Day 5: Bayesian Modelling in R part 1

Ioannis Ntzoufras

Associate Professor in Statistics

E-mail: ntzoufras@aueb.gr

Contents

- **0… Software & Bibliography**
- **1… Introduction to Bayesian Inference**
- **2… Markov Chain Monte Carlo**
- **5… The normal linear model**
  - **The conjugate case**
  - **The Gibbs Sampler**
  - **Using the arm package**
  - **Using the MCMCpack**
- **6… R2WinBUGS**
- **7… Variable Selection Using BAS**

0… Bibliography
R, WinBUGS and Related Software

Stand Alone MCMC programs and R

- **WINBUGS 1.4.3:** http://www.mrc-bsu.cam.ac.uk/software/bugs/ **[`R2WinBUGS`].**
- **OpenBUGS:** http://www.openbugs.net/w/FrontPage

  [`BRugs`, `R2WinBUGS`].
- **JAGS (Just Another Gibbs Sampler):** http://mcmc-jags.sourceforge.net/ **[`rjags`].**
- **STAN:** http://mc-stan.org/ **[`RStan`].**

 3

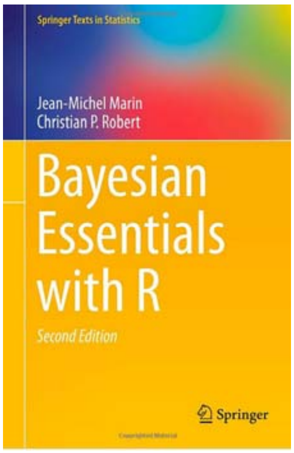
0… Bibliography
R, WinBUGS and Related Software

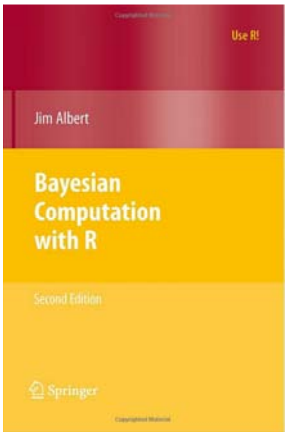
R packages

- **`MCMCpack`**: model-specific Markov chain Monte Carlo (MCMC) algorithms for wide range of models. Regression models and GLMs, measurement models (item response theory and factor models), changepoint models.
- **`arm`**: Bayesian inference using lm, glm, mer and polr objects.
- **`MCMCglmm`** : for fitting Generalised Linear Mixed Models using MCMC methods.
- **`BMA, BAS, BMS:`** Bayesian model averaging and variable selection for regression and glms
- **`Mombf:`** model selection based on non-local priors.
- **`BOA`** & **`CODA`**: MCMC Convergence diagnostics.

For more details and packages see http://cran.r-project.org/web/views/Bayesian.html

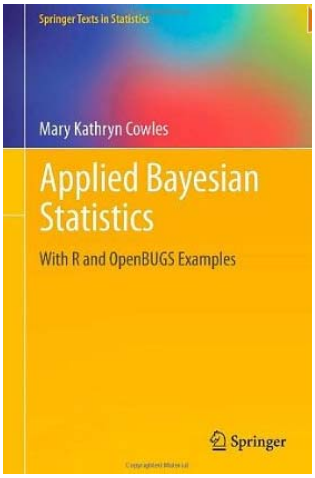
0… Bibliography
Bayesian Statistics with R

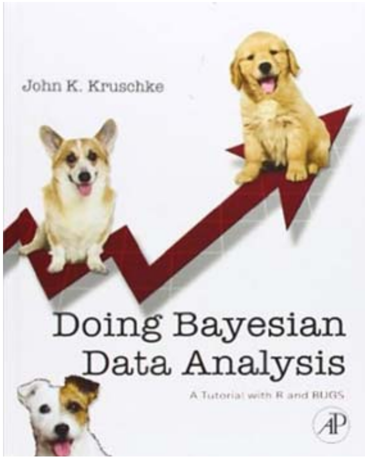




@ 2014, I. Ntzoufras @ Dep of Statistics, Athens University of Economics and Business 5

0… Bibliography
Bayesian Statistics with R and BUGS



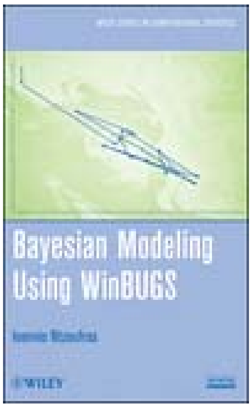


@ 2014, I. Ntzoufras @ Dep of Statistics, Athens University of Economics and Business 6

