Bayesian Models for Categorical Data

PETER CONGDON
Queen Mary, University of London, UK
Bayesian Models for Categorical Data
Bayesian Models for Categorical Data

PETER CONGDON
Queen Mary, University of London, UK
## Contents

### Preface

xi

### Chapter 1 Principles of Bayesian Inference

1.1 Bayesian updating 1
1.2 MCMC techniques 2
1.3 The basis for MCMC 4
1.4 MCMC sampling algorithms 6
1.5 MCMC convergence 10
1.6 Competing models 12
1.7 Setting priors 12
1.8 The normal linear model and generalized linear models 13
1.9 Data augmentation 17
1.10 Identifiability 19
1.11 Robustness and sensitivity 21
1.12 Chapter themes 23
References 23

### Chapter 2 Model Comparison and Choice

2.1 Introduction: formal methods, predictive methods and penalized deviance criteria 29
2.2 Formal Bayes model choice 30
2.3 Marginal likelihood and Bayes factor approximations 32
2.4 Predictive model choice and checking 35
2.5 Posterior predictive checks 37
2.6 Out-of-sample cross-validation 39
2.7 Penalized deviances from a Bayes perspective 41
2.8 Multimodel perspectives via parallel sampling 42
2.9 Model probability estimates from parallel sampling 46
2.10 Worked example 49
References 50
Chapter 3  Regression for Metric Outcomes  
3.1 Introduction: priors for the linear regression model  
3.2 Regression model choice and averaging based on predictor selection  
3.3 Robust regression methods: models for outliers  
3.4 Robust regression methods: models for skewness and heteroscedasticity  
3.5 Robustness via discrete mixture models  
3.5.1 Complete data representation  
3.5.2 Identification issues  
3.5.3 Dirichlet process mixture models  
3.6 Non-linear regression effects via splines and other basis functions  
3.6.1 Penalized random effects for spline coefficients  
3.6.2 Basis function regression  
3.6.3 Special spline functions  
3.7 Dynamic linear models and their application in non-parametric regression  
3.7.1 Some common forms of DLM  
3.7.2 Robust errors  
3.7.3 General additive models  
3.7.4 Alternative smoothness priors  
Exercises  
References

Chapter 4  Models for Binary and Count Outcomes  
4.1 Introduction: discrete model likelihoods vs. data augmentation  
4.1.1 Count data  
4.1.2 Binomial and binary data  
4.2 Estimation by data augmentation: the Albert–Chib method  
4.2.1 Other augmented data methods  
4.3 Model assessment: outlier detection and model checks  
4.3.1 Model assessment: predictive model selection and checks  
4.4 Predictor selection in binary and count regression  
4.5 Contingency tables  
4.6 Semi-parametric and general additive models for binomial and count responses  
4.6.1 Robust and adaptive non-parametric regression  
4.6.2 Other approaches to non-linearity  
Exercises  
References
Chapter 5 Further Questions in Binomial and Count Regression

5.1 Generalizing the Poisson and binomial: overdispersion and robustness 155
5.2 Continuous mixture models 157
  5.2.1 Modified exponential families 161
5.3 Discrete mixtures 163
5.4 Hurdle and zero-inflated models 167
5.5 Modelling the link function 171
  5.5.1 Discrete (DPP mixture) 172
  5.5.2 Parametric link transformations 173
  5.5.3 Beta mixture on cumulative densities 174
5.6 Multivariate outcomes 177
  5.6.1 The factor analysis approach 179
  5.6.2 Specific modelling of marginal and odds ratios 180
  5.6.3 Multivariate Poisson data 182
Exercises 189
References 192

Chapter 6 Random Effect and Latent Variable Models for Multicategory Outcomes

6.1 Multicategory data: level of observation and relations between categories 197
6.2 Multinomial models for individual data: modelling choices 198
6.3 Multinomial models for aggregated data: modelling contingency tables 202
  6.3.1 Conditional contingency tables: histogram smoothing and multinomial logit via Poisson regression 203
6.4 The multinomial probit 210
6.5 Non-linear predictor effects 213
6.6 Heterogeneity via the mixed logit 216
6.7 Aggregate multicategory data: the multinomial–Dirichlet model and extensions 219
6.8 Multinomial extra variation 221
  6.8.1 Uncertainty in Dirichlet parameters 223
6.9 Latent class analysis 226
Exercises 230
References 231

Chapter 7 Ordinal Regression

7.1 Aspects and assumptions of ordinal data models 235
7.2 Latent scale and data augmentation 236
### 7.3 Assessing model assumptions: non-parametric ordinal regression and assessing ordinality 243
### 7.4 Location-scale ordinal regression 247
### 7.5 Structural interpretations with aggregated ordinal data 251
### 7.6 Log-linear models for contingency tables with ordered categories 255
### 7.7 Multivariate ordered outcomes 258
### Exercises 263
### References 264

### Chapter 8 Discrete Spatial Data 267
#### 8.1 Introduction 267
#### 8.2 Univariate responses: the mixed ICAR model and extensions 268
#### 8.3 Spatial robustness 273
#### 8.4 Multivariate spatial priors 275
#### 8.5 Varying predictor effect models 279
#### Exercises 285
#### References 286

