



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.

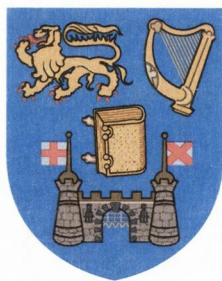
I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated 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.

Advances in Bayesian Model Development and Inversion in Multivariate Inverse Inference Problems

with application to palaeoclimate reconstruction

A thesis submitted to University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Statistics, School of Computer Science and Statistics,
University of Dublin, Trinity College

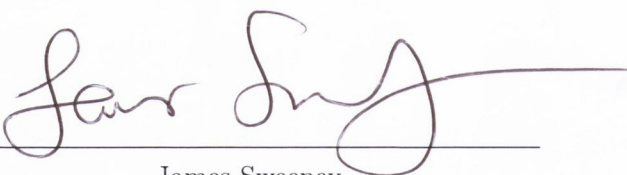


July 2012

James Sweeney

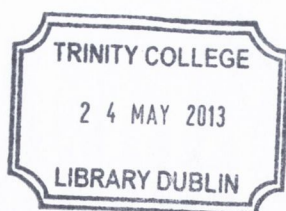
Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and James Sweeney.



James Sweeney

Dated: July 12th, 2012



Thesis 10049

Summary

An extremely challenging example of a multivariate inverse inference problem is the statistical reconstruction of palaeoclimate from fossil pollen data, which represents the motivating research problem considered in this thesis. The model training dataset, consisting of highly multivariate, zero-inflated compositional counts for vegetation, as well as measurements on several climate covariates, presents numerous challenges of model choice and inference. The addressing of these challenges provides the focus for the research contributions presented herein.

Specifically, a statistical inconsistency of existing spatial models for zero-inflated compositional counts data is identified and a parsimonious modelling solution to this problem developed. We discuss hierarchical or “nesting” structures for the decomposition of joint inference tasks, in the context of multivariate compositional data models, into the product of independent, less computationally challenging inference problems, and detail how the optimal decomposition structure can be learned from the data.

Outstanding issues of data analysis and model criticism prompt the development of a methodology for Bayesian residual analysis and outlier detection in the context of discrete, non-Gaussian count data. This methodology is built upon the use of Gaussian random effect terms as a surrogate for classical residuals and the harnessing of fast approximate Bayesian inference algorithms to provide computationally efficient implementations of the method. We demonstrate how the approach provides a visual method for the quick approximate validation of *a priori* model assumptions and the learning of underlying residual trend structure within the data.

The drawbacks of omitting influential climate covariates from forward models and the resulting impact on inverse stage inferences are detailed, and the existing forward modelling methodology for the palaeoclimate reconstruction problem is extended to incorporate an additional climate covariate. This prompts the development of a sampling scheme for model inversion which is demonstrated to be significantly faster than the competing, deterministic model inversion methods.

Finally, the above advances in Bayesian model development and the new method for computationally efficient inverse inference are applied to the palaeoclimate reconstruction problem and the progress over existing models and methods is demonstrated.

Acknowledgements

First and foremost I would like to express my thanks to my supervisor Professor John Haslett. John has been a wonderful supervisor, allowing me the space to make my own discoveries and yet closely attending when I needed direction. He has always been extremely helpful and patient, to too great a degree I must add, considering the extremely slow start I made to my research career. His knowledge of statistics is astounding and his continuing enthusiasm for my work has never ceased to amaze me. I owe all my statistical knowledge to John and I greatly appreciate the learning opportunities he has provided over the past couple of years. It is perhaps only now that I understand my great luck in having a supervisor who welcomes a discussion at a moments notice and who has always made time for me. I consider myself extremely fortunate to have been able to spend the past four years learning from John and I sincerely hope our collaboration will continue into the future.

Secondly, I would like thank Professor Håvard Rue for a very productive three months spent in Trondheim in the autumn of 2010. Much of the work contained in this thesis arose from my time in NTNU and a lot of it would not have been possible without his technical assistance. I acknowledge and thank the Norwegian Government, whose scholarship pool provided the funds for my trip to Trondheim.

I would like to thank my family, in particular my parents who have always been encouraging and supportive of my studies, perhaps too much so at times! I owe all that I have to them and I greatly appreciate all they have done and continue to do for me.

A thank you must go to Louis Aslett, without whose help and advice on computer matters I'm convinced half of the P.h.D. studies in the department would not be completed. I must also express my thanks to my wonderful group of friends who have put up with my disappearance for the past few months and who have always endeavoured to get me out of my pyjama's. A special mention to Adrian who is also finishing at this time; it appears our race to the bottom has ended in a draw! Thank you to Ronan, David and Aaron for your patience, having been burdened with living with someone in thesis writing mode for the past couple of months.

Finally and most importantly, I would like to express a very special thank you to my dear Rebecca Foley with whom I have shared the very many ups and downs of this time. Thank you for your patience and support.

Publications

Some of the material presented in this thesis has been taken from the author's following publication which is co-authored by Professor John Haslett.

Sweeney, J. and Haslett, J. (2011), Bayesian residual analysis in Poisson regression models, *Proceedings of the 26th International Workshop on Statistical Modelling*, **26**, pg 587-592.

Contents

Summary	vi
Acknowledgements	vi
Publications	vi
1 Introduction	1
1.1 Palaeoclimate Reconstruction Project	2
1.1.1 Motivation	2
1.1.2 Quantitative Reconstruction of Past Climate	3
1.1.3 Remaining Outstanding Challenges	4
1.2 Application Datasets	5
1.2.1 The RS10 Pollen Dataset	5
1.2.2 AMI Dataset	5
1.3 Overview of Chapters	6
1.4 Research Contributions	9
2 Palaeoclimate Reconstruction from Pollen Data	10
2.1 Palaeoclimate Reconstruction	10
2.2 Classical Approach	11
2.3 The Bayesian Approach	13
2.4 Advances in this Thesis	17
3 Statistical Methodology	19
3.1 Bayesian Inference	19

3.1.1	Bayesian Hierarchical Models	20
3.2	Markov Chain Monte Carlo	20
3.2.1	The Metropolis-Hastings Algorithm	21
3.2.2	Gibbs Sampling	22
3.2.3	Empirical Bayes	23
3.3	Integrated Nested Laplace Approximations	24
3.4	Spatial Prior Models	25
3.4.1	Gaussian Markov Random Fields	26
3.5	Modelling Discrete Data	28
3.5.1	Log-Link Functions	28
3.5.2	Spatial Zero-Inflated Models	29
3.5.3	Overdispersion Models	31
3.6	Inverse Problems	32
3.6.1	Toy Example	33
3.7	Model Validation for Inverse Problems	37
3.7.1	Cross Validation in the Inverse Sense	37
3.7.2	Mean Square Error of Prediction	41
4	Bayesian Residual Analysis for some Non-Gaussian Response Models	43
4.1	Bayesian Residual Analysis: An Overview	44
4.2	Gaussian Random Effects as a Tool for Residual Analysis	46
4.2.1	A Methodology for Residual Analysis and Outlier Detection	48
4.2.2	Model Framework	49
4.2.3	Fast Approximate Bayesian Inference for Posterior Random Effects	50
4.2.4	Residual Analysis and Outlier Detection	51
4.3	Properties Of the Proposed Methodology	55
4.3.1	Outlier Detection Properties in the Poisson Setting	56
4.3.2	Outlier Detection Properties in the Binomial Setting	60
4.3.3	Quick Approximate Model Validation	62
4.3.4	Strengths and Weaknesses of the Methodology	66
4.4	Application: Heart Attack Dataset	67

4.4.1	Model	68
4.4.2	Results	68
4.5	Conclusions	71
5	Models for Multivariate Observational Data	74
5.1	Multivariate Observational Data	75
5.1.1	Multivariate Observational Models	76
5.1.2	Univariate Models	77
5.1.3	Model Inversion	78
5.2	Multivariate Counts Data	79
5.2.1	Modelling Overdispersion	80
5.2.2	Sensitivity to Dependence Structure in the Latent Field	81
5.3	Dependence in the Likelihood	84
5.3.1	Multinomial Likelihood Function	84
5.3.2	Modelling the Multinomial Response	85
5.3.3	Sensitivity to Dependence Structure in the Likelihood	86
5.4	Decomposing Models Involving Multinomial Likelihoods	89
5.4.1	Hierarchical or “Nesting” Structures	91
5.4.2	Choice of Nested Comparisons	93
5.4.3	Choosing the “Best” Nesting Structure	95
5.5	Addressing Zero/N-inflation of the Multinomial Response	98
5.5.1	Statistical Inconsistency of Zero-Inflated Models for Binomial Data	98
5.5.2	Sensitivity to Zero/N-Inflation	100
5.6	Conclusions	102
6	Spatial Prior Models and Computationally Efficient Inverse Inference	105
6.1	Spatial Covariates	106
6.1.1	The Impact of Spatial Covariate Omission	107
6.2	Spatial Prior Models	110
6.2.1	Discrete Approximations to Continuous Spatial Fields	110
6.2.2	Random Walk Prior Models in \mathbb{R}^1	112

6.2.3	Random Walk Prior Models in Several Spatial Dimensions	117
6.3	Fast Inverse Prediction Given New Data	121
6.3.1	Numerical Evaluation Of Posteriors	122
6.3.2	Sampling Scheme for Computationally Efficient Inverse Inference	123
6.3.3	Performance of the Approach for Univariate Y	127
6.3.4	Extension to the Multivariate setting	134
6.4	Conclusions	135
7	Application: The Palaeoclimate Reconstruction Project	137
7.1	Bayesian Palaeoclimate Reconstruction Project	137
7.1.1	The RS10 Dataset	137
7.2	Forward Modeling and Inference Methodology	138
7.2.1	Modeling the Latent X_i	139
7.2.2	Modeling the Response	141
7.2.3	Inference	142
7.2.4	Results	143
7.2.5	Residual Analysis and Outlier Detection	144
7.3	Model Inversion	152
7.3.1	Inverse Inference	152
7.3.2	Results: 2D Climate Application	153
7.3.3	Results: 3D Climate Application	167
7.4	Fossil Climate Reconstruction at Glendalough	175
7.4.1	Discussion	177
7.5	Conclusions	181
8	Conclusions and Further Work	183
8.1	Conclusions	183
8.2	Further Work	185
8.2.1	Analysis of the RS10 Pollen Dataset	186
8.2.2	4 Dimensional Climate Space	186
	Bibliography	188

Chapter 1

Introduction

One of the primary aims in statistical modelling is to measure the influence of variables (called covariates) on the observed measure(s) of interest (the response). The aim may be to extract some subtle understanding of potentially complex relationships from the data, or to use calibrated models for prediction given “new data” as in inverse problems. Typically, several models are proposed and discrimination between the various models may be based on any number of criteria, possibly leading to the choice of an “optimal” model.

However, both time and computational constraints place limits on the number and complexity of models that can be considered. Where the observed response is non-Gaussian in nature the detection of “outliers”, or fast model validation through the inspection of posterior residuals, can be tremendously difficult. The burden of inference can enforce compromises in model complexity, resulting in the omission of important predictor variables from proposed data models. Additionally, the inversion of calibrated multivariate models for prediction can become computationally challenging, primarily due to evaluations of complex multidimensional integrals, a problem that worsens given increasingly sophisticated models.

These problems arise in the context of the palaeoclimate reconstruction project - the main application of interest in this thesis, which motivates much of the work contained herein. The huge, multivariate, spatially referenced counts dataset, available for model fitting, provides several computational and methodological challenges to the statistical modeller.

The aim of this thesis is to address many of these challenges. This includes the development of a richer class of models for palaeoclimate reconstruction. This will require extensions to existing statistical modelling methodology for large spatial regression problems, as well as developing computationally efficient algorithms for prediction given the fitted models. A further aim is the development of a methodology for the criticism of the training dataset, namely methods for the fast detection of outliers and model criticism for discrete, non-Gaussian count observations.

In order to provide a further introduction to the motivation for these aims, in the following

we describe the palaeoclimate reconstruction project and provide an overview of the remaining, outstanding challenges. We also introduce the datasets we propose to use, provide an outline of the structure of the thesis and summarise the research contributions made.

1.1 Palaeoclimate Reconstruction Project

1.1.1 Motivation

As politicians and scientists become increasingly aware of the potential catastrophic results of extreme changes in the Earth's climate, much time and effort has been devoted to the development of sophisticated global climate models for exploring possible future climate outcomes. However, future climate is inherently unknown; it is impossible to validate the speculative future climate predictions produced.

Conversely, palynological or proxy-based reconstructions of palaeoclimate provide a vital source of information from which we can draw inferences on the past - such inferences present a mechanism for the validation of proposed climate models, but also provide an invaluable insight into the Earth's history. For example, the analysis of high resolution oxygen-isotope records, obtained from Greenland ice cores, indicate that there were numerous *rapid* climate fluctuations during the last glacial period (Figure 1.1).

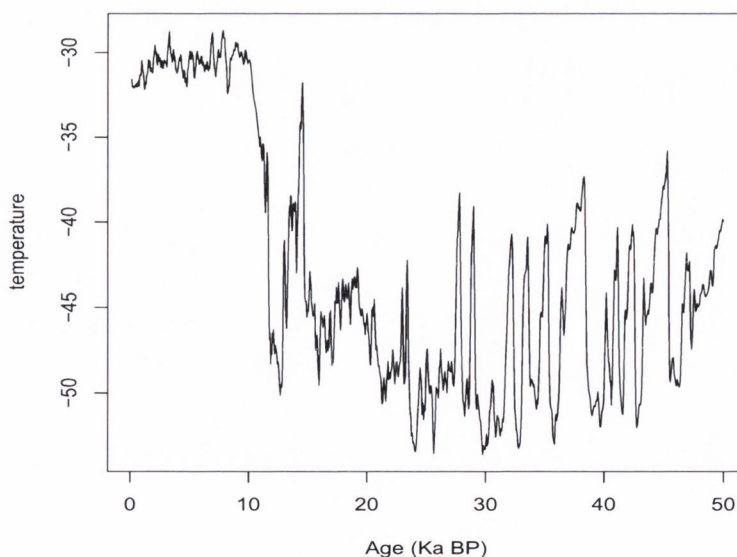


Figure 1.1: Temperature reconstruction at a site in Greenland (Grootes & Stuiver 1995). $\delta^{18}\text{O}$ is used as a proxy for temperature providing the basis for reconstructions of climate in Greenland over the past 50,000 years.

As the recording of environmental data using instrumental sources is a relatively recent occurrence, long term information must be derived from other, indirect, proxy indicators. Several (organism based) proxies for climate exist, including; chironomids (non-biting midges), diatoms (algae), beetles, tree rings and pollen produced by local vegetation, the central premise being that past climate can be inferred from fossil samples of these sources, albeit with some uncertainty. Huntley (1993) notes that many reconstructions of palaeoclimate using such proxies are qualitative, seeking to use the fossil proxy information to classify past climate in terms of similar modern day climatic conditions, or “biomes”. These are essentially subjective descriptions of past climate in terms such as “artic”, “boreal”, “humid”, “arid”; for a description of these biomes and an example of the qualitative reconstruction of palaeobiomes from fossil pollen data, see Allen et al. (2000).

Quantitative reconstructions of the palaeoclimate are to be preferred; such approaches attempt to use analytical tools to provide objective and repeatable reconstructions of past climate, enabling incorporation of data from many different sources and the simultaneous use of multiple proxies.

1.1.2 Quantitative Reconstruction of Past Climate

Haslett et al. (2006) presented a framework for the quantitative, pollen based reconstruction of palaeoclimate at single site locations. The framework involved the specification of a Bayesian hierarchical model for pollen-climate interaction; the response consisted of compositional count vectors of pollen data, with two recorded aspects of climate as covariates. The approach involved the splitting of the reconstruction problem into two distinct stages; at the initial, forward stage, the proposed model for pollen-climate interaction was calibrated using the modern training dataset. At the inverse stage, the inferred model was “inverted” and used to make inferences on past climate given fossil pollen data obtained from lake sediment cores (a broader discussion of the approach is available in Section 2.3).

The authors acknowledged the deferral of several substantial issues in the paper, chief amongst them being the quality of the modern training data. Inference procedures were sampling based; this resulted in the necessity to run models for several weeks and the computational drawbacks encountered by the approach made it difficult to criticize both the data and various other aspects of the methodology. Most importantly, the use of leave-one-out cross validation for model criticism was too computationally expensive to consider, thus the most vital statistic for the approach, the predictive accuracy, was impossible to calculate.

Salter-Townshend (2009) made significant advances in addressing many of the computational issues introduced in Haslett et al. (2006). Through the use of fast approximate Bayesian inference algorithms, Salter-Townshend was able to greatly reduce the time taken at the forward, or model fitting, stage. This facilitated exploration of various aspects of the methodology and concluded in refining of the model, addressing issues including zero-inflation of the counts

dataset. Hierarchical structures, obtained from expert opinion, were also introduced, enabling decomposition of the model likelihood. An approximate, leave-one-out cross validation algorithm was developed, which was used to determine the predictive accuracy of the approach; for the final model, averaged over the whole dataset, the 95% highest posterior density regions for climate contained the true climate locations only 74% of the time.

Ad hoc measures, such as the “Gaussian blurring” of posteriors for climate, in light of data uncertainty, were used to increase the predictive accuracy of the approach further. Some criticism of the training data was attempted; reference distributions for outlier detection were defined, however, posterior sampling distributions for the measures were unknown and thus critical metrics with which to determine outliers, unavailable. Furthermore, due to time constraints, inference procedures were empirical Bayes based; this resulted in the posterior uncertainty in model hyperparameters being unaccounted for at the inverse stage.

Perhaps the most significant result produced in Salter-Townshend (2009), was the conclusion that the use of only two climate covariates, temperature (MTCO) and growing season (GDD5), was not enough for accurate modelling of the pollen response, and was partly the reason for the poor predictive accuracy of the base model.

1.1.3 Remaining Outstanding Challenges

In the context of the palaeoclimate reconstruction project, there remain a number of significant outstanding challenges. To summarise the previous section, the most significant of these include:

- The development of richer models for the palaeoclimate problem, including the extension of existing models to incorporate further climate variables and the construction of likelihood models which capture unique features of the pollen dataset such as N-inflation of the compositional counts.
- The development of a methodology for the fast, computationally efficient detection of outliers in the modern pollen training dataset and methods for explicit criticism of the forward models.
- The development of fast integration schemes for prediction at the inverse stage, which account for all posterior uncertainty in model parameters.

The main aspect of these remaining challenges are acknowledged to be computational in nature and form the background for many of the extensions in the statistical methodology proposed in this thesis. These contributions are outlined in further detail in Section 1.4.

1.2 Application Datasets

Two datasets are used in this thesis which both motivate the research contributions contained herein and provide examples for the application of proposed methodological contributions. In the following a brief introduction is given to each.

1.2.1 The RS10 Pollen Dataset

The RS10 dataset of Allen et al. (2000) consists of a collection of modern (m) sample pollen counts along with associated measurements of several climate covariates, obtained from numerous locations across the northern hemisphere. The sample counts at a given site are provided by examining the pollen composition of a sample of sediment obtained from the upper 5 to 10mm of the surface of the lake bed. The pollen counts, denoted $Y^m = \{y_1^m, \dots, y_n^m\}$, $n = 7742$, are recorded for as many plant types as are distinguishable, though in this thesis information on only 28 plant types is considered. Present-day climate covariates for each site, $C^m = \{c_1^m, \dots, c_n^m\}$, are typically estimated by extrapolating information available from local weather stations. Five main climate covariates are provided, with the ones most frequently utilised in climate models being: a measure of the length of the growing season (GDD5), a measure of winter temperature (MTCO) and a measure of the amount of moisture available locally to plants (AET/PET). The two remaining variables relate to summer temperature (MTWA) and an additional temperature sum measure (GDD0). Taken together, this collection of data is referred to as the modern, or training, dataset. Further details of the model training dataset relevant to the reconstruction problem are discussed in Section 7.1.1.

Additional information, namely fossil pollen counts (denoted Y^f), are obtained from cores of lake bed sediment; typically, a core is divided up into N (potentially non-uniform) slices, with a sample pollen composition provided for each slice; the prehistoric climate (denoted C^f) corresponding to the deposition of the fossil pollen is inherently unknown, though time scale information (with uncertainty), is available from the carbon dating of a number of the pollen slices. In Chapter 7, a fossil pollen core from Glendalough upper, a lake site located in the Wicklow mountains in Ireland, is used for illustrative purposes.

1.2.2 AMI Dataset

A supplementary dataset considered in this thesis, with the purpose of displaying the application of developed methodologies outside of the palaeoclimate context, is the AMI data set. The data set, obtained from Souza & Migon (2010), consists of demographic and medical variables observed upon patient admission to hospital following an Acute Myocardial Infarction (AMI), more commonly referred to as a “heart attack”. A sample of 546 outpatients was observed at Procardíaco Hospital in Rio de Janeiro, Brazil; the recorded response variable is binary,

regarding patient outcome after hospital admission for AMI, namely survival (0) or death (1). There were 73 deaths recorded in total during the study. Information is also available for a number of predictor variables which are hypothesized to be important in predicting in-hospital mortality. There were 11 variables recorded in total for each patient.

The motivation for collection of the data was the development of models for the accurate prediction of patient outcome given the recorded AMI predictor variables. This has an obvious economic benefit in that clinical resources are potentially prioritized for those patients identified as having a higher probability of a negative outcome. Misrecording of the predictor variables corresponding to even a few of the patients can have a negative effect on model fitting and prediction accuracy; with this in mind, Souza & Migon (2010) aimed to identify patients for whom admission or additional data was possibly misrecorded.

1.3 Overview of Chapters

In the following we provide a brief outline of each chapter in this thesis.

Chapter 2: Palaeoclimate Reconstruction from Pollen Data

A brief review of the literature on palaeoclimate reconstruction from pollen data is presented. The process by which fossil climate is reconstructed, using both Bayesian and non-Bayesian methods is detailed, and the current weaknesses of each respective methodology outlined. We focus in particular on the work of Haslett et al. (2006) and Salter-Townshend (2009), who provide the starting points for much of the work contained herein.

Chapter 3: Statistical Methodology

The reconstruction of palaeoclimate from fossil pollen data requires drawing on a wide range of statistical methods and in this chapter we present many of these methods. Statistical models for the non-parametric Bayesian modelling of spatially referenced count data are presented, and procedures for making inference on the parameters of these models outlined. A simple toy example is used to describe the framework of inverse Bayesian inference problems and diagnostic methods for the evaluation of inverse model performance are detailed. Some gaps in the existing methodology are identified and the solutions developed in this thesis are introduced.

Chapter 4: Bayesian Residual Analysis for some Non-Gaussian Response Models

Methods for Bayesian residual analysis and outlier detection, under the assumption of discrete, non-Gaussian count outcomes, are identified as an outstanding challenge. A brief review of the existing literature on Bayesian residual analysis is provided and weaknesses of the current

methodology outlined. This prompts the development of an alternative methodology for outlier detection and residual analysis, based on the incorporation of Gaussian random effect terms into models for residual analysis purposes. The random effects are used as a “surrogate” for classical residuals and a number of simulation studies are used to assess the relative strengths and weaknesses of the proposed methodology. Finally, the power of the proposed approach is illustrated by application to the AMI data set, introduced in Section 1.2.2 above.

Chapter 5: Models for Multivariate Observational Data

In this chapter we focus on models for highly multivariate observational data and investigate the implications of erroneous *a priori* modelling choices. Specifically, we evaluate the impact of ignoring dependence structure between (correlated) model components at the forward stage, on the resulting (inverse) predictive performance of the calibrated models at the inverse stage. Hierarchical or “nesting” structures for compositional data are also introduced, which facilitate the decomposition of multivariate compositional data models into a series of less computationally challenging, univariate models for which inference tasks are computationally much simpler. We detail how the optimal structure for the nesting can be learned from the data. Finally, a statistical inconsistency of existing zero-inflated models for Binomial count data is highlighted. This prompts the development of a simple model extension to address this issue.

Chapter 6: Spatial Prior Models and Computationally Efficient Inverse Inference

The primary focus in Chapter 6 lies in the specification of spatial prior models for the forward stage and the construction of algorithms for computationally efficient inverse inference given calibrated models at the inverse stage. It is demonstrated that failure to include all relevant spatial variables in the forward models, upon which the response depends, will lead to erroneous inferences in spatial prediction. The factors impeding the specification of GMRF-based spatial prior models in several spatial dimensions are detailed. These obstacles are surmounted via the development of an approach for constructing spatial prior models on irregularly shaped regions in several spatial dimensions. Finally, a fast sampling-based scheme for making inverse inferences in calibration problems is developed, which results in a substantial speedup of model inversion tasks. We detail how the computational advantages of the scheme becomes more pronounced for increasing dimension of both C and Y .

Chapter 7: Applications in Palaeoclimate Reconstruction

The motivating research problem considered in this thesis concerns the reconstruction of fossil climate from fossil pollen data. In this chapter we apply the methodologies and algorithms, developed in previous chapters, to the RS10 pollen and climate dataset. We illustrate how the

incorporation of an additional climate variable, AET/PET, into the forward models, in addition to fully accounting for the compositional nature of the data collection procedure, leads to a model with substantially increased inverse predictive accuracy. The proposed methodologies for residual analysis and outlier detection are also applied and possible sources for the slight loss in predictive accuracy of the best-fitting model are identified. Climate reconstructions at Glendalough, for a number of different forward models, both marginal and nested, are compared to independent reconstructions from the statistical literature.

Chapter 8: Results and Conclusions

In Chapter 8 the results from preceding chapters are summarised and discussed. Outstanding issues and challenges of the work presented in this thesis are outlined and possible solutions to these remaining challenges are proposed.

1.4 Research Contributions

The following are the main contributions made by the research contained in this thesis:

1. A novel, Bayesian approach to residual analysis in settings where response variables are both non-Gaussian and discrete is presented, with a focus on the Poisson and Binomial regression model setting. The proposed approach has distinct advantages over existing methods as regards both computational speed, due to the harnessing of fast approximate Bayesian inference algorithms, and the automatic provision of a metric by which to determine potential outliers. It is also demonstrated that exploratory tools from classic Gaussian residual analysis may be harnessed to gain an extra insight into underlying model dynamics at no extra computational cost.
2. Hierarchical (or nesting) structures for models of zero inflated compositional counts data, initially introduced in Salter-Townshend (2009), are explored in greater detail - it is demonstrated that hierarchical structures can provide a full, but not necessarily unique, decomposition of Multinomial model likelihoods. We additionally illustrate how the ‘best’ nesting structure can be learned from the data. Methodological contributions include the development of a zero/ N inflated likelihood for modelling zero-inflated Binomial counts.
3. A fast, sampling based, inference procedure for prediction at the inverse stage in calibration problems is developed, which is demonstrated to be significantly computationally faster than existing deterministic integration algorithms. This has general application in the paradigm of inverse problems.
4. The richer class of models specifically developed for the palaeoclimate reconstruction project are applied. It is demonstrated that the inclusion of an extra climate covariate results in a class of models which have significantly superior predictive accuracy compared to existing, competing models. Finally, a number of exciting new results for the reconstruction of the palaeoclimate at a site at Glendalough in Ireland are presented.

Chapter 2

Palaeoclimate Reconstruction from Pollen Data

The primary objective of this chapter is to provide a brief review of the existing methodology for pollen based palaeoclimate reconstruction, the main motivating problem considered in this thesis. The highly multivariate nature of the data sets used for the statistical reconstruction of past climate provide a number of immense challenges to the statistical modeller, several of which are discussed in detail in the following sections. Although this chapter will focus explicitly on the existing palaeoclimate reconstruction literature, the statistical challenges we introduce and propose to address arise in a wide variety of statistical problems.

2.1 Palaeoclimate Reconstruction

As previously mentioned, palynological reconstructions of palaeoclimate provide a vital source of information from which we can make inferences on past climate. According to Huntley (2001), the main advantage of palaeovegetation based reconstruction of the palaeoclimate is the *multivariate* nature of the response data set; the data set contains information on a wide range of plant types (“taxa”), with each separate plant taxon providing information on multiple aspects of climate. Here, each taxon (plant type) can consist of one or more species, an entire genus or sometimes even a family of several plant species. As each particular plant taxon prefers slightly different climatic conditions (Huntley 1993), the pollen composition of the fossil record can be analysed to make insightful inferences about past climate.

The multivariate nature of the fossil record is necessary as reconstructions of local climate based on single plant taxa can be noisy or unpredictable; for example, Subally & Quézel (2002) note that the *Artemisia* genus can flourish under a number of contrasting climatic conditions - the *Artemisia* genus consists of about 300 species of plants, some of which conversely favour either extremely hot or extremely cold, arid climatic conditions. The benefit of a multivariate

pollen data set is that inferences on climate made by jointly analysing a number of plant taxa can be used to narrow the range of climate conditions that could have occurred thereby reducing the uncertainty in climate predictions.

All of the climate reconstruction methods discussed in this thesis exploit the uniformitarian principle, specifically, “the present is the key to the past”. The general framework of such an approach may be stated as follows; modern pollen data with known modern climate variables is used to calibrate models for pollen-climate interaction. The calibrated models are then “inverted” and used to make inferences on the climate corresponding to fossil pollen, obtained from fossil sediment cores, for which climatic conditions at the time of pollen deposition are unknown. There are a number of limitations to this approach, the principal one being that, in some cases, there exists no modern climate analogue for the pollen composition being analysed. Huntley (2001) discusses this limitation in greater detail and provides an introduction to several others.

In the following sections we focus on the statistical approaches used in model calibration for the pollen-climate relationship. The various estimation methods which appear in the palaeoclimate reconstruction literature can be divided into two distinct strands which we denote, as per Haslett et al. (2006), as “classical” (non-Bayesian) and Bayesian.

2.2 Classical Approach

According to Holden et al. (2008), almost all of the reconstruction methods currently employed by palaeolimnologists apply frequentist or non-Bayesian statistical methods. Furthermore, the majority of these statistical approaches are based on observing and attempting to model the empirical relationships between modern pollen taxa and climate (Birks et al. 2010), making use of the “uniformitarian principle” as introduced in the previous section. In the following we focus on one modelling strategy in particular, the “response surface” method for quantitative climate estimation. It is the method most similar in spirit in a Bayesian sense to the work of Haslett et al. (2006) and Salter-Townshend (2009), which provide the starting points for much of the work contained in this thesis.

Huntley (1993) provides an overview of the response surface method for the reconstruction of past climates from fossil pollen data. Given some observed modern pollen counts Y_i^m along with associated measurements on a number of climate predictor variables C^m , our interest lies in inferring the smooth spatial surface $X_i(C)$ which describes the underlying relationship between pollen and climate for the i^{th} plant taxon. $X_i(C)$ is denoted the “response surface” for taxon i and describes the way in which the pollen counts of the i^{th} taxon vary with climate, C . The pollen counts are thus an indirect observation of the “response” at a given location in climate space, perturbed by both non-climatic environmental and random variation. From here on the model fitting stage will be denoted the “forward stage”; correspondingly, the “inverse

stage” will denote where calibrated models are “inverted” and used to provide quantitative inferences on unknown climates for fossil pollen data. This notation and terminology is adopted for the remainder of this thesis.

Noting the multimodal response to climate of some observed pollen proportions, Bartlein et al. (1986) use cubic polynomials of order two and three as bases in the estimation of response surfaces for eight pollen taxa in two climate dimensions. A response surface is fit to the pollen percentages of each taxon with least squares used to estimate the regression effects. Only the non-zero pollen counts are used to estimate model parameters and an implicit constraint is employed by the authors in that only counts data from the climatological range of each species is used in model fitting; because of the global nature of the polynomial bases used for the response surfaces, the authors are required to transform the pollen percentages and omit the observed zeroes to prevent strange model behaviour at the boundaries of climate space. It is perhaps for this reason that no climate reconstructions are produced; estimated pollen compositions for the eight taxa used in model fitting are instead compared to their observed pollen compositions.

Prentice et al. (1991) attempt to surmount this model fitting problem by using non-linear calibration methods to infer non-parametric response surfaces. Response surfaces are obtained for thirteen different pollen taxa in three climate dimensions by using a locally weighted averaging technique; because the surfaces are fit locally rather than globally, their method does not suffer from the boundary effects experienced by Bartlein et al. (1986). Quantitative climate reconstructions are provided using the calibrated response surfaces; climate values are inferred for the fossil pollen data by “scanning” predicted pollen percentages, obtained from the calibrated response surfaces and comparing them to the observed pollen percentages. The ten “nearest” climates in pollen composition to the observed climates are identified using a squared chord distance dissimilarity measure. In order to address the multimodality of the output, the final (single) inferred climate value is taken as the centroid of the ten proposed climate values, each weighted by the inverse squared distances. An average chord distance measure is used to assess model fit.

Allen et al. (2000) use similar methods to provide reconstructions of the fossil climate corresponding to a fossil pollen core obtained from a site at Monticchio in Southern Italy. Qualitative methods are first used to classify each of the possible fossil climates into plant biomes. The qualitative results are then used to motivate quantitative reconstructions of the palaeoclimate - each fossil pollen sample is compared to predicted samples produced by the inferred response surfaces with the 10 closest modern climate analogues identified, subject to the constraint that the proposed climates must arise from the same biome identified by qualitative methods. As in Prentice et al. (1991), a squared distance metric is then used to provide single climate values from the 10 identified modern climate analogues.

However, Salter-Townshend (2009) notes the unsatisfactory nature of the squared distance

method used by both sets of authors to infer single climate values at the inverse stage. The deficiencies of the approach are illustrated through a hypothetical example where the 10 identified modern climate analogues for a fossil sample occur at contrasting extremes of climate space. The modern analogues are thus separately indicating that climates in the “centre” of the climate space are infeasible given the fossil pollen data. Conversely, if the squared chord distances for each modern analogue are roughly equal, using the inversely weighed centroid of the modern analogues will infer a final climate value which is located in the centre of the climate space, wholly disagreeing with the separate inferences of each analogue!

A further substantial modelling issue involves the decision of which environmental factors to include in the forward models; according to Beerling et al. (1995), the modelling approach at the forward stage is open to the criticism that the mechanism determining a species distribution may not involve the climate variables which are used to develop the model. Huntley (1993) discusses this subject in detail, proposing that reconstructions from models involving three particular aspects of climate, degree of winter cold (MTCO), growing season warmth (GDD5) and a measure of moisture (AET/PET) are to be preferred. Huntley (1993) also details how the models must account for interaction between the various climate covariates, for example a warmer summer will lead to a requirement for more moisture. As a result we can conclude that inferences derived from approaches which calibrate models on individual climate variables separately, such as ter Braak (1995), may be misleading.

The main weakness of classical methods is that there seems to be no consistent way to make statements of uncertainty in the quantitative reconstructions that are produced. Palaeoclimate reconstructions are presented in terms of single climate values that are estimated from multimodal outputs with only cursory measures of uncertainty provided. This deficiency is noted by Holden et al. (2008) who use the interesting phrase “the major weakness of these (sic classical) approaches is that they do not explicitly model the uncertainty associated with individual reconstructions”, a sentiment also expressed in Haslett et al. (2006) and Birks et al. (2010).

2.3 The Bayesian Approach

The Bayesian paradigm provides the solution to the problems introduced in the previous section; statistical parameters are presented as random variables and a likelihood function is used to make probabilistic statements about the random variables in light of the observed data. The main advantage of the Bayesian methodology is that all uncertainties, both uncertainty in model parameters and uncertainty in the data, can be accounted for in a coherent and consistent manner.

In a series of papers, Vasko et al. (2000), Toinvonen et al. (2001) and Korhola et al. (2002) appear to be the first to set out a detailed, Bayesian statistical modelling approach to climate

reconstruction from proxy data. The particular climate proxy considered by the authors are species of chironomids, a type of non biting midge. A parametric form is adopted for the climate response surfaces; unimodal Gaussian curves are used to model the chironomid response to one aspect of climate, the mean July temperature. Toinvonen et al. (2001) adopt a Poisson likelihood for the observed chironomid counts, ignoring the compositional nature of the counts induced by the data collection process. Vasko et al. (2000) extend this approach, accounting for Multinomial structure in the data; the model is evaluated by comparing estimated chironomid proportions, given inferred model parameters, to the observed proportions. Korhola et al. (2002) use this approach to provide a climate reconstruction, given fossil chironomid counts, for a site in northern Fennoscandia and compare the output to results obtained from a number of classical approaches.

However, as previously discussed, the unimodal assumption for climate response surfaces may not be suitable for all ecological processes. Each pollen/chironomid taxon may be comprised of a number of subspecies and thus may have more than one preference of environmental conditions. With this in mind, Bhattacharya (2006) relaxes the unimodal assumption for the response surfaces, using a flexible weighed mixture of Gaussian functions to model the chironomid response to climate. The predictive accuracy of the approach is evaluated in an inverse sense, using a leave-one-out cross validation measure with the author observing a substantial improvement in prediction accuracy given multimodal climate response surfaces.

For more recent work on climate reconstructions under the Bayesian paradigm see Paciorek & McLachlan (2009), who use a Bayesian framework to analyze forest composition from fossil pollen data with the aim of estimating the composition of ancient forests or Li et al. (2010), who use a Bayesian hierarchical model to incorporate three separate proxies for climate along with the use of external forcings to model large scale temperature evolution.

In the following, we introduce the work of Haslett et al. (2006) and Salter-Townshend (2009), who provide the background for many of the methodological and computational contributions that are presented in this thesis.

Haslett et al. (2006)

Haslett et al. (2006) present a Bayesian approach to palaeoclimate reconstruction which is similar in spirit to the response surface method of Huntley (1993). The authors consider a subset of the data (14 pollen taxonomic groups are selected from expert opinion from the total set of 48 taxa) and attempt to reconstruct prehistoric climate for two climate covariates, MTCO and GDD5, at a site in Ireland. The approach, as with standard response surface methods, involves the splitting up of the reconstruction problem into two distinct stages, the forward stage and the inverse stage. In the following we describe the statistical model as presented in the paper.

In the following we denote, by (C^m, Y^m) , the modern climate data (climate measurements C^m and pollen counts Y^m) and by (C^f, Y^f) their fossil data equivalents. At the forward stage, the authors make the explicit approximation that the pollen response surfaces, X , are inferred independent of the fossil pollen data (see Equation 2.2), justifying this assumption by noting that the fossil pollen on its own contains very little information on the latent X . Rougier (2008), see “cutting feedback”, provides additional support for this hypothesis.

$$\pi(X|Y^m, C^m, C^f, Y^f) \approx \pi(X|Y^m, C^m) \quad (2.1)$$

$$= \frac{\pi(Y^m|X, C^m)\pi(X)}{\pi(Y^m, C^m)} \quad (2.2)$$

Haslett et al. (2006) approximate continuous climate space in two dimensions by a fine discrete grid of dimension 51×51 with a “buffer region” employed to reduce the computational burden; a non-parametric conditional autoregression (CAR) model (Besag 1986) is used as a prior for X to ensure smoothness and address the possible multi-modal nature of the latent responses. The local structure in the model, provided by the CAR prior on X , allows the authors to harness the computational advantages of Markov random fields.

However, the normalising constant, $\pi(Y^m, C^m)$, in Equation 2.2 is not known in closed form; Haslett et al. (2006) perform approximate numerical integration using a Metropolis Hastings Markov chain Monte Carlo algorithm. Due to the number of latent parameters introduced by the model (around ten thousand), the model is slow to run and the authors readily admit that convergence of the chains is far from assured. Indeed the high dimensionality of the approach is acknowledged to be a source of much computational burden and leads to several compromises in model complexity.

Specifically, the smoothness parameter of the latent response surfaces, κ , is fixed *a priori*. Additionally, a compound (Dirichlet) multinomial structure is used to model the overdispersed compositional counts which the authors note is perhaps “overrestrictive”; one common δ parameter is used to model the overdispersion across all taxonomic groups. Haslett et al. (2006) admit that this model may not sit well with scientific theory - at many sites certain taxa are completely missing, a feature which is not allowed by the model. The authors also note the extensive number of zeroes in the observational dataset and acknowledge the need to treat zero-inflation of the pollen counts explicitly.

Due to the computational burden imposed by the sampling based inference procedure, model validation tools such as leave-one-out cross validation are inaccessible; rerunning the model for each left out set of datapoints and evaluating the accuracy of the resulting climate predictions given the “truth” is infeasible in finite computing time. A saturated cross validation method (the model is validated using the same counts that are used to train the model) is used instead to evaluate the predictive accuracy of the approach. Using the samples obtained

from the forward stage, the authors determine that approximately 96% and 97% of MTCO and GDD5 values lie inside their corresponding 95% highest posterior density regions.

Salter-Townshend (2009)

Model validation does not play a big role in Haslett et al. (2006). The reliance on sampling based algorithms for inference restricts the comparison of multiple models and the identification of observations which are not well captured by the fitted model. Through the use of fast approximate Bayesian inference algorithms (the INLA algorithm, see Section 3.3), Salter-Townshend (2009) is able to surmount this obstacle - model fitting is reduced from weeks in the case of Haslett et al. (2006) to minutes, with full Bayesian inference on all unknown model parameters. With the computational obstacle of model fitting overcome, Salter-Townshend (2009) is able to criticise and develop various aspects of the methodology.

The use of fast approximate inference algorithms enables the relaxation of a number of the computational concerns of Haslett et al. (2006); existing models are extended to include explicit modelling of the zero-inflation present in the data, at very little extra computational cost. The extra zeroes are addressed through the development of a parsimonious model which captures zero-inflation of the count observations at the cost of inferring an extra, zero-inflation parameter for each taxa. However, whilst Salter-Townshend (2009) identifies “nesting” or hierarchical structures (see Section 5.4) as a valuable tool for decomposing joint models involving Multinomial likelihoods into the product of independent components, the author does not show that such structures provide the correct full likelihood in the context of zero-inflated count outcomes. Furthermore, a point missed by Salter-Townshend (2009) is that, in the context of zero-inflated Multinomial count data, zero-inflation of the counts corresponding to one plant taxa possibly results in “N” inflation of the counts corresponding to another. Thus the proposed models for palaeoclimate reconstruction are not “symmetric” (see Section 5.5.1) and lead to statistically inconsistent inferences in model fitting. One contribution in this thesis is the addressing of this issue; we construct a parsimonious model which explicitly accounts for N-inflation in the compositional counts leading to consistency in the inferences produced for model parameters.

Whilst Salter-Townshend (2009) considers data criticism and model validation in a much more substantial manner than Haslett et al. (2006), the focus remains on the inverse stage of the calibration problem. Reference measures, such as the *root mean squared error of prediction* (RMSEP) are utilised with the explicit aim of identifying observations which do not agree well with the fitted model. However, the empirical properties of the resulting RMSEP values are unknown and hence critical bounds by which to detect outliers unavailable. Furthermore, as the focus is on inverse predictive performance of the fitted models, little or no effort is made to evaluate a-priori modelling assumptions at the forward stage; the identification of model weaknesses is difficult due to the lack of suitable model and data criticism tools for inverse

problems (Salter-Townshend 2009). A critical contribution in this thesis is the development of methods for fast model validation through the analysis of posterior random effect terms and the provision of explicit critical bounds by which outliers can be quickly detected, see Chapter 4 for further details. For the first time explicit, objective criticism of both the RS10 model training dataset and the forward models is possible.

A key advance in Salter-Townshend (2009) is the development of a fast leave-one-out cross validation algorithm by which the accuracy of the climate predictions produced by the calibrated model can be verified. For each left-out count the 95% highest posterior density region for climate is constructed and the resulting HPD region evaluated to determine if the true climate location is contained within it. The use of such a demanding statistic for model validation is required in order to effectively test the predictive ability of the model for its later uses in fossil climate reconstruction. However, the predictive accuracy of the final model of Salter-Townshend (2009) is shown to be just 74%. Suggesting uncertainty in the data as one possible cause of the poor model predictive accuracy, ad-hoc measures such as the “Gaussian blurring” of the climate posteriors, are used to increase predictive performance, essentially, posteriors on climate are convolved with a Gaussian kernel of fixed bandwidth which results in more conservative 95% HPD regions. However, Salter-Townshend (2009) readily admits that the appropriateness of this approach is not thoroughly investigated.

Conversely, in this thesis we illustrate that the poor model predictive performance in Salter-Townshend (2009) is due to the failure to fully account for the compositional nature of the pollen counts and the use of models which do not include all important climate covariates, specifically, Huntley (1993) recommend that at least three aspects of climate, namely GDD5, MTCO and AET/PET should be used in climate models for the specific plant taxa which are available in the RS10 pollen training dataset. We extend the current modelling methodology to incorporate an additional climate covariate; this results in a large increase in the “size” of the climate space and greatly impacts on the speed of the deterministic numerical integration algorithms of Salter-Townshend (2009) for model inversion. A further contribution in this thesis is the construction of a fast sampling-based scheme for model inversion, which results in substantial time savings in making inferences on the fossil climates corresponding to fossil pollen data at any given site.

2.4 Advances in this Thesis

Since the “proof of concept” paper of Haslett et al. (2006) and the thesis of Salter-Townshend (2009) the contributions contained in this thesis to the modelling of the RS10 dataset may be summarized as follows.

1. Extension of the existing 2 dimensional climate models to incorporate a further climate covariate, AET/PET (using Section 6.2.3).

2. Extension of the partially nested structures of Salter-Townshend (2009) to the lowest levels (using Section 5.4 and applied in Section 7.3.2).
3. Explicit modelling of the N-inflation in the pollen counts data (using Section 5.5).
4. Development of methods for criticism of the training dataset and the fast validation of the forward models (Section 4.2 and 7.2.5).
5. Construction of a fast, sampling-based scheme for inversion of the forward model for fossil climate prediction (Section 6.3).

Chapter 3

Statistical Methodology

The primary interest of the work detailed in this thesis is palaeoclimate reconstruction. Specifically, statistical models are constructed for the pollen-climate relationship and inferences made on the parameters of these models given model training data. The fitted models then are evaluated in terms of their fit to the data and inverted to make inferences on the unknown climate corresponding to some fossil pollen data.

Each of these tasks involve drawing on a wide range of statistical methods and in this chapter we detail many of these methods. We begin with a brief introduction to Bayesian inference and proceed to introduce details of statistical modelling methodology relevant to the work contained in this thesis.

3.1 Bayesian Inference

The Bayesian analyst, given the observed data, is concerned with learning about some unknown parameters corresponding to the processes which produced the data. These unknown parameters, denoted $X = \{x_1, \dots, x_n\}$, are treated as random variables under the Bayesian framework. One of the main advantages of the Bayesian methodology is that all uncertainties, both uncertainty in model parameters and uncertainty in the data, can be accounted for in a coherent and consistent manner.

The prior distribution, which we denote $\pi(X)$, is a key part of Bayesian inference. Prior knowledge about parameter values can be obtained from any number of sources, including, information derived from previous studies on the subject or based on expert opinion. Alternatively, non-informative priors may be specified, reflecting situations where the analyst cannot provide much background information on the parameters in question. We use the probability density function $\pi(X)$ to express these beliefs.

The observed data, which we denote $Y = \{y_1, \dots, y_n\}$, is modelled using a likelihood

function $\pi(Y|X)$ which is parameterised by the latent parameters X . This represents the probability distribution of the data Y , given X , and is used to calculate the joint probability of observing the data in question as a function of the parameters.

Posterior information regarding the unknown parameters X , subsequent to observing data Y , can be summarized through the use of Bayes' theorem.

$$\pi(X|Y) = \frac{\pi(X)\pi(Y|X)}{\pi(Y)} \quad (3.1)$$

$$\propto \pi(X)\pi(Y|X) \quad (3.2)$$

$$\propto \text{prior} \times \text{likelihood} \quad (3.3)$$

The posterior is proportional to the product of the prior and the likelihood and reflects all that is known about X in light of the observed data. Bayes's theorem provides a mechanism for combining both these sources of information.

3.1.1 Bayesian Hierarchical Models

The framework for many of the models used in this thesis is that of a Bayesian hierarchical model (Gelman et al. 2003, Chapter 2). The use of Bayesian hierarchical models allows the latent parameters X , to be dependent on further *hyperparameters*, θ as follows:

$$Y \sim \pi(Y|X) \quad (3.4)$$

$$X \sim \pi(X|\theta) \quad (3.5)$$

$$\theta \sim \pi(\theta) \quad (3.6)$$

The observable outcomes Y are modelled conditionally on the latent parameters X which themselves are specified in terms of the hyperparameters θ . The hyperparameters themselves may be incorporated into the hierarchical structure, modelled as random variables and estimated from the data, by the addition of extra levels to the model hierarchy.

3.2 Markov Chain Monte Carlo

Any features of the posterior distribution are legitimate for Bayesian inference (Gilks et al. 1996), these include the calculation of posterior moments, marginal densities or highest posterior density regions, for example. All of these integration based summaries can be expressed in terms of posterior expectations of functions of X by drawing samples $\{X^{(i)}, i = 1, \dots, n\}$

from $\pi(X|Y)$ and approximating:

$$E[f(X)|Y] = \frac{\int f(X) \pi(X) \pi(Y|X) dX}{\int \pi(X) \pi(Y|X) dX} \quad (3.7)$$

$$\approx \frac{1}{N} \sum_{i=1}^N f(X^{(i)}) \quad (3.8)$$

However, in many circumstances, posterior distributions are not known in closed form, generally due to the impossibility of analytically evaluating normalising constants. As a result, direct sampling from the posterior distribution is not possible and independent samples are unavailable. A solution to this problem is provided by the use of Markov chains which provide a method of generating dependent samples from the posterior distribution of interest. These samples can be used for Monte Carlo integration purposes; this is then Markov chain Monte Carlo. Markov chain Monte Carlo (MCMC) methods provide algorithms for the drawing of (correlated) samples from highly complex or multi-dimensional distributions from which direct sampling is impossible. Two such algorithms include the Metropolis-Hastings algorithm and the Gibbs sampler. A very simple summary of these methods is provided in the following, for a more comprehensive introduction see Gilks et al. (1996).

3.2.1 The Metropolis-Hastings Algorithm

Metropolis-Hastings algorithms (introduced in Metropolis et al. (1953) and generalised in Hastings (1970)) are an important class of MCMC sampling algorithm which generate a Markov chain of samples from any target probability distribution of interest, such as the posterior distribution in Equation 3.7. The main advantage of the algorithm is that it circumvents the problem of calculating the normalizing constant of the target distribution, thus the distribution need only be known up to a proportionality constant.

The basic idea of the algorithm is as follows; suppose the target distribution from which we wish to sample is $\pi(X)$. Given the current state X^t , the algorithm begins by first drawing a candidate state, X' , from a proposal density $q(\cdot|X^t)$. Generation of the candidate state X' is only dependent on the previous state X^t and this candidate is accepted with probability $\alpha(X', X^t)$ where:

$$\alpha(X^t, X') = \min \left(1, \frac{\pi(X')q(X^t|X')}{\pi(X^t)q(X'|X^t)} \right) \quad (3.9)$$

If the candidate state is accepted, the next value in the Markov chain, X^{t+1} , is set as X' , otherwise X^{t+1} remains in the same state and is set as X^t . In order to generate samples, the

algorithm is allowed to iterate from an initial starting state until convergence is achieved (i.e. until the samples increasingly appear to be dependent samples from the stationary distribution $\pi(X)$), from this point on the algorithm is run until the required number of samples are obtained. Though the set of samples from the Markov chain, $X^{(1)}, \dots, X^{(N)}$ are generally dependent, the autocorrelation between the samples can be studied to obtain a subset of the samples that are approximately independent.

It is important to choose the proposal distributions carefully, for while any choice of $q(\cdot|\cdot)$ will yield the correct stationary distribution of the Markov chain (Gilks et al. 1996), the mixing and convergence properties of the algorithm are dependent on the proposal distribution chosen. If the proposed moves between states are small, the probability of acceptance of a candidate value will be relatively high and consequently the chain will take a long time to explore the target distribution. Conversely, if the proposed moves between states are too large, the acceptance rate will be quite low and the chain will fail to move, greatly reducing the number of effective samples available for inference. In both these extreme cases the algorithm can be said to “mix slowly”, indicating that the chain moves slowly around the support of the target distribution (Gilks et al. 1996).

An additional issue is deciding when convergence of the Markov chain has occurred. The random walk can remain for many iterations in a region that has been influenced by the starting point of the chain potentially leading to the misleading conclusion that the chain has converged. A general solution to this problem, as noted by Gilks et al. (1996), is the running of multiple parallel chains, each with different initial starting points. Though increasing the computational burden, this provides a practical measure for determining if convergence has occurred.

3.2.2 Gibbs Sampling

Gibbs sampling (see Gelfand & Smith (1990)) is another MCMC technique which occurs as a special case of the Metropolis-Hastings algorithm. Suppose $X = \{x_1, \dots, x_n\}$ and $\pi(x_1, \dots, x_n)$ is the target joint probability distribution of interest. The idea behind Gibbs sampling is that we can set up a Markov chain simulation algorithm from the joint posterior distribution by successfully simulating individual parameters from the set of n conditional distributions. Using the Gibbs sampler, parameters are updated component wise using the proposal distribution:

$$q_i(x'_i|x_i^t, X_{-i}^t) = \pi(x'_i|X_{-i}^t) \quad (3.10)$$

This proposal is Markovian as x'_i depends solely on X_{-i}^t . Considering this proposal distribution in the context of the Metropolis-Hastings algorithm, we note that α in Equation 3.9 always

equals unity; x_i^{t+1} is *always* set equal to x_i^t . As a result, accept/reject steps are unnecessary; each $x_i^{(t+1)}$ is effectively a sample point from the marginal distribution $\pi(x_i)$. The algorithm thus proceeds by iteratively sampling from the full conditionals in a random sequence until the required number of samples has been obtained.

However, despite the absence of an accept/reject step, Gibbs sampling can also suffer from mixing and convergence issues. *Single-site* updating can be highly disadvantageous if parameters are highly dependent in the posterior $\pi(X)$ (Rue & Held 2005), for example in problems involving spatial regression models. Furthermore, if n is high dimensional, issues may arise in the speed at which the algorithm explores the full target distribution; it can be difficult to determine when the algorithm has converged. These issues plagued the forward modelling stage in Haslett et al. (2006); due to the number of random variables introduced by the non-parametric modelling approach, MCMC chains had to be run for a number of weeks to provide samples from the posterior and convergence was difficult to assess.

3.2.3 Empirical Bayes

In Section 3.1.1 we introduced the concept of a Bayesian hierarchical model. Given the hierarchical model framework, the data Y depend on some unknown parameters X , which are further dependent on some hyperparameters θ . At the lowest level of the model, the hyperparameters are drawn from an appropriate second stage prior, however, at some stage in the model framework the remaining parameters must be treated as known. A method of avoiding this assumption is the use of *empirical Bayes* methods (Carlin & Louis 2000). In an empirical Bayes analysis, the data are used to estimate the prior parameters and we then proceed as if these parameters are known. The empirical Bayes approach essentially involves the replacement of the integration step in Equation 3.7 by a maximization, using the maximized values of θ to make inferences on unknown parameters at higher stages of the hierarchical model.

$$\hat{\theta}(Y) = \max_{\theta} \left[\frac{\pi(Y, X, \theta)}{\pi(X|\theta, Y)\pi(Y)} \right] \quad (3.11)$$

The value of θ used in parameter inference, $\hat{\theta}$ is obtained by maximizing the marginal posterior distribution $\pi(\theta|Y)$ over θ . Inference on the remaining model parameters is thus based on the estimated posterior distribution $\pi(X|Y, \hat{\theta})$:

$$\pi(X|Y) = \frac{\int \pi(Y|X, \theta)\pi(X|\theta)\pi(\theta)d\theta}{\int \int \pi(Y|X, \theta)\pi(X|\theta)\pi(\theta)dX d\theta} \quad (3.12)$$

$$\approx \frac{\pi(Y|X, \hat{\theta})\pi(X|\hat{\theta})}{\int \pi(Y|X, \hat{\theta})\pi(X|\hat{\theta})dX} \quad (3.13)$$

The drawback of such an approach, as discussed in Gelman et al. (2003), is that the resulting predictions do not account for all parameter uncertainty and thus posterior density regions will generally be less conservative than those obtained by a fully Bayesian approach. However, if the posterior distribution of $\pi(\theta|Y)$ is reasonably “peaked” the approximation can be quite accurate.

3.3 Integrated Nested Laplace Approximations

In Section 3.2 we discussed how Markov Chain Monte Carlo methods can be used to obtain simulation based summary statistics from nearly any target posterior distribution of interest. In this section we consider an alternative to MCMC for fully Bayesian inference on model parameters, the deterministic, integrated nested Laplace approximation (INLA) method of Rue et al. (2009).

To recap, interest lies in making inference on a number of unknown parameters (X, θ) corresponding to a target posterior distribution of interest, such as that considered in Equation 3.15.

$$\pi(X, \theta|Y) = \pi(X|Y, \theta)\pi(\theta|Y) \tag{3.14}$$

$$= \frac{\pi(Y|X, \theta)\pi(X, \theta)}{\int_{\theta} (\int_X \pi(Y|X, \theta)\pi(X, \theta)dX)d\theta} \tag{3.15}$$

The INLA algorithm proceeds by using a Laplace approximation for $\pi(\theta|Y)$; essentially this involves replacing the denominator in Equation 3.16 by the Gaussian approximation $\tilde{\pi}_G(X|Y, \theta)$ to the full conditional $\pi(X|Y, \theta)$, evaluated at the modal value of X , $X^*(\theta)$. $\tilde{\pi}_G(X|Y, \theta)$ is thus approximated as a multivariate Gaussian distribution with mean $\mu(\theta)$ and covariance matrix $\Sigma(\theta)$ (i.e. $MVN(\mu(\theta), \Sigma(\theta))$), with (θ) indicating the explicit dependence of the approximation on θ . If additional Markov structure is specified for X , X is a Gaussian Markov Random Field (Section 3.4.1) and the Markov structure is carried through to $\tilde{\pi}_G(X|Y, \theta)$.

$$\pi(\theta|Y) = \frac{\pi(Y, X, \theta)}{\pi(X|Y, \theta)} \tag{3.16}$$

$$\approx \frac{\pi(Y, X, \theta)}{\tilde{\pi}_G(X|Y, \theta)} \Big|_{X=X^*(\theta)} \tag{3.17}$$

In order to make marginal statements about the latent field posterior X , the posterior for the model hyperparameters, $\pi(\theta|Y)$, is evaluated on a discrete grid; the approximate

marginal posterior for X can then be obtained by summing over the discrete values of the hyperparameters. In the following the Δ_i represent area weights which ensure the probability density of $\tilde{\pi}(\theta|Y)$ sums to 1 (see Equation 3.18 - 3.20).

$$\pi(X|Y) = \int \pi(X|\theta, Y)\pi(\theta|Y)d\theta \quad (3.18)$$

$$\approx \int \tilde{\pi}(X|\theta, Y)\tilde{\pi}(\theta|Y)d\theta \quad (3.19)$$

$$= \sum_i \tilde{\pi}(X|\theta_i, Y)\tilde{\pi}(\theta_i|Y) \times \Delta_i \quad (3.20)$$

Thus the posterior for each $x_i \in X$ is represented as a weighed mixture of Gaussians. If the marginal posterior of an individual x_i is required to a greater degree of accuracy, a Laplace approximation can be used in a similar manner to Equation 3.17. For further details we defer to Rue et al. (2009).

An important point to note however, is that the INLA methodology, much like methods based on Monte Carlo sampling techniques, is not without its issues. The numerical algorithms, used to explore the space of θ , require initial starting positions and are thus subject to “getting stuck” in local modes. Additionally, the posterior for model hyperparameters, $\pi(\theta|Y)$, is represented on a discrete grid - as a result, models are limited in the number of hyperparameters that can be considered (Rue et al. 2009) suggest $\dim(\theta) < 6$), providing an explicit constraint on the complexity of models that can be constructed when using the INLA algorithm.

