Fast Approximate Inverse Bayesian Inference in non-parametric Multivariate Regression

with application to palaeoclimate reconstruction

A thesis submitted to the University of Dublin, Trinity College

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Statistics, School of Computer Science and Statistics

University of Dublin, Trinity College

April 2009

Michael Salter-Townshend

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 Michael Salter-Townshend.

Michael Salter-Townshend

Dated: May 7, 2009

Abstract

Bayesian statistical methods often involve computationally intensive inference procedures. Sampling algorithms represent the current standard for fitting and testing models. Such methods, while flexible, are computationally intensive and suffer from long run times and high potential sampling error. New methods for fitting non-parametric approximations offer a fast and accurate alternative. Essentially, a multivariate Gaussian distribution is used to approximate the posterior of the model parameters.

Cross-validation is a useful tool in model validation which is an important aspect of statistical inference. Sampling based methods require many re-runs and are impractical for this task. A new method is developed in this thesis that performs fast cross-validation using the Gaussian approximations.

Study of the palaeoclimate provides insight into long-term climate variability. This represents the motivating problem for the work in this thesis. A probabilistic forward model for vegetation given climate is fitted to modern training data using Bayesian methods. The model is then inverted and inference on climate given fossil pollen counts may be performed; this is referred to as the inverse model and cross-validation is preferred in this context.

Highly multivariate models may sometimes be broken down into a sequence of independent smaller problems, which may then be dealt with more easily in parallel. Procedures for assessing the performance of this approach are developed for the inverse problem via fast cross-validation.

Spatial models for counts data with an over-abundance of zeros are developed and synergy with the Gaussian approximation method is demonstrated. Finally, the novel inference methods and new counts models are applied to the palaeoclimate training dataset and progress over the existing methods is demonstrated.

Acknowledgements

My first acknowledgement is to my supervisor, Professor John Haslett. Although this is no surprise (I have yet to see an acknowledgements section for a thesis that did not start with a supervisor) I must emphasize that John has done a good deal more for me than most supervisors would and a great more than his job description would suggest.

John is a functioning collection of seemingly contradictory things; he is both enthusiastic and patient, wise and humble, serious and fun. He has begun to teach me that it is possible to appreciate the big picture and the minute details of a complex problem at the same time. Without this unique set of attributes I have no doubt that I would not have lasted long in statistical research.

John has guided me through a crisis of confidence when I felt that I had nothing to contribute. He has been both friend and mentor. However, it is in the day-to-day supervisory capacity that he has excelled most. When I look back at the range of errors and shortcuts I have attempted to get past him it is bewildering how he has managed to supervise me with a smile and to gently guide me back towards the correct path.

Secondly, I thank Professor Håvard Rue, of NTNU in Trondheim, Norway. I visited Håvard in Norway twice in 2007 for a total of three productive months and was made to feel welcome. Håvard's knowledge of statistics seems almost limitless and the rapidity with which he has always replied to inquiries is astounding. His new methods and approximations are central to this thesis and will no doubt become used in a wide variety of statistical problem. Håvard contributed much discussion in the formulation of the zero-inflated model presented in the thesis.

On a more personal level, I must thank my patient and understanding girlfriend Emma who has put up with this foray into research from the beginning. She has

not only accepted my self afflicted impoverishment but has fed and clothed me on occasion. Most importantly, I have never had difficulty in leaving work in the lab as coming home to Emma is coming home to the most beautiful girl in the world, which tends to re-focus my attention...

I thank my family, in particular my parents, who have part financed my PhD time and are always proud and encouraging. I have always been assured that there is no problem on which I cannot seek their help and advice.

I acknowledge Science Foundation Ireland and Enterprise Ireland for the funding that paid for my time with John. The Norwegian Government Scholarship pool paid for one of the visits to Trondheim.

Michael Salter-Townshend

University of Dublin, Trinity College

April 2009

Contents

List of Tables

List of Figures

Publications

Some of the work presented in this thesis has been taken from the author's following publications, which are co-authored as stated

Haslett J., Bhattacharya S., Whiley M., Salter-Townshend M., Wilson S., Allen J.R.M., Huntley B., and Mitchell F. (2006). Bayesian Palaeoclimate Reconstruction *J. R. Statist. Soc. A.* , **169, Part 3**, 1–36.

Salter-Townshend M., and Haslett J. (2006) Zero-Inflation of Compositional Data, *Proceedings of the 21st International Workshop on Statistical Modelling*, **21, 2006**, 448–456.

Chapter 1

Introduction

Highly multivariate statistical problems may lead to slow inference procedures. One example of such a problem involves palaeoclimate reconstruction from fossil pollen data, which is an example of an inverse problem. Existing explorations of this challenging area of study often involve a trade off between model complexity and speed of inference. Fast approximate Bayesian inference methods offer a solution. In addition, an extension of the methodology allows for model validation to be performed quickly for the inverse problem. Conversely, the large scale of the palaeoclimate project offers a real challenge to the emerging approximate inference engine.

The Royal Statistical Society read paper "Bayesian Palaeoclimate Reconstruction", Haslett et al. (2006) presented work on high resolution pollen based reconstruction of the palaeoclimate at a single location since the last ice-age. This paper outlined the basic concepts involved in performing fully Bayesian inference on unknown climates given modern and fossil pollen data and modern climatic data. The work was a detailed "proof of concept"; extensions and improvements to the statistical methodology were considered, both in the paper and in the subsequent printed discussion.

The main crux of the methodology in that work was acknowledged to be computational; indeed the computational burden imposed compromises on the modelling. The work presented herein represents extensions in the statistical methodology and advances in the computations involved as developed by the author. These contributions are outlined in Section 1.4.

1.1 Palaeoclimate Reconstruction Project

The Bayesian palaeoclimate reconstruction project is an ongoing initiative to build upon existing classical approaches to the reconstruction of prehistoric climates, using fossil pollen data. Specifically, the project seeks to handle all uncertainties quantitatively and coherently in a fully Bayesian framework and to combine different types of information to reduce these uncertainties.

1.1.1 The RS10 Pollen Dataset

The primary dataset for the palaeoclimatology reconstruction project is the RS10 dataset of Allen et al. (2000). A collection of modern pollen surface sample counts $Y^m = \{y_1^m, \ldots, y_M^m\}; M = 7742$ are taken from the uppermost 5 to $10mm$ of lake bed sediment at numerous locations in the northern hemisphere. Along with covariates in the form of local contemporary climate measurements L^m, they comprise the modern dataset. This is also referred to as the training data. Sample fossil pollen counts Y^f are extracted from cores taken from lake or mire sediment. Measures of the prehistoric climate variables L^f at the time of deposition of these fossil pollen spores are unknown; the central premise of palynological palaeoclimate reconstruction is that these climates may be inferred from the pollen data, albeit with some uncertainty. Both the modern and fossil pollen data comprise counts of numerous plant types (taxa; see below). There is therefore a vector of counts reported at each sampling location. The length of this vector is equal to the number of distinguishable pollen spore types.

1.1.2 Response Surfaces

Reconstruction of past climates involves using a multivariate regression type model in which the proportion of the i^{th} species in the training pollen data set y_i^m is an indirect observation of a latent "response" to the corresponding modern climate variables L^m. This response is defined as the propensity to contribute pollen to the dataset in the given climatic conditions and is modelled as a smooth function of climate, fitted by reference to the pollen counts data. At the first stage, the training data is used to calibrate the model for the response of vegetation to climate. At the

second stage, the regression is "inverted" and applied to assemblages of fossil data, which yields a quantitative reconstruction of climate. The two stages are referred to as the forward and the inverse parts of the model respectively. This is known as the *response surface method*.

Huntley (1993) argues that at least some species may have multiple optima and hence the response function may be multimodal. Non-parametric modelling of the response function is therefore advocated. This is due at least in part to the fact that the pollen data is in fact sorted into plant "taxa" rather than individual species. Each taxon consists of one or more species; sometimes an entire genus or even an entire family comprising several plant species are categorized simply as a single taxon. A given taxon may then contain multiple species that thrive and fail at dissimilar climates. This is because the pollen data are categorized visually and multiple related species frequently produce pollen spores that are indistinguishable to the eye.

1.1.3 The Classical Approach

There is a considerable literature on palaeoclimate reconstruction from such palynological data using the response surface methodology in the botany community (see for example Bartlein et al. (1986), Huntley (1993) and Allen et al. (2000)). These reconstructions use various estimation methods, essentially attributing to a fossil pollen assemblage the modern climate that has the "closest" matching pollen assemblage.

The main disadvantage of the classical methods is that there is no consistent way to make statements of uncertainty in the reconstructions. There are other serious issues; for example the RS10 (response surface 10) method of Allen et al. (2000) finds the 10 climates that correspond to the fitted responses that are closest to the fossil pollen assemblage. These climates are then averaged and the average is returned as the estimated palaeo-climate. A crude measure of uncertainty is also reported as the "average chord distance"; the average distance between the vector of fossil pollen proportions and the vectors that are derived from the fitted response surfaces. An immediate problem occurs as a result of this simplified analysis. For example, tundra and steppe vegetation can produce very similar pollen assemblages, yet occur

under very different climatic regimes. For a given fossil pollen assemblage, some of the 10 closest modern day responses may be steppe and some tundra; the averaged corresponding climates will lie in between in the climate space. This reconstructed climate may be in a region of climate that does not produce pollen assemblages anything like the fossil assemblage; it may even correspond to a climate that simply does not occur.

1.1.4 The Bayesian Approach

Unlike the classical approach, the Bayesian paradigm deals with all sources of uncertainty in a coherent manner. The unknown statistical parameters X are treated as random variables and a likelihood function $\pi(Y|X)$ is used to express the relative probabilities of obtaining different values of this parameter when a particular dataset Y has been observed. Prior probability densities $\pi(X)$ are placed on the unknown parameters to reflect any subjective beliefs held before observation of data. The posterior density $\pi(X|Y)$ is delivered via Bayes theorem; it is a normalised product of the prior and likelihood densities and reflects the updated beliefs in light of the data.

$$\begin{aligned} \pi(X|Y) &= \frac{\pi(X)\pi(Y|X)}{\pi(Y)} \\ &\propto \pi(X)\pi(Y|X) \\ &\propto \text{prior} \times \text{likelihood} \end{aligned} \tag{1.1}$$

Bayes theorem constructs the posterior density $\pi(X|Y)$ which is a summary of all knowledge about the parameter X subsequent to observing Y. The posterior distribution is a comprehensive inference statement about the model variables X. Any summary of the posterior distribution is useful e.g. moments, quantiles, highest posterior regions and credible intervals.

The Bayesian model presented in Haslett et al. (2006) is briefly described next. The forward stage of the model infers the latent response of vegetation to climate given the modern pollen counts and corresponding modern climate data. The inverse stage then uses knowledge of the latent responses to infer climate from fossil counts data.

Forward Problem

In this stage of the inference, the modern training data of pollen counts and associated climatic data are used to inform probabilistic statements on the unobservable response of vegetation to climate. The vectors of pollen counts at each location are modelled as indirect observations of the unknown responses [1] to the climate at that location. Building upon the notation already used in this section, the responses are labeled X There is a vector of X responses at each point in the climate space, one for each taxon. Each individual taxon then has a set of responses across the climate space referred to as the taxon *response surface*; jointly over all taxa these are denoted by X.

Bayes theorem is used as above to construct the posterior for the response surfaces given the modern data, $\{Y^m, L^m\}$.

$$\pi(X|Y^m, L^m) = \frac{\pi(X)\pi(Y^m|L^m, X)}{\pi(Y^m)} = \frac{\pi(X)\pi(Y^m|L^m, X)}{\int_X \pi(Y^m|L^m, X)\pi(X)dX} \tag{1.2}$$

The integral in the denominator is typically not tractable analytically. In Haslett et al. (2006) numerical integration was performed approximately using a Metropolis-Hastings Markov Chain Monte Carlo algorithm.

Inverse Problem

The second stage of the Bayesian inference procedure is the calculation of posterior probability distributions on the unknown palaeoclimates L^f, given the posterior for the latent surfaces $\pi(X|Y^m, L^m)$ (derived in the first stage of the model) and the fossil pollen counts Y^f. This is the inverse problem (also known as multivariate non-linear calibration; ter Braak discussion of Haslett et al. (2006)).

Sampled responses X are passed to an MCMC algorithm for sampling from the posterior for palaeoclimate L^f given fossil pollen Y^f and the surfaces X.

[1] Response here refers to the unobservable response of vegetation to climate only; it is the propensity to contribute pollen to the pollen assemblage. The pollen counts are then an indirect observation of this response, perturbed by non-climatic environmental conditions and random variation.

$$\begin{aligned}
\pi(L^f|data) &= \pi(L^f|Y^m, Y^f, L^m) \\
&= \int \pi(L^f, X|Y^m, Y^f, L^m)dX \\
&= \int \pi(L^f|X, Y^m, Y^f, L^m)\pi(X|Y^m, Y^f, L^m)dX \\
&= \int \pi(L^f|X, Y^f)\pi(X|Y^m, Y^f, L^m)dX \qquad (1.3)
\end{aligned}$$

As the fossil pollen counts alone (without knowledge of the climate at which they occurred) contribute little, or even no, information to the posterior for response surfaces given data, $\pi(X|Y^m, L^m, L^f)$ is approximately equal to $\pi(X|Y^m, L^m)$:

$$\int \pi(L^f|X, Y^f)\pi(X|Y^m, Y^f, L^m)dX \simeq \int \pi(L^f|X, Y^f)\pi(X|Y^m, L^m)dX \qquad (1.4)$$

The fully Bayesian approach is to solve the left-hand side of Equation (1.4); the right-hand side is an approximation that is common to most inverse problems.

In fact, a positive feedback mechanism may occur if the fossil counts are left in Equation (1.4); removing them may in fact lead to a more accurate fit. This is referred to as "cutting feedback"; Rougier (2008) states that cutting feedback, although technically a violation of coherence, may be presented in terms of best-input. The model is trained using only the data about which the analyst is confident.

Essentially, fitting the responses using the modern training data, for which counts and climates are available, *and* the fossil data, for which counts only are available, may lead to unwanted positive feedback due to the fossil counts. The model training will begin by placing fossil counts in a region of climate space; given this selected region, the response surface appears to fit well. But the response surface was built using those counts in that region. The initial choice has been strengthened, despite the fact that is was an arbitrary choice. Training the model only on the modern data, for which climate is known, is preferred for this reason.

Integration over the latent surfaces was via Monte Carlo integration in Haslett et al. (2006): samples $X_i, \ldots, X_n$ are drawn from the posterior using the first stage (forward problem) and passed to the second stage (inverse problem):

$$\int \pi(L^f|X, Y^f)\pi(X|Y^m, L^m)dX \simeq \sum_{i=1}^{n} \pi(L^f|X_i, Y^f) \qquad (1.5)$$

An alternative to MCMC for this task is proposed in this thesis. Namely, the suite of approximation techniques referred to as INLA are applied and expanded for this purpose. The implications of imposing any new modelling procedures and algorithms on the forward stage are considered primarily in terms of the impact on the inverse stage.

1.2 Computational Challenges

The most pressing challenges encountered in the Bayesian palaeoclimate project to date involve the intensive computations necessary to carry out inference on the parameters of interest. This is due mainly to an over-reliance on the computationally intensive Markov Chain Monte Carlo algorithm. The large number of parameters required in the complex modelling leads to serious concerns about the mixing the algorithm achieves in the unknown parameter space. Linked to this is the problem that convergence is far from assured, even after runs of the order of weeks (see Haslett et al. (2006)). Additional detail in the model is prohibited due to memory and computation concerns.

As discussed briefly in Section 1.1.1, the unobservable response surfaces must be modelled non-parametrically. One way to achieve this is by discretising climate space on a regular grid and modelling the response as a random variable at each node. High resolution is desirable, requiring the use of a very fine discrete grid on the climate variable space. This results in a very large number of latent variables. The paper Haslett et al. (2006) dealt with a model including the order of 10^4 unknowns; and this is for a inference performed on a greatly reduced dataset using a simplified model.

1.3 Overview of Chapters

A brief outline of the research presented by chapter in this thesis follows.

Chapter 2: Literature Review and Statistical Methodology

A brief review of the literature on palaeoclimate reconstruction is presented. Progress towards the standard set by Haslett et al. (2006) is charted and Bayesian statistical methods for inverse inference are summarised. The remaining weaknesses in the current methodology are outlined and the techniques used to overcome these in this thesis are introduced.

Chapter 3: Models with Known Parameters

In order to separate modelling issues from issues of inference in the forward problem, this chapter focusses on models with known parameters. Various model choices and the implications of these choices are presented. Some new statistical models are detailed. The novel contributions of this chapter are the methods for determining the decomposability of large, multivariate models into separate, independent modules and the nested compositional model. Issues related to the decomposition of models are introduced and discussed.

Chapter 4: INLA Inference and Cross-Validation

The single biggest reduction of the computations required for a full Bayesian inference to be performed on the palaeoclimate dataset are due to the approximation techniques presented in this chapter. By approximating the posterior for the unknown variables in the forward problem with a tractable multivariate density, Markov Chain Monte Carlo may be avoided entirely. The posterior distribution for the responses may be expressed analytically and thus issues regarding mixing and convergence are avoided. The approximation error is typically lower than the Monte Carlo error and the computations required are dramatically reduced. This allows for the addition of extra detail to the model and dramatically increases the potential of the application of the procedure to larger datasets, examples of which are explored.

There are many modelling choices made throughout the application in Chapter 6. These choices must be supported and alternatives explored. Cross validation is a useful tool in evaluating the fit of the Bayesian models to the data. This is typically done by computationally intensive Markov Chain Monte Carlo; it requires many repeated runs and, for large problems such as the palaeoclimate reconstruc-

tion project, this may become computationally overwhelming and is therefore not considered. Approximation methods are once again discussed in this context along with further application to the pollen / climate dataset.

Chapter 5: Inference Methodology

An investigation is carried out into the implications of decomposing the counts data vectors and carrying out marginal inferences on each component of the vector sequentially. This is at best identical to a joint inference on all components at once and at worst an approximation to it. The accuracy of the approximation is expressed as a function of several impacting factors and the conditions for exact reproduction of the joint posterior from the marginal posteriors (perfect decomposition) are described.

Details related to performing inference using the techniques already developed in previous chapters are presented. This chapter serves as a platform towards the application to real data in Chapter 6.

Chapter 6: Application: the Palaeoclimate Reconstruction Project

The motivating problem for the research conducted in this thesis is to improve and advance the palaeoclimate reconstruction project. Therefore, a chapter is devoted to applying the work developed in previous chapters to the RS10 pollen and climate dataset. The various approximation algorithms allow for a richer modelling of the forward problem (the response of vegetation to climate) than was previously possible. The modelling of zero-inflation represents a fundamental change in the model. Practical issues regarding the application of the new model and use of the approximations are identified and discussed. Results are presented and compared with the results derived from previous approaches. A fast cross-validation methodology for the inverse problem is central to this chapter.

Chapter 7: Conclusions and Further Work

The results from the preceding chapters are summarised and discussed. An appraisal of the work to date is conducted and outstanding issues and challenges are identified. Solutions to these remaining challenges are suggested and alternative methodologies

briefly outlined.

1.4 Research Contributions

The following is a statement of the main contributions to the palaeoclimate reconstruction project by the author as presented in this thesis:

1. Investigation is conducted into the accuracy lost in sequential modelling of individual plant taxa responses to climate.

2. Nested compositional counts models for the palaeoclimate dataset are introduced. It is demonstrated that knowledge of the nesting structure is crucial to performing accurate inferences.

3. A fast Bayesian inference procedure on the forward stage of the palaeoclimate reconstruction model is demonstrated. This allows for far richer models to be developed and, more importantly, validated.

4. A model for parsimonious modelling of zero-inflation of the counts data that is compatible with the INLA methodology is presented.

5. A fast inverse cross-validation methodology using INLA is developed. This is a novel extension to the technique and is demonstrated through application to the pollen and climate dataset.

Chapter 2

Literature Review and Statistical Methodology

In order to set the context of the work in this thesis, a brief palaeoclimate reconstruction literature review is conducted in Section 2.1. Gaps in the existing methodology are identified and solutions developed in this thesis are introduced.

The contributions are relevant to a wider statistical methodology beyond palaeoclimate reconstruction; Section 2.2 discusses Bayesian methods that are relevant to the methodology developed in this thesis. Section 2.4 introduces explicit modelling of zero-inflated counts data. Section 2.5 defines inverse regression and demonstrates the generic challenge of such problems with a simple example. Section 2.6 begins the discussion of how models of this type are evaluated and compared, focusing on inverse problems.

2.1 Palaeoclimate Reconstruction Literature Review

Although the contributions made in this thesis to both statistical modelling and inference are applicable to a variety of problems, it is most natural to set them in the context of the motivating problem of statistical palaeoclimate reconstruction using pollen data.

Throughout the later chapters, existing methodology is referenced as required. Therefore, a brief and focussed review of the palaeoclimate literature only is con-

ducted here in order to motivate and frame the work in this thesis.

Detailed reviews are already available; see ter Braak (1995) for a review of non-Bayesian palaeoecology, Haslett et al. (2006) for a review of Bayesian and non-Bayesian palaeoclimate reconstruction and Bhattacharya (2004) for details on Bayesian inference in inverse problems with a focus on palaeoclimate reconstruction. It is not a worthwhile exercise to reproduce these in detail here; an overview, drawing directly from these and other sources is sufficient. Details may be found in the references.

The outline for the literature review is as follows:

- Section 2.1.1 provides a brief review of non-Bayesian estimation methods in the palaeoclimate literature. As per Haslett et al. (2006), these are referred to as "classical". This section is relies on reference to the existing reviews in ter Braak (1995), Haslett et al. (2006), Bhattacharya (2004).

- Section 2.1.2 deals mainly with the methodology of Haslett et al. (2006). Related Bayesian approaches are also discussed. Challenges and shortcomings in these techniques are identified.

Unfortunately, the terminology used in palaeoclimate statistics has become somewhat confused. ter Braak (1995) categorises non-Bayesian approaches into two distinct paradigms, which he terms "classical" and "inverse". The former refers to regression of ecological data on climate. The latter is vice-versa; hence the label inverse as cause and effect have been inverted. "Classical" reconstruction may be thought of as building a forward (cause implies effect) model and subsequently inverting the model to find cause given effect. This use of "inverse" reconstruction involves the simpler task of regression of cause (climate) on effect (ecology).

Haslett et al. (2006) and Bhattacharya (2004) do not consider the ter Braak (1995) definition of "inverse" modelling and use the term "classical" to refer to all non-Bayesian approaches. "Inverse" modelling in these works refers to the inversion of a forward model, Bayesian or otherwise. "Forward" models are equivalent to the models calibrated on the modern data in the "classical" approach of ter Braak (1995).

This is the terminology adopted here; thus quotation marks for these definitions

of "forward", "inverse" and "classical" will be dropped from here on; the ter Braak (1995) definition of "inverse" will be referred to as classical inverse.

2.1.1 Classical Approach

ter Braak (1995) notes that palaeoclimate reconstruction is a highly non-linear multivariate calibration problem. Although climate reconstruction from modern and fossil pollen is taken as the only worked example, the author notes that the techniques carry over immediately to calibration in other areas of palaeoecology.

He uses the interesting phrase

> "the present day calibration is used to infer the past climate"

to broadly describe the way that all statistical climate reconstruction techniques work. The contribution of this thesis mainly lies in the calibration of such data (spatial, compositional, zero-inflated counts). The focus is in building and assessing the models.

It is worth noting that although Krutchkoff (1967) claims the superiority of this definition of the classical inverse method in predictive power, ter Braak (1995) shows that this approach is only slightly better when samples are from a large central part of the distribution of the training set. The inversion of the forward model is considerably better at the extremes. The classical inverse method also treats each climate variable separately and independently; a surprising and illogical model. In the Bayesian context, it is more natural to build forward (cause implies effect) models and invert using Bayes rule.

The classical palaeoclimate modelling approach may be split into three approaches:

1. Response surfaces; polynomials and non-parametric
2. Analogue method; k nearest neighbours
3. Least squares type methods in classical inverse sense

The second two are classical inverse methods and are not considered further (the third method is a direct calibration of climate on pollen). The response surface method is the closest in spirit to the approach introduced in the Bayesian sense by

Haslett et al. (2006) and developed here. Response surface methods typically use least squares based methods to regress pollen on climate; this relationship is then inverted to produce inference on fossil climate given pollen. Bartlein et al. (1986) used cubic polynomials in two climate dimensions fitted to observed percentages of eight pollen types. The authors encountered two difficulties with their approach:

1. some pollen type exhibited multimodal responses
2. the polynomials lacked flexibility and behaved strangely at the edge of the sample climate space.

Both of these problems were addressed through switching from fitting cubic polynomial response functions to non-parametric responses. Prentice et al. (1991) used local weighted averaging to fit smooth non-parametric surfaces to the data. This technique has since been followed by Huntley (1993) and others and is the closest non-Bayesian equivalent to the model of Haslett et al. (2006).

This method posed the question of what to do with the problem of multiple modern analogues. In fact, this problem is common for inverse problems (see Section 2.5.1). In the method of Allen et al. (2000), the locations in climate space of the ten "nearest" response surface to the compositional fossil vector were averaged. This was an attempt to provide a single location as the most likely reconstructed climate. However, it can be a *most* unsatisfactory approach; in the simplest example, a plant type that is abundant in the centre of climate space will send the signal "not close to centre" when the fossil record has low pollen counts of this type. The ten nearest response surface values will come from the edges of climate space. Averaging these ten locations in climate space will then reconstruct the centre; i.e. the very place that the signal most strongly rejects!

2.1.2 Bayesian Approach

A Bayesian approach offers a solution to the above problems. Uncertainty is handled in a consistent manner and full posterior distributions on random variables of interest may be summarised in any way desired. So, for the above example, the posterior distribution would be multimodal with lowest probability assigned to the area in

which the pollen type is scarce; an honest assessment of belief in light of the low signal.

The Bayesian paradigm (Section 2.2) has been applied to palaeoclimate reconstruction; however, the literature is "very small and scattered" (Haslett et al. (2006)). The first detailed Bayesian methodology comes from a series of papers by a group in the University of Helsinki (Vasko et al. (2000), Toivonen et al. (2001) and Korhola et al. (2002)). However, they work with a single climate variable and use a unimodal response with a functional form, invoking *Shelford's law of tolerance*, which states that a species thrives best at a particular value of an environmental variable (optimum) and cannot survive if this variable is too high or too low.

Such a response model is inappropriate for many applications of ecology model. For example, Huntley (1993) shows that, for pollen data, multimodal responses in several climate dimensions are common. This is a result of species indistinguishably; most pollen spore types represent several species or even an entire genus.

More recent Bayesian work by Holden et al. (2008) also invoke Shelford's law. This allows them to avoid MCMC based inference. In that paper, zero-inflation of the data is explicitly modelled; presence and abundance when presence are modelled as functions of a single underlying spatial process. This model is related to the model of Salter-Townshend and Haslett (2006).

