Bayesian Statistical Modelling

Second Edition

PETER CONGDON
Queen Mary, University of London, UK

John Wiley & Sons, Ltd

Bayesian Statistical Modelling

WILEY SERIES IN PROBABILITY AND STATISTICS
ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Bayesian Statistical Modelling

Second Edition

PETER CONGDON
Queen Mary, University of London, UK

John Wiley & Sons, Ltd

Copyright © 2006 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The Publisher is not associated with any product or vendor mentioned in this book.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, Ontario, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-01875-0 (HB)
ISBN-10 0-470-01875-5 (HB)

Typeset in 10/12pt Times by TechBooks, New Delhi, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

Contents

Preface

This book updates the 1st edition of Bayesian Statistical Modelling and, like its predecessor, seeks to provide an overview of modelling strategies and data analytic methodology from a Bayesian perspective. The book discusses and reviews a wide variety of modelling and application areas from a Bayesian viewpoint, and considers the most recent developments in what is often a rapidly changing intellectual environment.

The particular package that is mainly relied on for illustrative examples in this 2nd edition is again WINBUGS (and its parallel development in OPENBUGS). In the author's experience this remains a highly versatile tool for applying Bayesian methodology. This package allows effort to be focused on exploring alternative likelihood models and prior assumptions, while detailed specification and coding of parameter sampling mechanisms (whether Gibbs or Metropolis-Hastings) can be avoided – by relying on the program's inbuilt expert system to choose appropriate updating schemes.

In this way relatively compact and comprehensible code can be applied to complex problems, and the focus centred on data analysis and alternative model structures. In more general terms, providing computing code to replicate proposed new methodologies can be seen as an important component in the transmission of statistical ideas, along with data replication to assess robustness of inferences in particular applications.

I am indebted to the help of the Wiley team in progressing my book. Acknowledgements are due to the referee, and to Sylvia Fruhwirth-Schnatter and Nial Friel for their comments that helped improve the book.

Any comments may be addressed to me at p.congdon@qmul.ac.uk. Data and programs can be obtained at ftp://ftp.wiley.co.uk/pub/books/congdon/Congdon_BSM_2006.zip and also at Statlib, and at www.geog.qmul.ac.uk/staff/congdon.html. Winbugs can be obtained from http://www.mrc-bsu.cam.ac.uk/bugs, and Openbugs from http://mathstat.helsinki.fi/openbugs/.

Peter Congdon
Queen Mary, University of London
November 2006

CHAPTER 1

Introduction: The Bayesian Method, its Benefits and Implementation

1.1 THE BAYES APPROACH AND ITS POTENTIAL ADVANTAGES

Bayesian estimation and inference has a number of advantages in statistical modelling and data analysis. For example, the Bayes method provides confidence intervals on parameters and probability values on hypotheses that are more in line with commonsense interpretations. It provides a way of formalising the process of learning from data to update beliefs in accord with recent notions of knowledge synthesis. It can also assess the probabilities on both nested and non-nested models (unlike classical approaches) and, using modern sampling methods, is readily adapted to complex random effects models that are more difficult to fit using classical methods (e.g. Carlin *et al.*, 2001).

However, in the past, statistical analysis based on the Bayes theorem was often daunting because of the numerical integrations needed. Recently developed computer-intensive sampling methods of estimation have revolutionised the application of Bayesian methods, and such methods now offer a comprehensive approach to complex model estimation, for example in hierarchical models with nested random effects (Gilks *et al.*, 1993). They provide a way of improving estimation in sparse datasets by borrowing strength (e.g. in small area mortality studies or in stratified sampling) (Richardson and Best 2003; Stroud, 1994), and allow finite sample inferences without appeal to large sample arguments as in maximum likelihood and other classical methods. Sampling-based methods of Bayesian estimation provide a full density profile of a parameter so that any clear non-normality is apparent, and allow a range of hypotheses about the parameters to be simply assessed using the collection of parameter samples from the posterior.

Bayesian methods may also improve on classical estimators in terms of the precision of estimates. This happens because specifying the prior brings extra information or data based on accumulated knowledge, and the posterior estimate in being based on the combined sources of information (prior and likelihood) therefore has greater precision. Indeed a prior can often be expressed in terms of an equivalent ‘sample size’.

Bayesian Statistical Modelling. Second Edition P. Congdon

Bayesian analysis offers an alternative to classical tests of hypotheses under which p-values are framed in the data space: the p-value is the probability under hypothesis H of data at least as extreme as that actually observed. Many users of such tests more naturally interpret p-values as relating to the hypothesis space, i.e. to questions such as the likely range for a parameter given the data, or the probability of H given the data. The Bayesian framework is more naturally suited to such probability interpretations. The classical theory of confidence intervals for parameter estimates is also not intuitive, saying that in the long run with data from many samples a 95% interval calculated from each sample will contain the true parameter approximately 95% of the time. The particular confidence interval from any one sample may or may not contain the true parameter value. By contrast, a 95% Bayesian credible interval contains the true parameter value with approximately 95% certainty.

1.2 EXPRESSING PRIOR UNCERTAINTY ABOUT PARAMETERS AND BAYESIAN UPDATING

The learning process involved in Bayesian inference is one of modifying one's initial probability statements about the parameters before observing the data to updated or posterior knowledge that combines both prior knowledge and the data at hand. Thus prior subject-matter knowledge about a parameter (e.g. the incidence of extreme political views or the relative risk of thrombosis associated with taking the contraceptive pill) is an important aspect of the inference process. Bayesian models are typically concerned with inferences on a parameter set $\theta = (\theta_1, \ldots, \theta_d)$, of dimension d, that includes uncertain quantities, whether fixed and random effects, hierarchical parameters, unobserved indicator variables and missing data (Gelman and Rubin, 1996). Prior knowledge about the parameters is summarised by the density $p(\theta)$, the likelihood is $p(y|\theta)$, and the updated knowledge is contained in the posterior density $p(\theta|y)$. From the Bayes theorem

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y), \qquad (1.1)$$

where the denominator on the right side is the marginal likelihood $p(y)$. The latter is an integral over all values of θ of the product $p(y|\theta)p(\theta)$ and can be regarded as a normalising constant to ensure that $p(\theta|y)$ is a proper density. This means one can express the Bayes theorem as

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

The relative influence of the prior and data on updated beliefs depends on how much weight is given to the prior (how 'informative' the prior is) and the strength of the data. For example, a large data sample would tend to have a predominant influence on updated beliefs unless the prior was informative. If the sample was small and combined with a prior that was informative, then the prior distribution would have a relatively greater influence on the updated belief: this might be the case if a small clinical trial or observational study was combined with a prior based on a meta-analysis of previous findings.

How to choose the prior density or information is an important issue in Bayesian inference, together with the sensitivity or robustness of the inferences to the choice of prior, and the possibility of conflict between prior and data (Andrade and O'Hagan, 2006; Berger, 1994).

Table 1.1 Deriving the posterior distribution of a prevalence rate π using a discrete prior

Possible π values	Prior weight given to different possible values of π	Likelihood of data given value for π	Prior times likelihood	Posterior probabilities
0.10	0.10	0.267	0.027	0.098
0.12	0.15	0.287	0.043	0.157
0.14	0.25	0.290	0.072	0.265
0.16	0.25	0.279	0.070	0.255
0.18	0.15	0.258	0.039	0.141
0.20	0.10	0.231	0.023	0.084
Total	1		0.274	1

In some situations it may be possible to base the prior density for θ on cumulative evidence using a formal or informal meta-analysis of existing studies. A range of other methods exist to determine or elicit subjective priors (Berger, 1985, Chapter 3; Chaloner, 1995; Garthwaite *et al.*, 2005; O'Hagan, 1994, Chapter 6). A simple technique known as the histogram method divides the range of θ into a set of intervals (or 'bins') and elicits prior probabilities that θ is located in each interval; from this set of probabilities, $p(\theta)$ may be represented as a discrete prior or converted to a smooth density. Another technique uses prior estimates of moments along with symmetry assumptions to derive a normal $N(m, V)$ prior density including estimates m and V of the mean and variance. Other forms of prior can be reparameterised in the form of a mean and variance (or precision); for example beta priors $\text{Be}(a, b)$ for probabilities can be expressed as $\text{Be}(m\tau, (1-m)\tau)$ where m is an estimate of the mean probability and τ is the estimated precision (degree of confidence in) that prior mean.

To illustrate the histogram method, suppose a clinician is interested in π, the proportion of children aged 5–9 in a particular population with asthma symptoms. There is likely to be prior knowledge about the likely size of π, based on previous studies and knowledge of the host population, which can be summarised as a series of possible values and their prior probabilities, as in Table 1.1. Suppose a sample of 15 patients in the target population shows 2 with definitive symptoms. The likelihoods of obtaining 2 from 15 with symptoms according to the different values of π are given by $({}^{15}{}_{2})\pi^2(1-\pi)^{13}$, while posterior probabilities on the different values are obtained by dividing the product of the prior and likelihood by the normalising factor of 0.274. They give highest support to a value of $\pi = 0.14$. This inference rests only on the prior combined with the likelihood of the data, namely 2 from 15 cases. Note that to calculate the posterior weights attaching to different values of π, one need use only that part of the likelihood in which π is a variable: instead of the full binomial likelihood, one may simply use the likelihood kernel $\pi^2(1-\pi)^{13}$ since the factor $({}^{15}{}_{2})$ cancels out in the numerator and denominator of Equation (1.1).

