Introduction to MCMC and Bayesian Hierarchical Models

Leonhard Knorr-Held
and
Günter Rasser

Imperial College London, UK and
University of Munich, Germany

(with many thanks to Clare Marshall, Sylvia Richardson, Nicky Best, Andrew Thomas, Peter Green and David Spiegelhalter.)

Tutorial:
Biometrisches Kolloqium, Homburg (Saar)
March 19 2001

Outline

1. Introduction to Bayesian Inference
2. Markov chain Monte Carlo
3. Bayesian Hierarchical Models

1. Introduction to Bayesian Inference

Introduction

Bayesians and frequentists differ in their concept of probability

- In Bayesian statistics, probability is used as a fundamental measure of uncertainty
  e.g. probability that it will rain tomorrow
  e.g. probability that an unknown quantity lies within a specified range

  $\rightarrow$ subjective probability
  - ! subjective probabilities should be "coherent", i.e. obey the law of probability
  - ! they should be constructed with scientific judgement
- Frequentists think of probabilities as frequencies observed in a long run of repeated experiments

Bayesian inference

Distinguish fundamentally:

- Observable quantities X, i.e. the data
- Unknown quantities θ

  These can be statistical parameters, missing data, mismeasured data ...
  $\rightarrow$ parameters are treated as random variables

  $\rightarrow$ 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

Components of Bayesian inference

- The prior distribution $p(\theta)$

  Expresses uncertainty or information available at the start of the study about unknown quantities (variables) by means of a probability distribution

- The likelihood $p(X|\theta)$

  Relate all variables into a 'full probability model' that summarises current knowledge on the random phenomenon

- The posterior distribution $p(\theta|X)$

  After observation of some variables (the data), use Bayes theorem to obtain conditional probability distributions for unobserved quantities of interest
  Expresses our uncertainty about θ after seeing the data
  Bayes theorem tells us how to calculate this

Bayes theorem

Provable from axioms of probability

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

If A_i is a set of mutually exclusive and exhaustive events (*i.e.* $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$), then

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)}.$$

For general inference, we use the expression:

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

Summarising posterior distributions

Given a posterior distribution $p(\theta|X)$, we may want to use different types of posterior summaries

- mean, standard deviations, medians, quantiles etc
- posterior probability of exceeding certain thresholds, $p(\theta > \theta_0|X)$
- credibility intervals

! Note that these have a direct probabilistic interpretation:
e.g. posterior probability that a relative risk (RR) lies between 1.13 and 1.97 of 0.95

They are different from classical confidence intervals:

95% of the 95% confidence intervals would contain the true value of RR in a long run of repeated experiments

Some examples

Binomial response - continuous prior

Data: r successes from n independent trials

Likelihood:

$$\binom{n}{r}\theta^r(1-\theta)^{n-r} \quad \propto \quad \theta^r(1-\theta)^{n-r}$$

Prior:

$$\begin{aligned}\theta &\sim \text{Beta}(a,b)\\ &\equiv \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1}\\ &\propto \theta^{a-1}(1-\theta)^{b-1}\end{aligned}$$

For a beta distribution,

$$\begin{aligned}\text{mean } m &= a/(a+b),\\ \text{variance } s^2 &= m(1-m)/(a+b+1)\end{aligned}$$

For $a=b=1$, corresponds to the Uniform[0,1]

Posterior:

$$\begin{aligned}p(\theta \mid r,n) &\propto p(r \mid \theta,n)p(\theta)\\ &\equiv \theta^r(1-\theta)^{n-r}\theta^{a-1}(1-\theta)^{b-1}\\ &\propto \theta^{r+a-1}(1-\theta)^{n-r+b-1}\\ &\propto \text{Beta}(r+a,\ n-r+b)\end{aligned}$$

Hence the posterior mean:

$$E(\theta|r,n) = (r+a)/(n+a+b)$$

a and b are equivalent to observing a priori a successes in $a+b$ trials $\rightarrow$ can be elicited

With fixed a and b, when n and $n-r$ grow, $E(\theta|r,n) \rightarrow r/n$ the maximum likelihood estimator, and the variance tends to zero

This is a general phenomenon, as n grows, the posterior distribution gets more concentrated and the likelihood dominates the prior

Normal response - known variance

Suppose that we have one observation:

$$x \sim \text{N}(\mu,\sigma^2)$$

where μ is unknown. For simplicity, we assume σ^2 is known.

Suppose we have a Normal prior distribution:

$$\mu \sim \text{N}(\nu,\tau^2)$$

where ν and τ^2 are fixed.