0… Bibliography
WinBUGS Books (1)

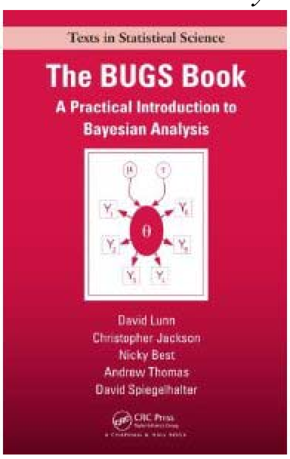
Ntzoufras, I. (2009). *Bayesian Modelling Using WinBUGS*. Wiley.

Book's web-site
http://stat-athens.aueb.gr/~jbn/winbugs_book



0… Bibliography
WinBUGS Books (2)

Lunn D., Jackson C., Best N., Thomas A. and Spiegelhalter D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Texts in Statistical Science, Chapman & Hall/CRC



0… Bibliography

Bayesian Data Analysis Books

- Gelman A., Carlin J.B., Sten H.S. and Rubin D.B. (2013). *Bayesian Data Analysis*. 3rd edition. London: Chapman and Hall.
- Carlin B. and Louis T. (2008). *Bayesian Methods for Data Analysis*. 3rd edition, London: Chapman and Hall.
- Christensen R., Johnson, W.O., Branscum A.J. and Hanson T.E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Chapman & Hall/CRC Texts in Statistical Science.
- Marin J.M. and Robert C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer Texts in Statistics.
- Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*, Wiley Series in Probability and Statistics, Wiley-Blackwell.
- Bolstad W.M. (2007). *Introduction to Bayesian Statistics*, 2nd Edition, Wiley-Blackwell.

0… Bibliography

Bayesian Modelling Books

- Books of P.D. Congdon:
  1. (2010). *Applied Bayesian Hierarchical Methods*. Chapman and Hall/CRC.
  2. (2007). *Bayesian Statistical Modelling*. 2nd Edition. Willey and Sons.
  3. (2003). *Applied Bayesian Modelling*. Wiley-Blackwell
  4. (2005). *Bayesian Models for Categorical Data*. Wiley-Blackwell.
- Gelman A. and Hill J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Analytical Methods for Social Research, Cambridge University Press.
- Dey D., Ghosh S.K. and Mallick B.K. (2000). *Generalized Linear Models: A Bayesian Perspective*, Chapman & Hall/CRC Biostatistics Series, CRC Press.

1…Introduction to Bayesian Inference

1.1. The Bayesian Paradigm

1.2. Posterior distribution

1.1 The Bayesian Paradigm

The usual classical approach

- is based on the likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$
- $\boldsymbol{\theta}$ parameter vector => unknown parameters that we wish to estimate
- Estimation of $\boldsymbol{\theta}$ is achieved via some estimators with some good statistical properties such as unbiasness
- Usually we obtain "good" estimators by maximising the likelihood function (maximum likelihood estimators or MLEs)
- EXAMPLE: for $Y_i \sim N(\mu,\sigma^2)$ we estimate μ using the sample mean given by $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

The Bayesian approach

Ladies and Gentlemen I present you

THE POSTERIOR DISTRIBUTION

$$f\,(\boldsymbol{\theta}|\boldsymbol{y})$$

The Bayesian approach

- Assumes that the parameters are random variables and not fixed unknowns.
- Specifies the prior distribution $f(\theta)$
- Inference is based on the posterior distribution $f\,(\boldsymbol{\theta}|\boldsymbol{y})$ which combines information coming from both the prior distribution and the likelihood (i.e. the data)

The Bayesian approach

Advantages

- Pure probability based approach
- Can incorporate information coming from experts or from previous studies (meta-analysis) via the prior.

