Lattice Date

Bayesian Estimation
Hierarchical Models

Dra. ROSA ABELLANA

DEPARTMENT PUBLIC HEALTH. UNIVERSITY OF BARCELONA

BAYESIAN STATISTICS

Classical framework unknown parameter, θ**, are fixed non-random quantities**. The estimate of this parameter is performed by a function of the sample, the estimator is a random variable.

Bayesian framework the unknown parameter, θ**,** it is **a random variable** and it has a distribution of probability **p(**θ**)**. This distribution, is the **initial distribution or prior distribution.**

- Prior Distribution: expresses our uncertainty about θ before seeing the data.

Information of the data on θ, collect by the likelihood function $p(x|\theta)$.

Bayesian inference. Use the theorem of Bayes:

$$\underbrace{p(\theta \mid x)}_{\text{A posteriori distribution of the parameter}} = \frac{p(x \mid \theta)p(\theta)}{p(x)} \alpha \quad \underbrace{p(x \mid \theta)}_{\text{Likelihood function}} \; \overbrace{p(\theta)}^{\text{Prior distribution}}$$

•$p(\theta|x)$ The **posterior distribution** is the final result of the Bayesian analysis and all the information on the parameter, it is obtained from this distribution. Posterior expresses our uncertainty about after seeing the data.

Point estimation

Given the posterior distribution p(θ|y), the Bayesian estimate of θ can be

Posterior mean $\hat{\theta} = \mathrm{E}(\theta \mid \mathrm{y})$

Posterior median $\hat{\theta} : \int_{-\infty}^{\hat{\theta}} \mathrm{p}(\theta \mid \mathrm{y}) \mathrm{d}\theta = 0.5$

Posterior mode $\hat{\theta} : \mathrm{p}(\hat{\theta} \mid \mathrm{y}) = \sup_{\theta} \mathrm{p}(\theta \mid \mathrm{y})$

Estimation for interval

The same that the median if we want to calculate the percentile $\alpha/2$ and the 1-$\alpha/2$ of p(θ|y), that is to say the points q_l and q_u

$$\int_{-\infty}^{q_L} \mathrm{p}(\theta \mid \mathrm{y}) \mathrm{d}\theta = \alpha/2 \qquad \int_{q_u}^{\infty} \mathrm{p}(\theta \mid \mathrm{y}) \mathrm{d}\theta = 1 - \alpha/2$$

Therefore $P(q_l < \theta < q_U) = 1 - \alpha$.

Ex: Bayesian inference using a Normal distribution

Known variance, unknown mean

We suppose have a sample of data of a Normal $x_i \sim N(\theta,\sigma^2)$ (i=1,..n). Assuming that σ^2 is known and prior distribution of θ is $N(\mu,\sigma^2/n_0)$

The posterior distribution:

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \prod_i p(x_i \,|\, \theta)\; p(\theta) \\ &\propto \exp\left[-\frac{\sum_i (x_i-\theta)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\theta-\mu)^2 n_0}{2\sigma^2}\right] \end{aligned}$$

Therefore, deleting all the terms that do not depend of θ

$$p(\theta|\boldsymbol{x}) = \mathrm{N}\left(\frac{n_0\mu + n\overline{x}}{n_0+n}, \frac{\sigma^2}{n_0+n}\right)$$

Concentration of trihalometranes (THMs) in dringking water

We suppose that we want to estimate the average concentration of THM in a zone.

We have 2 independent measures x_1 and x_2. The mean is 130 μg/l.

We suppose that the error of measure has a standard deviation of σ_e = 5 μg/l

Which is the estimate of the average concentration, θ?

Standard analyses

Estimate would be $\bar{x} = 130 \mu g / l$

The estimate of the standard error would be $\sigma_e / \sqrt{n} = 5 / \sqrt{2} = 3.5 \mu g / l$

Interval of confidence to 95% $\bar{x} \pm 1.96 \times \sigma_e / \sqrt{n}$ $132.1 \text{ a } 136.9 \mu g / l$

We suppose that you have historical data on concentrations levels of THM. The average of the concentration is 120 μg/l and the standard deviation is 10 μg/l

Prior distribution for $\theta \sim N(120, 10^2)$

If we express the standard deviation of the prior $\sigma_e / \sqrt{n_0}$

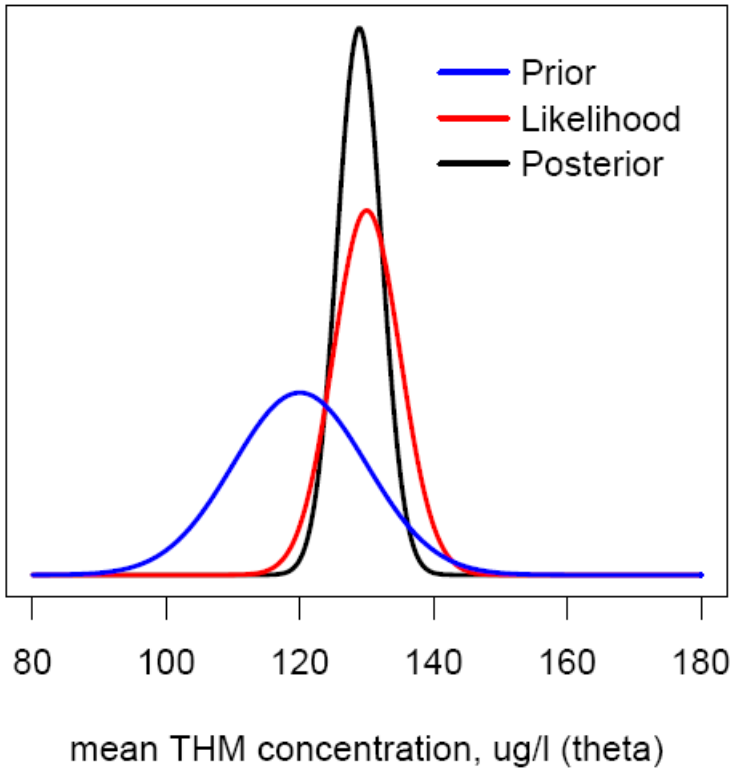
Then $n_o = \left(\sigma_e / 10\right)^2 = 0.25$

