- Frequentists compute confidence intervals with a given “coverage” probability.

  -- “coverage” probability is interpreted as the proportion of such intervals expected to contain unknown parameter under repetition of the experiment.

- Bayesians use intervals of highest posterior density.

  -- The HPD interval is the minimum width interval containing 95% (say) posterior probability.

Chapter 2

Bayes Theorem

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

Simple example

Example: What is the probability that a fair dice rolled a 1, given that it rolled an odd number?

By Bayes theorem, using *A*=1 and *B*=odd, we have

$$P(1|odd)=\frac{P(1\cap odd)}{P(odd)}=\frac{P(1)}{P(odd)}=\frac{\frac{1}{6}}{\frac{1}{2}}=\frac{1}{3}$$

Frequentists can interpret this probability under an experiment where a dice is rolled repeatedly, but even-numbered rolls are discarded.

Bayesian use of Bayes

$P(A|B)=\dfrac{P(A\cap B)}{P(B)}$ appears in the form $\pi(\theta\,|\,y)=\dfrac{f(y,\theta)}{f(y)}$

- θ denotes the unobservable quantities.
  - Often just the parameters, but in some models will include random effects, process errors, predictions etc.
- *y* denotes the data.
- *f(y,θ)* is the joint density of unobservables and data.
- *f(y)* is the (marginal) density of the data.
- *π(θ | y)* is the posterior density of the unobservables given the data.

Priors and likelihoods

The fundamental theorem of probability:

$$f(y,\theta)=\pi(\theta)\,f(y\,|\,\theta)$$

where

- The prior density of θ, is denoted *π(θ).*
- *f(y |θ)* is the model for the data (e.g., regression, ANOVA, binomial, Poisson, etc.). This gives the density function for the data. This is also known as the likelihood function.

Thus, the posterior can be obtained as

$$\pi(\theta\,|\,y)=\frac{f(y,\theta)}{f(y)}=\frac{f(y\,|\,\theta)\pi(\theta)}{f(y)}$$

Priors

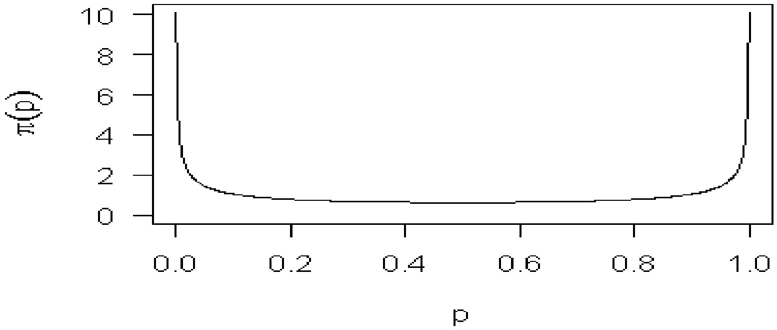
Prior information must be specified for every parameter in the model.

$$\pi(\theta) = \pi(\theta_1, \theta_2, ..., \theta_p)$$

where $\pi(\theta)$ is a joint density function.

Example: Binomial Priors

If we observe Binomial(n,p) data, the "reference" prior $\pi(p)$ is:

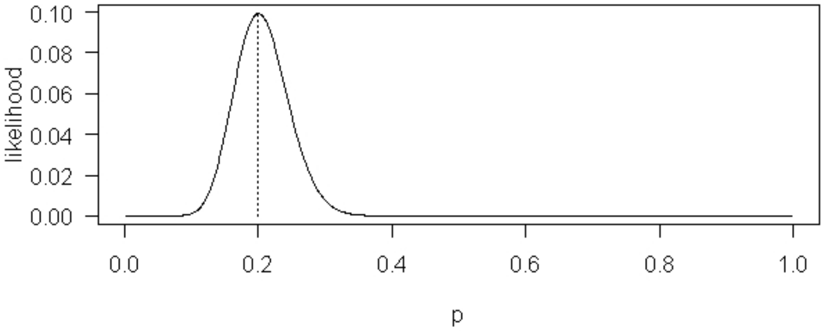


$$\pi(p) = \frac{1}{\pi\sqrt{p(1-p)}}$$

Example: Binomial likelihood

If we observe y=20 successes from 100 trials then the likelihood for *p*=(prob of success) is

$$f(20 \mid p) = \binom{100}{20} p^{20}(1-p)^{80}$$



Model likelihood *f(y)*

Bayes formula: $\pi(\theta \mid y) = \dfrac{f(y,\theta)}{f(y)} = \dfrac{f(y \mid \theta)\pi(\theta)}{f(y)}$

The denominator of Bayes formula, *f(y)*, is often called the model likelihood or marginal likelihood.

- *f(y)* acts solely as a "normalizing constant" (it does not involve θ) and can usually be ignored when working with $\pi(\theta \mid y)$.
- *f(y)* is important for model comparison.

The formula for f(y) is

$$f(y) = \int f(y,\theta)d\theta = \int f(y \mid \theta)\pi(\theta)d\theta$$

Example 1: Binomial data

Observe Y=20 from a Binomial(100,*p*) experiment.