Often, a prior amounts to a form of modelling assumption or hypothesis about the nature of parameters, for example, in random effects models. Thus small area mortality models may include spatially correlated random effects, exchangeable random effects with no spatial pattern or both. A prior specifying the errors as spatially correlated is likely to be a working model assumption, rather than a true cumulation of knowledge.

In many situations, existing knowledge may be difficult to summarise or elicit in the form of an 'informative prior', and to reflect such essentially prior ignorance, resort is made to non-informative priors. Since the maximum likelihood estimate is not influenced by priors, one possible heuristic is that a non-informative prior leads to a Bayesian posterior mean very close to the maximum likelihood estimate, and that informativeness of priors can be assessed by how closely the Bayesian estimate comes to the maximum likelihood estimate.

Examples of priors intended to be non-informative are flat priors (e.g. that a parameter is uniformly distributed between $-\infty$ and $+\infty$, or between 0 and $+\infty$), reference priors (Berger and Bernardo, 1994) and Jeffreys' prior

$$p(\theta) \propto |I(\theta)|^{0.5},$$

where $I(\theta)$ is the information[1] matrix. Jeffreys' prior has the advantage of invariance under transformation, a property not shared by uniform priors (Syverseen, 1998). Other advantages are discussed by Wasserman (2000). Many non-informative priors are improper (do not integrate to 1 over the range of possible values). They may also actually be unexpectedly informative about different parameter values (Zhu and Lu, 2004). Sometimes improper priors can lead to improper posteriors, as in a normal hierarchical model with subjects j nested in clusters i,

$$\begin{aligned} y_{ij} &\sim N(\theta_i, \sigma^2), \\ \theta_i &\sim N(\mu, \tau^2). \end{aligned}$$

The prior $p(\mu, \tau) = 1/\tau$ results in an improper posterior (Kass and Wasserman, 1996). Examples of proper posteriors despite improper priors are considered by Fraser *et al.* (1997) and Hadjicostas and Berry (1999).

To guarantee posterior propriety (at least analytically) a possibility is to assume just proper priors (sometimes called diffuse or weakly informative priors); for example, a gamma Ga(1, 0.00001) prior on a precision (inverse variance) parameter is proper but very close to being a flat prior. Such priors may cause identifiability problems and impede Markov Chain Monte Carlo (MCMC) convergence (Gelfand and Sahu, 1999; Kass and Wasserman, 1996, p. 1361). To adequately reflect prior ignorance while avoiding impropriety, Spiegelhalter *et al.* (1996, p. 28) suggest a prior standard deviation at least an order of magnitude greater than the posterior standard deviation.

In Table 1.1 an informative prior favouring certain values of π has been used. A non-informative prior, favouring no values above any other, would assign an equal prior probability of $1/6$ to each of the possible prior values of π. A non-informative prior might be used in the genuine absence of prior information, or if there is disagreement about the likely values of hypotheses or parameters. It may also be used in comparison with more informative priors as one aspect of a sensitivity analysis regarding posterior inferences according to the prior. Often some prior information is available on a parameter or hypothesis, though converting it into a probabilistic form remains an issue. Sometimes a formal stage of eliciting priors from subject-matter specialists is entered into (Osherson *et al.*, 1995).

[1] If $\ell(\theta) = \log(L(\theta|y))$ is the likelihood, then $I(\theta) = -E\left\{\frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_i}\right\}$.

If a previous study or set of studies is available on the likely prevalence of asthma in the population, these may be used in a form of preliminary meta-analysis to set up an informative prior for the current study. However, there may be limits to the applicability of previous studies to the current target population (e.g. because of differences in the socio-economic background or features of the local environment). So the information from previous studies, while still usable, may be downweighted; for example, the precision (variance) of an estimated relative risk or prevalence rate from a previous study may be divided (multiplied) by 10. If there are several parameters and their variance–covariance matrix is known from a previous study or a mode-finding analysis (e.g. maximum likelihood), then this can be downweighted in the same way (Birkes and Dodge, 1993). More comprehensive ways of downweighting historical/prior evidence have been proposed, such as power prior models (Ibrahim and Chen, 2000).

In practice, there are also mathematical reasons to prefer some sorts of priors to others (the question of conjugacy is considered in Chapter 3). For example, a beta density for the binomial success probability is conjugate with the binomial likelihood in the sense that the posterior has the same (beta) density form as the prior. However, one advantage of sampling-based estimation methods is that a researcher is no longer restricted to conjugate priors, whereas in the past this choice was often made for reasons of analytic tractability. There remain considerable problems in choosing appropriate neutral or non-informative priors on certain types of parameters, with variance and covariance hyperparameters in random effects models a leading example (Daniels, 1999; Gelman, 2006; Gustafson *et al.*, in press).

To assess sensitivity to the prior assumptions, one may consider the effects on inference of a limited range of alternative priors (Gustafson, 1996), or adopt a ‘community of priors’ (Spiegelhalter *et al.*, 1994); for example, alternative priors on a treatment effect in a clinical trial might be neutral, sceptical, and enthusiastic with regard to treatment efficacy. One might also consider more formal approaches to robustness based on non-parametric priors rather than parametric priors, or via mixture (‘contamination’) priors. For instance, one might assume a two-group mixture with larger probability $1 - q$ on the ‘main’ prior $p_1(\theta)$, and a smaller probability such as $q = 0.2$ on a contaminating density $p_2(\theta)$, which may be any density (Gustafson, 1996). One might consider the contaminating prior to be a flat reference prior, or one allowing for shifts in the main prior’s assumed parameter values (Berger, 1990). In large datasets, inferences may be robust to changes in prior unless priors are heavily informative. However, inference sensitivity may be greater for some types of parameters, even in large datasets; for example, inferences may depend considerably on the prior adopted for variance parameters in random effects models, especially in hierarchical models where different types of random effects coexist in a model (Daniels, 1999; Gelfand *et al.*, 1996).

1.3 MCMC SAMPLING AND INFERENCES FROM POSTERIOR DENSITIES

Bayesian inference has become closely linked to sampling-based estimation methods. Both focus on the entire density of a parameter or functions of parameters. Iterative Monte Carlo methods involve repeated sampling that converges to sampling from the posterior distribution. Such sampling provides estimates of density characteristics (moments, quantiles), or of probabilities relating to the parameters (Smith and Gelfand, 1992). Provided with a reasonably large sample from a density, its form can be approximated via curve estimation (kernel density) methods; default bandwidths are suggested by Silverman (1986), and included in implementations such as the Stixbox Matlab library (pltdens.m from http://www.maths.lth.se/matstat/stixbox). There is no limit to the number of samples T of θ that may be taken from a posterior density $p(\theta|y)$, where $\theta = (\theta_1, \ldots, \theta_k, \ldots, \theta_d)$ is of dimension d. The larger is T from a single sampling run, or the larger is $T = T_1 + T_2 + \cdots + T_J$ based on J sampling chains from the density, the more accurately the posterior density would be described.

Monte Carlo posterior summaries typically include posterior means and variances of the parameters. This is equivalent to estimating the integrals

$$E(\theta_k|y) = \int \theta_k p\,(\theta|y)\mathrm{d}\theta, \tag{1.2}$$

$$\begin{aligned}\mathrm{Var}(\theta_k|y) &= \int \theta_k^2 p\,(\theta|y)\mathrm{d}\theta - [E(\theta_k|y)]^2 \\ &= E\left(\theta_k^2|y\right) - [E(\theta_k|y)]^2. \end{aligned} \tag{1.3}$$

Which estimator $d = \theta_e(y)$ to choose to characterise a particular function of θ can be decided with reference to the Bayes risk under a specified loss function $L[d, \theta]$ (Zellner, 1985, p. 262),

$$\min_d \int L[d, \theta]p(y|\theta)p(\theta)\mathrm{d}\theta,$$

or equivalently

$$\min_d \int L[d, \theta]p(\theta|y)\mathrm{d}\theta.$$

The posterior mean can be shown to be the best estimate of central tendency for a density under a squared error loss function (Robert, 2004), while the posterior median is the best estimate when absolute loss is used, namely $L[\theta_e(y), \theta] = |\theta_e - \theta|$. Similar principles can be applied to parameters obtained via model averaging (Brock *et al.*, 2004).

