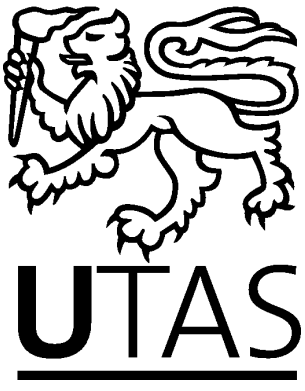
NONLINEAR APPLICATIONS OF MARKOV CHAIN MONTE CARLO

by

Gregois Lee, B.Sc.(ANU), B.Sc.Hons(UTas)

Submitted in fulfilment of the requirements
for the Degree of Doctor of Philosophy

Department of Mathematics
University of Tasmania
2010



I declare that this thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due acknowledgement is made in the text of the thesis.

Signed: _____
Gregois Lee

Date: _____

Signed: _____
Gregois Lee

Date: _____

ACKNOWLEDGEMENTS

To Simon Wotherspoon for making this probable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Statistics and Computing

In the early 20th century data analysis was constrained by computability. Calculations were performed by hand, providing real practical limits on the types of problems which were tractable. Salsburg (2002) provides a calculation showing that at least 8 months of 12-hour days would have been required for R. A. Fisher to have produced the tables in his "Studies in Crop Variation I" (Fisher, 1921) with the mechanical means at his disposal. It is hardly surprising that the emphasis during this period remained on linear models – problems soluble by ordinary least squares, with the tools at hand.

It was not until the 1960s that nonlinear regression began to appear regularly in the literature, and no accident that this eventuates concurrently with the appearance of machines to automate iterative calculations. The heavier computational burden had previously been insurmountable. But even after the advent of early computers, great emphasis was placed on the development of algorithms which could make *efficient* use of limited hardware resources – processors were slow and memory limited. Research into algorithms became synonymous with efficiency, and the attendant O-notation. The *Fast Fourier Transform* of Cooley and Tukey (1965) provides the archetypal example of the era. The explicit reference to speed in the title underscores the imperative.

In the early 21st century, the situation has improved markedly. Computing power is cheap and relatively abundant, and software is designed with re-use of objects and systems integration in mind. There has been a co-evolution of research into modelling methods. Modelling frameworks have diversified, and are now capable of representing a much broader range of observable phenomena. Informed by Tukey's observation "Far better an approximate answer to the right question, ... than an exact answer to the wrong question" (Tukey, 1962), we build models which more accurately reflect our understanding of reality. Increasingly, we are asking the right questions.

Indeed, since the 1970s there has been rapid development in methods which extend the general linear model, stimulated by the development of Generalized Linear Models (GLMs) (McCullagh and Nelder, 1989). These allow response residuals to be modelled using alternatives to the Gaussian distribution and conditional expectations to be related to covariates via a link function $\eta(\boldsymbol{\theta})$, rather than a direct linear relationship. The general linear model can then be seen as a GLM with an identity link function and a normally distributed response. The principal appeal of the framework is that the adoption of the exponential family as the basis guarantees the likelihood to be log-concave and unimodal so that estimation is straightforward. The adoption of GLMs has greatly extended the realm of linear models and vastly enhanced the scope of linear statistical modelling applications. What remains conspicuously absent is concurrent progress of a similar order in pursuit of *nonlinear* models, where the properties of the solution surface are more complex.

The adoption of Markov Chain Monte Carlo (MCMC) techniques by the statistical community represents a significant new chapter in stochastic modelling. MCMC methods provide a flexible and powerful base from which realistic stochastic models can be built. They are particularly important because models developed in this framework need not have analytical tractability. Provided that the relationships between the component parts are specified, samples can be obtained from the density of the resulting model, allowing estimation and inference from non-standard distributions. This is an enormously empowering development. Most importantly, it promotes the construction of more realistic models. Data no longer need be forced into overly simplistic models just because they are the only soluble forms. Models can now be developed to fit available data.

The development of MCMC tools have fundamentally altered the way that statisticians go about their business. But it is not merely statisticians who benefit. Greater accessibility of realistic modelling methods has led to statistical modelling taking a firm hold in primary research across a wide range of applied disciplines. The exchange of purely deterministic models in favour of more realistic stochastic models represents a paradigm shift in the foundation of science.

1.2 Structure of the Thesis

This thesis considers the application of Markov Chain Monte Carlo (MCMC) methods to problems which extend the general linear model in various nonlinear ways. The framework in which these applications are developed is exclusively Bayesian, though the the methods themselves are equally applicable, if perhaps less commonly used, in a likelihood context. Geyer (1995) and Diebolt and Ip (1995), and the references therein, provide details of non-Bayesian applications of MCMC.

1.2.1 Chapter Outlines

Chapter 2 provides an overview of Bayesian methodology and nomenclature. We begin by establishing the historical context of the development of Bayesian methods and the recent rise in popularity of their use. This is followed by the introduction of Bayes' theorem and an example illustrating its use. The prior distribution is then examined in greater detail, with a discussion of relative states of prior information, the use of conjugate priors, and further development of the example. We then turn to a discussion of the posterior distribution, its role in Bayesian inference and estimation, and how results are reported and interpreted.

Chapter 3 provides an overview of Markov Chain Monte Carlo (MCMC) methods. In particular the Gibbs Sampler and the Metropolis-Hastings algorithm are introduced. These computational techniques underpin the implementation of all the methods explored in later chapters. The concept of partitioning the posterior into manageable parts is fundamental to implementing an appropriate sampling scheme. In cases where this can be achieved with the full conditional distribution available in an analytical form, Gibbs Sampling provides a very efficient mechanism for producing posterior samples. In cases where a closed-form solution is not available, a Metropolis-Hastings sampling scheme can be implemented. This involves generating samples from a proposal distribution, and subjecting proposed points to a rejection filter such that candidates are accepted with probability matching that of the target posterior distribution. The details of how this can be achieved are described before moving on to a discussion of convergence issues in MCMC. Gelman and Rubin's potential scale reduction factor $\widehat{R}$ is described. Finally, we give a brief account of visual inspection of the MCMC chain trace and its use as a qualitative aid in assessing whether the support of the posterior has been appropriately sampled.

Having established the conceptual and computational framework for the thesis, we turn to application of these tools.

Chapter 4 demonstrates the use of Bayesian MCMC in the nonlinear regression context. This allows the model mean to take the form of a parametric nonlinear function. We begin by reviewing Least Squares parameter estimation in nonlinear models, and propose an MCMC alternative. A simple example is developed to illustrate the use of the MCMC method. Next, we undertake a detailed evaluation of the method's performance in the context of growth curve models. The main focus of the chapter is a comparative analysis with the Least Squares results provided by Ratkowsky (1983). We show that the MCMC method offers results comparable to those obtained under Least Squares across a range of nonlinear regression problems, and offers several significant advantages.

Chapter 5 describes a method for transforming the response using Gibbs Sampling. That is, we consider nonlinear transformations of the response such that the criteria required by the general linear model are met by the transformed variable, and modelling may proceed under the general linear framework. The method si-

multaneously estimates the response transformation along with parameters to fit an assumed linear model. This approach is of particular interest because it incorporates uncertainty in the choice of transformation into the modelling process, in contrast with many other methods of transformation currently in use. We demonstrate the successful application of the method by reconsidering an example put forward by Box and Cox (1964), and provide a comparative analysis with the results suggested by their method.

Chapter 6 extends the method introduced in the previous chapter to consider cases where the response can be represented as the sum of monotonic functions of covariates: Monotonic Additive Models. This is another example of relaxing linearity as the assumed form of the conditional mean response, but here the relationship of the response to each covariate is nonparametric in form.

In principle, our approach is similar to that of nonparametric additive models (Hastie and Tibshirani, 1990), in that we seek to find a series of nonparametric functions representing the relationship of each covariate to the response and combine these in an additive framework. Yet the mechanism by which we achieve this end is quite novel, and bears no relationship to the class of additive models considered by Hastie and Tibshirani (1990) beyond its conceptual structure. We develop a series of examples using simulated and real data to critically evaluate the performance of the method.

To extend the techniques of earlier chapters for use in multivariate or mixed-effects models, a method for modelling the correlation structure between estimands is required. In Chapter 7 we develop techniques for estimating correlations using variants of the Gibbs sampler. The principal method uses a rejection sampling strategy based upon carefully selected Gamma distributions. Another method is provided for use in cases where this approach can be demonstrated to operate with low efficiency. Both methods are tested against simulated data to illustrate their relative merits. A detailed account of the practical issues involved in the successful implementation of the techniques is provided.

Finally, Chapter 8 summarises the results presented in the thesis, offers concluding remarks, and identifies areas where the techniques developed in the thesis can be extended in future research. We propose enhancements to individual methods and applications where they may be applied in concert.

1.2.2 `R` Source Code

Many of the examples in the thesis could be implemented using variants of the BUGS (Bayesian inference Using Gibbs Sampling) environment (Lunn et al., 2000; Thomas et al., 2006; Plummer, 2003). However, examples which feature strong correlation between parameters mix poorly under the Gibbs sampler, and require very long run times to produce reasonable estimates. In particular, Chapters 4 & 7 investigate problems where it would be unrealistic to ignore cases of highly correlated estimates.

Rather than distract the reader by chopping and changing between development environments we elected to conduct all analyses and development of source code in the preparation of this thesis in the `R` statistical environment (R Development Core Team, 2009). The source code used in producing results presented herein is provided in Appendix A.

Chapter 2

Bayesian Analysis

2.1 Historical Context

Early Development

In 1763 the *Philosophical transactions of the Royal Society of London* published a posthumous article by the reverend Thomas Bayes, with the title *Essay Towards Solving a Problem in the Doctrine of Chances* (reprinted in *Biometrika* using notation familiar to the modern reader as Barnard and Bayes, 1958). In essence, the article presented a model for inductive logic, *inverse probability*, using observational data to enhance an observer's prior beliefs in the probability of an event occurring. The theory was accepted at the time of it's publication, and further developed into what can be recognised as the foundations of modern probability theory by the eminent French mathematician Pierre-Simon Laplace in the late 18th and early 19th century (Stigler, 1990; Hald, 1998; Dale, 1999).

Bayes' mathematics remain unchallenged. But the logician George Boole (1854, republished as Boole, 2008) is attributed with being the first to call into question the *philosophical* validity of allowing subjective criteria to enter into the probability calculus, and sparking a controversy that troubled statisticians for over a century.

The 20th Century

In the context of the ensuing debate, R. A. Fisher was at pains to justify the foundations of the emerging discipline of statistics in opposition to Bayesian principles (Fisher, 1922, 1925b). The result was what is today recognised as a *frequentist* approach; that is, a system of statistical procedures which are founded upon, and justified relative to, consideration of *all* samples which could conceivably arise in a given context. Despite Fisher's insight that for statistics to be useful, it must be capable of deriving results from sample sizes that researchers can realistically achieve (see, for example, Fisher, 1925a, 1935, re-issued as Fisher et al. (1990)), it is common for frequentist theory to be justified asymptotically – in light of infinite sample sizes.

Harold Jeffreys, a contemporary of Fisher, was interested in establishing a consistent probabilistic basis for the scientific method. His outlook was Bayesian, and his scientific output prodigious. A succinct summary of Jeffreys' contributions to statistics and how these relate to the work of Fisher and Bayes is given in Jeffreys (1974). Jeffreys' (ultimately) influential works *Scientific Inference* (1931, re-issued as Jeffreys, 2007) and *Theory of Probability* (1961, re-issued as Jeffreys, 1998) were instrumental in re-kindling Bayesian methods in statistics in the latter half of the 20th century. Other important contributions following on from the work of Jeffreys were made by L. J. Savage (Savage, 1954), Dennis Lindley (particularly Lindley, 1965a,b), Bruno de Finetti (de Finetti, 1974, 1975), and Edwin Jaynes (Jaynes, 2003), among others. Press (2003) contains biographical sketches of these authors and synopses of their work.

A Recent Renaissance

In recent years the ubiquity of unprecedented desktop computing power has enabled Bayesian methods to flourish, both in the statistical community and outside it. Berger (2000) describes the increase in Bayesian activity at the turn of the millennium, as indexed by the number of published research articles, the number of books, and the extensive number of Bayesian applications appearing in applied disciplines. This trend has continued to date. In addition to mainstream statistical tomes (prominent examples include Gelman et al., 2003; Carlin and Louis, 2008), texts clearly targeted at beginning undergraduate students are emerging (see, for example, Bolstad, 2007), with the implication that in some quarters at least, Bayesian methods are being incorporated into students' foundation in statistics. Significantly, practitioners in applied disciplines have embraced the Bayesian approach, with texts commonly appearing, for example, in fields such as clinical studies (Spiegelhalter et al., 2004; Broemeling, 2007; Grobbee and Hoes, 2008), ecology (McCarthy, 2007; Bolker, 2008; Royle and Dorazio, 2008; King et al., 2009), economics (Lancaster, 2004; Geweke, 2005; Greenberg, 2007; Koop et al., 2007), epidemiology (Banerjee et al., 2003; Moyé, 2007; Lawson, 2008), finance (Singpurwalla, 2006; Scherer and Martin, 2007; Rachev et al., 2008), and the social sciences (Gelman and Hill, 2007; Gill, 2007; Jackman, 2009). Clearly, Bayesian methods have come of age.

2.2 Introduction

Bayesian analysis differs from its frequentist counterpart in two important respects:

i) from the Bayesian viewpoint *all* estimands are random variables with an associated probability distribution, and

ii) Bayesian analysis provides a formal mechanism for admitting *subjective* information into the model structure.

The implications of these attributes are profound. Bayesian analysis is fundamentally different from frequentist analysis, both in terms of mechanics and interpretation. This chapter will outline these differences and their implications.

2.3 Bayes' Theorem

Given a vector of observations $\boldsymbol{y}$ whose probability distribution $p(\boldsymbol{y}|\boldsymbol{\theta})$ is conditional upon the parameter vector $\boldsymbol{\theta}$, which itself has probability distribution $p(\boldsymbol{\theta})$, then

$$p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{y}), \tag{2.1}$$

where the notation $p(a|b)$ denotes the probability of an event a occurring subject to the condition that b occurs. As we will generally be interested in the distribution of the parameter vector, given the observed data, we write *Bayes' Theorem* as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{2.2}$$

where $p(\boldsymbol{y})$ is a normalisation constant, ensuring that the expression on the right hand side of (2.2) integrates, or sums, to unity.

From (2.2) we see that the distribution of the parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$ is the product of two terms: the conditional probability of observing the data $\boldsymbol{y}$ given the parameters $\boldsymbol{\theta}$, and the probability distribution of those parameters, $p(\boldsymbol{\theta})$. Viewed as a function of $\boldsymbol{\theta}$, the first term is the *likelihood* function for the parameters $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$. The second term is referred to as the *prior distribution* of $\boldsymbol{\theta}$, to reflect the idea that the information contained in this term is independent from, and can be considered as arising prior to, observation of the data. Finally, the product $p(\boldsymbol{\theta}|\boldsymbol{y})$ is termed the *posterior* distribution, reflecting the idea that it represents the state of knowledge resulting from the modification of prior knowledge regarding $\boldsymbol{\theta}$ by the information contained in the observations $\boldsymbol{y}$.

Example

Suppose we have a single observation y, the realisation of a normally distributed random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ with known variance σ^2. The likelihood function for $\theta = \mu$ is then

$$l(\theta; y) = p(y|\theta) = \mathcal{N}(y|\mu, \sigma^2) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$

where $\mathcal{N}(y|\mu, \sigma^2)$ indicates the likelihood of observing the data y conditional on the values of the parameters μ and σ^2. As the likelihood is a function of the parameters, *Maximum Likelihood* methods seek to estimate θ such that it maximises the

probability of observing the data y.

Next we require a distribution to represent our prior beliefs regarding possible values for the value of θ. If we consider that a reasonable choice is a normal distribution with mean η and variance ϕ, we may write $p(\theta) = \mathcal{N}(\mu|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\eta, \phi)$, are *hyper-parameters* in the model. That is, they provide structure to the model by informing how model parameters are to be constrained. It is now apparent that we are considering a *hierarchy* of model parameters: here μ is constrained by the choice of $\boldsymbol{\lambda}$. It is possible to estimate all elements of the parameter hierarchy from the data (see, for example, Gelman et al., 2003), but for the sake of exposition we do not consider that case here. With the likelihood and prior in place we are now ready to apply Bayes' Theorem

$$\begin{aligned} p(\theta|y) &= \frac{\mathcal{N}(y|\mu, \sigma^2)\, \mathcal{N}(\mu|\lambda)}{p(y)} \\ &\propto \mathcal{N}(y|\mu, \sigma^2)\, \mathcal{N}(\mu|\eta, \phi^2) \\ &= \mathcal{N}\left(\mu \,\middle|\, \frac{\phi^2}{\sigma^2 + \phi^2}\, y + \frac{\sigma^2}{\sigma^2 + \phi^2}\, \eta,\ \frac{\sigma^2\phi^2}{\sigma^2 + \phi^2}\right). \end{aligned} \tag{2.3}$$

That is, the distribution of μ given the observation y is also a normal distribution, with the mean and variance shown in (2.3). The expected value $E(\theta|y)$ is the weighted average of the observation value y and the prior mean η, with weights dependent on the respective levels of uncertainty associated with the prior and the likelihood.

The variance associated with θ is constructed in terms of its reciprocal, the *precision* $\tau = \frac{1}{\sigma^2}$, which is more convenient to work with in a Bayesian context and offers an intuitive interpretation here: the posterior precision is the sum of the precisions from the likelihood and the prior, $\frac{1}{\sigma^2} + \frac{1}{\phi^2}$.

2.4 The Prior Distribution

The prior distribution provides the formal mechanism through which subjective information can be included in the modelling process. It allows freedom to nominate a distribution of values associated with the parameter vector $\boldsymbol{\theta}$, consistent with the range of values that $\boldsymbol{\theta}$ is believed likely to assume.

In the example of §2.3 we saw that the expected value $E(\theta|y)$ was the precision weighted average of the of the prior mean and the observed data, and that the weights were determined by the relative precision of these distributions. Influence over the location of $p(\theta|y)$ was provided by the choice of the prior mean η, and the *degree* to which the preference for that value was exerted was provided by the prior

precision, $\tau = \frac{1}{\phi^2}$.

This ability to incorporate prior information into the probability calculus in a carefully controlled formal manner has been cited (Lindley, 1965a,b; de Finetti, 1974, 1975; Berger, 2006, for example) as a major advantage of Bayesian analysis. However, it is precisely this feature which has formed the focal point for contention with frequentist stalwarts. The power to influence analytical outcomes implies a responsibility to understand and evaluate the implications of such choices. The remainder of this section will consider the issues associated with choosing appropriate prior distributions.

2.4.1 Prior Propriety

The prior distribution $p(\boldsymbol{\theta})$ provides a probability model of plausible values for the parameter vector $\boldsymbol{\theta}$. A fundamental property we expect of any probability function is

$$\left.\begin{array}{l} \int p(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta} \\ \sum p(\boldsymbol{\theta}) \end{array}\right\} = 1,$$

regardless of whether $\boldsymbol{\theta}$ is continuous or discrete. Situations arise in which we may wish to express a lack of preference for *any* value of θ. Yet if we try to take $p(\boldsymbol{\theta})$ to be uniform over the entire real line, the prior

$$p(\boldsymbol{\theta}) = c > 0, \quad -\infty < \boldsymbol{\theta} < \infty,$$

is not a proper probability density since the integral

$$\int_{-\infty}^{\infty} p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = c \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{\theta}$$

does not exist for any value c. Although the result is not a probability density in a strict sense, such distributions are sometimes employed as prior distributions to express indifference between values of $\boldsymbol{\theta}$ in some local region where the likelihood function attains appreciable density. These distributions are termed *improper* priors to reflect their degenerate nature. The posterior distributions which arise from improper priors are frequently proper probability densities, allowing some flexibility in the specification of priors without impediment to subsequent inference.

2.4.2 Non-informative Priors

A prior distribution which does not favour any particular value of θ over any other may be said to be "non-informative" for θ. Such distributions have the appeal that posteriors arising from their use are free from the subjective influence of the prior. For this reason they may sometimes be used as a "reference" prior; a benchmark

against which the sensitivity of posterior outcomes for other prior distributions may be evaluated.

Box and Tiao (1973) provide an example of non-informative priors supposing a sample $\boldsymbol{y} \sim \mathcal{N}(\theta, \sigma^2)$, where σ is known. The likelihood for θ is

$$l(\theta|\sigma, y) \propto \exp[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2], \tag{2.4}$$

and under this scenario a non-informative prior is locally uniform in θ

$$p(\theta) \propto c. \tag{2.5}$$

However, if the quantity of primary interest were instead $\kappa = \theta^{-1}$, the likelihood becomes

$$l(\kappa|\sigma, y) \propto \exp[-\frac{n}{2\sigma^2}(\kappa^{-1} - \bar{y})^2], \tag{2.6}$$

and since

$$p(\kappa) = p(\theta)\Big|\frac{d\theta}{d\kappa}\Big| = p(\theta)\theta^2 \propto \kappa^{-2}, \tag{2.7}$$

the corresponding non-informative prior for κ is proportional to κ^{-2}. In general, if a prior distribution is locally uniform for some (monotonic) function of the parameter(s) of interest $\phi(\theta)$, then the corresponding non-informative prior for θ is proportional to $|d\phi/d\theta|$.

2.4.3 Vague Priors

In practice we need not be overly concerned with strictly non-informative priors, provided that the prior is *relatively* uninformative when compared to the information contained in the data. The prior should include all plausible values for θ but need not be concentrated around the true value, because information regarding θ in the data will typically outweigh any reasonable prior probability specification.

Example (continued)

Consider an extension of the example presented in §2.3 where a sample of size n is available. Since the sample mean $\bar{y}$ is sufficient for θ, $p(\theta|\boldsymbol{y}) = p(\theta|\bar{y})$, and since $p(\bar{y}|\theta) = \mathcal{N}(\theta, \sigma^2/n)$,

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \mathcal{N}\left(\theta \,\Big|\, \frac{(\sigma^2/n)\mu + \phi^2\bar{y}}{(\sigma^2/n) + \phi^2}, \frac{(\sigma^2/n)\phi^2}{(\sigma^2/n) + \phi^2}\right) \\ &= \mathcal{N}\left(\theta \,\Big|\, \frac{\sigma^2\mu + n\phi^2\bar{y}}{\sigma^2 + n\phi^2}, \frac{\sigma^2\phi^2}{\sigma^2 + n\phi^2}\right). \end{aligned} \tag{2.8}$$

From (2.8) it is again clear that the posterior is a weighted average of the prior and data values, but now it is also apparent that the relative weighting is proportional to the number of observations n. As the number of observations increases the relative influence of the prior distribution is diminished. If n is sufficiently large, the influence exerted by the prior distribution is overwhelmed and becomes negligible.

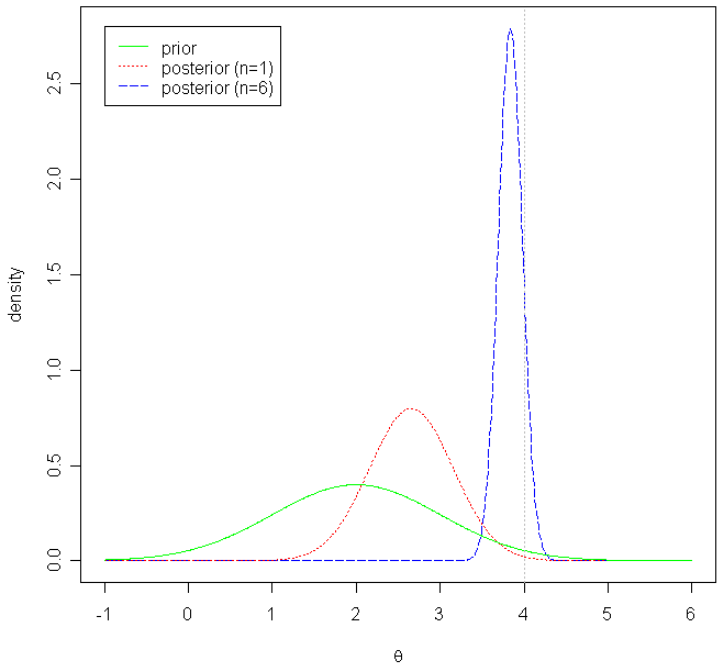


Figure 2.1: Data Driven Posterior.

Where the prior is relatively uninformative, n does not need to be very large. Figure 2.1 shows the relative influence of the prior $p(\theta) \sim \mathcal{N}(2, 2)$ on the posterior distributions resulting from simulated observational data $\boldsymbol{y} \sim \mathcal{N}(4, 1)$ for $n = 1$ and $n = 6$, respectively. A single observation from these data is sufficient to establish that the mean of the posterior is considerably larger than that of the prior. With 6 observations, the posterior has become focused near the true value of $\theta = 4$ with far greater precision.

2.4.4 Conjugate Priors

In the example of §2.3 the application of Bayes theorem using a normal prior led to a posterior which was also a normal distribution. Raiffa and Schlaifer (1961) pointed out that some priors give rise to posteriors from the same family of distributions. Formally, we may write that if $\mathcal{P}$ is a class of prior distributions for θ, and $\mathcal{F}$ is a class of sampling distributions, then $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \in \mathcal{P} \quad \text{for all} \quad p(.|\boldsymbol{\theta}) \in \mathcal{F} \quad \text{and} \quad p(.) \in \mathcal{P}. \tag{2.9}$$

Of particular interest are the *natural* conjugate families for the prior, which have the same functional form as the likelihood. This property then offers a very convenient structure for performing iterative calculations with Bayes Theorem. The conjugate prior gives rise to a posterior distribution from the same density family. Thus the posterior resulting from iteration t can be regarded as the prior during a subsequent iteration $t + 1$. Successive modification of the posterior by subsequent iterative application of Bayes Theorem is referred to as *Bayesian learning*, reflecting the update of prior beliefs in light of the data.

2.5 The Posterior Distribution

The left-hand side of (2.2) results from combining information available in the observed data $\boldsymbol{y}$, with any prior knowledge regarding the range of likely values for the parameters, via the likelihood function $l(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ and prior distribution $p(\boldsymbol{\theta})$, respectively. The distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is commonly referred to as the *posterior* distribution for $\boldsymbol{\theta}$, because it expresses the state of knowledge regarding parameter values after modifying prior beliefs by the information available in the observations.

2.5.1 Estimation

In cases where the posterior can be written in closed analytical form, summaries may be obtained directly from the properties of the posterior distribution (see, for example, Gelman et al., 2003). However, we will commonly be interested in problems for which no analytical derivation of the posterior is available, and in general we determine the posterior distribution via simulation. Specific details of the techniques employed for this purpose are deferred to Chapter 3. In general a numerical procedure will produce a matrix containing a sample of arbitrary size from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, where each of the m rows of the matrix are simulation draws for each of k parameters $\theta_j, j = 1, \ldots, k$, and n unobserved data points $y_i, i = 1, \ldots, n$. Estimates are obtained by calculating summary statistics from the simulated posterior samples.

2.5.2 Inference

The posterior distribution comprises the current state of knowledge regarding the distribution of $\boldsymbol{\theta}$ conditional upon the data. Because the posterior incorporates *all* of the available information regarding $\boldsymbol{\theta}$, inference in the Bayesian context simply requires extracting summary information from the posterior distribution for the quantities of interest. One may choose such summaries arbitrarily. For example, we will typically report basic summary quantities such as the mean, median, and various quantiles for each of the parameters of interest θ_j, simply by calculating the

requisite quantity from the appropriate column of the simulated posterior. However, we are also able to report any summary quantity for an arbitrary function $f(\boldsymbol{\theta})$ of the parameters θ_j, should we be interested in some transformation of the parameters.

2.5.3 Visualisation

It is also useful to examine univariate or pairwise bivariate graphical displays of the posterior summaries. Such graphics can be a valuable aid when critically assessing the appropriateness of a given model, in addition to inferential reporting. Again, these are directly available from the simulated posterior sample.

2.5.4 Reporting Results

Because the posterior distribution contains all of the information arising from a Bayesian calculation, it provides a complete reference source for all quantities of interest. For precisely that reason, it can be quite difficult to convey the posterior in an undigested form, especially as we are commonly interested in results from complex multi-parameter models. It is usual to report quantities which summarise relevant attributes of the posterior. Often, this may be achieved by taking simple functions of the quantity of interest, with the mean and various quantiles being common examples. However, Bayesian interval estimation and interpretation is markedly different from the frequentist approach.

Credible Intervals

The term *credible interval* first appears in the literature in Edwards et al. (1963), and refers to the Bayesian analog of the frequentist confidence interval for a univariate posterior quantity. In particular, for some predetermined α value a $1-\alpha$ credible interval (a, b) for θ, given the data $\boldsymbol{y}$, may be determined as

$$1 - \alpha = P(a < \theta < b|\boldsymbol{y}) = F(b) - F(a) \tag{2.10}$$

where $F(.)$ denotes the posterior CDF. Bayesian credible intervals and frequentist confidence intervals usually have identical endpoints in cases where the prior is uninformative. However, the frequentist confidence interval merely provides an assurance that $100(1-\alpha)\%$ of the intervals so constructed will contain the true parameter value, it makes no claim about the validity of any particular interval. By contrast, the Bayesian credible interval is constructed directly from the posterior for the parameter of immediate interest, and has the advantage of being directly interpretable as a probability statement. The $(1-\alpha)$ credible interval contains the true value of θ with probability $(1-\alpha)$.

Credible intervals generalise to *credible regions* for higher dimensions but since the results in this thesis are chiefly concerned with summarising univariate marginal

posterior quantities, such generalisations are not pursued. The interested reader may refer to Box and Tiao (1973).

While (2.10) allows the determination a credible interval (a, b) with credibility level $(1-\alpha)$, it does not specify uniquely defined endpoints. We describe two alternative methods for determining specific univariate credible interval estimates.

Highest Posterior Density Intervals

The $100(1-\alpha)\%$ *Highest Posterior Density* (HPD) interval for θ is the set

$$C = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}|\boldsymbol{y}) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant such that

$$p(C|\boldsymbol{y}) \geq 1-\alpha.$$

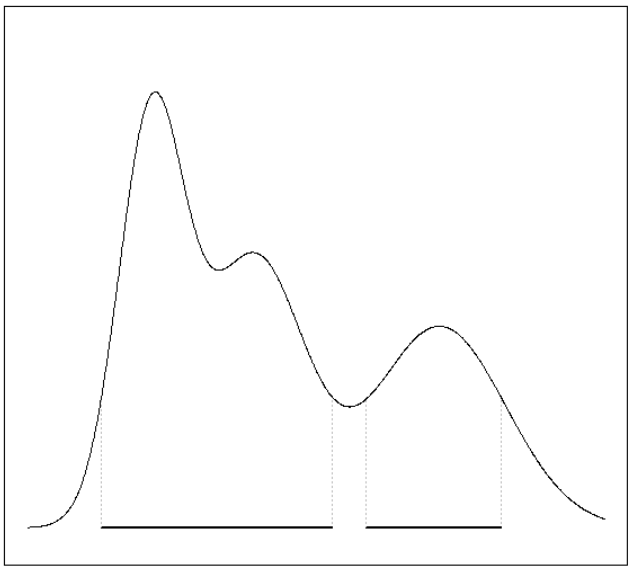


Figure 2.2: Disjoint Highest Posterior Density Interval.

Thus, the HPD interval provides the most likely values for θ, determined at the $1-\alpha$ level. It will therefore also be the shortest credible interval which can be formed at that probability level. The definition allows the interval so formed to be disjoint in cases of multi-modal posteriors where α is sufficiently large. For example,

the schematic diagram provided as Figure 2.2 shows the 80% HPD interval for a mixture of normal distributions. The rigour provided by the HPD interval comes at a computational cost. Determination of $k(\alpha)$ against an arbitrary distribution requires an iterative method to calculate the interval endpoints.

Equal Tail Intervals

A simpler approach to determine a specific credible interval can be found by taking the interval such that "tails" of equal probability are excluded from the extremities of the marginal distribution for the quantity of interest.

Frequentist confidence intervals rely on an assumption of normality and this provides symmetry: equal probability tails are supported by intervals of equal length on the support. No such symmetry is assumed in the Bayesian context, where the posterior distribution can take any form. In general the posterior will not be symmetrical and excluding "tails" of equal probability implies the exclusion of unequal segments at each end of the interval. For multi-modal or highly skewed posteriors, intervals based on the exclusion of equal tails may provide considerable differences in support for θ from their HPD counterparts. However, for unimodal, moderately skewed distributions the difference between the intervals produced by the two methods will be slight.

Thus, while recognising that HPD intervals are valuable in cases where the additional computational burden is warranted, equal-tail intervals are used as the default throughout this thesis. This also allows us to take advantage of the fact that equal-tail intervals are readily available from the simulated posterior quantiles at no additional computational expense.

2.6 Conclusion

In this chapter the basic principles of Bayesian analysis were introduced. After briefly outlining the historical context of Bayesian methods, Bayes Theorem was introduced, and issues of interest regarding the prior and posterior distributions were discussed, including the utility of conjugate forms and the nature of Bayesian inference and reporting. In summary, the following points are of interest:

- Bayesian methods have a (relatively) long history and well established pedigree. Despite controversies surrounding their use in the 20th century, there is no doubt regarding the mathematics of Bayes' theorem.
- Bayesian methods deal with unknown quantities probabilistically. That is, unknown parameters are considered as random variables rather than fixed, unobservable quantities.

- Bayesian methods allow the practitioner to incorporate any prior information regarding the probable values of a parameter into the probability calculus in a formal way. This powerful instrument is not available when using frequentist methods. Significantly, it reflects the way in which practitioners of the scientific method generate knowledge about the world in real everyday situations.

- In cases where prior input has potential to influence outcomes, and the extent to which this is appropriate is in question, the Bayesian practitioner can elect to repeat calculations using other prior distributions, including those which provide little or no information about likely parameter values. In this way the degree to which prior beliefs impact results can be readily and reliably assessed and reported.

- Bayesian analysis uses a single tool as the mechanism for calculating quantities of interest. Bayes theorem provides a consistent systematic approach to statistical inference. This contrasts with the frequentist approach which uses a variety of methods depending on the context of the problem at hand, providing confusion to the novice user and making the tool set more difficult to master.

- Bayesian analysis allows direct interpretation of probability statements regarding parameters of interest. This is a compelling reason for the use of Bayes theorem, as casual users of statistics will interpret confidence intervals in this way regardless of the mechanism used to generate them.

Having established the credentials of Bayesian methodology, we now turn to examine their computational implementation.

Chapter 3

Bayesian Computation

3.1 Introduction

The choice of computational method for Bayesian analysis is largely dependent upon the form of the posterior distribution. When the posterior is of known standard form sampling may be conducted directly by generating random deviates from the appropriate density. Such cases frequently arise from the adoption of conjugate priors as discussed in §2.4.4.

In higher dimensional problems it may be possible to partition the parameter vector, for example as $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\phi})$, and factorise the posterior into manageable subcomponents

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \;=\; p(\boldsymbol{\gamma}, \boldsymbol{\phi} \mid \boldsymbol{y}) \;=\; p(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, \boldsymbol{y})\, p(\boldsymbol{\phi} \mid \boldsymbol{y}), \tag{3.1}$$

simulating each independently. However, as problems grow more complex posterior forms tend to be non-standard and more sophisticated methods for constructing samples must be employed.

In general, Bayesian methods are implemented via a Markov Chain Monte Carlo (MCMC) scheme, with the specific details of the scheme tailored to suit the nature of the target density. Samples can be drawn from arbitrary posterior distributions but the efficiency with which this may be achieved varies, depending upon the posterior form.

The present chapter aims to provide a brief sketch of MCMC methods, sufficient to establish context for the applications found in later chapters. More detailed accounts of MCMC techniques appear in numerous texts. For example, Geyer (1992a); Gilks et al. (1995a) provide a general overview of the fundamental techniques and offer many suggestions for tuning these in various contexts. Gelman et al. (2003) discuss computational aspects of Bayesian analysis integrated with an applied analytical development. The work of Robert (1995) and Robert and Casella (2004), while

not restricted solely to Bayesian analysis, covers computational issues relevant to implementing MCMC in great detail, and thereby offers guidance in developing novel applications. Congdon (2001) offers a plethora of worked examples from a variety of disciplines, all from a Bayesian MCMC perspective. Marin and Robert (2007) offer up to date practical advice on a range of applications.

3.2 Markov Chain Monte Carlo

The crux of MCMC methods is that under fairly general conditions it is possible to construct a Markov chain which converges to a target density equivalent to the posterior distribution of the model in question (see, for example, Roberts, 1996). The posterior can then be described to any desired level of accuracy by sampling from the chain for sufficiently large number of iterations and using the ergodic average

$$\mathrm{E}(f(X)) = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t) \tag{3.2}$$

to form summary statistics for any quantity of interest. Bias introduced from samples taken prior to convergence of the Markov chain is usually avoided by discarding an initial *burn in* count of iterations $t \leq m$. Laws of large numbers ensure that for sufficiently large n, sample properties will reflect those of the population from which the sample is taken. That is

$$\lim_{t\to\infty} \frac{1}{t} \sum_{t=m+1}^{n} X_t = \int_{-\infty}^{\infty} x f(x)\, dx = \mathrm{E}(X), \tag{3.3}$$

where E(.) is the expectation operator. The next two sections describe, with increasing generality, how to construct such a Markov chain.

3.3 Gibbs Sampling

The Gibbs Sampler is an MCMC scheme which relies on partitioning the posterior into subcomponents, and simulating each of these in turn, conditional on the remainder. The scheme was introduced by Geman and Geman (1984), and brought to prominence in the statistical literature by Tanner and Wong (1987) and Gelfand and Smith (1990). Casella and George (1992), Smith and Gelfand (1992) and Gelfand (2000) provide accessible entry points into the literature.

