Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Importance Resampling MCMC: a methodology for cross-validation in inverse problems and its applications in model assessment

Sourabh Bhattacharya

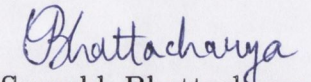
Trinity College Dublin
Ireland

Thesis submitted for the degree of Doctor of Philosophy

2004

Declaration

This thesis is entirely my own work. The contents of this thesis have not been submitted as an exercise for a degree at this or any other University. The contents of this thesis may be lent or copied by the Library at Trinity College Dublin.



Sourabh Bhattacharya

Acknowledgements

I would like to express my most sincere gratitude to my supervisor, Professor John Haslett. I believe he is the most sincere and caring supervisor. This thesis simply would not have materialised without him.

The Enterprise Ireland grant SC/2001/171 is gratefully acknowledged.

I thank all my dear friends, colleagues and everyone else. This thesis is the outcome of the support and good wishes of everybody.

Finally, I must mention of my family who played the key role. Their sacrifice is behind my enjoyment of this thrilling bit of research for three years.

Abstract

This thesis presents a methodology for implementing cross-validation in the context of Bayesian modelling of situations we loosely refer to as 'inverse problems'. It is motivated by an example from palaeoclimatology in which scientists reconstruct past climates from fossils in lake sediment. The inverse problem is then to build a model with which to make statements about climate, given sediment. One natural aspect of this is to examine model fit via cross-validation. In MCMC studies this can be computationally burdensome and our procedure has attractive properties in this respect. We demonstrate that, in addition, it is possible to take advantage of the flexibility inherent within the method to make it suitable for exploring multimodal distributions. We also propose to construct useful reference distributions using data obtained from cross-validation. Our proposals are illustrated using a simulated data set and several real data sets.

Contents

Chapter 1

Introduction

Leave-one-out cross-validation (simply, cross-validation in this thesis) involves running several similar versions of the same statistical model; in some situations 'brute force' implementation can be computationally burdensome. In this thesis we present, for use within Markov chain Monte Carlo (MCMC) (see, for example, Tierney (1994)) using the Metropolis-Hastings algorithm, a proposal mechanism based on Sampling/Importance Resampling (SIR) (see, for example, Rubin (1988); we however, refer to this method as Importance Resampling (IR) as Stern and Cressie (2000)). This is customised to exploit the basic similarities. We take advantage of recent work showing that IR can be efficiently and robustly implemented by sampling 'without replacement'. We argue that this proposal mechanism can deliver dramatic computational savings in the case of inverse problems. The motivating problems arise in palaeoclimatology, which we briefly describe below.

1.1 Palaeoclimatology: an overview

Quantitative reconstructions of palaeoclimate have intrinsic value as a source of insight into Earth's history. Such reconstructions are also an essential basis for evaluating the performance of the general circulation models (GCMs) used to explore the potential future climatic consequences of anthropogenic changes to the Earth system.

One aspect of palaeoclimatology is the attempt to reconstruct details of palaeo-environment by using fossil organisms in lake sediment. In simple terms, climate influences ecology and ecology influences the biotic assemblage of organisms that are to be found fossilised in lake sediment; thus such assemblages, called proxies for climate, can be used to reconstruct some details of past climate. Pollen, chironomids (non-biting midges), tree rings, diatoms, chrysophytes, coleoptera, ostracods, mosses, radiolaria, foraminifera, etc. are examples of such proxies. Typical assemblages involve counts of the frequency of fossil organisms from 20–100 different species. Very many counts are zero and the total count is several hundred. However, the information about past climate retained by such proxy data may be quite subtle (see West (1996)). Also, the data recorded are typically very noisy and the current age dating technology is of limited precision, in spite of recent advances. Many environmental reconstructions are qualitative and are presented in terms such as "cool", "temperate", "moist", "dry", etc.

1.1.1 Classical approaches

Imbrie and Kipp (1971) presented, for the first time, a procedure for the quantitative reconstruction of past environmental variables from fossil assemblages. Statistical models connecting climate to fossil assemblages are calibrated on modern data, which we also refer to as the training data and applied to fossil data. Birks (1998) reviews a number of developments in classical procedures used in environmental reconstruction from biotic assemblages.

Statistically, the problem may be stated as follows. For each of a large number of samples, n^m modern and n^f fossil, compositional data describing fossil assemblages, $p^m = \{p_i^m; i = 1, \cdots, n^m\}$ and $p^f = \{p_i^f; i = 1, \cdots, n^f\}$, are available for study. For the modern samples climate data $X^m = \{x_i^m; i = 1, \cdots, n^m\}$ are also available as covariates; the climate values at the fossil sites are missing. The objective in palaeoclimate study is to estimate these missing values and thus to reconstruct the pre-historic climate.

The most widely used modelling procedure builds on a multivariate regression type model

in which the proportion of the ith species in the training data set, p_i^m, is proportional to a 'response function' of the climate variable. The response function, which is typically unimodal, describes the way in which the proportion of the species varies with the climate. The regression is then 'inverted' and applied to assemblages of fossil data, which yields a corresponding quantitative reconstruction of climate. The weighted averaging partial least squares (WA-PLS)(ter Braak (1995)) is a widely used procedure for this purpose and has close links to correspondence analysis.

The unimodal nature of the response function implies that there exists an 'optimum' or 'ideal' climate, at which the species thrives the most, and the rate at which the species proportion diminishes from the optimum is controlled by the 'tolerance' of the species. The importance of the above assumption has been noted long ago. For instance, Shelford's law of tolerance states that a species not only requires a certain minimum amount of resource but also that species do not tolerate more than a certain maximum amount of the resource. A more general law states that each species thrives best at a particular optimum value of an environmental variable and cannot survive when the value is either too low or too high. For more discussion on unimodality, see ter Braak (1987). Vasko et al. (2000), Korhola et al. (2002) use unimodal functions as a basis for building their model. However, Huntley (1993) argues that at least some species may have multiple optima and hence the response function may be multimodal. He recommends non-parametric modelling of the response function. The WA-PLS method does not explicitly assume any parametric form of the response function; however, the unimodality assumption is invoked implicitly.

The main disadvantage of classical methods is that there seems to be no consistent way to make statements of uncertainty in the reconstructions thus obtained. Thus they do not provide any adequate basis for incorporating useful scientific information, for example, temporal smoothness of past climates.

1.1.2 The Bayesian approach

Unlike the classical paradigm, the Bayesian paradigm (see, for example Bernardo and Smith (1994)) is flexible enough to cope with all kinds of uncertainties. In this paradigm, the statistical parameter is treated as a random variable and a likelihood function is used to express the relative plausibility of obtaining different values of this parameter when a particular data have been observed.

Some past applications of the Bayesian paradigm to palaeoclimatology are van Deusen and Reams (1996), who use an AR(2) model, West (1996), who provides overview of many generic issues in Bayesian palaeoclimatology. Some recent work is by Trudinger (2000), Trudinger et al. (2002), Trudinger et al. (2002), Vasko et al. (2000), Toivonen et al. (2001), Korhola et al. (2002), Whiley et al. (2004).

West did not directly address palaeoclimate reconstruction, working instead with a proxy time series. He draws attention to the fact that proxy series inevitably correspond to random observation times and involve temporal and spatial aggregation. He identifies the difficult issues associated with uncertainty regarding observation times arising from problems in radiocarbon dating.

Vasko et al. (2000), Toivonen et al. (2001), Korhola et al. (2002) seem to be the first to use detailed statistical model based approach to palaeoclimate reconstruction, putting the Bayesian approach in the context of some of the classical procedures discussed in Birks (1995). They considered a one dimensional climate variable and data on non-biting midges, called chironomids; this allowed them to use the simple unimodal response function.

Instead of chironomids, Whiley et al. (2004) use data on pollen and two climate variables, GDD5 (growing degree days above 5^0C) and MTCO (mean temperature of the coldest month) and attempt to reconstruct the two climates that prevailed over Glendalough in Ireland about 15,000 calendar years ago. Most pollen species represent at least a group of species, sometimes an entire genus, and in some cases several genera or an entire family comprising numerous plant species. As a result, even if each species' relationship to a climate variable

were unimodal, one can hardly expect a unimodal relationship for a pollen species each with distinctive ecology and distribution. In practice, the form of the concurrent relationship of a pollen taxon to two or more climate variables is typically complex and multi-modal. Also unlike in the case of the chironomid data of Vasko et al. (2000), the total count of species in each pollen assemblage is not always available in Whiley et al. (2004). This forced the latter to make assumptions that may be questionable.

Figure 1.1 presents a summary of realisations of GDD5 and MTCO, given the fossil pollen data and the modern training data. Dates are given as radio-carbon years 'before present' (RCYBP); following Whiley et al. (2004) we denote 1000 RCYBP by 1 ka BP. Superimposed is a reconstruction using a classical 'response surface (RS)' methodology well known in pollen palaeoclimatology; see Huntley (1993). It is seen that the climate changed rapidly at $\sim$ 9 ka BP. Although broadly in agreement, in the case of MTCO reconstruction there is some divergence between the findings of the two methods just before this change. Huntley's method, using the same data, suggest a very rapid and extreme cooling at about 9.393 ka BP; this is known in the literature as the 'Younger Dryas' event; see Dansgaard et al. (1989). This particular Bayesian model offers only weak evidence for this; see the insert, where it is seen that the MTCO reconstruction is weakly bimodal.

1.2 Model checking

An important problem in statistical analysis is to determine whether the underlying model fits the data. For forward problems, both the classical and the Bayesian literature, especially the former, are rich with discussions on such issue. However, we are interested in model assessment in inverse problems. There seems to be no literature in this context. A brief review on model checking in forward problems is presented in Chapter 3. One criticism of classical procedures of assessing model fit is that they are based on asymptotic theory (for example, the $\mathcal{X}^2$ goodness-of-fit test) and could be poor for small data sizes. Bayesian approaches do not need asymptotics; however, they can be computationally challenging,

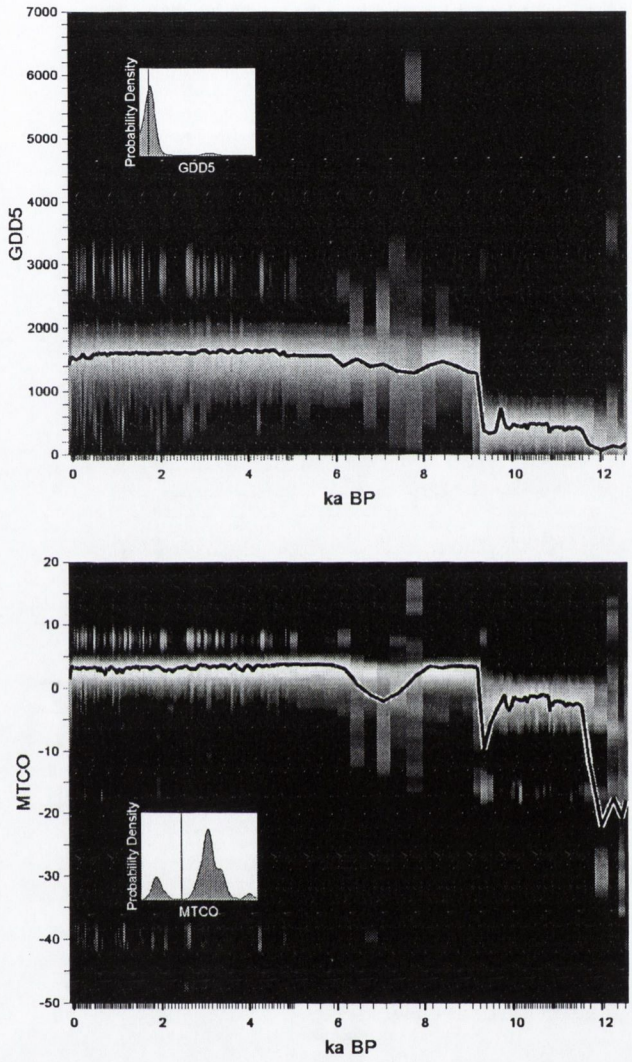


Figure 1.1: Reconstructions of GDD5 and MTCO with no temporal smoothness of past climates assumed. Overlaid are RS reconstructions. The inset is of the KDE of marginal posterior GDD5 and MTCO for the 107th historical sample, dated 9.393 ka BP; the vertical line denotes the corresponding point prediction by the RS method. Tick marks indicate the approximate radio-carbon dates of the 150 fossil samples.

in case the number of parameters is large. This is the case in Vasko et al. (2000) and Whiley et al. (2004). In the case of Vasko et al. (2000), a $\mathcal{X}^2$-like deviation measure has a very high magnitude, suggesting that the model does not fit the data adequately. As we shall show in the case of Whiley et al. (2004), the magnitude of such a statistic is comparatively quite small. However, the uncertainties in the estimates were high, which can be attributed to multimodal solutions. This can again be traced back to data quality. Such high uncertainties do not necessarily speak in favour of the null model in spite of the small magnitude of such discrepancy measure. In another example, presented in Diggle et al. (1998), assessment of residual contamination from nuclear weapons testing on a South Pacific island is sought. Here the sampling method generates spatially indexed Poisson counts conditional on an unobserved spatially varying intensity of radioactivity. In this case as well, we shall see that a reasonable discrepancy measure has quite small magnitude, but the model can still be called in question due to high uncertainty. To quantify such uncertainty suitable 'reference distributions' are needed. In this thesis, apart from proposing a method for implementing cross-validation, we also propose to construct useful reference distributions using data obtained from cross-validation.

The thesis is structured as follows. In Chapter 2 we review cross-validation and implementation details and point out technical differences between cross-validation in forward and inverse problems and argue that the latter kinds of problems are more challenging technically. In Chapter 3 we propose and evaluate a methodology for assessing fit of inverse problems. This is based on construction of reference distributions from samples obtained from cross-validation using IRMCMC. In Chapter 4 we introduce our proposed methodology, IRMCMC, and demonstrate that it provides solutions to the technical difficulties associated with cross-validation of inverse problems. The asymptotic accuracy of IRMCMC is discussed in Chapter 5. Applications of our methodologies to the palaeoclimate problems presented in Vasko et al. (2000) and Whiley et al. (2004) are discussed in Chapters 6 and 7 respectively. Sensitivity analysis in inverse problems is discussed with an example in Chapter 8. In

Chapter 9, IRMCMC has been applied to a geostatistical problem which is inverse in nature but requires forward implementation. This of course shows that our methods are useful for forward as well as inverse problems. We finally conclude in Chapter 10.

Chapter 2

Cross-validation techniques in Bayesian problems

The examples discussed in Chapter 1 motivate the interest of this thesis. These interests centre on the evaluation of the 'fit' of Bayesian models to the data. We are particularly interested in models that we describe as 'inverse' and on measures on cross-validation. The current dominant methodology is the computationally intensive MCMC. Model evaluation by cross-validation can be computationally overwhelming for high-dimensional models. The model of Diggle et al. (1998) has 160 parameters. In Vasko et al. (2000), the model consists of 3319 parameters and in the case of Whiley et al. (2004), the number of parameters is 9623. The central contribution of this thesis is an efficient algorithm for cross-validation. This is based on IR. We refer to our method as IRMCMC.

Cross-validation in forward problems has been the focus in a number of research papers; see, for example, Stone (1974), Stone (1977), Allen (1974), Geisser (1975). Generalized cross-validation is discussed in Golub et al. (1979) and Wahba (1980); a further discussion appears in Wahba (1990). See also Hastie and Tibshirani (1990), Chapter 3. For the use of cross-validation in model selection, see Brieman (1992), Brieman and Spector (1992), Zhang (1992); see also Crawford (1989). Recent reviews appear in Gelfand et al. (1992), Shao

(1993). See also Gelfand and Dey (1994), Gelfand (1996), Geisser and Eddy (1979). Bernardo and Smith (1994) discuss how cross-validation approximates the formal Bayes procedure of computing expected utilities (Good (1952)). Draper (1996) advocates cross-validation for model assessment. Vehtari (2001) uses cross-validation to compute expected utility for model assessment and comparison. In the case of disease-mapping model, Stern and Cressie (2000) attempt to reduce computational labour involved in MCMC based cross-validation using importance weighting. Marshall and Spiegelhalter (2003) provides an improvement on the method of Stern and Cressie (2000) through replication of both random effects and data.

However, we do not know of any paper that discusses cross-validation in the context of inverse problems. It will be argued that cross-validation in inverse problems is technically much more challenging than in forward problems. In the following we discuss these concepts, initially separately but always with a view to application of cross-validation to the problems described in Chapter 1. To fix attention, we first introduce a simple running example, based on Poisson regression.

Data $X = \{x_i\}$ and $Y = \{y_i\}$, $i = 1, \cdots, 10$ are available. Here we take the y_i as counts, corresponding to values of a continuous variable x_i. We adopt the Poisson model $P(\theta x_i)$ for the counts y_i, independently of the other cases; θ is an unknown parameter. The objective is to 'predict' x from a future y. We describe this as 'inverse regression', to contrast it with the much simpler forward application, the prediction of y for a future x. For the purposes of model validation, we need to contrast each of the ten x_i with the corresponding 'leave-one-out posterior distribution' $\pi(\cdot|X_{-i}, Y_{-i}, y_i)$; here X_{-i} and Y_{-i} stand for the data omitting the corresponding pair x_i, y_i. Figure 2.1, which displays artificial data simulated from the Poisson model, shows that the observed x_8 falls within a high density region of the leave-one-out posterior distribution. In fact, such region is also the 95% highest posterior density (HPD) credible region; this is defined below.

DEFINITION

The $100(1-\alpha)\%$ HPD credible region for x is the subset C of the parameter space of x of

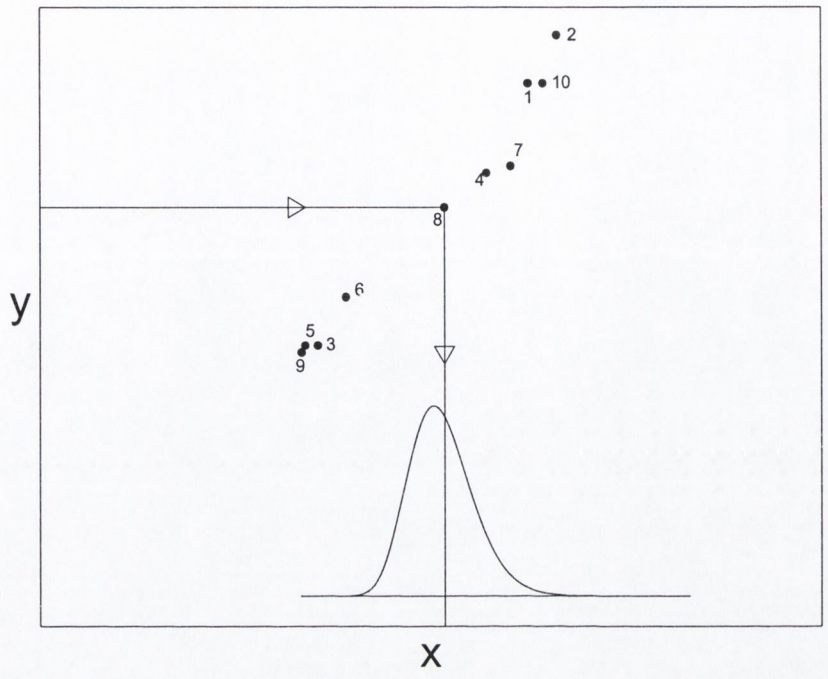


Figure 2.1: Artificial Poisson regression data; the case number is displayed alongside the datum. Case 8 illustrates cross validation with Poisson regression.

the form

$$C = \{x : \pi(x \mid X_{-i}, Y) \geq k(\alpha)\},$$

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

$$\pi(C \mid X_{-i}, Y) \geq 1 - \alpha.$$

Ideally, all the observed values of x should fall within, say, 95% HPD region of the relevant posterior distribution. HPD regions could be disjoint; see, for example, Figures 6.1, 7.2, 7.3.

2.1 Cross-validation in forward and inverse problems

The Poisson problem would be a forward problem had interest been in prediction of y and not x. In the Bayesian context, the leave-one-out posterior distribution of y at each case i

would be of interest. This is given by

$$\pi(y \mid X_{-i}, Y_{-i}, x_i) = \int p(y \mid x_i, \theta)\pi(\theta \mid X_{-i}, Y_{-i})d\theta \tag{2.1}$$

which is equivalent to the expectation of $p(y \mid x_i, \theta)$ with respect to the posterior $\pi(\theta \mid X_{-i}, Y_{-i})$.

But our interest is in the 'inverse' of the above problem. In other words, we are interested in learning about the leave-one-out posterior distribution of x in each case. This is given by

$$\pi(x \mid X_{-i}, Y_{-i}, y_i) = \int \pi(x \mid y_i, \theta)\pi(\theta \mid X_{-i}, Y)d\theta \tag{2.2}$$

Observe that the leave-one-out posterior distribution of x is the expectation of $\pi(x \mid y_i, \theta)$ with respect to the posterior $\pi(\theta \mid X_{-i}, Y)$.

However, two technical differences distinguish the two problems. In (2.1), assuming n cases, we observe that, $\pi(\theta \mid X_{-i}, Y_{-i}) \propto \pi(\theta)\Pi_{j\neq i}^{n} p(y_j \mid x_j, \theta)$. Now, since the right hand side of the preceding expression is always available, the functional form of $\pi(\theta \mid X_{-i}, Y_{-i})$ is always available. But in (2.2), $\pi(\theta \mid X_{-i}, Y) = \int \pi(\theta, x \mid X_{-i}, Y)dx$ and since the integral may not be available analytically, the functional form of $\pi(\theta \mid X_{-i}, Y)$ may be unknown. Again, the functional form of $p(y \mid x_i, \theta)$ in (2.1) is available, since p is known. But in (2.2), $\pi(x \mid y_i, \theta) = \pi(x, \theta)p(y_i \mid x, \theta)/\int \pi(x, \theta)p(y_i \mid x, \theta)dx$ and, since the integral in the denominator may be analytically intractable, $\pi(x \mid y_i, \theta)$ may be available only up to a proportionality constant. The above technical differences between forward and inverse problems are important. In the toy Poisson regression example, the leave-one-out posterior (2.2) is available analytically, but in the real problems discussed in Chapter 1, such analytic solutions are unavailable; it seems that implementation of MCMC runs leaving out each case in turn is the only possible solution for cross-validation in such problems. In Section 2.1.1 we review the basics of MCMC and also point out the difficulties involved in application of the method to cross-validation and the need to develop alternative methods. In the next section we review MCMC and associated concepts.

2.1.1 MCMC: an overview

Since the appearance of the article Gelfand and Smith (1990), MCMC methods have revolutionised statistical computing. Although the methodology finds most use in Bayesian statistics, other statistical aplications are also simultaneously influenced. For example, calculation of p-values in exact conditional inference (Diaconis and Sturmfels (1998)) and maximisation of intractable likelihood functions associated with generalised linear mixed models (McCulloch (1997)) may be achieved via MCMC.

Metropolis-Hastings algorithms, introduced by Metropolis et al. (1953)), are an important class of MCMC algorithms (see, for example, Hastings (1970), Smith and Roberts (1993), Tierney (1994), Gilks and Roberts (1996)). Given essentially any probability distribution (the "target distribution"), these algorithms provide a way to generate a Markov chain $\theta^{(0)}, \theta^{(1)}, \cdots$, having the target distribution as a stationary distribution.

Specifically, supposing that the target distribution has density f. Then, given $\theta^{(t)}$, a "proposed value" θ' is generated from some prespecified density $q(\theta^{(t)}, \theta)$ and then accepted with probability $\alpha(\theta^{(t)}, \theta')$, given by

$$\alpha(u, v) = \min\left\{\frac{f(v)}{f(u)}\frac{q(v, u)}{q(u, v)}, 1\right\} \tag{2.3}$$

If the proposed value is accepted, we set $\theta^{(t+1)} = \theta'$; otherwise, we set $\theta^{(t+1)} = \theta^{(t)}$.

In applying Metropolis-Hastings algorithms, it is necessary to choose the proposal density $q(u, v)$. Typically, this is chosen from some family of distributions, for example, normal distributions centered at u. Then there is a need to select the variance of the normal distribution appropriately to achieve some level of optimality in the performance of the algorithm. Roberts and Rosenthal (2001) demonstrate that for extremely small variance the algorithm will take a long time to converge to its stationary distribution. On the other hand, if the variance is taken to be extremely large, the proposed moves will generally be rejected and the algorithm will stay fixed for large number of iterations. Roberts and Rosenthal (2001) argue that there exist "good" values for the variance, between these two extremes, where the algorithm performs optimally. See also Gelman et al. (1996), Tanner (1996).

For the implementation of MCMC it is desirable to know how long the algorithm takes to get sufficiently close to the stationary distribution. In other words, what is the "burn-in" period? However, Jones and Hobert (2001) remark that burn-in is not strictly necessary but most variance estimation techniques are more effective when the Markov chain is stationary. More details on the burn-in debate can be found at C. J. Geyer's Web page at www.stat.umn.edu/~charlie/. There it has been argued why burn-in is pointless.

Another question of interest is to know how long the MCMC should continue. Put another way, it is desirable to know when the estimates based on the output are sufficiently accurate. To address this question evaluation of the standard deviation of the estimates (Monte Carlo error) is necessary. However, this is not straightforward due to the dependence within the Markov chain. Jones and Hobert (2001) discuss promising methods for evaluating the Monte Carlo error in this case. However, in most practical problems, a mixture of intuition, experience and *ad hoc* methods are used to determine the length of an MCMC run.

2.1.2 Application of MCMC to cross-validation

Since in the case of cross-validation the posterior density of interest is of the type $\pi(x, \theta \mid X_{-i}, Y)$, which is proportional to the prior times the likelihood, MCMC seems to be a natural tool for its exploration by sample generation. Marginalisation of the posterior (for example, (2.2)) can be done simply by ignoring realisations of parameters which are not of interest (θ in the case of (2.2)).

Assuming that $\theta = (\theta_1, \cdots, \theta_p)$ where p is the dimensionality of θ and x is unidimensional, and assuming further that the conditional $\pi(x \mid \theta_1, \cdots, \theta_p, X_{-i}, Y)$ and for $j = 1, \cdots, p$, the conditionals $\pi(\theta_j \mid x, \theta_1, \cdots, \theta_{j-1}, \theta_{j+1}, \cdots, \theta_p, X_{-i}, Y)$ are known up to normalising constants, an MCMC algorithm may proceed as follows.

- Select appropriate initial values for θ and x. Let them be $\theta^{(0)}$ and $x^{(0)}$ respectively. Then, for $t = 1, 2, \cdots$, do
- For $k = 1$ to p

- Propose a new value $\theta_k^{(t+1)}$ from some known 'proposal distribution' $Q_k(\cdot \mid \theta_k^{(t)})$
- Compute

$$\alpha_k = \min\left\{\frac{\pi(\theta_k^{(t+1)}|\theta_1^{(t+1)},\cdots,\theta_{(k-1)}^{(t+1)},\theta_{(k+1)}^{(t)},\cdots,\theta_p^{(t)},x^{(t)},X_{-i},Y)}{\pi(\theta_k^{(t+1)}|\theta_1^{(t+1)},\cdots,\theta_{(k-1)}^{(t+1)},\theta_{(k+1)}^{(t)},\cdots,\theta_p^{(t)},x^{(t)},X_{-i},Y)}\frac{Q_k(\theta_k^{(t)}|\theta_k^{(t+1)})}{Q_k(\theta_k^{(t+1)}|\theta_k^{(t)})}, 1\right\} \tag{2.4}$$

- Accept $\theta_k^{(t+1)}$ with probability α_k, else accept $\theta_k^{(t)}$.

- Propose a new value $x^{(t+1)}$ from some known proposal distribution $Q(\cdot \mid x^{(t)})$
  - Compute

$$\alpha = \min\left\{\frac{\pi(x^{(t+1)}|\theta_1^{(t)},\cdots,\theta_p^{(t)},X_{-i},Y)}{\pi(x^{(t)}|\theta_1^{(t)},\cdots,\theta_p^{(t)},X_{-i},Y)}\frac{Q(x^{(t)}|x^{(t+1)})}{Q(x^{(t+1)}|x^{(t)})}, 1\right\} \tag{2.5}$$

  - Accept $x^{(t+1)}$ with probability α, else accept $x^{(t)}$.

In the above algorithm each parameter is updated sequentially using the Metropolis-Hastings step. Recall that the performance of an MCMC algorithm depends on the proposal distributions chosen and hence much care is needed in the choice of $\{Q, Q_k; k = 1, \cdots, p\}$.

For other variations that might help improve performance of MCMC, see, for example, Geyer (1991a), Rue (2001). For MCMC algorithms in the case of particular inverse problems see Higdon et al. (2003); see also Oliver et al. (1997), Lee et al. (2002), Goodman and Sokal (1989), Liu and Sabatti (1999). Since none of these MCMC algorithms involve the concept of IR we call these 'regular MCMC' to distinguish them from our proposed algorithm IRMCMC which makes use of IR. Observe that in forward problems, y need not be updated by the Metropolis-Hastings step since given θ, direct simulation of y is possible in most cases. A real example is given in Chapter 9.

Clearly, this exercise could certainly be computationally burdensome in the extreme, particularly when p is very large and when the acceptance probabilities given by (2.4) and (2.5) are computationally demanding. This, in fact, is the case for all our real-life examples we present in this thesis. Moreover, in such examples, careful choice of the set of proposal distributions, $\{Q, Q_k; k = 1, \cdots, p\}$ in each case is not feasible, since the number of cases, n, is large; usually the same set of proposal distributions would be used 'assuming' that

they fit all cases. Provided that the leave-one-out posteriors of θ in each case do not differ much, a fact that seems to hold in all our examples and is explained in detail in Chapter 5 (but see Chapter 4, Section 4.3), it is not unreasonable to keep the proposal distributions $\{Q_k; k = 1, \cdots, p\}$ unaltered. However, this may not be a valid assumption in the case of Q, the proposal distribution of x, since different y_i are expected to predict different x_i and hence posteriors of x in different cases are expected to be different; in fact we would anticipate most of the mass in posteriors to not overlap. In Chapter 6 it is explained, with a real example, that when a particular case has bimodal solution, regular MCMC designed without particular attention to that case, can fail to explore it adequately. Simulation studies, reported in the same chapter support this conclusion.

To reduce the computational burden involved with cross-validation of forward Bayesian models, importance sampling (Geweke (1989), Robert and Casella (1999), Geyer (1991b)) may be employed. We discuss this in the next section. We also discuss that even if this may be suitable for cross-validation of forward problems, this is not in general applicable to inverse problems.

2.2 Importance sampling: an overview

The idea of importance sampling can be explained as follows. Suppose that interest lies in estimation of the expectation of a function $h(\theta)$ with respect to a distribution $f(\theta)$. Suppose further that a sample $\theta^{(1)}, \cdots, \theta^{(N)}$ is available from another distribution $g(\theta)$. Then writing the expected value of $h(\theta)$ may be estimated as

$$\hat{E}_N(h(\theta)) = \frac{1}{N} \sum_{\ell=1}^{N} h(\theta^{(\ell)}) w(\theta^{(\ell)}) \tag{2.6}$$

where $w(\theta) = f(\theta)/g(\theta)$ is the 'weight' of θ. The distribution g can be almost any density for $E_N(h(\theta))$ to converge to the true expected value. However, the variance of the above estimator is finite only when $\int h^2(\theta) \frac{f^2(\theta)}{g(\theta)} d\theta < \infty$ (see Robert and Casella (1999)). Hence importance sampling distributions with tails lighter than those of f (that is, those with

unbounded ratios f/g) are inappropriate for importance sampling, since in such cases, the variances of the corresponding estimators given by (2.6) will be infinite for many functions h. Also, in the case of unbounded w, the weights $w(\theta^{(\ell)})$, because of their high variability, will attach too much importance to a few realisations of θ. Finiteness of the variance of (2.6) is ensured by distributions g with tails thicker than f. Geweke (1989) gives sufficient conditions that ensure finiteness of the variance.

The choice of g that minimises the variance of (2.6) is given by

$$g^*(\theta) = \frac{|h(\theta)|f(\theta)}{\int |h(u)|f(u)du} \tag{2.7}$$

The result is due to Rubinstein (1981); see also Geweke (1989). Observe however, that the above optimality condition requires the knowledge of the integral, where the interest lies. Hence an alternative is to use the estimator

$$\hat{E}_N(h(\theta)) = \frac{\sum_{\ell=1}^N h(\theta^{(\ell)})w(\theta^{(\ell)})}{\sum_{\ell=1}^N w(\theta^{(\ell)})} \tag{2.8}$$

$$= \frac{\sum_{\ell=1}^N h(\theta^{(\ell)})|h(\theta^{(\ell)})|^{-1}}{\sum_{\ell=1}^N |h(\theta^{(\ell)})|^{-1}} \tag{2.9}$$

In the above, $\theta^{(\ell)} \sim g \propto |h|f$; note that the above estimator also converges to the true expectation by the Strong Law of Large Numbers. But since it is biased, it is not optimal unlike (2.7). Nevertheless, Casella and Robert (1998) have shown that (2.9) may perform better in some settings; see also Van Dijk and Kloeck (1984). For our purpose we shall always refer to the form given by (2.8).

2.2.1 Application of importance sampling to cross-validation

In forward cross-validation problems, Gelfand et al. (1992), Gelfand (1996) proposed the re-use of samples drawn from $\pi(\theta \mid X, Y)$. We refer to $\pi(\theta \mid X, Y)$ as the saturated posterior since it involves the complete available dataset. For each cross-validation, the sample available from the saturated posterior (which, in fact, is the importance sampling distribution) can be used to estimate the leave-one-out posterior distribution of y as in (2.8).

Typically, however, importance weights with respect to the saturated posterior density are not available in inverse problems. This is because the weight function in this case, given by

$$w_i(\theta) = \frac{\pi(\theta \mid X_{-i}, Y)}{\pi(\theta \mid X, Y)} = \frac{\int \pi(x, \theta \mid X_{-i}, Y) dx}{\pi(\theta \mid X, Y)}, \tag{2.10}$$

may not be available if the integration on the right hand side of the above expression is not tractable analytically. In fact, even if it is available, the saturated posterior is likely to be a poor importance sampling density unless the prior on x is sufficiently strong; we illustrate this fact in the context of the Poisson example.

Keeping the basic structure of the Poisson regression unchanged, if we put a $Gamma(\alpha x_i, \alpha)$ prior(with expected value x_i and variance x_i/α) on x, then $\pi(\theta \mid X_{-i}, Y) \propto \pi(\theta)\theta^{\sum_{j=1}^n y_j} \exp(-\theta \sum_{j \neq i} x_j)/(1 + \theta/\alpha)^{y_i + \alpha x_i}$. Also, the saturated posterior, $\pi(\theta \mid X, Y) \propto \pi(\theta)\theta^{\sum_{j=1}^n y_j} \exp(-\theta \sum_{j=1}^n x_j)$. Hence the importance weight function with the saturated posterior $\pi(\theta \mid X, Y)$ as the importance sampling density, as given by the ratio of the above two functional forms is $w_{i,\alpha}(\theta) = \exp(\theta x_i)/(1 + \theta/\alpha)^{y_i + \alpha x_i}$. Then for all $\theta \in C$, where the set C is such that$\pi(C \mid X, Y) = 1$, $w_{i,\alpha}(\theta) \uparrow 1$ as $\alpha \uparrow \infty$, indicating that all the importance weights approach a constant value as the prior strength increases. But if $\alpha \downarrow 0$, $w_{i,\alpha}(\theta) \downarrow 0$ for $\pi(\cdot \mid X, Y)$-almost all θ, signifying that all importance weights tend to be extremely small as the prior strength of x diminishes. This can cause numerical instability. Hastings (1970) also indicates that importance sampling can be a difficult method to use if values of the weights are all extremely small.

It is to be observed that adoption of the saturated posterior as the importance sampling density is akin to placing unit prior mass on x_i and hence the above facts are not surprising. We remark that in realistic situations one can hardly put a very strong prior on x to make importance weights $w_{i,\alpha}$ useful. Hence, even if the functional form of $\pi(\theta \mid X_{-i}, Y)$ is available, in practice it is still not advisable to use the saturated posterior as importance sampling distribution. Our central contribution, illustrated in Chapter 4, is the use of $\pi(x, \theta \mid X_{-i^*}, Y)$ as an importance sampling density, for some particular i^*.

Another difficulty with the importance sampling approach described above is that the

normalizing constant of $\pi(x \mid y_i, \theta)$ in (2.2) may be unknown. This means that $h(\theta)$ of (2.8) is not known completely. This prevents use of the importance sampling estimate given by (2.8). Hence the leave-one-out posterior distribution may not be obtained by any easy means. To overcome such difficulties we propose to combine very fast and easily implementable MCMC runs with IR. Specifically, we propose to leave out case i^*, to sample (by MCMC) realizations of (x, θ) and finally to draw a subsample of θ from the realized θ values using appropriate importance weights. Then, given each re-sampled θ, regular MCMC may be used to realise x. Details of this procedure is discussed in Chapter 4.

In the next section we discuss IR and its applicability to cross-validation of inverse problems.

2.3 IR with and without replacement and their application to cross-validation

The idea of IR can be expressed very simply. As in Section 2.2, but adopting a slightly different notation to suit the purpose of this section, let us suppose that interest lies in the distribution $f(\psi)$, where $\psi = (x, \theta)$. We also suppose that a sample $\hat{\psi} = \{\hat{\psi}^{(1)}, \cdots, \hat{\psi}^{(N)}\}$ is available from another distribution $g(\psi)$. Then, a sample of size K, selected from the available sample $\hat{\psi}$, with probabilities proportional to weights given by (2.6) approximates $f(\psi)$. We denote this sample of size K by $\tilde{\psi} = \{\tilde{\psi}^{(1)}, \cdots, \tilde{\psi}^{(K)}\}$.

An important technical question is whether IR should be used with or without replacement when $K << N$. Most of the references to IR in the literature use sampling with replacement. See, for example, Gelfand et al. (1992), Newton and Raftery (1994), O'Hagan and Forster (2004). However, it is to be noted in this context that choosing the sample from the original MCMC sample 'with replacement' may be unreliable if some importance weights are extremely large, since the same value of ψ may appear most of the time in the final sample, precluding adequate representation of the target posterior. This can be partially

avoided by resampling 'without replacement' where repetitions are disallowed. This has also been recommended by Stern and Cressie (2000), Gelman et al. (1995). Skare et al. (2003) show that sampling with replacement gives a larger sampling error than sampling without replacement when compared with respect to the total variation norm. This follows from the fact that the posterior has zero probability for having some $\tilde{\psi}^{(k)}$ equal; hence rejecting such proposals, by sampling without replacement, will give a lower total variance error. This we state below.

Theorem 1 *Let $\hat{p}^N$ be a probability distribution of a sample $\tilde{\psi}$ of size K drawn from a larger sample $\hat{\psi}$ of size N by means of some algorithm. Furthermore, let $\hat{p}^N$ be the conditional probability distribution of $\tilde{\psi}$ given $\Omega_0 = \{\tilde{\psi} : \tilde{\psi}^{(i)} \neq \tilde{\psi}^{(j)}, 1 \leq i < j \leq K\}$. If $K > 1$, then we have*

$$\| \hat{p}^N - f \|_{TV} \geq \| p^N - f \|_{TV}, \tag{2.11}$$

where $\| \hat{p}^N - f \|_{TV} = \sup_A(\hat{p}^N(A) - f(A))$. If $\| p^N - f \|_{TV} < 1$, then the inequality in (2.11) is strict. Furthermore, if $\hat{p}^N$ is induced by a IR type algorithm without replacement, then

$$\| \hat{p}^N - f \|_{TV} \geq \| p^N - f \|_{TV} + 0.5K(K-1)N^{-1} + O(N^{-2}) \tag{2.12}$$

Note that the theorem is applicable to any problem of any dimensionality. Skare et al. (2003) observe that, in the case of IR with replacement, convergence with N does not occur with respect to the relative supremum norm since the probability that $\tilde{\psi}^{(k_1)} = \tilde{\psi}^{(k_2)}$ for $k_1 \neq k_2$ is positive and the joint probability density is not defined with respect to the Lebesgue measure. In the total variation norm there may, however, still be convergence, and the dependencies between the $\tilde{\psi}^{(k)}$ variables indicate that the convergence will be $O(N^{-1})$. Skare et al. (2003) also show that the pointwise convergence rate of IR without replacement is $O(N^{-1})$ and also give an expression for the constant in the leading term.

In this thesis, our concern is, however, to draw θ by IR and not x. But this may be thought of as drawing $\psi = (x, \theta)$ by IR without replacement and then ignoring x to retain θ only. Thus the theorem of Skare et al. (2003) applies to our case. To make their theorem

more transparent, we contrast IR without replacement with IR with replacement by a simple example.