Therefore, the prior distribution to prior $\theta \sim N(120, \sigma_e^2/0.25)$

And the posterior distribution of θ

$$p(\theta_{zq}|\boldsymbol{x}) = \text{Normal}\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \frac{5^2}{0.25 + 2}\right)$$
$$= \text{Normal}(128.9,\ 3.33^2)$$

Obtaining an interval for θ of 122.4 to 135.4 μg/l – 95%- Credibility interval-


Prior
Likelihood
Posterior
80
100
120
140
160
180
mean THM concentration, ug/l (theta)

Bayesian inference using counts data: Estimate of the risk of an illness in a region.

We are interested in estimating the rate or relative risk for some Poisson data

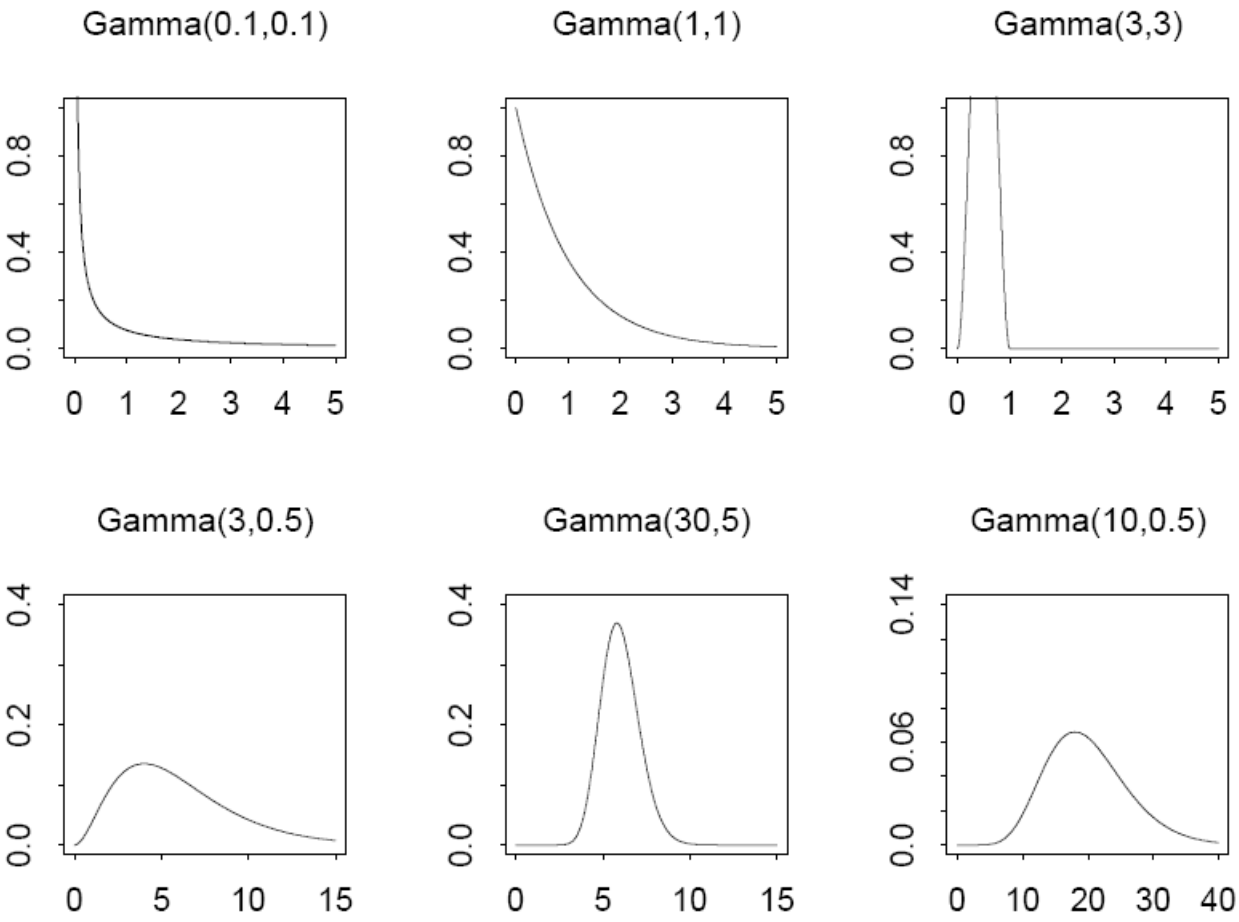
We suppose that we observe x=5 cases of leukemia in a region and that the numeral of cases expected standardized by age-gender is E=2.8.

We assume that the likelihood for x is a Poisson (λE), where λ is the unknown relative risk

$$p(x|\lambda, E) = \frac{(\lambda E)^x e^{-\lambda E}}{x!}$$

We assume that the distribution a priori for λ is a Gamma(a,b)

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$



X~Gamma(α,β)

E(x)= α/β
Var(X)= α/β²

The posterior distribution for λ will be

$$p(\lambda|x,E) \propto \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}\frac{(\lambda E)^x e^{-\lambda E}}{x!}$$
$$\propto \lambda^{a+x-1}e^{-(b+E)\lambda} \propto \text{Gamma}(a+x, b+E)$$

Vague Prior Information about λ

λ~Gamma(0.1,0.1)

E(λ)=0.1/0.1=1

Var(λ)=0.1/0.1^2=10

95th percentile= 5.8

Posterior distribution Gamma(5.1,2.9)

Posterior mean for λ

E(λ|x)=5.1/2.9=1.76

CI(95%):[0.58, 3.58]