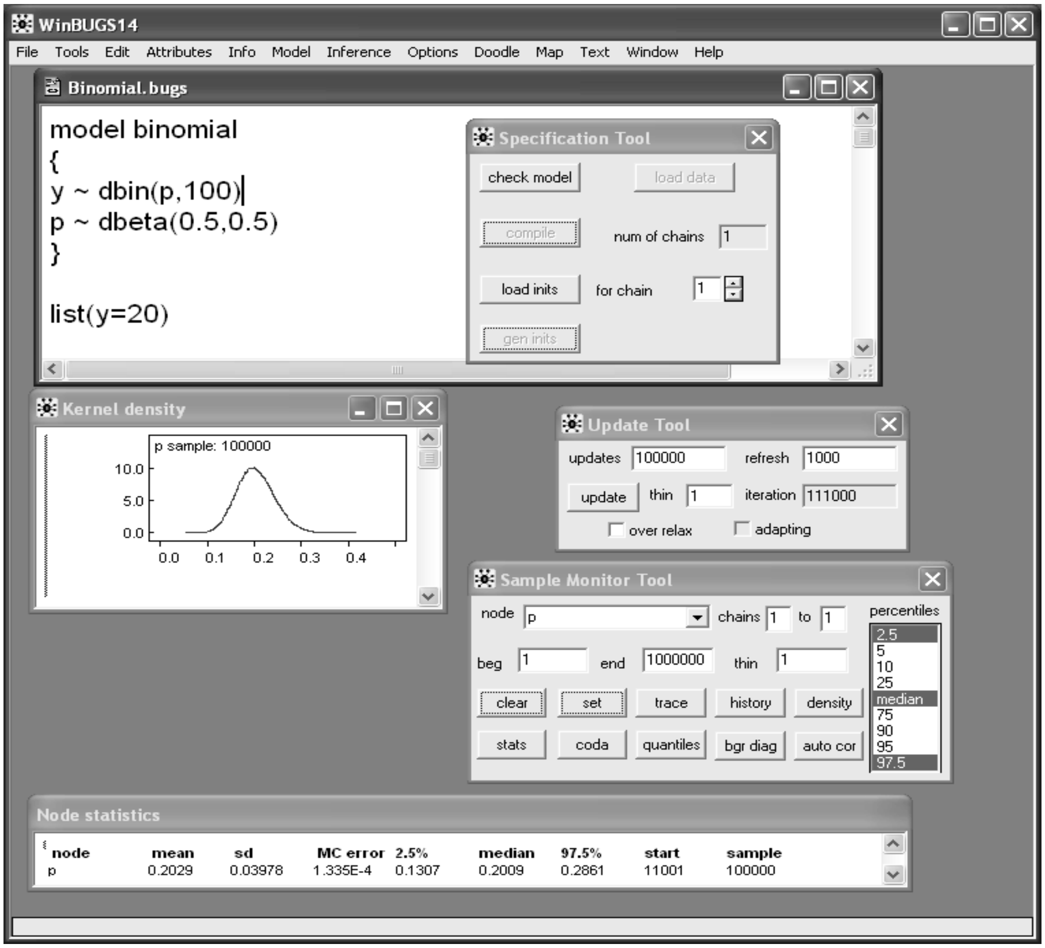
We need to calculate $\pi(p \mid y) = \dfrac{f(y \mid p)\pi(p)}{f(y)}$

where $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$, 0<p<1

and $f(y,p) \propto p^{20}(1-p)^{80}$

So, $\pi(p|y) \propto p^{19.5}(1-p)^{79.5}$, 0<p<1

The trick is to recognize that $\pi(p|y)$ is the density function of a Beta(20.5,80.5) distribution.



```
model binomial
{
y ~ dbin(p,100)
p ~ dbeta(0.5,0.5)
}

list(y=20)
```

num of chains 1

for chain 1

updates 100000 refresh 1000

thin 1 iteration 111000

□ over relax □ adapting

node p chains 1 to 1

beg 1 end 1000000 thin 1

```
node   mean    sd       MC error  2.5%    median  97.5%   start  sample
p      0.2029  0.03978  1.335E-4  0.1307  0.2009  0.2861  11001  100000
```

Example 2: IID Normal data

Here, the data (Y_1, Y_2, Y_3, …, Y_n) are independent and identically distributed as $N(\mu,\sigma^2)$. [This is conditional on μ and σ^2.]

To keep the calculus manageable, it is assumed that σ^2 is known, so that μ is only unknown

The prior on μ is $N(\nu,\phi^2)$.

Example concluded

The formula for $\pi(\mu \mid y)$ corresponds to a normal density. Specifically,

$$\mu \mid y \sim N\left(\nu^*, \phi^{2^*}\right)$$

where

$$\nu^* = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\phi^2}\nu}{\frac{n}{\sigma^2} + \frac{1}{\phi^2}}$$

$$\phi^{2^*} = \left(\frac{n}{\sigma^2} + \frac{1}{\phi^2}\right)^{-1}$$

Note that the posterior mean is a weighted average of the prior mean and sample mean.

IID example in WinBUGS

The model is

$$Y_i \mid \mu, \sigma^2 \sim N(\mu,\sigma^2),\ i=1,\ldots n, \quad \text{and} \quad \mu \sim N(\nu,\phi^2).$$

```
model IIDNormal
{
  for(i in 1:n) { y[i] ~ dnorm(mu,prec.y) }
  prec.y <- 1/sigma2
  mu ~ dnorm(nu,prec.mu)
  prec.mu <- 1/phi2
}
...plus a few details we shall see later
```

The general case

$$\pi(\theta \mid y) = \frac{f(y,\theta)}{f(y)} = \frac{f(y \mid \theta)\pi(\theta)}{f(y)}$$

In general, the calculus required to work with the above formula is formidable (to say the least). In special cases, if the prior is chosen to “match” the likelihood, then the calculus is manageable. These are known as conjugate priors.

Until the advent of MCMC, the computational difficulties were a major disadvantage of Bayesian modeling. Now, it is the other way around, with MCMC permitting the easy fitting of models that may be intractable to frequentist statistics!

Chapter 3

Bayesian Inference

- **Point Estimation**
- **Credible Intervals**
  - Use central intervals, or intervals of highest posterior density.
- **Hypothesis testing**

Point Estimation

Posterior mean:

The “best” estimator (in terms of minimizing squared error loss) of an unknown parameter is simply the expected value of the parameter under its posterior distribution.