For some particular i and i^* let $\pi(x, \theta \mid X_{-i}, Y) = \phi(x; 5, 1) \times \phi(\theta; 2, 1)$ and $\pi(x, \theta \mid X_{-i^*}, Y) = \phi(x; 0, 1) \times \phi(\theta; 2, 1.05)$, where $\phi(\cdot; \mu, \sigma)$ is the normal density with mean μ and standard deviation σ. Observe that although the marginal $\pi(x \mid X_{-i^*}, Y)$ poorly approximates $\pi(x \mid X_{-i}, Y)$, $\pi(\theta \mid X_{-i^*}, Y)$ approximates $\pi(\theta \mid X_{-i}, Y)$ well. To estimate $\pi(x, \theta \mid X_{-i}, Y)$ we apply IR to realised values of (x, θ) drawn from $\pi(x, \theta \mid X_{-i^*}, Y)$ with importance weight function $w(x, \theta) = \{\phi(x; 5, 1) \times \phi(\theta; 2, 1)\} / \{\phi(x; 0, 1) \times \phi(\theta; 2, 1.05)\}$.

We suppose that the interest is to estimate the probabilities (i) $P(2 < x < 3) = 0.0214$ and (ii) $P(\theta > 2) = 0.5$, both probabilities being with respect to $\pi(x, \theta \mid X_{-i}, Y)$. One way to proceed is to draw samples of (x, θ) from the correct density $\pi(x, \theta \mid X_{-i}, Y)$ and compute an estimate using the samples. We however do this by drawing (x, θ) from the importance sampling density $\pi(x, \theta \mid X_{-i^*}, Y)$ and use IR to sample approximately from $\pi(x, \theta \mid X_{-i}, Y)$.

For both probabilities (i) and (ii) we can envisage two estimators, (a) $\hat{p}_1$ and (b) $\hat{p}_2$ corresponding to IR with and without replacement respectively. We expect to illustrate the theorem of Skare et al. (2003) by showing that IR without replacement performs better than IR with replacement. We, however, remark that for (i) both estimates $\hat{p}_1$ and $\hat{p}_2$ are expected to be quite biased since the marginal of x with respect to the target density is poorly approximated by the marginal of x with respect to the importance sampling density. Thus we do not recommend estimating characteristics of x using IR but we certainly recommend doing the same in the case of θ since the marginal is well-approximated by the marginal of the importance sampling density. But in both cases it will be demonstrated that IR without replacement performs better than IR with replacement. This is in accordance with the theorem of Skare et al. (2003).

Here we take the initial sample size from $\pi(x, \theta \mid X_{-i^*}, Y)$ to be $N = 100$. Clearly, there would occur extremely large importance weights. We computed estimates of means $(\hat{p}_1)$, standard errors $(\hat{SE})$ and mean square errors $(\hat{MSE})$ of the estimators based on 1,000

replications. Results corresponding to estimation of $P(\theta > 2) = 0.5$ are presented in Table 2.2 and those corresponding to $P(2 < x < 3) = 0.214$ are given in Table 2.1. Note that for (i) (that is, $P(2 < x < 3)$), although both estimates $\hat{p}_1$ and $\hat{p}_2$ seem to be biased, $\hat{SE}$ and $\hat{MSE}$ of $\hat{p}_2$ are less than that of $\hat{p}_1$. In fact, unlike $\hat{p}_1$, the bias of $\hat{p}_2$ is steadily decreasing with subsample size K; we use $K = 2, 4, 6, 8, 10$.

The case of (ii), that is $P(\theta > 2)$ is in accord with the procedure recommended in this thesis. In this case, very clearly the bias of the estimators $\hat{p}_1$ and $\hat{p}_2$ is very much less. This is expected since the marginal of θ is well-approximated by the θ-marginal of the importance sampling density. In this case, $\hat{SE}$ and $\hat{MSE}$ of both the estimators are decreasing with K. Estimator $\hat{p}_2$ is undoubtedly better than $\hat{p}_1$ since the former has less $\hat{SE}$ and $\hat{MSE}$. We remark that both N and K are quite small in this context.

We also present in detail results of an experiment with an initial sample of size $N = 50,000$ and a subsample of size $K = 5000$.

The results of this experiment are presented in Figure 2.2. Panel (a) shows a plot of the cumulative normalized weights. Ideally, the weights should have been constant, with zero variance, in which case the plot would have been a straight line. But clearly this is not the case and the weights have high variability. Panels (b) and (c) compare the exact and the IR (with replacement)-approximated densities of x and θ respectively. Clearly, both densities are very poorly approximated; in fact, as seen in panel (a), they are characterized by many repetitions of some points that have high weights. In panels (b) and (d) (note that (d) uses without replacement) samples are not even drawn from a major part of the region to which the true distribution gives high probability. This is expected since $\pi(x \mid X_{-i^*}, Y) = \phi(x; 0, 1)$ clearly is a poor approximation of $\pi(x \mid X_{-i}, Y) = \phi(x; 5, 1)$. But the without replacement strategy in the case of θ, shown in panel (e) performs very well although the variance of the weights is very high.

We remark in this context that binomial and hypergeometric distributions represent with and without replacement strategies respectively, albeit with equal weights. If a random

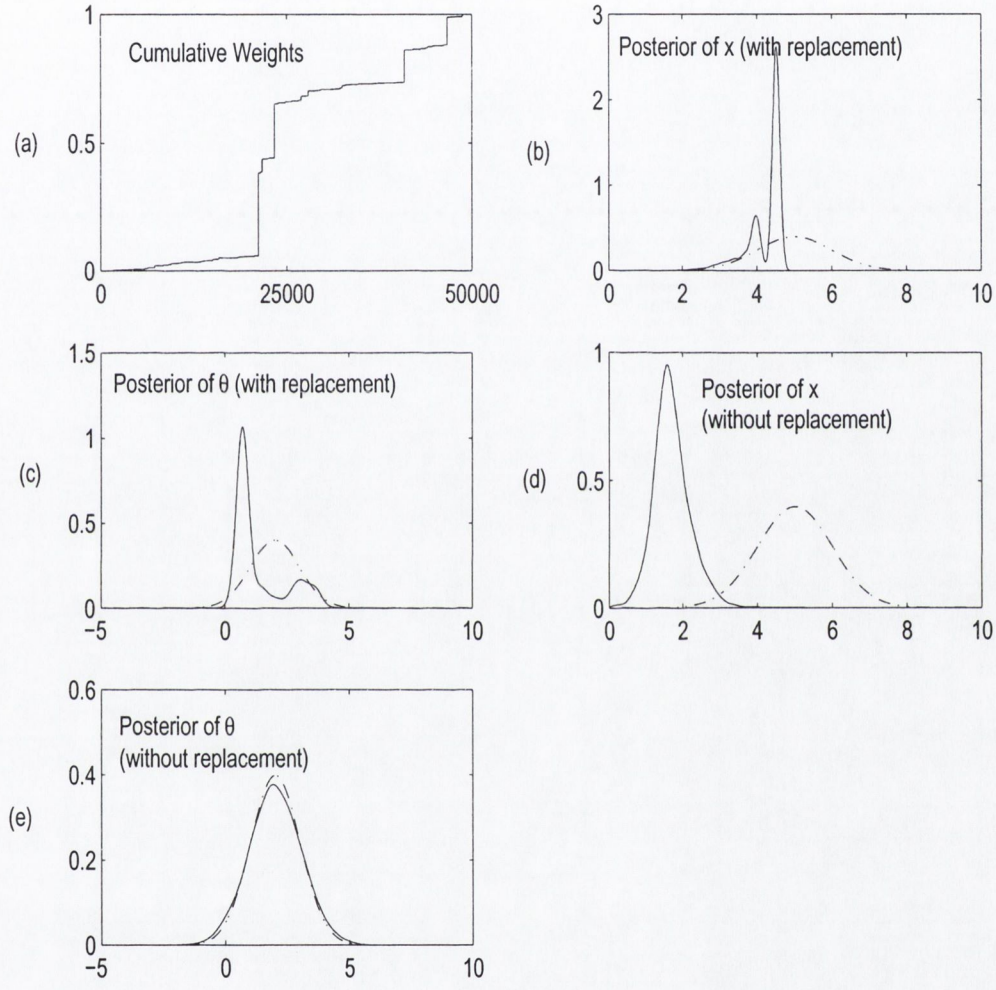


Figure 2.2: Demonstration of IR with and without replacement strategies.

Table 2.1: Assessment of the performance of IR with and without replacement in estimation of $P(2 < x < 3) = 0.0214$.

	With Replacement			Without Replacement		
K	$\hat{p}_1$	$\hat{SE}(\hat{p}_1)$	$\hat{MSE}(\hat{p}_1)$	$\hat{p}_2$	$\hat{SE}(\hat{p}_2)$	$\hat{MSE}(\hat{p}_2)$
2	0.657	0.427	2.767	0.608	0.355	2.658
4	0.676	0.390	2.762	0.464	0.280	2.462
6	0.649	0.392	2.729	0.340	0.222	2.338
8	0.683	0.380	2.763	0.266	0.175	2.278
10	0.663	0.380	2.738	0.209	0.144	2.243

sample of size K is drawn from a population of size N, of which a proportion p bears an attribute of interest, then the number of such attribute-bearers follow the hypergeometric distribution; see, for example, Stuart and Ord (1986). The mean value in this case is given by p, that is the sample proportion is an unbiased estimator of the population proportion. The variance of the sample proportion is given by $\frac{N-K}{N-1}\frac{p(1-p)}{K}$, which differs from the binomial case by the factor $(N-K)/(N-1)$ which is less than one for $K > 1$. Hence the variance in sampling without replacement is smaller than sampling with replacement. According Stuart and Ord (1986) this is intuitively obvious since extreme sampling possibilities are higher in the latter case.

We remark in this context that IR with replacement is asymptotically unbiased but unfortunately can have higher variance. On the other hand, IR without replacement has lower variance and is better than IR with replacement most especially when the importance weights are highly variable. However, the former is clearly biased. For example, in the case where $K = N$, IR without replacement is akin to the false assertion that the importance sampling density is the same as the target density.

From the above examples it is clear that realisations of θ may be re-used (without replacement is recommended), but not realisations of x. This observation is very much in accordance with the regular MCMC case described in Section 2.1.1 where it has been ob-

Table 2.2: Assessment of the performance of IR with and without replacement in estimation of $P(\theta > 2) = 0.5$.

	With Replacement			Without Replacement		
K	$\hat{p}_1$	$\hat{SE}(\hat{p}_1)$	$\hat{MSE}(\hat{p}_1)$	$\hat{p}_2$	$\hat{SE}(\hat{p}_2)$	$\hat{MSE}(\hat{p}_2)$
2	0.511	0.439	2.374	0.504	0.351	2.311
4	0.494	0.407	2.347	0.508	0.258	2.254
6	0.498	0.392	2.335	0.496	0.197	2.226
8	0.504	0.387	2.331	0.495	0.171	2.217
10	0.493	0.375	2.322	0.500	0.157	2.212

served that the proposal distribution for θ may be kept fixed in all cases but that for x may need to vary with each case. The rationale behind this is explained in detail in Chapter 5.

An alternative to leave-one-out cross-validation may be k-fold cross-validation. We review this next, borrowing heavily from Vehtari (2001).

2.4 k-fold cross-validation

Instead of leaving out just one observation at a time, it is also possible to omit several, say, approximately n/k observations at the same time and predict them with the help of the remaining $n - n/k$ observations. This can be repeated k times. Thus instead of dividing the data set into n groups, each consisting of a single observation, the data is partitioned into k groups, each group consisting of approximately n/k observations. Clearly, if $k << n$, huge computational savings can be achieved. Vehtari (2001) note that values of k between 8 and 16 seem to have good balance between the increased accuracy and increased computational load.

However, since k-fold cross-validation involves a smaller training data set than the complete training data set, estimates used for model validation based on this procedure are usually biased. See Vehtari (2001) for details. In leave-one-out cross-validation, where $k = n$, the bias is usually negligible. However, Vehtari (2001) demonstrate that k-fold cross-

validation may work better than leave-one-out cross-validation in cases where the data are dependent and it is more reasonable to leave out several data points at a time.

Vehtari (2001) note that in the case of time series with unknown finite range dependencies, k-fold cross-validation may be combined with the h-block cross-validation proposed by Burman et al. (1994). In h-block cross-validation, instead of leaving out the ith data point only, a block of h cases from either side of it is removed as well. The value of h depends on the dependence structure which could be estimated from autocorrelations. Burman and Nolan (1992) show that $h = 0$ could be used in the case of a stationary Markov process. However, exact properties of the process are usually unknown in real-life problems. Vehtari (2001) recommend k-fold cross-validation when computational methods for implementation of leave-one-out cross-validation fail. Since in this thesis we try to provide a reliable method for the implementation of leave-one-out cross-validation, we shall not pursue k-fold cross-validation.

In Chapter 3 we discuss an approach for assessing goodness-of-fit of inverse Bayesian models using reference distributions derived from leave-one-out posteriors of x.

Chapter 3

Cross-validation and model assessment in inverse problems using reference distribution approach

3.1 Introduction

The computation of the pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ is part of a process of model evaluation (and indeed selection). It is natural to ask (a) how to summarise these pairs to measure 'fit' and (b) how to determine whether the summarised measure indicates good fit. These two questions are the focus of this section. But our main contribution here is to (b), not (a). In this thesis we make no general claims about the appropriateness of any particular measure but only provide some examples. We anticipate that in general the choice of such measure will be dependent on the problem. Since the central contribution of this thesis is an efficient method of computing the pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$, given any inverse problem, by using the pairs an appropriate reference distribution of the summarised measure can be studied for model assessment purpose. Formal Bayesian decision theory can then be applied on the observed measure using the reference distribution. In this chapter, their use will be

demonstrated with simple problems.

We emphasise that in forward problems it is natural to compute pairs $\{y_i, \pi(y \mid X, Y_{-i})\}$ instead of $\{x_i, \pi(x \mid X_{-i}, Y)\}$. This in fact deals more directly with the probabilistic model of the forward part, denoted by $p(y \mid x, \theta)$; we will demonstrate with examples that the forward pairs $\{y_i, \pi(y \mid X, Y_{-i})\}$ provide more efficient model criticism than the inverse pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$. This is because the inverse pairs require an additional explicit prior specification on x, denoted by $\pi(x)$. Thus the probabilistic model $p(y \mid x, \theta)$ and the prior $\pi(x)$ are somewhat confounded in tests based on the inverse pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$.

To our knowledge there is no literature on tests based on the inverse part $\{x_i, \pi(x \mid X_{-i}, Y)\}$; however, model evaluation based on the forward part $\{y_i, \pi(y \mid X, Y_{-i})\}$ has found place in the literature. We discuss this below along with some other relevant ideas on model assessment. But in this thesis our focus is on $\{x_i, \pi(x \mid X_{-i}, Y)\}$ pairs since (a) we are concerned with applications of inverse problems and (b) for palaeoclimate problems, which motivate this thesis, the dimension of y is much greater than that of x; this makes summarisation of the x-part far more natural than summarisation of the y-part.

In what follows, we review some of the literature on model assessment in forward problems and, for inverse problems, we discuss different choices of the measure and describe how its reference distributions can be computed generically and cheaply. We illustrate the construction and use of reference distributions using variants of the Poisson regression model.

3.2 Overview of model assessment in forward problems

One approach for checking the fit of a model is by examining the marginal distribution of the data (Box (1980)). Specifically, if the marginal density of Y, defined by

$$p(Y \mid X) = \int p(Y \mid X, \theta)\pi(\theta)d\theta \tag{3.1}$$

is small, then Y is unlikely under the given model.

A problem with this approach is that $p(Y \mid X)$ is improper if $\pi(\theta)$ is. Since improper

priors are quite common in statistical practice, this means that the marginal may not have a probabilistic interpretation in a very wide range of statistical problems.

It is worth noting that computation of the marginal given by (3.1) is a difficult problem and very carefully chosen importance sampling densities will be needed; in general the estimates are unstable. Newton and Raftery (1994) suggest a combination of the prior and the posterior as an importance sampling density that ensures stability; but it has the undesirable requirement of simulation from both the prior and the posterior.

Gelfand et al. (1992) suggest an alternative approach, based on pairs $\{y_i, \pi(y \mid X, Y_{-i})\}$. The relevant distribution, called the conditional predictive distribution, is given by (2.1). Note that, unlike the marginal density $p(Y \mid X)$, the conditional predictive distribution is proper whenever the posterior $\pi(\theta \mid X_{-i}, Y_{-i})$ is. In addition, the collection of conditional predictive densities $\{p(y_i \mid X, Y_{-i}); i = 1, \cdots, n\}$ is equivalent to $p(Y \mid X)$ when both exist (Besag (1974)), which means that the former could be used even when the latter is undefined.

We note that the approach of Gelfand et al. (1992) is not easily applicable to inverse problems where, instead of $p(y_i \mid X, Y_{-i})$, interest should be on $\pi(x_i \mid X_{-i}, Y)$. This has been already explained in detail in Section 2.1. Moreover, the summaries proposed in Gelfand et al. (1992) appear to lack any proper foundational basis. Theoretical support has been attempted in Gelfand and Dey (1994) but everything breaks down when the number of parameters tends to infinity as the data size tends to infinity. In the case of Vasko et al. (2000) and Diggle et al. (1998) the number of parameters increase with the data size and such methods are not applicable there.

An alternative to the approach of Gelfand et al. (1992) is due to Rubin (1984) who proposed constructing discrepancy measures that attempt to measure departures of the observed data from the assumed model (likelihood and prior distribution). Gelman et al. (1995) stress that the choice of the discrepancy measure can (and should be) tuned to the aspect of the model whose fit is in question. In order to be computable in the classical framework, test statistics must be functions of the observed data alone. But Gelman et al. (1996) point out

that, for Bayesian model checking, generalised test statistics $D(Y, X, \theta)$ that depend on the parameters as well as the data may also be used.

One class of discrepancy measures is omnibus measures of fit. Gelman et al. (1995) recommend as an omnibus goodness of fit measure the statistic

$$D(Y, X, \theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i \mid x_i, \theta))^2}{Var(y_i \mid x_i, \theta)} \tag{3.2}$$

The distribution of this statistic can then be compared with that of $D(Y^*, X, \theta)$, where Y^* is a replication of the data with the same parameters that produced the data Y. In other words, Y^* is a simulation from the model using the same parameter θ that has given rise to the original data Y. The posterior predictive distribution of Y^* is defined as

$$p(Y^* \mid X, Y) = \int p(Y^* \mid X, \theta)\pi(\theta \mid X, Y)d\theta \tag{3.3}$$

Classical goodness of fit tests for the null hypothesis that the data Y come from the given model are based on (3.2) where θ would generally be replaced with its maximum likelihood or minimum chi-squared estimate. There are analytic results in classical statistics establishing the asymptotic distributions as chi-squared under the null hypothesis; see, for example, Rao (1965). The posterior predictive distribution provides a suitable reference distribution for any sample size (Stern and Cressie (2000)).

A convenient summary measure of the extremeness of $D(Y, X, \theta)$ with respect to $D(Y^*, X, \theta)$ is the tail area,

$$\begin{aligned} p_D &= P\{D(Y^*, X, \theta) > D(Y, X, \theta) \mid X, Y\} \\ &= \int P\{D(Y^*, X, \theta) > D(Y, X, \theta) \mid X, \theta\}\pi(\theta \mid X, Y)d\theta \end{aligned} \tag{3.4}$$

Observe that in the case where the distribution of $D(Y^*, X, \theta)$ is independent of θ, p_D is exactly equal to the frequentist p-value. As such, (3.4) is sometimes called the Bayesian p-value (Carlin and Louis (1996)). Meng (1994) stresses that it is important to remember that the use of (3.4) is not advocated in model choice: p_D is not the probability that the model is correct, and that Bayesian p-values should not be compared across models. Rather,

they serve only as measures of discrepancy between the assumed model and the observed data, and hence provide information concerning model adequacy.

However, Carlin and Louis (1996) note that this model checking strategy uses the data twice: once to compute the observed statistic $D(Y, X, \theta)$ and again to obtain the posterior predictive reference distribution. Bayarri and Berger (1999) demonstrate with examples that using data twice is undesirable. Specifically, in such cases, even with arbitrarily strong evidence against the null model, the p-value does not tend to zero. Also, the posterior predictive p-values do not generally have a uniform distribution under the null hypothesis, not even asymptotically (Bayarri and Berger (1999), Robins et al. (1999)). Bayarri and Berger (1999) have developed a related approach based on posterior distributions that condition on only part of the information in the data rather than using the full posterior distribution (saturated posterior distribution) to define the reference distribution. Their p-values are uniformly distributed under the null hypothesis and are also not as conservative as the posterior predictive p-values. However, Stern and Cressie (2000) point out that their approach requires more calculation and can be quite difficult to apply for the kinds of complex models that are most challenging to check in practice.

We now introduce our proposal of contructing suitable reference distributions using pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ and attempt to address question (b) in the context of inverse problems.

3.3 A new procedure for model assessment in inverse problems

For notational convenience, in this section, we denote by $\hat{x}_i$, the random variable corresponding to $\pi(\cdot \mid X_{-i}, Y)$.

Using an inverse cross-validation approach, we first simulate, for each $i = 1, \cdots, n$, N realisations from the distribution $\pi(\hat{x}_i \mid X_{-i}, Y)$. Let the simulated values be denoted by $\{\hat{x}_i^{(1)}, \cdots, \hat{x}_i^{(N)}\}$. The simulation will typically done by IRMCMC; see Chapter 4. We define

an observed discrepancy measure D^{obs}, that measures the discrepancy between observed and predicted values. In our case, the observed values are $\{x_i\}$. We usually take predicted values as $\{\tilde{x}_i\}$, where $\tilde{x}_i$ is the mode of the distribution $\pi(\cdot \mid X_{-i}, Y)$. The fact that the distributions $\pi(\hat{x}_i \mid X_{-i}, Y)$ could be multimodal motivated us to use the posterior mode instead of the posterior expectation of $\hat{x}_i$. However, we do not claim optimality of the mode as a summary of the posterior distribution. Some examples of D^{obs} are as follows:

$$D_1^{obs} = \sum_{i=1}^{n} \frac{(x_i - \tilde{x}_i)^2}{Var(\hat{x}_i \mid X_{-i}, Y)} \tag{3.5}$$

$$D_2^{obs} = \sum_{i=1}^{n} \frac{|x_i - \tilde{x}_i|}{\sqrt{Var(\hat{x}_i \mid X_{-i}, Y)}} \tag{3.6}$$

$$D_3^{obs} = \max_{1 \le i \le n} \left\{ \frac{|x_i - \tilde{x}_i|}{\sqrt{Var(\hat{x}_i \mid X_{-i}, Y)}} \right\} \tag{3.7}$$

$$D_4^{obs} = x_i \tag{3.8}$$

Corresponding to each of these observed discrepancy measures, generically denoted by D^{obs}, we also define discrepancy variables, generically denoted by D^{var}, where observed x_i in each of (3.5), (3.6) (3.7) and (3.8) is replaced by $\hat{x}_i^{(j)}; j = 1, \cdots, N$.

We make no argument on the merits and demerits of the above discrepancy measures. However, note that while measures D_1^{obs}, D_2^{obs} and D_3^{obs} provide summaries of distances between the observed values x_i and the corresponding modes of the leave-one-out posteriors $\pi(x \mid X_{-i}, Y)$, the measure D_4^{obs} is just the observed value for case i and thus is different from all other measures in the sense that it is not an overall measure of fit. Rather it provides insight specifically into the case i. For example, it can be used to check whether or not x_i is an outlier with respect to the underlying model. In this context we note that there may exist measures corresponding to which no reference distribution may be easily available. For instance, a measure D_5^{obs} may be defined as the number of x_i that fall within the 95% credible region of the corresponding leave-one-out posteriors. In this case there seems to exist no easily computable reference distribution.

Note that D^{var} has a distribution dictated by the n posterior distributions $\{\pi(\hat{x}_i \mid$

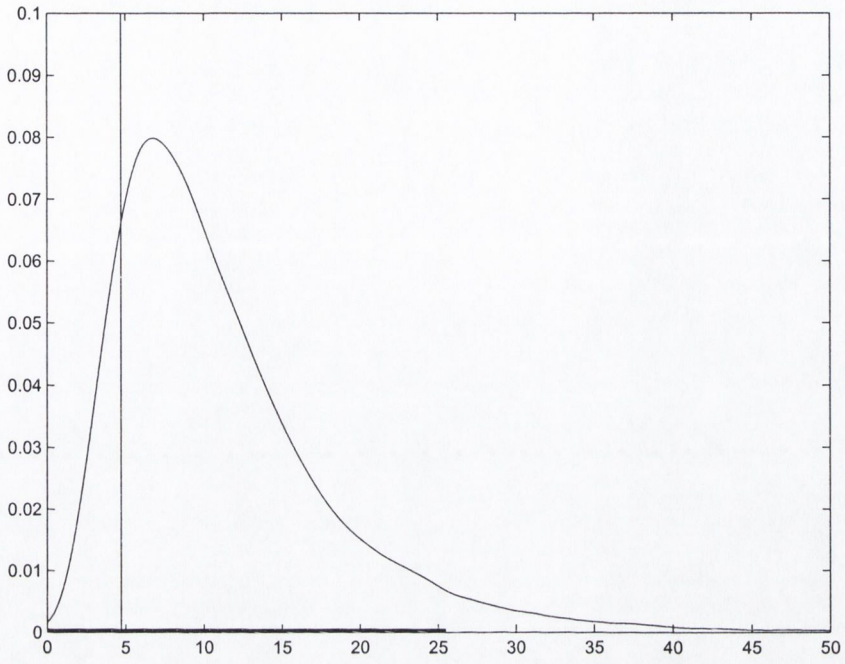


Figure 3.1: Demonstration of goodness of fit test with the artificial Poisson regression data at 10 sites; density of D_1^{var} shown, the vertical line denotes D_1^{obs}. Thick horizontal line denotes 95% credible region.

$X_{-i}, Y); i = 1, \cdots, n\}$. This again determines the joint posterior distribution of all $\hat{x}_i$'s given by $\pi(\hat{x}_1, \cdots, \hat{x}_n \mid Y)$. We denote the distribution of D^{var} by $\pi^*(\cdot \mid Y)$ and refer to the approach as the reference distribution approach. The model will be said to fit the data if D^{obs} is consistent with the reference distribution $\pi^*(D^{var} \mid Y)$. In other words, D^{obs} falling within the appropriate credible region of $\pi^*(D^{var} \mid Y)$ will indicate adequate fit of the underlying model to the data.

For an illustration of what is involved, we consider the Poisson regression example. In this context, Figure 3.1 shows the density of D_1^{var} and the corresponding observed D_1^{obs}, shown by the vertical line. Observe that D_1^{obs} falls within 95% credible region of D_1^{var}, as would have been expected, since the data arise from a known model. This indicates that the model fits the data well. Similarly, both D_2^{obs} and D_3^{obs} indicated that the model fits the data well. The measure D_4^{obs} is simply the observed value at a particular case and the

reference distribution D_4^{var} is simply the leave-one-out posterior for that case. Using this we have already found in Chapter 2 that for all 10 cases the observed value x_i fell within the 95% HPD regions of the corresponding leave-one-out posteriors. Hence all D_1^{obs}, D_2^{obs}, D_3^{obs} and D_4^{obs} indicate correctly that the underlying model fits the data.

3.4 Merits and demerits of the reference distribution approach

Our methodology is simple and we have attempted to provide guidelines when to accept or reject the model in question. It is also straightforward to extend our procedure to multivariate situations. However, this does not seem important for the applications we consider in this thesis (a bivariate situation does arise in Chapter 7, but marginal analysis leads to clear-cut conclusions) and hence we restrict ourselves to univariate cases only.

The key point of our proposal is that it does not recommend acceptance or rejection of a model by noting the magnitude of an observed discrepancy measure alone. This magnitude, combined with the sampling distribution of the discrepancy measure (the reference distribution), helps make the decision in a consistent manner. Indeed, the ability to construct a reference distribution for any appropriate discrepancy measure is the main novelty of our proposed method. It will be shown in Section 3.6 that assessing goodness-of-fit based on observed discrepancy measure alone may be undesirable.

Our approach is particularly useful in inverse problems, especially in problems where the dimensionality of y is much greater than the dimensionality of x. In such cases, summarising the x-part is far more easy than summarising the y-part. In the problems of Vasko et al. (2000) and Whiley et al. (2004) the dimensionalities of $\{x, y\}$ are $\{1, 52\}$ and $\{2, 14\}$ respectively and in such cases our approach seem very appropriate.

An important point to note is that there are no parameterisation problems in our approach with reference distributions since all parameters other than x are integrated out. No

asymptotic theory, of any sort, is needed to make this approach work. Also note that we do not use p-values, Bayesian or otherwise, in our approach. The problem of using data twice is avoided here by adopting the leave-one-out cross-validatory approach.

Computation of the pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ for each i needed to compute the observed discrepancy measure and the corresponding reference distribution appears very demanding at first sight, particularly if there are a large number of cases. However, the central contribution of this thesis is IRMCMC, a method that seems to be highly suitable for computing the pairs cheaply and efficiently. We recommend IRMCMC for the computational needs of this model assessment proposal in inverse problems. It is also clear that our procedure can be applied to forward problems. In fact, application to forward problems may be computationally cheaper, since typically MCMC will not be needed to simulate from $p(\cdot \mid x_i, \theta)$. But we emphasise that in this thesis, inverse, not forward, problems are of interest.

A possible drawback of this is that it seems difficult to decide which discrepancy measure to use. We are not aware of any literature that discusses this issue in the context of inverse problems; however, in the forward context see Meng (1994), Box (1980), Rubin (1984), Gelman et al. (1993), Gelman et al. (1996). However, the choice might be application specific (see, for example, Stern and Cressie (2000) in the forward context) and rather than a drawback this could be viewed as flexibility offered by this approach. Since we are not interested in the question concerning the most appropriate measure, most of our results will be based on (3.5), only because similar measures has been widely used in the forward cases, both in the Bayesian and the classical statistical literature. See, for example, O'Hagan and Forster (2004) and the references therein.

It has been argued that the reference distribution approach in inverse problems, which is the main contribution in this chapter, has some desirable properties and that the computational challenge involved may be overcome by IRMCMC. We now illustrate the approach by applying it on various problems involving repeated computer-simulated data and mainly noting the percentage of times it gives the correct answer. However, we acknowledge that

since we obtain only point estimates of true percentages, our evaluation procedure may not be completely adequate.

In the following illustrations we emphasise that experimental evaluation sheds more light on the particular choice of discrepancy measure. Even on any particular choice of discrepancy measure experimental replications can shed limited light. But since here we are concerned with simulation studies, where the true models and their properties are completely known, we may suppose that the point estimates provide useful evidence on the general performance of our approach based on reference distributions. Besides, we provide other relevant experimental details to supplement the inadequacy of the point estimates.

3.5 Illustration of inverse model assessment with the reference distribution approach

We now demonstrate our proposed methodology with four examples, the first being a forward problem. But we consider this foward regression problem mainly to contrast it with later examples on inverse regression problems. The forward example concerns the problem when the data actually comes from a Geometric distribution but is modeled as Poisson distribution; we assume that in this case our interest is in checking the forward aspect of the model, that is, if the 'y'-part is modeled correctly. Here we shall use pairs $\{y_i, \pi(y \mid X, Y_{-i})\}$.

The second example is the 'inverse' of the first problem, that is, it concerns the 'x'-part of the data, using pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$, other facts remaining the same as in the first example. Since in this case the interest is in the inverse, it is required to put a prior on x. It will be argued that checking model assumptions on y using pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ is less efficient than $\{y_i, \pi(y \mid X, Y_{-i})\}$.

The third example concerns an inverse problem where it is assumed that the forward aspect of the model is guessed correctly. Here goodness-of-fit test is akin to testing the prior on x. In this case the pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ seem natural.

In the fourth example we discuss our methodology in the context of a variable selection problem, assuming that the forward part has been guessed correctly but that the inverse part has been incorrectly modeled; the true model form being quadratic but modeled as linear or cubic.

In none of our examples do we claim optimality of any particular discrepancy measure. Throughout all illustrations the results based on the discrepancy measure D_1^{obs} and the corresponding reference distribution D_1^{var} will be presented.

Example 1: Forward regression

We consider a problem where the data actually comes from a Geometric distribution but has been modeled in reality as involving the Poisson distribution. In other words, given $x_1, \cdots, x_{10}$, which are drawn randomly from $U(1,2)$ (that is, uniform distribution on the interval $(1,2)$), data $y_i \sim$ Geometric(p_i), where $p_i = 1/(1+\theta x_i)$. A sample data set given $\theta = 1$ is shown in Figure 3.2. It is assumed, for purposes of illustration, that the data has been modeled as Poisson(θx_i). A uniform improper prior has been put on θ, that is, $\pi(\theta) = 1$; $\theta > 0$.

Note that, had y_i been Poisson(θ), then $E(y_i) = \theta x_i = Var(y_i)$. But for the Geometric case, $E(y_i) = \theta x_i$ but $Var(y_i) = E(y_i)(1+\theta x_i)$. In this example $x_i \in (1,2)$ and $\theta > 0$. Since x_i are bounded, for θ close to zero $Var(y_i) \approx E(y_i)$ and we can expect Poisson and Geometric distributions to agree. However, if θ is large, then $Var(y_i) >> E(y_i)$ and the two distributions are expected to disagree.

Sample reference distributions and observed discrepancy measures for $\theta = 0.1$ and $\theta = 15$ are shown in Figure 3.3 and 3.4 respectively. From the figures it is clear that Geometric and Poisson both fit the model well in the case where $\theta = 0.1$ but the Poisson distribution does not fit the data when $\theta = 15$. We also considered 1000 simulations from the Geometric distributions with the above set-up with different values of θ and applied our methodology in each case to assess the goodness-of-fit of the Poisson model to the Geometric data. Subsequently the true model has also been applied on the data to contrast with the fit achieved

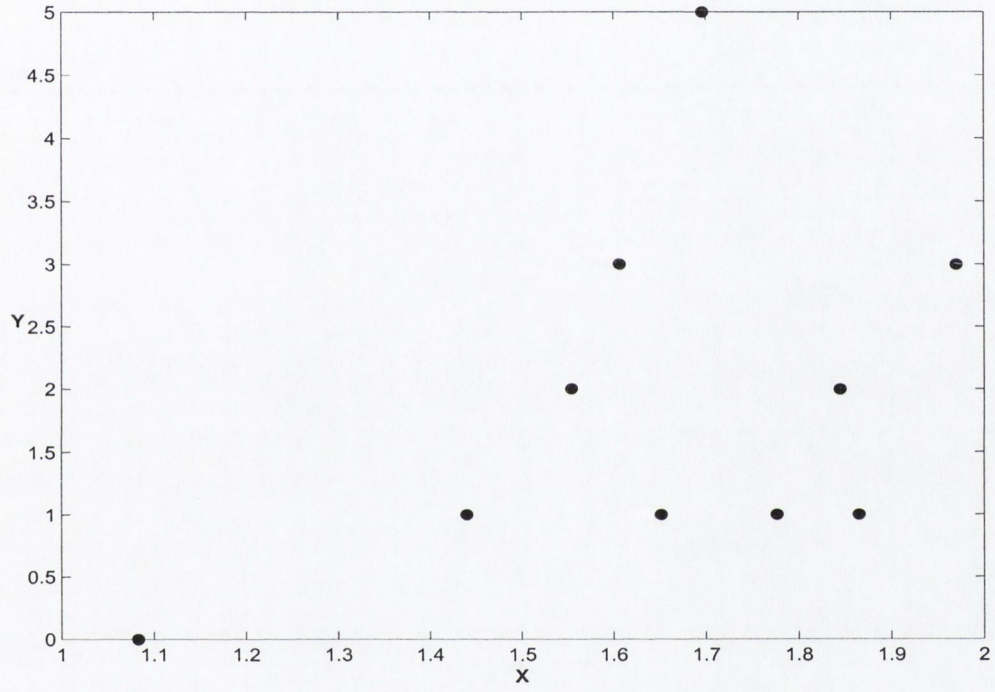


Figure 3.2: Data arisen from the Geometric model: $y_i \sim \text{Geometric}(p_i)$, where $p_i = 1/(1 + \theta x_i)$. Here $\theta = 1$.

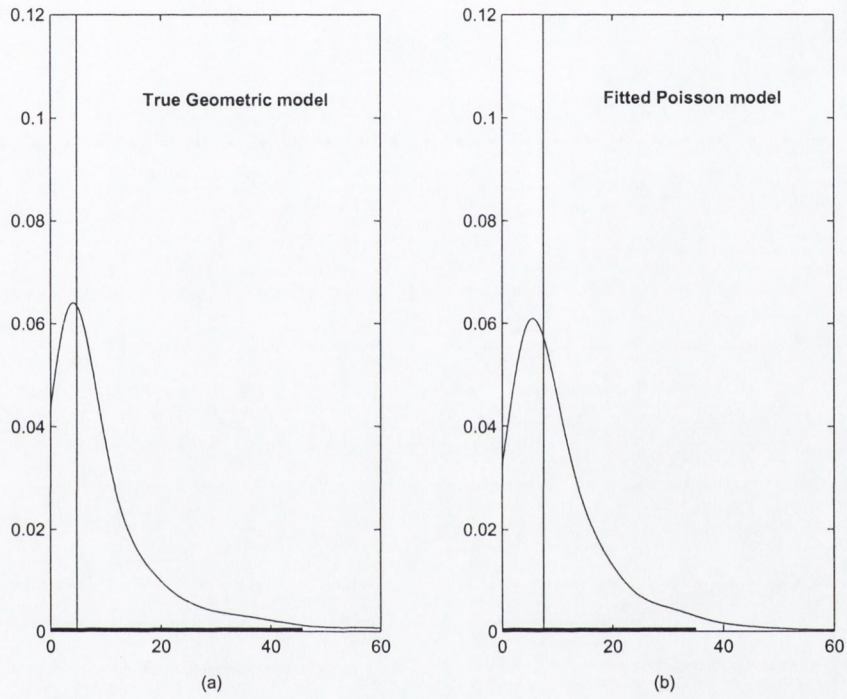


Figure 3.3: Result of goodness-of-fit test when true model is Geometric with $\theta = 0.1$. Reference distributions are shown in panels (a) and (b); vertical lines denote observed values. Since the observed values in both cases lie within the approximately 97% credible regions (denoted by the thick horizontal line), both Poisson and Geometric fit the data well.

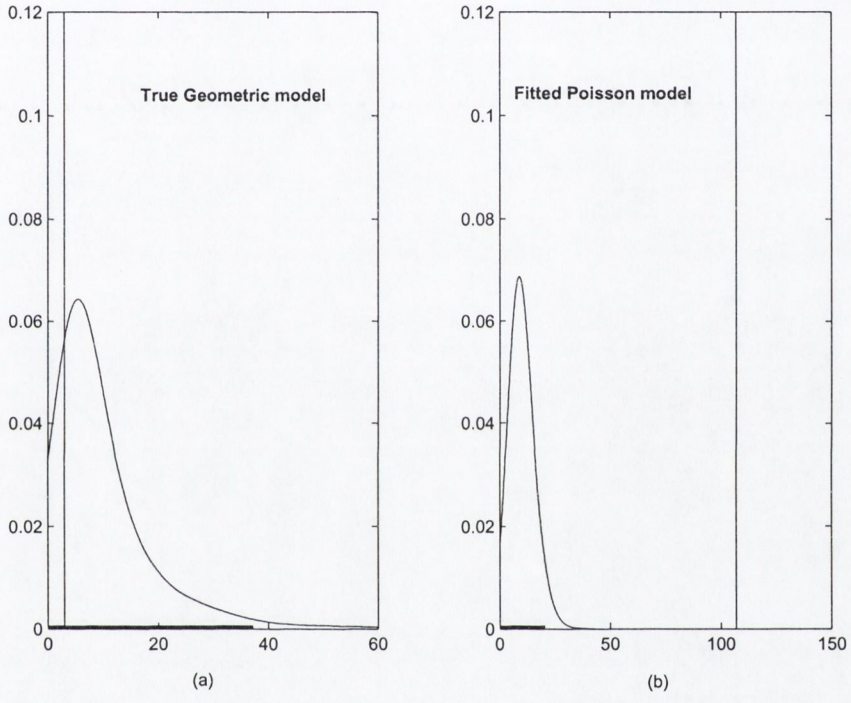


Figure 3.4: Result of goodness-of-fit test when true model is Geometric with $\theta = 15$. Panel (a) suggests that the Geometric distribution fits the data. But panel (b) suggests that the Poisson distribution does not fit the data.

Table 3.1: Forward problem: assessment of Poisson model fit when the true model is Geometric.

Parameter(θ)	Poisson agreement (%)	Geometric agreement (%)
0.1	97.0	99.7
1.0	59.7	99.0
3.0	13.2	98.8
5.0	3.3	98.5
7.0	0.8	99.1
15.0	0.0	97.5

by the Poisson model. The results are given in Table 3.1. For example, for $\theta = 0.1$ and the true model is Geometric, the erroneous Poisson model is accepted 97% times (false positive) and the true Geometric model is rejected 0.03% times (false negative).

In Example 2 we will contrast this with the inverse case.

Observe that, as θ increases, the Poisson model agrees less and less with the Geometric model. This is because the mean and variance of the Geometric distribution drift apart as θ increases. In fact, the percentages of agreement by the Poisson model decrease quite fast. It will be pointed out that in the inverse case the decrease is relatively slow in comparison. Note that when the Geometric model is applied to the data it fits the data very well in all cases. This is to be expected since it is the true model. In the inverse case it will be seen that the percentages of agreement by the Geometric model are comparatively slightly less.