Haslett et al. (2006)

Recognizing the issue with multimodal responses, Haslett et al. (2006) applied the non-parametric response surfaces approach of Huntley (1993) in a Bayesian context. A 50×50 regular grid was employed across a two dimensional climate space on which modern pollen data were placed; a Gaussian Markov prior on the non-parametric responses defined on this grid ensured the smoothness of the latent responses. This created a model flexible enough to deal with any type of smooth response function.

Although only a subset of the data was examined (14 taxonomic groups were selected by expert opinion from the total set of 48 taxa), the non-parametric approach led to around ten thousand latent random variables. As the posterior for these parameters is only available up to a normalising constant, sampling algorithms (Section 2.2.2) were employed to sample from the posterior. Empirical summary

statistics were then used in lieu of theoretical ones.

Computation was found to be the main challenge of the methodology; this in turn led to restrictions on both the complexity of the model and, more importantly, in the validation procedures used to test and compare models. Due to these shortcomings, the paper was presented as a "rather detailed proof of concept" Haslett et al. (2006). A section on issues deferred details some of the shortcomings and the printed discussion with the paper addresses several others.

The work contained in this thesis seeks to address some of these difficulties. Alternative inference techniques are employed, novel to the problem of palaeoclimate reconstruction, in place of the computationally overwhelming sampling algorithms used in Haslett et al. (2006). These techniques yield a normalised posterior on all parameters in closed form; see Section 2.3 for an introduction and Section 4.1 for full details. Assumptions made necessary by computational concerns may be relaxed, leading to a richer model posterior and the availability of more rigorous testing.

The final paragraph of the rejoinder from the authors of Haslett et al. (2006) to the contributed written discussion of the paper ended as follows:

> "Zero inflation is a particular challenge ... There may well be sampling procedures for the [parameters] that are more efficient than simple random sampling. In short, there remain many methodological challenges."

In fact, while the authors acknowledge the need to treat the zero counts specially, the model they employ is one for overdispersion only, once again sacrificing model sophistication to computational efficiency.

The avoidance of intensive sampling algorithms allows for more sophisticated models to be developed. In particular, this thesis presents a new model for spatial zero-inflated counts data. The new model is flexible, yet simple. It offers a far more satisfactory account for the extra zeros in the data, yet remains parsimonious.

Model validation did not play a big role in Haslett et al. (2006); leave-one-out cross-validation was presented as a focussed evaluation of the model's capability to reconstruct climate. A counts vector plus climate space location pair is left out of the modern dataset. The model is trained on the remaining data and the left-out climate is reconstructed using the trained model and the left-out counts vector. This is repeated for each data pairing and summaries of the ability of the model to

"predict" the data give a measure of model fit.

However, this would require fitting the model several thousand times. With running times of several weeks, after which the authors concede that "convergence to the correct posterior is far from assured", repeating the procedure even a dozen times is undesirable to say the least. Therefore, the authors use an approximate cross-validation shortcut; the model, as fit to the entire dataset is used as an approximation to the fit for each left-out point.

In contrast to this, constructing closed form posteriors, using new closed form techniques requires only a few minutes of run time. The development of these techniques is not a contribution in this thesis although application to the area of palaeoclimate reconstruction is novel. Re-fitting the model for each left-out point is now a realistic exercise. Another contribution in this thesis is to quickly correct the entire fitted model to account for leaving out a datapoint, rather than re-fit the model, thus achieving a fast inverse cross-validation.

2.2 Relevant Bayesian Methods

The Bayesian analyst is concerned with learning from a dataset about some unknown parameters. In the Bayesian framework, these parameters are treated as random variables and prior probability distributions are placed on these parameters. These reflect the analyst's beliefs before seeing the data; they can be subjective and informed by personal and / or expert opinion, informed by previous analysis of other datasets or totally uninformative, reflecting a complete ignorance or lack of belief.

The data are modelled using a likelihood function. This is a probability distribution for the data, given the parameters. Using Bayes rule (given in Equation (1.1)), the prior and likelihood are multiplied to give an un-normalised posterior. This posterior, once normalised, gives a probabilistic distribution on the updated beliefs in light of the data. All useful summaries of knowledge subsequent to observation of the data may be calculated directly from the posterior distribution.

One of the main advantages of using Bayesian methodology is that uncertainty in the data and the parameters can be treated in a consistent way. Existing beliefs can be built in to the priors so that the posterior reflects not only the information

carried in the data but, for example, expert opinions too.

2.2.1 Bayesian Hierarchical Model

The general type of model considered in this thesis is a Bayesian hierarchical model (see Bernardo and Smith (1994) chapter 4). Hierarchical models have two or more levels of dependency. The hyperparameters θ specify the distribution of the latent parameters X which in turn specify the parameters of the likelihood functions for the data Y (this notation will remain consistent throughout). The hyperparameters themselves may in turn be modelled as random variables with a hyperprior.

$$\begin{aligned} Y &\sim \pi(Y|X) \\ X &\sim \pi(X|\theta) \\ \theta &\sim \pi(\theta) \end{aligned} \tag{2.1}$$

The level of data is called the first level, the parameters of the likelihood are level two and so forth.

2.2.2 Markov Chain Monte Carlo

Normalisation of the posterior is one of the primary challenges to implementation of the Bayesian method. In recent years, the use of Markov Chain Monte Carlo (MCMC) methods has become almost ubiquitous in Bayesian statistical inference, due largely to the availability of cheap and powerful computing resources (Gelfand and Smith (1990)). One common algorithm for performing MCMC based inference is the Metropolis-Hastings rejection sampling algorithm.

Iterative Metropolis-Hastings algorithms (introduced in Metropolis et al. (1953) and generalised in Hastings (1970)) generate a Markov chain of samples from any target probability distribution (such as the posterior in Equation (2.4)). The samples of the Markov chain can then be used to form any desired summary of the target distribution. The desired distribution only needs to be known up to a proportionality constant and therefore the denominator in Equation (2.4) is not required.

A very simple summary is provided here; for details of the Metropolis-Hastings and other MCMC algorithms, see Gilks et al. (1996). Suppose a target distribu-

tion has density $f(X)$. Then, given a sample value of X_t, a proposed value X' is generated from a pre-specified proposal density $q(X'|X_t)$ and then accepted with probability $\alpha(X_t, X')$, given by

$$\alpha(X_t, X') = \min\left\{1, \frac{f(X') \quad q(X_t|X')}{f(X_t) \quad q(X'|X_t)}\right\} \tag{2.2}$$

If the proposed value is accepted, the next sampled value X_{t+1} is set to X'. Otherwise, X_{t+1} is set to X_t. If a symmetric proposal distribution is used (for example a random walk or an independence sampler), then the q terms in Equation (2.2) cancel. Looping over this procedure n times produces a Markov chain of samples $X_1, \ldots X_n$ from $f(X)$, after convergence.

Constructing the algorithm is usually not difficult. Challenges are encountered when attempting to construct an algorithm that will return enough independent samples from the posterior to facilitate accurate inferences. MCMC methodology is plagued by the following issues.

1. Mixing: Due to the frequently high dimensionality of the space of parameters of interest a large number of independent samples is required to adequately describe the posterior distribution. The variables are often strongly dependent on each other and therefore sequential samples from the Markov chain are highly correlated. One-at-a-time updates suffer from poor mixing due to these correlations, unless an efficient multivariate proposal distribution can be constructed to perform joint updates. Mixing refers to the ability of the algorithm to explore the full support of the target distribution. Independence samplers for high dimensional spaces will lead to high rejection rates in the Metropolis-Hastings algorithm, greatly reducing the number of effective samples.

2. Convergence and burn-in: The Markov chain requires an initial set of values. A set of burn-in iterations of the algorithm is necessary to ensure convergence to the stationary distribution of the chain (the target distribution). This burn-in period can be very long and for high dimensional parameter spaces, convergence is difficult (if not impossible) to assure. Gilks et al. (1996) provides details.

In most practical problems a mixture of intuition, experience and *ad hoc* methods are used to determine the length of an MCMC run required to generate a sufficient sample from the posterior. This is particularly challenging for the forward stage of the palaeoclimate reconstruction project due to the non-parametric modelling approach which necessitates a very large number of highly correlated variables.

2.2.3 Directed Acyclic Graphs

A directed acyclic graph, or DAG, is a powerful graphical tool for model building and illustration. A graph with directed arcs and no directed cycles is used to represent the model. The direction of the arcs gives the conditional dependencies.

Hierarchical models are most easily expressed using a DAG, where arrows between each variable denote dependence. If, for example, a variable a depends on another variable b in a statistical model, then the DAG for this is simply $a \to b$.

Throughout this thesis, the structure of all models considered have an overall DAG corresponding to

$$\theta \to X \to Y \tag{2.3}$$

Each of the three levels may consist of multiple variables with extra dependency structure, suppressed in this simplified graph.

Bayesian inference procedures, such as those listed below, allow for inference on the distribution of the parameters given the data via Bayes rule:

$$\pi(X, \theta|Y) = \frac{\pi(Y|X, \theta)\pi(X, \theta)}{\int_{X,\theta} \pi(Y|X, \theta)\pi(X, \theta)d(X, \theta)} \tag{2.4}$$

Bayesian Networks are probabilistic DAGs whose nodes represent random variables. The directed arcs of the DAG denote the conditional dependencies of the model represented. For example, Figure 2.1 shows a DAG for a hierarchical model;

2.2.4 Gaussian Markov Random Fields

It is common for the latent parameters to be modelled as a latent multivariate Gaussian field, particularly in spatial statistics (Banerjee et al. (2003), Finkenstdt et al. (2006)). This allows for a stochastic dependence between the latent parameters

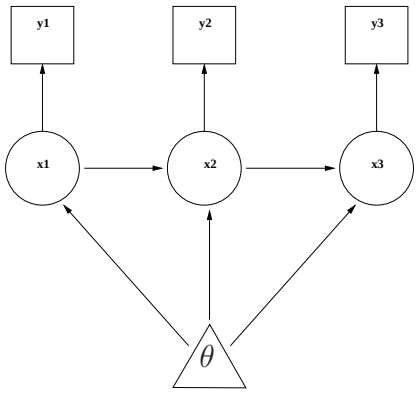


Fig. 2.1: **An example directed acyclic graph, or DAG. The entire hierarchical model is specified graphically; trivariate X depends on θ and each element is linked with the one to the right so that the model of X is, e.g. auto-regressive of order one, given θ. Data Y are observational modes that depend exclusively on the latent parameters X.**

X to be incorporated into the model. One or more of the set of hyperparameters may be used to model the degree of covariance between these latent variables.

For a vector X defined in a discrete location space, a labeled graph $\mathcal{G} = (\mathcal{V}, \omega)$ defines the Markov structure of X. $\mathcal{V} = \{1, \ldots, n\}$ indexes the locations and ω is the set of edges (dependency connections from one node to another) for each node of the graph. There is no edge between nodes i and j iff $x_i \perp x_j | x_{-\{i,j\}}$.

Definition 1 *$x = \{x_1, \ldots, x_n\}^T$ is a **GMRF** w.r.t. a labeled graph $\mathcal{G} = (\mathcal{V}, \omega)$ with mean μ and precision matrix Q iff its density has the form*

$$\pi(x) = (2\pi)^{-\frac{n}{2}} |Q|^{\frac{1}{2}} exp\left(-\frac{1}{2}(x-\mu)^T Q (x-\mu)\right) \tag{2.5}$$

and

$$Q_{ij} \neq 0 \Leftrightarrow \{i,j\} \in \omega for\ all\ i \neq j \tag{2.6}$$

Assigning a Markov structure to the latent field renders both the prior and posterior for the latent parameters to be a Gaussian Markov Random Field (GMRF) w.r.t. a graph $\mathcal{G}$. The reason for using a Markov structure, as opposed to defining a variogram in continuous space, is for computational savings associated with the sparseness of the matrices required; see later in Section 4.1.

If the precision (inverse covariance) matrix for the latent field is sparse, then fast numerical algorithms may be employed; details of this are described in Section 4.1. These Gaussian Markov random fields model the response surfaces of previous sections as stochastically smooth across the location space.

If each node of the graph has an edge to all other nodes then the graph is said to be *fully connected*. Assigning a regular Markov structure to the graph breaks many of the edges resulting in a sparse precision matrix.

Discrete and Finite Space

The use of Markov random fields requires the location space to be defined on a discrete grid. Use of a fine grid blurs the distinction between discrete and continuous space. The data for locations may then be shifted to the nearest gridpoint or left as continuous and calculations at these locations may be evaluated cheaply as weighted averages of the values at the surrounding gridpoint values.

Intrinsic GMRFs

Intrinsic GMRFs are defined by an improper log-density. No mean is specified and the the precision matrix cannot therefore be inverted to give the covariance matrix.

Intrinsic GMRF priors are often used for the parameters describing the latent surfaces. This allows for the specification of prior beliefs on the smoothness of the surfaces without specifying a prior mean.

Following Rue and Held (2005):

Definition 2 *Let Q be a symmetric, positive semi-definite matrix with rank $n-k > 0$, where $k > 0$ is the dimension of the null space of Q. $x = \{x_1, \ldots, x_n\}^T$ is an* ***improper GMRF*** *of rank $n-k > 0$ with parameters (μ, Q) if its improper density is*

$$\pi(x) = (2\pi)^{-\frac{(n-k)}{2}} (|Q|^*)^{\frac{1}{2}} exp\left(-\frac{1}{2} x^T Q x\right) \tag{2.7}$$

where $|Q|^*$ is the generalized determinant of Q (the product of the non-zero eigenvalues).

Intrinsic GMRFs are improper; the precision matrices are not of full rank and cannot therefore be inverted to give a covariance matrix (see Rue and Held (2005)

Chapter 3). In fact, the precision matrix for an intrinsic GMRF does not formally exist, however following Rue and Held (2005) the $n \times n$ matrix Q with rank $n - k$ (and $k > 0$) is referred to as the precision matrix of the intrinsic GMRF.

An intrinsic GMRF of k^{th} order is an improper GMRF of rank $n - k$, where $\sum_j Q_{ij} = 0$ for all i. Hence, the conditional mean of x_i is the weighted mean of its neighbours, but has no specified overall level.

Random Walk

A convenient prior on a vector X whose indices are one dimensional may be derived from the random walk. For example, the first order random walk in one dimension is constructed from independent increments of X, defined on n discrete points (nodes on the graph $\mathcal{G}$).

$$x_i - x_{i-1} \overset{iid}{\sim} \mathcal{N}(0, \kappa^{-1}) \tag{2.8}$$

which implies that

$$x_j - x_i \sim \mathcal{N}(0, (j-i)\kappa^{-1}) \tag{2.9}$$

for $i < j$. The full, joint density for X is then derived from its $n - 1$ increments $(\pi(x_1|x_0), \ldots \pi(x_n|x_{n-1})$ where $\pi(x_i|x_{i-1}) \sim \mathcal{N}(x_{i-1}, \kappa^{-1})$ given by Equation (2.8) as (again, following Rue and Held (2005))

$$\begin{aligned} \pi(X|\kappa) &\propto \kappa^{(n-1)/2} exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(Dx_i)^2\right) \\ &= \kappa^{(n-1)/2} exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2\right) \\ &= \kappa^{(n-1)/2} exp\left(-\frac{1}{2}X^T Q X\right) \end{aligned} \tag{2.10}$$

where $Q = \kappa S$, κ is a hyperparameter of the hierarchical model and S is the structure matrix given by

$$S = \begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & -1 & 2 & -1 & & \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 1 & \end{pmatrix} \tag{2.11}$$

Similar structure matrices may be constructed based on higher order random walks in any dimension of space, subject to edge effects. The imposition of a spatial structure models the latent surfaces as stochastically smooth using just a single hyperparameter κ. This hyperparameter controls the degree of spatial smoothing.

Link Functions

The parameters of many likelihood functions, particularly for counts data, require non-negative parameters. Link functions are therefore used to transform the unrestricted latent field variables into positive numbers.

For example, if the data are modelled as Poisson, then the rates (positive) may be modelled as the exponents of multivariate Gaussian distributed random variables.

$$\begin{aligned} y_i &\sim Poisson(y_i; \lambda_i) \\ \lambda_i &= exp(x_i) \\ X &\sim GMRF(X; \mu_X, Q_X) \end{aligned} \tag{2.12}$$

In this example the hyperparameters θ are comprised of the means and precisions of the latent field, μ_X and Q_X.

2.3 Integrated Nested Laplace Approximations

The Bayesian inference problem for models such as the one depicted graphically in Figure 2.1 is to infer posterior (i.e. given the data) probability distributions for the latent variables X and hyperparameters θ. i.e. to find

$$\pi(X,\theta|Y) = \frac{\pi(Y|X,\theta)\pi(X,\theta)}{\int_{X,\theta}\pi(Y|X,\theta)\pi(X,\theta)d\{X,\theta\}} \tag{2.13}$$

which is more conveniently expressed as

$$\pi(X|\theta,Y)\pi(\theta|Y) = \frac{\pi(Y|X,\theta)\pi(X|\theta)\pi(\theta)}{\int_\theta\left(\int_X\pi(Y|X,\theta)\pi(X|\theta)dX\right)d\theta} \tag{2.14}$$

The most common approach is to use MCMC to sample from the posterior for X and θ. This approach, while hugely popular, is not without its drawbacks (Sections 2.1.2 and 2.2.2). New techniques introduced in Rue and Held (2005) and developed further in Rue et al. (2008) offer a fast approximation called Integrated Nested Laplace Approximations (INLA).

Starting with the identity

$$\pi(\theta|Y) = \frac{\pi(X,\theta|Y)}{\pi(X|\theta,Y)} \tag{2.15}$$

Replacing the denominator with a normalised Gaussian approximation evaluated at the mode ($X^*(\theta)$) yields

$$\pi(\theta|Y) \simeq \left.\frac{\pi(X,\theta|Y)}{\tilde{\pi}_G(X|\theta,Y)}\right|_{X=X^*(\theta)} \tag{2.16}$$

This is known as the Laplace approximation for the hyperparameters. The Gaussian approximation for the latent field posterior, $\tilde{\pi}_G(X|\theta,Y)$, is demonstrated in detail in Section 4.1.

The basic procedure for INLA type inference on Bayesian hierarchical models is as follows:

- The posterior for the hyperparameters is approximated using the Laplace approximate in Equation (2.16).
- The posterior for the smooth latent field, given the data and hyperparameters, is approximated by a GMRF at gridded / discrete values of the hyperparameters.
- The approximate marginal posterior for the latent field, given the data only is found by summing over the discrete values of hyperparameters

- If the marginal value for a particular latent parameter (location in the field) is required to a greater degree of accuracy, a Laplace approximation is built using a similar procedure to Equation (2.16).

Full details of how this is achieved and the relative strengths and weaknesses of the method are examined in Section 4.1. It is sufficient here to note that implementation of these new methods is novel in the context of palaeoclimate research. They allow for increased sophistication in the forward model and more rigorous sensitivity analysis and model validation.

Contributions to the actual INLA methodology in this thesis consist of a method for performing fast updates to the *entire* posterior to correct for leaving out data; this has an immediate application in cross-validation in the inverse sense (see Section 2.6.2). Local corrections are sufficient for cross-validation in the forward sense as the location is known in this case.

2.4 Spatial Zero-Inflated Models

Many counts datasets include zero-inflation; there are an excessive number of zero counts. Of particular interest is spatial data that exhibit such an overabundance of zeros. If these zero counts are ignored then information is lost. If zeros are modelled as arising in the same manner as the non-zero data, then statistical inference carried out on the dataset will be biased by them.

There are several methods for modelling data with many of these extra zeros that fall into three broad categories (see Ridout et al. (1998)):

1. Mixed Models

2. Hurdle models

3. Zero-modified distributions

The first technique, mixture models, accounts for zero-mean random effects. The parameters of the likelihoods for the counts are mixed with a distribution centred on zero. The resulting overdispersed likelihood will allow for some of the desired additional probability on zero counts, however, the variance will be increased and

additional probability will also be placed on other counts far from the expected value.

Hurdle models (Mullahy (1986) a.k.a. two-part models Heilbron (1994)) provide for a two part likelihood. The first defines the probability of observing a zero count and the second part models only the positive counts. For example, if the positive counts are modelled using the zero-truncated Poisson then the count y is distributed as:

$$\pi(y) = \begin{cases} \pi_0 & y = 0 \\ \frac{(1-\pi_0)e^{-\lambda}\lambda^y}{(1-e^{-\lambda})y!} & y > 0 \end{cases} \tag{2.17}$$

where π_0 is the probability of observing a zero count.

Unfortunately, the mean of the zero-truncated distribution is dependent on the form of the non-zeros probability. For example, if a Negative-Binomial distribution with the same mean as the above Poisson was truncated at zero then the means of the truncated distributions will differ. This inconsistency will compound any modelling errors and lead to biases in the inferences (Ridout et al. (1998)).

Zero-modified distributions are very similar to hurdle models; the key difference is that the zeros may still arise from the process that generates the positive counts as well as from a zero-only process. For example, the zero-inflated Poisson is given by

$$\pi(y) = \begin{cases} 1 - q + qe^{-\lambda} & y = 0 \\ \frac{qe^{-\lambda}\lambda^y}{y!} & y > 0 \end{cases} \tag{2.18}$$

where $1-q$ is the probability of observing an *essential zero* count; i.e. a count arising from the process that generates only zero counts.

These are also referred to as *structural zeros* in the literature, with zeros arising from the process that also generates the positive counts referred to as *non-essential zeros* or *sampling zeros*.

This is equivalent to the Poisson hurdle model with $\pi_0 = 1 - q + q\text{Likelihood}(0)$. However, this relationship will of course vary with the choice of non-zero-inflated distribution.

The general form for a zero-modified counts distribution is

$$\pi(y) = \begin{cases} 1 - q + qL(0) & y = 0 \\ qL(y) & y > 0 \end{cases} \tag{2.19}$$

where L is the counts likelihood for the non-zero-inflated version of the distribution. It is a mixture of the non-zero-inflated likelihood and a point mass at zero.

These latter are the most flexible class of distributions for modelling zero-inflated counts data and they are the focus of the work presented here; this is because the pollen data are most accurately described by the mixture of a point mass at zero and a counts likelihood that may still return a zero. The term zero-inflated will be reserved for this method of modelling extra zeros from here on.

2.4.1 Single Process Model for Zero-Inflation

A zero-inflated distribution of counts has an extra parameter over the non-zero-inflated version. For spatial problems, modelled non-parametrically, this doubles an already large number of parameters in the model. As computational overhead is already one of the main challenges to Bayesian analysis of such models, it is desirable to reduce the number of free parameters.

If the parameter governing the point mass at zero and a parameter of the not strictly zero counts part (e.g. the mean) are related, then a more parsimonious model may reduce the number of parameters in the spatial model by half.

In the context of hurdle models, Heilbron (1994) calls this a *compatible model*. An analagous model for a zero-inflated Poisson is introduced by Lambert (1992), wherein the log of the Poisson rate is modelled as

$$log(\lambda) = B\beta \tag{2.20}$$

for some covariate matrix B. The probability of an essential zero and is given by

$$logit(1 - q) = -\tau B\beta \tag{2.21}$$

Lambert advocates such a model based on the absence of prior information about the relationship between the two variables. Lambert proposes the use of such models when the covariates affecting the two variables are the same.

Salter-Townshend and Haslett (2006) showed that using such a functional link in such models not only reduces the number of parameters by half, but that in the context of spatial data analysis ignoring such relationships, if they exist, may lead to a substantive loss in accuracy of inference. In that paper, the "probability of potential presence" q is modelled as being equal to the Binomial parameter for zero-inflated Binomially distributed counts.

Positive Power Link

A more flexible model may be readily achieved through the addition of a single extra parameter α. A power law functional relationship such as

$$q = p^{\alpha} \tag{2.22}$$

with $\alpha > 0$ provides a simple and intuitive, yet flexible model. This is the zero-inflation model that is used for the remainder of the thesis.

If dealing with rates λ rather than proportions, a relationship based on a transformation of the rates to the $[0, 1]$ interval is required, such as $q = \left(\frac{\lambda}{1+\lambda}\right)^{\alpha}$

This is related to Lambert (1992)'s model; solving for q in Equations (2.20) and (2.21) gives

$$q = \frac{1}{1 + \lambda^{-k}} \tag{2.23}$$

Equation (2.22) is monotonic for positive α; an increase in p implies an increase in q. This has the effect of limiting the model with the constraint that as the rate or proportion increases, the probability of observing an extra zero decreases. This flexible yet simple model is one of the lesser contributions of the work described in this thesis. Of course, this model should only be applied to data which exhibit such a relationship or when such a feature is desirable in the model. Justification for this model for the motivating pollen and climate dataset is given in Chapter 6 and so this is the model used in the rest of the thesis.

2.5 Inverse Regression

As per Section 2.1, inverse regression may either mean regressing cause on effect directly (referred to here as classical inverse methods) or regressing effect on cause (the forward model) and then inverting the model to provide estimates of cause given effect. The latter is the approach taken here; response surfaces are fitted using modern data on climate and pollen assemblage pairs. This calibrated model is then inverted to predict (or reconstruct) climate given an assemblage for which there is no climatic data.

2.5.1 Non-parametric Response Surfaces

One important aspect of palaeoclimate reconstruction is the fitting and use of response surfaces (Bartlein et al. (1986), Huntley (1993) and Allen et al. (2000)). The essential issues in the Bayesian modelling of these are presented in terms of a simple hierarchical model. The random variation in the observations (that is, the likelihood) is Gaussian, and all precision parameters are taken to be known. For illustration purposes a simple toy model is presented. Initially in this chapter a univariate model is used, subsequently generalising to multivariate cases. In later chapters, assumptions such as known precisions, are relaxed. The distinction between the forward and inverse stages (see Section 1.1) is stated and illustrated. The procedure is critically analysed and an inverse performance metric is introduced.

The basic idea is presented in Figure 2.2. Pollen counts $\tilde{Y} = \{\tilde{y}_j; j = 1, \ldots, 10\}$ on a single plant taxon are available at 10 regularly spaced points having known climates $\tilde{L} = \{\tilde{l}_j\}$. A model is fitted to these training data, represented by the smooth red line with associated uncertainty interval (dashed red line); this is the forward stage. It models the response of a single taxon to changes in one-dimensional climate. Note that response is measured indirectly; it is a latent variable. In the context of the pollen data, response is the propensity to produce pollen, as a function of climate.

A new count $\tilde{y}_{new}$ is introduced and inferences are made on the associated unobserved climate l_{new}; this is the inverse stage. The figure presents two examples of $\tilde{y}_{new}$. The model adopted is such that for one of these the inference on l_{new} is represented by a unimodal density and for the other the density is bimodal. This

potential mutlimodality is a consequence of the non-monotonic shape of the response surface.

The forward fitting stage is a form of non-parametric regression, in which the only requirement is that the response surface is smooth. A Bayesian approach involving a Gaussian process prior is implemented. In Section 2.5.2 a simple example is used to present the details. These are trivial if, as assumed there, the variance parameters are known and the likelihoods are Gaussian. The inverse stage, even for this toy model, is not trivial. Nevertheless, for this model, with known parameters, it is simple to compute.