3.4 Spatial Prior Models

In this section we focus on prior models for X where we assume X is a smooth spatial process indexed by some location vector, details of which are suppressed in the following. The use of multivariate Gaussian spatial priors for X is very common in point referenced spatial regression problems such as those involving environmental data (Bannerjee et al. 2004). X is defined as a multivariate Gaussian process with mean vector μ and $n \times n$ covariance matrix $\Sigma(\theta)$, where the individual elements in $\Sigma(\theta)$ describe the spatial covariance between each of the latent x_i . The degree of covariance between the x_i is governed by underlying model hyperparameters θ , which parameterise $\Sigma(\theta)$. In the following, for simplicity of notation, we suppress the explicit θ dependence; the spatial prior for X may thus be written as $X \sim MVN(\mu, \Sigma)$.

In a typical analysis the dimension of the covariance matrix is directly related to the number of recorded observations - we may be interested in inferring a smooth latent x_i for each observation y_i . However, as the dimensionality of the available dataset increases, such prior specifications become too computationally demanding to work with. This is due to the

$\mathcal{O}(n^3)$ cost of inverting large dense $n \times n$ covariance matrices Σ in order to evaluate probability densities. Bannerjee et al. (2004) refer to this as “the big n problem”.

In the following we discuss one possible solution for surmounting this difficulty. We consider the setting where additional Markov structure is applied to the precision (inverse covariance) matrix Q , leading to the harnessing of fast, sparse matrix algorithms to obtain vast computational savings.

3.4.1 Gaussian Markov Random Fields

Suppose as previously that X has a multivariate Gaussian distribution with mean μ and precision matrix Q . If we define a labeled graph $G = (V, \epsilon)$, where $V = (1, \dots, n)$ and ϵ be such that there is no edge between node i and j iff $x_i \perp x_j | X_{-ij}$. Then we say X is a GMRF wrt G .

Definition 3.4.1 *A random vector $X = (x_1, \dots, x_n)^T$ is called a **GMRF** wrt a labeled graph $G = (V, \epsilon)$ with mean μ and positive definite precision matrix Q , iff its density has the form:*

$$\pi(X) = (2\pi)^{-\frac{n}{2}} |Q|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(X - \mu)^T Q (X - \mu)\right) \quad (3.21)$$

and

$$Q_{ij} \neq 0 \iff \{i, j\} \in \epsilon \ \forall \ i \neq j.$$

A Gaussian Markov random field may thus be simply described as a random vector following a multivariate Gaussian distribution with additional Markov properties. The Markov properties are contained in the precision matrix Q ; if x_i and x_j are conditionally independent given the “rest” (X_{-ij}), then the corresponding Q_{ij} entry in the precision matrix Q is zero. Given a defined graph structure, many of the entries in Q are zero; the harnessing of computationally efficient algorithms for operations on sparse matrices will thus significantly reduce computation time (see Rue & Held (2005)).

GMRF structures can be quite natural, such as situations where the data are defined on a lattice as in image analysis problems, however most approaches based on GMRFs approximate continuous space by a discrete grid with observations “pushed” to the nearest gridpoint. This is not strictly necessary, Lindgren et al. (2011) have developed methods to extend this theory to the continuous plane, details of which are omitted here. With regard to the “big n problem”, the dimensionality of the latent process reduces to the dimensionality of the grid; this can result in vast computational savings in addition to those available from the sparse structure implied by the Markov properties.

Intrinsic Gaussian Markov Random Fields

In many applications, including those considered in this thesis, intrinsic Gaussian Markov random fields (IGMRFs) are of particular interest. Intrinsic GMRFs are always improper since the precision matrix is not of full rank and therefore cannot be inverted to give the covariance matrix. Intrinsic GMRFs have extensive use as a prior on the smoothness of the latent surfaces in spatial regression models, for their application in a variety of statistical problems see Rue & Held (2005, Chapters 4 & 5).

Definition 3.4.2 *Let Q be an $n \times n$ symmetric, positive semi-definite matrix with rank $n - k > 0$. Then $X = (x_1, \dots, x_n)^T$ is an **Improper GMRF** of rank $n - k$ with parameters (μ, Q) , if its density is*

$$\pi(X) = (2\pi)^{-\frac{n-k}{2}} (|Q|^*)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(X - \mu)^T Q (X - \mu)\right) \quad (3.22)$$

$|Q|^*$ denotes the generalised determinant which is formed as the product of the $n - k$ non-zero eigenvalues. The parameters (μ, Q) no longer represent the mean and the precision since they do not formally exist. Following Rue & Held (2005) we continue to adopt this terminology for convenience. The expected value of the individual x_i , μ_i is undefined, however, the conditional mean $E(x_i | X_{-i})$ can be expressed as a weighed average of the neighbouring x_j where the neighbourhood structure of each x_i is defined by G .

The precision matrix, Q for an intrinsic GMRF in one dimension is quite easily constructed from first principles using a simple random walk. Given an equally spaced grid of length n and assuming independence of the increments, the precision matrix corresponding to an IGMRF of order one on the line can be easily obtained by specifying a Gaussian prior with mean zero for the pairwise forward differences, i.e. $x_{i+1} - x_i \sim N(0, \kappa^{-1})$.

$$\pi(X|\kappa) \propto \kappa^{\frac{n-1}{2}} \exp\left(-\frac{\kappa}{2} \sum_{i=1}^{n-1} (\Delta x_i)^2\right) \quad (3.23)$$

$$= \kappa^{\frac{n-1}{2}} \exp\left(-\frac{\kappa}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2\right) \quad (3.24)$$

$$= \kappa^{\frac{n-1}{2}} \exp\left(-\frac{1}{2} X^T Q X\right) \quad (3.25)$$

We set $Q = \kappa R$ where κ is a hyperparameter governing the smoothness of the latent process. The higher the value of the precision parameter the smoother the resulting spatial process will be. Additionally, R is the structure matrix given by:

$$X \sim \text{GMRF}(\mu, Q) \quad (3.29)$$

Using this transformation, the unconstrained $(-\infty, +\infty)$ latent field variables are transformed into positive random numbers $[0, +\infty)$ which are compatible with the Poisson likelihood model. As per the preceding section, the latent field is modelled non-parametrically as a GMRF. In the context of model parameters that are constrained to lie on the unit interval $[0, 1]$, such as probability values, the logit transform, $p_i = \frac{\exp(x_i)}{1 + \exp(x_i)}$, provides the appropriate transformation.

3.5.2 Spatial Zero-Inflated Models

A common feature of ecological datasets, such as the RS10 pollen dataset, is their tendency to contain many zero values. If standard statistical models such as the Poisson are used to model such data, the excess of zero observations can greatly impact parameter estimation; statistical inferences derived from the model will be biased by them (Salter-Townshend 2009).

Ridout et al. (1998) discuss a number of solutions to this problem; here we focus on one of these in particular, the use of *zero-modified distributions* to account for the excess zeroes. Such models assume that the observed data arise from a process which consists of two distinct states; a zero state from which only zero counts are observed and a Poisson state from which the non-zero counts and some of the zero counts are observed (Hall 2000). As per Ridout et al. (1998), zero counts arising from the zero state are referred to as “structural” zeroes whereas those arising from the Poisson process are referred to as “sampling” zeroes.

The resulting zero-inflated Poisson likelihood for the data is:

$$\pi(y_i | \lambda_i, q_i) = \begin{cases} (1 - q_i) + q_i \text{Poisson}(0; \lambda_i) & y = 0 \\ q_i \text{Poisson}(y_i; \lambda_i) & y > 0 \end{cases} \quad (3.30)$$

In Equation 3.30, q_i represents the probability that an observed count arise from a Poisson model with rate parameter λ_i . Alternatively, in the context of a zero observation, $(1 - q_i)$ represents the probability that the count arises from a state that produces only zeroes. Where the response is spatially referenced, both the rate parameters of the Poisson model and the zero-inflation probabilities of the zero state may be separately modelled as a function of underlying model covariates. A Bayesian hierarchical model for such data may be presented as:

$$y_i \sim \text{ZIPoisson}(y_i; \lambda_i, q_i) \quad (3.31)$$

$$\lambda_i = \exp(x_i) \quad (3.32)$$

$$q_i = \text{logit}(z_i) \quad (3.33)$$

$$X \sim \text{GMRF}(\mu_x, Q_x) \quad (3.34)$$

$$Z \sim \text{GMRF}(\mu_z, Q_z) \quad (3.35)$$

The zero-inflated distribution has an additional spatial field in comparison with the non-zero-inflated Poisson model - the additional field Z , through the use of a logit transform, is used to control the probabilities which govern the point mass at zero. As a result, the number of latent parameters requiring inference at the forward stage is doubled for the zero-inflated model.

Of course alternative models to the Poisson may be used to model the observed counts. Other probability distributions for count data, such as the Negative Binomial, can be modified to account for excess zeroes in a similar manner. However, an important point to note is that these methods are not compatible with models for compositional data and can lead to statistically inconsistent, erroneous parameter estimates. We present and develop a solution to this issue in Section 5.5.

Single Process Models for Zero-Inflation

Of particular interest in this thesis are *single process models* for zero-inflation, where the latent spatial field governing the zero-inflation probabilities, Z , is assumed to be a function of the latent spatial field governing the rate parameters of the mean, X . Parsimonious models may then be developed which greatly reduce the computational burden of model fitting - Salter-Townshend & Haslett (2006) detail the use of such models, in situations where the latent parameters are modelled non-parametrically, reduces the number of latent parameters requiring inference by half. Essentially, the probability of presence of a given count observation is modelled as:

$$q_i = \left(\frac{\exp(x_i)}{1 + \exp(x_i)} \right)^\alpha \quad (3.36)$$

The probability of presence, for a given count observation, is treated as a monotonic function of the underlying latent field for positive α ; specifically, as the value of the underlying latent random variable x_i increases in value, the probability of observing a zero observation decreases. Additionally, if $\alpha = 0$, the zero-inflated model simply reduces to its non-zero inflated version.

A point to note is that this model should only be applied to data for which the assumption of such a relationship can be justified. For examples of applications of this model in a general context see Lambert (1992), who considers zero-inflation of manufacturing defects, or Hall

(2000) who considers the use of such models in an ecological setting. Salter-Townshend (2009) provides justification for the application of this model in the context of the palaeoclimate reconstruction problem relevant to this thesis.

3.5.3 Overdispersion Models

Counts data which display excess variation over that expected by a given statistical model are said to be *overdispersed*. For example, overdispersion in the context of Poisson count observations may be identified by the empirical variance being substantially greater than the empirical mean; Poisson models with a single rate parameter governing both the mean and the variance will perform poorly in such settings.

The solution to this problem is the use of overdispersed likelihoods; Gaussian random effect terms may be incorporated into models to explicitly model the additional source of variation as follows:

$$y_i \sim \text{Poisson}(y_i; \lambda_i) \quad (3.37)$$

$$\lambda_i = \exp(x_i + u_i) \quad (3.38)$$

$$u_i \sim N(0, \sigma^2) \quad (3.39)$$

$$X \sim \text{GMRF}(\mu, Q) \quad (3.40)$$

The addition of the random effect terms results in a model which is overdispersed with regard to the spatial component X . However, as the Gaussian random effects are not conjugate to the Poisson likelihood, this results in at least a doubling of the number of latent parameters in the model - a random effect term must be inferred for each datum as well as an additional hyperparameter σ^2 . However, as we will discuss in detail in Section 4.2.3, the use of the INLA algorithm of Rue et al. (2009) provides a solution to this problem - the computational speed of the algorithm facilitates quick approximate inference on all model parameters including the random effect terms.

Alternatively, the assumption that the random effects arise from a distribution conjugate to the Poisson likelihood, such as the Gamma distribution, leads to a modelling situation which is convenient in terms of computation. As per Salter-Townshend (2009) the model may be written as:

$$y_i \sim \text{Poisson}(y_i; \lambda_i) \quad (3.41)$$

$$\lambda_i = \text{Gamma}(\lambda_i, \delta, (1 - p_i)/p_i) \quad (3.42)$$

$$p_i = \frac{\delta}{\delta + e^{x_i}} \quad (3.43)$$

$$X \sim \text{GMRF}(\mu, Q) \quad (3.44)$$

As the Gamma distribution of the λ_i 's is conjugate to the Poisson likelihood, the above model can be simplified by analytically integrating out the λ_i 's (see Salter-Townshend (2009) for further details) to obtain:

$$\pi(y_i|x_i) = \Gamma \frac{(\delta + y_i)}{y_i! \Gamma(\delta)} p_i^\delta (1 - p_i)^{y_i} \quad (3.45)$$

This is the *Negative-Binomial* model which carries a single extra parameter over the simple Poisson model, namely an overdispersion parameter, δ , which governs the degree of overdispersion observed. As compared to the model with Gaussian random effects in Equation 3.37 - 3.40, inference on the parameters of the Negative Binomial model is computationally less intensive by virtue of the substantially reduced inference task due to the “integration out” of the Gamma overdispersion.

One final important caveat, as noted by Ridout et al. (1998), is that the source of overdispersion must be carefully analysed - overdispersion due to an excess of zeroes in the observational dataset is most suitably modelled with a zero-modified distribution. The use of overdispersed models in this setting will perform poorly in an obvious manner, resulting in an underestimation of the mean and an over-inflation of the variance. In Table 7.3 we observe this result in the context of the palaeoclimate reconstruction problem, where the lack of explicit models for the zero/N-inflation of pollen counts leads to a substantially overestimated overdispersion parameter and a deterioration in model performance in terms of prediction.

3.6 Inverse Problems

The main class of problems considered in this thesis are statistical calibration or *inverse inference* problems. Such problems can be decomposed into two distinct stages; at an initial model fitting or “forward stage”, training data is used to calibrate models for the relationship between some observed response Y and the recorded covariate(s) C . At the “inverse stage”, the calibrated models are used *inversely* to make inference on the unknown covariate(s), corresponding to new observations, for which such information is unknown.

A typical example of such a problem is provided by the radiocarbon dating problem, see Buck et al. (2006). Specifically, at the forward stage the “calibration curve”, which describes the relationship between calendar age and radiocarbon age, is estimated from a set of high precision calibration data. At the inverse stage, the calendar age of a fossil artifact may

then be determined, with quantifiable uncertainty, given the calibrated model and an estimate of the radiocarbon age of the fossil.

The radiocarbon dating problem provides a simple example of a *univariate inverse inference* problem; both the data Y and calendar age C are univariate - Y represents the set of radiocarbon ages and C the corresponding measurements of calendar age. As such, this particular problem provides a useful vehicle for displaying subtle details of inverse inference problems. In the following section we present a simple, illustrative toy example.

3.6.1 Toy Example

We create a simple toy example which is similar in spirit to the radiocarbon dating problem introduced above. Ten counts, $Y' = (y'_1, \dots, y'_{10})$, are observed at random discrete spatial locations, $C' = (c'_1, \dots, c'_{10})$, on an equally spaced grid, C of length 100; this is the model training dataset. Subsequent to their recording, an additional observation, y_{new} , is discovered for which the associated spatial location c_{new} is unknown. The broad idea of the problem is visually presented in Figure 3.1.

Interest ultimately lies in making inferences on the unknown spatial location of the new count observation. The first step is to calibrate a model for the response-covariate relationship given the model training data; this is the forward stage.

Forward Stage

At the forward stage, we construct a model for the response-covariate relationship. Each observation $y'_i \in Y'$, is assumed to be an indirect observation of the latent response surface X at location $c'_i \in C'$. The problem effectively corresponds to making inferences about an unknown, smooth function X , some of whose values are observed with error.

As the nature (shape) of the response surface is not known, the most appropriate course of action is to assume a non-parametric form; the only requirement is that the response surface is a smooth function over the location space C . We assume the observed data are conditionally Gaussian distributed with known noise parameter σ^2 . A simple model for the data can be written as follows:

$$y'_i = X(c'_i) + \epsilon_i \tag{3.46}$$

$$X \sim \text{GMRF}(Q) \tag{3.47}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{3.48}$$

The data observations are taken to be conditionally independent given X , where X is a

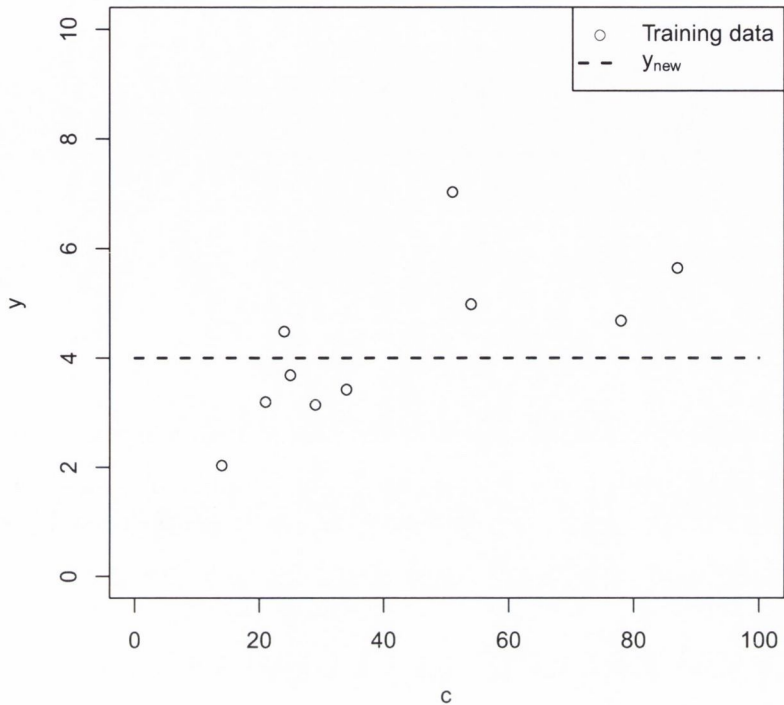


Figure 3.1: A simple example of an inverse inference problem. Model training data, consisting of ten observational counts along with associated known spatial location is recorded. Interest ultimately lies in making inference on the unknown spatial location of a new data count, y_{new} .

smooth spatial process defined at each of the 100 spatial locations. The model in Equation 3.47 can be rewritten as $Y' = AX + \epsilon$ where A is a 10×100 projection matrix which projects X , defined in 100 dimensional space, to the ten dimensional space of Y' . The projection matrix A simply consists of 0's and 1's as follows:

$$A[i, c_i] = \begin{cases} 1 & i = 1 : 10 \\ 0 & \text{otherwise} \end{cases} \quad (3.49)$$

For example, the first observation y'_1 is observed at location $c'_1 = 14$. Consequently the 1st row of A is a row of zeroes with the only non-zero entry at location 14; this entry has value 1. The expression of the model in this format allows us to define a prior for X which is defined across the 100 spatial locations of C . We specify an IGMRF prior of order 2 for X . This prior model is fully parameterised given the precision matrix Q_x for which the precision, or

smoothness parameter κ , is assumed known. See Lindgren & Rue (2008) for details on the structure of Q_x . In the following, Q_y is the diagonal inverse covariance matrix of the observed data with values σ^{-2} along the diagonal and zeroes elsewhere. For simplicity the value of σ^2 is assumed known.

The posterior distribution of X is proportional to a product of the likelihood times the prior:

$$\pi(X|Y', C', C) \propto \pi(Y'|X, C')\pi(X|C) \quad (3.50)$$

$$\propto \exp\left(-\frac{1}{2}(Y' - AX)^T Q_y (Y' - AX)\right) \exp\left(-\frac{1}{2}X^T Q_x X\right) \quad (3.51)$$

$$\propto \exp\left(-\frac{1}{2}(X - \mu)^T Q (X - \mu)\right) \quad (3.52)$$

As the GMRF prior for X is conjugate to the Gaussian likelihood for the data, the posterior distribution for X , given the fixed model hyperparameters is known exactly. Through some simple matrix manipulations, the posterior for X can be recognised as multivariate normal with mean parameters μ and posterior precision matrix Q where:

$$\mu = (Q_x + A^T Q_y A)^{-1} A^T Q_y Y' \quad (3.53)$$

$$Q = (Q_x + A^T Q_y A) \quad (3.54)$$

The posterior mean of X along with 95% HPD regions is presented in Figure 3.2.

Given the calibrated model, our interest turns to making inferences on the unknown spatial location, c_{new} , corresponding to the newly observed y_{new} . This is the inverse stage.

Inverse Stage

The process for *inverting* the calibrated model, $\pi(X|Y', C')$, to make inferences on c_{new} given y_{new} may be formalized as follows:

$$\pi(c_{\text{new}}|y_{\text{new}}, Y', C') = \int \pi(c_{\text{new}}, X|y_{\text{new}}, Y', C') dX \quad (3.55)$$

$$= \int \pi(c_{\text{new}}|X, y_{\text{new}}, Y', C') \pi(X|y_{\text{new}}, Y', C') dX \quad (3.56)$$

$$\approx \int \pi(c_{\text{new}}|X, y_{\text{new}}) \pi(X|y_{\text{new}}, Y', C') dX \quad (3.57)$$

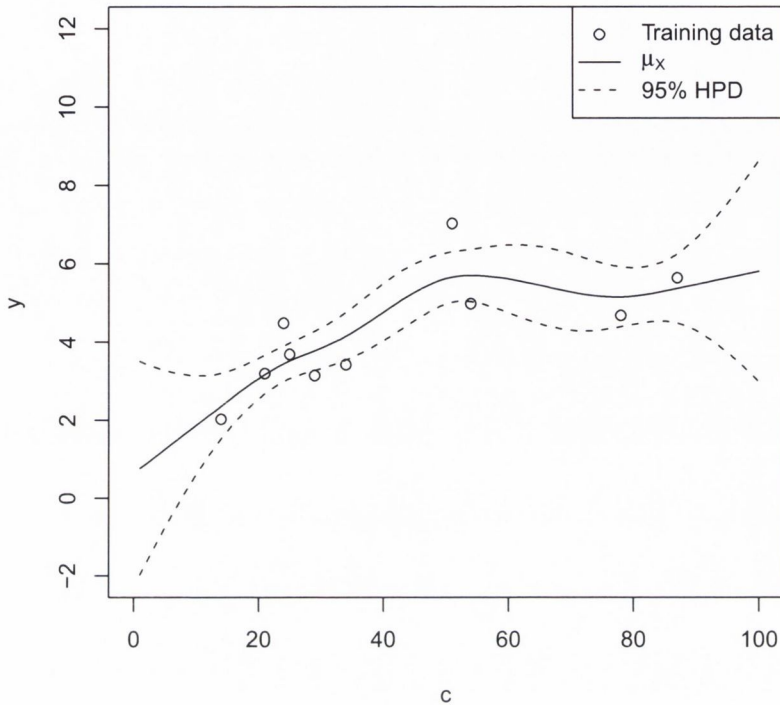


Figure 3.2: Posterior mean, μ , of the smooth response surface along with 95% HPD regions. μ_x is a smooth spatial function over C .

$$\propto \int \pi(y_{\text{new}}|c_{\text{new}}, X)\pi(c_{\text{new}})\pi(X|y_{\text{new}}, Y', C')dX \quad (3.58)$$

$$= K \int \pi(y_{\text{new}}|c_{\text{new}}, X)\pi(c_{\text{new}})\pi(X|Y', C')dX \quad (3.59)$$

Here K is a normalising constant which ensures Equation 3.59 sums to unity. As typically very little information is known *a priori* about the value of c_{new} , a uniform prior is placed on each of the spatial locations which comprise C . The inverse stage of the inverse problem can provide a number of computational difficulties. The “integration out” of X can be computationally taxing when the likelihood in Equation 3.59 is not conjugate to the posterior for the latent surface $\pi(X|Y', C')$. Fortunately, in the simple example presented here, they are conjugate and thus the integral in Equation 3.59 can be computed analytically.

In order to present a suitable posterior, the normalising constant, K , in Equation 3.59 is also of concern. This problem is addressed by discretizing the location space to a finite number of spatial locations. Thus K can be obtained by calculating the probability density at each grid location and dividing through by the sum of their values to provide a normalised posterior

which sums to 1. This posterior is shown in Figure 3.3.

The simplicity of the example presented here detracts from important details of the problem. It is quite evident that the information gleaned from the model is directly related to the shape of the response surface; if the response surface is “too” smooth then, in certain situations, the inferences derived from the model will be uninformative - for example, producing prediction regions which are uniform across the location space. Conversely, response surfaces which are not smooth or “wiggly in nature, may produce posteriors that are multimodal and uninformative in a completely contrasting manner. We refer the interested reader to Salter-Townshend (2009) (page 37) or Buck et al. (2006) for additional details.

One last important point to note, is that the problems considered in this thesis are not univariate, as in the above example, but *multivariate*; in the context of the motivating palaeoclimate reconstruction problem there are several response surfaces each of which have their own observed counts data. The simplest manner of addressing the multivariate inverse problem, is to treat each set of counts as independent for both the forward and inverse stages. Inference can then be performed on each response surface separately. The inverse predictive distribution given all the calibrated models is then the product of the predictive distributions for each of the individual models. This decomposition of multivariate models into the product of separate univariate models provides many computational conveniences at both the forward and inverse stages, and is discussed in great detail in Chapter 5.

3.7 Model Validation for Inverse Problems

In this thesis we primarily consider problems of inverse inference - in the context of the motivating palaeoclimate application, model training data is used to calibrate models for the interaction between climate (covariate) and pollen (response); the calibrated models are then used inversely, to make inferences on the unobserved climates corresponding to sets of fossil pollen counts. As the ultimate objective is to use calibrated models for prediction in an inverse sense; model evaluation metrics should focus on the inverse stage of the problem.

3.7.1 Cross Validation in the Inverse Sense

Leave-one-out cross-validation is a tool used commonly for evaluating model fit to data. However, cross-validation in the context of inverse problems, is subtly different to cross-validation in the forward sense. Specifically, in inverse cross-validation, the ability of the model to predict the known climate location c_i , given y_i and the remainder of the training dataset (Y_{-i}, C_{-i}) is assessed; this contrasts with cross-validation in the forward sense, where the ability of the model to predict y_i given c_i and (Y_{-i}, C_{-i}) is of interest.

As detailed by Bhattacharya & Haslett (2007) and Salter-Townshend (2009), there are

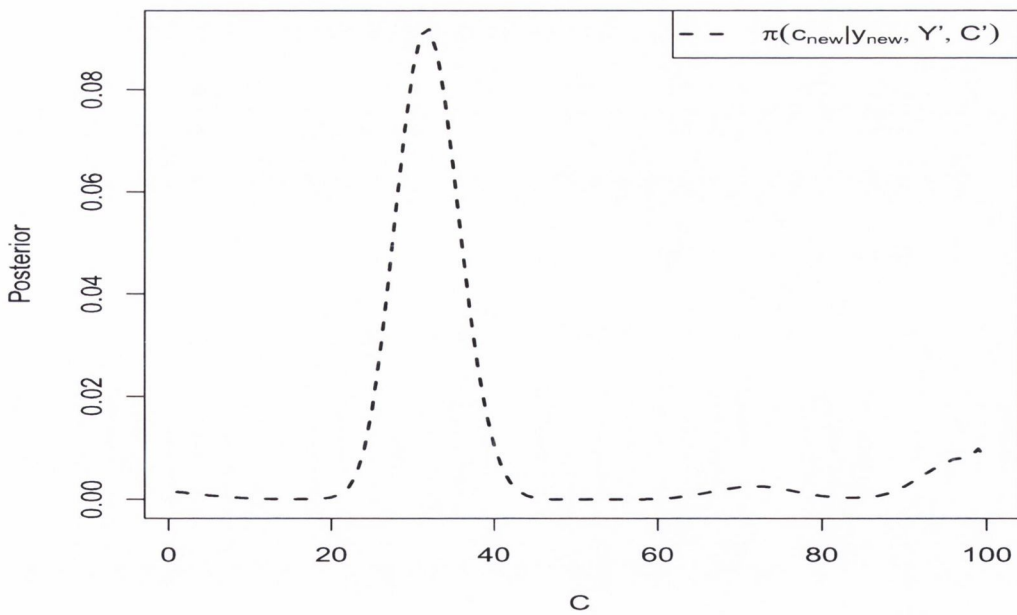
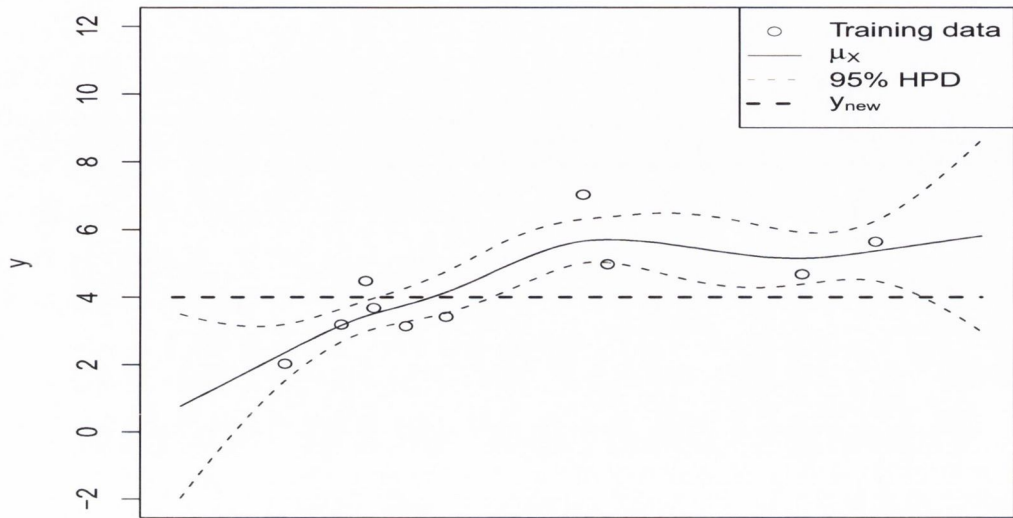


Figure 3.3: Above: Plot of calibrated model with training data also plotted. Below: Posterior distribution for spatial location given the new count datum. The posterior probability on location is observed to be higher for locations where the value of the response crosses the calibrated lines.

several problems involved with use of leave-one-out cross-validation for model validation in the context of inverse problems; given the model calibration dataset (C', Y') , the forward models must be refit given the omission of each set of pairs (c'_i, y'_i) . However, in situations where inference procedures are sampling based, this task can be extremely computationally intensive; in the context of the palaeoclimate reconstruction project, large, computationally complex models must be refit for each left out datum which is infeasible given finite computing time.

Through the use of the Integrated Nested Laplace Approximation (INLA) algorithm for fast approximate parameter inference, Salter-Townshend (2009) was able to greatly reduce this computational burden at the forward stage - fast updates could be made to model posteriors for the left out points and thus leave-one-out cross validation could be achieved both quickly and *exactly*. One of the primary model validation metrics, in an inverse sense, presented in Salter-Townshend (2009), is the percentage of observations lying outside the 95% *highest posterior density (HPD) region*.

Percentage Outside 95% Highest Posterior Density Region

The particular cross-validation statistic that is used extensively throughout this thesis, is the percentage of training data that fall outside the 95% highest posterior density region. The predictive distribution here, is the leave-one-out cross validation posterior predictive distribution for the omitted location given all other locations and count data.

As per Salter-Townshend (2009), we denote by Δ , the percentage of observations whose spatial location is outside the 95% highest posterior density region of their inverse predictive density. Essentially, if the model fits the data then the expected value of Δ across all observations is 5%. The following procedure to obtain the HPD regions is obtained from Salter-Townshend (2009):

1. The HPD region is initialized to contain none of the locations.
2. The discrete location of highest probability mass is selected and added to the HPD region
3. If the total mass of the HPD region is less than 95%, the location of the next highest probability mass is selected and added to the HPD region.
4. Step 3 is repeated until the total probability mass of the HPD region is greater than or equal to 95%.

As noted by Salter-Townshend (2009), this means that the HPD region will contain 95% **or more** of the total probability mass. Therefore the expected value of Δ is $\leq 5\%$. Returning to the inverse posterior for location given a sample fossil count presented in Figure 3.3, the corresponding HPD region on climate is presented in Figure 3.4.

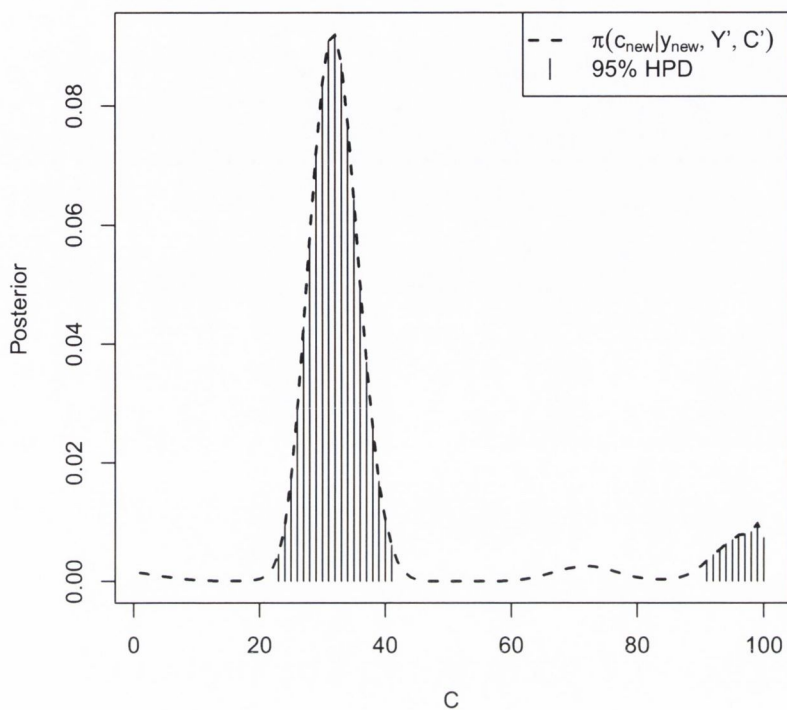


Figure 3.4: 95% highest posterior density for climate location given a new count. The location space C is discretized to a regular grid; the 95% HPD region for location contains 95.12% of the probability mass and the 95% HPD region is seen to contain two disjoint probability regions.

As the ultimate objective is the accurate prediction of ancient climates corresponding to fossil pollen counts, the leave-one-out cross validation measure is perhaps the most relevant metric for comparing and contrasting the performance of different models. However, this metric provides little insight in terms of the underlying predictive performance of each model - cross-validation simply evaluates whether the 95% HPD predictive region for the omitted climate location given all other counts contains the true climate location, but does not give an insight into the accuracy of the placement of the resulting posterior, or the range and multimodality of the prediction regions produced. This prompts the consideration of additional metrics for the evaluation of model performance, one such metric is the *mean squared error of prediction*.

3.7.2 Mean Square Error of Prediction

In the context of palaeoclimate reconstruction, a number of authors have considered the use of the mean squared error of prediction (*MSEP*) as a metric for evaluating model performance, see for example Vasko et al. (2000) or ter Braak (1995). In contrast with the binary nature of cross-validation metrics where predictive regions either contain the true climate location or do not (0 or 1), the *MSEP* provides a measure for evaluating the placement of posterior predictive distributions and thus provides an important metric for the comparison of multiple different models.

As location space is always discretised onto a grid with N_g gridpoints, the *MSEP* is simply obtained as the expectation of the square of the difference between the new location, c_{new} and the predicted location under the posterior for the inverse stage given the new count c_{new} which we denote here as c .

$$MSEP = E [||c - c_{\text{new}}||^2] \quad (3.60)$$

$$= \sum_{k=1}^{N_g} \pi(c_k) ||c_k - c_{\text{new}}||^2 \quad (3.61)$$

Here $||c_k - c_{\text{new}}||^2$ is the squared distance between location c_k on the grid and the true location c_{new} and $\pi(c_k)$ represents the associated posterior predictive mass on climate at location c_k , at the inverse stage, given the new count y_{new} , i.e. $\pi(c_k) = \pi(c_k | y_{\text{new}}, Y', C')$. In certain cases it may be more convenient to work with the *root mean square error of prediction* (*RMSEP*). This is simply obtained as the square root of the *MSEP*, i.e. $RMSEP = \sqrt{MSEP}$.

Since we consider climate models in this thesis which may incorporate differing number of climate covariates, the *MSEP* is always rescaled to lie between 0 and 1; in d dimensional climate space, this is achieved by dividing through the calculated *MSEP* by the the associated d .

However, a notable flaw with the *MSEP* as a model comparison metric is that it does not provide a method of determining whether model performance is poor due to the presence of multiple modes, or due to spurious mislocation of the climate posterior with regard to the true location. For example, in Figure 3.5, two contrasting climate predictions for c_{new} are produced which obtain approximately the same value for the *MSEP*.

The 95% HPD regions of both climate predictions produced in Figure 3.5 contain the true spatial location. However, clearly one of the predictions is to be preferred, that of $\pi_2(c_{\text{new}} | y_{\text{new}}, Y', C')$. Simple statistics such as the distance of the true climate location to the mode are insufficient on their own, but are potentially useful in this situation in providing

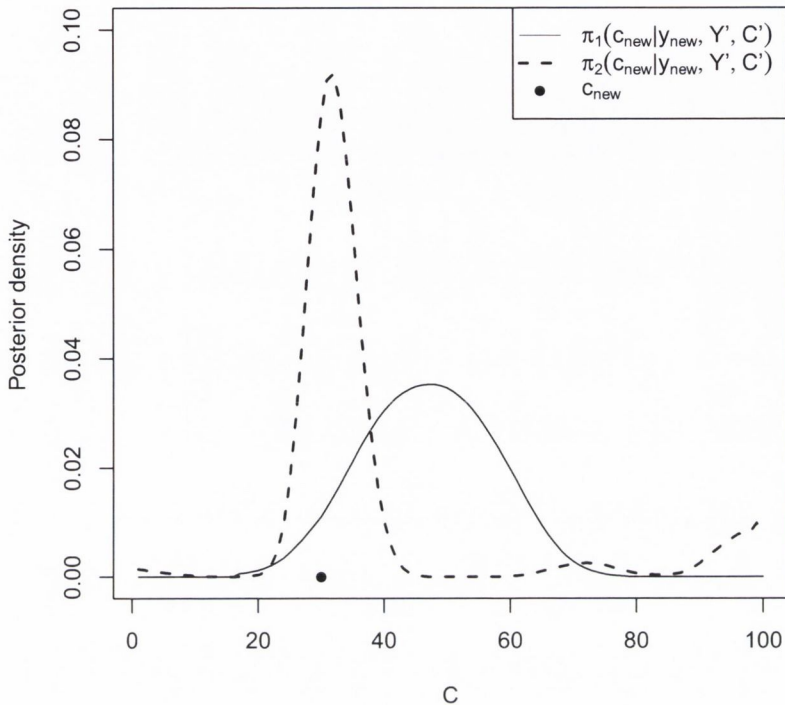


Figure 3.5: Mean squared error of prediction (*MSEP*) for two notional methods to climate reconstruction. The *MSEP* for each approach is 0.040 and 0.039 respectively.

an additional measure to differentiate between the *MSEPs*. In this thesis we denote this metric by $D_{\text{mode}} = \|c_{\text{mode}} - c_{\text{new}}\|$, which represents the absolute distance between the true location c_{new} and c_{mode} , which is the grid location with the highest predictive probability mass for c_{new} given the new count. This metric is additionally scaled by the dimensionality of the space d , in order to provide values of D_{mode} which are defined on $[0, 1]$. For example, in three climate dimensionals this metric is calculated as:

$$D_{\text{mode},i} = \sqrt{\frac{(c_{i1}^{\text{mode}} - c_{i1}^{\text{new}})^2 + (c_{i2}^{\text{mode}} - c_{i2}^{\text{new}})^2 + (c_{i3}^{\text{mode}} - c_{i3}^{\text{new}})^2}{3}} \quad (3.62)$$

where the c_i are scaled to lie between zero and one. This produces a metric defined on the correct spatial dimension. However the model with the “best” *MSEP* may not necessarily be the most with the best predictive accuracy in terms of leave-one-out cross-validation, as we observe in the following chapters.

Chapter 4

Bayesian Residual Analysis for some Non-Gaussian Response Models

In many data studies the identification of observations which deviate significantly from the fitted model is of paramount interest. However, Bayesian residual analysis, in the context of discrete, non-Gaussian, count observations, has long been recognised as a challenging problem (Albert & Chib 1995). This is owing to a number of factors; the sampling distributions of proposed residuals tend to be unknown, impeding the objective detection of outliers and the use of fast model validation tools. Furthermore, the use of simulation based inference procedures for parameter inference can be tremendously slow, placing time constraints on the number, and complexity, of models that the Bayesian analyst may consider.

The novel contributions in this chapter relate to model choice and data criticism. Specifically, we propose solutions to the above issues involving Bayesian residual analysis in the non-Gaussian setting, focusing in particular on residual analysis for Poisson and Binomial regression models. We build upon existing theory regarding the use of posterior random effect terms as a “surrogate” for classical residuals. We develop a methodology for outlier determination, based on Gaussian approximations to posterior random effects, providing metrics by which to systematically identify possible outliers. We propose a visual approach to assess *a priori* model assumptions and harness the fast approximate Bayesian inference algorithms of Rue et al. (2009) to provide computationally efficient implementations of our methods, enabling the quick comparison of multiple models. Additionally, we highlight how the investigation of posterior random effects for residual trend or patterns potentially provides a rich source of information on underlying, complex relationships hidden within the data. Much of the work in this chapter is reported in Sweeney & Haslett (2011).

This chapter is organized as follows; Section 4.1 provides a brief review of some of the existing literature for Bayesian residual analysis, with a focus on the Poisson, Binary and Binomial regression model setting. In Section 4.2 we introduce a framework for outlier detection in the

presence of discrete, non-Gaussian count observations and provide a solution to the inference issues which arise as a result of the approach. In Section 4.3, we discuss the properties and limitations of the approach as derived from a number of simulation studies. In Section 4.4, we apply our developed residual analysis methodology to a dataset from the medical literature.

4.1 Bayesian Residual Analysis: An Overview

In the context of Gaussian data, Chaloner & Brant (1988) propose a Bayesian approach for the detection of outliers based on the posterior distribution of the error terms in regression models. The approach may best be explained in the context of a simple regression problem; consider the non-parametric regression model $y_i = f(x_i) + \epsilon_i$ where ϵ_i is a random sample from $N(0, \sigma^2)$ and f a smooth non-parametric function of the predictor variables $X = \{x_1, \dots, x_n\}$. *A priori*, each model residual, $\epsilon_i(f, \sigma^2) = y_i - f(x_i)$, is assumed to arise from an $N(0, \sigma^2)$ distribution. In light of data, the distribution of each ϵ_i follows from the posterior distributions of f and σ^2 respectively, reflecting the posterior uncertainty in model parameters.

A given observation is flagged as outlying if the related posterior distribution of $\epsilon_i(f, \sigma^2)$, $\pi(\epsilon_i|Y)$, is located far from zero (Albert & Chib 1995). The posterior probability of such an event is $Pr_i = Pr(|\epsilon_i| > k\sigma|Y)$ and observations with values of $Pr_i > 2\Phi(-k)$ are determined as outliers. Access to standard, classical Gaussian residual theory provides the critical bounds for outlier detection, namely values of $k = 1.96$ and $2\Phi(-1.96) = .05$.

Bayesian residual analysis in the Gaussian setting is generally quite simple; critical regions for the detection of possibly aberrant observations are easily obtained from the well established theory on outlier detection. Additionally, plots of the posterior residuals, $\pi(\epsilon_i|Y)$, help provide an insight into underlying latent trends or patterns within the data that may have been overlooked at the model formulation stage.

In contrast, where the response data are non-Gaussian in nature, the definition of model residuals and their analysis is much less clear cut. This is due to a number of factors; in the context of binary data, Souza & Migon (2010) note that classical residuals, such as Pearson or deviance residuals, suffer from unknown sampling distributions; residual plots are thus difficult to interpret. As a result, outlier detection is confined to highlighting residuals which appear “large” in magnitude, a relatively subjective measure.

A Bayesian framework for outlier detection in the binary data setting is provided by Albert & Chib (1995). Given the binary regression model $E(y_i) = p_i = F(x_i\beta)$, the authors propose the model residual: $r_i = y_i - p_i$, noting that the continuous valued posterior distributions for each r_i can be visually evaluated to learn about outlying observations. Each posterior r_i has support on the interval $(y_i - 1, y_i)$ and outlier detection involves the identification of posteriors, $\pi(r_i|Y)$, which tend towards the “extremes” (-1 or 1) of their respective support region.

Observations with large values of $Pr(|r_i| > k|y)$ are identified as outlying, however, as in the classical setting, there is great difficulty in assigning appropriate values of k . The prior distributions of the r_i are unknown owing to the discrete nature of the response variable thus sampling distributions for the r_i are unavailable. According to the authors, a value of $k = .75$ may be suitable for outlier detection purposes, however, whilst this results in the detection of observations that do not agree well with the chosen model, there is no theory by which to determine that such a choice of k is *always* appropriate. Furthermore, there is no elaboration by the authors on the properties of the approach, specifically, the number of potential outliers that one should generally “expect”.

There are several papers in the literature which consider the incorporation of random effect terms into models for outlier detection purposes. Marshall & Spiegelhalter (2007) present a simulation based approach to identifying outliers in Bayesian hierarchical models with random effect components, illustrating the approach by application to mortality comparisons between hospitals. Outlier detection is based on predictive model criticism, involving the examination of possible conflict between the predictive prior and the likelihood. The computational nature of the approach, which is MCMC based, involves the development of approximate leave-one-out cross validation methods for outlier detection and proposes measures for evaluating whether prior assumptions regarding the distribution of the random effects are appropriate.

Albert & Chib (1995) suggest the use of “tolerance random variables”, $Z_i = x_i^T \beta + \epsilon_i$. Whilst the prior distributions of the ϵ_i (random effects) are Gaussian, their posterior distributions conditional on data are not. Remarking that the posteriors appear “Gaussian like”, the authors attempt to identify outliers using standard Gaussian residual theory. However, in our practical experience, posterior distributions for residuals corresponding to count observations of zero frequently exhibit substantial skewness. Furthermore, there tends to be significant skewness in posterior random effects where the global variance parameter is large. The authors do not discuss this issue and avoid it by fixing the value of the variance parameter of the random effects *a priori*.

Souza & Migon (2010) also promote the inclusion of random effect terms in binary regression models for outlier detection purposes. The prior distribution for each random effect term is specified as a two-component scale mixture of normals; $\gamma_i | k_i \sim N(0, [(1 - k_i)\sigma^2 + ck_i\sigma^2])$, $c > 1$ and $k_i | \pi \sim Bern(\pi)$. Observations with large $p_{k_i} = Pr(k_i = 1)$ are identified as outliers, namely observations with high posterior probability of requiring an extra random effect to capture excess variability. According to the authors, their results correspond well to those obtained using the methods of Albert & Chib (1995), however, there remain issues regarding the determination of appropriate critical values for the p_i . Approximately 5%, or 27 of the 546 observations, are identified as outliers using the critical bound chosen - as such, it appears that Gaussian residual theory is used to motivate the choice of critical values. However, no justification for such an assumption is provided and inference procedures, as in Albert & Chib

(1995) are sampling based, leading to long run times for parameter inference.

Thall & Vail (1990) use independent random effects to capture patient and visit level effects in a longitudinal study for epilepsy treatment. By examining residuals comparing the fitted and observed counts of each subject at each visit, they identify a number of patients who had particularly large counts relative to the fitted model. Breslow & Clayton (1993) improve on this, using more complex models to identify patients with especially low counts, that were not so apparent in Thall and Vail’s analysis. However, in both papers, outlier detection is (subjectively) based on the visual analysis of residual plots. A similar paper in the same vein, by Perperoglou & Eilers (2009), promotes the use of distribution free deviance effects to model overdispersion of count outcomes in a number of discrete count data studies. The authors are not motivated by residual analysis per se, but note that analysis of the estimated deviance effects may suggest patterns in the data that can be captured by modified models.

In the context of the motivating palaeoclimate reconstruction project, Salter-Townshend (2009) defines reference distributions for the detection of outliers in the modern training dataset used for model calibration. However, the posterior sampling distributions of the reference measures are unknown and thus critical measures with which to determine outliers are unavailable. Several assumptions of the model are not examined, including the distributional assumptions of the random effect terms used to model overdispersion of the count observations. The model criticism tools considered concern the predictive accuracy of the approach, a measure from which it is difficult to derive information on the possible inadequacy of the model - for example, Salter-Townshend (2009) notes a correlation between poor model prediction accuracy and increasing altitude, a possibly spurious result as we will later show in Chapter 7; the altitude covariate is perhaps in fact acting as a proxy for moisture availability.

All of the introduced approaches for Bayesian residual analysis in the non-Gaussian data setting are subject to the same constraint; it is difficult to provide critical bounds by which to identify outliers, resulting in outlier detection by the visual analysis of posterior random effects or through the use of ad-hoc reference measures. Inference procedures tend to be sampling based; this imposes computational constraints on the Bayesian analyst as the fitting and comparison of multiple models for the data is time consuming and thus rarely considered. Additionally, model validation tends to be computationally expensive, being based on cross validation measures; there is little consideration of the use of the posterior random effects as a fast model validation tool.

4.2 Gaussian Random Effects as a Tool for Residual Analysis

In many data studies, the variability in the observed data is typically greater than that which can be captured by the usual exponential family probability models. In the context of Poisson count observations, this “overdispersion” can be detected by comparing the empirical means

and the variances of the observations; a significant inequality in these values indicates that the counts are overdispersed. As an example, Figure 4.1 presents some overdispersed Poisson count observations, simulated using the data generating process in Equations 4.1 - 4.2.

$$u_i \sim N(0, 1) \tag{4.1}$$

$$y_i \sim \text{Poisson}(e^{2+.02x+u_i}) \tag{4.2}$$

One solution to this overdispersion problem, as introduced in Section 3.5.3, is to incorporate Gaussian random effect terms, $U = \{u_1, \dots, u_n\}$, into the model to capture the overdispersion. Given the modelling of this extra source of variability by the random effect terms, interest returns to the identification of observations which still seem to deviate significantly from the fitted model.

As previously mentioned, several authors have considered the examination of the posterior random effects, $\pi(u_i|Y)$, for outlier detection purposes. Common to all approaches is the problem of objectively obtaining measures or bounds for the automatic detection of outliers given the posterior random effects $\pi(u_i|Y)$. Expressed in mathematical form, a given observation is said to be an outlier if $\pi(|u_i| > k|Y)$ is greater than some critical measure $f_{\text{crit}}(k)$ for a chosen value of k . The major difficulty in existing approaches concerns the selection of appropriate values of k and in obtaining the associated critical values $f_{\text{crit}}(k)$ as the sampling distributions of the posterior random effects tend to be unknown.

In Albert & Chib (1995), outliers are detected amongst the (non-Gaussian) posterior random effects through the use of standard Gaussian residual theory. k is set equal to 1.96 and $f_{\text{crit}}(k)$ is thus equal to $2\Phi(-1.96)$. *A priori*, each $\pi(u_i)$ is Gaussian, however, conditional on the data each $\pi(u_i|Y)$ can be *extremely* non-Gaussian, especially in situations where y_i is zero. The authors do not address this issue, nor the properties of the chosen value of k . The explicit critical measure obtained from Gaussian residual theory is also used in an ad-hoc manner; only observations with outlying probabilities *significantly* greater than the critical measure are identified as outliers.

An outlier detection method based on the examination of the posterior random effects can be computationally quite demanding. It is not possible to “integrate out” the random effect terms as the Gaussian distribution of the random effects is not conjugate to the non-Gaussian likelihoods that are typically used for count observations. Furthermore, the time and computational expense of this approach is proportional to the size of the data set - a random effect must be included in the model for each observation, increasing the number of unknown parameters that must be inferred. This constrains the fitting, criticism and comparison of multiple models in studies that are data “rich” with problems exacerbated if inference procedures are sampling based.

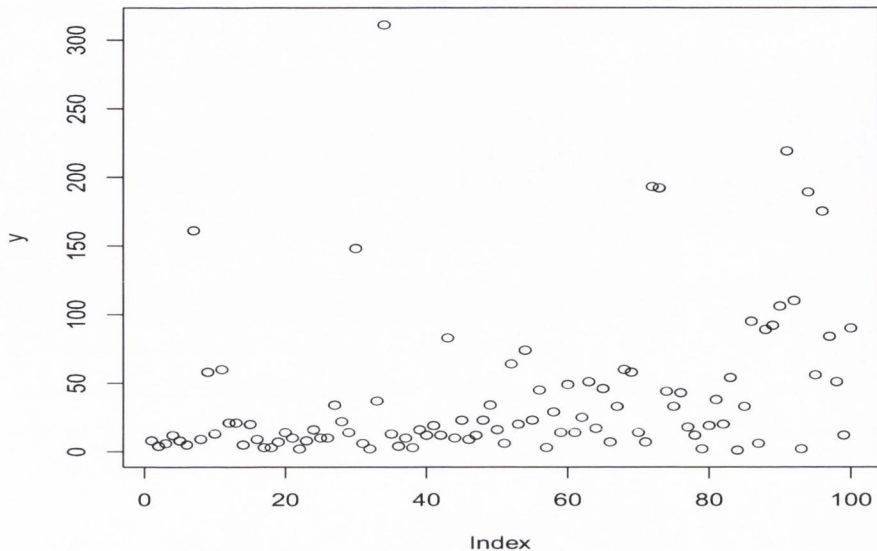


Figure 4.1: Overdispersed Poisson count observations, simulated using the data generating process in Equations 4.1 - 4.2.

4.2.1 A Methodology for Residual Analysis and Outlier Detection

In this thesis we propose to address the above issues through the formulation of a methodology for residual analysis and outlier detection, similarly built upon the use of posterior random effect terms as a “surrogate” for classical residuals. We propose to address the issue of objective bounds for outlier detection by expressing each posterior random effect in approximate Gaussian form, thus providing an explicit link to standard Gaussian residual theory. In Section 4.3 we examine the properties of this approximation. We illustrate how the posterior random effect terms can be quickly analyzed, in a classical residual analysis manner, to learn about residual patterns or trends within the data potentially masked by the discrete nature of the response variable. These features are considered as novel contributions in this thesis.

The primary advantage of the proposed methodology, as compared to existing methods, will be its computational efficiency; through the use of the INLA algorithm of Rue et al. (2009) we propose to address the time constraints imposed by the use of MCMC based methods. This will facilitate the quick implementation and comparison of multiple models for the data, providing a visual method for model validation and enabling the fast evaluation of model assumptions using classical residual analysis tools. Indeed the harnessing of the INLA algorithm for fast approximate inference on model parameters is critical to the success of the approach.

In the following we present a framework for Bayesian residual analysis in the non-Gaussian setting, as presented in Sweeney & Haslett (2011). In Section 4.2.2 we present a flexible frame-

work for modelling the covariate-response relationship within which most popular statistical models can be constructed. In Section 4.2.3 we detail how the INLA algorithm provides a mechanism for fast, computationally efficient inference on the unknown random effect terms which is critical to the success of the proposed approach. In Section 4.2.4 we present an objective approach to outlier detection based on standard Gaussian residual theory, facilitated by the expression of each posterior random effect in Gaussian form.

4.2.2 Model Framework

We present the model of Sweeney & Haslett (2011) in the flexible generalized linear model (GLM) framework (Gelman et al. 2003). Potential overdispersion of the observed counts is accounted for by the addition of a (mean zero) Gaussian random effect component to the non-parametric predictor. In the following let y_i represent the observed response, x_i the observed predictor variable(s), g a nonspecified link function and f some smooth non-parametric function of the predictor variables.

$$E(y_i|w_i) = g(w_i) \tag{4.3}$$

$$w_i = f(x_i) + u_i \tag{4.4}$$

$$u_i \sim N(0, \sigma^2) \tag{4.5}$$

Through the formulation of the inference task via a Bayesian hierarchical model, the model can be made as simple or as complex as necessary; the addition of extra stages to the model hierarchy is simple and easy to implement. Let U represent the vector of random effects $U = \{u_1, \dots, u_n\}$, $Y = \{y_1, \dots, y_n\}$ the vector of observations, $X = \{x_1, \dots, x_n\}$ the corresponding vector of covariates and θ , the vector of model hyperparameters.

$$\pi(f, U, \theta|Y, X) \propto \pi(Y, f, U, X, \theta) \tag{4.6}$$

$$\propto \pi(Y|f, U, X, \theta)\pi(U, f|X, \theta)\pi(\theta) \tag{4.7}$$

$$= \prod_{i=1}^n \pi(y_i|u_i, f, X, \theta)\pi(U, f|X, \theta)\pi(\theta) \tag{4.8}$$

The prior distribution of a given random effect, u_i is specified as $N(0, \sigma^2)$ with appropriate priors used for f and θ . As the joint posterior, $\pi(f, U, \theta|Y, X)$ is usually not known in closed form, obtaining marginal posterior statements requires the evaluation of complex integrals either, numerically, or through the use of sampling based inference procedures. In the context of large datasets, the addition of random effect terms to the model can be burdensome in

the extreme; as previously mentioned, a posterior random effect must be inferred for each observation which, given large amounts of data, will introduce computational issues.

To illustrate these computational issues, we briefly return to the motivating palaeoclimate reconstruction problem. The RS10 dataset of Allen et al. (2000), considered in Chapter 7, contains over 200,000 observations. The model framework for residual analysis we propose requires that a random effect be inferred for each datapoint; as a result numerical evaluation of the posterior is completely infeasible. Additionally, simulation based inference procedures such as MCMC will encounter the usual range of issues regarding correlation between samples, convergence of the sampling chains and burn-in periods. The use of MCMC for parameter inference in the context of the palaeoclimate reconstruction project was shown to be too slow to consider (Salter-Townshend 2009) for models that were much less computationally demanding than the ones considered later in this thesis.

4.2.3 Fast Approximate Bayesian Inference for Posterior Random Effects

In order to sidestep the computational issues of simulation based inference of unknown model parameters, we take advantage of the fast approximate Bayesian inference algorithms of Rue et al. (2009), as introduced in Section 3.3. In the following we briefly describe the framework of the approach in the context of the introduced model:

$$\pi(\theta|Y, X) \propto \frac{\pi(Y|f, U, X, \theta)\pi(U, f|X, \theta)\pi(\theta)}{\pi(U, f|\theta, Y, X)} \quad (4.9)$$

$$= \frac{\prod_{i=1}^n \pi(y_i|u_i, f, x_i, \theta)\pi(U, f|X, \theta)\pi(\theta)}{\pi(U, f|\theta, Y, X)} \quad (4.10)$$

$$\approx \frac{\prod_{i=1}^n \pi(y_i|u_i, f, x_i, \theta)\pi(U, f|X, \theta)\pi(\theta)}{\tilde{\pi}_G(U, f|\theta, Y, X)} \Big|_{\{U, f\}=\{U, f\}^*(\theta)} \quad (4.11)$$

Equation 4.11 is more commonly known as the Laplace approximation. A Gaussian approximation is fit to the joint posterior distribution of the random effect terms, U , and the non-parametric function of the predictor variables, f conditional on θ . $\tilde{\pi}_G(U, f|\theta, Y, X) |_{\theta=\theta^*}$ is the Gaussian approximation to the full conditional of $\{U, f\}$ and $\{U, f\}(\theta)$ is the mode of the full conditional for $\{U, f\}$ for a given value of θ . This simplifies to an empirical Bayes type approach if we fit this approximation using the modal value of θ which maximizes the posterior density of $\pi(\theta|Y, X)$; numerical search algorithms such as the Newton Raphson algorithm provide a method for locating the modal value. We defer to Rue et al. (2009) for additional details.