Example 2: Inverse regression

In the first example we considered a problem involving the Geometric distribution as the true model but modeled as Poisson distribution. There assessment of the model fit used pairs $\{y_i, \pi(y \mid X, Y_{-i})\}$. In this example we consider the same problem but now we focus on the pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ instead. We remind the reader that herein lies our interest. In contrast to the previous forward example, here we need to put a prior on x (in addition to the prior on θ, which we assume the same as in the previous example). We put the correct prior on x; that is $x \sim U(1, 2)$ (recall that in Example 1 x has been drawn randomly from

$U(1,2)$).

We describe the experiment in detail contrasting results with the forward case, whenever applicable. Below we present the details of our experiments, elaborating on the two cases, namely, $\theta = 15$ and $\theta = 0.1$.

Case 1: θ=15

Results of a sample experiment is presented in Figure 3.5.

In the case of the Geometric distribution the observed discrepancy measure falls within the approximately 97% credible interval of the reference distribution suggesting good fit. This is not the case for the Poisson distribution, clearly suggesting that the Poisson model does not fit the data. Figure 3.6 shows that the leave-one-out posterior of x corresponding to the Geometric model gives high density to the observed value of x but the same observed value is given low density by the leave-one-out posterior with respect to the Poisson distribution.

We considered 1000 simulations from the Geometric distributions with the above set-up and applied our methodology in each case to assess the goodness-of-fit of the Poisson model to the Geometric data. The Poisson model could fit the Geometric data only 1.4% times. However, when the Geometric model is assumed, agreement took place 97.6% times. This is clearly to be expected since the data arise from the Geometric distribution.

We now discuss the situation where Poisson and Geometric models agree.

Case 2: θ=0.1

We present below results of a sample experiment.

As seen in Figure 3.7 it seems clear that Geometric and Poisson models both fit the data. In fact, with repeated simulations in this case, both Poisson and the Geometric model fitted the data well 97.3% times.

Figure 3.8 shows that the leave-one-out posterior of x under the true Geometric model and the fitted Poisson model are very similar.

Apart from the two detailed cases we reported with $\theta = 0.1$ and 15, we provide in

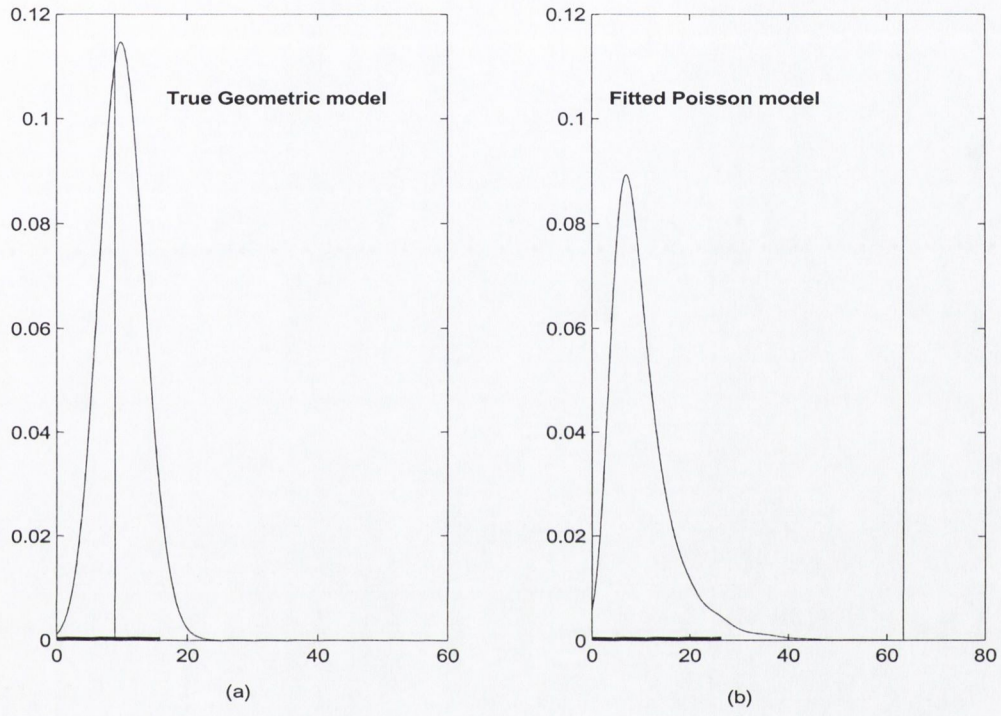


Figure 3.5: Result of goodness-of-fit test when true model is the Geometric distribution with $\theta = 15$. Panel (a) shows the observed discrepancy measure (vertical line) and the reference distribution corresponding to a Geometric distribution fitted to the data. The observed value is included in the approximately 97% credible interval (denoted by the thick horizontal line). The observed discrepancy measure and the reference distribution given the Poisson distribution is shown in (b). Since the vertical line does not fall within the thick horizontal line, we conclude that Poisson distribution does not fit the data.

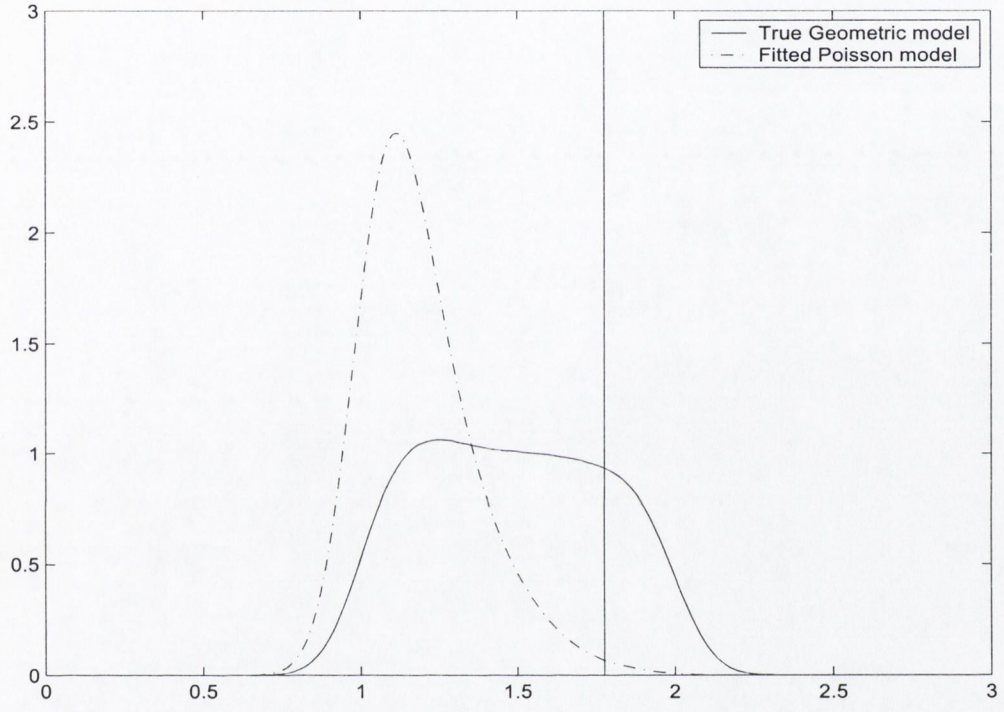


Figure 3.6: Leave-one-out posterior of x for a specific case corresponding to the Geometric distribution gives higher density to the observed value (denoted by the vertical line) compared to the Poisson distribution. Here Geometric distribution (with $\theta = 15$) is clearly more appropriate than the Poisson distribution.

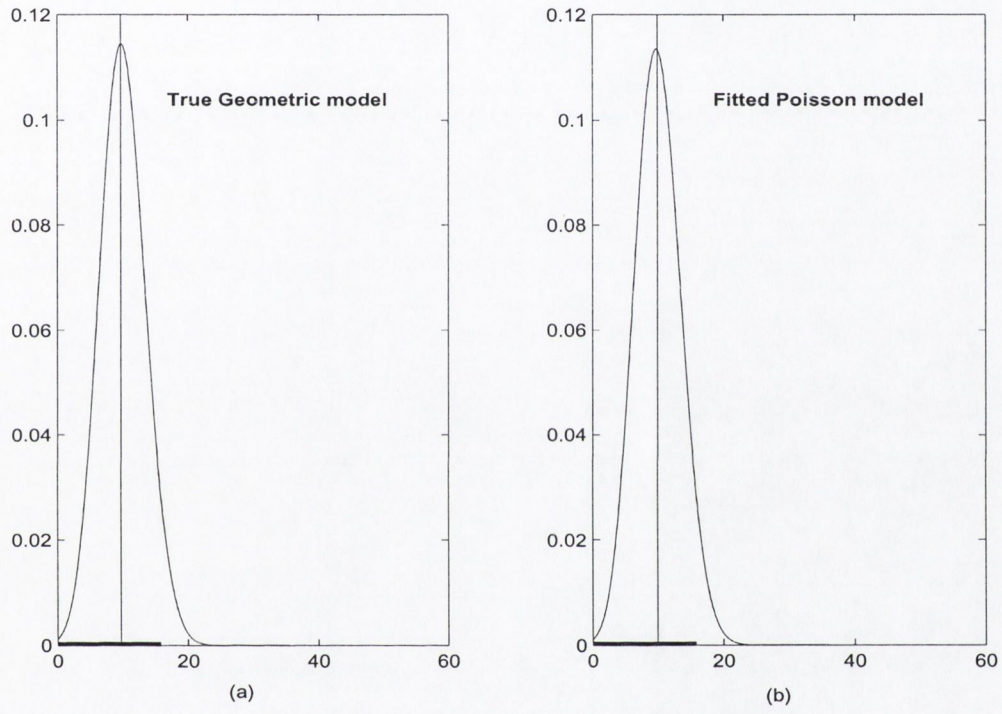


Figure 3.7: Result of goodness-of-fit test when true model is the Geometric distribution with $\theta = 0.1$. Panels (a) and (b) show that both Geometric and Poisson model fit the data well.

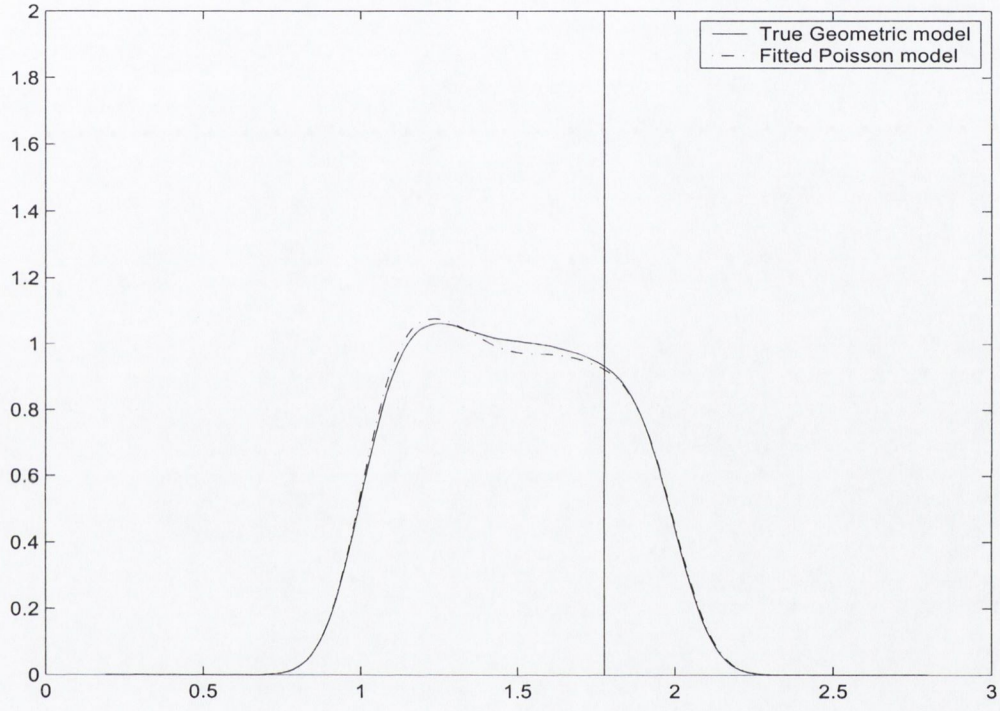


Figure 3.8: Leave-one-out posteriors of x for a specific case when the true model is Geometric with $\theta = 0.1$. Both posteriors, which are very similar, give high density to the observed datum (denoted by the vertical line).

Table 3.2: Inverse problem: assessment of Poisson model fit when the true model is Geometric.

Parameter(θ)	Poisson agreement (%)	Geometric agreement (%)
0.1	97.3	97.3
1.0	89.3	97.1
3.0	63.5	97.5
5.0	34.7	97.7
7.0	18.4	97.8
15.0	1.4	97.6

Table 3.2 abridged results of varying the parameter θ. The table clearly shows that Poisson disagrees more and more with the Geometric model as θ, and thus the difference between mean and variance of the Geometric model, increases. In other words, the percentage of false positives decreases fast as θ increases. On the other hand, the percentage of false negatives do not show any appreciable change with θ. But here one must note the contrast between Table 3.1, of the forward case, and Table 3.2, corresponding to the inverse case. In the former table the percentage of disagreement of the Poisson model with the Geometric data increases faster than in the table corresponding to the inverse case. Also, the percentages of agreement of the Geometric model with the true Geometric data is higher in the forward case. All these observations suggest that it is the forward problem that is more suited to model checking in the forward 'y'-part. This is very clearly in accordance with the discussion in Section 3.1.

This is because the prior on x confounds issues regarding the forward part. This has also been discussed in Section 3.1.

We now introduce an example to check the prior assumption on x in an inverse problem.

Example 3: Inverse regression – prior on x

For $i = 1, \cdots, 10$, data y_i come from the true model Poisson(θx_i). Data $X = \{x_i; i = 1, \cdots, 10\}$ are drawn randomly from an exponential distribution with mean λ. The parameter

Table 3.3: Assessment of inverse model fit.

Exponential mean(λ)	Agreement percentage
0.5	56.0
1.00	74.4
3.00	90.4
10.00	95.4

θ is selected randomly from the interval $(0, 1)$.

Given the above set up we now assume that it is known to us that $y_i \sim \text{Poisson}(\theta x_i)$, but that the (prior) distribution of x_i (which is actually exponential) is unknown. We also suppose that our ultimate interest lies in the prediction of x, and not y. As a result, here we are interested in checking the inverse(or x) aspect of the model. We test whether a uniform improper prior is appropriate for x.

We evaluate our approach with several different true values of λ. Note that for an exponential distribution with mean λ, the variance is λ^2; since uniform improper priors can be said to have infinite variance, we can expect the fitted model to agree with the true model when λ is large and disagree when λ is small. We summarise our findings in Table 3.3.

It is seen in the above table that the percentage of times the fitted model agrees with the true model is increasing with the value of λ. That this is not surprising is discussed above.

Figure 3.9 shows that as λ increases, the leave-one-out posterior of x with exponential prior with mean λ converges to the leave-one-out posterior of x with the uniform prior.

Example 4: Variable selection

In addition to the above three examples, we have also conducted a variable selection study, assuming the true model to be Poisson with mean $\theta = \theta_1 x_i + \theta_2 x_i^2$. Here the true values of θ_1 and θ_2 are 0.5 and the x_i were drawn randomly from $U(0, 10)$. As in the previous examples, here also we will present results based on D_1^{obs} and D_1^{var} only.

Given the above set up, three cases: (a) $\theta = \theta_1 x_i$; (b) $\theta = \theta_1 x_i + \theta_2 x_i^2$ and (c) $\theta = \theta_1 x_i + \theta_2 x_i^2 + \theta_3 x_i^3$. Clearly, except (b), others are incorrect.

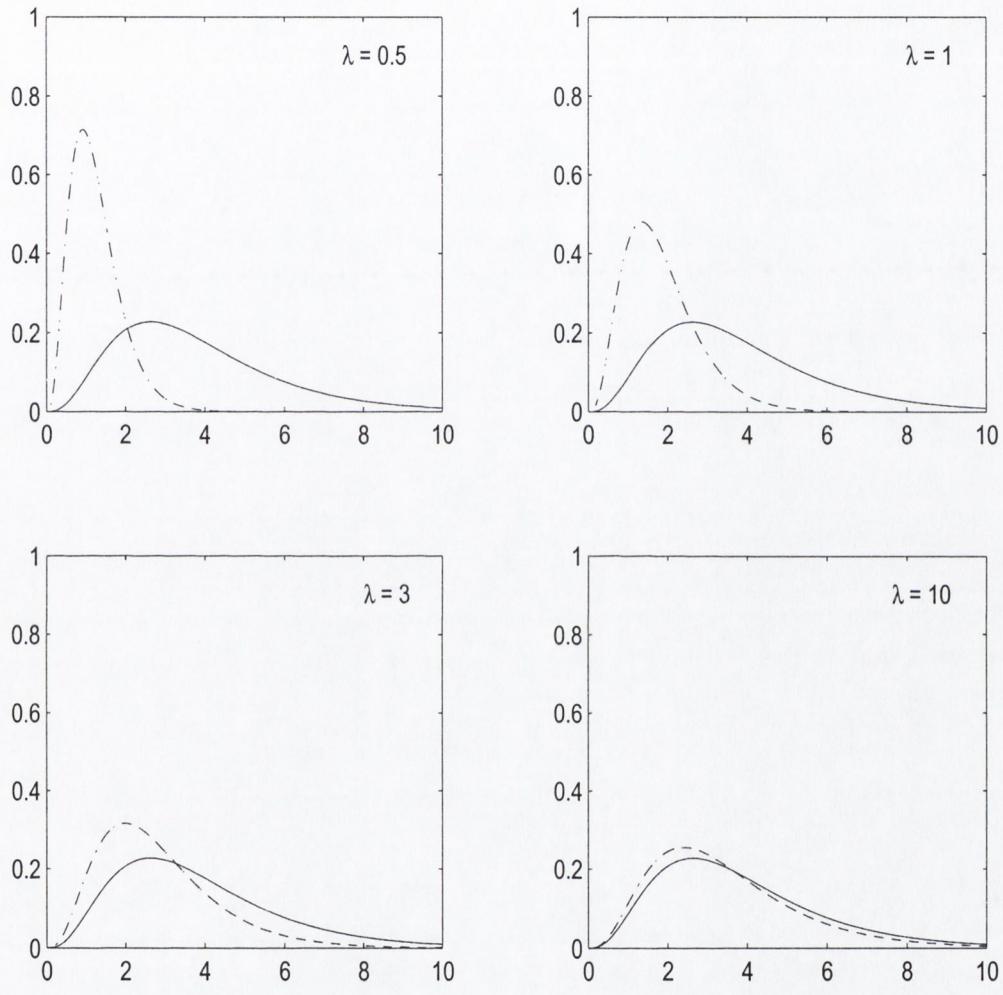


Figure 3.9: Leave-one-out posterior of x for a specific case as λ varies (dot-dashed line) compared with the posterior of x with uniform prior (in other words, $\lambda \to \infty$) (solid line). The two distributions agree more and more as λ increases.

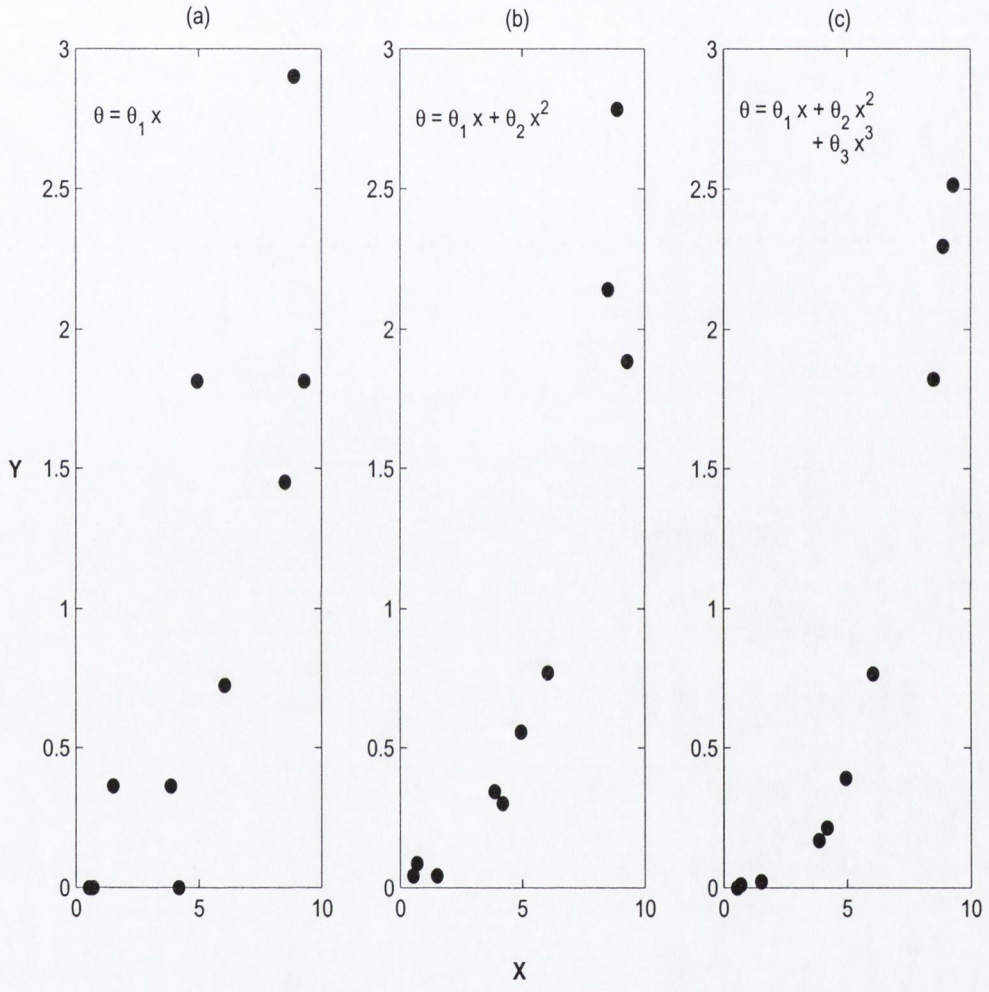


Figure 3.10: Scatter plots corresponding to models (a), (b) and (c) presented in panels (a), (b) and (c) respectively.

Displayed in Figure 3.10 are scatter plots corresponding to the above three models. Although the three plots are different, some similarity is exhibited between panels (b) and (c) which correspond to the quadratic and the cubic models respectively. It will be seen in our analysis that the two models indeed may agree.

For each of the three models (a), (b) and (c), with the simulation procedure repeated 1000 times, we implement our approach based on D_1^{obs} and D_1^{var} by simulating from the leave-one-out posteriors $\pi(x \mid X_{-i}, Y)$, corresponding to uniform priors for all variables. Case (b) was adjudged the correct model 95% times, cases (a) and (c) agreed with the true model 39% and 84% times respectively.

It is not at all surprising that (c) turns out to be far better than (a); this is because (a) wrongly assumes that $\theta_2 = 0$ but (c) does not neglect the quadratic term. In fact, in addition, (c) considers an extra cubic term. Noting that the true model (b) can be written as $\theta = \theta_1 x_i + \theta_2 x_i^2 + 0 \times x^3$, the true value of θ_3 in (c) can be said to to be zero.

Figure 3.11 displays the leave-one-out posteriors of x corresponding to each model, for each of the 10 data points, given a sample simulation case.

The figure clearly shows that although (b) and (c) agree well, (a) diagrees with them for many cases. The reference distributions of the discrepancy measures corresponding to the three models and their respective observed discrepancies are shown in Figure 3.12.

The observed values in the cases of (b) and (c) lie within the approximate 97% credible region, but that in the case of (a) lie far away from the appropriate credible region, suggesting clear rejection of the model.

We next discuss the use of reference distributions in detecting overfitting in models.

3.6 Overfitting models and reference distributions

In Chapters 7 and 9 it will be seen that even when the observed discrepancy measures are small, this does not necessarily lead to acceptance of the models in question. In such cases, basing decisions solely on the smallness of the magnitudes of the observed discrepancy

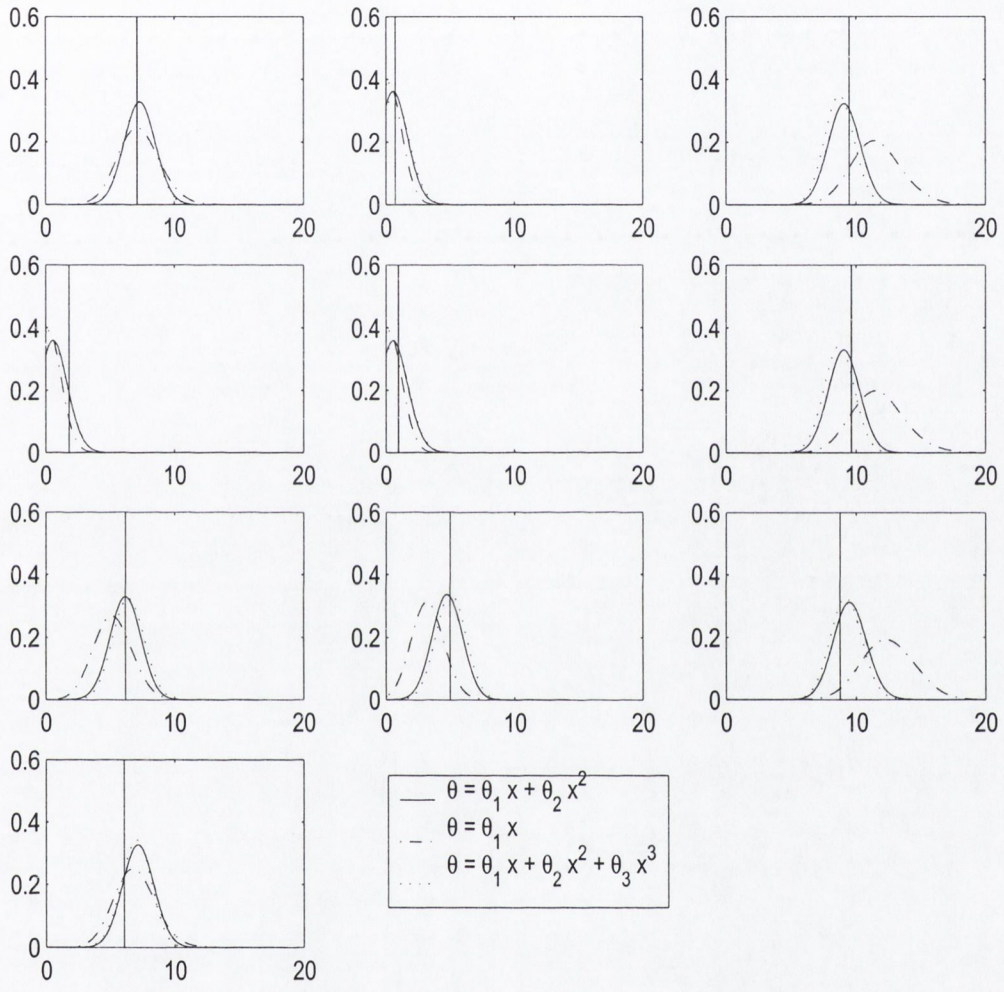


Figure 3.11: Leave-one-out posteriors of x for each of the three models for a sample data set. Vertical lines denote observed values.

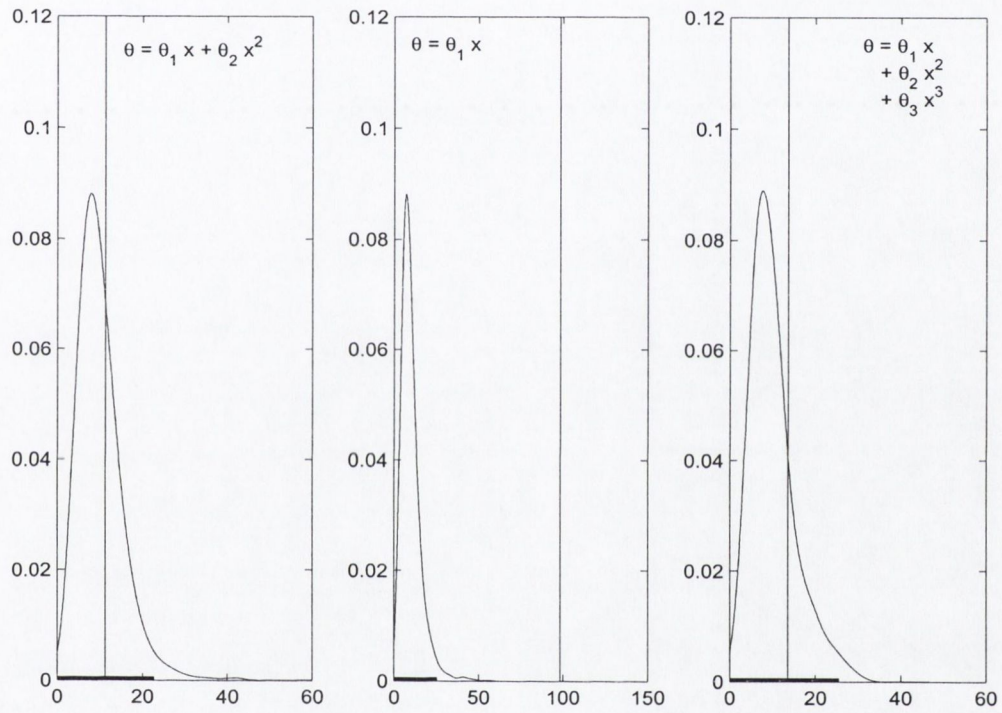


Figure 3.12: Reference distributions of discrepancy measures for each of the three models. Vertical lines denote observed discrepancy.

measures may be quite misleading. Below we illustrate this with examples. We begin with a very simple forward problem involving no regression; here we will be concerned with pairs $\{y_i, \pi(y \mid Y_{-i})\}$. In the next example we consider an inverse problem which is in accord with our interest in this thesis and where the pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$ will be considered.

Example 5: Overfit in simple forward model

Suppose that data $Y = \{y_1, \cdots, y_{10}\}$ actually came from a normal distribution with mean 0 and variance 10 but has been modeled in reality as $N(0, \tau)$; variance τ being treated as unknown. It may be assumed that the prior density of τ is inverse gamma with parameters α, β, given by

$$\pi(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{-(\alpha+1)} \exp(-\beta/\tau); \;\; \tau > 0, \;\; \alpha > 0, \;\; \beta > 0$$

Here

$$E(\tau \mid Y) = \frac{\beta'}{\alpha' - 1}$$

and

$$Var(\tau \mid Y) = \frac{\beta'^2}{(\alpha' - 1)(\alpha' - 2)}$$

where $\alpha' = \alpha + 5$ and $\beta' = \beta + \sum_{i=1}^{10} y_i^2/2$. In this case, if $\beta >> 10$, then the posterior variance of τ will be very large in comparison to the true variance and this in turn will make the prediction uncertainty very high. In this situation the model will overfit the data in the sense that although the observed values will be matched very well by the model (since the expected value is modeled correctly), high prediction uncertainty will remain. In other words, observed discrepancies will be small but the model may still be unacceptable.

Figure 3.13 presents a case with $\alpha = 3$ and $\beta = 1000$. Here the observed discrepancy being 0.5, is quite small. But since the reference distribution does not support such small value the model is to be rejected.

Example 6: Overfit in inverse regression

Now we demonstrate the situation with an inverse problem. Given θ and x_i; $i = 1, \cdots, 10$

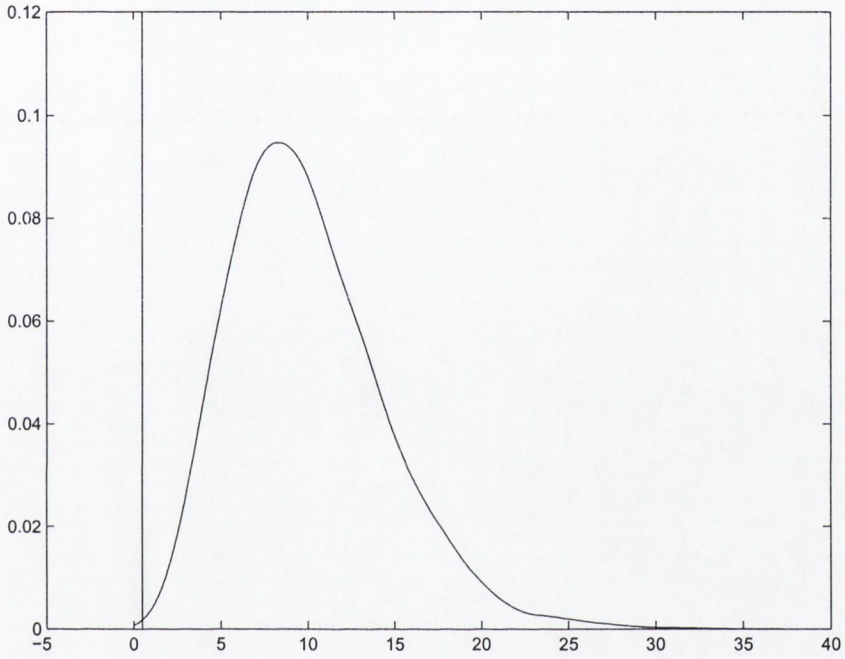


Figure 3.13: Demonstration of overfitted case. Here considering solely the observed discrepancy measure (denoted by the vertical line) wrongly leads to acceptance of the overfitted model; considering it with respect to the reference distribution leads to the correct decision (that is rejection of the model).

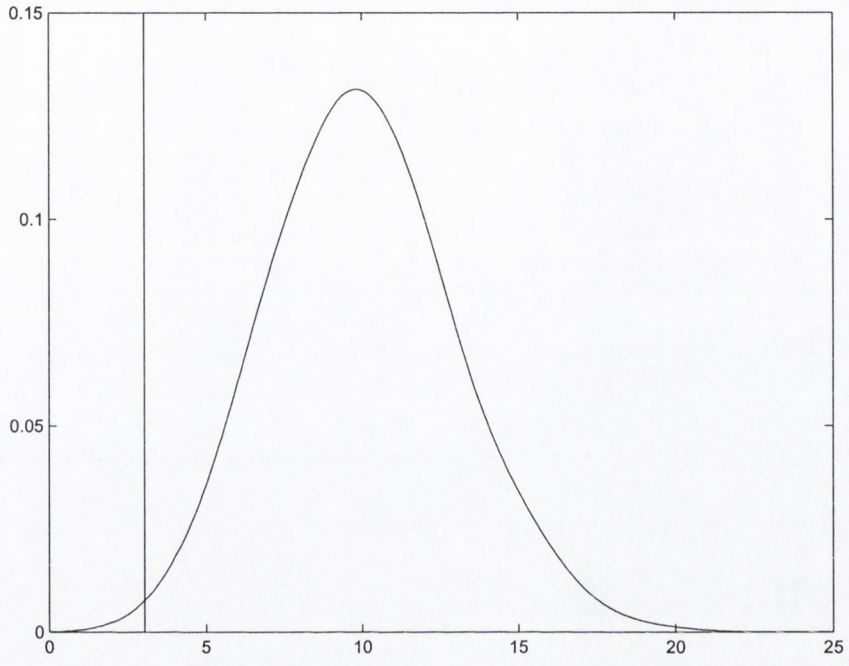


Figure 3.14: Demonstration of overfitted situation in an inverse problem involving Poisson and Geometric model. Here considering solely the observed discrepancy measure (denoted by the vertical line) wrongly leads to acceptance of the overfitted model; considering it with respect to the reference distribution leads to the correct decision (that is rejection of the model).

which arise from a uniform distribution, suppose that $y_i \sim \text{Poisson}(\theta x_i)$. But suppose that y_i has been modeled as a Geometric distribution with parameter $p_i = 1/(1 + \theta x_i)$. Here although the expected value of y_i under both models are same, given by θx_i, the variance under the Geometric model given by $\theta x_i(1 + \theta x_i)$ is greater than in the Poisson case, where it is given by θx_i. Thus, for certain values of θ the Geometric model may overfit the data which actually comes from the Poisson model. Figure 3.14 presents a case with $\theta = 15$.

Compared to Example 1 where the observed discrepancy measure was 0.5, here the observed discrepancy measure is greater, 3.04, but still small enough with respect to the reference distribution in this case. Thus the Geometric model is to be considered as a poor fit to the observed Poisson data.

3.7 Conclusions

We have proposed an approach based on reference distributions of discrepancy measures to assess goodness of model fit in inverse problems. This is based on pairs $\{x_i, \pi(x \mid X_{-i}, Y)\}$, the observed x-values and their leave-one-out posteriors. It has been argued and demonstrated with the help of Examples 1 and 2 that this approach is less efficient when testing the underlying probability model of the forward part y is of primary interest. However, even though this is less natural, when y is of much higher dimensionality than the inverse x, the reference distribution approach may be much more easy.

The inverse approach can be explicitly used to check the prior on x; Example 3 demonstrates this. Variable selection on the inverse side is demonstrated in Example 4.

The reliability of reference distributions in detecting overfitting models is demonstrated in Examples 5 and 6. Two real examples in this regard are given in Chapters 7 and 9.

We have completely avoided the question of recommending an optimal choice of discrepancy measure. This we recognise as future work.

Chapter 4

An Importance Resampling proposal for cross-validation in MCMC

We now introduce our proposed algorithm and discuss its implementation and relevant technical details. For all demonstration purposes in this chapter attention will be restricted to the Poisson example introduced in Chapter 1.

4.1 Proposal

We first explain our method schematically.

Cases	$1, \cdots, i^*, \cdots, n$
Omit case i^*	$\downarrow$
Realisations by regular MCMC	$(\theta^{(1)}, x^{(1)}), \cdots, (\theta^{(N)}, x^{(N)})$
For each $i(\neq i^*)$ compute weights	$w_{i^*,i}^{(1)}, \cdots, w_{i^*,i}^{(N)}$
IR	$\downarrow$
Use $w_{i^*,i}^{(\ell)}$ *to resample* θ*'s*	$\tilde{\theta}^{(1)}, \cdots, \tilde{\theta}^{(k)}, \cdots, \tilde{\theta}^{(K)}$
Regular MCMC with fixed θ	$\downarrow$
For each k*, sample* M *values of* x	$\hat{x}_i^{(1)}, \cdots, \hat{x}_i^{(M)}$

Pool MK values of x.

More precisely, our proposed procedure can be stated in the following manner.

1. Choose an initial case i^*. Use $\pi(x, \theta \mid X_{-i^*}, Y)$ as the importance sampling density. Methods of choosing i^* and the necessity of making such a choice are discussed in Section 4.3.

2. From this density sample values $(x^{(\ell)}, \theta^{(\ell)}); \ell = 1, \cdots, N$, for sufficiently large N. Typically, regular MCMC will be used for sampling.

3. For $i \in \{1, \cdots, i^* - 1, i^* + 1, \cdots, n\}$ do

   a. Compute, for each sample value, importance weights $w_{i^*,i}^{(\ell)} = w_{i^*,i}(x^{(\ell)}, \theta^{(\ell)})$, where the importance weight function is given by

   $$w_{i^*,i}(x, \theta) = \frac{\pi(x, \theta \mid X_{-i}, Y)}{\pi(x, \theta \mid X_{i^*}, Y)} \propto \frac{L(Y, X_{-i}, x, \theta)}{L(Y, X_{-i^*}, x, \theta)}, \tag{4.1}$$

   $L(Y, X, \theta)$ being the likelihood of the observed data under the model. Note that the importance weights are independent of the prior on (x, θ); hence, unlike the case with the saturated posterior, the strength of prior has no effect on the reliability of the results.

   b. For $k \in \{1, \cdots, K\}$

      (i) Sample without replacement $\tilde{\theta}^{(k)}$ from $\theta^{(1)}, \cdots, \theta^{(N)}$ where the probability of sampling $\theta^{(\ell)}$ is proportional to $w_{i^*,i}^{(\ell)}$.

      (ii) Draw M times from $\pi(x \mid y_i, \tilde{\theta}^{(k)})$. Note that although low-dimensional, it may not be straightforward to simulate directly from the above conditional and regular MCMC may be needed.

   c. Store the $K \times M$ draws of x as the posterior for x_i as $\hat{x}_i^{(1)}, \cdots, \hat{x}_i^{(KM)}$.

Our proposed steps 1, 2 and 3a eliminate the problem of obtaining the functional form of $\pi(\theta \mid X_{-i}, Y)$. The need to know the normalizing constant of $\pi(x \mid y_i, \theta)$ is avoided in step

3b(ii) by recommending regular MCMC. We next demonstrate that IRMCMC is MCMC with a novel proposal mechanism.

4.2 IRMCMC is MCMC with a special proposal mechanism

We demonstrate in this section that IRMCMC is really a version of MCMC with a special proposal mechanism.