This is formalised as follows, for each new count, where the Gaussian random function $X(L)$ models the response surface:

$$\begin{aligned}
\pi(l_{new}|\tilde{Y},\tilde{L},\tilde{y}_{new}) &= \int \pi(l_{new},X|\tilde{Y},\tilde{y}_{new},\tilde{L})dX \\
&= \int \pi(l_{new}|X,\tilde{Y},\tilde{y}_{new},\tilde{L})\pi(X|\tilde{Y},\tilde{y}_{new},\tilde{L})dX \\
&= \int \pi(l_{new}|X,\tilde{y}_{new})\pi(X|\tilde{Y},\tilde{y}_{new},\tilde{L})dX \\
&\simeq \int \pi(l_{new}|X,\tilde{y}_{new})\pi(X|\tilde{Y},\tilde{L})dX \\
&\propto \int \pi(\tilde{y}_{new}|X,l_{new})\pi(l_{new})\pi(X|\tilde{Y},\tilde{L})dX \qquad (2.24)
\end{aligned}$$

There are a number of interesting features of such a problem, many of which are not immediately obvious. These are most usefully demonstrated via investigation of an example toy problem, which is simple yet similar in spirit to the palaeoclimate reconstruction problem. This also serves to introduce some modelling choices which are retained throughout much of this thesis.

2.5.2 Toy Problem Example

Counts data $\tilde{Y}$ are available at N_d locations $\tilde{L}$. The response $X = X(L)$ is unobserved and treated here as a random function, defined on a fine regular grid (L) of 100 points across the location space. The interest here is its conditional distribution given the training data. The Bayesian formulation of the problem is then

$$\pi(X|\tilde{Y},\tilde{L}) = K\pi(\tilde{Y}|X,\tilde{L})\pi(X) = K\prod_{j=1}^{10}\pi(\tilde{y}_j|X(\tilde{l}_j))\pi(X) \qquad (2.25)$$

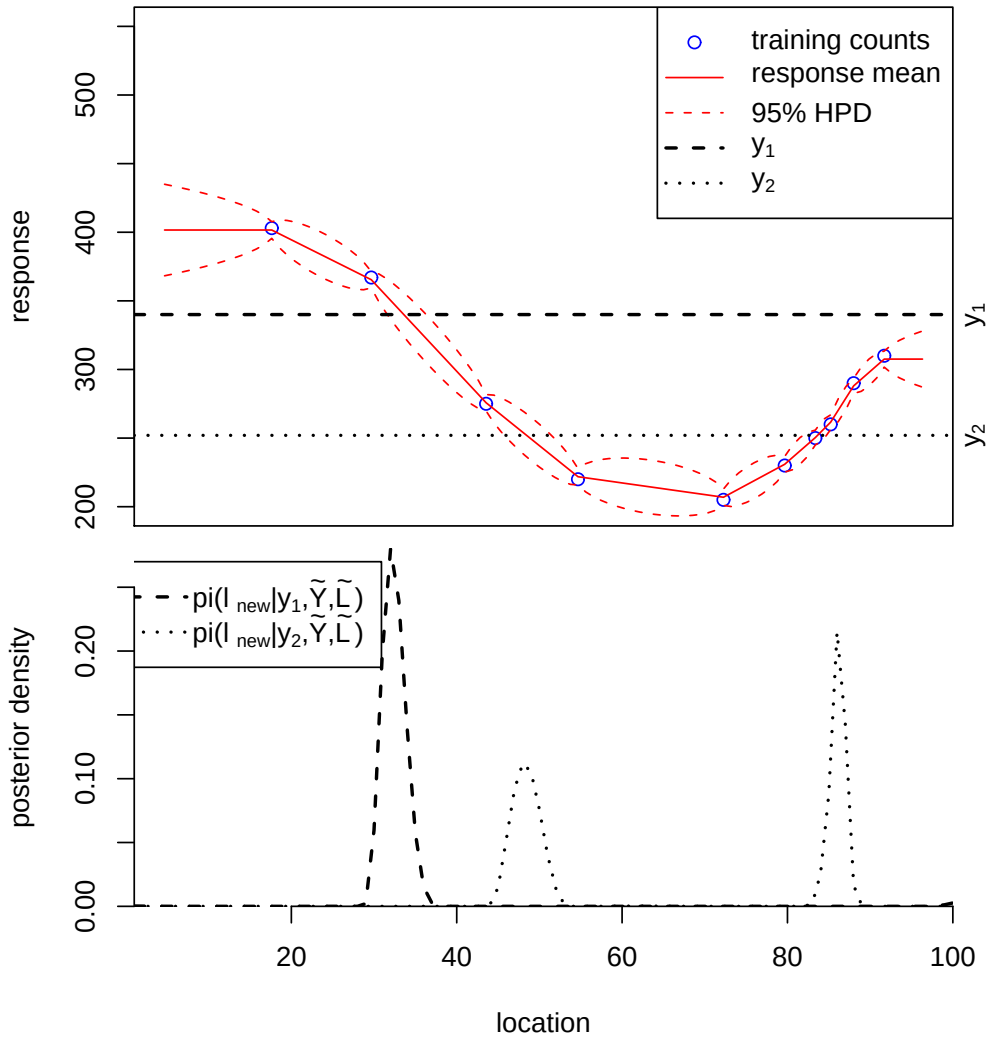


Fig. 2.2: **The basic concepts of the response surface methodology are illustrated. The top plot shows the counts data and results of the forward stage of the inference. The posterior mean and 95% highest posterior bounds are plotted against climate / location. The lower plot shows the posterior densities for climate given two new counts (inverse stage) and the forward stage results.**

where X is the latent response and K is the normalising constant. The data are conditionally independent, given the latent surface. In such Bayesian approaches K is source of much technical difficulty, even when $\pi(X)$ is Gaussian. However when the likelihood is also Gaussian, as here, K is available analytically. For this example, the surface $X(l)$ is the likelihood mean at $l \in L$ and σ_Y^2 is the variance. The prior $\pi(X)$ and the likelihood $\pi(\tilde{Y}|X, \tilde{L})$ are thus conjugate, leading to a Gaussian posterior, provided the prior and likelihood precision matrices, Q_X and Q_Y respectively, are known.

Across all gridpoints (L) in the location space, the likelihood contributes

$$\pi(Y|X) \propto exp\left(-\frac{1}{2}(X-Y)^T Q_Y (X-Y)\right) \tag{2.26}$$

where

$$Y(L) = \begin{cases} \tilde{Y} & L = \tilde{L} \\ 0 & L \neq \tilde{L} \end{cases} \tag{2.27}$$

and the likelihood precision matrix Q_Y is diagonal and of dimension $N_L \times N_L$, where N_L is the number of gridpoints. The diagonal entries are:

$$Q_Y[j,j] = \begin{cases} \frac{1}{\sigma_Y^2} & L_j \in \tilde{L}_j \\ 0 & L_j \notin \tilde{L}_j \end{cases} \tag{2.28}$$

The prior on X is an intrinsic Gaussian Markov random field (GMRF), given by

$$\pi(X) \propto exp\left(-\frac{1}{2}(X)^T Q_X (X)\right) \tag{2.29}$$

with the precision matrix Q_X given by

$$Q_X = \kappa \begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & -1 & 2 & -1 & & \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 1 & \end{pmatrix} \tag{2.30}$$

The model has two scalar parameters; the positive scalar prior precision parameter κ and the likelihood variance σ_Y^2. The higher the κ value used, the greater the smoothness of the latent surface $X(L)$. The decision to model the surface on a regular grid rather than specifying a continuous model defined at the datapoints is due to the desirable properties of GMRFs as discussed in Section 4.1 and is not discussed here.

The Markov property is inherited by the posterior which is a multivariate Gaussian with mean and precision matrix given by

$$\begin{aligned} \mu &= (Q_X + Q_Y)^{-1} Q_Y Y \\ Q &= Q_X + Q_Y \end{aligned} \tag{2.31}$$

Using this analytical form for the posterior of the latent surface given the data, the inverse stage posterior of Equation (2.24) may be computed numerically. The locations are discrete so the posterior for unknown location given a new count is defined only on a finite number of possible gridpoints. The posterior is therefore a probability mass function and normalisation is provided by rescaling the unnormalised product of the prior and likelihood functions such that the total is unity. A uniform prior $\frac{1}{N_L}$ is imposed here and the likelihood of the new count $\tilde{y}_{new}$ at any given location L is $N(X(L), \sigma_Y^2)$. The integral over the unidimensional latent surface is performed analytically. Sample calculations are provided in Table 2.1 for a given $\tilde{y}_{new}$.

Table 2.1: **Sample calculations from the inverse stage of the toy problem given a new count of 340 and the forward stage results shown in Figure 2.2.**

location	28	29	30	31	32	33	34	35	36	37
prior	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
likelihood	$3.58e^{-9}$	$3.95e^{-4}$	$9.70e^{-3}$	$3.22e^{-2}$	$4.55e^{-2}$	$3.90e^{-2}$	$2.28e^{-2}$	$9.32e^{-3}$	$2.55e^{-3}$	$4.16e^{-4}$
product	$3.58e^{-11}$	$3.54e^{-6}$	$9.70e^{-5}$	$3.22e^{-4}$	$4.55e^{-4}$	$3.90e^{-4}$	$2.28e^{-4}$	$9.32e^{-5}$	$2.46e^{-5}$	$4.21e^{-6}$
posterior	$2.20e^{-8}$	$2.18e^{-3}$	$5.96e^{-2}$	$1.98e^{-1}$	$2.80e^{-1}$	$2.40e^{-1}$	$1.40e^{-1}$	$5.73e^{-2}$	$1.51e^{-2}$	$2.59e^{-3}$

The model parameters κ and σ_Y^2 effect the forward and therefore the inverse stage results. Inferences on the same dataset as in Figure 2.2, but using different

values of κ and σ_Y^2 are presented in Figure 2.3.

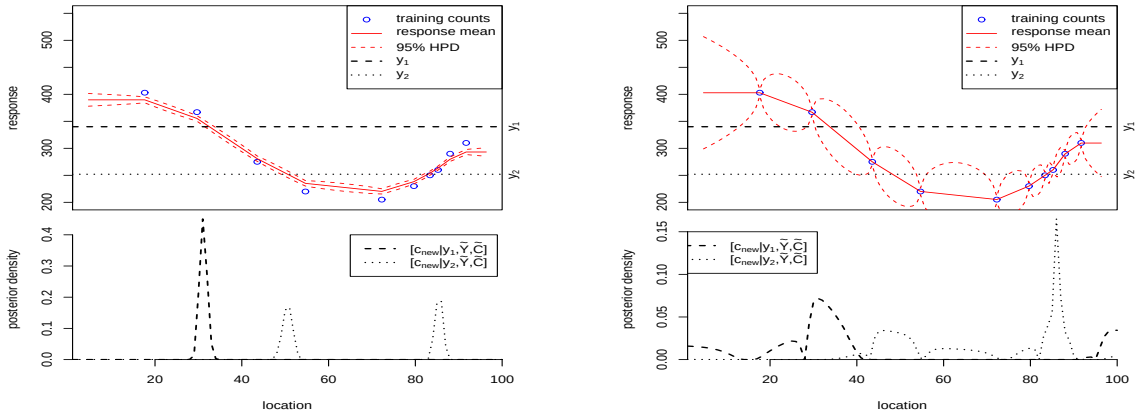
Impact of Model Parameters

Figure 2.3 illustrates the effect of varying the model parameters κ and σ_Y^2. Comparisons with the parameters used and results obtained for Figure 2.2 are made here. In Figure 2.3(a), κ is an order of magnitude larger. This induces a greater degree of smoothness in the latent surface and tightens the bounds of the 95% highest posterior density (HPD) region of the forward stage.

In Figure 2.3(b) κ is an order of magnitude smaller. The surface parameters linearly interpolate the data and uncertainty is high away from the datapoints. This leads to a highly multimodal posterior for the inverse stage. As κ goes to zero (no smoothing), the forward stage posterior tends toward the likelihood. The forward model then informs only at the datapoints and the inverse stage will yield a uniform mass function for new counts not close to training data counts. For new counts close to one or more training counts, the inverse stage posterior will be spiked at the associated training data locations.

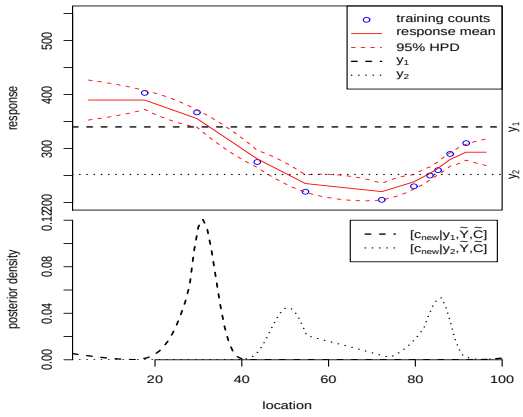
In Figure 2.3(c) σ_Y^2 is an order of magnitude larger. The likelihood density has a larger spread, as does the forward stage posterior. The prior on X begins to dominate the posterior for X. The inverse stage posterior then has a larger variance and the amplitude of the minor mode associated with y_2 grows relative to the major mode. As σ_Y^2 increases, the inverse stage posterior flattens out.

In Figure 2.3(d) σ_Y^2 is an order of magnitude smaller. As there is more data on the major mode of the posterior for location given y_2, this serves to increase that mode; consequently, the minor mode is reduced. As σ_Y^2 decreases, the forward stage posterior becomes dominated by the likelihood at the datapoints. Away from the datapoints, the prior dominates. Inferences on new counts that are close to training data counts will have inverse stage posteriors with sharp peaks at the associated training data locations, $\tilde{L}$.

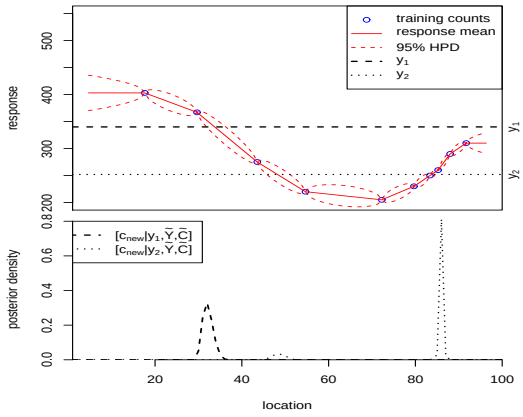


(a)

(b)



(c)



(d)

Fig. 2.3: **Results for inference using the same model and data as in Figure 2.2 but with one of either either κ or σ_Y^2 changed. Compared with the values used in constructing Figure 2.2:**

(a) κ is an order of magnitude larger. This induces a greater degree of smoothness in the latent surface.

(b) κ is an order of magnitude smaller. The surface parameters linearly interpolate the data and uncertainty is high away from the datapoints in the forward stage.

(c) σ_Y^2 is an order of magnitude larger. The likelihood density has a larger spread, as does the forward stage posterior.

(d) σ_Y^2 is an order of magnitude smaller.

2.6 Model Validation

MCMC is the dominant technique for Bayesian inference; as discussed in Section 2.2.2, the main crux of this methodology is computation. Cross-validation is a common technique for assessing the model fit to data. This requires re-fitting the same model for many subsets of the dataset in order to form reference distributions for the left out points. Model evaluation by cross-validation is very unsuited to the brute force approach of simply conducting many MCMC re-runs.

Model evaluation techniques in Bayesian problems are a well documented research area. However, Bhattacharya (2004) observes that "there seems to be no literature in this context [model assessment in inverse problems]". As stated previously, inverse problems refers to studies in which the forward model (cause implies effect; input drives output) is inverted, with the objective of predicting (or reconstructing) input given output.

2.6.1 Inverse Predictive Power

Ultimately, the objective of the investigation is to make accurate inferences on unobserved climates given fossil counts. Therefore, the performance metrics introduced below focus on the inverse stage of the problem. Pairs of 'new' data $(\tilde{y}_{new}, \tilde{l}_{new})$ are generated and the ability of the model to predict $\tilde{l}_{new}$ given $(\tilde{Y}, \tilde{L})$ pairs and $\tilde{y}_{new}$ is evaluated. The predictive distribution here is simply the posterior for the inverse stage, $\pi(l_{new} = \tilde{l}_{new}|\tilde{Y}, \tilde{L}, \tilde{y}_{new}) = \pi(l_{new}|data)$. The performance is summarised by a statistic on a function of the inverse stage posterior $\pi(l_{new}|\tilde{Y}, \tilde{L}, \tilde{y}_{new})$ and the 'unobserved' location $\tilde{l}_{new}$. As the posterior for location is often multimodal and seldom symmetric, simple statistics such as the distance between the modal prediction and $\tilde{l}_{new}$ will be insufficient.

For ease of notation, $\pi(l_{new}|data)$ is denoted $\pi(L)$. The desirable properties of the performance metric $D = D(\pi(L), \tilde{l}_{new})$ are listed here.

- Tends to zero for perfect prediction:
  $\lim_{[\pi(\tilde{l}_{new})\to 1]} D \to 0$
- Non-negativity: $D \geq 0$ for all cases.

One such metric is the expectation of the square of the difference between the new location $\tilde{l}_{new}$ and the predicted location under the posterior for the inverse stage given the new count $\tilde{y}_{new}$. For a location space discretised into N_L gridpoints, this is readily computed as

$$D(\pi(L), \tilde{l}_{new}) = E[||l_{new} - \tilde{l}_{new}||^2] = \sum_{j=1}^{N_L} \pi(l_j)||l_j - \tilde{l}_{new}||^2 \tag{2.32}$$

where $||l_{new} - \tilde{l}_{new}||$ is the distance between l_{new} and $\tilde{l}_{new}$. D is then the mean-squared error of prediction.

If the location space is rescaled to lie between 0 and 1 then this metric will lie between between 0 and 1 (or 0 and $\sqrt{(}2)$ for 2-dimensional location space).

$$g(l) = (l - min(L))/(max(L) - min(L)) \tag{2.33}$$

$$D(\pi(L), \tilde{l}_{new}) = E[(g(l_{new}) - g(\tilde{l}_{new}))^2] = \sum_{j=1}^{N_L} \pi(l_j)(g(l_j) - g(\tilde{l}_{new}))^2 \tag{2.34}$$

This metric will tend to unity as the predictive probability mass function tends to unity at the greatest distance in the location space from $\tilde{l}_{new}$ and will tend to zero as the predictive probability mass function tends to unity exactly at $\tilde{l}_{new}$.

2.6.2 Cross-Validation

Measurement of predictive performance is closely related to cross-validation of the model using the training dataset. In leave one-out-cross-validation, the ability of the model to predict each $\tilde{l}_i$ (rather than a new point), given the remainder of the training data $(\tilde{Y}, \{\tilde{l}_{j\neq i}; j = 1, \ldots, N_d\})$ is assessed. This step is repeated for each datum and a discrepancy measure summarises the validity of the initial analysis. This is in fact cross-validation for the inverse problem, which is discussed in detail in Section 4.2.

Cross-Validation in Bayesian Inverse Regression

One method for speeding up the brute force approach of repeated MCMC samplers is the use of importance sampling. The "saturated posterior" refers to the model

trained on all of the modern data. As leaving out a single point will only have a small effect on this distribution, the proposal density for a new MCMC run uses the saturated posterior as an importance sampler. The importance weights are easily calculable for the forward problem as they are proportional to the correct posterior to saturated posterior ratio. This ratio, having the same prior and likelihood terms, is expressible as the inverse of univariate likelihood of the left out datum, given the sample parameters (see, for example, Gelfand (1996)).

Cross-validation in the inverse sense is a more difficult challenge. The importance weights now typically involve an intractable integration (see Section 4.2 for details). In the inverse case, a prior for the input must also be constructed. This complicates the analysis.

Being the first attempt to address the issue of cross-validation in inverse problems and its applications to model assessment, Bhattacharya and Haslett (2008) provides an important benchmark. In that work, the authors use importance resampling (advocating without replacement) to approximate the posterior distribution of the parameters and a pre-chosen datum given the data minus the left-out datum. The importance weights are quick to compute and proposal densities are constructed to maximize the efficiency of the predictive distribution MCMC sampler. This technique is referred to as Importance Re-sampling MCMC (IRMCMC) for (leave-one-out) cross-validation in inverse problems. The difference between this resampling approach and the forward problem importance sampling is that for the first MCMC run, a selected datapoint is left out and regarded as a random variable. The integrations in the calculation of the importance weights for the resampling stage are no longer intractable. This observation is the key to the procedure (see Section 4.2.2).

An efficient sampling algorithm is thus constructed for the cross-validation. However, multiple sampling runs are still required. Bhattacharya and Haslett (2008) presents an example where brute force MCMC re-runs for each of 62 data sample sites takes 16 hours. IRMCMC achieves comparable, if not better, results in less than 40 minutes. For the example in Haslett et al. (2006), brute-force MCMC replications would take many years.

A new method for cross-validation in the inverse sense is presented and developed

in Section 4.2. Essentially, fast alterations are made to the saturated posterior for the forward model to correct for left out data. As the locations (inputs) are already made to lie on a finite grid, the marginal likelihood may be computed for all possible values of the left out point; thus MCMC is entirely avoided.

In Chapter 6 it is demonstrated that cross-validation in the inverse sense for this same problem is thereby reduced to less than 1 hour using the techniques presented in this thesis that synergize with the INLA method. In fact, that run-time is for a superset of the data with a more sophisticated hierarchical model and formal estimation of several hyperparameters that were necessarily preset ad-hoc in previous MCMC / IRMCMC attempts.

2.7 Conclusions

Palaeoclimate reconstruction is an example of an inverse problem. Existing attempts to infer climates from ecological data involve a trade off between model complexity and speed of inference. The Bayesian framework is preferred due to its ability to honestly model uncertainty by treating unknown parameters as random variables. This is particularly important when inverting forward models to obtain the inverse posteriors.

Sampling based methods, such as MCMC, are the standard methodology in Bayesian inference. Non-parametric models, with high numbers of random variables, may be poorly suited to these computationally intensive methods. Inference is labourious and this negatively impacts the level of complexity of the models. Fast, approximate methods for performing Bayesian inference allows the development of more sophisticated models for the palaeoclimate reconstruction problem. These include, for example, zero-inflated models.

Model comparison and validation, requiring many similar versions of the same models, is an important aspect of any statistical study. Validation of inverse problems occupies only a limited literature. MCMC type inference is unsuitable to this challenge. Existing approximate inference techniques must be extended to meet this challenge. Conversely, the palaeoclimate problem presents an interesting and challenging test of the approximate inference methodology.

2.7.1 Advances in this Thesis

Since the "proof of concept" paper Haslett et al. (2006), the work contained in this thesis to modelling the RS10 dataset may be summarised as follows:

1. Explicit modelling of the zero-inflation in the pollen counts data (Sections 2.4 and 6.3).

2. Estimation of model hyperparameters: i.e. smoothness of the latent responses across climate, degree of overdispersion and zero-inflation power (using Section 4.1.3).

3. Fast inversion of the forward model (due to the discrete / finite climate space).

4. Fast cross-validation in the inverse sense using the extension to INLA in Section 4.2.3.

5. Use of summary statistics for model comparison and validation for the inverse problem as introduced in Section 2.6.1 and Section 3.1.

Chapter 3

Models with Known Parameters

This chapter deals with modelling issues as distinct from any challenges relating to statistical inference of latent fields and hyperparameters. In this chapter, model parameters are taken to be known and inference details for the forward problem are suppressed, to be dealt with in later chapters. Specifically, the forward stage of the model is taken to have all known parameters; inference on the inverse stage is used to assess different forward models.

The novel contributions contained in this chapter relate to model choice. Specifically, the following questions are addressed: Under what circumstances a large, multivariate model, such as those required for the motivating palaeoclimate problem, may be broken down to produce a series of independent, smaller and more manageable inferential tasks? How might one proceed with such a decomposition? How might the validity or accuracy of the decomposition assessed? Finally, when a model may not be decomposed directly, are there augmentations to the data that might facilitate decomposition?

When dealing with highly multivariate datasets, such as the RS10 pollen and climate dataset, several modelling choices present themselves. These consist of choices for modelling the latent parameters of the hierarchical model, the hyperparameters and the choice of likelihood model for the data given the parameters. It is necessary to make clear the motivations and justifications for each of these choices. This requires the use of "model fit" techniques. Cross validation is the tool selected in this work, the details of which are presented in Section 4.2.

Section 3.1 introduces the type of inverse problem investigated in this thesis for

a single spatial process generating counts across locations.

Section 3.2 sets out the motivation for decomposing large, joint models into independent modules. A definition of decomposable models is given and conditions under which models may and may not be exactly disjoint-decomposed are presented. Finally, sources of interaction preventing decomposition are discussed.

A fully-Gaussian case in Section 3.3 allows for the introduction of several key modelling issues in a Normal context. The tractability and familiarity of the multivariate normal model are used to present modelling issues that apply in a wider context of multivariate modelling. Specifically, non-decomposable models are developed in the context of the multivariate normal model.

Departures from normality in Section 3.4 introduce additional issues related to counts data. Novel models for such data are also introduced in this latter section in the form of specialised likelihood functions, such as zero-inflated data models.

Section 3.5 deals with the constrained space associated with compositional data analysis. Some pitfalls of analysis on this space are described. Finally, novel models for specifying complex yet decomposable models for compositional data are presented.

Finally, conclusions are made from the work detailed in the preceding sections. These conclusions are carried forward into the later chapters.

3.1 The Univariate Problem

Returning to the toy problem described in Section 2.5.2, a univariate process varies smoothly across a location space. That section demonstrated the effect of differing model hyperparameters on the inverse problem. The inverse predictive distributions were found to be multimodal due to the shape of the response surface. The shape of the response surface also influences the degree of accuracy to which the inverse problem (prediction of location given count) may be solved.

To recap, the goal is to infer unknown location l given training counts $\tilde{Y}$ with training locations $\tilde{L}$ and a new count y_{new}.

As per Equation (2.24),

$$\pi(l_{new}|\tilde{Y},\tilde{L},\tilde{y}_{new}) \propto \int \pi(\tilde{y}_{new}|X,l_{new})\pi(l_{new})\pi(X|\tilde{Y},\tilde{L})dX \tag{3.1}$$

A flat prior is used for $\pi(l_{new})$. The posterior for location l is therefore proportional to the likelihood for count y at that location. The first task is to find $\pi(X|\tilde{Y}, \tilde{L})$, the details of which are suppressed in this chapter.

In the following examples, the toy data comprises a single smooth response surface which generates univariate counts data. There is a single count at each of 10 sampling locations.

Easy and Difficult Inverse Prediction Conditions

The immediate objective here is to identify the conditions for which inference in the inverse stage of the model is difficult. From the discussion above, the factors which influence the ability of the model to accurately predict climate given a new count and training data are the shape of the underlying surface and the prior and likelihood precision parameters (κ and $1/\sigma_Y^2$).