A $100(1-\alpha)\%$ credible interval for θ_k is any interval $[a, b]$ of values that has probability $1-\alpha$ under the posterior density of θ_k. As noted above, it is valid to say that there is a probability of $1-\alpha$ that θ_k lies within the range $[a, b]$. Suppose $\alpha = 0.05$. Then the most common credible interval is the equal-tail credible interval, using 0.025 and 0.975 quantiles of the posterior density. If one is using an MCMC sample to estimate the posterior density, then the 95% CI is estimated using the 0.025 and 0.975 quantiles of the sampled output $\{\theta_k^{(t)}, t = B+1, \ldots, T\}$ where B is the number of burn-in iterations (see Section 1.5). Another form of credible interval is the $100(1-\alpha)\%$ highest probability density (HPD) interval, such that the density for every point inside the interval exceeds that for every point outside the interval, and is the shortest possible $100(1-\alpha)\%$ credible interval; Chen *et al.* (2000, p. 219) provide an algorithm to estimate the HPD interval. A program to find the HPD interval is included in the Matlab suite of MCMC diagnostics developed at the Helsinki University of Technology, at http://www.lce.hut.fi/research/compinf/mcmcdiag/.

One may similarly obtain posterior means, variances and credible intervals for functions $\Delta = \Delta(\theta)$ of the parameters (van Dyk, 2002). The posterior means and variances of such functions obtained from MCMC samples are estimates of the integrals

$$\begin{aligned} E[\Delta(\theta)|y] &= \int \Delta(\theta)p(\theta|y)\mathrm{d}\theta, \\ \mathrm{var}[\Delta(\theta)|y] &= \int \Delta^2 p(\theta|y)\mathrm{d}\theta - [E(\Delta|y)]^2 \\ &= E(\Delta^2|y) - [E(\Delta|y)]^2. \end{aligned} \tag{1.4}$$

Often the major interest is in marginal densities of the parameters themselves. The marginal density of the kth parameter θ_k is obtained by integrating out all other parameters

$$p(\theta_k|y) = \int p(\theta|y)\mathrm{d}\theta_1\mathrm{d}\theta_2\cdots d\theta_{k-1}d\theta_{k+1}d\theta_d.$$

Posterior probability estimates from an MCMC run might relate to the probability that θ_k (say $k = 1$) exceeds a threshold b, and provide an estimate of the integral

$$\Pr(\theta_1 > b|y) = \int_b^\infty \int_{..} \int p(\theta|y)\mathrm{d}\theta. \tag{1.5}$$

For example, the probability that a regression coefficient exceeds zero or is less than zero is a measure of its significance in the regression (where significance is used as a shorthand for ‘necessary to be included’). A related use of probability estimates in regression (Chapter 4) is when binary inclusion indicators precede the regression coefficient and the regressor is included only when the indicator is 1. The posterior probability that the indicator is 1 estimates the probability that the regressor should be included in the regression.

Such expectations, density or probability estimates may sometimes be obtained analytically for conjugate analyses – such as a binomial likelihood where the probability has a beta prior. They can also be approximated analytically by expanding the relevant integral (Tierney *et al.*, 1988). Such approximations are less good for posteriors that are not approximately normal, or where there is multimodality. They also become impractical for complex multiparameter problems and random effects models.

By contrast, MCMC techniques are relatively straightforward for a range of applications, involving sampling from one or more chains after convergence to a stationary distribution that approximates the posterior $p(\theta|y)$. If there are n observations and d parameters, then the required number of iterations to reach stationarity will tend to increase with both d and n, and also with the complexity of the model (e.g. which depends on the number of levels in a hierarchical model, or on whether a nonlinear rather than a simple linear regression is chosen). The ability of MCMC sampling to cope with complex estimation tasks should be qualified by mention of problems associated with long-run sampling as an estimation method. For example, Cowles and Carlin (1996) highlight problems that may occur in obtaining and/or assessing convergence (see Section 1.5). There are also problems in setting neutral priors on certain types of parameters (e.g. variance hyperparameters in models with nested random effects), and certain types of models (e.g. discrete parametric mixtures) are especially subject to identifiability problems (Frühwirth-Schnatter, 2004; Jasra *et al.*, 2005).

A variety of MCMC methods have been proposed to sample from posterior densities (Section 1.4). They are essentially ways of extending the range of single-parameter sampling methods to multivariate situations, where each parameter or subset of parameters in the overall posterior density has a different density. Thus there are well-established routines for computer generation of random numbers from particular densities (Ahrens and Dieter, 1974; Devroye, 1986). There are also routines for sampling from non-standard densities such as non-log-concave densities (Gilks and Wild, 1992). The usual Monte Carlo method assumes a sample of independent simulations $u^{(1)}, u^{(2)}, \ldots, u^{(T)}$ from a target density $\pi(u)$ whereby $E[g(u)] = \int g(u)\pi(u)\mathrm{d}u$ is estimated as

$$\overline{g}_T = \sum_{t=1}^{T} g\big(u^{(t)}\big).$$

With probability 1, $\overline{g}_T$ tends to $E_\pi[g(u)]$ as $T \to \infty$. However, independent sampling from the posterior density $p(\theta\ |y)$ is not feasible in general. It is valid, however, to use dependent samples $\theta^{(t)}$, provided the sampling satisfactorily covers the support of $p(\theta\ |y)$ (Gilks *et al.*, 1996).

In order to sample approximately from $p(\theta|y)$, MCMC methods generate dependent draws via Markov chains. Specifically, let $\theta^{(0)}$, $\theta^{(1)}$, ... be a sequence of random variables. Then $p(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(T)})$ is a Markov chain if

$$p\big(\theta^{(t)}|\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t-1)}\big) = p\big(\theta^{(t)}|\theta^{(t-1)}\big),$$

so that only the preceding state is relevant to the future state. Suppose $\theta^{(t)}$ is defined on a discrete state space $S = \{s_1, s_2, \ldots\}$, with generalisation to continuous state spaces described by Tierney (1996). Assume $p(\theta^{(t)}|\theta^{(t-1)})$ is defined by a constant one-step transition matrix

$$Q_{i,j} = \Pr\big(\theta^{(t)} = s_j|\theta^{(t-1)} = s_i\big),$$

with t-step transition matrix $Q_{i,j}(t) = \Pr(\theta^{(t)} = s_j|\theta^{(0)} = s_i)$. Sampling from a constant one-step Markov chain converges to the stationary distribution required, namely $\pi(\theta) = p(\theta|y)$, if additional requirements[2] on the chain are satisfied (irreducibility, aperiodicity and positive recurrence) – see Roberts (1996, p. 46) and Norris (1997). Sampling chains meeting these requirements have a unique stationary distribution $\lim_{t\to\infty} Q_{i,j}(t) = \pi_{(j)}$ satisfying the full balance condition $\pi_{(j)} = \sum_i \pi_{(i)} Q_{i,j}$. Many Markov chain methods are additionally reversible, meaning $\pi_{(i)} Q_{i,j} = \pi_{(j)} Q_{j,i}$.

With this type of sampling mechanism, the ergodic average $\overline{g}_T$ tends to $E_\pi[g(u)]$ with probability 1 as $T \to \infty$ despite dependent sampling. Remaining practical questions include establishing an MCMC sampling scheme and establishing that convergence to a steady state has been obtained for practical purposes (Cowles and Carlin, 1996). Estimates of quantities such as (1.2) and (1.3) are routinely obtained from sampling output along with 2.5th and

[2] Suppose a chain is defined on a space S. A chain is irreducible if for any pair of states $(s_i, s_j) \in S$ there is a non-zero probability that the chain can move from s_i to s_j in a finite number of steps. A state is positive recurrent if the number of steps the chain needs to revisit the state has a finite mean. If all the states in a chain are positive recurrent then the chain itself is positive recurrent. A state has period k if it can be revisited only after the number of steps that is a multiple of k. Otherwise the state is aperiodic. If all its states are aperiodic then the chain itself is aperiodic. Positive recurrence and aperiodicity together constitute ergodicity.

97.5th percentiles that provide equal-tail credible intervals for the value of the parameter. A full posterior density estimate may also be derived (e.g. by kernel smoothing of the MCMC output of a parameter). For $\Delta(\theta)$ its posterior mean is obtained by calculating $\Delta^{(t)}$ at every MCMC iteration from the sampled values $\theta^{(t)}$. The theoretical justification for this is provided by the MCMC version of the law of large numbers (Tierney, 1994), namely that

$$\sum_{t=1}^{T} \frac{\Delta\left(\theta^{(t)}\right)}{T} \rightarrow E_\pi[\Delta(\theta)],$$

provided that the expectation of $\Delta(\theta)$ under $\pi(\theta) = p(\theta|y)$, denoted by $E_\pi[\Delta(\theta)]$, exists.

The probability (1.5) would be estimated by the proportion of iterations where $\theta_j^{(t)}$ exceeded b, namely $\sum_{t=1}^{T} 1(\theta_j^{(t)} > b)/T$, where $1(A)$ is an indicator function that takes value 1 when A is true, and 0 otherwise. Thus one might in a disease-mapping application wish to obtain the probability that an area's smoothed relative mortality risk θ_k exceeds zero, and so count iterations where this condition holds, avoiding the need to evaluate the integral

$$\Pr(\theta_k > 0|y) = \int_{..} \int_0^\infty {}_{..}\int p(\theta|y)\mathrm{d}\theta$$

where the k^{th} integral is confined to positive values.