The posterior mean is a.k.a. Bayes estimator.

Example: In the IID Normal example, the Bayes estimator of μ is ν^*.

Point Estimation

Posterior mean:

- Easy to obtain from MCMC software.
- The Bayes estimator is not parameterization invariant.

Point Estimation

Posterior mode:

The value of θ that maximizes $\pi(\theta \mid y)$.

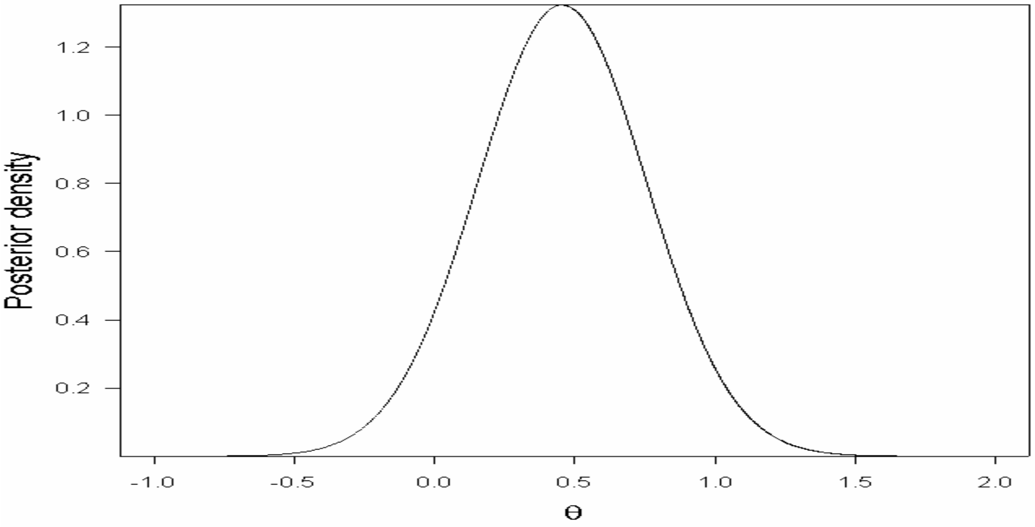
In the IID Normal example, the posterior mode of μ is also ν^*.

.

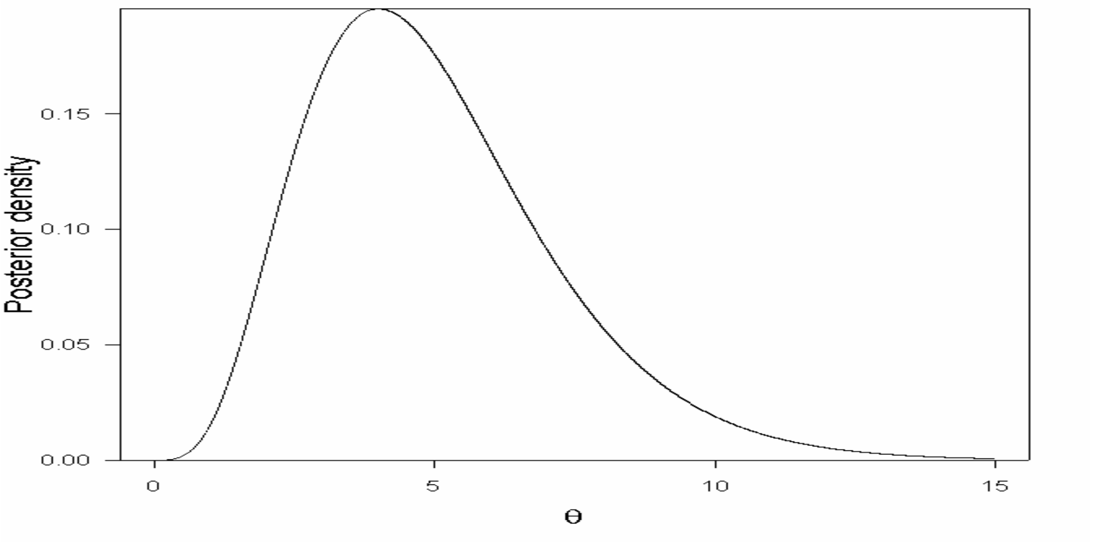
Credible Intervals

Use central intervals or interval of highest posterior density.

HPD Interval (symmetric posterior)



HPD Interval (skewed posterior)



Hypothesis testing

From Gelman et al. (2003, Bayesian Data Analysis, 2nd Ed., p. 250)

“The perspective of this book has little role for the non-Bayesian concept of hypothesis tests…. In order for a Bayesian analysis to yield a nonzero probability for a point null hypothesis, it must begin with a nonzero prior probability for that hypothesis; in the case of a continuous parameter, such a prior distribution (…) usually seems contrived.”

Model Checking

Some frequentist concepts are used in the context of posterior predictive model checking, whereby the observed data are compared to predictive outcomes.

Model comparison/selection

There are Bayesian equivalents of frequentist model selection tools (such as Akaike's Information Criterion, AIC).

The most widely used is the Deviance Information Criterion, DIC.

We'll leave DIC, and other model comparison techniques until later.

Chapter 4

Prior distributions

- Reference priors
  - Jeffreys' rule
  - Improper priors
- Vague priors
- Informative priors
- Hierarchical models
- Sensitivity to prior

Reference priors

- Historically, considerable research effort has focused at obtaining “non-informative” priors (Note: flat priors are not non-informative in general). However, like the holy grail, this much sought after prize has proved extremely elusive. It is now more common to use the terminology “reference prior” to denote a prior that is considered a default prior (for the particular model in question).
- The most widely used method for obtaining a (potential) reference prior is *Jeffreys's rule*.
- Reference priors are frequently *improper*.

Jeffreys' rule

Jeffreys' rule is motivated by the desire that inference should not depend on how a model is parameterized.