(SMR=5/2.8=1.78)

Strong Prior Information about λ

λ~Gamma (48,40)

E(λ)=48/40=1.2

Var(λ)=48/40^2 =0.03

95th percentile= 1.5

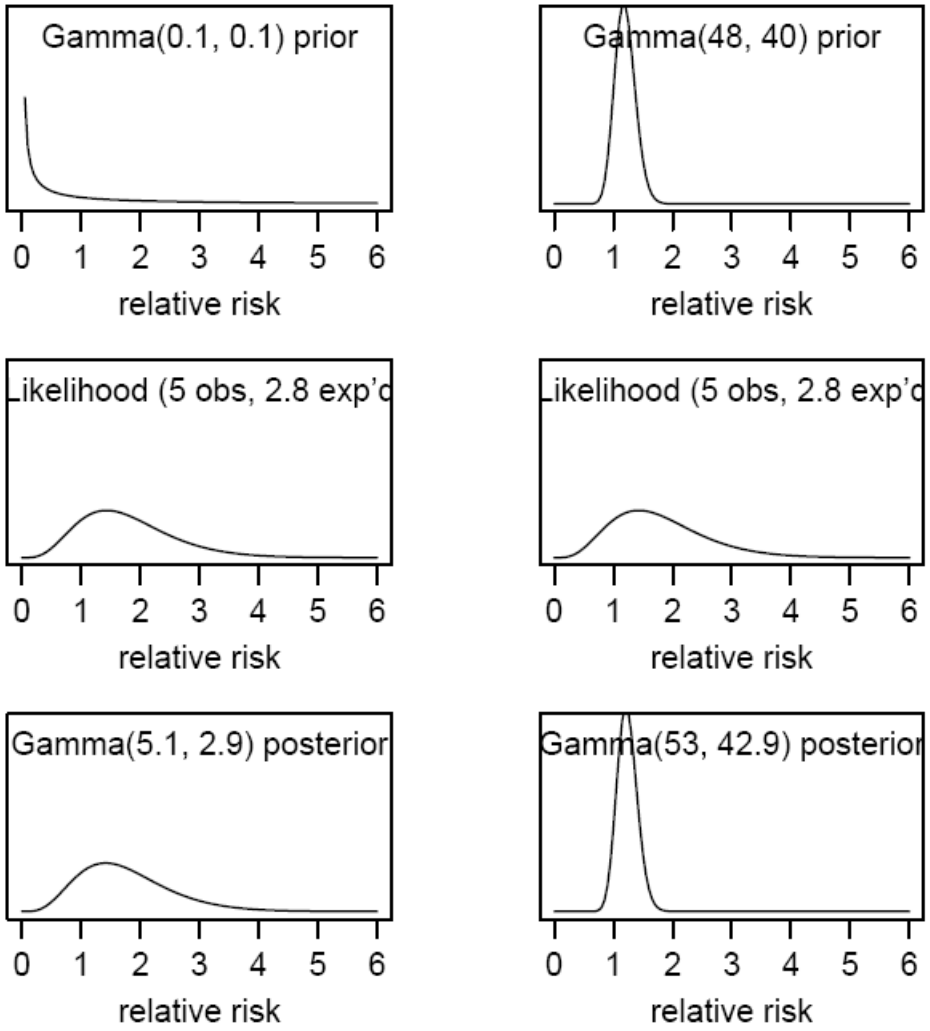
Posterior distribution Gamma(53,42.9)

Posterior mean for λ

E(λ|x)=53/42.9=1.24

CI(95%): [0.92, 1.56]

(SMR=5/2.8=1.78)


Gamma(0.1, 0.1) prior
0 1 2 3 4 5 6
relative risk
Gamma(48, 40) prior
0 1 2 3 4 5 6
relative risk
Likelihood (5 obs, 2.8 exp'd
0 1 2 3 4 5 6
relative risk
Likelihood (5 obs, 2.8 exp'd
0 1 2 3 4 5 6
relative risk
Gamma(5.1, 2.9) posterior
0 1 2 3 4 5 6
relative risk
Gamma(53, 42.9) posterior
0 1 2 3 4 5 6
relative risk

Prior distributions of the parameters

When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood

Likelihood	Parameter	Prior	Posterior
Normal	Mean	Normal	Normal
Normal	Accuracy	Gamma	Gamma
Binomial	Prob Event	Beta	Beta
Poisson	Rate or Mean	Gamma	Gamma

Conjugate prior distributions are mathematically convenient

Computations for non-conjugate priors are harder, but possible using MCMC

Don’t have any information about the parameters. To be objective, select a **non informative prior.**

- Give the same probability to all the values using a uniform distribution.

- Consider a proportional distribution to a constant

$P(\theta) \sim 1$

Improper distribution $(\int p(\theta)d\theta \neq 1)$

Commonly the posterior distribution will be proper

Inference is based on the likelihood $p(x|\theta)$.

Recommendations

Location parameters (e.g. mean, coefficients of regression)

$$\theta \sim \text{Unif}(-100, 100)$$
$$\theta \sim \text{Normal}(0, 100000)$$

Parameters scale $\tau=\sigma^{-2}$
(inverse variance, precision)

$$\tau \sim \text{Gamma}(\epsilon, \epsilon)$$

Sensitivity analysis plays a crucial role in assessing the impact of particular prior distributions, whether elicited, derived from evidence, or reference, on the conclusions of the analysis.