Disadvantages

- Subjectivity (via the prior)
- Difficulties in computing or interpreting the posterior distribution

The Bayesian approach

Posterior distribution is calculated using BAYES THEOREM

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

$$\propto f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

Posterior $\propto$ Likelihood x Prior

[proportional]

A simple example: *Posterior distribution of the mean of the normal distribution*

1. Data/Likelihood: $Y_i \sim N(\mu, \sigma^2)$
   σ^2 here is assumed to be known and constant
2. Prior: $\mu \sim N(\mu_0, \sigma_0^2)$
3. Posterior:

$$f(\boldsymbol{\theta} \mid \mathbf{y}) = N\left(w\bar{y} + (1-w)\mu_0,\ w\frac{\sigma^2}{n} \right)$$

$$w = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2 / n}$$

 17

1.2. Posterior distribution

Analytical Calculation of the posterior distribution is sometimes difficult

- **1970s**: Conjugate priors resulting in posteriors of the same type (and known form)
- **1980s**: Asymptotic approximations of the posterior
- **1990s**: Obtaining random samples from the posterior using Markov Chain Monte Carlo (MCMC) methods.

 18

2. Markov Chain Monte Carlo (MCMC) Methods

Introduction

2.1. Metropolis-Hastings Algorithm

2.2. Gibbs Sampling

2. Markov Chain Monte Carlo (MCMC) Methods

Existed in the past in physics

- **1954** Metropolis *et al*. (Metropolis Algorithm)
- **1970** Hastings (Metropolis-Hastings Algorithm)
- **1984** Geman and Geman (Gibbs Sampling)
- **1990** Smith *et al*. (Implementation of MCMC methods in Bayesian problems)
- **1995** Green (Reversible Jump MCMC)

2. Markov Chain Monte Carlo (MCMC) Methods

What is the idea:

Since we cannot analytically calculate the posterior distribution then we generate a random sample from this distribution and estimate the posterior

- Describe the posterior using posterior summaries estimated by the generated sample (e.g. posterior mean or variance)
- Plot marginal posteriors
- Estimate posterior dependencies using sample correlations etc.

2. Markov Chain Monte Carlo (MCMC) Methods

The logic:

We construct a Markov chain which has a stationary distribution the posterior distribution of interest

Every iteration (step) of the algorithm depends only on the previous one.

We use this chain to "generate" a sample from the stationary (target) distribution

2. Markov Chain Monte Carlo (MCMC) Methods

The procedure

- We specify some arbitrary initial values $\theta^{(0)}$ for the parameters θ
- For t=1,2, …, T we generate random values $\theta^{(t)}$ according to our algorithm
- When the chain has *converged* then we have values from the stationary distribution
- We eliminate the initial K values to avoid any possible effect due to the arbitrary selection of initial values. (*Burn-in* period)

2. Markov Chain Monte Carlo (MCMC) Methods

Terminology

- **Initial values**: Starting values $\theta^{(0)}$ of the parameter vector θ. They are used to initialize the algorithm.
- **Iteration**: Refers to one iteration of the algorithm => to one observation of the generated sample
- **Burn–in Period**: The period (and the number of iterations) until the algorithm stabilizes and starts to give random values from the posterior distribution

2. Markov Chain Monte Carlo (MCMC) Methods

Terminology (2)

- **Convergence**: When the chain is giving values from the stationary (target) distribution
- **Convergence diagnostics**: Tests to assure convergence
- **MCMC output**: The simulated sample

2. Markov Chain Monte Carlo (MCMC) Methods

Terminology (3)

- MCMC algorithms are based on Markov chains
  => the generated sample is not IID
  => i.e. there is *autocorrelation* between the subsequently generated values (as in time series data)
- We are interested to eliminate this autocorrelation
  1. We monitor autocorrelations using ACF plots
  2. If there are significant ACs of order L
     => we keep 1 iteration every L
- **Thin**: is the number of iterations we eliminate in order to keep one iteration.

Thinning can be also used to save storing space.

2. Markov Chain Monte Carlo (MCMC) Methods

ALGORITHMS

- METROPOLIS-HASTINGS ALGORITHM
- GIBBS SAMPLING
- MANY OTHERS MORE ADVANCED (too much for this sort course)

2.1. Metropolis–Hastings Algorithm

- If we are in t iteration of the algorithm
  => set $\boldsymbol{\theta}^{cur}=\boldsymbol{\theta}^{(t-1)}$ i.e. the current values of $\boldsymbol{\theta}$.
- Generate a new proposed (or candidate) values $\boldsymbol{\theta}^{prop}$ from a proposal distribution $q(\boldsymbol{\theta}^{prop}|\boldsymbol{\theta}^{cur})$.
- Calculate $a=\min\left\{1,\dfrac{f(\boldsymbol{\theta}^{prop}\mid \boldsymbol{y})q(\boldsymbol{\theta}^{cur}\mid \boldsymbol{\theta}^{prop})}{f(\boldsymbol{\theta}^{cur}\mid \boldsymbol{y})q(\boldsymbol{\theta}^{prop}\mid \boldsymbol{\theta}^{cur})}\right\}$
- Set $\boldsymbol{\theta}^{(t)}=\boldsymbol{\theta}^{prop}$ with probability α και $\boldsymbol{\theta}^{(t)}=\boldsymbol{\theta}^{cur}$ with probability $(1-\alpha)$