For notational convenience in this section we use the shorthand notation $\pi_i(\cdot)$ for $\pi(\cdot \mid X_{-i}, Y)$. For the IRMCMC methodology, the proposal kernels of θ and x are given by

$$\begin{aligned} Q_{1N}(\theta^{(t+1)} \mid x^{(t)}, \theta^{(t)}) &= P(t : x^{(t)} \in \mathcal{S}) \pi_N(\theta^{(t+1)} \mid x^{(t)}, X_{-i}, Y) \\ &+ P(t : x^{(t)} \notin \mathcal{S}) \delta_{\{\theta^{(t)}\}}(\theta^{(t+1)}) \end{aligned} \tag{4.2}$$

$$\begin{aligned} Q_2(x^{(t+1)} \mid \theta^{(t+1)}, x^{(t)}) &= P(t : x^{(t)} \in \mathcal{S}) \pi(x^{(t+1)} \mid \theta^{(t+1)}, X_{-i}, Y) \\ &+ P(t : x^{(t)} \notin \mathcal{S}) q(x^{(t+1)} \mid x^{(t)}) \end{aligned} \tag{4.3}$$

In the above, δ is the indicator function. $P(t : x^{(t)} \in \mathcal{S})$ denotes the probability that t is a stopping time. In other words, t is a stopping time if $x^{(t)}$ takes value in the set $\mathcal{S}$. $\pi_N(\theta \mid X_{-i}, Y)$ is the empirical distribution of $\pi(\theta \mid X_{-i}, Y)$. Observe that the empirical distribution function in this case is given by

$$F_{i,N}(\theta) = \frac{\sum_{\ell=1}^{N} \delta_{(-\infty,\theta]}(\theta^{(\ell)}) w_i(\theta^{(\ell)})}{\sum_{\ell=1}^{N} w_i(\theta^{(\ell)})}$$

Clearly, by the ergodic theorem the above empirical distribution function converges almost surely to the true distribution function as N is made infinitely large. q is a distribution that may or may not depend on values $x^{(t)}$ and $\theta^{(t+1)}$. Note that we suppress $\theta^{(t+1)}$ in the notation. This is because we generally choose the distribution to be independent of $\theta^{(t+1)}$ (normally, this is a random walk).

We now have a closer look at the proposal kernels Q_1 and Q_2. Q_1 says that if t is a stopping time propose a new value, $\theta^{(t+1)}$, from the empirical distribution $\pi_N(\theta^{(t+1)} \mid X_{-i}, Y)$; if not

then set $\theta^{(t+1)} = \theta^{(t)}$. A new value $\theta^{(t+1)}$ will be proposed by IR without replacement. The interpretation of Q_2 is similar to Q_{1N}. It says that if t is a stopping time propose a new value $x^{(t+1)}$ from the distribution $\pi_i(x \mid \theta^{(t+1)})$ (which will be typically done by MCMC) and if not then propose $x^{(t+1)}$ from any distribution q that may (or may not) depend on the current value $x^{(t)}$. Observe that in our case τ is a stopping time if $\tau = M$; that is $P(\tau = M) = 1$ and $P(\tau \neq M) = 0$. Note that our proposal implies keeping θ fixed for M consecutive realisations of x. This is a deterministic definition; however, randomness may be introduced by agreeing to stop the chain when the Monte Carlo error falls below a certain level. We demonstrate this in Section 4.5 (see also Jones and Hobert (2001)).

Denoting the acceptance probability of $\theta^{(t+1)}$ given $\theta^{(t)}$ and $x^{(t)}$ by $\alpha_N(\theta^{(t+1)} \mid x^{(t)}, \theta^{(t)})$, and using (4.2), we observe that $\alpha_N(\theta^{(t+1)} \mid x^{(t)}, \theta^{(t)}) \to 1$ as $N \to \infty$, for $\pi(\cdot \mid X_{-i}, Y)$-almost all $\left(\theta^{(t)}, x^{(t)}\right)$.

On the other hand, the acceptance probability of $x^{(t+1)}$ given $\theta^{(t+1)}$ and $x^{(t)}$ depends on whether on not t is a stopping time. If t is a stopping time, then the acceptance probability, $\alpha(x^{(t+1)} \mid \theta^{(t)}, x^{(t)}) = 1$, and $\beta(x^{(t+1)} \mid \theta^{(t)}, x^{(t)})$ otherwise, where

$$\beta(x^{(t+1)} \mid \theta^{(t)}, x^{(t)}) = \min\left\{\frac{\pi(x^{(t+1)} \mid \theta^{(t+1)}, X_{-i}, Y)}{\pi(x^{(t)} \mid \theta^{(t+1)}, X_{-i}, Y)} \frac{q(x^{(t)} \mid x^{(t+1)})}{q(x^{(t+1)} \mid x^{(t)})}, 1\right\} \tag{4.4}$$

The above facts show that IRMCMC is indeed a version of MCMC with special proposal kernels. The Markov chain can be written as

$$\begin{aligned}
&\{(x^{(1)}, \tilde{\theta}^{(1)}), (x^{(2)}, \tilde{\theta}^{(1)}), \quad \cdots, \quad (x^{(M)}, \tilde{\theta}^{(1)}),\\
&(x^{(1+M)}, \tilde{\theta}^{(2)}), (x^{(2+M)}, \tilde{\theta}^{(2)}), \quad \cdots, \quad (x^{(2M)}, \tilde{\theta}^{(2)}),\\
&\qquad\qquad \cdots\\
&(x^{(K+M)}, \tilde{\theta}^{(K)}), (x^{(K+M)}, \tilde{\theta}^{(K)}), \quad \cdots, \quad (x^{(KM)}, \tilde{\theta}^{(K)}), \cdots\}.
\end{aligned}$$

Note that the realisations of θ are piecewise constant in nature. To clarify this concept we perform a small simulation study involving the Poisson regression problem. We choose $M = 100$ and $K = 10$. Shown in Figure 4.1 is the sample path of our algorithm for a

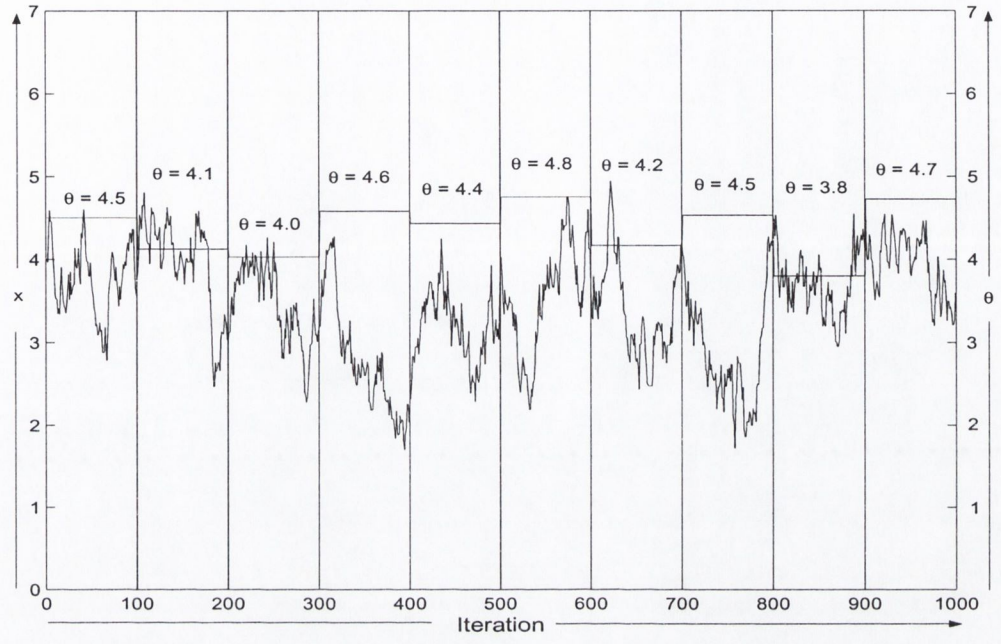


Figure 4.1: IRMCMC: sample path of x and θ in Poisson regression problem. Observe that θ is piecewise constant

particular case. To be noted is the fact that θ remains piecewise constant for $M = 100$ realisations of x. Vertical lines are used to clarify the piecewise constant nature of the realisations of θ. The advantages of the piecewise constant nature of θ will be spelt out in Section 4.4.

A regular MCMC run for adequate exploration of $\pi(x \mid y_i, \tilde{\theta}^{(t)})$ is easy to implement since the dimensionality of x is low and hence an optimal proposal distribution can be constructed easily. Note that an initial burn-in is essential if MCMC is used to draw $x^{(t)}$ from $\pi(\cdot \mid y_i, \theta^{(1)})$. Corresponding to $\tilde{\theta}^{(t)}$, for $t > 1$ the last realization of x corresponding to $\tilde{\theta}^{(t-1)}$ could be used as the initial value and hence no burn-in would be necessary. However, it is also possible to use independent initial values corresponding to each $\tilde{\theta}^{(t)}$ and to discard a certain number of initial realisations of x as burn-in. This can be very useful in detecting multimodal solutions and we use this idea to implement the real cross-validation exercises described in this thesis.

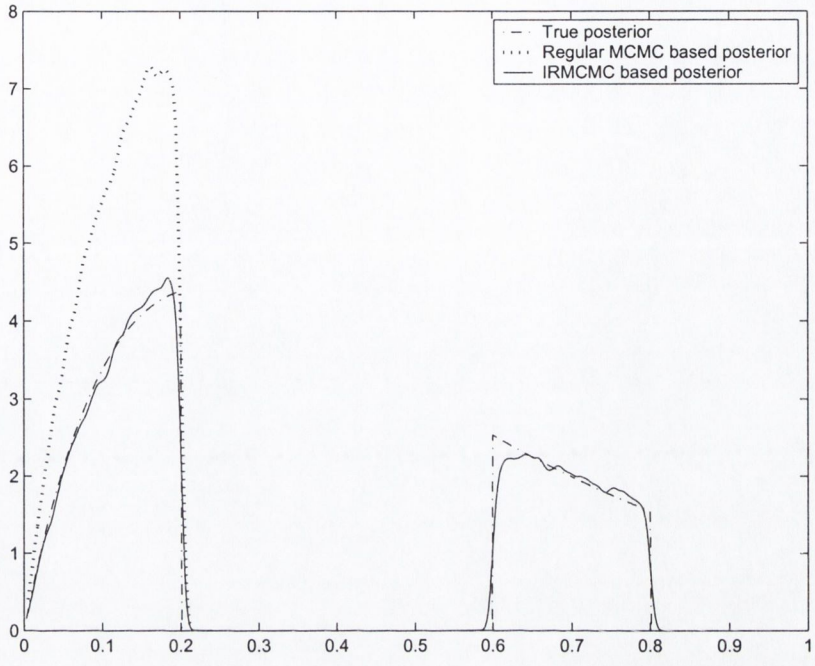


Figure 4.2: IRMCMC explores true posterior thoroughly; regular MCMC gets stuck at one modal region.

We denote this version of IRMCMC as 'IRMCMC with restarts'. Note that it is possible to select independent initial values for regular MCMC runs corresponding to each θ from the true posterior itself by rejection sampling (for an account of rejection sampling see Ripley (1987), Section 3.2, and Devroye (1986), Section II.3). When the dimensionality of x is low, this is not difficult to achieve and could also be a way to avoid the convergence issues of multiple regular MCMC chains.

Figure 4.2 shows a toy example (based on the already introduced Poisson regression example) where regular MCMC with a random walk proposal having an inadequately small step size completely fails to explore a bimodal solution. On the other hand, IRMCMC with restarts succesfully explores both the modes quite accurately.

It is also useful to mention that in cases where it is easy to sample directly from $\pi(x \mid y_i, \tilde{\theta}^{(k)})$ one can realise an almost *iid* sample from the true posterior of x at the cost of no burn-in. In most forward problems such direct simulation is possible (for an illustration, see

Chapter 9). It is certainly also possible in inverse problems, the Poisson regression problem being a toy example. But it does not seem possible in the case of the two real examples of inverse problems described in this thesis. We thus continue to recommend regular MCMC for its generality.

Note also that typically a regular MCMC (and burn-in) is needed to construct the empirical distribution function F_{iN}. Observe also that we do not need to evaluate the distribution function but only need to draw samples from F_{iN} without replacement.

It remains now to propose a procedure to select i^* appropriately.

4.3 Selection of appropriate importance sampling distribution

It will be shown in Chapter 5 that for large data sets, any choice of i^* is reasonable. However, it is important to provide a procedure to select i^* when the data size is finite. In this section, we propose and study properties of several procedures. We also demonstrate that even an 'optimal choice' of i^* is little better than choosing i^* at random, having excluded 'extreme' cases.

It follows from the discussion in Chapter 2 that in order to sample adequately from the target distribution using IR, it is desirable that the importance sampling distribution resembles the target distribution as closely as possible. Thus, for any cross-validation problem, i^* should be chosen in such a manner that the difference between $\pi(x, \theta \mid X_{-i^*}, Y)$ and $\pi(x, \theta \mid X_{-i}, Y)$ are as small as possible for all i. Since the prior for (x, θ) remains fixed for all cases, the difference between the likelihoods for different cases is the sole cause for difference between the posteriors. From (4.1) it follows that ideally $L(Y, X_{-i^*}, x, \theta) = L(Y, X_{-i}, x, \theta)$ for all i over the entire range of (x, θ). Clearly this is not possible. However, an adequate approximation to equality of weights is to choose i^* such that $L(Y, X_{-i^*}, x, \theta)$ is roughly 'central' in the set of $L(Y, X_{-i}, x, \theta)$. We elaborate with the Poisson example for which the

importance weight function of case i with case j as the importance sampling density is given by

$$w_{j,i}(x,\theta) = x^{y_i - y_j} \exp\{\theta(x_i - x_j)\}. \tag{4.5}$$

We now discuss different versions of 'centrality'. These are all based on pairwise distance measures, $d(i,j)$, the distance being defined in some sense between posteriors of (x,θ) corresponding to cases i and j. Given the pairwise distances, we select case i^*, where $i^* = \arg\min \sum_i d(i,j)$. We begin by an approach motivated by the Kullback-Leibler (KL) distance between two distributions and that $d_{KL}(i,j)$ is the measure based on KL- distance between distributions corresponding to cases i and j, to within an additive constant. Details are given below.

4.3.1 A Kullback-Leibler (KL) motivation for the selection of i^*

The importance weight function of case i when case j is the importance sampling density, is given by

$$w_{j,i}(x,\theta) = \frac{p(y_i|x,\theta)}{p(y_i|x_i,\theta)} \frac{p(y_j|x_j,\theta)}{p(y_j|x,\theta)} \tag{4.6}$$

We however, recognise our task as to select that $\pi(x,\theta \mid X_{-i^*}, Y)$ which is 'closest overall' to the remaining $\pi(x,\theta \mid X_{-i}, Y)$, for $i^* \neq i$. From such pairwise distances it is possible to determine the case which is closest to the centroid. One proposal is $i^* = \arg\min \sum_{i=1}^n d_{KL}(i,j)$. It has been shown in Chapter 5 that for large n, any choice of $i^* \in \{1,\cdots,n\}$ will suffice. However, it is important to study situations where n is small.

Denoting the expectation with respect to $\pi(x,\theta \mid X_{-j}, Y)$ by E_j, we note however, that,

$$E_j[w_{j,i}(x,\theta)] = \int \frac{\pi(x,\theta \mid X_{-i}, Y)}{\pi(x,\theta \mid X_{-j}, Y)} \pi(x,\theta \mid X_{-j}, Y) dx d\theta = 1$$

Observe that $E_j[\log\{w_{j,i}(x,\theta)\}] = \log\{E_j[w_{j,i}(x,\theta)]\} = 0$ if and only if the weights are equal. We describe $d_{KL}(i,j) = |E_j[\log\{w_{j,i}(x,\theta)\}]|$ as a measure of difficulty in using $\pi(x,\theta \mid X_{-j}, Y)$ as a basis for estimating $\pi(x,\theta \mid X_{-i}, Y)$. It is also possible to use $d_{KL}(i,j) = (E_j[\log\{w_{j,i}(x,\theta)\}])^2$. Note that if $w_{j,i}(x,\theta)$ includes normalizing constants of the posteriors

$\pi(x, \theta \mid X_{-i}, Y)$ and $\pi(x, \theta \mid X_{-j}, Y)$, then $-E_j[\log\{w_{j,i}(x, \theta)\}]$ is in fact the KL distance between these posteriors. We note that $d_{KL}(i, j)$ is not symmetric in (i, j). We thus propose $\hat{d}_{KL}(i, j) = (d_{KL}(i, j) + d_{KL}(j, i))/2$ as such a measure.

We now define the posterior $\pi(x, \theta \mid X_{-i^*}, Y)$ with minimum $\hat{d}_{KL}(i, i^*)$ to be the 'closest' to a given $\pi(x, \theta \mid X_{-i}, Y)$. We can not assert that least variability of importance weights $w_{i^*,i}(x, \theta)$ is ensured, nor that any measure of variability in weights will yield an estimate of least 'distance' in any sense.

For the Poisson regression case, it follows from (4.5) that

$$d_{KL}(i, j) = \mid (y_i - y_j)E_j(\log(x)) + E_j(\theta)(x_i - x_j) \mid . \tag{4.7}$$

Using the above equation, $\hat{d}_{KL}(i, j)$ can easily be calculated.

It is clear from (4.7) that the above theory can be used formally only when the posteriors $\pi(x, \theta \mid X_{-i}, Y)$ are known or samples already available from them. However, this is not the case in reality. One approximation is to use the absolute value (or the square) of

$$E_j[\log(w_{j,i}(x, \theta)] \approx E_j[\log\{w_{j,i}\,(E_j(x), E_j(\theta))\}] \tag{4.8}$$

A natural alternative for θ is to use expected values from the saturated posterior, which preceeds the cross-validation stage naturally. For x, we propose to approximate $E_j(x)$ by x_j. Thus it follows from (4.6) that,

$$E_j[\log(w_{j,i}(x, \theta)] \approx E_j\left[\log \frac{p(y_i|x_j, \theta)}{p(y_i|x_i, \theta)}\right], \tag{4.9}$$

which equals zero when $x_i = x_j$. Note that this is close in spirit to the KL-distance between $p(y_i \mid x_i, \theta)$ and $p(y_i \mid x_j, \theta)$. Also observe that in (4.9) expectation E_j is equivalent to taking expectation with respect to $\pi(\theta \mid X_j, Y)$ only. Thus,

$$d_{KL}(i, j) = |(y_i - y_j)\log(x_j) + (x_i - x_j)E(\theta \mid X, Y)|, \tag{4.10}$$

where $E(\theta \mid X, Y)$ is the expected value of θ with respect to the saturated posterior. In the Poisson regression case, $E(\theta \mid X, Y) = (\sum_{k=1}^{n} y_k)/(\sum_{k=1}^{n} x_k)$. Thus it is simple to compute

$\hat{d}_{KL}(i,j)$ for all j, given $E(\theta \mid X, Y)$. We may thus define the case that is 'closest' to a given $\pi(x, \theta \mid X_{-i}, Y)$.

Note that a separate regular MCMC is needed to evaluate $E(\theta|X,Y)$ in addition to a regular MCMC run corresponding to $\pi(x, \theta \mid X_{-i^*}, Y)$. This is computationally very burdensome when θ is of very high dimensionality (for example, in Whiley et al. (2004)). It may not be worth the effort if other simpler methods perform as good or better. Moreover, it will be demonstrated in Section 4.3.3 that the effects of approximations used to compute $\hat{d}_{KL}$ may not be negligible and may adversely affect performance.

In the next subsection we provide simpler alternative approaches to select i^*.

4.3.2 Other measures of centrality to determine i^*

An adequate approximation to euqality of weights is to choose i^* such that $L(Y, X_{-i^*}, x, \theta)$ is roughly 'central' in the set of $L(Y, X_{-i}, x, \theta)$. We elaborate with the Poisson example for which the importance weights are given by (4.5). We propose two distance measures whose minimisation offers different versions of centrality :

$$\text{(a)} \quad d_1(i) = \sum_{j=1}^{n} \left(\frac{\mid x_j - x_i \mid}{S_X} + \frac{\mid y_j - y_i \mid}{S_Y} \right) \tag{4.11}$$

$$\text{(b)} \quad d_2(i) = \sqrt{\sum_{j=1}^{n} \left(\frac{(x_j - x_i)^2}{S_X^2} + \frac{(y_j - y_i)^2}{S_Y^2} \right)} \tag{4.12}$$

where S_X and S_Y denote sample standard deviations of the X column and the Y column respectively. Note that the above measures can be easily extended to situations where x_i and y_i are multivariate. In the case where $x_i = (x_{i1}, \cdots, x_{ip})$ and $y_i = (y_{i1}, \cdots, y_{iq})$, the distance measures are defined by

$$\text{(c)} \quad d_1(i) = \sum_{j=1}^{n} \left(\sum_{k=1}^{p} \frac{\mid x_{jk} - x_{ik} \mid}{S_{X_k}} + \sum_{k=1}^{q} \frac{\mid y_{jk} - y_{ik} \mid}{S_{Y_k}} \right) \tag{4.13}$$

$$\text{(d)} \quad d_2(i) = \sqrt{\sum_{j=1}^{n} \left(\sum_{k=1}^{p} \frac{(x_{jk} - x_{ik})^2}{S_{X_k}^2} + \sum_{k=1}^{q} \frac{(y_{jk} - y_{ik})^2}{S_{Y_k}^2} \right)} \tag{4.14}$$

In the above S_{X_k} and S_{Y_k} denote sample standard deviations of the k^{th} column of X and Y respectively.

With the above proposition, we then have $i^* = \arg\min\{d_k(i); 1 \leq i \leq n\}$, for $k = 1, 2$. Note that unlike measures based on d_{KL}, no knowledge is required of any quantity to be estimated and thus seems far more reasonable and simpler to compute. We next demonstrate with the Poisson regression example that the measures d_1 and d_2 may outperform the procedure motivated by KL distance. In fact, we show that even a random choice of i^* from $\{1, \cdots, n\}$ may perform more adequately than the latter.

4.3.3 Comparison between methods of choosing an appropriate i^*

In this section we use the Kolmogorov-Smirnov (KS) measure to evaluate the performance of IRMCMC with respect to different choice of i^* in the case of the Poisson problem. The KS measure is defined by

$$\sup_{z \in R} \mid G_n(z) - G(z) \mid . \tag{4.15}$$

In (4.15), G is the true distribution function of the posterior of x corresponding to case i has been omitted and G_n denotes the empirical distribution function defined as

$$G_n(x) = \frac{1}{N} \sum_{\ell=1}^{N} \delta_{(-\infty, x]}(\hat{x}_i^{(\ell)})$$

where δ denotes the indicator function. For details and related issues see Lehmann (1986), Billingsley (1995), Rao (1965). Recall that the true distribution is easily available in this toy problem.

For a fixed value of θ, we simulate 500 replicates of (X, Y) such that, for $i = 1, \cdots, 10$, $y_i \sim P(\theta x_i)$. For each of the 500 replications, the P-values associated with the observed KS-measure are computed for the 10 cases. This has been repeated for different ways of selecting i^*. For the procedure motivated by the KL distance we can envisage four versions, each version shedding different light on the basic issue of selecting an appropriate i^*.

(1) KL-1: Approximate version of d_{KL} given by (4.10) is used. Implementation of this version seems to be feasible and sensible in practice.

(2) KL-2: Exact version of d_{KL} is used with normalising constants, making d_{KL} a KL-distance, is used. This is given by

$$d_{KL}(i,j) = |(y_i - y_j)E_j(\log(x)) + (x_i - x_j)E_j(\theta)| + \log(c_i) - \log(c_j),$$

where c_i is the normalising constant of $\pi(x, \theta \mid X_{-i}, Y)$, given by

$$c_i = \frac{(\sum_{k\neq i} x_k)^{(\sum_{k\neq i} y_k)}}{\Gamma(\sum_{k\neq i} y_k)\Gamma(y_i + 1)}.$$

$E_j(\theta)$ is given by $\sum_{k\neq j} y_k / \sum_{k\neq j} x_k$ and $E_j(\log(x))$ has been evaluated numerically. Note that in this simple Poisson regression case, where analytical solution is available, such form of d_{KL} might be used. But in reality such analytical solutions may not be available. However, this simple problem where analytical solutions are available will help us expose the fact that the approximation (4.10), although seems realistic and easy to compute, may not be sufficiently accurate and hence the performance of IRMCMC may be affected in that case.

(3) KL-3: In this case, the distance d_{KL} has been made independent of x by integrating it out. Here

$$d_{KL}(i,j) = |(y_j - y_i)E_j(\log(\theta)) - (y_j - y_i)E_j(\theta)|.$$

In the above, it has been assumed that $E_j(\log(\theta)) \approx E(\log(\theta) \mid X, Y)$ and $E_j(\theta) \approx E(\theta \mid X, Y)$. Also note that the normalising constants corresponding to cases i and j are not considered.

This version involves the assumption that x can be integrated out analytically, which is unrealistic, but is helpful in demonstrating that the random variable x involved in the measure d_{KL} can adversely affect selection of an appropriate i^*.

(4) KL-4: This is similar to KL-3 in essence but uses exact values of $E_j(\log(\theta))$ and $E_j(\theta)$ instead of approximations and normalising constants (here $c_i = (\sum_{k\neq i} x_k)^{(\sum_{k\neq i} y_k)}/\Gamma(\sum_{k\neq i} y_k)$) taken into account making d_{KL} a KL-distance. It will be demonstrated that this is the

best version; however, this will be unavailable in practice since analytical solutions are needed.

Apart from the performances of the above four versions of the procedure motivated by the KL distance, the performances of d_1, d_2 and the method of simple random selection are also considered and compared. It will be demonstrated that even with very small samples there is very little difference to choose between all the candidates for i^*, but that very simple measures seem to offer a choice that is easy to compute. We remark that both d_1 and d_2 seemed to exhibit similar performances; indeed in the realistic applications described in Chapters 6, 7 and 9, both yielded same results.

We say that IRMCMC corresponding to a given value of i^* satisfactorily approximates the target distribution of x at case i if the P-value for that case is greater than 0.05. In each replication, the number of P-values (note that there are ten P-values in each replication) exceeding 0.05 is noted. Let us denote this number corresponding to the r^{th} replication by $\mathcal{N}_r$. Ideally, $\mathcal{N}_r = 10$ for each r. However, this may not always be the case in reality.

The proportion of times $\mathcal{N}_r = 10$ with respect to the proposed measures, may be used to compare the performances of the measures. This is given by

$$\hat{P}(\mathcal{N}_r = 10) = \sum_{r=1}^{500} \delta_{\{\mathcal{N}_r=10\}}(\mathcal{N}_r)/500,$$

where δ denotes the indicator function. A high value of $P(\mathcal{N}_r = 10)$ indicates satisfactory performance of IRMCMC, given a particular distance measure.

Using the above criterion, Table 4.1 compares the performances of different ways of selecting i^* for different values of θ.

Observe that all seven proposals perform quite adequately, the proportons $\hat{P}(\mathcal{N}_r = 10)$ being high. That this holds despite the fact that the variabilities of the simulated data sets change as θ changes demonstrates the considerable robustness of IRMCMC. It is particularly satisfying to note that even a randomly selected i^* performs very adequately. This is in accordance with the fact proved in Chapter 5, that the posteriors of θ corresponding to different omitted data points are (asymptotically) similar.

Table 4.1: Assessment of the performances of methods of selecting an appropriate i^*.

$\hat{P}(\mathcal{N}_r = 10)$							
θ	KL-1	KL-2	KL-3	KL-4	d_1	d_2	Random
0.5	0.974	0.980	0.934	0.980	0.882	0.894	0.908
1.0	0.978	0.972	0.932	0.986	0.970	0.932	0.938
3.0	0.948	0.964	0.976	0.992	0.984	0.968	0.944
5.0	0.912	0.924	0.968	0.998	0.966	0.974	0.946
7.0	0.934	0.948	0.966	0.988	0.976	0.960	0.934
9.0	0.934	0.938	0.964	0.992	0.986	0.980	0.936
11.0	0.900	0.938	0.970	0.994	0.984	0.984	0.964
13.0	0.892	0.904	0.938	0.994	0.994	0.982	0.970
15.0	0.910	0.926	0.968	1.000	0.986	0.962	0.948

However, compared to other procedures, the performance of KL-1 is the poorest. The exact version of KL-1, denoted by KL-2, performs better than KL-1, indicating that crude approximations involved in KL-1 might have adversely affected its performance.

The version KL-3, which corresponds to approximation after integrating out the random variable x, perform better than both KL-1 and KL-2. This is not unexpected, since in both KL-1 and KL-2 the random variable x, which can be regarded as a nuisance parameter while resampling θ only, is retained. This causes loss of efficiency of the procedure. Since only posteriors of θ, not x, are of interest, while resampling θ, and since an appropriate choice of i^* is needed only to ensure efficiency of the resampling procedure, the choice of i^* should not explicitly depend on x. Since KL-3 avoids this problem, it performs much better than both KL-1 and KL-2.

The version KL-4 is the exact version of KL-3, and hence outperforms KL-3. The fact that x has been integrated out and that exact solutions have been used help KL-4 perform excellently. Clearly, this is the best performer among all seven proposals.

Note that the distance measures d_1 and d_2 perform better than all procedures other than KL-4. This is because the measures use information from the data only for their determination of i^* and involve no unknown parameters and consequently no approximations.

This makes them safe from unreliable approximations that could have made them inefficient. Neither do they involve the undesirable random variable x in the distance calculation. Thus they perform better than KL-1, KL-2, KL-3. Since d_1 and d_2 use information from the data and the proposal of the random choice of i^* use absolutely no information, they also perform better than the random choice proposal. On the flip side, since they do not use information on the posterior of θ, which is of interest, they perform less efficently than KL-4, which rightly use information on θ. Thus d_1 and d_2 seem to be compromise between the most desirable and less desirable characteristics. However, since KL-4 is unrealistic in practice, we recommend d_1 and d_2 as reliable and realistic measures of determining i^* appropriately. Since both perform adequately, we arbitrarily recommend d_1.

In Chapter 6 it will be seen that in the case of Vasko et al. (2000) the importance weights are independent of Y. Hence, in that case, i^* is such that $x_{i^*} = median(X)$ if d_1 is used and $x_{i^*} \approx mean(X)$ if d_2 is used. Note that there may not exist any i^* such that $x_{i^*} = mean(X)$.

We next illustrate with an example the performance of these measures when an extreme observation is present in the data set.

4.3.4 Choice of i^* in the case of an extreme observation

In this section, we demonstrate with an example that IRMCMC corresponding to both the distance measures may fail to perform adequately at an extreme case.

Figure 4.3 presents a data set with an extreme observation at case 2. The above method with respect to both measures d_1 and d_2 clearly showed that case 2 is extreme. Both the distance measures d_1 and d_2 give $i^* = 5$ as the minimiser. The distance measures d_1 and d_2 are shown in Table 4.2. The performance of IRMCMC with case 5 omitted is adjudged by the KS measure; the results are present in the first three columns of Table 4.2. IRMCMC performs adequately at all cases except at the case with the extreme observation, where it performs very poorly. This is because leaving out case 5 approximates the posterior of x at case 2 very poorly; see Figure 4.4. The reason is that the support of the posterior of θ

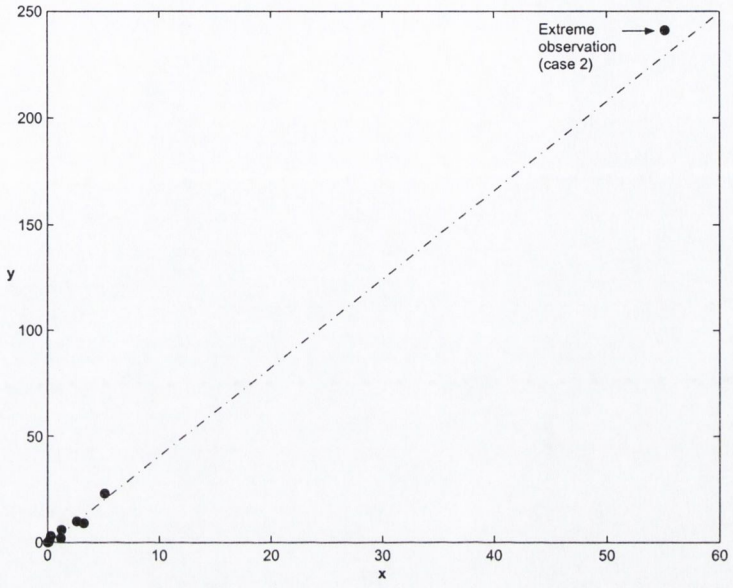


Figure 4.3: Data including an extreme observation.

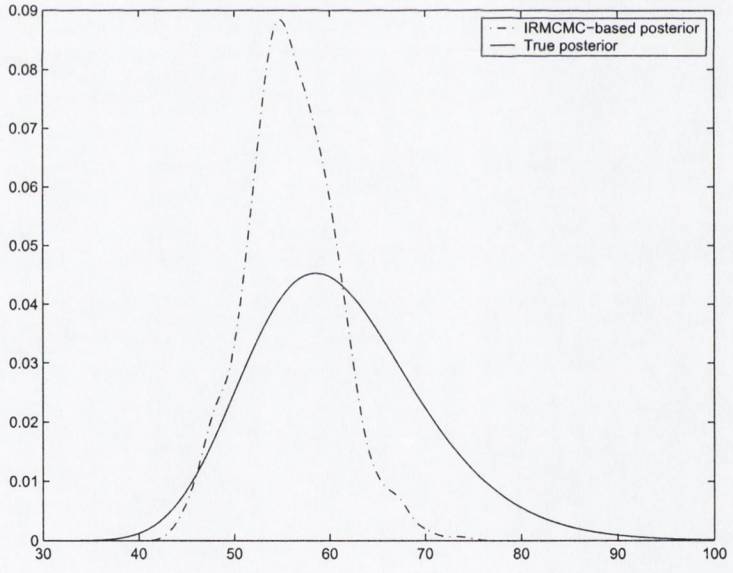


Figure 4.4: The true posterior of x at case 2 and the IRMCMC-approximated posterior with case 5 left out are shown. IRMCMC does not adequately represent the parameter space in this case.

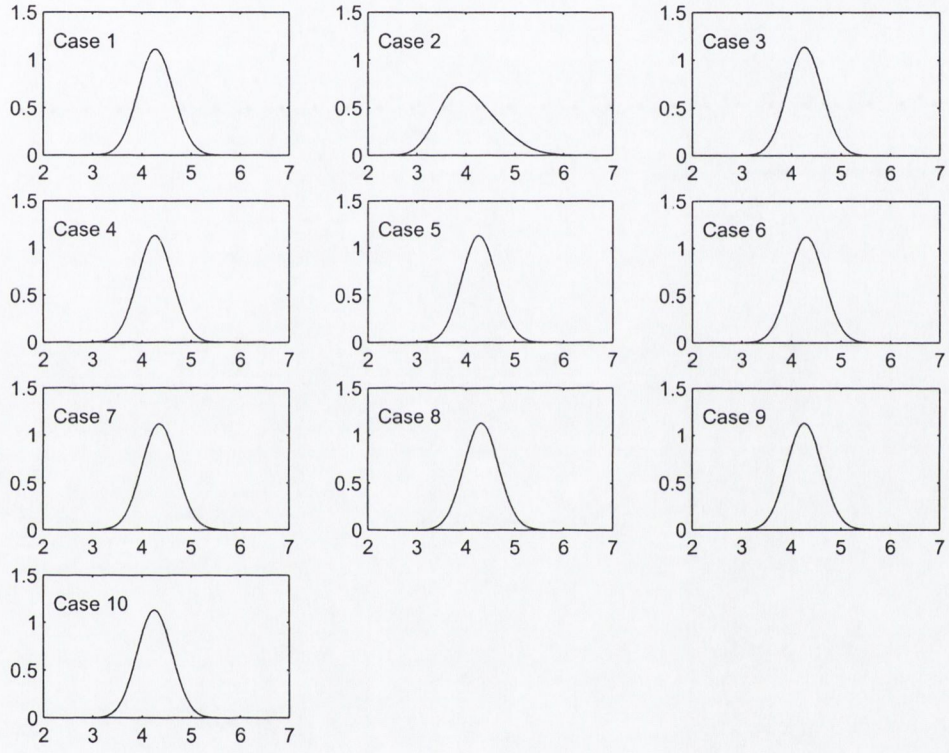


Figure 4.5: Leave-one-out posterior densities of θ. The support of the density at case 2 includes those at all other cases.

Table 4.2: Assessment of the performance of IRMCMC using the KS measure in the case where an extreme observation is present.

Case 5 omitted			Case 2 omitted		Distance	
Case	Obs KS	P-value	Obs KS	P-value	d_1	d_2
1	0.005	1.000	0.180	0.000	10.318	7.306
2	0.381	0.000	0.007	0.979	59.727	42.234
3	0.007	0.977	0.041	0.000	8.391	5.954
4	0.003	1.000	0.044	0.000	8.437	5.984
5	0.003	1.000	0.106	0.000	8.055	5.726
6	0.006	0.999	0.137	0.000	8.364	5.926
7	0.005	0.999	0.150	0.000	8.464	6.040
8	0.003	1.000	0.076	0.000	8.084	5.764
9	0.005	1.000	0.057	0.000	8.281	5.879
10	0.003	1.000	0.078	0.000	8.165	5.810

at the extreme case includes that of case 5 (in fact, those of all the other cases; see Figure 4.5). The smaller support of case 5 does not allow adequate represention of the complete parameter space of θ at case 2 which in turn causes poor approximation of the posterior of x at case 2.

Apparently it seems that leaving out the extreme case, i.e., taking $i^* = 2$ may suffice. This is because the support in this case properly contains those at other cases and hence no representation problems arise. However, a closer examination suggests otherwise; as explained below.

While we can here choose 'robust' i^*, it might represent a very inefficient choice since the variance of the leave-one-out posterior of θ for the extreme case is high compared to other leave-one-out posteriors of θ. Therefore, if the size of the initial sample is not extremely large, we might fail to adequately represent the remaining other posteriors. That this indeed is the case is shown in columns 4 and 5 of Table 4.2. All cases apart from the extreme case (which is, in fact, the importance sampling density) are very poorly approximated by IRMCMC.

Figure 4.6 compares a true leave-one-out posterior of x with the IRMCMC-approximated

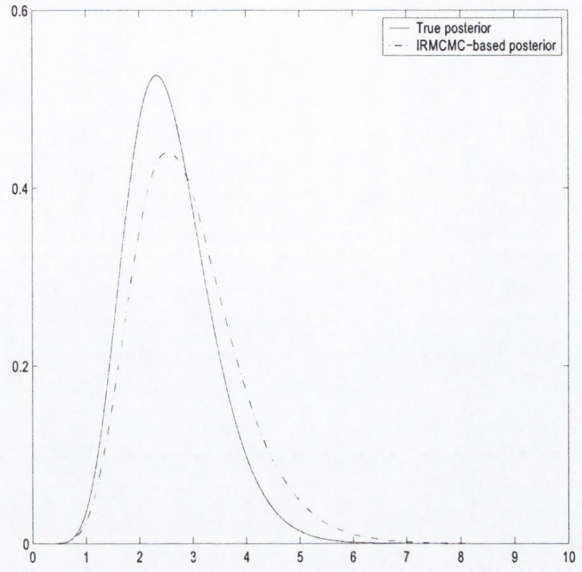


Figure 4.6: In the case of an extreme case IRMCMC fails to approximate the true posterior adequately.

posterior given that the extreme case has been omitted. Note that IRMCMC over-represents the right tail area of the true target posterior.

There is therefore a trade-off, quite generally, between robustness and efficiency. Recall that the main reason for IRMCMC is efficiency (excellent mixing comes as bonus). The method of choosing i^* by minimising distance measures d_1 and d_2, which seeks to choose that distribution as the importance sampling distribution which is closest to all other leave-one-out posteriors, thus makes sense.

4.3.5 Conclusions

We have thus demonstrated that it does not matter much which i^* is chosen, even when the data size, n, is very small, providing that sensible steps are taken to avoid using extrema. At a very extreme case, the leave-one-out posterior of θ will have a much wider support compared to leave-one-out posteriors at other cases, thus making the importance sampling

Table 4.3: Ratio of the KS measure corresponding to IRMCMC and regular MCMC.

Case	$R_{1,1000}$	$R_{1,2000}$	$R_{1,3000}$	$R_{1,4000}$	$R_{1,5000}$	$R_{50,100}$
1	0.337	0.739	1.670	1.807	3.560	1.028
2	0.391	0.317	1.674	0.500	1.243	1.741
3	1.431	1.289	2.299	0.706	1.240	0.799
4	0.712	1.331	2.216	1.055	0.932	0.393
5	0.754	0.625	0.824	2.511	2.684	0.326
6	0.626	0.243	1.838	2.616	2.314	1.209
7	0.905	0.661	0.521	4.181	0.512	1.083
8	–	–	–	–	–	–
9	0.395	2.976	1.115	1.157	1.729	1.160
10	1.509	1.512	2.195	1.200	0.817	0.894

distribution a poor approximation to the target distribution. In fact, all approximate methods for computing 'deletion diagnostics' (see, for example Haslett and Dillane (2004)) will fail in extreme cases and common sense needs to be used. Also, we remark that the example of the extreme cases provided here is very extreme indeed. Deleting the very extreme observation avoids the problem. We have conducted simulation studies with much less extreme observations (not reported) and in such cases any i^* seems quite satisfactory.

It is however clear that IRMCMC can fail, if there are unsuspected extreme cases (or cases that are influential in some unsuspected way). But (a) we can flag such cases by computing distances and (b) if the data size is large we may not have many such cases. In general it seems reasonable to recommend regular MCMC to explore extreme cases and IRMCMC for exploring the remaining cases.