This principle extends to empirical estimates of the distribution function, $F()$ of parameters or functions of parameters. Thus the estimated probability that $\Delta \leq h$ for values of h within the support of Δ is

$$\hat{F}(d) = \sum_{t=1}^{T} \frac{1\left(\Delta^{(t)} \leq h\right)}{T}.$$

The sampling output also often includes predictive replicates $y_{\text{new}}^{(t)}$ that can be used in posterior predictive checks to assess whether a model's predictions are consistent with the observed data. Predictive replicates are obtained by sampling $\theta^{(t)}$ and then sampling y_{new} from the likelihood model $p(y_{\text{new}}|\theta^{(t)})$. The posterior predictive density can also be used for model choice and residual analysis (Gelfand, 1996, Sections 9.4–9.6).

1.4 THE MAIN MCMC SAMPLING ALGORITHMS

The Metropolis–Hastings (M–H) algorithm is the baseline for MCMC schemes that simulate a Markov chain $\theta^{(t)}$ with $p(\theta|y)$ as its stationary distribution. Following Hastings (1970), the chain is updated from $\theta^{(t)}$ to θ^* with probability

$$\alpha\left(\theta^*|\theta^{(t)}\right) = \min\left(1, \frac{p\left(\theta^*|y\right) f\left(\theta^{(t)}|\theta^*\right)}{p\left(\theta^{(t)}|y\right) f\left(\theta^*|\theta^{(t)}\right)}\right),$$

where f is known as a proposal or jumping density (Chib and Greenberg, 1995). $f(\theta^*|\theta^{(t)})$ is the probability (or density ordinate) of θ^* for a density centred at $\theta^{(t)}$, while $f(\theta^{(t)}|\theta^*)$ is the probability of moving back from θ^* to the original value. The transition kernel is $k(\theta^{(t)}|\theta^*) = \alpha(\theta^*|\theta^{(t)}) f(\theta^*|\theta^{(t)})$ for $\theta^* \neq \theta^{(t)}$, with a non-zero probability of staying in the current state, namely $k(\theta^{(t)}|\theta^{(t)}) = 1 - \int \alpha(\theta^*|\theta^{(t)})f(\theta^*|\theta^{(t)})\mathrm{d}\theta^*$. Conformity of M–H sampling to the Markov chain requirements discussed above is considered by Mengersen and Tweedie (1996) and Roberts and Rosenthal (2004).

If the proposed new value θ^* is accepted, then $\theta^{(t+1)} = \theta^*$, while if it is rejected, the next state is the same as the current state, i.e. $\theta^{(t+1)} = \theta^{(t)}$. The target density $p(\theta|y)$ appears in ratio form so it is not necessary to know any normalising constants. If the proposal density is symmetric, with $f(\theta^*|\theta^{(t)}) = f(\theta^{(t)}|\theta^*)$, then the M–H algorithm reduces to the algorithm developed by Metropolis *et al.* (1953), whereby

$$\alpha\left(\theta^*|\theta^{(t)}\right) = \min\left[1, \frac{p(\theta^*|y)}{p(\theta^{(t)}|y)}\right].$$

If the proposal density has the form $f(\theta^*|\theta^{(t)}) = f(\theta^{(t)} - \theta^*)$, then a random walk Metropolis scheme is obtained (Gelman *et al.*, 1995). Another option is independence sampling, when the density $f(\theta^*)$ for sampling new values is independent of the current value $\theta^{(t)}$. One may also combine the adaptive rejection technique with M–H sampling, with f acting as a pseudo-envelope for the target density p (Chib and Greenberg, 1995; Robert and Casella, 1999, p. 249). Scollnik (1995) uses this algorithm to sample from the Makeham density often used in actuarial work.

The M–H algorithm works most successfully when the proposal density matches, at least approximately, the shape of the target density $p(\theta|y)$. The rate at which a proposal generated by f is accepted (the acceptance rate) depends on how close θ^* is to $\theta^{(t)}$, and this depends on the dispersion Σ or variance σ^2 of the proposal density. For a normal proposal density a higher acceptance rate would follow from reducing σ^2, but with the risk that the posterior density will take longer to explore. If the acceptance rate is too high, then autocorrelation in sampled values will be excessive (since the chain tends to move in a restricted space), while a too low acceptance rate leads to the same problem, since the chain then gets locked at particular values.

One possibility is to use a variance or dispersion estimate V_θ from a maximum likelihood or other mode finding analysis and then scale this by a constant $c > 1$, so that the proposal density variance is $\Sigma = cV_\theta$ (Draper, 2005, Chapter 2). Values of c in the range 2–10 are typical, with the proposal density variance $2.38^2 V_\theta/d$ shown as optimal in random walk schemes (Roberts *et al.*, 1997). The optimal acceptance rate for a random walk Metropolis scheme is obtainable as 23.4% (Roberts and Rosenthal, 2004, Section 6). Recent work has focused on adaptive MCMC schemes whereby the tuning is adjusted to reflect the most recent estimate of the posterior covariance V_θ (Gilks *et al.*, 1998; Pasarica and Gelman, 2005). Note that certain proposal densities have parameters other than the variance that can be used for tuning acceptance rates (e.g. the degrees of freedom if a Student t proposal is used). Performance also tends to be improved if parameters are transformed to take the full range of positive and negative values $(-\infty, \infty)$ so lessening the occurrence of skewed parameter densities.

Typical random walk Metropolis updating uses uniform, standard normal or standard Student t variables W_t. A normal random walk for a univariate parameter takes samples $W_t \sim N(0, 1)$ and a proposal $\theta^* = \theta^{(t)} + \sigma W_t$, where σ determines the size of the jump (and the acceptance rate). A uniform random walk samples $U_t \sim \text{Unif}(-1, 1)$ and scales this to form a proposal $\theta^* = \theta^{(t)} + \kappa U_t$. As noted above, it is desirable that the proposal density approximately matches the shape of the target density $p(\theta|y)$. The Langevin random walk scheme is an

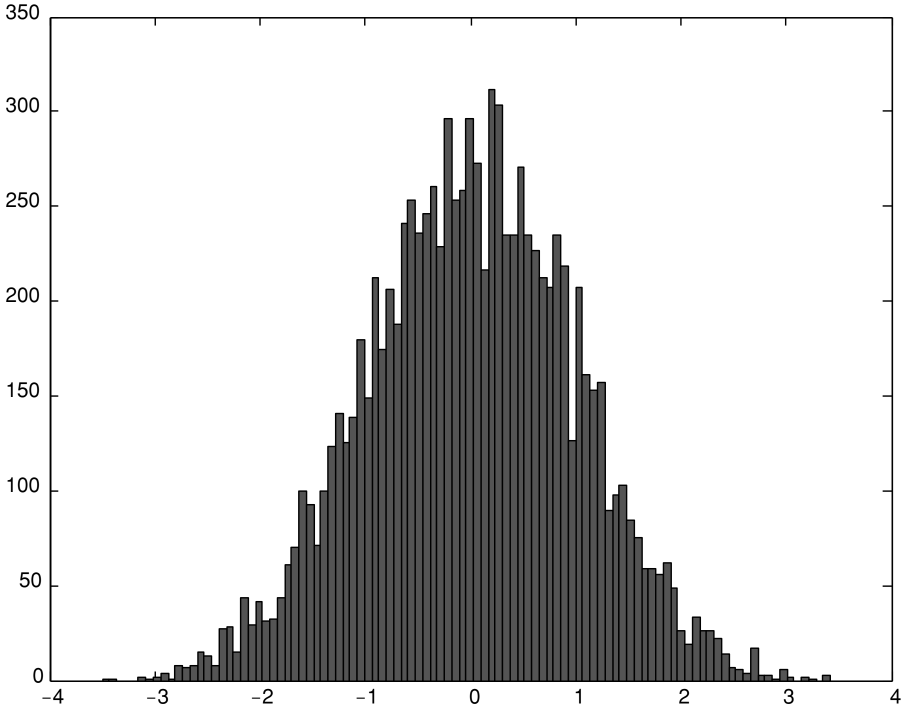

Figure 1.1 Uniform random walk samples from a $N(0, 1)$ density.

example of a scheme including information about the shape of $p(\theta|y)$ in the proposal, namely $\theta^* = \theta^{(t)} + \sigma(W_t + 0.5\nabla\log(p(\theta^{(t)}|y))$ where ∇ denotes the gradient function (Roberts and Tweedie, 1996).

As an example of a uniform random walk proposal, consider Matlab code to sample $T = 10\,000$ times from a $N(0, 1)$ density using a $U(-3, 3)$ proposal density – see Hastings (1970) for the probability of accepting new values when sampling $N(0, 1)$ with a uniform $U(-\kappa, \kappa)$ proposal density. The code is

```
N = 10000; th(1) = 0; pdf = inline('exp(-x^2/2)'); acc=0;
  for i=2:n        thstar = th(i-1) + 3*(1-2*rand);
       alpha = min([1,pdf(thstar)/pdf(th(i-1))]);
  if   rand <= alpha   th(i)=thstar; acc=acc+1;
  else th(i)=th(i-1); end
  end
  sprintf('acceptance rate  %4.0f',100*acc/n)
  hist(th,100);
```

The acceptance rate is around 49% (depending on the seed). Figure 1.1 contains a histogram of the sampled values.