2.1. Metropolis–Hastings Algorithm

- Note that for the calculation of α we do not need to know the normalizing constant since

$$a = \min\left\{1, \frac{f(\boldsymbol{\theta}^{prop} \mid \boldsymbol{y}) q(\boldsymbol{\theta}^{cur} \mid \boldsymbol{\theta}^{prop})}{f(\boldsymbol{\theta}^{cur} \mid \boldsymbol{y}) q(\boldsymbol{\theta}^{prop} \mid \boldsymbol{\theta}^{cur})}\right\}$$

$$= \min\left\{1, \frac{\{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{prop}) f(\boldsymbol{\theta}^{prop}) / f(\boldsymbol{y})\} q(\boldsymbol{\theta}^{cur} \mid \boldsymbol{\theta}^{prop})}{\{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{cur}) f(\boldsymbol{\theta}^{cur}) / f(\boldsymbol{y})\} q(\boldsymbol{\theta}^{prop} \mid \boldsymbol{\theta}^{cur})}\right\}$$

$$= \min\left\{1, \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{prop}) f(\boldsymbol{\theta}^{prop}) q(\boldsymbol{\theta}^{cur} \mid \boldsymbol{\theta}^{prop})}{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{cur}) f(\boldsymbol{\theta}^{cur}) q(\boldsymbol{\theta}^{prop} \mid \boldsymbol{\theta}^{cur})}\right\}$$

2.1. Metropolis–Hastings Algorithm

- Note that for the calculation of α we do not need to know the normalizing constant since

$$a = \min\left\{1, \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{prop}) f(\boldsymbol{\theta}^{prop}) q(\boldsymbol{\theta}^{cur} \mid \boldsymbol{\theta}^{prop})}{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{cur}) f(\boldsymbol{\theta}^{cur}) q(\boldsymbol{\theta}^{prop} \mid \boldsymbol{\theta}^{cur})}\right\}$$

α depends on

- The likelihood
- The prior
- The proposal

2.1. Metropolis–Hastings Algorithm

Random walk Metropolis

- Usual choice for the proposal:

$$q(\theta^{prop} \mid \theta^{cur}) = N(\theta^{cur}, c^2).$$

- We propose a new value θ^{prop} with mean equal to the current value of the chain and variance controlled by c^2.
- c^2 is also called ***tuning parameter*** since it affects the convergence of the chain and must be tuned appropriately.
- The acceptance probability is simplified to

$$a = \min\left\{1, \frac{f(\boldsymbol{\theta}^{prop} \mid \boldsymbol{y})}{f(\boldsymbol{\theta}^{cur} \mid \boldsymbol{y})}\right\} = \min\left\{1, \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{prop}) f(\boldsymbol{\theta}^{prop})}{f(\boldsymbol{y} \mid \boldsymbol{\theta}^{cur}) f(\boldsymbol{\theta}^{cur})}\right\}$$

  due to the symmetry of the proposal

2.1. Metropolis–Hastings Algorithm

Random walk Metropolis

Tuning of c^2

It affects the convergence of the chain and must be tuned appropriately.

- Small values make the chain to move slowly
  => Propose values very close to the current values
  => accept them with high probability
  => High autocorrelations
- Large values make the chain to move less but with bigger moves
  => Propose values away from the current values
  => reject them with high probability
  => The chain may stack to the same set of values for a long time
  => High autocorrelations

2.1. Metropolis–Hastings Algorithm

Random walk Metropolis

Tuning of c^2 – Optimal acceptance

- Roberts et al. (1997), Neal and Roberts (2008)
  - 23% for multidimensional problems
  - 45% for univariate cases
- Any choice of c^2 from 20–40% should be fine

  "there is little to be gained by fine tuning of acceptance rates"

  (Roberts and Rosental, 2001)

2.2. Gibbs Sampling

- If we are in *t* iteration of the algorithm

  => set $\boldsymbol{\theta}^{cur}=\boldsymbol{\theta}^{(t-1)}$ i.e. the current values of $\boldsymbol{\theta}$.