More than one parameter of interest $\theta=(\theta_1, \theta_2)$

- Joint posterior distribution
- Marginal posterior distribution

First find the **joint posterior** distribution of all the parameters,

$$p(\theta_1, \theta_2 \mid x) \propto p(x \mid \theta_1, \theta_2) \cdot p(\theta_1, \theta_2)$$

Next obtain the marginal **distribution marginal** of each parameters

$$p(\theta_1 \mid x) = \int p(\theta_1, \theta_2 \mid x) d\theta_2$$

$$p(\theta_2 \mid x) = \int p(\theta_1, \theta_2 \mid x) d\theta_1$$

The method of estimate calls **Fully Bayesian**

The Bayesian inferences centers around the posteriori distribution of the parameters

$$p(\boldsymbol{\theta}|x) \propto p(x|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

Where θ typically is a vector of parameters $\theta=\{\theta_1, \theta_2 \ldots\ldots \theta_k\}$

P(x| θ) and p(θ) will often available in closed form, but p(θ|x) is usually not analytically tractable. We want to

Posterior marginal $p(\theta_i|x) = \int\int \ldots \int p(\boldsymbol{\theta}|x)\ d\boldsymbol{\theta}_{(\ i)}$

Where $\theta_{(-i)}$ is the vector of parameters excluding θ_i

Calculate the properties of p(θ_i|x), such as $E(\theta_i) = \int \theta_i\ p(\theta_i \mid x)\, d\theta_i$

To avoid to resolve analytically these integrals can use simulation such as Chains of Markov Monte Carlo, (MCMC)

Exemple: Linear regression

$$Y_i = \beta X_i + \varepsilon_i \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \text{N}(0,1/\tau)$$

Unkown parametres in the linear regression model β and τ

Belief prior

$$p(\beta) \sim Normal(\mu_\beta, 1/\tau_\beta) \qquad p(\tau) \sim Gammal(\alpha_\tau, \delta_\tau)$$

Likelihood funtion

$$L(\beta, \tau) = \frac{\tau}{(2\pi)^{n/2}} exp\left\{-\tau \sum_{i=1}^{n} \left(\frac{(Y_i - \beta X_i)^2}{2}\right)\right\}$$

Posterior function

$$p(\beta, \tau) \propto L(\beta, \tau)\, p(\beta) p(\tau)$$

Posterior density

If τ is fixed, the density of β conditional on τ is a normal density

$$\boldsymbol{p(\beta|\tau, Y) \sim Normal(\mu_{\beta*}, 1/\tau_{\beta*})}$$

$$\boldsymbol{\tau_{\beta*} = \tau_{\beta} + \tau \sum X_i^2} \qquad \boldsymbol{\mu_{\beta*} = \frac{1}{\tau_{\beta}} \left(\beta \tau_{\beta} + \tau \sum X_i \; Y_i \right)}$$

If β is fixed, the density of τ conditional on β is a gamma density

$$\boldsymbol{p(\tau|\beta, Y) \sim Gamma(\alpha_{\tau\square}, \delta_{\tau*})}$$

$$\boldsymbol{\alpha_{\tau*} = n + \alpha_{\tau}} \qquad \boldsymbol{\delta_{\tau\square} = \frac{\delta_{\tau} + \sum (Y - \beta X)^2}{2}}$$

General Monte Carlo Integration

Integrals are evaluated by the integration of Monte Carlo that uses the values simulated of all the unknown parameters, which are generated from a chain of Markov, that is to say, that the generation of a new value for a parameter conditions to the previous value.

Sample of θ_k from a process of Markov satisfies that

$$p\left(\theta_k^{(t+1)} \mid \theta_k^{(t)}, \ldots \theta_k^{0}\right) = p\left(\theta_k^{(t+1)} \mid \theta_k^{(t)}\right)$$

$\theta_k^{(0)}, \theta_k^{(1)} \ldots \theta_k^{(t+1)}$ A Markov chain, and it is a sample of the posterior marginal distribution $p(\theta_k|x)$

For each parameter will have

$$\left[\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)}\right] \quad \left[\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_n^{(1)}\right] \quad \ldots\ldots, \left[\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_n^{(t)}\right] \sim p(\theta \mid x)$$

We can estimate

$$E\left[g(\theta_1)\right] = \int g(\theta_1) p(\theta_1 \mid x) d\theta_1 \approx \frac{1}{T} \sum_{i=1}^{T} g\left(\theta_1^{(i)}\right)$$

The algorithm MCMC more used and known is the Gibbs sampling, which works from the full conditioned distributions

Sampling of Gibbs

We suppose that have n parameters with distributions conditioned

$$p(\theta_1 \mid \theta_2, \ldots \theta_n) \quad p(\theta_2 \mid \theta_1, \ldots \theta_n) \qquad p(\theta_n \mid \theta_2, \ldots \theta_{n-1})$$

1) Choose initial values. $\left[\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)}\right]$

2) Sample $\theta_1^{(1)}$ from $\theta_1^{(1)} \sim p\left(\theta_1 \mid \theta_2^{(0)}, \ldots \theta_k^{(0)}, x\right)$

   Sample $\theta_2^{(1)}$ from $\theta_2^{(1)} \sim p\left(\theta_2 \mid \theta_1^{(1)}, \ldots \theta_k^{(0)}, x\right)$

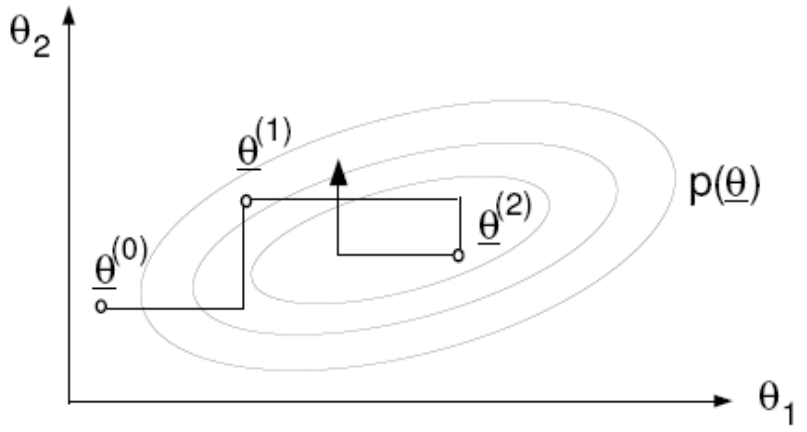
   ....