Example: if instantaneous mortality is *m*, then the annual survival rate is $s=e^{-m}$. Some modelers might use *m,* while others might use *s.* Inference should not depend on this arbitrary choice of parameterization.

Jeffreys' rule: The reference prior is obtained as the square root of the determinant of the information matrix for the model.

Jeffreys' rule

Jeffreys' rule is widely accepted for single parameter models, but its use is somewhat more controversial, and often subject to modification, in multi-parameter models….and can also be a chore to calculate.

Example:

- The Jeffreys' prior for the mean of normally distributed data is the flat prior, $\pi(\mu)=1$, and for the standard deviation is the inverse prior $\pi(\sigma)=1/\sigma$ (this is equivalent to *log*(σ) being flat).

Jeffreys' rule

In general, the flat prior is the Jeffreys' prior for "location" parameters and the inverse prior is the Jeffreys' prior for "scale" parameters.

The above priors make intuitive sense:

- If one is totally ignorant of a location parameter, then it could take any value on the real line with equal prior probability.
- If totally ignorant about the scale of a parameter, then it is as likely to lie in the interval 1-10 as it is to lie in the interval 10-100. This implies a flat prior on the log scale.

Improper priors

Priors such as $\pi(\mu)=1$, $\pi(\sigma)=1/\sigma$ are *improper* because they do not integrate to 1. That is, the area under the prior density is not unity (and, in fact, is infinity).

In most cases, improper priors can be used in Bayesian analyses without major problems. However, things to watch out for are:

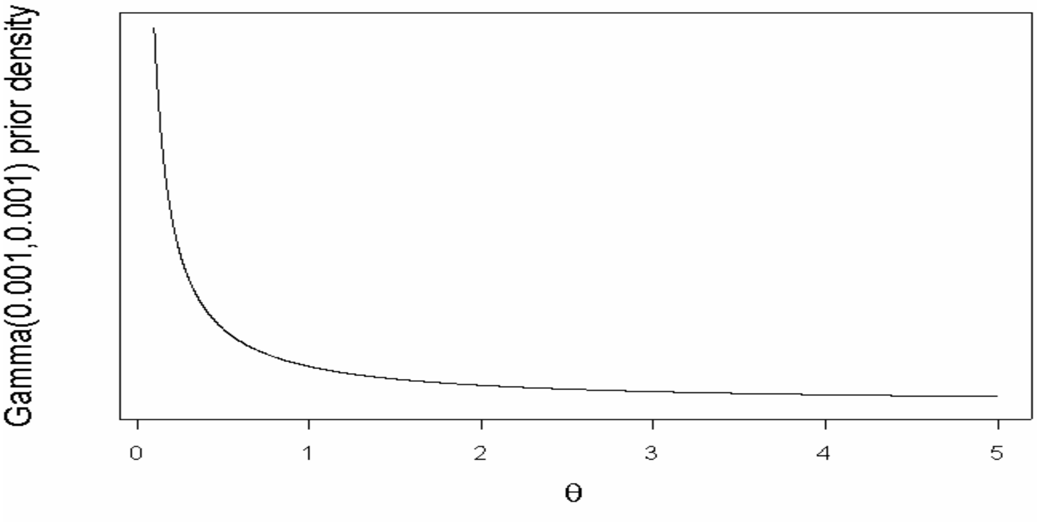
- In a few models, the use of improper priors can result in improper posteriors.
- Use of improper priors makes model selection and hypothesis testing difficult.
- WinBUGS does not allow the use of improper priors.

Vague priors

Essentially, these are densities with high spread, such as a normal density with extremely large variance. These give similar prior value over a large range of parameter values.

- In WinBUGS, the flat prior can be approximated by a vague normal density prior, with mean=0 and variance=1,000,000, say.
- The inverse prior, $\pi(\sigma)$=1/σ, can be approximated by a Gamma density (with very small shape parameter and rate parameters).

Gamma approximating prior



Informative priors

As the name suggests, informative priors convey information concerning prior preference for certain values of the parameters.

Where does this prior information come from?

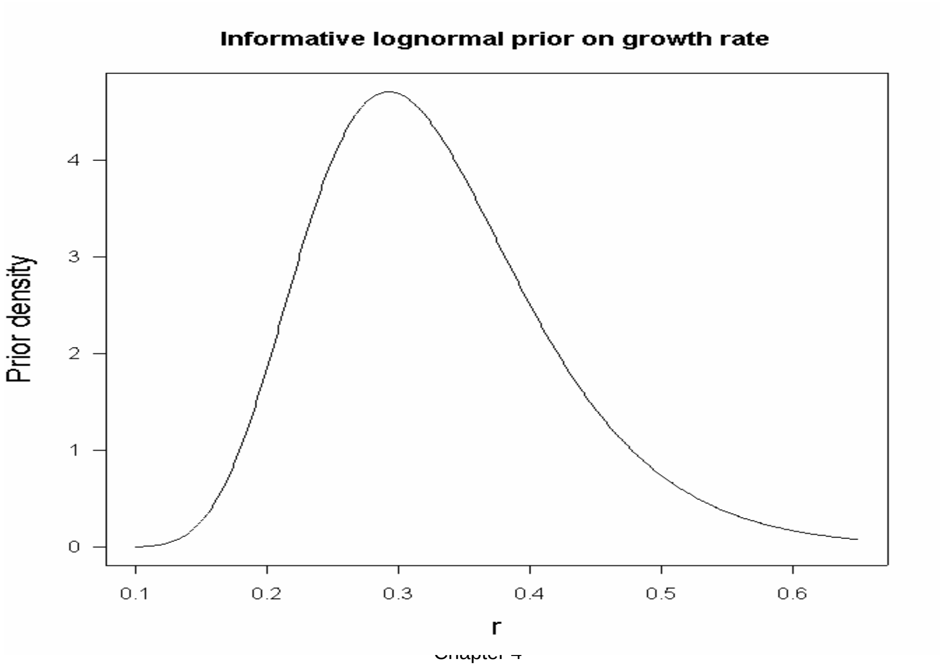
- “Expert” opinion.
- Previous experiments of a similar nature. E.g., in fisheries, prior information about population parameters of a fish stock can be obtained from previous work done on other stocks of the same species.
  - This can often be done formally using meta-analysis or hierarchical Bayesian modeling of the existing data.