### Chapter 9 Time Series Models for Discrete Variables 289
#### 9.1 Introduction: time dependence in observations and latent data 289
#### 9.2 Observation-driven dependence 291
##### 9.2.1 Binary data and Markov Chain models 293
##### 9.2.2 Observation-driven models for individual multicategory data 297
##### 9.2.3 Time series of aggregate multinomial data 298
#### 9.3 Parameter-driven dependence via DLMs 300
##### 9.3.1 Measurement error models 302
##### 9.3.2 Conjugate updating 304
#### 9.4 Parameter-driven dependence via autocorrelated error models 309
#### 9.5 Integer autoregressive models 311
#### 9.6 Hidden Markov models 313
#### Exercises 315
#### References 317

### Chapter 10 Hierarchical and Panel Data Models 321
#### 10.1 Introduction: clustered data and general linear mixed models 321
#### 10.2 Hierarchical models for metric outcomes 322
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.3</td>
<td>Hierarchical generalized linear models</td>
<td>324</td>
</tr>
<tr>
<td>10.3.1</td>
<td>Augmented data sampling for hierarchical GLMs</td>
<td>328</td>
</tr>
<tr>
<td>10.4</td>
<td>Random effects for crossed factors</td>
<td>333</td>
</tr>
<tr>
<td>10.5</td>
<td>The general linear mixed model for panel data</td>
<td>336</td>
</tr>
<tr>
<td>10.5.1</td>
<td>Time dependence</td>
<td>338</td>
</tr>
<tr>
<td>10.5.2</td>
<td>Longitudinal categorical data</td>
<td>340</td>
</tr>
<tr>
<td>10.6</td>
<td>Conjugate panel models</td>
<td>346</td>
</tr>
<tr>
<td>10.7</td>
<td>Growth curve analysis</td>
<td>350</td>
</tr>
<tr>
<td>10.8</td>
<td>Multivariate panel data</td>
<td>352</td>
</tr>
<tr>
<td>10.9</td>
<td>Robustness in panel and clustered data analysis</td>
<td>358</td>
</tr>
<tr>
<td>10.10</td>
<td>APC and spatio-temporal models</td>
<td>362</td>
</tr>
<tr>
<td>10.11</td>
<td>Space–time and spatial APC models</td>
<td>364</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>371</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>374</td>
</tr>
<tr>
<td>11.1</td>
<td>Introduction: types of missing data</td>
<td>379</td>
</tr>
<tr>
<td>11.2</td>
<td>Density mechanisms for missing data</td>
<td>381</td>
</tr>
<tr>
<td>11.3</td>
<td>Auxiliary variables</td>
<td>385</td>
</tr>
<tr>
<td>11.4</td>
<td>Predictors with missing values</td>
<td>387</td>
</tr>
<tr>
<td>11.5</td>
<td>Multiple imputation</td>
<td>391</td>
</tr>
<tr>
<td>11.6</td>
<td>Several responses with missing values</td>
<td>395</td>
</tr>
<tr>
<td>11.7</td>
<td>Non-ignorable non-response models for survey tabulations</td>
<td>397</td>
</tr>
<tr>
<td>11.7.1</td>
<td>The differential non-response model</td>
<td>398</td>
</tr>
<tr>
<td>11.7.2</td>
<td>Alternative parameterizations of the differential non-response model</td>
<td>401</td>
</tr>
<tr>
<td>11.7.3</td>
<td>Ecological inference: imputing missing interior data in $R \times C$ tables</td>
<td>403</td>
</tr>
<tr>
<td>11.7.4</td>
<td>Cross-tabulations with missing data on any factor</td>
<td>407</td>
</tr>
<tr>
<td>11.8</td>
<td>Recent developments</td>
<td>409</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>410</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>412</td>
</tr>
</tbody>
</table>

### Chapter 11 Missing-Data Models

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.1</td>
<td>Introduction: types of missing data</td>
<td>379</td>
</tr>
<tr>
<td>11.2</td>
<td>Density mechanisms for missing data</td>
<td>381</td>
</tr>
<tr>
<td>11.3</td>
<td>Auxiliary variables</td>
<td>385</td>
</tr>
<tr>
<td>11.4</td>
<td>Predictors with missing values</td>
<td>387</td>
</tr>
<tr>
<td>11.5</td>
<td>Multiple imputation</td>
<td>391</td>
</tr>
<tr>
<td>11.6</td>
<td>Several responses with missing values</td>
<td>395</td>
</tr>
<tr>
<td>11.7</td>
<td>Non-ignorable non-response models for survey tabulations</td>
<td>397</td>
</tr>
<tr>
<td>11.7.1</td>
<td>The differential non-response model</td>
<td>398</td>
</tr>
<tr>
<td>11.7.2</td>
<td>Alternative parameterizations of the differential non-response model</td>
<td>401</td>
</tr>
<tr>
<td>11.7.3</td>
<td>Ecological inference: imputing missing interior data in $R \times C$ tables</td>
<td>403</td>
</tr>
<tr>
<td>11.7.4</td>
<td>Cross-tabulations with missing data on any factor</td>
<td>407</td>
</tr>
<tr>
<td>11.8</td>
<td>Recent developments</td>
<td>409</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>410</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>412</td>
</tr>
</tbody>
</table>
Preface