   Sample $\theta_k^{(1)}$ from $\theta_k^{(1)} \sim p\left(\theta_k \mid \theta_1^{(1)}, \ldots \theta_{k-1}^{(0)}, x\right)$

3) Repeat step 2 many 1000 s(or n) of times

The conditional distributions are called **full conditionals** because they are conditioned to all the other parameters

Example with k=2



2) Sample $\theta_1^{(1)}$ From $\theta_1^{(1)} \sim p\left(\theta_1 \mid \theta_2^{(0)}, x\right)$

Sample $\theta_2^{(1)}$ From $\theta_2^{(1)} \sim p\left(\theta_2 \mid \theta_1^{(1)}, x\right)$

Sample $\theta_2^{(2)}$ From $\theta_1^{(2)} \sim p\left(\theta_1 \mid \theta_2^{(1)}, x\right)$

Sampling of Gibbs ⟶ **We stroll for the sample space**

First samples are discarded

⇩

Burning process

⇩

Check if the chains came from posterior distribution

⇩

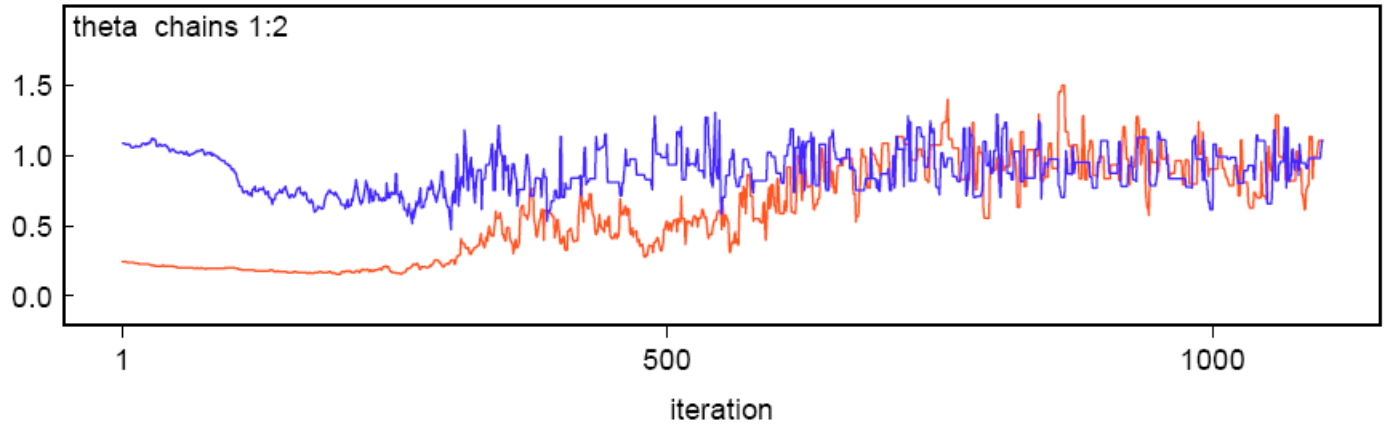
Convergence to posterior distribution.

Checking convergence

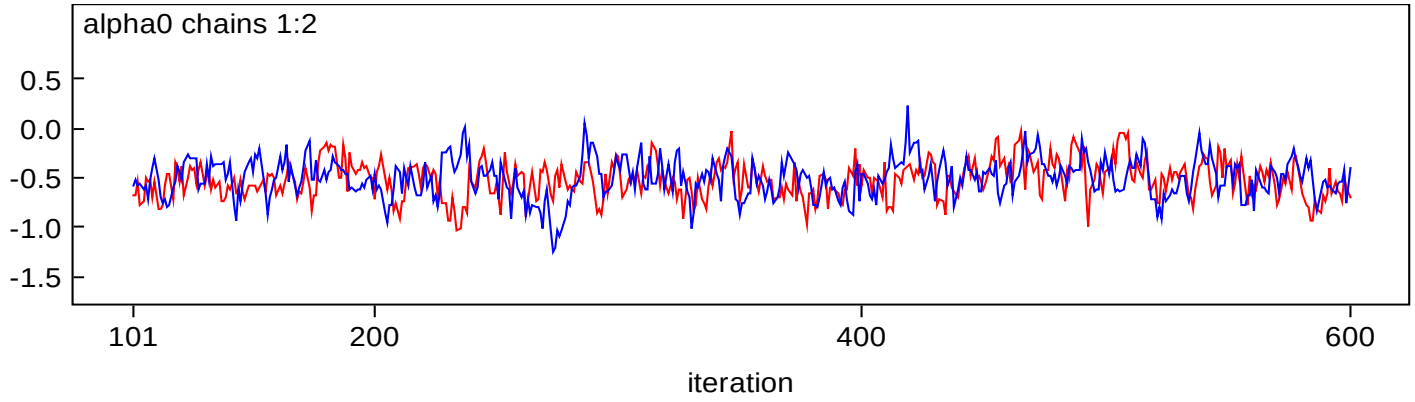
•Convergence is to a distribution not to a single value.