Specifying informative priors

Rather than trying to directly specify values for the parameters of a prior density, it is often easier to express probability bounds, from which the parameters can then be obtained.

Example: The “expert” may specify that the growth rate of a population is between 0.2 and 0.5 with 90% prior probability, and be happy with a lognormal shaped prior density. A quick bit of math shows that these quantiles correspond to a lognormal with mean of -1.15 and standard deviation of 0.28 on the log scale.

Specifying informative priors



Hierarchical priors

These are two-stage priors, in the sense that a prior is placed on a prior.

Example: The IID Normal(μ,σ^2) example (with known σ^2) used a N(ν,τ^2) prior on μ. The values of ν and τ^2 are specified after due consideration of the prior information (if any) known about μ.

A hierarchical prior for this example would place priors on the values of ν and τ^2. This prior is known as a hyper-prior, and its parameters are known as hyper-parameters.

Hierarchical priors are more flexible than non-hierarchical priors, and make the posterior less sensitivity to the prior.

Hierarchical models

Hierarchical models use hierarchical priors to perform meta-analyses, whereby a number of related experiments are performed and it is desired to combine information.

The relative ease of implementation, and interpretability, of Bayesian hierarchical models (c.f. frequentist “empirical-Bayes” mixture models) is a major strength of the Bayesian approach.

More on this later.

Sensitivity to prior

After doing a Bayesian analysis, it is never long before the inevitable question is asked – “How would the posterior change if you used a different prior?”

- The most common and straightforward approach is to repeat the analysis using a handful of alternative priors.
- Substantial high-powered theoretical research has investigated the sensitivity of the posterior to “arbitrary” changes to the prior, but this work is currently of little help to the practitioner.
- Simpler results exist for specific changes to the prior.

Simple sensitivity example

Recall the IID Normal example where we saw that the posterior mean was given by

$$\nu^* = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\nu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

From this we can see that the change (derivative) in the posterior mean with respect to the prior mean is

$$\frac{d\nu^*}{d\nu} = \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} = \frac{\tau^{*^2}}{\tau^2}$$

which is just the ratio of posterior variance to prior variance.

Sensitivity more generally

The sensitivity result from the previous slide can be generalized.

For example, it can be shown that, if parameter μ has a N(ν, τ^2) prior then the derivative of the posterior mean of any parameter θ with respect to ν is the posterior covariance of μ and θ, divided by τ^2 .

Chapter 5

Bayesian computation

- The problem of high dimensional integration
- Conjugate priors
- Importance sampling
- Markov Chain Monte Carlo
    - Metropolis-Hastings algorithm
    - Gibbs sampler
    - MCMC diagnostics

High dimensional integration

Recall the formula for the posterior distribution:

$$\pi(\theta \mid y) = \frac{f(y,\theta)}{f(y)} = \frac{f(y \mid \theta)\pi(\theta)}{f(y)}$$

In general θ is a vector of unobserved quantities ("parameters") and in some models its dimension could be in the 100's or 1000's.

Several aspects of Bayesian inference require integration with respect to $\pi(\theta \mid y)$ and θ.

High dimensional integration

Doing model comparison requires calculation of model likelihoods, *f(y)* .

$$f(y) = \int f(y,\theta)d\theta = \int f(y\,|\,\theta)\pi(\theta)d\theta$$

A bit of calculus shows that *f(y)* can be expressed as an integral with respect to the posterior density,

$$f(y) = \left(\int \frac{1}{f(y\,|\,\theta)} \pi(\theta\,|\,y)d\theta \right)^{-1}$$

High dimensional integration

Point estimation and interval calculation require the (marginal) density of individual elements of θ, which again requires integration with respect to the posterior density.

For example, the Bayes estimator of θ_i (the ith element of θ) is

$$\begin{aligned} E(\theta_i\,|\,y) = \int \theta_i \pi(\theta\,|\,y)d\theta &\equiv \int \theta_i \pi(\theta\,|\,y)d\theta_1 ... d\theta_p \\ &= \int \theta_i \pi(\theta_i\,|\,y)d\theta_i \end{aligned}$$

where

$$\pi(\theta_i\,|\,y) = \int \pi(\theta\,|\,y)d\theta_1 d\theta_2 ... d\theta_{i-1} d\theta_{i+1} ... d\theta_p$$

Conjugate priors

In simple models the integration problem can be avoided by choosing a particular type of prior.

- the prior density of θ, $\pi(\theta)$, is conjugate if $\pi(\theta \mid y)$ will belong to the same "statistical family".
  - This was the case with the IID Normal example where both prior and posterior were normally distributed.

Sampling the posterior

Instead of trying to tackle the integration problem – simulate a whopping great sample (e.g., 10,000's) from the (joint) posterior $\pi(\theta \mid y)$.

- The Bayes estimator of a parameter can then be approximated (to arbitrary precision) from its average over the samples from the posterior.
- Intervals can be obtained from drawing histograms of the sample values.
- Model likelihoods can be approximated (to arbitrary precision) by the inverse of the average of $1/ f(y \mid \theta)$ over the samples from the posterior.

Sampling the posterior

Two widely used approaches for sampling from $\pi(\theta \mid y)$:

Importance sampling:

- Simulate values from a simple density that is similar to $\pi(\theta \mid y)$, and do an adjustment (re-weighting).