This book continues the themes in my two earlier Wiley books in seeking to make modern Bayesian methods accessible via a practically oriented exposition, with statistical computing and applied data analysis at the forefront. As before, I have focused on the WINBUGS package, which has now reached a wide degree of acceptance in application to the ever expanding corpus of methodological work and applications based on Monte Carlo Markov Chain techniques. I hope that the applied focus will help students and researchers alike, including those with primary disciplinary interests outside statistics (e.g. psychology, geography and epidemiology). Nevertheless a wide range of simple or more advanced modelling techniques are discussed and I hope the approach is also helpful for courses in Bayesian data analysis and statistical computing. I have sought to review recent Bayesian methodology for categorical outcomes (binary, count and multinomial data) but my take on this will obviously emphasize some themes more than others: particular aspects that I have focused on include non-parametric and non-linear regression models, model choice, time series and spatio-temporal models. Missing data models are also considered. As with my earlier Wiley books, a set of worked examples with documented code forms part of the book’s presentation and can be downloaded from the publisher’s website ftp://www.wiley.co.uk/pub/books/congdon, the WINBUGS site or STATLIB.

Peter Congdon
1.1 BAYESIAN UPDATING

Bayesian inference differs from classical inference in treating parameters as random variables and using the data to update prior knowledge about parameters and functionals of those parameters. We are also likely to need model predictions and these are provided as part of the updating process. Prior knowledge about parameters and updated (or posterior) knowledge about them, as well as implications for functionals and predictions, are expressed in terms of densities. One of the benefits of modern Monte Carlo Markov Chain (MCMC) sampling methods (e.g. Chib and Greenberg, 1995; Tierney, 1994; Gelfand and Smith, 1990; Gilks et al., 1996a; Smith and Roberts, 1993) is the ease with which full marginal densities of parameters may be obtained. In a regression model the parameters would be regression coefficients and possible variance parameters, and functionals of parameters might include elasticities (in econometrics) or effective dose (in biometrics).

The new Bayesian sampling-based estimation techniques obtain samples from the posterior density, either of parameters themselves, or functionals of parameters. They improve considerably on multiple integration or analytical approximation methods that are infeasible with large numbers of parameters. Nevertheless many issues remain in the application of sampling-based techniques, such as obtaining convergence, and choice of efficient sampling method. There are also more general
problems in Bayesian methods such as choice of priors (and possible sensitivity of inferences to alternative choices).

The basis for Bayesian inference may be derived from simple probability theory. Thus the conditional probability theorem for events $A$ and $B$ is that

$$
\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}
$$

Replacing $B$ by observations $y$, $A$ by a parameter set $\theta$ and probabilities by densities results in the relation

$$
p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.1}
$$

where $p(y|\theta)$ is the likelihood of $y$ under a model and $p(\theta)$ is the prior density, or the density of $\theta$ before $y$ is observed. This density expresses accumulated knowledge about $\theta$, or, viewed another way, the degree of uncertainty about $\theta$. It may also include working model assumptions (e.g. assumptions about the nature of error structures); for example, one model might assume uncorrelated errors over time or space and another model assume correlated errors.

Classical analysis via maximum likelihood focuses on the likelihood $p(y|\theta)$ without introducing a prior, whereas fully Bayesian analysis updates the prior information about $\theta$ with the information contained in the data. The denominator $p(y) = \int p(y|\theta)p(\theta)d\theta$ in (1.1) defines the ‘marginal likelihood’ or ‘prior predictive density’ of the data and may be set to be an unknown constant $c$. So posterior inferences about $\theta$ under (1.1) can be equivalently stated in the relation

$$
p(\theta|y) = \frac{p(y|\theta)p(\theta)}{c}
$$

or

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

This can be stated as ‘posterior is proportional to likelihood times prior’.

### 1.2 MCMC TECHNIQUES

The basis of modern Bayesian inference regarding $p(\theta|y)$ is the use of iterative MCMC methods that involve repeated sampling from the posterior distribution, using long, possibly multiple, chains of parameter samples. One is then interested in posterior summaries of parameters or functionals from the MCMC output in the form of expectations, densities or probabilities. These summaries typically include posterior means and
variances of the parameters themselves, or of functions $\Delta = \Delta(\theta)$ of the parameters, which analytically are

$$E(\theta_k|y) = \int \theta_k p(\theta|y) d\theta$$

$$\text{Var}(\theta_k|y) = \int \theta_k^2 p(\theta|y) d\theta - [E(\theta_k|y)]^2$$

$$= E(\theta_k^2|y) - [E(\theta_k|y)]^2$$

$$E[\Delta(\theta)|y] = \int \Delta(\theta) p(\theta|y) d\theta$$

$$\text{Var}[\Delta(\theta)|y] = \int \Delta^2 p(\theta|y) d\theta - [E(\Delta|y)]^2$$

$$= E(\Delta^2|y) - [E(\Delta|y)]^2$$

Often the major interest is in marginal densities of the parameters themselves. Let the model dimension be $d$, so that $\theta = (\theta_1, \ldots, \theta_d)$. Then the marginal density of the $j$th parameter $\theta_j$ is obtained by integrating out all other parameters

$$p(\theta_j|y) = \int p(\theta|y) d\theta_1 d\theta_2 \ldots \theta_{j-1}\theta_{j+1} \ldots \theta_d$$

The predictive density for new or replicate data useful in model checking and comparison is

$$p(y_{\text{new}}|y) = \int p(y_{\text{new}}, \theta|y) d\theta = \int p(y_{\text{new}}|y, \theta)p(\theta|y) d\theta$$