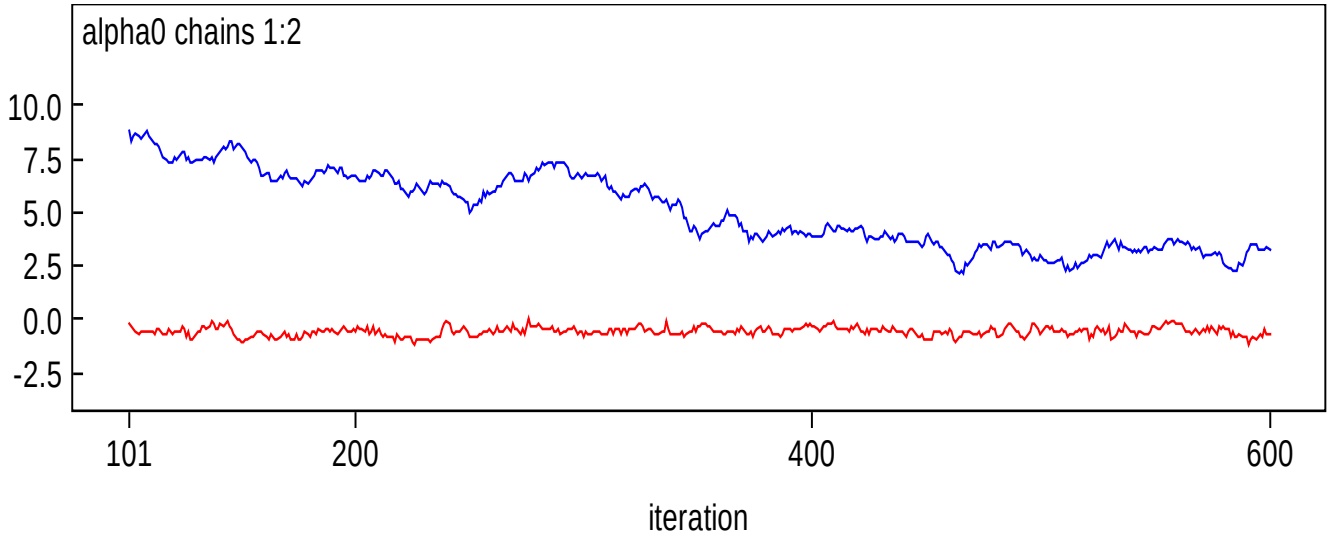
•Once convergence reached, samples should look like a random scatter about a stable mean value.

•Practice, is to *run* multiple chains for the same model but with different initial points, and visually inspect “trace” plot for convergence



Example convergence and no convergence





Statistician of Gelman-Rubin

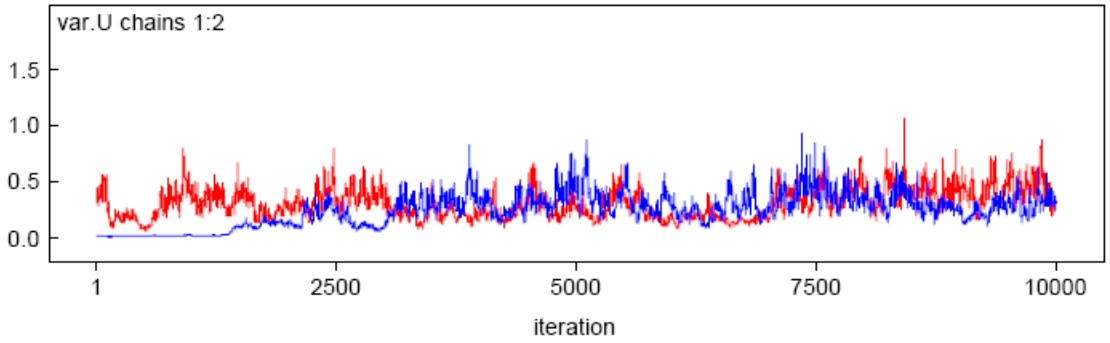
Formally, can use the Gelman-Rubin statistician based in the ratio of the *between* and *within* chains variance (ANOVA)

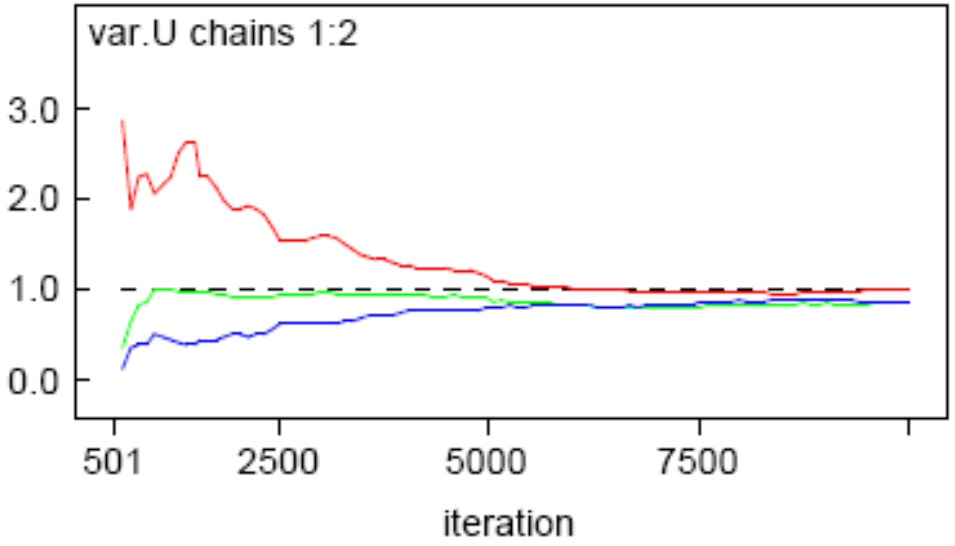
Gelman: the best to identify the no convergence, starting from different initial values,

Intuitively, the behavior of all the chains would have to be the same.

Or, the variance within the chain would have to be the same that the variance along the chain

Example of Gelman and Rubin with the Winbugs





B pooled

W average of width

Gelman-Rubin statistic

- Generate multiple chains starting at *over-dispersed* initial values.