We now demonstrate with an example that IRMCMC may explore the posterior (mix) faster than regular MCMC.

4.4 Comparison of IRMCMC and regular MCMC with respect to mixing

Table 4.3 displays $R_{K,M}$, the ratio of the KS measure corresponding to IRMCMC and regular MCMC respectively, each having a run of length $K \times M$. The proposal mechanism of x has been kept the same for both regular MCMC and IRMCMC. In particular, a normal random walk proposal kernel with variance close to the optimum variance has been used. A large (> 1) value of $R_{K,M}$ indicates that regular MCMC is more accurate than IRMCMC given respective values of K and M. On the other hand, a ratio less than one says otherwise. We observe that for $K = 1$, IRMCMC is generally more accurate than regular MCMC when M is relatively small but for large M, regular MCMC seems better. This is because the posterior correlation between x and θ makes the autocorrelation in the regular MCMC samples of x higher than in the IRMCMC case for each fixed θ and makes small sample sizes less adequate in the former case. This has been supported by our experiments (not shown). The last column, which gives the ratios when $K = 50$ and $M = 100$, show that IRMCMC and regular MCMC produce quite similar results in some cases, but the former is significantly better than the latter in some other cases. This is in accord with the previous columns and suggests that it is worth taking $M > 1$, of moderate value, and $K > 1$, but relatively small.

This provides insight into the real advantage of IRMCMC over regular MCMC. For even where x and θ are low dimensional, IRMCMC mixes as fast as, and sometimes faster than, regular MCMC. But the computational cost of IRMCMC does not increase with the dimensionality of θ. Thus in situations where θ is very high-dimensional, simulating just a few values of θ by IR and running an MCMC chain for the low-dimensional variable x is clearly very much less expensive than simulating a large number of them by regular MCMC methods.

The above example, although demonstrating the superiority of IRMCMC over regular

MCMC, does not provide any way to determine N, K and M. For a proper choice of the initial sample size N see Jones and Hobert (2001). Once N is determined it remains to choose suitable values for K and M. Rubin (1987) suggests that $N/K = 20$ should be adequate, although he does mention an adaptive scheme for selecting N/K which may warrant further study. In the next section we describe a 'quick and generic' method of determining sample sizes K and M. In accordance with the theory provided in Section 4.2 we propose to vary M rather than keep it fixed.

4.5 Determination of the run length of IRMCMC

Our proposal of determining K and the run length of each regular MCMC, for fixed θ, is based on the assessment of Monte Carlo error. The subsample size of θ for each case could be selected as in the *iid* case, i.e., K should be such that the variance of the sample mean of θ falls below a certain pre-specified level.

Next, for each $k = 1, \cdots, K$ and given $\tilde{\theta}^{(k)}$, let $\bar{h}_{n_k}$ denote the ergodic average of a function of interest $h(x)$; the average taken over realised MCMC samples of x. We assume that the following central limit theorem holds (see Jones and Hobert (2001) for conditions under which the assumption holds)

$$\sqrt{n_k}\left(\bar{h}_{n_k} - E_{\tilde{\theta}^{(k)}}h\right) \xrightarrow{D} N(0, \sigma_k^2) \tag{4.16}$$

where $E_{\tilde{\theta}^{(k)}}h$ is the expected value of h given $\tilde{\theta}^{(k)}$. Convergence in distribution is denoted by '$\xrightarrow{D}$'.

Then given an estimate of σ_k^2, it is possible to get an asymptotic standard error for $\bar{h}_{n_k}$. We break up the run of the sampler into batches of equal size that are assumed to be approximately independent. Specifically, we suppose that for given $\tilde{\theta}^{(k)}$ a Markov chain is run for $n_k = a_k \times b$ iterations where b is large enough so that the quantities

$$S_j = \frac{1}{b}\sum_{k=(j-1)b}^{jb-1} h(x^{(k)})$$

are approximately independently $N(E_\pi h, \frac{\sigma_k^2}{b})$ for $j = 1, \cdots, a_k$. The batch means estimate of σ_k^2 is

$$\hat{\sigma}_k^2 = \frac{b}{a_k - 1} \sum_{j=1}^{a_k} (S_j - \bar{h}_{n_k})^2 \tag{4.17}$$

The desired accuracy for $\bar{h}_{n_k}$ can be specified by specifying the asymptotic variance and running the regular MCMC algorithm as long as the variance is greater than specified. The iteration at which the variance falls below this may be regarded as a random stopping time. See Section 4.2 for the connection of stopping time with IRMCMC. However, our intention is to specify an accuracy level after combining MCMC runs of x for each $\tilde{\theta}^{(k)}$. One way to do this is to consider the pooled mean

$$\bar{h}_N = \frac{1}{K} \sum_{k=1}^{K} \bar{h}_{n_k}$$

where $N = \sum_{j=1}^{K} n_k$. Assuming that $\bar{h}_{n_k}$ are approximately independent, it follows that $\bar{h}_n$ is $N(E_\pi h, \sigma^2)$, where

$$\sigma^2 = \frac{1}{K^2} \sum_{k=1}^{K} \frac{\sigma_k^2}{n_k}$$

If we now consider a particular definition of a_k, given by

$$a_k = \inf \left\{ m \in \{1, 2, \cdots\} : \frac{\sigma_k^2}{mbK} \leq \epsilon \right\}, \tag{4.18}$$

then it implies that $\sigma^2 < \epsilon$. Observe that the performances of two diferent algorithms can be compared at each case using the total run length N. Also, an overall measure of performance may be obtained by the total run length corresponding to all cases. We next compare the performances of regular MCMC and IRMCMC in the context of the Poisson regression using the above theory.

Table 4.4 shows the run lengths needed by regular MCMC and IRMCMC to achieve an accuracy of $\epsilon = 0.001$ in (4.18). For IRMCMC, case 8 has been omitted. In each case IRMCMC requires a shorter run than regular MCMC to attain the same level of accuracy. The total run length for 10 cases needed by IRMCMC is much less that needed by regular MCMC. This superiority of IRMCMC over regular MCMC is in agreement with the evidence

Table 4.4: Comparison between regular MCMC and IRMCMC using run lengths necessary to attain a given accuracy.

Case	MCMC	IRMCMC
1	28590	13210
2	44460	33910
3	450	240
4	16690	4640
5	600	160
6	4770	2350
7	15910	8180
8	13000	13000
9	680	230
10	28360	14680
Total	153510	90600

provided in Section 4.2. In the case of high dimensional problems, where $\theta = (\theta_1, \cdots, \theta_p)$ and $x = (x_1, \cdots, x_q)$; p and q being the dimensionality of θ and x respectively, this method is not directly applicable. In such cases, Monte Carlo errors of $\bar{\theta} = \sum_{j=1}^{p} \theta_j/p$ and $\bar{x} = \sum_{j=1}^{q} x_j/q$ may be considered. We remind the reader here that the cost of implementation of IRMCMC does not rise with the dimensionality of θ.

Thus we have proposed a method, to which we refer as IRMCMC, that is much more advantageous than regular MCMC in terms of both computation speed and mixing. We have demonstrated that IRMCMC is particularly more advantageous than regular MCMC in high dimensional problems.

The choice of case i^* may require further discussion; we do not claim to have given any optimality condition with respect to the distance measure used. But clearly, when the size of the data is sufficiently large, then any value of $i^* \in \{1, \cdots, n\}$ is appropriate. In such a case, IR could be replaced by simple random sampling. This will be discussed in Chapter 5.

Chapter 5

Asymptotics associated with cross-validation in inverse problems

In this chapter we demonstrate that, in the case of large data size, any case can be omitted to obtain the initial sample needed for IRMCMC. This we do by showing that for large samples the leave-one-out posteriors of θ as well as the saturated posterior are nearly the same. This is important since in some situations importance weights, even with respect to leave-one-out posteriors may not be available. We provide such an example in Chapter 8. In such cases, relying on asymptotic theory, we propose to implement simple random sampling from the saturated posterior.

The concept of approximate equality of posteriors of θ for large data size is based on 'consistency' of the posteriors. The phenomenon of concentration of the posteriors around the true value as more and more data come in is called consistency. In Section 5.1 we explain this in some detail and explain that the posteriors (leave-one-out and the saturated) of θ concentrates around θ_0, the true value of θ, as the data size n tends to infinity. So, for large data sets, the posteriors are nearly the same and hence realisations obtained from any posterior can safely be re-used for all others. In fact, IR may be replaced by simple random sampling. Our exposition in Section 5.1 hinges on the assumption that θ is a finite

dimensional parameter.

5.1 Consistency

In Bayesian analysis, one starts with a prior knowledge (sometimes imprecise) expressed as a distribution on the parameter space and updates the knowledge according to the posterior distribution given the data. It is therefore important to know whether the updated knowledge becomes more and more precise as data are collected indefinitely. This requirement is called consistency of the posterior distribution. Although it is an asymptotic property, consistency is important since the violation of consistency is clearly undesirable and one may have serious doubts against inferences based on an inconsistent posterior distribution.

DEFINITION. A posterior distribution π_n is said to be consistent at θ_0 if for every neighbourhood U of θ_0, $\pi_n(U) \to 1$ almost surely under the law determined by θ_0 (in short, a.s. $[\theta_0]$).

The above formulation of consistency is from the point of view of classical Bayesians who believe in an unknown true model. Indeed, θ_0 in our case can be thought of as the unknown true parameter which gives rise to the observed data. Subjective Bayesians however, reject ideas involving an unknown true model and think only in terms of the predictive distribution of a future observation. Nevertheless, consistency is important to subjective Bayesians as well. Blackwell and Dubins (1962) and Diaconis and Freedman (1986) show that consistency is equivalent to intersubjective agreement, meaning that two Bayesians will ultimately have very close predictive distributions.

Ibragimov and Has'minskii (1981), henceforth referred to as IH, have used a very general framework for parametric models that include both regular and non-regular problems. In fact IH verify their conditions for various classes of non-regular problems and some stochastic processes as well. Within their framework, Ghosh and Ramamoorthi (1994) provide a nec-

essary and sufficient condition for a suitably normed posterior to have a limit in probability. The theory of IH is based on the following three basic conditions on the likelihood ratios described below. These conditions hold for almost all parametric problems of interest and the iid assumption is not important here at all.

To describe these conditions, fix a parameter point θ_0 and consider $Z_n(u)$ as the ratio of the likelihoods at the points $\theta_0 + \psi_n u$ and θ_0, where ψ_n is the appropriate normalizing factor. For example, $\psi_n = n^{-1/2}$ in the regular (smooth) cases.

Conditions (IH)

(IH1) For some constants $M, m, \alpha > 0$,

$$E|Z_n^{1/2}(u_1) - Z_n^{1/2}(u_2)|^{1/2} \leq M(1 + R^m)|u_1 - u_2|^{\alpha}$$

for every u_1, u_2 with $|u_1|, |u_2| \leq R$ and $R > 0$.

(IH2) $EZ_n^{1/2}(u) \leq \exp[-g_n(\| u \|)]$, where $g_n : (0, \infty) \to (0, \infty)$ is an increasing function satisfying $\lim_{n \to \infty, y \to \infty} y^N \exp[-g_n(y)] = 0 \forall N \geq 0$.

(IH3) Finite dimensional distributions of the stochastic process $Z_n(\cdot)$ converge to the corresponding finite dimensionals of another precess $Z(\cdot)$.

Roughly speaking, (IH1) implies the mean-square continuity of the process $Z_n^{1/2}(\cdot)$ whereas (IH2) controls the tail behaviour. Condition (IH3) is the most important one, deciding the limiting distribution. IH obtained the limiting distribution of Bayes estimates and of the maximum likelihood estimate (MLE) in terms of the process $Z(\cdot)$. Thus, in particular, Bayes estimates are consistent in the frequentist sense. In fact, much more is true – Bayes estimates are always (asymptotically) efficient while the MLE need not always be so. Thus even as merely frequentist estimators, Bayes estimates are at least as good as any other competitor.

We next argue that consistency implies that for a large sample size any two leave-one-out posteriors of θ are nearly the same.

5.1.1 Consistency and its connection with robustness with respect to the choice of i^*

In the above general set up is very convenient for studying asymptotic properties of the posterior distribution and we can extract many useful information about the posterior (see Ghosal et al. (1995), Ghosh et al. (1994)). First, the posterior is consistent under (IH1) and (IH2) if the prior density is positive and continuous at θ_0 and grows at most like a polynomial. In fact, with large probability, most of the mass of the posterior is concentrated in a neighbourhood of size ψ_n around θ_0. As a result of a theorem in Ghosal et al. (1995), in finite dimension, consistency implies that the posterior based on any two reasonable priors are almost the same. Thus in large samples, Bayesian methods are insensitive to the choice of the prior.

We show that any two leave-one-out posteriors of θ are asymptotically equivalent but those of x are not. This we prove by showing that the joint leave-one-out posterior with case i omitted is asymptotically equivalent to the saturated posterior. Our result hinges on the observation that the saturated posterior, $\pi(\theta \mid X, Y)$, can be thought of as a leave-one-out posterior with any arbitrary case i omitted. The corresponding random variable x has a prior that puts all the mass on x_i. (see also Chapter 2, Section 2.2); the prior on (x, θ) being

$$\pi^{(1)}(x, \theta) = \delta_{x_i}(x)\pi(\theta) \tag{5.1}$$

where $\pi(\theta)$ denotes a prior density that is positive and continuous at θ_0. δ_{x_i} denotes point mass on x_i. In this chapter, for notational convenience, we denote the saturated posterior by $\pi_n(x, \theta)$. Note that the posterior depends on n through $X = \{x_i\}$ and $Y = \{y_i\}$; $i = 1, \cdots, n$. The leave-one-out posterior with case i omitted where x has a prior that does not give point mass to x_i will be denoted by $\pi_n^{(i)}(x, \theta)$. The prior in this case will be denoted by $\pi^{(2)}(x, \theta)$. We assume that $\pi^{(2)}$ is positive and continuous at the true value (x_i, θ_0).

We denote the parameter space by $\mathcal{X} \times \Theta$ and assume that this is a separable metric space endowed with its Borel σ-algebra $\mathcal{B}(\mathcal{X} \times \Theta)$. For each $(x, \theta) \in \mathcal{X} \times \Theta$, $P_{(x,\theta)}$ is a probability

measure on a measurable space $(\mathbf{Y}, \mathcal{A})$, such that for all $A \in \mathcal{A}$, $(x, \theta) \to P_{(x,\theta)}(A)$ is $\mathcal{B}(\mathcal{X} \times \Theta)$-measurable.

We find it convenient to think of $y_1, y_2, \cdots$ as the co-ordinate random variables defined on $\Omega = (\mathbf{Y}^\infty, \mathcal{A}^\infty)$ and $P^\infty_{(x,\theta)}$ as the product measure defined on Ω.

Below we present our main result in the form of a theorem.

Theorem 2

$$\lim_{n\to\infty} \int_{\mathcal{X}\times\Theta} |\pi_n(x,\theta) - \pi_n^{(i)}(x,\theta)| dx d\theta = 0 \quad a.s.\ [\theta_0] \tag{5.2}$$

PROOF: Observe that

$$\begin{aligned}
&\int_{\mathcal{X}\times\Theta} |\pi_n(x,\theta) - \pi_n^{(i)}(x,\theta)| dx d\theta \\
&= \int_{\mathcal{X}-\{x_i\}\times\Theta} |\pi_n(x,\theta) - \pi_n^{(i)}(x,\theta)| dx d\theta \\
&+ \int_{\Theta} |\pi_n(x_i,\theta) - \pi_n^{(i)}(x_i,\theta)| d\theta \\
&= \int_{\mathcal{X}-\{x_i\}\times\Theta} \pi_n^{(i)}(x,\theta) dx d\theta \qquad (5.3) \\
&+ \int_{\Theta} |\pi_n(x_i,\theta) - \pi_n^{(i)}(x_i,\theta)| d\theta \qquad (5.4)
\end{aligned}$$

Since by consistency $\pi_n^{(i)}(x,\theta)$ converges almost surely to the point mass at (x_i, θ_0), the integral (5.3) trivially tends to zero almost surely as $n \to \infty$. All we need to show is that the integral (5.4) tends to zero almost surely as $n \to \infty$. We need to show that for ω in a set of $P^\infty_{(x_i,\theta_0)}$ measure one

$$\int_{\Theta} \pi_n^{(i)}(x_i,\theta) \left| 1 - \frac{\pi_n(x_i,\theta)}{\pi_n^{(i)}(x_i,\theta)} \right| d\theta \to 0 \tag{5.5}$$

Now,

$$\frac{\pi_n(x_i,\theta)}{\pi_n^{(i)}(x_i,\theta)} = \frac{\pi(\theta)}{\pi^{(2)}(x_i,\theta)} \frac{\int \pi^{(2)}(x,\theta) L(Y, X_{-i}, x, \theta) dx d\theta}{\int \pi(\theta) L(Y, X_{-i}, x_i, \theta) d\theta} \tag{5.6}$$

Let $\delta > 0$, $\eta > 0$ and $\epsilon > 0$ be given. We choose a neighbourhood $U_1 \times U_2$ of (x_i, θ_0), such that for all $(x,\theta) \in U_1 \times U_2$, $|\pi^{(2)}(x_i,\theta_0) - \pi^{(2)}(x,\theta)| < \delta$, $|\pi^{(1)}(x_i,\theta_0) - \pi^{(1)}(x_i,\theta)| < \delta$

and $\left|\frac{\pi_1(x_i,\theta)}{\pi_2(x_i,\theta)} - \frac{\pi_1(x_i,\theta_0)}{\pi_2(x_i,\theta_0)}\right| < \delta$. By consistency there exists Ω_0, $P^{\infty}_{(x_i,\theta_0)}(\Omega_0) = 1$, such that for $\omega \in \Omega_0$,

$$\frac{\int_{U_1\times U_2} \pi^{(j)}(x,\theta)L(Y,X_{-i},x,\theta)dxd\theta}{\int_{\mathcal{X}\times\Theta} \pi^{(j)}(x,\theta)L(Y,X_{-i},x,\theta)dxd\theta} \to 1 \text{ for } j = 1,2 \tag{5.7}$$

Now fix $\omega \in \Omega_0$, choose n_0 such that for $n > n_0$,

$$\int_{\mathcal{X}\times\Theta} \pi^{(j)}(x,\theta)L(Y,X_{-i},x,\theta)d\theta \leq (1-\eta)^{-1}\int_{U_1\times U_2} \pi^{(j)}(x,\theta)L(Y,X_{-i},x,\theta)d\theta \text{ for } j = 1,2 \tag{5.8}$$

For given $\zeta > 0$ and for $n > n_0$, we must also have that

$$\left|\frac{\int_{U_1\times U_2} L(Y,X_{-i},x,\theta)dxd\theta}{\int_{U_1\times U_2} L(Y,X_{-i},x_i,\theta)d\theta} - 1\right| < \zeta \tag{5.9}$$

For $(x,\theta) \in U_1 \times U_2$,

$$\begin{aligned}
&\left(\frac{\pi(\theta_0)}{\pi^{(2)}(x_i,\theta_0)} - \delta\right)(1-\eta)\left(\frac{\pi^{(2)}(x_i,\theta_0)-\delta}{\pi(\theta_0)+\delta}\right)(1-\zeta) \\
\leq\ & \frac{\pi_n(x_i,\theta)}{\pi_n^{(i)}(x_i,\theta)} \\
\leq\ & \left(\frac{\pi(\theta_0)}{\pi^{(2)}(x_i,\theta_0)} + \delta\right)(1-\eta)^{-1}\left(\frac{\pi^{(2)}(x_i,\theta_0)+\delta}{\pi(\theta_0)-\delta}\right)(1+\zeta)
\end{aligned} \tag{5.10}$$

Hence, for δ, η, ζ small,

$$\left|\frac{\pi_n(x_i,\theta)}{\pi_n^{(i)}(x_i,\theta)} - 1\right| < \epsilon \tag{5.11}$$

Thus, for $n > n_0$,

$$\begin{aligned}
&\int_{\Theta} \left|\pi_n(x_i,\theta) - \pi_n^{(i)}(x_i,\theta)\right| d\theta \\
\leq\ & \int_{U_2} \pi_n^{(i)}(x_i,\theta)\left|\frac{\pi_n(x_i,\theta)}{\pi_n^{(i)}(x_i,\theta)} - 1\right| d\theta + 2\eta \\
\leq\ & \epsilon(1-\eta) + 2\eta
\end{aligned}$$

Hence the proof of (5.2) is complete.

Note that this shows that for large samples it is immaterial which site is omitted. Samples of θ drawn corresponding to any case is re-usable. However, since the corresponding leave-one-out posterior distributions of x converge to the true value x_i and hence samples of x should not be re-used.

In the next chapter we discuss assessment of model fit and model comparison. We propose a new method for model assessment based on cross-validation with IRMCMC and also explain that it can be used to compare different models.

Chapter 6

Application of IRMCMC to the chironomid model

So far we have illustrated the theory of IRMCMC with the help of toy examples. However, it is important to show that it works also in the case of practical problems.

Vasko et al. (2000) reported a regular MCMC cross-validation exercise for a data set comprising multivariate counts y_i on $m = 52$ species of chironomid at $n = 62$ lakes (sites) in Finland. The unidimensional x_i denote mean July air temperature. As species respond differently to summer temperature, the variation in the composition provides the analyst with information on summer temperatures. This information is exploited to reconstruct past climates from count data derived from fossils in the lake sediment; see Korhola et al. (2002).

The cross-validation exercise was computationally challenging, requiring 62 separate regular MCMC exercises and involved a parameter θ of dimension 3318. However, implementation of cross-validation by regular MCMC is not infeasible in this case. But the problem seems to be an ideal real life problem where the performance of IRMCMC can be tested by making comparison with regular MCMC. In Chapter 7 we discuss the model proposed by Whiley et al. (2004) where it shall also be argued that reguar MCMC based cross-validation

is clearly infeasible.

In the case of Vasko et al. (2000), our MCMC implementation took 16 hours. In contrast, the IRMCMC implementation took 16 minutes for the initial run and 20 minutes for the remaining 61. Additionally, IRMCMC drew attention to the bimodality of one of the posteriors, a point completely missed by the MCMC implementation. We provide details of this below. But first we explain the high dimensionality of θ and our implementation of cross-validation by IRMCMC.

6.1 Model description

In Vasko et al. (2000), the vector y_i of counts at site i followed the multinomial distribution,

$$(y_i \mid y_{i+}, \mathbf{p}_i) \sim Multinomial(y_{i+}, \mathbf{p}_i). \tag{6.1}$$

Here $y_i = (y_{i1}, \cdots, y_{im})$, $y_{i+} = \sum_{k=1}^{m} y_{ik}$ and $\mathbf{p}_i$ is an (unobserved)vector of relative abundances $(p_{i1}, \cdots, p_{im})$, of dimensionality $(m-1) = 51$. We denote the multinomial likelihood as

$$L(y_i \mid y_{i+}, \mathbf{p}_i) = \frac{(y_{i+})!}{\prod_{k=1}^{52} y_{ik}!} \prod_{k=1}^{52} p_{ik}^{y_{ik}} \tag{6.2}$$

The unobserved $\{\mathbf{p}_i; i = 1, \cdots, n\}$, thus provide 62×51 parameters, even before temperature x_i is related to the relative abundances. Vasko et al. (2000) related these via a Dirichlet model,

$$(\mathbf{p}_i \mid x_i, \Psi_1, \cdots, \Psi_{52}) \sim Dirichlet(\Lambda_i). \tag{6.3}$$

where the kth component λ_{ik} of Λ_i was modelled as $\lambda_{ik} = \lambda(x_i, \Psi_k)$, for a simple function λ of x_i and of $\Psi_k = (\alpha_k, \beta_k, \gamma_k)$, a 3-component parameter vector associated with the kth species. Vasko et al. (2000) chose a simple unimodal function of these species specific parameters, given by

$$\lambda(x_i, \Psi_k) = \alpha_k \exp\left[-\left(\frac{x_i - \beta_k}{\gamma_k}\right)^2\right] \tag{6.4}$$

The mode, β_k, represents the value of temperature at which the species k is most abundant. Tolerance of the species is denoted by γ_k and α_k is a scaling factor. There are thus an additional 3×52 parameters, yielding 3318 in total. We write $\theta = \{\mathbf{p_1}, \cdots, \mathbf{p_{62}}, \Psi_1, \cdots, \Psi_{52}\}$. For choice of priors on ψ_k, see Vasko et al. (2000). With the above notation, the density of the Dirichlet distribution (6.3) is given by

$$\begin{aligned} \pi(\mathbf{p}_i \mid x_i, \Psi_1, \cdots, \Psi_{52}) &= \frac{\Gamma(\sum_{k=1}^{52} \lambda_{ik})}{\prod_{k=1}^{52} \Gamma(\lambda_{ik})} \prod_{k=1}^{52} p_{ik}^{\lambda_{ik}-1} \\ &= \frac{\Gamma(\sum_{k=1}^{52} \lambda(x_i, \Psi_k))}{\prod_{k=1}^{52} \Gamma(\lambda(x_i, \Psi_k))} \prod_{k=1}^{52} p_{ik}^{\lambda(x_i,\Psi_k)-1} \end{aligned} \tag{6.5}$$

Thus, the joint posterior with site i^* omitted can be written as

$$\begin{aligned} \pi(x, \theta \mid X_{-i^*}, Y) &\propto \pi(x) \left(\prod_{k=1}^{52} \pi(\Psi_k) \right) \\ &\times \pi(\mathbf{p}_{i^*} \mid x, \Psi_1, \cdots, \Psi_{52}) L(y_{i^*} \mid y_{i^*+}, \mathbf{p}_{i^*}) \\ &\times \prod_{i \neq i^*, i=1}^{62} \pi(\mathbf{p}_i \mid x_i, \Psi_1, \cdots, \Psi_{52}) L(y_i \mid y_{i+}, \mathbf{p}_i) \\ &= \pi(x) \left(\prod_{k=1}^{52} \pi(\Psi_k) \right) \\ &\times \frac{\Gamma(\sum_{k=1}^{52} \lambda(x, \Psi_k))}{\prod_{k=1}^{52} \Gamma(\lambda(x, \Psi_k))} \prod_{k=1}^{52} p_{i^*k}^{\lambda(x,\Psi_k)+y_{i^*k}-1} \\ &\times \prod_{i \neq i^*, i=1}^{62} \frac{\Gamma(\sum_{k=1}^{52} \lambda(x_i, \Psi_k))}{\prod_{k=1}^{52} \Gamma(\lambda(x_i, \Psi_k))} \prod_{k=1}^{52} p_{ik}^{\lambda(x_i,\Psi_k)+y_{ik}-1} \end{aligned} \tag{6.6}$$

In the next section we describe in detail cross-validation of this model using IRMCMC.

6.2 Cross-validation of the model using IRMCMC

It follows directly from the leave-one-out joint posterior distribution (up to a proportionality constant) given by (6.6), that the importance weight function corresponding to site i, leaving out site i^*, is given by

$$w_{i,i^*}(x, \theta) = \frac{\pi(\mathbf{p}_i \mid x, \Psi_1, \cdots, \Psi_{52})}{\pi(\mathbf{p}_{i^*} \mid x, \Psi_1, \cdots, \Psi_{52})} \frac{\pi(\mathbf{p}_{i^*} \mid x_{i^*}, \Psi_1, \cdots, \Psi_{52})}{\pi(\mathbf{p}_i \mid x_i, \Psi_1, \cdots, \Psi_{52})} \tag{6.7}$$

It is clear from (6.5) that the normalizing constants involved in the first ratio of the above expression cancel. Also observe that the importance weights are independent of the count data y_i, since the individual multinomial likelihood terms consistent with Equation 6.1 are independent of x_i and hence cancel in the ratio. We remark that this was not the case with the Poisson regression example. Hence, in this case, as noted in Section 4.3, i^* is such that $x_{i^*} = median(X)$ or $mean(X)$, depending on the distance measure chosen. However, in our case, the distribution of X is approximately symmetrical; hence it is immaterial which distance function is selected. From the set of medians of X, we arbitrarily choose site 38. For the implementation of IRMCMC, we performed many experiments with varying N, K and M. However, the differences in results were insignificant. We report results for $N = 5000, K = 50, M = 100$.

The regular MCMC needed for the posterior with site i omitted has been implemented by first drawing each $\mathbf{p}_i$ from the Dirichlet distribution with parameters $(\lambda_{ik} + y_{ik})$. Observe that during this step, other parameters are held fixed. Ψ_k are proposed from a trivariate normal random walk distribution and accepted or rejected in accordance with the Metropolis-Hastings acceptance probability. The random variable x is then proposed from a univariate normal random walk distribution. The variances of the proposal distributions have been approximately optimised for adequate exploration of the posterior; see Chapter 2.

After obtaining the output from the regular MCMC case, we select realisations of Ψ_k by IR. Using them, we explore the conditional joint density of $(x, \mathbf{p_i})$ by MCMC. Note that $\mathbf{p}_i$ depends on the unknown x, so we consider it jointly with x for the MCMC purpose. We remark here that since $\mathbf{p}_i$ are conditionally independent, realisations of $\{\mathbf{p}_j; j \neq i\}$ are not necessary. Restarts are used to conduct MCMC for each realisation of Ψ_k. The updating of $\mathbf{p}_i$ in this case has been done by direct simulation from the Dirichlet distribution with parameters $(\lambda(x, \Psi_k) + y_{ik})$and hence does not require evaluation of acceptance probability as needed in a Metropolis-Hastings algorithm. Thus this step is computationally very efficient. Only x needs to be updated in accordance with the Metropolis-Hastings procedure; but since

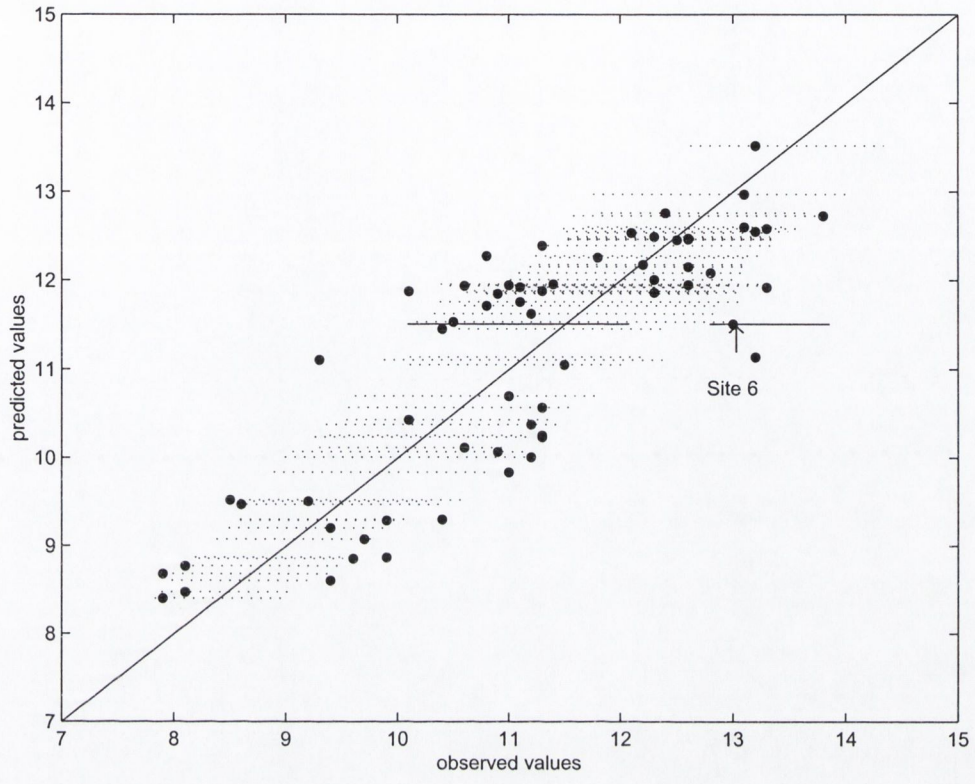


Figure 6.1: Cross-validation using IRMCMC. Dotted lines denote 95% HPD credible intervals. At site 6, the 95% HPD credible interval is disjoint.

this is a one dimensional random variable, it causes absolutely no computational burden.

95% HPD credible regions (see Chapter 2) and the associated observed temperatures are presented in Figure 6.1. Observe, that at site 6, the 95% HPD credible region is disjoint. Note that more than 35% of the observed temperatures fell outside the 95% HPD credible regions, suggesting poor model fit. Now we present a detailed comparison between IRMCMC and regular MCMC.

Figures 6.2, 6.3 and 6.4 show the density estimates of temperature obtained by regular MCMC and IRMCMC for the 62 sites. Distance values with respect to d_1, scaled to lie within [0,1] are also shown in each case. Note that IRMCMC and regular MCMC closely agree with each other at all cases except site 6. In fact, in that case regular MCMC explores a unimodal posterior but IRMCMC explores a distribution with two modes, one being a

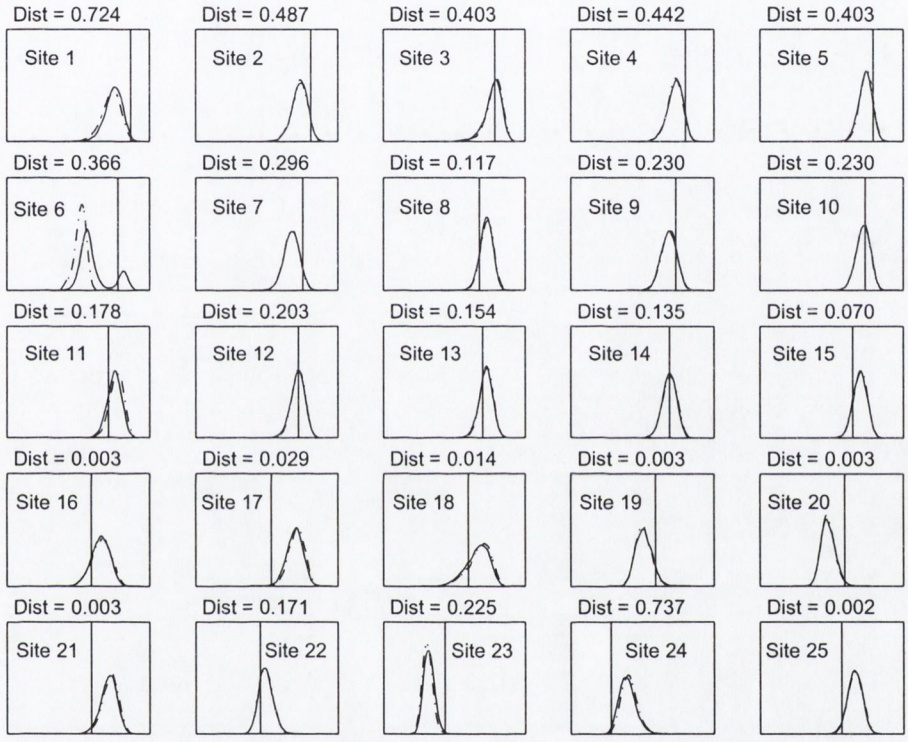


Figure 6.2: Comparison between MCMC (dotted line) and IRMCMC (solid line) for sites 1-25. Distance d_1 scaled to lie within [0,1] are shown in each case. Vertical line denotes observed value.

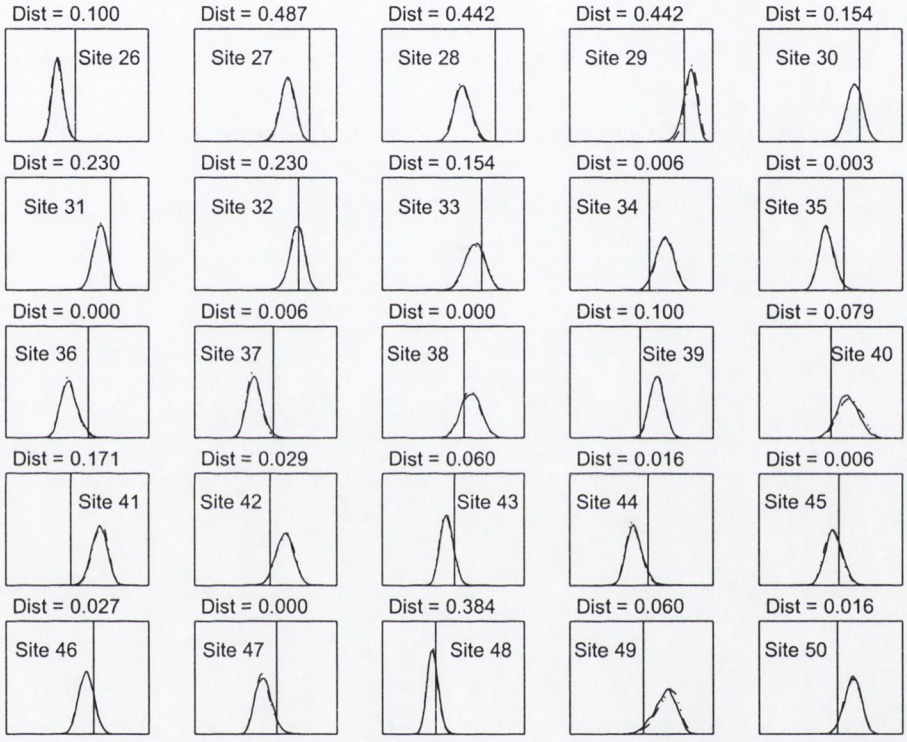


Figure 6.3: Comparison between MCMC (dotted line) and IRMCMC (solid line) for sites 26-50. Distance d_1 scaled to lie within [0,1] are shown in each case. Vertical line denotes observed value.

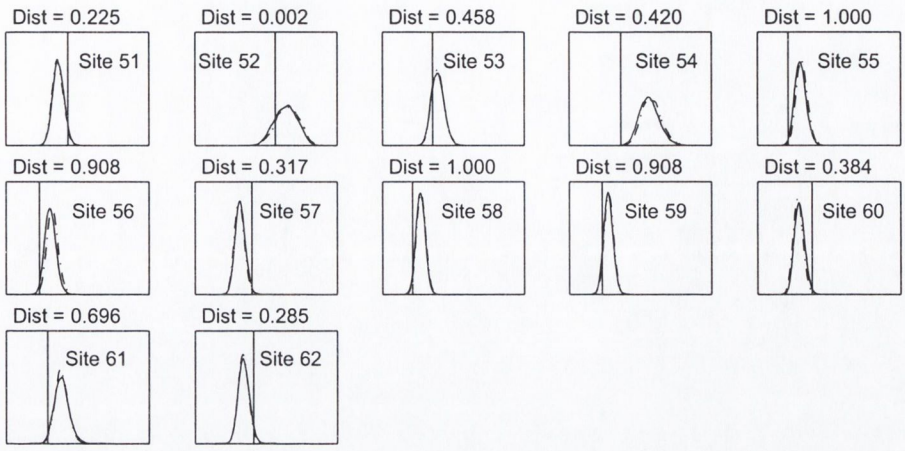


Figure 6.4: Comparison between MCMC (dotted line) and IRMCMC (solid line) for sites 51-62. Distance d_1 scaled to lie within [0,1] are shown in each case. Vertical line denotes observed value.

minor mode, which in fact explains the observed datum as indicated by the vertical line. Indeed, a bimodal posterior is not unexpected at that site since the species composition is dominated by the presence of a very large count of a particular species. The situation may be interpreted as the abundant species and the remaining species have preference for disjoint regions of the climate space. Thus they send conflicting signals for these two regions, resulting in bimodality. For more details on bimodality, see Chapter 7. In Section 6.2.2 we describe a simulation study to justify our claim of bimodality under these circumstances.

Distributions of normalised importance weights are displayed in Figure 6.5. Clearly, the distributions in most cases are far from uniform and many big jumps are noted. It is worthwhile to note that the distribution of the importance weights played no important role in the accuracy of IRMCMC. This is clear from the agreement between regular MCMC and IRMCMC. Certainly, this is not unexpected, as IR without replacement provides a great deal of protection against extreme weights and realisations of θ, not x, are re-used. This has been explained in detail in Chapters 2 and 4. We recall that the leave-one-out posteriors of θ are expected to be very similar, but the dissimilar leave-one-out posteriors of x cause the

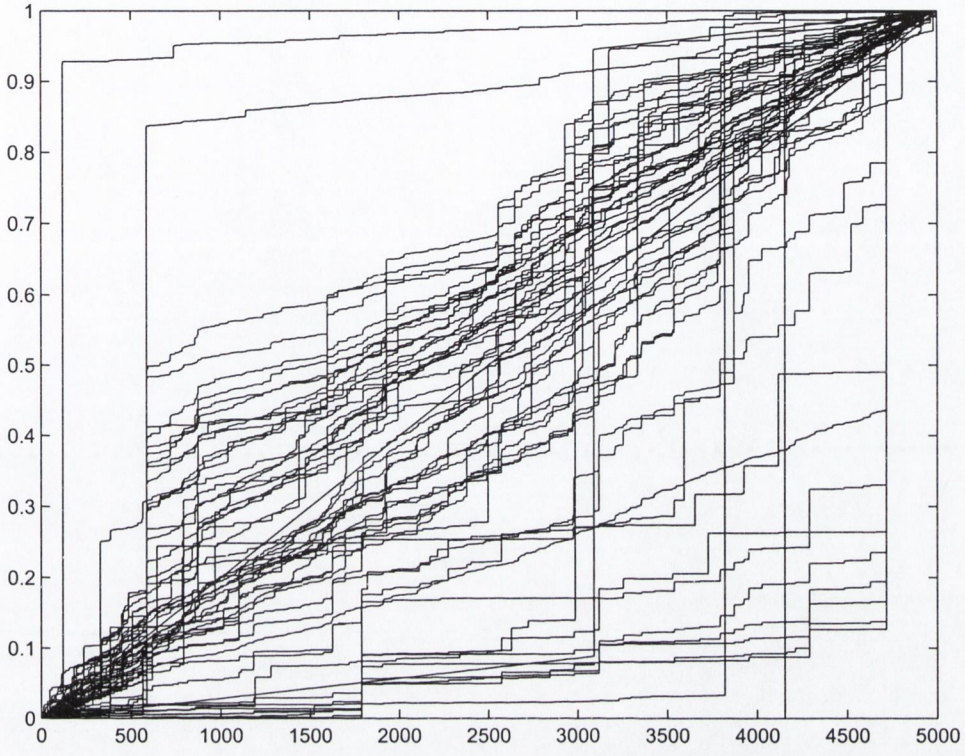


Figure 6.5: Cumulative normalised importance weights.

distributions of importance weights to be different from uniform.