Otherwise, if the dimension of the model hyperparameters, θ , is not too great (typically the models in this thesis contain at most four), the joint posterior distribution of θ , given the data, can be computed on an (arbitrarily fine) discrete grid. Probability weights can then

be calculated which enable us to represent the posterior distribution for each random effect, $\pi(u_i|Y)$ as a weighed mixture of Gaussians.

$$\tilde{\pi}(u_i|Y) = \sum_k \tilde{\pi}_G(u_i|\theta_k, Y) \times \pi(\theta_k|Y) \times \Delta_k \quad (4.12)$$

$\tilde{\pi}_G(u_i|\theta_k, Y)$ is available from $\tilde{\pi}_G(u_i, f(x_i)|\theta_k, Y)$ and Δ_k are area weights which ensure the posterior probability distributions for each random effect sum to one.

In Figure 4.2 we fit Gaussian approximations ($\tilde{\pi}_G$) to posterior random effects for a number of count values where $\pi(u) \sim N(0, 1)$. We observe that the error between the Gaussian approximation to the posterior and the true posterior reduces as the value of the count observation increases. This is to be expected, asymptotic statistical theory dictates that Gaussian approximations to the Poisson distribution become more accurate with increasing value of the rate parameter.

The R-INLA package of Rue et al. (2009) provides software which implement the approximations introduced in this section. Through the use of numerical algorithms for inference on the (low dimensional) model hyperparameters and the harnessing of algorithms for fast operations on sparse matrices, the software facilitates quick, approximate inference on unknown model parameters. Full posterior inference on the random effect terms can thus be carried out in seconds or minutes whereas previously, using MCMC methods, this would have taken hours or even days. This facilitates the fast fitting, criticism and comparison of multiple models for the observed data and is crucial to the success of the proposed methodology for residual analysis.

4.2.4 Residual Analysis and Outlier Detection

As identified in Section 4.1, the provision of objective critical bounds for systematic outlier detection is a recurrent problem in non-Gaussian data studies. In the following we address this issue by taking advantage of the Gaussian approximation to the posterior distribution of the random effect terms to obtain objective critical measures for outlier detection from Gaussian residual theory. In Section 4.3 we investigate the properties associated with this choice of critical measures.

Suppose we say the i^{th} observation is an outlier if $Pr_i = Pr(|u_i| > k|Y)$ is sufficiently “large” (i.e. greater than $f_{crit}(k)$) where k is a critical value supplied from an available reference distribution.

$$Pr(|u_i| > k|Y) = \int Pr(|u_i| > k\sigma|Y)\pi(\sigma^2|Y)d\sigma^2 \quad (4.13)$$

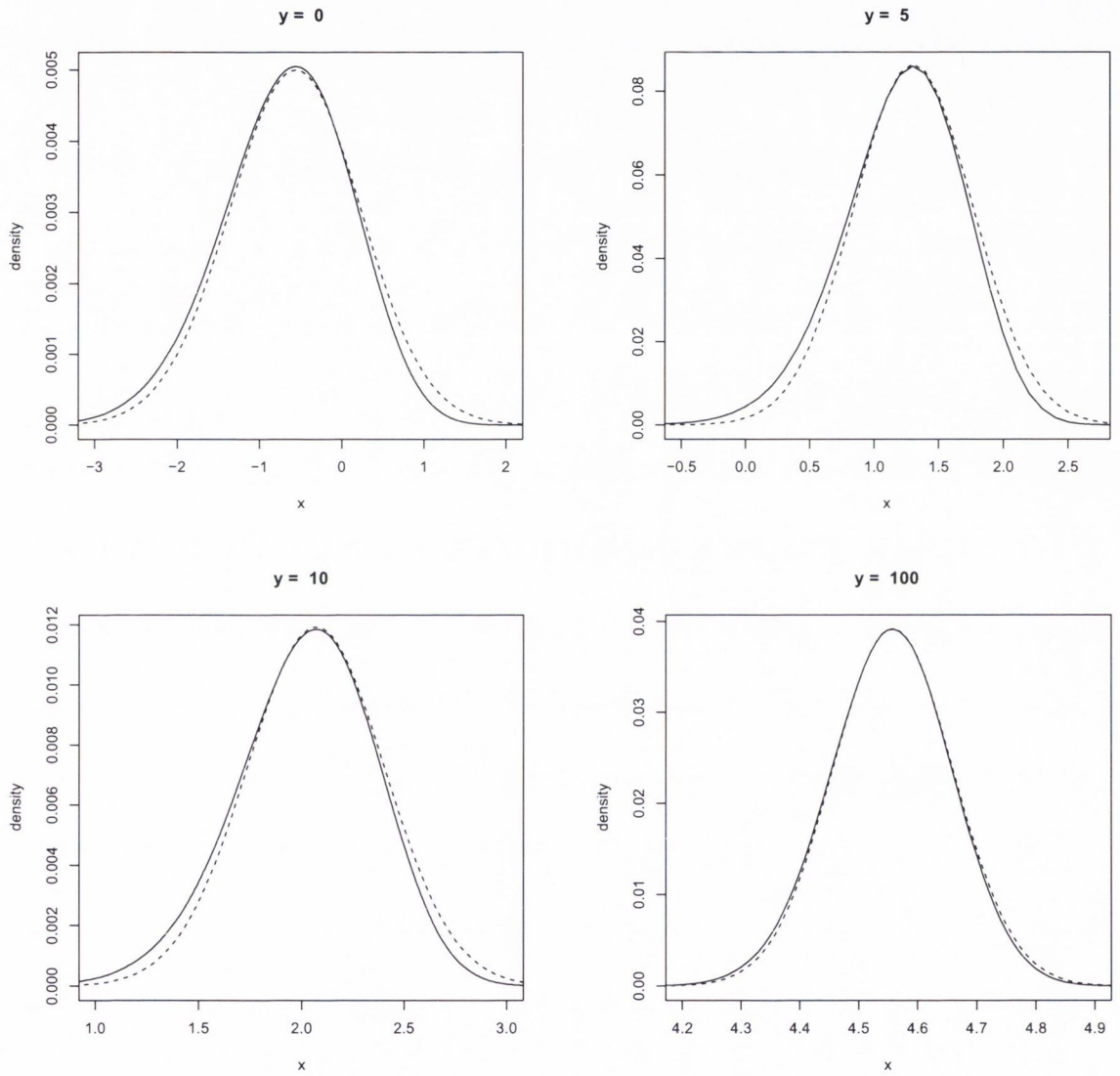


Figure 4.2: Comparison of $\tilde{\pi}_G(u|y)$ (---) and $\pi(u|y)$ (—) for a number of count values. We observe that the Gaussian approximation to the posterior becomes more accurate as the value of the count observation increases.

$$\approx \int Pr_G(|u_i| > k\sigma|Y)\pi(\sigma^2|Y)d\sigma^2 \quad (4.14)$$

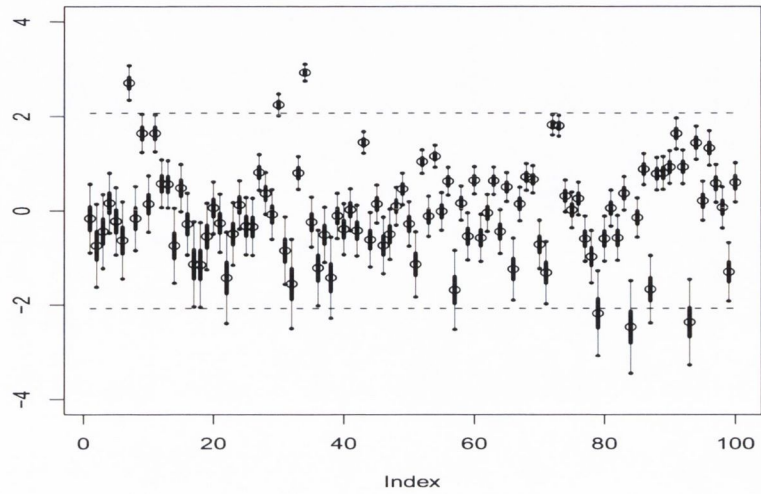
$$= E_{\sigma^2}[1 - Pr_G(u_i(\sigma) < k\sigma|Y) + Pr_G(-u_i(\sigma) < k\sigma|Y)] \quad (4.15)$$

A priori, each $u_i \sim N(0, \sigma^2)$ where $\sigma^2 \in \theta$. The probability that a given u_i is “outlying”, for a given value of σ^2 , is calculated from the Gaussian approximation to the full conditional. The integral in Equation 4.13 can be replaced with a summation due to the representation of $\pi(\sigma^2|Y)$ on a discrete grid. In the above, $Pr(u_i < k\sigma|Y) = \Phi(k\sigma, \mu_i, \tau_i^2)$ where $\pi(u_i|\sigma^2, Y) \approx N(\mu_i(\sigma^2), \tau^2(\sigma^2))$ and k is available from standard Gaussian residual theory. The resulting probability, $Pr(|u_i| > k|Y)$ can be compared to $2\Phi(-k)$ to identify “suspicious” observations. Thus a given random effect may be detected as outlying due to the magnitude of its mean or the “size” of the variance surrounding it. Uncertainty in the variance parameter, σ^2 of the random effect terms can be “integrated out” by repeating this process for all values of σ^2 with the corresponding posterior probabilities obtained from $\pi(\sigma^2|Y)$.

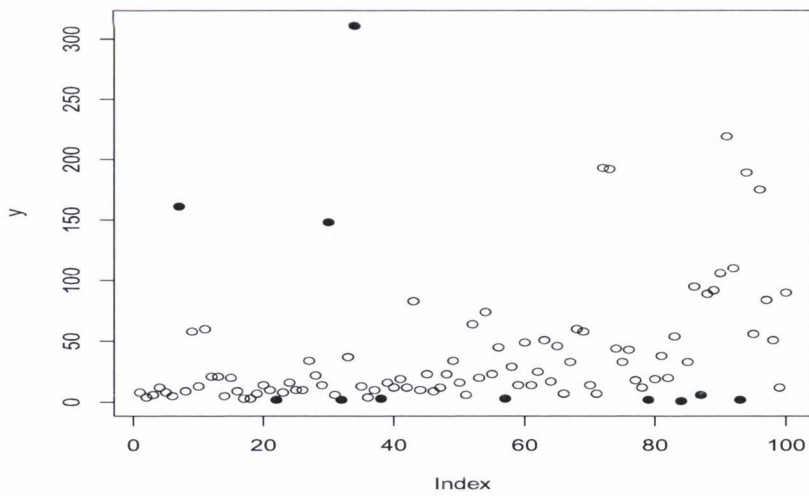
In order to provide an example of the method, the posterior random effects corresponding to the count observations in Figure 4.1, are inferred given the known model parameters. In Figure 4.3, the posterior random effect for each observation is plotted along with the 95% error bounds. Under standard Gaussian residual theory and given the data generating process, we would expect that approximately 5% of observations should be detected as potentially outlying. However, we see that 11% of the observations are identified as having outlying probabilities that are significant. We also note that the majority of the counts detected appear to be those corresponding to small values of y_i . In Section 4.3.1, we investigate this result in further detail, providing an explanation for the (excess) number of observations detected.

Under the proposed framework, the posterior random effects are used as a surrogate for classical residuals to aid in outlier detection. However, they may additionally be used in an exploratory fashion (as in classical residual analysis) for model criticism purposes. In Section 4.3 we will illustrate how the examination of the mean posterior random effects, $E(U|Y)$, may help identify patterns within the data masked by the discrete nature of the response variable. In Figure 4.4 we present a quantile-quantile plot of the mean posterior random effects corresponding to the posterior random effects in Figure 4.3 (a). We see that the mean posterior random effects seem to reflect well the underlying data generating process which was Gaussian.

Plots of the posterior random effects such as the quantile-quantile plot presented in Figure 4.4 provide a quick visual method of evaluating *a priori* model assumptions. As we will later see in Section 4.3.3, the quantile-quantile plots can be used to provide an informative insight into underlying model properties and detect model inadequacy. This is considered one of the novel contributions in this thesis.



(a)



(b)

Figure 4.3: (a) Boxplots of the posterior random effects corresponding to the count observations presented in Figure 4.1. Posterior distributions of the random effects which significantly cross the dashed lines ($\pm 1.96\sigma$) are considered outliers. σ here is fixed at its posterior modal value, as in an empirical Bayesian analysis fashion. In (b), the observations (\bullet), corresponding to the outliers detected in (a) are identified.

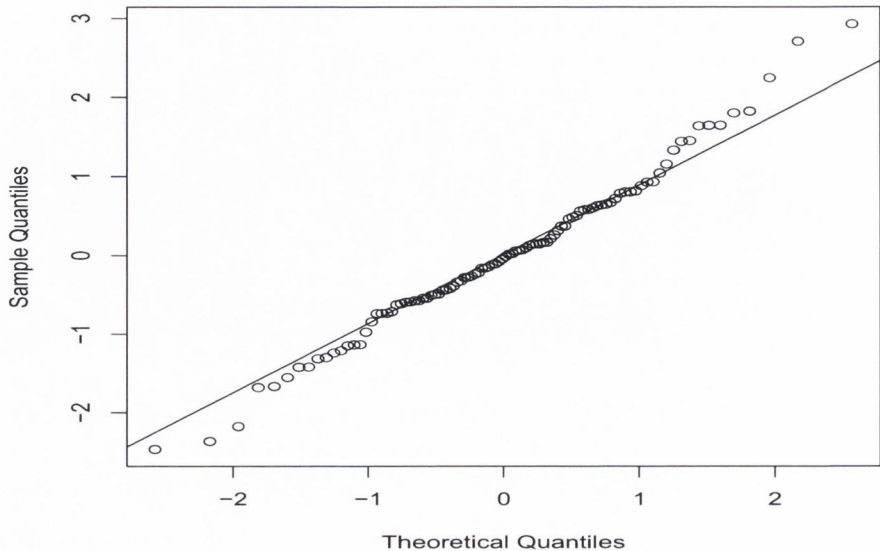


Figure 4.4: Quantile-quantile plot of the inferred mean posterior random effects, $E(U|Y)$. The sample quantiles of the random effects seem to match well the theoretical quantiles of a Gaussian distribution.

To summarise, the framework for residual analysis and outlier detection we present in this thesis can be considered as an approach which facilitates significant exploration of the observed data. In contrast to MCMC based methods for parameter inference, the use of fast approximate Bayesian inference algorithms allows the fast fitting of multiple models for the observed data. Through the expression of the posterior random effect terms in approximate Gaussian form, we are able to quickly identify outliers, using critical measures obtained from Gaussian residual theory. We are also able to use the posterior random effect terms as a fast model validation tool; analysis of the posterior random effects may identify residual, uncaptured trend patterns, helping to suggest more appropriate models for the data in question.

4.3 Properties Of the Proposed Methodology

The proposed methodology for residual analysis in the non-Gaussian setting is based upon the examination of posterior random effect terms which we use as a surrogate for classical residuals. In order to identify outliers, we approximate each posterior random effect by a weighed mixture of Gaussians. In this thesis we propose that such an approximation facilitates access to standard Gaussian residual theory. We do not expect the posterior random effects to have the same distributional properties as the Gaussian distribution, but based on empirical analysis the results seem to coincide greatly, especially where count observations are large.

In this section we propose to investigate the properties of this approach. Section 4.3.1 focuses on deriving, via simulation studies, the outlier detection properties of the proposed approach in the Poisson setting; in Section 4.3.2 we consider a similar analysis for the Binomial and Bernoulli response setting. Finally, in Section 4.3.3 the use of the posterior random effects as a model validation tool is investigated. Whilst the simulated studies in the following appear quite specific in nature, they are motivated by the data applications considered later in this thesis.

4.3.1 Outlier Detection Properties in the Poisson Setting

Returning to the simple Poisson example presented in Figure 4.3, we note that 11% of the observations are recorded as have significant outlying probabilities, more than would be expected under standard Gaussian residual theory. In analysing the plots of the posterior random effects, we observe that the 95% highest posterior density intervals are significantly wider for random effects corresponding to low count values than for large count values, reflecting the fact that the Gaussian approximation to a given posterior is more accurate for large count values. This perhaps explains why the majority of the count observations suggested as outlying are those with low values - a given random effect may be considered outlying due to the magnitude of its mean or the uncertainty surrounding it (Albert & Chib 1995).

The discrete nature of the response variable in the non-Gaussian setting masks the underlying random effect; as a result there is more uncertainty in posterior distributions of random effect terms in the non-Gaussian setting as compared to the Gaussian setting. We should therefore “expect” to identify more observations as potentially outlying than would be expected under standard Gaussian residual theory.

In the following we investigate this claim. 200 count observations are simulated for each of 15 values of $\mu \in [0, 5]$ according to the data generating process in Equation 4.16 - 4.17. Given the known model parameters, the proposed methodology for outlier detection is used to identify the number of outliers occurring amongst the simulated counts for each value of μ . These numbers are recorded and this process repeated 50 times in order to derive the outlier detection properties of the approach. As the value of μ increases, the simulated count observations will also increase in value. In Figure 4.2 we observed that the posterior random effects became more “Gaussian like” as the value of the count observation increased; as a result we propose that the number of outliers detected given increasing μ should tend to the number expected under classical Gaussian residual theory.

$$u_i \sim N(0, 1) \tag{4.16}$$

$$y_i \sim \text{Poisson}(e^{\mu+u_i}) \tag{4.17}$$

Figure 4.5 presents the simulation results. As expected, for increasing value of μ , we observe that the number of outliers detected tends towards the number expected under classic Gaussian residual theory. Conversely, we also note that the performance of the approach deteriorates as $\mu \rightarrow 0$. This is due to an increase in the number of zero or low count observations which arise, these observations are “more likely” to be identified as outlying due to the magnitude of the uncertainty in the posterior distributions of the corresponding random effects.

A further important factor in the detection of outlying observations is the *global* posterior variance $\pi(\sigma^2|Y)$ of the random effects. In order to calculate the outlying probabilities for individual random effects, we must integrate over $\pi(\sigma^2|Y)$ (see Equation 4.15).

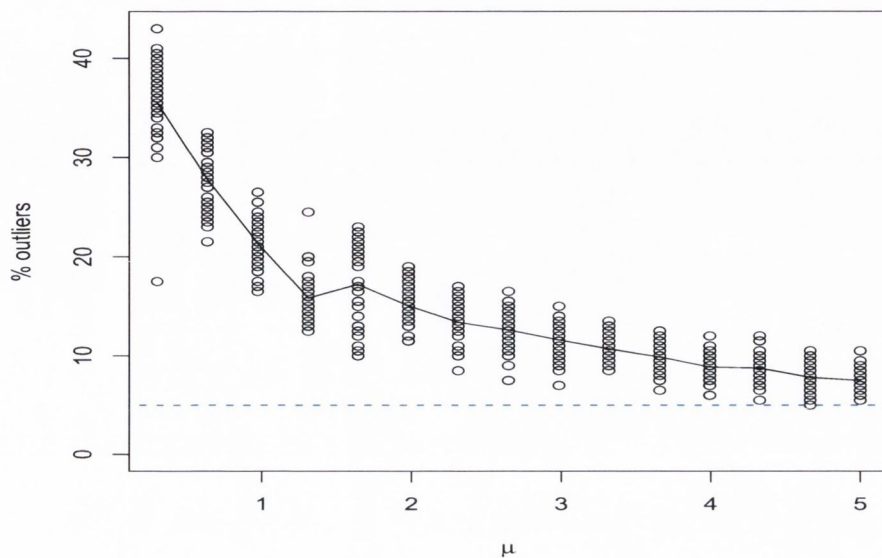
In order to assess the effect of σ^2 on outlier detection, 200 count observations are simulated for each of 15 values of $\sigma^2 \in [0.1, 4]$, according to the data generating process in Equation 4.18 - 4.19. The mean parameter μ is set equal to 4 to reduce the number of zero or low count observations which may mask deficiencies of the approach. Given the known model parameters, the proposed methodology for outlier detection, presented in Equation 4.13 - 4.15 previously, is used to identify the number of outliers occurring amongst the simulated counts for each value of σ^2 . The performance of the approach is evaluated in terms of two critical bounds, variously $k = \Phi^{-1}(.95)$ and $k = \Phi^{-1}(.975)$. This process is repeated 50 times in order to derive the outlier detection properties of the approach conditional on σ^2 .

$$u_i \sim N(0, \sigma^2) \tag{4.18}$$

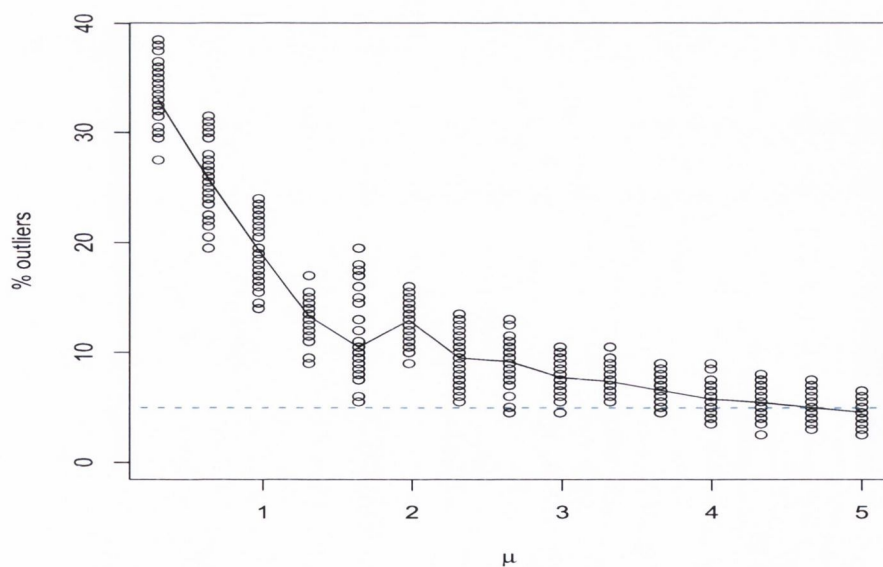
$$y_i \sim \text{Poisson}(e^{4+u_i}) \tag{4.19}$$

In Figure 4.6 the results of the simulation study are plotted. We observe that the number of outliers detected is relatively stable across a range of values of σ^2 . In a general sense, we observe that on average twice as many observations are identified as outlying in the Poisson count setting as would typically be expected under standard Gaussian residual theory. These results perhaps provide a further explanation for the percentage of outliers (11%) identified using 5% Gaussian error bounds in the simple example presented earlier in this chapter.

Additionally, we observe the deterioration of the approach as $\sigma^2 \rightarrow 0$. As previously mentioned, the prior distribution of each random effect is specified as $N(0, \sigma^2)$. As $\sigma^2 \rightarrow 0$, the posterior distribution of a given random effect, $\tilde{\pi}_G(u_i|Y) \sim N(\mu_i, \tau_i^2)$, reverts to the prior mean of 0. However, the reduction in posterior variance of the random effects is not linear with respect to the reduction in σ^2 for small values of σ^2 . The discrete nature of the response variable masks the underlying random effects, thus as σ^2 decreases in magnitude the uncertainty in the posterior random effects due to the non-Gaussian response becomes more prominent and results in an increase in the number of posterior random effects identified as potentially outlying. The most important result obtained from Figure 4.6 is that if the data

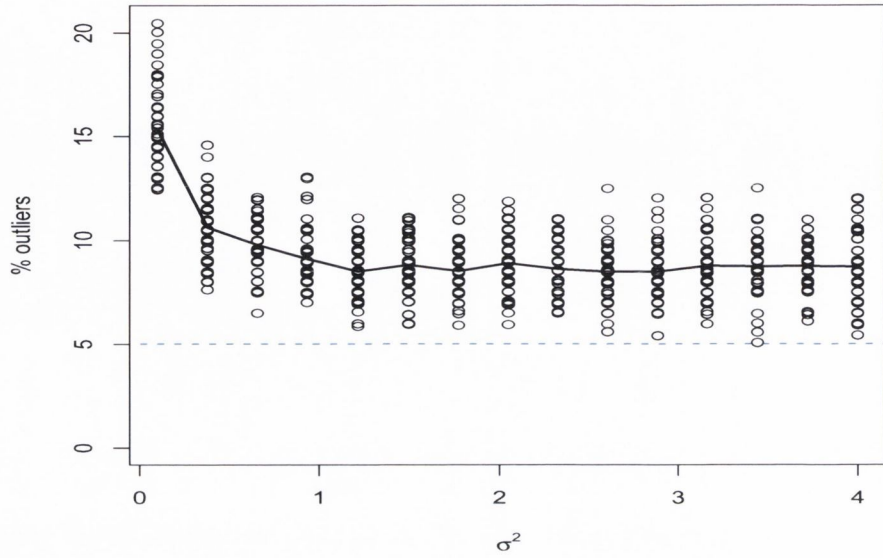


(a)

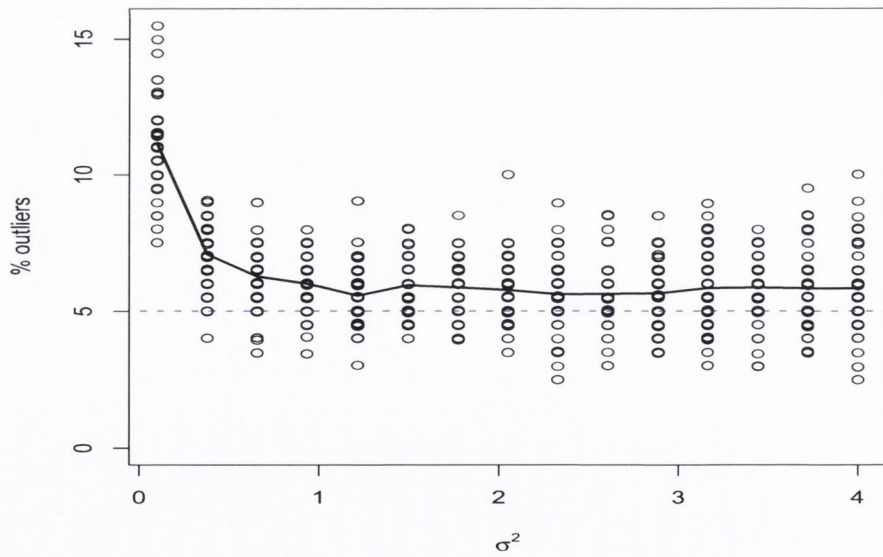


(b)

Figure 4.5: (a) A plot of the percentage of observations detected as outliers for a number of simulated datasets when (a) $k = \Phi^{-1}(.95) = 1.96$ and (b) $k = \Phi^{-1}(.975) = 2.24$, across a range of values of μ . Each \circ represents the number of outliers detected in an independent simulation for a given value of μ . The dashed line (--) represents the 5% error line.



(a)



(b)

Figure 4.6: (a) A plot of the percentage of observations detected as outliers for a number of simulated datasets when (a) $k = \Phi^{-1}(0.95) = 1.96$ and (b) $k = \Phi^{-1}(0.975) = 2.24$, across a range of values of σ^2 . Each \circ represents the number of outliers detected in an independent simulation for a given value of σ^2 . The dashed line (--) represents the 5% error line.

truly are Poisson distributed (i.e. overdispersion of the counts is negligible) the proposed approach performs poorly.

4.3.2 Outlier Detection Properties in the Binomial Setting

With a view to the application of the developed methodology for outlier detection to the AMI dataset of Souza & Migon (2010) in Section 4.4, we attempt to obtain a broad outline of the properties of the proposed approach in the context of Binomial response data.

The first simulation study considered concerns the performance of the approach with regard to increasing values of the sum total, N , of the Binomial counts. The discrete nature of the response variable will mask the underlying random effect, hence there will be more uncertainty in posterior distributions of random effect terms in the Binomial setting as compared to the Gaussian setting. Therefore, as in the Poisson setting, we should “expect” to identify more observations as potentially outlying than would be expected under standard Gaussian residual theory. In the following we use simulated data to assess the validity of this claim.

Given the data generating process in Equation 4.20 - 4.22, 200 count observations are simulated for each of 15 values of $N \in [1, 1000]$. Given the known model parameters, the proposed methodology for outlier detection is used to identify the number of outliers occurring amongst the simulated Binomial counts for increasing value of N . These numbers are recorded and this process repeated a large number of times (50) in order to approximately derive the outlier detection properties of the approach for each value of N . An intercept term μ_i is simulated from a Uniform(3, -3) distribution in order to evaluate the properties of the approach over a broad range of values of $p_i = \text{logit}(\mu_i + u_i)$.

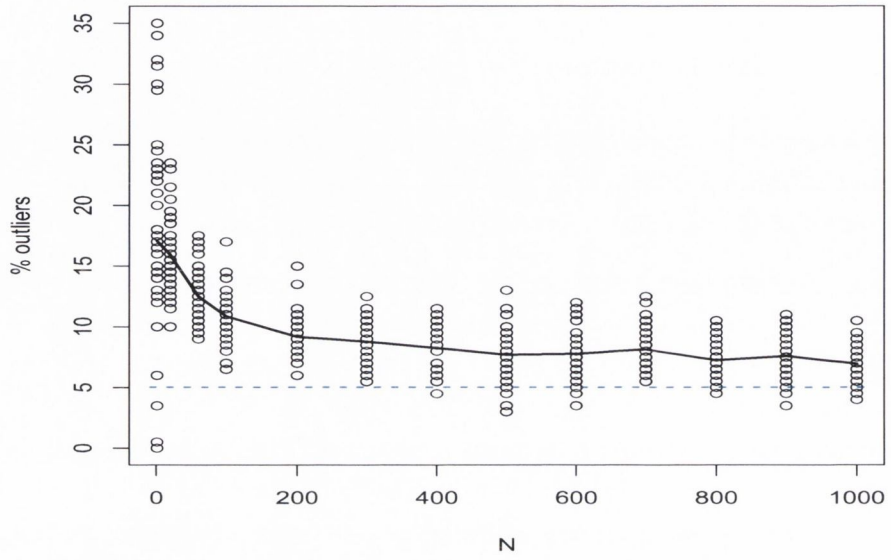
$$u_i \sim N(0, 1) \tag{4.20}$$

$$y_i \sim \text{Binomial}(N, \text{logit}(\mu_i + u_i)) \tag{4.21}$$

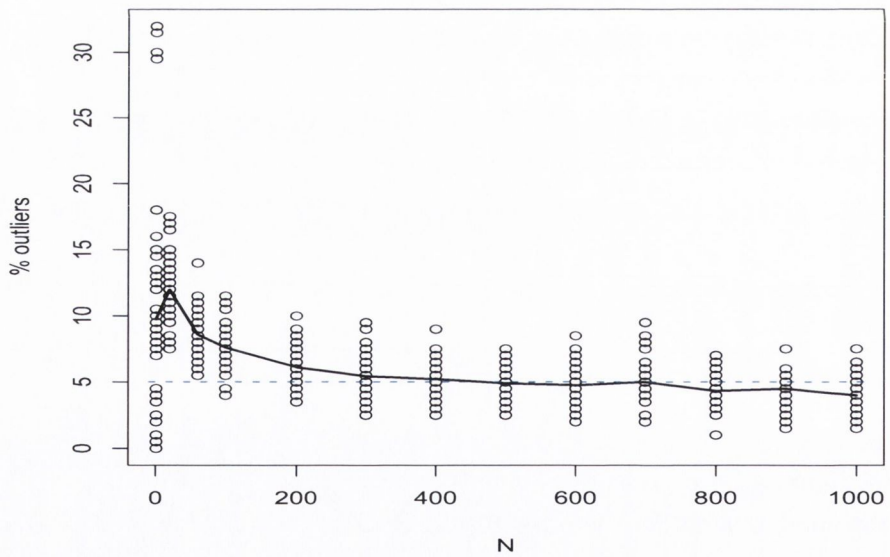
$$\mu_i \sim \text{Uniform}(-3, 3) \tag{4.22}$$

In Figure 4.7 we present the simulation results. We observe that, for increasing value of N , the number of outliers detected tends towards the number expected under classic Gaussian residual theory. Conversely, we also note that the performance of the approach deteriorates as N decreases in value, with the highest average number of outliers detected when $N = 1$, reflecting poor performance of the approach in settings where the response is binary.

As in the Poisson setting considered in the preceding section, a further important factor in the detection of outlying observations is the *global* posterior variance $\pi(\sigma^2|Y)$ of the random effects. Fixing the value of N to be 1000, we evaluate the performance of the proposed methodology for outlier detection by examining the number of outliers detected for varying



(a)



(b)

Figure 4.7: (a) A plot of the percentage of observations detected as outliers for a number of simulated datasets when (a) $k = \Phi^{-1}(.95) = 1.96$ and (b) $k = \Phi^{-1}(.975) = 2.24$, across a range of values of N . Each \circ represents the number of outliers detected in an independent simulation for a given value of N . The dashed line (--) represents the 5% error line.

values of σ^2 .

As before, 200 count observations are simulated for each of 15 values of $\sigma^2 \in [0.1, 6]$ according to the data generating process in Equation 4.23 - 4.25. Given the known model parameters, the proposed methodology for outlier detection is used to identify the number of outliers occurring amongst the simulated Binomial counts for each value of σ^2 . Once more this process is repeated 50 times.

$$u_i \sim N(0, \sigma^2) \tag{4.23}$$

$$y_i \sim \text{Binomial}(1000, \text{logit}(\mu_i + u_i)) \tag{4.24}$$

$$\mu_i \sim \text{Uniform}(-3, 3) \tag{4.25}$$

In Figure 4.8 we observe that the number of outliers detected is relatively stable across a range of values of σ^2 , though the performance worsens as σ^2 tends towards zero. This reflects that the use of the posterior random effect terms as a surrogate for classical residuals appears to have good properties for outlier detection, save in circumstances where the overdispersion in the observed data becomes increasingly small in magnitude, i.e. the data truly are Binomial in nature.

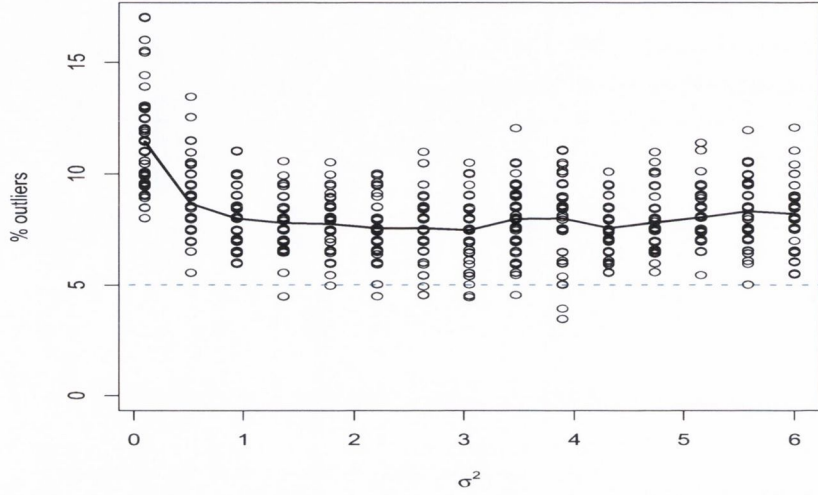
In Section 4.3.4 we discuss a number of conclusions regarding the properties of the proposed methodology, as derived from the simulation studies presented here.

4.3.3 Quick Approximate Model Validation

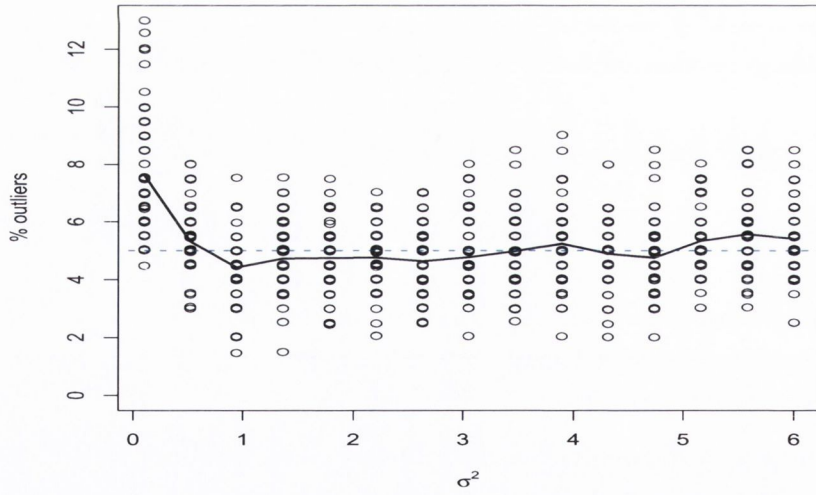
The posterior random effects have a further important use as a model validation tool, facilitating a quick approximate measure for determining if the *a priori* model assumptions are appropriate. In the following simulation studies, we illustrate how the posterior random effects can be examined for model misspecification and how they may further help suggest appropriate alternative distributions for the random effect terms.

Returning to Figure 4.4, we observe that a quantile-quantile plot of the mean posterior random effects well reflects the underlying data generating process which is Gaussian. However, the question arises as to what form the quantile-quantile plots of the posterior u_i take if the “true” distribution of the random effects is other than Gaussian.

We investigate this scenario by simulating some Poisson count observations where the u_i are generated from a Gamma distribution and model fitting proceeds as if the u_i are Gaussian in nature. In Figure 4.9, we plot the quantile-quantile plot of the $E(U|Y)$ which result. We observe that the quantile-quantile plot verifies that model misspecification has occurred; the Gamma distributed nature of the u_i can be clearly detected.

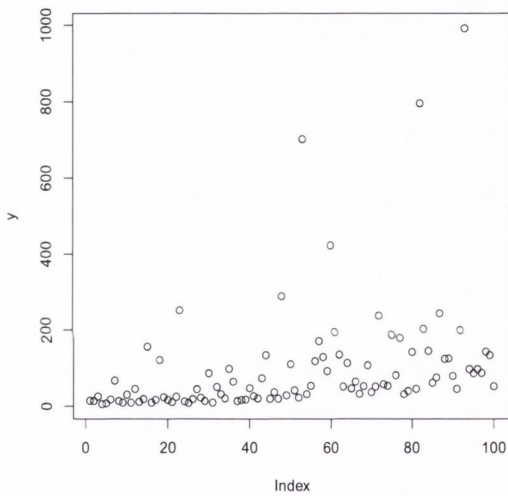


(a)

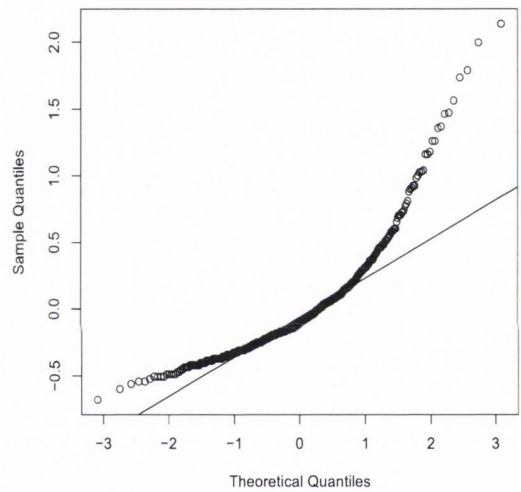


(b)

Figure 4.8: (a) A plot of the percentage of observations detected as outliers for a number of simulated datasets when (a) $k = \Phi^{-1}(.95) = 1.96$ and (b) $k = \Phi^{-1}(.975) = 2.24$, across a range of values of σ^2 . Each \circ represents the number of outliers detected in an independent simulation for a given value of σ^2 . The dashed line (—) represents the 5% error line.



(a)



(b)

Figure 4.9: (a) Plot of simulated count observations where $u_i \sim \text{Gamma}(1, 2.5)$ and $y_i \sim \text{Poisson}(e^{3+u_i})$ (b) The corresponding quantile-quantile plot of the mean posterior random effects, inferred given the *a priori* assumption that the random effects were in fact Gaussian. The (true) Gamma distributed nature of the random effects can be observed.

Frequently, omission of model covariates, particularly in the case of categorical data, results in residual trend patterns in the posterior random effects. To provide an example of this, the u_i are simulated from a Gaussian mixture distribution with different means and variances; $u_i \sim p \times N(1, .5) + (1 - p) \times N(3, 1)$ with $p = .5$. Poisson counts are generated conditional on the simulated random effects; in Figure 4.10 (a) the count observations which result can be observed. Once more, model parameters are inferred, conditional on the *a priori* assumption that the random effects are Gaussian distributed.

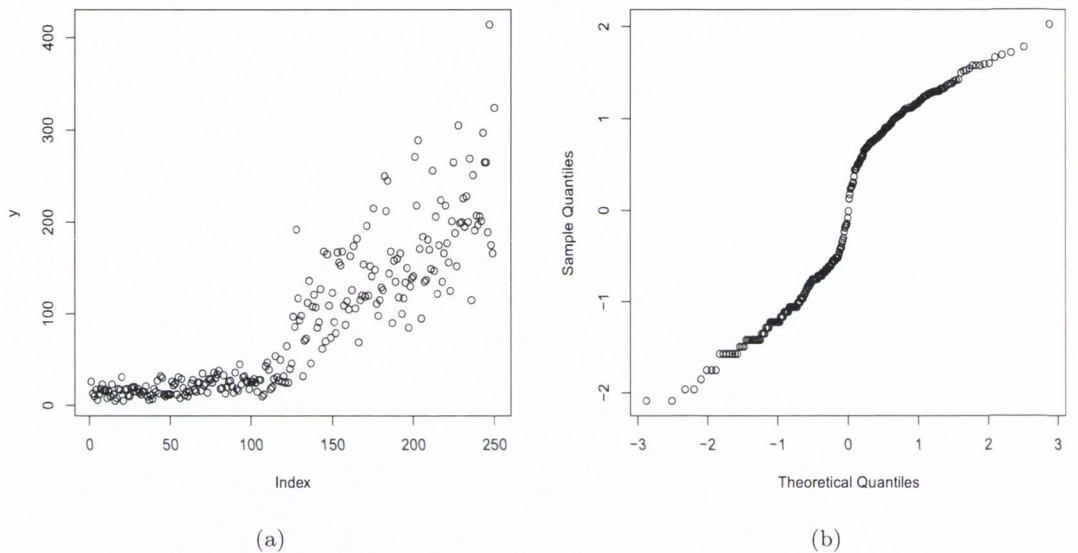


Figure 4.10: (a) Plot of the simulated count observations generated using random effects that arise from a mixture distribution and (b) the corresponding quantile-quantile plot of the mean posterior random effects where the *a priori* distribution of the random effects is univariate Gaussian.

Figure 4.10 (b) illustrates that the posterior u_i can be used to detect this misspecification; the quantile-quantile plot of the posterior random effects indicates a two tier structure in the data due to the omission of the covariates. There is residual trend in the $E(U|Y)$ as the model is incorrectly specified. The random effects are generated using a process comprised of a mixture of Gaussians with differing means and variances which the assumed prior model for the random effects, $u_i \sim N(0, \sigma^2)$, is unable to adequately capture. In parameter inference, the global variance parameter, σ^2 of the random effects is overestimated as a result of the mixture distribution, the single variance parameter inflated in attempting to account for the excess variability in the counts. As a result, less outliers are detected on average than would be expected given the simulation studies considered previously.

Although the focus in the above was on residual analysis in the Poisson count setting,

this approach can be performed in the exact same manner for posterior random effects in the context of Binomial trial outcomes. In Section 4.4.2 we detail the use of the above methods for quickly assessing model properties in the context of a Binary regression problem. Later on, in Chapter 7, we detail the usefulness of the approach for fast approximate model validation in the context of zero/N-inflated Binomial response data.

4.3.4 Strengths and Weaknesses of the Methodology

Whilst the simulation experiments provided in the preceding sections do not provide an extensive and rigorous assessment of the outlier detection properties of the proposed methodology, the experiments provide a wealth of information on the relative strengths and weaknesses of the approach. We illustrated that the posterior random effect terms can suitably be used for outlier detection with the experiments in this section indicating the stable properties of the approach under the various scenarios considered. The major strengths of the proposed methodology include its computational speed, due to the harnessing of the INLA algorithm for inference on model parameters, and the manner in which random effects can be used in an exploratory fashion. The fast nature of the inference procedure used for parameter inference means that multiple models for the data in question can be quickly fit and discriminated between. The posterior random effects provide a method for doing so; they can be visually examined, at no extra computational cost, to quickly assess *a priori* model assumptions providing a fast, approximate model validation tool. This analysis may also be used to suggest more appropriate models for the data in question or detect the presence of unmodelled covariates.

However, we observe that the proposed approach does not appear to have good outlier detection properties in the context of low Poisson or Binomial counts data. As previously mentioned, such observations are “more likely” to be identified as potentially outlying due to the magnitude of the uncertainty in the posterior distributions of the corresponding random effects. Ultimately, the major weakness of the approach lies with the overdispersion of the data. If no overdispersion is present, the posterior distribution of σ^2 , which models the variance of the random effect terms, will tend to zero. Furthermore, the posterior random effect terms, which are a function of σ^2 will also tend to zero. If the data are only slightly overdispersed, it is difficult to accurately infer the variance parameter σ^2 , typically resulting in its underestimation. As a result misleading inferences will be obtained by comparing the posterior random effect terms to critical bounds which are a function of σ^2 . Figure 4.6 (a) illustrates this problem, the number of outliers detected increases as σ^2 tends to zero reflecting the poor performance of the approach.

In the absence of overdispersion, the proposed methodology falls down and alternative methods for outlier detection and residual analysis must be pursued.

4.4 Application: Heart Attack Dataset

In this section our proposed framework for residual analysis is demonstrated on the heart attack dataset of Souza & Migon (2010). Our main motivation is to predict outliers within the available dataset and compare our results with those obtained by the authors. The Bernoulli nature of the response variable provides a particularly challenging problem for the developed methodology, however, we illustrate how the proposed approach helps identify outliers in a much more systematic manner than those considered in the paper. Additionally, we demonstrate how visual analysis of the posterior random effect terms provides a novel insight into underlying model dynamics that potentially explain model failings.

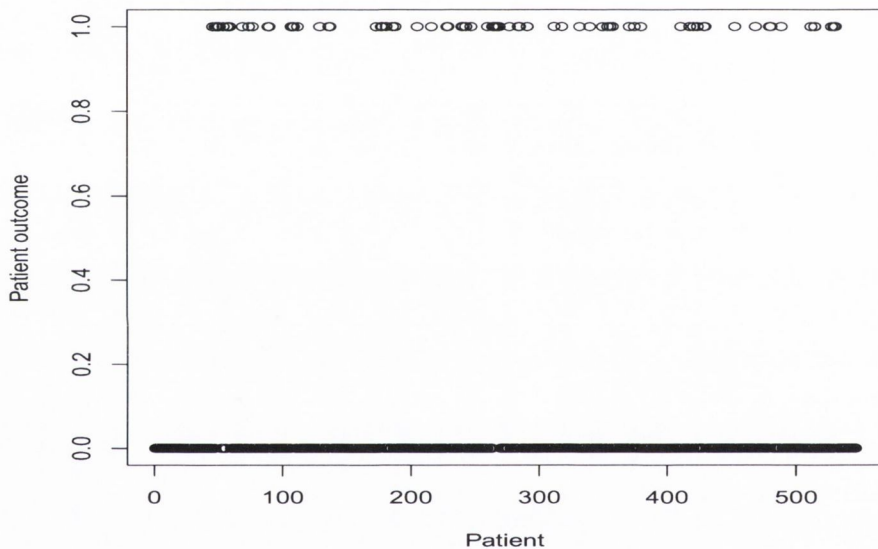


Figure 4.11: Survival outcomes for patients admitted to hospital following a heart attack. A sample of 546 outpatients with 73 deaths was observed with the primary endpoint of the trial being in-hospital death from any cause.

The primary aim of Souza & Migon (2010), was the development of a Bayesian binary regression model to predict the probability of patient death after hospital admission following a heart attack. A secondary aim of the study was the identification of outliers, namely patients for whom data upon admission was possibly misrecorded, leading to models with enhanced predictive performance.

The dataset, introduced in Section 1.2.2, consists of demographic and medical variables observed upon patient admission to hospital. The resulting patient outcomes, namely survival (0) and death (1), are presented in Figure 4.11. Information on 11 predictor variables for in-hospital mortality is available. These include continuous variables such as age, measured in

years and dichotomous variables such as sex, smoking history, heart attack history, diabetes and whether a history of arterial hypertension exists. Additional information is derived from electromagnetic examination of the patient subsequent to hospital admission; this includes a categorical measure for the severity of the heart attack which we henceforth denote “Killip”.

4.4.1 Model

Souza & Migon (2010) undertake a variable selection procedure, making use of Bayes factors to discriminate among competitive models as well as examining them for their predictive accuracy. The model with the best predictive ability, as determined by the authors, included the 9 variables age, sex, heart attack history, history of arterial hypertension, smoking habit, the Killip class on admission and the interactions age \times hypertension, sex \times hypertension and hypertension \times heart attack history. An additional random effect term for each observation, u_i , was incorporated into the model to capture any outstanding, unexplained variability.

In Souza & Migon (2010) three different prior specifications are proposed for the u_i . Our proposed approach has most in common with model M_1 of the authors where $u_i \sim N(0, \sigma^2)$. The resulting model for the data is:

$$z_i = X_i^T B + u_i \quad (4.26)$$

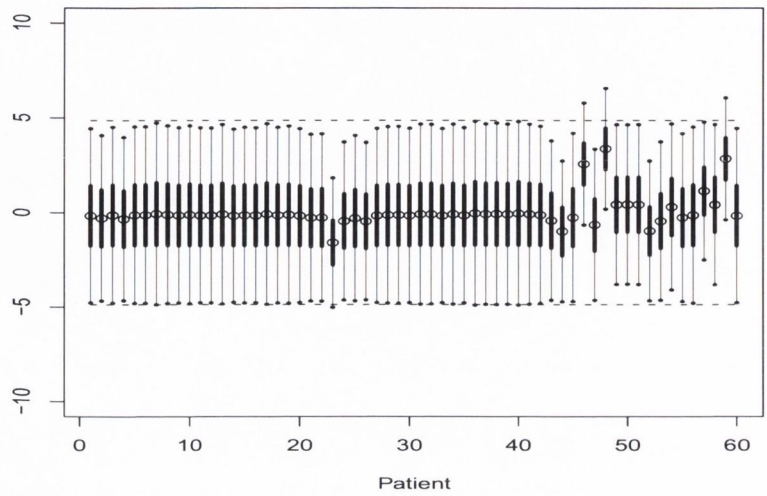
$$\pi(z_i|\theta) \sim N(X_i^T B, \sigma^2) \quad (4.27)$$

$$\pi(y_i|z_i) \sim \text{Bernoulli}\left(\frac{e^{z_i}}{1 + e^{z_i}}\right) \quad (4.28)$$

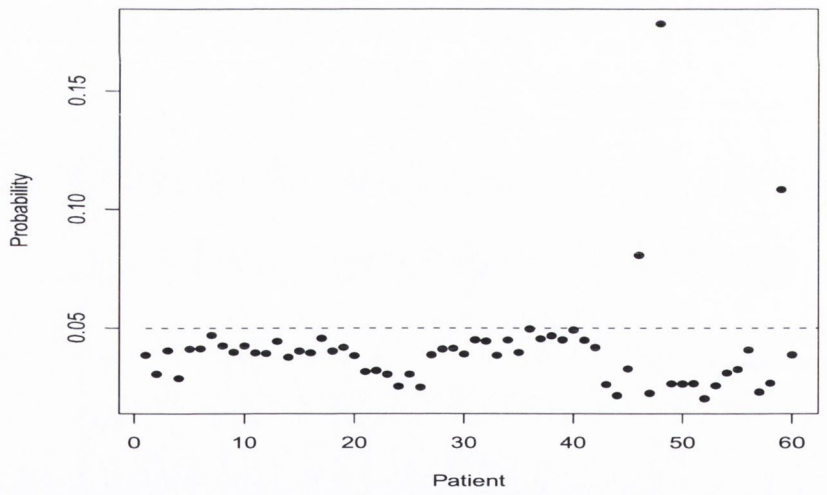
$B = \{\beta_0, \beta_1, \dots, \beta_8\}$ are a set of regression parameters corresponding to the set of covariates, $X_i = \{1, x_{i1}, \dots, x_{i8}\}$ and θ represents the underlying model hyperparameters including σ^2 . As noted by Fong et al. (2007) and Roos & Held (2011), posterior inference on the variance parameter of the random effect terms in logistic regression models is strongly influenced by the prior specification; there is very little information in the Bernoulli counts and hence the model is very prior sensitive (Simpson et al. 2011). As a result, an informative, $\sigma^2 \sim \Gamma(6, 1)$ prior is placed on the variance parameter of the random effects. Conversely, each of the regression parameters is supplied with a non-informative prior.

4.4.2 Results

Figure 4.12 plots the modal posterior random effects of the first 60 observations; the results correspond well to those produced in Figure 1 of Souza & Migon (2010). One of the main benefits of our approach is the expression of the posterior random effect terms as mixtures of Gaussians - as a result we can harness Gaussian residual theory to provide critical regions by



(a)



(b)

Figure 4.12: (a) Posterior random effects (o) for the first 60 patients with 50% (—) & 95% (-) highest posterior density regions. $\pm 1.96\sigma$ bounds are represented by (--). Posterior distributions which *significantly* cross the error bounds are identified as outliers. σ here is fixed at its posterior modal value of 2.48.

(b) The resulting posterior outlying probabilities (•) as well as the 5% critical bound (--). Observations with outlying probabilities greater than the bound are identified as outliers.

which to assess outlying activity. In Figure 4.12 we plot the posterior outlying probabilities corresponding to the first 60 observations. There are three outliers identified which corresponds exactly with the results obtained by Souza & Migon (2010). However, most importantly, the run time for their MCMC based approach was two hours; our results were obtained in a matter of seconds. The outlying probabilities through the use of an objective critical bound provided by Gaussian residual theory; none of the outliers are subjectively chosen.

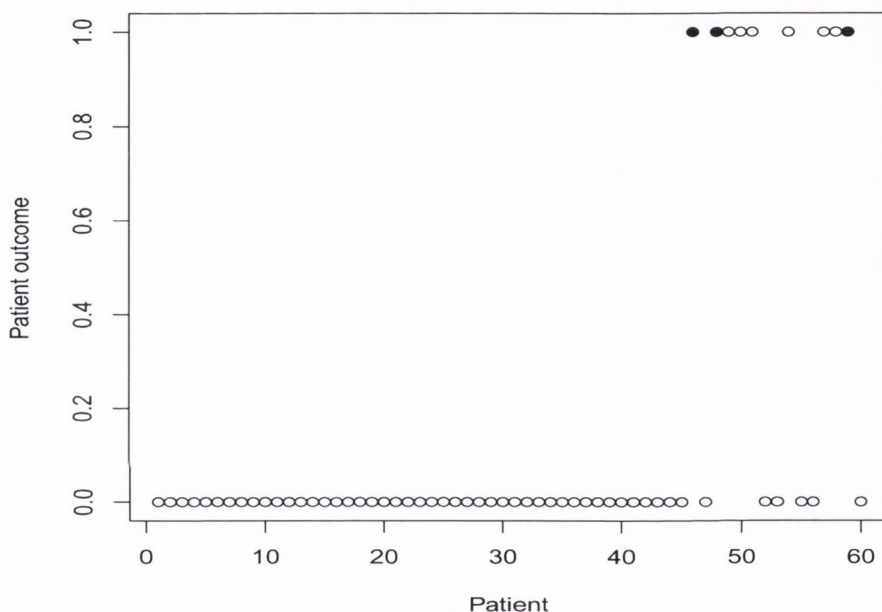


Figure 4.13: Survival outcomes (\circ) for the first 60 patients with 3 identified outliers (\bullet).

This contrasts with the results of Souza & Migon (2010), who identify as outliers the 27 observations (all corresponding to deaths) with the largest probabilities of requiring an extra random effect to capture excess variability. This corresponds to exactly 5% of the observed dataset; the justification for the ad-hoc cut off point used in outlier determination is weak and appears motivated by the expectation, under standard Gaussian residual theory, that approximately 5% of the observations may be considered potentially outlying. If the subjective cut-off point of the authors is lowered slightly, an additional 13 observations are included in the outlying dataset.

Conversely, our approach flags 39 of the 546 observations as requiring further investigation which includes all 27 counts identified by Souza & Migon (2010). This represents approximately 7% of the complete dataset - this represents approximately half the amount of observations we would expect given the simulation studies previously considered in Figure 4.7 (a) for $N = 1$. A possible reason for the flagging of less observations, than would be expected

under the Gaussian framework, is provided by visual inspection of the posterior random effect terms. In Figure 4.14(a) we provide a quantile-quantile of the mean posterior random effects. We observe that the posterior random effects do not follow a Gaussian distribution but appear to be generated from a mixtures distribution, perhaps reflecting the possible absence of an important predictor covariate in the model.

Using the Mclust (Fraley & Raftery 2007) package in R for cluster analysis, the best fitting model identifies 4 Gaussian clusters in the posterior random effects. In Figure 4.14(b) we simulate a number of random effects from the mixture distribution identified by the Mclust package which we overlay on the inferred posterior random effects. The results are seen to correspond almost exactly.

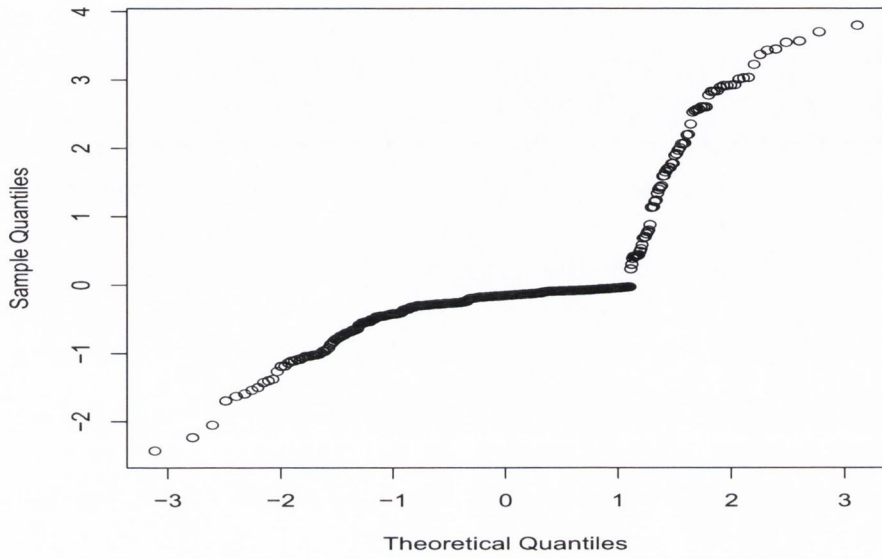
The quantile-quantile plots provide a quick method of model validation, indicating that a more appropriate model for the residual trend in the random effect terms is a mixture of four Gaussians - interestingly, the best fitting model of the authors was one which employed a bivariate mixture for the *a priori* random effects. However, as previously stated, the trend structure observed in the model residuals may also indicate the presence of an explanatory variable which was not recorded at the patient admission stage.

4.5 Conclusions

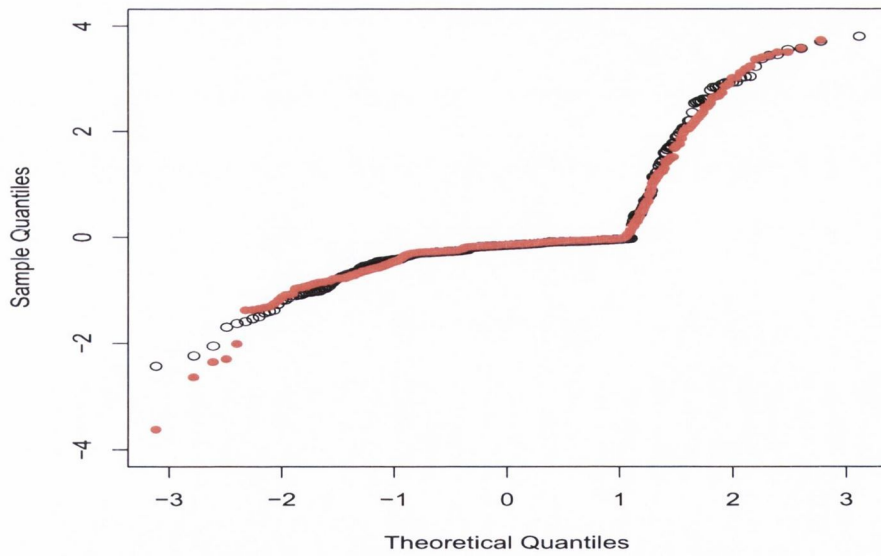
Existing methods for Bayesian residual analysis in the presence of discrete, non-Gaussian count observations are subject to the same issue - it is difficult to define critical bounds by which outliers can be objectively and systematically identified. Furthermore, though several authors consider the use of Gaussian random effect terms for outlier detection purposes, none appear to consider their use as a tool for quick, visual validation of the fitted models.