Suppose we have a posterior which is the k dimensional joint probability distribution $p(\boldsymbol{\theta}|y), \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, and interest lies in determining properties of the marginal density

$$f(\theta_1) = \int \dots \int f(\theta_1, \theta_2, \dots, \theta_k)\, \mathrm{d}\theta_2 \dots \mathrm{d}\theta_k. \tag{3.4}$$

Gibbs sampling provides a method for obtaining a sample from $f(\theta_1)$ without having to evaluate (3.4) directly. And, by (3.3), taking a sufficiently large sample implies that any desired characteristic from $f(\theta_1)$ can be determined with arbitrary precision.

Writing $\boldsymbol{\theta}_{-j}$ for the vector with the jth component deleted $(\theta_1, \theta_2, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k)$, the sample from $f(\theta_1)$ is obtained by taking successive samples from each subcomponent of the posterior $p(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{y})$, and iterating such that

$$p(\theta_j^t \mid \boldsymbol{\theta}_{-j}^{t-1}, \boldsymbol{y}) = p(\theta_j^t \mid \theta_1^t, \theta_2^t, \dots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \dots, \theta_k^{t-1}, \boldsymbol{y}), \tag{3.5}$$

where the superscript t denotes the iteration count, until the required sample size is reached. That is, each subcomponent of the posterior is updated conditionally using the most recent values from the remaining subcomponents.

Gibbs sampling is efficient, since every sample generated is known to come from the target distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, but is obviously only available for use in cases where the full conditional posterior can be determined analytically. In cases where the posterior is not available in this form a more general, albeit less efficient, sampling strategy must be employed.

3.4 Metropolis-Hastings Sampling

The Metropolis-Hastings algorithm (3.6) provides a general regime allowing samples to be drawn from an arbitrary target distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. At each iteration a candidate point θ^* is sampled from a *proposal* distribution $q(\boldsymbol{\theta}|\boldsymbol{y})$ and subjected to an acceptance test designed to admit proposed points into the sample with probability proportional to the density of $p(\boldsymbol{\theta}|\boldsymbol{y})$ at θ^*. As with the Gibbs sampler, the process is iterated until a sample of the required precision is obtained.

The Metropolis algorithm was developed by Metropolis et al. (1953) and generalized by Hastings (1970). Müller (1991) and Tierney (1994) provided articles which brought the process to the attention of the statistical mainstream. Chib and Greenberg (1995) and Gilks et al. (1995b) provide accessible introductory treatments.

The Metropolis-Hastings algorithm may be written as

$$
\begin{array}{l}
\text{Initialise } \theta^0 \\
\quad \text{Loop } \{ \\
\qquad \text{Sample } \theta^* \text{ from } q(\theta^* \mid \theta^{t-1}) \\
\qquad \text{Sample } u \text{ from } \mathcal{U}(0,1) \\
\qquad \text{If } u \leq \alpha(\theta^*, \theta^{t-1}) \\
\qquad\quad \theta^t = \theta^* \\
\qquad \text{else} \\
\qquad\quad \theta^t = \theta^{t-1} \\
\qquad \text{Increment } t \\
\quad \}
\end{array}
\tag{3.6}
$$

where $\mathcal{U}(0,1)$ is the standard unit interval Uniform distribution, superscripts denote iteration counts, and

$$\alpha(\theta^*, \theta^{t-1}) = \min\left(1, \frac{p(\theta^*)\ q(\theta^{t-1}|\theta^*)}{p(\theta^{t-1})\ q(\theta^*|\theta^{t-1})}\right) \tag{3.7}$$

defines the criteria for the acceptance test.

If we restrict our attention to the case of a symmetrical proposal distribution, as considered by Metropolis et al. (1953),

$$q(\theta^{t-1}|\theta^*) \ = \ q(\theta^*|\theta^{t-1}) \tag{3.8}$$

and (3.7) simplifies to

$$\alpha(\theta^*, \theta^{t-1}) = \min\left(1, \frac{p(\theta^*)}{p(\theta^{t-1})}\right), \tag{3.9}$$

from which the mechanics of the rejection process can be clearly understood. When a proposed candidate θ^* is closer to the mode of $p(\theta|y)$ than the current parameter value θ^{t-1}, it will always be accepted since $p(\theta^*) \geq p(\theta^{t-1})$. When the proposal candidate is less probable than the current value it is accepted with probability appropriate to ensure that samples are drawn from $p(\theta|y)$.

Of course, when $p(\theta^*) \leq p(\theta^{t-1})$ the acceptance rate is directly proportional to the ratio of the terms in this inequality, so choice of proposal distribution is critical for efficient sampling. Indeed, the efficiency of Gibbs sampler can now be plainly seen, since the proposal and target distributions are equivalent for the full conditional case, and rejection never occurs.

The additional generality of asymmetry in $q(.)$ supplied by Hastings (1970) modifies the acceptance rate using the ratio of ratios

$$\frac{p(\theta^*) \ / \ q(\theta^*|\theta^{t-1})}{p(\theta^{t-1}) \ / \ q(\theta^{t-1}|\theta^*)} \tag{3.10}$$

which normalise the numerator and denominator according to the degree of asymmetry in the proposal distribution and provide the full generality of (3.7).

The Metropolis-Hastings algorithm describes how samples from an MCMC simulation may be obtained to provide inference regarding an arbitrary posterior, assuming that the Markov chain converges to the required target distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. It can be shown, Robert and Casella (see, for example, 2004) that the stationary distribution of the Markov chain so generated *is* the required target distribution.

3.5 Diagnosing Convergence

Convergence of an MCMC chain to the target distribution is difficult to establish. There is no specific test which can be performed to indicate that convergence has been achieved, and in practice diagnosis tends to be negatively defined: one looks for signs of non-convergence and in the absence of these assumes that the chain has satisfactorily converged.

A number of diagnostics to aid the detection of convergence have been put forward, starting with Heidelberger and Welch (1981), Schruben (1982) and Heidelberger and Welch (1983). Establishing convergence metrics remained a controversial topic in the MCMC literature throughout the 1990's, with a prominent suggestion put forward by Gelman and Rubin (1992, modified, and generalised, in Brooks and Gelman (1998)), and criticised by Geyer (1992a,b). Other input to the debate was provided by Geweke (1992), Raftery and Lewis (1992, 1995), Gelman (1995), Cowles and Carlin (1996), and Kass et al. (1998), and the many references therein.

3.5.1 Gelman and Rubin's $\widehat{R}$

Gelman and Rubin (1992) propose a general approach to monitoring the convergence of MCMC output using $m > 1$ chains with overdispersed starting values. Chains are diagnosed as having converged when the influence of the initial values is no longer evident. That is, the output from the multiple chains is effectively indistinguishable. The specific diagnostic tool used for this purpose uses a comparison of the within- and between-chain variances, essentially providing an anova style statistic.

There are two estimates for the variance of the stationary distribution represented by the MCMC output, the mean of the m within-sequence variances s_i^2

$$W = \sum_{i=1}^{m} s_i^2/m \tag{3.11}$$

and the empirical variance from all chains combined

$$\hat{\sigma}^2 = \frac{(n-1)W}{n} + \frac{B}{n}, \tag{3.12}$$

where

$$B/n = \sum_{i=1}^{m} (\bar{x}_{i.} - \bar{x}_{..})/(m-1), \tag{3.13}$$

the variance between the m sequence means $\bar{x}_{i.}$. If the chains have converged, both estimates are unbiased. If not, (3.11) will be an underestimate, since the chains have not had sufficient opportunity to explore the full support of the posterior, and (3.12) will overestimate the variance, since the started values of the chain were chosen to be overdispersed.

Then, using the assumption that the posterior can be approximated by a normal distribution with estimated mean and variance, a t-distribution can be used to construct a Bayesian credible interval with mean

$$\hat{\mu} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij},$$

variance

$$\hat{V} = \hat{\sigma^2} + \frac{B}{mn},$$

and degrees of freedom

$$d = \frac{2\hat{V}^2}{\mathrm{var}(\hat{V})},$$

with $\mathrm{var}(\hat{V})$ estimated by the method of moments. Finally, the convergence diagnostic itself is the ratio of the current variance estimate $\hat{V}$ to the within-chain variance estimate W, with an adjustment factor to account for the additional variance in the t-distribution,

$$\widehat{R} = \frac{\hat{V}}{W} . \frac{d+3}{d+1}. \tag{3.14}$$

This provides an estimate of the factor by which the scale of the current distribution for x might be reduced if the chain were allowed to continue indefinitely $n \to \infty$. If

$\widehat{R}$ is substantially greater than unity, there is reason to believe that further iteration of the chain will improve inference regarding the posterior target. That is, Bayesian credible intervals based on the t-distribution have the potential to shrink by a factor of $\widehat{R}$. Brooks and Gelman (1998) updated the original diagnostic of Gelman and Rubin (1992) to the form indicated in (3.14), and generalized this for use with multiple parameters simultaneously.

3.5.2 Discussion

Despite the reassurances implicit in diagnostic statistics such as Gelman and Rubin's $\widehat{R}$, a simple quantitative diagnostic to detect Markov chain convergence proves to be an elusive problem. Cowles and Carlin (1996), for example, surveyed thirteen convergence diagnostics and found that each of them failed to detect the type of non-convergence they were designed to identify in two simple models. Kass et al. (1998) discuss a range of issues related to the implementation of MCMC including the difficulties associated with convergence diagnosis, and use of the $\widehat{R}$ diagnostic presented in Gelman and Rubin (1992). The discussion raises a number of difficulties to the development of robust convergence diagnostics related to qualitative differences in chain behaviour associated with starting values, model misspecification, and uncertainties in model choice; highlighting the fact that diagnosing convergence is not a straightforward issue.

The position adopted in this thesis is one of eclecticism. In keeping with Neal from Kass et al. (1998) we elect to run a small number of chains and monitor these using Gelman and Rubin's $\widehat{R}$, and in addition visually inspect the chain trace from each simulated variable for signs of poor mixing or divergence, under the assumption that careful visual monitoring of qualitative trends is an instructive supplement to diagnostic metrics. The implementation of the updated version of Gelman and Rubin's $\widehat{R}$ (3.14) provided in the `coda` package (Plummer et al., 2009) of the `R` statistical environment was used throughout, in combination with visual inspection based on the principles outlined in the next section.

3.5.3 Mixing

In order for the sample generated from an MCMC scheme to reflect the properties of the posterior distribution, it must visit the entire support of the posterior in proportions reflecting the posterior density. The degree to which this is achieved can be assessed by inspection of the MCMC chain trace, which thus provides a useful qualitative measure of performance. As Geyer (1992a) points out, inspection of the chain trace will only identify some forms of problem behaviour, but in practice we find it a useful supplement to quantitative diagnostics. Examples of typical situations of interest are shown in Figure 3.1.

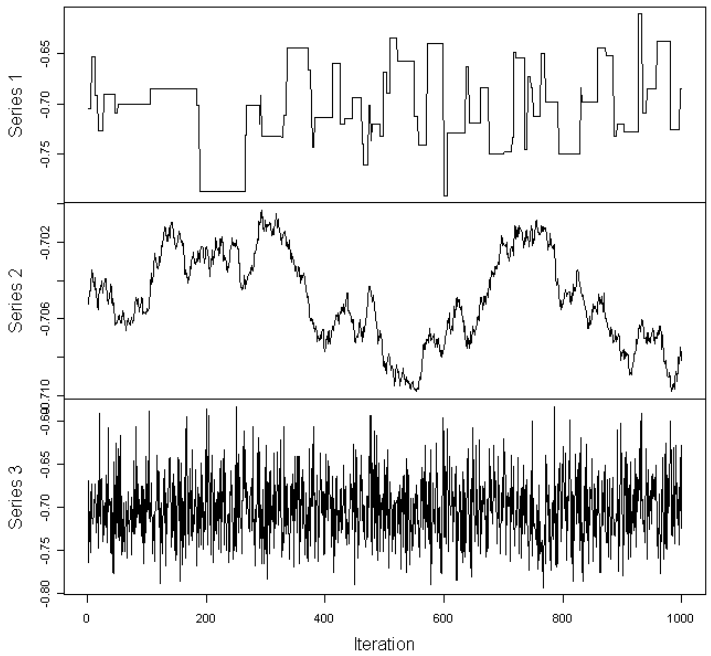


Figure 3.1: Metropolis Proposal Tuning.

Metropolis-Hastings Sampling

Figure 3.1 shows three MCMC chain traces of 1000 iterations, illustrating the impact of choice of proposal distribution upon the posterior samples obtained. The top panel shows evidence of *poor mixing*. This is characterised by the blocky appearance resulting from many candidate points being rejected between infrequent movements of the chain across the posterior support. This sample is the result of choosing a proposal distribution with variance larger than posterior target. Many points are proposed in the tails of the posterior and fail to gain acceptance under (3.7). Thus the chain retains the same value for many iterations, as indicated by the horizontal portions of the chain trace.

In principle, samples generated from a scheme displaying this behaviour would still accurately reflect the properties of the target distribution if the simulation was allowed to continue for a large enough number of iterations. However, pragmatism constrains the process. We require estimates from finite sample sizes, and because the sampling here is inefficient it is preferable to generate samples using a more suitable proposal distribution. That is, we *tune* the proposal to the posterior.

The middle panel shows the opposite extreme. The meandering appearance of the trace results from a proposal distribution with variance much smaller than the posterior. Proposed points are frequently accepted, but only arise within a region of

the posterior which is proximal to the current value of the chain. A chain with this characteristic fails to visit the support of the posterior in appropriate proportions.

The lower panel shows 1000 samples from the same posterior with a well tuned proposal distribution. Samples are obtained from the entire support of the posterior in proportion to the posterior density.

Gibbs Sampling

In the case of Gibbs sampling, the posterior distribution is the proposal distribution, the probability of acceptance of any candidate point is 1, and the situation in the top panel of Figure 3.1 can never arise, because rejection never occurs. However, behaviour similar to that indicated by the middle panel can still arise, particularly when the simulated estimands are are highly correlated. In such cases movements in any direction other than the main axis of the posterior are relatively improbable, and so sampling schemes which update one variable at a time can become "trapped" in restricted regions of the posterior for an unacceptable number of iterations. Solutions for dealing with such contingencies will be introduced on a case wise basis as necessary throughout.

We will frequently employ more than one chain to assist in the assessment of mixing. Doing so provides a check that the regions of the support visited by both chains are approximately equal, and in satisfactory circumstances these chains will provide independent samples from the posterior.

3.6 Conclusion

This chapter has provided background material to the implementation of Markov Chain Monte Carlo sampling schemes. We saw that in cases where the posterior could be appropriately partitioned and full conditional distributions could be analytically determined, Gibbs sampling provides an efficient means of drawing samples from marginal components of the posterior. In cases where these requirements could not be met, Metropolis-Hastings sampling allows samples to be drawn from an arbitrary posterior, at the expense of some efficiency. After an initial outline of the desirability of detecting convergence, a detailed description of Gelman and Rubin's convergence diagnostic was provided. We opt for checking the value of this metric in addition to visual inspection of the MCMC chain trace. The advantage of the latter is that it allows for a qualitative assessment of the mixing characteristics of the chains. The concept of "tuning" a proposal distribution to achieve adequate mixing of the MCMC chain was also described for use in cases employing Metropolis-based sampling regimes, and a similar problem scenario identified for the Gibbs sampling case. We now turn to the application of these procedures in problems of substantive interest.

Chapter 4

Nonlinear Regression Models

4.1 Introduction

Nonlinear regression extends the general linear model by allowing the expected conditional response to take a nonlinear form. This flexibility presents a dilemma. Once arbitrary parametric functions are admissible for the model mean, many alternative *parameterisations* of the basic functional form will be available. On what basis are alternative candidates best selected?

Historically, researchers have taken a necessarily pragmatic view of this problem: parameterisations were chosen on the basis of enabling numerical methods to converge, and preference given to those parameterisations which produced approximately normal sampling distributions for the estimators, as these were required to minimise inferential bias in asymptotically justified confidence intervals.

Early accounts investigating nonlinear regression are due to Beale (1960), Hartley (1961), Marquardt (1963), Hartley and Booker (1965), Jennrich (1969), and Gallant (1975). Ratkowsky (1983), Gallant (1987), and Ross (1990) offer practical advice, Bates and Watts (1988) and Seber and Wild (2003) offer more comprehensive treatments.

This chapter presents Bayesian MCMC as an alternative to Least Squares methods for nonlinear regression. In many situations, the basic MCMC apparatus presented in the introductory chapters can be applied in a straightforward way. We re-examine the growth curve models presented in Ratkowsky (1983) to establish the applicability of the MCMC approach, and explore the details of problem cases to establish the limitations of naïve application of the technique, and provide guidance in surmounting obstacles. MCMC provides advantages when faced with difficulties of choosing between alternative parameterisations. We show that the availability of posterior samples obtained under one parameterisation allows estimation and inference regarding alternative parameterisations.

4.2 Linear Regression Models

Throughout most of the 20th century the general linear model formed the mainstay of statistical practice, and it continues to be of central importance. Indeed, many of the modelling frameworks which have come to fruition in recent decades are extensions to the general linear form, arising from relaxing or generalizing one or more of the assumptions upon which the general linear model rests

1. The mean response is a linear function of the predictors.
2. Model residuals are conditionally independent.
3. Model residuals are distributed with conditional mean zero.
4. Model residuals have constant conditional variance.
5. Model residuals are conditionally normal in distribution.

The criteria 2–5 are often represented more compactly as

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{4.1}$$

where ε_i is the ith model residual, the $\sim$ symbol is read "distributed as", *iid* stands for identically and independently distributed, and $\mathcal{N}(0, \sigma^2)$ is the normal distribution with mean zero and standard deviation σ.

It is common to suppress the conditionality of the criteria when writing the model, so that we encounter the linear model as

$$y_i = \beta_0 + \sum \beta_j x_j + \varepsilon_i \tag{4.2}$$

where $y_i, i = 1, \ldots, n$, is the ith observation, x_j, $j = 1, \ldots, p-1$,is the jth covariate, $\beta_k, k = 0, \ldots, p-1$ is the kth parameter to be estimated, and the model residuals ε_i meet the criteria specified in (4.1). Model fitting focuses upon estimation of, and inference regarding, the parameters $\boldsymbol{\beta}$.

The residual sum of squares (RSS) is a measure of model fit, and represents the variability in the data which remains unexplained by the model. Specifically, RSS is the squared sum of deviations from the model mean

$$RSS = S(\boldsymbol{\beta}) \; = \; \sum_{i=1}^{n} \varepsilon_i^2 \; = \; \sum_{i=1}^{n} [Y_i - (\beta_0 + \sum \beta_j x_j)]^2, \tag{4.3}$$

where $\boldsymbol{\beta}$ is chosen so as to minimise (4.3).

Alternatively, we might choose to view (4.2) in terms of the likelihood that parameters take on particular values. The criteria (4.1) determine the likelihood function for (4.2)

$$l(\boldsymbol{\beta}, \sigma) \propto \sigma^{-n} e^{-S(\beta)/2\sigma^2}, \tag{4.4}$$

where the model parameters $\boldsymbol{\beta}$ are chosen to maximise the likelihood. From (4.3) and (4.4) we see that both are functions of the model parameters $\boldsymbol{\beta}$, and that minimising the residual sum of squares (4.3) is equivalent to maximising the likelihood function (4.4).

Least Squares estimates for $\boldsymbol{\beta}$ are obtained by setting partial derivatives of (4.3) equal to zero with respect to each $\boldsymbol{\beta}_k$, $k = 0, 1, \ldots, p-1$,

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = 0 \tag{4.5}$$

and solving the resultant *normal equations* for the respective $\boldsymbol{\beta}_k$.

The key feature of (4.2) is that the terms involving $\boldsymbol{\beta}$ are additive: the model is linear in its parameters. This in turn implies that the solutions to the normal equations are linear combinations of the observations. Writing the general solution in matrix notation

$$\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{4.6}$$

one can plainly see that the parameter estimates $\hat{\boldsymbol{\beta}}$ rely only upon $\boldsymbol{X}$ and $\boldsymbol{Y}$.

4.3 Nonlinear Regression Models

Nonlinear regression allows for criterion one from §4.2 to be relaxed so that the model mean is no longer required to be a linear function of the covariates. To emphasise this difference, it is customary to adopt a different notation, and we commonly see nonlinear models written as

$$y_u = f(\boldsymbol{\xi}; \boldsymbol{\theta}) + \varepsilon_u \tag{4.7}$$

where $y_u, u = 1, \ldots, n$, is the uth observation, $\boldsymbol{\xi}$ is a vector of covariate values, $f(\boldsymbol{\xi}; \boldsymbol{\theta})$ is an arbitrary function of the covariates parameterised by the vector $\boldsymbol{\theta}$, and the model residuals ε_u meet the criteria specified in (4.1). Model fitting focuses upon estimation of, and inference regarding, the model parameters $\boldsymbol{\theta}$.

As previously, (4.1) ensures equivalence between Least Squares and Maximum Likelihood estimates of $\boldsymbol{\theta}$. However, in the nonlinear case the Least Squares estimate $\hat{\boldsymbol{\theta}}$ requires minimisation of the residual sum of squares

$$S(\boldsymbol{\theta}) = \sum_{u=1}^{n} \{y_u - f(\boldsymbol{\xi}, \boldsymbol{\theta})\}^2, \tag{4.8}$$

and here the normal equations take the form

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_k} = \sum_{u=1}^{n} \{y_u - f(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})\} \left[\frac{\partial f(\boldsymbol{\xi}, \boldsymbol{\theta})}{\partial \theta_k} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0. \tag{4.9}$$

Unlike the linear case, (4.9) is a function of the model parameters $\boldsymbol{\theta}$. Thus, when the model is nonlinear in $\boldsymbol{\theta}$, the normal equations are also nonlinear in $\boldsymbol{\theta}$. This renders them much more difficult to solve.

Least Squares estimation of model parameters in nonlinear regression often relies on iterative numerical techniques, such as the Gauss-Newton (Hartley, 1961; Draper and Smith, 1998) or Newton-Raphson methods (Marquardt, 1963; Chambers, 1973; Ratkowsky and Dolby, 1975). These techniques estimate $\boldsymbol{\theta}$ by repeatedly solving linearized forms of $f(.)$ in restricted local regions around the current estimated values $\hat{\boldsymbol{\theta}}$. It is often necessary to provide routines with starting values approximating the final estimates to enable convergence, and determining these is something of an art.

Moreover, the estimators obtained using these procedures are known to be biased, with the extent of that bias determined by what Bates and Watts (1980, 1988) (following Beale, 1960) describe as the "intrinsic nonlinearity" of the model – data combination. Additional bias may be introduced by the choice of parameterisation of $f(.)$ in (4.7) (Bates and Watts, 1981; Cook and Witmer, 1985; Cook and Goldberg, 1986). Further, confidence intervals for these estimators rely on asymptotic assumptions of normality (Clarke, 1987; Chen and Jennrich, 1995), which may only be reasonably approximated by sample sizes which are beyond those typically available to researchers in biology, agriculture, and other applied fields. In summary, nonlinear regression poses considerable challenges to the non-specialist. Indeed, Ratkowsky (1983) provides a book length treatment on how to achieve reasonable results.

4.4 Nonlinear Regression using MCMC

Given the established difficulties of nonlinear regression, and the promise of a general purpose modelling framework such as MCMC, it seems natural to consider how we might apply MCMC to nonlinear parameter estimation.

We developed a basic MCMC procedure based on the Metropolis-Hastings algorithm (3.6), with a univariate normal proposal distribution for each parameter. A very simple adaptive mechanism was used to tune the proposal to the posterior as described in the next section. We begin by considering a simple example to establish

the basic use of the method.

4.4.1 Example: Biochemical Oxygen Demand

Biochemical oxygen demand (BOD) is used as a measure of environmental pollution caused by anthropogenic wastes. Typically a small quantity of the waste material is mixed with pure water and sealed in a container which is incubated at a fixed temperature for a small number of days. The reduction of the dissolved oxygen in the water allows calculation of the BOD, in units of mg/l, during the incubation period.

#	Day	BOD
1	1	0.47
2	2	0.74
3	3	1.17
4	4	1.42
5	5	1.60
6	7	1.84
7	9	2.19
8	11	2.17

Table 4.1: Biochemical Oxygen Demand Data
Source: Bates and Watts (1988)

Bates and Watts (1988) consider the BOD data provided in Table 4.1 and fit the function

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \alpha(1 - e^{-\beta\boldsymbol{x}}), \tag{4.10}$$

where f is predicted biochemical oxygen demand, and x is time. We demonstrate an MCMC parameter estimation procedure by fitting (4.10) to these data.

Initialisation

Figure 4.1 shows the BOD data with a curve depicting the fit of the initial parameter values passed to the MCMC procedure: the mean BOD value $\alpha_0 = 1.45$, and a rate parameter $\beta_0 = 1$. Obviously these starting values provide a poor fit to the data. A starting value for the precision is also required, here $\tau_0 = 4$.

Operation

The MCMC method consists of two phases: an adaptive phase (see, for example, Gilks et al., 1994; Gilks and Roberts, 1995; Fearnhead, 2008) and a sampling phase.

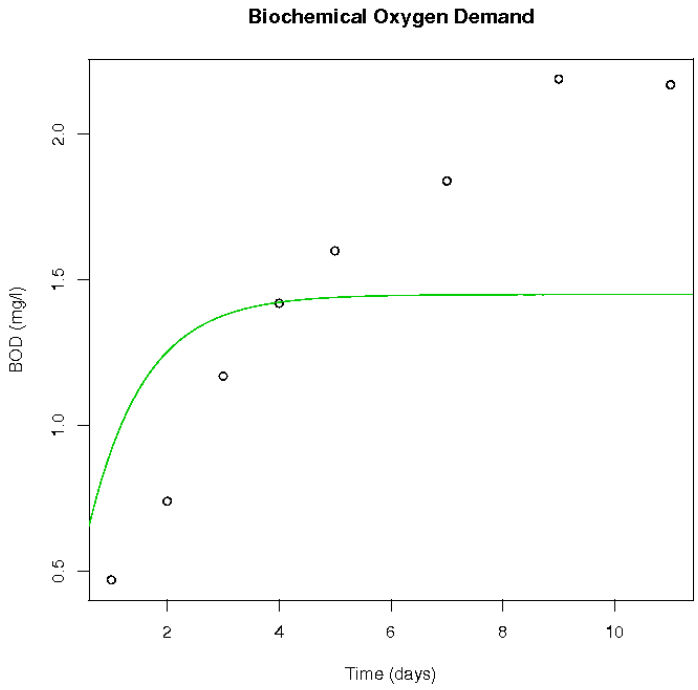


Figure 4.1: BOD Data with Initial Parameter Value Fit.

Values of arguments supplied to the function for this example are given in parentheses, and are the default values used throughout.

The adaptive phase employs a simple procedure to tune the proposal distribution to the posterior. First the chain is initialised using the starting values $\boldsymbol{\theta}_0 = (\alpha_0, \beta_0, \tau_0)$. The covariance of the Markov chain is calculated after k_1 (1000) iterations and used to update the proposal distribution. This process is repeated r (5) times during the adaptive phase. The simplicity of this approach avoids the potential for sampling biases which can be introduced via more elaborate adaptive mechanisms (see, for example, Gelfand and Sahu, 1994; Robert and Casella, 2004). Gelman et al. (1996) offer advice on improvements for mixing but this thread is not taken up here.

The sampling phase uses the final parameter estimates from the adaptive phase as starting values, and generates a sample of length k_2 (5000) from the posterior distribution. We routinely monitor Gelman and Rubin's $\widehat{R}$ as a check against failure to converge. A number of chains c (2) are run as an additional check against non-convergence, as described in §3.5.3. Assuming no evidence of malfunction, ck_2 (10000) samples from the posterior are obtained during the sampling phase, from which arbitrary summary statistics can be calculated, as described in §2.5.

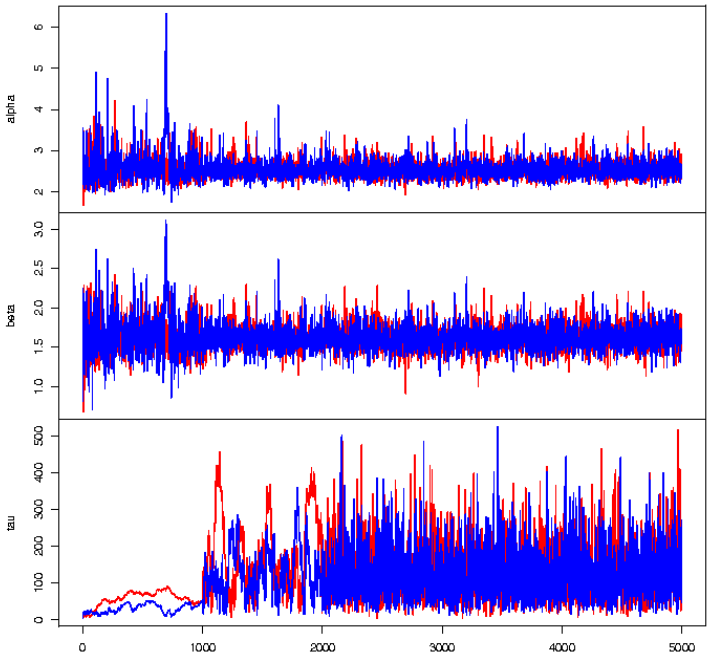


Figure 4.2: Adaptive MCMC Trace: BOD Model.

Diagnostics

During either phase of the MCMC procedure diagnostic plots can be used to visually check for signs of non-convergence. Figures 4.2 and 4.3 show the chain traces for the adaptive and sampling phases of the BOD example.

The first 1000 iterations of Figure 4.2 show obvious symptoms of instability, most clearly indicated by the divergent chain traces for τ. The chains for both α and β also show maximum variability during this portion of the trace. At iteration 1001, the chains have been restarted with an updated proposal distribution using the covariance of the previous 1000 iterations. This improves the efficiency of acceptance, and results in reduced variability in the traces for α and β, and better mixing for the precision parameter τ. A second restart at iteration 2001, with the proposal using the covariance of the values generated in iterations $[1001, 2000]$, appears sufficient to have stabilised the sampling process, none of the parameters indicate problematic symptoms past this point.

Figure 4.3 shows the sample trace obtained by running the MCMC routine using a proposal based on the covariance estimate from the final 1000 iterations of the adaptive phase as starting values. Again, no symptoms of non-convergence are evident, and 10,000 posterior samples have been obtained for each parameter. Note also the change in scale for parameters α and β between this and the previous figure

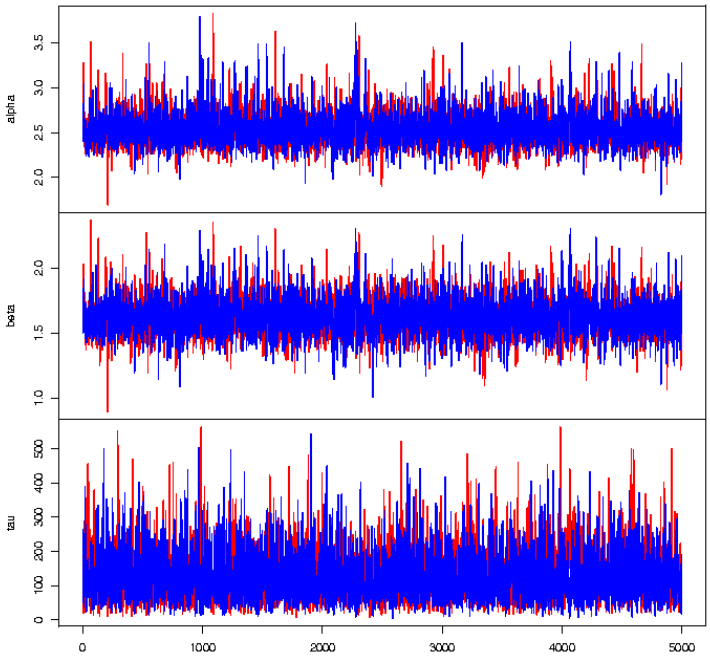


Figure 4.3: MCMC Posterior Sample Trace: BOD Model.

– use of the adaption strategy has allowed the procedure to focus precisely on the posterior estimates.

At the conclusion of the sampling phase Gelman and Rubin's $\widehat{R}$ estimates of the potential scale reduction factor for α, β and τ were 1.04, 1.01 and 1.01 respectively. Further reporting of these values will be suppressed unless particular problems warrant their inclusion in the discussion.

Results

Table 4.2 provides the MCMC parameter estimates for the BOD model. Values in the first column were generated using equation (4.10) and the Nonlinear Least Squares (NLS) routine from the `R` package `nlme` (Pinheiro et al., 2008) to provide a (Gauss-Newton) least squares comparison. Note that the residual variance provided by NLS has been converted to a measure of precision, using $\hat{\tau} = \frac{1}{\hat{\sigma}^2}$. The practice of reporting precision rather than variance will be maintained throughout.

The fit of the model to the data is provided as Figure 4.4. The 95% estimates shown are the curves associated with the 2.5% and 97.5% quantiles listed in Table 4.2.

		MCMC		*Quantiles*		
	NLS	*Mean*	σ	50%	2.5%	97.5%
$\hat{\alpha}$	2.4979	2.5318	0.1861	2.5129	2.2200	2.9544
$\hat{\beta}$	1.5972	1.6158	0.1437	1.6097	1.3468	1.9177
$\hat{\tau}$	228.6265	129.3753	73.6373	114.9239	27.5127	309.5821

Table 4.2: Summary Statistics: BOD Data Model.

Discussion

One of the key advantages of using MCMC to simulate the posterior distribution is that the posterior samples remain available for subsequent use. As discussed in §2.5.2, inference in the Bayesian context is simply a matter of summarising posterior quantities of interest. The summary statistics seen in Table 4.2, for example, were generated directly from the posterior samples. However, the availability of these samples also provides the ability to visualise the sampling distributions of the model parameters.

Figure 4.5 shows the pairwise marginal scatterplots of the posterior distribution for the BOD model. From these plots we can see that α and β are highly correlated, that α shows greater positive skew than β and that the joint posterior distribution of these parameters exhibits curvature, a feature that shall become important later in the chapter. The distinctive conical shape associated with τ in these diagrams indicates that high values of precision (low variance) are correlated with the posterior mode.

Having established the utility of the MCMC method, we now turn to consider its performance relative to Least Squares in the context of growth curve models.

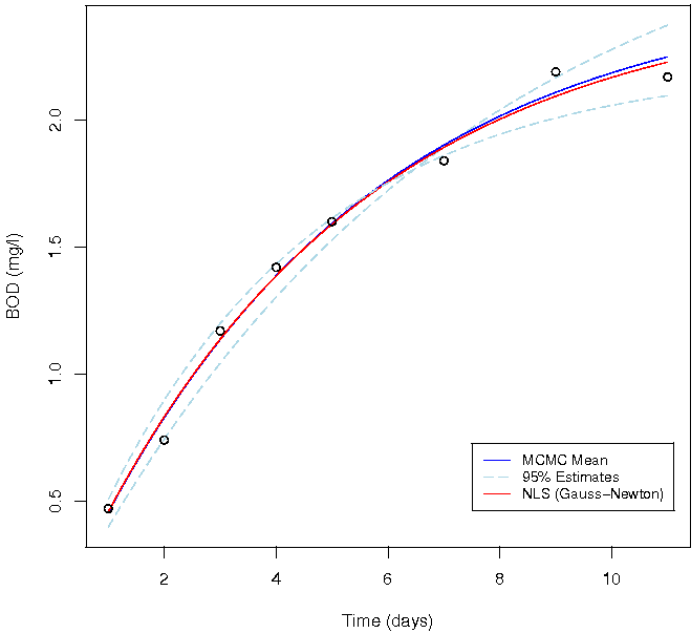


Figure 4.4: Fitted BOD Model.

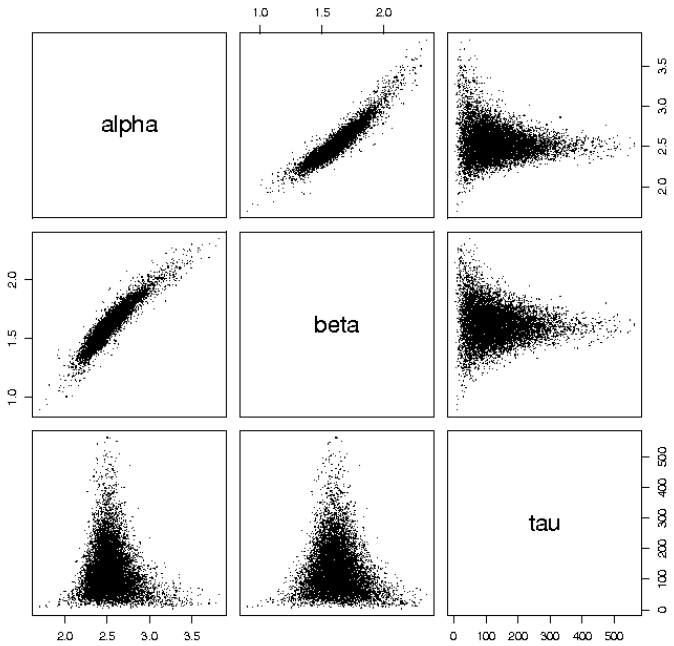


Figure 4.5: Pairwise Marginal Scatterplots: BOD Model.

4.5 Growth Curve Models using MCMC

4.5.1 Ratkowsky's Regression Strategy

Ratkowsky (1983) developed a strategy for nonlinear regression based on the distinction made by Bates and Watts (1980, 1988) between *intrinsic* nonlinearity, attributable to the functional form of the model, and *parameter effects* nonlinearity, attributable to the chosen parameterisation within that functional form. Hougaard (1982) and Kass (1984), among others, have also considered parameterisation issues.