While it is possible for the proposal density to relate to the entire parameter set, it is often computationally simpler in multi-parameter problems to divide θ into D blocks or components, and use componentwise updating. Thus let $\theta_{[j]} = (\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_D)$ denote the parameter set omitting component θ_j and $\theta_j^{(t)}$ be the value of θ_j after iteration t. At step j of iteration $t+1$ the preceding $j-1$ parameter blocks are already updated via the M–H algorithm while $\theta_{j+1}, \ldots, \theta_D$ are still at their iteration t values (Chib and Greenberg, 1995). Let the vector of partially updated parameters be denoted by

$$\theta_{[j]}^{(t,t+1)} = \left(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \ldots, \theta_D^{(t)}\right).$$

The proposed value θ_t^* for $\theta_j^{(t+1)}$ is generated from the jth proposal density, denoted by $f(\theta_j^*|\theta_j^{(t)}, \theta_{[j]}^{(t,t+1)})$. Also governing the acceptance of a proposal are full conditional densities $p(\theta_j^{(t)}|\theta_{[j]}^{(t,t+1)})$ specifying the density of θ_j conditional on other parameters $\theta_{[j]}$. The candidate value θ_j^* is then accepted with probability

$$\alpha\left(\theta_j^{(t)}, \theta_{[j]}^{(t,t+1)}, \theta_j^*\right) = \min\left[1, \frac{p\left(\theta_j^*|\theta_{[j]}^{(t,t+1)}\right) f\left(\theta_j^{(t)}|\theta_j^*, \theta_{[j]}^{(t,t+1)}\right)}{p\left(\theta_j^{(t)}|\theta_{[j]}^{(t,t+1)}\right) f\left(\theta_j^*|(\theta_j^{(t)}, \theta_{[j]}^{(t,t+1)}\right)}\right].$$

1.4.1 Gibbs sampling

The Gibbs sampler (Casella and George, 1992; Gelfand and Smith, 1990; Gilks *et al.*, 1993) is a special componentwise M–H algorithm whereby the proposal density for updating θ_j equals the full conditional $p(\theta_j^*|\theta_{[j]})$ so that proposals are accepted with probability 1. This sampler was originally developed by Geman and Geman (1984) for Bayesian image reconstruction, with its potential for simulating marginal distributions by repeated draws recognised by Gelfand and Smith (1990). The Gibbs sampler involves parameter-by-parameter or block-by-block updating, which when completed forms the transition from $\theta^{(t)}$ to $\theta^{(t+1)}$:

1. $\theta_1^{(t+1)} \sim f_1\left(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_D^{(t)}\right)$;
2. $\theta_2^{(t+1)} \sim f_2\left(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_D^{(t)}\right)$;

-
-
-

D. $\theta_D^{(t+1)} \sim f_D\left(\theta_D|\theta_1^{(t+1)}, \theta_3^{(t+1)}, \ldots, \theta_{D-1}^{(t+1)}\right)$.

Repeated sampling from M–H samplers such as the Gibbs sampler generates an autocorrelated sequence of numbers that, subject to regularity conditions (ergodicity, etc.), eventually ‘forgets’ the starting values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_D^{(0)})$ used to initialise the chain, and converges to a stationary sampling distribution $p(\theta|y)$.

The full conditional densities may be obtained from the joint density $p(\theta, y) = p(y|\theta)p(\theta)$ and in many cases reduce to standard densities (normal, exponential, gamma, etc.) from which sampling is straightforward. Full conditional densities can be obtained by abstracting out from the full model density (likelihood times prior) those elements including θ_j and treating other components as constants (Gilks, 1996).

Consider a conjugate model for Poisson count data y_i with exposures t_i and means λ_i that in turn are gamma distributed, $\lambda_i \sim \mathrm{Ga}(\alpha, \beta)$,

$$p(\lambda_i|\alpha, \beta) = \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \beta^{\alpha} / \Gamma(\alpha).$$

Assume priors $\alpha \sim E(a)$, $\beta \sim \mathrm{Ga}(b, c)$ where a, b and c are preset constants (George *et al.*, 1993). The posterior density of the $n + 2$ parameters $\theta = (\lambda_1, \ldots, \lambda_n, \alpha, \beta)$, given y is proportional to

$$\mathrm{e}^{-a\alpha}\beta^{b-1}\mathrm{e}^{-c\beta}\left\{\prod_{i=1}^{n}\exp(-t_i\lambda_i)\lambda_i^{y_i}\right\}\left\{\prod_{i=1}^{n}\lambda_i^{\alpha-1}\exp(-\beta\lambda_i)\right\}\left[\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right]^{n},$$

where all constants (such as the denominator $y_i!$ in the Poisson likelihood) are combined in the proportionality constant. The full conditional densities of λ_i and β are obtained as $\mathrm{Ga}(y_i + \alpha, \beta + t_i)$ and $\mathrm{Ga}(b + n\alpha, c + \sum_{i=1}^{n} \lambda_i)$, respectively. The full conditional density of α is

$$f(\alpha|y, \beta, \lambda) \propto \mathrm{e}^{-a\alpha}\left[\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right]^{n}\left(\prod_{i=1}^{n}\lambda_i\right)^{\alpha-1.}$$

This density cannot be sampled directly, though techniques such as adaptive rejection sampling (Gilks and Wild, 1992) may be used. Alternatively, a Metropolis step may be included to update α while other parameters are sampled from their full conditionals, an example of a Metropolis within Gibbs procedure (Brooks, 1999).

Figure 1.2 contains a Matlab code applying the latter approach to the well-known data on failures in 10 power plant pumps, also analysed by George *et al.* (1993). The number of failures is assumed to follow a Poisson distribution $y_i \sim \mathrm{Poisson}(\lambda_i t_i)$, where λ_i is the failure rate, and t_i is the length of pump operation time (in thousands of hours). Priors are $\alpha \sim E(1)$, $\beta \sim \mathrm{Ga}(0.1, 1)$. The code includes calls to a kernel-plotting routine, and a Matlab adaptation of the coda routine, both from Lesage (1999); coda is the suite of convergence tests originally developed in S-plus (Best *et al.*, 1995). Note that the update for α is in terms of $\nu = g(\alpha) = \log(\alpha)$, and so the prior for α has to be adjusted for the Jacobean $\partial g^{-1}(\nu)/\partial\nu = \mathrm{e}^{\nu} = \alpha$.

```
[time,y] = textread('pumps.txt','%f%f')
n=10;T=10000; B=1000;lam=ones(n,1);beta=0.9*ones(1,T); acc=0;
scale=0.75;a.alph=0.1;  nu=-0.4*ones(1,T);a.beta=0.1;  b.beta=1;
alph(1)=exp(nu(1));
for t=1:T for i=1:n
loglam(i,t)=log(lam(i,t));end
P=exp(nu(t)-a.alph*alph(t)+n*alph(t)*log(beta(t))...
  -n*gammaln(alph(t))+(alph(t)-1)*sum(loglam(1:n,t)));
nustar=nu(t)+ scale*randn;
alphstar=exp(nustar);
Pstar=exp(nustar-a.alph*alphstar+n*alphstar*log(beta(t))...
  -n*gammaln(alphstar)+(alphstar-1)*sum(loglam(1:n,t)));
if (rand <= Pstar/P)   alph(t+1)=exp(nustar); acc=acc+1;
else                   alph(t+1)=alph(t); end
```

```
% update parameters from full conditionals
for i=1:n
lam(i,t+1)=gamrnd(alph(t+1)+y(i),1/(beta(t)+time(i)));end
beta(t+1)=gamrnd(a.beta+n*alph(t+1),1/(b.beta+sum(lam(1:n,t+1))));
% accumulate draws for coda input
for i=1:n pars(t,i)=lam(i,t);end
pars(t,n+1)=beta(t); pars(t,n+2)=alph(t); end
 sprintf('acceptance rate alpha %5.1f',100*acc/T)
 hist(beta,100); pause; hist(alph,100); pause;
 [hbeta,smbeta,xbeta] = pltdens(beta); plot(xbeta,smbeta); pause;
 [halph,smalph,xalph] = pltdens(alph); plot(xalph,smalph); pause;
 for i=1:12  for t=B+1: T
 parsamp(t-B,i)=pars(t,i); end
 end
 coda(parsamp)
```

Figure 1.2 Matlab code: nuclear pumps data Poisson–gamma model.

Figure 1.3 shows the histogram of β obtained from a single-chain run of 10 000 iterations, and its slight positive skew. Single-chain diagnostics (with 1000 burn-in iterations excluded) are satisfactory with lag 10 autocorrelations under 0.10 for all unknowns. The acceptance rate for α is 38%.

1.5 CONVERGENCE OF MCMC SAMPLES

There are many unresolved questions around the assessment of convergence of MCMC sampling procedures (Brooks and Roberts, 1998; Cowles and Carlin, 1996). One view is that a single long chain is adequate to explore the posterior density, provided allowance is made for dependence in the samples (e.g. Bos, 2004; Geyer, 1992). Diagnostics in the coda routine include those obtainable from a single chain, such as the relative numerical efficiency (RNE) (Geweke, 1992; Kim *et al.*, 1998), Raftery–Lewis diagnostics, which indicate the required sample to achieve a desired accuracy for parameters, and Geweke (1992) chi-square tests.