In cases where weights were extremely close to zero, we replaced them by a small positive quantity in such a manner that the weights sum to one. This is clearly a questionable practice, but very close agreement of IRMCMC and regular MCMC confirms that it does not lead to unreliability. In the cases of maximum distance sites (sites 55 and 58; see Figure 6.4), all the weights were very close to zero. In such cases we made the *ad hoc* assumption that the weights are all equal. For finite number of parameters and large data size, the theorem proved in Chapter 5 justifies the above assumption. But note that in this problem, the number of parameters increases with the data size. But the very close agreement of IRMCMC with regular MCMC, even for this non-regular problem, suggests a considerable robustness of our approach.

It is important to point out that since the Dirichlet distribution is a natural conjugate

prior of multinomial distribution it is indeed possible to simplify the model by analytically integrating out $\{\mathbf{p}_i\}$ and reducing computational labour. The simplified posterior with site i^* omitted, with $\theta = \{\Psi_1, \cdots, \Psi_{52}\}$ is then given by

$$\pi(x, \theta \mid X_{-i^*}, Y) \propto \pi(x) \prod_{k=1}^{52} \pi(\Psi_k) \prod_{i=1}^{62} \frac{\Gamma(\sum_{k=1}^{52} \lambda_{ik})}{\prod_{k=1}^{52} \Gamma(\lambda_{ik})} \frac{\prod_{k=1}^{52} \Gamma(\lambda_{ik} + y_{ik})}{\Gamma(\sum_{k=1}^{52} \lambda_{ik} + y_{ik})} \tag{6.8}$$

But our experiments, using exactly the same run length N and same proposal distributions for x and Ψ_k as has been used for the larger problem given by (6.6), suggested that retaining the latent variables $\{\mathbf{p}_i\}$ improves the mixing behaviour of our regular MCMC algorithm needed for the site omitted. In fact, the latent variables allow regular MCMC to visit interesting regions of the θ-space. For instance, the bimodality problem (which is really to be expected; see Section 6.2.2 for justification) was much less prominent even with IRMCMC, when the simplified model was used. It is to be noted in this context that mixing is often improved by augmenting the state space vector to include additional components. See, for example, Edwards and Sokal (1988), Besag and Green (1993).

Next we discuss model assessment using data obtained by cross-validation.

6.2.1 Results of cross-validation and assessment of model fit

It has been found that more than 40% of the observed data lie outside the 95% HPD credible regions, suggesting poor fit of the model to the data. Moreover, the HPD credible regions also provide insight into fit of the datum at individual sites. Hence, this procedure of assessing model fit is quite informative. However, a formal approach to assessing model fit is also desirable. In this section, we also apply the more formal Bayesian hypothesis testing procedure discussed in Chapter 3. There it has been also argued that the approach of looking at individual sites can be regarded as a special case of the Bayesian hypothesis testing procedure using a suitable reference distribution.

An application of the Bayesian hypothesis testing procedure gives

$$p = \pi^* \left(\left| \frac{D_1^{var} - D_1^{obs}}{\sqrt{V_{\pi^*}(D_1^{var} \mid Y)}} \right| \leq \epsilon \mid Y \right) \approx 0.$$

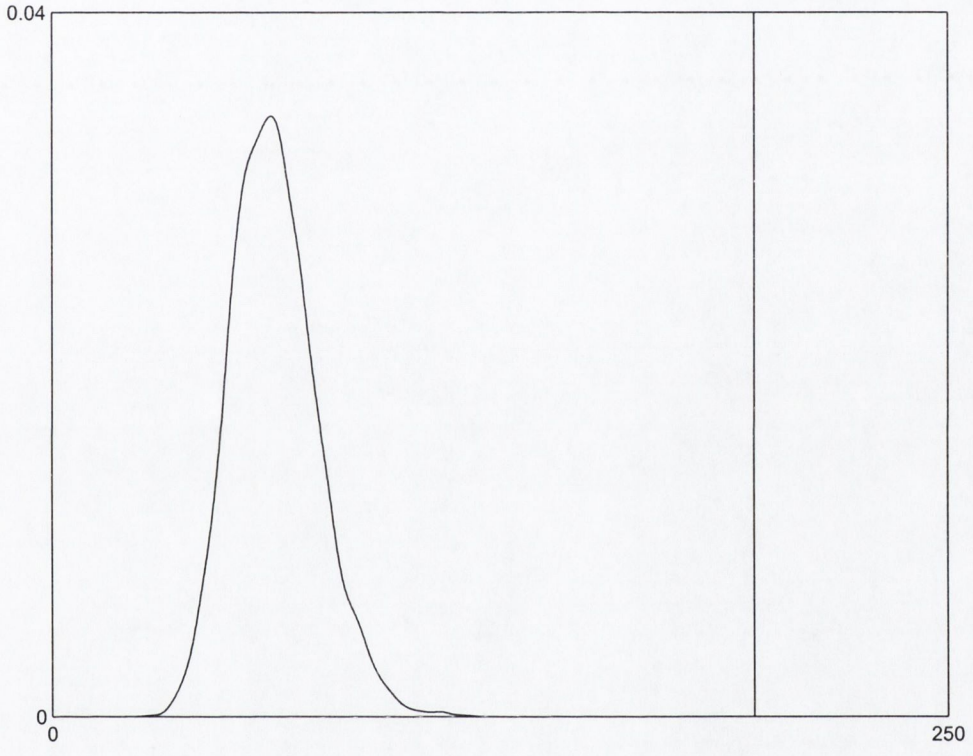


Figure 6.6: Distribution of D_1^{var}; the vertical line is the observed value, D_1^{obs}.

indicating that the model does not fit the data. Observed D_1^{obs} and the distribution of D_1^{var} are shown in Figure 6.6. Note that D_1^{obs} is located far from the mode of D_1^{var}. This is a consequence of the fact that many observed temperature values are far from the modes of the respective posterior density; this is clearly seen in Figures 6.2, 6.3 and 6.4.

Anticipating that the unimodal model used to describe λ_{ik} in (6.4) is inappropriate, we used another modelling approach where, rather than unimodal functions, the proportions were modelled as mixtures of normal densities; the number of components being unknown. We viewed the parameters associated with each component of the mixture as samples arisen from the Dirichlet Process (see, for example, Ferguson (1974), Ferguson (1983), Escobar and West (1995)). This way of modelling automatically induces a prior on the number of components; see Antoniak (1974). In the posterior analysis, the number of components for each species was found to be greater than one with high probability. From the cross-validation exercise with IRMCMC it was found that 82% of the observed values fell within 95% highest posterior density regions. Leave-one-out posteriors obtained using regular MCMC were in agreement with those obtained using IRMCMC. However, the hypothesis test described in Chapter 3 was not satisfied. Next we discuss a simulation study to discuss bimodality.

6.2.2 Simulation study for demonstrating bimodality

For our purpose, we consider a simulated dataset of counts on 5 taxa each present at 10 sites. The data set is shown in Table 6.1. Associated with simulated samples are simulated values of a climate variable. We start off by computing the posterior distribution of the environmental variable at site 1 *twice* with two different initial values for the regular MCMC; one chain starts at 7.5 and the other at 13.3. The densities of the climate variable constructed using the two individual chains are shown in Figure 6.7. Both the densities obtained from the two individual regular MCMC chains are almost the same and there is no reason to expect the posterior distribution to be bimodal.

Now, in the same simulated dataset we increase the value of $y_{11} = 31$ (the count of species

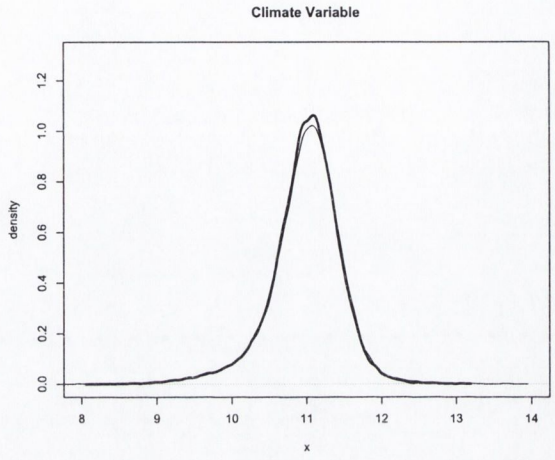


Figure 6.7: Thick line denotes the density of climate variable corresponding to the regular MCMC with initial value 13.3; thin line is the same corresponding to the chain with initial value 7.5.

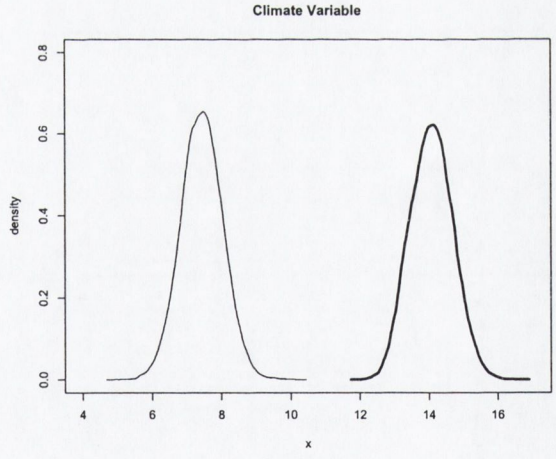


Figure 6.8: Thick line denotes the density of climate variable corresponding to the regular MCMC with initial value 13.3; thin line is the same corresponding to the chain with initial value 7.5.

Table 6.1: Simulated palaeo data

Case	Climate	Species 1	Species 2	Species 3	Species 4	Species 5
1	11.266	31	15	19	81	67
2	11.773	23	2	18	70	39
3	10.762	29	28	9	46	44
4	12.294	50	0	38	111	42
5	13.092	54	0	19	90	49
6	11.857	38	9	30	60	57
7	10.810	14	8	12	40	29
8	11.094	43	15	5	38	39
9	11.283	76	16	27	81	87
10	10.582	18	13	16	74	46

1 at site 1) to 531. In this new dataset so formed, the value of y_{11} could now be viewed as an observation that conflicts with count data on the other species of the same site. As before two chains were run and the resultant densities are shown in Figure 6.8.

We notice that the two different runs explore two disjoint regions unlike the case with the initial dataset. Each chain explores the region associated with the corresponding modes, failing to jump from one modal region to another. The shortcoming of the regular MCMC methodology in exploring bimodal distributions becomes clear in this case.

IRMCMC is now applied to this problem. Also, for comparison with IRMCMC we run an MCMC but with a different proposal mechanism for the climate variable. The proposal density, based on experimental evidence, is given by

$$\frac{1}{2}N(8.5, 1.3^2) + \frac{1}{2}N(13.5, 1.3^2)$$

where $N(8.5, 1.3^2)$ denotes normal density with mean 8.5 and standard deviation 1.3; $N(13.5, 1.3^2)$ is defined analogously. Since the proposed values of x do not depend on the previous values, the above proposal distribution is also called the independence sampler (see, for example, Tierney (1994)).

The results are displayed in Figure 6.9. We find that MCMC with the independence sampler does explore both the modal regions, but it has to be borne in mind that the

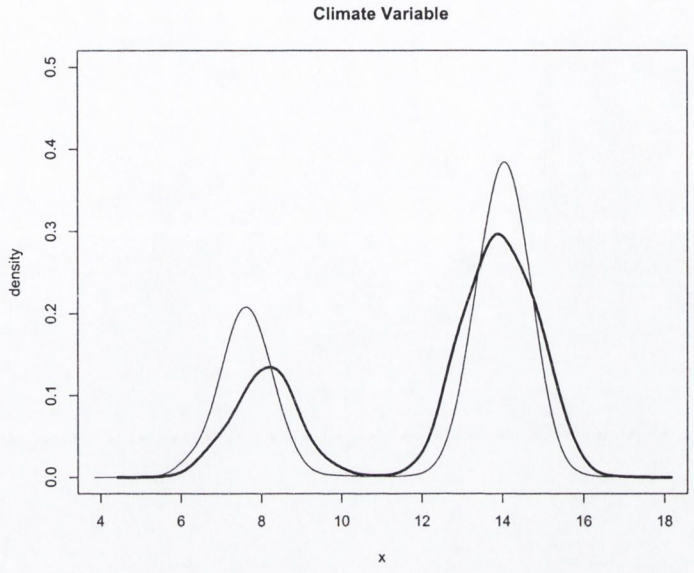


Figure 6.9: Thick Line: IRMCMC-approximated density of environmental variable at site 1. Thin Line: regular MCMC-approximated density.

proposal density has been constructed on the basis of experimental evidence. Such a luxury many not be affordable in practice. However, IRMCMC with restarts explores both the modal regions spontaneously, without requiring any modification. Thus the above simulation study demonstrates two facts; firstly, the occurrence of bimodality may shed light on some conflicting information present in the data set and secondly, it shows the effectiveness of IRMCMC in detecting such situations.

Coming back to the real case of Vasko et al. (2000), we remark that with *a priori* knowledge of bimodality, a suitable MCMC proposal mechanism could have been adopted to explore the posterior predictive distribution at site 6. But the problem was unsuspected and only a later examination of the data revealed it. IRMCMC discovered bimodality without requiring any special proposal mechanism.

6.3 Conclusion

In this problem IRMCMC was seen primarily as a way of achieving computational shortcut. But in practice, it demonstrates its capability of exploring bimodality (that had remained unnoticed by regular MCMC) which in fact provides insight into an interesting issue on conflict between two different kinds of information.

In the next chapter we apply our cross-validation techniques to a more complex and challenging palaeoclimate problem.

Chapter 7

Pollen based reconstruction: a more difficult case

In this chapter we apply our procedures to a very much larger and more complicated model, concerned with pollen based reconstruction as outlined in Chapter 1. For the purposes of cross-validation, there are 7815 compositional records on fourteen taxa, and two climate variables. The model is large, with 9623 parameters. A single cross-validation exercise with regular MCMC takes about 5 hours, on the hardware at our disposal. A crude estimate of the time for the remaining 7814 cross-validation exercises is about five years using brute-force; it is thus infeasible. With IRMCMC we can complete the exercise in about 3 hours.

We outline the model below and show why it is large and difficult. We then implement cross-validation, discussing the performance of IRMCMC. Finally we consider issues of model fit.

7.1 Model description

The compositional pollen data arise from counts. Each vector of proportions represents each of a set of distinguishable taxa in a sample extracted from the sediment. In fact, for many cases, only the proportions were reported and not the total count. We have used total count

$= 400$ in these circumstances, and have thus created count values.

For the purposes of this study there are $m = 14$ different taxa, of which 13 are vegetation taxa, together with an 'Other' category. Among the 13 taxa, there is also one composite taxon Artemisia plus Chenopodiaceae. This we denote by 'Art.+Chen'. This yields a vector y_i of counts, with elements y_{ik} for $k = 1, \cdots, m$.

In Whiley et al. (2004), for each sample, modern or fossil, $y_i \mid \mathbf{p}_i, n_i \sim Multinomial(n_i, \mathbf{p}_i)$ independently; here $\mathbf{p}_i$ denotes the underlying composition of the pollen assemblage in the sediment sample and n_i the total count. The elements p_{ik} refer to the k^{th} taxon at the i^{th} site.

The Dirichlet mixture of multinomials provides a natural and convenient model for such variation, being conjugate to the multinomial. Note that the conjugacy allows us to integrate out the $\mathbf{p}_i$; in Chapter 6 we did not avail of this step since the $\mathbf{p}_i$ seemed to improve mixing behaviour of the regular MCMC. However, in this case, retaining the above variables would make computation burdensome in the extreme, even for a single run of regular MCMC, with a particular case omitted. Hence we were forced to integrate out $\mathbf{p}_i$; the marginal distribution of y_i is then the compound multinomial distribution (see Dey and Matai (2002)), given by

$$\pi(y_i \mid \gamma_i, \delta, n_i) = \frac{n_i!\Gamma(\delta)}{\Gamma(n_i + \delta)} \prod_{k=1}^{m} \left(\frac{\Gamma(y_{ik} + \delta\gamma_{ik})}{\Gamma(\delta\gamma_{ik}) y_{ik}!} \right) \tag{7.1}$$

Under this parameterisation, $E(p_{ik}|\gamma_{ik}, \delta, n_i) = \gamma_{ik}$. The δ parameter has a simple interpretation as controlling 'extra-multinomial' dispersion. Note that in this formulation $\sum_k \gamma_{ik} = 1$. For computational reasons the constraint $\gamma_{ik} \geq 10^{-10}$ has also been imposed. Whiley et al. (2004) model the γ_{ik} as functions in a real two-dimensional climate space as $\gamma(x_i, \phi_k)$ to relate the climate x to the propensity of the pollen from the k^{th} taxon to occur in lake sediments. The two dimensional climate space used in this case is shown in Figure 7.1. The climate space consists of 740 points, each associated with a 13 dimensional random variable. Thus, there are $740 \times 13 = 9620$ parameters associated with the climate space. Hence, these, along with δ and the two (assumed) random variables GDD5 and MTCO make a total of

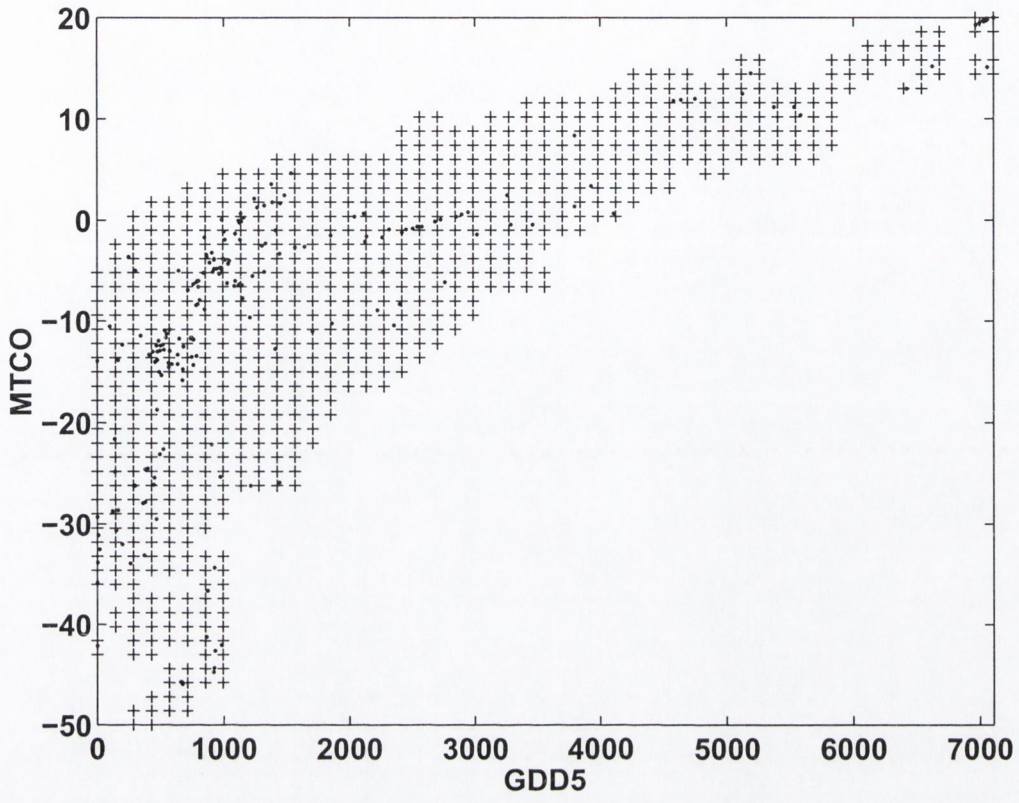


Figure 7.1: Aberrant points with respect to simultaneous credible regions are shown as black dots; plus signs indicate those lattice points which form the support of the response surfaces.

9623 parameters. For explicit modelling details see Whiley et al. (2004).

In order to assess the validity of the above palaeoclimate model, Whiley et al. (2004) removed 61 sites from the training data and constructed estimates for the climate at these sites. The 61 sites chosen were from lakes in Spain (47 sites), Italy (10 sites), Scotland (3 sites) and Norway (1 site). For these sites, the original pollen count data were available. Also, the samples were analysed by palaeoclimatologists, thus already providing some information. However, the subset was small and quite biased towards the regions from which the samples were selected. Apart from this, the procedure of Whiley et al. (2004) is likely to be much more biased than a proper 61-fold cross-validation; see Chapter 2 for a discussion on k-fold cross-validation. In the next section we describe our full leave-one-out cross-validation using

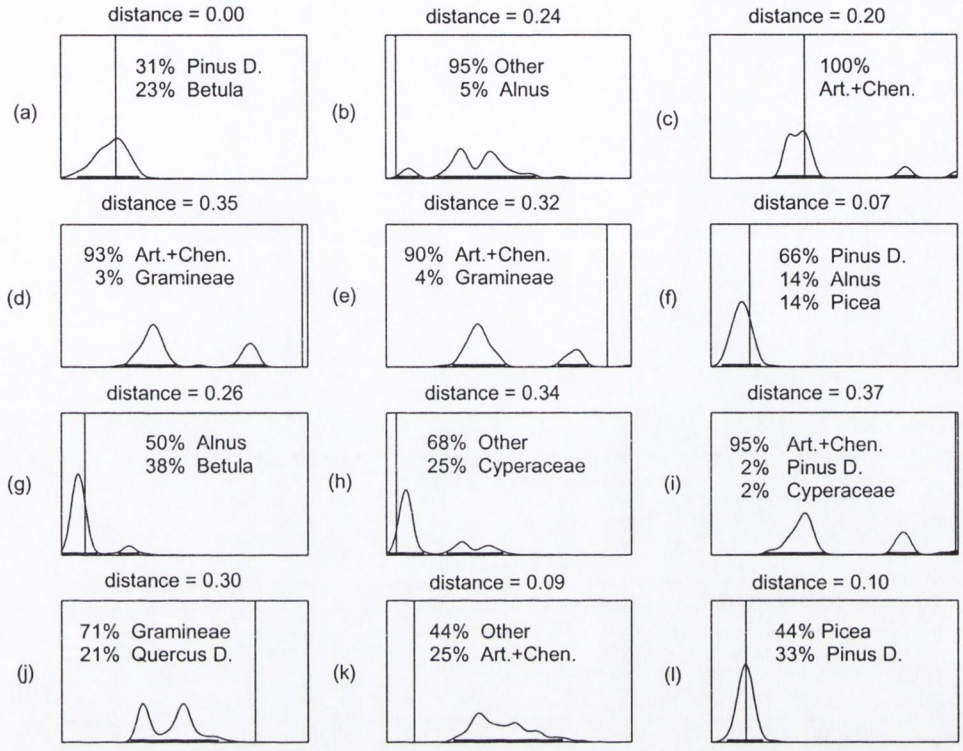


Figure 7.2: Densities of GDD5. Thick black lines denote 95% HPD credible regions; vertical lines denote observed value. Distance d_1 scaled to lie between 0 and 1 and percentages of dominant species in each case are also shown

IRMCMC.

7.2 Cross-validation by IRMCMC

The key first step in IRMCMC is to choose one case, i^*, from the 7815 cases and to study $\pi(x, \theta | X_{-i^*}, Y)$ via MCMC, storing the realisations of θ as well as the climate variables. Here we choose $i^* = 5354$; see discussion below. Figure 7.2 (a) and 7.3 (a) present the leave-one-out posterior distributions of GDD5_{5354} and MTCO_{5354} respectively, given (X_{-5354}, Y). Superimposed are the reported value of GDD5_{5354} and MTCO_{5354}. The second step is to perform variations on this for each of the remaining 7814 cases, by *re-utilising* the stored

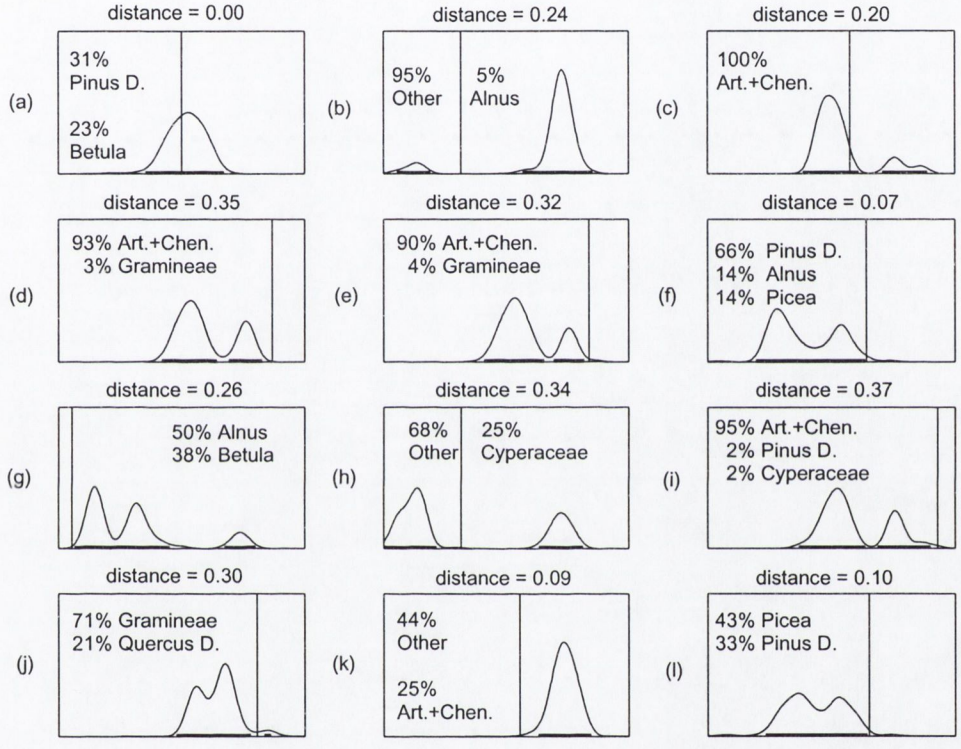


Figure 7.3: Densities of MTCO. Thick black lines denote 95% credible regions; vertical lines denote observed value. Distance d_1 scaled to lie between 0 and 1 and percentages of dominant species in each case are also shown

values of θ; see details below. Other panels in Figures 7.2 and 7.3 present further examples comparing the leave-one-out posterior distributions with the observed values of GDD5 and MTCO. Note that several of the posterior distributions are multi-modal. For these, the 95% highest posterior density regions can comprise disjoint intervals; see Figures 7.2 and 7.3.

Using an algorithm given by Besag et al. (1995), we also compute simultaneous credible regions for GDD5 and MTCO for each of the 7815 sites; however they are not HPD credible regions. It has been found that 4027(51.5%), and 7188(92.0%) points fall in the 50% and 95% simultaneous credible regions respectively. The blackened points in Figure 7.1 are the points that lie outside the respective 95% simultaneous credible regions. Although a small percentage of observed values lie outside the 95% credible regions, a considerably high percentage of observed values lie outside the 50% credible region. Table 7.2 shows the number (and the percentages) of the observed climate values falling within the individual HPD regions. The conclusion drawn from this table does not differ from the one drawn from the simultaneous credible regions, thereby raising doubts about the reliability of the model. We discuss, in the next section, the modelling implications of these findings.

The second step involves, for each case, (a) the use of IR to choose a number of values ($M = 100$) of θ and (b) for, each of these, the use of MCMC with restarts to explore the leave-one-out posterior distribution $\pi(x|X_{-i}, Y, \theta)$ for *fixed* θ. As the dimensionality of x is only two, each MCMC stage is very fast. For each realised θ, we store an MCMC sample of size $K = 10$. The IR stage is conducted to ensure that the simulation is indeed conditional on (X_{-i}, Y) despite being based on θ values conditional on (X_{-5354}, Y). The weights used in the IR stage, which follow from (7.1) are

$$\prod_{j=1}^{14} \frac{\Gamma(y_{i^*j} + \delta\gamma(x_{i^*}, \phi_j))}{\Gamma(y_{ij} + \delta\gamma(x_i, \phi_j))} \frac{\Gamma(y_{ij} + \delta\gamma(x, \phi_j))}{\Gamma(y_{i^*j} + \delta\gamma(x, \phi_j))} \times \frac{\Gamma(\delta\gamma(x, \phi_j))}{\Gamma(\delta\gamma(x, \phi_j))} \frac{\Gamma(\delta\gamma(x_i, \phi_j))}{\Gamma(\delta\gamma(x_{i^*}, \phi_j))} \tag{7.2}$$

It is clear from (7.2) that the importance weights are constant if $x_{i^*} = x_i$ and $y_{i^*} = y_i$. To ensure this as best as possible we recall the distance measures in Chapter 4, minimisation of

which offers different versions of centrality. For this problem, the measures are given by:

$$\text{(a) } d_1(\ell) = \sum_{i=1}^{7815}\left(\sum_{j=1}^{2}\frac{|x_{ij}-x_{\ell j}|}{S(x_{.j})}+\sum_{j=1}^{14}\frac{|y_{ij}-y_{\ell j}|}{S(y_{.j})}\right) \tag{7.3}$$

$$\text{(b) } d_2(\ell) = \sqrt{\sum_{i=1}^{7815}\left(\sum_{j=1}^{2}\frac{(x_{ij}-x_{\ell j})^2}{S^2(x_{.j})}+\sum_{j=1}^{14}\frac{(y_{ij}-y_{\ell j})^2}{S^2(y_{.j})}\right)}$$

In the above, $S(t_{.j})$ denotes sample standard deviation of the j^{th} column of the variable t. We have, then, $i^* = \arg\min\{d_k(\ell); 1 \leq \ell \leq 7815\}$ Minimisation of d_1 with respect to yields $i^* = 5354$ but when d_2 is minimised, values 5968, 7786 and 7787 are obtained. However, distributions of $w_{i,5354}(x,\theta)$ for each of $i = 5968, 7786, 7787$ were approximately the same (not shown) which suggests that it is immaterial which of the 4 sites obtained should be left out. For our purpose, we arbitrarily select $i^* = 5354$. It has been observed in Chapter 1 that IRMCMC may not perform excellently for cases having large distances. However, in this case, a close agreement between regular MCMC and IRMCMC even for the case having the maximum distance (site 82; see Figure 7.4) shows that IRMCMC is quite reliable. Cumulative normalised weights corresponding to sites having maximum distances are shown in Figure 7.5. They seem reasonably close to uniformity. As in Chapter 6, IR without replacement performs quite robustly even for this much larger and complicated problem. Also, the final results were not sensitive to different choices of N, M and K.

It is important to ensure the performance of the regular MCMC for site i^* as best as possible. The values of the response surfaces at the lattice points and δ were updated using a random walk proposals with nearly optimum step size. The optimisation was done using information gained from an initial run. GDD5 and MTCO were updated by randomly proposing from a random walk and an independence sampler. This allows efficient exploration of the posterior in question. Although site $i^* = 5354$ does not involve multimodality (see the first panels of Figures 7.2 and 7.3), we point out that in case multimodality is the solution, then our optimisation method does lead to its efficient exploration, as in the case of site 82; see Figure 7.4.

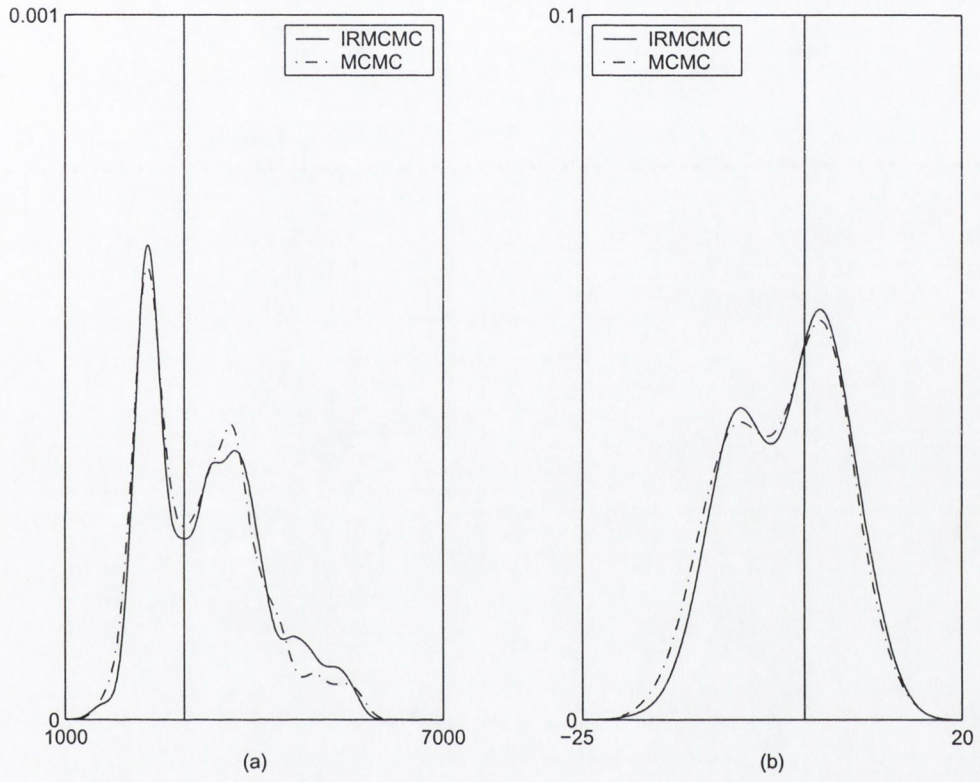


Figure 7.4: Comparison of regular MCMC and IRMCMC for the site 82, the site with maximum distance; (a) corresponds to GDD5 and (b) corresponds to MTCO. Kernel density estimates of GDD5 and MTCO based on both MCMC and IRMCMC are presented. Vertical lines denote observed value.

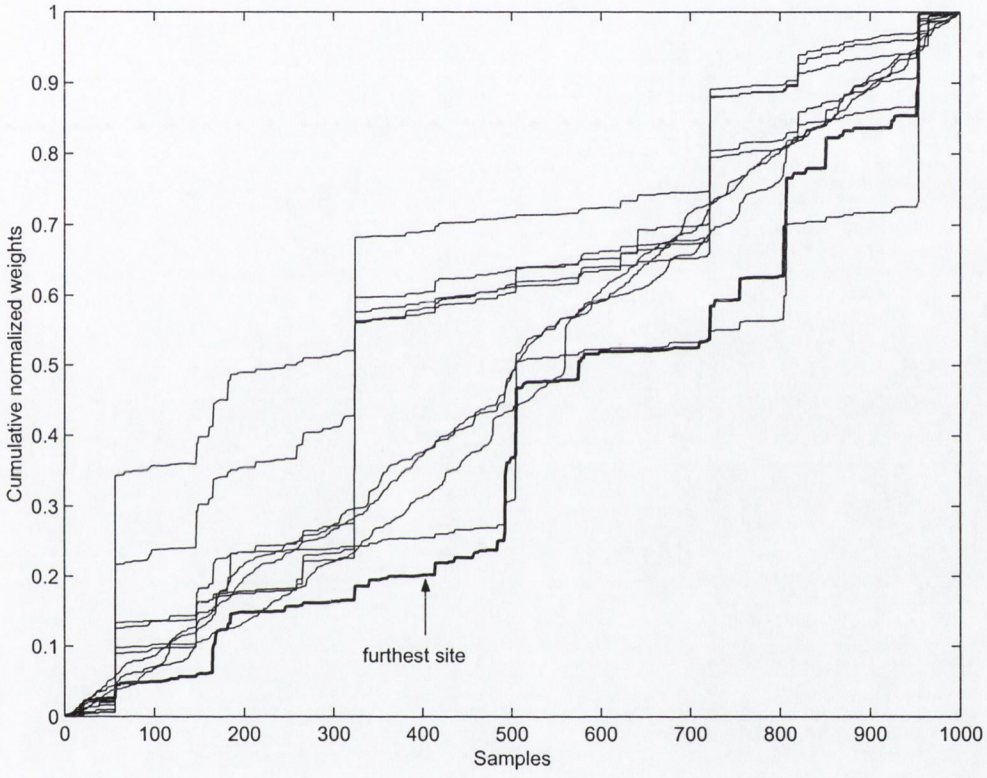


Figure 7.5: Cumulative normalised importance weights corresponding to sites having highest distance values. The thick line corresponds to site 82, the site with maximum distance.

Table 7.1: Simultaneous credible regions of GDD5 and MTCO

50% CI	95% CI
GDD5	MTCO
4027(51.5%)	7188(92.0%)

Table 7.2: Individual highest posterior density regions of GDD5 and MTCO

50% CI		95% CI	
GDD5	MTCO	GDD5	MTCO
4757(60.9%)	4908(62.8%)	7474(95.6%)	7587(97.1%)

7.3 Results of cross-validation

Tables 7.1 show the number (and the percentages) of observed GDD5 and MTCO falling within 50% and 95% simultaneous credible regions. On the other hand, Table 7.2 shows results corresponding to the individual highest posterior density regions of GDD5 and MTCO. Observe that although very few are aberrant with respect to 95% credible regions, a large number of data points are aberrant with respect to 50% credible regions. This indicates that a large number of observations do not fall within the main modal region showing impreciseness of the predictions. The Bayesian hypothesis test described in Chapter 3 is applied individually to GDD5 and MTCO. Distributions of discrepancy measures D_1^{var} along with observed D_1^{obs} corresponding to both GDD5 and MTCO are shown in Figure 7.6.

Both observed discrepancy measures fail to lie within the high density regions of the respective sampling distributions, implying a clear rejection of the model. Obviously, any measure constructed using the joint density of the two climate variables will also fail in this case. It is noteworthy that the observed values are much smaller than the minimum value that has significant density, suggesting that the predicted climates (modes of the posterior densities) and the observed climates agree well with each other but large credible regions

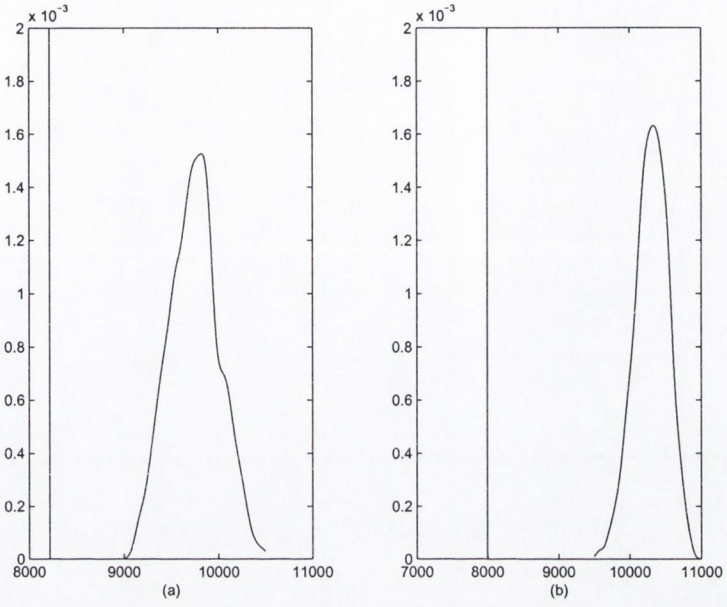


Figure 7.6: (a) Distribution of D_1^{var}, corresponding to GDD5. (b) Distribution of D_1^{var}, corresponding to MTCO. In each case, observed value D_1^{obs} is denoted by the vertical line.

(indicating high uncertainty) are responsible for such poor fit. This is mainly due to the multimodal nature of the densities. However, it is useful to investigate causes of such poor fit, which we do below.

Figures 7.2 and 7.3 display some densities of GDD5 and MTCO with corresponding observed values overlaid, shown by vertical lines. Also shown are the percentages of the most abundant species (dominant species) in each case and the distance measured by the function d_1 in (7.3). Panel (a) of both figures correspond to the minimum distance (see Section 7.2). To be noted is that observed values in this case as well as in the case of maximum distance (see Figure 7.4) are well predicted.