Posterior probabilities might relate to the probability that $\theta_j$ exceeds a threshold $b$, and involve integrals of the form

$$\Pr(\theta_j > b|y) = \int_{b}^{\infty} p(\theta_j|y) d\theta_j$$

Such expectations, densities or probabilities may be obtained analytically for conjugate analyses, such as a binomial likelihood where the probability has a beta prior. Results can be obtained under asymptotic approximations (Bernardo and Smith, 1994), similar to those used in classical statistics, or by analytic approximations (e.g. Laplace) based on expanding the relevant integral (Kass et al., 1988). Such approximations tend to be less good for posteriors that are not approximately normal or where there is multimodality. An alternative strategy facilitated by contemporary computer technology is to use sampling-based approximations based on the Monte Carlo principle. One such sampling method is
importance sampling (Geweke, 1989; McFadden, 1989), and other precursors of modern Bayesian sampling include data augmentation for Bayes inference in missing-data problems (Tanner and Wong, 1987).

1.3 THE BASIS FOR MCMC

The canonical Monte Carlo method assumes a sample of independent $d$-dimensional simulations $u^{(1)}, u^{(2)}, \ldots, u^{(T)}$ from a target density $\pi(u)$ whereby $E[g(u)] = \int g(u)\pi(u)du$ is estimated as

$$\bar{g}_T = \sum_{t=1}^{T} g(u^{(t)})$$

With probability 1, $g_T$ 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., 1996b). In order to sample approximately from $p(\theta|y)$, MCMC methods generate pseudorandom-dependent draws via Markov chains. Specifically let $\theta^{(0)}, \theta^{(1)}, \ldots$ be a sequence of random variables. Then $p(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(T)})$ is a Markov chain if

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

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

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

with $t$-step transition matrix $Q_{t,j}(t) = \Pr(\theta^{(t)} = s_j|\theta^{(0)} = s_i)$. Sampling from a constant one-step Markov chain converges to a stationary distribution $\pi(\theta) = p(\theta|y)$ if additional requirements\(^1\) on the chain are satisfied (irreducibility, aperiodicity and positive recurrence) – see Roberts (1996, p 46) and Norris (1997). Sampling chains meeting these

---

\(^1\)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 only be revisited after a 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.
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 \( \bar{g}_T \) 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 2.5% and 97.5% percentiles that provide credible intervals for the value of the parameter. A full posterior density estimate may be derived also (e.g. by kernel smoothing of the MCMC output of a parameter). For \( \Delta(\theta) \) its posterior mean is obtained by calculating \( \Delta^{(i)} \) at every MCMC iteration from the sampled values \( \theta^{(i)} \). The theoretical justification for this is provided by the MCMC version of the law of large numbers (Tierney, 1994), namely that

\[
\sum_{i=1}^{T} \Delta(\theta^{(i)})/T \to E_{\pi}[\Delta(\theta)]
\]

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

The probability (1.4) would be estimated by the proportion of iterations where \( \theta_j^{(i)} \) exceeded \( b \), namely \( \sum_{i=1}^{T} 1(\theta_j^{(i)} > b)/T \), where \( 1(A) \) is an indicator function which takes value 1 when \( A \) is true, 0 otherwise. Thus one might in a disease mapping application wish to obtain the probability that an area’s smoothed relative mortality risk \( \theta_k \) exceeds zero, and so count iterations where this condition holds, avoiding the need to evaluate the integral

\[
\Pr(\theta_k > 0) = \int_{0}^{\infty} p(\theta_k|y) d\theta_k
\]

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

\[
\hat{F}(d) = \sum_{i=1}^{T} 1(\Delta^{(i)} \leq d)/T
\]

The sampling output also often includes predictive replicates \( y_{new}^{(i)} \) 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 $\theta^{(i)}$ and then sampling $y_{\text{new}}$ from the likelihood model $p(y_{\text{new}} | \theta^{(i)})$. The posterior predictive density can also be used for model choice and residual analysis (Gelfand, 1996, sections 9.4–9.6).

### 1.4 MCMC SAMPLING ALGORITHMS

The Metropolis–Hastings (M–H) algorithm is the baseline for MCMC sampling schemes and is based on a binary transition kernel. Following Hastings (1970), the chain is updated from $\theta^{(i)}$ to $\theta^*$ with probability

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

with transition kernel $\alpha(\theta^* | \theta^{(i)}) f(\theta^* | \theta^{(i)})$, where $f$ is known as a proposal or jumping density (Chib and Greenberg, 1995). If the proposed update is rejected the next state is the same as the current state. The 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 $\theta^*$ is to $\theta^{(i)}$, and this depends on the variance $\sigma^2$ assumed in the proposal density. For a normal proposal density a higher acceptance rate follows from reducing $\sigma^2$, but with the risk that the posterior density will take longer to explore. 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.