Ratkowsky argued that model parameterisations which perform in a "close to linear" fashion are preferable. The sampling distributions of the least squares estimators approximate normality more closely in such cases, are therefore less biased, and provide a more reliable basis for the calculation of confidence intervals. A significant additional benefit is that convergence of the numerical routines employed is often facilitated when the sampling distributions of the parameters approximate normal distributions.

These arguments are appealing in the context of Least Squares, where approximately normal sampling distributions are required for the asymptotic theory of inference to hold. However, MCMC allows this requirement to be relaxed. The credible intervals introduced in §2.5.4 can be obtained from the any posterior distribution, regardless of form. Moreover, as we have seen, the availability of posterior samples allows ready calculation of any summary quantity desired. Significantly, this provides the ability to transform estimates between alternative parameterisations of a model, leaving the practitioner free to explore model parameterisations motivated by criteria other than the necessity that it may be the only mathematically tractable option.

4.5.2 Model Functions and Data

Ratkowsky (1983) considers five nonlinear growth curve types; two three-parameter cases: the Gompertz (4.11) and logistic (4.12) model functions

$$\boldsymbol{y} = \alpha \exp[-\exp(\beta - \gamma \boldsymbol{x})], \tag{4.11}$$

and

$$\boldsymbol{y} = \frac{\alpha}{1 + \exp(\beta - \gamma \boldsymbol{x})}, \tag{4.12}$$

and three four-parameter cases, the Morgan-Mercer-Flodin (MMF) (4.13), Richards (4.14), and Weibull-type (4.15) model functions

$$\boldsymbol{y} = \frac{\beta\gamma + \alpha \boldsymbol{x}^{\delta}}{\gamma + \boldsymbol{x}^{\delta}}, \tag{4.13}$$

#	$\boldsymbol{x}$	$\boldsymbol{y}$
1	0.50	1.30
2	1.50	1.30
3	2.50	1.90
4	3.50	3.40
5	4.50	5.30
6	5.50	7.10
7	6.50	10.60
8	7.50	16.00
9	8.50	16.40
10	9.50	18.30
11	10.50	20.90
12	11.50	20.50
13	12.50	21.30
14	13.50	21.20
15	14.50	20.90

Table 4.3: Bean Data
Source: Heyes and Brown (1956), cited in Ratkowsky (1983)

$$\boldsymbol{y} = \frac{\alpha}{[1 + \exp(\beta - \gamma \boldsymbol{x})]^{\frac{1}{\delta}}}, \tag{4.14}$$

and

$$\boldsymbol{y} = \alpha - \beta \exp(-\gamma \boldsymbol{x}^{\delta}), \tag{4.15}$$

where $\boldsymbol{y}$ is a vector of responses, $\boldsymbol{x}$ is a vector of covariate data, and $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$, or $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta)$, is a vector of parameters to be estimated.

#	$\boldsymbol{x}$	$\boldsymbol{y}$
1	0.00	1.23
2	1.00	1.52
3	2.00	2.95
4	3.00	4.34
5	4.00	5.26
6	5.00	5.84
7	6.00	6.21
8	8.00	6.50
9	10.00	6.83

Table 4.4: Cucumber Data
Source: Gregory (1956) cited in Ratkowsky (1983)

#	x	y
1	1.00	16.08
2	2.00	33.83
3	3.00	65.80
4	4.00	97.20
5	5.00	191.55
6	6.00	326.20
7	7.00	386.87
8	8.00	520.53
9	9.00	590.03
10	10.00	651.92
11	11.00	724.93
12	12.00	699.56
13	13.00	689.96
14	14.00	637.56
15	15.00	717.41

Table 4.5: Onion Data
Source: Gregory (1956) cited in Ratkowsky (1983)

Each of these model functions was applied to four datasets related to vegetative growth processes, providing an initial set of 20 cases of interest. These data are provided in Tables 4.3 - 4.6, and will be referred to as the Bean (Table 4.3), Cucumber (Table 4.4), Onion (Table 4.5) and Pasture (Table 4.6) data respectively.

A similar approach will be followed here. Each of the 20 initial data – model function combinations were evaluated using the MCMC procedure described in §4.4.1. The results are categorised by the model function type listed above.

#	x	y
1	9.00	8.93
2	14.00	10.80
3	21.00	18.59
4	28.00	22.33
5	42.00	39.35
6	57.00	56.11
7	63.00	61.73
8	70.00	64.62
9	79.00	67.08

Table 4.6: Pasture Data
Source: Ratkowsky (1983)

4.5.3 Three Parameter Models

Logistic & Gompertz Models

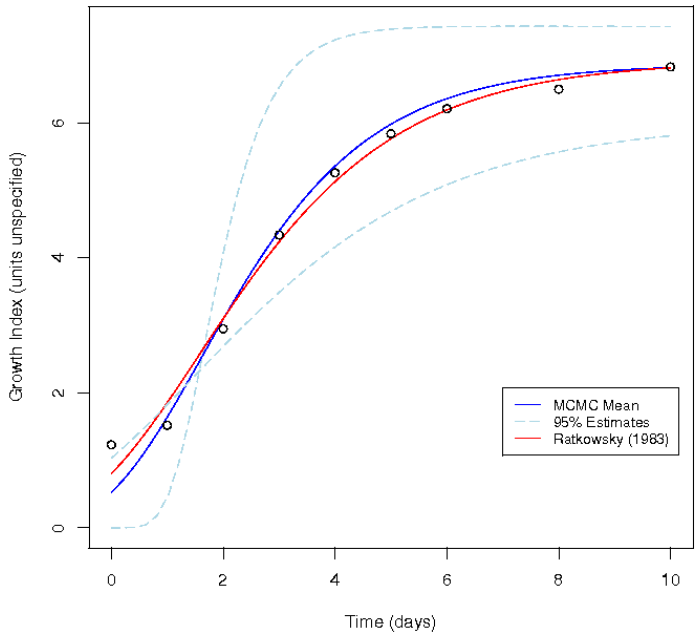


Figure 4.6: Fitted Gompertz-Cucumber Model.

MCMC parameter estimates using the Gompertz (4.11) and logistic (4.12) model functions for each data set are provided in Tables 4.7 and 4.8. The results obtained by Ratkowsky (1983) are provided in the first column for comparison, with variances converted to precisions as previously noted. In each case the parameter estimate reported by Ratkowsky (1983) falls within the 95% credible interval obtained using the MCMC procedure. These tables reaffirm that the MCMC procedure is capable of producing results comparable with those obtained by nonlinear Least Squares.

Assessing Model Fit

In Tables 4.7 and 4.8 the precision estimates $\hat{\tau}$ obtained using MCMC are comparable with those reported by Ratkowsky (1983). However, using precision as the sole criterion of model fit is problematic, and it needs to be interpreted with care. For example, Tables 4.7 and 4.8 show the precision estimates for the Cucumber models to be large relative to the other datasets, across both model functions and both estimation methods. By contrast, the precision estimates for the Onion data suggest that both methods of estimation produce models which fail to fit the data well. Visualising the fit of the models corresponding to the Gompertz cases in Figures

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	22.5100	22.4333	0.9108	22.3869	20.7853	24.4382
$\hat{\beta}$	2.1060	2.2127	0.3635	2.1646	1.6675	3.0485
$\hat{\gamma}$	0.3880	0.4067	0.0644	0.3996	0.3050	0.5582
$\hat{\tau}$	0.9533	0.9358	0.3910	0.8840	0.3321	1.8462
Cucumber Data:						
$\hat{\alpha}$	6.9250	6.8711	0.3426	6.8948	6.0185	7.4349
$\hat{\beta}$	0.7680	0.9432	0.9230	0.7861	0.5672	2.5750
$\hat{\gamma}$	0.4930	0.5839	0.4445	0.5050	0.3911	1.5406
$\hat{\tau}$	16.1551	15.8105	10.1483	13.8672	1.1566	40.6648
Onion Data:						
$\hat{\alpha}$	723.1000	720.8077	23.9323	719.8647	676.1035	770.6005
$\hat{\beta}$	2.5000	2.5910	0.3574	2.5577	1.9896	3.4062
$\hat{\gamma}$	0.4500	0.4663	0.0628	0.4608	0.3609	0.6055
$\hat{\tau}$	0.0009	0.0009	0.0004	0.0008	0.0003	0.0017
Pasture Data:						
$\hat{\alpha}$	82.8300	83.6054	6.3747	82.9106	72.5160	97.6067
$\hat{\beta}$	1.2240	1.2341	0.0942	1.2244	1.0769	1.4427
$\hat{\gamma}$	0.0370	0.0373	0.0050	0.0371	0.0288	0.0482
$\hat{\tau}$	0.2755	0.2863	0.1609	0.2563	0.0633	0.6907

Table 4.7: Summary Statistics: Gompertz Models.

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	21.5100	21.4892	0.4514	21.4892	20.6151	22.4068
$\hat{\beta}$	3.9570	4.0229	0.3160	3.9984	3.4709	4.7196
$\hat{\gamma}$	0.6220	0.6333	0.0534	0.6297	0.5402	0.7511
$\hat{\tau}$	1.9305	1.9191	0.7955	1.8044	0.6992	3.7551
Cucumber Data:						
$\hat{\alpha}$	6.6870	6.6747	0.1804	6.6727	6.3374	7.0305
$\hat{\beta}$	1.7450	1.7731	0.1738	1.7625	1.4579	2.1589
$\hat{\gamma}$	0.7550	0.7706	0.0796	0.7650	0.6303	0.9447
$\hat{\tau}$	28.3286	25.7820	14.7332	23.0487	5.2821	62.0883
Onion Data:						
$\hat{\alpha}$	702.9000	701.9848	15.0409	701.6988	671.9047	732.7782
$\hat{\beta}$	4.4430	4.5246	0.3959	4.4986	3.8082	5.3896
$\hat{\gamma}$	0.6890	0.7023	0.0636	0.6990	0.5889	0.8411
$\hat{\tau}$	0.0013	0.0013	0.0005	0.0013	0.0005	0.0026
Pasture Data:						
$\hat{\alpha}$	72.4600	72.6086	2.2411	72.4511	68.7950	77.1946
$\hat{\beta}$	2.6180	2.6270	0.1072	2.6235	2.4255	2.8501
$\hat{\gamma}$	0.0670	0.0675	0.0041	0.0675	0.0595	0.0759
$\hat{\tau}$	0.7463	0.7335	0.4181	0.6560	0.1512	1.7412

Table 4.8: Summary Statistics: Logistic Models.

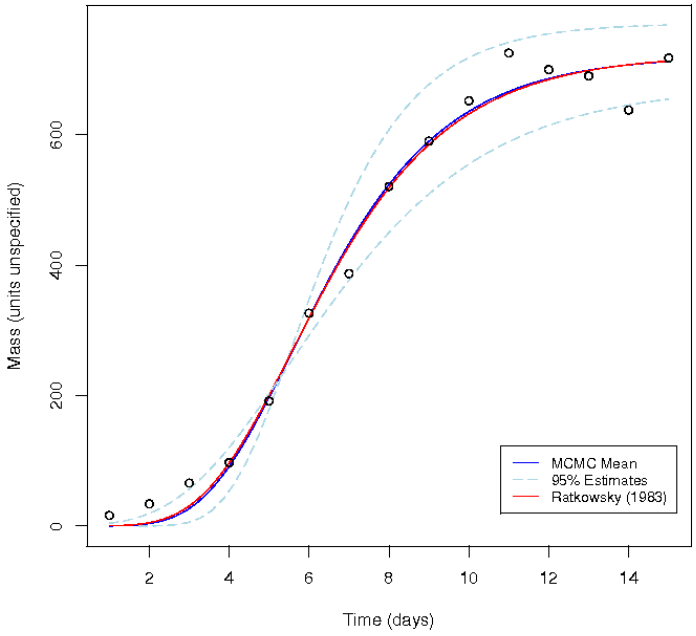


Figure 4.7: Fitted Gompertz-Onion Model.

4.6 and 4.7 reveal this interpretation to be flawed. The Onion models appear to fit better than the Cucumber models.

Faced with this apparent contradiction it is tempting to think that posterior variability should provide a useful guide. After all, the wide 95% credible interval associated with $\hat{\tau}$ in the Gompertz – Cucumber case suggests that we might expect the model to fit less well. However, the Logistic – Cucumber model also reports a wide credible interval for $\hat{\tau}$, and inspection of the fitted model for that case (Figure 4.8) reveals no suggestion of ill fit. These examples highlight the fact that it is inadvisable to make comparisons between nonlinear models on the basis of simple summary statistics. As Seber and Wild (2003) point out, the relative magnitudes of residual variances vary with the model – data combination on a casewise basis in nonlinear regression.

Diagnostics

The previous section revealed the fit of the Gompertz – Cucumber model (Figure 4.6) to be unsatisfactory. Because the MCMC posterior samples are available, diagnosis of the problem is straightforward.

Scatterplots of the pairwise marginal posterior slices for the Gompertz – Cucumber model parameters are provided in Figure 4.9. High values of precision are asso-

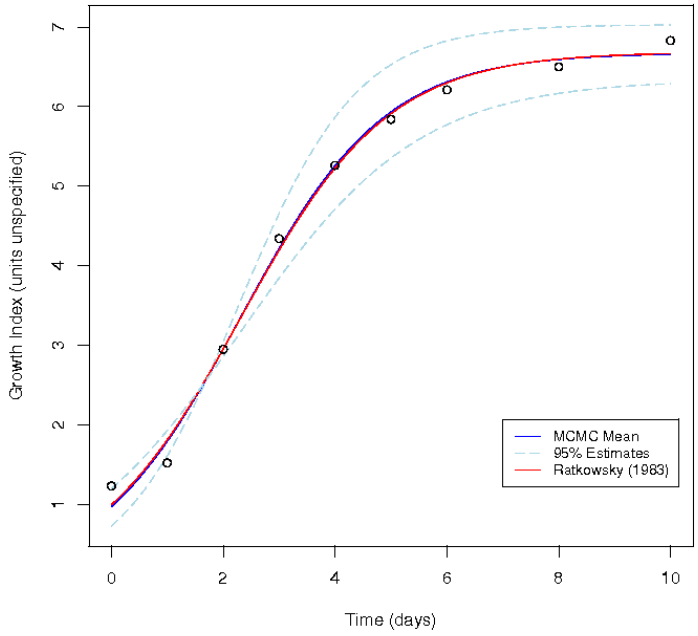


Figure 4.8: Fitted Logistic-Cucumber Model.

ciated with the mode of each parameter estimate, as expected. But there is also a large low density region distant from the mode corresponding to very low precision values. Inspection of the MCMC sample trace provides further insight. Figure 4.10 reveals a major excursion away from the main density of the posterior by one chain around iteration 3000. This is correlated with a period of low precision and, along with some later minor excursions, is responsible for the large low density areas observed in the pairwise marginal plots. Armed with these diagnostic aids, we may choose to re-run the procedure to produce a more satisfactory posterior sample.

However, because these excursions are minor aberrations among 10000 samples, we might expect the existing median estimates to be robust to their influence. The model fit using the MCMC median values from Table 4.7 is provided in Figure 4.11, and attests the adequacy of these values.

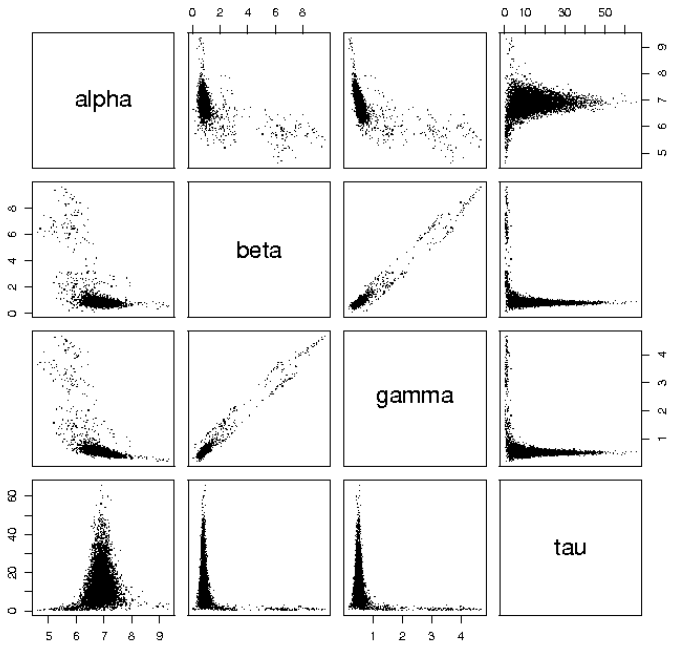


Figure 4.9: Pairwise Marginal Scatterplots: Gompertz-Cucumber Model.

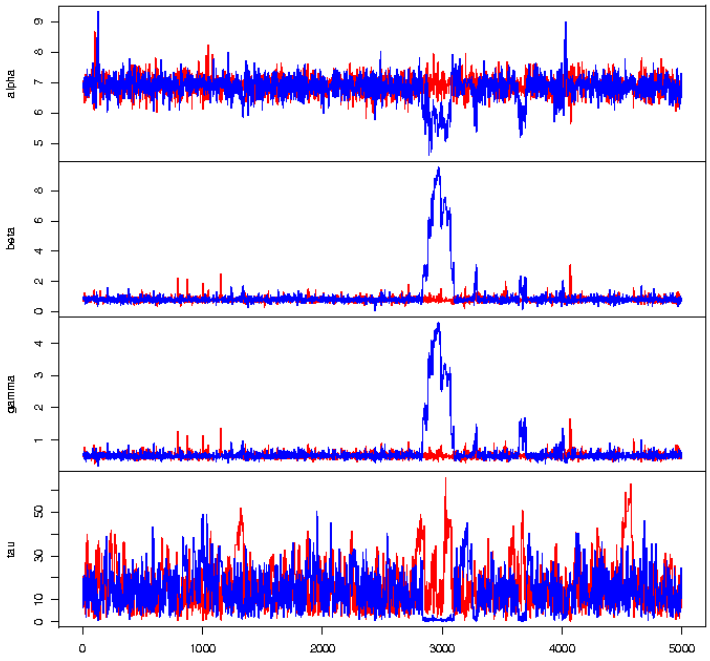


Figure 4.10: MCMC Posterior Sample Trace: Gompertz-Cucumber Model.

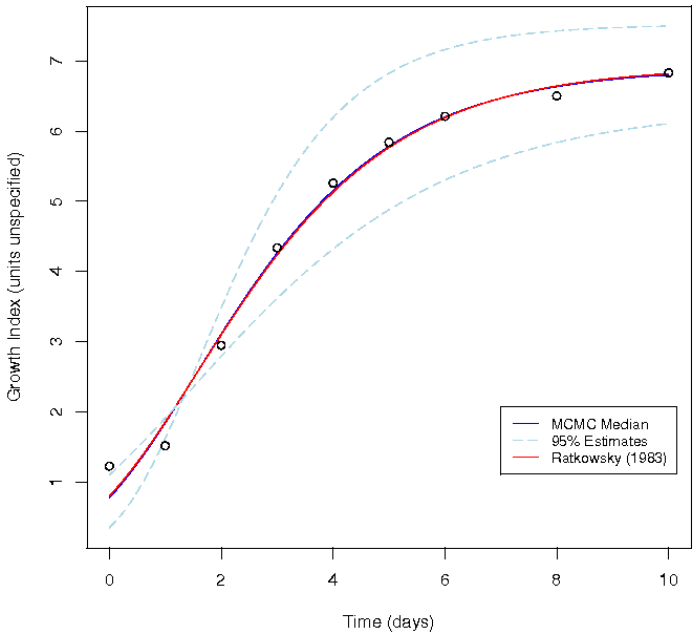


Figure 4.11: (Median) Fitted Gompertz-Cucumber Model.

4.5.4 Four Parameter Models

Estimates from the four parameter models (4.13) – (4.15) are provided in Tables 4.9, 4.10 and 4.11. The performance of each model function will be evaluated in turn. In the interests of brevity exhaustive details of all 20 cases will be omitted, other than reporting the summary statistics in the tables listed above. Instead, we focus on cases where the MCMC method did not perform well, with a view to diagnosis.

Morgan-Mercer-Flodin Models

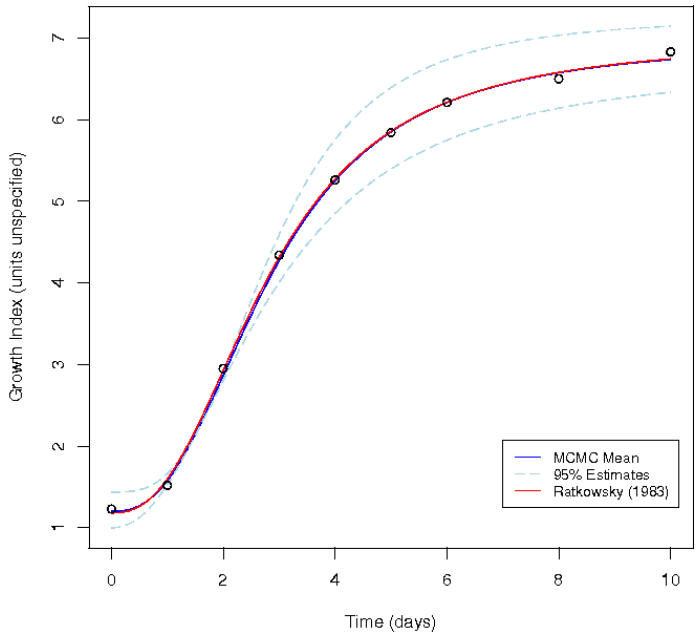


Figure 4.12: Fitted MMF-Cucumber Model.

The estimates for the MMF model function are provided in Table 4.9. Direct comparison of the estimates reported in the first two columns may be misleading. The mean MCMC estimates for many of the parameters appear to approximate their NLS counterparts only roughly, and $\hat{\gamma}$ is frequently out by a factor of two. However, visualising the fit of the models against the data reveals that the differences are not as severe as might be expected based on the comparative estimates. The Cucumber models (Figure 4.12), for example, are in excellent agreement despite the large reported difference in values for $\hat{\tau}$. By contrast, the fitted models for the Onion data report identical precisions and are shown in Figure 4.13. The MCMC fit appears less satisfying than that based on the Least Squares estimates. We shall address this shortcoming in §4.5.5.

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	22.0800	21.8398	0.5025	21.8219	20.9195	22.9150
$\hat{\beta}$	1.6530	1.8032	0.4111	1.8086	0.9670	2.6094
$\hat{\gamma}$	5586.0000	12214.0748	5005.5759	12470.5180	2704.3017	20495.2275
$\hat{\delta}$	4.5600	4.9107	0.2973	4.9832	4.1717	5.2934
$\hat{\tau}$	1.7271	1.8047	0.7447	1.6888	0.6777	3.5440
Cucumber Data:						
$\hat{\alpha}$	6.9860	6.9620	0.1408	6.9637	6.6778	7.2636
$\hat{\beta}$	1.1810	1.2016	0.1051	1.1974	0.9996	1.4356
$\hat{\gamma}$	12.9600	14.3788	4.1638	13.5018	9.7855	24.9164
$\hat{\delta}$	2.4750	2.5443	0.2193	2.5152	2.1828	3.0785
$\hat{\tau}$	208.3333	109.9113	65.7183	96.8482	18.4222	257.4667
Onion Data:						
$\hat{\alpha}$	723.9000	736.8373	23.2472	735.3257	695.4091	785.6300
$\hat{\beta}$	33.3500	28.3925	16.1151	27.1367	2.3796	63.6523
$\hat{\gamma}$	6266.0000	3472.0489	2189.9940	2564.8439	1011.2856	8507.0643
$\hat{\delta}$	4.6410	4.2297	0.3313	4.1789	3.6286	4.8351
$\hat{\tau}$	0.0010	0.0010	0.0004	0.0010	0.0004	0.0019
Pasture Data:						
$\hat{\alpha}$	80.9600	77.8206	5.3891	76.9990	70.3627	89.0825
$\hat{\beta}$	8.8950	9.9395	1.5540	9.9826	6.6335	12.8589
$\hat{\gamma}$	49577.0000	255654.2848	162767.8104	242685.0136	19253.3617	586740.4717
$\hat{\delta}$	2.8280	3.1966	0.2689	3.2682	2.5313	3.5450
$\hat{\tau}$	0.3690	0.3711	0.2277	0.3248	0.0658	0.9157

Table 4.9: Summary Statistics: Morgan-Mercer-Flodin Models.

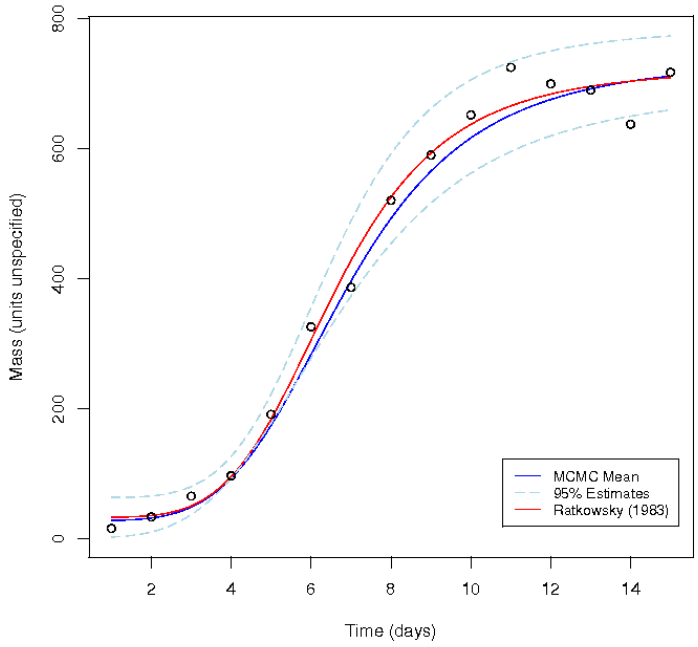


Figure 4.13: Fitted MMF-Onion Model.

In the meantime, we note that the MMF model function provides a better fit to the Onion data than that provided by the Gompertz model function in §4.5.3. In particular, the additional parameter has allowed for the tail at the lower extent of the data to be more adequately incorporated into the model. This supports the idea that some model functions are simply better suited to the nuances of particular data than others.

Richards Models

The Richards model estimates (Table 4.10) are clearly less reliable than their MMF counterparts. Other than the asymptote parameter $\hat{\alpha}$ none of the Least Squares estimates fall within the MCMC 95% credible intervals, despite some of these spanning several orders of magnitude. Evidently the MCMC method has failed. Figure 4.14 shows the model fit for the Cucumber data, and is representative of the (omitted) remainder.

Pairwise marginal scatterplots of parameter estimates for the Cucumber data are provided in Figure 4.15. The coral-like structures in the β, γ and δ plots indicate that the MCMC chain was not successful in exploring the entire support of the posterior. Reasonable results cannot be expected under such circumstances. This is a form of the poor mixing behaviour that was discussed in §3.5.3, and requires

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	21.2000	20.3238	0.7187	20.3202	18.9200	21.7626
$\hat{\beta}$	5.6910	848.7764	283.8791	851.2890	303.6976	1390.5505
$\hat{\gamma}$	0.7770	97.7261	34.4409	96.9855	35.4533	164.5484
$\hat{\delta}$	1.6190	316.3795	126.1428	301.3518	100.8983	603.3485
$\hat{\tau}$	1.9920	0.5348	0.2371	0.5013	0.1822	1.0937
Cucumber Data:						
$\hat{\alpha}$	6.6840	6.0204	0.5829	6.1188	4.6977	7.0175
$\hat{\beta}$	1.7800	6525.0989	7312.8947	1593.2456	270.5109	22919.5780
$\hat{\gamma}$	0.7590	2253.8263	2479.6127	498.2850	61.5660	7478.4430
$\hat{\delta}$	1.0170	4195.7634	5430.6056	846.3076	158.8584	18209.8394
$\hat{\tau}$	23.5849	2.5193	2.9389	1.0434	0.1765	10.5079
Onion Data:						
$\hat{\alpha}$	699.6000	679.8403	19.6350	679.3874	642.0294	719.9728
$\hat{\beta}$	5.2770	910.1372	273.4602	944.7644	331.7866	1408.4570
$\hat{\gamma}$	0.7600	102.5406	30.9412	105.3308	37.7974	158.1207
$\hat{\delta}$	1.2790	310.2726	100.7502	315.9978	110.6082	502.8827
$\hat{\tau}$	0.0013	0.0006	0.0002	0.0005	0.0002	0.0011
Pasture Data:						
$\hat{\alpha}$	69.6200	65.0017	2.6766	64.4884	60.7064	71.1110
$\hat{\beta}$	4.2550	509.6861	286.2745	601.4125	4.4371	833.8269
$\hat{\gamma}$	0.0890	8.4686	4.7672	9.9435	0.0900	14.0599
$\hat{\delta}$	1.7240	262.6247	148.1082	308.5982	1.7786	456.3105
$\hat{\tau}$	0.8264	0.3255	0.5242	0.1193	0.0301	1.8972

Table 4.10: Summary Statistics: Richards Models.

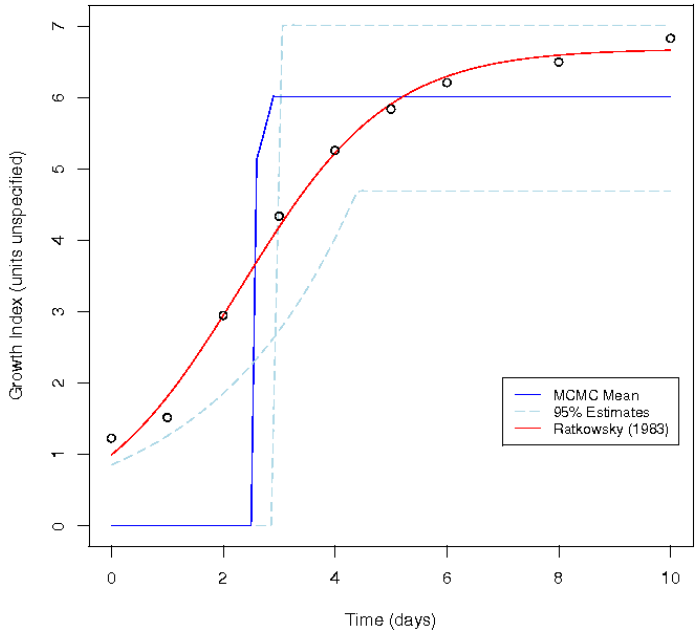


Figure 4.14: Fitted Richards-Cucumber Model.

the proposal to be "tuned" to adequately visit the support of the posterior. We will return to discuss solutions to this problem in §4.5.5, after results from the Weibull-type models have been considered.

Weibull-type Models

MCMC estimates for the Weibull-type model function (Table 4.11) appear to be in reasonable agreement with those obtained by Ratkowsky (1983). With the exception of the precision estimate $\hat{\tau}$ for the Pasture dataset, the Least Squares estimates all fall within the 95% credible intervals provided by MCMC, and these appear satisfactorily narrow.

However, visualisation of the fitted models reveals inadequacies not obvious by inspection of the summary statistics. The MCMC estimates fit the Cucumber data well, and the Bean and Onion data only approximately. Figure 4.16 shows the worst case – the predicted fit to the Pasture data – where the MCMC parameter estimates fail to even approximate the shape of the data.

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	21.1000	21.2243	0.4664	21.1883	20.3885	22.2627
$\hat{\beta}$	19.8100	20.1125	0.7440	20.0411	18.8346	21.8060
$\hat{\gamma}$	0.0018	0.0028	0.0017	0.0023	0.0008	0.0074
$\hat{\delta}$	3.1800	3.0396	0.2951	3.0412	2.4567	3.6180
$\hat{\tau}$	2.0202	1.9614	0.8831	1.8143	0.6735	4.1420
Cucumber Data:						
$\hat{\alpha}$	6.6560	6.6730	0.2035	6.6567	6.3227	7.1045
$\hat{\beta}$	5.5490	5.5764	0.3050	5.5574	5.0227	6.2507
$\hat{\gamma}$	0.1180	0.1230	0.0308	0.1201	0.0686	0.1950
$\hat{\delta}$	1.7630	1.7580	0.2130	1.7499	1.3414	2.2160
$\hat{\tau}$	37.3134	32.3002	20.5344	28.1196	5.7023	83.9186
Onion Data:						
$\hat{\alpha}$	695.0000	699.4017	15.4481	698.6912	670.6295	735.6980
$\hat{\beta}$	673.5000	687.1938	24.3655	684.5834	648.3933	760.4275
$\hat{\gamma}$	0.0015	0.0026	0.0017	0.0022	0.0008	0.0076
$\hat{\delta}$	3.2620	3.0716	0.2857	3.0845	2.4447	3.6110
$\hat{\tau}$	0.0014	0.0014	0.0006	0.0013	0.0005	0.0028
Pasture Data:						
$\hat{\alpha}$	69.9600	65.6773	5.1034	67.0003	51.4912	70.5215
$\hat{\beta}$	61.6800	64.6102	3.3045	63.5829	61.4388	74.9134
$\hat{\gamma}$	0.0001	0.0074	0.0250	0.0008	0.0001	0.0772
$\hat{\delta}$	2.3780	1.8238	0.3950	1.8922	0.8862	2.4987
$\hat{\tau}$	0.5952	0.0597	0.0414	0.0404	0.0046	0.1205

Table 4.11: Summary Statistics: Weibull-type Models.

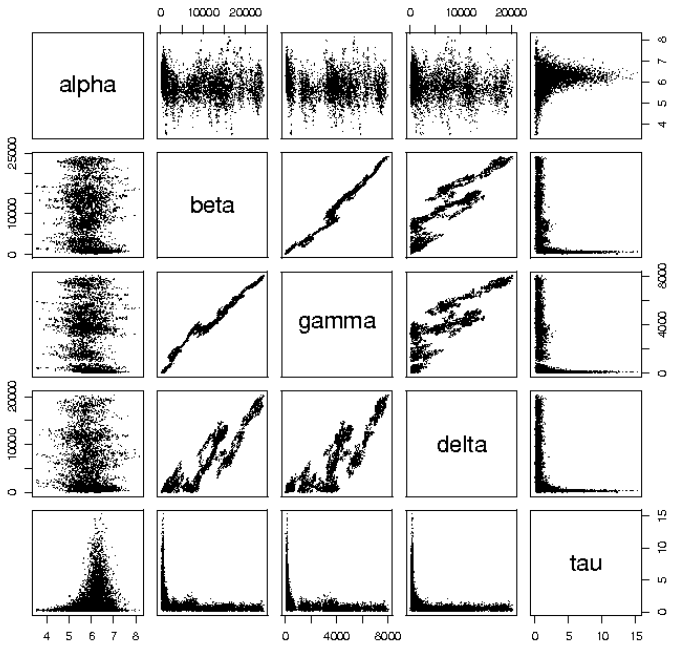


Figure 4.15: Pairwise Marginal Scatterplots: Richards-Cucumber Model.

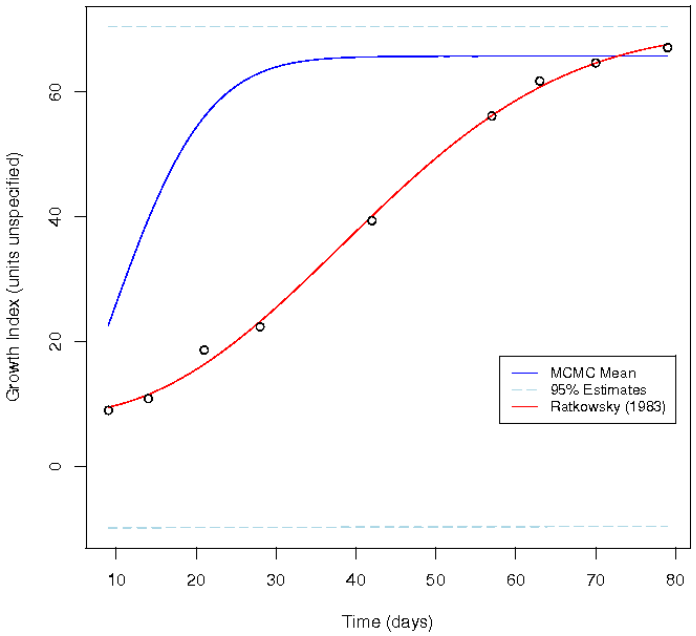


Figure 4.16: Fitted Weibull-Pasture Model.

4.5.5 Troubleshooting

The MCMC method worked well for the three parameter models but less well for some of the four parameter cases. To address these shortcomings we review the components of nonlinearity for the specific data – model combinations of concern, and use these to identify solutions.

Curvature Components

Table 4.12 provides estimates of the *intrinsic* and *parameter effects* curvature reported by Ratkowsky (1983), based on the measures of Bates and Watts (1980, 1988). Details of how these figures are calculated are omitted, but available elsewhere (Hamilton et al., 1982; Bates and Watts, 1988). For present purposes it suffices to know that the figures are a measure of the degree to which the nonlinear solution surface approximates a plane in the local region of estimation, with lower numbers indicating more nearly adequate approximations. Parameter effect curvature is a measure of transformable nonlinearity – an alternative parameterisation of the model may render the solution surface more adequately approximated by the plane in a region local to the parameter estimates. Intrinsic curvature is an inherent property of the data – model combination and is not amenable to transformation by alternative parameterisation.

	Model	*Gompertz*	*Logistic*	*MMF*	*Richards*	*Weibull*
Data	*Curvature*	(4.11)	(4.12)	(4.13)	(4.14)	(4.15)
Pasture	IN	0.090	0.073	0.180	0.267	0.130
	PE	2.324	0.644	90.970	6.679	42.675
Onion	IN	0.234	0.131	0.257	0.330	0.271
	PE	0.700	0.379	31.319	6.271	16.371
Cucumber	IN	0.121	0.118	0.103	0.332	0.188
	PE	0.633	0.351	1.154	14.811	1.878
Bean	IN	0.232	0.107	0.210	0.295	0.232
	PE	0.880	0.372	24.934	4.268	13.253

Table 4.12: Curvature Components: Intrinsic (IN) and Parameter Effects (PE). (Reproduced from Ratkowsky (1983)).