Markov chain Monte Carlo (MCMC)

- Simulate from a Markov chain which has $\pi(\theta \mid y)$ as its equilibrium distribution.

Importance sampling

Example: The Bayes estimator of θ_i can be written

$$E(\theta_i \mid y) = \int \theta_i \pi(\theta \mid y) d\theta = \int \frac{\theta_i \pi(\theta \mid y)}{p(\theta)} p(\theta) d\theta$$

where $p(\theta)$ is any density that is easy to sample from.

The statistical interpretation of the above formula is that the Bayes estimator can be obtained by sampling (tens of thousands of) of θ values from the distribution with density function $p(\theta)$, and calculating the average of

$$\frac{\theta_i \pi(\theta \mid y)}{p(\theta)}$$

Markov chain Monte Carlo

A Markov chain refers to a random process where the value at time (or iteration) *t* depends on the value at time *t-1* but not on any earlier values.

Methods such as the Metropolis-Hastings algorithm implement a Markov chain to generate a (very long) sequence of random θ values. The algorithm is constructed such that the generated values come from the desired "target" distribution, which in the Bayesian context is $\pi(\theta \mid y)$.

Metropolis-Hastings algorithm

If we have just generated value $\theta^{(k)}$, the Metropolis-Hastings algorithm proceeds by using a simple "proposal density" $p(\theta,\theta^{(k)})$ to suggest that the next value be θ^*. This suggested value is accepted with probability

$$\Pr(set\ \ \theta^{(k+1)} = \theta^*) = \min\left(1, \frac{p(\theta^{(k)}, \theta^*)\pi(\theta^* \mid y)}{p(\theta^*, \theta^{(k)})\pi(\theta^{(k)} \mid y)}\right)$$

If θ^* is not accepted then we "stay put", and set $\theta^{(k+1)} = \theta^{(k)}$.

Gibbs sampler

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm in which (the vector) $\theta^{(k+1)}$ is obtained from $\theta^{(k)}$ by updating the vector elements one at a time.

WinBUGS is the Windows based version of the **B**ayesian analysis **U**sing **G**ibbs **S**ampler software.

MCMC software

The BUGS software (including WinBUGS) undoubtedly provides the easiest and most widely used implementation of MCMC. It is reasonably general, but there are some classes of models that it can not cope with.

Other software includes routines written in a variety of languages (e.g., C++). The Automatic Differentiation Modeler Software (ADMB) is more powerful and general than BUGS, but is non-free and is far harder to use. (ADMB can also do pure maximum likelihood analysis, and hence permits both Frequentist and Bayesian modeling.)

MCMC properties

- The MCMC algorithm needs to “burn in” – to become independent of the starting value.
  - Usually enough to discard the first few thousand values.

- Successive values may be highly correlated. That is, the sequence of values from the chain are not generally independent.

- Badly behaved posterior densities (e.g., bi-modal) may result in mixing problems whereby some parts of $\pi(\theta \mid y)$ are not properly sampled.

MCMC diagnostics

Several software packages are available for checking the MCMC output.

- WinBUGS provides exploratory diagnostic tools and windows for seeing trace plots and histograms of the sampled values.

- CODA (Convergence Diagnostics Analysis) and BOA (Bayesian Output Analysis) packages are available for R and Splus on a variety of platforms. These perform more sophisticated checks.
  - WinBUGS provides the option to save the MCMC output in the form required by CODA.

Chapter 6

The Gibbs Sampler

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm. It generates a sample from an arbitrarily complex multidimensional distribution by sampling from each of the univariate full conditional distributions in turn.

Bivariate example

To keep things simple, consider sampling from an arbitrary two dimensional distribution p(θ_1, θ_2). (In our context, this will be the posterior distribution for a Bayesian model with two parameters, π(θ_1, θ_2 | *y*).)

It is assumed that the two univariate conditional densities p(θ_1 | θ_2) and p(θ_2 | θ_1) can be sampled from.

Bivariate example

The Gibbs sampler argument proceeds as follows –

If we are able to generate a value θ_1 from the marginal density p(θ_1) , then a value θ_2 sampled from the density p(θ_2 | θ_1) will have density function p(θ_2) and the pair of values (θ_1, θ_2) has joint density p(θ_1, θ_2).

Similarly,

If we are able to generate a value θ_2 from the marginal density p(θ_2) , then a value θ_1 sampled from the density p(θ_1 | θ_2) will have density function p(θ_1) and the pair of values (θ_1, θ_2) has joint density p(θ_1, θ_2). .

Bivariate example

If ($\theta_1^{(k)}$, $\theta_2^{(k)}$) is the sampled point at iteration k then the Gibbs sampler obtains the next point, ($\theta_1^{(k+1)}$, $\theta_2^{(k+1)}$) by generating $\theta_1^{(k+1)}$ from the density p(θ_1 | $\theta_2^{(k)}$) and $\theta_2^{(k+1)}$ from the density p(θ_2 | $\theta_1^{(k+1)}$) .

The “burn in” period is required so that the If ‘s on the previous page are satisfied.

Higher dimensional parameters

The same idea works in p (>2) dimensions, but each univariate density sampled from is a “full conditional” density, whereby all other parameters are conditioned upon.

For example, the value of $\theta_1^{(k+1)}$ is obtained by sampling from the univariate density

$$p(\theta_1 \mid \theta_2^{(k)}, \theta_3^{(k)}, ..., \theta_p^{(k)})$$

and then the value of $\theta_2^{(k+1)}$ is obtained by sampling from the univariate density

$$p(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^{(k)}, ..., \theta_p^{(k)})$$

…and so on.