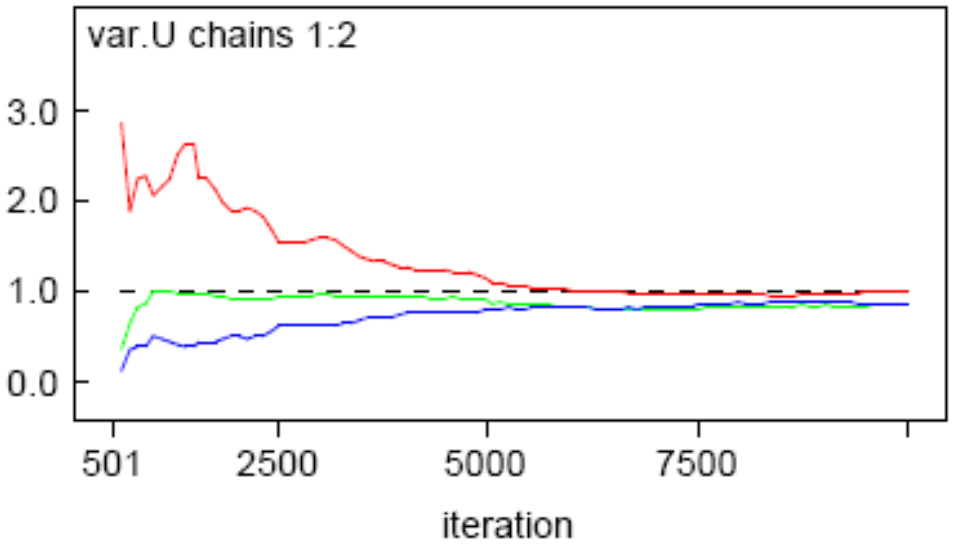
- Denote the number of chains generated by *M* and the length of each chain by 2*T*.

- Take as a measure of posterior variability the width of the 100(1- *a*)% credible interval for the parameter of interest (in WinBUGS, *a* = 0.2).

- From the final *T* iterations we calculate the empirical credible interval for each chain. We then calculate the average width of the intervals across the *M* chains and denote this by *W*

- Calculate the width *B* of the empirical credible interval based on all *MT* samples pooled together.

The ratio $R = B / W$ of pooled to average interval widths should be greater than 1 if the starting values are suitably overdispersed; it will also tend to 1 as convergence is approached, and so we might assume convergence for practical purposes if $R < 1.05$, say.



Generalized linear mixed models mixed via Fully Bayesian

Generalized linear mixed model, the mean conditioned to the random effects, μ^b, of the outcome variable, Y, is related to the explanatory variables by the link function , g(.), :

$$g\left(\mu^b\right) = \eta^b = X\beta + Zb$$

where X is design matrix of the fixed effects,
β is the parameters of the fixed effects,
Z is design the matrix of the random effects
and b is a vector of dimension n that contains the random effects

$$b \sim \mathrm{NMV}(0, \Sigma)$$

This joint distribution of the random effects is expresses as

$$f[b \mid \delta] \longrightarrow$$ It allowed to define spatial correlation

And the conditional distribution of each random effect denotes

$$f\left[b_i \mid b_j \; j \neq i, \delta\right]$$

In this model the unknown parameters are β, b, and δ, to which will assign them a priori distribution,

$$f[\beta], f[b \mid \delta], f[\delta]$$

Based to the beliefs of the researcher.

The **joint posterior distribution** of the unknown parameters is obtained by theorem of Bayes :

$$f\left[b,\delta,\beta \mid Y\right] \propto f\left[Y \mid b,\delta,\beta\right]\cdot f\left[b \mid \delta\right]\cdot f\left[\delta\right]\cdot f\left[\beta\right]$$

And the **marginal posterior distributions** of the parameters are defined

$$f\left[\beta \mid Y\right] = \int_{\beta}\int_{b_1}\int_{b_2}\ldots\int_{b_N}\int_{\delta} f\left[b,\delta,\beta \mid Y\right] d\beta\, db_1\, db_2 \ldots db_N d\delta$$

$$f\left[b_i \mid Y\right] = \int_{\beta}\int_{b_1}\ldots\int_{b_{i-1}}\int_{b_{i+1}}\ldots\int_{b_N}\int_{\delta} f\left[b,\delta,\beta \mid Y\right] d\beta db_1 db_{i-1} db_{i+1} \ldots db_N d\delta$$

$$f\left[\delta \mid Y\right] = \int_{\beta}\int_{b_1}\int_{b_2}\ldots\int_{b_N} f\left[b,\delta,\beta \mid Y\right] d\beta\, db_1\, db_2 \ldots db_N$$

To avoid to resolve analytically these integrals, simulation by Chains of Markov Monte Carlo, (MCMC) (Gilks, Richardson and Spiegelhalter, 1996).

Conditional distributions are:

$$f\left[\beta \mid b, \delta, Y \cdot\right] \propto f\left[\beta\right] \cdot \prod_{i=1}^{N} l\left[Y_i ; \beta, \delta\right]$$

$$f\left[b_i \mid b_j \; j \neq i, \delta, \beta, Y\right] \propto f\left[b_i \mid b_j \; j \neq i\right] \cdot l\left[Y_i ; \beta, \delta\right]$$

$$f\left[\delta \mid b, \beta, Y\right] \propto f\left[\delta\right] \cdot f\left[b \mid \delta\right]$$

Gibbs sampling process:

1) Chose initial values for the parameters of interest:

$$\delta^{(0)}, b_1^{(0)}, \ldots b_N^{(0)}, \beta^{(0)}$$

2) Sample a new value of $\delta^{(1)}$ from his conditioned distribution. The conditioned parameters are substituted by the initial values:

$$f\left[\delta \mid b^{(0)}, \beta^{(0)}, Y\right] \propto f\left[\delta\right] \cdot f\left[b^{(0)} \mid \delta\right]$$