From Table 4.12 it is apparent that parameter effects nonlinearity is always larger than the intrinsic nonlinearity, often substantially so. The MMF model function (4.13) demonstrates very large parameter effects curvature for all but the Cucumber dataset. The Weibull-type function (4.15) displays a similar pattern to the MMF function, with parameter effects curvature values approximately halved. This

suggests that alternative parameterisations may offer opportunities for improved parameter estimates in these cases. By contrast, the Richards model function (4.14) shows the greatest intrinsic nonlinearity, across the entire range of data. This indicates that reparameterisation may be less successful for the Richards model function.

Also evident are the relatively low values of both curvature measures across the range of three parameter models compared to their four parameter counterparts. In summary, it seems that the MCMC method has performed well in cases which feature low curvature component values, and less well in cases which exhibit higher degrees of nonlinearity. These are precisely the cases identified by Ratkowsky (1983) as being problematic for Least Squares.

Ratkowsky (1983) argued that choosing parameterisations of the model function to minimise parameter effects curvature resulted in "close to linear" models, with approximately normal sampling distributions for the Least Squares estimators. To this end he nominated a number of alternative parameterisations of the model functions identified in Table 4.12 as featuring strong parameter effects curvature. We examine these alternative parameterisations and continue the comparative analysis using the MCMC method.

Because the focus of this chapter is on providing a comparative analysis, we continue to use uninformative priors with the alternative parameterisations of the model functions. If we were interested in evaluating informative priors, the priors would need to be adjusted to suit each alternative parameterisation considered, as suggested in §2.4.2.

Morgan-Mercer-Flodin Models

Ratkowsky (1983) suggested (4.16) as a reparameterisation for the MMF model function (4.13). Taking the exponential of the γ parameter reduces the parameter effects curvature from 31.3 to 1.2.

$$\boldsymbol{y} = \frac{\beta \exp(\gamma) + \alpha \boldsymbol{x}^{\delta}}{\exp(\gamma) + \boldsymbol{x}^{\delta}} \tag{4.16}$$

Re-running the MCMC estimation process using the reparameterised MMF model (4.16) results in the estimates shown in Table 4.13, which now resemble their Least Squares counterparts very closely. The fit of the reparameterised model to the Onion data is provided in Figure 4.17. The improvement over the previous fit (Figure 4.13) is obvious.

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	22.0800	22.0667	0.7449	22.0063	20.7606	23.7558
$\hat{\beta}$	1.6530	1.6772	0.4854	1.6793	0.7021	2.6430
$\hat{\gamma}$	5585.90	7204.06	3.4117	6479.51	850.734	114691.4
$\log(\hat{\gamma})$	8.6280	8.8824	1.2272	8.7764	6.7461	11.6568
$\hat{\delta}$	4.5600	4.6933	0.6558	4.6397	3.5326	6.1812
$\hat{\tau}$	1.7271	1.7154	0.7248	1.6146	0.6133	3.3705
Cucumber Data:						
$\hat{\alpha}$	6.9860	6.9839	0.1517	6.9795	6.6961	7.2974
$\hat{\beta}$	1.1810	1.1870	0.1055	1.1862	0.9788	1.4000
$\hat{\gamma}$	12.9604	13.2699	1.2218	13.0946	9.1862	20.6187
$\log(\hat{\gamma})$	2.5619	2.5855	0.2003	2.5722	2.2177	3.0262
$\hat{\delta}$	2.4750	2.4976	0.1952	2.4872	2.1300	2.9254
$\hat{\tau}$	208.3333	116.1218	72.7156	101.2977	20.1144	293.4719
Onion Data:						
$\hat{\alpha}$	723.9000	722.0931	26.0697	721.1746	673.9311	776.3938
$\hat{\beta}$	33.3500	37.9107	19.6793	36.7288	4.4723	80.9006
$\hat{\gamma}$	6266.041	10091.01	4.36536	8462.501	993.2675	338608.6
$\log(\hat{\gamma})$	8.7429	9.2194	1.4737	9.0434	6.9010	12.7326
$\hat{\delta}$	4.6410	4.8866	0.7797	4.7973	3.6381	6.7250
$\hat{\tau}$	0.0010	0.0010	0.0004	0.0009	0.0003	0.0020
Pasture Data:						
$\hat{\alpha}$	80.9600	83.6873	12.9151	81.1935	68.3074	118.2702
$\hat{\beta}$	8.8950	8.8378	2.4422	8.9183	3.5247	13.6827
$\hat{\gamma}$	49577.9	72482.5	14.2008	49503.6	14869.8	232104900
$\log(\hat{\gamma})$	10.8113	11.1911	2.6533	10.8098	7.3045	19.2627
$\hat{\delta}$	2.8280	2.9189	0.7360	2.8276	1.7499	5.1156
$\hat{\tau}$	0.3690	0.3624	0.2218	0.3207	0.0573	0.8641

Table 4.13: Summary Statistics: Reparameterised MMF Models.

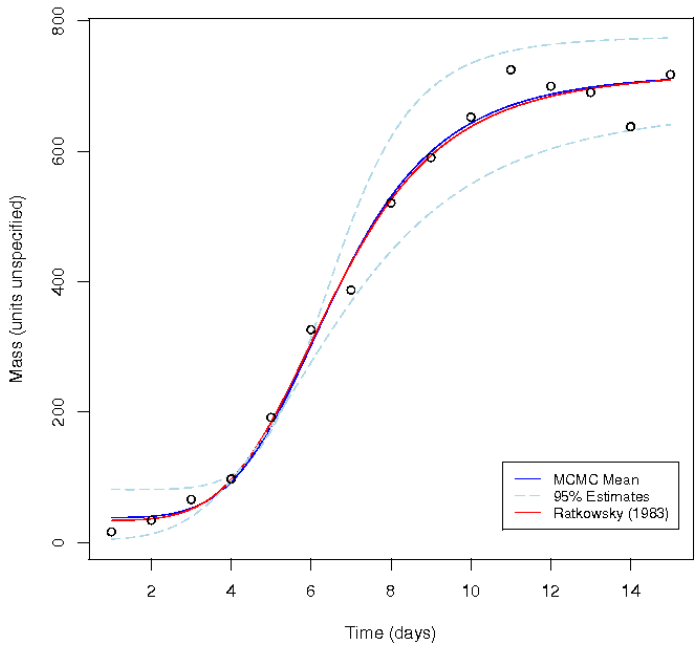


Figure 4.17: Fitted Reparameterised MMF-Onion Model.

Richards Models

Table 4.12 shows that the Richards model function (4.14) has the highest intrinsic curvature across all of the data considered. In addition, it has the lowest parameter effects curvature of the four parameter models for all but the Cucumber data. These factors suggest that reparameterisation may not provide a solution for the Richards model function. Indeed, Ratkowsky (1983) showed that a number of alternative parameterisations performed even less well.

Given this prognosis, we re-fit the original form using simulated data to test the idea that more data may provide sufficient information regarding posterior structure. We used (4.14) with Ratkowsky's Least Squares estimates for the model mean and variance and generated 18 data points – twice the original number. The MCMC procedure was re-run using this simulated data.

The model fit can be seen against the simulated data in Figure 4.18. The red curve is the Least Squares fit to the original data, the deterministic component from which the simulated data were generated. The model appears to fit well, showing evidence that the MCMC estimates are sensitive to nuances of the data. Even this modest increase in available data has allowed the MCMC method to succeed. We repeated the simulation exercise for the Pasture, Onion and Bean data and observed similar results. MCMC appears to be viable for cases which feature intrinsic curvature,

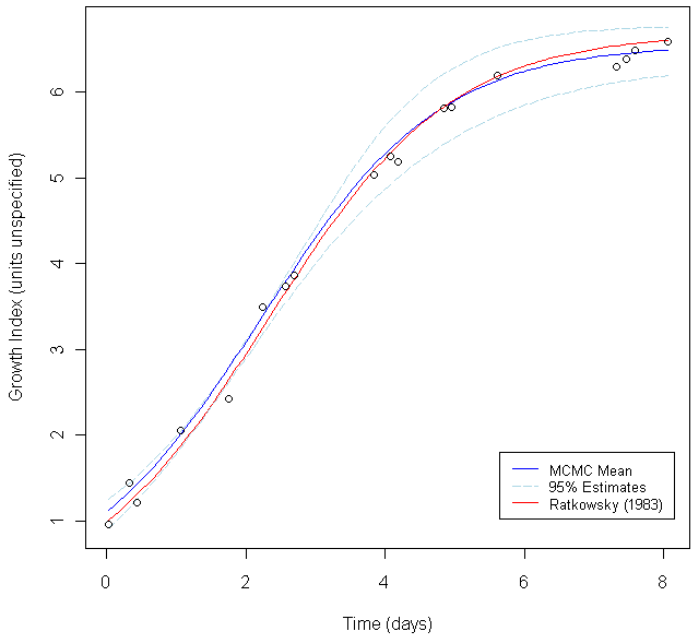


Figure 4.18: Fitted Simulated Richards-Cucumber Model.

provided sufficient data are available to provide information regarding the structure of the posterior.

Weibull-type Models

Table 4.12 showed the Weibull-type model function to have similar characteristics to the MMF model function. Ratkowsky (1983) suggested

$$\boldsymbol{y} = \exp(\alpha) - \exp[\beta - \exp(-\gamma)\boldsymbol{x}^{\delta}] \tag{4.17}$$

as a replacement for the original Weibull-type model function (4.15). When applied to the Pasture dataset, the reparameterised version provides a parameter effects curvature measure of 1.43, compared to the previous value of 42.68.

Results from reparameterised function (4.17) are provided in Table 4.14. A plot of the fitted Weibull-Pasture model using this function appears in Figure 4.19. As observed with the reparameterised MMF model function, the new MCMC estimates very closely resemble their Least Squares counterparts.

		MCMC		*Quantiles*		
	Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Bean Data:						
$\hat{\alpha}$	3.0493	3.0474	0.0202	3.0471	3.0079	3.0894
$\hat{\beta}$	2.9862	2.9812	0.0357	2.9814	2.9078	3.0522
$\hat{\gamma}$	6.3368	6.5176	0.7026	6.4573	5.3168	8.0765
$\hat{\delta}$	3.1800	3.2729	0.3580	3.2427	2.6563	4.0609
$\hat{\tau}$	2.0202	1.9879	0.8419	1.8714	0.6899	3.9603
Cucumber Data:						
$\hat{\alpha}$	1.8955	1.8917	0.0284	1.8923	1.8346	1.9454
$\hat{\beta}$	1.7136	1.7017	0.0543	1.7049	1.5775	1.8023
$\hat{\gamma}$	2.1371	2.2283	0.3153	2.1870	1.7513	2.9941
$\hat{\delta}$	1.7630	1.8420	0.2616	1.8086	1.4442	2.4683
$\hat{\tau}$	37.3134	31.9001	19.9819	28.0000	5.1600	80.8659
Onion Data:						
$\hat{\alpha}$	6.5439	6.5422	0.0207	6.5420	6.5021	6.5834
$\hat{\beta}$	6.5125	6.5067	0.0392	6.5068	6.4268	6.5848
$\hat{\gamma}$	6.4890	6.6697	0.7747	6.6019	5.3420	8.3829
$\hat{\delta}$	3.2620	3.3531	0.3845	3.3222	2.7014	4.2046
$\hat{\tau}$	0.0014	0.0014	0.0006	0.0013	0.0005	0.0029
Pasture Data:						
$\hat{\alpha}$	4.2479	4.2529	0.0493	4.2467	4.1731	4.3760
$\hat{\beta}$	4.1220	4.1278	0.0774	4.1208	3.9933	4.3095
$\hat{\gamma}$	9.2103	9.3063	1.1950	9.2381	7.1716	11.9885
$\hat{\delta}$	2.3780	2.4008	0.3017	2.3901	1.8174	3.0789
$\hat{\tau}$	0.5952	0.5914	0.3632	0.5206	0.1051	1.4998

Table 4.14: Summary Statistics: Reparameterised Weibull Models.

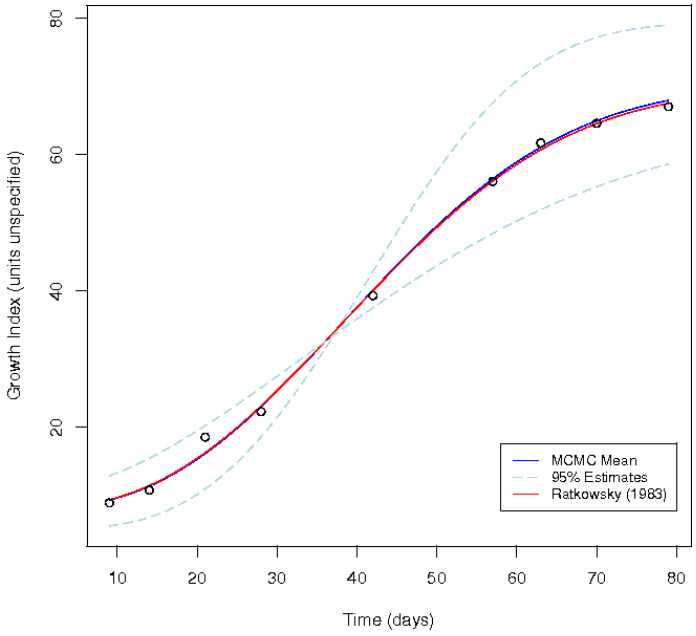


Figure 4.19: Fitted Reparameterised Weibull-Pasture Model.

4.6 Discussion

4.6.1 Back Transformation

One of the chief benefits of having simulated posterior samples available is the ability to summarise the posterior in any way that suits our needs. We have seen throughout this chapter that this is very useful: we may apply functions to these samples to obtain summary statistics, diagnose performance problems by viewing the MCMC chain trace and pairwise marginal scatterplots, or use the latter simply because visualisation provides a meaningful interpretative aid. However, there is yet another very useful application for posterior samples: transformation of posterior samples obtained under one parameterisation to produce samples from another.

For example, in §4.5.5 we found that the reparameterised Weibull function (4.17) allowed convergence by the MCMC routine, whereas the initial parameterisation (4.15) appeared problematic. If, however, there was some motivation leading us to prefer (4.15), samples from that parameterisation can be obtained simply by taking the appropriate transformation of the samples obtained under (4.17).

This has been done to produce Table 4.15, which shows remarkable agreement between the transformed MCMC estimates and those produced using Least Squares. Moreover, the transformed samples provide the same rich source of information re-

	MCMC		*Quantiles*		
Ratkowsky	*Mean*	σ	50%	2.5%	97.5%
Pasture Data:					
$\hat{\alpha}$	69.9600	70.4840	3.7678	69.9833	65.3389	79.2471
$\hat{\beta}$	61.6800	62.2217	4.9211	61.6026	54.8664	74.2035
$\hat{\gamma}$	0.0001	0.0002	0.0003	0.0001	0.0000	0.0008
$\hat{\delta}$	2.3780	2.4008	0.3017	2.3901	1.8174	3.0789
$\hat{\tau}$	0.5952	0.5914	0.3632	0.5206	0.1051	1.4998

Table 4.15: Summary Statistics: Back-Transformed Weibull-Pasture Data.

garding the posterior as directly simulated samples. Credible intervals associated with the transformed parameter estimates are also provided in Table 4.15.

4.6.2 Posterior Curvature

We have seen that our MCMC method performs well in cases which feature low curvature component values, in keeping with Least Squares. The idea that MCMC should so closely mimic the performance of Least Squares is initially somewhat surprising. After all, MCMC requires only that samples are obtained from the entire support of the posterior proportional to its density, not that the sampling distributions of the parameters approximate normality.

The reason for the similarity in performance lies in the fact that our implementation of MCMC mixes slowly in cases where posterior curvature is significant. In principle, MCMC would still arrive at reasonable estimates if the chain were allowed to run for a sufficiently long period. However, we have no way of knowing how long that may be, and prefer to construct more efficient alternatives. Insight into the mixing of the chain is offered by examination of the pairwise marginal scatterplots of the posterior distribution before and after reparameterisation. Examples based on the MMF – Onion models are provided in Figure 4.20 and Figure 4.21 respectively.

Consider the lower $\hat{\gamma}$ – $\hat{\delta}$ plot provided in Figure 4.20. Imagine the current value of the chain to be a point P in the lower left tail of the posterior slice. P has low values for both γ and δ. Given the shape of the posterior in this region, a proposed point Q with low values for γ and low to mid-range δ values stand a reasonable chance of acceptance. However, even a slight increase in γ in this region would cause the proposed sample to fail the acceptance test (3.7) with virtual certainty. Conversely, when P resides in the the upper right of the posterior section, horizontal excursions in proposed points would be reasonably well tolerated and vertical excursions less so. The net result is that proposed points are frequently rejected and the chain mixes slowly.

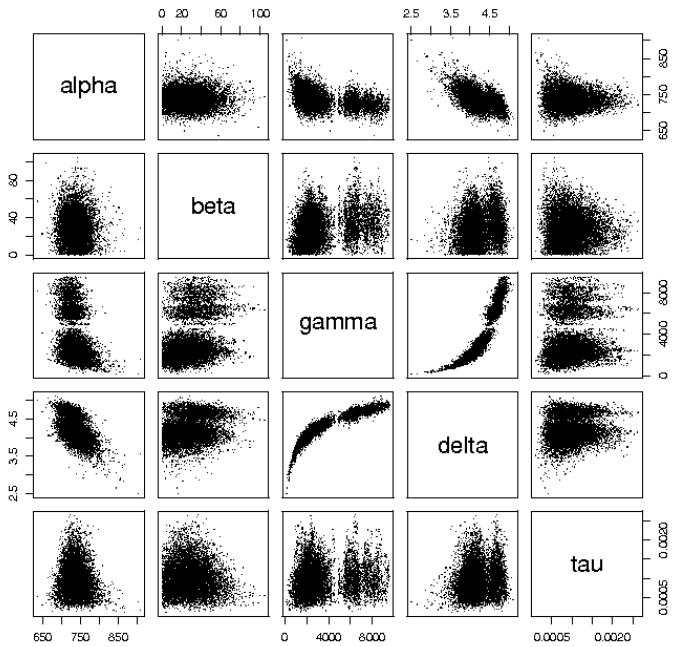


Figure 4.20: Marginal Scatterplots: MMF-Onion Model.

One can envisage an adaptive proposal regime which would better accommodate posterior curvature (see, for example, Gilks et al., 1994; Andrieu and Thoms, 2008; Cai et al., 2008), but that is not explored further here. Instead, we recognise that when using Metropolis sampling methods a transformed proposal distribution with the original data is equivalent to the original proposal with transformed data. In this case the alternative parameterisation (4.16) has effectively transformed the relationship between γ and δ and allows our existing method to adequately assess the support of the posterior. Despite the strong correlation between γ and δ the MCMC method has no trouble converging to the correct solution.

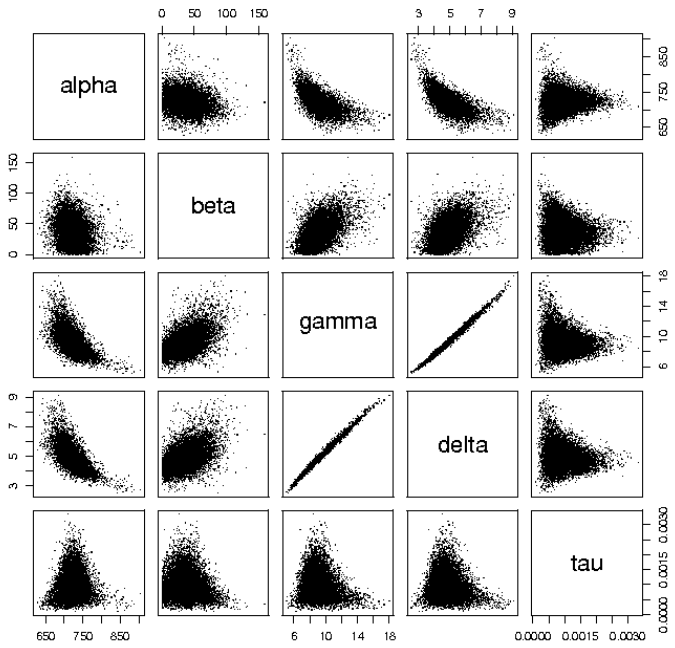


Figure 4.21: Marginal Scatterplots: Reparameterised MMF-Onion Model.

4.7 Conclusion

We have seen that MCMC is comparable to Least Squares for nonlinear regression. Cases for which our MCMC method performed less well were shown to be cases with strong curvature components, for which Least Squares also performs poorly. Reparameterisation of the model functions used adequately addressed these performance problems in cases where the curvature was a result of the original parameterisation employed. Again, these were the same adjustments identified by Ratkowsky (1983) as required to obtain adequate performance from Least Squares estimators.

While the estimates obtained under the two schemes are similar, MCMC offers some natural advantages. Interval estimates do not require approximate normality to avoid bias (Hartley, 1964; Box, 1971), and are immediately available by taking quantiles of the posterior samples. Thus, in all of the cases where reasonable parameter estimates were obtained, readily interpretable credible intervals are also available. This offers a significant improvement over the Least Squares point estimates offered by Ratkowsky (1983), as these require substantial extra effort to subsequently check sampling distributions and estimate bias in any intervals produced.

The availability of posterior samples was shown to provide additional advantages in the diagnosis of problem situations. MCMC chain traces and posterior sections were

both shown to provide useful diagnostic information, and the latter aided interpretation of the posterior. We also showed that back transformation allowed posterior samples obtained under one parameterisation to provide estimates and inference regarding alternative parameterisations. This offers another substantial advantage over Least Squares methods, allowing practitioners to quickly and easily explore alternative parameterisations and importantly, to avoid being forced to consider a restricted range of parameterisations simply because their sampling properties render them the only tractable option.

Chapter 5

Response Transformations

5.1 Introduction

In §4.2 we outlined the assumptions which underpin the general linear model. When these criteria are satisfied the task of the analyst is simplified, and models which summarise the data using only a small number of parameters result. This is the power of parametric modelling. Unfortunately real world data often fail to meet these fundamental requirements, and depending upon which of the criteria are violated the analyst may need to turn to more sophisticated modelling tools. However, if it is not possible to meet the criteria on the original scale of the data, it may be that there is a nonlinear transformation of the response which will provide an adequate remedy. In particular, data which are heteroscedastic or of questionable normality are often amenable to transformation. Analysts frequently wish to exhaust this possibility before abandoning the general linear model.

There are a number of methods by which a suitable transformation might be selected. A practitioner may simply select from a handful of candidate transformations based on an careful examination of the model residuals. Atkinson (1985) and Carroll and Ruppert (1988) provide book-length treatments of transformation methods and their use. Alternately, techniques such as Box–Cox transformation (Box and Cox, 1964) automatically select an optimal transformation from a parametric family. Still more sophisticated techniques such as ACE (Breiman and Friedman, 1985), AVAS (Tibshirani, 1988) and the method due to Kruskal (1965) construct a monotone transformation of the response so the requirements of the general linear model are more nearly met.

Unfortunately, all these methods for selecting a transformation suffer a common limitation - any subsequent inferences make no allowance for the uncertainty in the choice of transformation. The implications of this shortcoming have been controversial in the literature (Bickel and Doksum, 1981; Box and Cox, 1982; Hinkley and Runger, 1984; Rubin, 1984), and a method to address it would presumably be welcomed.

We develop a Bayesian MCMC approach that selects a monotone transformation of the response so that the transformed data best fits an assumed linear model. Because we jointly estimate the model coefficients and the transformation of the response, any inferences will reflect the uncertainty in the choice of transformation. What is most surprising about our approach is that it depends only on the order of the responses – it is a rank method.

5.2 The Method

The method is considerably easier to state in the absence of ties, so initially assume $y = (y_1, y_2, \ldots, y_n)$ is a vector of responses such that $y_i < y_j$ when $i < j$. We seek a strictly increasing transformation f of the responses such that

$$\begin{aligned} z_i &= f(y_i) \qquad\qquad i = 1, \ldots, n \\ \boldsymbol{z} &\sim \mathrm{N}(X\beta, \sigma^2 I) \end{aligned} \tag{5.1}$$

where $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ is the vector of transformed responses, X is a known design matrix and β is a vector of coefficients to be estimated. As f is arbitrary, we may fix the scaling of f and set $\sigma = 1$; that is, $\boldsymbol{z} \sim \mathrm{N}(0, 1)$.

Rather than determine f directly, we estimate the vector z of transformed responses. The prior for f will induce constraints on z. We choose a prior for f that is "uniform" in the sense that it places no further constraints on z other than the order constraints imposed by the monotonicity of f and the ordering of y.

If we also adopt an improper uniform prior for β, the posterior $p(z, \beta \,|\, y)$ is determined by (5.1) alone. As we are modelling the transformed responses, it would make little sense to adopt an informative prior for β.

We construct samples from the joint posterior for β and z by Gibbs sampling (Gilks et al., 1995b). To Gibbs sample from the posterior we require the conditional distributions of the z_i and β. As f is strictly increasing, $y_i < y_j$ implies $z_i < z_j$, so that the conditional distributions of the individual z_i are simply

$$\begin{aligned} z_1 \,|\, z_2, \beta &\sim \mathrm{N}\left((X\beta)_1, 1\right) \mathrm{I}(-\infty, z_2) \\ z_i \,|\, z_{i-1}, z_{i+1}, \beta &\sim \mathrm{N}\left((X\beta)_i, 1\right) \mathrm{I}(z_{i-1}, z_{i+1}) \qquad 1 < i < n \\ z_n \,|\, z_{n-1}, \beta &\sim \mathrm{N}\left((X\beta)_n, 1\right) \mathrm{I}(z_{n-1}, \infty) \end{aligned} \tag{5.2}$$

where $(X\beta)_i$ is the i-th component of $X\beta$, and $\mathrm{N}(\mu, \sigma^2)\,\mathrm{I}(a, b)$ denotes the $\mathrm{N}(\mu, \sigma^2)$ distribution truncated to the open interval (a, b). Since $p(\beta \,|\, z, y) = p(\beta \,|\, z)$, we have that

$$\beta \,|\, z \sim \mathrm{N}\left((X^T X)^{-1} X^T z, (X^T X)^{-1}\right). \tag{5.3}$$

Together these relations define a Gibbs sampling scheme for β and z. To draw samples from the posterior, we simply draw β and the components of z in succession

from their conditionals (5.2) and (5.3).

In the presence of ties, the order constraints on the z_i are more complex. When there are ties, there are $m < n$ distinct observations, $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m$. Order these so that $\tilde{y}_i < \tilde{y}_j$ when $i < j$, and define index sets

$$I_k = \{i \,|\, y_i = \tilde{y}_k\}$$

with cardinalities $n_k = |I_k|$.

We have two options, we may allow the procedure to break ties, or we may choose to enforce the ties.

If we allow the procedure to break ties, then f is multi-valued and we need only preserve the ordering among distinct values. In this case the conditional distributions of the z_i are

$$\begin{aligned} z_1 \,|\, Z_2, \beta &\sim \mathrm{N}\left((X\beta)_1, 1\right) \mathrm{I}(-\infty, \min Z_2) \\ z_i \,|\, Z_{i-1}, Z_{i+1}, \beta &\sim \mathrm{N}\left((X\beta)_i, 1\right) \mathrm{I}(\max Z_{i-1}, \min Z_{i+1}) \qquad 1 < i < n \\ z_n \,|\, Z_{n-1}, \beta &\sim \mathrm{N}\left((X\beta)_n, 1\right) \mathrm{I}(\max Z_{n-1}, \infty) \end{aligned} \tag{5.4}$$

where $Z_k = \{z_i \,|\, i \in I_k\}$. Again, together (5.3) and (5.4) define a Gibbs sampling scheme for β and z.

Alternately, if we choose to preserve the ties then f is univalent and there are $m < n$ distinct values $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_m$ such that $\tilde{z}_k = f(\tilde{y}_k)$ and

$$z_i = \tilde{z}_k \qquad \text{when } i \in I_k. \tag{5.5}$$

In this case the conditional distributions of the $\tilde{z}_i$ are simply

$$\begin{aligned} \tilde{z}_1 \,|\, \tilde{z}_2, \beta &\sim \mathrm{N}\left(n_1^{-1} \sum_{j \in I_1} (X\beta)_j, n_1^{-1}\right) \mathrm{I}(-\infty, \tilde{z}_2) \\ \tilde{z}_i \,|\, \tilde{z}_{i-1}, \tilde{z}_{i+1}, \beta &\sim \mathrm{N}\left(n_i^{-1} \sum_{j \in I_i} (X\beta)_j, n_i^{-1}\right) \mathrm{I}(\tilde{z}_{i-1}, \tilde{z}_{i+1}) \qquad 1 < i < n \\ \tilde{z}_n \,|\, \tilde{z}_{n-1}, \beta &\sim \mathrm{N}\left(n_n^{-1} \sum_{j \in I_n} (X\beta)_j, n_n^{-1}\right) \mathrm{I}(\tilde{z}_{n-1}, \infty) \end{aligned} \tag{5.6}$$

and conditions (5.3), (5.5) and (5.6) define a Gibbs sampling scheme for β and z.

5.2.1 Discussion

Although fixing $\sigma = 1$ fixes the scale of f, in general the model remains unidentifiable and the posterior is improper. Our modest requirement that f is strictly

increasing implies that if $z_i = f(y_i)$ is a suitable transformation, then so is any arbitrary translation $z_i = f(y_i) + k$, and k will be confounded with the intercept term in the model. But the situation is more complex than this – consider a simple two sample test where y consists of observations from two treatment groups A and B, and we wish to fit means μ_A and μ_B to the transformed observations. Then again, as we only require f to be strictly increasing, if the observations from the two groups do not interleave so that all B observations exceed the A observations, then we may choose f to make $\mu_B - \mu_A$ arbitrarily large.

The identifiability problem may be resolved by further constraining either β or z. However, Gelfand and Sahu (1999) suggest this may not be necessary, and that if we run a Gibbs sampler for which the posterior is improper but all the full conditionals are proper, then we may still be able use the output to obtain meaningful inferences for the identifiable components of the model. Indeed, this is our experience – while the chain does not converge in general, the components of the chain corresponding to identifiable parameters do converge and meaningful inferences can be made.

5.3 Example: Toxic Agents

Box and Cox (1964) describe an experiment investigating the survival times of rats to illustrate their transformation method. We re-consider that example here. Rats were divided at random into groups of size 4, with each group receiving one of three poisons and one of four treatments. Thus there were two factors, one of three and the other of four levels, in a replicated 3×4 factorial design applied to $n = 48$ rats. The data are provided in Table 5.1, and survival times are plotted against experimental factors in Figure 5.1.

Treatment	Poison I	Poison II	Poison III
A	0.31, 0.45, 0.46, 0.43	0.36, 0.29, 0.40, 0.23	0.22, 0.21, 0.18, 0.23
B	0.82, 1.10, 0.88, 0.72	0.92, 0.61, 0.49, 1.24	0.30, 0.37, 0.38, 0.29
C	0.43, 0.45, 0.63, 0.76	0.44, 0.35, 0.31, 0.40	0.23, 0.25, 0.24, 0.22
D	0.45, 0.71, 0.66, 0.62	0.56, 1.02, 0.71, 0.38	0.30, 0.36, 0.31, 0.33

Table 5.1: Toxic Agent Data, (Box and Cox, 1964)

From the boxplots we can see that both treatment and poison types appear to have differential effects on survival time. Treatments A and C appear to be more effective than B and D, and poison III appears to be the most effective of those on trial. There is also some indication that the variability of the response is related to the mean: both the treatments and poison types associated with shorter survival times also appear to be less variable.

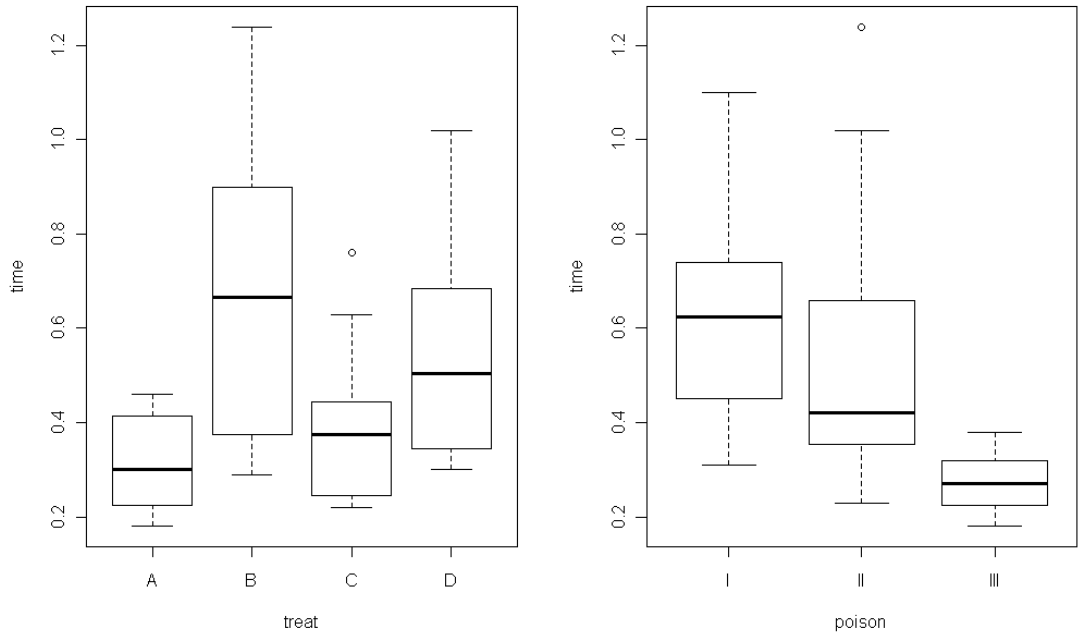


Figure 5.1: Survival Time by Experimental Factor

5.3.1 Model Fitting

We begin by considering the model

$$y_{tpj} = \mu + \tau_t + \pi_p + \varepsilon_{tpj}, \tag{5.7}$$

where μ represents a baseline response in the absence of treatments or poisons, τ_t represents the effect of the tth treatment, π_p the effect of the pth poison, and ε_{tpj} is the residual for the jth replicate given the tth treatment and the pth poison.

The analysis of variance associated with model (5.7) is shown in Table 5.2, from which it appears that both treatment and poison type are highly significant main effects.

```
            Df  Sum Sq Mean Sq F value    Pr(>F)
treat        3 0.92121 0.30707  12.273 6.697e-06 ***
poison       2 1.03301 0.51651  20.643 5.704e-07 ***
Residuals   42 1.05086 0.02502
```

Table 5.2: ANOVA: Poison Model (5.7)

5.3.2 Model Checking

We can assess the fit of the model by inspecting a plot of the standardised residuals against the fitted values of the model, as shown in the left-hand panel of Figure 5.2. A nonparametric locally weighted regression (loess) (Cleveland, 1979, 1981) smooth line has been added to aid interpretation.

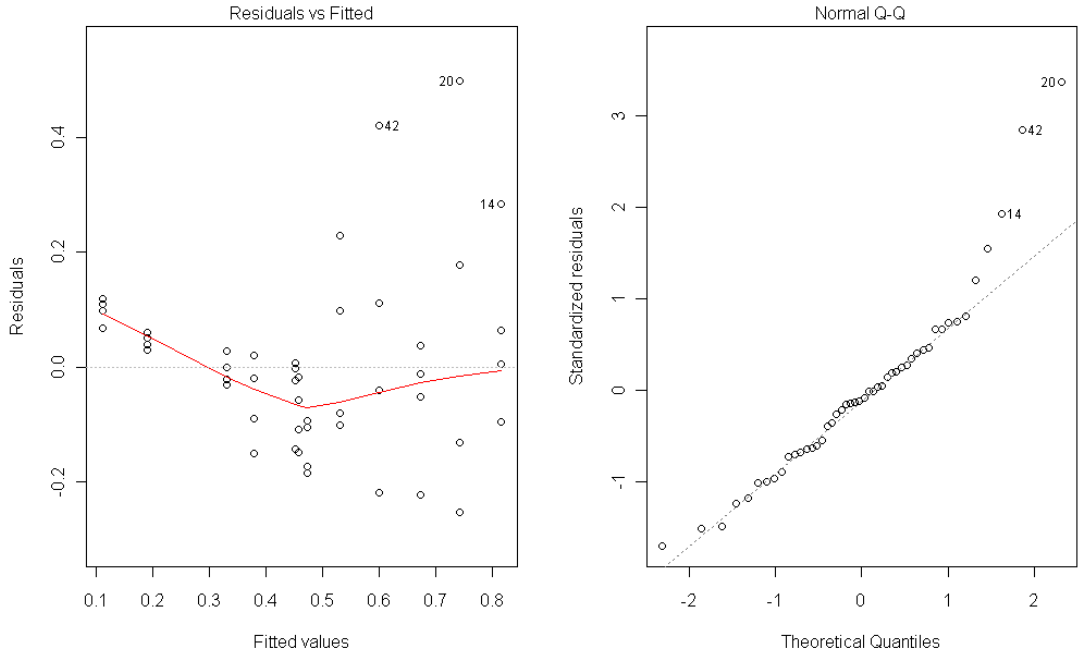


Figure 5.2: Diagnostic Plots: Poison Model (5.7)

The plot shows a striking increase in variance among the residuals as the mean fitted response increases. We can also see that the model under predicts survival for the shortest response times, and over-predicts in the mid-range where the residuals are mostly negative. Box et al. (1978), suggest that this indicates the presence of "transformable nonaddivity" (Tukey, 1949; Anscombe and Tukey, 1963) among the treatment and poison effects. The right-hand panel indicates that the residuals display positive skew relative to $\mathcal{N}(\bar{y}, s^2)$, the normal distribution with the parameters equal to the sample mean and variance.