WinBUGS implementation

In the Bayesian context, the univariate full conditional densities that WinBUGS is working with are

$$\pi(\theta_1 \mid \theta_2^{(k)}, \theta_3^{(k)}, ..., \theta_p^{(k)}, y)$$

These are, for a very large class of Bayesian models, relatively straightforward to deduce. This is essentially what WinBUGS is doing during the compilation phase.

WinBUGS implementation

In particular, for models that can be drawn by doodles, the full conditional density for any stochastic node is a function of only the parent and offspring nodes. To see this, note that Bayes formula says

$$\pi(\theta_1 \mid \theta_2, \theta_3, ..., \theta_p, y) = \frac{\pi(\theta_1, \theta_2, \theta_3, ..., \theta_p, y)}{\pi(\theta_2, \theta_3, ..., \theta_p, y)}$$

The above full conditional density is a function of only θ_1 because all other values are treated as constants, and so it is only terms in the joint density involving θ_1 that are of relevance to the full conditional.

Chapter 7

Introduction to WinBUGS

WinBUGS is the Windows version of the **B**ayesian analysis **U**sing the **G**ibbs **S**ampler software developed by the UK Medical Research Council and the Imperial College of Science, Technology and Medicine, London.

At the time of writing, WinBUGS is freely available at

`http://www.mrc-bsu.cam.ac.uk/bugs`

BUGS Background

The BUGS project began in 1989 in the Biostatistics Unit of the Medical Research Council, U.K.

- “Classic” BUGS
  - Thomas et al. (1992).
  - Batch mode operation.
- WinBUGS
  - Developed by MRC for Windows operating system in late 1990’s.
  - Current version (May 2006) is 1.4, with upgrade to version 1.4.1.
- OpenBUGS
  - Open source version of WinBUGS for running on Windows and Linux, as well as inside the R statistical package (BRugs).

BUGS and R

- CODA
  - R package for convergence diagnosis (Best et al., 1995).
  - Other similar R packages available, e.g., BOA.
- R2WinBUGS
  - R package for executing WinBUGS from R (Sturtz et al., 2005).
  - Uses the WinBUGS scripting language (the WinBUGS interface appears on the desktop).
- BRugs
  - R package for running OpenBUGs components from within R.

WinBUGS Structure

Essentially, the WinBUGS program is simply a syntactical representation of the model, in which the distributional form of the data and parameters are specified. (It does not require (in most cases) knowing the formulae of density functions).

For example, $y \sim Binomial(n, p)$

is written in WinBUGS as

```
y ~ dbin(p,n)
```

and $y_i \sim Binomial(n_i, p_i)\,,\;\; i = 1,...M$

is written in WinBUGS as

```
for(i in 1:M) {y[i] ~ dbin(p[i],n[i])}
```

WinBUGS Structure

NOTE: The interpretation of WinBUGS code is unlike that of other programming languages such as R.

In R: $y = y{+}1$ makes perfect sense.

In WinBUGS: y <- $y{+}1$ is nonsensical, because a datum (or parameter) can not equal itself plus unity.

If you can write the model down on paper, then you should be able to code it up in WinBUGS.

CAUTION: There is no guarantee that WinBUGS will "work".

WinBUGS: Practical 1

- Start up WinBUGS.
- Click File->New to open up a BUGS window.
- Type in the code:

```
model NormalPrior
{
      mu ~ dnorm(0,1)
}
```

- Click Model->Specification
  - Click "check model". Bottom left of screen should say "Model is syntactically correct (else it will provide an error message and the cursor will be positioned at the error.)
  - This example contains no data (so ignore the "load data" step, for now).
  - Click "compile".
  - Click "gen inits".
- WinBUGS is now ready to generate the MCMC sample.

WinBUGS: Practical 1

- Click Model->Update to open the Update Tool window.
  - Click “update”.
  - You’ve just generated 1000 samples from a Markov chain with a standard normal stationary distribution!
  - These first thousand samples have not been saved, which is good practice because the chain needs to burn in.
- Click Inference->Samples to start the Sample Monitor.
  - Type “mu” in the node box and click on “set”.
- Go to Update Tool and click “update”.
- Go to Sample Monitor Tool
  - Several choices of summary plots and statistics can now be selected.

WinBUGS: Practical 2

Repeat Practical 1, but with a non-normal distribution.

To see the choices of distribution:

- Click Help > User Manual.
- Scroll down and click on Contents.
- Follow the links:

  Model Specification > The BUGS language: stochastic nodes > Distributions

WinBUGS: Practical 3

The previous two examples did not include any data.

Here, we will assume that we observe a single observation $y \sim \mathrm{N}(\mu,1)$, where μ has a standard normal prior distribution. (In this case $\mu|y \sim \mathrm{N}(0.5y,0.5)$)

```
model NormalPrior
{
  mu ~ dnorm(0,1)
  y ~ dnorm(mu,1)
}
list(y=  )
```

Insert a number here

After checking the model syntax, use the mouse to highlight the word “list”, and click “load data” on the Specification Tool window. Then, proceed as before.

WinBUGS: Practical 4

Recall, the IID Normal example with known variance of the data:

```
model IIDNormal
{
  for(i in 1:10) { y[i] ~ dnorm(mu,1) }
  mu ~ dnorm(0,1)
}
list(y=c(1.64,1.10,1.33,0.27,0.61,
         0.25,-0.02,-0.08,0.43,-0.53 ))
```

WinBUGS example

More generally, in the case of ten (say), iid observations:

```
model IIDNormal
{
  for(i in 1:10) { y[i] ~ dnorm(mu,1) }
  mu ~ dnorm(0,1)
}
list(y=c(1.64,1.10,1.33,0.27,0.61,
         0.25,-0.02,-0.08,0.43,-0.53 ))
```

Moving along…

Now, let’s drop the assumption of known variance, and instead we shall assume that the data are IID Normal with unknown mean μ and unknown variance σ^2.