In this chapter, we have sought to address these issues through the development of a statistical methodology for residual analysis and outlier detection in the non-Gaussian data setting. This methodology is built upon the incorporation of Gaussian random effect terms into models to capture possible overdispersion in the counts data, with the posterior random effect terms treated as a “surrogate” for classical residuals. The approximation of the posterior random effect terms in Gaussian form provides access to standard Gaussian residual theory and to classical residual analysis tools such as quantile-quantile plots of the posterior random effects.

There are several benefits of the proposed methodology - the use of fast approximate Bayesian inference algorithms facilitates the quick fit and comparison of multiple models for the data in question. Objective critical bounds for systematic outlier detection are available from the harnessing of Gaussian residual theory. Additionally, the setting up of models in the generalised linear model framework demonstrates that the approach is possibly compatible with models in addition to the Poisson and Binomial models considered in this chapter; this



(a)



(b)

Figure 4.14: (a) Q-Q plot of $E(U|Y)$ (\circ) & (b) Q-Q plot of $E(U|Y)$ (\circ) overlain with simulations (\bullet) from a distribution comprised of a weighed mixtures of four Gaussians with parameters estimated from mclust.

theory is the subject of ongoing investigation.

Whilst a rigorous mathematical formulation of the properties of the proposed methodology has not been presented here, a series of directed simulated data studies have provided an indication of certain properties or features of the approach. Performance failings are very evident, if the overdispersion present in the data is small in magnitude, the random effect terms are masked by the discrete nature of the response and are difficult to detect. The outlier detection properties of the approach in such settings is observed to deteriorate. Conversely, in the presence of overdispersion which is significant in magnitude, the approach appears to have good residual analysis and outlier detection properties. In the context of the presented Poisson and Binomial regression examples, the other features which affect performance appear to be the values of the counts under consideration, with studies involving many low count values providing a particular challenge to the methodology.

An application to a real dataset has shown the power of the developed approach for Bayesian residual analysis in the non-Gaussian setting. This application represents a challenging test of the methodology, due to the binary nature of the response variable. Inference on all model parameters is conducted in a matter of seconds and the provision of an automatic, explicit bound for outlier detection purposes represents an advance over the methods of Souza & Migon (2010). Finally, the use of quantile-quantile plots of the mean posterior random effects, provide a quick visual method of determining that the *a priori* modelling assumption of a univariate Gaussian distribution for the random effect terms is inappropriate.

Chapter 5

Models for Multivariate Observational Data

Statistical calibration problems involving highly multivariate observational data sets are *multivariate inverse inference* problems. Specifically, the forward stage involves the calibration of multivariate statistical models for the relationship between model covariates and the highly multivariate response. The calibrated models are then used inversely, to make inferences on the unobserved covariates corresponding to new data for which such information is unknown.

However, the statistical modelling of multivariate data poses many challenges of computation and model choice; fully Bayesian inference on the parameters of multivariate models for the highly multivariate response can be tremendously slow or even infeasible. This problem becomes acute in the presence of high dimensional data sets, such as the RS10 pollen and climate data set, enforcing compromises in the complexity of models that are ultimately considered. As we will observe in the following sections, these compromises in model choice can adversely impact the prediction accuracy of calibrated models at the inverse stage.

The novel contributions in this chapter relate to statistical modelling. Specifically, we examine the introduced issues regarding the statistical modelling of multivariate response data and extend existing modelling methodology to address these issues. We illustrate that the optimal hierarchical (nesting) structure for compositional counts data, in terms of prediction accuracy at the inverse stage, can be learned from the observed data. We detail the statistical inconsistency of existing zero-inflation models for compositional count data and address this issue via the construction of a modelling framework that considers both zero and N-inflation of the counts simultaneously. Inference details are suppressed throughout this chapter and model parameters are assumed known save to mention details of a computational nature; the focus is the impact of different model choices at the forward stage on the accuracy of the predictions produced at the inverse stage.

This chapter is organized as follows; in Section 5.1 we consider models for multivariate

response data. We discuss the assumption of conditional independence of model components, which enables the decomposition of multivariate joint models into a series of separate univariate models and detail how this assumption introduces many computational conveniences at both the forward and inverse stage of the multivariate inverse inference problem.

In Section 5.2 and Section 5.3 we discuss models for multivariate counts data and explore the effect of dependence structure at various model levels on the accuracy of the predictions produced at the inverse stage. We focus in particular on models for compositional data, where dependence between model components is introduced by the data collection process. Given the assumption of conditional independence of model components, we illustrate that the omission of dependence structure leads to prediction intervals that are “too narrow”, resulting in a deterioration in the prediction accuracy of calibrated models.

In Section 5.4 we discuss hierarchical or “nesting” structures for the statistically efficient decomposition of multivariate joint models involving compositional data into smaller, univariate inference tasks. We investigate a number of hierarchical structures for model decomposition and detail how the “best” hierarchical structure, in terms of predictive power at the inverse stage, can be learned from the data.

In Section 5.5 we examine hierarchical models for zero inflated compositional data. We detail how, in the Multinomial setting, zero-inflation of counts corresponding to one group can lead to N-inflation of another. A specialised likelihood function is introduced to account for this extra source of variability.

5.1 Multivariate Observational Data

In a typical spatial regression problem, interest generally lies in making inferences on the latent, unobserved response surfaces which describe the relationship between some recorded covariates (spatial location) and the observed, multivariate response. Conversely, in this thesis, the latent response surfaces themselves are not of substantial interest. We are instead primarily focused on *multivariate inverse inference*; the inferred multivariate response surfaces are required in order make inferences, inversely, on the unobserved covariates (fossil climate) corresponding to a set of “new“ (fossil pollen) responses. However, the initial (forward stage) which involves the calibration of multivariate models for the covariate-response relationship can be extremely challenging; with regard to the palaeoclimate reconstruction problem, the calibration dataset, available for model fitting, is highly multivariate.

In Equation 5.2 we present an example of a highly multivariate dataset which will serve as a tool to simply explain features of the inverse problem in following sections. The dataset consists of multivariate observations $Y = (Y_1, \dots, Y_m)$ which are (spatially) indexed by the (possibly multivariate) climate $C = (c_1, \dots, c_n)$. In the context of the motivating palaeoclimate reconstruction problem, each column $Y_i \in Y$ represents the observed pollen response to

recorded C for each plant taxon.

$$\text{Data} = (Y_1, \dots, Y_m, C) \quad (5.1)$$

$$= \left(\begin{array}{ccc|c} y_{11} & \cdots & y_{m1} & c_1 \\ \vdots & \vdots & \vdots & \vdots \\ y_{1n} & \cdots & y_{mn} & c_n \end{array} \right) \quad (5.2)$$

5.1.1 Multivariate Observational Models

To recap, at the forward, model fitting stage the objective is to calibrate proposed models for the response-covariate interaction. If we assume that the observations in Equation 5.2 are realizations from a multivariate Gaussian distribution with unknown mean, X and observed with error ε , a suitable model for the multivariate dataset may be of the form presented in Equation 5.3.

In the context of the motivating palaeoclimate climate problem the model has the following interpretation; each column of $X = (X_1, \dots, X_m)$ represents the smooth, unknown spatial response surface governing the climate-pollen response for each plant taxon, with each ε_{ij} an independent, non-spatial error term.

$$Y = X(C) + \varepsilon \quad (5.3)$$

$$\left(\begin{array}{ccc} y_{11} & \cdots & y_{m1} \\ \vdots & \vdots & \vdots \\ y_{1n} & \cdots & y_{mn} \end{array} \right) = \left(\begin{array}{ccc} X_1(c_1) & \cdots & X_m(c_1) \\ \vdots & \vdots & \vdots \\ X_1(c_n) & \cdots & X_m(c_n) \end{array} \right) + \left(\begin{array}{ccc} \varepsilon_{11} & \cdots & \varepsilon_{m1} \\ \vdots & \vdots & \vdots \\ \varepsilon_{1n} & \cdots & \varepsilon_{mn} \end{array} \right) \quad (5.4)$$

$$\pi(X|Y, C) = \pi(Y|X, C)\pi(X|C) \quad (5.5)$$

Accounting for sources of correlation structure in highly multivariate Y may require the use of extremely complex models. In the context of compositional data, implicit correlation structure is introduced by the data collection process; the constraint that the rows of Y in Equation 5.2 must sum to a specified total N introduces dependence structure in the multivariate Y . Equally, this sum constraint also introduces correlation between the multivariate response surfaces of X ; if the counts for an individual plant taxon increase, simultaneously the counts must decrease for all others, implying a strong negative correlation structure across

the response surfaces.

With regard to the latent field, if we assign a multivariate Gaussian prior distribution to X , $X \sim MVN(\mu, \Sigma_X)$, possible dependence structure can be captured through the covariance matrix Σ_X . The choice of a complex interaction structure for Σ_X generally results in a huge increase in the number of hyperparameters required to fully specify the model, corresponding to the inclusion of extra correlation parameters which govern the correlation structure between individual taxa. This potentially results in large dense matrices for Σ_X , further resulting in an increase in computation time with regard to matrix operations involving Σ_X .

Inference procedures for models with complex dependence structures in the latent field X or in the multivariate observations Y can be tremendously slow due to the sheer magnitude of the number of parameters we must jointly infer. Indeed, the computational burden of jointly inferring all model parameters and covariance structures, in the context of large datasets, may be computationally infeasible, thus imposing trade offs between model complexity and finite machine computation time.

For example, with regard to the motivating palaeoclimate reconstruction problem, the model calibration dataset considered in this thesis consists of 7742 observations for each of 28 different plant taxa, providing 216776 observations in total. The use of standard geostatistical modelling methods will encounter the “big n” problem (Bannerjee et al. 2004); the consideration of all response surfaces jointly will require the manipulation of a covariance matrix Σ_X of dimension 216776×216776 , with the use of complex covariance structures between the latent response surfaces rendering this an extremely dense matrix. The computational cost of inferring the parameters of multivariate joint model is thus too computationally expensive to consider, enforcing the consideration of simpler models or approximation methods which greatly reduce the complexity of the inference task.

5.1.2 Univariate Models

The simplest approach to dealing with this challenging problem is to decompose the multivariate joint model into a subset of smaller, independent univariate models. At the forward stage this involves the decomposition of the multivariate joint inference problem for the latent field, $\pi(X|Y, C)$, into a sequence of independent inference tasks, i.e. $\pi(X|Y, C) = \prod_i^m \pi(X_i|Y_i, C)$ - the response surface for each plant taxa is thus inferred independently of the others, Salter-Townshend (2009) denotes this approach as “inference-via-the-marginals”. The multivariate model introduced in Equation 5.3 is thus decomposed into the one presented below; inference on the model parameters for each subset can then be carried out in isolation, drastically reducing the computational complexity of model fitting.

$$Y_i = X_i(C) + \varepsilon_i \quad (5.6)$$

$$\begin{pmatrix} y_{i1} \\ \vdots \\ y_{in} \end{pmatrix} = \begin{pmatrix} X_i(c_1) \\ \vdots \\ X_i(c_n) \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in} \end{pmatrix} \quad (5.7)$$

$$\pi(X|Y, C) = \prod_{i=1}^m \pi(Y_i|X_i, C)\pi(X_i|C) \quad (5.8)$$

This approach is based on the assumption of conditional independence of model components; the multivariate observations Y are assumed to be conditionally independent given the latent field X , i.e. $\pi(Y|X) = \prod_i^m \prod_j^n \pi(y_{ij}|X_i(c_j))$, in turn the latent parameters for individual plant taxa are assumed conditionally independent given the spatial location, $\pi(X|C) = \prod_i^m \pi(X_i|C)$. There are no interaction terms included in the model for modelling possible dependence between the univariate models at the likelihood level or in the latent field (the covariance matrix Σ_X is thus block diagonal).

If the individual model components are truly independent, then inferences on the individual model components in isolation will be equivalent to those obtained when making inference on the the full model jointly. However, if interaction exists at any level between the (assumed independent) model components, inference on the parameters of the univariate models will only act as an approximation to joint inference on the full model, possibly leading to statistically inefficient, or sub-optimal, parameter inferences. As we will observe in the following sections, the quality of this approximation is dependent on the strength of the true dependence structure.

5.1.3 Model Inversion

The primary interest in this thesis is the use of calibrated models, inversely, for prediction given new data for which the true model covariates are unknown. Specifically, given the model training data (Y, C) , the goal is to infer the unknown climate c^{new} corresponding to a set of new, univariate pollen responses $Y^{\text{new}} = (y_1^{\text{new}}, \dots, y_m^{\text{new}})$.

Written explicitly:

$$\pi(c^{\text{new}}|Y^{\text{new}}, Y, C) \propto \int_X \pi(Y^{\text{new}}|X, c^{\text{new}})\pi(X|Y, C)dX \quad (5.9)$$

If the multivariate integral in Equation 5.9 is not known analytically, as is frequently the case in this thesis, computationally intensive sampling algorithms may be used for its evaluation. However, given the conditional independence assumption introduced in the previous section, the multivariate joint probability distributions, $\pi(Y^{\text{new}}|X, c^{\text{new}})$ and $\pi(X|Y, C)$ in Equation 5.9 decompose into the product of independent parts, further enabling the multivariate joint integration step to be decomposed into the product of smaller, less computationally intensive univariate integrals in Equation 5.12 which can be evaluated deterministically. This results in extensive time savings at the inverse stage.

$$\pi(c^{\text{new}}|Y^{\text{new}}, Y, C) = \prod_i^m \pi(c^{\text{new}}|y_i^{\text{new}}, Y, C) \quad (5.10)$$

$$= \prod_i^m \int_{X_i} \pi(c^{\text{new}}, X_i|y_i^{\text{new}}, Y, C) dX_i \quad (5.11)$$

$$\propto \prod_i^m \int_{X_i} \pi(y_i^{\text{new}}|X_i, c^{\text{new}}) \pi(X_i|Y, C) dX_i \quad (5.12)$$

The decomposition of joint multivariate models into a series of independent univariate models is based on the assumption of the conditional independence of model components. The resulting absence of dependence structure at the likelihood level or between the latent response surfaces in the multivariate joint model introduces many computational conveniences at both the forward and inverse stages. However, as we explore in the following sections, if this model decomposition is used erroneously, inefficiencies will occur in the statistical inferences derived from the model.

5.2 Multivariate Counts Data

In the previous section, models for multivariate observational datasets were introduced where the underlying distribution of the observations was assumed Gaussian in nature. If the distribution of the observational dataset is other than Gaussian, as is the case for the motivating palaeoclimate application, this introduces an additional complexity to model inference tasks; posterior distributions for the parameters of the latent field are typically unavailable in closed form. In the following we consider the case where the observational dataset, introduced in Equation 5.2, consists of multivariate Poisson count observations.

We extend the multivariate model, introduced in Equation 5.1.1, to the Poisson count setting. The individual counts are modelled as Poisson distributed and linked to the latent field X through the use of log-link function. The latent field is modelled as previously, with a multivariate Gaussian prior used to capture spatial smoothness. The hierarchical model for

the data is:

$$Y \sim \text{Poisson}(\lambda) \tag{5.13}$$

$$\lambda = \exp(X) \tag{5.14}$$

$$X \sim \text{MVN}(\mu, \Sigma_X) \tag{5.15}$$

Counts data that exhibits signs of overdispersion are of particular interest in this thesis; with regard to the motivating palaeoclimate reconstruction problem, the pollen count observations available for model fitting, appear to contain more variability than that expected by the usual exponential family models. As per Section 3.5.3, if the empirical variance of the observed counts is significantly greater than the empirical mean, this may indicate the presence of excess variability and the counts are said to be “overdispersed”.

5.2.1 Modelling Overdispersion

As introduced in Section 3.5.3, a simple method of accounting for this excess variability is through the addition of mean-zero non-spatial Gaussian random effects U , one for each count observation, to the model. The addition of random effect terms to the model thus induces overdispersion of the latent field with regard to the spatial component X .

For example, say the data Y is Poisson with rate parameters λ and λ , constrained to be non-negative, is comprised of $\exp(X + U)$, the hierarchical model for the data in question is:

$$Y \sim \text{Poisson}(\lambda) \tag{5.16}$$

$$\lambda = \exp(X + U) \tag{5.17}$$

$$X \sim \text{MVN}(\mu, \Sigma_X) \tag{5.18}$$

$$U \sim \text{MVN}(0, \Sigma_U) \tag{5.19}$$

The computational problems of this particular approach are as follows; the incorporation of random effects terms into the model leads to a substantial increase in the number of latent parameters requiring inference. In the presence of large amounts of data and in the context of multiple taxa being considered simultaneously, joint models for the data, which additionally incorporate dependence structure in the latent field, are much too computationally expensive to consider. One solution to this computational problem is to consider the overdispersed count data for each taxa independently of the others and fit univariate models instead, at the cost of disregarding the dependence structure that can be built into joint models for the data in question.

In the following section we focus on the scenario where dependence structure exists in the latent field, specifically, we aim to explore the effect of ignoring dependence structure in the latent field and the resulting impact on prediction accuracy of calibrated models at the inverse stage.

5.2.2 Sensitivity to Dependence Structure in the Latent Field

In the following we build upon the work of Salter-Townshend (2009) and use simulated data to illustrate the effect of (unaccounted for) dependence structure in the latent field at the forward stage on the predictive distributions produced at the inverse stage. Overdispersed Poisson count observations Y are generated given a latent field, X , comprised of 10 identical response surfaces (i.e. the same response surface is used 10 times) defined on a regular grid of 100 spatial locations and multivariate random effects U . Given the simulated data, the first step is to calibrate a model for the location-response relationship.

A hierarchical model for the data in question is:

$$Y \sim \text{Poisson}(\lambda) \tag{5.20}$$

$$\lambda = \exp(X + U) \tag{5.21}$$

$$X \sim \text{MVN}(\mu, \Sigma_X) \tag{5.22}$$

$$U \sim \text{MVN}(0, \Sigma_U) \tag{5.23}$$

Each row (j) of Y is a vector of counts of length 10, (y_{j1}, \dots, y_{j10}) , indexed by univariate spatial location c_i . The prior for X is a multivariate Gaussian process of dimension 10 with dependence structure between the individual response surfaces included in Σ_X . The random effects are treated as independent within an individual taxon but dependent across taxa with covariance structure across taxa modelled in Σ_U .

The modelling of dependence structure in the latent field between individual response surfaces and taxon random effects is not a simple task. Dependence structure, modelled by the inclusion of interaction terms in the respective covariance matrices, greatly increases the computational burden of the inference task by virtue of the large number of covariance parameters which we must additionally infer. Furthermore, the inclusion of dependence structure in the multivariate model for univariate model components introduces the constraint that all model parameters must be jointly inferred.

The computational cost of inferring all model parameters jointly, even in the simple example considered here is extensive. However, the assumption of conditional independence of model components greatly reduces the computational burden - dependence structure is ignored, the response surfaces which encompass the latent field are not independent given location but

are modelled as such in order to simplify the inference task and facilitate decomposition of the inference problem. The latent parameters corresponding to each individual taxa are thus inferred independently of all others.

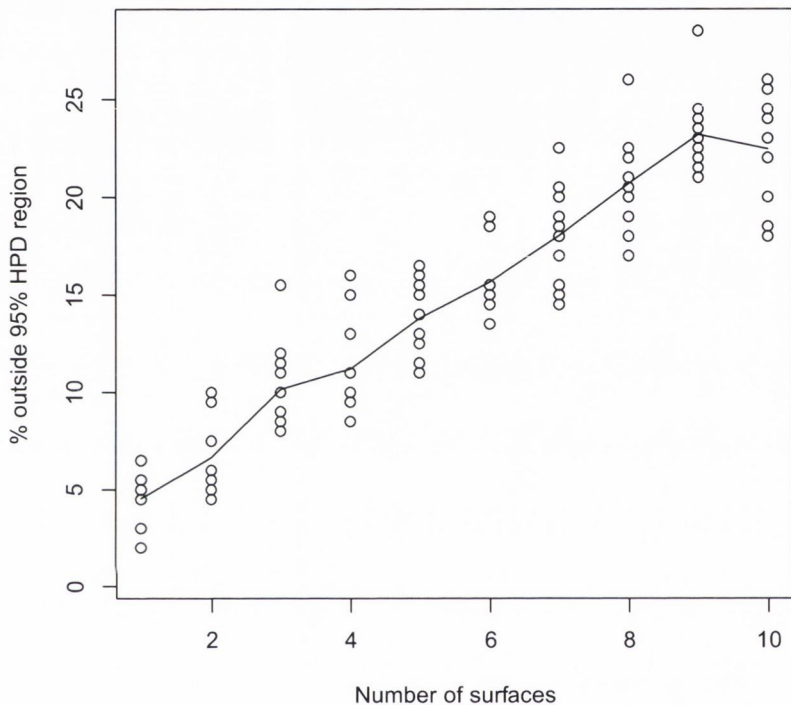


Figure 5.1: A plot of the percentage of locations falling outside their 95% predictive distribution as the number of surfaces jointly considered increases. Interaction is related to the use of the same, unimodel, latent response surface to generate all counts and through dependence in the random effects used for overdispersion.

In Figure 5.1 we present the result of this modelling choice. The incorrect assumption of independence of model components, given known model parameters, is manifested in the plot as an increase in the number of observations falling outside their 95% highest posterior density predictive (HPD) distribution region for climate; furthermore, the error rate increases with each additional taxon considered. Using the decomposed model, the assumption is made that there is no dependence structure between the response surfaces which comprise the latent field or between taxon random effects.

Conversely, by virtue of the data generating procedure there is a strong dependence structure between the latent parameters corresponding to each taxa. The same smooth response surface and random effects are used to simulate the counts for each plant taxa, the latent

parameters for each additional taxa considered are thus fully correlated with correlation equal to 1. The univariate approximation to the multivariate model does not account for this dependence structure, treating the data and latent parameters for each taxa as conditionally independent, which they are clearly not. This is manifested as a deterioration in predictive power at the inverse stage.

However, other model evaluation statistics help provide a useful insight into the performance of the model decomposition used. We note that, in spite of the number of observations detected as falling outside the 95% HPD region (Δ) increasing with each additional taxa considered, the absolute distance of the mode of the predictions on location produced to the true locations decreases, on average, with each additional taxa considered. Furthermore, the mean squared error of prediction is also seen to decrease.

These results seem to indicate that the climate predictions produced become increasingly accurate with each additional taxa considered. However, the erroneous treatment of the spatial response surfaces as conditionally independent given spatial location leads to posterior distributions for location that are not sufficiently conservative. The 95% posterior distributions on spatial location given the observed data are too narrow, brought about by not modelling the inherent dependence structure introduced by the data generation process.

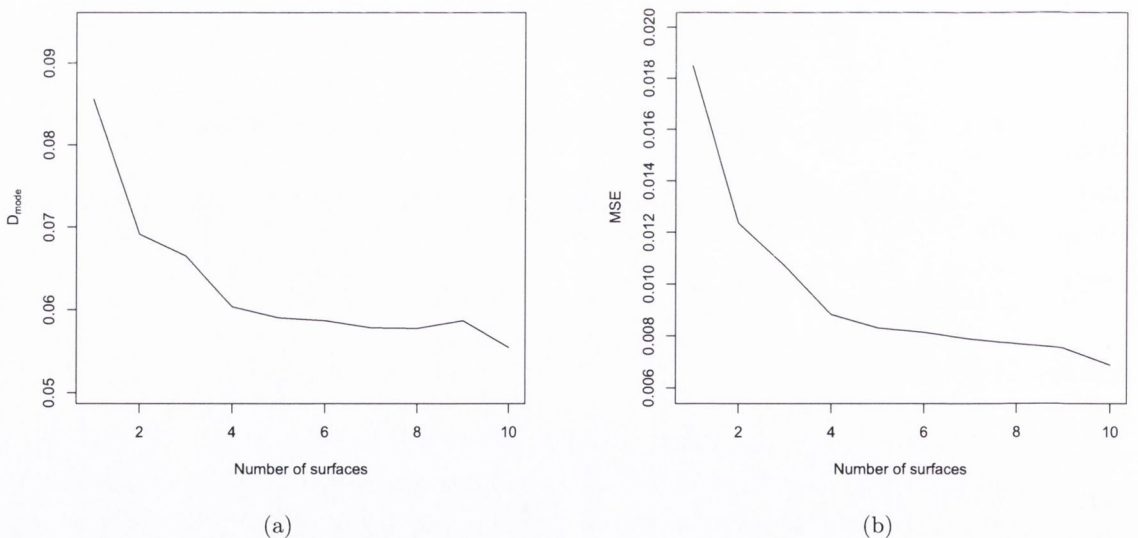


Figure 5.2: Plots of the (a) $MSEP$ and (b) the distance to the mode (D_{mode}) for increasing number of taxa considered. For increasing number of taxa the placement of the inverse predictive distributions becomes increasingly accurate and peaked, as indicated by the decreasing $MSEP$ and D_{mode} statistics.

The simplest approach to inference on the parameters of large multivariate model is to

perform inference on the taxon specific components of the latent field independently for both the forward and inverse stage of the problem, ignoring possible dependence structure in the latent field. In this section we have observed that the predictive posteriors produced at the inverse stage are quite sensitive to this approach. As previously discussed, a potential source of dependence structure in the latent field is the analysis of data which is subject to sum constraints, such as compositional data. In the following section we examine this issue further.

5.3 Dependence in the Likelihood

In the previous section, the impact of (unaccounted for) dependence structure in the latent field at the forward stage, on the predictions produced at the inverse stage was illustrated. One possible manner in which this dependence structure can arise is through the data collection process, which we refer to as *dependence in the likelihood*; in the context of compositional data, the sum constraint on the observed data introduces a strong implied dependence structure into the latent field which likelihood models for the data must take into account.

In the following sections we discuss this issue further, introducing Multinomial models for the observational data and investigating the pitfalls of conditional independence assumptions between individual model components in the presence of sum constraints.

5.3.1 Multinomial Likelihood Function

The Multinomial distribution is a multivariate likelihood with known degree of dependence (covariance) between model components. As opposed to situations where the data are only presented in terms of proportions, if the totals for each row of the counts vector are known, then there is extra information available in the data as opposed to the base compositional data setting. The probability distribution of the counts Y , given the total N is given by the Multinomial distribution.

The Multinomial likelihood can be expressed as the product of independent Poisson distributions constrained by conditioning on the sum being equal to the total count N . In Equation 5.24 we present the likelihood kernel for the Multinomial likelihood function and illustrate how this can be re-expressed in terms of the product of independent Poisson distributions.

$$\pi(Y|P, N) \propto \prod_{i=1}^m p_i^{Y_i} \tag{5.24}$$

$$\equiv \frac{\prod_{i=1}^m e^{-\lambda_i} \lambda_i^{Y_i}}{e^{-(\sum_{i=1}^m \lambda_i)} (\sum_{i=1}^m \lambda_i)^Y} \tag{5.25}$$

$$= \frac{\prod_{i=1}^m \text{Poisson}(Y_i, \lambda_i)}{\text{Poisson}(N, N)} \quad (5.26)$$

Here $\lambda = \{\lambda_1, \dots, \lambda_m\}$ represent the rate parameters for each of the independent Poisson distributions where each $\lambda_i = Np_i$ and simple mathematical workings show that the correct full joint Multinomial likelihood is returned from the product of the marginals subject to the constraint. Of course the p_i are not known; each y_i represents an indirect observation of the unobserved, underlying latent field and can be used to make inferences on the unknown p_i .

5.3.2 Modelling the Multinomial Response

The representation of the Multinomial likelihood function as the product of independent Poissons provide conveniences as regards statistical models for the Multinomial response. As opposed to the probabilities P , which are constrained to lie between zero and one, the λ_i 's are constrained only to lie on the positive real line; through the use of a log-link (see Section 3.5.1) function, the λ_i 's can be directly related to the unconstrained latent field X , which is indexed by the spatial location C .

A typical hierarchical model is:

$$Y \sim \frac{\prod_{i=1}^m \text{Poisson}(Y_i, \lambda_i)}{\text{Poisson}((\sum_{i=1}^m \lambda_i), N)} \quad (5.27)$$

$$\lambda_i = \exp(X_i) \quad (5.28)$$

$$X_i \sim MVN(\mu_{X_i}, Q_{X_i}) \quad (5.29)$$

Dependence is assumed only to arise through the likelihood, thus independent multivariate normal priors may be used for each of the m latent response surfaces X_i which make up the latent field X . However, though the representation of the Multinomial as the product of independent Poisson distributions introduces certain modelling conveniences, model parameters must still be inferred jointly. The likelihood cannot be decomposed into the product of conditionally independent parts due to the sum constraint.

It is quite simple to extend the modelling framework introduced in Equation 5.27-5.29 to account for overdispersion of the Multinomial count outcomes. Though computationally burdensome, the addition of Gaussian random effect terms to capture the overdispersion is easily done through the addition of another level to the model hierarchy:

$$Y \sim \frac{\prod_{i=1}^m \text{Poisson}(y_i, \lambda_i)}{\text{Poisson}((\sum_{i=1}^m \lambda_i), N)} \quad (5.30)$$

$$\lambda_i = \exp(X(c_i) + U_i) \quad (5.31)$$

$$X_i \sim MVN(\mu_{X_i}, Q_{X_i}) \quad (5.32)$$

$$U_i \sim MVN(0, \Sigma_U) \quad (5.33)$$

The response surfaces for each plant taxa are assumed *a priori* independent; the non-spatial random effects, used to model count overdispersion are assumed to consist of independently identically distributed Gaussian random effects and as a result Σ_U in Equation 5.33 is taken to be diagonal. As previously discussed, this supplies a flexible modelling strategy where individual taxa are allowed to display differing degrees of overdispersion whilst preserving the Multinomial likelihood, at the cost of inferring a large number of random effects (one for each count observation) in addition to an extra variance hyperparameter for each of the m plant taxa.

This contrasts greatly with the approach of Haslett et al. (2006) who use a compound-Multinomial model for modelling overdispersed pollen count vectors. Essentially a Dirichlet process prior on the probabilities P is mixed with a Multinomial likelihood for the count outcomes to form a compound-Multinomial distribution for the likelihood. As noted by Haslett et al. (2006), this results in the rather unsatisfactory constraint that the overdispersion experienced by each taxa is fixed to be *the same* across all taxa though maintaining the conditionally independent structure of the Multinomial likelihood.

Inference and Modelling Issues

The use of Multinomial models for the observed data introduces many inference and modelling problems. Though the Multinomial likelihood may be re-expressed as the product of independent Poisson distributions, the sum constraint, which corrects for the compositional nature of the data, requires that all model parameters be jointly inferred. In the presence of large amounts of data (the palaeoclimate reconstruction problem consists of $n = 7742$ observations for each of $m = 28$ plant taxa) the computational cost of simultaneously inferring all unknown model parameters is computationally infeasible. As a result approximations, in terms of likelihood models for the observed data, must be considered.

5.3.3 Sensitivity to Dependence Structure in the Likelihood

One such approximation to the joint (Multinomial) model likelihood is to simply approximate the Multinomial likelihood as the product of independent Poisson distributions without accounting for the sum constraint. Essentially:

$$\pi(Y|\lambda, N) \propto \frac{\prod_{i=1}^m e^{-\lambda_i} \lambda_i^{Y_i}}{e^{-(\sum_{i=1}^m \lambda_i)} (\sum_{i=1}^m \lambda_i)^Y} \quad (5.34)$$

$$\approx \prod_{i=1}^m e^{-\lambda_i} \lambda_i^{Y_i} \quad (5.35)$$

We denote this the ‘marginal model’ - as the likelihood is now taken to be the product of independent Poisson likelihoods, the specification of a prior model for the latent field that does not consider dependence structure between the latent response surfaces enables the decomposition of the joint inference task; inference on the model parameters for each taxa can be completed separately. The joint model is thus approximated by the product of the marginal models. In the following we investigate the properties of this approximation.

A special case of the Multinomial model is the case where $m = 2$, which is the Binomial model. In the following we generate some overdispersed Binomial data, conditional on the latent field and re-infer model parameters given the true model (Binomial) and the marginal model, where the counts are treated as independently Poisson distributed.

At each of 100 spatial locations, a set of Binomial counts, Y are generated with probabilities given by a logit transform of the latent field $w_i = x(c_i) + u_i$ where X is a smooth spatial surface defined over the location space and u_i is a simulated random effect from a $N(0, 1)$ distribution. Due to the random effects, the counts are overdispersed, displaying more variability than simple Binomial data.

Given the simulated data, model parameters are inferred for the correct Binomial model and the approximate marginal model and the calibrated models used to infer the location of further simulated model validation data for which the true location is known. The number of observations which lie outside their 95% predictive region for location given count is then compared for the two approaches with this process completed a large number of times (200), in order to build up a profile of the prediction properties of each approach. The results are presented in Figure 5.3

We observe in Figure 5.3 that, for the true Binomial model, approximately 5% of observations lie outside their 95% predictive regions for location given observation. Conversely, the corresponding error statistic for the Poisson marginal approximation to the Binomial is approximately 27% reflecting the result that the approximation of the Binomial likelihood by independent Poisson distributions fares poorly in this setting. The Poisson approach treats each set of counts as two independent Poisson observations whereas the Binomial approach acknowledges the fact that there is one less degree of freedom in each set of counts; given the total N , only one of the counts provides any information.

Additional information about the vagaries of the approximation can be learned from studying the performance of the approximate and joint approach with regard to the distance metrics, D_{mode} and the mean squared error of prediction, $MSEP$. In Figure 5.4 (a), we observe that the error for the D_{mode} statistic for both the Binomial model and the marginal Poisson model seem to produce predictions for location that are equally well located, i.e. they have similar

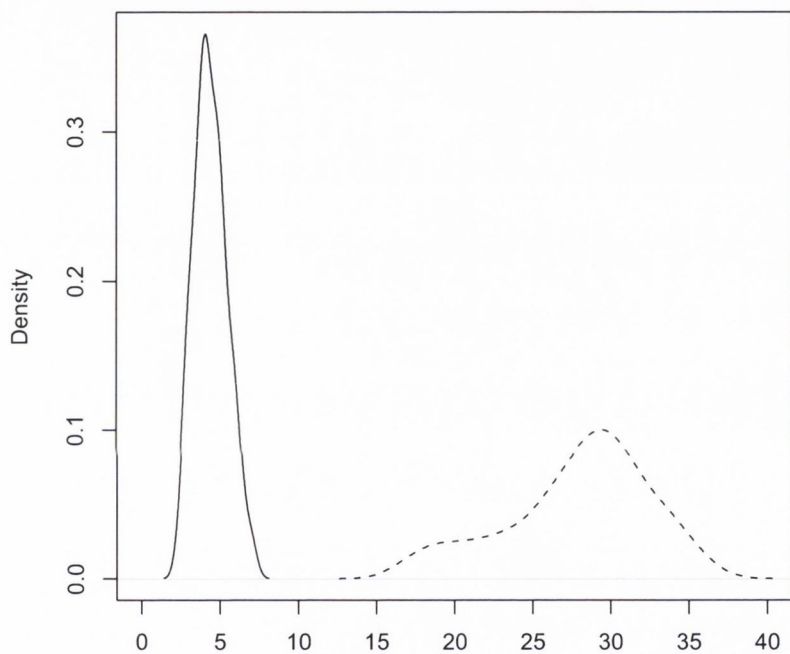


Figure 5.3: A plot of the percentage of locations falling outside their 95% predictive HPD distribution regions for the true model (solid line) where the Binomial nature of the data is addressed (sum constraint accounted for), versus the marginal model (dashed line) where the the counts are treated as independent, overdispersed Poisson observations. The error statistic is significantly higher for the Poisson model, which does not account for the strong dependence structure in the simulated data, implied by the sum constraint.

accuracy in predicting modes. In Figure 5.4 (b), we learn the cause of the poor predictive accuracy at the inverse stage; the posterior distributions for location produced by the marginal Poisson model, given each set of new counts, treat each count as independent information. As a result, the posterior predictive distribution for location is more peaked for the Poisson model than the Binomial model equivalent - the predictive distributions for location produced at the inverse stage are thus “too narrow” resulting in a deterioration in prediction accuracy.

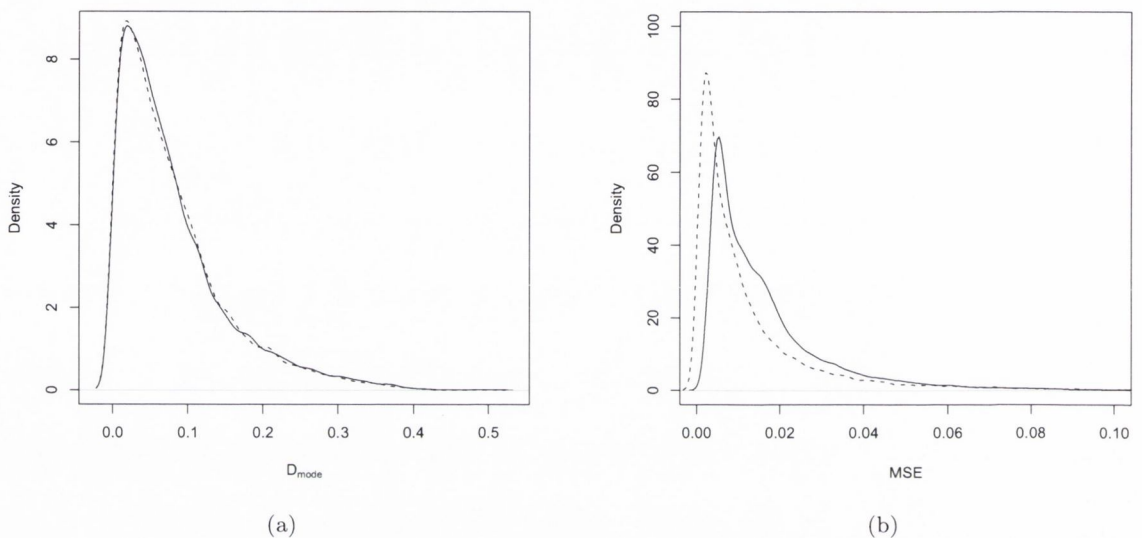


Figure 5.4: (a) Plots of the distance to the mode D_{mode} and (b) the $MSEP$ for the joint (solid line) and marginal (dashed line). Both models seem to predict the ‘true’ climate with equivalent accuracy, however the predictive posteriors on location produced using the marginal Poisson model are not conservative enough

In Figure 5.5, we additionally observe that the error statistic Δ , of the Poisson approximation to the Binomial is strongly dependent on the degree to which the data are overdispersed. The σ^2 parameter of the random effect terms is set as .5, i.e. $u_i \sim N(0, .5)$, representing a lower degree of overdispersion than the example considered in Figure 5.3. With the lower degree of overdispersion the accuracy of the decomposition is seen to improve substantially.

5.4 Decomposing Models Involving Multinomial Likelihoods

The previous section illustrated that naive decompositions of highly multivariate models, such as the assumption of the conditional independence of model components in the presence of Multinomial data, can lead to poor inference outcomes at the inverse stage. The sum constraint implied by the data collection process requires that the latent parameters for all model

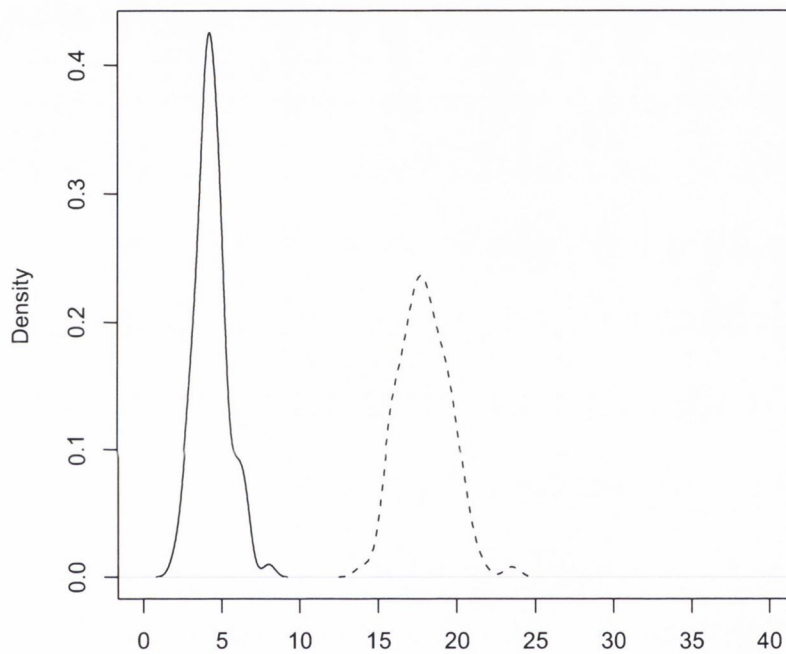


Figure 5.5: A plot of the percentage of locations falling outside their 95% predictive HPD distribution regions for the true model (solid line) where the Binomial nature of the data is addressed (sum constraint accounted for), versus the approximate model (dashed line) where the counts are treated as independent, overdispersed Poisson observations. The error statistic is lower for the Poisson approximation to the Binomial model than the example presented in Figure 5.5 due to a lower degree of overdispersion.

components must be jointly inferred.

In this section we study the inferentially efficient methods of Rodríguez (2007) for *decomposing* large multivariate models, involving Multinomial data, into the product of smaller, independent univariate models which represent much less challenging inference tasks.

5.4.1 Hierarchical or “Nesting” Structures

One possible strategy for decomposing joint models involving Multinomial response data, is to define a hierarchy of nested comparisons between subsets of the responses. As we will see in the following, the approach is quite attractive from a computational viewpoint; using this method, the multivariate joint model can be decomposed into a series of *disjoint* univariate models, the parameters of which can be inferred separately.

To illustrate how the approach might work, we provide a simple example in the guise of the motivating palaeoclimate problem. Given the (notional) plant species “grassy”, “shrub” and “tree”, we observe pollen counts $Y = \{y_1, y_2, y_3\}$, which are constrained to sum to a total N . A simple graph of the data is presented in Figure 5.6

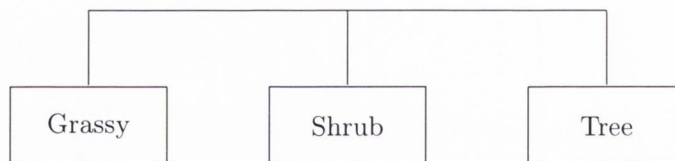


Figure 5.6: A simple example of some Multinomial pollen counts where there are 3 pollen categories, grassy, shrub and tree.

The sum total N is explicitly known, thus the data at the lowest level may be considered as a set of Multinomial responses. As we observed in Section 5.3.3, the use of likelihood models for Multinomial response data which do not take in account the sum constraint can lead to erroneous inferences at the inverse stage. The data must be treated as a set of constrained observations and unknown parameters in the joint model inferred jointly.

In the following we let $\{p_1, p_2, p_3\}$ denote the probability that an observed pollen count is from a tree species, a shrub species or a grass species respectively. $\{y_1, y_2, y_3\}$ are the corresponding, observed pollen counts. The Multinomial likelihood, neglecting constants, may be written as follows:

$$\pi(y|p) \propto p_1^{y_1} p_2^{y_2} p_3^{y_3} \quad (5.36)$$

If we consider that the responses can be ordered sequentially, then nesting structures provide an avenue for decomposing the Multinomial likelihood. For example, say we consider that pollen from trees and shrub species are somewhat related given that both comes from “woody” type plants; we group these two plant types into a subset which we entitle “woody”. As a result the joint problem in Figure 5.6 decomposes into a sequence of two independent problems in Figure 5.7. We first decide whether a grassy or woody type pollen spectra has been observed. If woody pollen is observed, we are then interested in the category, tree or shrub, into which the observed response falls. The important point to note is that, given the known total for woody, the observed counts for tree and shrub are conditionally independent of the corresponding count for grassy.

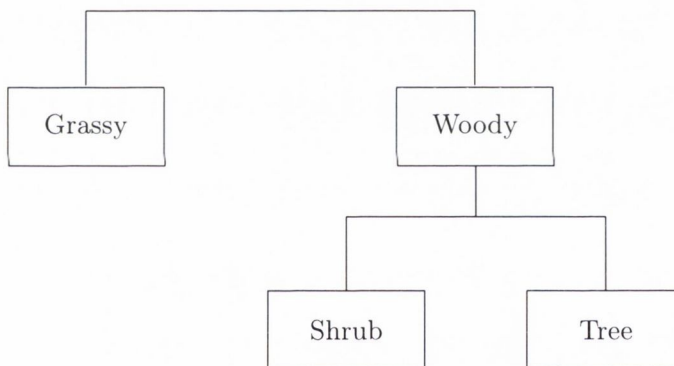


Figure 5.7: A simple, sequential ordering of the pollen data.

Let $(p_1 + p_2)$ represent the probability that an observed count is from a woody type species. To observe the effect that this sequential structure has on the likelihood, we multiply and divide through Equation 5.36 by $(p_1 + p_2)^{y_1+y_2}$.

$$\pi(y|p) \propto \frac{(p_1 + p_2)^{y_1+y_2}}{(p_1 + p_2)^{y_1+y_2}} p_1^{y_1} p_2^{y_2} p_3^{y_3} \quad (5.37)$$

$$= \left(\frac{p_1}{p_1 + p_2} \right)^{y_1} \left(\frac{p_2}{p_1 + p_2} \right)^{y_2} (p_1 + p_2)^{y_1+y_2} p_3^{y_3} \quad (5.38)$$

We reparameterize as follows; let $q_1 = (p_1 + p_2)$ denote the probability that an observed

count is woody and $q_2 = (p_1 + p_2)$ denote the conditional probability of observing a tree pollen count given that a woody type pollen count has been observed. Equation 5.38 may be rewritten as:

$$\pi(y|p) \propto q_1^{y_1}(1 - q_1)^{y_2} | (q_2)^{y_1+y_2}(1 - q_2)^{y_3} \quad (5.39)$$

The decomposed likelihood in Equation 5.38 may be recognised as the product of the likelihood kernels for two independent Binomial likelihoods. The first component, involving q_1 , represents the probability of observing a woody pollen count with $(1 - q_1)$ representing the probability of observing a grassy type pollen count. The second component, involving q_2 , represents the probability of observing a tree species pollen count *conditional* on a woody type pollen count being observed.

The Multinomial likelihood in Equation 5.36 is thus rewritten as the product of independent Binomial likelihoods; the parameters of the separate Binomial likelihoods are not shared. Through the use of a logit transform, the constrained probabilities in Equation 5.38 can be linked to the unconstrained latent field X . If no interaction is assumed between the individual components (response surfaces) of X , the multivariate joint model involving the Multinomial likelihood splits into the product of independent univariate models; inference on the unknown parameters corresponding to each model can thus be made separately, greatly simplifying the inference task.

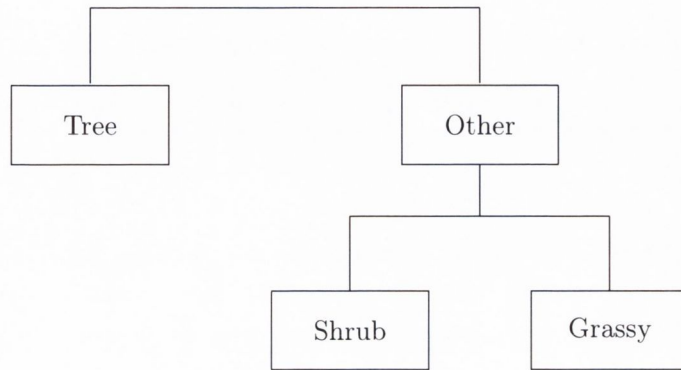
5.4.2 Choice of Nested Comparisons

An obvious point that was overlooked in the previous section is that a clear hierarchical structure existed in the simple example. In the absence of a clear hierarchical structure, the complete set of possible nested comparisons include contrasting:

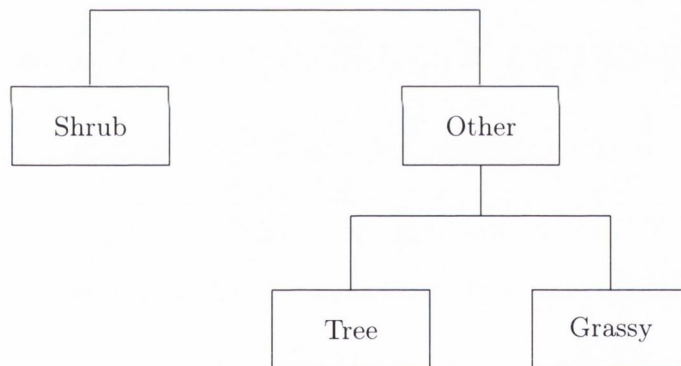
$$\begin{aligned} \{\text{Grassy}\} & \text{ versus } \{\text{Tree, Shrub}\} \\ \{\text{Tree}\} & \text{ versus } \{\text{Grassy, Shrub}\} \\ \{\text{Shrub}\} & \text{ versus } \{\text{Tree, Grassy}\} \end{aligned}$$

The first of these is presented in Figure 5.7 with the latter two presented in Figure 5.8. All decompositions of the likelihood will yield the correct joint likelihood. However, as noted by Rodríguez (2007), any choice of nesting contrasts can be selected for modelling, though only orthogonal comparisons will lead to a factorisation of the model likelihood into the product of independent parts. The authors use the interesting phrase that “the choice of comparisons

should be based on the logic of the situation”.



(a)



(b)

Figure 5.8: The two alternative nesting structures to the one presented in Figure 5.7.

As all decompositions will lead to the correct Multinomial likelihood, we can see that there is no one “unique” decomposition of the model likelihood. However, only the “true” nesting structure will lead to a statistically efficient decomposition of the model.

This is explained as follows; say the pollen of tree species and shrub species is fully correlated with a value of -1 and both are uncorrelated with the pollen count for the grassy species conditional on their sum. If a model decomposition such as that in Figure 5.8 (a) is chosen, the shrub and tree pollen counts will not be independent conditional on the sum of the grassy and shrub pollen counts. As a result, in making inference on the parameters of

the independent model components, there will be residual uncaptured dependence structure in the data, resulting in spurious correlations being inferred between model components, as per Aitchison (1986).

In the following section we will observe how this dependence structure manifests itself as a deterioration in predictive power at the inverse stage.

5.4.3 Choosing the “Best” Nesting Structure

A number of questions arise with regard to the use of nesting structures to decompose joint models where the likelihood is Multinomial. Firstly, how sensitive are the predictions at the inverse stage to the nesting structure chosen? Secondly, if the predictions at the inverse stage are sensitive to the nesting structure chosen, is it possible to infer the “best” or most appropriate nesting structure?

Simulated toy data, where the true nesting structure for the generated data is known, provides the easiest way to answer the posed questions. In order to investigate the statistical efficiency of the different nested comparisons introduced in the previous section, we simulate data from the model corresponding to the nesting structure presented in Figure 5.7 as follows:

At 200 random locations on a grid of length 100, Binomial counts (Y_G, Y_W) are generated with probability $\{P_G, P_W\}$, with the sum constraint $N = 1000$. The probabilities are obtained, given the known spatial location, by a logit transform of a latent smooth overdispersed Gaussian field $W = X + U$. The overdispersion of the underlying latent field is set as $\sigma^2 = 1$, i.e. each $u_i \in U \sim N(0, 1)$. These simulated observations constitute the counts at the first level of the model hierarchy, i.e. the simulated data represents counts of “woody or grassy” pollen.

At the second level, given the counts Y_W , a further set of Binomial counts (Y_T, Y_S) are generated in a similar manner (i.e. are also overdispersed Binomial counts conditional on an underlying overdispersed smooth latent surface). The simulated data at this level represent counts of “tree or shrub” pollen, conditional on the presence of woody type pollen at the upper nest level.

The data generating procedure may be presented as follows:

$$\{Y_G, Y_W\} \sim \text{Binomial}(1000, \{P_G, P_W\}) \quad (5.40)$$

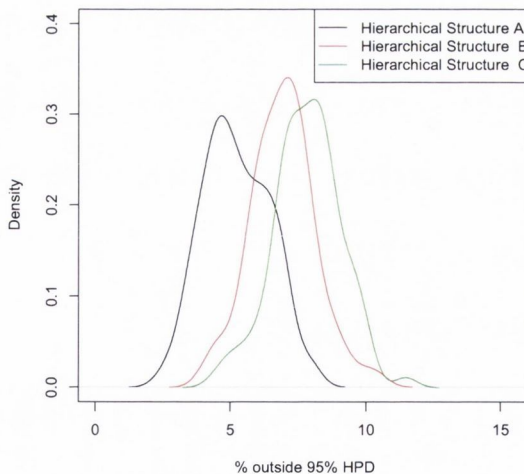
$$\{Y_T, Y_S\} \sim \text{Binomial}(Y_W, \{\frac{P_T}{P_W}, \frac{P_S}{P_W}\}) \quad (5.41)$$

The probabilities at each level in the model hierarchy are constrained to sum to one. A further point to note is that, in the absence of the known nesting structure, at the lowest

level, $Y_1, Y_2, Y_3 \sim \text{Multinomial}(P_G, P_T, P_S)$. Thus at the lowest level the simulated counts simply represent a set of overdispersed Multinomial count observations. As the joint model decomposes into the product of independent univariate models given the choice of nesting structure, inference on the latent parameters for each univariate model can be completed independently of the other.

For each set of simulated data, generated using the process outlined in Equation 5.40 - 5.41, the model parameters corresponding to each of the three nesting structures detailed in Figure 5.7 and Figure 5.8 are inferred. The calibrated models are then used to calculate the leave-one-out cross validation prediction accuracy for each nesting approach as well as a number of model fit measures. This process is completed a large number of times (100) in order to build up a profile of the predictive accuracy of the inferred models for each choice of nested comparisons.

We observe that the leave-one-out cross validation metric for the nesting structure (A) used to generate the data has an average error rate of 5%. Specifically, the number of observations which fall outside the 95% HPD region for location given data is approximately 5% given the inferred model parameters. Conversely, the corresponding statistics for the alternative nesting structures B & C have average error (Δ) statistics of approximately 7% and 8% respectively.



(a)

Figure 5.9: Prediction accuracy of the three different nesting structures. The model with the correct nesting structure (A) is shown to have the best predictive performance in terms of leave-one-out cross-validation prediction accuracy

Thus we may conclude that the use of nesting structures other than the “true” nesting structure do not provide a disjoint decomposition of the full joint model - the existence of residual

uncaptured dependence structure can be observed by the deterioration in prediction accuracy at the inverse stage of models which use nesting structures which do not fully decompose the model likelihood.

Model criticism tools provide a method of evaluating various aspects of the approach. In Figure 5.10, we observe that the absolute (distance) of the mode of the prediction produced, to the correct location is smallest for the approach involving the correct hierarchical structure *A*, than for the alternative nesting structures. Secondly, we observe that the posteriors on location for the correct hierarchical structure are on average slightly more conservative than those of the alternative, erroneous nesting structures. These have residual, spurious dependence structure left in the model, thus impairing their predictive accuracy at the inverse stage - if residual dependence is not modelled at the forward stage, this manifests itself in posterior predictive distributions on climate which are ‘narrower’ than the correct posterior predictive distributions.

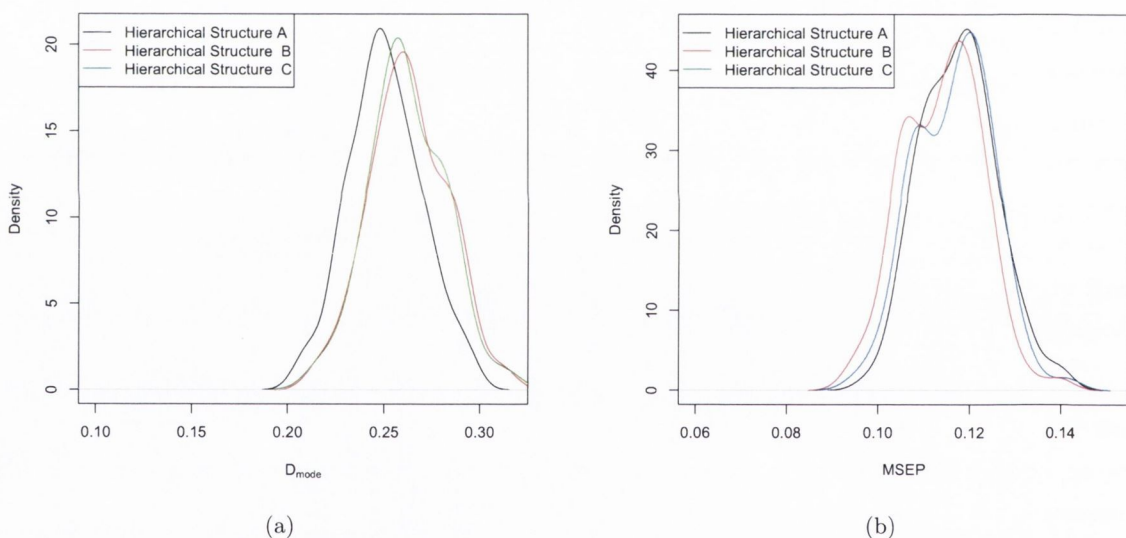


Figure 5.10: Plots of distance to the mode and expected distance to the mode for each of the three nesting structures. The true model (*A*) is shown to have the largest *MSEP*, on average, of the three models considered.

The method provided here for locating the “best” hierarchical is based on the analysis of inverse predictive power. For all possible combination of the observed taxa, the inverse predictive ability of each nesting decomposition can be compared to obtain the model with the best predictive power. This is analogous to the approach of Marden (1992), who studied the use of nesting orthogonal contrasts to analyze some rank data with the ultimate goal of finding the set of orthogonal contrasts which best captured the main features of the data. However, as

the number of observed groups increases, the number of permutations which we must evaluate in order to identify the most appropriate nesting structure increases exponentially. We address this issue in Chapter 7, detailing how expert opinion can be availed of to “narrow” the number of permutations to be considered.

5.5 Addressing Zero/N-inflation of the Multinomial Response

In practice, the multivariate compositional data sets that we wish to model cannot be adequately modelled using standard statistical families such as the Multinomial. For example, as mentioned in Section 3.5.2, a common feature of ecological data sets, such as the RS10 dataset, is their tendency to contain many zero counts - if statistical models are used which do not account for a possible excess in the number of zeros, inferences derived from the data are likely to be erroneous. Ridout et al. (1998) note that the erroneous nature of these inferences is predictable; the use of standard Poisson models for data that contain an excess of zeros will lead to an underestimation in the rate parameter of the Poisson model as well as posteriors on model parameters which are not sufficiently conservative.

The additional zeros may be modelled through the use of *zero-modified* distributions (see Section 3.5.2). In a zero-modified distribution, as per Hall (2000), the observed data are assumed to arise from one of two distinct states, a zero state from which only zero counts are observed and an alternative state from which all of the non-zero counts and a few of the zero counts are observed - the alternative state can be modelled through the use of standard statistical families such as the Poisson or the Binomial. However, it transpires that the use of such zero-modified distributions, in the context of data which is subject to sum constraints, can lead to inconsistent parameter estimates. In the following, we discuss this issue in further detail and propose a solution to this problem.