Relative numerical efficiency compares the empirical variance of the sampled values to a correlation-consistent variance estimator (Geweke, 1999; Geweke *et al.*, 2003). Numerical approximations of functions such as (1.4) based on T samples will have the same accuracy as ($T \times$ RNE) samples based on iid (independent, identically distributed) drawings directly from the posterior distribution. The method of Raftery and Lewis (1992) provides an estimate of the number of MCMC samples required to achieve a specified accuracy of the estimated quantiles of parameters or functions; for example, one might require the 2.5th percentile to be estimated to an accuracy ± 0.005, and with a certain probability of attaining this level of accuracy (say, 0.95). The Raftery–Lewis diagnostics include the minimum number of iterations needed to estimate the specified quantile to the desired precision if the samples in the chain were independent. This is a lower bound, and may tend to be conservative (Draper, 2006). The Geweke procedure considers different portions of MCMC output to determine whether they can be considered as coming from the same distribution; specifically, initial and final portions of a chain of sampled parameter values (e.g. the first 10% and the last 50%) are compared, with tests using sample means and asymptotic variances (estimated using spectral density methods) in each portion.

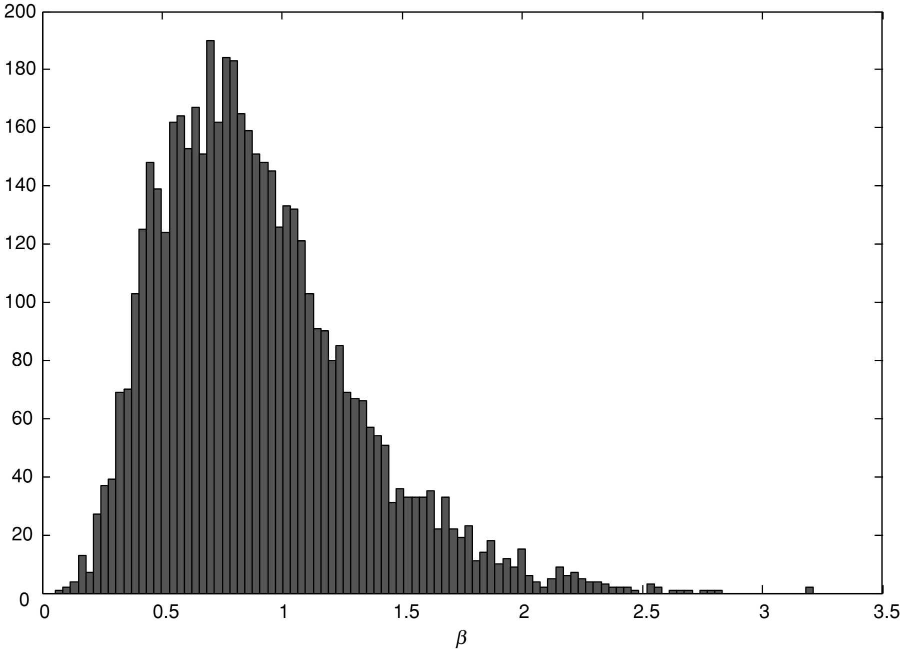

Figure 1.3 Histograms of samples of beta.

Many practitioners prefer to use two or more parallel chains with diverse starting values to ensure full coverage of the sample space of the parameters, and so diminish the chance that the sampling will become trapped in a small part of the space (Gelman and Rubin, 1992, 1996). Single long runs may be adequate for straightforward problems, or as a preliminary to obtain inputs to multiple chains. Convergence for multiple chains may be assessed using Gelman–Rubin scale-reduction factors that compare variation in the sampled parameter values within and between chains. Parameter samples from poorly identified models will show wide divergence in the sample paths between different chains, and variability of sampled parameter values between chains will considerably exceed the variability within any one chain. To measure variability of samples $\theta_j^{(t)}$ within the jth chain ($j = 1, \ldots, J$) define

$$w_j = \left(\theta_j^{(t)} - \overline{\theta}_j\right)^2/(T-1),$$

defined over T iterations after an initial burn-in of B iterations. Ideally the burn-in period is a short initial set of samples where the effect of the initial parameter values tails off; during the burn-in the parameter trace plots will show clear monotonic trends as they reach the region of the posterior.

Variability within chains W is then the average of the w_j. Between-chain variance is measured by

$$B = \frac{T}{J-1}\sum_{j=1}^{J}(\overline{\theta}_j - \overline{\theta})^2$$

where (θ) is the average of the $\overline{\theta}_j$. The potential scale reduction factor (PSRF) compares a pooled estimator of var(θ), given by $V = B/T + TW/(T-1)$ with the within-sample estimate W. Specifically the PSRF is $(V/W)^{0.5}$ with values under 1.2 indicating convergence.

Another multiple-chain convergence statistic is due to Brooks and Gelman (1998) and known as the Brooks–Gelman–Rubin (BGR) statistic. This is a ratio of parameter interval lengths, where for chain j the length of the $100(1-\alpha)\%$ interval for parameter θ is obtained, namely the gap between 0.5α and $(1-0.5\alpha)$ points from T simulated values. This provides J within-chain interval lengths, with mean I_U. For the pooled output of TJ samples, the same $100(1-\alpha)\%$ interval I_P is also obtained. Then the ratio I_P/I_U should converge to 1 if there is convergent mixing over different chains. Brooks and Gelman also propose a multivariate version of the original G–R ratio, which, a review by Sinharay (2004) indicates, may be better at detecting convergence in models where identifiability is problematic; this refers to practical identifiability of complex models for relatively small datasets, rather than mathematical identifiability. However, multiple-chain analysis can also be a useful check on unsuspected mathematical non-identifiability, or on model priors that are not constrained to produce unique labelling. Fan *et al.* (2006) consider diagnostics based on score statistics for parameters θ_k; for likelihood $L = p(y \mid \theta)$, or target density $\pi(\theta) = p(\theta|y)$, define score functions $U_k = \partial\pi/\partial\theta_k$, and then obtain means m_k and variances V_k of U_{kj} statistics obtained from chains $j = 1, \ldots, J$. Then $X^2 = J\, m_k^2/V_k$ is asymptotically chi-squared with d degrees of freedom under convergence.

The following Matlab program obtains univariate PSRFs and the multivariate PSRF for an augmented data probit analysis of the shopping data used in Example 4.9. Two chains are run for $T = 1000$ iterations with a burn-in of 50 iterations, with flat priors on the regression parameters. All scale factors obtained are very close to 1. The main program and the Gelman–Rubin functions called are as follows:

```
[y,Inc,Hsz,WW] = textread('shop.txt','%f %f %f %f'); n=84;
for i=1:n X(i,1)=1; X(i,2)=Inc(i); X(i,3)=Hsz(i); X(i,4)=WW(i); end
beta = [0 0 0 0]'; Lo = -10.* (1-y); Hi =10.* y; T=1000; burnin=50;
for ch=1:2 for t=1:T
% truncated normal sample between Lo and Hi
  Z = rand_nort(X * beta, ones(size(X * beta)), Lo, Hi);
  sigma=inv(X' * X); betaMLE = inv(X' * X)* X' * Z;
  beta = rand_MVN(1, betaMLE, sigma)';
for j=1:4 betas(t,j,ch)=beta(j); end
end
end
[PSRF] = GRpsrf(betas,T,4,2)
[MPSRF] = GRmpsrf(betas,T,4,2)

function [PSRF] = GRpsrf(th,T,d,J)
W = zeros(1,d); B = zeros(1,d); mn = mean(reshape(mean(th),d,J)');
for j=1:J
  dw = th(:,:,j) - repmat(mean(th(:,:,j)),T,1);
  db = mean(th(:,:,j))- mn;
  W = W + sum(dw.*dw); B = B + db.*db; end
```

```
    W = W / ((T-1) * J); S = (T-1)/T * W + B/(J-1);
PSRF = sqrt((J+1)/J * S ./ W - (T-1)/J/T); end

function [MPSRF] = GRmpsrf(th,T,d,J)
W = zeros(d); B = zeros(d); mn = mean(reshape(mean(th),d,J)');
for j=1:J
    dw = th(:,:,j) - repmat(mean(th(:,:,j)),T,1);
    db = mean(th(:,:,j))- mn;
    W = W + dw'*dw; B = B + db'*db; end
W = W / ((T-1) * J); B = B / (J-1); V = sort(abs(eig(W\B)));
MPSRF = sqrt( (T-1)/T + V(end) * (J+1)/J); end
```

Parameter samples obtained by MCMC methods are correlated, which means extra samples are needed to convey the same information. The extent of correlation will depend on a number of factors including the form of parameterisation, the complexity of the model and the form of sampling (e.g. block or univariate sampling of parameters). Analysis of autocorrelation in sequences of MCMC samples amounts to an application of time series methods, in regard to issues such as assessing stationarity in an autocorrelated sequence. Autocorrelation at lags 1, 2 and so on may be assessed from the full set of sampled values $\theta^{(t)}$, $\theta^{(t+1)}$,$\theta^{(t+2)}$, ..., or from subsamples K steps apart $\theta^{(t)}$, $\theta^{(t+K)}$, $\theta^{(t+2K)}$, ..., etc. If the chains are mixing satisfactorily then the autocorrelations in the one-step apart iterates $\theta^{(t)}$ will fade to zero as the lag increases (e.g. at lag 10 or 20). Non-vanishing autocorrelations at high lags mean that less information about the posterior distribution is provided by each iterate and a higher sample size T is necessary to cover the parameter space. Slow convergence will show in trace plots that wander, and that exhibit short-term trends rather than rapidly fluctuating around a stable mean.