Apart from the minimum and maximum distance cases, most other cases shown in the figures are clearly aberrant. We find that for these the composition is often dominated by a few taxa. The percentages corresponding to the dominant taxa are shown in each panel. It is also to be noted that many posterior distributions are multi-modal. Strongly bimodal posteriors can arise in at least two ways. One is when the record is dominated by two taxa,

say A and B, each of which has a strong preference for mutually exclusive regions of climate space. Such records are rare, but when they arise, they send strong and conflicting signals for these two regions. Another is where the signal is dominated by A but B is missing, despite the fact that B would normally be encountered, together with A, in a sub-region of the climate space preferred by A. Such a record is also rare, but also expresses strong and conflicting signals, manifest in dis-allowing a portion of the space signalled by A. There are however different issues, arising in particular from the construction of the composite 'Other' and the 'Art.+Chen.' categories. The 13 taxa including the composite ones were chosen to be particularly suitable for Glendalough, for they dominate in the region of climate space known to have been traversed by the Glendalough climate over the study period. But some of the 7815 records are dominated by 'Art.+Chen.' and 'Other'. An extreme example arises in panel (c) where 'Art.+Chen.' is the only taxon present; another is seen in panel (h) where 68% of the record comprises 'Other'. The signal here is probably meaningless.

In this connection, we recall from Chapter 1 that the reconstruction of MTCO corresponding to the Younger Dryas event was inaccurate; see Figure 1.1. Observe that the scenario there seems very much in keeping with the above discussion of conflicting signals. We also observe in Figure 7.3 that quite a few aberrant observed MTCO values lie at the extreme tail region of the corresponding leave-one-out density. In such cases, the restriction of climate to the lattice seems to have ruled out some plausible *a posteriori* guesses of climate values.

Chapter 8

Application of IRMCMC to sensitivity analysis in inverse problems

So far we have explained the use of IRMCMC in cross-validation problems. In this chapter we investigate its use in sensitivity analysis in the case of inverse problems. To our knowledge, other than Whiley et al. (2004) there exists no work that addresses this 'inverse' sensitivity analysis. But Whiley et al. (2004) address this problem by regular MCMC which is computationally burdensome in the extreme and is also very unlikely to explore multimodal solutions adequately. In this chapter we propose a very efficient methodology, which actually is a version of IRMCMC, that responds to both these challenges positively.

8.1 Sensitivity analysis in forward problems

Suppose that data $X^m = \{x_i; i = 1, \cdots, n\}$ and $Y^m = \{y_i; i = 1, \cdots, n\}$ are available; it is assumed that $y_i \sim P(\theta x_i)$, as in Chapter 2. Recall that the problem is forward if interest lies in the prediction of y and not x. In such a case, given the model, only a prior on θ needs to be specified. In the next section it will be seen that in the case of inverse problems, where

interest lies in the prediction of x and not y, a prior needs to be specified on x as well. In this section however, we will focus on forward problems where there is only one unknown parameter, θ, to which a prior must be assigned.

Possible mis-specification of the prior distribution on θ is usually investigated by sensitivity analysis. This has been an important element of the philosophies of a number of Bayesians (see, for example, Berger (1985) and the references therein). Loosely, sensitivity analysis involves trying different reasonable priors and scrutinising the resultant posterior quantities. Re-thinking is necessary if, due to different prior assumptions, the posterior quantities are changed in a way that has practical impact on interpretations or decisions.

However, recomputing posterior quantities for all reasonable priors could be computationally extremely expensive, particularly for high-dimensional problems. Importance sampling is generally recommended for sensitivity analysis (see, for example, Athreya et al. (1996), Doss (1994)). For a prior π_0 on the parameter θ, samples are generated from the posterior $\pi_0(\theta \mid X^m, Y^m)$, usually by regular MCMC. Then for another prior of interest, π_1, importance weights of posteriors $\pi_1(\theta \mid X^m, Y^m)$ are computed with respect to the 'initial posterior' $\pi_0(\theta \mid X^m, Y^m)$. This is given by

$$w_1(\theta) = \pi_1(\theta \mid X^m, Y^m)/\pi_0(\theta \mid X^m, Y^m) \propto \pi_1(\theta)/\pi_0(\theta) \tag{8.1}$$

These weights are then used to compute approximations to posterior quantities of $\pi_i(\theta \mid X^m, Y^m)$, for example, expectation of an appropriate function $h(\theta)$ as in (2.6).

The above technique of sensitivity analysis is appropriate to forward problems. However, such technique is inapplicable in the case of inverse problems, as we demonstrate below.

8.2 Sensitivity analysis in inverse problems

In addition to the set up described in the above Poisson regression problem, suppose that a further set of observations Y^f are available from the model but the corresponding X^f are unavailable. The interest is to learn about the set of unknown values, X^f; θ is treated as

a nuisance parameter (see, for example, Berger et al. (1999)). Here we assume that an appropriate prior, given by $\pi(\theta)$ is assigned to θ.

Our interest in this case is to check sensitivity of the priors on X^f and not on θ. The priors on X^f may or may not depend on θ. A convenient way to proceed is to propose priors $\{\pi_0(\cdot \mid \theta),\ \pi_i(\cdot \mid \theta); i = 1, \cdots, k\}$ on X^f, where $\pi_0(\cdot \mid \theta)$ is the prior of main interest and $\{\pi_i(\cdot \mid \theta); i = 1, \cdots, k\}$ are considered variations of the former. The interest is then to check sensitivity with respect to the posterior $\pi_i(X^f \mid X^m, Y^m, Y^f) = \int \pi_i(X^f, \theta \mid X^m, Y^m, Y^f) d\theta$. However, since the functional form of the posteriors involves integrating out the nuisance parameter θ (which may be high-dimensional), this may not be available, even up to a constant. Hence importance weights, given by

$$\begin{aligned} w_i(X^f) &= \pi_i(X^f \mid X^m, Y^m, Y^f) / \pi_0(X^f \mid X^m, Y^m, Y^f) \\ &= \frac{\int \pi_i(X^f, \theta \mid X^m, Y^m, Y^f) d\theta}{\int \pi_0(X^f, \theta \mid X^m, Y^m, Y^f) d\theta} \end{aligned}$$

will not be available here, unlike in the forward case, which was given by (8.1).

In principle, it is possible to apply IRMCMC to such inverse problems. In other words, one can realise from $\pi_0(X^f, \theta \mid X^m, Y^m, Y^f)$, typically by regular MCMC, a sample of (X^f, θ). Then using importance weights of the form given by

$$\begin{aligned} w_{0,i}(X^f, \theta) &= \frac{\pi_i(X^f, \theta \mid X^m, Y^m, Y^f)}{\pi_0(X^f, \theta \mid X^m, Y^m, Y^f)} \\ &\propto \frac{\pi_i(X^f \mid \theta)}{\pi_0(X^f \mid \theta)} \end{aligned} \tag{8.2}$$

one can realise a subsample of θ realisations corresponding to the posterior $\pi_i(\theta \mid X^m, Y^m, Y^f)$. Given each realised θ, realisations of X^f can be obtained from $\pi_i(X^f \mid Y^f, \theta)$, typically by regular MCMC. The realised X^f can then be said to be samples from the target posterior $\pi_i(X^f \mid X^m, Y^m, Y^f)$. This method can be repeated for each $i = 1, \cdots, k$. Note that this is exactly the IRMCMC proposal as described in Chapter 4.

Thus IRMCMC may be straightforwardly applied to inverse problems exactly as in cross-validation. There is no issue now as to the choice of importance sampling distribution: it is

the posterior $\pi_0(X^f, \theta \mid X^m, Y^m, Y^f)$ computed with reference to the prior $\pi_0(\cdot \mid \theta)$. Nothing more needs to be said.

There is however a variation, which is the subject of this chapter. For it is not necessarily the case that the priors $\pi_0(X^f \mid \theta)$ and $\pi_i(X^f \mid \theta)$ are of the same dimension. In Chapter 7 we provided cross-validation details of the palaeoclimate problem of Whiley et al. (2004). In Section 8.4 we describe sensitivity analysis of the problem and point out that priors $\pi_0(X^f \mid \theta)$ and $\pi_i(X^f \mid \theta)$ have different dimensions for a particular i. It is thus not possible to compute weights as in (8.2). We propose an approximation, discussed in the following section. This approximation is particularly appropriate for large data sets, as here.

8.3 Proposal

Our proposal is to generate a sample of θ realisations from some distribution of θ that adequately approximates all the target posteriors of θ given by $\pi_i(\theta \mid X^m, Y^m, Y^f) = \int \pi_i(X^f, \theta \mid X^m, Y^m, Y^f) dX^f$. We can then resample from the available sample of θ-realisations simply by random sampling, *without computing any importance weights.* We thus consider a special case of IRMCMC where the weights are equal. This solves the dimensionality problem since we no longer need to compute importance weights. Here are the details.

Note that, for any $i = 0, 1, \cdots, k$,

$$\begin{aligned} \pi_i(x \mid X^m, Y^m, Y^f) &= \int \pi_i(x, \theta \mid X^m, Y^m, Y^f) d\theta \\ &= \int \pi_i(x, \mid Y^f \theta) \pi_i(\theta \mid X^m, Y^m, Y^f) d\theta \qquad (8.3) \\ &\approx \int \pi_i(x_i \mid Y^f, \theta) \pi(\theta \mid X^m, Y^m) d\theta \qquad (8.4) \end{aligned}$$

In (8.3), $\pi_i(\theta \mid X^m, Y^m, Y^f) = \int \pi_i(X^f, \theta \mid X^m, Y^m, Y^f) dX^f$ Note that, in (8.4), we assumed that the approximation $\pi(\theta \mid X^m, Y^m, Y^f) \approx \pi(\theta \mid X^m, Y^m)$ holds, the latter being the saturated posterior. For large data size, this is indeed a valid assumption, as shown by the theoretical exposition in Chapter 5. We first generate and store a sufficiently large sample of θ from the saturated posterior $\pi_i(\theta \mid X^m, Y^m)$, typically by regular MCMC; we

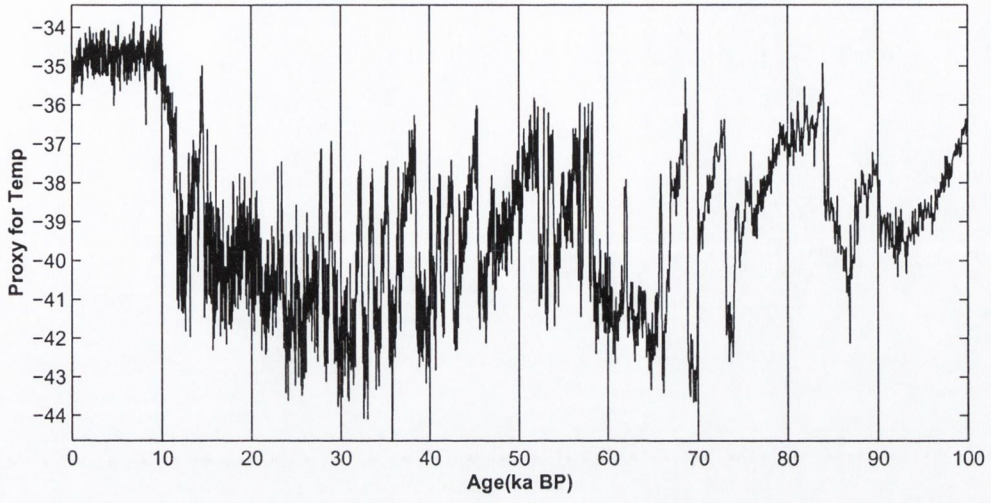


Figure 8.1: Greenland ice core data.

subsequently generate realisations of X^f by (a) drawing θ randomly from the store, and (b) generating realisations from $\pi(X^f \mid Y^f, \theta)$. In our application X^f has dimension 300; it is still substantial but it is much less than the dimension of (X^f, θ).

Observe that our procedure completely solves any integrability issues by doing random sampling from a sample generated from the saturated posterior. This procedure does not rely on importance weights, hence problems faced by traditional procedures do not appear here. Also, as demonstrated in Chapter 4, regular MCMC needed to simulate from $\pi_i(x \mid Y^f, \theta)$ may be started with different initial values which may help explore multimodal solution more efficiently. We demonstrate these advantages in the case of the palaeoclimate model of Whiley et al. (2004).

Before discussing the application of our proposed method to the case of Whiley et al. (2004) we review temporal smoothness issues described in their paper.

8.4 Temporal smoothness issues in the palaeoclimate reconstruction model

In Whiley et al. (2004) the modern climate values denoted by X^m are known but the prehistoric climates X^f are not. Climate change exhibits some degree of smoothness in time. Loosely speaking, climate changes can be characterised as 'small', mostly, but occasionally very large. Whiley et al. (2004) model this smoothness stochastically by specifying an appropriate family of priors. Light can be shed upon this by an examination of the ice core data (Fig 8.1). This has been obtained from the stable oxygen isotope $\delta^{18}O$ data from the ice of the GISP2 core drilled near the summit of the Greenland ice sheet (see, for example Stuiver et al. (1995)). This provides a basis for estimating the temporal properties of the climate system since the end of the last ice age. The data principally reflect atmospheric temperature over Greenland; their temporal characteristics can thus be argued to be relevant to the reconstruction of temperature in Glendalough in Ireland, although they represent a locality some distance from Ireland. In this section the bivariate terms x_j^f may be more readily denoted as $x(t_j)$, where the t_j are in ^{14}C years before present (BP). Here $j = 1$ denotes the deepest (oldest) sample and $j = 150$ the most recent. In the following we suppress the superscript ^{f}.

The variation in Figure 8.1 can be adequately modelled by a random walk. But the normal scores plot of the increments strongly suggests that the variation is much longer tailed than the normal distribution. Whiley et al. (2004) investigate the appropriateness of several different priors for the dependence between the unknown $x(t_j)$. However, they assume that each is of the form $x(t_j) = x(t_{j-1}) + A(t_j)\varepsilon(t_j)$, where $A_j = A(t_j)$ is a matrix; recall that climate has two dimensions in this case. The distributions of the innovations $(\varepsilon_{j1}, \varepsilon_{j2})$ are what distinguish different priors. They investigate the sensitivity of the posterior distribution of reconstructed climates considering the distribution of the innovations to be Student's t distribution with $d = 8$ degrees of freedom, the Cauchy distribution and the Gaussian

distribution. They also consider a prior with complete temporal independence, that is the prior here is simply uniform on the climate space shown in Figure 7.1 with climates at different times being independent of each other. In all cases, GDD5 and MTCO are assumed to be distributed independently of each other. It is further assumed that $A_j = |t_j - t_{j-1}|\Delta$ where Δ is diagonal with $(\Delta_{GDD5}, \Delta_{MTCO})$ being the diagonal elements. Thus we may look upon Δ as a two-dimensional random variable.

Thus, for priors with temporal smoothness,

$$\pi^f(x|\Delta) = \pi^f_{GDD5}(x|\Delta)\pi^f_{MTCO}(x|\Delta) \tag{8.5}$$

Specifically, using the random walk model it can be written as

$$\pi_{GDD5}(x|\Delta) = \prod_{j=2,150} g_d(x_{GDD5}(t_j)|\mu_{GDD5}(t_j, t_{j-1}), \sigma_{GDD5}(t_j, t_{j-1})),$$

and similarly for MTCO. By $g_d(u|\mu, \sigma)$ we denote the value of the pdf of a t distribution with d degrees of freedom, evaluated at $\sigma^{-1}(u - \mu)$; we set $\mu = x_{GDD5}(t_{j-1})$ and $\sigma^2 = \Delta^2_{GDD5}(t_j - t_{j-1})^2$. Note that the Cauchy is a t distribution with one degree of freedom and Gaussian is a t distribution with infinite degrees of freedom.

We have demonstrated IRMCMC as a more efficient alternative to regular MCMC, both in terms of computational speed and exploration of the posterior. Using regular MCMC, θ may be simulated from a posterior corresponding to a specific temporal smoothness prior and re-weighed towards posteriors corresponding to the remaining temporal smoothness priors, using appropriate importance weights. Regular MCMC may then be used for the simulation of climate variables and Δ.

However, the idea of reweighing the samples of θ using appropriate importance weights is rendered invalid when, among the possible priors on climate, the prior with no temporal smoothness is also considered. Observe that the prior with complete temporal independence does not consist of the two-dimensional random variable Δ and so has dimension two less than the other priors with temporal dependence. In other words, the joint posterior of the parameters, when no temporal smoothness in climate is assumed, is given by $\pi(x, \theta \mid$

X^m, Y^m, Y^f); but with the above temporal smoothness assumptions, it is given by $\pi(x, \theta, \Delta \mid X^m, Y^m, Y^f)$, which are of different dimensionality. Hence it is not possible to compute importance weights of the temporal smoothness priors with respect to the prior with complete temporal independence; and vice versa.

In the next section, we demonstrate that simple random sampling of the realisations of θ obtained the saturated posterior may be serve as an easily available alternative to the more precise IR, which is unavailable. Note that the former may be regarded as IR with equal weights.

8.5 Sensitivity analysis of the palaeoclimate model using IRMCMC

The posterior of the unknown fossil climates x with temporal smoothness assumed can be written as

$$\begin{aligned} \pi(x \mid X^m, Y^m, Y^f) &= \int \pi(x, \Delta, \theta \mid X^m, Y^m, Y^f) d\Delta d\theta \\ &= \int \pi(x, \Delta \mid \theta, Y^f) \pi(\theta \mid X^m, Y^m, Y^f) d\Delta d\theta \end{aligned} \tag{8.6}$$

In the case where no temporal smoothness is assumed, the posterior of x is given by

$$\begin{aligned} \pi(x \mid X^m, Y^m, Y^f) &= \int \pi(x, \theta \mid X^m, Y^m, Y^f) d\theta \\ &= \int \pi(x \mid \theta, Y^f) \pi(\theta \mid X^m, Y^m, Y^f) d\theta \end{aligned} \tag{8.7}$$

Observe that Δ is present in (8.6) and absent in (8.7). To conduct sensitivity analysis we propose to simulate realisations of θ obtained from the saturated posterior (using regular MCMC). Then, for fixed θ, we simulate (x, Δ) from $\pi(x, \Delta \mid \theta, Y^f)$ in the cases of priors with temporal dependence and x from $\pi(x, \Delta \mid \theta, Y^f)$ in the case of the independence prior. Recall that this is the basic idea of IRMCMC. But since importance weights are unavailable in this case, we rely on asymptotics instead of exact computation using IR, assuming implicitly

that the importance weights are all equal. This may not be an unreasonable assumption since, given the enormous size of the modern training data (X^m, Y^m) (7815 cases), the fossil data Y^f provides no extra information about θ. For large sample sizes this assumption may indeed be valid as demonstrated in the context of leave-one-out cases in Chapter 5. In the next section we compare results obtained by IRMCMC and regular MCMC.

8.6 Results of sensitivity analysis

The real advantage of using IRMCMC is the computational speed. As noted earlier, the total time taken to compute the posteriors corresponding to all four priors using four regular MCMC runs took about 22 hours. But with our method, a single regular MCMC run is needed. Random sampling from the samples of θ and simulating large samples from posteriors given fixed θ can be done at almost no computational cost. This also allows exploration of the posterior much more adequately than a corresponding regular MCMC algorithm during a given time. This has been demonstrated experimentally, with the help of the Poisson regression example, in Chapter 4. In our case, computation of the saturated posterior took about 5 hours 28 minutes and then using and the remaining computation taking just about 12 minutes. The results of sensitivity analysis using IRMCMC and regular MCMC are shown in Figures 8.2 and 8.3 respectively.

In general it seems that IRMCMC and regular MCMC are not in perfect agreement with each other and results with respect to the former seems to exhibit more variability. Although it is difficult to assert which of the two methods give more reliable results it is arguable that due to independence of realisations of θ in the former case, IRMCMC explores the solution space more reliably than regular MCMC and hence results obtained by IRMCMC are more believable. A more prominent evidence is given below.

To ascertain which prior is appropriate for temporal smoothness, Whiley et al. (2004) focus attention on two aspects of rapid climate change, well defined from many other sources of the palaeoclimatology literature. These are the rapid cooling at the start of the Younger

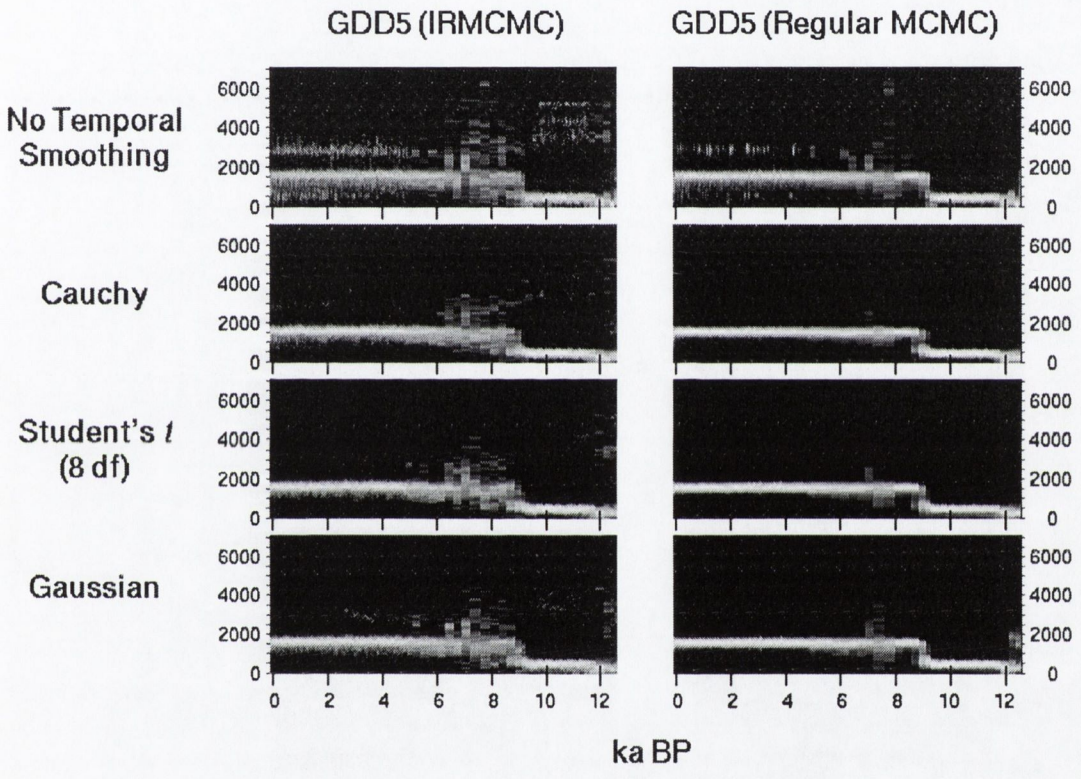


Figure 8.2: Reconstructions of GDD5 using IRMCMC and regular MCMC.

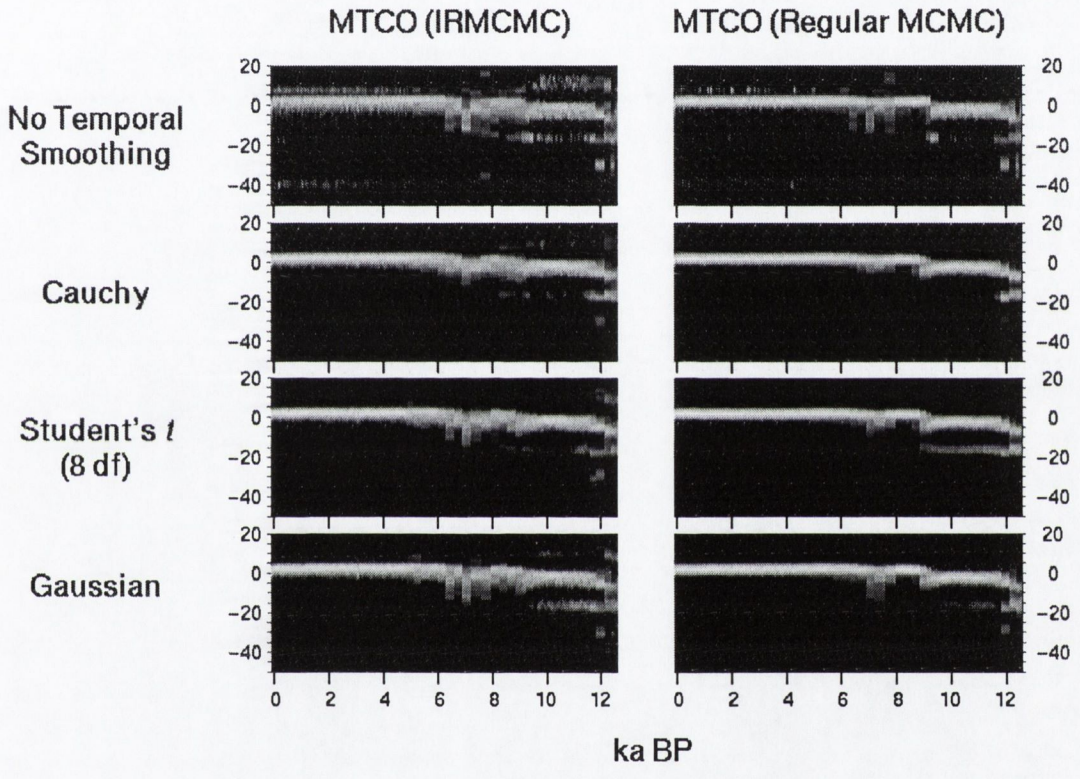


Figure 8.3: Reconstructions of MTCO using IRMCMC and regular MCMC.

Dryas event (ca 10.8 ka BP), an event in the late glacial when the warming into the post glacial was interrupted and rapid warming at the end of this event (10.0 ka BP). Some literature speculates that each of these climate changes occurred within a few decades (Dansgaard et al. (1989)). Although the temporal resolution of the fossil pollen record can not define such detail, Whiley et al. (2004) attempt address this in Figures 8.4 and 8.6 by presenting estimates of P(GDD5>1000 day degrees) and P(MTCO<-4^0), in each case for the relevant period of time. They use regular MCMC to conduct the sensitivity analysis. On the other hand, Figures 8.5 and 8.7 present the same estimates constructed using IRMCMC. Observe that in Figure 8.4 the three different priors for the smoothness of climate change give a very consistent signal; the warming was most probably between 8.75 and 9.25 ka BP and this signal is about 500 years different from that of the independence model. From Figure 8.6 only a weak signal of the Younger Dryas events is seen, that is, of sharp climate cooling about 500 years before rapid warming. It is also seen in Figure 8.4 that all models that involve temporal smoothness arrive at almost the same conclusion.

However, results obtained using IRMCMC as seen in Figure 8.5 seem to give a more consistent signal than that seen in the regular MCMC analysis, since all priors, including the independence model model, are in agreement with each other about the warming period. A much stronger signal about the Younger Dryas event is visible in Figure 8.7.

Thus this chapter demonstrates that IRMCMC is useful not only for cross-validation. Even for sensitivity analysis it can be a valuable tool. We have also suggested tentatively that, compared to regular MCMC, our procedure is capable of more adequate exploration of the posteriors of interest.

In this chapter, because of dimensionality problems, we have used simple random sampling of θ, samples of which have been obtained from the saturated posterior. However, when dimensions of all posteriors are the same, IRMCMC should certainly be used with appropriate weights.

In the next chapter we apply IRMCMC to a forward problem.

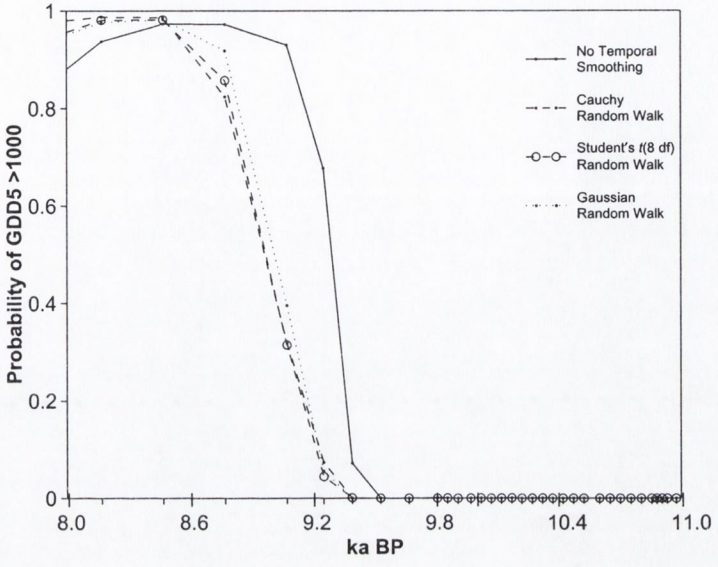


Figure 8.4: Probability that GDD5 is greater than 1000 day degrees (obtained using regular MCMC). Tick marks indicate the approximate radio-carbon dates of the fossil pollen samples.

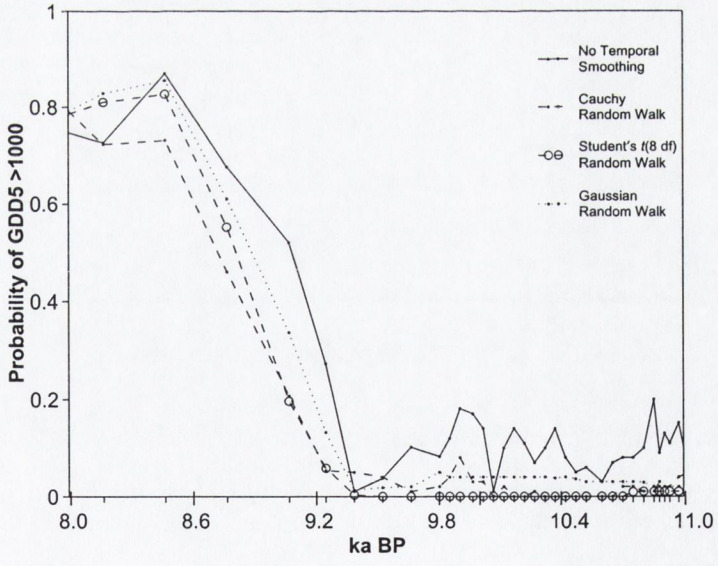


Figure 8.5: Probability that GDD5 is greater than 1000 day degrees (obtained using IRM-CMC). Tick marks indicate the approximate radio-carbon dates of the fossil pollen samples.

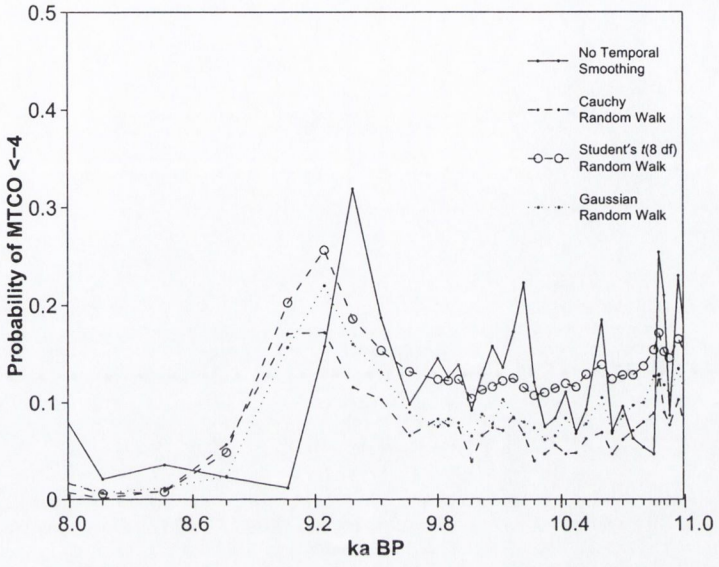


Figure 8.6: Probability that MTCO is less than 4^0 (obtained using regular MCMC). Tick marks indicate the approximate radio-carbon dates of the fossil pollen samples.

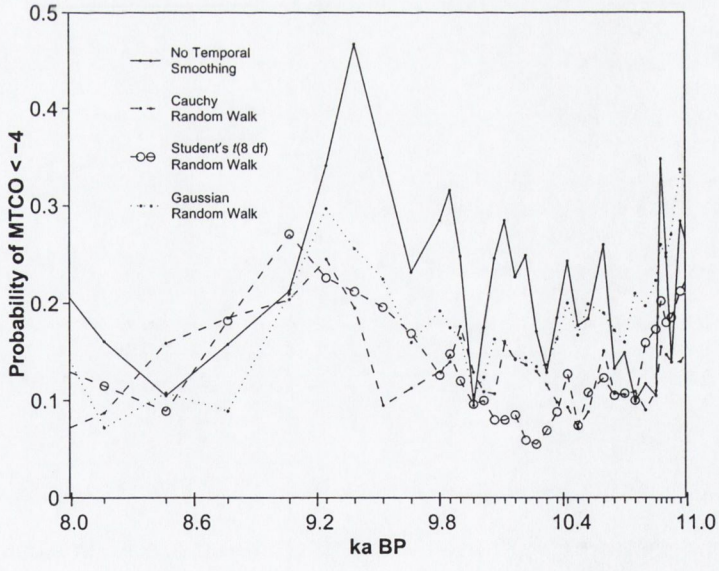


Figure 8.7: Probability that MTCO is less than 4^0 (obtained using IRMCMC). Tick marks indicate the approximate radio-carbon dates of the fossil pollen samples.

Chapter 9

Application of IRMCMC to a forward problem

So far we have considered the applications of IRMCMC to inverse problems only and forward problems have received no attention. It has been explained in Chapter 2 that forward problems are straightforward compared to inverse problems. In this chapter we introduce a problem which is inverse in nature but the cross-validation of which requires forward implementation. We show thhat the simple random sampling from the saturated posterior (that is, IR with equal weights) is sufficiently robust. We further demonstrate that in this forward problem the MCMC requirement in Step 3b(ii) of Chapter 4 can be replaced by direct simulation. The model, which arose in the context of a geostatistical problem, has been the considered by Diggle et al. (1998). The problem they consider is the assessment of residual contamination from nuclear weapons testing on a South Pacific island, in which the sampling method generates spatially indexed Poisson counts conditional on an unobserved spatially varying intensity of radioactivity. We give a fuller account of the problem before proceeding with cross-validation.

9.1 A geostatistical problem

Rongelap Island, which is located in the Pacific Ocean approximately 2500 miles south-west of Hawaii, experienced contamination due to fall-out from a nuclear weapons testing programme that forced the former inhabitants of the island to live in self-imposed exile on a much smaller island of Mejatto.

To investigate whether Rongelap can safely be resettled the current levels of ^{137}Cs contamination at a set of $n = 157$ locations over the island have been examined by in-situ gamma-counting. According to the well-established theory of radioactive emissions, the counts y_i at the $n = 157$ locations can be treated approximately as realisations of mutually independent Poisson random variables with expectations $M_i = t_i\lambda(l_i)$, where t_i denotes the length of time over which the counts are recorded and $\lambda(l)$ measures the ^{137}Cs radioactivity at location l. Diggle et al. (1998) assume that $\log\lambda(l) = \beta + S(l)$, where $S(\cdot)$ is a zero-mean, stationary Gaussian process with variance σ^2 and isotropic correlation function given by

$$\rho(u) = \exp[-(\alpha u)^{\delta}] \tag{9.1}$$

The parameter β represents the mean log-intensity over the island. Anticipating identifiability problem and based on a pilot regular MCMC run we fixed $\delta = 0.5$ for the purpose of this chapter; see also pp323 of Diggle et al. (1998).

One objective of the Rongelap survey is to estimate $\lambda(l)$. Interest is also in non-linear functions of $\lambda(l)$ such as $\max\lambda(l)$, the location associated with this maximum and with the regions of the island where radioactivity is above a specific threshold. Thus the problem is inverse; the interest basically lies in $s_i = S(l_i)$ which can be likened to x_i in the earlier problems. But the main difference of this problem with the earlier ones is that s_i are unobserved unlike x_i. Thus for assessing model fit via cross-validation it is needed to predict y_i (and not s_i since it is unobserved) after leaving out the observation. Hence forward implementation is needed for cross-validation although the problem is inverse in nature.

Diggle et al. (1998) assess the fit of the model through the variogram(for definition, see

Diggle et al. (1998)). By simulating independent replicated spatial samples from the fitted model, each with identical locations to those of the data, they examine the variability of the empirical variogram and construct tolerance intervals for it. However, their method uses the data twice; Bayarri and Berger (1999) demonstrate that using data twice is undesirable, see also Chapter 3. In the next section we discuss the assessment of the model fit via leave-one-out cross-validation using IRMCMC.

9.2 Cross-validation of the geostatistical model using IRMCMC

Denoting $(y_1, \cdots, y_n)$ by Y, $(s_1, \cdots, s_n)$ by S and letting Y_{-i}, S_{-i} stand for all but the ith element of Y and S respectively, we note that for each i it is needed to compute

$$\begin{aligned} \pi(y_i \mid Y_{-i}) &= \int \pi(y_i, \theta, S \mid Y_{-i}) d\theta dS \\ &= \int p(y_i \mid s_i, \beta) \pi(\theta, S \mid Y_{-i}) d\theta dS \qquad (9.2) \\ &= \int p(y_i \mid s_i, \beta) \pi(s_i \mid \theta, S_{-i}) \pi(\theta, S_{-i} \mid Y_{-i}) d\theta dS \qquad (9.3) \end{aligned}$$

In (9.2) and (9.3), $p(\cdot \mid s_i, \beta)$ denotes Poisson probability mass function with parameter depending upon s_i and β.

For the purpose of IR, using methods prescribed in Chapter 4 we leave out the case with minimum distance, $i^* = 121$. To investigate robustness with respect to our proposal we also compare the results with those obtained after leaving out the case with maximum distance, $i^* = 151$ and the saturated posterior, $\pi(\theta, S \mid Y)$ (note that importance weights with respect to the saturated posterior can be computed in this case). All the three posteriors $\pi(\theta, S \mid Y_{-121})$, $\pi(\theta, S \mid Y_{-151})$ and $\pi(\theta, S \mid Y)$ are almost identical, as shown in Figures 9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7 and 9.8. This suggests considerable overall robustness to the choice of reasonable importance sampling densities and in fact simple random sampling would be as effective as importance resampling. Observe that, given s_i and β, MCMC is not needed

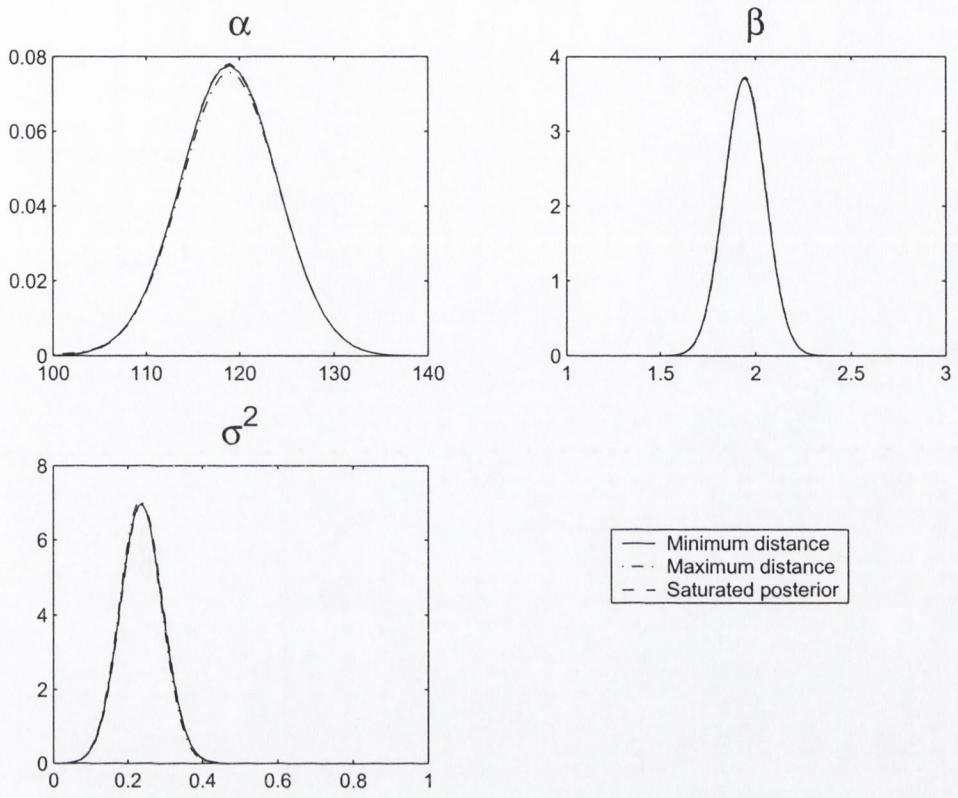


Figure 9.1: Comparison of densities of parameters with respect to the saturated posterior and cases with minimum and maximum distances omitted.

to simulate from $p(\cdot \mid s_i, \beta)$ as direct simulation is possible. Different choices of N, M and K did not play any significant role as in problems dealt with in the earlier chapters. We present results corresponding to $N = 5000$, $M = 50$ and $K = 100$.