3) Sample a new value of the first random effect, $b_1^{(1)}$

$$f\left[b_1 \mid b_j^{(0)}\ j \neq 1, \delta^{(0)}, \beta^{(0)}, Y\right] \propto f\left[b_1 \mid b_j^{(0)}\ j \neq 1\right] \cdot l\left[Y_i; b^{(0)}, \beta, \delta^{(0)}\right]$$

Until the random effect n-th $b_N^{(1)}$

$$f\left[b_N \mid b_j^{(0)}\ j \neq N, \delta^{(0)}, \beta^{(0)}, Y\right] \propto f\left[b_N \mid b_j^{(0)}\ j \neq N\right] \cdot l\left[Y_i; b^{(0)}, \beta, \delta^{(0)}\right]$$

4) Finally, sample a new value of the fixed effect β from:

$$f\left[\beta \mid b^{(0)}, \delta^{(0)}, Y \vdots \cdot\right] \propto f[\beta] \cdot \prod_{i=1}^{N} l\left[Y_i; \beta, b^{(0)}, \delta^{(0)}\right]$$

Example: Disease mapping

BUGS code

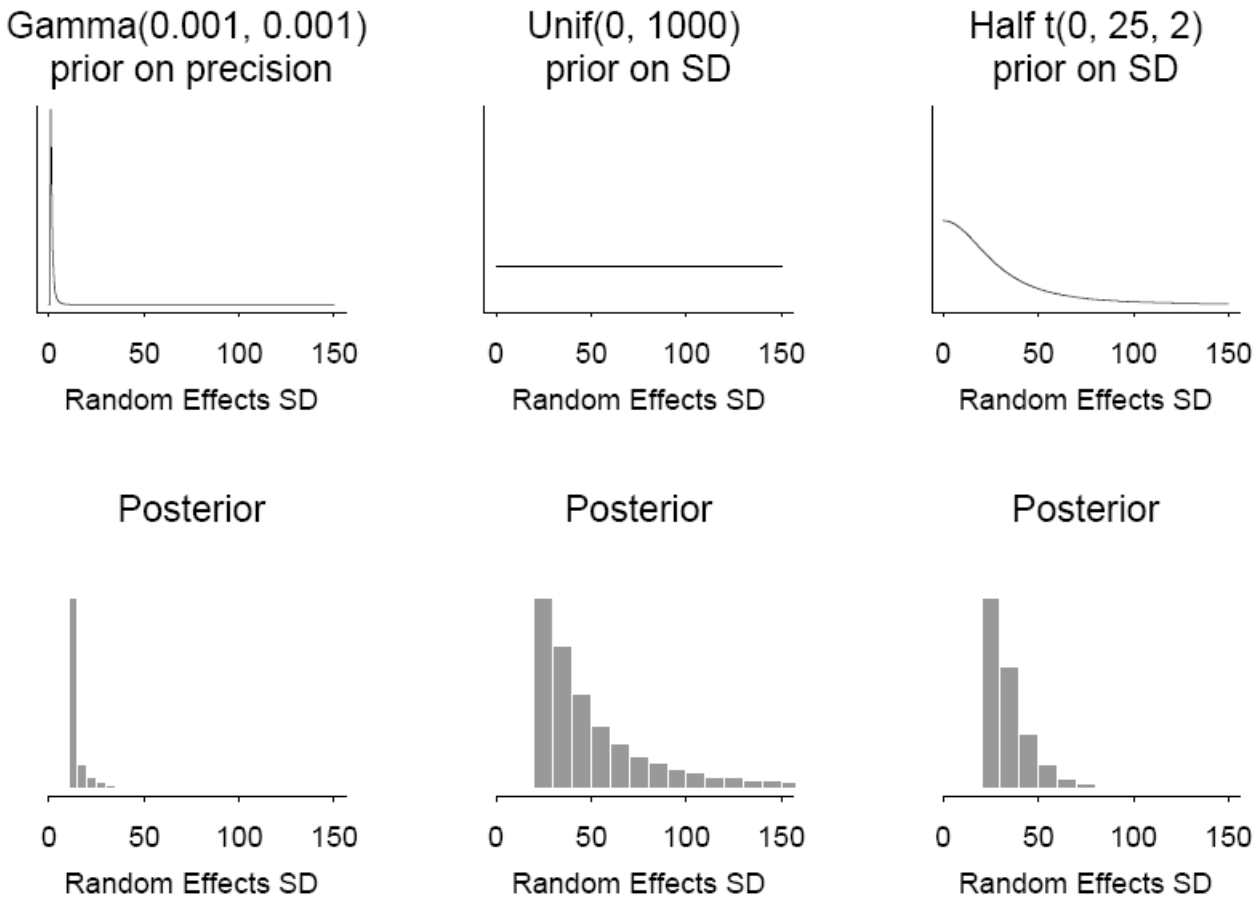
```
model {
 for(i in 1 : I) {
  O[i]  ~ dpois(mu[i])
  log(mu[i]) <- log(E[i]) + alpha + theta[i]
  theta[i] ~ dnorm(0, tau)  # area-specific random effects

  R[i] <- exp(alpha + theta[i])  # area-specific relative risk

  SMR[i] <- O[i] / E[i]     # Observed SMRs - this is just a function of
                            # the data, not of stochastic parameters, but
                            # is included in the model code so that values
                            # of SMR can be inspected and mapped
 }
 # Priors:
 alpha  ~ dflat()           # uniform prior on overall intercept

 tau ~ dgamma(0.001, 0.001) # prior on precison of area random effects
 sigma <- 1/sqrt(tau)       # between-area sd of random effects
}
```

Prior distributions for the variance of the random effects



Bibliography

Books

Banerjee S Carlin BP, Gelfrand To.E. (2004) Hierarchical Modelling and Analysis for Spatial Dates. Chapman & Hall /CRC. **Chapter 5**

Mollié To. (1999). Bayesian and Empirical Bayes Approaches Tone disease mapping in Disease Mapping and Risk Assessment for Public Health edited by Lawson