Problems of convergence in MCMC sampling may reflect problems in model identifiability due to overfitting or redundant parameters. Running multiple chains often assists in diagnosing poor identifiability of models. This is illustrated most clearly when identifiability constraints are missing from a model, such as in discrete mixture models that are subject to 'label switching' during MCMC updating (Frühwirth-Schnatter, 2001). One chain may have a different 'label' to others and so applying any convergence criterion is not sensible (at least for some parameters). Choice of diffuse priors tends to increase the chance of poorly identified models, especially in complex hierarchical models or small samples (Gelfand and Sahu, 1999). Elicitation of more informative priors or application of parameter constraints may assist identification and convergence.

Correlation between parameters within the parameter set $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ also tends to delay convergence and increase the dependence between successive iterations. Reparameterisation to reduce correlation – such as centring predictor variables in regression – usually improves convergence (Zuur *et al.*, 2002). Robert and Mengersen (1999) consider a reparameterisation of discrete normal mixtures to improve MCMC performance. Slow convergence in random effects models such as the two-way model (e.g. repetitions $j = 1, \ldots, J$ over subjects $i = 1, \ldots, I$)

$$y_{ij} = \mu + \alpha_i + u_{ij}$$

with $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $u_{ij} \sim N(0, \sigma_u^2)$ may be lessened by a centred hierarchical prior, namely $y_{ij} \sim N(\kappa_i, \sigma_u^2)$ and $\kappa_i \sim N(\mu, \sigma_\alpha^2)$ (Gelfand *et al.*, 1995; Gilks and Roberts, 1996). For three-way nesting with

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + u_{ijk}$$

with $\beta_{ij} \sim N(0, \sigma_\beta^2)$, the centred version is $y_{ijk} \sim N(\zeta_{ij}, \sigma_u^2)$, $\phi_{ij} \sim N(\kappa_i, \sigma_\beta^2)$, and $\kappa_i \sim N(\mu, \sigma_\alpha^2)$. Vines *et al.* (1996) suggest sweeping for the subject effects, so that

$$y_{ij} = \nu + \rho_i + u_{ij},$$

where $\rho_i = \alpha_i - \overline{\alpha}$, $\nu = \mu + \overline{\alpha}$, so that $\sum_{i=1}^{I} \rho_i = 0$, with $\rho_i \sim N(0, \sigma(1 - 1/I)$.

Scollnik (2002) considers WINBUGS implementation of this prior.

1.6 PREDICTIONS FROM SAMPLING: USING THE POSTERIOR PREDICTIVE DENSITY

In classical statistics the prediction of out-of-sample data z (for example, data at future time points or under different conditions and covariates) often involves calculating moments or probabilities from the assumed likelihood for y evaluated at the selected point estimate θ_m, namely $p(y|\theta_m)$. In the Bayesian method, the information about θ is contained not in a single point estimate but in the posterior density $p(\theta|y)$ and so prediction is correspondingly based on averaging $p(z|y, \theta)$ over this posterior density. Generally $p(z|y, \theta) = p(z|\theta)$, namely that predictions are independent of the observations given θ. So the predicted or replicate data z given the observed data y is, for θ discrete, the sum

$$p(z|y) = \sum_\theta p(z|\theta)p(\theta|y)$$

and is an integral over the product $p(z|\theta)p(\theta|y)$ when θ is continuous. In the sampling approach, with iterations $t = B + 1, \ldots, B + T$ after convergence, this involves iteration-specific samples of $z^{(t)}$ from the same likelihood form used for $p(y|\theta)$, given the sampled value $\theta^{(t)}$.

There are circumstances (e.g. in regression analysis or time series) where such out-of-sample predictions are the major interest; such predictions may be in circumstances where the explanatory variates take different values to those actually observed. In clinical trials comparing the efficacy of an established therapy as against a new therapy, the interest may be in the predictive probability that a new patient will benefit from the new therapy (Berry, 1993). In a two-stage sample situation where m clusters are sampled at random from a larger collection of M clusters, and then respondents are sampled at random within the m clusters, predictions of populationwide quantities or parameters can be made to allow for the uncertainty attached to the unknown data in the $M - m$ non-sampled clusters (Stroud, 1994).

1.7 THE PRESENT BOOK

The chapters that follow review several major areas of statistical application and modelling with a view to implementing the above components of the Bayesian perspective, discussing worked examples and providing source code that may be extended to similar problems by students and researchers. Any treatment of such issues is necessarily selective, emphasising particular methodologies rather than others, and particular areas of application. As in the first edition of *Bayesian Statistical Modelling*, the goal is to illustrate the potential and flexibility of Bayesian approaches to often complex statistical modelling and also the utility of the WINBUGS package in this context – though some Matlab code is included in Chapter 2.

WINBUGS is *S* based and offers the basis for sophisticated programming and data manipulation but with a distinctive Bayesian functionality. WINBUGS selects appropriate MCMC updating schemes via an inbuilt expert system so that there is a blackbox element to some extent. However, respecifying or extending models can be done simply in WINBUGS without having to retune the MCMC sampling update schemes, as is necessary in more direct programming in (say) R, Matlab or GAUSS. The labour and checking required in direct programming increases with the complexity of the model. However, the programming flexibility offered by WINBUGS may be more favourable to some tastes than others – WINBUGS is not menu driven and pre-packaged, and does make greater demands on the researcher's own initiative. A brief guide to help new WINBUGS users is included in an appendix, though many online WINBUGS guides exist; extended discussion of how to use WINBUGS appears in Scollnik (2001), Fryback *et al.* (2001), and Woodworth (2004, Appendix B).

Issues around prior elicitation and sensitivity to alternative priors may to some viewpoints be downplayed in necessarily abbreviated worked examples. In most applications multiple chains are used with convergence assessed using Gelman–Rubin diagnostics, but without a detailed report of other diagnostics available in coda and similar routines. The focus is more towards illustrating Bayesian implementation of a range of modelling techniques including multilevel models, survival models, time series and dynamic linear models, structural equation models, and missing data models. Any comments on the programs, data interpretation, coding mistakes and so on would be appreciated at p.congdon@qmul.ac.uk. The reader is also referred to the website at the Medical Research Council Biostatistics Unit at Cambridge University, where a highly illuminating set of examples are incorporated in the downloadable software, and links exist to other collections of WINBUGS software.

REFERENCES

Ahrens, J. and Dieter, U. (1974) Computer methods for sampling from gamma, beta, Poisson and binomial distributions. *Computing*, **12**, 223–246.

Andrade, J. and O'Hagan, A. (2006) Bayesian robustness modelling using regularly varying distributions. *Bayesian Analysis*, **1**, 169–188.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York.

Berger, J. (1990) Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, **25**, 303–328.

Berger, J. (1994) An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

Berger, J. and Bernardo, J. (1994) Estimating a product of means. Bayesian analysis with reference priors. *Journal of American Statistical Association*, **89**, 200–207.

Berry, D. (1993) A case for Bayesianism in clinical trials. *Statistics in Medicine*, **12**, 1377–1393.

Best, N., Cowles, M. and Vines, S. (1995) *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.3.* MRC Biostatistics Unit: Cambridge.

Birkes, D. and Dodge, Y. (1993) *Alternative Methods of Regression* (*Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*). John Wiley & Sons, Ltd/Inc: New York.

Bos, C. (2004) Markov Chain Monte Carlo methods: implementation and comparison. *Working Paper*, Tinbergen Institute & Vrije Universiteit, Amsterdam.

Brock, W., Durlauf, S. and West, K. (2004) Model uncertainty and policy evaluation: some theory and empirics. *Working Paper*, No. 2004-19, Social Systems Research Institute, University of Wisconsin-Madison.

Brooks, S. (1999) Bayesian analysis of animal abundance data via MCMC. In *Bayesian Statistics 6*, Bernardo, J., Berger, J., Dawid, A. and Smith, A (eds). Oxford University Press: Oxford, 723–731.

Brooks, S. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–456.

Brooks, S. and Roberts, G. (1998) Assessing convergence of Markov Chain Monte Carlo algorithms. *Statistics and Computing*, **8**, 319–335.

Carlin, J., Wolfe, R., Hendricks Brown, C. and Gelman, A. (2001) A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics*, **2**, 397–416.

Casella G. and George, E. (1992) Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

Chaloner, K. (1995) The elicitation of prior distributions. In *Bayesian Biostatistics*, Stangle, D. and Berry, D. (eds). Marcel Dekker: New York.