If the proposal density is symmetric, with $f(\theta^* | \theta^{(i)}) = f(\theta^{(i)} | \theta^*)$, then the Hastings algorithm reduces to an algorithm used by Metropolis et al. (1953) for indirect simulation of energy distributions, whereby

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

A particular symmetric density in which $f(\theta^* | \theta^{(i)}) = f(|\theta^{(i)} - \theta^*|)$ leads to the random walk Metropolis (Gelman et al., 1996). While it is possible for the proposal density to relate to the entire parameter set, it is often computationally simpler to divide $\theta$ into blocks or components, and use componentwise updating, where updating is used in a generic sense allowing for possible non-acceptance of proposed values.
Thus let $\theta_{[j]} = (\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d)$ denote the parameter set omitting $\theta_j$ and $\theta_j^{(t)}$ be the value of $\theta_j$ after iteration $t$. At step $j$ of iteration $t + 1$ the preceding $j - 1$ parameters 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

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

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

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

The Gibbs sampler (Gelfand and Smith, 1990; Gilks et al., 1996a; Casella and George, 1992) is a special componentwise M–H algorithm whereby the proposal density for updating $\theta_j$ is 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 full potential for simulating marginal distributions by repeated draws recognised by Gelfand and Smith (1990). The Gibbs sampler involves parameter-by-parameter updating which when completed forms the transition from $\theta^{(t)}$ to $\theta^{(t+1)}$:

1. $\theta_1^{(t+1)} \sim f_1(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_d^{(t)});$  
2. $\theta_2^{(t+1)} \sim f_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_d^{(t)});$  
   •  
   •  
   d. $\theta_d^{(t+1)} \sim f_d(\theta_d|\theta_1^{(t+1)}, \theta_3^{(t+1)}, \ldots, \theta_{d-1}^{(t+1)}).$

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 initialize 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 \( \theta_j \) and treating other components as constants (Gilks, 1996). Consider a conjugate model for Poisson count data \( y_i \) with means \( \mu_i \) that are themselves gamma distributed; this is a model appropriate for over-dispersed count data with actual variability \( \text{Var}(y) \) exceeding that under the Poisson model. Suppose \( \mu_i \sim \text{Ga}(\alpha, \beta) \), namely

\[
f(\mu_i|\alpha, \beta) = \mu_i^{\alpha-1} \exp(-\beta \mu_i) / \Gamma(\alpha)
\]

and further that \( \alpha \sim E(a) \), and \( \beta \sim G(b, c) \), where a, b and c are preset constants; this prior structure is used by George et al. (1993). So the posterior density of \( \theta = (\mu_1, \ldots, \mu_n, \alpha, \beta) \) given \( y \) is proportional to

\[
e^{(-a\alpha)} \beta^{b-1} \prod_{i=1}^{n} \exp(-\mu_i) \mu_i^{y_i} \left\{ \prod_{i=1}^{n} \mu_i^{\alpha-1} \exp(-\beta \lambda_i) \right\}^{\frac{1}{\beta^a / \Gamma(\alpha)}}
\]

where all constants (such as the denominator \( \prod y_i! \) in the Poisson likelihood) are combined in the proportionality constant. It is apparent that the conditional densities of \( \mu_i \) and \( \beta \) are \( \text{Ga}(y_i + \alpha, \beta + 1) \) and \( \text{Ga}(b + n\alpha, c + \sum \mu_i) \) respectively. The full conditional density of \( \alpha \) is

\[
f(\alpha|y, \beta, \mu) \propto \exp(-a\alpha) \left[ \beta^a / \Gamma(\alpha) \right]^{\frac{n}{\prod \mu_i}}^{\alpha-1}
\]

This density is non-standard but log-concave and cannot be sampled directly (as can the gamma densities for \( \mu_i \) and \( \beta \)). However, adaptive rejection sampling (Gilks and Wild, 1992) may be used.

As examples of how M–H sampling might be carried out in practice, consider a Poisson density with unknown mean \( \mu \) and data \( y = (y_1, \ldots, y_n) \). One possible reference prior for \( \mu \) is \( p(\mu) = 1/\mu \) so the posterior density on \( \mu \) can be written

\[
p(\mu|y) \propto \prod_{i=1}^{n} \left[ \exp(-\mu) \mu^{y_i} / \mu \right]
\]

Suppose \( \mu \) is transformed to \( \theta = \log(\mu) \). Then the posterior in \( \theta \) includes a Jacobian adjustment, \( \partial \mu / \partial \theta = \exp(\theta) = \mu \), which cancels with \( p(\mu) \), and so

\[
p(\theta|y) \propto \prod_{i=1}^{n} \exp(-e^\theta + y_i \theta) = \exp \left( -ne^\theta + \sum_{i=1}^{n} y_i \theta \right)
\]
If the prior is placed directly on $\theta$ rather than $\mu$, e.g. $\theta \sim N(a, b)$ where $a$ and $b$ are known, a Jacobian adjustment is not needed. Then

$$p(\theta | y) \propto \exp \left( -n\theta + \sum_{i=1}^{n} y_i \theta \right) \exp \left[ -\frac{(\theta - a)^2}{2b} \right]$$

In either case one might use a symmetric normal proposal density to generate potential new values $\theta^*$ via (1.6). This density is centred on $\theta_1$ with variance $\sigma_p^2$ that might be set at a default such as $\sigma_p^2 = 1$ and increased or reduced in order to obtain better acceptance rates. Alternatively it might be based on $\text{Var}(\theta)$ from a maximum likelihood (ML) analysis but with variance increased, namely $\sigma_p^2 = K \text{Var}(\theta_{ML})$, $K > 1$.