5.5.1 Statistical Inconsistency of Zero-Inflated Models for Binomial Data

In Section 3.5.2, we introduced the concept of zero-modified distributions for the modelling of zero-inflated count data. To recap, a standard zero-inflated model in the context of Binomial count outcomes may be presented as:

$$\pi(y) = \begin{cases} (1 - q) + q \times \text{Binomial}(0, N, p) & y = 0 \\ q \times \text{Binomial}(y, N, p) & y > 0 \end{cases} \quad (5.42)$$

Thus, with probability q , an observed zero count arises from a Binomial distribution with parameters N and p . Alternatively, with probability $(1 - q)$, the observed count arises from a distribution with a point mass at zero. However, the zero-modified Binomial distribution is not

symmetric. Consider the following example; say we have observed the pair of Binomial counts (y_1, y_2) which are constrained to sum to the total N . Inferences regarding p , derived from the model in Equation 5.42, are dependent on whether we designate y_1 or y_2 as the response. To see this, consider the setting where $y_1 = 0$ and thus $y_2 = N$ - if y_1 is designated as the response, the probability of observing this pair of observations is $\pi(y_1, N) = (1-q) + q(1-p)^N$. Conversely, if y_2 is designated as the response, the corresponding probability is $\pi(y_2, N) = q(1-p)^N$. These probabilities are not equal - as a result, the use of such a model will lead to inconsistency in statistical inferences.

To illustrate the impact that the choice of response has on parameter estimation, we create the following simple example. Given a smooth response surface P , defined on a regularly spaced grid of length 100, 500 pairs of count outcomes (y_{1i}, y_{2i}) are generated from the zero-inflated Binomial model in Equation 5.42 with $q = p^\alpha$, $\alpha = .3$. In total, 46 of the 500 Y_1 counts are zero whereas none of the Y_2 counts are zero. The generated counts for Y_1 are presented in Figure 5.11 (a). Model parameters are estimated under two scenarios; (1) Y_1 is regarded as the response and (2) Y_2 is regarded as the response. The expected value of P obtained using both approaches is plotted in Figure 5.11 (b).

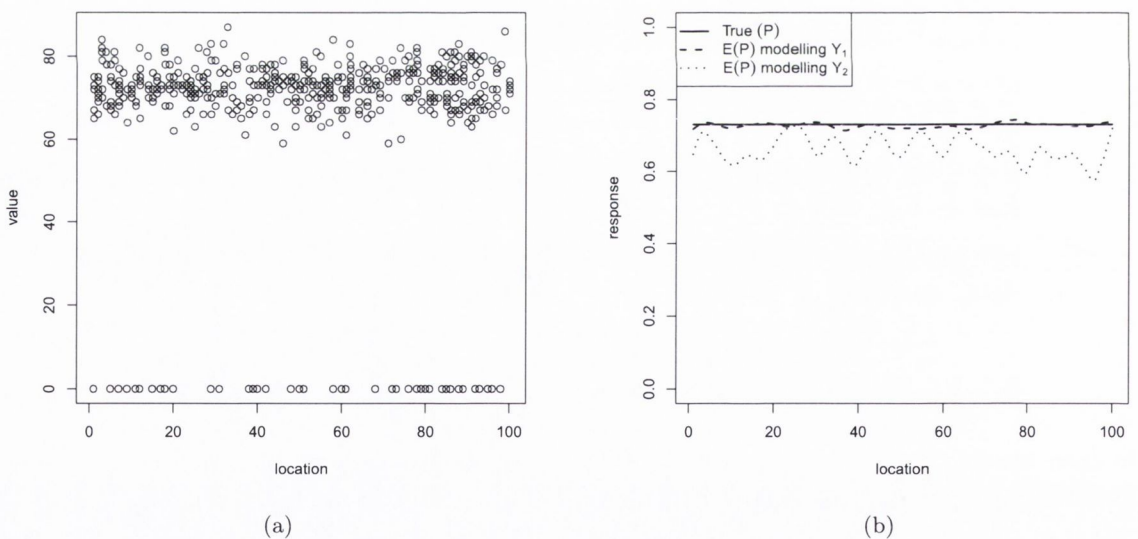


Figure 5.11: (a) Simulated zero-inflated Binomial counts data. (b) The response surface P , estimated both using Y_1 as the response and subsequently Y_2 . Statistical inconsistency can be observed in the results obtained.

If Y_2 is modelled as the response the zero-inflation parameter must be zero as there are no zero observations observed for Y_2 ; the approach thus decomposes to a simple Binomial model which cannot handle the excess variability in the counts due to the extra N 's in Y_2 ;

these correspond to the zeros of Y_1 . Conversely, the model parameters estimated when Y_1 is designated as the response are approximately correct; this is to be expected as this was the model from which the data were generated.

The statistical inconsistency in results can be clearly observed. The mean parameter is underestimated for the zero-inflated model when Y_2 is designated as the response; furthermore model parameter estimates are not consistent - chosen models for the data are subject to the problem that inferences derived from the model are dependent upon the counts which we choose to analyse. We refer to this problem as “N-inflation” of the counts data - in a Binomial model, zero-inflation of one set of counts will lead to N-inflation in the other.

In the following section we introduce a simple extension to the likelihood model introduced above which corrects for this problem. The identification of the statistical inconsistency of existing zero-inflation models and the development of a solution to this problem is regarded as a novel contribution in this thesis.

5.5.2 Sensitivity to Zero/N-Inflation

An extension of the standard zero-inflated Binomial model to model N-inflation is:

$$\pi(y) = \begin{cases} (1 - q_1)q_2 + q_1q_2 \times \text{Binomial}(0, N, p) & y = 0 \\ q_1(1 - q_2) + q_1q_2 \times \text{Binomial}(N, N, p) & y = N \\ q_1q_2 \times \text{Binomial}(y, N, p) & 0 < y < N \end{cases} \quad (5.43)$$

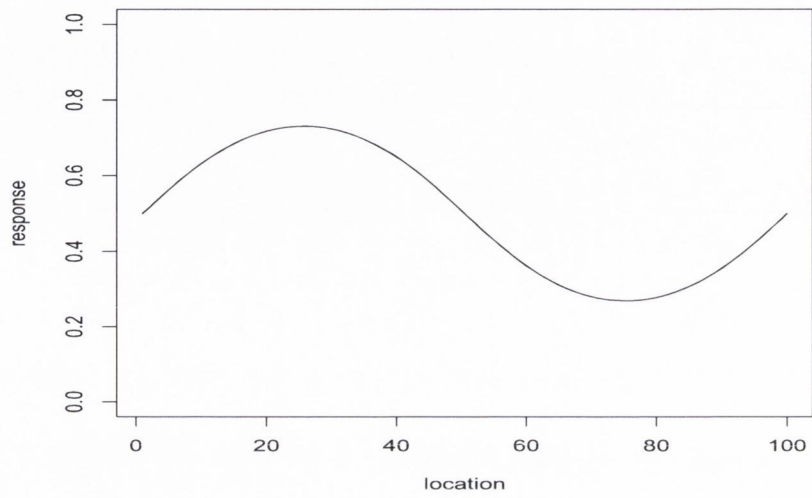
where:

$$q_1 = p^{\alpha_1}; q_2 = (1 - p)^{\alpha_2}$$

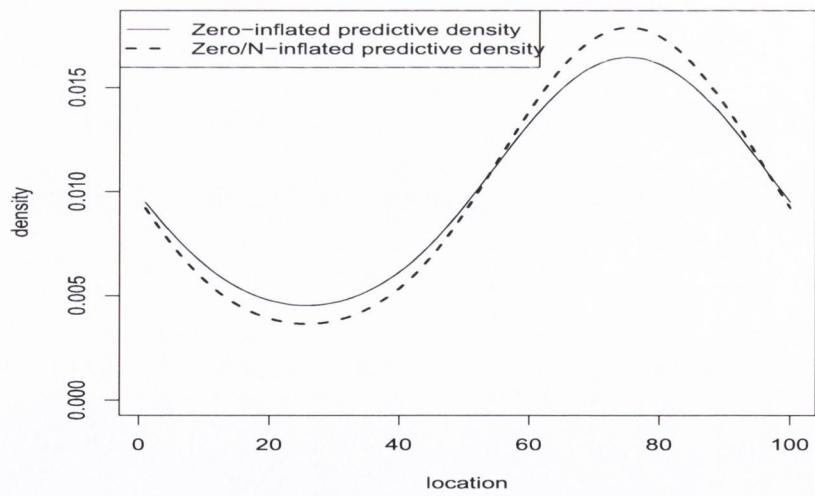
Through the use of a logistic transformation, the probability can be made a function of the underlying spatial field X . $p = \left(\frac{\exp(x)}{1 + \exp(x)}\right)$. If $\alpha_2 = 0$ then it is clear that this model decomposes into the zero-inflated Binomial model. Thus the zero/N-inflated model has one additional parameter, a parameter governing ‘N-inflation’, over the zero-inflated model. The constructed likelihood is compatible with the INLA algorithm.

In Figure 5.12, we observe the difference that this likelihood makes in spatial prediction. Given a generated count of zero from a zero/N-inflated Binomial with known parameters, the posterior on spatial location produced by the zero/N-inflated model is more peaked than the corresponding posterior predictive posterior obtained by a zero-inflated model, reflecting that there is additional information available from the knowledge that if y is zero then $N - y$ is non-zero. In the presented example $\alpha_1 = \alpha_2 = .3$ and $N = 1000$.

In Figure 5.13 we present the striking contrast in predictive distributions produced by the zero/N-inflated and zero-inflated models for a count of N . Given the zero-inflated model, all



(a) Response surface



(b) Comparison of the predictive distribution obtained when using a zero/N-inflated likelihood model or a zero-inflated likelihood model for a count of zero ($y = 0$)

Figure 5.12

counts of N are inferred to arise at the highest points on the response curve - this is due to the result that most counts of N will be generated in this region of space. However, as previously mentioned, in the context of Binomial counts data, zero-inflation of one set of counts will lead to N-inflation of the paired counts. Thus the zero/N-inflated model recognises that the count of N may actually be an artifact of zero-inflation and responds accordingly, resulting in the correct predictive distribution which is much less peaked.

In the context of the motivating palaeoclimate reconstruction problem, in Chapter 7 we illustrate that failure to explicitly account for the N-inflation present in the RS10 training dataset leads to a deterioration in predictive accuracy of the calibrated models at the inverse stage. Through explicit modelling of the N-inflation present in the pollen counts, model prediction accuracy is shown to be substantially increased. Further discussion on this subject is deferred to Chapter 7.

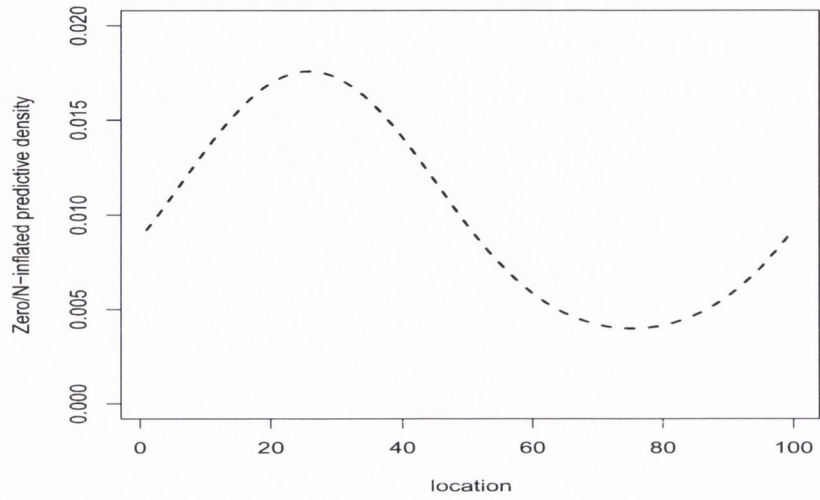
5.6 Conclusions

Multivariate statistical models for highly multivariate observational datasets introduce issues of computation and inference at the forward modelling stage. In the presence of dependence structure, either in the data or through model specification, the parameters of the multivariate statistical model must be jointly inferred. However, in many cases, such as in the context of the motivating palaeoclimate reconstruction problem, joint inference on all model parameters is simply infeasible due to the sheer number of parameters requiring inference.

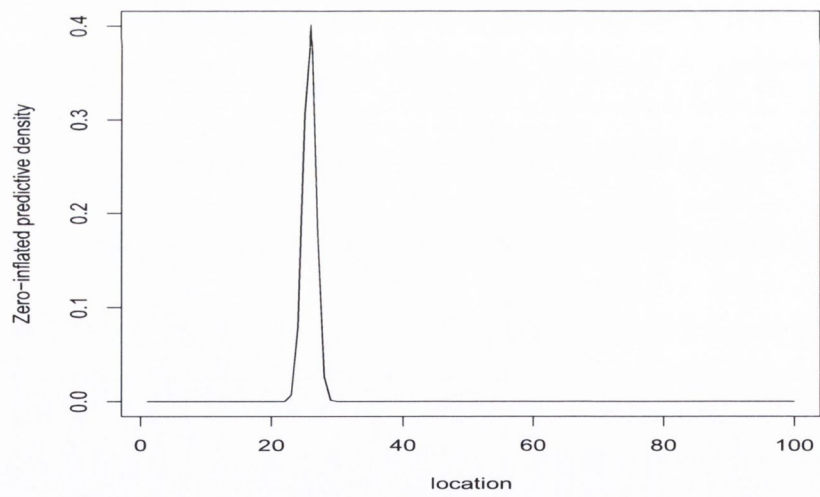
One solution to this problem is to decompose the multivariate joint model into a series of separate, independent univariate models, greatly simplifying inference tasks. This decomposition is based on the assumption of conditional independence of the multivariate model components; the multivariate observations are assumed to be conditionally independent given the multivariate latent field and the response surfaces which comprise the latent field assumed conditionally independent given spatial location. If this decomposition is appropriate, inferences on the parameters of each of the univariate models in isolation will be equivalent to those obtained making inferences on the full model jointly.

However, naive use of this decomposition, such as in settings where there are sum-to-total constraints on the observational data, will lead to poor inference outcomes at the inverse stage. A simulated example was used to demonstrate this result; the modelling of Binomial counts data as conditionally independent Poisson observations was shown to result in a model with substantially poorer predictive performance, as compared to the correct Binomial model. This was due to inverse predictive posteriors on spatial location which were not sufficiently conservative, brought about by the Poisson model erroneously treating each pair of Binomial counts as two separate, independent pieces of information.

One possible strategy for decomposing a computationally intensive joint model, involving



(a)



(b)

Figure 5.13: Predictive distributions for the (a) zero/ N -inflated and the (b) zero-inflated likelihood model for the count value $y = N = 1000$. Note the change in density values of the y axis - the zero/ N -inflated model recognises that there's not much information contained in the count of N .

Multinomial response counts, into a series of separate univariate models is to define a hierarchy of nested comparisons between subsets of the responses. Essentially, the joint likelihood, involving dependent Multinomial counts, may be decomposed into the product of conditionally independent, less computationally intensive inference tasks involving separate Binomial likelihoods, resulting in computational conveniences at both the forward and inverse stages. Though every proposed nesting structure will lead to a full decomposition of the Multinomial model likelihood, only the “true” nesting structure will lead to a statistically efficient decomposition of the joint model. In terms of inverse problems, a simulated example was used to demonstrate that the optimal nesting structure for any set of count observations can be identified by fitting all possible nesting structures and identifying the one which best meets the model fitting criterion of choice, in this case Δ , a measure of the inverse predictive performance. In situations where the evaluation of all possible nesting structures is too onerous to consider, the use of expert opinion can be used to narrow the number of possible permutations of taxa.

Finally, statistical models for Binomial counts data, which do not simultaneously account for an overabundance of zeros in each of the pair of Binomial counts, are subject to inconsistency in parameter estimates - the over-abundance of zeros for one taxa results in N-inflation of the counts corresponding to the other. A parsimonious modelling solution to this problem is developed, and a simple example is used to demonstrate the impact of the use of the inconsistent likelihood model in the setting where N-inflation is present in the data. The resulting inverse inferences were shown to be highly erroneous.

Chapter 6

Spatial Prior Models and Computationally Efficient Inverse Inference

The novel contributions in this chapter relate to computationally efficient inference at both the forward and inverse stages of spatial calibration problems. We outline a method of obtaining approximately correct precision structures for intrinsic GMRF models on irregularly shaped spatial domains, resulting in vast computational savings at the forward stage. An additional contribution is the development of a sampling based algorithm for computationally efficient inference at the inverse stage. We detail how the algorithm facilitates the ‘integration out’ of hyperparameter uncertainty at no extra computational cost, as compared to empirical Bayes based methods and substantially reduces the time and computational cost of prediction at the inverse stage.

This chapter is organized as follows; in Section 6.1 we explore the effect of ignoring influential model covariates in spatial prior models at the forward stage, on the accuracy of the predictions produced at the inverse stage. We illustrate that if the constructed models do not include all spatial covariates upon which the response depends, inferences at the inverse stage will be erroneous.

In Section 6.2 we discuss spatial prior models for the forward stage of univariate inverse inference problems. We discuss models for discrete spatial variation, examine the use of random walks as spatial priors and detail their implementation in several spatial dimensions. Additionally, we detail a method of obtaining approximately correct precision matrices for intrinsic GMRF models on irregularly shaped spatial domains.

In Section 6.3 we propose an algorithm for computationally efficient inference at the inverse stage. The discretization of space involves the specification of a finite collection of random

variables, at each of which the posterior predictive regions must be evaluated in order to obtain normalising constants. We detail a sampling algorithm which helps to efficiently integrate out hyperparameter uncertainty in inverse problems.

6.1 Spatial Covariates

In a typical spatial regression problem, interest generally lies in making inferences on the latent, unobserved spatial response surface, X , which describes the relationship between some recorded spatial covariates, C and the observed response Y . In this chapter we focus on the setting where the response is *univariate*, i.e. we have an observational dataset Y consisting of the n univariate observations (y_1, \dots, y_n) .

However, whilst Y and spatially referenced X are taken as univariate in the following, the spatial variables C will remain *highly multivariate*. For example, in a regression problem such as the spatial distribution of iron ore at various locations in a mine, the data set may contain information on a number of spatially referenced variables such as longitude, latitude and *depth*. In this setting $C = (C_1, C_2, C_3)$ where each row $c_i \in C$ describes a location in three dimensional space and the observed response, namely the quantity of iron ore at a particular location in the mine, is a function of each of these three spatial variables jointly. Thus, the failure to construct models which include all three variables will lead to misleading inferences being derived from fitted models at both the forward and inverse stages.

In the context of the motivating palaeoclimate application, Huntley (2001) alludes to this problem, citing an earlier study by Conolly & Dahl (1970), where inappropriate climate variables were used to model the pollen response to climate of a particular plant species, *Rubus chamaemorus*. Specifically, the authors concluded that the spatial distribution of this particular plant was highly correlated with the maximum summer temperature of a given region; subsequent studies revealed the spurious nature of this statement, concluding that other variables such as the temperature of winter months and in particular, the variability in moisture available for plant respiration, were far more important in determining the species range. The strong, spurious correlation between the species range and extreme summer temperature deduced by the authors, was in fact a manifestation of the correlation between extreme summer temperature and winter temperature, a variable which had been omitted in model construction.

In this thesis our primary interest is in *inverse inference*. Calibrated models themselves are not of intrinsic interest, our interest lies instead in making inferences on the underlying, unobserved spatial covariates corresponding to new data for which such information is unknown. With this in mind, in the following section we illustrate the impact of the omission of important spatial variables during model construction on the inferences derived using the calibrated model at the inverse stage.

6.1.1 The Impact of Spatial Covariate Omission

If models for the observed data are constructed which do not include the relevant spatial variables used to generate the response, inferences derived from the model will be erroneous.

To illustrate this point, we construct a simple toy example as follows; a single Gaussian distributed datum y^{new} , is generated at a single location in two dimensional space $(c_1^{\text{new}}, c_2^{\text{new}})$, given a smooth latent response surface X and known model parameters, i.e.

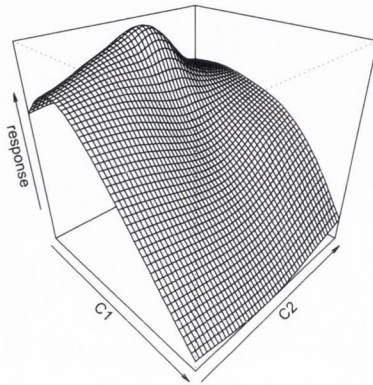
$$y^{\text{new}} \sim N(X(c_1^{\text{new}}, c_2^{\text{new}}), 1) \quad (6.1)$$

Here $C = (C_1, C_2)$; the two dimensional surface $X(C_1, C_2)$, presented on a 50×50 grid, is plotted in Figure 6.1 (a), with the corresponding one dimensional (marginalised) projections ($X_1(C_1)$ and $X_2(C_2)$) of the surface also displayed (Figure 6.1 (b) and Figure 6.1 (c) respectively). As $X(C_1, C_2)$ is defined on a regular 50×50 grid, $X_1(C_1)$ and $X_2(C_2)$ are easily obtained by marginalising over $X(C_1, C_2)$ in the C_1 and C_2 directions.

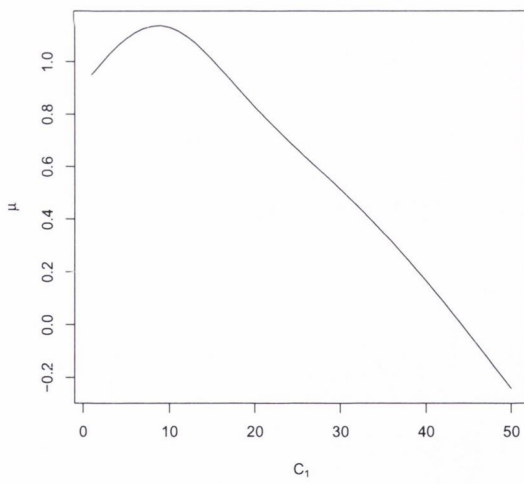
The generated value of y^{new} is dependent on a smooth function of both C_1 and C_2 due to the data generating mechanism. Hence, the use of models which do not account for this spatial interaction will lead to erroneous inferences at the inverse stage. In Figure 6.2 (a - c) this result is observed: the posterior predictive distribution, $\pi_{\text{joint}}(c_1^{\text{new}}, c_2^{\text{new}} | y^{\text{new}})$ ($= \int_X (c_1^{\text{new}}, c_2^{\text{new}}, X | y^{\text{new}}) dX$) is presented along with its marginalised posteriors ($\pi_{\text{joint}}(c_1^{\text{new}} | y^{\text{new}})$ & $\pi_{\text{joint}}(c_2^{\text{new}} | y^{\text{new}})$) and compared to the corresponding uni-dimensional predictive posteriors $\pi_{\text{ind}}(c_2^{\text{new}} | y^{\text{new}})$ and $\pi_{\text{ind}}(c_1^{\text{new}} | y^{\text{new}})$. These are marginally obtained as $\pi_{\text{ind}}(c_i^{\text{new}} | y^{\text{new}}) = \int_{X_i} (c_i^{\text{new}}, X_i | y^{\text{new}}) dX_i$ where $i = 1, 2$ - the dependence of X on both C_1 and C_2 is ignored and the uni-dimensional smooths (Figure 6.1 (b - c)) are used to provide reconstructions for each spatial dimension separately.

We observe spurious multimodality and mislocation of the posterior produced for c_2^{new} using the model which does not account for the spatial interaction of X over C_1 and C_2 . The highest values of posterior probability mass are spuriously placed at the edges of location space. In contrast, given the (correct) joint model, the highest values of the posterior probability mass are centred on approximately the correct spatial locations. This, being the model from which y^{new} was generated gives the correct predictive distribution.

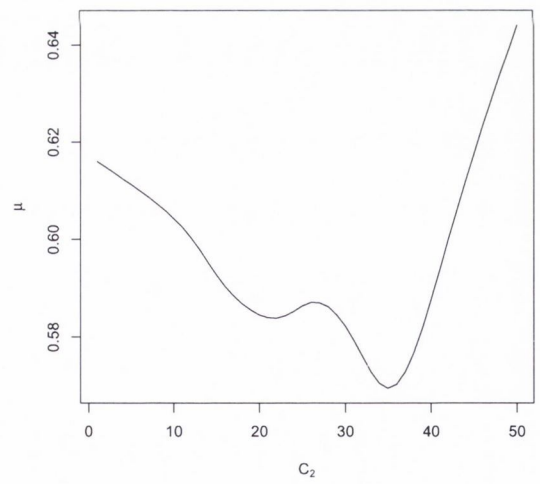
This simple toy example provides the following conclusion; if models for the data are constructed which do not include all spatial covariates upon which the data depend, erroneous inferences are produced at the inverse stage. In Chapter 7, in the context of the palaeoclimate reconstruction problem, we illustrate that failure to incorporate all influential climate variables into the forward models results in a deterioration in prediction accuracy and mislocation of climate posteriors at the inverse stage.



(a)

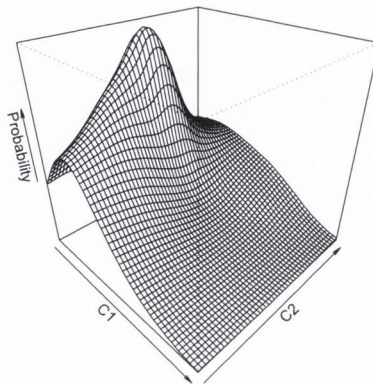


(b)

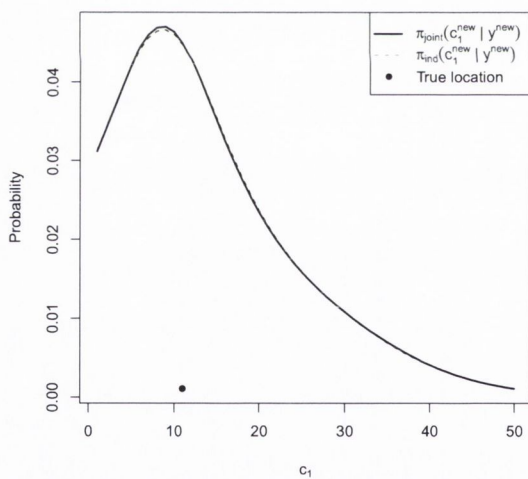


(c)

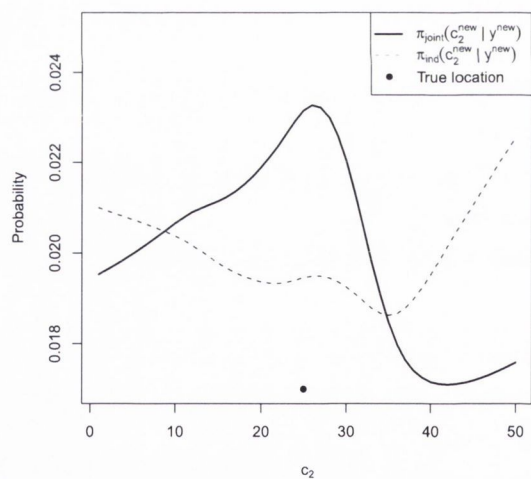
Figure 6.1: (a) Simulated two dimensional spatial surface $X(C_1, C_2)$, defined on a 50×50 grid. (b) & (c) represent the respective unidimensional marginals of the two dimensional surface ($X_1(C_1)$ and $X_2(C_2)$) with interaction of X over the spatial covariates marginalised out.



(a)



(b)



(c)

Figure 6.2: (a) Posterior distribution on spatial location $\pi((c_1^{\text{new}}, c_2^{\text{new}})|y^{\text{new}})$, given the simulated count $y^{\text{new}} = 3$, in two dimensional space, evaluated using the joint model. In (b) & (c) we represent $\pi_{\text{ind}}(c_1^{\text{new}}|y^{\text{new}})$ and $\pi_{\text{ind}}(c_2^{\text{new}}|y^{\text{new}})$, obtained using the uni-dimensional surfaces, along with the marginalised posteriors of $\pi_{\text{joint}}((c_1^{\text{new}}, c_2^{\text{new}})|y^{\text{new}})$. The true location from which the count was generated is represented by the black dot (\bullet).

6.2 Spatial Prior Models

In the preceding section, we used the known response surfaces at the forward stage to examine the effect of the omission (from models) of important spatial covariates on the inferences produced at the inverse stage. In this particular section we relax the assumption of known model parameters and focus on spatial prior models for the latent response surface X .

As previously mentioned in Section 3.4, the use of multivariate Gaussian random field priors for X is very common in point referenced spatial regression problems. X is defined as a multivariate Gaussian process with mean vector μ and $n \times n$ covariance matrix $\Sigma(\theta)$, where the individual elements in $\Sigma(\theta)$ describe the spatial covariance between each of the latent x_i . The degree of covariance between the x_i is governed by both the underlying model hyperparameters θ (which parameterise $\Sigma(\theta)$) and the “distance” between spatial locations $c_i \in C$.

However, in using Gaussian random field models, the dimension of the covariance matrix is directly related to the number of recorded observations; as the dimensionality of the dataset under consideration increases, such prior specifications can quickly become too computationally intensive to consider. This is due to the necessity of inverting large dense $n \times n$ covariance matrices Σ in order to evaluate probability densities, where n is the number of unique spatial locations at which data is observed. Lindgren et al. (2011) provide a brief discussion on several approaches which try to address or avoid this issue - one such method involves the discretization of space and the harnessing of computationally efficient models for discrete or lattice based data.

6.2.1 Discrete Approximations to Continuous Spatial Fields

In situations where the training dataset is *large*, the use of spatial models based on Gaussian random fields may not be feasible. An example of one such dataset is the RS10 dataset considered in this thesis; there are 7742 distinct observations for each separate plant taxon. The use of spatial models based on Gaussian random fields involves the manipulation of extremely *dense* covariance matrices of dimension 7742×7742 . For even the simplest models, the computational cost of the approach is prohibitive.

One solution to this problem is to discretize the spatial region under consideration. The main benefit is computational; the discretization of the spatial region facilitates access to Gaussian Markov random field models for X . As previously mentioned, GMRFs are parameterized by a mean μ and a precision (inverse covariance) matrix Q which describes neighbourhood structure; due to the Markov properties of GMRFs, Q is sparse and thus the use of numerical algorithms for operations on sparse matrices facilitates fast computing time.

However, the dimension of the spatial region under consideration can have a large impact on the computational efficiency of Gaussian Markov random field models; the computational

benefits of the approach reduce with each additional spatial dimension considered. Rue & Martino (2006) note that the computational effort involved with manipulating Q matrices is dependent on the size of the matrix and the neighbourhood structure. For example, for a two dimensional GMRF defined on a square lattice of size $m \times m$, the cost of factorising Q is $\mathcal{O}(n^{3/2})$ where $n = m^2$ (Lindgren et al. 2011). For a GMRF defined in three spatial dimensions, such as on the unit cube, the equivalent factorisations have a computational cost of $\mathcal{O}(n^{5/3})$ where $n = m^3$ (Rue & Martino 2006). The reason for this is that Q is much less sparse in three spatial dimensions than two; the number of non-zero terms in Q induced by the local neighbourhood structure goes from $\mathcal{O}(n \log(n))$ in two spatial dimensions to $\mathcal{O}(n^{4/3})$ in three.

A further issue with the discretization of space is that the number of latent parameters requiring inference increases as a function of the spatial region being discretized, i.e. approximating \mathbb{R}^2 space by a square lattice of length m results in m^2 latent parameters (the node points) which must be inferred. Approximating \mathbb{R}^3 space by a square lattice in three dimensions results in m^3 latent parameters. This power law increase in the number of latent parameters is known as the *curse of dimensionality* (Bishop 2006) and is illustrated graphically in Figure 6.3.

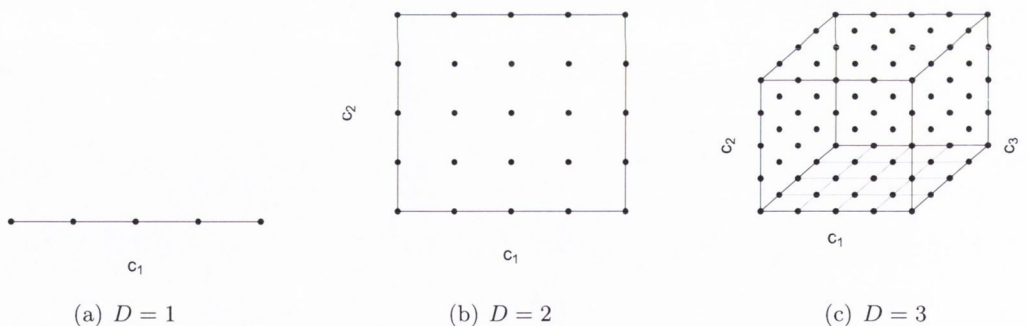


Figure 6.3: Illustration of the curse of dimensionality; the number of points required to discretize a regular space grows exponentially with the dimensionality D of the space.

Due to this curse of dimensionality, approaches based on the discretization of space are essentially constrained to consider at most three spatial dimensions, depending on the resolution of the discretization. Though the number of latent parameters may ultimately end up being significantly greater than the number of observations, the computational benefits of the sparse nature of the precision matrices involved in GMRF models mean that the approach is preferable in situations where the observational dataset is large. In the following we make extensive use of illustrative examples involving spaces of one or two dimensions as operating in this space makes it relatively easy to illustrate introduced concepts graphically.

$$X(c) = .04c + 3\sin(.01c\pi) \quad (6.5)$$

A Gaussian model is assumed for Y , i.e. each y_i is generated randomly from $N(X(c_i), 1)$. The generated data, along with X is presented in Figure 6.4.

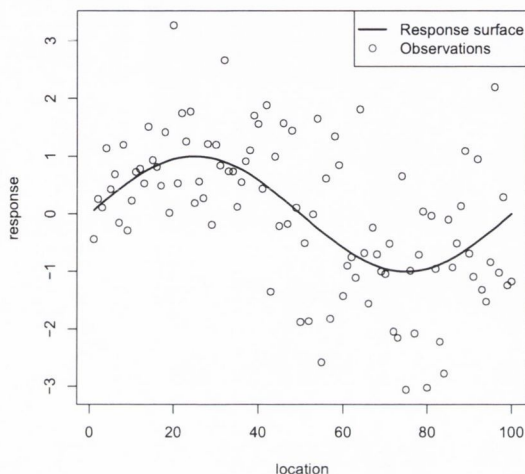


Figure 6.4: Simulated Gaussian response data, generated using a smooth spatial response surface (pictured).

In the following we assume that the nature (shape) of the response surface is not known, but is to be inferred from the observed data. A simple model for the data is:

$$y_i = X(c_i) + \varepsilon_i \quad (6.6)$$

$$X \sim \text{GMRF}(Q_x) \quad (6.7)$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad (6.8)$$

The prior model for X is an intrinsic GMRF of order 2. In the following we compare and contrast the posteriors which result from models for X which utilise (a) the correct precision matrix Q and (b) the precision matrix Q^* , obtained using convolution methods. In order to compare like with like, model hyperparameters are fixed *a priori*, specifically $\sigma^2 = 1$ and $\kappa = 300$. As both the likelihood and the prior are of multivariate Gaussian form, the posterior for X is of known distributional form; $\pi(X|Y) \sim \text{MVN}((Q_y + Q_x)^{-1}Q_y Y, (Q_y + Q_x)^{-1})$ where Q_y is a diagonal matrix consisting of values of σ^{-2} along the diagonal and zeroes elsewhere.

Intuitively, the use of a precision matrix with incorrect neighbourhood structure at the boundary will result in error in posterior X at the boundary. In Figure 6.5(b) we plot the diagonal of the posterior variance matrix $V = (Q_y + Q_x)^{-1}$ obtained where Q_x is equal to Q and Q^* respectively. We observe that the posterior variances produced with Q^* as the prior precision matrix results in an underestimation of the posterior variances as compared to the correct posterior variances obtained using Q . Furthermore, in Figure 6.5(a) we observe that the posterior mean of X is erroneous at the boundary and the 95% HPD regions are too “tight”.

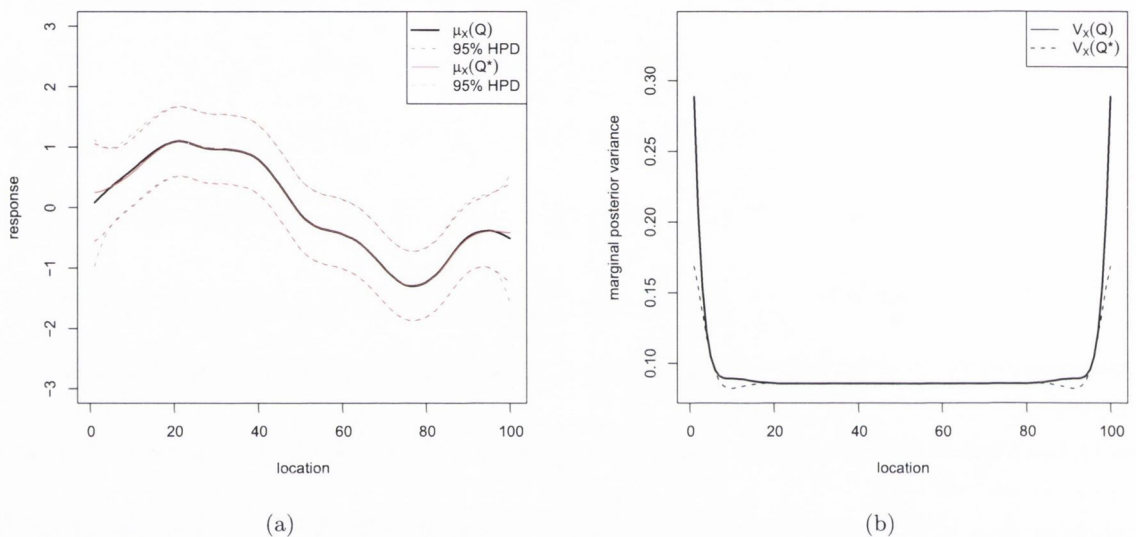


Figure 6.5: Comparison of the (a) posterior response surfaces and (b) posterior variances produced using spatial prior models with Q and Q^* as the prior precision matrices. The precision structure of Q^* at the boundary is incorrect and is reflected in erroneous estimation of the posterior mean and variance at boundary regions.

However, we also note that the correct posterior means and variances of X are obtained in the interior of the spatial region. This motivates the following proposal; the spatial region under consideration, “the region of interest”, is extended to incorporate a buffer region on either side. This results in an increase in the size of X due to the buffer region and the similarly extended prior precision matrix Q^* will still have incorrect neighbourhood structure, however, the buffer region is constructed to be *large enough* that there is no boundary effect in the region of interest.

This idea is analogous to a proposal by Besag & Higdon (1999), who account for “edge effects” in the precision structure of random walk on the lattice by extending the region of interest with “additional layers”. The authors note that the structure of the precision matrix

is correct in the interior and converges quickly to the correct values on its boundary as the number of layers increases. Kneib (2006) refers to this approach “as a restriction of the infinite lattice to the finite case without correcting for the boundaries”.

In Figure 6.6 we observe the improvement in the posterior inferences for X as a result of the incorporation of a buffer region. The discretized space of length 100 is extended to include an additional 10 gridpoints on each side. Q^* is thus of dimension 120×120 . In Figure 6.6 (b) we observe that the underestimation in the posterior variances at the boundaries of the region of interest has been corrected. Additionally, we observe that the posterior mean of X and the corresponding 95% HPD regions essentially overlap with those of the correct approach. In general, the greater the buffer region used, the lesser the resulting error in the posterior response will be, as per Besag & Higdon (1999).

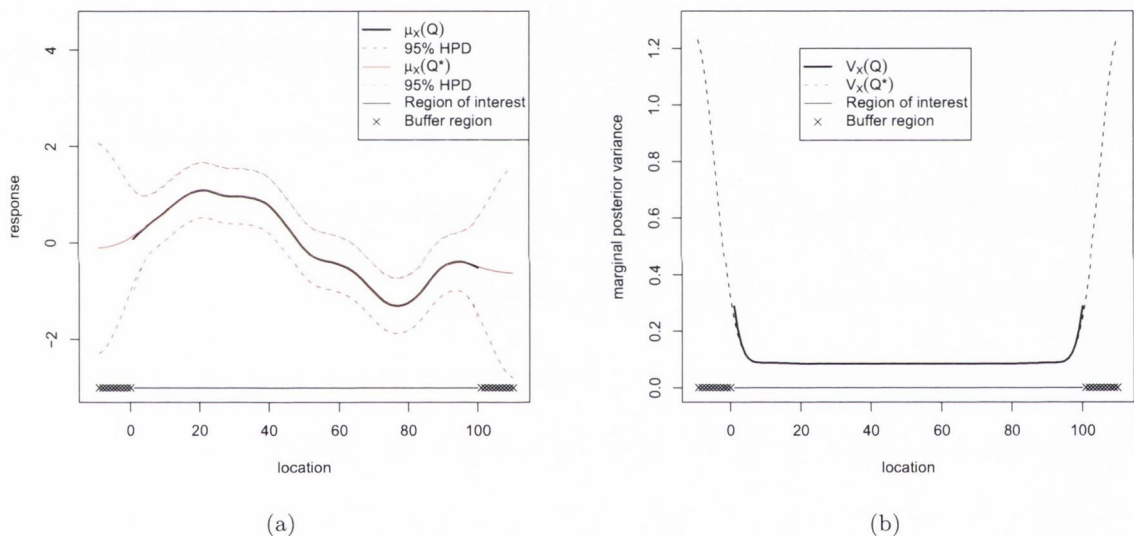


Figure 6.6: Comparison of the (a) posterior response surfaces and (b) the posterior variances produced using spatial prior models with Q and Q^* as the prior precision matrices. With the incorporation of a buffer region, the posterior for X is approximately correct in the region of interest.

The discussion thus far has revealed a modelling/computational strategy that seems to produce a good way of obtaining approximately correct precision structures for Q . The precision structure of a first order random walk for a regularly spaced lattice in any spatial dimension is easy to define, being a simple function of its nearest neighbours - if the region of interest is extended to incorporate a buffer region, the precision structure of higher order random walks can be obtained by convolution methods, which will have the correct precision structure in the interior. Whilst the extension of the region of interest increases the number of latent

parameters that must be inferred in the uni-dimensional setting, this approach can result in vast computational savings in the context of irregularly spaced regions of interest in higher spatial dimensions.

6.2.3 Random Walk Prior Models in Several Spatial Dimensions

In practice, the spatial regression problems we consider in this thesis are not confined to single spatial dimensions; the observational dataset usually contains information on a number of spatial covariates. In the following we discuss the construction of Markovian spatial prior models, based on random walks of order two, in both two and three spatial dimensions.

Random Walk Prior Models in \mathbb{R}^2

Kneib (2006) discusses the construction of a bivariate random walk on a square lattice of length m in two dimensions, noting the most commonly used neighbourhood structure is based on the four nearest neighbours; the dependence structure and coefficients of the precision matrix at the interior in presented in Figure 6.7. Boundary conditions are not a problem in the first order setting as the neighbourhood structure for a given node is a simple function of its first order neighbours; the resulting Q matrix will have a rank of $m^2 - 1$.

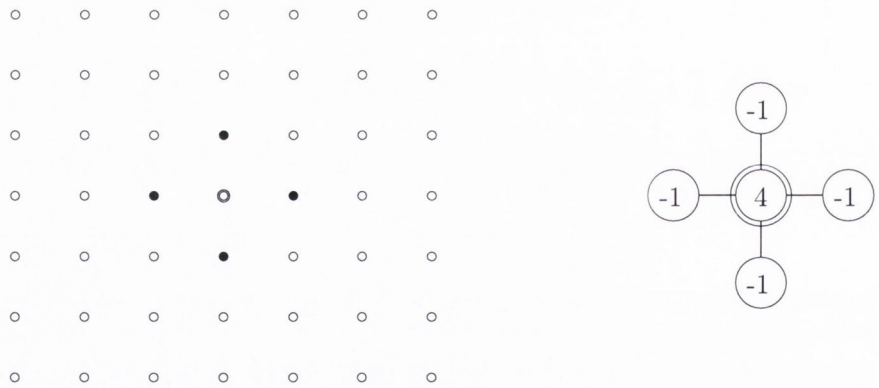


Figure 6.7: Dependence structure and coefficients of the precision matrix for a first order random walk in two spatial dimensions based on the four nearest neighbours.

Since our main requirement in modelling the spatially referenced observations is that the resulting response surface is *smooth*, we typically work with bivariate random walks of order two. Rue & Held (2005) provide a method for obtaining the precision structure of a random walk of order 2 on a lattice, based on an approximation to the *biharmonic* differential operator:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 = \frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial x^2\partial y^2} + \frac{\partial^4}{\partial y^4} \quad (6.9)$$

where the biharmonic differential operator is a two-dimensional extension of the squared second order derivative. Kneib (2006), approximates the derivatives in Equation 6.9 by difference operators based on the twelve nearest neighbours resulting in a precision matrix with non-zero elements defined by:

$$-\left(\Delta_{(1,0)}^2 + \Delta_{(0,1)}^2\right)^2 = -\left(\Delta_{(1,0)}^4 + \Delta_{(1,0)}^2\Delta_{(0,1)}^2 + \Delta_{(0,1)}^4\right) \quad (6.10)$$

where $\Delta_{1,0}$ represents the forward difference in direction (1,0) and similarly for $\Delta_{0,1}$. This approach yields the precision structure and coefficients in Figure 6.8. As noted by Kneib (2006), the neighbourhood structure at the boundaries can be obtained by careful modification of the biharmonic differential operator (see Kneib (2006) for details) which must be incorporated in the precision structure by hand (Rue & Held 2005).

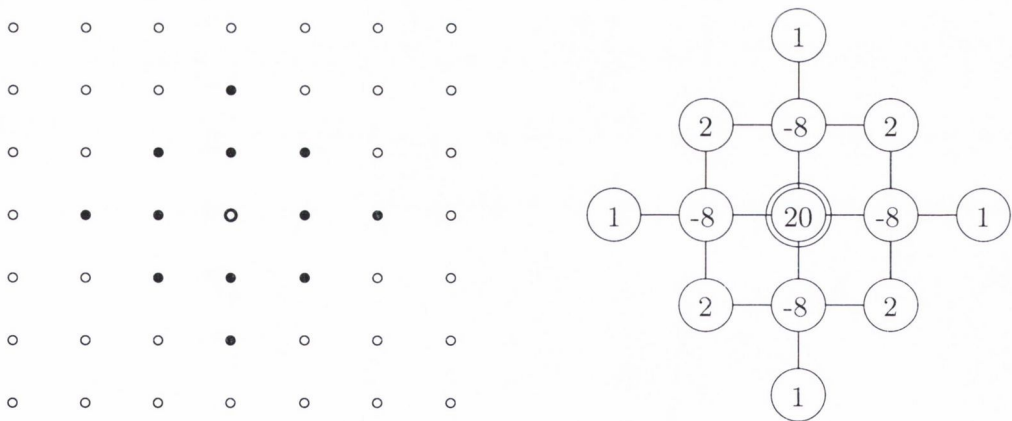


Figure 6.8: Dependence structure and coefficients of the precision matrix for a two dimensional second order random walk based on an approximation to the biharmonic differential operator.

If continuous two-dimensional space is approximated by a regular lattice, this approach can be used to obtain the neighbourhood structure of the precision matrix of an intrinsic GMRF of order two. However, an important point to note is that this approximation of continuous space by a square lattice can be somewhat wasteful if the region of interest is an irregular

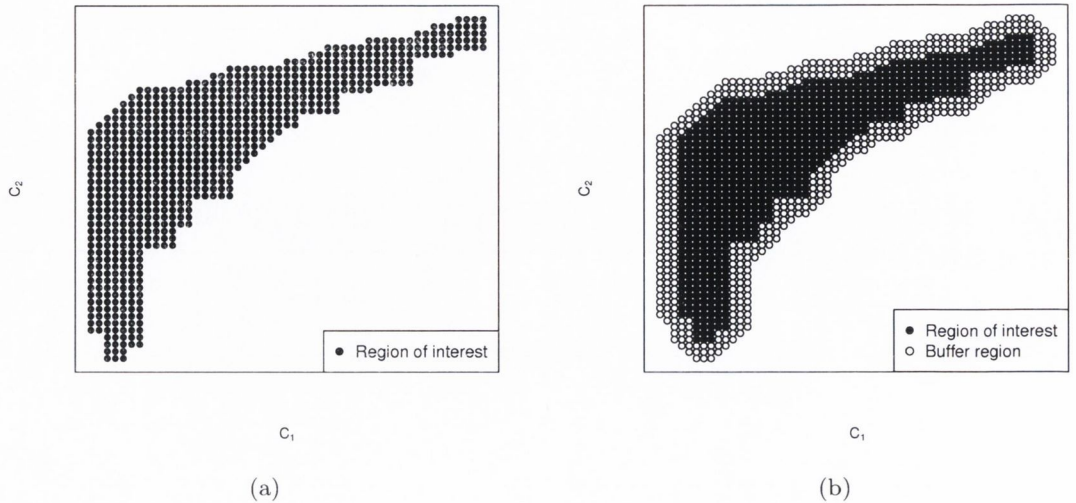


Figure 6.9: (a) Irregularly shaped region of interest for which boundary conditions are difficult to calculate and (b) the incorporation of a buffer region enables the specification of the correct neighbourhood structure at the boundary.

subset of the spatial domain, as for example in Figure 6.9 (a). If the two dimensional space is approximated by a lattice of resolution 50×50 , considerable effort will be expended in making inference on latent parameters outside of the region of interest.

A discretization of the continuous space in Figure 6.9 (a) on a square lattice of resolution $m \times m$ results in m^2 latent parameters which comprise the discretized response surface over the spatial domain. The reason for working on a square lattice is that corrections to the second order neighbourhood structure of the precision matrix at the boundary regions are known and the precision matrix with correct rank can be obtained. In contrast, the derivation of the correct neighbourhood structure for the irregularly shaped region of interest is onerous in the extreme, requiring a large number of careful modifications to the biharmonic differential operator.

An alternative to discretizing the continuous space by a two-dimensional regular lattice is once more to “constrain the infinite space to the finite space” (Besag & Higdon 1999). A convex hull of locations, including an additional buffer region of length 3 in each direction, is defined around the “region of interest”, as presented in Figure 6.9 (b). The precision structure of a random walk of order one on the irregularly shaped region, incorporating the buffer region, can be obtained as a function of its nearest neighbours (see Figure 6.7). An approximation to the correct precision structure of Q can then be obtained by convolving the lower order precision structure matrix with itself as in Equation 6.4.

In the simple example presented here $m = 50$; the dimension of Q is 2500×2500 as

compared to 1205×1205 for Q^* . Q^* has the correct precision structure in the region of interest and $\dim(Q^*) \ll \dim(Q)$, resulting in a large speeding up of inference tasks at the forward stage; regions of space which lie outside the region of interest are “cut out” from the spatial analysis. However, an important point to note is that the normalising constant must be carefully amended to ensure that the normalising constant of the resulting GMRF prior for X based on Q^* is correct. This is obtained as the product of the non-zero eigenvalues of Q^* .

Random Walk Prior Models in \mathbb{R}^3

The neighbourhood structure of the second order random walk on a three dimensional lattice is intuitively found by extending the biharmonic differential operator to three spatial dimensions:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right)^2 = \frac{\partial^4}{\partial x^4} + 2 \frac{\partial^4}{\partial x^2 \partial y^2} + 2 \frac{\partial^4}{\partial x^2 \partial z^2} + 2 \frac{\partial^4}{\partial y^2 \partial z^2} + \frac{\partial^4}{\partial y^4} + \frac{\partial^4}{\partial z^4} \quad (6.11)$$

As per Altas et al. (2002), the derivatives in Equation 6.11 can be approximated by difference operators based on the 24 nearest neighbours resulting in a precision matrix with non-zero elements defined by:

$$Q_{ij} = \begin{cases} 42 & d_{ij}^2 = 0 \\ -12 & d_{ij}^2 = 1 \\ 2 & d_{ij}^2 = 2 \\ 1 & d_{ij}^2 = 4 \\ 0 & d_{ij}^2 > 4 \end{cases} \quad (6.12)$$

where d_{ij}^2 is the squared distances between node i and node j on a 3D discrete grid, specifically $d_{ij}^2 = (i_x - j_x)^2 + (i_y - j_y)^2 + (i_z - j_z)^2$ where the (x, y, z) subscripts denote the location of node i and node j in each respective space.

The coefficients presented in Equation 6.12 represent the neighbourhood structure at an interior point in the region of interest. However, restrictions at the boundary are onerous to compute and cumbersome to incorporate. An additional point to note is that discretizations of continuous space in \mathbb{R}^3 will exacerbate the problems introduced in the preceding section; due to the curse of dimensionality a regular discretization of \mathbb{R}^3 to a lattice of dimension $m \times m \times m$ results in m^3 latent parameters; if the region of interest is irregular in shape, considerable effort will be expended in making inference on latent parameters which lie outside the region of interest, greatly slowing inference tasks if not rendering them infeasible.

A solution to this problem is once more via the convolution method - we define the region of interest and add additional layers to form a buffer region around the region of interest in order to negate edge effects. The neighbourhood structure of the first order random walk on the irregularly shaped lattice is obtained by conditioning on the first order neighbours. The first order precision structure matrix can then be convolved with itself to obtain the precision structure of a second order random walk - this matrix will have the correct precision structure at the interior and approximations to the correct structure at the boundaries which are far from the region of interest. This idea is analogous to the extension of the work of Besag & Higdon (1999), which proposes a similar approach in two spatial dimensions, to consider three spatial variables concurrently.

This approach results in vast computational savings - in Chapter 7, we detail how this method reduces the number of latent parameters in a three dimensional spatial smoothing problem from 125000 to approximately 40000. Additionally, the dimension of the precision matrices which we must manipulate is reduced accordingly, resulting in inferences procedures being speeded up by several orders of magnitude.

6.3 Fast Inverse Prediction Given New Data

So far we have given only cursory mention to the inverse stage of the calibration problem. Thus, the main objective in the remainder of this chapter is to study features of the inverse stage in detail. In particular, we are motivated to explore computationally efficient methods of making inference *inversely*.

To briefly recap, at the inverse stage of a given calibration problem, interest lies in making inferences on the unknown spatial location c^{new} corresponding to a newly observed datum y^{new} given the calibrated model $\pi(X, \theta|Y, C)$. As before, X represents the smooth latent response and θ the vector of hyperparameters which parameterise the model. For simplicity in notation in the following, we omit specific reference to the training dataset (Y, C) and thus $\pi(X, \theta|Y, C)$ simplifies to $\pi(X, \theta)$. c^{new} and y^{new} as also relabeled as c and y respectively.

At the inverse stage we introduce and ‘integrate out’ the latent variables (X, θ) in order to evaluate $\pi(c|y)$:

$$\pi(c|y) = \int_{\theta} \int_X \pi(c, X, \theta|y) dX d\theta \quad (6.13)$$

$$= K \int_{\theta} \int_X \pi(y|c, X, \theta) \pi(c) \pi(X, \theta) dX d\theta \quad (6.14)$$

In order to obtain the normalising constant K in Equation 6.14, the continuous space under consideration is discretized to a finite number, n of spatial locations $C = (c_1, \dots, c_n)$. K is

then obtained by evaluating the unnormalised posterior at each of the n spatial locations and dividing through by the sum of their values to provide a normalised posterior which sums to 1, i.e.

$$K = \sum_{i=1}^n \int_{\theta} \int_x \pi(y|x_i, \theta) \pi(c_i) \pi(x_i|\theta) \pi(\theta) dx_i d\theta \quad (6.15)$$

where $x_i = X(c_i)$ is univariate. In this thesis $\pi(x_i|\theta)$ is always of univariate Gaussian form and the posterior for θ is discretized to a regular grid due to the use of the INLA algorithm for approximate Bayesian inference on model parameters (see Section 3.3). If the likelihood for the newly observed count y is Gaussian, the unnormalised posterior in Equation 6.14 is then available analytically by summing over the discretized posterior of θ .

Conversely, if the likelihood for the observed counts is non-Gaussian, the integral in Equation 6.14 is no longer tractable. The posterior at each grid location must then be evaluated by alternative methods; options include the use of sampling based algorithms or numerical integration algorithms such as quadrature. In the following we discuss the computational drawbacks of numerical integration of the posterior in Equation 6.14 and provide the motivation for a sampling based solution.

6.3.1 Numerical Evaluation Of Posteriors

The inverse stage is necessarily computational in nature - in order to obtain the normalising constant for $\pi(c|y)$, the unnormalised posterior must be evaluated at each and every spatial location on the discretized grid. However, much like the forward stage, the inverse stage is subject to the curse of dimensionality. The number of equally spaced points required to discretize a d dimensional space in \mathbb{R}^d increases as a power law function of d - this can be observed in Figure 6.10 below:

In Figure 6.10, for the provided examples, we observe that $\pi(c|y)$ is significantly non-zero at only a very small subset of the spatial locations. As a result, in using numerical integration algorithms for posterior evaluation, we will expend significant effort evaluating posteriors on location which are essentially zero. This problem is exacerbated if uncertainty in model hyperparameters is additionally taken into account.

The ‘integration out’ of the latent parameters (X, θ) in Equation 6.15 is completed as follows; conditional on each value of $\theta_k \in \pi(\theta)$ where $\pi(\theta)$ is represented on a discrete grid, $\pi(X|\theta_k) \sim GMRF(\mu(\theta_k), Q(\theta_k))$ where $\pi(X|\theta_k) = \pi(x_1, \dots, x_n|\theta_k)$; n here is the number of discrete grid points. The evaluation of $\pi(c|y)$ at each point c_i on the discrete grid requires numerically integrating out the corresponding univariate x_i for each value $\theta_i \in \theta$:

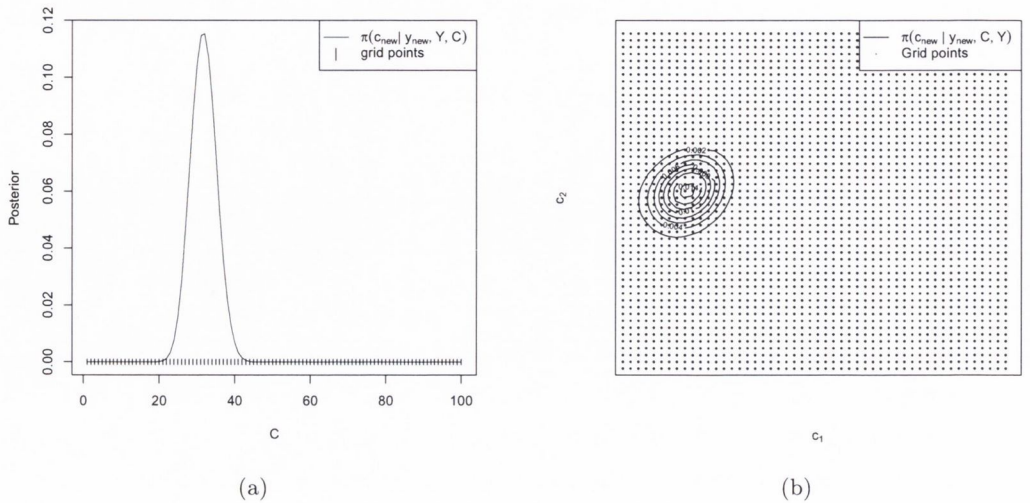


Figure 6.10: The number of gridpoints at which the posterior must be evaluated increases with each additional spatial dimension considered. In (a) $\pi(c^{\text{new}}|y^{\text{new}})$ is only significantly non-zero at 20 of the 100 spatial locations (a continuous line is plotted for the probability density). In (b) we present a two-dimensional example where $\pi(c^{\text{new}}|y^{\text{new}})$ is only significantly non-zero at a small subsection of locations on a 50×50 lattice.

$$K = \sum_{i=1}^n \sum_{k=1}^m \int_x \pi(y|x_i, \theta_k) \pi(c_i) \pi(x_i|\theta_k) \pi(\theta_k|Y) dx_i \quad (6.16)$$

For a fixed number of quadrature points, the computational cost of this numerical integration is $\mathcal{O}(nm)$ where n is the number of discrete locations which comprise the space under consideration and m represents the number of discrete points at which $\pi(\theta)$ is evaluated. In the context of the palaeoclimate reconstruction problem, if model specification requires the use of a large numbers of hyperparameters and several spatial dimensions are considered, inference procedures at the inverse stage will be slowed tremendously. This problem may be slightly alleviated if an empirical Bayes approach is employed and the value(s) of θ are fixed at their posterior mode, reducing the computational cost to $\mathcal{O}(n)$. However, if posterior uncertainty in θ is significant, the resulting posteriors on spatial location will be erroneous in location or spread.