Chen, M., Shao, Q. and Ibrahim, J. (2000) *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag: New York.

Chib, S. and Greenberg, E. (1994) Bayes inference in regression models with ARMA(p,q) errors. *Journal of Econometrics*, **64**, 183–206.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithm. *The American Statistician*, **49**, 327–345.

Cowles, M. and Carlin, B. (1996) Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Daniels, M. (1999) A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, **27**, 567–578.

Devroye, L. (1986) *Non-Uniform Random Variate Generation*. Springer-Verlag: New York.

Draper, D. (in press) *Bayesian Hierarchical Modeling*. Springer-Verlag: New York.

Fan, Y., Brooks, S. and Gelman, A. (2006) Output assessment for Monte Carlo simulations via the score statistic. *Journal of Computational and Graphical Statistics*, **15**, 178–206.

Fraser, D., McDunnough, P. and Taback, N. (1997) Improper priors, posterior asymptotic normality, and conditional inference. In *Advances in the Theory and Practice of Statistics*, Johnson, N. and Balakrishnan, N. (eds). John Wiley & Sons, Ltd/Inc.: New York, 563–569.

Frühwirth-Schnatter, S. (2001) MCMC estimation of classical and dynamic switching and mixture models, *Journal of the American Statistical Association*, **96**, 194–209.

Frühwirth-Schnatter, S. (2004) Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal*, **7**, 143–167.

Fryback, D., Stout, N. and Rosenberg, M. (2001) An elementary introduction to Bayesian computing using WinBUGS. *International Journal of Technology Assessment in Health Care*, **17**, 96–113.

Garthwaite, P., Kadane, J. and O'Hagan, A. (2005) Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, **100**, 680–700.

Gelfand, A. (1996) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 145–161.

Gelfand A. and Sahu, S. (1999) Gibbs sampling, identifiability and improper priors in generalized linear mixed models. *Journal of the American Statistical Association*, **94**, 247–253.

Gelfand, A. and Smith, A. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelfand, A., Sahu, S. and Carlin, B. (1995) Efficient parameterization for normal linear mixed effects models. *Biometrika*, **82**, 479–488.

Gelfand, A., Sahu, S. and Carlin, B. (1996) Efficient parametrization for generalized linear mixed models. In *Bayesian Statistics 5*, Bernardo, J., Berger, J., Dawid, A.P. and Smith, A.F.M. (eds). Clarendon Press: Oxford, 165–180.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, A. and Rubin, D. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Gelman, A. and Rubin, D. (1996) Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research*, **5**, 339–355.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian Data Analysis* (1st edn) (Texts in Statistical Science Series). Chapman & Hall: London.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

George, E., Makov, U. and Smith, A. (1993) Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, **20**, 147–156.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, Bernardo, J., Berger, J., Dawid, A. and Smith, A. (eds). Clarendon Press: Oxford.

Geweke, J. (1999) Using simulation methods for Bayesian econometric models: inference, development and communication. *Econometric Reviews*, **18**, 1–126.

Geweke, J., Gowrisankaran, G. and Town, R. (2003) Bayesian inference for hospital quality in a selection model. *Econometrica*, **71**, 1215–1238.

Geyer, C. (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, **7**, 473–511.

Gilks, W. (1996) Full conditional distributions. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 75–88.

Gilks, W. and Roberts, G. (1996) Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 89–114.

Gilks, W. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Gilks, W., Clayton, D., Spiegelhalter, D., Best, N., McNeil, A., Sharples, L. and Kirby, A. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society, Series B*, **55**, 39–52.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 1–20.

Gilks, W., Roberts, G. and Sahu, S. (1998) Adaptive Markov Chain Monte Carlo. *Journal of the American Statistical Association*, **93**, 1045–1054.

Gustafson, P. (1996) Robustness considerations in Bayesian analysis. *Statistical Methods in Medical Research*, **5**, 357–373.

Gustafson, P., Hossain, S. and MacNab, Y. (in press) Conservative priors for hierarchical models. *Canadian Journal of Statistics.*

Hadjicostas, P. and Berry, S. (1999) Improper and proper posteriors with improper priors in a Poisson–gamma hierarchical model. *Test*, **8**, 147–166.

Hastings, W. (1970) Monte-Carlo sampling methods using Markov Chains and their applications. *Biometrika*, **57**, 97–109.

Ibrahim, J. and Chen, M. (2000) Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.

Jasra, A., Holmes, C. and Stephens, D. (2005) Markov Chain Monte Carlo Methods and the label switching problem in Bayesian mixture modeling. *Statististical Science*, **20**, 50–67.

Kass, R. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Kim, S., Shephard, N. and Chib, S. (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **64**, 361–393.

Lesage, J. (1999) *Applied Econometrics using MATLAB*. Department of Economics, University of Toledo: Toledo, OH. Available at: www.spatial-econometrics.com/html/mbook.pdf.

Mengersen, K.L. and Tweedie, R.L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, **24**, 101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Norris, J. (1997) *Markov Chains*. Cambridge University Press: Cambridge.

O'Hagan, A. (1994) *Kendall's Advanced Theory of Statistics: Bayesian Inference* (Vol. 2B). Edward Arnold: Cambridge.

Osherson, D., Smith, E., Shafir, E., Gualtierotti, A. and Biolsi, K. (1995) A source of Bayesian priors. *Cognitive Science*, **19**, 377–405.

Pasarica, C. and Gelman, A. (2005) Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Technical Report*, Department of Statistics, Columbia University.

Raftery, A. and Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics* (Vol. 4), Bernardo, J., Berger, J., Dawid, A. and Smith, A. (eds). Oxford: Oxford University Press, 763–773.

Richardson, S. and Best, N. (2003) Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, **14**, 129–147.

Robert, C. (2004) Bayesian computational methods. In *Handbook of Computational Statistics* (Vol. I), Gentle, J., Hãrdle, W. and Mori, Y. (eds). Springer-Verlag: Heidelberg, Chap. 3.

Robert C. and Casella, G. (1999) *Monte Carlo Statistical Methods*. Springer-Verlag: New York.

Robert, C.P. and Mengersen, K.L. (1999) Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Computational Statistics and Data Analysis*, 325–343.

Roberts, G. (1996) Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 45–59.

Roberts, G. and Rosenthal, J. (2004) General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1**, 20–71.

Roberts, G. and Tweedie, R. (1996) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.

Roberts, G., Gelman, A. and Gilks, W. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110–120.

Scollnik, D. (1995) Simulating random variates from Makeham's distribution and from others with exact or nearly log-concave densities. *Transactions of the Society of Actuaries*, **47**, 41–69.

Scollnik, D. (2001) Actuarial modeling with MCMC and BUGS. *North American Actuarial Journal*, **5**, 96–124.

Scollnik, D. (2002) Implementation of four models for outstanding liabilities in WinBUGS : a discussion of a paper by Ntzoufras and Dellaportas. *North American Actuarial Journal*, **6**, 128–136.

Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman & Hall: London.

Sinharay, S. (2004) Experiences with Markov Chain Monte Carlo Convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics,* **29**, 461–488.

Smith, A. and Gelfand, A. (1992) Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, **46**(2), 84–88.

Spiegelhalter, D., Freedman, L. and Parmar, M. (1994) Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society*, **157**, 357–416.

Spiegelhalter, D., Best, N., Gilks, W. and Inskip, H. (1996) Hepatitis: a case study in MCMC methods. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 21–44.

Stroud, T. (1994) Bayesian analysis of binary survey data. *Canadian Journal of Statistics*, **22**, 33–45.

Syverseen, A. (1998) Noninformative Bayesian priors. Interpretation and problems with construction and applications. Available at: http://www.math.ntnu.no/preprint/statistics/1998/S3-1998.ps

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–1762.

Tierney, L. (1996) Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S. and Spiegelhalter, D. (eds). Chapman & Hall: London, 59–74.

Tierney, L., Kass, R. and Kadane, J. (1988) Interactive Bayesian analysis using accurate asymptotic approximations. In *Computer Science and Statistics: Nineteenth Symposium on the Interface*, Heiberger, R. (ed). American Statistical Association: Alexandria, VA, 15–21.

van Dyk, D. (2002) Hierarchical models, data augmentation, and MCMC. In *Statistical Challenges in Modern Astronomy III*, Babu, G. and Feigelson, E. (eds). Springer: New York, 41–56.

Vines, S., Gilks, W. and Wild, P. (1996) Fitting Bayesian multiple random effects. models. *Statistics and Computing*, **6**, 337–346.

Wasserman, L. (2000) Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society, Series B*, **62**, 159–180.

Woodworth, G. (2004) *Biostatistics: A Bayesian Introduction*. Chichester: John Wiley & Sons, Ltd/Inc.

Zellner, A. (1985) Bayesian econometrics. *Econometrica*, **53**, 253–270.

Zhu, M. and Lu, A. (2004) The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education*, **12**, 1–10.

Zuur, G., Garthwaite, P. and Fryer, R. (2002) Practical use of MCMC methods: lessons from a case study. *Biometrical Journal*, **44**, 433–455.