Then the posterior distribution for μ is

$$\begin{aligned}p(\mu \mid \mathbf{x}) &\propto p(\mu)\,p(\mathbf{x} \mid \mu)\\ &\propto \exp\left[-\frac{1}{2}\left\{\frac{(\mu-\nu)^2}{\tau^2}+\frac{(x-\mu)^2}{\sigma^2}\right\}\right]\end{aligned}$$

The part inside $\{\ldots\}$

$$\begin{aligned}&= \mu^2\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right) - 2\mu\left(\frac{\nu}{\tau^2}+\frac{x}{\sigma^2}\right) + \text{const}\\ &= \left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\mu - \frac{\frac{\nu}{\tau^2}+\frac{x}{\sigma^2}}{\frac{1}{\tau^2}+\frac{1}{\sigma^2}}\right)^2 + \text{const}\end{aligned}$$

So

$$p(\mu \mid x) \sim \text{N}(\mu_1,\tau_1^2)$$

with

$$\mu_1 = \frac{\frac{\nu}{\tau^2}+\frac{x}{\sigma^2}}{\frac{1}{\tau^2}+\frac{1}{\sigma^2}}$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau^2}+\frac{1}{\sigma^2}$$

Note that there are 2 equivalent expressions for the posterior mean μ_1:

$$\mu_1 = \nu + (x-\nu)\frac{\tau^2}{\sigma^2+\tau^2}$$

$$\mu_1 = x - (x-\nu)\frac{\sigma^2}{\sigma^2+\tau^2}$$

that have different interpretations: in the first, the posterior mean is the prior mean ν adjusted towards the observed data x, in the second, the data is "shrunk" towards the prior mean.

Both formulation show the compromise between prior mean and observed value, with weights proportional to the precisions (the inverse variances)

Predictive distributions

For future data Z, we may be interested in predictive distribution $p(Z|x)$, given by

$$p(Z|X) = \int p(Z|X,\theta)\ p(\theta|X)\ d\theta,$$

which generally simplifies to

$$p(Z|X) = \int p(Z|\theta)\ p(\theta|X)\ d\theta.$$

Normal case (one observation)

$$p(z|x) \propto \int \exp\left(-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right)\left(\exp -\frac{1}{2}\frac{(\mu-\mu_1)^2}{\tau_1^2}\right) d\mu$$

$\Rightarrow$ the predictive distribution of z is still a normal with mean

$$E(z|x) = E(E(z|x,\mu)|x) = E(\mu|x) = \mu_1$$

and variance

$$\begin{aligned}\mathrm{var}(z|x) &= E(\mathrm{var}(z|x,\mu)|x) + \mathrm{var}(E(z|x,\mu)|x)\\ &= E(\sigma^2|x) + \mathrm{var}(\mu|x) = \sigma^2 + \tau_1^2\end{aligned}$$

So our predictive interval is centred at the posterior mean of μ and the width of the interval depends on both the uncertainty in the estimation of μ **and** the measurement error.

Normal case, multiple observations

Suppose that we have a random sample, $x_1, \ldots, x_n$. Then $\bar{x} = \sum x_i/n$ summarises all the information in the sample on μ, with $p(\bar{x}|\mu) \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$

Similarly we can deduce:

$$p(\mu \mid \bar{x}) \sim \mathrm{N}(\mu_n, \tau_n^2)$$

with

$$\mu_n = \frac{\frac{\nu}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}},\ \frac{1}{\tau_n^2} = \frac{1}{\tau^2} + \frac{n}{\sigma^2}$$

This can be equivalently written:

$$\mu_n = w\nu + (1-w)\bar{x}, \quad \tau_n^2 = \frac{\sigma^2}{n}(1-w)$$

where

$$w = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}.$$

As $n \to \infty$,

$$w \to 0,\ \mu_n \to \bar{x},\ \tau_n^2 \to \sigma^2/n \ (\text{ M.L. estimates})$$

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

which does not depend on the prior.

In the frequentist setting, the MLE $\hat{\mu} = \bar{x}$ and

$$p(\hat{\mu} \mid \mu) = p(\bar{x}|\mu) = \mathrm{N}(\mu, \frac{\sigma^2}{n}),$$

whereas in the Bayesian framework, the "dual statement" is made:

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

In summary, on these examples, we see

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of the prior s.d. and the s.e.(MLE)

and 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.

These observations are generally true, when the MLE exists and is unique.

Priors

When the posterior is in the same family as the prior then we have what is known as *conjugacy*. This has the advantage that prior parameters can usually be interpreted as a *prior sample*.

Examples include:

Prior	Likelihood	Posterior
Normal	Normal (μ unknown)	Normal
Inv. Gamma	Normal (σ^2 unknown)	Inv. Gamma
Beta	Binomial	Beta
Gamma	Poisson	Gamma

Unfortunately conjugate priors do not exist for all likelihoods, and could be restrictive.

- Non-conjugate priors can also be used
  - Maths is harder
  - We will return to these in the hierarchical modelling sections

Where does the prior come from ?

- it is in principle subjective
- it might be elicited from experts
- it might be more convincing to be based on historical data, *e.g.* a previous clinical trial. But the assumed relevance is still a subjective judgement.
- there are various "objective" approaches.

Uniform priors

Set $p(\theta) \propto 1$

- This is typically improper ($\int p(\theta)d\theta \neq 1$).
- The posterior will still usually be proper.
- Inference is based on the likelihood $p(x \mid \theta)$.
- It is not really objective, since a flat prior $p(\theta) \propto 1$ on θ does not correspond to a flat prior on $\phi = g(\theta)$, but to $p(\phi) \propto \mid g'(\theta) \mid^{-1}$

'Non-informative' priors

Also known as *vague, ignorance, flat, diffuse* and so on

- Usually there exists a prior that reproduces the classical results
- Often improper
- Can be quite unrealistic
- However, always useful to report likelihood-based results as arising from a 'reference' prior.