6.3.2 Sampling Scheme for Computationally Efficient Inverse Inference

The use of sampling based inference procedures may provide a solution to the problems raised in the preceding section - as opposed to exact numerical evaluation of the posterior at each

of the n spatial locations, m times in a fully Bayesian analysis, sampling algorithms can be utilised so that samples are only obtained from regions of space where the probability is non-zero. Therefore, sampling based methods should be more robust with regard to problems such as the curse of dimensionality and, in particular, the inversion of forward models with numerous hyperparameters.

An MCMC scheme which provides a method for achieving this aim is the *Metropolis within Gibbs algorithm*. The development of a methodology for increasing computational efficiency in prediction at the inverse stage of calibration problems, such as the palaeoclimate reconstruction problem, is considered a novel contribution in this thesis.

Metropolis Within Gibbs Algorithm

The target distribution from which we wish to sample, $\pi(c|y)$, is generally intractable. However, this issue is resolved by augmenting the target distribution with the latent variables (X, θ) with the resulting augmented target distribution $\pi(c, X, \theta|y)$ tractable. As θ only depends on (c, y) indirectly, $\pi(\theta|X, c, y) \approx \pi(\theta)$. The full posterior may be rewritten as:

$$\pi(c, X, \theta|y) \propto \pi(y|c, X, \theta)\pi(c)\pi(X|c, \theta)\pi(\theta) \quad (6.17)$$

If the conditional distributions $\pi(c|X, \theta, y)$ and $\pi(X|\theta, c, y)$ are of known distributional form, the required samples can be obtained using a Gibbs sampling step, by sampling from each of the full conditional distributions in turn. Denoting $X(c)$ by x , the broad outline of the Gibbs sampling step is presented as follows:

1. Choose arbitrary starting values $x^{(0)}$, $\theta^{(0)}$ and $c^{(0)}$.
2. For $i = 1, \dots, N$:
 - Sample $c^{(i)} \sim \pi(c|x^{(i-1)}, \theta^{(i-1)}, y)$
 - Sample $x^{(i)} \sim \pi(x|\theta^{(i-1)}, c^{(i)}, y)$.
 - Sample $\theta^{(i)} \sim \pi(\theta)$

Here each $c^{(i)}$ is univariate and represents the climate location sampled at iteration i from the set of all c values which comprise the discretized grid. Each $x^{(i)}$ is also univariate and represents a sample from the multivariate Gaussian spatial field defined at each of the n discretized climate locations.

However, in the problems considered in this thesis we are unable to sample directly from the conditional distributions in the above (save $\pi(\theta)$) as they tend to be unknown. This issue

is resolved by using a Metropolis - Hastings step to generate the required samples; this is then the *Metropolis within Gibbs algorithm*. In the following we discuss the construction of proposal distributions for x and c .

Proposal Scheme for c

Proposal schemes for c are complicated by the fact that $\pi(c|y)$ is generally nonstandard, often multimodal (Haslett et al. 2006) consisting of disjoint probability regions (Bhattacharya 2004) with no paths between them. For example:

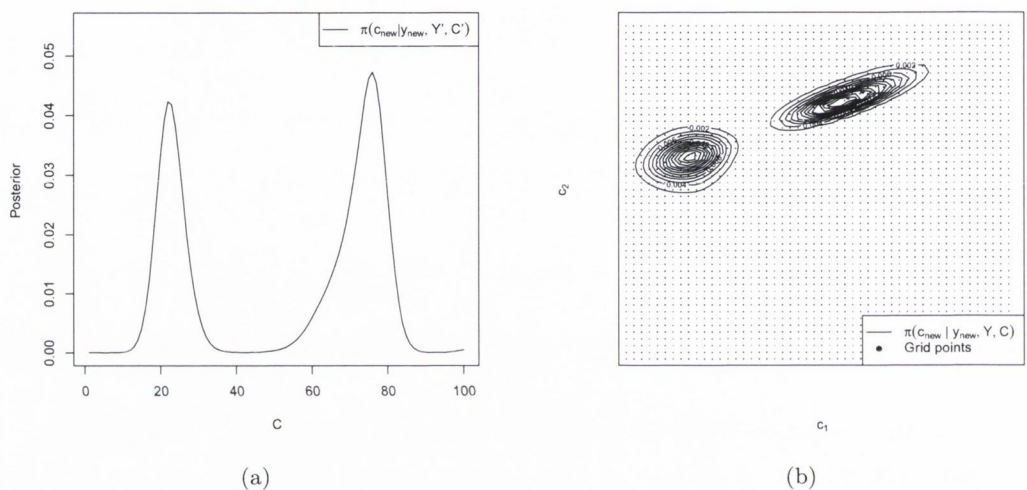


Figure 6.11: Examples of multimodal posterior distributions for spatial location in $d = 1$ and $d = 2$ climate dimensions.

As a result, proposal schemes based on random walks which propose ‘local’ moves for c , are subject to the problem of getting stuck in local modes - the starting value specified for c is obviously important in this setting. Conversely, proposal distributions which propose large moves will suffer from low acceptance rates due to the significant number of locations at which probability mass is non-zero, as for example in Figure 6.11 (b) above. This problem worsens with increasing spatial dimension or ‘peaked-ness’ of the target posteriors.

The “best” idea in a sense, as noted by Gilks et al. (1996), is to obtain a proposal distribution q , which closely resembles the target posterior - conditional on the modal hyperparameters $\theta = \theta^*$, we propose to utilise the Laplace approximation (see Tierney (1994)) to $\pi(c|y, \theta^*)$ as the foundation for our proposal density for $\pi(c|y)$. Essentially, we evaluate the Laplace approximation to $\pi(c|y, \theta^*)$ at each spatial location $(c_1, \dots, c_n) \in C$, to obtain the required proposal distribution, i.e.

For $i = 1, \dots, n$ do

$$\tilde{\pi}_{LA}(c_i|y) \approx \frac{\pi(y, x_i, c_i, \theta^*)}{\tilde{\pi}_G(x|y, c_i, \theta^*)} \Big|_{x_i=x_i^*(\theta^*)} \quad (6.18)$$

In the above the distribution of x_i corresponding to the mode of θ , $\pi(x_i|\theta^*)$ is integrated out via the Laplace approximation at every spatial location $c_i \in C$, to obtain an approximation to $\pi(c_i|y, \theta^*)$. This is completed at each location on the grid and the probability values are summed over C to provide a normalised distribution which sums to unity.

The procedure for obtaining this approximation is outlined in Bishop (2006, pages 213-215), or alternatively in Rue & Held (2005, pages 167-170) and thus is not explored in further detail here. The Laplace approximation in this setting can be quite accurate, or indeed exact if the underlying response is Gaussian. However, as noted by Bjornkamp (2011) - the quality of this approximation is subject to the quality of the Gaussian approximation to $\pi(x_i|y, c_i, \theta^*)$. Bjornkamp (2011) notes that the Laplace approximation is subject to poor performance if $\pi(x_i|y, c_i, \theta^*)$ is skewed, which frequently occurs in the presence of low value or zero counts for y .

Additionally, the proposal region must be *bounded* (Albert 2007), i.e. $\pi(c|y)/\tilde{\pi}_{LA}(c|y, \theta^*) > 0 \forall c$. However, as we are only evaluating the Laplace approximation at the mode of θ , the approximation will tend to produce a proposal density for c that tends to be less conservative than the target posterior or even slightly erroneous in location. We correct for this by raising $\tilde{\pi}_{LA}(c|y, \theta^*)$ to the power of α and renormalising, where $\alpha \in (0, 1)$, resulting in a proposal distribution that is more diffuse than the target. As a general rule of thumb, we propose using a value of $\alpha = .5$. In Section 6.3.3 below, we discuss a method for determining if q is sufficiently bounded.

Using q obtained via the Laplace approximation, samples may be obtained using a Metropolis-Hastings step as follows:

1. Sample a candidate value $c^* \sim q(c^*|c^{(i-1)})$
2. Compute the ratio:

$$R = \frac{\pi(c^*|\dots)q(c|c^*)}{\pi(c|\dots)q(c^*|c)}$$

3. If $\min(R, 1) > \text{Uniform}(0, 1)$ then $c^{(i)} = c^*$ else $c^{(i)} = c^{(i-1)}$
4. Repeat for a number of Metropolis-Hastings steps, if required, to reduce correlation in the samples

In the following we use an *independence sampler* (Tierney 1994) for q , i.e. $q(c^*|c) = q(c^*)$ and $q(c|c^*) = q(c)$ - the probability of each independent move can be easily obtained from the

proposal density. As we will observe in the following sections, the Laplace approximation as a proposal density closely approximates the target posterior and thus independent samples from q have high acceptance rates. The generated samples can then be fed back into the Metropolis-within-Gibbs algorithm presented above to provide samples from the target posterior.

Proposal Scheme for X

Proposal schemes for $x = X(c)$ are much simpler to construct as, conditional on each value $\theta_k \in \theta$, each $\pi(x|c, y, \theta_k)$ is continuous. Additionally, $\pi(x|c, y, \theta_k)$ is ‘Gaussian-like’ being univariate and unimodal, though in the presence of zero values it may exhibit skewness (Bjornstrom (2011)). A simple proposal density for x is the *random walk proposal* density.

The random walk proposal density is symmetric, i.e. $q(x^*|x) = q(x|x^*)$ thus the acceptance probability simplifies to the ratio of the target density evaluated at the proposed new and old values. As noted by Rue & Held (2005), a typical example of a random walk proposal is the addition of a mean zero normal distribution to the current value of x , i.e. $x^* = x + z$ where $z \sim \mathcal{N}(0, \sigma_d^2)$.

The broad outline of the approach is as follows:

1. Sample a candidate value $x^* \sim q(x^*|x^{(i-1)})$

2. Compute the ratio:

$$R = \frac{\pi(x^*|\dots)}{\pi(x|\dots)}$$

3. If $\min(R, 1) > \text{Uniform}(0, 1)$, then $x^{(i)} = x^*$ else $x^{(i)} = x^{(i-1)}$

4. Repeat for a number of Metropolis steps, if required, to reduce correlation in the samples

σ_d^2 is a scale parameter which affects the mixing of the algorithm and can be tweaked in order to increase acceptance rates - Roberts et al. (1997) provide analytical results, proposing that an acceptance rate around 50% in the univariate setting is appropriate. If a ‘good’ value for σ_d^2 is not known *a priori*, the algorithm can be initialized with $\sigma_d^2 = 1$ and its value adjusted accordingly during a burn-in period until the desired acceptance rate is approximately achieved.

6.3.3 Performance of the Approach for Univariate Y

In the following we use a simple toy example to evaluate the performance of the proposed sampling algorithm with regard to fast model inversion in the univariate setting.

Toy Example

Model training data, comprising univariate counts $Y = (y_1, \dots, y_5)$ observed at 5 distinct spatial locations, is presented in Figure 6.12. At the forward stage the individual counts are modelled as Poisson distributed and linked to the underlying latent field X through the use of a log-link function where X is defined on an equally spaced grid of length 100.

The prior for X is an intrinsic GMRF prior of order 2 with precision matrix $Q = \kappa R$ (see Section 6.2.2 for details on the structure of R) and the prior precision parameter κ is assigned a non-informative $\Gamma(1, .00005)$ prior. The resulting hierarchical model for the data is:

$$y_i \sim \text{Poisson}(\lambda_i) \tag{6.19}$$

$$\lambda_i = \exp(x_i) \tag{6.20}$$

$$X \sim \text{IGMRF}(\kappa R) \tag{6.21}$$

$$\kappa \sim \Gamma(1, .00005) \tag{6.22}$$

The INLA algorithm is used for approximate Bayesian inference on the unknown model parameters. In Figure 6.13 (a), the posterior distribution of κ , $\pi(\kappa|Y)$, is presented on an equally spaced grid of length 50. The posterior for the latent field, $\pi(X|Y)$, having marginalised over κ (see Equation 3.20) is presented in Figure 6.13 (b). As the focus here is on the inverse stage, additional details of model fitting are omitted.

In the following, we evaluate the proposed sampling-based algorithm for model inversion by comparing its performance, in terms of computational speed and inverse predictive accuracy, to model inversion via quadrature for a number of toy examples.

Comparison of Numerical Integration vs Sampling for Inverse Prediction

Given a new count, $y = 10$, for which the corresponding spatial location is unknown, we observe (see Figure 6.14 (a)) that the Laplace approximation to the posterior $\tilde{\pi}_{LA}(c|y, \kappa^*)$ is slightly erroneous but nonetheless provides a good approximation to the target distribution $\pi(c|y)$. The proposal distribution, $q(c)$, required for the Metropolis-within-Gibbs sampling scheme, is then obtained by raising the Laplace approximation $\pi_{LA}(c|y, \kappa^*)$ to the power of $\alpha = .5$ and renormalising. σ_d^2 , initialized with value .1, is rescaled during the first 50 iterations such that the acceptance rate in the Metropolis steps for x is approximately 50%, as per Roberts et al. (1997). Finally, the starting value for c is randomly sampled from $\tilde{\pi}_{LA}(c|y, \kappa^*)$ and the scheme proceeds by iteratively sampling from $\pi(c|x, \theta, y)$, $\pi(x|\theta, c, y)$ and $\pi(\theta)$ in turn, to provide the required samples from $\pi(c|y)$.

In Figure 6.14 (b - c) we plot histograms of the resulting samples generated - intuitively, the

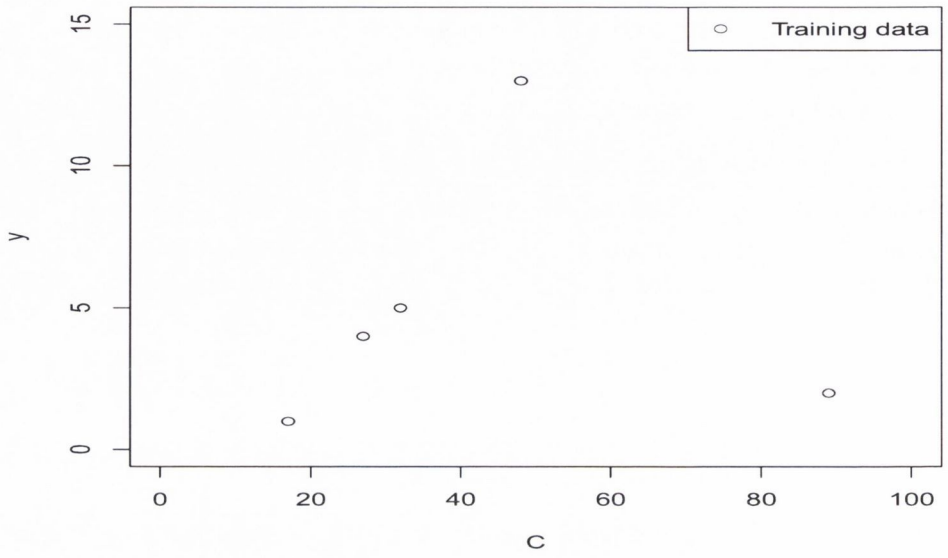
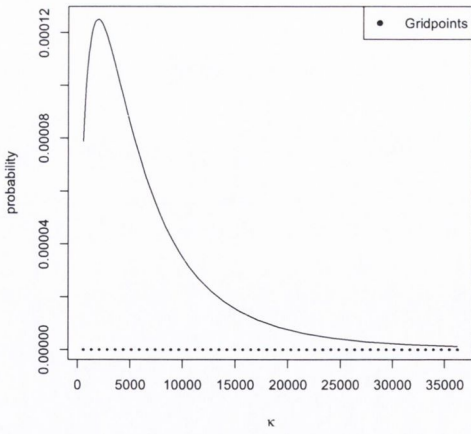
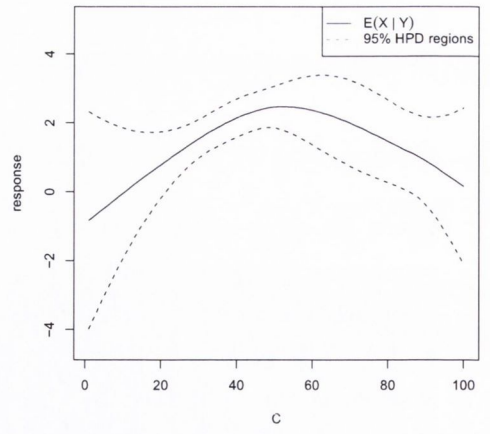


Figure 6.12: Training data.



(a)



(b)

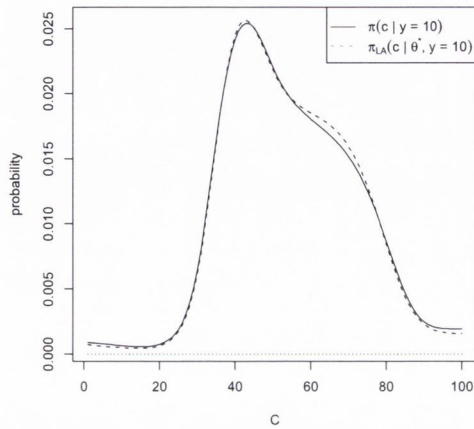
Figure 6.13: (a) $\pi(\kappa|Y)$ presented on a grid of length 50 and (b) the inferred response $\pi(X|Y)$.

approximations to the target posterior improve in accuracy for increasing number of samples. The time cost comparison is as follows; evaluating $\pi(c|y = 10)$ via deterministic Laguerre quadrature with 25 evaluation points and incorporating all posterior uncertainty in κ , takes approximately .22 seconds. Conversely, using sampling based methods, the time taken to generate 5000 samples from the target posterior is .03 seconds, increasing to .07, .18 and .34 seconds for 10000, 25000 and 50000 samples respectively - the time taken to produce the samples increases linearly in the number of samples. The acceptance rate is around 70% for the independence sampler proposal density provided by the scaled Laplace approximation to $\pi(c|y, \kappa^*)$ and 5 Metropolis-Hastings steps are used for each Gibbs step to provide samples for c that are approximately independent, as illustrated by the autocorrelation plot of the generated samples in Figure 6.15 (b).

In the simple examples presented above, the “true” posterior predictive distribution is defined over a relatively “wide” region. As a result, large numbers of samples are required for accurate approximations to the posterior. However, it is clear that extensive time savings can be made at the inverse stage, via the proposed sampling scheme, if the target posteriors are relatively “peaked”, as in Figure 6.17 (b) and Figure 6.17 (d). 10000 (approximately) independent samples are sufficient to accurately approximate the target posterior $\pi(c|y = 0)$, as observed in .06 seconds (Figure 6.17 (b)). Similarly, the time taken to generate the required number of samples (5000) for $\pi(c|y = 50)$ is .03 seconds (Figure 6.17 (d)). These times compare extremely favourably to the .22 seconds required for exact evaluation of the each posterior using 25 point Laguerre quadrature.

However, it must also be noted that the use of sampling algorithms for inverse prediction have one important caveat; it is difficult to determine when “sufficient” samples have been obtained to accurately approximate target posteriors. In practice the number of samples generated is not determined by the Monte Carlo error but by practical considerations such as time constraints - for example in the palaeoclimate reconstruction problem considered in Chapter 7, samples must be generated for each of 7742 independent reconstruction problems. Appreciable and significant uncertainty may exist in the target posteriors produced given the constrained numbers of samples, as highlighted by a comparison of the sample histograms in Figure 6.14(b - d).

Of course a further concern is the evaluation of situations where the proposal density is not sufficiently bounded. In Figure 6.16 we illustrate that poor proposal distributions, in this case due to an overly peaked proposal distribution for c , can once more be identified by viewing the autocorrelation plot of the generated samples. Proposal distributions that are too “diffuse” or conservative may be identified in a similar manner as acceptance rates will be similarly affected.



(a)

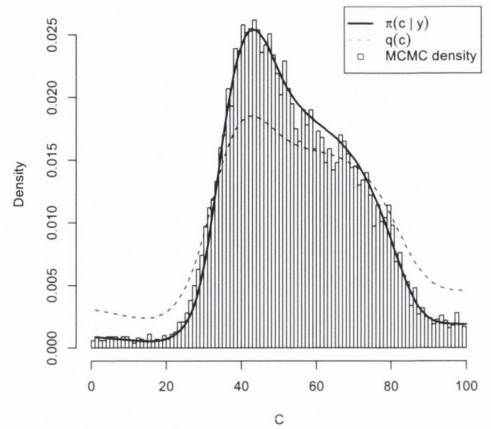
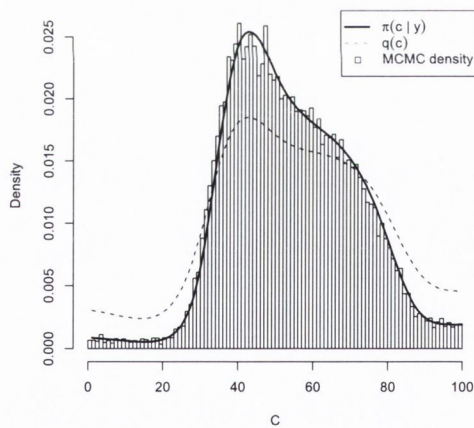
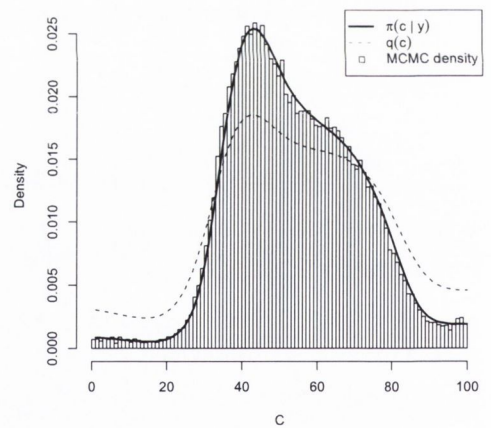
(b) $N = 10000$ (c) $N = 25000$ (d) $N = 50000$

Figure 6.14: Comparison of predictive posteriors, obtained using via deterministic and sampling based methods. N represents the number of samples generated. In the example considered here $y = 10$. The accuracy of the sampling-based approximation to $\pi(c|y)$ demonstrably improves as the number of samples increases.

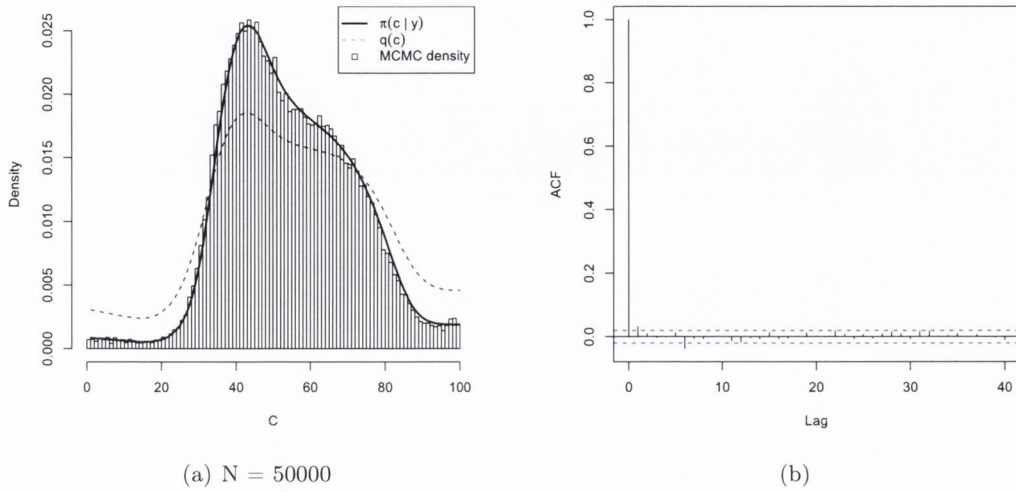


Figure 6.15: Due to the excellent approximation to $\pi(c|y = 10)$, provided by $\tilde{\pi}_{LA}(c|y, \kappa^*)$ and the use of multiple (5) Metropolis steps for each iteration, there appears to be little or no correlation between the generated samples for c .

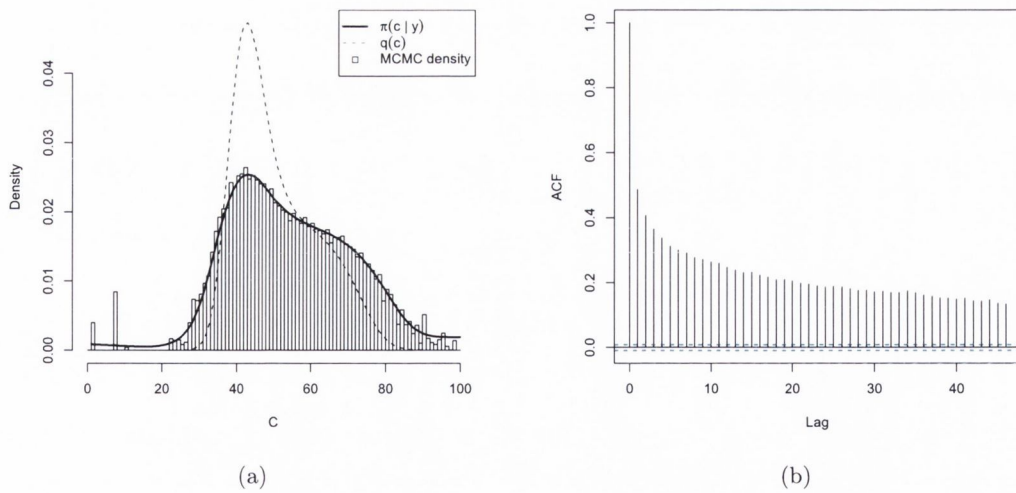
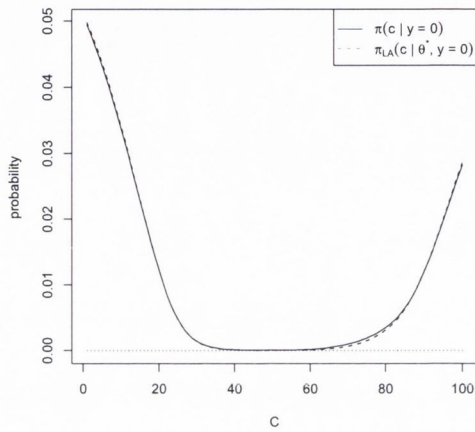
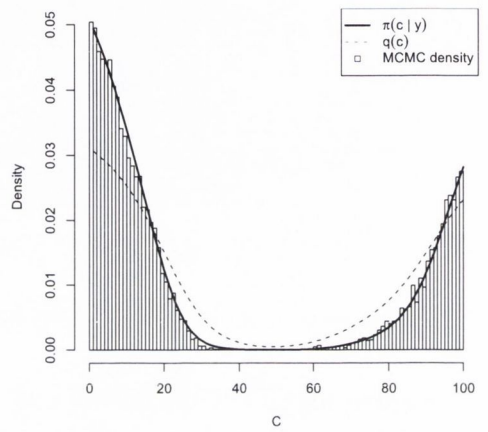


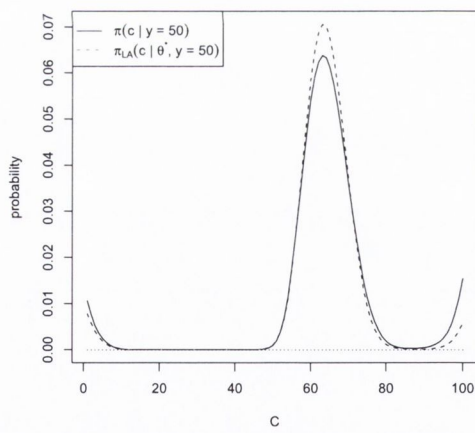
Figure 6.16: Sample autocorrelation plots of the generated samples provide a method of detecting poor proposal distributions for c . In (a) we observe that the proposal distribution, $q(c)$, is not sufficiently conservative, resulting in a high rate of autocorrelation between the generated samples for c .



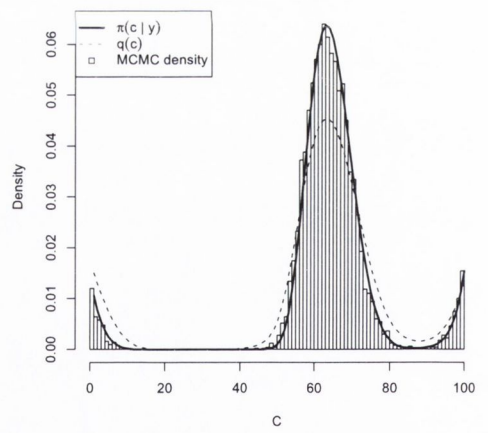
(a)



(b) N=10000



(c)



(d) N=5000

Figure 6.17: Fewer samples are required to accurately approximate target posteriors that are relatively peaked. In (b) $y = 0$ and in (d) $y = 50$.

6.3.4 Extension to the Multivariate setting

In practice, the problems considered in this thesis are multivariate in nature, both in terms of the data Y and the climate C .

Multivariate Y

For numerical integration algorithms, the computational cost of model inversion increases as a linear function of the number of taxa which are jointly considered. For example, if the computational cost of model inversion in the univariate setting, where $Y = y$, is $\mathcal{O}(nm)$ (see Section 6.3.1 above), the computational cost when Y is multivariate of length p (i.e. $Y = (y_1, \dots, y_p)$) is $\mathcal{O}(nmp)$. Due to the (assumed) conditional independence of each taxon given climate, the (multivariate) model inversion problem decomposes into the product of p separate, univariate model inversion problems.

For the sampling-based scheme, the cost of obtaining the Laplace approximation is also linear in the number of taxa, however, this is typically a much smaller calculation than the numerical integration equivalent. This is because the Laplace approximation is only evaluated once at each location, conditional on the modal hyperparameters of each individual taxa, i.e. $\pi_{LA}(c|Y, \theta^*) = \prod_{j=1}^m (\pi_{LA}(c|y_j, \theta_j^*))$. Here θ_j^* represents the modal hyperparameters of taxon j . If Y is multivariate, X is also, but due to the conditional independence assumption, the generation of samples for multivariate X simply decomposes into the generation of univariate samples for each of the p components of X independently.

The computational advantages enjoyed by the sampling-based approach become more pronounced as the dimension of Y increases. To highlight this point, we re-examine a simulated example considered in Section 5.2.2 previously. In that particular section, a simple multivariate regression problem was used to illustrate how posteriors on climate become increasingly peaked with each additional taxa considered, as indicated by reducing average values of the *MSEP* (see Figure 5.2 (a)). The important point to note however is that, as the target posteriors become increasingly peaked, the number of samples required for accurate approximations reduces and thus the sampling scheme will increase in its computational efficiency.

In Figure 6.18, this point is visually demonstrated. Two examples are considered where (a) Y is multivariate of dimension 5 (i.e. there are 5 separate plant taxa and (b) multivariate of dimension 10. The time taken to exactly evaluate the posteriors on climate in each respective setting, using deterministic Laguerre quadrature with 25 evaluation points, is 1.13 seconds and 2.27 seconds respectively.

For the sampling-based approach, 10,000 samples in the 5 taxon setting are sufficient to accurately approximate the target posterior. As the climate posterior in the 10 taxon setting is substantially more peaked, only 5,000 samples are required. The time taken to produce each set of samples is just .27 seconds and .25 seconds respectively. The introduction of

additional taxa at the inverse stage results in the proposed sampling scheme becoming more computationally efficient due to the tightening of the predictive regions of target posteriors.

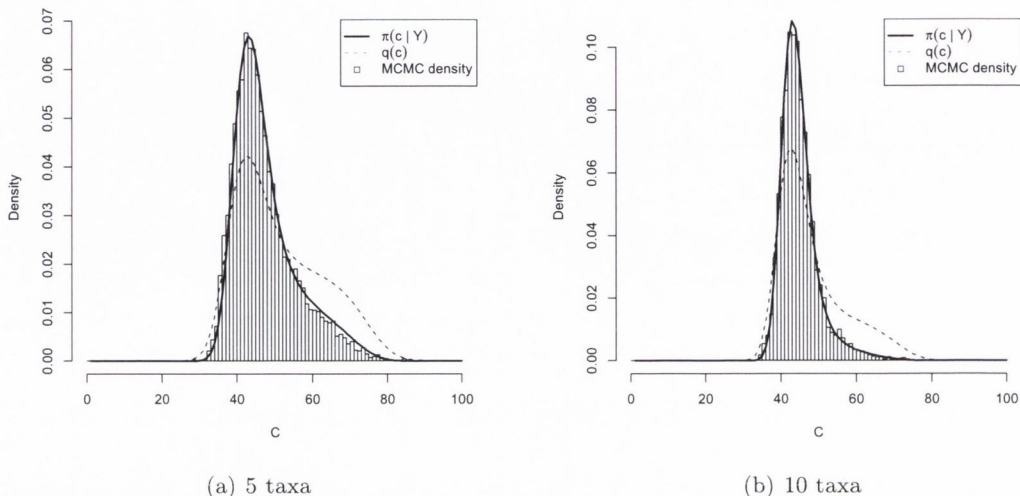


Figure 6.18: In (a) Y is multivariate of length 5 and in (b) Y is multivariate of length 10. Observe that the posterior distributions become more “peaked” for increasing dimension of Y and thus less samples are required to accurately approximate the target posterior.

A final point to note is that the computational advantage of the sampling approach becomes even more pronounced for increasing dimension of C . This is because, as per Figure 6.11 (a) and Figure 6.11 (b), the percentage of spatial locations at which the target posterior is non-negligible is generally much less in two spatial dimensions than one. Thus, the use of the sampling scheme, which avoids visiting spatial locations at which the target posterior is essentially zero, results in substantial time savings. This claim is evidenced in Section 7.4; it is demonstrated how the use of the sampling-based scheme helps half the time taken for climate reconstruction at the Glendalough site in the two dimensional climate setting and reduces the time taken by a factor of approximately 20 in the three dimensional climate setting, as compared to numerical integration based methods.

6.4 Conclusions

If statistical models are constructed which do not include all spatial covariates upon which the response depends, inferences at the inverse stage will be erroneous. This result was highlighted via a simple toy example, where the failure to account for the interaction of a smooth surface over a set of spatial covariates was manifested in spurious multimodality and mislocation of inverse predictive posteriors.

In situations where the observational dataset is too large for the consideration of spatial models based on Gaussian random fields, due to the “big n problem”, discretizations of the space under consideration to a regular grid and the use of GMRF models with sparse precision structures provide a solution. However, both the dimensionality of the discretization of space chosen and the neighbourhood structure specified have a large impact on the computational efficiency of this approach. Furthermore, due to the curse of dimensionality, the maximum number of spatial dimensions for which we can consider the use of such methods is 3 at most, due to the large number of random variables introduced by regular discretizations of space.

Extensive computational savings can be made by “cutting out” regions of space which are not of interest. However, problems then arise regarding the specification of the precision structures of intrinsic GMRFs of order 2 on the irregularly shaped spatial domain. We provide an approximate solution to this problem as follows: first a buffer region is incorporated around the region of interest. The precision structure of a random walk of order one on the augmented space is then easily obtained as a function of its first order neighbours. An approximation to the correct second order structure can then be obtained as a convolution of the first order structure with itself. This structure matrix will have the correct neighbourhood structure at the interior but incorrect structure at the boundary, which is constructed however, to be far from the region of interest.

The discretization of multidimensional space results in slow inference procedures at the inverse stage. This is because model inversion via deterministic integration algorithms require the evaluation of the inverse predictive posterior at every grid location on the discretized space. Numerical methods for the evaluation of posteriors will waste substantial effort evaluating probabilities at many grid locations that are essentially zero.

A Metropolis-within-Gibbs sampling scheme is constructed to address this issue, with the proposal density for moves on the discretized spatial domain obtained from a scaled Laplace approximation to the target posterior. The Laplace approximation has to be evaluated at every gridpoint but is however, much more computationally efficient than integration via quadrature, requiring the computational equivalent of 9 quadrature points for the calculation at each node. The sampling algorithm avoids the repeated sampling from locations with negligible probability mass, which blights deterministic integration approaches. Furthermore, the computational advantages enjoyed by the sampling-based scheme become more pronounced with increasing dimension of Y and C .

Chapter 7

Application: The Palaeoclimate Reconstruction Project

In this chapter we harness many of the models and methods developed in the preceding chapters to extend existing models for palaeoclimate reconstruction. Our main contributions include the extension of existing models to incorporate an additional climate variable, the development of a methodology for fast criticism of the model training data and the development of a computationally efficient algorithm for *quickly* obtaining climate samples at the inverse stage. A fossil climate reconstruction at Glendalough is utilised to display the striking differences in climate reconstructions obtained by the incorporation of an additional climate covariate into the forward model.

7.1 Bayesian Palaeoclimate Reconstruction Project

In the following we present the work contributed by this thesis to the ongoing Bayesian palaeoclimate reconstruction project, as described in Haslett et al. (2006) and Salter-Townshend (2009). We begin with a brief recap of the reconstruction dataset, previously introduced in Section 1.2.1, and detail further important features of the dataset absent from the previous discussion.

7.1.1 The RS10 Dataset

The RS10 dataset (Allen et al. 2000) consists of a collection of 7742 sample pollen proportions for each of 28 distinct plant taxa, as well as associated measurements for a number of climate variables. In this chapter we consider models in terms of three of the climate variables; the number of growing degree days above 5°C (GDD5), the mean temperature of the coldest month (MTCO) and an estimate of the ratio of the actual to potential evapotranspiration

(AET/PET). As detailed by Huntley (1993), these three climatic variables provide the main constraints which govern the geographical ranges of individual plant species.

However, there are additional features of the reconstruction dataset which impact on model performance; as noted by Haslett et al. (2006), at very many of the sampling locations the count totals used to produce the pollen proportions are unknown. As a result, the rather unsatisfactory decision is made to express the observed counts at each location as a per mille of the sum at that location; a total count sum of 1000 is then assumed. Furthermore, the climate measurements at a specific site location are not known explicitly, but interpolated from information available from the nearest meteorological weather station. Salter-Townshend (2009), citing the uncertainty in climate locations as the main factor in poor model predictive performance, attempted to account for this uncertainty in an ad-hoc manner at the inverse stage. However, the work contained in this chapter will reveal that the compositional nature of the data and the covariates used in the palaeoclimate reconstruction models are of far greater importance.

Finally, though the extension of the existing palaeoclimate reconstruction models of Haslett et al. (2006) and Salter-Townshend (2009) to three climate dimensions is considered a novel contribution in this thesis, an additional climate variable MTWA - the *mean temperature of the warmest month* is not accounted for in the forward models. Later in this chapter we identify signals that suggest this variable may be of importance for accurate prediction, and that the forward models should possibly be amended for its inclusion. In Section 8.2.2 we detail the difficulties of this task.

7.2 Forward Modeling and Inference Methodology

At the forward stage, models are calibrated for the relationship between the chosen climate variables and the observed response. Here the training dataset Y consists of $n = 7742$ climate referenced (C) observations for each of $N_T = 28$ plant taxa. For simplicity in notation in the following, we suppress explicit reference to C .

In Section 5.1.1, we detailed how multivariate forward models for the palaeoclimate problem are too computationally expensive to consider, instead, as per Section 5.1.2, univariate models are used to model each plant taxon separately - the joint forward model for the palaeoclimate problem is thus decomposed into the product of conditionally independent univariate components:

$$\pi(X, \theta|Y) = \pi(X|\theta, Y)\pi(\theta|Y) \tag{7.1}$$

$$= \prod_{i=1}^{N_T} \pi(X_i|Y_i, \theta_i)\pi(\theta_i|Y_i) \tag{7.2}$$

Here, as previously, we denote by Y_i the pollen counts corresponding to taxon i and by X_i the latent response surface of taxon i ; θ_i denotes the model hyperparameters relevant to taxon i . The use of separate, independent univariate models for each taxa introduces many computational conveniences at the forward stage - the forward model for each plant taxa can be fitted independently, greatly reducing inference tasks. This is the by-taxa model of Salter-Townshend (2009).

However, Salter-Townshend (2009) detailed how the use of this decomposition results in erroneous inferences at the inverse stage of the palaeoclimate problem - unmodelled, residual dependence structure in the latent field (see Section 5.2.2) or at the likelihood level (Section 5.3), is manifested as poor predictive performance of the calibrated models. Recognising the compositional nature of the training dataset as a factor, nesting structures (Section 5.4), which still facilitated decomposition of the forward stage, were utilised in order to reduce the error. However, the introduced nesting structure does not decompose the joint model exactly, on account of the failure of the author to explicitly model sources of N-inflation in the RS10 dataset.

Our contributions in this thesis to the forward modeling stage involve the extension of the nesting structure introduced in Salter-Townshend (2009) (see Figure 7.11) to the lowest levels. Using the methods developed in Section 5.4.3, the appropriate nesting structures for the lowest levels are learned from the data. The joint likelihood for the ($N_T = 27$) groups comprising the nested structure can be written in the form of Equation 7.2 above, and the forward model for each group can then be fitted independently as in the case of the marginals, or by-taxa model.

Furthermore, using the methods detailed in Section 6.2.3, we extend the models of Salter-Townshend (2009) to consider an additional climate variable, overcoming the additional computational burden through the use of *buffer regions*. Statistical consistency issues regarding the nested likelihood are addressed by the creation of a symmetric zero/N-inflated Binomial likelihood model (Section 5.5.2). In the following, we introduce the models used in this chapter for the latent X_i and the data response and briefly discuss details of inference.

7.2.1 Modeling the Latent X_i

In this thesis, climate models in both two and three dimensions are considered for each X_i . In the two-dimensional climate setting, the two climate variables considered are GDD5 and MTCO. As per Haslett et al. (2006), climate space is discretized to a 50×50 grid with climate locations “pushed” to their nearest grid location. This results in 2500 latent parameters, which comprise the response surface of a given plant taxa. The prior on each latent surface is an intrinsic GMRF of order two:

$$\pi(X_i) \sim \text{GMRF}(Q) \tag{7.3}$$

The intrinsic GMRF prior model is parameterised by the precision matrix Q , where $Q = \kappa R$. R is a structure matrix, obtained from a second order random walk in two climate dimensions (see Section 6.2.3) and κ is a scaling parameter which governs smoothness. The structure of R at the interior in the two dimensional setting is described in Figure 6.7 with the appropriate corrections at the boundary available in Kneib (2006).

Models in three climate dimensions, involving the AET/PET variable in addition to the climate variables included in the two dimensional setting above, are complicated by the huge increase in the number of latent parameters introduced by the discretization of climate space. The discretization of space to a regular grid of dimension $50 \times 50 \times 50$ in three climate dimensions results in 125,000 latent parameters - added to the reduced sparsity of the structure matrix in three dimensions and the cost of manipulating matrices of this size, the computational cost of this approach is prohibitive. In Section 6.2.3, we detailed how the computational cost may be mitigated by recognizing that many of the latent parameters in the three dimensional space are *not of interest*. A region of interest is defined (the region of interest for two dimensional climate space is presented in Figure 6.9) and a buffer region is then incorporated around the region of interest in order to negate edge effects. The resulting precision matrix, $Q^* = \kappa R^*$ where R^* is the structure matrix of a second order random walk in three spatial dimensions, obtained by convolution methods (see Section 6.2.3), will have the correct neighbourhood structure at the interior, and approximations to the correct precision structure at the boundary.

X_i in three dimensions is once more modeled as a GMRF:

$$\pi(X_i) \sim \text{GMRF}(Q^*) \tag{7.4}$$

The structure of R^* at the interior is described in Equation 6.12. In this chapter, a buffer region of length three nodes in each spatial direction is utilised. Using this approach, parts of space in which we not interested are ‘cut out’, reducing inference tasks considerably; this results in a substantial decrease in the number of latent parameters requiring inference, from 125,000 to around 40,000, resulting in vast computational savings of several orders of magnitude at the forward stage.

7.2.2 Modeling the Response

As a full 63.15% of the counts in the observational dataset are zero, only zero-modified likelihood families (Section 3.5.2) for the observed data are considered. As per Salter-Townshend (2009), we use spatial zero-inflated models where the probability of presence for a given pollen count is linked to a function of the underlying latent response surface. Additional heterogeneity in the count observations is modeled through the incorporation of random effect terms (Section 3.5.3) into the model - a random effect is included for each observation.

The three zero-inflated models for the response considered in this chapter are:

Zero-inflated Negative Binomial model

$$\pi(y_{i,j}|x_{i,j}, \alpha_i, \delta_i) = \begin{cases} (1 - q_{i,j}) + q_{i,j}\text{NegBin}(0; p_{i,j}, \delta_i) & y_{i,j} = 0 \\ q_{i,j}\text{NegBin}(y_{i,j}; p_{i,j}, \delta_i) & y_{i,j} > 0 \end{cases} \quad (7.5)$$

where $p_{i,j} = \frac{\delta_i}{\delta_i + e^{x_{i,j}}}$ and $q_{i,j} = \left(\frac{e^{x_{i,j}}}{1 + e^{x_{i,j}}}\right)^{\alpha_i}$; this is the by-taxon model presented in Salter-Townshend (2009). There are three hyperparameters governing the zero-inflated Negative Binomial model for each taxon; a precision parameter κ_i , a zero-inflation parameter α_i and an overdispersion parameter δ_i , which models overdispersion of the count observations. The Negative Binomial model is essentially a Poisson model with Gamma distributed overdispersion, i.e. $\lambda_{i,j} \sim \gamma(\delta_i, (1 - p_{i,j})/p_{i,j})$ (see Section 3.5.3).

Zero-inflated Gaussian Overdispersed Poisson model

$$\pi(y_{i,j}|x_{i,j}, \alpha_i, \sigma_i^2) = \begin{cases} (1 - q_{i,j}) + q_{i,j}\text{Poisson}(0; \lambda_{i,j}) & y_{i,j} = 0 \\ q_{i,j}\text{Poisson}(y_{i,j}; \lambda_{i,j}) & y_{i,j} > 0 \end{cases} \quad (7.6)$$

where $\lambda_{i,j} = e^{x_{i,j} + u_{i,j}}$ and $q_{i,j} = \left(\frac{e^{x_{i,j} + u_{i,j}}}{1 + e^{x_{i,j} + u_{i,j}}}\right)^{\alpha_i}$. Overdispersion of the pollen counts is modeled via Gaussian random effect terms, i.e. $u_{i,j} \sim N(0, \sigma_i^2)$. As the Gaussian random effect terms are not conjugate to the Poisson likelihood, each random effect must be inferred at the forward stage, adding to the inference burden. There are three hyperparameters governing the zero-inflated overdispersed Poisson model for each taxon; a precision parameter κ_i , a zero-inflation parameter α_i and an overdispersion parameter σ_i^2 , which models the variance of the random effect terms.

Zero/N-inflated Gaussian Overdispersed Binomial model

$$\pi(y_{i,j} | \dots) = \begin{cases} (1 - q_{i,j})r_{i,j} + q_{i,j}r_{i,j}\text{Binomial}(0, N_{i,j}, p_{i,j}) & y_{i,j} = 0 \\ q_{i,j}(1 - r_{i,j}) + q_{i,j}r_{i,j}\text{Binomial}(N_{i,j}, N_{i,j}, p_{i,j}) & y_{i,j} = N_{i,j} \\ q_{i,j}r_{i,j}\text{Binomial}(y_{i,j}, N_{i,j}, p_{i,j}) & 0 < y_{i,j} < N_{i,j} \end{cases} \quad (7.7)$$

where $p_{i,j} = \frac{e^{x_{i,j}+u_{i,j}}}{1+e^{x_{i,j}+u_{i,j}}}$, $q_{i,j} = \left(\frac{e^{x_{i,j}+u_{i,j}}}{1+e^{x_{i,j}+u_{i,j}}}\right)^{\alpha_{1i}}$ and $r_{i,j} = \left(\frac{1}{1+e^{x_{i,j}+u_{i,j}}}\right)^{\alpha_{2i}}$. $N_{i,j}$ is the total count at location j for group i - through the use of nesting structures, the data are grouped into sets of taxonomically related groups; for the nested compositional model presented in Section 7.3.2, there are 27 such groups. There are four hyperparameters governing the zero/N-inflated Binomial model for each group; a precision parameter κ_i , two zero-inflation parameters (α_{1i}, α_{2i}) and an overdispersion parameter σ_i^2 , which models the variance of the random effect terms. If α_{2i} is set equal to zero, this model simplifies to the statistically inconsistent standard zero-inflated Binomial model (see Section 5.5.1).

The first two models do not take into account the compositional nature of the data and are referred to as *marginal models* in the following. Additionally, as the zero/N-inflated Binomial model is based upon nesting structures for the grouping of pollen data, it is referred to as the *nested model*. Each of the models considered are overdispersed with regard to the spatial response. For simplicity in discussion in the following we drop the ‘‘Gaussian overdispersed’’ term, from herein the zero-inflated Gaussian overdispersed Poisson model will be referred to as the zero-inflated Poisson model. Similarly the zero-inflated Gaussian overdispersed zero/N-inflated Binomial model will be referred to as the zero/N-inflated Binomial model

7.2.3 Inference

Due to the assumption of decomposable joint models, the forward models for each taxon are fit independently of the rest. As the number of hyperparameters is relatively low (at most 4 for the nested model), the INLA algorithm of Rue et al. (2009) (Section 3.3) can be used for forward stage inference. This provides a quick approximate method for obtaining closed form posteriors for the latent X_i . For an assessment of the algorithm in the context of application to the palaeoclimate reconstruction problem see Salter-Townshend (2009).

Treatment of Hyperparameters

Salter-Townshend (2009), in the context of the palaeoclimate reconstruction problem, illustrated that approximations to $\pi(\theta_i | Y_i)$ by point masses at the posterior modal values of θ_i , resulted in very little loss of information at the inverse stage - this is the empirical Bayes (Section 3.2.3) approximation to fully Bayesian inference. Intuitively, the quality of this ap-

proximation is due to the large numbers of observations for model training purposes - there are 7742 count observations for each of the 28 plant taxa at the forward stage.

An implicit constraint on inference procedures for models in three climate dimensions is the issue of storage - if each posterior $\pi(\theta_i|Y_i)$ is represented on a coarse grid of length 50, the storage cost of $\pi(X_i, \theta_i|Y)$ for just one plant taxon is of the order of 350 megabytes. The total cost of storing the equivalent results for all 28 pollen taxa is around 10GB, potentially monopolizing computer resources. Due to this storage constraint and the quality of the empirical Bayes approximation (Salter-Townshend 2009), inference procedures for θ_i in the following are empirical Bayes based.

7.2.4 Results

The hardware used is a dedicated Linux cluster with 12 $3.33GHz$ processors and 96GB of RAM; the multi-core nature of the machine allows multiple models, 12 at a time, to be fit in parallel. The R-INLA package of Rue et al. (2009) is used for parameter inference; through the use of numerical algorithms for inference on the (low dimensional) model hyperparameters and the harnessing of algorithms for fast operations on sparse matrices, the software facilitates quick, computationally efficient approximate inference on unknown model parameters. In Table 7.1 the average time (in seconds) taken to fit each model for a single plant taxon incorporating either two or three climate covariates is presented.

Model	Hyperparameters	2D	3D
Zero-inflated Negative Binomial	3	67	27,386
Zero-inflated Poisson	3	105	41,322
Zero/N-inflated Binomial	4	131	116,595

Table 7.1: Average time taken (in seconds) for empirical Bayes based inference for each taxon model in two and three climate dimensions.

In Table 7.1 we observe that the zero-inflated Negative Binomial model has the shortest fitting time for each set of climate covariates. This is due to the “integration out” of the random effect terms, included to model overdispersion of the counts data, which must be inferred for both the zero-inflated Poisson model and the zero/N-inflated Binomial model (see Section 7.2.2). Due to the almost twenty-fold increase in the number of latent parameters introduced by the incorporation of an additional climate covariate, and the reduced sparseness of precision matrices in three dimensions, inference for the 3D forward models takes substantially longer than the 2D setting - for example, empirical Bayes based inference on the nested forward model for a single plant taxon takes around a day and a half.

The calibrated models reveal a number of features of the training dataset:

1. Strong prior distributions had to be specified for the precision parameter of the latent

response surfaces to ensure smoothness of the posterior response surfaces. Exploratory analysis, including the refitting of forward models for several values for the prior parameters revealed that a prior distribution of $\pi(\kappa_i) \sim \Gamma(300, .1)$ produced posterior response surfaces with sufficient smoothness.

2. For all models, including the (Gamma overdispersed) Negative Binomial model, the overdispersion parameters are significantly non-zero. This illustrates that there is extensive variability in the pollen counts over and above that expected by the zero-inflated models, even with the explicit modelling of the excess zeroes.
3. The use of zero/N-inflated Binomial models as compared to zero-inflated Binomial models results in a significant decrease in the inferred overdispersion parameters. This is due to explicit modeling of the N-inflation present in the data.

Model validation in the forward sense did not play a major role in Salter-Townshend (2009) or Haslett et al. (2006). Model criticism was instead confined to the evaluation of the predictive properties of calibrated models in the inverse sense - measures of prediction accuracy, such as the *MSEP*, were used for model comparison. Little or no attempt was made to detect observations which did not well fit the calibrated models at the forward stage.

In contrast, one of the primary contributions to the palaeoclimate reconstruction project presented in this thesis is the development of richer models which facilitate quick approximate methods for the evaluation of the performance of the fitted models. This method is based on analysis of posterior random effect terms which are included in models to account for overdispersion of the pollen counts. Using the methods developed for residual analysis and outlier detection in Chapter 4, we illustrate how the posterior random effect terms, in the context of Gaussian overdispersion, can be quickly evaluated to learn about the underlying data mechanisms and the suitability of *a priori* model assumptions. Through the harnessing of Gaussian residual theory, outliers among the very many count observations can be automatically identified - for the first time, explicit outlier detection is possible for the pollen counts of individual taxa at the forward stage. Of course it must be stated that this approach is only rendered feasible by the computational speed of the INLA algorithm.

7.2.5 Residual Analysis and Outlier Detection

For the sake of brevity in the following, we constrain our discussion to the results obtained by the zero/N-inflated Binomial model in three climate dimensions. As previously discussed, Gaussian random effect terms, one for each observation, are incorporated into the forward models for each taxa to account for possible overdispersion of the pollen counts. The use of this model in the context of the palaeoclimate problem requires the specification of an explicit nesting structure for the forward problem. The nesting structure relevant to the following is presented in Section 7.3.2.

We begin with an evaluation of posterior model properties, for a number of plant taxa, through the visual analysis of quantile-quantile plots of the mean posterior random effect terms. As detailed in Section 4.3.3, the expression of the posterior random effect terms in Gaussian form provides for a quick, visual method of establishing whether *a priori* distributional assumptions are appropriate - the observation of trends in the posterior random effect terms other than Gaussian may indicate possible model misspecification at the forward stage.

Quantile-Quantile Plots

In Figure 7.1, the quantile-quantile plots of the mean posterior random effect terms are presented for an arbitrary six (for sake of brevity) of the (27) available plant taxa. The random effects are assigned an *a priori* Gaussian distribution; as per Section 4.3.3, if the *a priori* distribution is appropriate, the posterior random effect terms should also display distributional behaviour that is approximately Gaussian in nature.

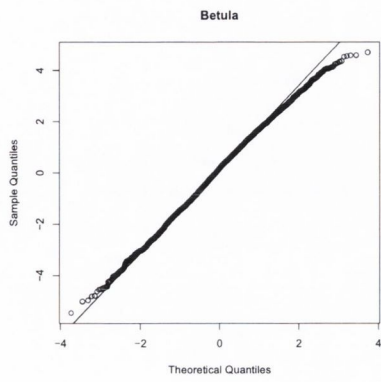
After controlling for the excess zeroes, the *a priori* Gaussian distribution for the random effect terms appear appropriate as determined by visually analysing the quantile-quantile plots. This is a significant modelling achievement considering that each set of pollen observations consists of very many Binomial counts which exhibit signs of both zero and N-inflation. However, the posterior random effects for each taxon do exhibit some signs of tail behaviour greater than that expected by Gaussian theory.

This is confirmed upon viewing the sample density of the mean posterior random effects, corresponding to each of the plant taxa in Figure 7.1, which are presented in Figure 7.2. We observe tail behaviour in all the sample density plots, perhaps indicating the presence of an unmodelled explanatory variable as per Section 4.2.4. For 4 of the 6 taxa there is a noticeable right skew towards the positive axis, indicating that several of the posterior random effects are larger than expected. Overall, the *a priori* Gaussian distribution for the overdispersion appears reasonable.

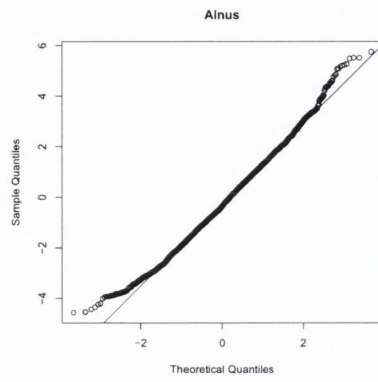
Of course an important feature of the methodology for residual analysis presented in Chapter 4 was the possibility of explicit outlier detection. For the sake of brevity in the following, we focus our attention on outlier detection for one of the plant taxa - the *Cedrus* taxon, though the analysis can be readily applied to any of the other taxa. Our primary aim is to determine whether a recurring trend can be established in the detected outliers.

Case Study: Outlier Detection for *Cedrus*

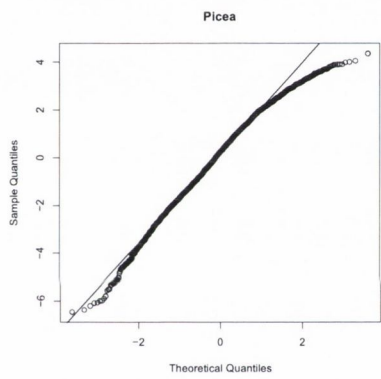
In this section, the power of the developed methodology for outlier detection in Chapter 4 is illustrated by application to the zero/N-inflated counts of the *Cedrus* taxon. Specifically, Gaussian residual theory is harnessed in order to provide explicit critical bounds by which outliers may be automatically detected amongst the posterior random effect terms. The counts



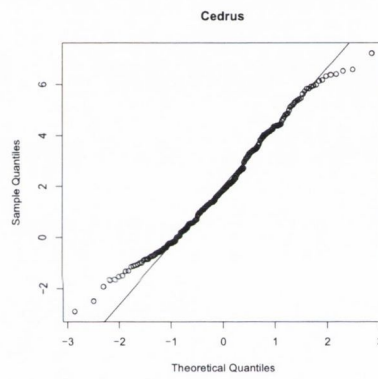
(a)



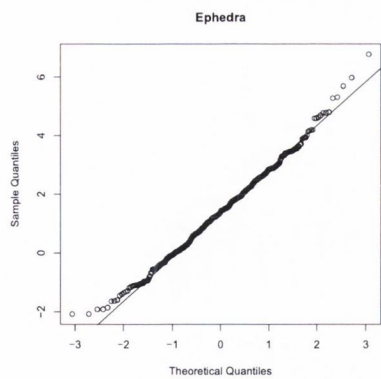
(b)



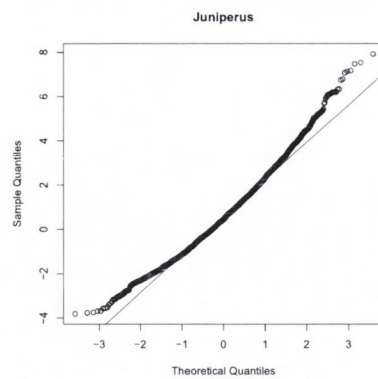
(c)



(d)



(e)



(f)

Figure 7.1: Quantile-quantile plots of the mean posterior random effects for a number of plant taxa.