5.3.3 Box-Cox Transformation

Box and Cox (1964) suggest that the inadequacies of the model can be corrected by a transformation of the response. In particular they put forward the following transformation schema for data $y > \lambda_2$

$$y^{(\lambda)} = \begin{cases} ((y+\lambda_2)^{\lambda_1} - 1)/\lambda_1, & \lambda_1 \neq 0 \\ \log(y+\lambda_2), & \lambda_1 = 0. \end{cases} \tag{5.8}$$

Since the survival times $y > 0$, setting $\lambda_2 = 0$ reduces (5.8) to the more familiar single parameter form

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0, \end{cases} \tag{5.9}$$

where the value of λ is chosen to render the model residuals most normal and homoscedastic. The boxcox() function from the MASS package (Venables and Ripley, 2002) distributed with the R statistical environment (R Development Core Team, 2009) plots the profile log likelihood for the transformation parameter λ. This is shown in Figure 5.3.

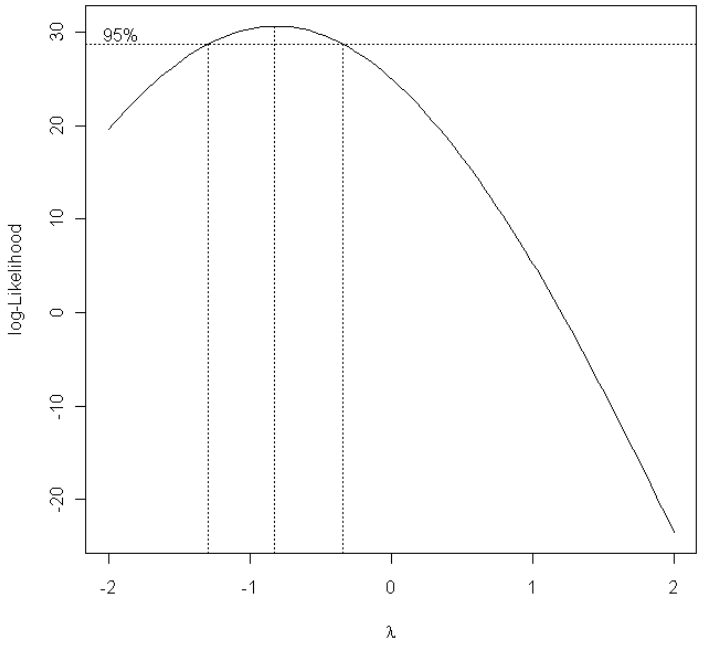


Figure 5.3: Profile Log Likelihood, Box-Cox Transformation (5.9)

The domain for λ is shown as $[-2, 2]$, offering a range of easily interpreted parameter values. The square- ($\lambda = \frac{1}{2}$) and cube-roots ($\lambda = \frac{1}{3}$), for instance, and their inverses ($\lambda = -\frac{1}{2}, \lambda = -\frac{1}{3}$), all fall within this domain. For our example, it is clear that the original scale of the response ($\lambda = 1$) is a poor choice. A log transformation ($\lambda = 0$) also seems inappropriate, as this falls outside the 95% confidence interval for the profile log likelihood of λ. The maximum value appears to be around -0.8, however, it is not easy to interpret this value, and it is usual practice to adopt the closest

value which offers an accessible interpretation. The most readily interpretable choice appears to be $\lambda = -1$, corresponding to fitting a linear model to the inverse response

$$y^{(\lambda)} = 1/y_{tpj} = \mu + \tau_t + \pi_p + \varepsilon_{tpj} \tag{5.10}$$

Box and Cox (1964) suggest that this inverse response model can be interpreted as the "rate of dying", measured in units of time^{-1}.

5.3.4 Evaluating the Transformed Model

Model (5.10) was fit to the data with the results shown in Table 5.3. Again, both of the main effects are highly significant. Because the scale of the response has been altered by the transformation, the only valid comparisons that can be made between Table 5.2 and Table 5.3 are those based on the F values. We note that the transformation has strengthened the case for significant main effects by more than a factor of two for each experimental factor.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
treat	3	20.414	6.805	27.982	4.192e-10	***
poison	2	34.877	17.439	71.708	2.865e-14	***
Residuals	42	10.214	0.243			

Table 5.3: ANOVA: Reciprocal Transformed Poison Model

The diagnostic plots for the transformed model are shown in Figure 5.4. There has been considerable improvement. The left hand plot shows that the residual variance is substantially improved, though the loess smooth line still suggests some evidence of curvature. The right-hand plot shows that the skewness of the residuals has also been reduced.

5.3.5 MCMC Transformation

We now revisit the example with an MCMC approach based on the method presented in §5.2.

Model Fitting

We note the presence of ties in the y_i and begin by electing to preserve these, employing (5.3) and (5.4) and using $M = 1000$ iterations of the Gibbs Sampler thinned at every $k = 10$th iteration, to fit

$$z_{tpj} = f(y_{tpj}) + \varepsilon_{tpj} = \mu + \tau_t + \pi_p + \varepsilon_{tpj} \tag{5.11}$$

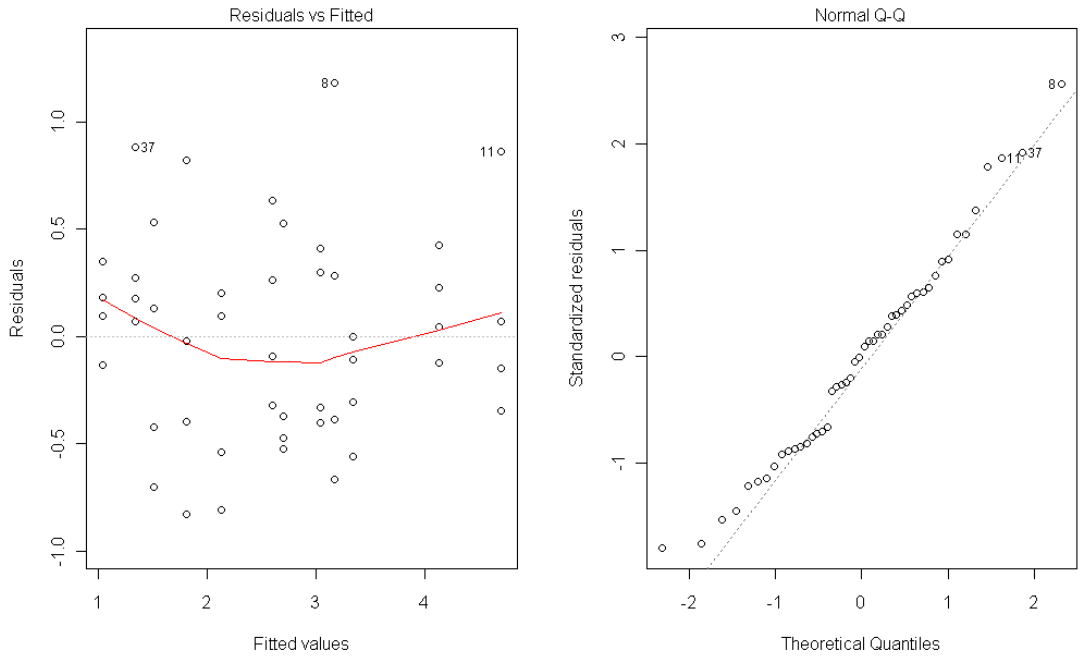


Figure 5.4: Diagnostic Plots: Transformed Poison Model (5.10)

where $f(\boldsymbol{y})$ is an arbitrary monotonic function of the response, μ is the mean response in the reference group, here those rats receiving `poison I` using `treatment A`, τ_t and π_p are adjustments to the mean for the t-th treatment, $t = 2, 3, 4$, and p-th poison, $p = 2, 3$, and $\varepsilon \sim \mathcal{N}(0, 1)$.

Results

The MCMC transformation model (5.11) produces the ANOVA results shown in Table 5.4, and the estimates shown in Table 5.5. Table 5.4 is based on the mean posterior estimates for the individual z_i, providing equivalent results to the ANOVA table produced under the reciprocal model (5.10), again showing highly significant main effects. By comparing the F-values to those in Table 5.3 we can see that this transformation provides comparable, and slightly enhanced, significance for each main effect relative to the inverse transformation suggested by the Box–Cox method.

The estimates for the model coefficients and associated 95% credible interval limits are provided in Table 5.5. As discussed in §5.2.1, the mean of the transformed response is confounded with any translation of the estimated transformation f. The 95% credible interval for this value includes zero, and $\mu = 0$ is the most reasonable interpretation for this term.

```
            Df  Sum Sq Mean Sq F value    Pr(>F)
treat        3  85.974  28.658  28.009 4.135e-10 ***
poison       2 151.235  75.618  73.906 1.752e-14 ***
Residuals   42  42.973   1.023
```

Table 5.4: ANOVA: MCMC Transformed Poison Model

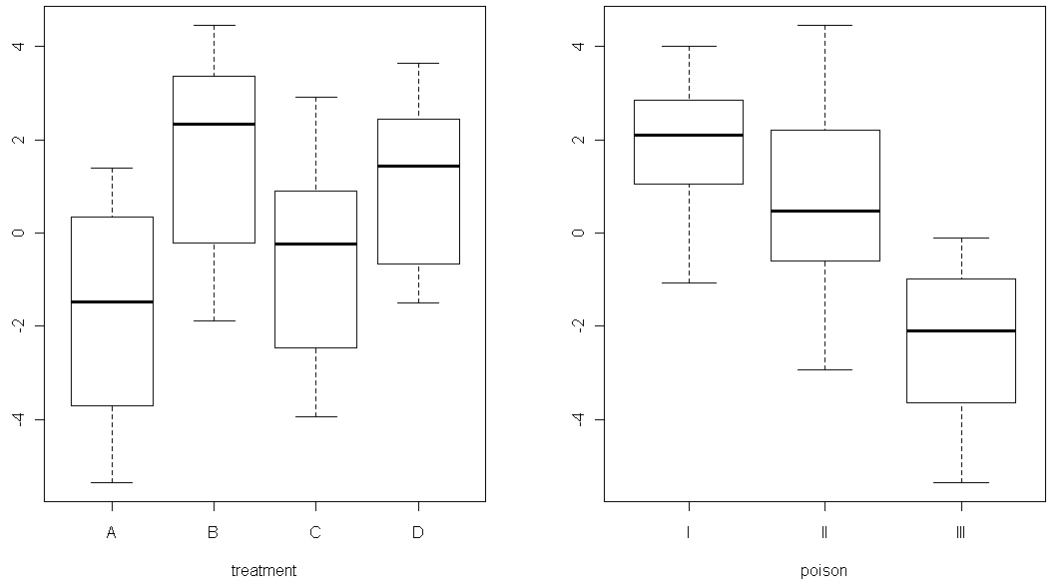


Figure 5.5: Transformed Survival Time by Experimental Factor

The remaining estimates are the relative adjustments attributed to the `treatment` and `poison` factors. These can be readily interpreted by inspecting boxplots of the transformed response mean estimates z_i by treatment factor, as shown in Figure 5.5. Relative adjustments to an unidentifiable intercept initially appears to be unhelpful, but because the transformed response estimates z_i are available, calculating the means for the reference groups is straightforward. Here, the mean corresponding to `treatment A` $\tau_1 = -1.7322$, and that corresponding to `poison I` $\pi_1 = 1.8583$. The mean estimates listed in Table 5.5 are adjustments to these values, as can be seen by comparing these values to the boxplot means in Figure 5.5.

The estimated transformation $z_i = \hat{f}(y_i)$ is shown against the ordered data y_i in Figure 5.6. The median estimate is shown in red, with the mean estimate in blue. Obviously these correspond closely with one another, an indication of posterior symmetry. The transformed points z_i by which the estimate of $f(.)$ is determined are overlaid. As we elected to preserve ties there are 34 points corresponding to the 34 unique values of $y_i, i = 1, \ldots, 48$. Bands corresponding to the 95% credible interval estimates for $f(.)$ are shown in light blue.

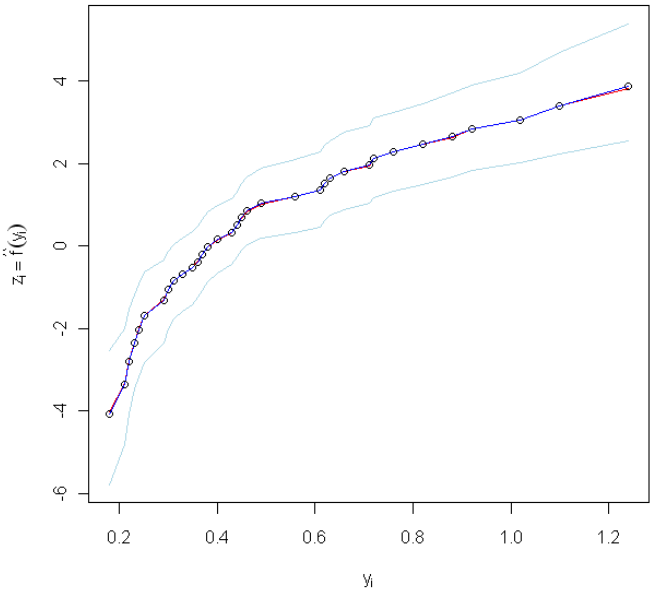


Figure 5.6: MCMC Estimated Response Transformation: Model (5.11)

Figure 5.7 shows a plot of the empirical quantiles of the transformed response against the quantiles of the normal distribution with parameters equal to the mean and variance of the transformed sample. Apparently the transformation has been quite successful at rendering the response normal. In particular the tails of the distribution have been shortened relative to the Box–Cox suggested inverse transformation.

Finally, the residual versus fitted values plot for the MCMC transformation model (5.11) is shown in Figure 5.8. Rather strikingly, there is now no indication of any unexplained trend between the residuals and the conditional mean of the model, and again this appears to be an improvement over the Box–Cox result. We can also see from the plot that the estimated transformation has successfully stabilised the variance.

Breaking Ties

If we were prepared to allow ties in the original data to be broken, model (5.11) can be fit using (5.3), (5.5) and (5.6). The estimated transformation resulting from this fit is shown in Figure 5.9. While the result is very similar to that shown in Figure 5.6, close scrutiny reveals that $f(.)$ is now determined by 48 unique $\boldsymbol{z}$ values.

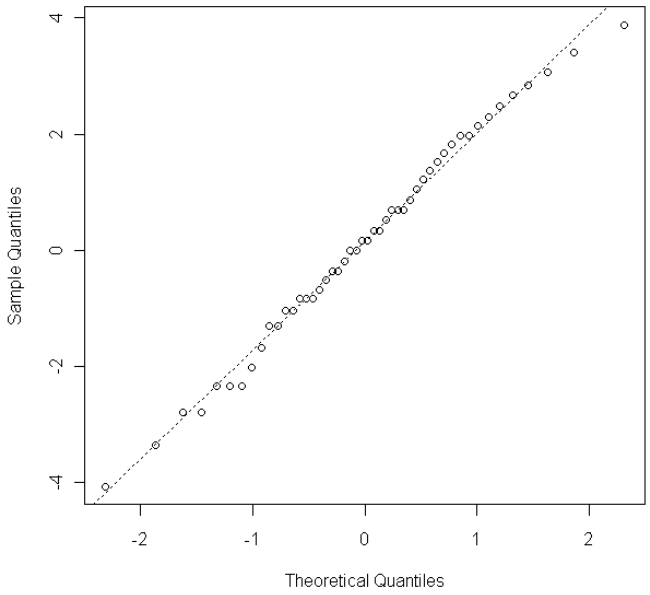


Figure 5.7: Normal Quantile Quantile Plot: MCMC Estimated Transformation, Model (5.11)

	$\hat{\mu}$	$\hat{\tau_2}$	$\hat{\tau_3}$	$\hat{\tau_4}$	$\hat{\pi_2}$	$\hat{\pi_3}$
mean	-0.3722	3.4611	1.2235	2.7275	-1.0115	-4.1743
sd	1.2388	0.5731	0.4561	0.5352	0.3921	0.5832
2.5%	-2.8188	2.3859	0.3336	1.7005	-1.8088	-5.3564
50%	-0.3012	3.4569	1.2278	2.7398	-1.0117	-4.1624
97.5%	1.7395	4.5699	2.1170	3.8059	-0.2570	-3.0749

Table 5.5: Parameter Estimates: Transformed Poison Model (5.11)

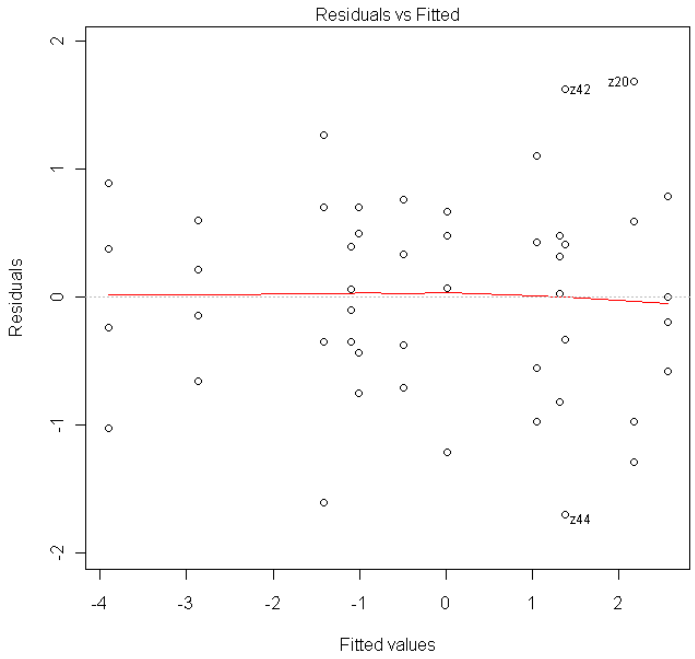


Figure 5.8: Residuals vs Fitted Values: MCMC Model (5.11)

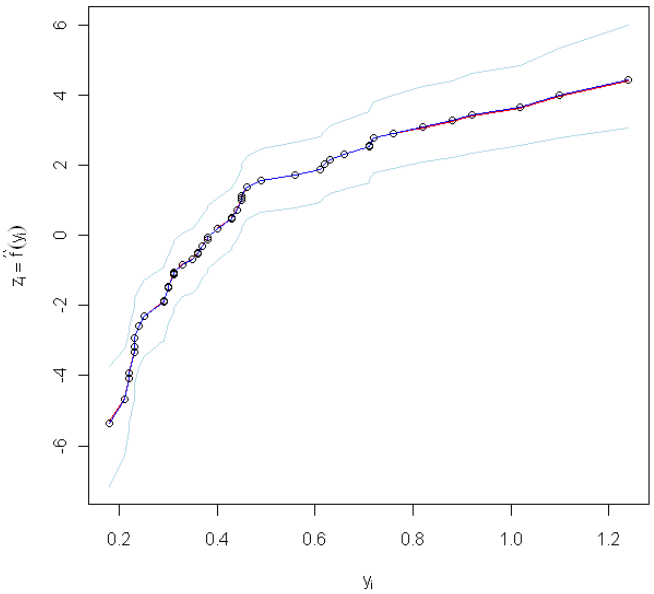


Figure 5.9: MCMC Estimated Response Transformation: Model (5.11)

5.4 Conclusion

In this chapter we developed and demonstrated an MCMC approach to the transformation of response data. This is a nonlinear application of MCMC which enables the assumptions of linear models to be satisfied in cases where the data are amenable to transformation. The transformation was chosen to maximise conformity to the assumptions of the general linear model, and thereby broaden the scope of its use to include data which meet the requirements of the model for some scale other than that on which the data were collected. Thus, this is a nonlinear application filling a niche between linear models and the parametric nonlinear models of the previous chapter.

We have shown that the method performs favourably compared to the well known Box–Cox transformation method, by re-examining an example put forward in the original paper describing their technique (Box and Cox, 1964). Our method provided improved significance of the experimental factors relative to the Box–Cox method. This attests its ability to enhance the detectability of differences between factors which may be confounded in the observations on the original scale of the data. Importantly, our method also reduced the systematic variation observed in the residuals of the fitted model, ensuring that the criteria for the general linear model are more nearly met, and allowing estimation and inference to proceed within that framework without bias.

The method determines the transformation such that the transformed data fit an assumed linear model. The approach offers the advantage that the specified model and transformation of the response data are estimated jointly. Therefore the transformation ensures that the residual component of the model $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and that all subsequent inference appropriately reflects the uncertainty in the choice of transformation.

Chapter 6

Monotonic Additive Models

6.1 Introduction

The general linear model assumes that the response can be expressed as a linear combination of the covariate data. Yet many regression situations feature covariates which have nonlinear relationships with the response. It is therefore useful to have strategies which allow relaxation of the linearity criterion specified in §4.2. In Chapter 4 we pursued such nonlinear relationships by investigating models with parametric means. We considered both many-to-one and one-to-many relationships between model functions and data and discovered that some model functions seemed better able to represent particular data than others. This suggests that the choice of any *particular* parametric form may restrict the expression of nuance in the response – covariate relationship. In particular, this is likely to be true in the multiple regression context, where any given model function is unlikely to represent all such relationships equally well. Therefore, it can be instructive to explore the relations between the response and covariates using nonparametric methods.

In Chapter 5 we determined monotonic nonlinear functions of the response, chosen specifically to maximise conformity with the general linear model. In this chapter we adapt the method to provide estimates of the functional relationship between the response and individual covariates. This allows us to construct models in which these functions are combined in an additive fashion; that is, models where the response is viewed as the sum of transformations of the covariates. Because our method preserves the rank order of the data, these are *Monotonic Additive Models*.

This modelling strategy is more flexible than either the general linear model or parametric nonlinear regression, and allows relationships between variables to be explored in a context less inhibited by assumptions. Yet it still produces models which are readily interpretable. Individual functional relationships provide a sense of the marginal relationship between predictors and the response.

While conceptually similar to the general class of additive models (Ezekiel, 1924; Friedman and Stuetzle, 1981; Stone, 1985; Hastie and Tibshirani, 1990; Simonoff, 1996; Shively et al., 1999; Ruppert et al., 2003), in that the response is modelled as the sum of functions of the independent variables, our technique bears no common heritage or other resemblance to those methods. Indeed, our approach differs in some important respects. There is no concern regarding how to choose an appropriate representation for the functional relationships employed. This avoids many of the decisions inherent in other nonparametric approaches to the exploration of covariate relationships. There is no need, for example, to choose between adopting particular choices of basis, knot sequences and smoothing penalty structures. Importantly, it also avoids problems of inference arising from uncertainty in these choices. Finally, our approach offers a distinct advantage in situations where it is reasonable to assume, or required to enforce, that functions of the covariates should be monotonic. There is no need to impose additional constraints to obtain a reasonable fit. This feature suggests the method as a natural fit to many data arising in applied disciplines.

6.2 Method Details

The modelling approach is similar to that presented in the previous chapter, except that the focus has shifted to estimating functions of covariates, rather than responses. As previously seen, the method is considerably easier to state in the absence of ties, so we initially assume that the k-th covariate $\boldsymbol{x_k} = (x_{k1}, x_{k2}, \ldots, x_{kn})$ is sorted such that $x_{ki} < x_{kj}$ when $i < j$, and seek to fit the model

$$\boldsymbol{y} = \beta_0 + \sum \beta_k f_k(\boldsymbol{x_k}) + \boldsymbol{\varepsilon} \tag{6.1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ are coefficients, and the f_k, are strictly monotonic transformations representing the marginal relationship between the kth covariate and the response.

As in Chapter 5, we choose priors for each f_k that place no further constraints on $\boldsymbol{z}_k$ other than the ordering imposed by the monotonicity of f_k and the ordering of x_k. And, as previously, rather than determine the transformation f_k directly we estimate the vector z_k of transformed covariates

$$\begin{aligned} &\boldsymbol{z}_k = f_k(x_{ki}) \qquad k = 2, \ldots, p, \quad i = 1, \ldots, n \\ &\boldsymbol{z}_k \sim \mathcal{N}(\mu_k, \sigma^2 \boldsymbol{I}) \end{aligned} \tag{6.2}$$

where μ_k is the mean for the kth covariate transformation. That is, in (5.1) we considered the transformation of the response $\boldsymbol{z} = f(\boldsymbol{y}), z_i \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$, and now we consider z_k as a function of the kth covariate, $\boldsymbol{z}_k = f(\boldsymbol{x_k}), z_k \sim \mathcal{N}(\mu_k, \sigma^2 \boldsymbol{I})$. The vectors $\boldsymbol{z}_k$ are contained in the matrix $\boldsymbol{Z}$, whose first column is the vector $\mathbf{1}$ and whose remaining columns $\boldsymbol{Z}_k$, $k = 2, \ldots, p$ contain the elements $\boldsymbol{z}_{ki}$, $i = 1, \ldots, n$, which determine the estimated transformation of the kth covariate.

To construct samples from the joint posterior by Gibbs sampling we require the conditional distributions of the z_k, β and σ. The conditional distributions of the individual z_i (with k suppressed) are analogous to those presented in (5.2)

$$\begin{aligned} z_1 \,|\, z_2, \beta &\sim \mathrm{N}\left((\mu_k)_1, \sigma^2\right) \mathrm{I}(-\infty, z_2) \\ z_i \,|\, z_{i-1}, z_{i+1}, \beta &\sim \mathrm{N}\left((\mu_k)_i, \sigma^2\right) \mathrm{I}(z_{i-1}, z_{i+1}) \qquad 1 < i < n \\ z_n \,|\, z_{n-1}, \beta &\sim \mathrm{N}\left((\mu_k)_n, \sigma^2\right) \mathrm{I}(z_{n-1}, \infty) \end{aligned} \tag{6.3}$$

where $(\mu_k)_i$ is the i-th component of $\boldsymbol{\mu}_k$, and $\mathrm{N}(\mu_k, \sigma^2)\,\mathrm{I}(a, b)$ denotes the $\mathrm{N}(\mu_k, \sigma^2)$ distribution truncated to the open interval (a, b). The conditional distributions in (6.3) extend to cases where the covariates include ties in a manner analogous to (5.4), (5.5) and (5.6).

Finally, since $p(\beta \,|\, z, \tau, y) = p(\beta \,|\, z, \tau)$

$$\beta \,|\, z, \tau \sim \mathrm{N}\left((X^T X)^{-1} X^T z, \tau (X^T X)^{-1}\right). \tag{6.4}$$

Together the relations (6.2) – (6.4) define a Gibbs sampling scheme for β, z and σ, allowing us to estimate the components of the model (6.1).

6.3 Example: Simulated Data

To illustrate the process we begin with a simple example using simulated data.

6.3.1 Data Generation

The functions

$$f_1 = \frac{x_1^3}{3}, \tag{6.5}$$

and

$$f_2 = \arctan(x_2), \tag{6.6}$$

were used with $n = 100$ values drawn as uniform random deviates, $x_1 \sim \mathcal{U}(0, 2)$ and $x_2 \sim \mathcal{U}(-2, 2)$ to generate a response $\boldsymbol{y}$ as the sum of these functions with a normally distributed error component

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + \varepsilon_i, \tag{6.7}$$

with $\varepsilon_i \sim \mathcal{N}(0, \frac{1}{10})$. The data are shown against the respective covariates in Figures 6.1 and 6.2. It is obvious that the data do not have distributions which are conditionally normal. In Figure 6.1 the data tend to concentrate away from their mean

at the margins of their distribution, while in Figure 6.2 the data display a tendency to cluster along the lower boundary of the distribution and exhibit positive skew.

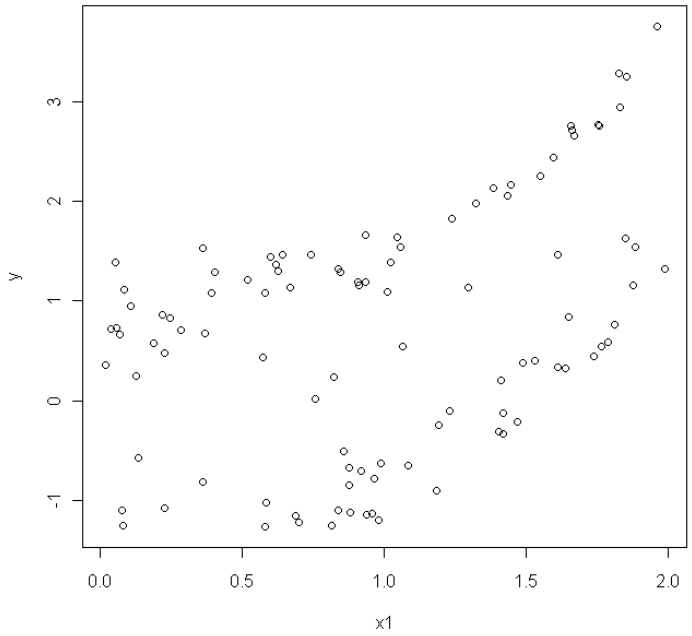


Figure 6.1: Simulated Data against Covariate x_1

6.3.2 Model Fitting

We used the Gibbs sampler described in §6.2 to fit the model

$$y_i = \beta_0 + \beta_1 f_1(x_{1i}) + \beta_2 f_2(x_{2i}) + \varepsilon_i \tag{6.8}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are parameters and $\boldsymbol{f} = (f_1, f_2)$ are functions to be estimated from the data. Here two chains were run for $k = 2000$ iterations and thinned to retain every 50th simulated value. After checking convergence diagnostics and visually inspecting the chain traces, the initial 200 values were discarded from each chain leaving 3600 posterior samples from which to calculate summary statistics.

6.3.3 Results

The parameter estimates for the model coefficients $\hat{\boldsymbol{\beta}}$ are provided in Table 6.1.

The estimates for functions f_1 and f_2 are shown against the covariates x_1 and x_2 in Figures 6.3 and 6.4. In each case the mean of the the estimated transformation

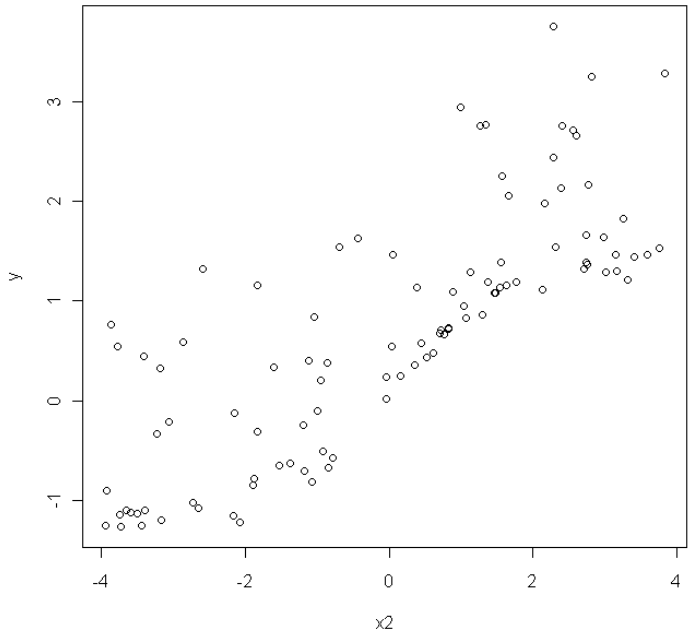


Figure 6.2: Simulated Data against Covariate x_2

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
mean	0.6740	0.9606	0.9777
2.5%	0.4187	0.9006	0.9371
50%	0.6668	0.9609	0.9776
97.5%	0.9263	1.0193	1.0161

Table 6.1: Parameter Estimates: Monotonic Additive Model (6.8)

is shown in blue, overlaid on the median estimate in red. The almost perfect correspondence between these estimates indicates symmetry in the posterior. Quantiles for the 2.5% and 97.5% estimates are shown in light blue, and the covariate values are shown as a rug plot along the horizontal axis. Finally, the original functions used to generate the data (6.5) and (6.6) provided as a reference, and are shown as the grey dashed lines in Figures 6.3 and 6.4 respectively.

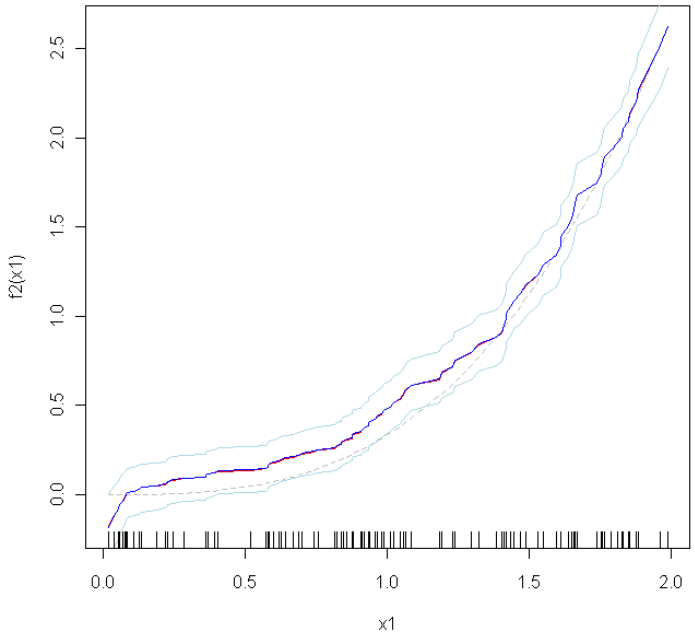


Figure 6.3: Estimated Monotonic Function f_1

Figure 6.3 shows a reasonable correspondence between the data generating function (6.5) and its estimate $f_1(x_1)$, with almost all of the original function falling within the limits of the 95% credible interval bands shown. The figure indicates that the estimate $f_1(x_1)$ provides accurate prediction for the upper values of x_1, but tends to over estimate the lower range for all but the smallest values. Figure 6.4 shows the estimate $f_2(x_2)$ to be in very close agreement with the data generating function (6.6), and is wholly contained within the 95% credible interval bands.

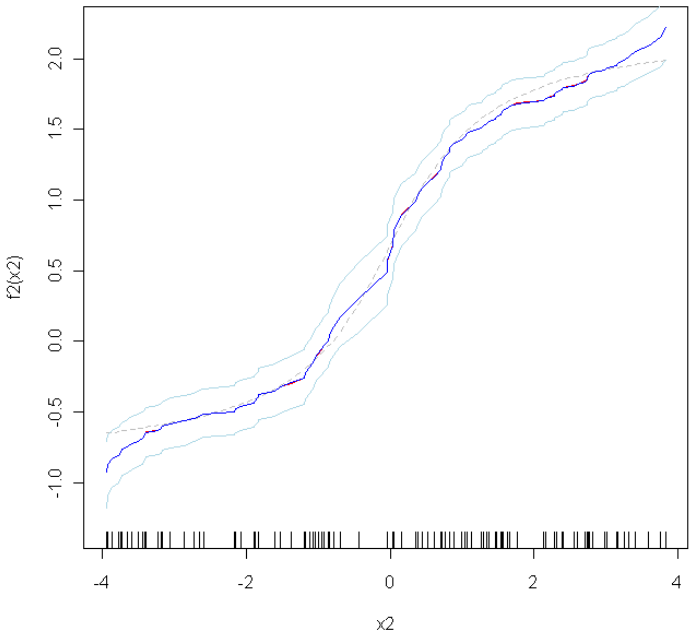


Figure 6.4: Estimated Monotonic Function f_2

6.4 Example: Black Cherry Trees

Atkinson (1985) considers a series of measurements made on a sample of felled black cherry trees, relating timber volume to the girth and height of the trees. Interest lies in predicting the volume of timber in unfelled trees based on easily measured metrics with a view to estimating the economic value of a forest area. As girth is more easily and accurately measured than height on unfelled trees, a model based on girth alone is considered preferable.

The data are provided in Table 6.2, and are also available in the `R` statistical environment, as the data set `trees`. Volume is measured in cubic feet, girth in inches (taken at 4 feet 6 inches; ∼1370 mm, above ground level), and height in feet (Ryan et al., 1976). Scatter plots of the response and covariates are plotted in Figures 6.5 and 6.6. Clearly, there is a strong relationship between tree girth and timber volume, though the relationship between tree height and volume is less clear. It seems reasonable to expect that volume is a monotonically increasing function of both girth and height.