  $\boldsymbol{\theta}^{cur}=(\theta_1^{cur},\theta_2^{cur}, \dots ,\theta_p^{cur})$
- Generate θ_1^{new} from $f(\theta_1|\theta_2^{cur},\dots, \theta_p^{cur},\mathbf{y})$
- Generate θ_2^{new} from $f(\theta_2|\theta_1^{new},\theta_3^{cur},\dots, \theta_p^{cur},\mathbf{y})$
- … … … … … … … … … … … … … … … …
- Generate θ_j^{new} from $f(\theta_j|\theta_1^{new},\dots,\theta_{j-1}^{new},\theta_{j+1}^{cur},\dots, \theta_p^{cur},\mathbf{y})$
- … … … … … … … … … … … … … … … …
- Generate θ_p^{new} from $f(\theta_p|\theta_1^{new},\dots, \theta_{p-1}^{new},\mathbf{y})$
- Set $\boldsymbol{\theta}^{(t)}=\boldsymbol{\theta}^{new}$

2.2. Gibbs Sampling

$f(\theta_j|\theta_1,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_p,\mathbf{y})$

- is called the full conditional of the posterior distribution
- it is frequently denoted by $f(\theta_j|\bullet)$ or $f(\theta_j|rest)$

2.2. Gibbs Sampling

Differences with Metropolis-Hastings algorithm

- $\boldsymbol{\theta}^{(t-1)} \neq \boldsymbol{\theta}^{(t)}$ – A new set of values is always generated
- The Gibbs sampler is a special case of MH with proposal $q() = f(\theta_j|\bullet)$
- Every time we update one parameter at a time (or a block of parameters)

2.2. Gibbs Sampling

$f(\theta_j|\bullet)$ may be unknown

- Use adaptive rejection sampling for log-convave distributions (Gilks & Wild, 1992)
- For generalized linear models (GLMs), posterior distributions are log-concave (Dellaportas & Smith, 1993)
- This is the main approach used in WinBUGS
- Metropolis steps for the unknown conditionals can be used

2.2. Gibbs Sampling

Advantages

- Simple to implement
- No tuning – automatic

Disadvantages

- Need to calculate conditional posteriors
- Some conditional posteriors may not be available
- No flexibility if high autocorrelations exist

2.2. Gibbs Sampling

Gibbs sampling for a Normal regression model

$$Y_i \sim N(\mu_i, \sigma^2) \text{ for } i=1,2,\ldots,n$$
$$\mu_i = \alpha + \beta X_i$$
$$\boldsymbol{\theta}=(\alpha, \beta, \sigma^2)^T$$

- **PRIORS**:

$$f(\boldsymbol{\theta})=f(\alpha, \beta, \sigma^2)= f(\alpha)\, f(\beta)\, f(\sigma^2)$$

  - $f(\alpha) \sim Normal(\mu_\alpha , \sigma_\alpha^2)$
  - $f(\beta) \sim Normal(\mu_\beta , \sigma_\beta^2)$
  - $f(\sigma^2) \sim Inverse\ Gamma(\gamma, \delta)$

    $\Rightarrow f(\tau)=Gamma(\gamma, \delta)$ for $\tau=1/\sigma^2$

Gibbs Sampling for normal regression

Full Conditional Posteriors

- $$\alpha \mid \beta, \sigma^2, \boldsymbol{y} \sim N\left(w_1(\bar{y}-b\bar{x})+(1-w_1)\mu_\alpha,\ w_1\frac{\sigma^2}{n}\right)$$

- $$\beta \mid \alpha, \sigma^2, \boldsymbol{y} \sim N\left(w_2\frac{\sum_{i=1}^n x_i y_i - an\bar{x}}{\sum_{i=1}^n x_i^2}+(1-w_2)\mu_\beta,\ w_2\frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$$

$$w_1=\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma^2/n} \qquad w_2=\frac{\sigma_\beta^2}{\sigma_\beta^2+\sigma^2/\sum_{i=1}^n x_i^2}$$

40

Gibbs Sampling for normal regression

Full Conditional Posteriors

$$\sigma^2 \mid \alpha, \beta, \boldsymbol{y} \sim \text{Inverse Gamma}\left(\frac{n}{2} + \gamma,\ \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)^2 + \delta \right)$$