If an informative prior on σ^2 is to be specified then an inverse gamma[*] distribution will typically be used. This corresponds to a gamma distribution on 1/ σ^2 .

If a “non-informative” prior on σ^2 is desired then it can be approximated by specifying a highly dispersed gamma distribution on 1/ σ^2 .

[*] The gamma distribution is a generalization of the χ^2 .

Moving along…

Note: For normal densities, Bayesian's typically work with 1/ σ^2 (precision) rather than σ^2 (variance). In WinBUGS, the normal density is specified as dnorm(mean,precision).

IID Normal, μ and σ^2 unknown

```
model IIDNormal2
{
  for(i in 1:10) { y[i] ~ dnorm(mu,prec) }
  var <- 1/prec
#Add priors
  mu ~ dnorm(0,1)
  prec ~ dgamma(0.001,0.001)  #Disperse gamma
}
#Data
list( y=c( 1.64,1.10,1.33,0.27,0.61,0.25,
          -0.02,-0.08,0.43,-0.53 )  )
#Inits
list( mu=0, prec=1)
```

WinBUGS syntax

In WinBUGS, the tilde sign ~ means “distributed as”. It is used to:

- Specify the distribution of the data.
- Specify the prior distributions.
- Values to the left of a ~ are called “stochastic”.

The left arrow <- corresponds to the “equals” sign. It is used in calculations, such as **`var<-1/prec`**

- Values to the left of a <- are called “logical”.

Linear regression: Lines example

```
model
{
  for(i in 1:N){
     Y[i] ~ dnorm(mu[i], tau)
     mu[i] <- alpha + beta*(x[i] - mean(x[]))
               }
  sigma <- 1/sqrt(tau)
  alpha ~ dnorm(0, 1.0E-6)
  beta ~ dnorm(0, 1.0E-6)
  tau ~ dgamma(1.0E-3, 1.0E-3)
}
list(x=c(1,2,3,4,5), Y=c(1,3,3,3,5), N=5 )
list(alpha = 0, beta = 0, tau = 1)
```

Chapter 8:
Bayesian Diagnostics

Jun Yu

Bayesian Diagnostics

- Convergence diagnostics.
- Posterior predictive checks.
- DIC, model selection, and complexity.
- Bayes factors
- Sensitivity analysis

Convergence diagnostics

- Primarily, to assess whether the MCMC chain has converged to a stationary distribution.
- We will use the CODA package in R. This implements diagnostic, based on multiple chains.
  1. Geweke diagnostic for stationarity.
  2. Heidelberger-Welch stationarity and run-length diagnostics.
  3. Raftery-Lewis run-length diagnostic (for quantiles).
- The CODA package also contains:
  1. Diagnostic plots
  2. Functions for manipulating BUGS output
  3. A function to find HPD intervals

A Stochastic Volatility Model

- Observation equation:

$$y_t = \exp(0.5h_t)\varepsilon_t, \varepsilon_t \sim iidN(0,1)$$

- State equation:

$$h_t = \mu + \phi(h_{t-1} - \mu) + \eta_t, \eta_t \sim iidN(0, \sigma_\eta^2)$$

Priors

$$\mu \sim N(0,10)$$

$$\phi^{*} \sim Beta(20,1.5)$$

$$\sigma_{\eta}^{2} \sim 1/\Gamma(2.5,0.025)$$

$$\phi^{*} = 0.5(\phi+1)$$

A SV model – WinBugs Code

```
{
#specify the observation equation

for (i in 1:N) {
   Ymean[i] <- exp(0.5*theta[i]);
   Yisigma2[i] <- 1/(Ymean[i]*Ymean[i]);
   Y[i]~ dnorm(0,Yisigma2[i]);
   }
```

A SV model -- Cont

```
#specify the state equation

theta0 ~ dnorm(mu,itau2);
thetamean[1] <- mu + phi*(theta0-mu);
theta[1] ~ dnorm(thetamean[1], itau2);
for (i in 2:N) {
    thetamean[i] <- mu + phi*(theta[i-1]-mu);
    theta[i] ~ dnorm(thetamean[i], itau2);
   }
```

A SV model -- Cont

```
#specify the priors

phistar ~ dbeta(20,1.5);
phi <- 2*phistar -1;
mu ~ dnorm(-10,0.04);
itau2 ~ dgamma(2.5,0.025);
tau <- sqrt(1/itau2);
}
```

A SV model -- Cont

```
#specify the sample size, initial value & data

list(N=1512)
list(phistar=0.975, mu=-10, itau2=50)
```

DIC, model selection, & complexity.

- Spiegelhalter et al. (2002) formalized the concept of the deviance information criterion, DIC, as a measure of model fit and complexity.
- The deviance is -2 times the log-likelihood

$$D(\theta) = -2\log(f(y \mid \theta))$$

- Define “Dbar” as the posterior mean of the deviance

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

- and “Dhat” as the deviance evaluated at some plug-in estimate of theta, typically the posterior mean of theta

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

DIC, model selection, & complexity.

- Using somewhat heuristic arguments, Spiegelhalter et al. (2002) argued that

$$p_D = \overline{D(\theta)} - D(\bar{\theta})$$

quantifies the “effective number of parameters” in the model

- Model goodness of fit is a trade-off between model fit and model complexity. The DIC is defined to be

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

- or equivalenly

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

DIC, model selection, & complexity.

- Spiegelhalter et al. (2002) developed DIC in the context of hierarchical models, but it has since been applied much more widely.

Bayes factors

- To choose between two models, the Bayes factor is the ratio of the marginal densities for the data under the two models

$$B_{12} = f_1(y) / f_2(y)$$

- where

$$f_i(y) = \int f_i(y \mid \theta_i)\pi(\theta_i)d\theta_i$$