The performance of the inference will also be greatly effected by the new data presented to the model. Recall, for example, the multimodal posterior arising from the inverse stage inference given a new count of $y_2 = 252$ in Figure 2.2: the inverse inference is placing large probability mass in another location; this must be classed as a poor inference and will occur (to varying degrees; see Figure 2.3) for any model parameters as it is due to the non-monotonic shape of the response function.

Conversely, given a new count of $\tilde{y}_{new} = 450$, the model will always place a unimodal inverse stage posterior on the very left of the location space; this is in fact the only area from which such a high count can have been generated, given the underlying surface. It is contrasting situations such as these that is the interest here; cross-validation and model fit are discussed in later chapters.

It is useful here to explore some interesting and challenging features of the inverse problem. The impact that the shape of the underlying latent response curve has on the posterior for location given count is demonstrated. Three examples are are presented in each of sections 3.1.1 and 3.1.2. Although both sections necessarily use examples given a new count, with unobserved location, Section 3.1.1 is new count value specific whereas Section 3.1.2 uses example response curves that will generate similar challenges regardless of the value of the new count.

3.1.1 Given New Counts Data

Loosely speaking, conditions may be subdivided into 3 categories; easy, medium and hard. These are due to encountering degrees of strongly informative data, weakly informative / uninformative data and misleading data respectively. These three categories are illustrated with examples in Figure 3.1 for fixed values of the model parameters, κ and σ_Y^2.

The easy case refers to a problem that, due to the shape of the response function, delivers a tight, unimodal posterior distribution for the location (climate) given a count. The medium case delivers a diffuse posterior as the fitted response function carries little informtation on location given count. Finally, the hard case delivers a multimodal location posterior with the true location not necessarily located under the major mode. The posterior distribution for location may in fact place much probabilty mass far from the correct location, resulting in misleading predictions.

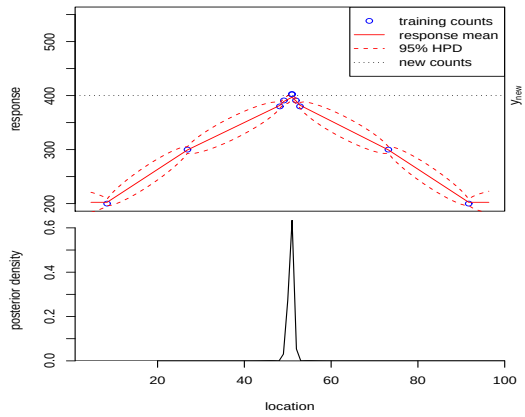
Figure 3.1(c) shows an important result; the performance statistic D (see Equation (2.32)) is not simply an indicator of the strength of the signal provided for the inverse stage but also of the uniqueness. Figure 3.1(c) carries a strong signal at the correct location, however the inverse stage posterior places higher mass at another, distant, location; this results in a poor performance rating.

3.1.2 Given Training Data Only

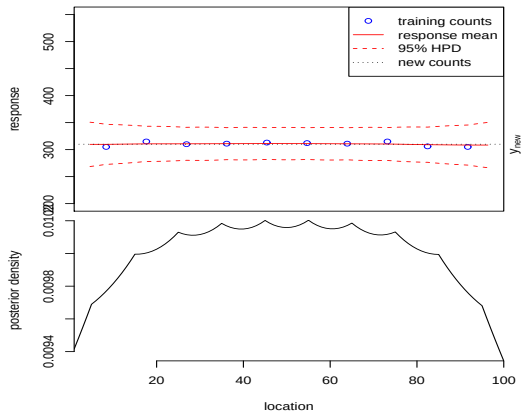
This is closer in spirit to cross-validation than the previous section. The goal here is to categorize training datasets based on the ability of the fitted model to predict locations given arbitrary new counts. Again, the categories are described as easy, medium and hard. Figure 3.2 depicts an example of each case.

3.1.3 Percentage Outside Highest Predictive Distribution Region

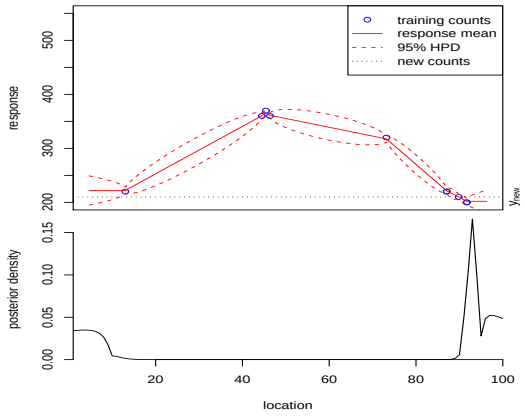
Another cross-validation statistic that will be used throughout this thesis is the number of training data that fall outside the 95% highest posterior predictive distribution. The predictive distribution here is the leave-one-out cross-validation posterior predictive distribution for the location given all other locations and the counts



(a) Easy Case



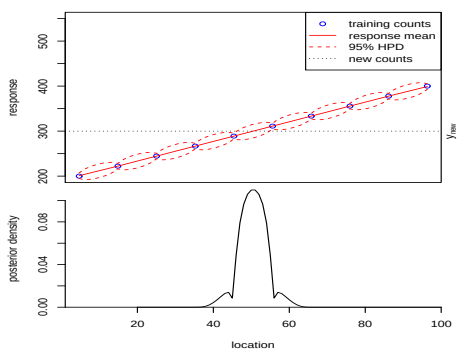
(b) Medium Case



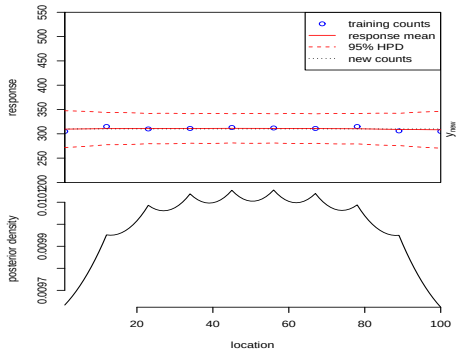
(c) Hard Case

Fig. 3.1: **Forward and inverse stage posteriors for 3 markedly different datasets. The ability of the model fitted to the training data to predict the new location given the new count, $\tilde{y}_{new}$, is severely effected by the training dataset. The new data generated for inference in the inverse stage of the problem is generated from the likelihood given a draw from the posterior for the forward stage. For comparison, values of $\tilde{y}_{new}$, $\tilde{l}_{new}$ and $D(\pi(L), \tilde{l}_{new})$ are provided in a table.**

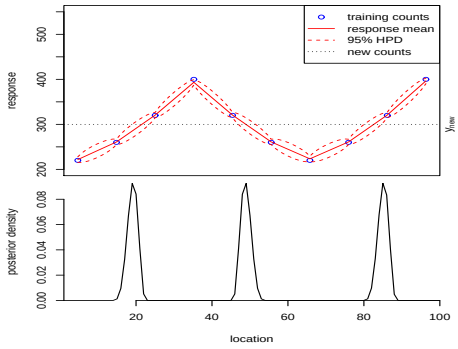
(a)	$\tilde{y}_{new}$ = 410	$\tilde{l}_{new}$ = 50	D = $9.30e-5$
(b)	$\tilde{y}_{new}$ = 310	$\tilde{l}_{new}$ = 50	D = 0.079
(c)	$\tilde{y}_{new}$ = 210	$\tilde{l}_{new}$ = 10	D = 0.516



(a) Easy Case



(b) Medium Case



(c) Hard Case

Fig. 3.2: **Forward and inverse stage posteriors for 3 markedly different datasets. The ability of the model fitted to the training data to predict the new location given any new count, $\tilde{y}_{new}$, is severely effected by the training dataset. The new data generated for inference in the inverse stage of the problem is generated from the likelihood given a draw from the posterior for the forward stage. For comparison, example values of $\tilde{y}_{new}$, $\tilde{l}_{new}$ and $D(\pi(L), \tilde{l}_{new})$ are listed.**

(a)	$\tilde{y}_{new}$ = 300	$\tilde{l}_{new}$ = 50	D = 0.001
(b)	$\tilde{y}_{new}$ = 310	$\tilde{l}_{new}$ = 43	D = 0.086
(c)	$\tilde{y}_{new}$ = 300	$\tilde{l}_{new}$ = 47	D = 0.076

These results are largely independent of the new data $(\tilde{y}_{new}, \tilde{l}_{new})$ supplied to the inverse stage, with the exception that in (b) locations closer to the centre will yield lower D values as the posterior, although vague, will be centred in the correct area. Additionally, (c) depends on whether the new count $\tilde{y}_{new}$ corresponds with 2 or 3 possible locations. For $(\tilde{y}_{new}, \tilde{l}_{new}) = (390, 35)$, D is 0.155.

data.

Definition 3 Δ *is the % of data lying outside the* 95% *highest posterior density region of their inverse predictive density*

If the model fits the data, then the expected value of Δ is 5%. This does not depend on the accuracy of the inverse predictions delivered. When the location space lies on a discrete grid it is not always possible to define a 95% HPD region. For a discretized space, the method for defining the 95% HPD regions is as follows:

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 next highest probability mass is selected and added to the HPD region.
4. Repeat step 3 until the total probability mass of the HPD region is greater than or equal to 95%.

This means that the HPD region contains 95% **or more** of the total probability mass. Therefore, the expected value of Δ is $\leq 5\%$.

The concern in this thesis is the case of more than one response surface, each of which generates its own counts data. The simplest approach to the inverse problem is to perform inference on each set of counts separately for both the forward and inverse stage of the problem. The inverse predictive distribution given all components is then the product of all inverse predictive distributions for each of the components of the counts assemblage. Such a model, expressible as non-overlapping separable parts, is discussed in the next section.

3.2 Disjoint-Decomposable Models

Complex, highly multivariate datasets that require multivariate models pose several challenges of computation and model choice (see also Section 5.1). One approach in dealing with these challenges is to *disjoint-decompose* the problem into disjoint,

independent modules. Each module requires a model to be fitted to a separate subset of the overall dataset and inference may be carried out separately on each subset, i.e. independently from the other subsets.

Definition 4 *Multivariate models which may be expressed exactly as the product of disjoint parts are said to be* ***disjoint-decomposable****.*

This is closely related to independence; a probability model that factorises into a number of (potentially multivariate) independent distributions, with no terms appearing in more than one distribution is disjoint-decomposable. There is no *interaction* between the margins of such models. However, many models that do not factorise in such a manner may be approximately disjoint-decomposable. This may allow for a far simpler inferential approach to be taken, with a post-hoc correction for the decomposition.

The difference between disjoint-decomposable and independence is subtle, but relevant in this work. If a model is comprised of independent modules then it is disjoint-decomposable. However, some models that are not expressible as the product of independent parts may decompose in practice.

The simplest example of this is with regard to compositional data. If counts proportions data Y with sum n are modelled with a Multinomial distribution then it is immediately clear that the data are dependent. However, the data may be modelled as independent Poisson counts and the product of the marginal likelihoods then gives an approximation to the true, joint likelihood. The approximation error can be corrected by dividing by the probability that the total count is n. This probability is available trivially as a Poisson count.

If inference on the parameters P of the Multinomial are modelled with a Dirichlet prior then the posterior is Dirichlet. Decomposing the model into independent Poisson likelihoods with independent Gamma priors yields a product of Gammas as the decomposed-model posterior. To correct the approximation error this product is then divided by the posterior distribution on the sum.

So, given a counts vector Y, constrained to sum to n with a Dirichlet prior with parameters α and a Multinomial likelihood, the Dirichlet posterior may be written as a product of Gammas, scaled by a Gamma distribution on the sum:

$$Dirichlet(P) \equiv \frac{\prod_i^N Gamma(P_i|\alpha_i, \sum_j^N \alpha_j)}{Gamma(1|\sum_j^N \alpha_j, \sum_j^N \alpha_j)} \tag{3.2}$$

$$Multinomial(Y|P,n) \equiv \frac{\prod_i^N Poisson(Y_i|P_i n)}{Poisson(n|n)} \tag{3.3}$$

$$Dirichlet(P|Y) \equiv \frac{\prod_i^N Gamma(P_i|\alpha_i + Y_i, \sum_j^N \alpha_j + Y_j)}{Gamma(1|\sum_j^N \alpha_j + Y_j, \sum_j^N \alpha_j + Y_j)} \tag{3.4}$$

If the model does not disjoint-decompose then the model corresponding to a decomposable version approximates the joint model. Inference on such a model will approximate joint inference on the non-decomposable model. This is dealt with in Chapter 5. The accuracy of the approximation of the decomposable model to a non-decomposable version depends on the level of interaction between the modules. Simple performance checks such as the one described in Section 3.2.2 may be used to determine the legitimacy of decomposing the model.

If the model does not decompose into univariate marginals, decomposition into smaller, more manageable multivariate marginals may still render the inference to be far simpler.

3.2.1 The Marginals Model

A simple and intuitive decomposition of the pollen dataset is by taxon (plant type). Under this model, each taxon response is modelled as conditionally independent, given the climate. The problem described in Section 3.1 is for a single response surface. The predictive distribution for climate is computed given the count for this single taxon. Repeating the inference procedure separately for each taxon marginally and taking the product of the climate predictive distributions yields a predictive probability distribution over all taxa, given the vector of taxa counts. A simple graph of such models is shown in Figure 3.3.

The marginal inference on each taxon, independently of the others, allows for many conveniences for the computational inference (discussed in Chapter 5) and in model design. As there are only a handful of hyperparameters associated with modelling each taxon and a single latent surface, both the model and the inference are simple. Interaction between taxa is not allowed in the marginal model so modelling interactions do not have to be considered.

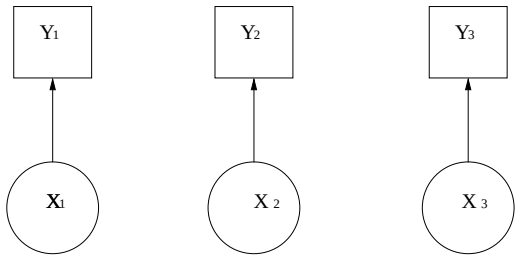


Fig. 3.3: **A decomposable model. The data Y_i are (potentially multivariate) dependent on parameters X_i (also potentially multivariate). The graph has no connections between modules $1, 2$ and 3; therefore the entire likelihood is expressible as the product of these three parts.**

The following section describes a method for testing the validity of the conditional independence assumption that is required to disjoint-decompose the model.

3.2.2 Non-Disjoint-Decomposable Models

If marginals / decomposable models are used erroneously, errors are incurred. The error increases with the number of dependent parts that are modelled independently.

Figure 3.4 demonstrates how, for toy data, a cross-validation error statistic increases with each additional component modelled. Each entry in a vector of observations occurring at multiple locations in a uni-dimensional space is modelled as an indirect observation of a latent parameter. These latent parameters vary smoothly across the location space and are modelled as GMRFs, independent of the other latent surfaces.

The data are generated from a latent field composed of 15 identical smooth surfaces defined on a regular grid of the location space. These surfaces are not independent given the locations, however they are modelled as such in order to facilitate decomposition of the problem. This induces the errors, manifested in the plot as an increase in the percentage of points falling outside their corresponding leave-one-out predictive distribution's 95% highest posterior density estimate with each surface added. This is the cross-validation Δ statistic introduced in Section 3.1 and has an expected value of less than or equal to 5%.

Such dependence can occur as a result of direct interaction or through the joint dependence on unobserved covariates. If the model needs to incorporate dependency

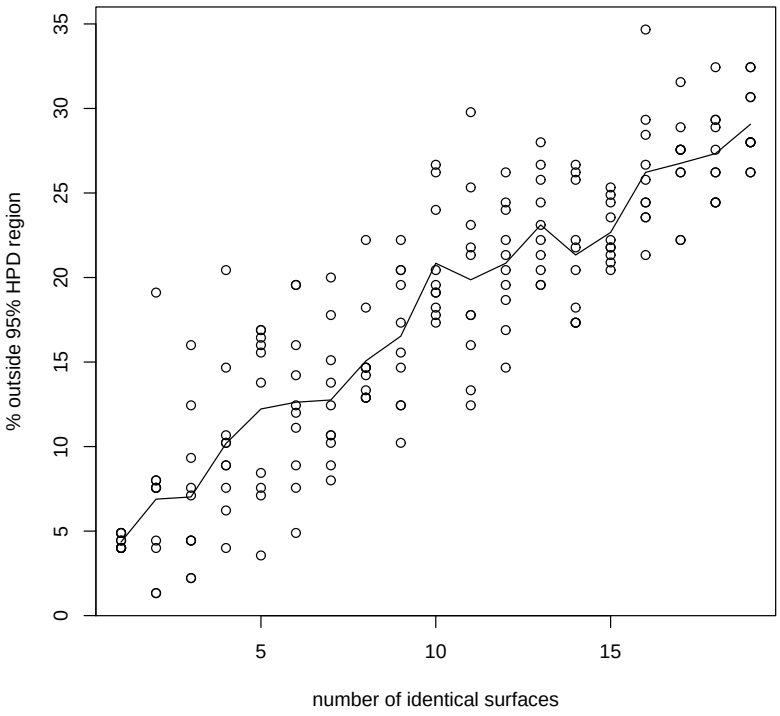
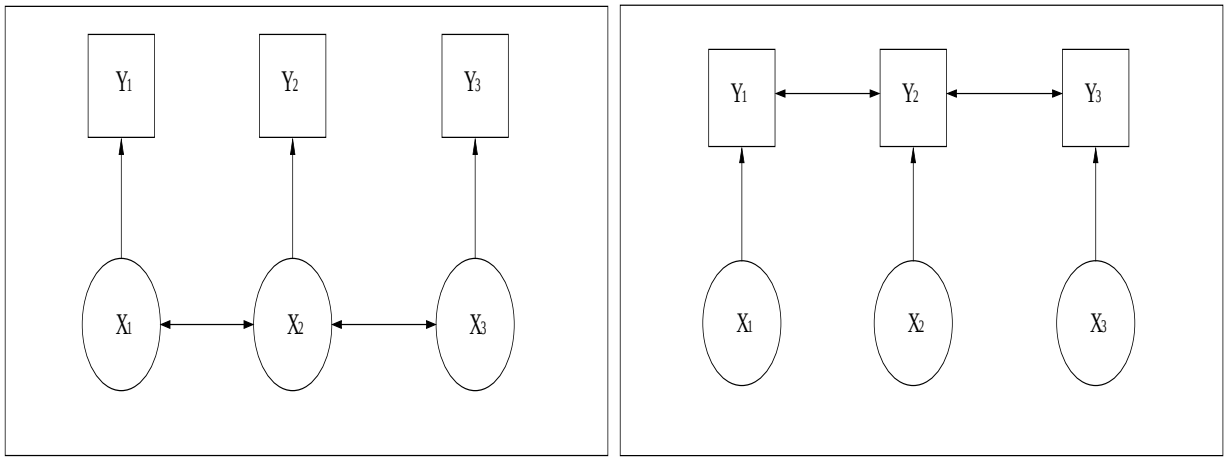


Fig. 3.4: **A plot of the percentage of locations lying outside their corresponding leave-one-out cross-validation predictive distributions 95% HPD region. Interaction in this case is with inter-surface correlation parameters equal to one (fully correlated), representing the most extreme level of interaction.**

A single smooth surface is used T times to generate random counts data at each of 100 discrete locations. These T counts at each location are then treated as independent information; the error rate associated with this mis-specification grows linearly with T which is on the horizontal axis. The graph is obtained by simulating counts data 10 times for each value of T on the horizontal axis. The line denotes the mean across these 10 replications.

The shape of the response surface dictates the slope of the line. If the responses are all linear then the error rate on the y-axis will not increase so the line is horizontal. Furthermore, if the correlation parameter is zero, no error is incurred and the line is again horizontal.



(a) Prior Dependence (b) Likelihood Dependence

Fig. 3.5: **Two non-decomposable models. The data Y_i are (potentially multivariate) dependent on parameters X_i (also potentially multivariate). Graph (a) has connections between modules $1, 2$ and 3 at the level of the parameters; this is expressed as non-zero interaction terms in the joint prior precision matrix. Graph (b) has connections between modules $1, 2$ and 3 at the level of the data; this is expressed as non-zero interaction terms in the joint likelihood precision matrix. Where to place these terms depends on the source of interaction in the model. In either case, the posterior for the parameters X will have non-zero interaction terms across the modules and hence cannot be disjoint-decomposed.**

between these multiple surfaces, an overall covariance model must be set up. This is most readily discussed in the Gaussian context of the following section.

These interactions may be modelled as either non-zero precision terms in the multivariate prior or in the likelihood. Figure 3.5 shows graphs for these two models.

3.2.3 Sources of Interaction

If a joint model is not decomposable, there must be non-zero off diagonal terms in either the prior or the likelihood precision matrix corresponding to inter-taxa dependence. It is the source of interaction that determines where these non-zero terms should occur. For multivariate Bayesian hierarchical models, such as the motivating palaeoclimate reconstruction problem, the following *potential* sources

of interaction across plant taxa are identified. Each source may prevent model decomposition.

Covariates

Additional, covariate information Z is sometimes available for large, multivariate datasets. For example, in the RS10 pollen and climate dataset, altitude, longitude and latitude are available. If, given climate, the propensity to produce pollen is thought to depend on one or more of these covariates then the response surfaces will not be independent.

A common approach to including this in the modelling is to model the counts data as dependent on the response surface values plus the covariate data times a vector of unknown regression parameters β.

$$y_i \sim \pi\left(y_i | x_i + z_i^T \beta\right) \tag{3.5}$$

If the data are conditionally independent given locations (climates) *and* covariates, then they should *not* be modelled as conditionally independent given the locations only. In this case, the interaction occurs indirectly through the covariates.

Competition

Interaction at the data level may occur as a result of direct competition. Interactions of this type are independent of the underlying response and should therefore be modelled in the likelihood precision matrix. An example in terms of the motivating pollen dataset would be competition for resources between plant types.

Constraints

The data collection mechanism may produce non-zero correlation between otherwise independent components of the data. For example, if the data were collected until a preset total were reached, then this would not only set an upper limit on the counts data; a negative correlation between the entries of the data vector would also be produced, due to the sum-to-total constraint.

Since the information carried by the data now resides in the relative proportions, rather than absolute values, the model should be for these proportions. Therefore,

a sum-to-unity constraint must apply to the latent parameters. For example, if the data likelihood model were Multinomial, then an appropriate distribution on the parameters of the Multinomial would be defined on the simplex (Section 3.5.1).

However, an important observation that was was missed in Haslett et al. (2006) is that such models may still be disjoint-decomposed, provided the data are in fact conditionally independent given the constraint. This is discussed in some detail in Section 3.5.7.

3.3 Multivariate Normal Model

If the data are modelled as Gaussian given Gaussian latent parameters then inversion of the model may be done analytically. Although the data of the motivating problem are integers (counts data) and thus are not suitably modelled as Gaussian, the familiar Gaussian framework does allow for some of the modelling nuances dealt with in this chapter to be introduced in an easily demonstrated context.

Furthermore, the inference procedure introduced in Section 4.1 is directly motivated by and related to the Gaussian model. Therefore, lessons learned from the multivariate normal model can be readily applied to a more general context of inference.

Constructing non-decomposable probability models requires explicit modelling of the covariance between interacting modules. Again, the familiar Gaussian context is useful to illustrate some important aspects of such models.

3.3.1 General Case Normal Models

There are various models for multiple response surfaces that do not disjoint-decompose as the product of their marginals. These involve models for which there are various forms of interaction, either between the counts data, given the responses, or between the responses themselves, given the locations. This is manifested in the model as non-decomposable multivariate joint likelihoods or non-decomposable multivariate joint priors, respectively.

It is very important to again note that even if such interactions exist, but are not built into the joint model, then the model will still disjoint-decompose. Inference

will be identical, albeit flawed, for the joint model and the product of the by-taxon marginals.

If interaction (dependence) is to be modelled, it arises in the Gaussian context either in the latent parameters precision matrix or in the likelihood precision matrix. In either case, the relevant precision matrix will have non-zero entries for between components (taxa) interactions. Note that in all cases considered, there are non-diagonal terms in the prior precision matrix for each taxon. These are intra-taxon terms and are a result of the prior belief on the smoothness across location space of the latent responses. Thus, for a model with multiple non-interacting / independent taxa, the overall precision matrix is **block-diagonal**. Each block is then the (independent) taxon-specific precision matrix.

This is best illustrated by means of a very simple example. Suppose there are two counts Y at each of two locations in a $1D$ location space $L = \{l_1, l_2\}$. This gives a total of four counts $Y = \{y_1^1, y_1^2, y_2^1, y_2^2\}$, where the subscript indexes the components of the counts vector and the superscript indexes the location. i.e. y_i^j is the count for the i^{th} component at the j^{th} location.

The model requires four latent variables $X = \{x_1^1, x_1^2, x_2^1, x_2^2\}$, using the same indexing as for the data. The fully Gaussian hierarchical model is then specified entirely by the following multivariate normal distributions:

$$\begin{aligned} Y &\sim MVN(Y; X, \Sigma_Y) \\ X &\sim MVN(X; \mu_X, \Sigma_X) \end{aligned} \tag{3.6}$$

where the hyperparameters θ are the prior mean vector μ_X and the prior covariance matrix Σ_X. In fact, it is more convenient to work with precision matrices rather than covariance matrices with precision $Q_X = \Sigma_X^{-1}$ the inverse of the covariance matrix (see Section 2.2.4). Similarly, $Q_Y = \Sigma_Y^{-1}$, when Σ_Y exists.

Specification of the full, joint model now involves specifying the hyperparameters $\{\mu_X, Q_X\}$ and the likelihood precision matrix Q_Y.

The posterior distribution for X is also multivariate normal with mean and precision matrix given by

$$\mu = (Q_X + Q_Y)^{-1}(Q_X\mu_X + Q_Y\mu_Y) \tag{3.7}$$

$$Q = Q_X + Q_Y \tag{3.8}$$

If either Q_X or Q_Y have non-zero terms for precision(x_1, x_2) or precision(y_1, y_2) respectively, then the posterior precision matrix Q will also carry these non-zero terms. Therefore, the posterior will not disjoint-decompose exactly and the product of posterior marginals will not equal the full joint posterior.

It is worth reiterating at this point that if both the prior and likelihood disjoint-decompose, then so does the posterior. This is easily seen in the context of the multivariate normal as if both terms on the right hand side of Equation (3.8) do not contain non-zero terms preventing decomposition, then their sum, giving the posterior precision matrix, will also not contain such terms. Hence, the product of (perhaps multivariate) marginals will yield exactly the joint posterior and the model thus disjoint-decomposes.

If a model with inter-surface interactions is required then there are four options. Specification is through one of:

1. known terms in the prior
2. known terms in the likelihood
3. unknown parameters in the prior
4. unknown parameters in the likelihood

The first two options are closely related in the context of multivariate normal models for both the prior and the likelihood. Which of these two to incorporate depend on the model and are problem specific. The source of interaction informs the choice of modelling interaction in the prior or in the likelihood. The last two options involve inference issues and are dealt with in Chapter 5.