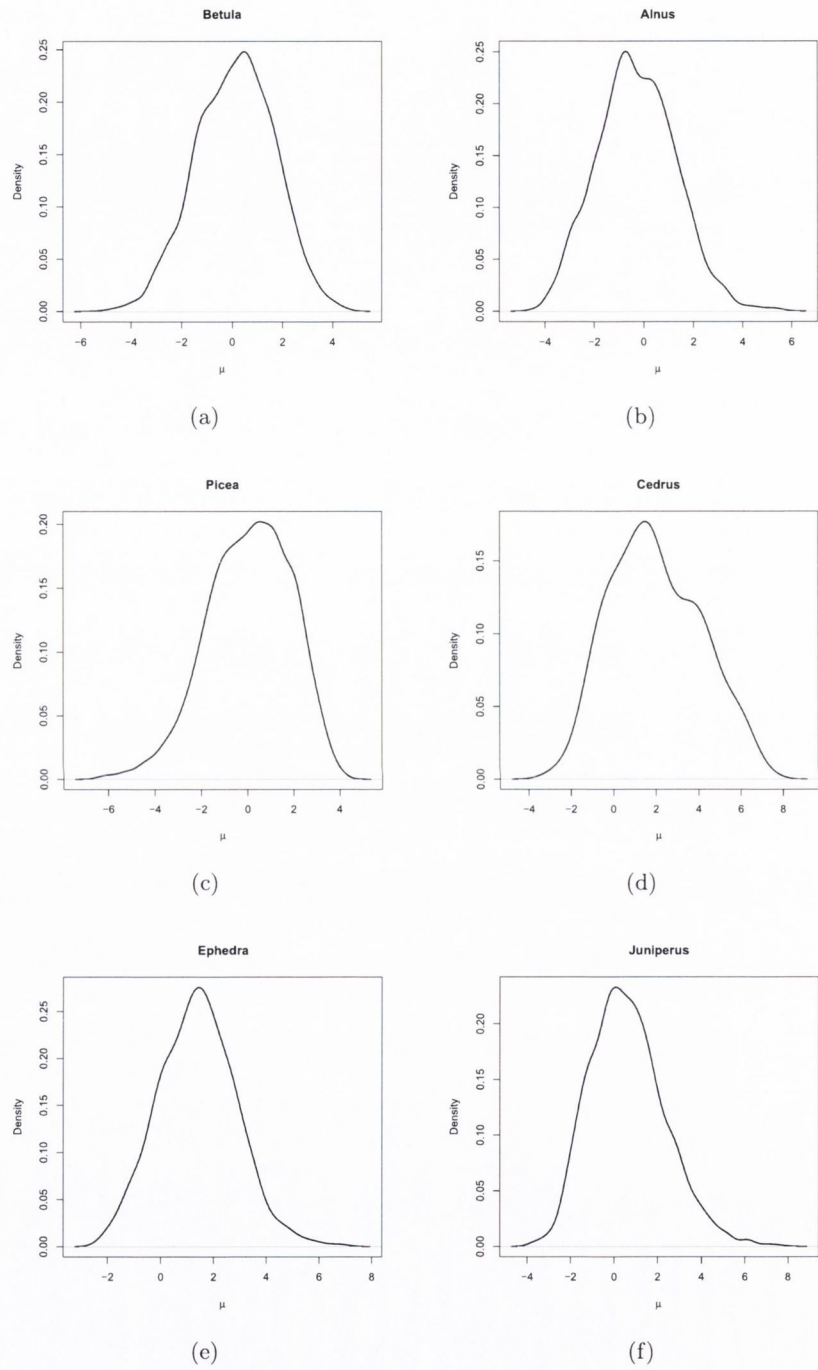


Figure 7.2: Comparison of the sample densities of the mean posterior random effects for a number of the plant taxa.

corresponding to the identified random effects may subsequently be investigated to evaluate the veracity of the claim of outlying behaviour.

Using the methods of Section 4.2.4, over 4.2% of the *Cedrus* dataset are identified as potentially outlying - this is less than typically expected, given the analysis conducted in Section 4.3 (here $\sigma_{Cedrus}^2 = 6.67$), as there are very many zero counts recorded for the *Cedrus* taxon. This suggests that the Gaussian overdispersion parameter is being overestimated to compensate for the slightly skew behavior observed in Figure 7.2 (d). In Figure 7.3 we plot the 95% HPD regions for a subset of the posterior random effect terms as well as 95% critical bounds obtained from standard Gaussian residual theory. As per Section 4.2.4, HPD regions which lie significantly outside the critical bounds are identified as possible outliers.

An important point to note is that the posterior random effects corresponding to all detected outliers are strictly positive in nature; this indicates that the counts at all identified sites are significantly larger than expected given the respective site locations. In the following we plot the detected outliers on a world map for illustrative purposes, and discuss features of the detected outliers.

Cedrus, or more commonly known as Cedar, is a genus of coniferous trees that are native to mountainous regions of the Mediterranean and the western Himalayas. In terms of favoured environmental conditions, the species fares best at altitudes of between 1500-3200 in the Himalayas or 1000-2000m in the Mediterranean, expressing a preference for temperate regions and dry, semi-arid sites.

Given this knowledge, we can conclude that the largest outlier, detected at a site in Northern Finland, is spurious in nature as the site lies outside the geographical range of the *Cedrus* species. As the *Cedrus* genus is a relative of the pine family, we speculate that pollen corresponding to a pine species native to the region may have been mistaken for *Cedrus* pollen. We observe a similar result at a site in the Alps in France; in Figure 7.13 (a), the *Cedrus* taxon shares a nest level with *Olea*, or Olive, which typically dominates the assemblage in this region. However for this site a low *Cedrus* count is the only observation for this nest, whereas pollen data corresponding to other pine species is quite plentiful. Furthermore, in the surrounding sites, little trace of *Cedrus* pollen is observed. This perhaps indicates misidentification of the pollen once more.

The next largest outliers correspond to three sites in the mountains of Kashmir region of India. Each of the *Cedrus* counts are low at these sites, being counts of 1, 4 and 10 respectively. Each of the sites are at altitudes of between 3680 - 5100 metres above sea level, which may be considered too extreme for the *Cedrus* species to flourish. The MTCO values recorded at each site are -12.9°C , -16°C and -16.8°C respectively, with these temperatures at the outer limit at which the *Cedrus* species can survive. At each site there are additional trace amounts of pollen corresponding to other tree species such as *Alnus* and *Abies* with the hardy shrub (grassy) species, *Artemisia* and *Chenopodiaceae* dominating the pollen assemblage. The

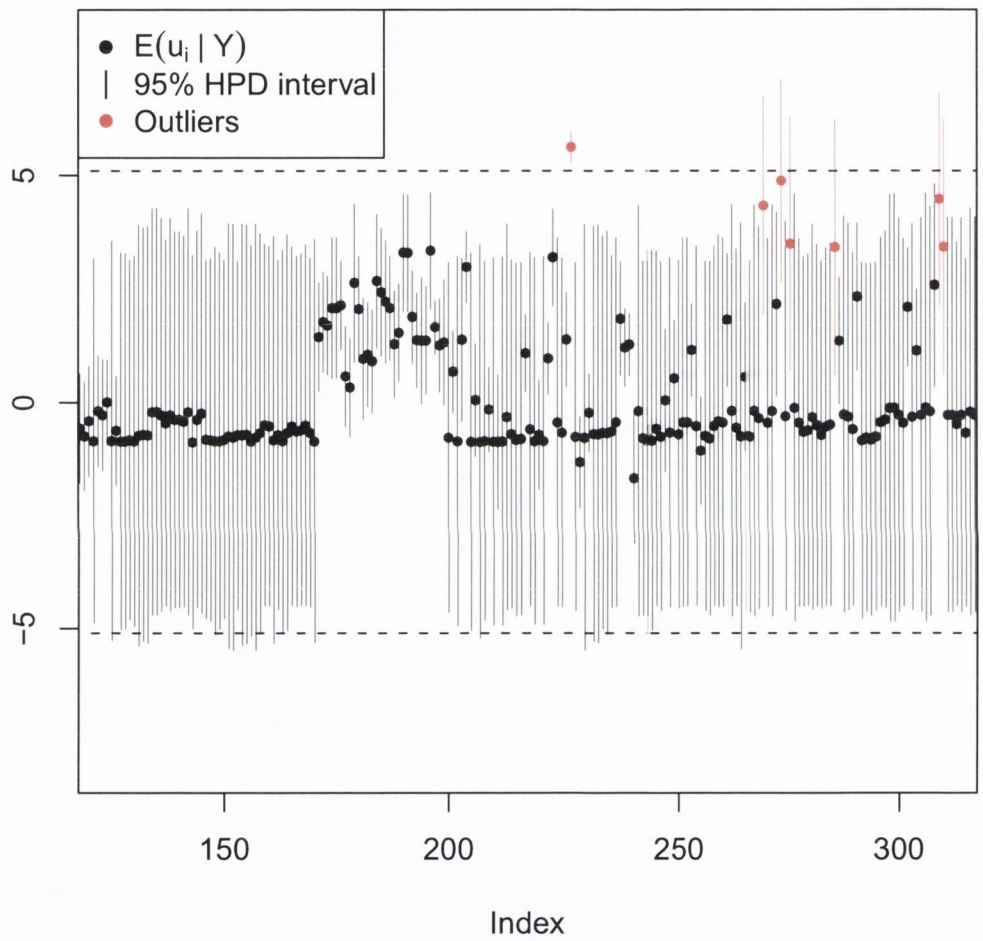


Figure 7.3: Mean posterior random effects and 95% HPD intervals corresponding to a subset of the Cedrus counts. Posterior random effects which significantly cross the dashed lines (critical bounds) are considered to be outliers.

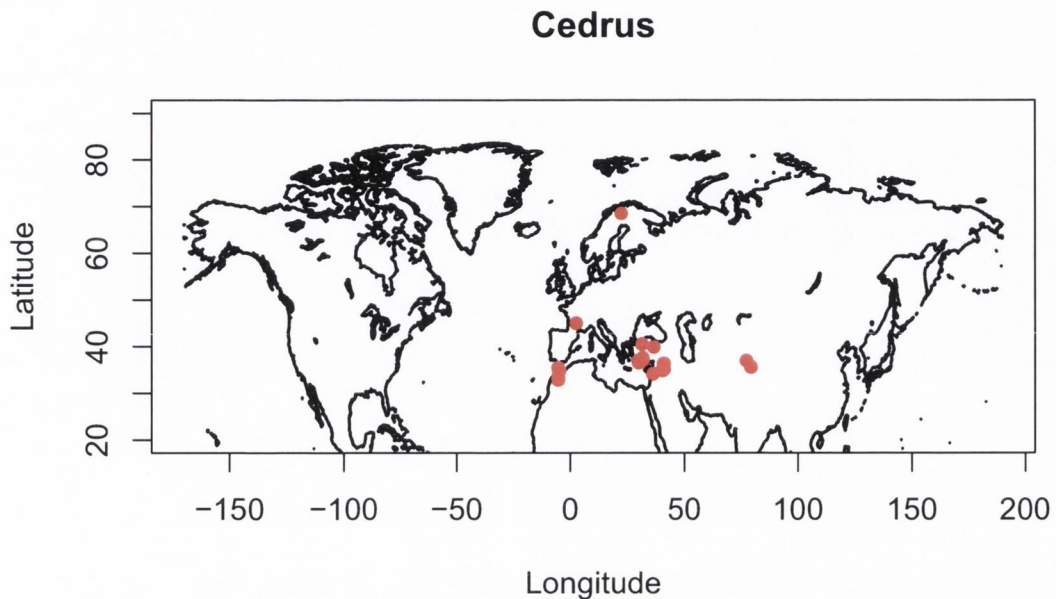


Figure 7.4: Map of the northern hemisphere presenting the site locations of the detected outliers. 4.24% of the *Cedrus* dataset is detected as potentially outlying.

explanation for the trace amounts of tree pollen at this extreme site may be due to features of the locality - pollen spores may be transferred from lower altitudes on the prevailing wind to each of the site locations. An alternative explanation is that the climate measurements are simply misrecorded. A firm conclusion cannot be made without obtaining expert advice.

No specific trend can be determined for the remaining detected outliers save to mention that the *Cedrus* counts appear large at each of the relevant sites, as indicated by the strictly positive posterior random effect terms. However, one clue is that these sites are located in regions of North Africa and the Mid-East which have very hot summer temperatures, to which a local *Cedrus* species, *Cedrus Atlantica*, is perhaps better adapted than respective *Olea* species. This result can be observed by plotting the sample density of the recorded MTWA's at which *Cedrus* pollen is observed versus the sample density of the MTWA's corresponding to the outlying observations - Figure 7.5 illustrates that outlying behaviour and extreme count observations, as determined by the analysis of the posterior random effects, appears correlated with MTWA. However, definite conclusions or explanations for this result once more require the provision of expert opinion.

The analysis in the above represents a substantial advance in comparison with what has been done previously. For example, Salter-Townshend (2009) attempted to identify outliers in the RS10 pollen dataset through the analysis of the pollen composition at sites for which large

values of the *RMSEP* were recorded. However, this required the analysis of 28 pollen counts jointly and it was difficult to determine the cause of the outlying behaviour. Furthermore, as the distribution of the *RMSEPs* was unknown, it was impossible to provide critical cut-off bounds or points; Salter-Townshend (2009) identified outliers in a subjective manner. Here, we avoid these obstacles via the analysis of posterior random effect terms which enable the systematic identification of outlying observations for each of the separate pollen taxa.

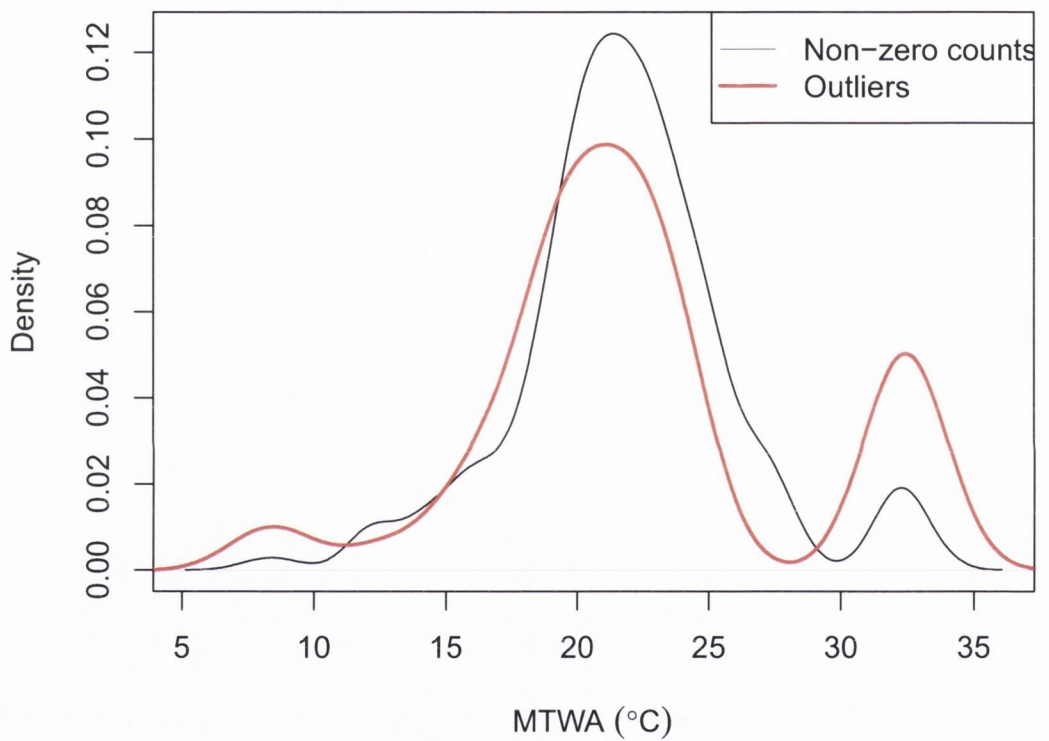


Figure 7.5: Sample density of the MTWA's across all locations at which *Cedrus* pollen is observed as compared to the sample density of the MTWA's recorded for the outlying observations. We observe more extreme MTWA measurements, on average, for the outlying observations

7.3 Model Inversion

The primary interest in this thesis is not so much in the forward models, but in the inverse use of the fitted models for prediction. In the following we focus on the inverse stage of the reconstruction problem and evaluate the predictive performance of the various models introduced in Section 7.2.2. In the following we detail our advances over existing methods.

7.3.1 Inverse Inference

At the inverse stage, forward models are inverted to make inferences on the unknown fossil climate corresponding to new sets of pollen counts. As climate space is discretized to a regular grid, normalised posteriors can be obtained by evaluating the posterior predictive mass at each discrete location on the grid and dividing through by the sum.

In Section 7.3.2 and Section 7.3.3, quadrature is used to numerically evaluate the inverse predictive densities in order to obtain exact 95% HPD regions for evaluating model accuracy - as fully decomposable models are assumed (see Equation 7.2), inversion of the model can be performed separately for each taxon. The joint posterior is then found as the normalised product of the independent inverse predictive densities for each taxa separately. As per Salter-Townshend (2009), a buffer region is employed in both 2 and 3 dimensional space to omit infeasible climate locations, and reduce the number of locations at which the posterior must be evaluated.

In Section 7.4, predictions for the palaeoclimate corresponding to fossil pollen at a site in Glendalough in Ireland are produced using the sampling based methods established in Chapter 6. We illustrate how the developed algorithm substantially reduces the computation time in obtaining samples from the posterior, especially with regard to the zero/N inflated Binomial model, as compared to numerical integration based methods.

Assessing Model Predictive Performance

The most challenging model comparison statistic, in the context of the palaeoclimate reconstruction problem, is the percentage of observations, Δ , which lie outside their corresponding leave-one-out inverse predictive distributions 95% highest posterior density (HPD) region. In order to obtain this statistic, models must be refit given the data minus each set of left out counts; the resulting inverse predictive distribution for climate is then analysed to determine if the true climate location lies within the 95% HPD region.

However, this is a *computationally intensive* task - on a standard computer with a 3.4HGz processor and 4GB of RAM, the time taken to refit the model is 8 minutes. In order to obtain the full leave-one-out cross validation statistic, this process must be repeated 7742 times, requiring the order of weeks for its computation. As a compromise, a saturated cross

validation metric is used instead; in Section 7.3.2 we illustrate that, due to the large numbers of observations available for model fitting, the differences between the inverse leave-one-out-predictive densities and saturated predictive densities are essentially negligible.

Other model comparison measures considered include the root mean squared error of prediction (*RMSEP*) and D_{mode} , a measure of distance between the true climate location and the location corresponding to the mode of the saturated inverse predictive distribution (see Section 3.7.2).

7.3.2 Results: 2D Climate Application

In this section we explore the results obtained using the 2-dimensional climate models, which were introduced in Section 7.2 previously. We begin by first discussing the performance of saturated cross validation measures for model validation in the context of the palaeoclimate problem.

Saturated Cross-Validation

An approximation to the full leave-one-out cross validation is provided by the use of saturated cross validation methods. Essentially, models are evaluated using the pool of data from which the models were fit - in the presence of large amounts of data, the saturated and leave-one-out cross validation measures should approximately agree; the omission of a single set of observations will have little effect on the inverse predictive densities produced. The benefit of this approach is that the models only need to be fit once, the calibrated models can then be used to evaluate model performance.

In Figure 7.6, we illustrate that the inverse predictive densities produced by the saturated model provide an exact approximation to the leave-one-out inverse predictive densities with essentially negligible error. For a random subset of 500 of the 7742 model training counts, the predictive performance statistics for the leave-one-out and saturated cross validations are compared for the zero-inflated Poisson model. We observe approximately zero error in the comparison of cross validation methods, justifying the use of the approximation in this setting; model validation is thus reduced from the order of weeks to the order of minutes.

Marginal Models

The two marginal models, introduced in Section 7.2.2 above, are fit to the pollen dataset. In using the marginal models, the pollen observations for each taxon are considered conditionally independent given the latent response surfaces, and the latent response surfaces (for each taxon) conditionally independent given climate location; as per Salter-Townshend (2009), model parameters for each taxon may thus be independently inferred. If this decomposition of

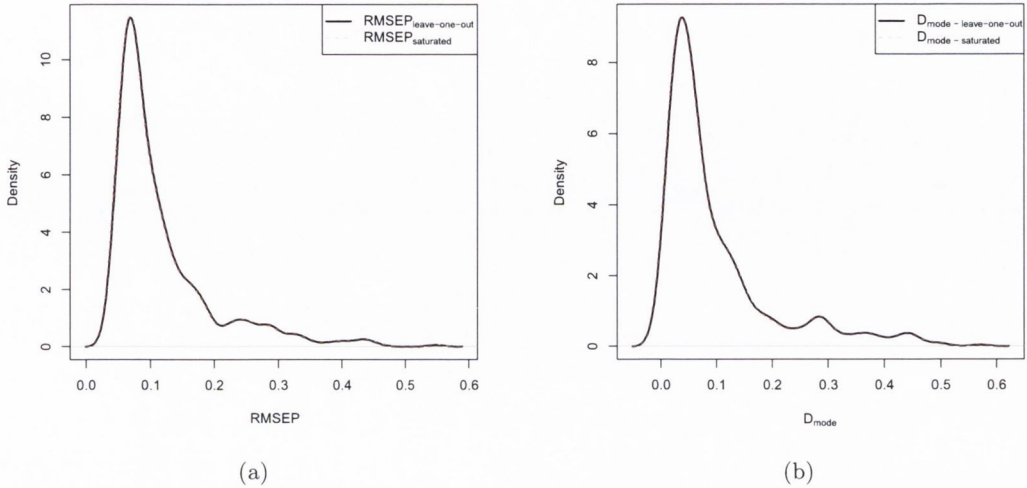


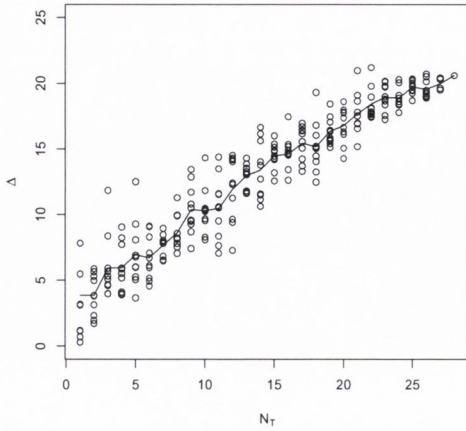
Figure 7.6: (a) Comparison of the (a) $RMSEP$ and (b) D_{mode} for a saturated versus leave-one-out cross validation. The error in the approximation is negligible due to large numbers of observations available for model training.

the joint model into independent components adequately reflects the underlying data generating process, the Δ statistic, namely the number of observations lying outside the saturated inverse cross validation predictive densities 95% HPD region, should be approximately 5% for any given set of taxa.

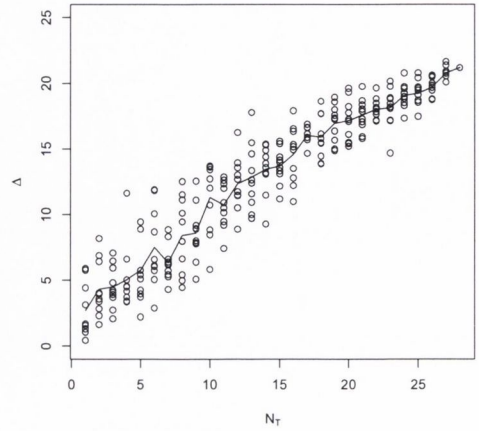
In Figure 7.7 we observe that this is not the case; for increasing number of taxa, the predictive accuracy of the approach is seen to deteriorate almost (approximately) linearly. As per Salter-Townshend (2009), this figure is obtained as follows; considering a single taxon, there are $\binom{28}{1}$ choices of one taxon, for two taxa there are $\binom{28}{2}$ combinations and so on. We take a random ten of these $\binom{28}{N_T}$ combinations for each of $N_T = (1, \dots, 27)$ and plot the mean of the Δ statistic obtained. The value of Δ for all 28 taxa is 20.59% and 21.19% for the zero-inflated Negative Binomial model and the zero-inflated Poisson model respectively.

Whilst model performance in terms of the number of observations lying within their respective cross-validation 95% HPD regions is seen to deteriorate with each additional taxon considered, in Figure 7.9 we observe that the predictive power in terms of the $RMSEP$ and D_{mode} conversely improves. These metrics provide a measure of evaluating the placement of climate posteriors with regard to the true known location and indicate that the inverse predictive posteriors on climate becomes increasingly accurate, in terms of location, with each additional taxa considered. As per Section 5.2.2, this indicates that the cross-validation predictive densities produced at the inverse stage are not sufficiently conservative.

As we observe in Table 7.2 and Figure 7.8, the predictive performance of the zero-inflated

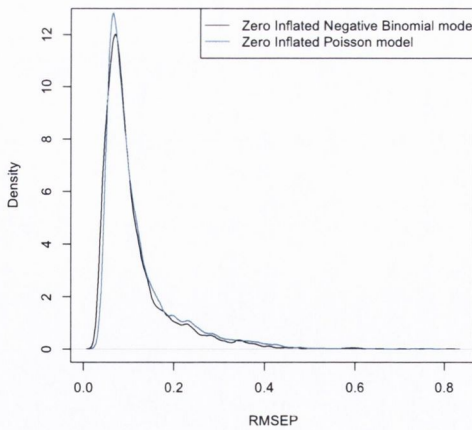


(a) Zero-inflated Negative Binomial Model

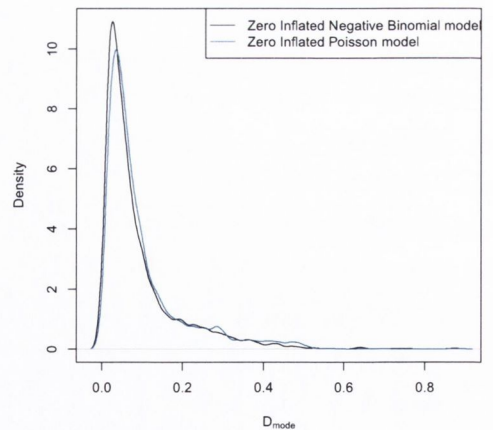


(b) Zero-inflated Poisson Model

Figure 7.7: Plot of Δ , the percentage of observations lying outside the corresponding saturated inverse cross-validation 95% HPD region for both the (a) zero-inflated Negative Binomial and the (b) zero-inflated Poisson model. The predictive performance of both models is seen to disimprove linearly with each additional taxa considered. The value of Δ for all 28 taxa is 20.59% and 21.19% for each respective approach, reflecting that the zero-inflated Negative Binomial model provides a slightly better fit to the data. Both models represent a poor fit to the data, as reflected by the Δ statistic for each which is substantially greater than 5%



(a)



(b)

Figure 7.8: (a) Model evaluation statistics comparing the performance of the Gamma overdispersed zero-inflated Negative Binomial model with that of the zero-inflated Gaussian overdispersed Poisson model.

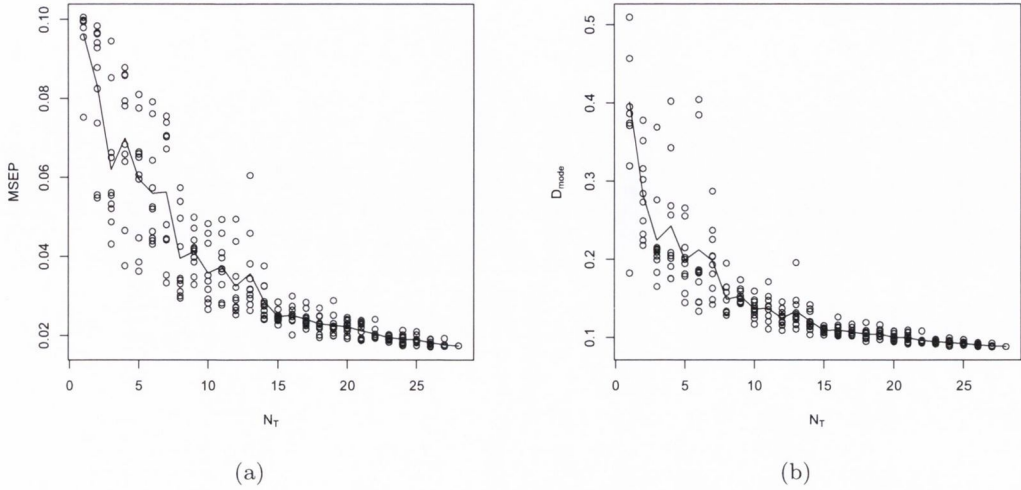


Figure 7.9: The predictive power statistics for the zero-inflated Negative Binomial model.

Negative Binomial model is slightly superior to that of the zero-inflated Poisson equivalent. In order to provide a possible explanation for this superiority in predictive performance we analyse the posterior random effect terms produced by the zero-inflated Poisson model. In Figure 7.10 we plot quantile-quantile plots of the mean posterior random effects of a number of the plant taxa. The plots indicate that the random effect terms display behaviour which is more common to that of the Gamma distribution (see Section 4.3.3), perhaps providing a reason for the slightly superior fit of the zero-inflated Negative Binomial model.

Model	Δ	\overline{RMSEP}	\overline{D}_{mode}
Zero-inflated Negative Binomial	20.59%	0.109	.088
Zero-inflated Poisson	21.19%	0.116	.097

Table 7.2: Comparison of predictive performance of the marginal models.

Given marginal models for the data, the plant taxa are assumed to be conditionally independent given climate at both the forward and inverse stages. However, the above analysis indicates that this assumption is erroneous, or at the very least that the plant taxa are not conditionally independent given *two dimensions* of climate. This is perhaps an unsurprising result - in Chapter 2 we detailed how two climate variables may not be sufficient to accurately model the pollen response. Additionally, as the data collection process is compositional in nature - pollen spores are counted until a predefined total is reached, approaches which consider each taxon response surface separately will be subject to erroneous inferences at the inverse stage.

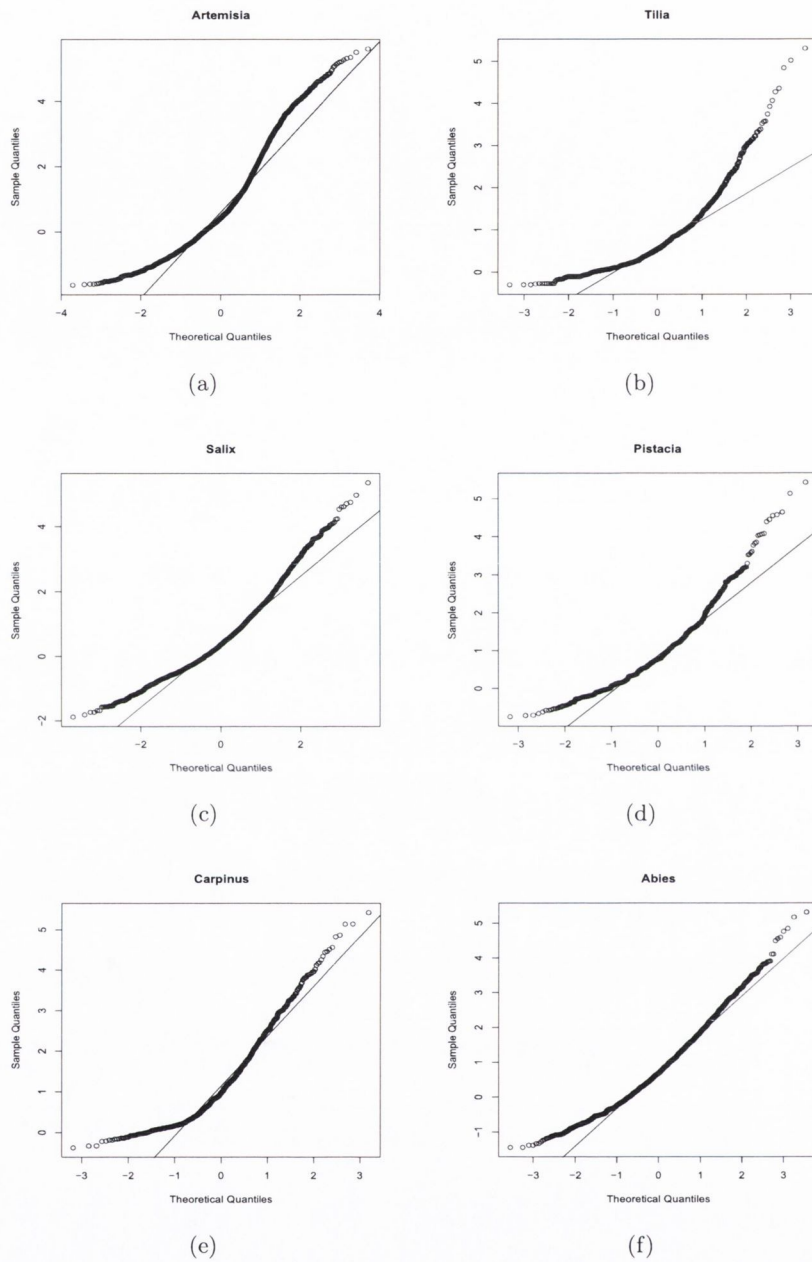


Figure 7.10: Quantile-quantile plots of the mean posterior random effects produced by the zero-inflated (Gaussian overdispersed) Poisson model for a number of plant taxa. The posterior random effect terms appear to exhibit behaviour that is Gamma distributed in nature (see Figure 4.9) despite the *a priori* specification of a Gaussian distribution for the random effects terms.

If the deterioration in performance is due to the compositional nature of the dataset, the use of hierarchical or nesting structures may provide a method of addressing this problem.

Nested Model

Salter-Townshend (2009) proposed to account for the compositional nature of the training dataset through the use of nesting structures (see Section 5.4). The plant taxa were grouped into smaller subsets of similar plants species - the nesting structure, obtained from expert opinion, is presented in Figure 7.11. As components of the separate nests or groups were considered conditionally independent given each nest total, Salter-Townshend (2009) was able to decompose the Multinomial joint problem into a series of conditionally independent Binomial problems. The nesting of taxa imposes a strict structure on the possible correlation structure between the individual taxa - in nests where there are just 2 taxa, the components of each taxa are fully negatively correlated due to the sum constraint.

However, as discussed in Section 5.5.1, the zero-inflated nature of the training dataset introduces a number of problems at the forward stage - as detailed in Section 5.5.2, models which do not account for N-inflation of the compositional counts data will produce statistically inconsistent results. Erroneous inferences are obtained in an obvious way - the use of a single zero-inflation parameter will not be able to capture all sources of heterogeneity in the observed counts, leading to an overestimation of overdispersion parameters. For the palaeoclimate reconstruction problem this result is observed in Table 7.3.

Furthermore, Salter-Townshend (2009) does not specify a fully nested structure - at several of the lowest nests in Figure 7.11 there are more than two taxa; for these levels, conditional independence of the individual taxa is assumed and zero-inflated Negative Binomial models fit to each taxa separately. However, the data are compositional in nature and the use of models which do not address the sum constraint will lead to erroneous inferences on model parameters and result in poor predictive performance at the inverse stage (see Section 5.3.3). In the context of the palaeoclimate reconstruction problem, in Figure 7.7 we observe that the predictive performance disimproves linearly with each additional taxa that is independently modelled.

In the following section we detail an extension of the partially nested structure of Salter-Townshend (2009) to full nesting of all levels.

Learning the Optimal Nesting Structure

In Section 5.4.3, we detailed how the optimal nesting structure, in the sense of cross-validation prediction accuracy, for a compositional dataset can be learned from the data. Essentially, each set of nesting structures are explored in turn until the “best” set of nested comparisons, here in terms of an inverse cross-validation metric, are identified. However, with regard to

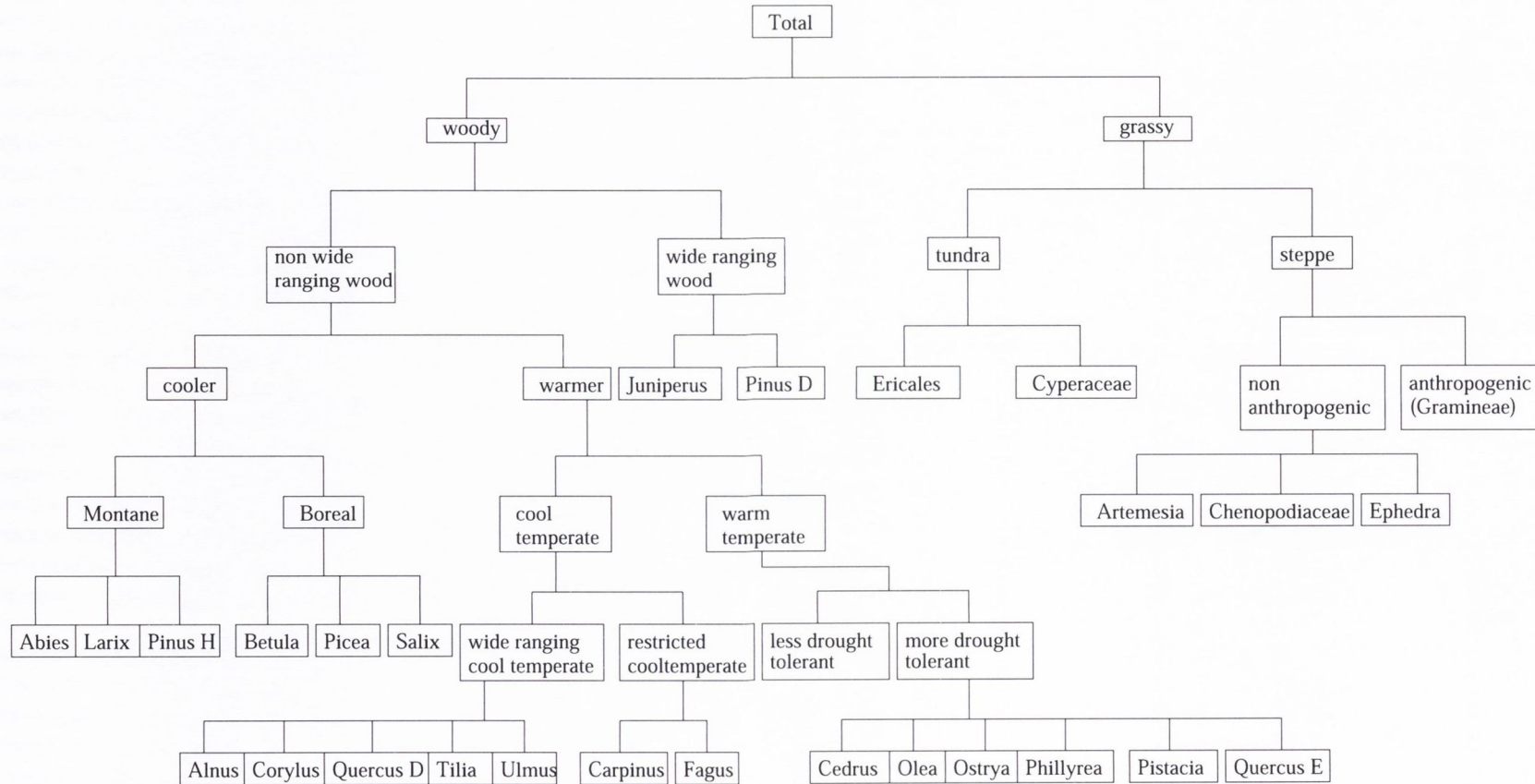


Figure 7.11: The nesting structure for the forward stage, as presented in Salter-Townshend (2009)

the palaeoclimate reconstruction problem, the exploration of suitable nesting structures for the 28 taxon dataset is constrained by the sheer number of taxa; the number of possible taxa reorderings that would need to be explored is of the order of 10^{29} for the simplest continuation ratio model type nesting structures (see Rodríguez (2007)).

As such, we attempt to cut down on the number of structures we must explore by first analysing the performance of the nested model presented in Figure 7.11 at the nest levels which contain just 2 taxa. If the nesting structure in Figure 7.11 is correct, the inverse predictive accuracy of each nest should be approximately 95% (i.e. $\Delta = 5\%$). There are 12 nests in Figure 7.11 which contain only two taxa; as the individual nests are conditionally independent given the sum totals for each nest and climate, each of the nests may be separately modelled using zero/N-inflated Binomial model presented above. Inversion of the model can also be performed separately and the joint posterior is found as the normalised product of the independent inverse predictive densities of each nest. In this case, the saturated cross-validation prediction accuracy is 90.9%, indicating that the partially nested structure is quite accurate in its decomposition. We proceed to investigate nesting structures for the lowest levels.

For the nests labeled “montane”, “boreal” and “non-anthropogenic” in Figure 7.11 there are three taxa at each lowest level - for each of these sets of taxa there are only three possible reordering structures. The three possible orderings are evaluated for each nest level and the optimal structure, in terms of saturated inverse cross-validation predictive accuracy, is presented in Figure 7.12, the inverse prediction accuracy of each structure is 96.12%, 95.08% and 95.67% respectively.

For the nests labeled “more drought tolerant” and “wide ranging cool temperature”, locating the optimal nesting structure is more computationally expensive due to the number of taxa in each nest, 5 and 6 respectively. The “best” nesting structure for each group, once more in terms of saturated cross validation prediction accuracy, are presented in Figure 7.13. The saturated cross-validation inverse predictive accuracy of the nesting structure in Figure 7.13 (a) is 90.65% as compared to 83.23% for the non-nested model. For Figure 7.13 (b) the nesting structure improves the saturated cross-validation predictive accuracy from 87.2% to 95.2%. These nesting structures are carried forward to the following sections.

One initial important comparison to make is between the performance of models which only account for zero-inflation in the Binomial split at each nest and models which account for zero/N-inflation of the counts - the zero-inflated model can be considered a subset of the zero/N-inflated model with the N-inflation parameter fixed to zero. In Table 7.3, we see that model performance, as expected, is superior for the statistically consistent zero/N-inflated model.

As the zero-inflated Binomial model does not account for the N-inflation present in the data, the overdispersion parameter is significantly overestimated. This represents the only

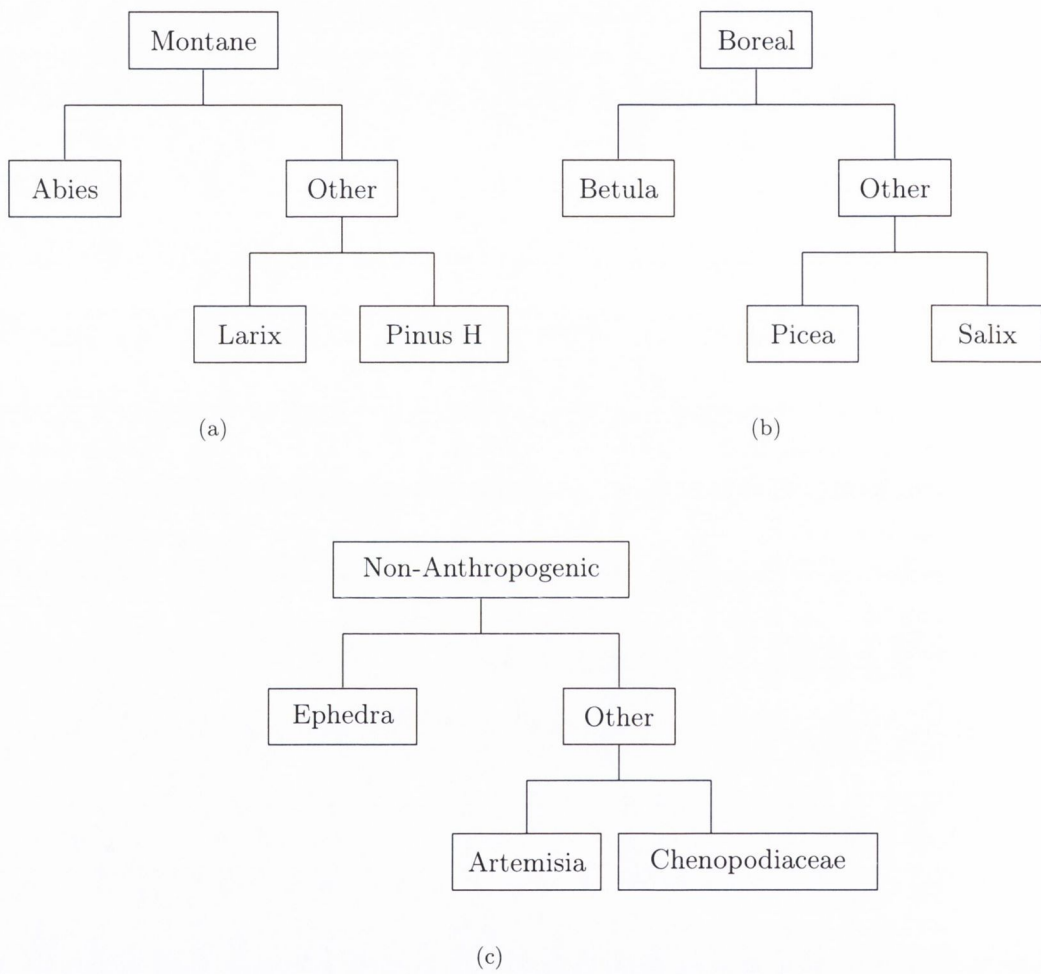
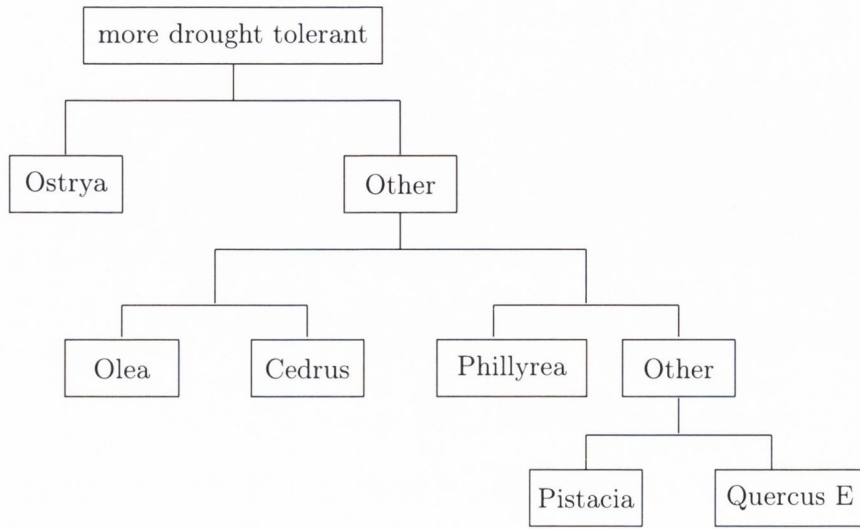
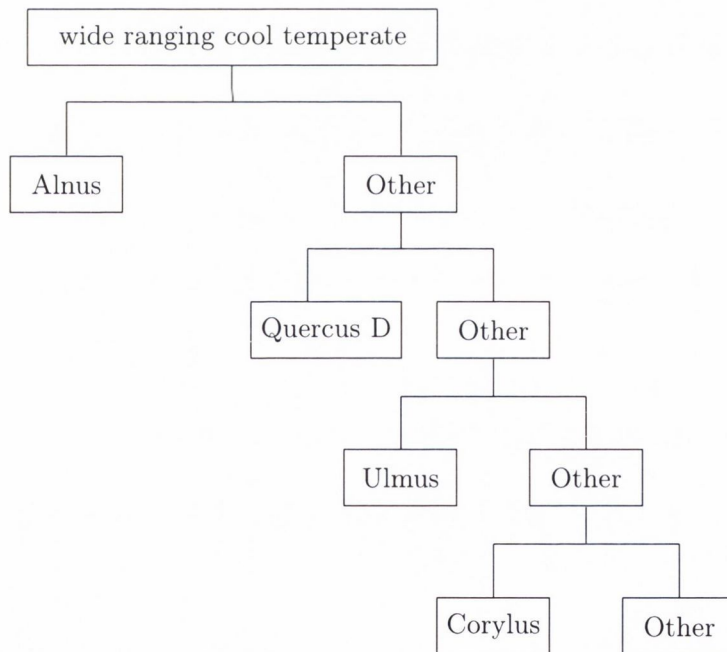


Figure 7.12: Optimal nesting structures for the lowest levels, in terms of inverse cross-validation predictive accuracy, as learned from the data.



(a)



(b)

Figure 7.13: Optimal nesting structures for the lowest levels, in terms of inverse cross-validation predictive accuracy, as learned from the data.

Model	Δ	$\bar{\alpha}_1$	$\bar{\alpha}_2$	$\bar{\sigma}^2$	\overline{RMSEP}	\overline{D}_{mode}
Zero-inflated Binomial	15.57%	0.215	—	8.65	.1255	.1014
Zero/N-inflated Binomial	13.23%	0.203	0.205	4.50	.1252	.0994

Table 7.3: Comparison of predictive performance of the zero-inflated and zero/N-inflated Binomial model.

difference between the models. The average inferred value of σ^2 across all 27 nests is 8.65, as compared to an average value of 4.5 for the zero/N-inflated model. The explicit modelling of sources of N-inflation in the data results in a model with improved predictive performance in terms of Δ - comparison of the average \overline{RMSEP} and \overline{D}_{mode} in Table 7.3 indicate that the zero/N-inflated model produces posteriors on climate which are more accurate in terms of location despite being slightly less conservative.

For the fully nested model, the parameters corresponding to each nest level can be independently fit due to the assumption of conditional independence of the each nest level given the sum constraint and climate. If this assumption is accurate, Δ across all 27 nests should be approximately 5%. However, the saturated cross-validation inverse predictive accuracy of the model is actually 13.23%. This indicates that there is residual unmodelled dependence structure, either in the nesting decomposition used or due to the omission of an important climate variable in the forward model. However it must also be noted that the extension of the partially nested model of Salter-Townshend (2009) and explicit modelling of N-inflation in the data results in a substantial increase in predictive accuracy; Δ reduces from 26.46%, given the partially nested zero-inflated Beta-Binomial model in Salter-Townshend (2009), to 13.23% given the fully nested zero/N-inflated Binomial model presented in this thesis.

Comparison Between Nested and Non-Nested Models

In Figure 7.14 we observe that the inverse predictive power of the nested model is not substantially different from that of the marginal models - both the \overline{RMSEP} and \overline{D}_{mode} are slightly larger on average, indicating that the inverse predictive densities produced by the nested model are more conservative. This is to be expected, as the nested model takes into account the compositional nature of the dataset - the fully nested model contains 27 groups whereas the by-taxon model considers there to be 28 independent taxa.

In Table 7.4 we observe that there is a substantial increase in saturated cross-validation inverse predictive accuracy for the nested model as compared to the non-nested models - Δ reduces from 20.59% for the best by-taxon model to 13.23% for the nested model at the cost of slightly more conservative prediction intervals as evidenced by the larger \overline{RMSEP} statistic.

In accounting for the compositional nature of the dataset, the posterior predictive densities produced by the nested model will necessarily be more conservative than the non-nested equiv-

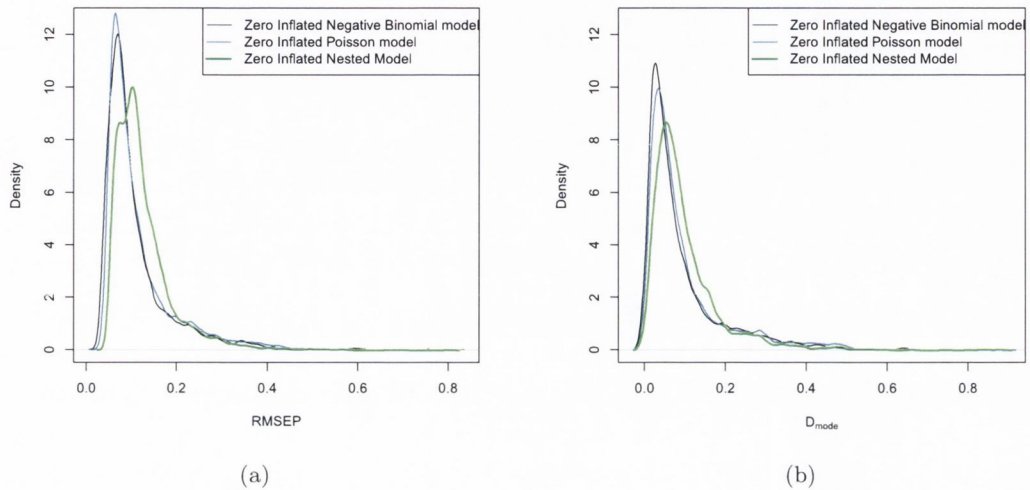


Figure 7.14: (a) Comparison of the sample density of the (a) $RMSEP$ and (b) D_{mode} for the three 2D models. On average, the zero-inflated Binomial model is the model with the poorest predictive power in terms of the $RMSEP$ and D_{mode} but the best in terms of Δ .

Model	Δ	$RMSEP$	\bar{D}_{mode}
Zero-inflated Negative Binomial	20.59%	0.1083	.0881
Zero-inflated Poisson	21.19%	0.1161	.0972
Zero/N-inflated Binomial	13.23%	.1252	.0994

Table 7.4: Comparison of results for the three models

alents. This is observed in Figure 7.15 where contour plots of 95% posterior inverse predictive density regions, produced by both the nested and non-nested model, are presented for a number of reconstruction examples. The marginal models are superior to the nested model in terms of predictive power but less favourable in terms of a leave-one-out cross-validation statistic - this illustrates the point that, in the context of the palaeoclimate reconstruction problem, the *RMSEP* may not be the most suitable measure for determining the “best” model.

Shortcomings of 2D Models

The two dimensional climate setting has revealed important features of the reconstruction problem. As per Salter-Townsend (2009), the treatment of the taxa as conditionally independent given 2D climate results in poor model performance, in terms of saturated cross-validation inverse predictive accuracy. As observed in the above, this is partially due to the compositional nature of the dataset; nested structures are used to improve the cross validation prediction accuracy from 79.41% ($\Delta = 20.59\%$) in terms of the zero-inflated Negative Binomial model to 86.77% ($\Delta = 13.23\%$) for the zero/N-inflated Binomial model. Whilst the compositional nature of the data is an important factor in the deterioration of the predictive performance of the marginal model, the saturated inverse predictive accuracy of 86.77% for the fully nested model reflects that are other factors which impact on model performance; given the “true” model, the Δ statistic should be approximately 5%.

Salter-Townshend (2009) proposed a link between poor model performance and increasing altitude and also provided evidence for a link between poor predictive power and AET/PET. For the best fitting 2D climate model, the zero/N-inflated Binomial model, we observe the same result, namely a correlation between poor model predictive performance and both altitude and AET/PET.

In Figure 7.16 we observe the nature of this correlation; the predictive accuracy of the calibrated models is observed to deteriorate for increasing altitude and decreasing AET/PET, indicating that the predictive performance of the nested model is not as accurate in arid climate regions as compared to regions with plentiful moisture. Increasing altitude appears to have an impact on prediction accuracy; however this is perhaps once more a manifestation of the lack of a moisture variable - plant species which may not be well suited to the climate conditions at sea level in hot arid regions may thrive at higher altitudes where the climate is slightly cooler and more moisture potentially available.

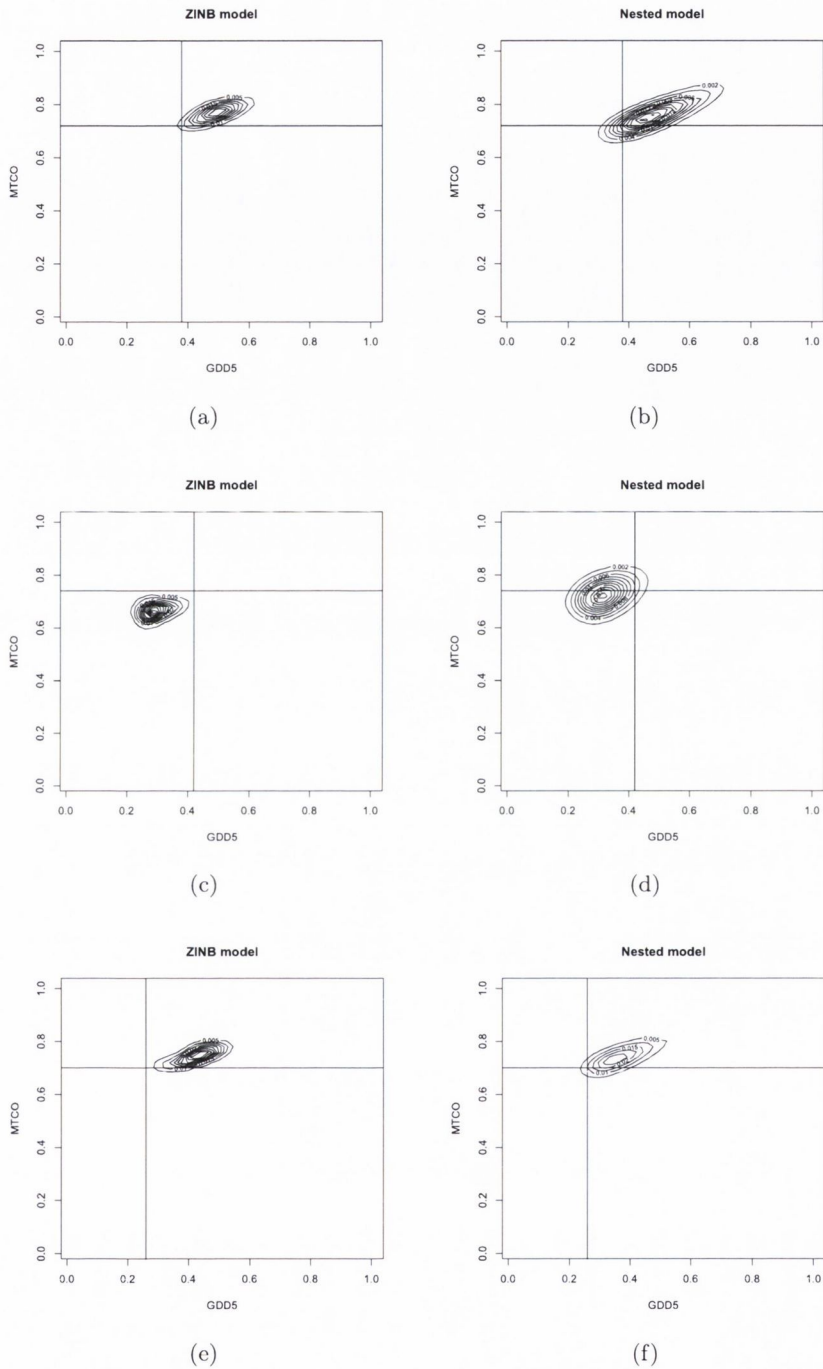


Figure 7.15: Comparison of the 95% inverse cross-validation predictive densities produced by the zero-inflated Negative Binomial (ZINB) model and the zero/N-inflated Binomial (nested) model for three count examples. The nested model is observed to produce predictive densities that are more conservative, as evidenced by the larger probability regions and smoother contours, acknowledging the compositional nature of the data. The true climate location is marked by the intersection of the two lines.