Jeffrey's priors

There have been many suggestions for specifying non-informative priors, a classical one being that of Jeffreys (1961).

Jeffrey's choice is that of

$$p(\theta) \propto I(\theta)^{1/2}$$

where $I(\theta)$ is Fisher information defined by:

$$I(\theta) = E\left(\frac{\partial \log p(X|\theta)}{\partial \theta}\right)^2$$

Since we have the transformation:

$$I(\theta) = I(g(\theta))(g'(\theta))^2,$$

one can see that Jeffreys's prior is invariant to reparametrisation. Besides, it has the appealing property of giving more weight to values of θ where the amount of information is larger.

Various 'non-informative' priors for the binomial parameter

Consider r successes from n trials: $r \sim B(n, \theta)$, then

$$\log p(r|\theta) = r \log \theta + (n - r) \log(1 - \theta) + \text{const}$$

and $I(\theta) = \frac{n}{\theta(1-\theta)}$.

Thus Jeffreys' prior is

$$p(\theta) \propto (\theta(1 - \theta))^{1/2},$$

which is a Beta distribution B(1/2, 1/2).

The Uniform density is a Beta(1,1) distribution.

A prior density that is uniform for logitθ is

$$p(\theta) \propto \theta(1 - \theta),$$

which is the improper B(0,0) distribution.

In practise, there will not be much difference between these alternatives, but the improper B(0,0) prior distribution leads to an improper posterior if $r = 0$ (or $n = 0$)!

Review

- Only need probability theory as basis for inference.
- Considerable progress has been made in computational issues
- No need to worry about significance tests, stopping rules, p-values.
- Models can be complex as reality demands (see the next lectures).
- Help to make choices explicit and accountable
- Tells us what we want to know: *how should this piece of evidence change what we currently believe?*
- Inferences need to be justified to an outside world (reviewers, regulatory bodies, the public and so on): in particular
  - Where did the prior come from?
  - Is the model for the data appropriate (diagnostics)?

2. Markov Chain Monte Carlo

Bayesian estimation methods

- Bayesian inference centres around the posterior distribution $\pi(\underline{\theta}|X)$ which has a known, although potentially complex functional form.
- Any features of the posterior are legitimate for Bayesian inference: moments, quantiles, highest posterior density regions (HPD's),...
- All can be expressed in terms of posterior expectations of functions of $\underline{\theta}$,

$$E[f(\underline{\theta})|X] = \frac{\int f(\underline{\theta})\pi(\underline{\theta})p(X|\underline{\theta})d\underline{\theta}}{\int \pi(\underline{\theta})p(X|\underline{\theta})d\underline{\theta}}$$

- In most applications, analytic evaluation of $E[f(\underline{\theta})|X]$ is impossible
- Alternatives include numerical evaluation, analytic approximations (e.g. Laplace), and Monte Carlo integration, including MCMC.

Heuristic view of simulation methods for Bayesian inference

- Imagine generating a random sample of values from a probability distribution (e.g. normal);
- Construct a histogram from the sample;
- If the sample is large enough, histogram can provide virtually complete information about the distribution from which these samples were drawn:
  - Mean, variance, percentiles of sample $\approx$ mean, variance, percentiles of original distribution.
- MCMC methods enable us to generate large samples from the posterior distributions of model parameters:
  - These samples can be summarised to estimate properties (e.g. mean, variance, percentiles) of the posterior distribution.

Monte Carlo integration

- Suppose we can draw samples from the joint posterior distribution for $\underline{\theta}$, *i.e.*

$$\underline{\theta}^{(1)}, \underline{\theta}^{(2)}, ..., \underline{\theta}^{(N)} \sim \pi(\underline{\theta}|\underline{x})$$

- Then

$$E(g(\underline{\theta})) = \int g(\underline{\theta})\pi(\underline{\theta}|\underline{x})d\underline{\theta}$$
$$\approx \frac{1}{N}\sum_{i=1}^{N} g(\underline{\theta}^{(i)}) = \bar{g}_N$$

this is Monte Carlo integration

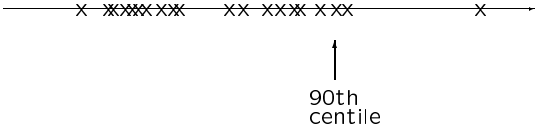
- Theorems exist which prove convergence in the limit as $N \to \infty$ even if the sample is dependent (crucial to the success of MCMC)

Assume we have a sample ($N = 20$) from $\pi(\theta_k \mid \underline{x})$

- Posterior mean

$$E(\theta_k) \approx \frac{1}{N}\sum_{i=1}^{N} \theta_k^{(i)}$$

- Quantiles



- Kernel density estimates

$$\hat{\pi}(\theta_k|\underline{x}) \approx \frac{1}{N}\sum_{i=1}^{N} h(\theta_k; \theta_k^{(i)})$$

How do we sample from the posterior?