6.4.1 Model Fitting

We fit the model

#	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21.0
13	11.4	76	21.4
14	11.7	69	21.3
15	12.0	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14.0	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16.0	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4
27	17.5	82	55.7
28	17.9	80	58.3
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

Table 6.2: Cherry Tree Data

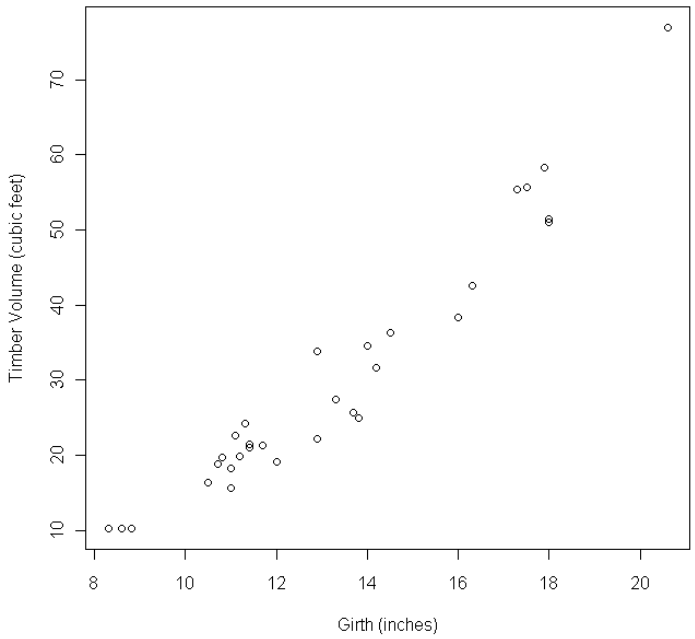


Figure 6.5: Black Cherry Trees: Timber Volume by Tree Girth

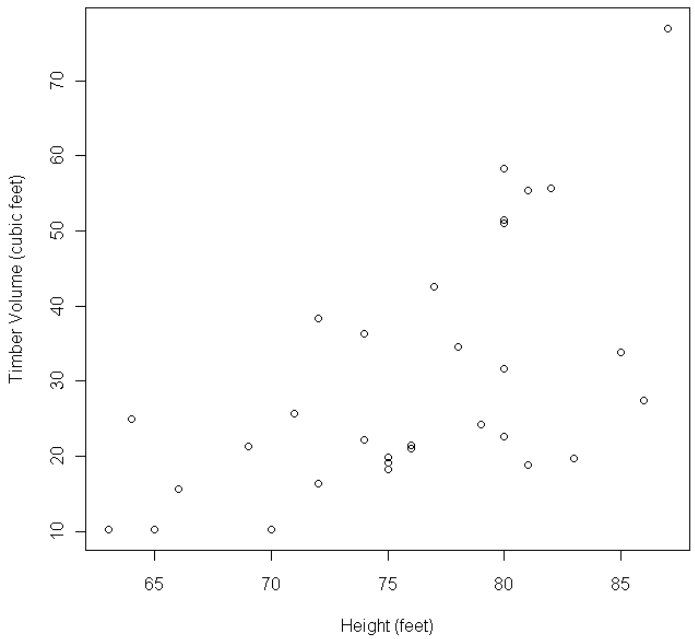


Figure 6.6: Black Cherry Trees: Timber Volume by Tree Height

$$y_i = \beta_0 + \beta_1 f_1(x_{1i}) + \beta_2 f_2(x_{2i}) + \varepsilon_i \tag{6.9}$$

where y_i is the timber volume x_{1i} is the girth, and x_{2i} the height, of the ith tree, $i = 1, 2, \ldots, n$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are coefficients, and $\boldsymbol{f} = f_1, f_2$ are transformations, to be estimated.

Two chains were run for $k = 2000$ iterations and thinned to retain every 50th simulated value. After checking convergence diagnostics and visually inspecting the chain traces, the initial 200 values were discarded from each chain leaving 3600 posterior samples from which to calculate summary statistics.

6.4.2 Results

The estimates for the model coefficients and associated 95% credible interval limits are provided in Table 6.3.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
mean	29.6550	0.9691	0.7075
2.5%	24.142	0.8635	0.3371
50%	29.592	0.9684	0.7000
97.5%	35.842	1.0746	1.1054

Table 6.3: Parameter Estimates: Trees Model (6.9)

The estimated transformations for tree girth and tree height are shown in Figures 6.7 and 6.8. In each case the mean (blue line) and median (red line) estimate of the transformation is shown, along with the 95% credible interval confidence bands. As previously, the coincidence of the mean and median estimates is a sign of symmetry in the posterior.

We can get a visual impression of the model fit by plotting the marginal model mean against the covariate x_k. These features are shown for timber volume against tree girth in Figure 6.9. If the model provides an adequate fit to the data the partial residuals should behave like a random sample with zero mean. We therefore expect these points to be randomly scattered about the curve. As no systematic pattern is evident in the figure, we have a positive diagnosis of model adequacy.

Finally, a plot of the actual timber volume versus the fitted values is provided in Figure 6.10. The points in blue are the results of fitting model (6.9), and the points in red are the results of fitting a penalized regression spline model with cross-validation as described in Wood (2006). Although the two methods use very

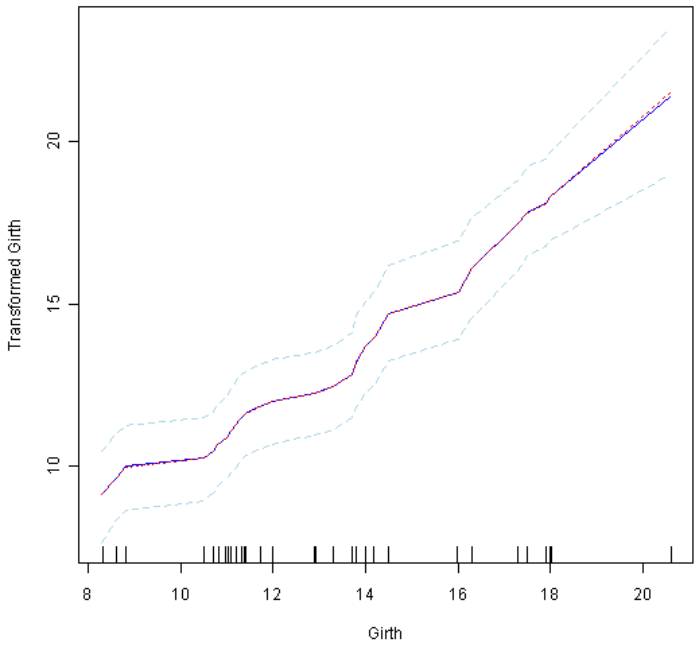


Figure 6.7: Estimated Transformation: Tree Girth

different assumptions and methodology, there is very good agreement in the final results.

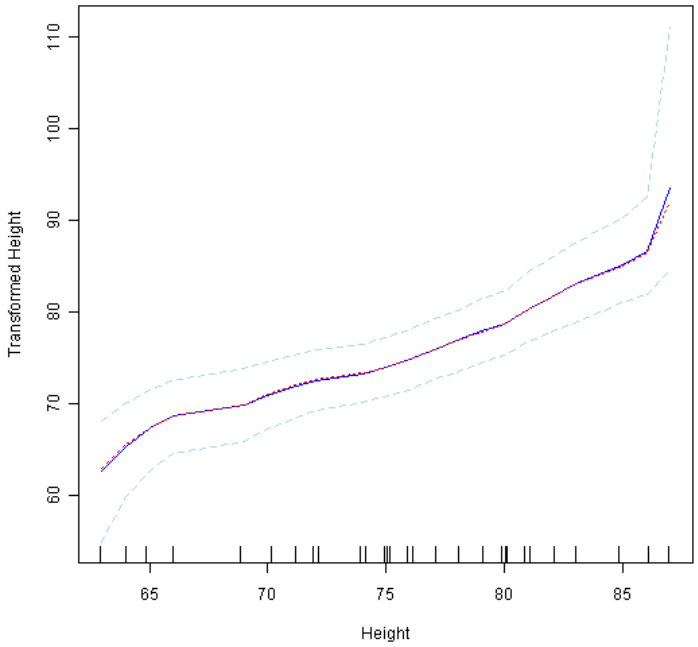


Figure 6.8: Estimated Transformation: Tree Height

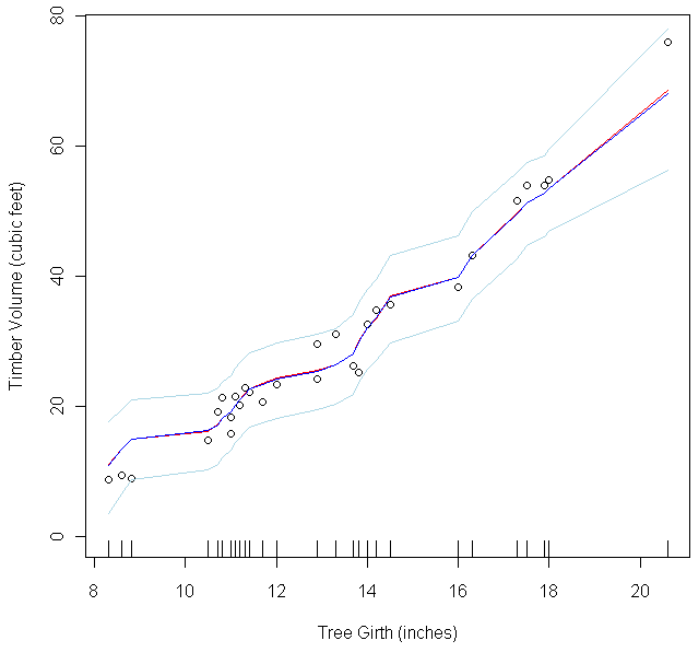


Figure 6.9: Fitted Mean: Timber Volume by Tree Girth

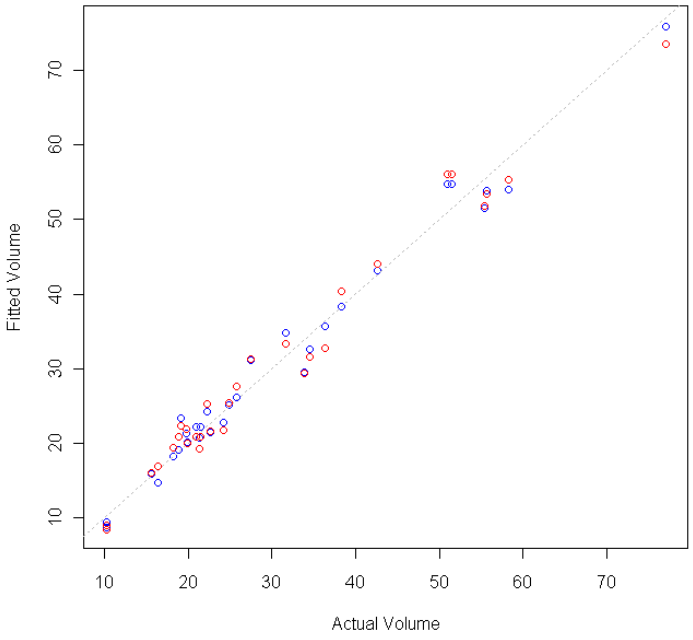


Figure 6.10: Actual Timber Volume by Fitted Volume

6.5 Example: US Temperature Data

Peixoto (1990) presents data for the average daily minimum temperature in January of 56 cites in the United States. The data are plotted in Figure 6.11. Loess smooth lines have been added to the scatterplots to aid interpretation of the relationships between the variables. We can see, for instance, that minimum temperature seems to have a linear relationship with latitude, but a nonlinear relationship with longitude.

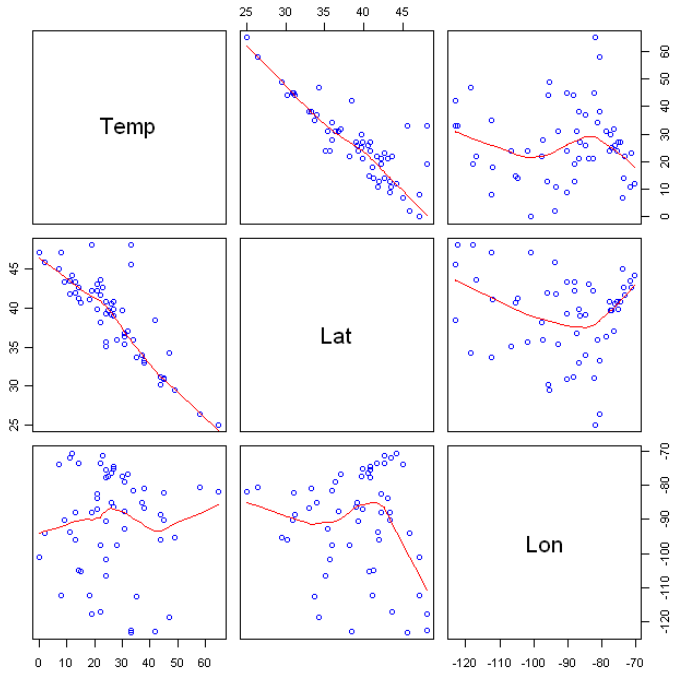


Figure 6.11: Pairwise Scatterplots: US Temperature Data

Peixoto (1990) demonstrates that an accurate model for minimum temperature is

$$\text{min.temp}_i = \beta_0 + \beta_1 \text{lat}_i + \beta_{21} \text{lon}_i + \beta_{22} \text{lon}_i^2 + \beta_{23} \text{lon}_i^3 + \varepsilon_i, \tag{6.10}$$

where $\texttt{lat}_i$ is the degrees of latitude, and $\texttt{lon}_i$ is the degrees of longitude, for the i-th city. So it seems that minimum temperature is well explained by a cubic function of longitude. As model (6.10) does not feature a term for an interaction between latitude and longitude, it is by definition an additive model. However, because the suggested relationship between temperature and longitude is not monotonic, we do not expect that the method presented in §6.2 to perform well for the longitude component of the model. Nevertheless, we pursue the exercise to illustrate the limitations of the method.

6.5.1 Model Fitting

We fit the model

$$y_i = \beta_0 + \beta_1 f_1(x_{1i}) + \beta_2 f_2(x_{2i}) + \varepsilon_i \tag{6.11}$$

where y_i is the minimum January temperature, x_{1i} is the latitude, and x_{2i} the longitude, of the ith city, $i = 1, 2, \ldots, n$, and estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and $\boldsymbol{f} = f_1, f_2$ from the data.

As with earlier examples in this chapter, two chains were run for $k = 2000$ iterations and thinned to retain every 50th simulated value. After checking convergence diagnostics and visually inspecting the chain traces, the initial 200 values were discarded from each chain leaving 3600 simulated posterior samples.

6.5.2 Results

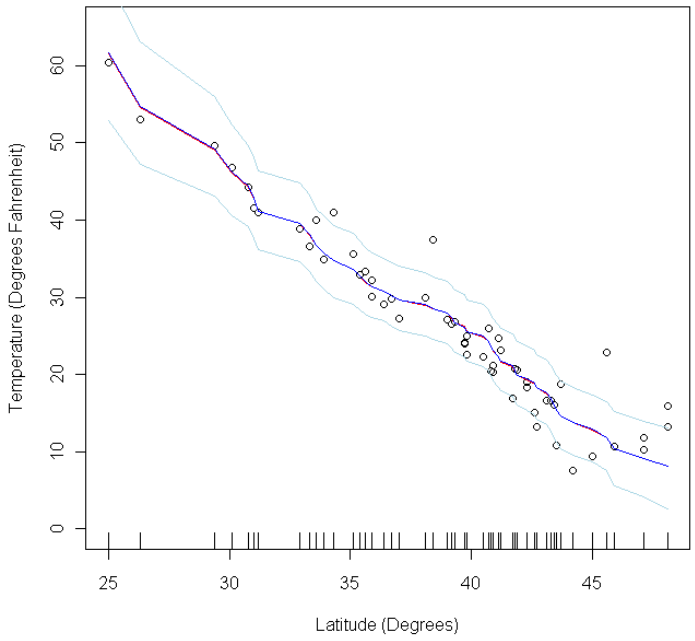


Figure 6.12: Estimated Latitude - Temperature Function: US Data

The estimates for the model coefficients and associated 95% credible interval limits are provided in Table 6.4.

The estimated transformations for latitude and longitude are shown in Figures 6.12 and 6.13. As previously, in each case the mean (blue line) and median (red line)

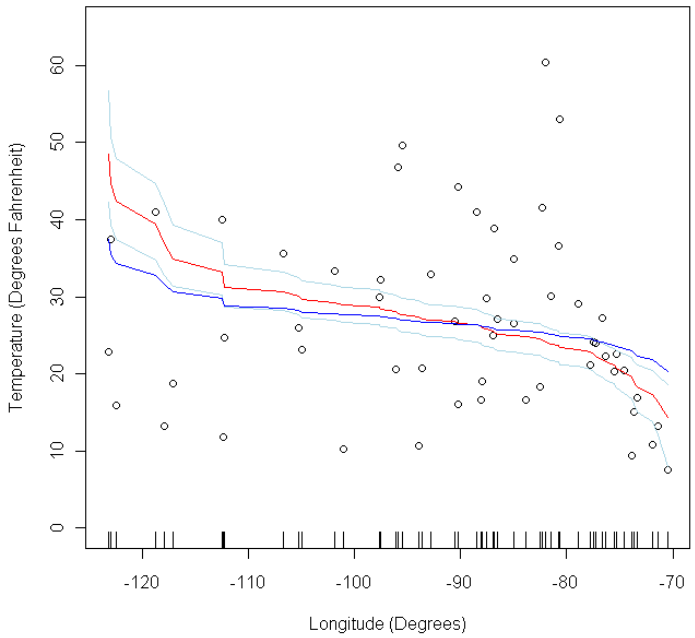


Figure 6.13: Estimated Longitude - Temperature Function: US Data

estimate of the transformation is shown, along with the 95% credible interval confidence bands.

We can see that the model appears to have successfully dealt with the linear relationship between temperature and latitude, but that the estimated transformation for longitude appears problematic, indicated by the divergence between the mean and median estimates. Because our method enforces strict monotonicity, it is unable to represent the cubic relationship between temperature and longitude.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
mean	26.494	-0.8329	-0.4130
2.5%	22.211	-0.9923	-0.6770
50%	26.560	-0.8304	-0.4114
97.5%	30.450	-0.6904	-0.1554

Table 6.4: Parameter Estimates: US Temperature Model

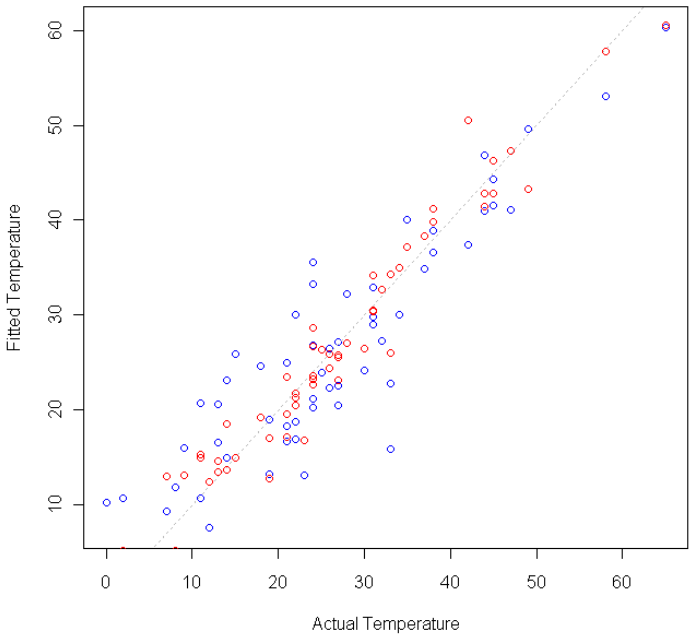


Figure 6.14: Fitted vs Actual Values: US Temperature Data

However, the suggested transformation is not wholly inappropriate. Figure 6.14 shows the fitted versus actual values for models (6.10) and (6.11) in red and blue respectively. Both models seem to provide a reasonable estimate for the mean, but the linear model (6.10) displays noticeably better fit, evident as less scatter of the red points around the line $y = x$.

6.6 Conclusion

In this chapter we developed and demonstrated a method for fitting Monotonic Additive Models. While conceptually similar to the general class of additive models considered by Hastie and Tibshirani (1990), our technique bears no common heritage or other resemblance to those methods. There is no need, for example, to choose between adopting particular choices of basis, knot sequences and smoothing penalty structures in fitting model by our method. Importantly, our technique also avoids problems of inference arising from uncertainty in these choices.

Monotonic Additive Models as presented in this chapter are an exploratory technique, allowing one to identify transformations which may improve model fit and interpretability. As we have seen, the method is not well suited for the estimation of marginal relationships where turning points are an important feature. But this limitation is also a strength, in that the approach has natural advantages over more

flexible techniques in situations where the assumption of monotonicity is reasonable. Such data occur frequently in a wide range of applied disciplines.

Chapter 7

Estimating Correlations

7.1 Introduction

Each of the earlier chapters have considered the application of a method based upon a univariate response with observations assumed independent, as specified in criterion two of §4.2. However, one frequently encounters data for which independence is not a reasonable assumption. Perhaps the most common example is that of multiple measurements on the same experimental unit. The repeated measures may occur when multiple attributes of the unit are measured at a single observation period, or when measures of the same attribute are repeated throughout time. Observations within units are likely to be more similar than observations between units, so the assumption of independence is no longer reasonable. Thus in cases where we may be interested in conducting a multivariate analysis of several variables, or a longitudinal analysis of a single variable, for example in a mixed-effects framework, we will also be interested in estimating correlations between estimands.

A reliable method for estimating correlations is fundamental to extending the methods presented in previous chapters into multivariate and mixed-effects contexts. Because the underlying methodology for each of those methods is Bayesian MCMC, it is natural that we develop a technique consistent with that framework. That we are able to do so reaffirms the versatility of the methodology.

There is an extensive literature describing covariance estimation, with the following references providing a sense of the range of approaches, as well as sources of many other articles of interest. Chib and Greenberg (1998) considered the estimation of covariance in multivariate probit models. Daniels and Kass (1999) develop a Bayesian treatment for estimating covariance matrices with small samples in a hierarchical modelling framework. Barnard et al. (2000) discuss decompositions of the covariance matrix, and illustrate this with an application which parallels developments in the present chapter, though the treatment is quite different. Daniels and Kass (2001) extend their earlier work to consider robustness of the estimators and efficiency in dealing with large matrices. Browne (2002) undertakes a comparative analysis between various generic MCMC approaches to models which include con-

strained covariance matrix estimation. Daniels and Pourahamdi (2002) develop a framework for developing conditionally conjugate prior distributions for covariance matrices, and emphasise the importance of these in longitudinal models.

In this chapter we develop MCMC techniques for estimating correlations using variants of the Gibbs sampler. Difficulties arise from the constraint that individual estimands must satisfy the positive definite requirement of the correlation matrix as a whole. The principal method we describe exploits this constraint and implements a rejection sampling strategy based upon carefully selected Gamma distributions. Another numerical method is provided for use in cases where this approach can be demonstrated to operate with low efficiency. Both methods are tested against simulated data to illustrate their relative merits.

7.2 Sampling Strategy

7.2.1 Introduction

Given observations $y_1, \ldots, y_n$ from a multivariate normal distribution

$$Y_i \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V})$$

we wish to estimate v_{ij}, the ij^{th} off diagonal element of the covariance matrix V, by Gibbs sampling.

Write the elements of $\boldsymbol{V}$ as $v_{ij} = v_{ji} = x$. Assuming a uniform prior, the posterior is proportional to

$$|\boldsymbol{V}(x)|^{-\frac{n}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{V}(x)^{-1}\boldsymbol{S})} \tag{7.1}$$

where $\boldsymbol{S} = \boldsymbol{R}^T\boldsymbol{R}$ is formed from the matrix of observation residuals

$$\boldsymbol{R} = \left[Y_1 - \boldsymbol{\mu} \,|\, Y_2 - \boldsymbol{\mu} \,|\, \ldots \,|\, Y_n - \boldsymbol{\mu}\right]^T.$$

In order to apply the Gibbs sampler to this problem we must find a way to draw random deviates $\boldsymbol{X}$ with density proportional to (7.1).

7.2.2 Implications of the Positive Definite Constraint

One of the difficulties we face in constructing a sampling distribution is that we wish to produce estimates for individual matrix entries without violating the constraints imposed by the nature of the covariance matrix as a whole. While these conditions add extra complexity to the problem, we can exploit properties of the constraints themselves to help focus our sampling regime.

Since $\boldsymbol{V}(x)$ is a covariance matrix, it is symmetrical and positive definite, and the determinant $|\boldsymbol{V}(x)| > 0$. Consider the expansion of the determinant as a sum of permutation products

$$|\boldsymbol{V}(x)| = \sum_{p} (-1)^k \prod \boldsymbol{x_{ij_p}}$$

where the matrix $\boldsymbol{V}(x)$ has elements x_{ij}, with row indices $i = 1, 2, \ldots, m$, and column indices $j = 1, 2, \ldots, m$. The determinant for this matrix may be written as the sum of the $p = m!$ products $\boldsymbol{x_{ij_p}} = (x_{1j_1} x_{2j_2} \ldots x_{mj_m})$, where $(j_1, j_2, \ldots, j_m)$ is a permutation of the set of index integers $\{1, 2, \ldots, m\}$. An *inversion* is defined as occuring for each position in a permutation sequence where a larger integer precedes a smaller one $j_p > j_q$, $p = 1, 2, \ldots, m$, and $q = p + 1, p + 2, \ldots, m$, with k equal to the sum of inversions in the sequence.

Since each $x_{ij} = x_{ji}$ can occur twice in each $\boldsymbol{x_{ij_p}}$, it is clear that the determinant $|\boldsymbol{V}(x)|$ is a quadratic function of the elements x and can be written

$$|\boldsymbol{V}(x)| = a_0 + a_1 x + a_2 x^2$$

for some a_0, a_1 and a_2. As $\boldsymbol{V}$ can only be positive definite on a finite real interval (x_c, x_d), $a_2 < 0$, and $|\boldsymbol{V}(x)|$ always has two real roots. Hence we can write

$$|\boldsymbol{V}(x)| = a(x - x_c)(x_d - x)$$

for some $a > 0$, and we must determine the interval (x_c, x) to appropriately constrain our sampling distribution.

7.2.3 Determining the Sampling Interval

The quadratic $|\boldsymbol{V}(x)|$ can be readily calculated from the Schur decomposition of a determinant

$$|\boldsymbol{A}| = |\boldsymbol{A}_{-j,-j}||A_{j,j} - \boldsymbol{A}_{j,-j} \boldsymbol{A}^{-1}_{-j,-j} \boldsymbol{A}_{-j,j}|$$

where negative subscripts denote the deletion of the corresponding row or column. We find that

$$|\boldsymbol{V}(x)| = |\boldsymbol{V}_{-i,-i}|(\, (V_{j,j} - \boldsymbol{u}^T_{-i} \boldsymbol{W}_{-i,-i} \boldsymbol{u}_{-i}) - (2\boldsymbol{W}_{-i,i} \boldsymbol{u}_{-i})x - (W_{i,i})x^2 \,) \qquad (7.2)$$

where the matrix $\boldsymbol{W} = \boldsymbol{V}^{-1}_{-j,-j}$ and the vector $\boldsymbol{u} = \boldsymbol{V}_{-j,j}$, noting that this requires $i < j$ for the indexing to remain consistent.

Given this quadratic we can determine the interval on which $\boldsymbol{V}(x)$ remains positive definite, (x_c, x_d). We now turn to the problem of constructing an appropriate sampling regime to exploit this information.

7.2.4 Refactoring the Trace

Applying Cramer's rule to $\boldsymbol{V}(x)^{-1}$ from the posterior distribution (7.1), it follows that the entries are quadratic in $x = x_{ij}$ (with index notation suppressed), and we can write the trace of the posterior in the form

$$\operatorname{tr}(\boldsymbol{V}(x)^{-1}S) = \frac{a_0 + a_1 x + a_2 x^2}{(x - x_c)(x_d - x)} = b_0 + \frac{2b_1}{x - x_c} + \frac{2b_2}{x_d - x} \tag{7.3}$$

for some b_0, b_1 and b_2.

7.2.5 Rejection Sampling

It follows from this form of the trace that the posterior (7.1) is proportional to

$$\begin{aligned} &|\boldsymbol{V}(x)|^{-\frac{n}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{V}(x)^{-1}\boldsymbol{S})} \\ &\quad \propto ((x - x_c)(x_d - x))^{-\frac{n}{2}} e^{-(b_1/(x-x_c) + b_2/(x_d - x))} \\ &\quad \propto \left((x - x_c)^{-\frac{n}{2}} e^{-b_1/(x-x_c)}\right) \times \left((x_d - x)^{-\frac{n}{2}} e^{-b_2/(x_d - x)}\right) \end{aligned} \tag{7.4}$$

which can be recognized as the product of an inverse gamma density in $x - x_c$ and an inverse gamma density in $x_d - x$.

Making the transformation $z = x - x_c$ in (7.4) yields

$$\left(z^{-\frac{n}{2}} e^{-b_1/z}\right) \times \left((x_d - x_c - z)^{-\frac{n}{2}} e^{-b_2 (x_d - x_c - z)^{-1}}\right)$$

So we can draw deviates with a density proportional to (7.4) by drawing z from an inverse gamma distribution

$$z^{-1} \sim \mathcal{G}\left(n/2 - 1, b_1\right) I(1/(x_d - x_c), \infty)$$

truncated above $x_d - x_c$, and then rejection sampling, retaining draws with a probability

$$K(x_d - x_c - z)^{-\frac{n}{2}} e^{-b_2/(x_d - x_c - z)}.$$

For maximal efficiency, the constant of proportionality K should be chosen as

$$K = \max_{z \in (x1, x2)} (x_d - x_c - z)^{\frac{n}{2}} e^{b_2/(x_d - x_c - z)}.$$

If instead we choose K as the global maximum, the rejection probability is

$$1 - (2b_2)^{\frac{n}{2}} (n(x_d - x_c - z))^{-\frac{n}{2}} e^{n/2 - b_2/(x_d - x_c - z)}$$

Alternately, we can make the transformation $z = x_d - x$ in (7.4) to form

$$\left((x_d - x_c - z)^{-\frac{n}{2}} e^{-b_1/(x_d - x_c - z)} \right) \times \left(z^{-\frac{n}{2}} e^{-b_2/z} \right).$$

Again we can draw deviates with a density proportional to (7.4) by drawing z from an inverse gamma distribution

$$z^{-1} \sim \mathcal{G}\left(n/2 - 1, b_2\right) I(1/(x_d - x_c), \infty)$$

truncated above $x_d - x_c$, and then rejecting sampling, retaining draws with probability

$$K(x_d - x_c - z)^{-\frac{n}{2}} e^{-b_1/(x_d - x_c - z)}.$$

Again, if we assume the global maximum occurs within the interval (x_c, x_d), the most efficient rejection probability is

$$1 - (2b_1)^{\frac{n}{2}} (n(x_d - x_c - z))^{-\frac{n}{2}} e^{n/2 - b_1 (x_d - x_c - z)^{-1}}.$$

To minimize the number of rejections, we should sample from the most concentrated of the two inverse gamma densities, and reject from the most diffuse.

7.3 Implementation

7.3.1 Example: Estimating a Single Variance Matrix Element

Consider the estimation of a single matrix element $x = v_{ij}$ from the $m \times m$ correlation matrix

$$\boldsymbol{V}(x) = \begin{pmatrix} 1.00000 & 0.53847 & -0.96159 \\ 0.53847 & 1.00000 & -0.73212 \\ -0.96159 & -0.73212 & 1.00000 \end{pmatrix}$$

under the assumptions that we have a sample of n observations with correlation structure $\boldsymbol{V}(x)$, and that each of the remaining off-diagonal elements are also known.

Choosing $n = 10$ we formed a matrix of observation residuals

$$\boldsymbol{R}_{10} = \boldsymbol{B}\boldsymbol{L}^T$$

where $\boldsymbol{B}$ is an $n \times m$ matrix of standard normal random deviates, $b_{ij} \sim \mathcal{N}(0,1)$, and $\boldsymbol{L}^T$ is the upper triangular Cholesky decomposition of $\boldsymbol{V}(x)$. The true correlation structure of the observations is therefore known to be $\boldsymbol{V}(x)$, and the sample correlations for $\boldsymbol{R}_{10}$ were

$$\begin{pmatrix} 1.0000 & 0.5488 & -0.9684 \\ 0.5488 & 1.0000 & -0.7349 \\ -0.9684 & -0.7349 & 1.0000 \end{pmatrix}.$$

Letting $i = 1$ and $j = 2$, we wish to estimate the value $x = v_{12} = 0.53847$.

Implementation

Noting that $i < j$, the Schur decomposition was obtained as described in §7.2.3, and solved for $(x_c, x_d) \approx (0.51702, 0.89098)$, the domain from which v_{12} may be sampled without violating the constraint that $\boldsymbol{V}(x)$ remain non-negative definite. The trace was re-factored as in (7.3), and values $b_1 \approx 0.08878$ and $b_2 \approx 2.39899$ found by interpolation. With these elements in place we turn to constructing samples for $x = v_{12}$.

To generate samples from (7.4):

1. If $b_1 < b_2$, take a sample of size N $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ from an Inverse Gamma distribution with parameters $\alpha = n/2 - 1$ and $\beta = b_2$, truncated below $c_0 = \frac{1}{x_d - x_c}$. (Conversely, if $b_2 < b_1$, take $\boldsymbol{z} \sim \mathcal{IG}(n/2 - 1, b_1)$, truncated below c_0).

2. Substitute z_i, $i = 1, 2, \ldots, N$ into the log posterior, and implement a rejection filter such that individual z_i values are rejected if the resultant log posterior value falls lower than a uniform random deviate $u \sim \mathcal{U}(0,1)$.

3. Translate the remaining z_i onto $(x_c,\ x_d)$ by addition to x_c $(b_1 < b_2)$ or subtraction from x_d $(b_2 < b_1)$ to produce x.

Taking $N = 5000$ and noting that $b_1 < b_2$, we drew

$$\boldsymbol{z}^{-1} \sim \mathcal{G}(\alpha, \beta)$$

truncated below $c = \frac{1}{x_2 - x_1} \approx 0.37396$, where $\alpha = n/2 - 1 = 4$ and $\beta = b_2 \approx 2.39899$.

The log posterior was evaluated at z_i, $i = 1, 2, \ldots, N$, subject to a rejection filter and translated onto $(x_c, x_d) \approx (0.51702, 0.89098)$ in the straightforward manner described in points 2 and 3 above.

Results

A histogram of the resultant samples is offered in Figure 7.1, with the true posterior density overlaid in blue. The target value $x = v_{12} \approx 0.5385$ is shown by the vertical perforated line in red. Table 7.1 shows summary statistics for the sampling distribution.

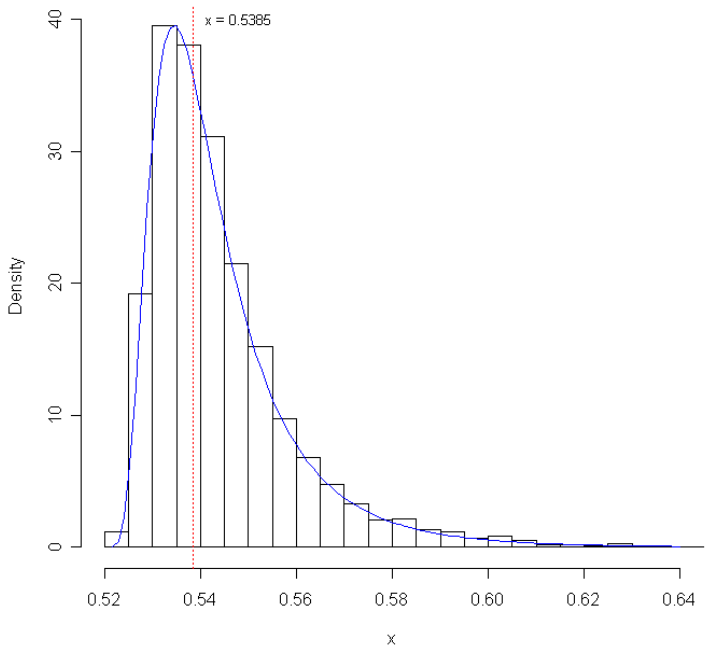


Figure 7.1: Bi-Gamma Sampling Distribution for $x = v_{12}$

		MCMC		*Quantiles*		
	Target	*Mean*	σ	50%	2.5%	97.5%
v_{12}	0.53847	0.54427	0.01552	0.54028	0.52701	0.58713

Table 7.1: Bi-Gamma Sampling Distribution Summary

Discussion

As can be seen from the results, the mean and 97.5% quantile reflect the strong positive skew of the sampling distribution, yet both the mean and median provide reasonable estimates and the standard deviation is attractively small. Most importantly, the histogram provides an accurate picture of the true posterior. While reassuring, this is hardly surprising given the information assumed known in this case.

7.3.2 Rejection Rates

As a result of the rejection filter implemented in the example of §7.3.1, 4761 x values were generated from an initial sample of $N = 5000$ z values, providing a sampling efficiency of slightly more than 95%. While this level of efficiency would seem to be acceptable, rejection rates allude to a potential practical problem with the sampling strategy. If the distribution of the log posterior occurs such that the majority of the probability mass is located below a sufficiently large negative value, many z values would be rejected and constructing a sample for x could become inefficient and intolerably slow.

To guard against the possibility of inefficiency through frequent rejection, an amendment to the procedure will now be proposed based on the Griddy-Gibbs sampler described by Ritter and Tanner (1992).

7.3.3 The Griddy-Gibbs Sampler

The Griddy-Gibbs technique constructs samples based on an empirically derived approximation of the conditional posterior CDF evaluated at a "grid" of predetermined points. The algorithm implemented here is an adaptation of that offered in Ritter and Tanner (1992) and is constructed as follows.

To obtain a sample value x:

1. Form an interval $(x_C, x_D) = (x_c + \varepsilon, x_d - \varepsilon)$ for some small positive number ε to avoid numerical instability at the endpoints of the positive definite sampling interval (x_c, x_d) identified in §7.2.3.
2. Construct a "grid" of M equispaced points x_i, $i = 1, 2, \ldots, M$ on (x_C, x_D).
3. Evaluate the posterior at x_i to form w_i, the relative weighted contributions of the x_i to the posterior density.
4. Obtain points on the empirical CDF as (x_i, p_i), $i = 1, 2, \ldots, M$ where p_i are the cumulative sum of $w_i / \sum w_i$.

5. Choose a standard uniform deviate $u \sim \mathcal{U}(0, 1)$.

6. Obtain x by finding the ordinate value for the intersection of u with the empirical CDF using interpolation on the line segment with endpoints (x_k, p_k) and (x_{k+1}, p_{k+1}) where $p_k \leq u \leq p_{k+1}$.

Refining the Sampling Interval

The Griddy-Gibbs procedure can be illustrated by reviewing the example of §7.3.1. Figure 7.2 shows the evaluation of the conditional posterior for v_{12} on (x_C, x_D) using $\varepsilon = 5.0 \times 10^{-8}$ and $M = 100$ to produce the unnormalised relative weights w_i on an equispaced grid. As can be seen from the figure, the density is concentrated on a relatively small portion of the gridded interval (x_C, x_D).