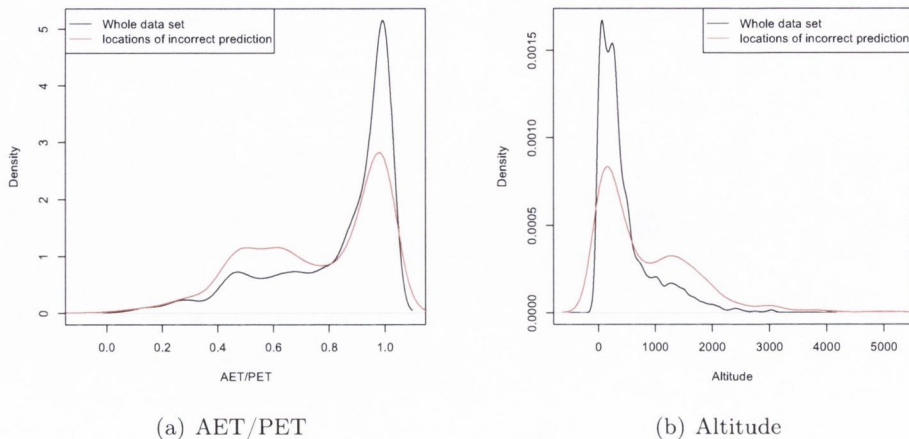


Figure 7.16: Comparison of the sample density plots of (a) AET/PET and (b) Altitude contrasting the sample density of the locations at which prediction accuracy is poor versus the sample density of all data locations. There appears to be a link between poor model predictive performance and both extreme altitude and extreme AET/PET.

7.3.3 Results: 3D Climate Application

In this section we evaluate the impact on prediction accuracy of the incorporation of an additional climate variable, AET/PET, into the forward models. As mentioned previously, making inferences on the parameters of the 3D models is computationally very costly - this is due to the sheer number of parameters introduced by the discretization of 3D space at the forward stage; empirical Bayes based inference on the parameters of the 3D model for each taxa takes the order of a day. For the sake of brevity in the following, we focus on the results of the 2 best fitting models (as identified in the 2D setting); the zero-inflated Negative Binomial model and the zero/N-inflated (nested) Binomial model.

Zero-Inflated Negative Binomial Model

The Δ statistic for the zero-inflated Negative Binomial model in 3D indicates that the model provides a better fit to the observed data than the equivalent 2D model with AET/PET omitted; Δ for the 3D model is 5.04% lower than that of the corresponding 2D model (15.54% vs 20.59%). As the only difference between the models is the inclusion of the AET/PET variable, this indicates that AET/PET is important for accurate climate prediction. This confirms the hypothesis of Huntley (1993) who argued that climate models conditioned on these 3 particular aspects of climate, at a minimum, are required.

While the predictive performance of the model is improved in the 3D setting, in Figure 7.17, we observe that the individual pollen taxa are still not conditionally independent given 3

aspects of climate. However this approximation is more appropriate than the $2D$ setting as evidenced by the reduced Δ statistic overall.

The primary reason for the increase in predictive accuracy appears to be that the climate posteriors produced by the $3D$ model are significantly more conservative than those of the $2D$ model. In Table 7.5, this is indicated by the value of \overline{RMSEP} being substantially larger than the predictive posteriors produced by the $2D$ models. In Figure 7.18 we illustrate this result graphically, plotting a number of cross-validation inverse predictive densities which help to emphasize the extent to which the predictive regions are more conservative.

In order to produce the $2D$ marginalised contour plots in Figure 7.18, the inverse predictive densities for the $3D$ model (MTCO, GDD5, AET/PET) are marginalised across the AET/PET dimension. This is easily achieved as the joint inverse predictive density on climate is defined on a discrete $3D$ ($50 \times 50 \times 50$) lattice - the marginalised $2D$ (MTCO,GDD5) predictive density of the $3D$ model is thus obtained by summing across the lattice in the AET/PET direction and renormalising.

Model	Δ	\overline{RMSEP}	$\overline{D}_{\text{mode}}$
2D Zero-inflated Negative Binomial	20.59%	.1083	.08962
3D Zero-inflated Negative Binomial	15.54%	.1458	.1122

Table 7.5: Comparison of results between the $2D$ and $3D$ marginal models

Zero/N-inflated Binomial Model

The use of the of zero/N-inflated Binomial models, in the context of Multinomial problems, requires the specification of a nesting structure. To this end, the optimal nesting structure identified in the $2D$ setting in Section 7.3.2 is carried forward to the $3D$ setting. Whilst a more appropriate approach is to re-infer the optimal nesting structure for the $3D$ model, the computational cost of model fitting render such an approach infeasible.

In Figure 7.19 we compare the sample densities of the model evaluation statistics from the $2D$ setting with the $3D$ equivalents. Given the incorporation of the additional AET/PET climate variable, the nested model in $3D$ produces posteriors on climate that that are significantly more conservative than the $2D$ version.

The superior predictive accuracy of the $3D$ model, as compared to the $2D$ setting is displayed in Table 7.6. We observe that the incorporation of the AET/PET climate variable into the forward models results in enhanced predictive performance; the Δ statistic reduces from 13.23% for the 2 dimensional model to 9.32% for the 3 dimensional model.

In Figure 7.20 we illustrate the extent to which the inverse predictive densities for the zero/N inflated Binomial model in $3D$ are more conservative than their $2D$ equivalents.

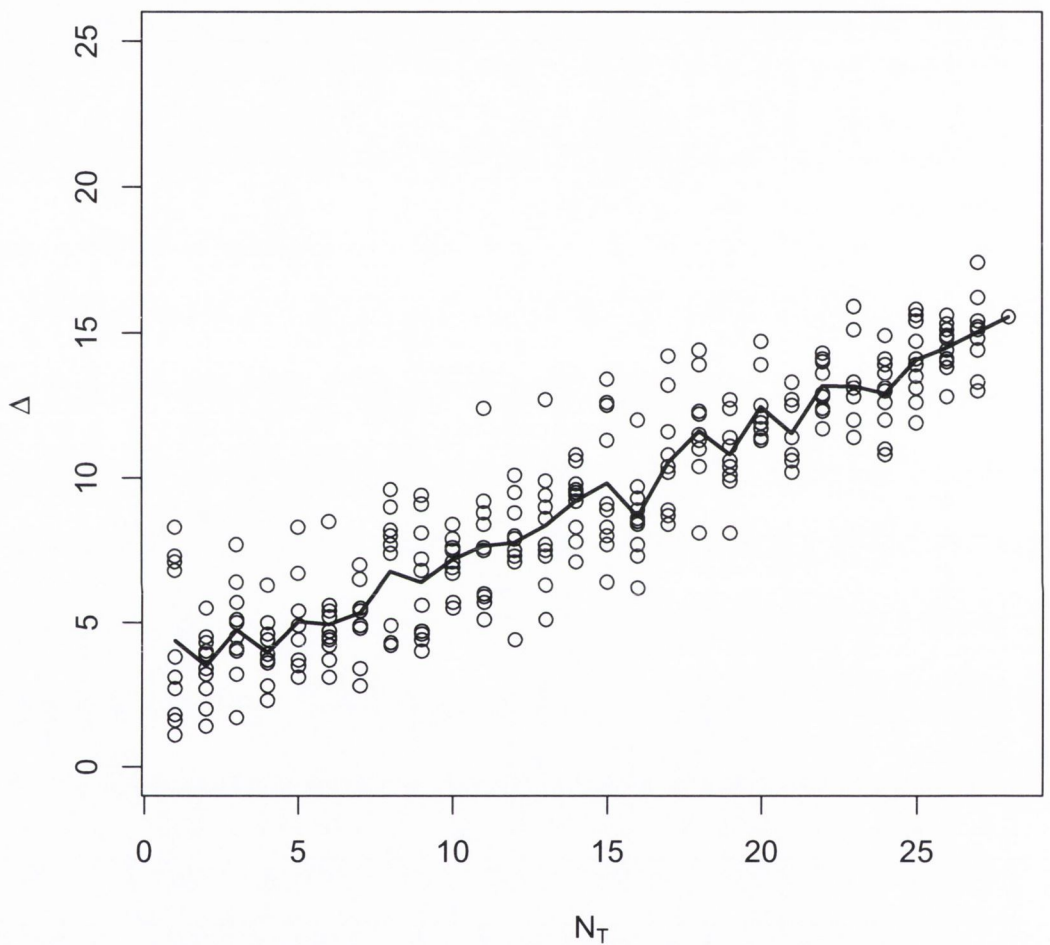


Figure 7.17: Plot of Δ , the percentage of observations which fall outside their respective saturated inverse cross-validation 95% HPD regions for climate given the fitted 3D models. As in the 2D setting, the predictive performance of the models, in terms of Δ , is seen to deteriorate linearly with each additional taxa considered, indicating the the plant taxa are not conditionally independent given models conditioned on 3 aspects of climate. Given the “true” model Δ should be approximately 5%.

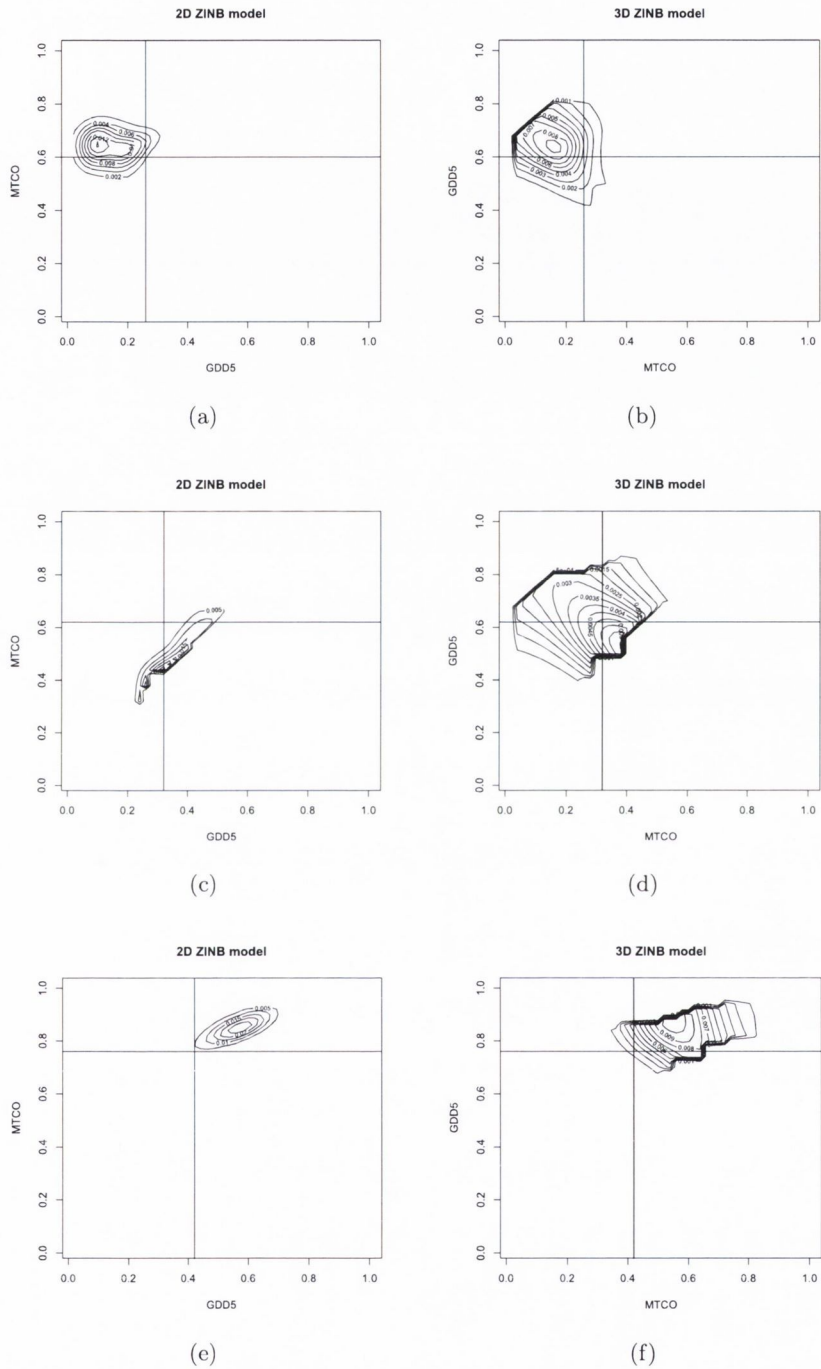


Figure 7.18: Comparison of the cross-validation 95% inverse predictive densities produced by the zero-inflated Negative Binomial (ZINB) model in both 2 and 3 climate dimensions. For comparison purposes, the joint climate posteriors in 3D are marginalised to the 2D setting. The buffer region used in the 2D setting to restrict implausible climates is clearly visible in the 3D reconstructions, resulting in boundary effects on the inverse predictive densities (denoted by the dark black regions in the contour plots).

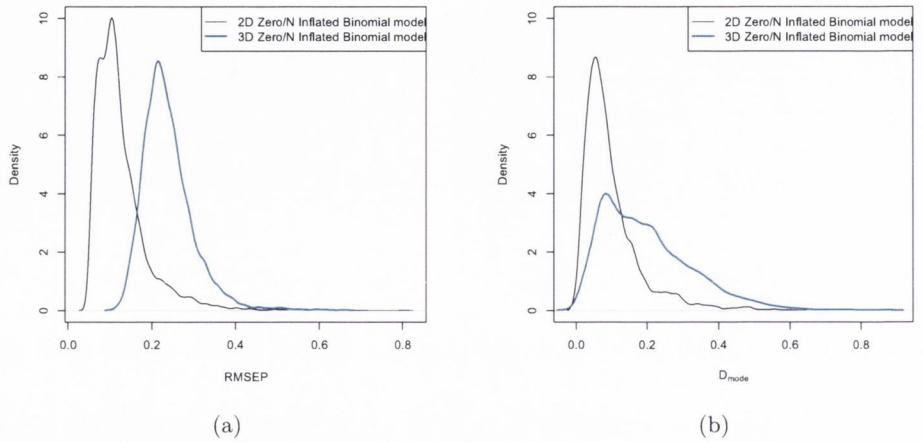


Figure 7.19: Sample density plots of the $RMSEP$ and D_{mode} for the zero/ N -inflated Binomial model in $2D$ and $3D$.

Model	Δ	\overline{RMSEP}	\overline{D}_{mode}
2D Zero/ N -inflated Binomial	13.23%	.1252	.0994
3D Zero/ N -inflated Binomial	9.32%	.2372	.1943

Table 7.6: Summary statistics for model fit and comparison for the zero/ N -inflated model for both 2 and 3 dimensions of climate. The zero/ N -inflated model in $3D$ is a better fit to the data in terms of Δ but provides more conservative posteriors on climate as evidenced by \overline{RMSEP}

Shortcomings of 3D Models

Whilst the predictive accuracy of the reconstruction models is observed to improve substantially given the incorporation of the AET/PET variable, model shortcomings remain. For example, for the model with the best predictive performance in 3D, the zero/N-inflated Binomial model, the saturated cross-validation accuracy, Δ , is 9.32%. Whilst this represents a substantial improvement in prediction accuracy over the 2D setting, its value is still significantly greater than 5%, indicating that there may remain unidentified factors which impact on model performance.

There are several possible sources for the loss in predictive power. Inference procedures are empirical Bayes based, thus all uncertainty in the model parameters is not taken into account. Furthermore, the forward model was calibrated using a dataset from which the detected outliers were not excluded. One other possible source for the loss in predictive power was identified in Section 7.2.5 previously; specifically, exploratory analysis of the posterior random effect terms of the *Cedrus* taxon revealed a link between outlying behaviour, in terms of larger than expected counts, and extreme MTWA values. This result is confirmed across all taxa in Figure 7.21 below; essentially, a sample density plot of the MTWA values recorded at each of the 7742 sites is compared to a sample density plot of the MTWA values corresponding to the 9.32% of observations in the RS10 dataset for which climate was incorrectly classified. A distinct pattern emerges - climate prediction accuracy is observed to be poor at sites with extreme MTWA values. In Section 8.2.2 we discuss a modelling approach which may facilitate the inclusion of the MTWA climate covariate in the forward models to address this issue.

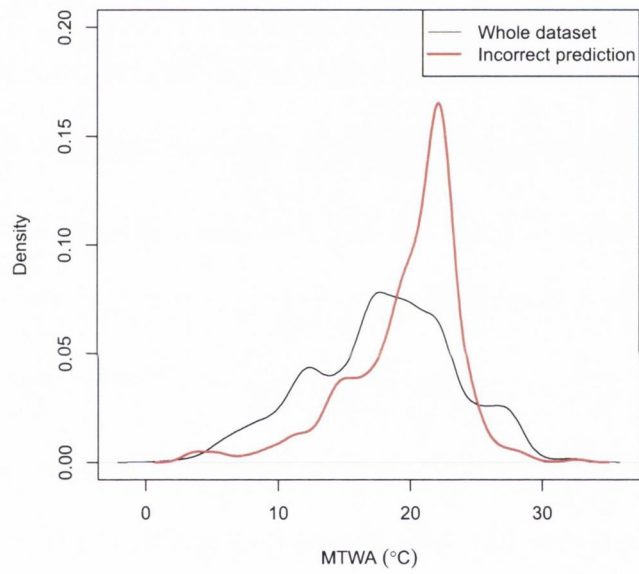


Figure 7.21: The sample density of 2 sets of MTWA from the RS10 dataset. The black line represents the entire dataset whilst the red line represents the sample density of those altitudes for which prediction was incorrect.

7.4 Fossil Climate Reconstruction at Glendalough

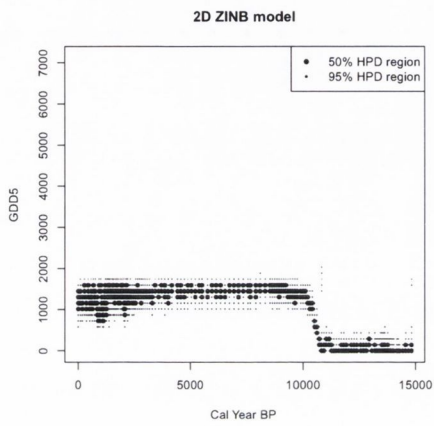
In the following we compare and contrast the fossil climate reconstructions produced by a number of the models introduced in the preceding sections - the specific purpose here is to highlight the impact of model choices at the forward stage on the climate predictions that are ultimately produced using the calibrated models at the inverse stage.

The fossil pollen core we use for illustrative purposes is from the Glendalough lake site in Wicklow, Ireland (Haslett et al. 2006). The data consists of 150 slices of lake sediment with sample pollen percentages for each of 28 plant taxa available for each slice. In order to transform the percentages into a format compatible with the calibrated forward models, the percentages are rescaled to a per mille basis and a total count sum of 1000 is then assumed. For simplicity in the following, we fix the calendar (year) ages of the slices at their maximum *a posteriori* (MAP) calendar ages, as obtained from Haslett & Parnell (2006), with the ages ranging from present day to approximately 15,000 years ago. Haslett et al. (2006) provide an estimate of present day GDD5 and MTCO values for Glendalough, giving estimates of 1772 degree days for GDD5 and an MTCO value of $4.4^{\circ}C$. Based on the estimated (modern) AET/PET values of a number of sites (in the RS10 training dataset) coincident to the Glendalough site, an AET/PET ratio of 1 seems reasonable.

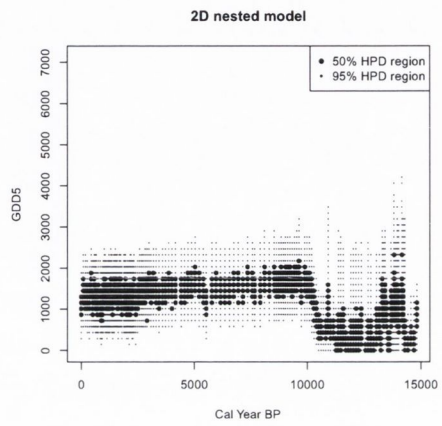
We consider two models for reconstruction purposes. These are the zero-inflated Negative Binomial model and the zero/N-inflated (nested) Binomial model. Reconstructions using both 2D and 3D versions of the models are produced; model inversion is via the novel sampling-based scheme introduced in Section 6.3.2. Use of the sampling scheme here results in a significant speed up in model inversion, reducing the time taken to provide the climate reconstructions by half in the context of the 2D zero/N-inflated Binomial model (12 minutes vs 25 minutes) and by a factor of ≈ 20 for the 3D models (1.15 hours versus 22 hours). 10,000 samples appear sufficient to produce accurate inferences on fossil climate for each respective model.

In Figure 7.22 (a - d) we present the climate reconstructions for GDD5 at Glendalough. The reconstructions produced by each model indicate the occurrence of an extreme climate event approximately 10,000 years ago and this event is known as the Younger Dryas (see Haslett et al. (2006) for further details). Three of the four models indicate the presence of an additional climate event approximately 14,000 years ago, though in the context of 3D zero-inflated model, evidence for this is weaker than the nested models. Note that, owing to the zero/N-inflated model accounting for the compositional nature of the pollen dataset, the resulting climate reconstructions contain much more uncertainty than the zero-inflated Negative Binomial model equivalents.

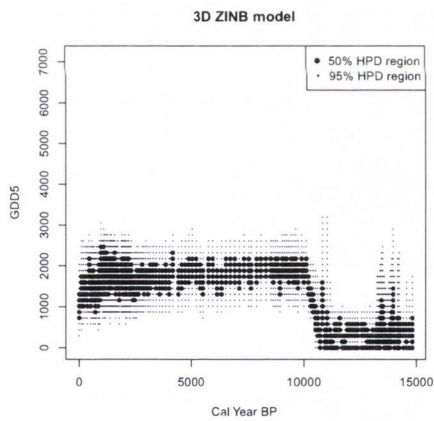
In the context of the MTCO reconstructions at Glendalough, in Figure 7.23 (a-d) we observe a much greater contrast in the climate reconstructions produced by each model, noting



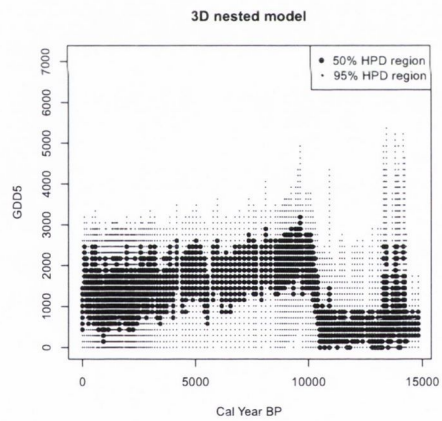
(a)



(b)



(c)



(d)

Figure 7.22: 50% & 95% HPD regions for the reconstruction of GDD5 at Glendalough for the zero-inflated Negative Binomial model and zero/N-inflated (nested) Binomial model in both $2D$ and $3D$.

a striking difference between the respective *2D* and *3D* reconstructions. With the inclusion of the AET/PET covariate, the inferred climate change corresponding to the Younger Dryas at Glendalough appears to be -35°C on average lower for the *3D* zero-inflated Negative Binomial model as compared to its *2D* equivalent. The corresponding difference for the zero/N-inflated model is around -15°C . Additionally, both *3D* models indicate the presence of a rapid warming and cooling event around 14,000 years before present - for the *2D* zero/N-inflated model, the evidence for this is much weaker. Crucially, the *2D* zero-inflated Negative Binomial model, as in Haslett et al. (2006), does not well reflect the Younger Dryas at all.

However, there is significant contrast between the respective *3D* reconstructions, in that the *3D* zero-inflated Negative Binomial model suggests that the change in MTCO during the Younger Dryas period was significantly more extreme than that proposed by the zero/N-inflated Binomial model. A further important point to note is that, whilst the *2D* zero/N-inflated Binomial model and the *3D* zero-inflated Negative Binomial model are broadly similar in terms of saturated cross-validation prediction accuracy, they produce fossil climate reconstructions that are substantially different; this is an important result, indicating the influence of the AET/PET climate variable on the MTCO reconstructions produced and validating its inclusion in the forward models.

Finally, as the *2D* models do not include the AET/PET variable, only two reconstructions are presented in Figure 7.24 (a - b). The *3D* zero-inflated Negative Binomial model indicates a substantial change in the moisture availability for plant uptake during the Younger Dryas, suggesting climate conditions similar to boreal or arctic type climatic conditions, and a substantial decrease in moisture availability. Conversely, reconstructions from the zero/N-inflated model suggest that the opposite was the case; while the 50% HPD regions appear more conservative pre-10,000 years ago than at present, they still place the majority of the highest predictive mass at AET/PET ratios above .6, indicating a far less harsh climate than that reconstructed by the zero-inflated Negative Binomial model.

In the following, a more in-depth evaluation of the produced climate reconstructions is provided. We identify the *3D* zero/N-inflated Binomial model as producing reconstructions which appear to best agree with reconstructions from other, independent sources and also provide a rationale for this result.

7.4.1 Discussion

In the preceding section we observed that vastly differing climates at Glendalough are reconstructed conditional on the forward model used. In the following, we discuss these climate reconstructions in further detail and discriminate between them by comparing them to reconstructions obtained from independent sources.

Firstly, in Figure 7.25 we present the temperature reconstruction at a site in Greenland

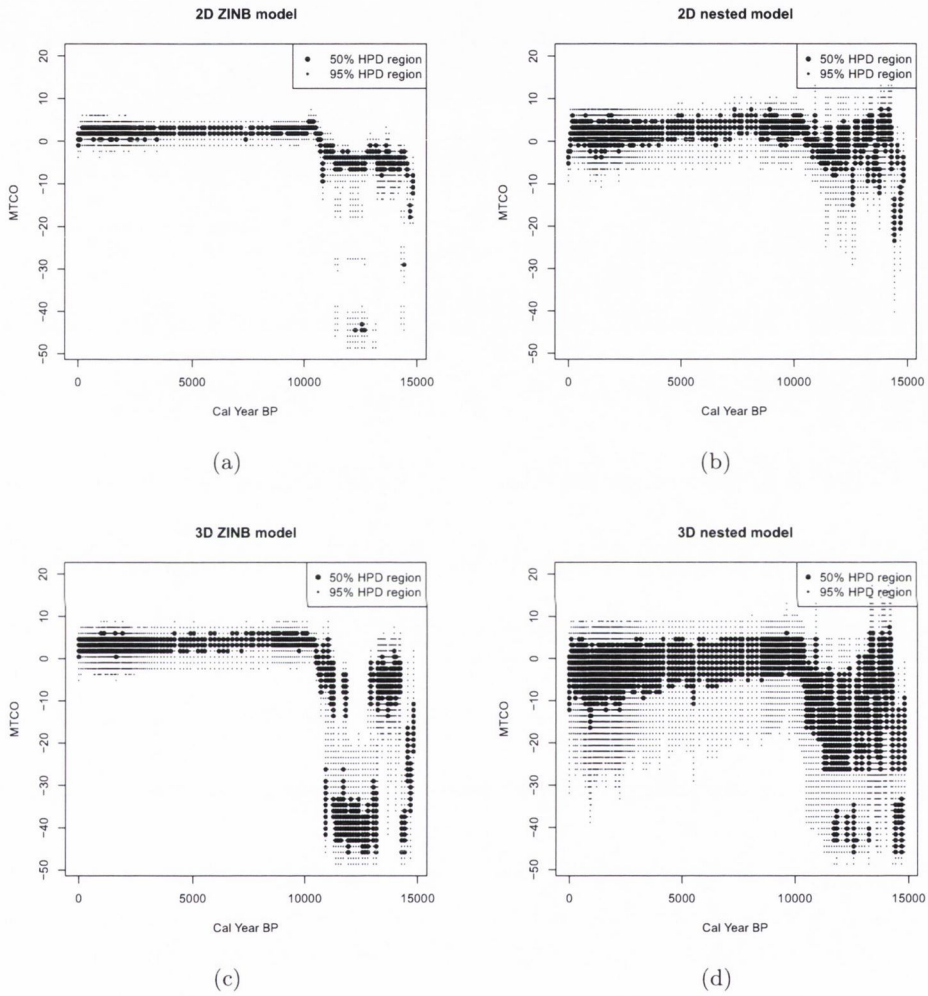


Figure 7.23: 50% & 95% HPD regions for the reconstruction of MTCO at Glendalough for the zero-inflated Negative Binomial model and zero/ N -inflated (nested) Binomial model in both $2D$ and $3D$. There are striking differences in the reconstructions produced by each model.

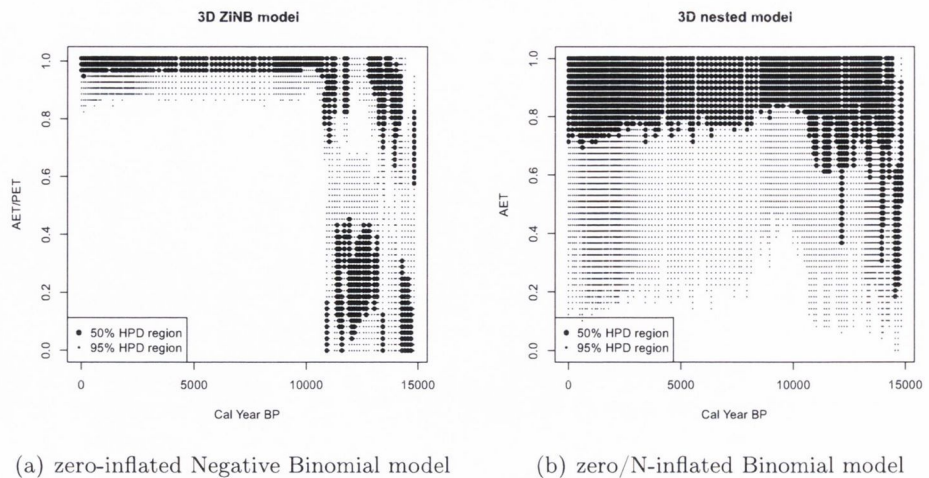


Figure 7.24: 50% & 95% HPD regions for the reconstruction of AET/PET at Glendalough for the zero-inflated Negative Binomial model and zero/N-inflated (nested) Binomial model in both $2D$ and $3D$.

obtained via analysis of high resolution oxygen-isotope records; this is the GISP2 dataset of Grootes & Stuiver (1995)). The reconstruction indicates there were rapid and extreme temperature changes in the period between 10,000 - 15,000 calendar years before present.

While the location of Greenland is (relatively) far from Ireland, climate reconstructions for locations across Europe, which also indicate large scale climate changes in this period, are separately provided by a number of authors; Seret et al. (1992) detail rapid temperature changes at Place des Vosges in France during this period, reconstructing climate using data from both beetles and pollen. Watts et al. (1996) and Allen et al. (1996) observe large-scale climate variability for sites located at Iberia in Spain and Monticchio in Italy (both reconstructions are pollen-based). Isarin (1997) reconstructs the temperature across Europe during the Younger Dryas, noting large-scale temperature drops during this period based on the analysis of periglacial features. Brooks & Birks (2000) (using chironomids) and Atkinson et al. (1987) (using beetles) also provide evidence for large scale climate changes during this period at sites in Britain.

Therefore, our first conclusion is that the palaeoclimate reconstructions produced by the $2D$ models are not plausible as they do not agree with other independent sources. The $2D$ zero-inflated Negative Binomial model does not identify the large scale climate changes which occurred during this period. Furthermore, based on the available literature, the changes indicated by the $2D$ zero/N-inflated Binomial model are not sufficiently extreme. These conclusions reflect the importance of the AET/PET climate variable for accurate reconstructions using pollen-based proxies.

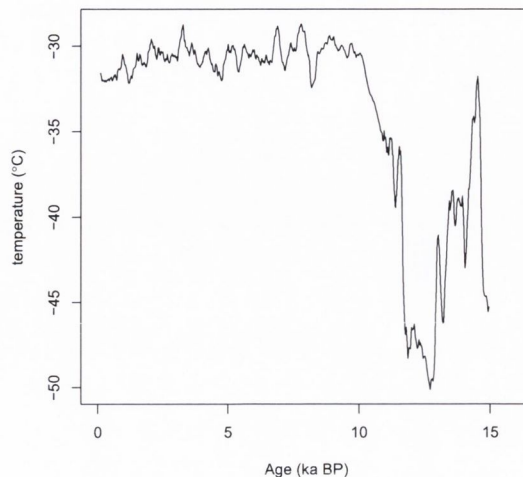


Figure 7.25: Temperature reconstructions at a site in Greenland for the past 15,000 years.

A further result which requires investigation are the substantial differences in the reconstructions of MTCO and AET/PET produced by the 3D models during the Younger Dryas period. According to Isarin (1997), average MTCO temperatures during this period, at latitudes similar to that of Glendalough, were between -15°C and -25°C . Atkinson et al. (1987) propose values between -20°C and -25°C for sites across Britain (including Ireland). Further evidence for this range of values is provided by O’Connell et al. (1999) who suggest that “winter temperatures were probably -20°C or lower” in Ireland during the Younger Dryas period. The inferences made by each respective author were derived from quantitative-based reconstruction methods.

These results indicate that the reconstruction for MTCO produced by the zero-inflated Negative Binomial model for the Younger Dryas period at Glendalough is in fact too cold. This perhaps, is a reflection of model inadequacies - the compositional nature of the data is not taken into account at either the forward or the inverse stage. The reconstruction is thus based on the premise that the 28 available pollen counts, many of which are zero, each provide independent pieces of information, which is untrue. Additionally, a large percentage of the fossil pollen counts during the Younger Dryas correspond to the *Betula* taxa. This perhaps provides a further explanation for the extreme temperatures reconstructed - Seret et al. (1992) notes a tendency for reconstruction models to produce climate reconstructions that are too cold in settings where *Betula* dominates the pollen assemblage.

One final conclusion we make is that, in the context of the various reconstruction models presented in this thesis, the zero/N-inflated Binomial model appears to produce fossil climate reconstructions which agree favourably with the reconstructions produced by several authors,

based on a variety of reconstruction methods, and is therefore the model to be preferred for reconstruction purposes.

7.5 Conclusions

In Salter-Townshend (2009), the predictive accuracy of the partially nested model was just 74.5% ($\Delta = 25.5\%$). Uncertainty in the recorded climates at individual sites was proposed as the cause of the poor prediction accuracy and ad-hoc, “Gaussian blurring” of posteriors was used to substantially improve this result. However, the work detailed in this chapter has indicated that this hypothesis is erroneous - the poor predictive performance experienced by Salter-Townshend (2009) appears to have been a manifestation of the failure to fully account for the compositional nature of the RS10 dataset and the omission of the AET/PET climate covariate from forward models. The evidence for this statement comes from the result that the saturated cross-validation inverse predictive accuracy of the 3D zero/N-inflated Binomial model, developed in this thesis, is approximately 91%, indicating that there is little impact of uncertainty in the data upon predictive performance.

Explicit criticism of the forward models was not considered in Haslett et al. (2006) or Salter-Townshend (2009). In contrast, here the use of Gaussian random effect terms to model overdispersion of the pollen counts provides a method for examining model fit at the forward stage. For the zero-inflated (Gaussian overdispersed) Poisson model in 2D, we observed that the *a priori* specification of a Gaussian distribution for the random effect terms was inappropriate for a number of the pollen taxa (Figure 7.10), with quantile-quantile plots of the mean posterior random effects indicating that a Gamma distribution is perhaps more suitable. Further evidence for this result is perhaps provided by the slightly better predictive performance of the zero-inflated (Gamma overdispersed) Negative Binomial model in 2D. However, in the context of the 3D zero/N-inflated model, the *a priori* assumption of Gaussian behaviour proved quite reasonable, as determined by the visual analysis of a number of quantile-quantile plots (Figure 7.1), though there appear to be some evidence of skew behaviour in the posterior random effect terms.

As previously illustrated, the posterior random effect terms are a flexible tool that may also be used for outlier detection: analysis of the posterior random effect terms for the *Cedrus* taxon allowed for a detailed investigation of potentially outlying observations within the training dataset. A number of outliers were identified, some of which were possibly due to mislabeling of pollen samples at the data collection stage. However, other detected outliers hinted at underlying model weaknesses; the random effects provide evidence that the addition of the MTWA climate variable to the forward models may be required for more accurate climate reconstruction. A future task involves the repeating of this analysis for each taxa with the aim of removing the spurious observations from the RS10 training dataset.

For all models applied to the RS10 training dataset, the overdispersion parameters are significantly non-zero. This indicates there is variability in the pollen dataset over and above that expected by the zero-inflated models. A further point to note is that there appears to be substantial N-inflation in the dataset. This result is confirmed by the significantly non-zero average value of α_2 across all taxa in Table 7.3. In attempting to account for the excess variability in the dataset due to this N-inflation, the overdispersion parameter of the zero-inflated Binomial model is, on average, significantly overestimated (Table 7.3). Note that the statistically consistent zero/N-inflated Binomial model displays superior predictive performance, in terms of Δ , to the statistically inconsistent zero-inflated Binomial model.

Examination of the Δ statistics of each of the marginal models reveal that, whilst the assumption of taxon independence in three climate dimensions appears to be more acceptable than the assumption of taxon independence given just two, the model fit is still quite poor ($\Delta = 15.84\%$). As the RS10 dataset is compositional in nature, it is speculated that this poor predictive accuracy is due to the lack of accounting for this compositional structure in the forward models.

There are striking differences in the reconstructions obtained for the fossil climate at Glendalough by both marginal and nested models in $2D$ and $3D$. The results obtained using the best fitting $3D$ zero/N-inflated model are shown to be the only one to agree with the existing literature; the use of a by-taxon model in $3D$ reconstructs climates during the Younger Dryas that are simply too cold to have existed in Ireland at that time. This is due to the predominance of *Betula* pollen in the fossil pollen assemblage during this period. An additional conclusion of the comparison of the reconstructions in both $2D$ and $3D$ is that the AET/PET climate variable is essential for the capturing of extreme climate events in the fossil pollen record.

The development of a fast sampling-based scheme for model inversion is shown to substantially reduce the time taken in model inversion. As compared to numerical integration methods based on quadrature, the sampling-based scheme results in a speedup of model inversion by a factor of 20 for the $3D$ zero/N-inflated Binomial model. The corresponding speedup in the $2D$ setting was only a factor of 2, indicating that the computational advantages of the sampling-based scheme are more pronounced for increasing C . If uncertainty in model hyperparameters was additionally taken into account, the computational and time savings of the sampling-based approach would be even greater.

Chapter 8

Conclusions and Further Work

The motivating application of the research contained in this thesis concerns the statistical reconstruction of fossil climate from fossil pollen data. There are many challenging features of this problem, both in terms of model fitting and in the inverse use of the fitted models for prediction. In seeking to address these challenges, several important research contributions have been made. In the following we summarise these contributions and reflect on several conclusions regarding the palaeoclimate reconstruction process, as derived from the applied work.

8.1 Conclusions

Model validation forms a crucial part of model development, involving the evaluation of *a priori* modelling assumptions and the analysis and identification of possibly spurious observations within the training dataset. However, in settings where the response consists of discrete, non-Gaussian count observations, such as in the case of the RS10 pollen dataset, these tasks are difficult to perform.

This problem has motivated the development of a methodology for Bayesian residual analysis and outlier detection in the non-Gaussian setting, based on the analysis of Gaussian approximations to posterior random effect terms. We conclude that the approach has distinct advantages over existing methods as regards both computational speed, crucially due to the harnessing of fast approximate Bayesian inference algorithms, and the automatic provision of metrics by which to systematically determine potential outliers. It is also demonstrated that exploratory tools from classic Gaussian residual analysis may be harnessed to gain an extra insight into underlying model dynamics, facilitating subtle criticisms of extraordinarily complex models. Application to two contrasting datasets in this thesis have revealed the power of the approach.

However, the weaknesses of the proposed methodology are also quite evident. The success,

or otherwise, of the approach is intrinsically linked to the degree of overdispersion in the data, resulting in poor model performance if the overdispersion is small in magnitude. Furthermore, the outlier detection properties of the approach, in the context of observations with low count values such as binary outcomes, or the presence of many zeroes, is shown not to perform well.

Forward models which fail to acknowledge the compositional nature of multivariate count datasets, where the counts are constrained to sum to a total, result in fitted models with poor prediction accuracy, as determined by Δ . The resulting inverse predictive posteriors are not always sufficiently conservative and/or erroneous in location. Hierarchical (nesting) structures are shown to provide a useful manner of addressing this problem, facilitating the decomposition of multivariate compositional data models into a series of separate univariate models, for which inference tasks are much less computationally challenging. Such structures provide a full, but not necessarily unique, decomposition of model likelihoods.

However, careful comparison of the leave-one-out, inverse cross-validation prediction accuracy of each of the possible nesting structures can help discern the optimal structure. In situations where the number of groups is extensive and thus the investigation and evaluation of all possible nesting structures is not possible, expert opinion can be used to reduce the permutations of nesting structures to a manageable number. This approach provides the nesting structure identified for the RS10 pollen dataset in this thesis.

Application of standard zero-inflation models to Binomial response data can lead to statistically inconsistent inferences in model fitting. It is observed that this statistical inconsistency results in the overestimation of overdispersion parameters and a reduction in the inverse predictive accuracy of the calibrated models due to the mislocation of predictive posteriors. A parsimonious model is developed which also addresses this “N-inflation” of the data. The model carries one extra hyperparameter over the standard zero-inflated setting. The application of the new model to the RS10 pollen dataset reveals a substantial improvement in inverse predictive accuracy.

Model inversion via numerical integration methods, in the context of large multivariate inverse inference problems, can be extremely slow and computationally wasteful. Predictive posteriors for climate, given the fossil pollen counts, are frequently only significantly non-zero at a small subset of the discretized space under consideration. Thus the evaluation of climate posteriors at all gridpoints, required in order to obtain normalising constants for numerical integration based model inversion, is disadvantageous.

This problem is addressed in this thesis via the development of a fast sampling based inference procedure for computationally efficient model inversion. Laplace approximations to the inverse predictive posteriors at each gridpoint are used to both detect locations for which the inverse posterior predictive density is negligible, and provide a proposal distribution for the sampling scheme. The dramatic time savings of the approach are illustrated with application to the Glendalough fossil pollen core; full inference on the unknown fossil climates at the

Glendalough site is reduced from 22 hours, using deterministic quadrature, to around an hour for the sampling-based scheme. This procedure has general application in the paradigm of inverse inference problems where model hyperparameters are placed on a discrete grid and the forward model posterior is Gaussian.

Significant advances have been made in this thesis as regards the modelling associated with the palaeoclimate reconstruction problem. The best fitting zero/ N -inflated Gaussian overdispersed (nested) Binomial model is considerably richer than existing models, enabling explicit criticism of both the training dataset and the fitted models. It is demonstrated that the inclusion of an extra climate covariate and the full addressing of the compositional nature of the RS10 dataset leads to a class of models which have significantly superior predictive accuracy as compared to existing approaches. The rate of successful climate prediction, being approximately 91% for the training dataset, is slightly less than desired, reflecting remaining modelling as well as data quality issues. The MTWA climate variable, which evidence suggests is required for more accurate climate prediction, is not currently included in the forward models. Furthermore, the removal of outliers from the training dataset and the updating of models in light of this has not yet been completed.

The reconstruction of the fossil climate at Glendalough has provided considerable information regarding the palaeoclimate reconstruction process. This provides the basis for a number of conclusions. Specifically, marginal models, which do not account for the compositional nature of the pollen data, will perform poorly in comparison to nested models at the reconstruction stage, potentially leading to erroneous and misleading inferences on climate and weaker reconstructions. Additionally, three climate variables, at a minimum, are required in the forward stage models, with the variables MTCO, GDD5 and AET/PET appearing to be particularly crucial for accurate inference on fossil climate. Interaction between these climate variables must also be accounted for at both the forward and inverse stages.

Evidence for these conclusions is provided by the comparison of the reconstruction output of the fitted models with a number of separate, independent reconstructions from the palaeoclimate reconstruction literature, obtained from a variety of proxy sources.

8.2 Further Work

Whilst the research presented in this thesis has contributed substantially to the palaeoclimate reconstruction project, several outstanding challenges remain. In the following the nature of these challenges is briefly outlined.

8.2.1 Analysis of the RS10 Pollen Dataset

A major contribution of the research contained in this thesis is the development of a methodology for forward model criticism and objective outlier detection through the analysis of posterior random effect terms. A sample application of this methodology to the *Cedrus* taxon, presented in Section 7.2.5, revealed model imperfections - several of the proposed outliers were identified as possibly arising due to the failure to explicitly account for the MTWA climate variable in the forward models. Others were identified as a possible manifestation of analyst error in pollen identification at the data collection stage.

However, no attempt is made in this thesis to correct the forward models in light of the data identified as erroneous. For a thorough analysis, the outliers identified for each individual plant taxon must be analysed on a case by case basis to differentiate truly spurious observations from those resulting due to labeling issues, a necessarily laborious task. This task also requires the provision of expert opinion in order to correctly reclassify mislabeled observations. The forward models must then be refit given the updated data set. Only preliminary work in this regard has begun.

8.2.2 4 Dimensional Climate Space

The saturated cross-validation inverse predictive accuracy of the best fitting 3 dimensional (nested) model in this thesis is approximately 91%, that is to say that nearly 91% of the climate observations in the model training dataset are contained within their respective saturated 95% HPD inverse posterior predictive density region. This result indicates that forward models based on 3 dimensions of climate alone are perhaps insufficient to fully describe pollen-climate interaction. Whilst the 3 climate variables considered in the forward models in this thesis are advocated as the most important for climate reconstruction by the botany community, the exploratory analysis presented in Section 7.2.5 identifies a further link between a deterioration in model predictive performance and high values of the MTWA climate variable.

The extension of 3 dimensional models to incorporate this additional climate variable is infeasible given the current modelling approach, which is based on the use of GMRF prior models. The discretization of climate space to a regular $50 \times 50 \times 50 \times 50$ grid in 4 climate dimensions, required for the specification of GMRF prior models, results in the order of 6 million latent variables for each pollen taxa. This number can be reduced dramatically by the “cutting out” of regions of space which are not of interest, as in the 3 dimensional climate setting. However the number of latent variables remaining will still be quite large. An additional obstruction is the nature of the neighbourhood structure of the second order intrinsic GMRF prior which will necessarily be much less sparse in $4D$ than $3D$.

Due to the dimensionality of the RS10 pollen dataset, the consideration of multivariate Gaussian prior models, involving the manipulation of dense covariance matrices of dimension

7742×7742 are equally infeasible. However, Gaussian predictive process models (Banerjee et al. 2008) provide a potential solution in this regard. Essentially, the multivariate spatial process defined at each of the 7742 data locations is projected to a lower dimensional subspace of “knots” which are constructed to be substantially fewer in number than the number of data locations. This results in a substantial speeding up of matrix manipulations involving covariance matrices at moderate cost in loss of inferential accuracy. For an broader discussion of the relative merits and demerits of the approach, we refer the interested reader to Banerjee et al. (2008). One final important point to note is that Eidsvik et al. (2010) has shown that the predictive process approach is also compatible with the INLA algorithm of Rue et al. (2009), retaining the ability to quickly fit and analyse forward models for the data.

Bibliography

- Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Monographs on statistics and applied probability, Chapman & Hall, London.
- Albert, J. (2007), *Bayesian Computation with R*, Springer.
- Albert, J. & Chib, S. (1995), 'Bayesian residual analysis for binary response regression models', *Biometrika* **82**(4), 747–759.
- Allen, J. R. M., Huntley, B. & Watts, W. A. (1996), 'The vegetation and climate of northwest Iberia over the last 14 000 yr', *Journal of Quaternary Science* **11**(2), 125–147.
- Allen, J. R. M., Watts, W. A. & Huntley, B. (2000), 'Weichselian palynostratigraphy, palaeovegetation and palaeoenvironment; the record from Lago Grande di Monticchio, southern Italy', *Quaternary International* **73/74**, 91–110.
- Altas, I., Erhel, J. & Gupta, M. M. (2002), 'High accuracy solution of three-dimensional biharmonic equations', *Numerical Algorithms* **29**, 1–19.
- Atkinson, T. C., Briffa, K. R. & Coope, G. R. (1987), 'Seasonal temperatures in Britain during the past 22,000 years, reconstructed using beetle remains', *Nature* **325**, 587–592.
- Banerjee, S., Gelfand, A. E., Finlay, A. O. & Sang, H. (2008), 'Gaussian predictive process models for large spatial data sets', *J. R. Statist. Soc. B (2008)* **70**(4), 825–848.
- Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Monographs on Statistics and Applied Probability Series, CRC Press.
- Bartlein, P. J., Prentice, I. C. & Webb, T. (1986), 'Climate response surfaces from pollen data for some eastern North American taxa', *Journal of Biogeography* **13**, 35–57.
- Beerling, D., Huntley, B. & Bailey, J. (1995), 'Climate and the distribution of *Fallopia japonica*: use of an introduced species to test the predictive capacity of response surfaces', *Journal of Vegetation Science* **6**, 269–282.
- Besag, J. (1986), 'On the statistical analysis of dirty pictures', *Journal of the Royal Statistical Society* **B**(48), 259–302.

- Besag, J. & Higdon, D. (1999), ‘Bayesian Analysis of Agricultural Field Experiments’, *Journal of the Royal Statistical Society: Series B* **61**(4), 691–746.
- Bhattacharya, S. (2004), Importance Resampling MCMC: a methodology for crossvalidation in inverse problems and its applications in model assessment, PhD thesis, Dept. of Statistics, School of Computer Science and Statistics, University of Dublin, Trinity College.
- Bhattacharya, S. (2006), ‘A Bayesian semiparametric model for organism based environmental reconstruction’, *Environmetrics* **17**, 763–776.
- Bhattacharya, S. & Haslett, J. (2007), ‘Importance Re-sampling MCMC for Cross-Validation in Inverse Problems’, *Bayesian Analysis* **2**(2), 385–408.
- Birks, H. J. B., Heiri, O., Seppä, H. & Bjune, A. E. (2010), ‘Strengths and Weaknesses of Quantitative Climate Reconstructions Based on Late-Quaternary Biological Proxies’, *The Open Ecology Journal* **3**, 68–110.
- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.
- Bjornkamp, B. (2011), ‘Approximating Probability Densities by Iterated Laplace Approximations’, Available at <http://arxiv.org/abs/1103.3508>.
- Breslow, N. E. & Clayton, D. G. (1993), ‘Approximate Inference in Generalised Linear Mixed Models’, *Journal of the American Statistical Association* **88**(421), 9–25.
- Brooks, S. J. & Birks, H. J. B. (2000), ‘Chironomid-inferred Late-glacial air temperatures at Whitrig Bog, southeast Scotland’, *Journal of Quaternary Science* **15**(8), 759–764.
- Buck, C. E., Aguilar, D. G. P., Litton, C. D. & O’Hagan, A. (2006), ‘Bayesian nonparametric estimation of the radiocarbon calibration curve’, *Bayesian Analysis* **1**(2), 265–288.
- Carlin, B. P. & Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Vol. 2nd ed., Chapman & Hall/CRC.
- Chaloner, K. & Brant, R. (1988), ‘A Bayesian approach to outlier detection and residual analysis’, *Biometrika* **75**(4), 651–659.
- Conolly, A. P. & Dahl, E. (1970), Maximum summer temperature in relation to the modern and quaternary distributions of certain arctic-montane species in the British Isles, in ‘Studies in the vegetational history of the British Isles’, Cambridge University Press, Cambridge, pp. 159–233.
- Eidsvik, J., Finley, A. O., Bannerjee, S. & Rue, H. (2010), Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models, Technical report, Norwegian University of Technology (NTNU).

- Fong, Y., Rue, H. & Wakefield, J. (2007), ‘Bayesian Inference for Generalised Linear Mixed Models’, *Biostatistics* **8**(1), 1–27.
- Fraley, C. & Raftery, A. E. (2007), ‘mclust: Model-based clustering / normal mixture modeling’, <http://www.stat.washington.edu/mclust>. R package version 3.1-1.
- Gelfand, A. E. & Smith, A. F. M. (1990), ‘Sampling-Based Approaches to Calculating Marginal Densities’, *Journal of the American Statistical Association* **85**, 398–409.
- Gelman, S. J., Carlin, J. B., Stern, H. & Rubin, D. (2003), *Bayesian Data Analysis*, (Second ed.), Chapman & Hall., London, UK.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), Bayesian model comparison via jump diffusions, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter, eds, ‘Markov Chain Monte Carlo in Practice’, Chapman & Hall, London, pp. 1–21.
- Grootes, P. M. & Stuiver, M. (1995), ‘Gisp2 Oxygen Isotope Data’, *NOAA/NGDC Paleoclimatology Program* .
- Hall, D. B. (2000), ‘Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study’, *Biometrics* **56**, 1030–1039.
- Haslett, J. & Parnell, A. (2006), ‘A simple monotone process with application to radiocarbon-dated depth chronologies’, *Journal of the Royal Statistical Society: Series C* **57**(4), 399–418.
- Haslett, J., Whitley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B. & Mitchell, F. J. G. (2006), ‘Bayesian palaeoclimate reconstruction’, *Journal of the Royal Statistical Society: Series A* **169**(3), 1–36.
- Hastings, W. K. (1970), ‘Monte carlo sampling methods using markov chains and their applications’, *Biometrika* **57**, 97–109.
- Holden, P. B., Mackay, A. W. & Simpson, G. L. (2008), ‘A Bayesian palaeoenvironmental transfer function model for acidified lakes’, *Journal of Palaeolimnology* **39**(4), 551–566.
- Huntley, B. (1993), ‘The Use of Climate Response Surfaces to Reconstruct Paleoclimate from Quaternary Pollen and Plant Macrofossil Data’, *Philosophical Transactions of the Royal Statistical Society* **341**, 215–224. Palaeoclimates and their Modelling.
- Huntley, B. (2001), ‘Reconstructing Past Environments From The Quaternary Palaeovegetation Record’, *Biology and Environment: Proceedings of the Royal Irish Academy* **101B**(1-2), 3–18.
- Isarin, R. (1997), ‘Permafrost Distribution and Temperatures in Europe During the Younger Dryas’, *Permafrost and Periglacial Processes* **8**, 313–333.

- Kneib, T. (2006), Mixed model based inference in structured additive regression, PhD thesis, Faculty of Mathematics, Computer Science and Statistics, LMU München.
- Korhola, A., Vasko, K., Toivonen, H. T. T. & Olander, H. (2002), ‘Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling’, *Quaternary Science Reviews* **21**, 1841–1860.
- Lambert, D. (1992), ‘Zero-inflated Poisson Regression with an Application to Defects in Manufacturing’, *Technometrics* **34**(1), 1–14.
- Li, B., Nychka, D. W. & Ammann, C. M. (2010), ‘The Value of Multiproxy Reconstruction of Past Climate’, *Journal of the American Statistical Association* **105**(491), 883–911.
- Lindgren, F. & Rue, H. (2008), ‘On the Second-Order Random Walk Model for Irregular Locations’, *Scandinavian Journal of Statistics* **35**, 691–700.
- Lindgren, F., Rue, H. & Lindström, J. (2011), ‘An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach’, *Journal of the Royal Statistical Society: Series B* **73**(4), 423–498.
- Marshall, E. C. & Spiegelhalter, D. J. (2007), ‘Identifying outliers in Bayesian hierarchical models: a simulation-based approach’, *Bayesian Analysis* **2**(2), 409–444.
- Metropolis, N., Rosenbluth, A., Rosenbluth, R., Teller, A. & Teller, E. (1953), ‘Equation of state calculations by fast computing machines’, *Journal of Chemical Physics* **21**(6), 1087–1092.
- O’Connell, M., Huang, C. C. & Eicher, U. (1999), ‘Multidisciplinary investigations, including stable-isotope studies, of thick Late-glacial sediments from Tory Hill, co. Limerick, western Ireland’, *Palaeogeography, Palaeoclimatology, Palaeoecology* **147**, 169–208.
- Paciorek, C. J. & McLachlan, J. S. (2009), ‘Mapping Ancient Forests: Bayesian Inference for Spatio-Temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record’, *Journal of the American Statistical Association* **104**(486), 608–622.
- Perperoglou, A. & Eilers, P. H. C. (2009), ‘Penalized regression with individual deviance effects’, *Computational Statistics and Data Analysis* **25**(2), 341–361.
- Prentice, I. C., Bartlein, P. J. & Webb, T. (1991), ‘Vegetation and climate change in eastern North America since the last glacial maximum’, *Ecology* **72**(6), 2038–2056.
- Ridout, M., Demeétrio, C. G. B. & Hinde, J. (1998), Models for count data with many zeroes, in ‘Proceedings of the XIXth International Biometric Conference’, pp. 179–192.
- Roberts, G. O., Gelman, A. & Gilks, W. R. (1997), ‘Weak convergence and optimal scaling of random walk Metropolis algorithms’, *The Annals of Applied Probability* **7**(1), 110–120.

- Rodríguez, G. (2007), ‘Lecture Notes on Generalized Linear Models’, Available at <http://data.princeton.edu/wws509/notes/>.
- Roos, M. & Held, L. (2011), ‘Sensitivity analysis in Bayesian generalized linear mixed models for binary data’, In submission.
- Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Vol. 104 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.
- Rue, H. & Martino, S. (2006), Approximate Bayesian Inference for Hierarchical Gaussian Markov Random Field Models, Technical report, Norwegian University of Technology (NTNU).
- Rue, H., Martino, S. & Chopin, N. (2009), ‘Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)’, *Journal of the Royal Statistical Society, Series B* **71**, 319–392.
- Salter-Townshend, M. (2009), Fast Approximate Inverse Bayesian Inference in non-parametric Multivariate Regression with application to palaeoclimate reconstruction, PhD thesis, Dept. of Statistics, School of Computer Science and Statistics, University of Dublin, Trinity College.
- Salter-Townshend, M. & Haslett, J. (2006), ‘Zero-inflation of compositional data’, *Proceedings of the 21st International Workshop on Statistical Modelling* **21**, 448–456.
- Seret, G., Guiot, J., Wansard, G., de Beaulieu, J. & Reille, M. (1992), ‘Tentative Palaeoclimatic Reconstruction Linking Pollen and Sedimentology in La Grande Pile (Vosges, France)’, *Quaternary Science Reviews* **11**, 425–430.
- Simpson, D., Lindgren, F. & Rue, H. (2011), Fast approximate inference with INLA: the past, the present and the future, Technical report, Norwegian University of Technology (NTNU).
- Souza, A. D. P. & Migon, H. S. (2010), ‘Bayesian outlier analysis in binary regression’, *Journal of Applied Statistics* **37**(8), 1355–1368.
- Subally, D. & Quézel, P. (2002), ‘Glacial or interglacial: *Artemisia*, a plant indicator with dual responses’, *Review of Palaeobotany and Palynology* **120**, 123–130.
- Sweeney, J. & Haslett, J. (2011), ‘Bayesian residual analysis in Poisson regression models’, *Proceedings of the 26th International Workshop on Statistical Modelling* **26**, 587–592.
- ter Braak, C. J. F. (1995), ‘Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-nearest neighbours, partial least squares and weighted averaging partial least squares) and classical approaches’, *Chemometrics and Intelligent Laboratory Systems* **28**, 165–180.

- Thall, P. F. & Vail, S. C. (1990), 'Some Covariance Models for Longitudinal Count Data with Overdispersion', *Biometrics* **46**, 657–671.
- Tierney, L. (1994), 'Markov chains for exploring posterior distributions', *The Annals of Statistics* **22**, 1701–1728.
- Toinonen, H. T. T., Manila, H., Korhola, A. & Olander, H. (2001), 'Applying Bayesian Statistics to Organism-Based Environmental Reconstruction', *Ecological Applications*, **11**(2), 618–630.
- Vasko, K., Toinonen, H. & Korhola, A. (2000), 'A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction', *Journal of Palaeoclimatology* **24**, 243–250.
- Watts, W. A., Allen, J. R. M., Huntley, B. & Fritz, S. C. (1996), 'Vegetation history and climate of the last 15,000 years at Laughi di Monticchio, southern Italy', *Quaternary Science Reviews* **15**, 113–132.