It is important to mention that we experienced very poor mixing with the usual sequential updating of the parameters due to very strong posterior correlation between the parameters. Reparameterisation of β and s_i to $\beta^* = \beta + \bar{s}$ and $s_i^* = s_i - \bar{s}$ respectively, where $\bar{s} = \sum_{i=1}^{n} s_i/n$ significantly reduced correlation and hence greatly improved mixing properties of the MCMC runs corresponding to the importance sampling densities. In other words, we transformed $(\alpha, \beta, s_1, \cdots, s_{n-1}, s_n)$ to $(\log(\alpha), \beta^*, s_1^*, \cdots, s_{n-1}^*, s_n^*)$, where $s_n^* = \bar{s}$ and finally β^* and s_i^* were transformed back to the original parameters α, β and s_i. For more on reparameterisation

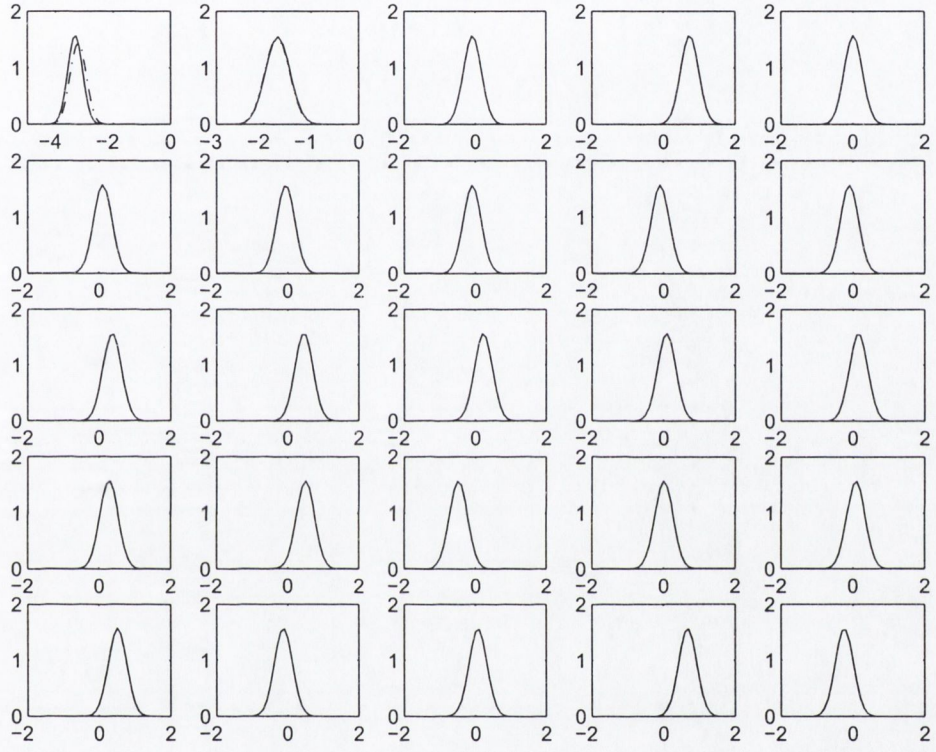


Figure 9.2: Comparison of $s_1 - s_{25}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted. Solid line, dash-dot line and dotted line indicate minimum distance omitted, maximum distance omitted and the saturated posterior respectively.

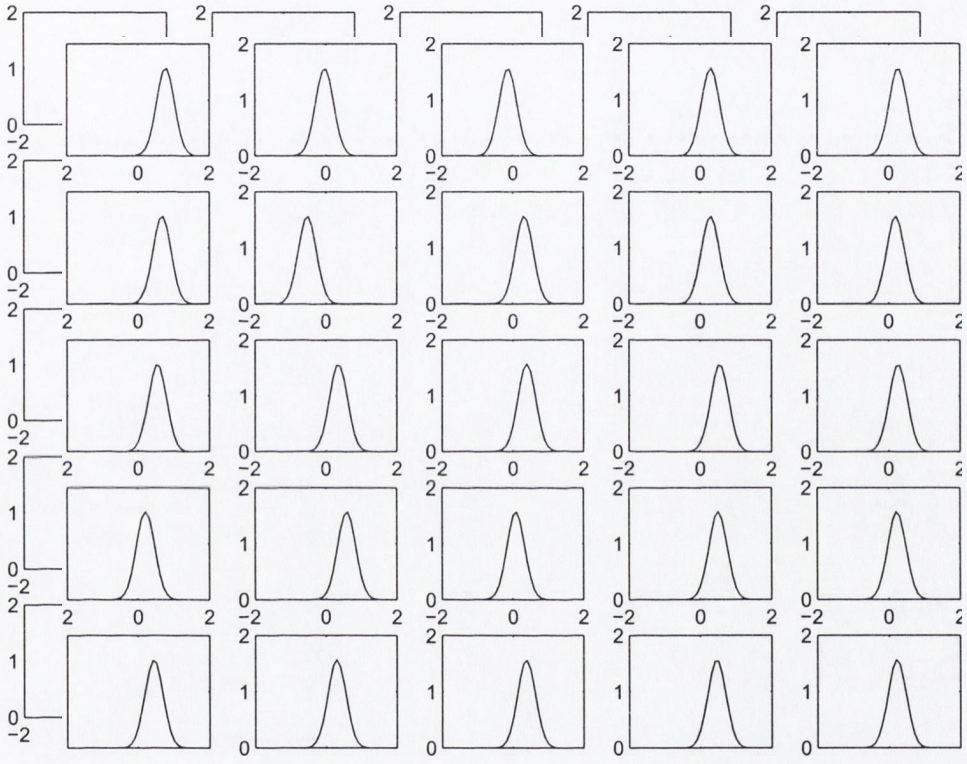


Figure 9.3: Comparison of $s_{26} - s_{50}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted.

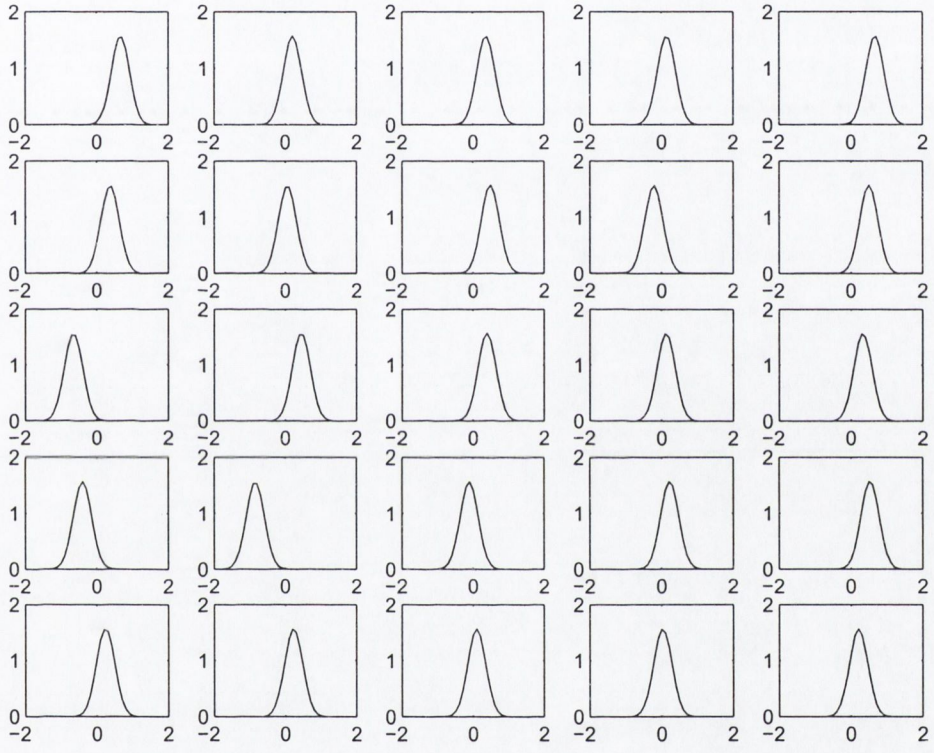


Figure 9.4: Comparison of $s_{51} - s_{75}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted.

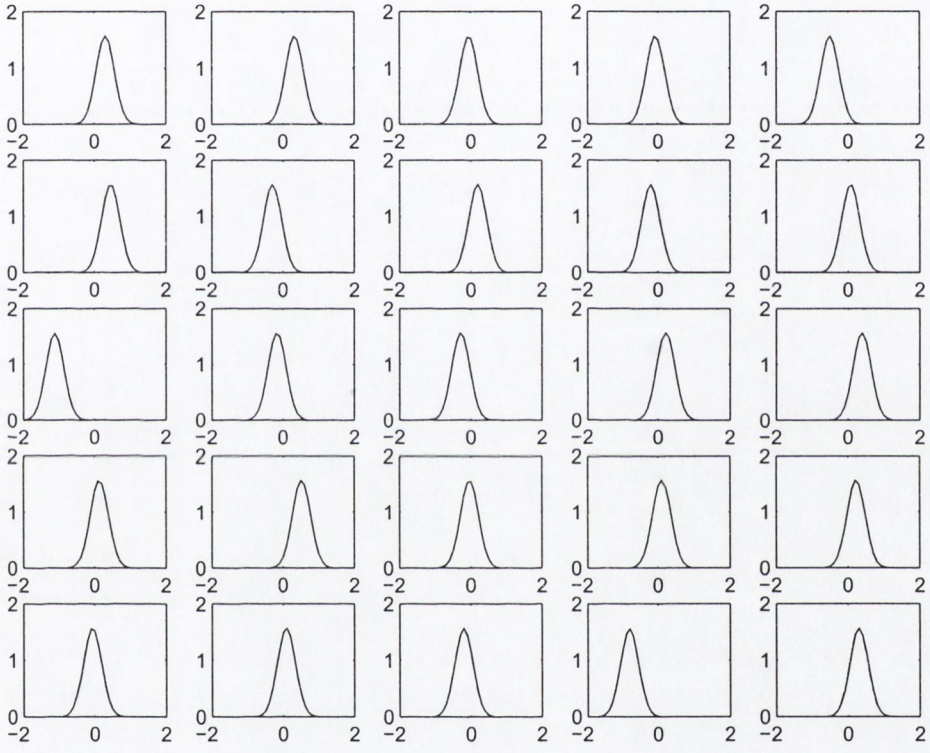


Figure 9.5: Comparison of $s_{76} - s_{100}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted.

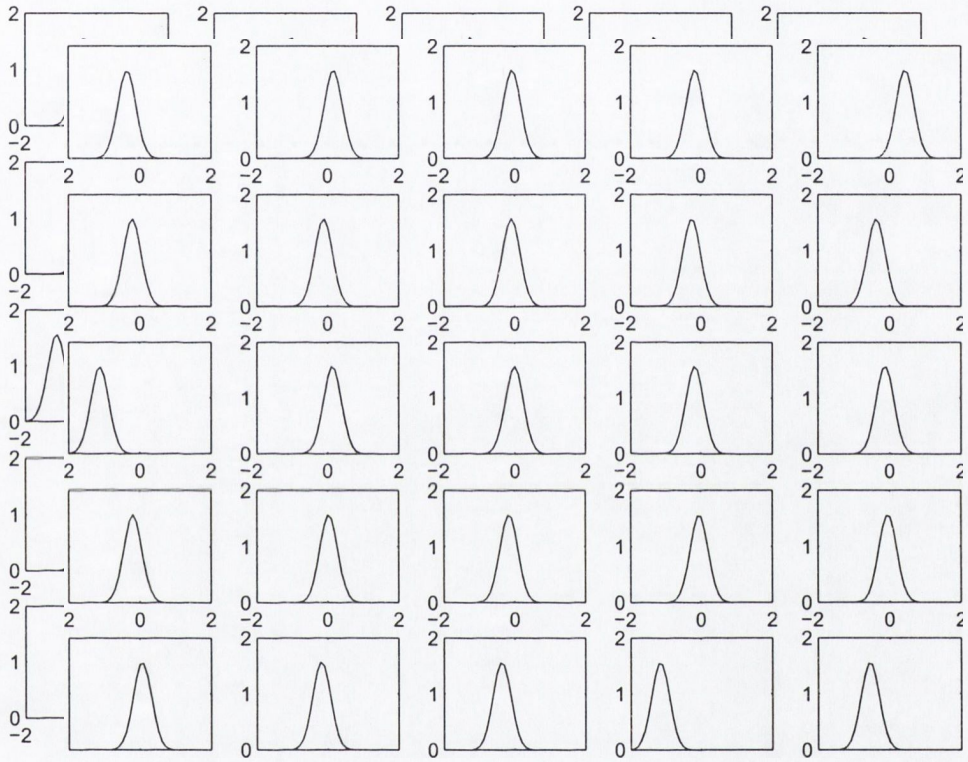


Figure 9.6: Comparison of $s_{101} - s_{125}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted.

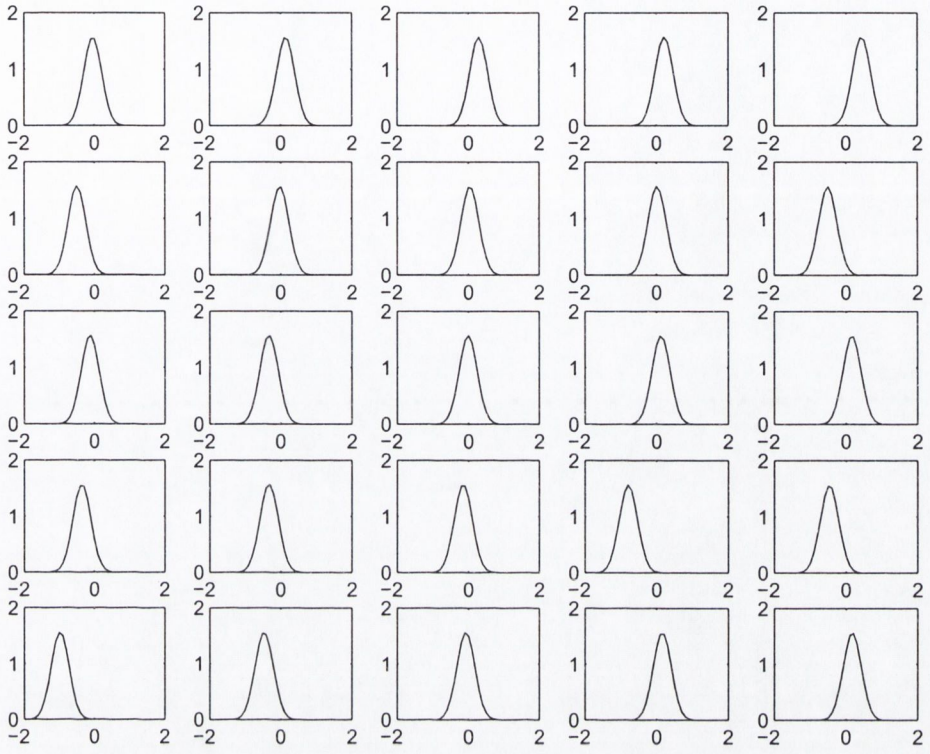


Figure 9.7: Comparison of $s_{126} - s_{150}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted.

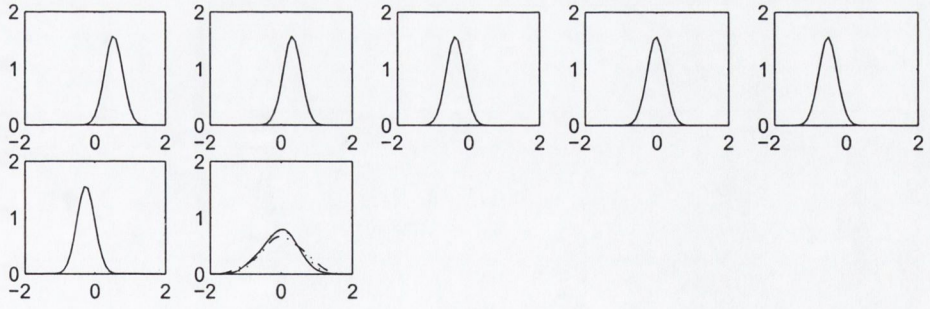


Figure 9.8: Comparison of $s_{150} - s_{157}$ with respect to the saturated posterior and cases with minimum and maximum distances omitted.

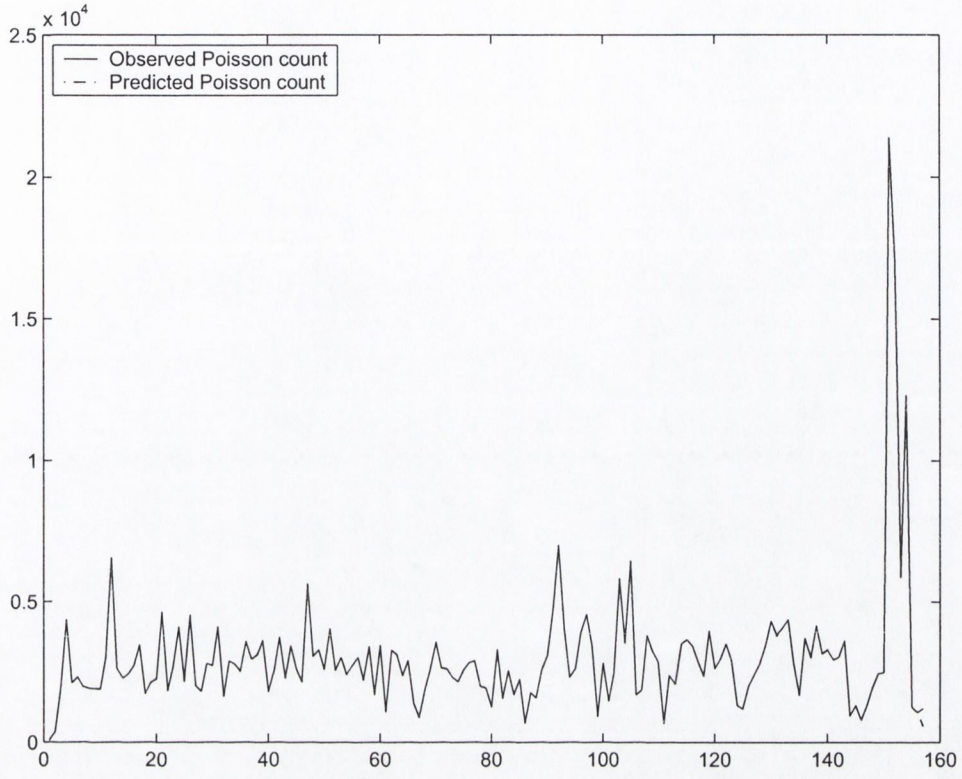


Figure 9.9: Index plots comparing observed and predicted Poisson counts.

see Gilks and Roberts (1996). This reparameterisation proved to be far more effective than the block updating method of Rue (2001).

9.3 Results of cross-validation

The results of cross-validation are displayed in Figure 9.9. Modes of the leave-one-out posterior distributions, which are taken as predicted values at cases omitted, agree closely with the corresponding observed counts. Also, all 157 observed values lie within their respective 95% credible regions. However, this does not necessarily indicate that the model fits the data satisfactorily since D_1^{obs} is not consistent with the reference distribution D_1^{var}; see Figure 9.10. Note that observed D_1^{obs} is smaller than the smallest value that receives significant mass. This indicates that although the predicted and the observed counts agree with each

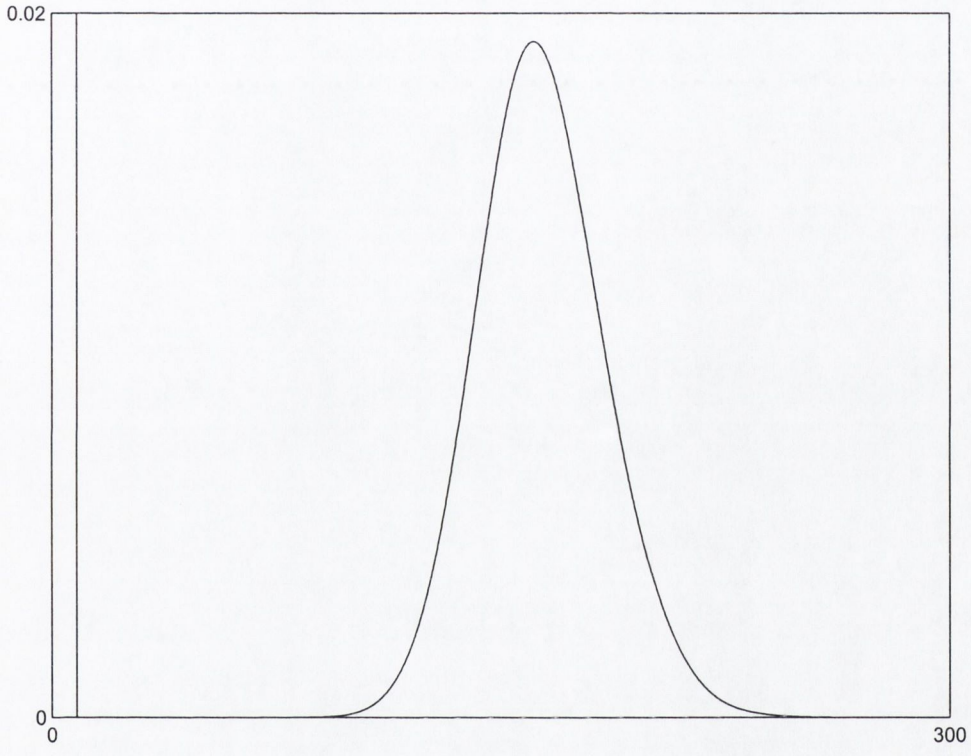


Figure 9.10: Distribution of D_1^{var}. The vertical line denotes observed D_1^{obs}.

other fairly well, uncertainties within the leave-one-out distributions are very high. A similar conclusion has been drawn in Chapter 7 in the case of the pollen based palaeoclimate reconstruction model. Diggle et al. (1998) also report impreciseness of the model.

Chapter 10

Conclusions and future work

In this thesis work we attempt to propose an alternative to regular MCMC for leave-one-out cross-validation in inverse problems. We call this new methodology IRMCMC.

The effectiveness of our proposal has been illustrated with a toy example and applications to three real, high-dimensional problems (inverse as well as forward). The superiority of IRMCMC over regular MCMC is clear; the former is many times more faster than the latter and is particlarly more efficient in exploring multimodal distributions. Besides, we have shown that IRMCMC may also be employed for sensitivity analysis and indicated its use in more general problems. We have also indicated that output obtained by IRMCMC in a cross-validation exercise can be used to construct useful reference distributions of omnibus measures of model fit. However, this needs further work. For example, it is not clear to us which omnibus measure is the most appropriate one.

The 'restart' mechanism involved in IRMCMC requires further discussion, as does the issue of the ability of IRMCMC to locate multiple modes. As an exploratory strategy, the idea of restarts using independent initial values scattered around the parameter space may be useful. This is particularly the case if the modes are used as starting points. In practice, the main modes will often be induced separately by the prior and by the likelihood and it will then be possible to locate them, by deterministic hill climbing, from knowledge of the posterior distribution (up to scale) (Besag and Green (1993)). But combining separate runs

into coherent inferences does not seems straightforward. Ideally, one would like each run to be sufficient in length that it samples all the modes frequently, in which case, multiple runs have no intrinsic merit (Besag and Green (1993)).

The choice of N, M and K requires further work, particularly in the case of multimodal distributions. In fact, the size of the initial regular MCMC sample, N, must be large enough to ensure that the parameter space of the entire set of leave-one-out posteriors is very adequately represented. In practice this seems very difficult to ensure and it is not clear how large N should be.

Moreover, even if the entire parameter space is adequately represented by the optimised regular MCMC for the initial run, there is the danger of inadequate representation in the IR step, even if the subsample size, K, is large enough (but how large is "large enough"?). In the pollen based palaeoclimate example, comparison of the performance of IRMCMC with regular MCMC for all 7815 leave-one-out posteriors was infeasible. Hence, although IRMCMC and regular MCMC were in agreement for the case with maximum distance, we provide no guarantee that IRMCMC performed satisfactorily for all leave-one-out posteriors. The issue of the choice of M, the size of the MCMC sample generated from $\pi(x \mid y_i, \theta)$, for fixed θ, seems less challenging. This is because the dimensionality of x is very low and a long enough MCMC run is not difficult to implement. In our examples the sizes of samples obtained from the leave-one-out posteriors are small considering the large number of parameters and multimodality issues. Clearly, further work is needed in this direction and very much longer runs are desirable. Given fixed θ, such very long runs are simple to conduct. But the challenge lies in running the initial regular MCMC for a very large number of iterations, given that obtaining a sample of size 1000 takes more than 5 hours in the case of the pollen example with the hardware at our disposal. We also do not claim to have properly optimised our initial regular MCMC algorithms and there is scope for improvement in this aspect. However, different reasonable choices of N, M and K with our proposal mechanism did not alter our results significantly.

The performance of IRMCMC depends on the importance sampling distribution chosen. Identifying available importance sampling distributions with the corresponding sample number, in Chapter 4 we offered a method of selecting the most adequate sample number (denoted by i^*) by minimising distance functions which provide suitable measures of centrality.

Apart from minimising distance functions another way of selecting an appropriate i^* may be suggested. A new data set $(\tilde{x}, \tilde{y})$ may be defined where each of the (possibly multivariate) terms of $\tilde{x}$ and of $\tilde{y}$ is given by the median in (X, Y) for the corresponding variable. We may use this artificially constructed case as i^*. However, the procedure of choosing an appropriate i^* may require further work. Indeed, when the variance of the data is very high, the central points will poorly approximate extreme data points and hence IRMCMC may produce unreliable results. In such cases carefully implemented regular MCMC may be more reliable. Another possibility may be to re-use samples obtained from a mixture of several leave-one-out posteriors, corresponding to diferent values of i^*, perhaps corresponding to maximum, minimum and the median distance. This may facilitate adequate representation of most of other leave-one-out posteriors. However, it is not clear how to obtain the importance weights in this case since the respective normalizing constants do not cancel in the ratio.

The behaviour of importance weights needs some discussion. The proposal of Skare et al. (2003) seems promising but it is computationally burdensome. The use of importance link functions (ILF) (MacEachern and Peruggia (2000)) may improve the behaviour of the importance weights in some cases but the procedure is not easy to apply in the case of complicated models (Vehtari and Lampinen (2002)). If the size of the data set is moderate, computationally intensive adaptive importance sampling methods (see, for example, Zlochin and Baram (2002)) may also be used. However, we have shown in Chapter 5 that, for large data size, careful selection of i^* is unimportant. This has also been seen in the case of the real example presented in Chapter 9.

It has been demonstrated in Chapter 8 that IRMCMC can be applied to general problems

as well. But this requires good importance sampling densities that are challenging to obtain in practice. In the case of cross-validation or sensitivity analysis, such appropriate densities seemed to be presenting themselves naturally. An appropriate choice of i^* was all that was required. However, the applicability of IRMCMC to problems different from those we considered very much depends on our ability to find sufficiently thick-tailed distributions that adequately represent the target distribution. For complex, high dimensional problems such a choice may be very difficult to make.

It has been shown in Chapter 5 that the distributions of θ are asymptotically equivalent whereas those of x are not. However, handling the posteriors become easy if they have a convenient asymptotic form. For example, in case the posteriors of θ are asymptotically normal, then direct simulation is possible. Even better, it may be possible to marginalise the posterior of x using some second order asymptotic theory, for example, Laplace approximation (see, for example, Tierney and Kadane (1986)). Shun and McCullagh (1995) consider the applicability of Laplace approximations to high dimensional problems. Although we do not use such asymptotic approximations in this thesis work we recognise such approach as a future possibility.

The fact that our method has been used for palaeoclimate reconstruction (although in the disguise of sensitivity analysis) makes it clear that IRMCMC stretches beyond cross-validation (and sensitivity analysis) to the realm of general simulation from any probability distribution consisting of at least two variables. More specifically, if $P(x, \theta)$ is a probability density of variables x and θ which can be factorised as $P(x, \theta) = P(x \mid \theta)P(\theta)$, and if there exists a density ($g(\theta)$, say) that closely mimics $P(\theta)$, the marginal density of θ, then it is possible to realise θ from $g(\theta)$ by MCMC or otherwise and reweigh the realisations towards $P(\theta)$, preferably by importance resampling and then sample x from $P(x \mid \theta)$. But this requires good importance sampling densities $g(\theta)$; this can be quite challenging. In the case of cross-validation or sensitivity analysis, such appropriate densities seemed to be presenting themselves naturally. Indeed, we only needed to choose from an already available

set of distributions, most of which were good approximations to the target distribution, at least asymptotically. But the applicability of IRMCMC to problems different from those we considered very much depends on our ability to find sufficiently thick-tailed distributions that adequately represent the target distribution. For complex, high dimensional problems such a choice may be very difficult to make. Geweke (1989) proposed importance sampling densities that work adequately for a very wide range of problems.

We remark that the possibilities for sensitivity analysis include using each θ for all models in turn. For example, in the case of the sensitivity analysis conducted in Chapter 8, each simulated value of θ may be used for each of the three random walk priors and the complete temporal independence prior. Although it is certainly possible that any given θ may not be acceptable for all models, the asymptotically valid assumption that the weights are equal may be useful in this connection. In such a case, it is possible to conduct sensitivity analysis at a much finer level of detail, with control for θ. This we also recognise as future work.

Bibliography

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics 16*, 125–127.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes With Applications to Nonparametric Problems. *The Annals of Statistics 2*, 1152–1174.

Athreya, K. B., H. Doss, and J. Sethuraman (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics 24*(1), 69–100.

Bayarri, M. J. and J. O. Berger (1999). P-values for composite null models. Technical report, Institute of Statistics and Decision Sciences, Duke University.

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

Berger, J. O., B. Liseo, and R. L. Wolpert (1999). Integrated likelihood methods for eliminating nuisance parameters (with discussion). *Statistical Science 14*(1), 1–28.

Bernardo, J. M. and A. M. Smith (1994). *Bayesian Theory.* New York: J. Wiley and Sons.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B 36*, 192–236.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian Computation and Stochastic Systems (with discussion). *Statistical Science 10*(1), 3–66.

Besag, J. and P. J. Green (1993). Spatial Statistics and Bayesian Computation. *Journal of the Royal Statistical Society. Series B(Methodological) 55*(1), 25–37.

Billingsley, P. (1995). *Probability and Measure.* New York: John Wiley and Sons.

Birks, H. (1995). Quantitative palaeoenvironmental reconstructions. In D. Maddy and J. Brew (Eds.), *Statistical Modelling of Quaternary Science Data, Technical Guide 5*, pp. 161–254. Cambridge, Quaternary Research Association.

Birks, H. (1998). Numerical tools in paleolimnology – progress, potentials and problems. *Journal of Paleolimnology 20*, 307–332.

Blackwell, D. and L. Dubins (1962). Merging of opinions with increasing opinions. *The Annals of Mathematical Statistics 33*, 882–886.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A 143*(4), 383–430.

Brieman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association 87*, 738–754.

Brieman, L. and P. Spector (1992). Submodel selection and evaluation in Regression. The X-Random Case. *International Statistical Review 60*, 291–319.

Burman, P., E. Chow, and D. Nolan (1994). A cross-validatory method for dependent data. *Biometrika 81*(2), 351–358.

Burman, P. and D. Nolan (1992). Data dependent estimation of prediction functions. *Journal of Time Series Analysis 13*(3), 189–207.

Carlin, B. P. and T. A. Louis (1996). *Bayes and empirical bayes methods for data analysis.* Chapman and Hall. Second Edition.

Casella, G. and C. P. Robert (1998). Post-processing Accept-Reject samples: recycling and rescaling. *Journal of Computational and Graphical Statistics 7*(2), 139–157.

Crawford, S. (1989). Extensions to the CART algorithm. *International Journal of Machine Studies 31*, 197–217.

Dansgaard, W., J. White, and S. Johnsen (1989). The abrupt termination of the younger dryas climate event. *Nature 339*, 532–534.

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

Dey, D. and T. Matai (2002). *Encyclopedia of Environmetrics*, Chapter Dirichlet-Multinomial Distribution, pp. 522–523. Wiley.

Diaconis, P. and D. Freedman (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics 14*, 1–67.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics 26*, 363–397.

Diggle, P., J. Tawn, and R. Moyeed (1998). Model-based geostatistics. *Applied Statistics 47*(3), 299–350.

Doss, H. (1994). Discussion: Markov Chains for Exploring Posterior Distributions. *Annals of Statistics 22*(4), 1728–1734.

Draper, D. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica 6*(4), 760–767. Discussion.

Edwards, R. G. and A. D. Sokal (1988). Generalisation of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D 38*, 2009–2012.

Escobar, M. D. and M. West (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Ferguson, T. S. (1974). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics 1*, 209–230.

Ferguson, T. S. (1983). Bayesian Density Estimation by Mixtures of Normal Distributions. In H. Rizvi and J. Rustagi (Eds.), *Recent Advances in Statistics*, pp. 287–302. New

York: Academic Press.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association 70*(350), 320–328.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), 153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, London, pp. 145–162. Chapman and Hall.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society B 56*(3), 501–514.

Gelfand, A. E., D. K. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling methods(with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 147–167. Oxford University Press.

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. R. Rubin (1995). *Bayesian Data Analysis.* Chapman and Hall. Second Edition.

Gelman, A., X. L. Meng, and H. Stern (1993). Bayesian model checking using tail area probabilities. Technical report, Department of Statistics, University of California, Berkeley.

Gelman, A., X. L. Meng, and H. S. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica 6*, 733–807.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5*, pp. 599–607. Oxford University Press.

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica 57*(6), 1317–1339.

Geyer, C. J. (1991a). Monte carlo maximum likelihood for dependent data. In E. Keramidas (Ed.), *Computer Science and Statistics: Proceedings of of the 23rd Symposium Imterface*, pp. 156–163.

Geyer, C. J. (1991b). Reweighting Monte Carlo Mixtures. Technical Report 568, University of Minnesota, School of Statistics.

Ghosal, S., J. K. Ghosh, and T. Samanta (1995). On convergence of posterior distributions. *Annals of Statistics 23*, 2145–2152.

Ghosh, J. K., S. Ghosal, and T. Samanta (1994). Stability and convergence of posterior in non-regular problems. In S. S. Gupta and J. O. Berger (Eds.), *Statistical Decision Theory and Related Topics V*, pp. 183–199. Springer-Verlag.

Ghosh, J. K. and R. V. Ramamoorthi (1994). Lecture notes on Bayesian asymptotics. Under preparation.

Gilks, W. R. and G. O. Roberts (1996). Strategies for improving MCMC. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, London, pp. 89–114. Chapman and Hall.

Golub, G., M. Heath, and G. Wahba (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics 21*, 215–224.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B 14*, 107–114.

Goodman, J. and A. D. Sokal (1989). Multigrid Monte Carlo method. *Physical Review Letters D 40*, 2035–2072.

Haslett, J. and D. Dillane (2004). Application of 'delete=replace' to deletion diagnostics for variance component estimation in the linear mixed model. *Journal of the Royal Statistical Society. Series B 66*, 131–143.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models.* London: Chapman and Hall.

Hastings, W. K. (1970). Monte Carlo Sampling Using Markov Chains and Their Applications. *Biometrika 57*(1), 97–109.

Higdon, D., H. Lee, and C. Holloman (2003). Markoc chian Monte Carlo-based approaches for inference in computationally intensive inverse problems. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 7.* Oxford University Press.

Huntley, B. (1993). The use of climate response surfaces to reconstruct palaeoclimate from quaternary pollen and plant macrofossil data. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences 341*, 215–223.

Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag: New York.

Imbrie, J. and N. Kipp (1971). The Agulhas current during the late pleistocene: Analysis of modern faunal analogs. *Science 207*, 64–66.

Jones, G. L. and J. P. Hobert (2001). Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo. *Statistical Science 16*(4), 312–334.

Korhola, A., K. Vasko, H. T. T. Toivonen, and H. Olander (2002). Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling. *Quaternary Science Reviews 21*, 1841–1860.

Lee, H., D. M. Higdon, Z. Bi, M. Ferreira, and M. West (2002). Markov Random Field Models for High-Dimensional Parameters in Simulations of Fluid Flow in Porous Media. *Technometrics 44*(3), 230–241.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses.* New York, Inc: Springer-Verlag.

Liu, J. and C. Sabatti (1999). Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid,

and A. F. M. Smith (Eds.), *Bayesian Statistics 6*, pp. 386–413. Oxford University Press.

MacEachern, S. N. and M. Peruggia (2000). Importance link function estimation for Markov chain Monte Carlo methods. *Journal of Computational and Graphical Statistics 9*(1), 99–121.

Marshall, E. C. and D. J. Spiegelhalter (2003). Approximate cross-validatory predictive checks in disease-mapping models. *Statistics in Medicine 7*(22), 1649–1660.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalised linear mixed models. *Journal of the Americal Statistical Association 92*, 162–170.

Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics 22*, 1142–1160.

Metropolis, N., A. Rosenbluth, R. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics 21*(6), 1087–1092.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap(with discussion). *Journal of the Royal Statistical Society. Series B 56*(1), 3–48.

O'Hagan, A. and J. Forster (2004). *Kendall's Advanced Theory of Statistics, Bayesian Inference (Vol 2B)*. London: Arnold.

Oliver, D. S., L. B. Cunha, and A. C. Reynolds (1997). Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology 29*, 61–91.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

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

Roberts, G. O. and J. S. Rosenthal (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science 16*(4), 351–367.

Robins, J. M., A. van der Vaart, and V. Ventura (1999). The asymptotic distribution of P-values in composite null models. Technical report, Department of Epidemiology, Harvard School of Public Health.

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 395–402. Oxford: New York.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics 12*, 1151–1172.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fractions of missing information are modest: the SIR algorithm, discussion of m. a. tanner and w. h. wong, The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 543–546.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method.* New York: J. Wiley.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B 63*, 325–338.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*(422), 486–494.

Shun, Z. and P. McCullagh (1995). Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society. Series B 57*(4), 749–760.

Skare, Ø., E. Bølviken, and L. Holden (2003). Improved Sampling-Importance Resampling and Reduced Bias Importance Sampling. *Scandivanian Journal of Statistics 30*, 719–737.

Smith, A. F. M. and G. O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society. Series B 55*, 3–24.

Stern, H. S. and N. Cressie (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine 19*, 2377–2397.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B 36*, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B 39*, 44–47.

Stuart, A. and J. K. Ord (1986). *Kendall's Advanced Theory of Statistics*. London: Charles Griffin and Company Limited.

Stuiver, M., P. M. Grootes, and T. F. Braziunas (1995). The GISP2 δ^{18}O climate record of the past 16,500 years and the role of the sun, ocean, and volcanoes. *Quaternary Research 44*, 341–354.

Tanner, M. A. (1996). *Tools for Statistical Inference*. New York, Inc.: Springer-Verlag.

ter Braak, C. J. F. (1987). *Unimodal models to relate species to environment*. Doctoral thesis, University of Wageningen.

ter Braak, C. J. F. (1995). Non-linear methods for multivariate statistical calibration and their use in paleoecology: a comparison of inverse and classical approaches. *Chemometrics and Intelligent Laboratory Systems 28*, 165–180.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions (with discussion). *Annals of Statistics 22*(4), 1701–1702.

Tierney, L. and J. B. Kadane (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association 81*(393), 82–86.

Toivonen, H. T. T., H. Mannila, A. Korhola, and H. Olander (2001). Applying

Bayesian statistics to organism-based environmental reconstruction. *Ecological Applications 11*(2), 618–630.

Trudinger, C. M. (2000). *The carbon cycle over the last 1000 years inferred from inversion of ice core data.* Doctoral thesis, Monash University.

Trudinger, C. M., I. G. Enting, and P. J. Rayner (2002). Kalman filter analysis of ice core data: method development and testing of statistics. *Journal of Geophysical Research 107*(20), 73–93.

Trudinger, C. M., I. G. Enting, P. J. Rayner, and R. J. Francey (2002). Kalman filter analysis of ice core data: double deconvolution of CO_2 and $\delta^{13}C$ measurements. *Journal of Geophysical Research 107*(20), 73–93.

van Deusen, P. and G. Reams (1996). Bayesian procedures for reconstructing past climate. In J. Dean, D. Meko, and T. Swetnam (Eds.), *Tree Rings, Environment and Humanity*, pp. 335–339. Arizona, Radiocarbon.

Van Dijk, H. K. and T. Kloeck (1984). Experiments with some alternatives for simple importance sampling in Monte Carlo integration. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics II*, North-Holland, Amsterdam.

Vasko, K., H. T. Toivonen, and A. Korhola (2000). A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction. *Journal of Paleolimnology 24*, 243–250.

Vehtari, A. (2001). *Bayesian Model Assessment and Selection Using Expected Utilities.* Doctoral thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of Computational Engineering.

Vehtari, A. and J. Lampinen (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation 14*(10), 2439–2468.

Wahba, G. (1980). Spline bases, regularization, and generalized cross validation for solving

approximation problems with large quantities of noisy data. In W. Cheney (Ed.), *Approximation Theory III*, pp. 905–912. Academic Press.

Wahba, G. (1990). Spline Functions for Observational Data. CBMS-NSF Regional Conference series, SIAM. Philadelphia.

West, M. (1996). Some statistical issues in paleoclimatology (with discussion). In J. Berger, J. Bernardo, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 5*. Oxford: Oxford University Press.

Whiley, M., J. Haslett, S. Bhattacharya, J. R. M. Allen, and B. Huntley (2004). Bayesian palaeoclimate reconstruction. Technical report, Trinity College, Dublin, Ireland. Available at http://www.tcd.ie/Statistics/staff/johnhaslett.shtml.

Zhang, P. (1992). Model selection via multifold cross-validation. Technical report, Department of Statistics, University of California, Berkeley.

Zlochin, M. and Y. Baram (2002). Efficient Nonparametric Importance Sampling for Bayesian Learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'2002)*, pp. 2498–2502.