- In general, we want samples from the joint posterior distribution $\pi(\underline{\theta}|\underline{x})$
- *Independent* sampling from $\pi(\underline{\theta}|\underline{x})$ may be difficult
- **BUT** $\{\underline{\theta}^{(t)}\}$ can be generated by any process which draws samples throughout the support of $\pi(\underline{\theta}|\underline{x})$ in the correct proportions
- One way is through a Markov chain with $\pi(\underline{\theta}|\underline{x})$ as its stationary distribution
- This is

Markov chain Monte Carlo (MCMC)

To summarize:

Bayesian posterior inference may be achieved via Monte Carlo integration using simulated values of all the unknown quantites $\underline{\theta}$ in the model generated from a Markov chain with $\pi(\underline{\theta}|\underline{x})$ as its stationary distribution

Aside: A little Markov chain theory

- Suppose we generate a sequence of random variables $Y^{(0)}, Y^{(1)}, ...$ such that at each time t, the next state $Y^{(t+1)}$ is drawn from a distribution

  $$P(Y^{(t+1)}|Y^{(t)})$$

  so that conditional on the value of $Y^{(t)}$, $Y^{(t+1)}$ is independent of $Y^{(t-1)}, ..., Y^{(0)}$
- $Y^{(0)}, Y^{(1)}, ...$ is a *Markov chain*
- P is the *transition kernel* of the chain.
- Subject to regularity conditions the chain will gradually "forget" $Y^{(0)}$ and converge to a unique stationary distribution ϕ, i.e.

  $$P[Y^{(t)} \in A|Y^{(0)}] \to \phi(A) \quad \text{as } t \to \infty, \ \forall \ Y^{(0)}$$
- As t increases $\{Y^{(t)}\}$ will look more and more like a *dependent* sample from $\phi(Y)$.
- Def^{n}: A probability distribution $\phi(.)$ in a Markov chain is stationary if

  $$\phi(Y^{(t+1)}) = \int \phi(Y^{(t)})P(Y^{(t+1)}|Y^{(t)})dY^{(t)}$$

  or, less formally, if $Y \sim \phi$ before the transition, then so it will afterwards (i.e. $\phi = P\phi$)

How do we design a Markov chain with $\pi(\underline{\theta}|X)$ as its unique stationary distribution?

- This is surprisingly easy and several standard 'recipes' are available.
- The key idea is most practical MCMC methods is *reversibility* or *detailed balance*
- Returning to our original notation – replace $\phi(.)$ by our posterior of interest $\pi(.|X)$ – detailed balance means

  $$\pi(\underline{\theta}^{(t)}|X)P(\underline{\theta}^{(t+1)}|\underline{\theta}^{(t)}) = \pi(\underline{\theta}^{t+1}|X)P(\underline{\theta}^{(t)}|\underline{\theta}^{(t+1)})$$
- All we need do is choose P so that this equality holds and we have a Markov chain with the desired stationary distribution.
  (Proof: take integrals of both sides w.r.t. $\underline{\theta}^{(t)}$)
- Note – detailed balance is sufficient but not necessary...

additional requirements are irreducibility and aperiodicity

The basic MCMC sampling methods

The Metropolis-Hastings sampler

- Draw a candidate new value θ^* from an arbitrary density *proposal* $q(\theta^*|\theta^{(t)})$
- For example, q may be as simple as a Normal or Uniform distribution centered at the current value, $\theta^{(t)}$.
- θ^* accepted as the next state of the chain (i.e. $\theta^{(t+1)} = \theta^*$) with probability

  $$\alpha(\theta^{(t)}, \theta^*) = \min\left\{1, \frac{\pi(\theta^*|X)q(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)}|X)q(\theta^*|\theta^{(t)})}\right\} \qquad (*)$$

  and otherwise θ is left unchanged and $\theta^{(t+1)} = \theta^{(t)}$.
- The transition kernel for this chain is

  $$P[\theta^{(t+1)}|\theta^{(t)}] = \overbrace{q(\theta^*|\theta^{(t)}) \cdot \alpha(\theta^{(t)}, \theta^*)}^{\text{Accept } \theta^*} + \overbrace{I(\theta^{(t+1)} = \theta^{(t)})[1 - \int q(\theta^*|\theta^{(t)})\alpha(\theta^{(t)}, \theta^*)d\theta^*]}^{\text{Rejection of all possible candidates } \theta^*}$$

Proof of detailed balance

- From $(*)$,

$$\pi(\theta^{(t)}|X)q(\theta^{(t+1)}|\theta^{(t)})\alpha(\theta^{(t)},\theta^{(t+1)}) = \pi(\theta^{(t+1)}|X)q(\theta^{(t)}|\theta^{(t+1)})\alpha(\theta^{(t+1)},\theta^{(t)})$$

- This, together with the expression for the kernel, yields,

$$\pi(\theta^{(t)}|X)P(\theta^{(t+1)}|\theta^{(t)}) = \pi(\theta^{(t+1)}|X)P(\theta^{(t)}|\theta^{(t+1)})$$

- So, we have detailed balance $\Rightarrow$ our stationary distribution is the req'd posterior $\pi(\theta|X)$.