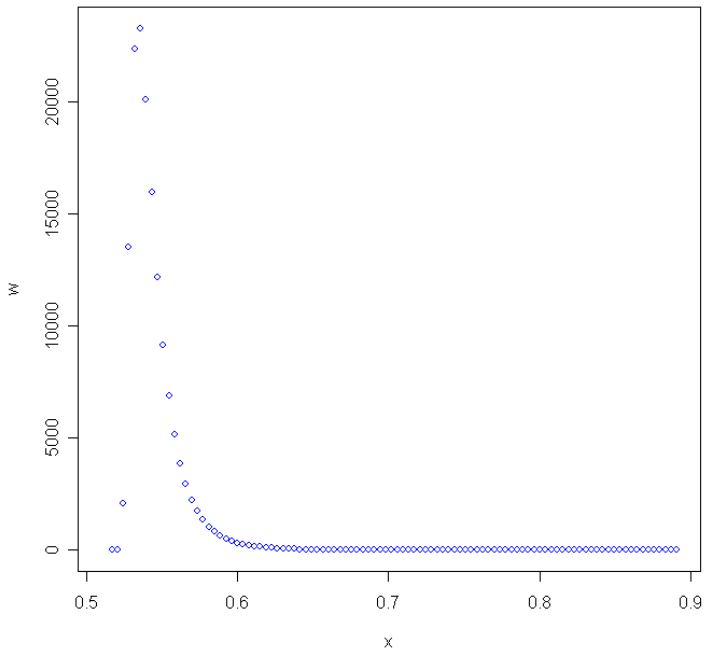


Figure 7.2: Evaluating the Posterior on a Grid

In practice, the sub-interval containing the majority of the density can be so narrow as to contain only a small number of the gridded points, thereby providing a poor empirical approximation to the CDF. Indeed, a sufficiently coarse grid spacing could conceivably produce $w_i \approx 0$, $\forall\, i = 1, 2, \ldots, M$.

To counter this possibility we further amended the Griddy-Gibbs procedure described above, introducing the following sequence of steps to replace step 2:

2a Select points (x_i, w_i) from the Highest Posterior Density sub-interval (x_c, x_d), $x_c \geq x_C$, $x_d \leq x_D$, such that $w_i \geq \delta$, for some small $\delta > 0$.

2b Form a new grid x_i, $i = 1, 2, \ldots, M$, on (x_c, x_d).

2c Evaluate the posterior at x_i to form w_i, the relative weighted contributions of the x_i to the posterior density.

Choosing $M = km$, $k \in \mathbb{Z}, m \in \mathbb{Z}$, steps 2a–2c can be iterated until the sub-interval (x_c, x_d) contains at least m points, ensuring the desired minimum level of smoothness in the Highest Posterior Density region.

From these Highest Posterior Density w_i the CDF of the conditional posterior density can be approximated as described in Step 3 above.

Implementation

Here we used $k = m = 10$, and $\delta = 5.0 \times 10^{-4}$ to produce Figure 7.3. Note the reduced range of the ordinate axis from that shown in Figure 7.2. The resultant HPD interval $(x_c, x_d) \approx (0.52079, 0.66143)$ in contrast to $(x_C, x_D) \approx (0.51702, 0.89098)$.

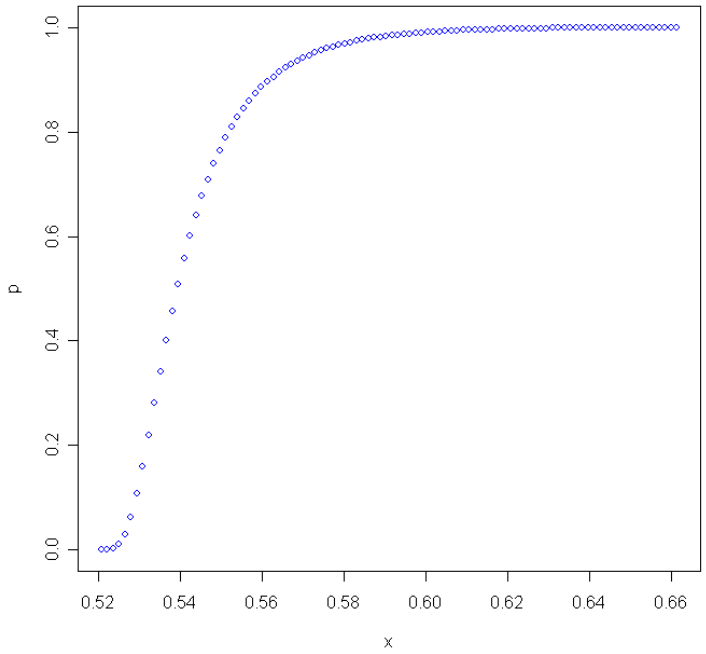


Figure 7.3: Approximating the Cumulative Distribution Function

Figure 7.4 illustrates the determination of a sample value for x. A standard uniform random deviate u is indicated by the perforated blue horizontal. The intersection of this value with the empirical CDF provides the required sample value x. Here the result has been obtained by linear interpolation on the corresponding line segment though more sophisticated approximations could be used if required.

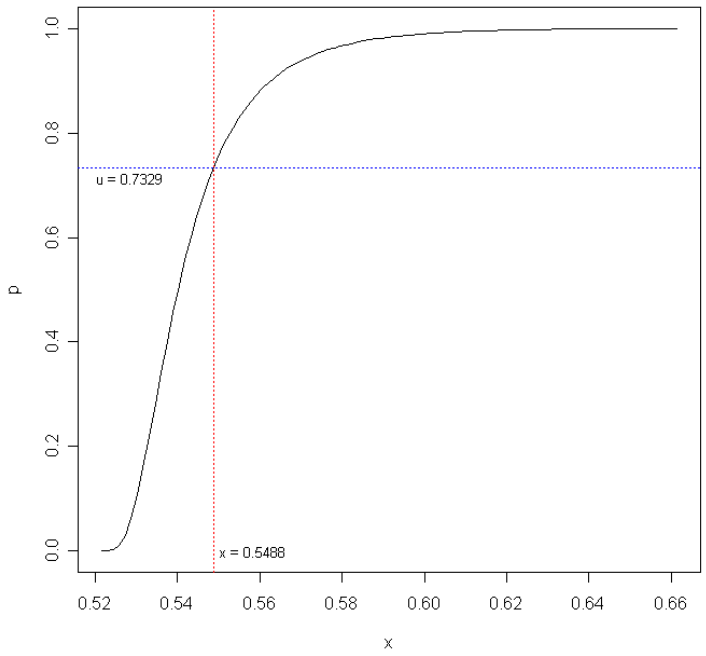


Figure 7.4: Transforming a Uniform Random Deviate via Griddy Gibbs

Results

Figure 7.5 shows the sampling distribution for $x = v_{12}$, produced by transformation of $N = 5000$ uniform random deviates via the Griddy-Gibbs procedure. Corresponding summary statistics are provided in Table 7.2. The similarity with the sampling distribution obtained via the Bi-Gamma procedure (Figure 7.1, Table 7.1) is obvious.

7.3.4 Example: Estimating Multiple Matrix Entries

We now consider the more general case where none of the off-diagonal matrix elements $x = v_{ij}$, $j > i$ are known and each must be estimated. That is, the previous example is extended to estimate the three upper diagonal elements of $\boldsymbol{V}(x)$.

		Griddy Gibbs			*Quantiles*	
	Target	*Mean*	σ	50%	2.5%	97.5%
v_{12}	0.53847	0.54445	0.01577	0.54031	0.52679	0.58629

Table 7.2: Griddy-Gibbs Sampling Distribution Summary

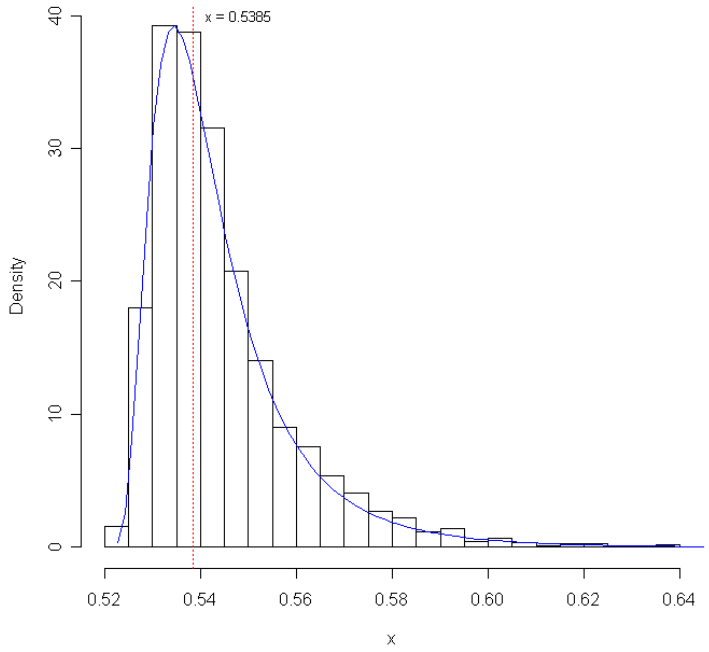


Figure 7.5: Griddy Gibbs Sampling Distribution for $x = v_{12}$

We proceed in a similar fashion to the single element case, except that at each of the N iterations a single value of x is estimated for each conditional v_{ij}. That is, at each iteration we loop though the upper diagonal elements of $\boldsymbol{V}(x)$ and estimate each in turn conditional on the value of the others by evaluating the steps detailed in §7.3.1. Using the notation of the Gibbs Sampler introduced in §3.3, we have

$$\boldsymbol{x}^t = (x_1^t, x_2^t, x_3^t) = (v_{12}^t, v_{13}^t, v_{23}^t),$$

where the superscripts indicate the iteration count t. From this we generate

$$\begin{aligned} x_1^{t+1} &\sim f_1(x_1 \mid x_2^t, x_3^t) \\ x_2^{t+1} &\sim f_2(x_2 \mid x_1^{t+1}, x_3^t) \\ x_3^{t+1} &\sim f_3(x_3 \mid x_1^{t+1}, x_2^{t+1}) \end{aligned}$$

where f_n, $n = 1, 2, 3$, are the full conditional distributions of the posterior.

Implementation

We retained the set of observation residuals R_{10} for this example, and the matrix of sample correlations provided in §7.3.1 was used as a starting point for the estimation of $x = v_{ij}$.

Results

Table 7.3 shows the summary statistics for $N = 5000$ samples for each of the upper triangular elements $x = v_{ij}$, $j > i$ in $\boldsymbol{V}(x)$. In each case the target value is contained within the 95% credible interval for $x = v_{ij}$. The credible intervals are broader than the those obtained under the previous examples for estimation of a single matrix entry, reflecting the additional uncertainty present when each element is estimated conditional to a set of constraints which are themselves uncertain.

		MCMC		*Quantiles*		
	Target	*Mean*	σ	50%	2.5%	97.5%
v_{12}	0.53847	0.48320	0.09839	0.49186	0.26790	0.64708
v_{13}	-0.96159	-0.96429	0.01067	-0.96636	-0.97820	-0.93785
v_{23}	-0.73212	-0.67439	0.07519	-0.68672	-0.79092	-0.48889

Table 7.3: Bi-Gamma Sampling Distribution Summary $x = v_{ij}$, $n = 10$

Table 7.4 shows the results of repeating the exercise with $n = 100$ observations. Clearly the additional information in the increased available data has greatly as-

sisted the accuracy of the results.

		MCMC		*Quantiles*		
	Target	*Mean*	σ	50%	2.5%	97.5%
v_{12}	0.53847	0.53524	0.03223	0.53329	0.47758	0.59859
v_{13}	-0.96159	-0.96447	0.00365	-0.96458	-0.97136	-0.95701
v_{23}	-0.73212	-0.71998	0.02128	-0.71871	-0.75903	-0.68075

Table 7.4: Bi-Gamma Sampling Distribution Summary $x = v_{ij}$, $n = 100$

7.4 Conclusion

The methods developed in earlier chapters represent a selection of tools which allow nonlinear models to be explored using MCMC. One common feature among them is that they assume independent univariate observations. In order to extend the toolbox into multivariate contexts we require a method for estimating correlations, and two schemes were presented here. This extension greatly enhances the potential for application of these tools in more complex and realistic nonlinear models with MCMC.

In this chapter we developed and demonstrated methods for the estimation of correlation elements based on the Gibbs Sampler. Using the roots of the quadratic function associated with a correlation matrix $\boldsymbol{V}$, we identified the interval to which sampling must be constrained in order to preserve the positive definite property of the matrix. By factoring the trace of $\boldsymbol{V}$ we showed that the posterior distribution was proportional to the product of two inverse Gamma densities, and described a regime for sampling from this distribution. The successful implementation of the sampling strategy was demonstrated for a single matrix element using simulated data. Because we are able to predict cases where the scheme has potential to be inefficient, we developed a second approach based on the Griddy-Gibbs sampler. The second method is more robust at the expense of greater computational overhead. Finally, we demonstrated the ability of the scheme to successfully estimate each element from a simulated correlation matrix.

Chapter 8

Conclusion

8.1 Synopsis

Stochastic modelling has become a fundamental feature of science in the 21st century. We are now better equipped to build realistic models than at any previous time in history. Widespread availability of high-performance desktop computers has allowed computational methods to flourish. Among these, Bayesian MCMC provides a central core upon which many applications can be built.

Bayes theorem provides a consistent systematic approach to statistical analysis. This contrasts with the frequentist approach which uses a variety of methods depending on the context of the problem at hand, providing confusion to the novice user and making the tool set more difficult to master. Moreover, Bayesian methods allow direct interpretation of probability statements regarding quantities of interest. This alone is compelling reason for their use in applied disciplines, as casual users of statistics will interpret confidence intervals in this way regardless of the mechanism used to generate them. Not least, Bayesian methods have become popular because they pave the way that practitioners wish to walk.

Markov Chain Monte Carlo methods provide the mechanism for implementing Bayes theorem computationally. Using carefully constructed Markov chains, samples can be constructed from arbitrary posterior distributions, regardless of their analytical tractability. This is an enormously empowering development, and has led to something of a revolution in computational statistics, among the statistical fraternity and applied disciplines alike. The debates of the 20th century regarding the regarding the philosophical validity of Bayes' methods have all but disappeared in the wake of pragmatism. Bayesian methods have seen extensive adoption simply because they work. And importantly, they work in cases where alternatives do not.

The confluence of Bayesian MCMC and the computing power to exploit them has provided unprecedented scope for realistic statistical modelling. Until the realisation of Bayesian MCMC, adequately addressing structural complexity was difficult in many applications and required the development of specific methodology and custom built software. The only feasible alternative was to coerce the data into the over-simplified framework of an existing method. Now, Bayesian MCMC provides a consistent unifying framework within which many problems can be analysed in their full complexity using generic software.

8.2 Thesis Summary

Throughout the thesis we have developed nonlinear extensions to the general linear model based on Bayesian MCMC. Where possible we have undertaken a comparative analysis with established methods. In all cases we have demonstrated that our MCMC methods perform comparably to existing methods, and often provide additional advantages.

In Chapter 4 we implemented a Bayesian MCMC method for parameter estimation in nonlinear regression and provided a comparative analysis of our approach with the Least Squares methods used by Ratkowsky (1983). We showed that our method performs equally well, and offers a number of distinct advantages. The assumption of normality is not a requirement for valid inference, and estimates are not biased in the absence of normality. The availability of posterior samples provides ease of estimation and inference through the ability to calculate arbitrary summary statistics against the posterior samples.

The availability of these samples was shown to provide additional advantages in the diagnosis of problem situations. MCMC chain traces and posterior sections both provide useful diagnostic information, and the latter also offers an aid to interpretation of the posterior. Finally, we demonstrated that our method allowed posterior samples obtained under one parameterisation to provide estimates and inference regarding alternative parameterisations by back transformation of posterior samples. This provides a substantial advantage over Least Squares methods, allowing practitioners to quickly and easily explore alternative parameterisations. Importantly, it avoids confinement to a restricted range of parameterisations simply because their sampling properties render them the only tractable option.

In Chapter 5 we developed a novel method for transforming the response using the Gibbs sampler. Our technique finds a nonlinear transformation of the response such that the criteria underpinning the general linear model are met by the transformed variable, so that modelling may proceed under that framework. Our method simultaneously estimates the response transformation along with parameters to fit an assumed linear model. A significant advantage of our approach is that it incorporates uncertainty in the choice of transformation into subsequent inference, in contrast with other transformation methods currently in use. We demonstrated the

successful application of our method by reconsidering an example put forward by Box and Cox (1964), and provided a comparative analysis with the results suggested by their method.

Our method provided improved significance of the experimental factors relative to the Box–Cox method. This attests its ability to enhance the detectability of differences between factors which may be confounded in the observations on the original scale at which the data were collected. Importantly, our method also reduced the systematic variation observed in the residuals of the fitted model, ensuring that the criteria for the general linear model were more nearly met, and allowing estimation and inference to proceed within that framework without bias.

In Chapter 6 we adapted our transformation method to provide a new class of models. We estimated the functional relationships between the response and individual covariates, and then used these to construct models in which the functions were combined in an additive fashion. That is, our approach considered the response as the sum of transformations of the covariates. While this is similar in concept to the general class of additive models considered by Hastie and Tibshirani (1990) among others, in that the response is modelled as the sum of functions of the independent variables, our technique is very different in substance. Nevertheless, we demonstrated that it provides comparable results to those models.

Our modelling strategy is very flexible and produces readily interpretable results by enabling visualisation of the functional relationships between individual predictors and the response. In addition, our approach offers a distinct advantage in situations where it is reasonable to assume that functions of the covariates should be monotonic. There is no need to impose additional constraints to obtain a reasonable fit. This feature suggests the method as a natural fit to many data arising in applied disciplines, where monotonicity is a desirable feature.

Finally, to extend the ideas presented in the previous chapters into multivariate or mixed-model contexts, a method for estimating correlations is required. In Chapter 7 we developed two methods for the estimation of correlations using variations of the Gibbs sampler.

Our principal method relies upon a clever exploitation of the primary constraint. We define a sampling interval by exploiting the fact that each estimand in a correlation matrix must be chosen to preserve the positive definite property of the matrix as a whole. Samples from this interval are then generated from a pair of carefully selected Gamma distributions. We are able to foresee that cases could conceivably arise in which this method would operate with low efficiency, and developed another method for use in that event. The second method is more robust at the expense of greater computational overhead. Both methods were tested against simulated data to illustrate their relative merits. We successfully demonstrated the ability of both schemes to successfully estimate each element from simulated correlation matrices.

Taken together, the methods described in this thesis form the basis of a toolkit for the exploration of nonlinear relationships. These tools offer an important supplement to existing modelling strategies and point to a number of directions for further development.

8.3 Further Research

8.3.1 Nonlinear Regression Models

The MCMC routines we have developed for nonlinear regression are modular, using models which are checked for conformity against a standardised constructor prior to the commencement of MCMC sampling. Implementation of variations on the nonlinear procedures described in Chapter 4 is therefore quite straightforward. There are a number of obvious variants of the models presented in that chapter. An extension to allow for models with multiplicative error structures, for example. Indeed, several extensions to the framework have already been developed (including multiplicative error models), but in the interest of presenting a single cohesive comparative analysis with the work of Ratkowsky (1983), details of these procedures were omitted from the main text.

Asymptotic regression was introduced by Stevens (1951), with other early accounts due to Patterson (1956), and Finney (1958). The appeal of the form is that many data seem to approach some limit asymptotically, and linear models are inadequate to reflect such limits. Pinheiro and Bates (2000) describe the implementation of the method available in the `R` statistical environment. We have also implemented a Bayesian MCMC method which provides estimation and inference in asymptotic regression models. Source code for our additional methods additions are provided in Appendix A, along with the code used to produce the results seen throughout the thesis.

8.3.2 Monotonic Additive Models

Additional work is required to develop wider utility among this class of models. A relatively simple extension of the modelling strategy would allow for the inclusion of indicator variables, so that separate functional relationships might be estimated for each level of some covariate factor, for example. A more challenging extension would be the ability to fit a surface as a function of several continuous covariates.

Both of these suggestions would benefit from a model comparison technique. However, it is not yet clear how to undertake model comparison and selection for these models.

If we were able to calculate Bayes factors (Kass, 1993; Kass and Raftery, 1995; Raftery, 1995) for these models, one could establish the significance of individual

model terms by making comparisons between models which differ by only a single term. Lack of significance in a coefficient β_k would indicate that inclusion of the k-th covariate was unnecessary, and allow model selection to proceed on that basis.

To compute Bayes factors we would need to ascertain the marginal likelihood for the model(s) in question (see, for example, Carlin and Chib, 1995; Chib, 1995). Because our approach takes advantage of the fact that the conditional distributions are readily available, no explicit reference to the likelihood is required. Until a method for calculating the likelihood is discovered, model comparison and selection based on Bayes factors remains out of reach.

8.4 Concluding Remarks

Over the course of the last decade the accessibility and use of MCMC tools has increased substantially. The original BUGS (Bayesian analysis Using the Gibbs Sampler) software has diversified into a family of tools which now includes WinBUGS (Lunn et al., 2000), OpenBUGS (Thomas et al., 2006), and JAGS (Just Another Gibbs Sampler) (Plummer, 2003); CODA (Plummer et al., 2009) has enabled a standardised format for post-processing MCMC posterior samples; with all of these tools (and many others) accessible from within the R statistical environment, providing interoperability and ease of use for a wide (and growing) range of MCMC tools. Yet for all these improvements Bayesian analysis remains out of reach of many. While the tools are available they are not yet perceived as being accessible outside a small, largely specialist audience. Lunn et al. (2009) provide a critique, and suggest future directions.

Models continue to increase in complexity and more closely reflect the real-world phenomena which inform their construction. Methods which extend models and modelling frameworks by allowing estimation and inference in nonlinear contexts will increasingly be in demand. The methods developed and demonstrated in this thesis provide a set of tools which begin to meet these needs, and identify others yet to be met. They represent a small step in a long journey. One hopes that at some stage in the future nonlinear methods will enjoy a level of accessibility and adoption comparable to their linear counterparts. Certainly nonlinear applications of MCMC will remain a key area of active research for the foreseeable future.

Appendix A

Source Code

A.1 Chapter 4

```
is.function1 <- function(f)
                      is.function(f) && length(formals(f)) == 1
is.function2 <- function(f)
                      is.function(f) && length(formals(f)) == 2

nls.additive <- function(f, y, X, label, start,
    constraint = function(p) p[length(p)] > 0,
    log.prior  = function(p) dgamma(p[length(p)], 0.01, 0.01, log=T))
{

  ## Argument checking
  if(!is.function2(f))
    stop("f must be a function of two arguments")

  if(missing(start))
    stop("Missing starting values")

  if(!is.function1(constraint))
    stop("Constraint must be a function of one argument")

  if(!constraint(start))
    stop("Starting value does not satisfy constraint")

  if(!is.function1(log.prior))
    stop("prior must be a function of one argument")


  ## Model components
  list(
       label = label,           ## Data-Model Identifier
```

```
      n     = length(y),       ## Number of observations
      n.p   = length(start),   ## Number of parameters
      names = names(start),    ## Parameter names
      start = start,           ## Starting point
      y = y,                   ## Response and covariates
      X = X,

      ## Predicted, fitted values and residuals:
      predict = f,
      fitted     = function(p) f(p,X),
      residuals  = function(p) y - f(p,X),
      constraint = constraint,
      log.prior  = log.prior,
      log.likelihood = function(p)
          sum(dnorm(y, f(p,X), 1/sqrt(p[length(p)]), log=T))
      )

} ## end nls.additive



nls.multiplicative <- function(f, y, X, start,
    constraint = function(p) TRUE,
    log.prior  = function(p) dgamma(p[length(p)], 0.01, 0.01, log=T))
{

  ## Argument checking
  if(!is.function2(f))
    stop("f must be a function of two arguments")

  if(missing(start))
    stop("Missing starting values")

  if(!is.function1(constraint))
    stop("Constraint must be a function of one argument")

  if(!constraint(start))
    stop("Starting value does not satisfy constraint")

  if(!is.function1(log.prior))
    stop("prior must be a function of one argument")

  ## Model components
  list(
       n     = length(y),       ## Number of observations
       n.p   = length(start),   ## Number of parameters
       names = names(start),    ## Parameter names
       start = start,           ## Starting point
```

```
       y = y,                    ## Response and covariates
       X = X,

       ## Predicted, fitted values and residuals:
       predict = f,
       fitted     = function(p) f(p,X),
       residuals  = function(p) y - f(p,X),
       constraint = constraint,
       log.prior  = log.prior,
       log.likelihood = function(p)
           sum(dlnorm(y, log(f(p,X)), 1/sqrt(p[length(p)]), log=T))
       )

} ## end nls.multiplicative



asymp.additive <- function(y, X, start,
      constraint = function(p) p[length(p)] > 0,
      log.prior  = function(p) dgamma(p[length(p)], 0.01, 0.01, log=T))
{

  ## Argument checking
  X <- drop(X)

  if(!is.vector(X))
    stop("Expected a single covariate")

  if(!is.function1(constraint))
    stop("Constraint must be a function of one argument")

  if(!is.function1(log.prior))
    stop("prior must be a function of one argument")


  ## If no starting point - use simple heuristics to generate one.
  if(missing(start)) {

    ## Make Asym slightly bigger/smaller than biggest/smallest
    y.last <- y[order(X)[length(y)]]

    if (y.last - min(y) < max(y) - y.last) {
      Asym <- min(y) - 0.1 * (max(y) - min(y))
    } else {
      Asym <- max(y) + 0.1 * (max(y) - min(y))
    }
```

```
    ## Use least squares on the logs to estimate lrc
    cfs <- coef(lsfit(X, log(abs(y - Asym))))
    lrc <- log(-cfs[2])

    ## Estimate R0 based on Asym and lrc
    R0    <- mean((y - Asym * (1-exp(-exp(lrc)*X))) / exp(-exp(lrc)*X))
    tau   <- 1 / var(y - Asym + (R0-Asym) * exp(-exp(lrc)*X))
    start <- c(Asym=Asym, R0=R0, lrc=lrc, tau=tau)
  }


  ## Model components
  list(
       n     = length(y), ## Number of observations
       n.p   = 4,         ## Number of parameters
       names = c("Asym", "R0", "lrc", "tau"), ## Parameter names
       start = start,     ## Starting point
       y = y,             ## Response and covariates
       X = X,
       fitted    = function(p)     p[1]+(p[2]-p[1]) * exp(-exp(p[3])*X),
       residuals = function(p) y - p[1]+(p[2]-p[1]) * exp(-exp(p[3])*X),
       constraint = constraint,
       log.prior  = log.prior,
       log.likelihood = function(p)
       sum(dnorm(y, p[1]+(p[2]-p[1]) * exp(-exp(p[3])*X), 1/sqrt(p[4]), log=T))
       )

} ## end asymp.additive




mcmc.metropolis <- function(model, covm, start,
                            n.chains=1, iters=1000, sub=10) {

  ## Extract model structure
  n.p        <- model$n.p
  loglik     <- model$log.likelihood
  logprior   <- model$log.prior
  constraint <- model$constraint

  if(!is.matrix(covm))   covm <- diag(rep(covm, length=n.p), n.p, n.p)
  if(missing(start))    start <- model$start
  if(!is.matrix(start)) start <- matrix(start, n.p, n.chains)

  L.covm <- chol(covm)
  dimnames(L.covm) <- NULL
```

```
  ## Initialize chains
  chains <- array(0, c(n.p, iters, n.chains))
  dimnames(chains) <- list(model$names, NULL, paste("chain", 1:n.chains))

  for(k.chain in 1:n.chains) {

    ## Initialize P and the log posterior
    P      <- as.vector(start[, k.chain])
    logp.P <- loglik(P) + logprior(P)

    for(k.iter in 1:iters) {
      for(k.sub in 1:sub) {

        ## Get next proposal point
        Q <- P + rnorm(n.p, 0, 1) %*% L.covm

        if(constraint(Q)) {

          ## Compute log posterior at Y
          logp.Q <- loglik(Q) + logprior(Q)

          ## Metropolis Hastings rule
          if(logp.Q - logp.P > log(runif(1))) {

            ## Proposal accepted - Q replaces P
            P <- Q
            logp.P <- logp.Q
          }
        }
      }
      ## Subsample the chain
      chains[, k.iter, k.chain] <- P
    }
  }
  list(model=model, chain=chains, last=chains[, iters,], covm=covm)

} ## end mcmc.metropolis




adapt.metropolis <- function(model,
      covm=c(rep(0.1,length(model$start)-1),1), start=model$start,
      n.chains=1, iters=200, sub=10, n.adapt=5, adapt.scale=0.5, plot=T)
{
  n.p <- model$n.p
```

```
  ## Concatenation of all chains
  chains <- array(NA, c(n.p, iters * n.adapt, n.chains))
  dimnames(chains) <- list(model$names, NULL, paste("chain", 1:n.chains))

  #label <- round(runif(1)*1000,0)
  label <- model$label

  ## Repeatedly simulate with updated covariance
  for(k in 1:n.adapt) {

    paste("start:", start)

    fit   <- mcmc.metropolis(model, covm, start,
                             n.chains=n.chains, iters=iters, sub=sub)
    chains[,((k-1)*iters+1):(k*iters),] <- fit$chain
    covm  <- adapt.scale * cov.chain(fit, drop=0)
    #print(paste("covm - iter", k,":", diag(covm)))
    start <- fit$last
    #print(paste("start - iter", k,":", start))

    if(plot) {
      filename <- paste(label,"-adapt-",k,".ps", sep="")
      postscript(filename, width=gfx.w, height=gfx.h, horizontal=FALSE,
                 paper="special",family="URWHelvetica")
      plot.chain(list(model=model, chain=chains), main="")
      dev.off()
    }
  }

  fit$covm <- covm
  fit

} ## end adapt.metropolis


#
## Sundry Ancillary Functions
#

## Extract a subset of an array chains.  We can drop a number of
## intial iterations, or select a subset of variable or chains.
subset.chain <- function(ch, drop=0, subset=NULL, chain=NULL) {

  dm <- dim(ch)
  if(is.null(subset)) subset <- 1:dm[1]
  if(is.null(chain))   chain <- 1:dm[3]
  ch[subset, (drop+1):dm[2], chain, drop=F]
}
```

```
## Collapse several chains down to a single matrix,
## with one column for each parameter and a row for each sample.
collapse.chain <- function(ch) {
  dm <- dim(ch)
  ch <- matrix(ch, dm[1], dm[2]*dm[3], dimnames=list(dimnames(ch)[[1]], NULL))
  t(ch)
}



## Compute basic parameter summaries from the chain
summary.chain <-  function(fit, drop=100, subset=NULL, chain=NULL,
                           digits=4, as.matrix=TRUE, rat=TRUE,
                           quantiles=c(0.5, 0.025, 0.975)) {

  summarize <- function(x) {
      c("Mean"=mean(x), "Std Dev"=sd(x), quantile(x, quantiles))
  }

  ch  <- collapse.chain(subset.chain(fit$chain, drop, subset, chain))
  tab <- t(apply(ch, 2, summarize))

  ## as.matrix allows xtable compatability
  if (!as.matrix)
    print(tab, digits=digits)
  else
    tab
}


## Compute chain correlations
summary.chain.cor <-  function(fit, drop=100, subset=NULL,
                               chain=NULL, digits=4) {

  ch  <- collapse.chain(subset.chain(fit$chain, drop, subset, chain))

  cat("\nCorrelations\n")
  crl <- cor(ch)
  crl[!lower.tri(crl)] <- NA
  print(crl[-1, -ncol(crl), drop=FALSE], digits=digits, na="")
}



## Plot the chain
plot.chain <- function(fit, drop=0, subset=NULL, chain=NULL, main=NULL, ...) {
```

```
  ch <- subset.chain(fit$chain, drop, subset, chain)
  k <- dim(ch)[1]

  ## establish colour scheme based on chain count
  c <- dim(ch)[3]
  if (c < 2) cols <- 4
  else {
    if (c==2) cols <- c(2,4)
    else cols <- seq(3,3+c-1)
  }

  opar <- par(oma=c(4, 0, 4, 0) + 0.1,
              mar=c(0, 5.1, 0, 1),
              mfcol=c(if(k > 5) 5 else k, (k-1) %/% 5 + 1))

  for(i in 1:k) {

    ylab <- dimnames(ch)[[1]][i]
    matplot(ch[i,,], type="l", lty=1, col=cols, axes=F, ylab=ylab,...)
    box()
    axis(side=2)
    if(i==k || i%%4==0) axis(side=1, outer=TRUE)
  }

  if (is.null(main))
    title(deparse(substitute(fit)), outer=TRUE)
  else
    title(main=main, outer=TRUE)

  par(opar)
}




## Constructs pairwise plots for a subset of parameters.
pairs.chain <- function(fit, drop=100, subset=NULL, chain=NULL, main=NULL) {

  if (is.null(main)) main.label <- deparse(substitute(fit))
  else               main.label <- main

  pairs(collapse.chain(subset.chain(fit$chain, drop, subset, chain)),
        pch=".", main=main.label )
}

## Constructs pairwise plots for a subset of parameters.
cov.chain <- function(fit, drop=100, subset=NULL, chain=NULL) {
```

```
  cov(collapse.chain(subset.chain(fit$chain, drop, subset, chain)))
}


## Posterior mean of a set of chains
mean.chain <- function(fit, drop=100, subset=NULL, chain=NULL) {
  apply(collapse.chain(subset.chain(fit$chain, drop, subset, chain)),
        2, mean)
}


## Simple plot for single covariate models
plot1.fit.chain <- function(fit, drop=100, chain=NULL, n=50, xpred=NULL,
                            main="Posterior Mean", ...) {

  x <- drop(fit$model$X)
  y <- fit$model$y

  if(!is.vector(x)) stop("Expect a single covariate model")
  cfs.mean <- mean.chain(fit, drop, NULL, chain)

  if(is.null(xpred)) xpred <- range(x)

  xp <- seq(min(xpred), max(xpred), length=n)
  yp <- fit$model$predict(cfs.mean, xp)
  X <- c(x,xp)
  Y <- c(y,yp)

  plot(X, Y, type="n", main=main,...)
  points(x,y)
  lines(xp,yp, col="blue")

}
```

A.2 Chapter 5

```
##
## 0. Truncated Normal Random Generator
##
## Common to all following functions.

rtnorm <- function(n, mu, sigma, lower=-Inf, upper=Inf) {

    z <- qnorm(runif(n, pnorm(lower,mu,sigma),
                        pnorm(upper,mu,sigma)),
               mu, sigma)

    z[z == Inf] <- lower[z == Inf]
    z[z ==-Inf] <- upper[z ==-Inf]

    pmin(pmax(lower,z),upper)
}

##
## 1. No ties
##
## Assumes no ties in y => sort by y and then red/black sample.

gibbs.order1 <- function(z, mu, sigma, y) {

  n <- length(y)

  ## Re-order everything by y
  ord <- order(y, z)
  mu <- mu[ord]
  sigma <- sigma[ord]
  z <- z[ord]

  ##
  ## Gibbs sample from truncated Normals
  ##

  ## The first z - only bounded above
  z[1] <- rtnorm(1, mu[1], sigma[1], upper=z[2])

  ## The last z - only bounded below
  z[n] <- rtnorm(1, mu[n], sigma[n], lower=z[n-1])

  ## All interior z with even indices.
  even <- seq(2, n-1, 2)
  z[even] <- rtnorm(length(even), mu[even], sigma[even],
```

```
                    z[even-1], z[even+1])

  ## All interior z with odd indices.
  odd <- seq(3, n-1, 2)
  z[odd] <- rtnorm(length(odd), mu[odd], sigma[odd],
                   z[odd-1], z[odd+1])

  ## Return in the original order
  z[order(ord)]
}

##
## 2. Preserving ties
##
## Assume tied values may be present and preserve them.

gibbs.order2 <- function(z, mu, sigma, y) {

  n <- length(y)

  ## Reduce y to ordered unique values
  y.u <- sort(unique(y))

  ## Compute indices k such that y = y.u[k]
  k <- match(y, y.u)

  ## Get the corresponding unique z.  In principle where y is tied, z
  ## should be tied, so taking the mean really is the same as choosing
  ## any one of the tied z.
  z.u <- as.vector(sapply(split(z,k), mean))

  ## Get mu and sigma for the unique values.  The tau.u is the sum of
  ## the precisions, and mu.u is the precision weighted mean.
  tau <- 1/sigma^2
  tau.u <- as.vector(sapply(split(tau, k), sum))
  mu.u <- as.vector(sapply(split(tau * mu, k), sum)) / tau.u
  sigma.u <- 1/sqrt(tau.u)

  ## There are no ties in the unique values so we can use gibbs.order1
  z.u <- gibbs.order1(z.u, mu.u, sigma.u, y.u)

  ## Expand out the duplicate values
  z.u[k]
}

##
## 3. Breaking ties
##
```

```
## Assume ties may be present and allow breakage.

gibbs.order3 <- function(z, mu, sigma, y) {

  ## Sort and add fake +/-Inf endpoints so that all of the original
  ## points have an upper and lower bound
  ord <- order(y)
  z <- c(-Inf, z[ord], Inf)
  y <- c(-Inf, y[ord], Inf)
  mu <- c(0, mu[ord], 0)
  sigma <- c(0, sigma[ord], 0)

  n <- length(y)

  ## Reduce y to ordered unique values
  y.u <- unique(y)

  ## Compute indices k such that y = y.u[k]
  k <- match(y, y.u)


  ## Get min and max of each group of ties
  mn <- as.vector(sapply(split(z,k), min))
  mx <- as.vector(sapply(split(z,k), max))

  ## All interior z with even k - will be bounded below by max of next
  ## group down, and bounded above by min of next group up.
  k.even <- k%%2==0
  k.even[c(1,n)] <- FALSE
  z[k.even] <- rtnorm(sum(k.even), mu[k.even], sigma[k.even],
                                   mx[k[k.even]-1], mn[k[k.even]+1])


  ## Get new min and max of each group of ties
  mn <- as.vector(sapply(split(z,k),min))
  mx <- as.vector(sapply(split(z,k),max))

  ## All interior z with odd k - will be bounded below by max of next
  ## group down, and bounded above by min of next group up.
  k.odd <- k%%2==1
  k.odd[c(1,n)] <- FALSE
  z[k.odd] <- rtnorm(sum(k.odd), mu[k.odd], sigma[k.odd],
                                 mx[k[k.odd]-1], mn[k[k.odd]+1])

  ## Return in orginal order without the fake endpoints
  z[order(ord)+1]
}
```