Prior Precision Matrix

The precision, or "degree of mutual agreement", of $x_{i_1}^{j_1}$ and $x_{i_2}^{j_2}$ is

$$prec(x_{i_1}^{j_1}, x_{i_2}^{j_2}) = q_X{}_{i_1,j_1}^{i_2,j_2} \tag{3.9}$$

As this is a four dimensional indexing, it is necessary to construct a system for indexing across both surfaces / counts and the location space using a single index to yield a two dimensional precision matrix. The convention adopted here is that

the subscript i and superscript j pairing becomes single index $(i-1)N_L + j$, where N_L is the total number of discrete points in the location space. Note that there is one surface per taxon for the pollen data example.

The overall precision matrix covering all possible (process, location) pairings for the latent parameters X is then:

$$Q_X = \begin{pmatrix} {q_X}_{1,1}^{1,1} & {q_X}_{1,1}^{1,2} & {q_X}_{1,1}^{2,1} & {q_X}_{1,1}^{2,2} \\ {q_X}_{1,2}^{1,1} & {q_X}_{1,2}^{1,2} & {q_X}_{1,2}^{2,1} & {q_X}_{1,2}^{2,2} \\ {q_X}_{2,1}^{1,1} & {q_X}_{2,1}^{1,2} & {q_X}_{2,1}^{2,1} & {q_X}_{2,1}^{2,2} \\ {q_X}_{2,2}^{1,1} & {q_X}_{2,2}^{1,2} & {q_X}_{2,2}^{2,1} & {q_X}_{2,2}^{2,2} \end{pmatrix} \tag{3.10}$$

Rows of Q_X denote the precisions for individual surfaces across locations and columns denote inter-surface precisions at a point in the location space. If the latent surfaces are modelled as conditionally independent, given location, then ${q_X}_{i_1,j_1}^{i_2,j_2} = 0$ for $i_1 \neq i_2$, regardless of j_1 and j_2.

$q_{i_1=i_2}$ are intra-surface parameters and may be reduced to a single hyperparameter via imposition of a regular structure as shown in Section 2.2.4.

Interaction between surfaces is modelled as being a local effect in the location space. i.e. precision, for a common location, between two surfaces is the same across the locations space. Interaction between surfaces at non-equal locations is modelled as zero. Consideration of whether the joint model disjoint-decomposes exactly now amounts to checking whether the posterior precision matrix is block-diagonal. (Diagonal implies total independence; block diagonality is a consequence of the conditional independence across taxa, but not across climate; see Section 3.2.1.)

Likelihood Precision Matrix

The data are typically modelled in Bayesian hierarchical models as being conditionally independent, given the level two parameters. This implies diagonal precision and covariance matrices Q_Y and Σ_Y. If interaction between components is at the data level, then this is no longer the case.

However, for the model with multivariate normal likelihood with multivariate normal prior, the multivariate normal posterior will have the exact same precision matrix whether the interactions are placed in the likelihood precision matrix or the prior precision matrix as it is simply a sum of the two. The posterior mean will be

slightly different; however it is simple to show that for any choice of inter-surface precision parameters in the likelihood, there exists a set of inter-surface parameters for the prior precision matrix that yield the same posterior mean.

In the following the notation is that the model with interactions in the likelihood precision matrix is given the superscript A and the model with the interactions in the prior takes the superscript B. If the prior is zero-mean and the posterior means under each model are taken to be equal:

$$\begin{aligned}
(Q_X^A + Q_Y^A)^{-1} Q_Y^A Y &= (Q_X^B + Q_Y^B)^{-1} Q_Y^B Y \\
(Q_X^B + Q_Y^B)(Q_X^A + Q_Y^A)^{-1} Q_Y^A Y &= Q_Y^B Y \\
Q_X^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A Y + Q_Y^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A Y &= Q_Y^B Y \\
Q_X^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A Y &= Q_Y^B Y - Q_Y^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A Y
\end{aligned}$$

One solution to which equation is

$$\begin{aligned}
Q_X^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A &= Q_Y^B - Q_Y^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A \\
Q_X^B &= (Q_Y^B - Q_Y^B (Q_X^A + Q_Y^A)^{-1} Q_Y^A)(Q_Y^A)^{-1} (Q_X^A + Q_Y^A)
\end{aligned}$$

The above demonstrates the similarity between modelling interaction at the prior and at the likelihood precision matrices.

3.3.2 Sensitivity to Dependence

Errors are incurred if a decomposable model is applied to data that have arisen from a model that does not disjoint-decompose. In the context of this section (Normal models with known parameters), this arises as setting the multivariate likelihood precision matrix terms that relate to between-module interactions to zero. If Q is the true, joint precision matrix, then $\tilde{Q}$ is the disjoint-decomposed model precision matrix with all interaction terms set to zero. If the model disjoint-decomposes, these two matrices will be identical. If not, the disjoint-decomposed model will be an approximation to the true model. The accuracy of this approximation will of course depend on the level of interaction between the modules that are treated as independent in the disjoint-decomposed model. A toy model demonstrates this as follows:

suppose "counts" data are generated a multivariate normal distribution with mean μ and precision matrix Q

$$Y \sim MVN(\mu, Q) \tag{3.11}$$

where the length of μ is the product of the number of discrete locations N_L and the number of surfaces N_T. These surfaces vary smoothly across locations so that given a new vector of counts, the model may be inverted and the location associated with these counts predicted. In this example, $N_T = 2$ and $N_L = 10$.

Using the same indexing system in Equation (3.10) to index over components and locations, the full precision matrix for two components in two locations will be decomposable as the product of marginals iff $q_{i_1,j_1}^{i_2,j_2} = 0$ for all $i_1 \neq i_2$. This equates to whether Q is in fact block-diagonal, with the number of blocks equal to the length of the data vector at each location N_T and the length of the side of each square block equal to the number of discrete locations N_L.

Provided interactions between counts are consistent throughout the location space, symmetric, and that there is no interaction between data at disparate locations, then there are $\binom{N_T}{2}$ interaction terms. Thus, for two such surfaces, there is a single parameter ρ governing interactions.

Intuitively, the closer this parameter is to zero, the closer Q is to being block-diagonal and thus the closer the joint model is to being decomposable. Replications of the above toy model, for varying values of ρ, the scalar interaction parameter, show this relationship empirically; see Figure 3.6. For each replication, two new randomly generated smooth response surfaces are generated; thus the results in the figure are generalised across all shapes of response surface. The counts data at location j come from random draw of a bivariate-Normal distribution with mean equal to the two responses at that point and precision matrix given by

$$Q_j = \begin{pmatrix} q & \rho q \\ \rho q & q \end{pmatrix} \tag{3.12}$$

It can be seen from Figure 3.6 that the greater the absolute value of the interaction parameters ρ, the more points fall outside their 95% HPD region for the inverse predictive distribution. The mean of the percentage outside the 95% HPD region is less than 5% at $\rho = 0$ due to the use of discrete HPD regions. Only regions of 95%

or more may be specified so that the percentage outside is expected to be 5% or less. Section 5.3.1 discusses this in more detail.

3.4 Counts Data

Non-Gaussian likelihood models are introduced and examined in this section. This leads to posteriors that are typically not available in closed form. The main concern is with multivariate counts data; treated as conditionally independently using models such as the Poisson and various related distributions or with multivariate counts likelihoods for the case of vectors of constrained counts data. Section 4.1 will show how closed forms for the posterior will be achieved for these non-Gaussian likelihoods and Gaussian priors.

3.4.1 Poisson Model

For all models that disjoint-decompose, each element of a constrained counts vector is modelled independently of the rest. In the simplest case for counts data, the counts are modelled as being Poisson distributed, with the rate parameter derived as a deterministic function of the underlying latent surface. Thus, for the hierarchical model, with log-link for the rate parameter and a GMRF prior

$$\begin{aligned} y_i &\sim Poisson(y_i; \lambda_i) \\ \lambda_i &= exp(x_i) \\ X &\sim GMRF(X; \mu_X, Q_X) \end{aligned} \tag{3.13}$$

However, richer likelihood models are often required. Among these are models for data that are *overdispersed* and data that are *zero-inflated*.

3.4.2 Scaled Poisson

If the "effort" spent on counting the data varies across the sampling space, then this will effect the expected and observed counts. A more general form of the Poisson distribution with an additional "effort" parameter allows direct modelling of this effect. For example, if the time t spent collecting counts data Y were to vary, then

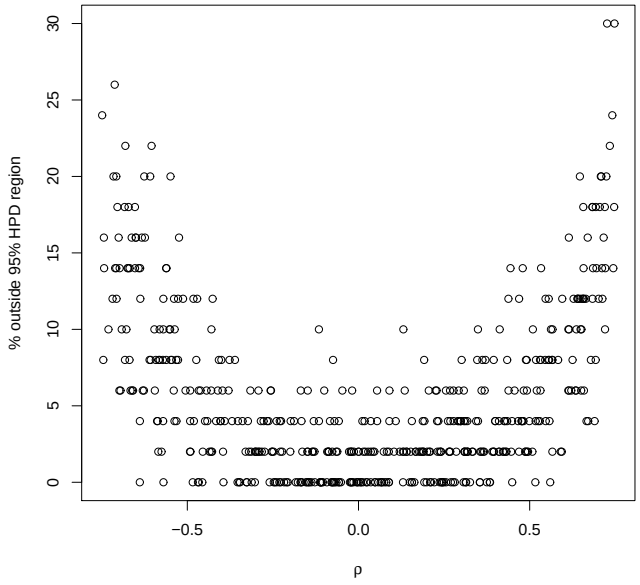


Fig. 3.6: **Replications of a multivariate normal model. Two response functions vary smoothly across a discrete space; a location is drawn at random and two "counts" are generated from a bivariate-Normal distribution with mean equal to the responses at that point and precision matrix given by Equation (3.12). The larger the absolute value of the interaction parameter ρ, the greater the error in the approximate decomposition of the joint model.**

The above result is taken across a range of randomly generated response surfaces; thus it is generalised w.r.t. the shape of the response. Figure 3.4 suggests an error rate of about 12% for two surfaces with a correlation of $\rho = 1$; however, this is for a particular response surface only.

the scaled Poisson distribution for those counts, with rate λ would be

$$y \sim \frac{(\lambda t)^y e^{-\lambda t}}{y!} \tag{3.14}$$

For the motivating pollen dataset, the time or effort spent in counting the data is unknown; what is available is the total count across all plant taxa. This can be used as an observed surrogate, or proxy, for the effort / time spent collecting the counts for each taxon. In fact, this total count imposes a constraint in the form of a strict upper limit on the count for each individual taxon. Counts thus constrained are typically modelled using the Binomial distribution, or a related distribution.

3.4.3 Overdispersion

Of particular interest here is data for which there is overdispersion; i.e. the single rate parameter of the Poisson distribution is insufficient as the variance is greater than the mean of the data, given the parameters. Zero-mean, normally distributed non-spatial random effects may be added to the parameters, resulting in overdispersion with respect to the spatial component of the latent surface.

The data are then indirect observations of a latent variable with two distinct parts; a spatially structured part X and a random effects part ϵ.

$$\begin{aligned} \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon) \\ \delta_i &= x_i + \epsilon_i \\ \Rightarrow \delta_i &\sim \mathcal{N}(x_i, \sigma_\epsilon) \end{aligned}$$

The data then depend on this new parameter δ_i

$$\begin{aligned} y_i|\delta_i &\sim \pi(y_i|\delta_i) \\ \Rightarrow y_i &\sim \int_{\epsilon_i} \pi(y_i|\delta_i)\pi(\delta_i|x_i, \sigma_\epsilon)d\epsilon_i \end{aligned} \tag{3.15}$$

This results in double the number of latent random variables in the model and at least one extra hyperparameter (the variance of the random effects), which is an undesirable situation. An alternative is to introduce random effects that are modelled using a distribution that is conjugate to the likelihood. The random effects may then be analytically integrate out, leaving a reparameterised, closed form likelihood that has a variance that is larger than the mean.

For example, the data Y is Poisson given the rates λ; the rates are a mixture of a spatially smooth part X, modelled as a GMRF, and a non-spatial random effect δ. If this zero-mean random effect component is such that the product of the spatial part and the random effect is Gamma distributed, with mean equal to the spatial part, then the hierarchical model is:

$$\begin{aligned} y_i &\sim Poisson(y_i;\lambda_i) \\ \lambda_i &\sim Gamma(\lambda_i;\delta,(1-p_i)/p_i) \\ p_i &= \frac{\delta}{\delta+e^{x_i}} \\ X &\sim GMRF(X;\mu_X,Q_X) \end{aligned} \tag{3.16}$$

This simplifies by integrating out the λ_is (suppressing the indices i):

$$\begin{aligned} \pi(y|x) &= \int_0^\infty Poisson(y;\lambda)Gamma(\lambda;\delta,(1-p)/p)d\lambda \\ &= \int_0^\infty \frac{\lambda^y}{y!}e^{-\lambda}\lambda^{\delta-1}\frac{exp(\frac{-\lambda p}{1-p})}{\Gamma(\delta)(\frac{1-p}{p})^\delta}d\lambda \\ &= \frac{1}{y!\Gamma(\delta)}p^\delta\frac{1}{(1-p)^\delta}\int_0^\infty \lambda^{\delta+y-1}exp(\frac{-\lambda}{1-p})d\lambda \\ &= \frac{1}{y!\Gamma(\delta)}p^\delta\frac{1}{(1-p)^\delta}(1-p)^{\delta+y}\Gamma(\delta+y) \\ &= \frac{\Gamma(\delta+y)}{y!\Gamma(\delta)}p^\delta(1-p)^y \end{aligned} \tag{3.17}$$

which is the *Negative-Binomial* distribution.

This counts distribution carries just a single extra parameter (δ) over the simple Poisson model. δ controls the degree of overdispersion.

3.4.4 Sensitivity to Zero-Inflated Likelihood

If a non-zero inflated likelihood is used for data that are zero-inflated, then inference on the parameters of that likelihood will be erroneous. Specifically, the extra zeros will reduce the unobserved mean parameter of the likelihood and will increase the variance. In the context of this Chapter (models with known parameters), the inverse predictive distributions will misplace probability mass if inversion is done using the wrong (i.e. non-zero-inflated) likelihood function.

This will occur in a predictable manner; the non-zero-inflated model will place inverse predictive probability mass for zero counts exclusively at the regions for

which the response function is lowest. Non-zero counts will be treated the same as for an equivalent zero-inflated likelihood and will therefore be correct. The degree of zero-inflation will thus govern the degree of the error of the non-zero-inflated model when applied to zero-inflated data to generate inverse predictive distributions.

A toy problem example is once again employed to illustrate this point. A smooth surface p in a uni-dimensional discrete space gives the (known) parameters for a zero-inflated Binomial likelihood:

$$\pi(y) = \begin{cases} 1 - q + qBin(0; p, n) & y = 0 \\ qBin(y; p, n) & y > 0 \end{cases} \tag{3.18}$$

where $q = p^{\alpha}$, $\alpha = 0.3$
and $n = 1000$ is the total count.

Results for the inverse predictive distribution given known, deterministic forward models are shown in Figure 3.7. The result of applying a non-zero-inflated model to zero-inflated data is clearly demonstrated; under the non-zero-inflated model, all zero-counts are inferred to arise exclusively at the lowest points of the response curve. This is because the non-zero-inflated model will generate most zeros in this are of the locations space. While the zero-inflated model still necessarily reconstructs the same location of least response as being most probable, the curve is not nearly so peaked. This, being the model from which data was simulated, gives the correct predictive distribution.

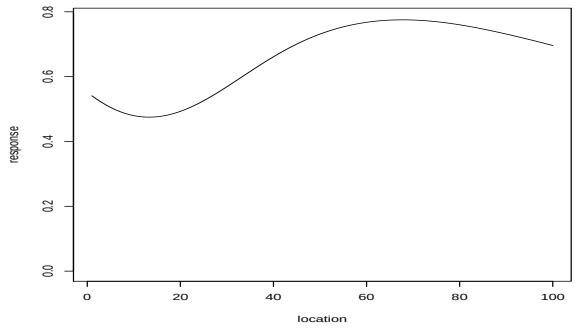
3.5 Compositional Data

If a vector $p = \{p_1, \ldots, p_N\}$ has all non-negative elements representing proportions of a whole then the vector is constrained to sum to unity

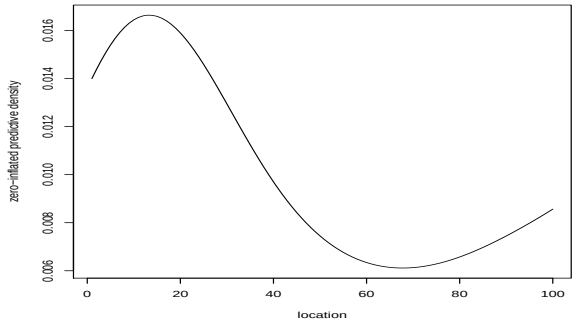
$$p_1 + \ldots + p_N = 1 \tag{3.19}$$

Such vectors of proportions are compositional data and are frequently and erroneously modelled using techniques developed for unconstrained spaces (Aitchison (1986); Aitchison and Egozcue (2005)).

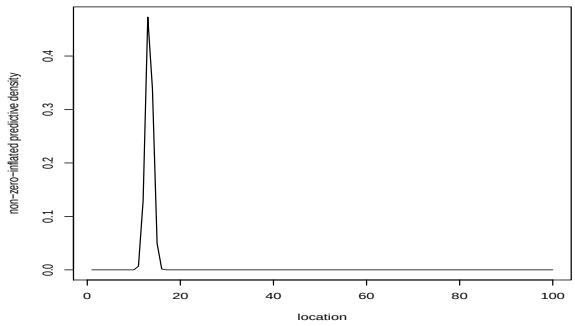
Due to the constraint, the data have one less degree of freedom than the length of the compositional vector. The full vector may be completely specified using the



(a) response surface



(b) Count = 0 predictive distribution; zero-inflated model



(c) Count = 0 predictive distribution; non zero-inflated model

Fig. 3.7: **The result of performing inverse predictions using a non-zero-inflated likelihood model for zero inflated counts data: Given a zero count, inverting the zero-inflated model (from which the data were generated, Figure (a), thus the correct model; Equation (3.18)) gives the correct inverse predictive distribution, Figure (b). The non-zero-inflated Binomial likelihood model cannot account for extra zero-counts, thus is places all the inverse predictive probability mass at the region of lowest response, Figure (c). Note the change of scale in the y-axes.**

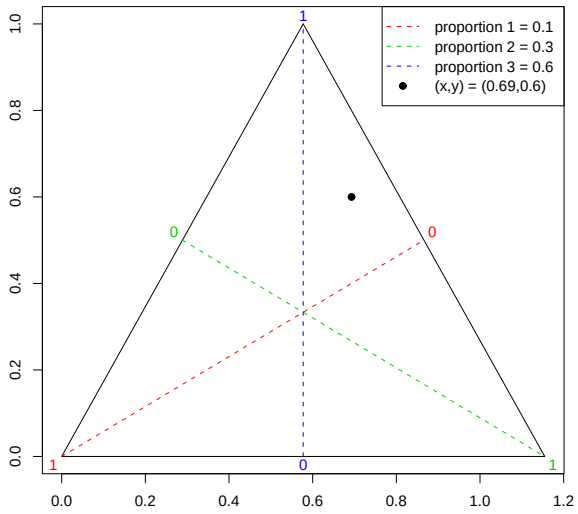


Fig. 3.8: **A proportions vector of length 3, under the sum to unity constraint is represented on a 2D simplex space, represented on a** *ternary diagram*. **An example vector with values for the proportions vector** $p = \{0.1, 0.3, 0.6\}$ **is shown. Any value of the compositional vector** p **of length** N **may uniquely determined by an** $N-1$ **vector and the sum-to-one constraint on** p**.**

components of any $N-1$ subvector (the left out value being determined as one minus the sum of the subcomposition). Any such subvector completely specifies the full composition.

3.5.1 The Simplex Space

The set of all possible vectors for a given length of composition is referred to as the *simplex space*. The simplex space for a compositional vector of length N with unit-sum constraint is then $N-1$ dimensional.

There are two contrasting approaches to modelling compositional data in a coherent manner that do not fall into the trap of using traditional, unconstrained multivariate statistics:

1. Model the data on a simplex space using new, emerging techniques developed specifically for use on compositional datasets.

2. Use traditional multivariate statistics on some transformation of the data to the real space and project results back to the constrained simplex space.

These two competing approaches require different models to be developed. The latter typically uses multivariate normal distributions and established Gaussian theory; this leaves choice of transformation and subvector as the main decisions to be made in the modelling sense. The former requires alternative distributions, defined directly on the simplex space.

Although progress has been made on statistics defined on the simplex space, the availability of a rich and established theory of multivariate analysis makes the use of the transformation technique the more appealing option. This is the approach advocated by Aitchison (1986) and is more widely adopted. In fact, distributions defined on the simplex space tend to have strong implied independence structures.

3.5.2 Dirichlet Distribution

The majority of distributions defined on the simplex sample space are of the Dirichlet class. Despite the sum-to-one constraint, this class has an inflexible covariance structure making it unsuitable for many applications. The Dirichlet is the multivariate generalisation of the Beta distribution with probability density given by

$$\pi(p|\eta) = \frac{\Gamma(\sum_{i=1}^{N})\eta_i}{\prod_{i=1}^{N}(\eta_i)} \prod_{i=1}^{N} p_i^{\eta_i - 1} \tag{3.20}$$

Although all N elements of the compositional vector p appear in the density calculation, the distribution itself is defined on the $N-1$ dimensional simplex as p_i is uniquely determined by p_{-i}. η are the parameters of the Dirichlet distribution.

The covariance between two components of a Dirichlet distributed compositional vector is

$$cov(p_i, p_j) = -\frac{\eta_i \eta_j}{(\sum_k \eta_k)^2(\sum_k \eta_k + 1)} \tag{3.21}$$

for $i \neq j$. Thus, for modelling positive covariances between components, the Dirichlet distribution is entirely unsuitable. The covariance structure in the Dirichlet is entirely due to the sum-to-one constraint; this is why the Dirichlet is said to have a strong implied independence structure. Given the constraint, the components are necessarily modelled as conditionally independent.

A Dirichlet with parameter vector η may be expressed as a product of Gamma distributions, with shape parameters η and rate parameters all equal to $\sum_i \eta_i$, conditioned on the sum = 1 following a Gamma distribution with shape and rate both equal to $\sum_i \eta_i$.

$$\pi(P;\eta) = \frac{\prod_{i=1}^{N_T} \text{Gamma}(p_i;\eta_i,\sum_k \eta_k)}{\text{Gamma}(1;\sum_k \eta_k,\sum_k \eta_k)} \tag{3.22}$$

If sampling from the Dirichlet is required, this can be achieved by sampling from the Gamma marginals and then rescaling such that the sum is one. In fact, this is the usual algorithm for sampling from a Dirichlet distribution.

3.5.3 Generalized Dirichlet Distribution

The Generalized Dirichlet distribution (Connor and Mosimann (1969)) has a more general covariance structure, achieved by doubling the number of parameters. If the number of components is N, then a Generalized Dirichlet (GD) distribution has two sets of parameters, each of length N:

$$GD(P;a,b) = \left[\prod_{i=1}^{N-1} B(a_i,b_i)^{-1}\right] p_N^{b_{N-1}-1} \prod_{i=1}^{N-1}\left[p_i^{a_i-1}\left(\sum_{j=1}^{N} p_j\right)^{b_i-1-(a_i+b_i)}\right] \tag{3.23}$$

In the Generalized Dirichlet distribution, one of the components is always negatively correlated with the rest. The other components may be positively or negatively correlated with each other (Wong (1998)). Labelling the component that is strictly negatively correlated with the rest as p_1, if $i,j > 1$ then $cov(p_i,p_j)$ may be positive or negative. However, for any index greater than the lower of i,j, the sign of the correlation will stay the same. i.e. :

$$\begin{aligned} i &> j > 1 \\ cov(x_i,x_j) &= +ve \\ \Rightarrow cov(x_k,x_j) &= +ve \qquad \forall k > j \end{aligned} \tag{3.24}$$

This also applies to negative correlations so that the covariance structure for the Generalized Dirichlet distribution is in fact quite limited (see Section 3.5.6).

3.5.4 Logistic-Normal Class of Distributions

One class of distributions on the N dimensional simplex that have a richer covariance structure than those provided by the Dirichlet class of distributions is defined by

distributions in an unconstrained multivariate space (e.g. $\Re^N$) with transformation to the simplex space.

This is the approach advocated and developed by Aitchison (1986). The proportions are transformed to the real space through the use of a transformation function. Standard multivariate analysis is carried out on these transformed variables (typically using multivariate Gaussian distributions) and the results are transformed back to the simplex space using the inverse of the original transforming function.

If N dimensional multivariate normally distributed real random vectors are transformed to the simplex space in N dimensions using the one-to-one inverse centered logratio transform:

$$p_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}; i = 1, \ldots N \tag{3.25}$$

then the compositional vectors on the simplex are said to have a centered Logistic-Normal distribution (Aitchison (1986), Aitchison and Egozcue (2005)). The transformation from the simplex space to the real space is the centered logratio transform is given by

$$x_i = log(p_i) - \frac{\sum_{j=1}^{N} log(p_j)}{N} \tag{3.26}$$

the second term being there to centre the real vector around zero. There are other, closely related transformations that give rise to similar distributions (such as the additive Logistic-Normal). The general term for such distributions is Logistic-Normal and the centered Logistic-Normal is the distribution used in this thesis.

The advantage of transforming to the real space is that standard multivariate statistical procedures and models based on the multivariate normal distribution are made available. This allows for rich, well developed models to be used for the real-space X parameters, before simply transforming back to the simplex space.

Arbitrarily rich covariance structures may be built for the compositional vector through specification of multivariate normal distributions on the real space. The entire battery of existing techniques for the Normal distribution may be employed.

Chapter 6 of Aitchison (1986), shows that for any Dirichlet distribution with parameters δ large enough that probability mass is highest at the centre of the simplex then there is a very close Logistic-Normal distribution with diagonal covariance / precision matrix.

The approach broadly used in this thesis is to define normal priors on an unconstrained space with transformation to the simplex. The motivation for this is not in fact in the conveniences arising in the Normality assumptions across the composition, but because modelling latent surfaces as Gaussian and Markov across the location space allows for the use of the specialist multivariate inference techniques introduced in Section 2.2 and detailed in Chapter 4.1.

Extra unknown parameters could be introduced to model the inter-dependence, but this is an inference question and is therefore dealt with in Chapter 4.1.