Proposal distributions

- Metropolis-Proposal

$$q(\theta^*|\theta) = q(\theta|\theta^*)$$

  is symmetric

- Independence Proposal

$$q(\theta^*|\theta) = q(\theta^*)$$

  does not depend on the current value θ

- Note:

$$q(\theta^*|\theta) = q(\theta^*|X) \quad \Rightarrow \quad \alpha \equiv 1$$

  iid samples from $\pi(\theta|X)$

The Gibbs sampler

Let our vector of unknowns $\underline{\theta}$ consist of k sub-components $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_k)$

1) Choose starting values $\theta_1^{(0)}$, $\theta_2^{(0)}$, ..., , $\theta_k^{(0)}$

2) Sample $\theta_1^{(1)}$ from $\pi(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., , \theta_k^{(0)}, X)$

   Sample $\theta_2^{(1)}$ from $\pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., , \theta_k^{(0)}, X)$

   .....

   Sample $\theta_k^{(1)}$ from $\pi(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, ..., , \theta_{k-1}^{(1)}, X)$

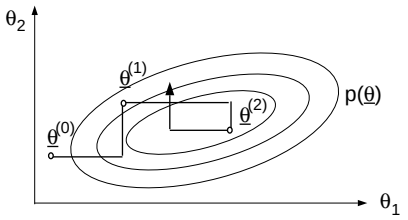
3) Repeat step 2 many 1000s of times
   – eventually obtain sample from $\pi(\underline{\theta}|X)$

We are using the kernel

$$P(\underline{\theta}^{(t+1)}|\underline{\theta}^{(t)}) = \prod_{i=1}^{k} \pi(\theta_i^{(t+1)}|\{\theta_j^{(t+1)}, j<i\}, \{\theta_j^{(t)}, j>i\}, X)$$

The distributions $\pi(\theta_i|\{\theta_j, j \neq i\}, X)$ are called full conditionals.

Gibbs sampling ctd.



- Sample $\theta_1^{(1)}$ from $\pi(\theta_1|\theta_2^{(0)}, X)$
- Sample $\theta_2^{(1)}$ from $\pi(\theta_2|\theta_1^{(1)}, X)$
- Sample $\theta_1^{(2)}$ from $\pi(\theta_1|\theta_2^{(1)}, X)$
- .....

$\underline{\theta}^{(n)}$ forms a Markov chain with (*eventually*) a stationary distribution $\pi(\underline{\theta}|X)$.

Example: Normal random sample

Suppose data $X_1, \ldots, X_n$ are a random sample from $N(\mu, \tau^{-1})$

Assume independent priors on μ and τ

$$\mu \sim N(\gamma, \kappa^{-1})$$

$$\tau \sim \Gamma(\alpha, \beta)$$

The posterior distribution $\pi(\mu, \tau|\underline{x})$ (up to a constant of proportionality) is

$$\sqrt{\tau} e^{\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\}} e^{\{-\frac{\kappa}{2}(\mu-\gamma)^2\}} \beta e^{-\beta\tau} (\beta\tau)^{\alpha-1}$$

The full conditional distributions can be derived as

$$\mu|\tau, \underline{x} \sim \mathrm{N}\left(\frac{\tau\sum x_i - \gamma\kappa}{n\tau+\kappa}, \frac{1}{n\tau+\kappa}\right)$$

$$\tau|\mu, \underline{x} \sim \Gamma\left(\alpha + n/2,\ \beta + \sum(x_i-\mu)^2/2\right)$$

Implement the Gibbs Sampler by alternately drawing μ and τ from these distributions.

Componentwise MH-algorithm

For $\underline{\theta} = \{\theta_1, \ldots, \theta_k\}$

Update elements θ_i seperately with MH-proposals conditionally on $\{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k\}$

As components of $\underline{\theta}$ are updated seperately this is a so-called single-site MCMC algorithm.

Gibbs is a special case of componentwise MH where the proposal density is just the full conditional so the acceptance probability at each sub-step is 1.

Hybrid sampler:

Combination of Gibbs and other MH-proposals

More generally block-updating may be considered.

Performance of MCMC methods

There are three main issues to consider

- Convergence (how quickly does the distribution of $\theta^{(t)}$ approach $\pi(\theta|X)$?)
- Efficiency (how well are functionals of $\pi(\theta|X)$ estimated from $\{\theta^{(t)}\}$?)
- Simplicity (how convenient is the method to use?)

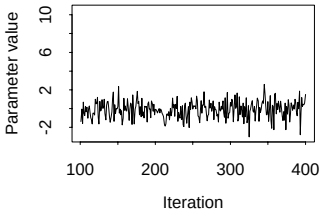
Note that computer effort should be measured in seconds, not iterations!

Gibbs Sampling is not superior to other methods on *any* of these criteria.

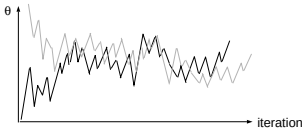
Checking convergence

This is the users responsibility!

- Note: Convergence is to target **distribution** (the required posterior), not to a single value.
- Once convergence reached, samples should look like a random scatter about a stable mean value.



One approach is to run many long chains with widely differing starting values.



Efficiency: How many iterations after convergence?