If the variance is too low, acceptance rates will be high but the chain will mix slowly (i.e. move slowly through the parameter space) and converge slowly to $p(\theta | y)$. If the variance is too large, the acceptance rate will be low because the proposed new parameter values will have low values of the ratios $p(\theta^* | y) / p(\theta_1 | y)$. For example, if the required acceptance rate is $\rho$ (e.g. $\rho = 0.45$), then one might run the sampler for a certain number $N$ of iterations and compare the actual number of proposals accepted $N_1$ with required number $N_2 = \rho N$ and revise the scale according as $N_1 \leq N_2$ or $N_1 > N_2$. If $N_1 \leq N_2$,

$$\sigma_{p, \text{new}} = \sigma_{p, \text{old}} / (2 - N_1 / N_2)$$

while if $N_1 > N_2$,

$$\sigma_{p, \text{new}} = \sigma_{p, \text{old}} / \left[ 2 - (N - N_1) / (N - N_2) \right]$$

Multiparameter updating will involve a multivariate density such as a multivariate normal with dispersion matrix $\Sigma_p$. For example, consider the mean $\mu$ and standard deviation $\sigma$ in a univariate Student $t$ density with known degrees of freedom $\nu$. Taking

$$p(\mu, \sigma) = 1 / \sigma$$

as one among several possible reference priors, the posterior density is

$$p(\mu, \sigma | y) \propto \frac{1}{\sigma} \prod_{i=1}^{n} \left[ 1 + \frac{(y_i - \mu)^2}{\nu \sigma^2} \right]^{-(\nu+1)/2}$$

Transforming to $\theta = \log \sigma$, the Jacobian is $\partial \sigma / \partial \theta = \sigma$ and the posterior density in $\mu$ and $\theta$ is

$$p(\mu, \theta | y) \propto \frac{1}{e^\theta} \prod_{i=1}^{n} \left[ 1 + \frac{(y_i - \mu)^2}{\nu e^{2\theta}} \right]^{-(\nu+1)/2}$$
This is the comparator density in (1.5) or (1.6), with a possible proposal density, when \( \theta = \log \sigma \), being a bivariate normal centred at \((\mu^{(i)}, \theta^{(i)})\).

1.5 MCMC CONVERGENCE

There are many unresolved questions around the assessment of convergence of MCMC sampling procedures (Cowles and Carlin, 1996). It is generally accepted to be preferable 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; Gelman, 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 the variability of sampled parameter values between chains will considerably exceed the variability within any one chain. To measure variability of samples \( \theta^{(i)}_j \) within the \( j \)th chain \((j = 1, \ldots, J)\) define

\[
V_j = \frac{\sum_{t=s+1}^{s+T} (\theta^{(i)}_j - \bar{\theta}_j)^2}{(T - 1)}
\]

over \( T \) iterations after an initial burn-in of \( s \) 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. Convergence is therefore assessed from iterations \( s + 1 \) to \( s + T \).

Variability within chains \( V_W \) is then the average of the \( V_j \). Between-chain variance is measured by

\[
V_B = \frac{T}{J-1} \sum_{j=1}^{J} (\bar{\theta}_j - \bar{\theta})^2
\]

where \( \bar{\theta} \) is the average of the \( \bar{\theta}_j \). The scale reduction factor (SRF) compares a pooled estimator of \( \text{Var}(\theta) \), given by \( V_P = V_B / T + TV_W / (T - 1) \), with the within-sample estimate \( V_W \). Specifically the SRF is \((V_P / V_W)^{0.5}\) with values under 1.2 indicating convergence.

Parameter samples obtained by MCMC methods are correlated, which means extra samples are needed to convey the same information.
Additionally, as in any iterative estimation, there may be a delay in seeking the region of the posterior density where the modal value is located. The extent of correlation, and the convergence towards the modal region, will depend on a number of factors including the form of parameterization, the sample size, the complexity of the model and the form of sampling (e.g. block or univariate sampling of parameters).

A more recently proposed convergence statistic is that 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 \( \theta \) is obtained, i.e. the gap between \( 0.5 \alpha \) 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 \( T_J \) samples, the same \( 100(1 - \alpha)\% \) interval \( I_P \) is also obtained. Then the ratio \( I_P / I_U \) should converge to one if there is convergent mixing over different chains.

Analysis of sequences of samples from an MCMC chain amounts to an application of time series methods, in regard to problems 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 \( \theta^{(t)}, \theta^{(t+1)}, \theta^{(t+2)}, \ldots \), or from subsamples \( K \) steps apart, \( \theta^{(t)}, \theta^{(t+K)}, \theta^{(t+2K)}, \ldots \), 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 fluctuating rapidly 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 so that obtaining a G–R statistic for some parameters is not sensible. 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 \( \theta = (\theta_1, \theta_2, \ldots, \theta_d) \) 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 (Gelfand et al., 1995; Zuur et al., 2002).

1.6 COMPETING MODELS

Generally there are several possible competing models for the data, differing in likelihood or prior specifications. It is necessary either to choose between them, or to have some way of averaging inferences over them in terms of their relative probability or likelihood. From the conditional probability rule one can obtain posterior model probabilities. Let $M_j$ be one among several possible alternative models $\{M_1, \ldots, M_J\}$. Then with $B = y$ and $A = M_j$, we have that

$$p(M_j|y) = \frac{p(y|M_j)p(M_j)}{p(y)}$$

where the denominator is equivalent to $\sum_k p(y|M_k)p(M_k)$. The formal Bayes model choice is based on posterior model probabilities $p(M_j|y)$ and associated quantities, such as the Bayes factor $B_{jk} = p(y|M_j)/p(y|M_k)$ comparing models $j$ and $k$. This and other methods for comparing and checking models are considered in Chapter 2.