Interest in this thesis is in models with some inter-process covariances that do not require either of the following:

1. Prior knowledge of inter-component covariance / precision terms

2. Complicated, highly parameterised models for such covariances

3.5.5 Multivariate, Constrained Likelihood Functions

Although the Multinomial is a multivariate likelihood, with known covariances between components, it can be expressed as a product of independent Poisson distributions, with parameters equal to $n \times P$, conditioned on the sum being equal to the total count n. This sum ($\sum_i y_i = n$) itself follows a Poisson distribution with rate parameter n.

$$\pi(Y; n, P) = \frac{\prod_{i=1}^{N_T} \text{Poisson}(y_i; \lambda_i)}{\text{Poisson}(n; n)} \tag{3.27}$$

with $\lambda_i = n \times p_i$ and $\text{Poisson}(y_i; \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda}}{y_i!}$

Thus, if a model using a decomposable likelihood function (such as the Multinomial) and a prior with conditional independence (such as the Dirichlet or indeed an Aitchison type prior with no inter-component covariance terms) is employed, then the posterior has conditionally independent components.

For the multivariate normal model, the prior and likelihood precision matrices in this case will be block-diagonal with no inter-component entries being non-zero. Thus, the posterior precision matrix, which is the sum of the prior and likelihood precision matrices, is also block-diagonal. The joint model, which is multivariate normal is therefore equal to a product of smaller multivariate normals, each describ-

ing a separate component of the composition. Thus, the model disjoint-decomposes exactly.

For constrained, multivariate counts data, the simplest example is of a Dirichlet prior with a Multinomial likelihood. Due to conjugacy, the posterior for the latent parameters is also Dirichlet. Inference on these parameters may therefore be carried out on each component separately and the joint distribution may be constructed from the marginals post-hoc by conditioning on the sum.

Compound-Multinomial Likelihood

In the paper Haslett et al. (2006), the Multinomial was mixed with a Dirichlet distribution and a compound-Multinomial (or Dirichlet-Multinomial) distribution was formed for the likelihood. Although overdispersed with respect to the Multinomial, this distribution still enforces the conditional independence assumption.

$$\pi(Y|P,\delta,n) = \frac{n!\Gamma(\delta)}{\Gamma(n+\delta)} \prod_{i=1}^{N} \left(\frac{\Gamma(y_i+\delta p_i)}{\Gamma(\delta p_i) y_i!} \right) \tag{3.28}$$

where n is the total count, δ is a scalar overdispersion parameter, N is the number of components of the composition, Y are the counts and P are the parameters of the Multinomial.

The derivation of this likelihood is as follows, starting with the mixing of a Multinomial with a Dirichlet with parameters δP:

$$\begin{aligned} \pi(Y|\delta,P,n) &= \int_{\phi} \pi(Y|\phi,n)\pi(\phi|P,\delta)d\phi \\ &= \int_{\phi} \frac{n!}{\prod_{i=1}^{N} y_i!} \prod_{i=1}^{N} \phi_i^{y_i} \frac{\Gamma(\delta)}{\prod_{i=1}^{N}\Gamma(\delta p_i)} \prod_{i=1}^{N} \phi_i^{\delta p_i - 1} d\phi \\ &= \frac{n!}{\prod_{i=1}^{N} y_i!} \frac{\Gamma(\delta)}{\prod_{i=1}^{N}\Gamma(\delta p_i)} \int_{\phi} \prod_{i=1}^{N} \phi_i^{\delta p_i + y_i - 1} d\phi \end{aligned} \tag{3.29}$$

The term inside the integral is an un-normalised Dirichlet distribution on ϕ with parameters $\delta P + Y$. Substituting the inverse normalising constant times this Dirichlet for this term and taking the normalising constant outside the integral gives

$$\pi(Y|P,\delta,n) = \frac{n!\Gamma(\delta)}{\Gamma(n+\delta)} \prod_{i=1}^{N} \left(\frac{\Gamma(y_i+\delta p_i)}{\Gamma(\delta p_i) y_i!} \right) \int_{\phi} \text{Dirichlet}(\phi; \delta P + Y) d\phi \tag{3.30}$$

The integral is over a valid probability distribution and is therefore equal to one, yielding Equation (3.28).

3.5.6 Nested Compositional Models

Section 3.5.2 shows how common Dirichlet class distributions on the simplex have a strong implied conditional independence structure. Section 3.5.5 shows how multivariate, constrained likelihood functions such as the Multinomial may be disjoint-decomposed. The Generalized Dirichlet described in Section 3.5.3 allows for the breaking of this independence structure. However, it has double the number of parameters and still only allows for positive correlation between one component with all the others.

Aitchison type models, using a transformation from unconstrained, multivariate normally distributed, vectors in the unconstrained space to the simplex space provide an extremely rich class of model. However, they require either a prior knowledge of all interaction parameters or a specification through a large array of hyperparameters.

The former is unsuited to the pollen dataset; there is no such prior knowledge and interactions between components of the data vectors (whether through actual interaction or due to joint dependence on unobserved covariates) should be modelled in the likelihood.

The latter leads to problems of inference; specifically, the inference method of Section 4.1 requires a low number of hyperparameters and the conditional independence of the data, given the parameters (latent random variables).

Nested models provide an interesting alternative; these are models in which there is more than one "level". [1] This is graphically illustrated in Figure 3.9. Each level itself comprises a composition; each component of the composition may then be split into another composition on another level. The *structure of the nesting* refers to the number of levels and the splitting of each component into sub-compositions at each level.

A big advantage of such models is that they are still independent (given the constraint) at each level, provided the nesting structure is known. Unconditionally, the lowest level may exhibit a richer correlation structure than is possible with a Dirichlet model.

[1] This use of the term nest and nested should not be confused with the nested in Integrated Nested Laplace Approximation as the two are unrelated.

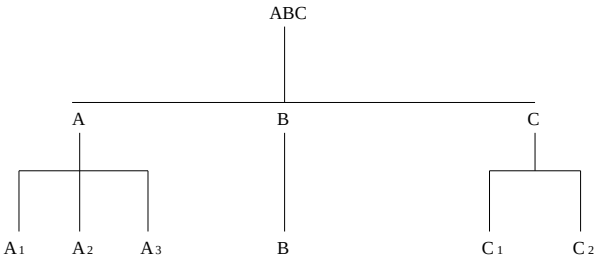


Fig. 3.9: **A two level nesting structure. At the top level is ABC; the first level splits this into A, B and C components. The second level splits A into 3 components, A_1, A_2 and A_3. The second level also splits C into C_1 and C_2, but B is not subdivided. There are thus 3 components at level 1 and a total of 6 at level 2. If the nesting structure was not known, then the vector $\{A_1, A_2, A_3, B, C_1, C_2\}$ would be modelled directly as components of ABC.**

Nested Multinomials

In Figure 3.9, for a Multinomial nesting, the total count is $A + B + C$; level 1 is a Multinomial of length 3 with components $\{A, B, C\}$ and parameters $\{P_A, P_B, P_C\}$, with $P_A + P_B + P_C = 1$. A is further split into 3 components on level 2 and these are also Multinomial with total count A and parameters $\{P_{A_1}, P_{A_2}, P_{A_3}\}$. C is Binomial (Multinomial of length 2) at the second level, with total count C and parameter(s) P_{C_1} ($P_{C_2} = 1 - P_{C_1}$). It is easy to show that the lowest level is then also Multinomial, with parameters $\{P_A P_{A_1}, P_A P_{A_2}, P_A P_{A_3}, P_B, P_C P_{C_1}, P_C P_{C_2}\}$ and total count $A + B + C$.

These two models then yield the same joint likelihood; i.e. a nested Multinomial is equivalent to a Multinomial on just the lowest level of each nest. Thus, knowledge of an existing nesting structure does not change the likelihood, if all nests are modelled as Multinomial.

Nested Dirichlets

For a nested Dirichlet model, the lowest level is not expressible as a Dirichlet. *Iso-probability contours* are a useful tool in illustrating the types of correlation structure achievable through various distributions. On the simplex, they demonstrate that, regardless of the parameters, the Dirichlet has strictly convex contours, due to the

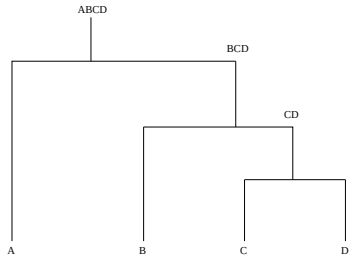


Fig. 3.10: **A three level nesting structure. At the top level is $ABCD$; the first level splits this into A and BCD. The second level splits BCD into two components, B and CD. The third level then splits CD into C and D.**

implied independence structure (Wong (1998)). The nested Dirichlet, however, can give rise to concave contours, showing positive correlations between components.

In the case of building priors for compositions, Wong (1998) shows that the Generalized Dirichlet allows for a more general covariance structure. Examination of the algorithm used to generate samples from the Generalized Dirichlet reveals an interesting result; the Generalized Dirichlet may be thought of as a series of two component nests. Indeed, Tian et al. (2003) touch on this briefly, noting that the nested Dirichlet is a special case of the Generalized Dirichlet.

However, this is strictly for the case of a nesting structure composed of nests of size 2 only; i.e. nested Betas (see Figure 3.10). This is most clearly seen by examination of the sampling algorithm for the Generalized Dirichlet (as per Wong (1998)):

$$\begin{aligned} p_1 &= rbeta(a_1, b_1) \\ sum &= p_1 \\ for \quad j &= 2, \ldots, N \\ &\{ \\ p_j &= rbeta(a_j, b_j)(1 - sum) \\ sum &= sum + p_j \\ &\} \end{aligned} \tag{3.31}$$

Thus Wong (1998) shows how constructing a prior, given knowledge of this nesting structure with binary splitting, gives rise to a Generalized Dirichlet for the lowest

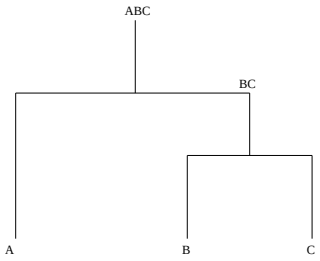


Fig. 3.11: **The simplest non-trivial nesting structure. At the top level is ABC; the first level splits this into A and BC. The second level splits BC into two components, B and C.**

level with the b parameters weighted by the number of splits to each component.

A more general nesting structure, such as that shown in Figure 3.9 cannot be written as a Generalized Dirichlet; however, it can be written as simply the product of the Dirichlet distributions at each nest. Thus constructing a prior with knowledge of the nesting structure is straightforward. A rich covariance structure for the lowest level is obtained, with the covariances entirely dictated by the nesting structure.

The comparison between the prior for the nested model and the equivalent non-nested model is pleasantly straightforward. For the simplest case, shown in Figure 3.11, the nested model has Dirichlet prior for the first level:

$$\pi(p_A, p_{BC}) \propto p_A^{\delta-1} p_{BC}^{2\delta-1} \tag{3.32}$$

the prior being centred on $\{1/3, 2/3\}$, since knowledge of the nesting structure dictates that BC must ultimately become two parts. The prior precision is dictated by the single δ parameter.

Similarly, the second level has Dirichlet prior:

$$\pi(p_B, p_C) \propto p_B^{2\delta-1} p_C^{2\delta-1} \tag{3.33}$$

The product of Equations (3.32) and (3.33) gives the nested model prior.

The model without knowledge of the nesting structure has prior equal to

$$\pi(p_A, p_B, p_C) \propto p_A^{2\delta-1} p_B^{2\delta-1} p_C^{2\delta-1} \tag{3.34}$$

Therefore the ratio of nested model to non-nested model is proportional to

$$\text{nested prior} : \text{non-nested prior} \propto \frac{(p_B + p_C)^{2\delta-1}}{p_A^{\delta}} \tag{3.35}$$

The interpretation of this simple ratio is that the nested model has a more flattened out probability distribution on the simplex. Greater variability has been achieved by recognizing that ABC does not split directly into three components, but undergoes two binary splits. More generally, the ratio will be proportional to a fraction with numerator equal to the product of intermediate priors. The denominator is a correction term for the weightings given to each component of an asymmetric split.

In fact this is problem specific, as some models will be constructed giving equal a-priori probability to A and BC in the above example. In this case, the result is the same but with no term below the line in Equation (3.35).

For unbiased priors (no knowledge of the nesting structure), this results in asymmetric priors on the simplex that are nonetheless centred about the middle of the simplex (see Figure 3.12). In this case, the isoprobability contours are a-priori convex. However, the posterior may be part concave, unlike the restrictive Dirichlet posterior obtained from a Dirichlet prior and Multinomial likelihood.

Nested Dirichlet-Multinomials

In the case of a Dirichlet prior and Multinomial likelihood for each nest, the comparison between the posterior for the nested model and the equivalent non-nested model is similar to the Dirichlet prior case of the previous subsection. The posterior is Dirichlet (due to conjugacy between the Dirichlet priors and Multinomial likelihoods at each nest). Therefore, the ratio between the nested model posterior and the non-nested posterior is similar to Equation (3.35) and is given by:

$$\text{nested posterior} : \text{non-nested posterior} \propto \frac{(p_B + p_C)^{y_B+y_C+2\delta-1}}{p_A^{\delta}} \tag{3.36}$$

where y_i is the count associated with component i.

Nested Compound-Multinomials

For an overdispersed compound-Multinomial model for the counts, the algebra is not as neat for the ratio of nested to non-nested models. The posterior for a Dirichlet

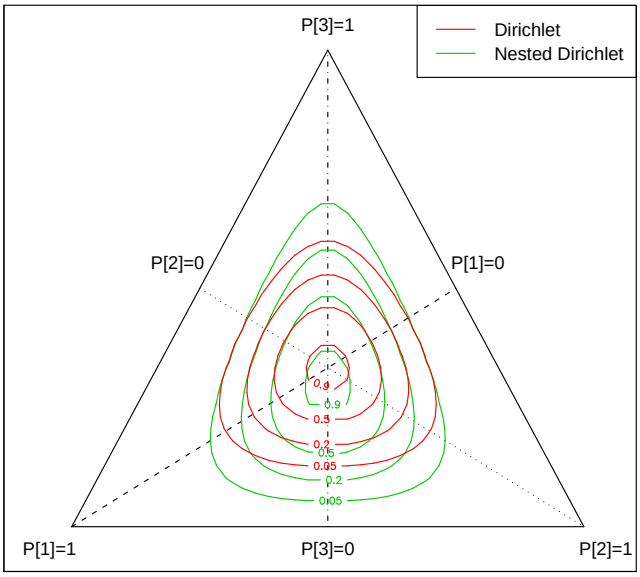


Fig. 3.12: **Dirichlet and Nested Dirichlet priors as per Wong (1998), with the nesting structure shown in Figure 3.11. In fact, Wong (1998) demonstrates this for Generalized Dirichlets, which in his context of constructing priors is equivalent to nested Dirichlets. For the nested model, the first split has prior $B(4, 8)$ and the second is $B(8, 8)$. This is equivalent to a Generalized Dirichlet(a, b) with $a = \{4, 8\}$ and $b = \{8, 8\}$. The non-nested model simply has Dirichlet$(8, 8, 8)$.**

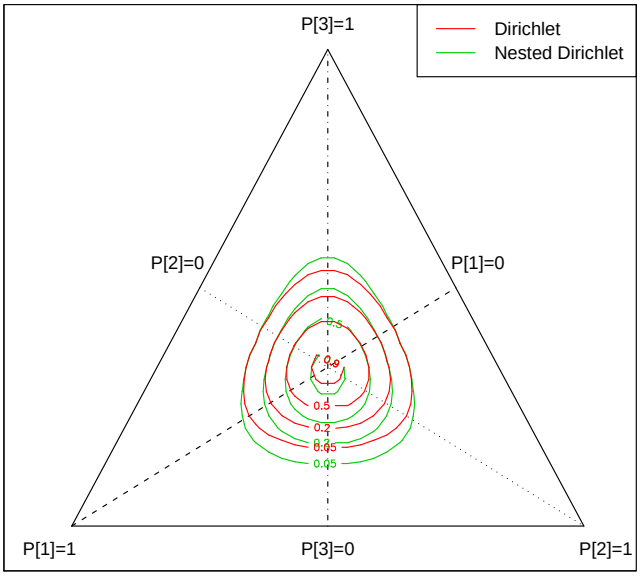


Fig. 3.13: **Dirichlet and Nested Dirichlet posteriors, with the nesting structure shown in Figure 3.11. Three counts, all equal to 5 are observed; priors are as per Figure 3.12 are assigned and isoprobability contours for the resultant posteriors, nested and non-nested are plotted.**

prior and a Multinomial likelihood provide a guideline; the overdispersed compound-Multinomial model for the counts is equivalent to an integral over auxiliary compositional parameters and is therefore an average of such models. The following section shows empirical results for the comparison between nested and non-nested compound-Multinomials in terms of the inverse predictive power when the data are indeed generated from a nested compound-Multinomial likelihood.

Sensitivity to Nesting Structures

A simple toy problem using data simulated from a known model illustrates the impact of a nesting structure. At regular locations in a 15×15 grid, 3 counts are distributed according to a Multinomial distribution of length 3. The 3 parameters of this distribution arise from a Dirichlet mixture of length 3 of a transformation to the simplex space of stochastically smooth fields X (i.e. the likelihood is compound-Multinomial). Each of these 3 counts is subdivided into 3 more counts; again according to a compound-Multinomial distribution.

Therefore, at each location in the $2-D$ regular lattice:

$$\begin{aligned}
\{Y_A, Y_B, Y_C\} &\sim Multinomial(1000, \{P_A, P_B, P_C\}) \\
\{Y_{A_1}, Y_{A_2}, Y_{A_3}\} &\sim Multinomial(Y_A, \{P_{A_1}, P_{A_2}, P_{A_3}\}) \\
\{Y_{B_1}, Y_{B_2}, Y_{B_3}\} &\sim Multinomial(Y_B, \{P_{B_1}, P_{B_2}, P_{B_3}\}) \\
\{Y_{C_1}, Y_{C_2}, Y_{C_3}\} &\sim Multinomial(Y_C, \{P_{C_1}, P_{C_2}, P_{C_3}\})
\end{aligned} \tag{3.37}$$

with all P parameter vectors being Dirichlet mixtures of corresponding compositional vectors ϕ.

$$P \sim Dirichlet(\delta\phi) \tag{3.38}$$

where δ is a scalar controlling the degree of overdispersion.

The vectors ϕ vary smoothly across the location space and the sum-to-unity constraint applies everywhere and at both levels.

The Dirichlet mixtures are proportional to $\prod_{i=1}^{3} p_i^{\delta_i - 1}$ where $\delta_i = \delta\phi_i$. The Multinomial part is proportional to $\prod_{i=1}^{3} p_i^{y_i}$. The compound-Multinomial is the product of these two and is therefore proportional to $\prod_{i=1}^{3} p_i^{\delta_i + y_i}$.

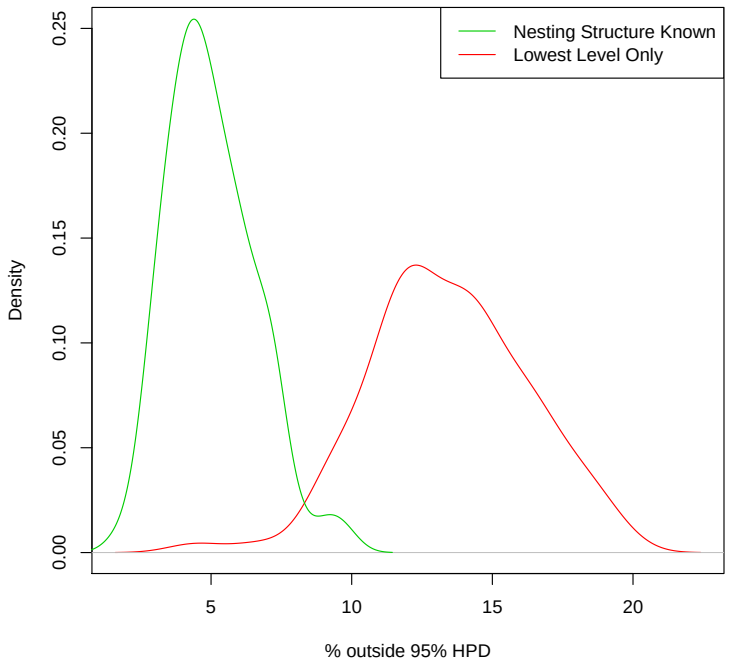


Fig. 3.14: Replications using known parameters P. The overdispersion parameter is set to 10^3 for all components at both levels of the nesting structure. The high degree of overdispersion results in a clear difference between the joint likelihood terms for the nested model and the lowest-level only model.

The full, joint density of the compositional vector at each level is proportional to the product of the 3 lower level compound-Multinomials times the compound-Multinomial for the first level. The non-nested model differs only due to the overdispersion (the likelihood for a nested Multinomial is itself a Multinomial; Section 3.5.6).

Comparison of Figures 3.14 and 3.15 show this result. In both cases, a look at the number of cases generated from the nested model that fall outside the corresponding 95% HPD region for the inverse predictive distribution under the non-nested model is greater than it should be. However, the error is less for a larger overdispersion parameter, corresponding to a lower extra, non-spatial variability. i.e. more overdispersion leads to higher incurred error when nesting is ignored.

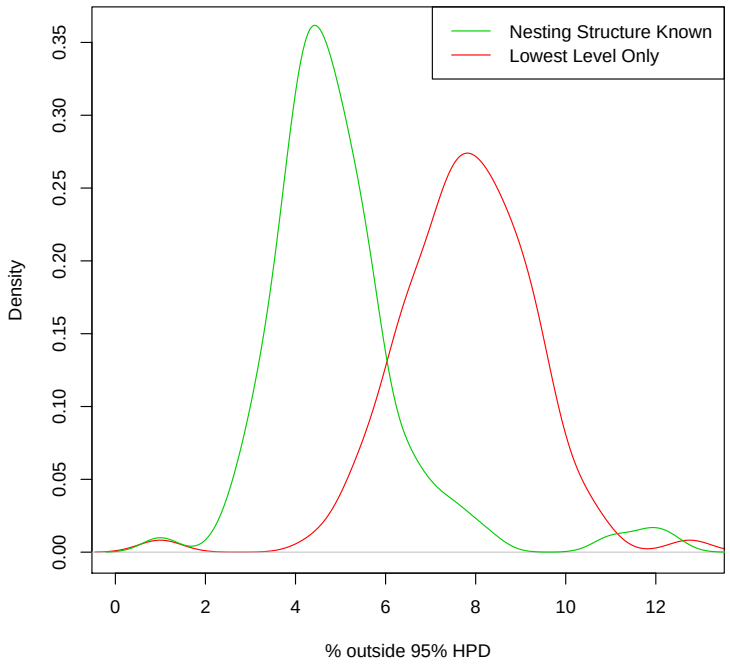


Fig. 3.15: **Replications using known parameters P. The overdispersion parameter is set to 10^5 for all components at both levels of the nesting structure. Note that this represents less overdispersion than Figure 3.14. The low degree of overdispersion results in a small difference between the joint likelihood terms for the nested model and the lowest-level only model.**

3.5.7 Disjoint-Decomposing Compositional Models

Inference using large, compositional models may be difficult or even impossible due to computation and memory constraints, or compatibility issues with inference techniques (see Chapter 4). In these circumstances, approximating the joint model with a decomposable model might provide more accurate results than attempting to use the non-decomposable model.

Although compositional models may appear to be inherently non-decomposable, they do disjoint-decompose provided the modules they are to be disjoint-decomposed into are conditionally independent, given the constraint. For example, given a Multinomial likelihood and a Dirichlet prior, the Dirichlet posterior may be expressed as a product of independent Gamma distributions divided by a Gamma distribution on the sum. Inference for the forward stage of the model may then take place in the unconstrained space with response surfaces fitted to the counts individually. The joint model is then constructed by conditioning on the probability of the sum being equal to unity. The distribution on the sum is a single, straightforward calculation.

An important detail to note here is that the joint problem has been disjoint-decomposed into the product of the unconstrained marginals. The product of the constrained marginals (Betas for the Dirichlet and Binomials for the Multinomial) do not give the correct joint model.

For the inverse predictive stage, given the fitted forward model, the inference cannot be disjoint-decomposed. However, this is typically a far smaller calculation and so can be done using the non-disjoint-decomposed joint model. Inversion of the model amounts to integrating out the latent parameters of the model at each point in the location space to yield the marginal likelihoods at each point.

If the forward stage yields closed form posteriors (as is always the case in this thesis; see Section 4.1), then simple Monte-Carlo integration for the multidimensional integral required for the sum delivers the required distribution. If the forward stage posteriors can only be sampled from (e.g. via MCMC), then MCMC may be used again for the inverse predictive distributions.

Two other options for decomposing joint compositional models are briefly discussed here. Both involve re-expressing the joint model as a product of independent parts; the inverse predictive distribution stage will also disjoint-decompose in these

cases, leading to a further gain in efficiency.

Product of Conditionals

For any joint probability distribution, the following identity holds:

$$\pi(A, B, C) = \pi(A)\pi(B|A)\pi(C|A, B) \tag{3.39}$$

which is a product of univariate distributions.

So, for example, a Multinomial likelihood of length N may be expressed *exactly* as a product of Binomials:

$$\pi(Y; P, n) = \pi(Y_1; n, p_1)\pi(y_2; n - y_1, \frac{p_2}{1-p_1}) \dots \pi(y_i; n - \sum_{j=1}^{i-1} y_j, \frac{p_i}{1 - \sum_{j=1}^{i-1} p_j}) \dots \pi(y_N; y_N, 1) \tag{3.40}$$

the final term being equal to one.

This chain decomposition model may always be used to write the joint model as a product of independent parts and thus disjoint-decompose it exactly. However, it requires knowledge of the conditional distributions in advance; interactions cannot be learned about during the inference procedure as hyperparameters that are unknown cannot appear in more than one module if inference on them is to be parallel.

Furthermore, when dealing with zero-inflated likelihoods, the zero-modification of the unconstrained marginals has a clear interpretation that is consistent with the theory presented in Section 2.4; Equation (3.40) does not have this appealing characteristic as clearly the modules conditioned on more terms will have decreasing zero-inflation.

3.6 Conclusions

A summary of the main points in this chapter is presented. Forward inference has been suppressed in order to focus on modelling issues in isolation. Some sensitivity to model analysis has been done using inference of the inverse problem, given known forward models. Aspects of the models introduced in this chapter are used throughout the later parts of the thesis.

3.6.1 Disjoint-Decomposition of Models

Multivariate models are said to disjoint-decompose exactly into a product of independent modules if there's no interaction terms between modules. Inverse inference may still need to be done using the joint model; for example if there is a sum-to-one constraint. Since the forward model is typically far larger in scale, this is not an issue.

Disjoint-decomposing by (multivariate) marginals is one example: Multiple smooth latent surfaces give rise to multiple types of count; the modules may each account for a single surface. This decomposition will be exact provided the joint model is that the surfaces are conditionally independent given the locations and that the counts are conditionally independent given the surfaces.

In order for a posterior distribution to disjoint-decompose exactly, both the prior and the likelihood should disjoint-decompose. When a joint model does not disjoint-decompose exactly, there might exist an approximation to the joint model that does disjoint-decompose. In this case, the model is said to approximately disjoint-decompose.

A goodness-of-fit measure Δ is used to determine the appropriateness of the disjoint-decomposable model or the quality of the approximation of using such a model when there exists non-zero inter-dependence across multiple spatially smooth surfaces.