- For all the samplers we have considered, C.L.T.'s exist for our ergodic averages,

  $$N^{1/2}(\bar{g}_N - E_{\pi|X}[g(\underline{\theta})]) \to N(0, \sigma^2) \quad \text{as } N \to \infty$$

  or equivalently,

  $$\bar{g}_N \to N(E_{\pi|X}[g(\underline{\theta})], \sigma^2/N)$$

- $\sigma/\sqrt{N}$ is the *Monte Carlo standard error*. i.e. `MC error` $\propto$ sd of the difference between the mean of the sampled values of θ and the true posterior mean.
- An algorithm is inefficient with σ^2/N is large in comparison to $V_{\pi|X}(g(\underline{\theta}))$
- Accuracy of the posterior estimates can be assessed by the Monte Carlo standard error for each parameter.

Estimating the MC standard error

The Batch-means estimator

- Divide our run of length N into b consecutive blocks of length k

  $$\tilde{\sigma}/\sqrt{N} = \frac{1}{b-1}\sum_{i=1}^{b}(\bar{g}_{k,i} - \bar{g}_{N,1})^2$$

  where $\bar{g}_{k,i}$ is the mean of batch i,

  $$\bar{g}_{k,i} = \frac{1}{k}\sum_{j=(i-1)k+1}^{ik} g(\underline{\theta}^{(j)})$$

- Efficiency = Posterior variance / σ^2
  $\Rightarrow$ want MC error small in relation to posterior standard deviation.
- Rule of thumb: run simulation until the MC error for each parameter $< 5\%$ of sample (posterior) standard deviation.

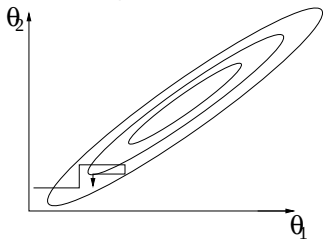
Poor mixing

- MCMC samplers often show poor mixing
  - i.e. sampler does not move rapidly throughout the support of the target distribution
- Slows convergence and increases Monte Carlo error variance
- Chains tend to be highly autocorrelated
- Often caused by high posterior correlations between model parameters

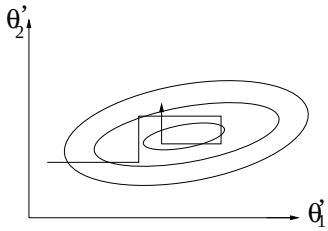
Possible remedies:

- Reparametrization
- Blocking of highly correlated components

Slow Mixing



Fast Mixing



Example for reparameterizing a model

- Consider simple regression model

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \alpha + \beta x_i \\ \alpha, \beta &\sim \text{flat priors} \end{aligned}$$

- Posterior correlation between α and β is

$$\rho_{\alpha\beta} = -\frac{\overline{x}}{\sqrt{\overline{x}^2 + \frac{1}{n}\sum_i (x_i - \overline{x})^2}}$$

- If $\overline{x} >>$ sd(x) $\Rightarrow \rho_{\alpha\beta} \to \pm 1$
- Remedy: standardize x_i about the sample mean $\overline{x}$:

$$\mu_i = \alpha' + \beta'(x_i - \overline{x})$$

- Posterior correlation $\rho_{\alpha'\beta'} = 0$

Some strengths of MCMC

- Freedom in modelling
  - in principle, no limits
  - well-adapted for models defined on sparse graphs
- Freedom in inference
  - in principle, no limits
  - can estimate arbitrary functions of model parameters (e.g. ranks, probabilities of threshold exceedence)
  - opportunities for simultaneous inference
- Allows/encourages sensitivity analysis
- Model comparison/criticism/choice
- Coherently integrates uncertainty
- Only available method for complex problems

Some weaknesses and dangers

- Order $\sqrt{N}$ precision
- Possibility of slow convergence and difficulties is diagnosis
- Risk that fitting technology runs ahead of statistical science
- Risk of undisciplined, selective presentation
- Difficulty of validating code

Some general comments on statistical computing and MCMC

The **great flexibility** provided by modern Bayesian methods has some downsides:

- **general purpose** software may be **inefficient** and **slow** for specific problems
- software designed for **specific problems** may be **efficient** but not applicable to other (extended) problems

It is very easy to cross the limits of any MCMC software; for more challenging problems one is almost always left with the only possibility:

CODE IT UP YOURSELF!

This is not as difficult as it seems, and may even provide **insight** into the problem and **satisfaction**.

However, there is now a number of interesting software around, e.g. WinBUGS or BayesX.

3. Bayesian Hierarchical Models

Introduction

Many statistical problems involve multiple parameters

Why ?

It is necessary to reflect the complexity of observables and different patterns of heterogeneity, dependence, mismeasurements ...