1.7 SETTING PRIORS

Priors may be chosen to encapsulate existing knowledge (e.g. based on results or parameter estimates from historical data) or to impart relatively little information in relation to that provided by the data. In the latter category are ‘reference priors’ that are constructed automatically without needing to choose tuning parameters. The latter are also sometimes known as objective priors (e.g. Casella and Moreno, 2002), whereas priors based to some degree on existing knowledge are known as subjective or elicited priors (Garthwaite et al., 2004). While objective priors have the benefit of being ‘off the shelf’ and of ‘letting the data speak for themselves’, they may also create problems in sampling-based estimation. Improper priors, for example, may lead to an improper joint posterior distribution even when all the full conditional posteriors are proper (Casella and George, 1992; Hobert and Casella, 1996).

In practice other principles may guide prior specification: for example, formal model choice (section 1.6 and Chapter 2) may be adversely affected by using diffuse or just proper priors. Complex models for
relatively small samples may require relatively informative priors in order to be identifiable from the data. Conjugate priors in which the posterior has the same form as the prior have advantages in tractability, and also in interpretation, since the prior can be interpreted in terms of a prior sample size or as pseudo data.

In many applications involving small or modest sample sizes posterior inferences about parameters or functionals of parameters may be affected by the prior used. Formal model choice via the Bayes factor may remain sensitive to prior specification for all sample sizes. In fact, the Bayes factor $B_{21}$ for model 2 vs. model 1 (where model 2 is more parameterized) tends to zero as the sample size $n$ increases. Also when model 2 contains an additional parameter and the diffuseness of the prior on that parameter is increased, $B_{21}$ tends to zero (Bartlett, 1957; Lindley, 1957).

On the other hand, problems with Bayes factor stability may be overstated. Although the Bayes factor may change with a change in the prior, it is pertinent to ask whether the ranking of the leading models changes when the prior is changed in reasonable ways. Possible alternatives to the formal Bayes factor include Bayes factors using minimal training samples from $y = (y_1, \ldots, y_n)$ to provide proper priors, examples being the fractional Bayes factor (O’Hagan, 1995) and the intrinsic Bayes factor (Berger and Pericchi, 1996a; 1996b).

A good principle for any application is to carry out a sensitivity analysis with a variety of prior specifications. This is especially so for parameters that are known to have sensitivity implications, e.g. the variances in hierarchical random effects models, or where the model is only weakly identified (e.g. see Chapter 8). A practical procedure implicit in the WINBUGS package and with some support in the literature (e.g. Besag et al., 1995) is the use of just proper, minimally informative priors. However, some argue that improper or just proper priors can be avoided by some subject matter reasoning. Thus Kadane and Lazar (2003, p 281) state that ‘if statisticians were to think about the reality underlying the parameter, they should always be able to describe it reasonably well using a prior distribution’.

1.8 THE NORMAL LINEAR MODEL AND GENERALIZED LINEAR MODELS

In this book the focus is on discrete outcome models $p(y|\theta)$ that generalize the tools used for continuous data, e.g. regression models, time series models and panel analysis. Such generalization usually starts
with non-hierarchical linear regression. The mainstay of regression
modelling for continuous data is the normal linear model whereby a
response variable \( y_i \), measured for subjects \( i = 1, \ldots, n \), is related linearly
to regressor variables (or predictors) \( x_{i1}, \ldots, x_{ip} \). Assuming that
predictors are measured without error, model discrepancies or measure-
ment errors in the \( y_i \) are expressed via errors \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \) assumed
to be independently normally distributed with a common variance \( \sigma^2 \).
Thus

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \\
\varepsilon_i \sim N(0, \sigma^2)
\]

Equivalently

\[
Y = X\beta + \varepsilon
\]

where \( Y \) is an \( n \times 1 \) column vector, \( \beta = (\beta_0, \ldots, \beta_p)' \) is a \( (p + 1) \times 1 \)
column vector of regression parameters and \( X \) is an \( n \times (p + 1) \) matrix of
predictor variables including an intercept \( x_{i0} = 1 \). Letting \( \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = X_i\beta \), where \( X_i \) is the \( i \)th row of \( X \), the
normal linear model may also be written in the form

\[
y_i \sim N(\eta_i, \sigma^2)
\]

However, for discrete responses (e.g. binary or count data) the
distribution of \( y_i \) will not usually be normal. Also, to ensure that
predictions \( \hat{y}_i \) are of the same form as \( y_i \) itself (e.g. if \( y_i \) is binary it is
necessary that \( \hat{y}_i \) be between 0 and 1), the mean \( \mu_i \) of the dependent
variable may need to be a transformation of the linear predictor \( \eta_i \) rather
than equalling it. In the class of generalized linear models (Nelder and
Wedderburn, 1972), \( y_i \) may follow one of a wide set of distributions
within the linear exponential family instead of being assumed normal. Thus

\[
y_i \sim p(y_i|\theta_i, \phi) \\
l(\theta_i, \phi|y) = \log[p(y_i|\theta_i, \phi)] = [y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i, \phi) \tag{1.7}
\]

where \( a, b \) and \( c \) are known functions that define the particular distribution, \( \theta_i = h(\eta_i) = h(X_i\beta) \) are unknown location parameters, and the scale
parameter \( \phi \) and hence \( a_i(\phi) \) may be known or unknown. Canonical
models are where \( \theta_i = \eta_i \). By using the relations \( \text{E}(\partial l/\partial \theta) = 0 \) and \( \text{E}(\partial^2 l/\partial \theta^2) + \text{E}(\partial l/\partial \theta)^2 = 0 \), one may show that \( \text{E}(y_i) = b'(\theta_i) \) and \( \text{Var}(y_i) = a_i(\phi) b''(\theta_i) \).