3.6.2 Zero-Inflated Models

If the data are zero-inflated, then an appropriate model for zero-inflation must be used. As noted in Section 2.4, mixing the non-zero-inflated likelihood with a point mass at zero gives a flexible model. The size of the mass placed at zero doubles the number of parameters, unless the zero-inflation and the count when present are controlled by the same spatial process in which case a single additional scalar hyperparameter may be all that is required.

Section 3.4.4 used a simple toy example to demonstrate the implications on the inverse inference of using a non-zero-inflated likelihood when the counts are in fact zero-inflated. In this case, inverse inference will be erroneous.

3.6.3 Nested Constrained Models

The Dirichlet-Multinomial type model disjoint-decomposes (given the distribution on the sum); however, the covariance structure is extremely limited. Richer covariance structures are only possible if interactions are either known or inferred so that the logistic-Normal class may be used with off-diagonal terms to model the covariance. This requires either prior knowledge of interaction or else a large number of additional parameters on which inference must be performed (detailed in Chapter 5).

Known nesting structures provide one alternative. They break the problem into separate modules but allow for a rich covariance structure nonetheless; and this without any additional parameters. The central concept of the nested compositional counts model is that counts y_A and y_B may not be conditionally independent given responses x_A and x_B; however, they may be independent given the sum $y_A + y_B$ and the responses $x_A/(x_A + x_B)$, $x_B/(x_A + x_B)$.

The inverse predictive distribution parts still needs joint modelling but the forward part, which is the more labourious, does not. Knowledge of the nesting structure must be known a-priori; this is the only requirement. Sequential inference of the forward models may be performed.

If there is a nesting structure, then inference on the inverse problem will be erroneous if the nesting is ignored. The degree to which the nested and non-nested versions of the model differ is affected by how overdispersed the counts data are.

Chapter 4

INLA Inference and Cross-Validation

This chapter deals with inference and model validation for conditionally independent counts. i.e. it assumes a disjoint-decomposable joint model and inference on the forward model may be performed sequentially for each component of vector assemblages (counts). Issues relating to the disjoint-decomposition of the joint model are dealt with later in Chapter 5.

The inference procedure developed by Rue et al. (2008) is introduced; although this thesis does not contribute substantially to the methodology of this new inference technique, the application to palaeoclimate reconstruction is novel and represents one of the first large applications of the method. The technique is presented in Section 4.1, including details pertaining directly to the palaeoclimate problem application. In fact, the problem is too large for even the INLA method.

Model evaluation and comparison for the inverse problem using cross-validation of the modern dataset was all but impossible using brute force MCMC in Haslett et al. (2006). An approximate cross-validation procedure developed in Bhattacharya (2004) and Bhattacharya and Haslett (2008) offers a faster sampling-based approach. An extension of the inference method of Rue et al. (2008) is developed in Section 4.2. This allows cross-validation in the inverse sense of the model to be performed extremely efficiently (many orders of magnitude faster than re-fitting the model for each left-out datum). Further savings are achieved using computational tricks that are presented along with implications to accuracy.

4.1 The Integrated Nested Laplace Approximation

The Integrated Nested Laplace Approximation (INLA; Rue et al. (2008)) is a new method of performing Bayesian inference on a particular class of problem. It is best suited to Bayesian hierarchical models for which there are a large number of parameters and a small number of hyperparameters, with a specific form of prior covariance on the parameters.

The forward model fitting required in the pollen based palaeoclimate problem is one such problem. In fact, the model as introduced in Haslett et al. (2006) is very well suited to inference via INLA. In Haslett et al. (2006), the model was limited due to computational concerns; computationally intensive MCMC chains were used to sample from the un-normalised posterior for the ten thousand latent parameters in the model. Even after several weeks, the authors admit that "convergence was far from assured".

In contrast with MCMC, the INLA method does not sample from the posterior. It approximates the posterior with a closed form expression. Therefore, problems of convergence and mixing are not an issue. In order to understand how the posterior is approximated, a number of steps are required. The first is a Gaussian Markov Random Field approximation to the posterior for the latent surface, given data and hyperparameters; this is discussed in Section 4.1.1. Subsequently, Section 4.1.3 shows how a simple approximation is built for the posterior of the hyperparameters, given data. Section 4.1.4 shows how more accurate approximations are built for single parameters, if required.

An exhaustive comparison with existing techniques for Bayesian inference, such as MCMC, is not given in this thesis. Rue et al. (2008) provides a more than adequate investigation of both the strengths and weaknesses of the method; it is therefore sufficient here to draw upon those findings. Observations on the suitability of the method to the motivating palaeoclimate problem are given in Chapter 6. Section 4.1.1 shows how the method can work even for uncommon, bimodal likelihoods such as zero-inflated models.

4.1.1 The Gaussian Markov Random Field Approximation

Multivariate normal priors are frequently assigned to the latent surfaces in a hierarchical model to induce a-priori smoothness of the non-parametric surfaces. This is particularly common in spatial statistics, but the technique can be used for any problem in which the only prior on a large set of parameters with locations / distances is that they vary smoothly (see Rue and Held (2005) for details and examples). The smoothness hyperparameter is taken as known in this section; Section 4.1.3 demonstrates the construction of the posterior for this and other model hyperparameters.

If the structure of the prior is Markov (defined on a regular grid), then the prior is a GMRF (Section 2.2.4). Assignment of such priors is common; in fact, this was the prior used for the response surfaces in Haslett et al. (2006).

When the likelihood for data Y given parameters X is expressible as a multivariate normal, then given a multivariate normal prior on X, the posterior $\pi(X|Y)$ is multivariate normal (due to self-conjugacy the normalising constant has an analytical solution):

$$\pi(Y|X) = MVN(Y; X, Q_Y) \tag{4.1}$$

$$\pi(X) = MVN(X; \mu_X, Q_X) \tag{4.2}$$

$$\pi(X|Y) = MVN(X; (Q_X + Q_Y)^{-1}(Q_X\mu_X + Q_Y\overline{Y}), Q_X + Q_Y) \tag{4.3}$$

The dimension of the precision matrices Q_X and Q_Y is the square of the dimension of X and Y.

If the likelihood has a diagonal covariance matrix, due to conditional independence given the parameters, then the precision matrix is also diagonal. Thus, the posterior precision matrix is the same as the prior precision matrix with terms added on to the diagonal.

When the likelihood is not expressible as a multivariate normal then the posterior is not multivariate normal. However, a simple and fast approximation to the log-likelihood leads to a Gaussian approximation for the posterior. The un-normalised posterior is expressible exactly as:

$$\pi(X|Y) \propto exp(-\frac{1}{2}X^T Q_X X + \sum_{j=1}^{N} log\pi(y_j|x_j)) \tag{4.4}$$

for a zero mean Gaussian prior with precision matrix Q_X.

If the log-likelihood is approximated with a second order Taylor series, then the approximation is quadratic. Hence, the posterior is expressible as a multivariate normal:

$$\begin{aligned} \pi(X|Y) &\simeq exp(-\frac{1}{2}(X-\mu)^T(Q_X + diag(c))(X-\mu) && (4.5)\\ &= \tilde{\pi}_G(X|\theta, Y) && (4.6) \end{aligned}$$

$\tilde{\pi}_G(X|\theta, Y)$ is then known as the Gaussian approximation, the $\tilde{\pi}$ denoting that it is an approximation and the $_G$ standing for Gaussian.

The posterior mode μ and the c terms must still be found. c contains the elements of the likelihood precision that are added to the prior precision diagonal to form the posterior precision matrix.

An important point to note here is that the log-likelihood is a function of the parameters X given the data Y. This is what is approximated with a quadratic in X. Likelihoods that are not approximately quadratic as functions of the data given the parameters (such as common probability mass functions like the Poisson, the Binomial, etc) are often adequately approximated as quadratic in X. See Figure 4.1.

The posterior mean μ is found by Newton-Raphson or a similar iterative search algorithm. The c terms in Equation (4.6) are simply the second order coefficients of the Taylor series expansion. An important caveat is that the data are modelled as conditionally independent given the parameters; each Taylor series expansion is univariate and thus a simple and fast calculation. Writing the log-likelihood as $f[x]$, the Taylor series to second order is:

$$\begin{aligned} f[x] &\simeq f[x_0] + f'[x_0](x-x_0) + \frac{1}{2}f''[x_0](x-x_0)^2 \\ &= a + bx - \frac{1}{2}cx^2 && (4.7) \end{aligned}$$

with $b = f'[x_0] - f''[x_0]x$ and $c = -f''[x_0]$ (a is not required). Thus $b = f'[x_0] + cx$.

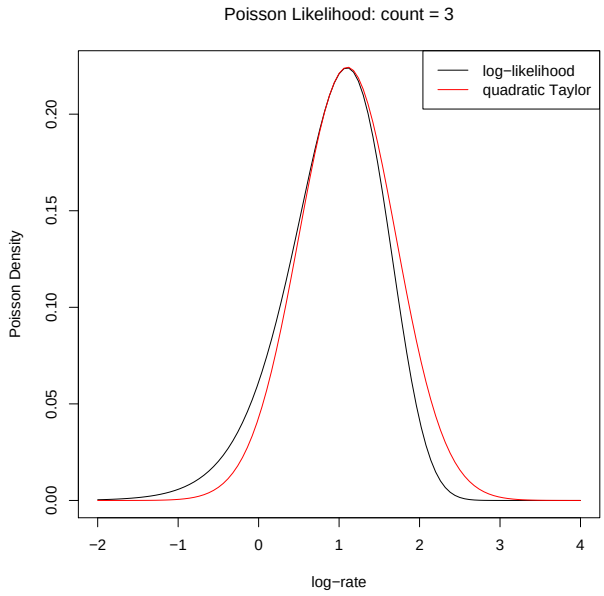


Fig. 4.1: **Although the Poisson with a low rate parameter is a mass function is poorly approximated by a quadratic, the Poisson for a given count (in this case 3) is adequately approximated by a Gaussian (quadratic) as a function of the log-rate.**

The search algorithm for finding the mean μ represents the bulk of the work in finding the Gaussian approximation. Equivalent to Newton-Raphson is iteratively solving the system of linear equations

$$(Q_X + diag(c))\mu = b \tag{4.8}$$

for μ.

The key to solving this equation fast is that the Markovian structure of Q_X (and thus $Q_X + diag(c)$) leads to the matrix being very sparse. A Cholesky decomposition of a matrix Q renders the matrix as the product of a lower triangular matrix with its own transpose:

$$Q = LL^T \tag{4.9}$$

Solving Equation (4.8) for μ then amounts to solving $Lv = b$ for v and then solving $L^T\mu = v$ for μ. All calculations are swift and potentially large matrices may be stored cheaply due to sparseness and lower diagonality.

The Taylor series expansion is most accurate if centred on the mode (mean) of the posterior. Therefore, at each iteration of the search algorithm, the Taylor series is recalculated. This still represents a small number of computational steps as the algorithm typically converges quickly (less than 10 iterations for even very large problems, such as 10^4 parameters with precision matrices of order $10^4 \times 10^4$).

Univariate Toy Example

A simple toy example illustrates the Gaussian approximation. For a single Poisson count $y = 3$ the objective is to construct the posterior probability distribution for the rate. As the rate is strictly positive, a log-link is used and the log-rate x is inferred.

A diffuse Gaussian prior on x with mean zero and precision κ of 0.001 (variance $= 1/\kappa$ of 10^3) is placed on the log-rate parameter. The likelihood is assumed to be Poisson. As the problem is univariate, numerical integration is easily used to determine the posterior for the log-rate. Gaussian approximations formed by expanding a second order Taylor series around 4 different values of the log-rate are shown for comparison in Figure 4.2.

The posterior distribution is then

$$\begin{aligned} \pi(x|y) &\propto \pi(x)\pi(y|x) \\ &= exp(-\frac{\kappa}{2}x^2 + yx - exp(x)) \end{aligned} \tag{4.10}$$

To approximate this we construct a quadratic Taylor expansion of the unnormalised log-likelihood $yx - exp(x)$ around a suitable x_0. The univariate approximation is now

$$\tilde{\pi}(x|y) \propto exp(-\frac{c+\kappa}{2}x^2 + bx) \tag{4.11}$$

which is in the form of a Normal distribution with mean $\frac{b}{c+\kappa}$ and variance $\frac{1}{c+\kappa}$ and from Equation (4.7):

$$\begin{aligned} c &= e^{x_0} \\ b &= yx_0 - e^{x_0} + cx_0 \end{aligned} \tag{4.12}$$

The mode x_0 is found by iterating the Taylor series expansion with setting $x_0 = (c + \kappa)b$. This is equivalent to using Newton-Raphson to find the mode of the posterior. Convergence for this example takes only two steps.

Even if the quadratic approximation to the log-likelihood is poor, the posterior approximation may be good. An example of this is when the likelihood is a zero-inflated Poisson (see Section 2.4 and Equation (2.18)) given by

$$\pi(y|x) = \begin{cases} 1 - q + qe^{\lambda} & y = 0 \\ q\text{Poisson}(y;\lambda) & y > 0 \end{cases} \tag{4.13}$$

with $q = logit^{-1}(x)$ and $\lambda = e^x$.

Although the likelihood as a function of count y is bimodal (a mixture of a point mass at zero and a Poisson), as a function of the log-rate x it is unimodal. However, it is very skewed and thus the quadratic approximation to the likelihood is poor; see Figure 4.3(a). The posterior, even under a diffuse prior, is necessarily much less skewed and thus the Gaussian approximation to the posterior is much better; Figure 4.3(b).

Multivariate Toy Example

A multivariate example shows the difference between inference using a GMRF approximation and MCMC sampling. A smooth (Gaussian) curve X is defined at regular points in a 1-dimensional space. Counts data are generated from a zero-inflated Poisson with rate parameter given by the exponential of the smooth curve and probability of potential presence given by the inverse logit:

$$\pi(y_j|x_j) = \begin{cases} 1 - q_j + q_j e^{\lambda_j} & y_j = 0 \\ q_j \frac{e^{-\lambda_j}\lambda_j^{y_j}}{y_j!} & y_j > 0 \end{cases} \tag{4.14}$$

with $q_j = \frac{e^{x_j}}{1+e^{x_j}}$ and $\lambda_j = e^{x_j}$.

An intrinsic GMRF prior is placed on X with precision matrix given by

$$Q = \kappa R \tag{4.15}$$

where R is as per Equation 2.11 and κ is set to ensure smoothness of the curve across the locations.

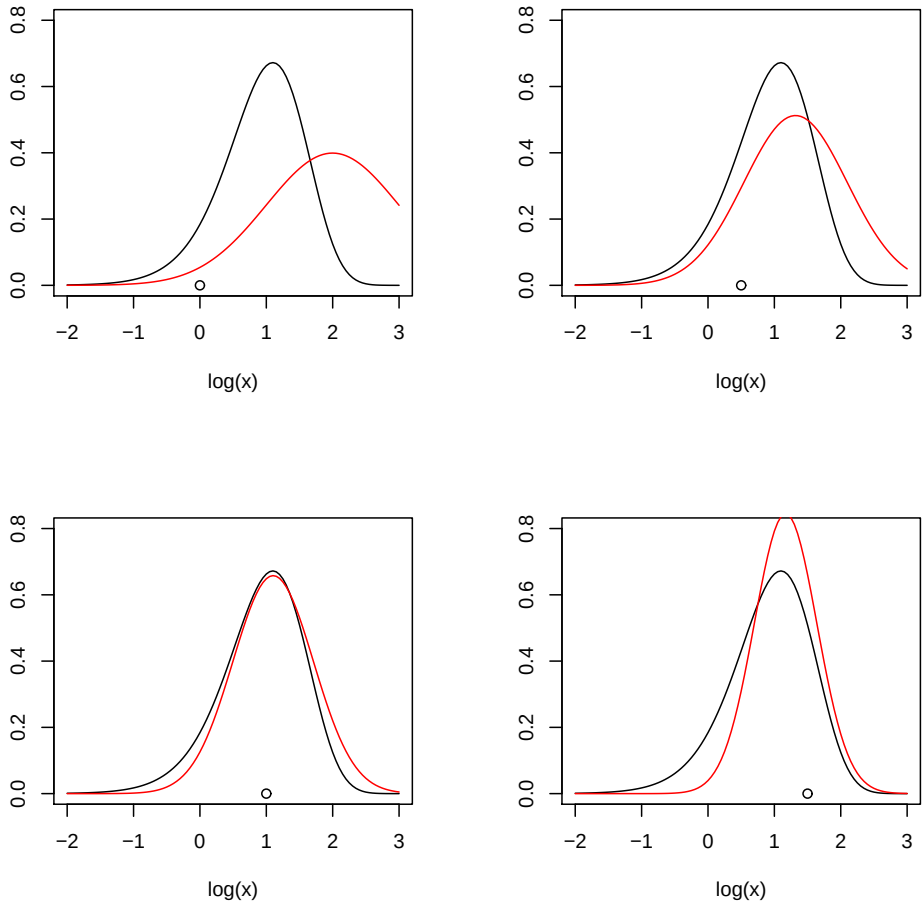
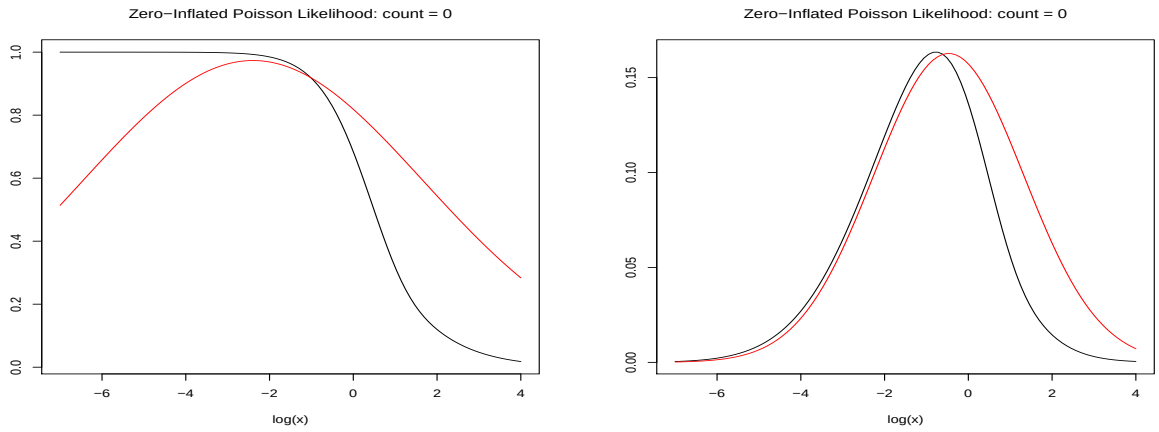


Fig. 4.2: **The effect of varying the point around which the Taylor series is expanded. The true posterior is shown in black along with the Gaussian approximation formed through a Taylor series expansion of the log-likelihood in red. A circle shows the location of the point around which the Taylor series is formed. The Taylor series is most accurate at the centre; thus the Gaussian approximation to the posterior is most accurate when the quadratic approximation to the likelihood is centred at the posterior mode (mean).**



(a) Likelihood

(b) Posterior

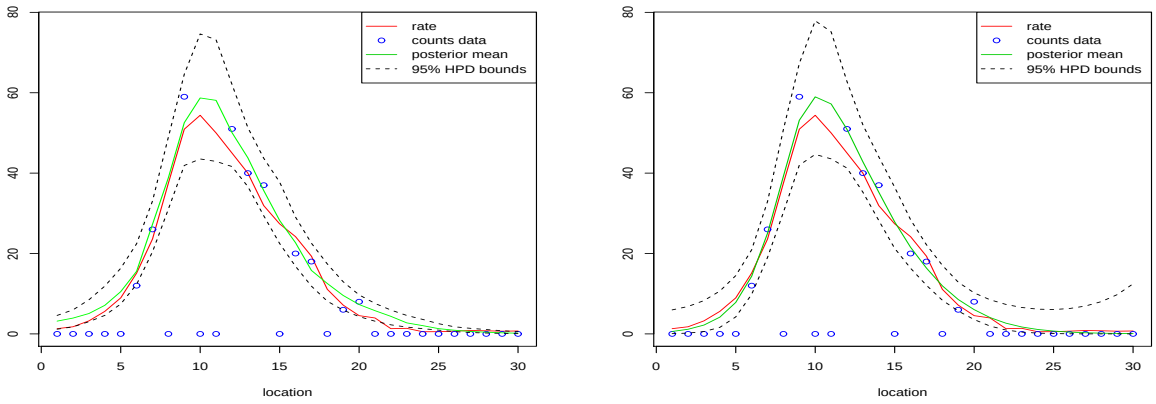
Fig. 4.3: **A zero-inflated Poisson likelihood (black line) given a count of zero is poorly approximated by a quadratic (red line) in (a). The posterior for the log-rate (black line) is more adequately approximated by a Gaussian (red line) in (b).**

The goal is then to infer the posterior $\pi(X|Y)$:

$$\pi(X|Y) \quad \propto \quad exp\left(-\frac{1}{2}X^T QX + \sum_{j=1} log\pi(y_j|x_j)\right) \tag{4.16}$$

Both a Metropolis Hastings MCMC algorithm and a GMRF approximation were coded and the results for both appear in Figure 4.4. There are 30 locations, regularly spaced, with counts data generated for each of them. To run an MCMC chain of length 3×10^5 for the 30 latent parameters took about 4 minutes. In contrast, it took just under a fifth of a second to fit the GMRF approximation. This represents a speedup in performance of around 3 orders of magnitude for similar results. Even then, the MCMC sampler was started in the correct place to avoid burn in and the code was optimised and tweaked to achieve a good acceptance rate for the proposals.

The most obvious difference between the two Figures is to the right hand side. It appears that the GMRF approximation is overestimating the response (which is very low) in this region. However, examination of the trace plot for the log of the response shows that in fact, the Metropolis-Hastings routine is mixing poorly in that region (Figure 4.5). This is an illustration of why it is not correct to assume an MCMC sampler is doing a better job of approximating the posterior than the



(a) MCMC (b) GMRF

Fig. 4.4: **Zero-inflated counts data (blue circles) arise from a rate (red line) which varies smoothly across discrete locations (x-axis). A Metropolis-Hastings MCMC algorithm samples from the true posterior, given an intrinsic multivariate normal prior on the log of the rate. The results are shown in (a) as the posterior mean (green solid line) and 95% highest posterior density bounds (black dashed lines) for the rate. A GMRF approximation to the posterior is similarly shown in (b).**

The MCMC chain was run for 10^5 iterations and then thinned by selecting every 10^{th} iteration.

Both inference methods give good results, reconstructing the "true" response (red line) well. However, the GMRF method achieves results around 3 orders of magnitude faster.

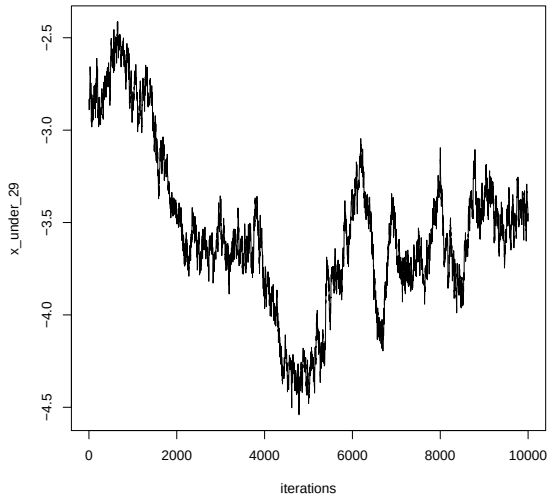


Fig. 4.5: **A trace plot of the log-rate at the second location from the right in Figure 4.4. Despite proposals that are tweaked to ensure good acceptance rates and thinning to** 10% **of a long run of length** 10^5 **iterations, the MCMC algorithm does not mix well. Mixing is not an issue for approximations to the posterior.**

GMRF approximation.

4.1.2 Spatial Zero-Inflated Counts Data

Section 4.1.1 shows how the Gaussian approximation may be applied even to models with mixture likelihoods, such as a zero-modified distribution. A multivariate example of this is shown in Section 4.1.1, where a GMRF prior and zero-inflated Poisson likelihood yield a posterior that is adequately approximated by a GMRF. In this example, the rate and the probability of potential presence are both functions of a single stochastic process.

It will be shown in Chapter 6 that such a model is suitable in the context of pollen based ecology. Where such a model is unsuitable, two separate processes should be modelled; one for the response and another for the probability of presence. However, such a model is unsuited to approximation with a GMRF. That each datum depends only on a single parameter is a central assumption (or modelling choice) in fitting such approximations. This is because approximating the likelihood with a quadratic

requires the function to be univariate in the latent parameter; a second order Taylor series expansion for a bivariate function has 8 terms, compared to 2 for an expansion of a univariate function.

In some cases, such as the incorporation of zero-mean random effects, the model may simply be reparameterised so that each datum depends on a single parameter, which is the sum of the spatial part plus the non-spatial random effect (see Rue and Held (2005), Chapter 1). This typically leads to a graph / GMRF of double the size and thus double the number of parameters, but is not a fundamental obstacle to the implementation of the approximation.

However, a similar reparameterisation does not exist for the zero-modified distribution; the likelihood necessarily depends on two distinct parameters. To see this, if the zero-inflated counts data depend on, say, x for the response when present and z for the probability of potential presence, then the second order Taylor series expansion of the log-likelihood $f[x, z]$ is

$$\begin{aligned} f[x,z] &\simeq f[x_0,z_0] + f_x[x_0,z_0](x-x_0) + f_z[x_0,z_0](z-z_0) \\ &+ \frac{1}{2}\left(f_{xx}[x_0,z_0](x-x_0)^2 + f_{zz}[x_0,z_0](z-z_0)^2\right) \\ &+ f_{xz}[x_0,z_0](x-x_0)(z-z_0) \end{aligned} \tag{4.17}$$

where f_x is the first order derivative w.r.t. x of f, f_z is the first order derivative w.r.t. z of f, f_{xx} is the second order derivative w.r.t. x of f, f_{zz} is the second order derivative w.r.t. z of f and f_{xz} is $\frac{df^2}{dxdz}$.

This cannot be re-expressed as a function of a single variable. The Taylor series remains multidimensional and is cumbersome. Equation (4.17) cannot be expressed as $(y-y_0)^T Q_Y (y-y_0)$ and is therefore incompatible with the GMRF approximation technique.

In the case of a single underlying process controlling both response and the probability of potential presence, modelling two separate spatial processes will necessarily perform worse than modelling the single process. For example, if MCMC methods are used, the two spatial parts will be very highly correlated (as they in fact arise from a single process), leading to poor mixing, as shown in Figure 4.6. This is a consequence of the overparameterisation of the model.