A.3 Chapter 6

```
gibbs.beta <- function(y,X,tau,beta0,Tau0) {
  V <- solve(tau * crossprod(X) + Tau0)
  mu <- V %*% (tau * t(X)  %*% y + Tau0 %*% beta0)
  mu + drop(rnorm(ncol(X)) %*% chol(V))
}

gibbs.tau <- function(y,X,beta,a,b) {
  r <- y - X %*% beta
  rgamma(1, a+length(r)/2, b+crossprod(r)/2)
}


rtnorm <- function(n,mu,sigma,lower=-Inf,upper=Inf) {
  z <- qnorm(runif(n,pnorm(lower,mu,sigma),
                    pnorm(upper,mu,sigma)),mu,sigma)

  z[z==Inf] <- lower[z==Inf]
  z[z==-Inf] <- upper[z==-Inf]
  pmin(pmax(lower,z),upper)
}


am.gibbs <- function(formula, data,
                     break.ties=FALSE, beta=NULL,
                     beta0=0, Sigma0=1000,
                     tau.a=0.001, tau.b=0.001,
                     iters=1000, thin=10) {

  ## Extract the response y and design matrix X
  mf <- model.frame(formula,data)
  y <- model.response(mf)
  X <- model.matrix(formula,mf)

  ## Number of beta, obs
  m <- ncol(X)
  n <- nrow(X)

  ## Setup prior
  beta0 <- rep(beta0, length=m)
  if(!is.matrix(Sigma0))
    Sigma0 <- diag(rep(Sigma0, length=m), m, m)
  Tau0 <- solve(Sigma0)

  ## Initialize beta
  if(is.null(beta)) beta <- beta0
```

```
  ## Allocate chain
  Zs <- array(0, c(n, m-1, iters))
  betas <- matrix(0, iters, m)

  ## Initialize transformed predictors
  Z <- X

  ## translm.gibbs:
  ##for(k1 in 1:iters) {
  ##  for(k2 in 1:thin) {
  ##    ##! Update z
  ##    z <- gibbs.order3(z, X %*% beta, sigma, y)
  ##    ## Update beta
  ##    beta <- gibbs.beta(y, X, 1, beta0, Tau0)
  ##  }
  ##  ch[k1,] <- c(beta, z)
  ##}

  for(k1 in 1:iters) {
    for(k2 in 1:thin) {
      tau <- gibbs.tau(y, Z, beta, tau.a, tau.b)
      beta <- gibbs.beta(y, Z, tau, beta0, Tau0)
      sigma <- rep(1/sqrt(tau), n)
      for(i in 2:m) {
        mu <- (y - Z[,-i] %*% beta[-i]) / sign(beta[i])
        ifelse(break.ties == TRUE,
               Z[,i] <- gibbs.order3(Z[,i], mu, sigma, X[,i]),
               Z[,i] <- gibbs.order2(Z[,i], mu, sigma, X[,i]))
      }
    }
    Zs[,,k1] <- Z[,-1]
    betas[k1,] <- beta
  }
  list(Z=Zs, beta=betas)
}
```

A.4 Chapter 7

```
##
## Truncated inverse gamma deviates
##

rinvgammat <- function(n, a, b, max) {

    1/qgamma(runif(n, pgamma(1/max,a,b), 1), a, b)
}

##
## Visualise truncated inverse gamma samples
##

vis.rinvgammat <- function(n, a, b, c, x0=0, x1=1.1*c) {

  xs <- seq(x0, x1, length=500)
  op <- par(mfrow=c(2,2))

  ## Plot 1: Gamma Density
  plot(dgamma(xs,a,b) ~ xs, type="l", ylab="",
       main=paste("Gamma (",round(a,2),",",round(b,2),") Density",sep=""))
  abline(v=1/c, lty=3, col="blue")

  ## Plot 2: Gamma CDF
  plot(pgamma(xs,a,b) ~ xs, type="l",
       main="Gamma Probability Function", ylab="")
  abline(h=pgamma(1/c, a, b), v=1/c, lty=3, col="blue")

  ## Plot 3: Uniform Sample (truncated at pgamma(1/c))
  X <- runif(n,pgamma(1/c,a,b),1)
  z <- 1/qgamma(X, a, b)

  xs <- seq(pgamma(1/c, a, b), 1, length=50)
  ys <- 1/qgamma(xs, a, b)

  plot(xs ~ ys, type="l", xlim=range(xs), ylim=range(ys),
       main=paste("Inverse Gamma Quantiles [",
                  round(pgamma(1/c,a,b),2),",1]", sep=""))
  abline(h=pgamma(1/c, a,b), lty=3, col="red")

  ## Plot 4: Inverse Deviates returned from Uniform Sample Quantiles
  plot(density(z, adjust=3, to=max(z)), xlim=c(0,1.1*max(z)))

  n/2*(1+log(2 * b[1]/(n*(dx-z)))) - b[1]/(dx-z)
  par(op)
```

```
}


##
## Post - evaluate posterior for fixed i,j over a range of x.
##

post <- function(i,j,V,S,n,x) {
  x <- as.vector(x)
  m <- length(x)
  p <- double(m)

  for(k in 1:m) {

    V[i,j] <- V[j,i] <- x[k]

    p[k] <- exp(-(n*log(det(V))
                  + sum(diag(solve(V,S))))/2)
  }
  p
}


##
## Convenience method returning upper triangular indices of a matrix
##

upper.tri.index <- function(M) {

  stopifnot(is.matrix(M))

  upr <- upper.tri(M)
  i <- row(M)[upr]
  j <- col(M)[upr]
  ij <- cbind(i,j)

  ij
}


##
## Posterior sample for V_{ij} for single, fixed (i,j) pair.
##

cov.bigamma <- function(i0,j0,V,S,n,N,vis=FALSE) {

  ## Ensure i < j
  ## (otherwise indexing in Schur decomposition is more complex)
```

```
i <- min(i0,j0)
j <- max(i0,j0)

## Schur decomposition to compute |V(x)|
## solve with one argument returns inverse
W <- solve(V[-j,-j])
## the jth row minus cols i and j. ie: an (m-2)-tuple vector
u <- V[j,-c(i,j)]
## coefficients are then
a0 <- V[j,j] - drop(u %*% W[-i,-i] %*% u)
a1 <- -2 * drop(W[-i,i] %*% u)
a2 <- -W[i,i]

## Compute x1 and x2, the roots of |V(x)|
s <- if(a1 >= 0) 1 else -1
q <- -(a1 + s * sqrt(a1^2 - 4 * a0 * a2))/2
xs <- sort(c(a0/q, q/a2))

## Compute b1 and b2 by interpolation:
x <- ((1:3) * xs[1] + (3:1) * xs[2]) / 4  ## 3-vector
X <- cbind(1, 2/(x-xs[1]), 2/(xs[2]-x))   ## 3-matrix
tr <- double(3)
Vx <- V
## for each value of x
for(k in 1:3) {
  ## substitute in position i,j
  Vx[i,j] <- Vx[j,i] <- x[k]
  ## solve system and calculate trace
  tr[k] <- sum(diag(solve(Vx,S)))
}
## b0 not important, drop
bs <- solve(X,tr)[-1]


## Now draw Vij by rejection:
dx <- xs[2] - xs[1]  ## interval length

if(bs[1] < bs[2]) {
  ## Alt code - truncated inverse gamma by rejection
  #z <- 1/rgamma(N,n/2-1,bs[1])
  #z <- z[z<dx]
  #logp <- griddy(N, n, bs[2], dx)
  z <- rinvgammat(N, n/2-1, bs[1], dx)
  logp <- n/2 * (1 + log(2 * bs[2]/(n*(dx-z)))) - bs[2]/(dx-z)
  logu <- log(runif(N))
  x <- xs[1] + z[logp > logu]
} else {
  ## Alt code 1 - truncated inverse gamma by rejection
```

```
    #z <- 1/rgamma(N,n/2-1,bs[2])
    #z <- z[z<dx]
    ## Alt code 2 - griddy Gibbs
    #logp <- griddy(N, n, bs[1], dx)
    z <- rinvgammat(N, n/2-1, bs[2], dx)
    logp <- n/2 * (1 + log(2 * bs[1]/(n*(dx-z)))) - bs[1]/(dx-z)
    logu <- log(runif(N))
    x <- xs[2] - z[logp > logu]
  }
  if (vis==TRUE) {
    par(mfrow=c(2,2))
    hist(z, xlim=range(z), breaks=50, freq=F, main="")
    hist(logp, xlim=range(logp), breaks=50, freq=F, main="", xlab="")
    hist(logu, xlim=range(logu), breaks=50, freq=F, main="", xlab="")
  }
  x
}


##
## Mulitple chain wrapper for cov.gibbs (see below)
##

m.chain <- function(V, S, n, N, n.chains=2, R=100) {

  ch <- array(0, dim = c(N, 2 * dim(upper.tri.index(V))[1], n.chains))

  for (k in 1:n.chains) {

    ch[,,k] <- cov.gibbs(V, S, n, N, R=R)
  }
  ch
}


##
## Sample covariances in heteroscedastic covariance matrix V
## assumes known variances, v_ii.
##
cov.gibbs <- function(V, S, n, N, R=100) {

  ## Arguments:
  ## V - covariance matrix
  ## S - cross product of the observation matrix
  ## n - number of observations
  ## N - sample count (suggestion only)
  ## R - max consecutive rejections
```

```
## Indices of uppr triangle, column wise
## ensures i < j, for Schur decomposition
ij <- upper.tri.index(V)

## Count of estimands: v_ij, i!=j
M <- dim(ij)[1]

## Allocate the chain
ch <- matrix(0,N,M)

## rejection counter
rs <- matrix(0,N,M)

for(k1 in 1:N) {
  if (k1%%100==0) print(k1)

  ## Loop over upper triangular elements
  for(k2 in 1:M) {

    i <- ij[k2,1]
    j <- ij[k2,2]

    ## Schur decomposition to compute |V(x)|
    W <- solve(V[-j,-j])
    u <- V[-c(i,j), j]
    a0 <- V[j,j] - drop(u %*% W[-i,-i] %*% u)
    a1 <- -2 * drop(W[-i,i] %*% u)
    a2 <- -W[i,i]

    ## Compute x1 and x2, the roots of |V(x)|
    s <- if(a1 >= 0) 1 else -1
    q <- -(a1 + s * sqrt(a1^2 - 4 * a0 * a2))/2
    xs <- sort(c(a0/q, q/a2))

    ## Compute b1 and b2 by interpolation
    x <- ((1:3) * xs[1] + (3:1) * xs[2])/4
    X <- cbind(1, 2/(x-xs[1]), 2/(xs[2]-x))
    tr <- double(3)
    Vx <- V
    for(k in 1:3) {
      Vx[i,j] <- Vx[j,i] <- x[k]
      tr[k] <- sum(diag(solve(Vx,S)))
    }
    (b <- solve(X,tr)[-1])

    ## Now draw Vij by rejection / Griddy Gibbs
    dx <- xs[2] - xs[1]
    c <- 0
```

```
      if (b[1] > b[2]) {

        z <- rinvgammat(1, n/2-1, b[1], dx)

        ## while log posterior of z is less than a random value
        while(n/2*(1+log(2 * b[2]/(n*(dx-z)))) - b[2]/(dx-z) < log(runif(1))) {
          c <- c + 1
          ## if it looks hopeless, switch to griddy gibbs
          if (c > R) {
            z <- griddy(i,j,V,S,xs,n)
            c <- -1
            break
          }
          ## otherwise continue sampling, until you get a real one
          z <- rinvgammat(1, n/2-1, b[1], dx)
        }
        V[i,j] <- V[j,i] <-  ifelse(c == -1, z, xs[1] + z)
        rs[k1,k2] <- c
      }
      else {

        z <- rinvgammat(1, n/2-1, b[2], dx)

        while(n/2*(1+log(2*b[1]/(n*(dx-z))))-b[1]/(dx-z) < log(runif(1))) {
          c <- c + 1
          if (c > R) {
            z <- griddy(i,j,V,S,xs,n)
            c <- -1
            break
          }
          z <- rinvgammat(1, n/2-1, b[2], dx)
        }
        V[i,j] <- V[j,i] <- ifelse(c == -1, z, xs[2] - z)
        rs[k1,k2] <- c
      }
    }
    ch[k1,] <- V[upper.tri(V)]
  }
  cbind(ch, rs)
}


##
## Same as cov.gibbs, but uses griddy Gibbs (see below)
##

cov.griddy <- function(V, S, n, N) {
```

```
  ## Indices of uppr triangle, column wise
  ## ensures i < j, for Schur decomposition
  ij <- upper.tri.index(V)
  M <- dim(ij)[1]

  ## Allocate the chain
  ch <- matrix(0,N,M)

  ## parameter store
  ts <- matrix(0,N,M)

  for(k1 in 1:N) {
    if (k1%%10==0) print(k1)

    ## Loop over upper triangular elements
    for(k2 in 1:M) {
      (i <- ij[k2,1])
      (j <- ij[k2,2])

      ##print(k2)

      ## Schur decomposition to compute |V(x)|
      (W <- solve(V[-j,-j]))
      (u <- V[-c(i,j), j])
      (a0 <- V[j,j] - drop(u %*% W[-i,-i] %*% u))
      (a1 <- -2 * drop(W[-i,i] %*% u))
      (a2 <- -W[i,i])

      ##as[k1,k2,] <- c(a0,a1,a2)

      ## Compute x1 and x2, the roots of |V(x)|
      s <- if(a1 >= 0) 1 else -1
      q <- -(a1 + s * sqrt(a1^2 - 4 * a0 * a2))/2
      (xs <- sort(c(a0/q, q/a2)))

      ## generate our deviate
      ts[k1,k2] <- system.time(x <- griddy(i,j,V,S,xs,n))[3]

      ## substitute into V(x)
      (V[i,j] <- V[j,i] <-  x)

    }
    ch[k1,] <- V[upper.tri(V)]
  }
  cbind(ch,ts)
}
```

```
##
## Sample the covariances in a heteroscedastic
## covariance matrix V. Assumes diagonal entries
## known. Update covariances in turn via Gibbs.
##

cov.gibbs.c <- function(V, S, n, N, r.max=100, debug=FALSE) {

  ## Indices of uppr triangle, column wise
  ## ensures i < j, for Schur decomposition
  ij <- upper.tri.index(V)
  M <- dim(ij)[1]

  ## Allocate the chain
  ch <- matrix(0,N,M)

  ## rejection counter
  rs <- matrix(0,N,M)

  ## parameter stores
  as <- array(0, dim=c(N,M,3))
  bs <- array(0, dim=c(N,M,2))
  xs <- array(0, dim=c(N,M,2))

  for(k1 in 1:N) {
    if (k1%%100==0) print(k1)
    ## Loop over upper triangular elements: column-wise!!
    for(k2 in 1:M) {
      i <- ij[k2,1]
      j <- ij[k2,2]

      if (debug == TRUE) { print(k2) }

      ## Schur decomposition to compute |V(x)|
      W <- solve(V[-j,-j])
      u <- V[-c(i,j), j]
      a0 <- V[j,j] - drop(u %*% W[-i,-i] %*% u)
      a1 <- -2 * drop(W[-i,i] %*% u)
      a2 <- -W[i,i]
      as[k1,k2,] <- c(a0,a1,a2)

      if (debug==TRUE) {
          print(paste(paste("a",0:2,":",sep=""),round(as[k1,k2,],5)))
      }

      ## Compute x1 and x2, the roots of |V(x)|
      s <- if(a1 >= 0) 1 else -1
```

```
q <- -(a1 + s * sqrt(a1^2 - 4 * a0 * a2))/2
xs[k1,k2,] <- sort(c(a0/q, q/a2))

if (debug==TRUE) {
    print(paste(paste("x", 1:2, ":", sep=""),round(xs[k1,k2,],5)))
}

## Compute b1 and b2 by interpolation
x <- ((1:3) * xs[k1,k2,1] + (3:1) * xs[k1,k2,2])/4
X <- cbind(1, 2/(x-xs[k1,k2,1]), 2/(xs[k1,k2,2]-x))
tr <- double(3)
Vx <- V
for(k in 1:3) {
  Vx[i,j] <- Vx[j,i] <- x[k]
  tr[k] <- sum(diag(solve(Vx,S)))
}
bs[k1,k2,] <- solve(X,tr)[-1]

if (debug==TRUE) {
    print(paste(paste("b", 1:2, ":", sep=""), round(bs[k1,k2,],5)))
}

## Now draw Vij by rejection / Griddy Gibbs
dx <- xs[k1,k2,2] - xs[k1,k2,1]
c <- 0
if (bs[k1,k2,1] > bs[k1,k2,2]) {

  z <- rinvgammat(1, n/2-1, bs[k1,k2,1], dx)

  ## while log posterior of z is less than a random value
  while(n/2 * (1 + log(2 * bs[k1,k2,2]/(n*(dx-z))))-bs[k1,k2,2]/(dx-z)
            < log(runif(1))) {

    c <- c + 1

    ## if it looks hopeless, switch to griddy gibbs
    if (c > r.max) {
      if (debug==TRUE) {
          print(paste(">",r.max,"rejections... calling griddy()"))
      }
      z <- griddy(i,j,V,S,n,xs[k1,k2,])
      c <- -1
      break
    }
    ## otherwise continue sampling, until you get a real one
    z <- rinvgammat(1, n/2-1, bs[k1,k2,1], dx)
  }
  rs[k1,k2] <- c
```

```
        v <- append(ch[,k2],ifelse(c == -1, z, xs[k1,k2,1] + z))
        V[i,j] <- V[j,i] <-  mean(v)/sd(v)
      }
      else {

        z <- rinvgammat(1, n/2-1, bs[k1,k2,2], dx)

        while(n/2 * (1+log(2 * bs[k1,k2,1]/(n*(dx-z)))) - bs[k1,k2,1]/(dx-z)
                  < log(runif(1))) {

          c <- c + 1

          if (c > r.max) {

            if (debug==TRUE) {
                print(paste(">",r.max,"rejections... calling griddy()"))
            }

            z <- griddy(i,j,V,S,n,xs[k1,k2,])
            c <- -1
            break
          }
          z <- rinvgammat(1, n/2-1, bs[k1,k2,2], dx)
        }
        rs[k1,k2] <- c
        v <- append(ch[,k2],ifelse(c == -1, z, xs[k1,k2,2] - z))
        V[i,j] <- V[j,i] <- mean(v)/sd(v)
      }
      if (debug==TRUE){ print(V) }
    }
    ch[k1,] <- V[upper.tri(V)]
  }
  cbind(ch,rs,as[,,1],as[,,2],as[,,3],bs[,,1],bs[,,2],xs[,,1],xs[,,2])
}

##
## Inverse Wishart Density based on Cholesky decomposition.
## Necessary because R treats chol() and solve() inconsistently
##

dinvwish <- function(L, v, S, det.S, debug=FALSE) {

  if (debug==TRUE) { print("Entering dinvwish()") }

  ## L is chol(W)
  ## v is degrees of freedom for L,
  ## here set equal to n: correct?
```

```
  k <- nrow(S)
  gammapart <- 1
  for (i in 1:k) {
    gammapart <- gammapart * gamma((v + 1 - i)/2)
  }

  denom <- gammapart * 2^(v * k/2) * pi^(k * (k - 1)/4)

  if (debug==TRUE) { print(paste("denom:", denom)) }

  ## det(W) is product of diagonal chol(W) squared
  detW <- (prod(diag(L)))^2

  if (debug==TRUE) { print(detW)  }

  ##hold <- S %*% chol2inv(L) ## <-- not equal to solve(S) %*% W !!!
  hold <- solve(S) %*% t(L) %*% L


  tracehold <- sum(hold[row(hold) == col(hold)])
  ##num <- det.S^(v/2) * detW^(-(v + k + 1)/2) * exp(-1/2 * tracehold)
  num <- det.S^(-v/2) * detW^((v - k - 1)/2) * exp(-1/2 * tracehold)

  if (debug==TRUE) {
      print(paste("num:", num)); print("Leaving dinvwish()")
  }

  return(num/denom)
}


##
## Griddy Gibbs
##

griddy <- function(i, j, V, S, x, n, N=1, m=10, plot=FALSE) {

  ## i,j - indices of current estimand
  ## V - current approx of the matrix V
  ## S - cross product of the observation matrix
  ## x - interval on which V is positive definite
  ## n - number of observation points
  ## N - number of samples required
  ## m - smoothing parameter

  ## modify endpoints for starting sequence to avoid
  ## nuerical instability in evaluating posterior
  epsilon <- 5e-08
```

```
delta <- 5e-04

xx <- c(x[1] + epsilon, x[2] - epsilon)

hpd <- FALSE

while(length(hpd[hpd==TRUE]) < m) {

  ## generate grid points
  (xs <- seq(xx[1], xx[2], length=m*10))

  ## evaluate posterior at xs
  w <- post(i, j, V, S, n, xs)
  w[is.nan(w)] <- 0

  if (plot==TRUE) {
    ##plot(w ~ xs, type="p", col="blue", xlim=c(x[1],-0.3), xlab="x")
    plot(w ~ xs, type="p", col="blue", xlim=c(xs[1],xs[length(xs)]), xlab="x")
    abline(v=V[i,j], lty=3, col="red")
    ##text(x = V[i,j] - 0.015, y = -1, labels=expression(v_12), cex=0.8)
    ##axis(side=1, V[i,j], tcl=-0.5, labels = expression(v[12]), cex=0.5)
  }

  ## Highest Posterior Denisty
  hpd <- w/max(w) > delta

  ## previous definition of HPD does not necessarily cover (0,1] well
  ## adding indices adjacent to HPD provides better coverage
  for (k in 2:(length(hpd)-1)) {

    if (hpd[k]==TRUE && hpd[k-1]==FALSE) {
      hpd[k-1] <- TRUE
    }
    if (hpd[k]==FALSE && hpd[k-1]==TRUE) {
      hpd[k] <- TRUE
      break
    }
  }
  (xx <- c(xs[hpd][1], xs[hpd][length(xs[hpd])]))
}

w <- w[hpd]
xs <- xs[hpd]


## Empirical CDF
p <- cumsum(w/sum(w))
```

```
  if (plot==TRUE) {
    ##plot(p ~ xs,type="p", col="blue", xlim=c(xs[1],-0.3), xlab="x")
    plot(p ~ xs,type="p", col="blue", xlim=c(xs[1],xs[length(xs)]), xlab="x")
    ##abline(v=V[i,j], lty=3, col="red")
  }

  ## generate uniform random deviate(s)
  (u <- runif(N))

  ## find and return the CDF transformed value
  z <- interpolate(xs,p,u)

  if (plot==TRUE) {
    plot.griddy(xs, p, z, u)
  }

  ## return value
  z
}

##
## subroutines called from griddy()
##

interpolate <- function(xs,ys,y) {

    ## return an x value for every y
    x <- double(y)

    for (k in 1:length(y)) {

        ## obtain index of x1
        i <- length(ys[ys <= y[k]])

        if (i != length(ys))
            x[k] <- xs[i] + (xs[i+1]-xs[i])/2 +
                (y[k]-ys[i])*(xs[i+1]-xs[i])/(ys[i+1]-ys[i])
        else
            x[k] <- x[k-1] + x[k-1]-x[k-2]
    }
    x
}

## visualise gribby gibbs operations

plot.griddy <- function(xs,p,x,y) {

    dx <- c((xs[-1]-xs[-length(xs)])/2, 0)
```

```
    xs <- xs+dx

    plot(p ~ xs, type="l", main="", ylab="p", xlab="x")

    abline(h=y, lty=3, col="blue")
    abline(v=x, lty=3, col="red")

    text(min(xs)+ diff(range(xs))/20, u-0.02,
         labels=paste("u =",round(u,4)),cex=0.8)
    text(z+diff(range(xs))/15, 0,
         labels = paste("x =",round(z,4)), cex=0.8)
}

##
## Convenience function for testing griddy():
## Determines the interval (x1,x2) from which v_ij
## of a variance matrix V may be sampled to comply
## with the non-negative definite constraint on V.
##

posdef.int <- function(i0,j0,V) {

  ## Ensure i < j (otherwise indexing
  ## Schur decomposition is more complex)
  i <- min(i0,j0)
  j <- max(i0,j0)

  ## Schur decomposition to compute |V(x)|
  ## solve with one argument returns inverse
  ## Hence W is the inverse of V[-j,-j]
  W <- solve(V[-j,-j])
  ## the jth row minus cols i and j.
  ## ie: an (m-2)-tuple vector
  u <- V[j,-c(i,j)]
  ## coefficients are then
  a0 <- V[j,j] - drop(u %*% W[-i,-i] %*% u)
  a1 <- -2 * drop(W[-i,i] %*% u)
  a2 <- -W[i,i]

  ## Compute x1 and x2, the roots of |V(x)|
  s <- if(a1 >= 0) 1 else -1
  q <- -(a1 + s * sqrt(a1^2 - 4 * a0 * a2))/2
  xs <- sort(c(a0/q, q/a2))

  xs
}
```

BIBLIOGRAPHY

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and Computing*, 18(4):343–373.

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, 5(2):141–160.

Atkinson, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford Statistical Science Series. Oxford University Press.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

Barnard, G. A. and Bayes, T. (1958). Studies in the history of probability and statistics: Ix. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4):293–315.

Barnard, J., McCulloch, R., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica*, 10:1281–1311.

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(1):1–25.

Bates, D. M. and Watts, D. G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares. *The Annals of Statistics*, 9(6):1152–1167.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

Beale, E. M. L. (1960). Confidence regions in non-linear estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22(1):41–88.

Berger, J. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association*, 95:1269–1276.

Berger, J. O. (2006). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311.

Bolker, B. M. (2008). *Ecological Models and Data in R*. Princeton University Press.

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*. Wiley-Interscience, 2nd edition.

Boole, G. (2008). *An Investigation of the Laws of Thought*. Merchant Books.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77(377):209–210.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*. Wiley-Interscience.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley.

Box, M. J. (1971). Bias in nonlinear estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(2):171–201.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598.

Broemeling, L. D. (2007). *Bayesian Biostatistics and Diagnostic Medicine*. Chapman & Hall/CRC.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Browne, W. J. (2002). MCMC algorithms for constrained variance matricies. Technical report, Institute of Education, University of London.

Cai, B., Meyer, R., and Perron, F. (2008). Metropolis-hastings algorithms with adaptive proposals. *Statistics and Computing*, 18(4):421–433.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484.

Carlin, B. P. and Louis, T. A. (2008). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, 3rd edition.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, New York.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *American Statistician*, 46(3):167–174.

Chambers, J. M. (1973). Fitting nonlinear models: Numerical techniques. *Biometrika*, 60(1):1–13.

Chen, J.-S. and Jennrich, R. I. (1995). Diagnostics for linearization confidence intervals in nonlinear regression. *Journal of the American Statistical Association*, 90(431):1068–1074.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *American Statistician*, 49(4):327–335.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Clarke, G. P. Y. (1987). Marginal curvatures and their usefulness in the analysis of nonlinear regression models. *Journal of the American Statistical Association*, 82(399):844–850.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, 35(1):54.

Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley & Sons.

Cook, R. D. and Goldberg, M. L. (1986). Curvatures for parameter subsets in nonlinear regression. *The Annals of Statistics*, 14(4):1399–1418.

Cook, R. D. and Witmer, J. A. (1985). A note on parameter-effects curvature. *Journal of the American Statistical Association*, 80(392):872–878.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Dale, A. I. (1999). *A History of Inverse Probability from Thomas Bayes to Karl Pearson*. Springer-Verlag, New York, 2nd edition.

Daniels, M. and Kass, R. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57:1173–84.

Daniels, M. J. and Kass, R. E. (1999). Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263.

Daniels, M. J. and Pourahamdi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89:553–566.

de Finetti, B. (1974). *Theory of Probability*, volume 1. John Wiley & Sons, New York.

de Finetti, B. (1975). *Theory of Probability*, volume 2. John Wiley & Sons, New York.

Diebolt, J. and Ip, E. H. S. (1995). *Markov Chain Monte Carlo in Practice*, chapter Stochastic EM: Method and Application, pages 259–273. Chapman & Hall/CRC.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley-Interscience, 3rd edition.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242.

Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association*, 19(148):431–453.

Fearnhead, P. (2008). Editorial: Special issue on adaptive monte carlo methods. *Statistics and Computing*, 18(4):341–342.

Finney, D. J. (1958). The efficiencies of alternative estimators for an asymptotic regression equation. *Biometrika*, 45(3–4):370–388.

Fisher, R. A. (1921). Studies in crop variation I: An examination of the yield of dressed grain from broadbalk. *Journal of Agricultural Science*, 11:107–135.

Fisher, R. A. (1922). On the theoretical foundations of mathematical statistics. *Philosophical Transactions of the Royal Society, A*, 222:309–368.

Fisher, R. A. (1925a). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Fisher, R. A. (1925b). Theory of statistical estimation. *Procedings of the Cambridge Philosophical Society*, pages 700–725.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

Fisher, R. A., Bennet, J. H., and Yates, F. (1990). *Statistical Methods, Experimental Design, and Scientific Inference: A Re-issue of Statistical Methods for Research Workers, The Design of Experiments, and Statistical Methods and Scientific Inference*. Oxford University Press.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823.

Gallant, A. R. (1975). Nonlinear regression. *American Statistician*, 29(2):73–81.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. John Wiley and Sons.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association*, 95(452):1300–1304.

Gelfand, A. E. and Sahu, S. K. (1994). On markov chain monte carlo acceleration. *Journal of Computational Graphics and Statistics*, 3:261–267.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. (1995). *Markov Chain Monte Carlo in Practice (edited by W R Gilks and S Richardson and D J Spiegelhalter)*, chapter Inference and Monitoring Convergence, pages 131–143. Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall, 2nd edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). *Efficient Metropolis Jumping Rules*, chapter 5, pages 599–608. Oxford University Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

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.

Geweke, J. (1992). *Bayesian Statistics 4 (edited by J M Bernado and A P David and A F M Smith)*, chapter Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments. Clarendon Press - Oxford.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. John Wiley and Sons.

Geyer, C. J. (1992a). Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483.

Geyer, C. J. (1992b). [practical markov chain monte carlo]: Rejoinder. *Statistical Science*, 7(4):502–503.

Geyer, C. J. (1995). *Markov Chain Monte Carlo in Practice*, chapter Estimation and Optimization of Functions, pages 240–258. Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1995a). *Markov Chain Monte Carlo In Practice*. Chapman & Hall / CRC.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995b). *Markov Chain Monte Carlo in Practice*, chapter Introducing Markov Chain Monte Carlo, pages 1–20. Chapman & Hall.

Gilks, W. R., Roberts, G. O., and George, E. I. (1994). Adaptive direction sampling. *The Statistician*, 43:179–189.

Gilks, W. R. and Roberts, G. R. (1995). *Markov Chain Monte Carlo in Practice*, chapter Strategies for Improving MCMC, pages 89–114. Chapman & Hall/CRC.

Gill, J. (2007). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman & Hall/CRC, 2nd edition.

Greenberg, E. (2007). *Introduction to Bayesian Econometrics*. Cambridge University Press.

Gregory, F. G. (1956). General aspects of leaf growth. In *The Growth of Leaves*, Proceedings of the 3rd Easter School in Agricultural Science, pages 3–17. University of Nottingham, Butterworths, London.

Grobbee, D. and Hoes, A. W. (2008). *Principles and Methods of Clinical Epidemiology*. Jones & Bartlett Publishers.

Hald, A. (1998). *A History of Mathematical Statistics 1750-1930*. Wiley, New York.

Hamilton, D. C., Watts, D. G., and Bates, D. M. (1982). Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. *The Annals of Statistics*, 10(2):386–393.

Hartley, H. (1961). The modified gauss-newton methods for the fitting of non-linear regression functions by least squares. *Technometrics*, 3:pp 269–280.

Hartley, H. O. (1964). Exact confident regions for the parameters in non-linear regression laws. *Biometrika*, 51:pp 347–343.

Hartley, H. O. and Booker, A. (1965). Nonlinear least squares estimation. *The Annals of Mathematical Statistics*, 36(2):638–650.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245. Special Issue on Simulation Modeling and Statistical Computing.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Heyes, J. K. and Brown, R. I. (1956). Growth and cellular differentiation. In *The Growth of Leaves*, Proceedings of the 3rd Easter School in Agricultural Science, pages 31–52. University of Nottingham, Butterworths, London.

Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79(386):302–309.

Hougaard, P. (1982). Parametrizations of non-linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):244–252.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*, volume Vol. 845 of *Wiley Series in Probability and Statistics*. John Wiley & Sons. ISBN 9780470011546.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1974). Fisher and inverse probability. *International Statistical Review*, 42(1):1–3.

Jeffreys, H. (1998). *Theory of Probability*. Oxford University Press, 3rd edition.

Jeffreys, H. (2007). *Scientific Inference*. Muller Press.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Kass, R. E. (1984). Canonical parameterizations and zero parameter-effects curvature. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(1):86–92.

Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, 42:551–560.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain monte carlo in practice: A roundtable discussion. *American Statistician*, 52(2):93–100.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

King, R., Gimenez, O., Morgan, B., and Brooks, S. (2009). *Bayesian Analysis for Population Ecology*. Chapman & Hall/CRC.

Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian Econometric Methods*. Cambridge University Press.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):251–263.

Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics*. Wiley-Blackwell.

Lawson, A. B. (2008). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Chapman & Hall/CRC.

Lindley, D. V. (1965a). *Introduction to Probability & Statistics from a Bayesian Viewpoint*, volume Part 1. Probability. Cambridge University Press.

Lindley, D. V. (1965b). *Introduction to Probability & Statistics from a Bayesian Viewpoint*, volume Part 2 Inference. Cambridge University Press.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Marin, J.-M. and Robert, Christian, P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer.

Marquardt, D. (1963). An algorithm for least squares estimation of non-linear parameters. *Journal of the Society for industrial and Applied Mathematics*, 11:pp431–441.

McCarthy, M. A. (2007). *Bayesian Methods for Ecology*. Cambridge University Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall: London.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, A. H., Teller, M. N., and Teller, E. (1953). Equations of state caluculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–92.

Moyé, L. A. (2007). *Elementary Bayesian Biostatistics*. Chapman & Hall/CRC.

Müller, S. (1991). A generic approach to posterior integration and gibbs sampling. Technical Report 91-09, Purdue University.

Patterson, H. D. (1956). A simple method for fitting an asymptotic regression curve. *Biometrics*, 12(3):323–329.

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician*, 44(1):26–30.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the R Core team (2008). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-90.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and Computing. Springer-Verlag.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria. ISSN 1609-395X.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2009). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-4.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics*. Wiley-Interscience.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rachev, S. T., Hsu, J. S. J., Bagasheva, B. S., and Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. Wiley.

Raftery, A. E. (1995). *Markov Chain Monte Carlo in Practice*, chapter Hypothesis testing and Model Selection, pages 163–188. Chapman & Hall/CRC.

Raftery, A. E. and Lewis, S. M. (1992). [practical markov chain monte carlo]: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, 7(4):493–497.

Raftery, A. E. and Lewis, S. M. (1995). *Markov Chain Monte Carlo in Practice (edited by W R Gilks and S Richardson and D J Spiegelhalter)*, chapter Implementing MCMC, pages 115–130. Chapman & Hall/CRC.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard Business School, Boston, Massachusetts.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling: A Unified Practical Approach*. Marcel Dekker.

Ratkowsky, D. A. and Dolby, G. R. (1975). Taylor series linearization and scoring for parameters in nonlinear regression. *Applied Statistics*, 24(1):109–122.

Ritter, C. and Tanner, M. A. (1992). Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, 87(419):861–868.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edition.

Roberts, G. O. (1996). *Markov Chain Monte Carlo in Practice*, chapter Markov Chain Concepts Related to Sampling Algorithms, pages 45–57. Chapman & Hall.

Ross, G. J. S. (1990). *Non-Linear Estimation*. Springer-Verlag.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press.

Rubin, D. B. (1984). Distinguishing between the scale of the estimand and the transformation to normality. *Journal of the American Statistical Association*, 79(386):309–310.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Ryan, T. A., Joiner, B. L., and Ryan, B. F. (1976). *Minitab Student Handbook*. Duxbury Press.

Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Owl Books.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons, New York.

Scherer, B. and Martin, R. D. (2007). *Introduction to Modern Portfolio Optimization with NuOPT, S-PLUS and S+Bayes*. Springer.

Schruben, L. W. (1982). Detecting initialization bias in simulation output. *Operations Research*, 30(3):569–590.

Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear Regression*. Wiley-Interscience, revised edition.

Shively, T. S., Kohn, R., and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, 94(447):777–794.

Simonoff (1996). *Smoothing Methods in Statistics*. Springer. Ltn 519.536 S599s.

Singpurwalla, N. D. (2006). *Reliability and Risk: A Bayesian Perspective*. Wiley.

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

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley.

Stevens, W. L. (1951). Asymptotic regression. *Biometrics*, 7(3):247–267.

Stigler, S. M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press. ISBN: 978-0674403413.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz., S. (2006). Making bugs open. *R News*, 6:12–17.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83(402):394–405.

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

Tukey, J. W. (1949). One degree of freedom for non-addivity. *Biometrics*, 5(3):232–242.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathemetical Statistics*, 33(1):1–67.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.