In epidemiology, multiple parameters involved in analyses of :

- "subject effect" in growth curves models
- "frailty" in correlated or familial survival data
- "centre effect" in meta-analyses
- relative risks for a disease outcome in different areas/age/time periods
- relative risks for different (tumour) sites in toxicological or occupational studies

Multiple parameter problems

How to make inference on multiple parameters $\{\theta_1, \ldots, \theta_I\}$ measured on I units (persons, centres, areas, ...) *which are related or connected by the structure of the problem ?*

- individual estimates of θ_i are likely to be highly variable (unless very large sample sizes)
- simulaneous testing procedures comparing each θ_i to a null value θ_0,
  while controlling for overall Type I errors, ignore the relatedness of $\theta_i \rightarrow$ over conservative

$\rightarrow$ express judgement of similarity of θ_i and integrate all analyses into a single model

Hierarchical models

> *"When a model has many parameters, it may be the case that we can consider them as a sample from some distribution. In this way we model the parameters with another set of parameters and build a model with different level of hierarchy."*
>
> (Schervish, 1995)

Connection between the notion of similarity of the θ_i and the representation of the θ_i as drawn from a distribution

$\rightarrow$ idea of **exchangeability**

Exchangeability

A sequence of random variables $x_1, x_2, \ldots$ is said to be exchangeable if, for each n and any permutation $\pi(1), \ldots, \pi(n)$:

$$p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

Def^n: Partial Exchangeability

The sequence $x_1, x_2, \ldots$ is partially exchangeable if it can be partitioned into subsequences $x_1^j, x_2^j, \ldots, j = 1, 2 \ldots$ such that the variables in each subsequence are exchangeable

Example : x_i^j where i represent treatment and j replication, there might be exchangeability between replications within a treatment, but not between treatments

Hierarchical Bayes model

Observables y,

Parameters $\theta = (\theta_1, \ldots, \theta_n)$

- likelihood $p(y|\theta)$ (1st level) models the structure of observables
- prior $p(\theta)$ is decomposed into conditional distributions $p(\theta|\eta_2)$ (2nd level), $p(\eta_2|\eta_3)$... expressing structural judgements (eg exchangeability ...) and a marginal distribution $p(\eta_m)$ such that

$$p(\theta) = \int p(\theta|\eta_2)p(\eta_2|\eta_3)\ldots p(\eta_{m-1}|\eta_m)p(\eta_m)d\eta_2 d\eta_3 \ldots d\eta_m$$

η_l are called the hyperparameters of level l

Beta-Binomial model

Suppose we observe I sets of binomial data, e.g.

- I=12 Hospitals performing cardiac surgery
- Number of surgical failures (deaths) per centre

	Hospital						
	A	**B**	**C**	....	**J**	**K**	**L**
No. of ops. n	47	148	119	....	97	256	360
No. of deaths r	0	18	8	....	8	29	24

How would you model these data?

One approach is to apply the same beta-binomial model to all the hospitals:

Likelihood:

$$\prod_{i=1}^{I} p(r_i|\pi, n_i) = \prod_{i=1}^{I} \text{Binomial}(n_i, \pi)$$

Prior:

$$p(\pi) = \text{Beta}(a, b)$$
(a, b known constants)

Posterior:

$$p(\pi \mid \underline{r}, \underline{n}) \propto \prod_{i=1}^{I} p(r_i \mid \pi, n_i)p(\pi)$$
$$= \text{Beta}(\sum_i r_i + a,\ \sum_i (n_i - r_i) + b)$$

BUT is it reasonable to assume a *common* probability π of failure for every hospital?

The beta-binomial model above assumes that each outcome (proportion of failures per hospital) is *independent and identically distributed* according to the binomial probability distribution with parameter π

- Does this model adequately describe the random variation in outcomes for each hospital?
- Are the hospital failure rates more variable than our model assumes?

Reasons for excess variation in response

- Individual heterogeneity *i.e.* systematic differences between units which are not attributable to binomial variation
  - this concept is often termed *frailty* in survival analysis
  - for binary/count data this is often termed *overdispersion*
- Repeated response measurements from the same unit tend to be *correlated*
  - $\Rightarrow$ 2 responses from the same unit will be more alike than 2 responses from different units
  - $\Rightarrow$ variation in responses is not completely random
- Failure to measure or include a relevant explanatory variable
- Inaccurate measurement of relevant explanatory variables

Modelling the excess variation

Perhaps we could modify our beta-binomial model to allow for a *different* failure probability, π_i for each hospital i:

$$p(r_i \mid n_i, \pi_i) = \text{Binomial}(n_i, \pi_i)$$

$$p(\pi_i) = \text{Beta(a, b)}$$

Interpretation

- π_i is the 'true' surgical failure rate in hospital i
- π_i's are viewed as a random sample from a common *population distribution*
  - $\Rightarrow$ hospital failure rates are assumed to be **similar** but not identical

  Is this reasonable?
- Beta(a, b) prior describes the distribution of surgical failure rates amongst the 'population' of hospitals

How would you specify values for a and b?

Hierarchical Bayes approach

- Assume a *joint probability model* for the entire set of parameters $(\underline{\pi}, a, b)$

  $\Rightarrow$ assign known prior distributions to a and b, e.g.

  $$a \sim \text{Exponential}(0.01)$$
  $$b \sim \text{Exponential}(0.01)$$

- Apply Bayes theorem to simultaneously estimate the joint posterior distribution of all the unknown quantities

Level 1	$r_i \sim \text{Binomial}(n_i, \pi_i)$
Level 2	$\pi_i \sim \text{Beta}(a, b)$
Level 3	Prior for a, b

Advantages of this approach

- The posterior distribution for each π_i
  - *'borrows strength'* from the likelihood contributions for *all* hospitals, via their joint influence on the estimate of the unknown population (prior) parameters a and b
  - reflects our full uncertainty about the true values of a and b

Such models are also called *Random effects* or *Multilevel* models

Poisson-Gamma model

Small area disease counts

Question:

is stomach cancer elevated in coastal areas where shellfish is contaminated by diarrheic shellfish poisoning toxins ?