To ensure predictions of the right form, a transform or link function \( g(\mu_i) \) is usually needed to link the mean \( \mu_i = \text{E}(y_i) \) to the regression term in the predictors \( X_i \). So assuming a linear regression term, one obtains

\[
g(\mu_i) = X_i \beta = \eta_i
\]

with the inverse link being

\[
\mu_i = g^{-1}(X_i \beta) = g^{-1}(\eta_i)
\]

Different types of link may be used for any particular distribution (e.g. for the binomial distribution, commonly used links are the logit, probit or extreme value). The normal linear model with an identity link is in fact a special case of the linear exponential family, in which \( \mu_i = \theta_i = X_i \beta \), and the relevant functions are \( a_i(\phi) = \sigma^2 \), \( b(\theta_i) = \theta_i^2/2 \), \( c(y_i, \phi) = -0.5 y_i^2/\sigma^2 + \log(2\pi\sigma^2) \).

The Poisson with log link between \( \eta_i \) and \( \mu_i \) leads to \( \theta_i = \log(\mu_i) = X_i \beta \), with relevant functions being \( a_i(\phi) = 1 \), \( b(\theta_i) = \exp(\theta_i) \), \( c(y_i, \phi) = -\log y_i \).

Thus

\[
y_i \sim \text{Po}(y_i|\mu_i)
\]

where

\[
\text{Po}(y_i|\mu) = \exp[y_i \log(\mu_i) - \mu_i - \log y_i!]
\]

The variance = mean relation of the Poisson is obtained by differentiating \( b(\theta_i) \). For the binomial with mean \( \pi_i \) and \( n_i \) subjects at risk in the \( i \)th group or trial the likelihood may be written as

\[
\binom{n_i}{y_i} \left[ \frac{\pi_i}{1 - \pi_i} \right]^{y_i} (1 - \pi_i)^{n_i - y_i}
\]

This is obtained from the exponential family with \( \theta_i = X_i \beta = \log[\pi_i/(1 - \pi_i)] \), \( a_i(\phi) = 1 \), \( b(\theta_i) = n_i \log[1 + \exp(\theta_i)] \), and

\[
c(y_i, \phi) = \log\binom{n_i}{y_i}
\]

Thus the logit function is the canonical link between \( \pi_i \) and \( \eta_i = \theta_i \).

Random variation in subject means (\( \mu_i \) in the Poisson, \( \pi_i \) in the binomial) may be introduced. If modelled in the regression link, this might take the form

\[
g(\mu_i) = X_i \beta + \alpha_i
\]
where $\alpha_i$ might be normal, or a mixture of normals. Alternatively conjugate mixing may be considered, as when Poisson means $\mu_i$ are assumed to be gamma distributed over subjects, and when binomial probabilities are taken to be beta distributed (Nelder and Lee, 2000). The motivations for such extensions often lie in heterogeneity greater than expected under the standard exponential family models or to obtain improved estimates of varying means for subjects that allow ‘pooling of strength’.

Random variation between subjects may be unstructured and exchangeable (the features of the density are unaffected by permutation of the subject identifiers). However, structured variation is likely for observations arranged in space or time (e.g. spatially correlated variation if the observation units are areas). Consider the mixed model in spatial epidemiology (Besag et al., 1991) for Poisson distributed event totals $y_i$ with means $\mu_i = \lambda_i E_i$, where $E_i$ are expected events and $\lambda_i$ are relative disease risks (see Chapter 8). Then

$$\log(\lambda_i) = X_i \beta + \alpha_{i1} + \alpha_{i2}$$

where $\alpha_{i1}$ and $\alpha_{i2}$ are white–noise and spatial errors respectively. The impact of unstructured random effects $\alpha_{i1}$ is to pool strength over all areas towards the overall mean. With structured errors, the configuration of areas affects the nature of the pooling of strength: smoothing is towards local rather than global averages.

Classical ML analysis of models following particular forms of the exponential family density (1.7) involves Newton–Raphson iteration or iteratively reweighted least squares, and usually relies on asymptotic normality to provide parameter densities or hypothesis tests. However, the asymptotic approximations may not hold for small or moderate sample sizes or for non-linear models (e.g. see Zellner and Rossi, 1984). Inferences based on asymptotic normality of the ML parameters can be affected, perhaps distorted, by choice of parameterization, by sample size, by the form of regression (non-linear vs. linear) and so on. An illustration by Pastor-Barriuso et al. (2003) is for a logit model subject to a change point in the impact of regressors. Bayesian analysis of discrete data follows the generalized linear model (GLM) structure but is not constrained to asymptotic normality to obtain posterior inferences.

---

$E_i$ denotes the expected deaths, taking account of the size of the population exposed to risk and its age structure but assuming a standard death rate schedule (e.g. national death rates).