Data:

16 areas selected for their potential high level of contamination

y_i : observed number of stomach cancers in area i,
E_i : expected number of stomach cancers in area i, adjusted for age, sex

Parameters:

θ_i : underlying relative risk of stomach cancer in area i

- 1st level – local variability (within area):

  $$y_i \sim \text{Poisson}(E_i\theta_i)$$

- 2nd level – exchangeability

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

In the first instance, suppose that α and β are known (from previous experience), then

$$p(\theta_i|\alpha, \beta, Y) \sim \text{Gamma}(y_i + \alpha, E_i + \beta)$$

$$E(\theta_i|\alpha, \beta, Y) = \frac{y_i + \alpha}{E_i + \beta}$$

Parameters θ_i are stabilised and shrunk towards the population mean α/β

In general, we model uncertainty in hyperparameters at a further hierarchical level

- 3rd level – hyperpriors (e.g. exponential distributions)

  $$\alpha \sim p(\alpha), \quad \beta \sim p(\beta)$$

The joint posterior distribution is :

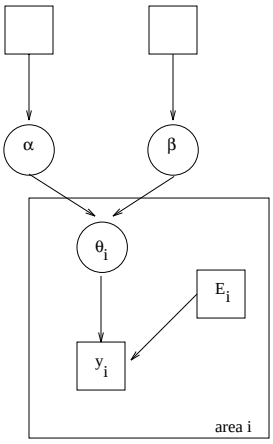
$$\prod_i p(y_i|\theta_i)p(\theta_i|\alpha, \beta)p(\alpha)p(\beta)$$

When computing $p(\theta_i|Y, E)$, the uncertainty in α, β and other $\theta_j s$ is integrated out.

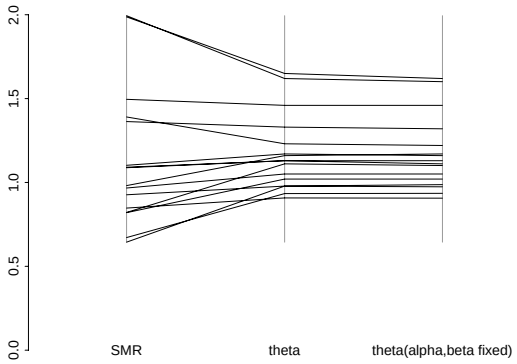
Average risk of stomach cancer in all areas is given by:

$$E\left(\frac{\alpha}{\beta} \mid Y\right) = \int \frac{\alpha}{\beta}\ p(\alpha, \beta \mid y_i, E_i)\ d\alpha\ d\beta$$

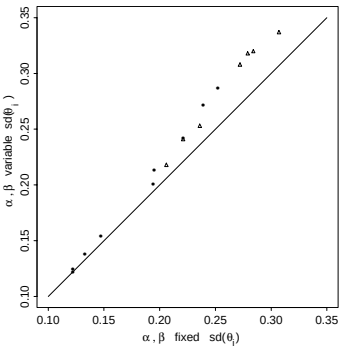
DAG for small area model



Comparing the dispersion of SMR and Bayesian estimates



Comparison of the posterior standard deviations obtained for the relative risk θ_i : α and β fixed versus α and β given exponential priors



Triangles indicate areas with $y_i < 10$

Exchangeability of $\{\theta_i\}$ rarely realistic: data often present **heterogeneity** or **clustering**

use e.g. prior with spatially correlated parameters

When relevant covariates are available, include those in regression-type approach

Computation for Bayesian hierarchical models

- Under the posterior distribution, the parameters are generally *dependent*, so we have to compute with a multivariate distribution, often in a high number of dimensions, with arbitrarily complex patterns of dependence.
- Sometimes posterior calculations can be done explicitely with suitable hyperpriors.
- But in most other cases, analytical integration or static simulation are not possible and inference in hierarchical models relies on MCMC algorithms.

Comparison with traditional approaches

- Classical approch to random effects/multilevel modelling uses quasi-likelihood (e.g. MQL, PQL)
  - Software includes SAS PROC MIXED, MLn
  - Methods are *approximate*, whereas Bayesian hierarchical models provide *exact* estimates
- For large samples (units and observations per unit), Bayesian and classical inference tend to give similar results
- For small samples
  - Little information available for estimating population variability
  - Classical methods can be unreliable
  - Bayesian approach allows inclusion of external knowledge (prior) which can help to stabilize the model

Conclusion

Many interlinked arguments to favor the use of hierarchical models:

- by breaking down the problem in layers, we are able to separate structural judgments on observables, on parameters and subjective information
- reduces the arbitrariness of hyperparameter choice $\rightarrow$ robustify the inference
- natural structure for expressing dependence, prior correlations
- through shrinkage and borrowing of strength, parameter estimates are stabilized
- computationally feasible via MCMC algorithms