A note on zero-inflated models lying on a grid: The GMRF approximation method requires all calculations to be defined on a grid across the locations space

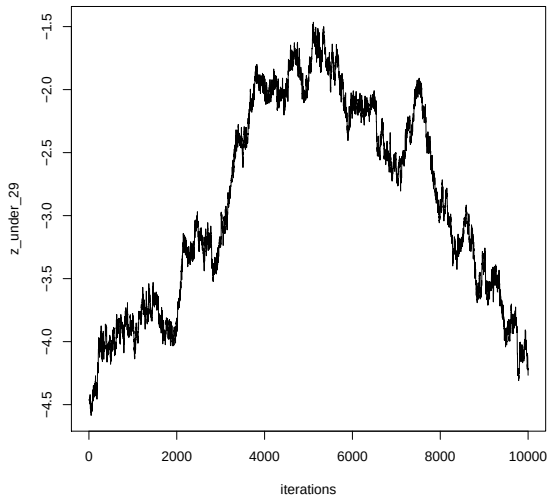


Fig. 4.6: **A trace plot of the logit of the probability of potential presence at the second location from the right in Figure 4.4 when two distinct processes are modelled. Despite proposals that are tweaked to ensure good acceptance rates and thinning to 10% of a long run of length 10^5 iterations, the MCMC algorithm does not mix well. The model is overparameterised as there are double the number of random variables.**

(due to the Markov structure). In Rue and Held (2005) and Rue et al. (2008), the authors advocate pushing each datapoint to the nearest gridpoint. This will typically keep the locations within the measurement error as the granularity of the grid may be very large. When two or more datapoints are closest to the same grid-point they are simply aggregated or averaged. However, this has implications for zero-inflated counts data as the distributions are not closed under addition. Therefore, in this thesis, the data are pushed to the nearest gridpoint but are kept separate so that there may be multiple counts per gridpoint.

4.1.3 Posterior for the Hyperparameters

In a hierarchical model, some of the hyperparameters may be unknown. They must be treated as random variables and inferred from the data. An MCMC algorithm will typically sample from the joint posterior for the parameters and the hyperparameters given the data $\pi(X, \theta|Y)$. This posterior may also be expressed as the product of the marginal posterior for the hyperparameters $\pi(\theta|Y)$ and the posterior for the parameters given the hyperparameters $\pi(X|\theta, Y)$:

$$\pi(X, \theta|Y) = \pi(\theta|Y)\pi(X|\theta, Y) \tag{4.18}$$

The objective then is to find $\pi(\theta|Y)$ and $\pi(X|\theta, Y)$. Re-expressing Equation (4.18) as

$$\pi(\theta|Y) = \frac{\pi(X, \theta|Y)}{\pi(X|\theta, Y)} \tag{4.19}$$

and using a GMRF approximation for the denominator $\pi(X|\theta, Y)$ gives the *Laplace approximation* for the posterior of the hyperparameters:

$$\pi(\theta|Y) \simeq \left.\frac{\pi(X, \theta|Y)}{\tilde{\pi}_G(X|\theta, Y)}\right|_{X=X^*(\theta)} \tag{4.20}$$

and writing the numerator in closed form, up to proportionality constant gives

$$\tilde{\pi}_L(\theta|Y) \propto \left.\frac{\pi(\theta)\pi(X|\theta)\prod\pi(Y_i|X_i, \theta)}{\tilde{\pi}_G(X|\theta, Y)}\right|_{X=X^*(\theta)} \tag{4.21}$$

where $\pi_L(\theta|Y)$ is called the Laplace approximation to the posterior.

Integration over the Hyperparameters

As stated previously, the type of problem to which the INLA routine is well suited is one in which the hyperparameters are essentially nuisance parameters. It is acceptable to treat them crudely, with a view to integrating them out entirely. Evaluating Equation (4.21) at a number of discrete points on a grid over all θ is the approach taken in Rue et al. (2008). The normalised approximate posterior $\pi_L(\theta|Y)$ is found by summing Equation (4.21) across the points and dividing by the sum.

The marginal posterior for the parameters is then simply the weighted sum over all the discrete points in θ space:

$$\tilde{\pi}(X|Y) = \sum_k \tilde{\pi}_G(X|\theta_k, Y) \times \tilde{\pi}(\theta_k|Y) \times \delta_k \tag{4.22}$$

where δ_k are area weights, depending on the not-necessarily uniform discrete spacings in the θ space. This is only possible for low dimensional θ, otherwise the numerical integration quickly becomes difficult as the number of gridpoint grows as N^D, where N is the number of points along a particular component of θ and D is the number of components of θ.

Rue et al. (2008) shows that this is a good approximation for a wide range of statistical problem. It is sufficient to note that the MCMC alternative is cumbersome to the point of rendering it impractical. The hyperparameters are often highly correlated with the parameters; e.g. the smoothness hyperparameter and associated latent field random variables. Thus a Metropolis-Hastings algorithm will suffer from serious mixing problems unless a tailored block-update proposal mechanism is built.

Approximation Using Modal Hyperparameters

A further approximation is to use the modal value of the hyperparameters. The hyperparameters are then considered known, having been inferred from the data for the purpose of constructing the marginal posterior for the parameters given the data (maximum a-posteriori selection of hyperparameters; see, for example Oakley and O'Hagan (2002)).

This approximation is motivated by the inverse problem application central to this thesis. In order to invert the model and construct inverse predictive distributions for unknown hyperparameters, the model must be inverted (inverse problem) for

each value on the discrete grid of hyperparameters. The marginal distribution for the posterior of the parameters given the data only is represented as a mixture of Gaussians (Equation (4.22)). This is itself modelled non-parametrically and a numerical integration step would be added to each marginal likelihood calculation for the inverse predictive procedure.

The decision on whether to integrate out the hyperparameters or to simply use the modal values of the hyperparameters is a trade off between speed and accuracy. The decision is context specific, depending not only on the data and the model used and the number of hyperparameters, but also the task that is being performed. For example, in order to perform actual inverse predictions given counts data the gain in accuracy associated with integrating over the hyperparameters may well justify the additional effort, which is not great.

In cross-validation terms, however, the increase in computation is enormous, as there will be k replications where k is the number of datapoints; furthermore, the modal approximation will place the bulk of the probability mass in the correct location so that the inverse predictive distributions are approximately equal for parameters given fixed (modal) hyperparameters and the marginal parameters. Therefore, use of hyperparameters fixed at their modal posterior values is advocated for the task of cross-validation. This is discussed for the pollen dataset in Section 6.4.1, with results.

4.1.4 Laplace Approximation for Parameters

If the posterior for an individual latent parameter (individual location on the non-parametric latent field) given the hyperparameters $\tilde{\pi}(x_j|\theta_k, Y)$ is required, then the simplest univariate approximation is the marginal Gaussian, taken from the joint GMRF approximation. This is already constructed when finding the Laplace approximation for the hyperparameters posterior and is therefore very cheap. The marginal variances are computed using a fast recursion algorithm applied to the already available Cholesky decomposition of the joint precision matrix (see Rue and Held (2005) or Rue et al. (2008)).

When more accurate approximations are required, there is a Laplace approximation for the marginal distributions of the parameters given by (Rue et al. (2008))

$$\tilde{\pi}_L(x_j|\theta_k, Y) \simeq \left.\frac{\pi(X|\theta_k, Y)}{\tilde{\pi}_{GG}(X_{-j}|x_j, Y, \theta_k)}\right|_{X_{-j}=X^*_{-j}(x_j\theta_k)} \tag{4.23}$$

where $\tilde{\pi}_{GG}(X_{-j}|x_j, Y, \theta_k)$ is a Gaussian approximation to the conditional posterior for all X except x_j, evaluated at the k^{th} point of the discrete grid over θ. This Laplace approximation involves re-fitting a GMRF approximation for each θ_k and each x_j and is thus expensive computationally.

Approximations to this approximation may be made, such as fitting a skew-normal via a third order Taylor expansion (called a Simplified Laplace Approximation in Rue et al. (2008)). Both the computational effort and the accuracy will fall between the Gaussian approximation and the full Laplace approximation. This leads to a method of checking the accuracy of the approximations: If the Kullback-Leibler divergence between Equation (4.22) using the Gaussian and the Simplified Laplace approximations is small, then both are deemed acceptable. Otherwise, the Kullback-Leibler divergence between the Simplified Laplace and Laplace approximation versions of Equation (4.22) are similarly compared. If these again differ, then the Laplace is the best estimate, but adequacy of the approximation is not determined (see again Rue et al. (2008)).

Of course, if the one of these more accurate approximations are calculated in order to assess the relative accuracy of the Gaussian approximation then they should simply be used instead. The computational convenience of the simpler Gaussian approximation is then lost. An intermediary solution is to calculate the more accurate approximations at a random subset of points; the accuracy of the Gaussian approximation at these locations then gives an indication of the accuracy across all locations. If the Gaussian approximation is thereby deemed to be of insufficient accuracy then the more accurate approximations are required.

4.1.5 Approximation for Parameters: Inverse Problem

The method of checking the accuracy of the approximations given above is for the forward problem. The interest here is in the inverse problem; if the Gaussian approximation and the more accurate (but more costly) Laplace type approximations give the same answer for the inverse predictive distributions, then there is no gain.

This is an indication that the Gaussian approximation is sufficiently accurate for the requirements of the inverse problem.

In order to invert the model, the marginal posteriors must be calculated at *all* locations. In regions of little data, the GMRF prior will dominate and the posterior will be close to GMRF (in regions of no data, the posterior will converge to the prior). Where there is an abundance of data, the Central Limit Theorem will apply and, again, the posterior will be approximately Gaussian. The Simplified Laplace involves a third order term for the Taylor series approximation to the posterior; this is the skew term. Only if the marginal posterior in a given location has significant skew will the Laplace approximation be closer to the actual posterior than the Gaussian approximation.

4.2 Cross Validation

Model comparison and evaluation is of fundamental concern in all statistical data analysis. As previously stated, the concern here is in the inverse problem (predicting input variable from a given model output variable). For this reason, model evaluation in the inverse sense is the main focus. Specifically, the ability of the model to successfully reconstruct or predict the correct location of a new count (or vector of counts) is what determines the usefulness of the model.

As a never ending supply of new data is not available and the model will ultimately be used to make unverifiable reconstructions, validation must be done on the same training data set to which the forward model is fitted. Cross-validation is a common technique for assessing the model fit to data. M-fold cross-validation refers to the practice of leaving out a subset of the data ($1/M$ of all the training dataset); the model is fitted to the remaining data ($(M-1)/M$ of the total) and the ability of the model to predict (or reconstruct) the remaining data is assessed. The concern here is with leave-one-out cross validation; this involves refitting the model to all but one datum and then assessing the ability of the model to predict the single left out point. This must be repeated M times, where M is the total number of datapoints (7742 in the motivating palaeoclimate problem). From here on, leave-one-out cross-validation will simply be referred to as cross-validation as

this will be the only form of cross-validation considered.

Performing cross-validation in the forward sense for the palaeoclimate problem involves fitting the model to all but a single pollen count (or assemblage) and then predicting that count (or assemblage), given the associated climate (location in climate space). Cross-validation in the inverse sense is more appropriate. The main reason for this is that the ultimate goal of the project is to make inferences on climate, given a vector of fossil pollen counts.

Interest here therefore lies in forming the leave-one-out inverse cross-validation predictive posteriors for the locations. These are then compared in some way with the actual observed values. So, for each observation j:

$$\pi(l_j|\tilde{Y}_{-j}, \tilde{L}_{-j}, \tilde{y}_j) \propto \int \pi(\tilde{y}_j|X, l_j)\pi(l_j)\pi(X|\tilde{Y}_{-j}, \tilde{L}_{-j})dX \tag{4.24}$$

In either case, the main difficulty in performing cross-validation is computational. For example, to perform exact cross-validation using a brute-force approach involves fitting the model M times, where M is the number of datapoints. The motivating palaeoclimate problem has just under eight thousand modern / training data collection sites. Given that an MCMC sample of the saturated posterior (the posterior of the parameters given all training data data) takes of the order of two weeks to run, eight thousand replications would take 300 years. Some existing techniques are introduced in Section 4.2.1. A new approach for the cross-validation in the inverse sense, made possible by the closed form of the forward model posteriors, is explored in Section 4.2.3. This represents a novel contribution to the INLA methodology.

Looking at Equation (4.24), the brute force approach is to compute the leave-one-out posterior for the responses $\pi(X|\tilde{Y}_{-j}, \tilde{L}_{-j})$ for each left-out datum pair $\{l_j, y_j\}$. The integral on the right hand side of Equation (4.24) must then be computed for each j and this represents further computational labour as it is an intractable integral. The contribution in this work is to deliver a method for fast updates to the saturated posterior (available in closed form via the INLA method) to be made. The leave-one-out posteriors are thus available without needing to recompute the INLA approximations for each leave-one-out subset of the data. This novel extension of the INLA methodology is detailed in Section 4.2.3 and is a consequence of the Multivariate-Normal form of the GMRF posterior for X.

The other challenge in performing cross-validation is how to summarise the model fit. There are many alternatives, some of which are described in Section 4.2.5.

4.2.1 Importance Resampling

There is a large literature on performing fast Bayesian cross-validations for the forward problem that focus on methods for speeding up the MCMC re-runs. One approach involves using importance resampling. (see, for example, a review in Bhattacharya and Haslett (2008)).

Using an idea that first appeared in Gelfand et al. (1992), samples from the saturated posterior are re-used so that MCMC is not needed for the leave-one-out cross-validations. Specifically, if the data are $\{Y, L\}$ (outputs and inputs to the model, respectively) and the Bayesian hierarchical model is fitted via MCMC sampling of the latent parameters X, then the saturated posterior is $\pi(X|Y, L)$. This gives the importance sampling distribution.

For a particular left out datum y_j, the posterior predictive distribution $\pi(y|l_j, L_{-j}, Y_{-j})$ is desired. This is calculated from the integral

$$\pi(y|l_j, L_{-j}, Y_{-j}) = \int_X \pi(y|l_j, X)\pi(X|L_{-j}, Y_{-j})dX \tag{4.25}$$

Integration is via MCMC. Following the notation in Vehtari and Lampinen (2002), if samples from $\pi(X|L_{-j}, Y_{-j})$ are available - denoted $\dot{X}^h$ - then a sample $\dot{y}^h$ from $\pi(y|l_j, \dot{X}_j^h)$ is a sample from $\pi(y|l_j, L_{-j}, Y_{-j})$. But $\pi(X|L_{-j}, Y_{-j})$ is the undesirable MCMC repetition. A sample from this distribution is more easily obtained through importance resampling; if X^h is a sample from the saturated posterior $\pi(X|L, Y)$, then samples $\dot{X}^h$ may be obtained by resampling $\dot{X}^h$ with importance weights given by

$$w_j[h] = \frac{\pi(X^h|L, Y_{-j})}{\pi(X^h|L, Y)} = \frac{1}{\pi(y_j|l_j, X^h)} \tag{4.26}$$

This is analytically available, since it is the likelihood function with known parameters. The weights are thus simple and quick to calculate and samples from the importance sampling distribution (the saturated posterior) are already available from the initial MCMC run.

4.2.2 Cross-Validation in Inverse Problems

When model inputs L are of lower dimension than outputs Y, cross-validation of the inverse problem may offer an appealing alternative to forward cross-validation. Discrepancy measures between the data and the corresponding predictive densities are computed; these measures are easier to construct and interpret for the lower dimensional variable. Furthermore, when the ultimate interest is in predicting l for a given y (the inverse problem), then inverse cross-validation is more appropriate.

However, there is very little literature on cross-validation in inverse problems. In fact, according to Bhattacharya (2004)

> "we do not know of any paper that discusses cross-validation in the context of inverse problems"

Therefore, the technique introduced in Bhattacharya (2004) and presented in detail in Bhattacharya and Haslett (2008), provides the only benchmark. In that paper, the authors point out that importance weights for the inverse problem are not tractable calculations. They suggest using a posterior for a single left-out point (obtained via regular MCMC) as the importance sampling density and show how to calculate the importance weights for the other points. Sampling without replacement is also advocated, to protect against highly variable importance weights. They call the algorithm Importance Re-sampling MCMC or IRMCMC.

Although there is still an MCMC step in the IRMCMC algorithm, it is of much lower dimension than re-running MCMC for each left out point. Another important detail is the choice of the initial left-out point; Bhattacharya and Haslett (2008) demonstrate methods for making this choice.

For cross-validation in inverse problems, IRMCMC not only runs much faster than doing multiple regular MCMC runs, but it also achieves superior mixing (and thus more accurate results). This is due to the low dimensionality of the inverse problem, compared to the forward problem. The MCMC runs in the IRMCMC algorithm explore the typically multimodal target distribution of l better as updates for X do not have to be run in parallel (see Bhattacharya and Haslett (2008) for discussion and examples). In fact, IRMCMC may be thought of as regular MCMC with a special proposal mechanism.

4.2.3 Fast Augmentation of the Multivariate Normal Moments

The INLA method introduced in Section 4.1 delivers closed-form approximations to the saturated posterior for the latent parameters X. The same weightings in Equation (4.26) may again be used to perform **forward** cross-validation without MCMC. The marginal saturated posterior $\pi(x_j|L,Y)$ (available analytically) is weighted using Equation (4.26) and the uni-dimensional integral in Equation (4.25 is computed using fast numerical methods, such as Gauss-Hermite quadrature (see Rue et al. (2008)).

This section introduces a novel method for fast **inverse** cross-validation using the INLA methods. When the saturated posterior for the latent parameters X is approximated with a GMRF (as per Section 4.1.1) then augmenting this posterior to correct for a left-out datum is straightforward. The posterior is entirely specified by the first two moments; fast updates to these to correct for left out data negates the need to re-fit the entire model.

If both the prior and the posterior are multivariate normal, then the posterior covariance matrix is given by

$$\Sigma = (Q + R)^{-1} \tag{4.27}$$

where Q is the prior precision matrix and R is the likelihood precision matrix and is diagonal.

The posterior covariance matrix for the case of leaving out the j^{th} point is similarly given by

$$\Sigma_{-j} = (Q + R - r_j I_j)^{-1} \tag{4.28}$$

where r_j is the precision of the j^{th} datum and I_j is a square matrix of zeros with a one at the jj^{th} entry.

If for some scalar γ_j,

$$\Sigma_{-j} = (I - \gamma_j \Sigma I_j)\Sigma \tag{4.29}$$

then the covariance matrix for all data except any left-out point j may be found without additional inversion of the precision matrix.

γ_j is found by post multiplying both sides of Equation (4.29) by Σ_{-j}^{-1} as per Equation (4.28) gives

$$\begin{aligned} I &= (I - \gamma_j \Sigma I_j)\Sigma(Q + R - r_j I_j) \\ &= (I - \gamma_j \Sigma I_j)(I - r_j \Sigma I_j) \\ &= I - r_j \Sigma I_j - \gamma_j \Sigma I_j + r_j \gamma_j \Sigma I_j \Sigma I_j \\ &= I - r_j \Sigma I_j - \gamma_j \Sigma I_j + r_j \gamma_j \Sigma_{jj} \Sigma I_j \\ &= I - (r_j + \gamma_j - r_j \gamma_j \Sigma_{jj}) \Sigma I_j \end{aligned} \tag{4.30}$$

which only holds if

$$\gamma_j = \frac{r_j}{r_j \Sigma_{jj} - 1} \tag{4.31}$$

The posterior mean is found via

$$\mu_{-j} = \Sigma_{-j} R_{-j} Y \tag{4.32}$$

which becomes

$$\begin{aligned} \mu_{-j} &= (I - \gamma_j \Sigma I_j)\Sigma(R - r_j I_j)Y \\ &= (I - \gamma_j \Sigma I_j)(\mu - r_j \Sigma I_j Y) \\ &= \mu - r_j \Sigma I_j Y - \gamma_j \Sigma I_j \mu + r_j \gamma_j \Sigma I_j \Sigma I_j Y \\ &= \mu - \gamma_j \Sigma I_j \mu + (r_j \gamma_j \Sigma I_j \Sigma I_j - r_j \Sigma I_j) Y \\ &= \mu - \gamma_j \Sigma I_j \mu + \gamma_j (r_j \Sigma I_j \Sigma I_j - \frac{r_j}{\gamma_j} \Sigma I_j) Y \\ &= \mu - \gamma_j \Sigma I_j \mu + \gamma_j (r_j \Sigma_{jj} \Sigma I_j - r_j \Sigma_{jj} \Sigma I_j + \Sigma I_j) Y \\ &= \mu - \gamma_j \Sigma I_j \mu + \gamma_j \Sigma I_j Y \\ &= \mu - \gamma_j \Sigma I_j (\mu - Y) \end{aligned} \tag{4.33}$$

Hence, for multivariate normal prior and likelihood, the augmented posterior moments are calculated using fast matrix-vector products with a few scalar calculations. In fact, inversion of the precision matrix may be avoided altogether; only

the marginal means and variances are required in cross-validation. The marginal variances are quickly calculated from the Cholesky decomposition of the (sparse) precision matrix $Q + R$. The updated mean is found by solving the equation

$$(Q + R)\nu = -I_j(\mu - Y) \tag{4.34}$$

for ν. Given the Cholesky decomposition of $(Q + R) = LL^T$, the solution is by forward-solving and then back-solving two sparse systems of linear equations (as per the solution of Equation (4.8) and is thus very fast. Now the updated mean is given by

$$\mu_{-j} = \mu - \gamma_j \nu \tag{4.35}$$

The marginal variances are the diagonal of the covariance matrix and are found via

$$\begin{aligned} diag(\Sigma_{-j}) &= diag((I - \gamma_j \Sigma I_j)\Sigma) \\ &= diag(\Sigma - \gamma_j \Sigma I_j \Sigma) \\ &= diag(\Sigma) - \gamma_j diag(\Sigma I_j \Sigma) \\ &= diag(\Sigma) - \gamma_j \Sigma_{j,} \times \Sigma_{,j} \end{aligned} \tag{4.36}$$

where the required elements of Σ are found directly from the Cholesky decomposition of the precision matrix $Q + R$ and $\{\Sigma_{j,}, \Sigma_{,j}\}$ denotes the j^{th} row and column of Σ respectively. [1]

In the more general case of non-Gaussian likelihoods for counts data, the posterior is approximated as a GMRF and the same shortcuts may be applied, with appropriate changes. Augmenting the posterior distribution of the latent parameters to correct for leaving out a single datum is again both fast and exact (given the saturated posterior approximation with a GMRF). Due to the transform generally required between the latent unconstrained parameters and the expectation of the counts data, the above equations require adaptation.

[1] The equations shown here for fast augmentation of the saturated posterior mean and variances were derived by Professor John Haslett.

Recall that in the context of the GMRF approximation the posterior mean for the latent parameters is given by

$$(Q + diag(c))\mu = b \tag{4.37}$$

where b is the first order coefficient in the Taylor series expansion of the log-likelihood and c is the second order coefficient. Solving this equation for μ is fast due to the sparse structure of the posterior precision matrix $Q + diag(c)$, which is decomposed as LL^T. Correcting the vectors b and c to leave out a single datum at position j is extremely fast as only the j^{th} element is changed. If there is only a single count at j, then both b_j and c_j are exactly zero. Otherwise, the Taylor expansion of the log-likelihood of the remaining counts at j is calculated to get b_{-j} and c_{-j} at j; but this is a univariate problem and thus very simple and fast.

The augmented mean is then a case of solving

$$(Q + diag(c_j))\mu_{-j} = b_{-j} \tag{4.38}$$

for μ_{-j} (note that the prior precision matrix Q is unchanged).

Furthermore, the multidimensional optimization step required to find the mode around which the Taylor expansions are computed may be entirely avoided. The Taylor series will be accurate provided the expansion is centred approximately around the mode. The saturated posterior provides a good approximation and is already available. Thus extra expensive, iterative searches for the mode (optimization) are not required.

Relationship to Existing Methods

Cross-validation of the forward problem is markedly different from the inverse problem. In the forward problem, the leave-one-out predictive distribution of a count given location is required; therefore only the marginal posterior of the latent parameter at that known location requires augmentation to account for the left out point. This is thus a univariate problem.

Furthermore, as the likelihood normalising constant is known, the ratio of the leave-one-out posterior to the saturated posterior is given by the (inverse of the) marginal likelihood, which is a known function. This fact is exploited in sampling

based cross-validation for the quick calculation of importance sampling weights (see Section 4.2.1).

Rue et al. (2008) also use this ratio for cross-validation of the forward problem in the context of the INLA method. The marginal posterior for the parameter X at location i when datum y_i is removed is quickly and easily found as

$$\pi(x_i|Y_{-i}) \propto \frac{\pi(x_i|Y)}{\pi(y_i|x_i)} \tag{4.39}$$

where $\pi(x_i|Y)$ is the marginal saturated posterior.

As this is univariate, integration and thus normalisation are available numerically.

The inverse problem is more difficult; dropping a datapoint now requires the construction of the predictive distribution across all locations, given the left out count and the updated posterior. The normalising constant must be found by calculating this across *all* possible locations, so that even if the cross-validation summary statistic is a function of the value of the predictive distribution at the correct location alone, it must still be calculated everywhere.

Inverse cross-validation is more closely related to the fast rank-one updates described in Rue et al. (2008) to compute the posterior multivariate normal mean, conditioned on a fixed value of the parameter at a single point. These fast updates are used for the computation of the Laplace approximation for the marginals of the model parameters (see Section 4.1.4). This task is, at first glance, very different from the task of inverse cross-validation; however, some of the same methodology is applied.

Both tasks could be completed through a re-fitting of the entire model but in both cases this computationally intensive option may be avoided by instead correcting (updating) the moments of the saturated posterior. As in that paper, a multidimensional optimization step is avoided; the saturated posterior mean (mode) is used as an approximation to the updated mode. As shown above, the Taylor series coefficients b and c are recalculated quickly as they are only required for a single point; centering the expansion around the approximated mode avoids the multidimensional optimization.

It should be noted here that these fast updates to the posterior moments assume

that the joint posterior, across all locations, is available; this is only the case for the GMRF approximation to the latent parameters. If the Laplace approximation is used for all the latent parameters, this gives disjoint, albeit potentially more accurate, approximations for the posterior marginals only. This approximation is *not* amenable to fast updating of the saturated posterior for leave-one-out cross-validation as updates must be performed on a closed-form joint posterior for X.

Local Corrections

The calculations in Equations (4.29) and (4.40) may be sped up by observing that the effect of removing a datapoint is local. Rue et al. (2008) calls the area effected by changes to a point / location the "region of interest". The moments need only be changed within this region of interest, so that correcting the saturated posterior involves only a few, fast calculations. The region of interest may be found by working out from the location of the left-out point until changes to the moments drop below a certain threshold or by using preset distances to specify the region. Depending on the size of the preset and fixed region, the latter involves fewer calculations and is thus faster, but may represent an approximate cross-validation if the region is set too small.

Inversion of the Posterior Precision Matrix

Solving Equation (4.38) using the same method as per finding the saturated posterior requires a fresh Cholesky decomposition of the posterior precision matrix; unfortunately L_{-j} is not calculable directly from L. In addition, the marginal variances given by Equation (4.36) require computation of an entire row (or column; the matrix is symmetric) of the joint covariance matrix. Repeating for each left out point therefore ultimately involves computing the entire saturated posterior covariance matrix.

If the entire saturated posterior covariance matrix is calculated then updates to this matrix are given using Equation (4.29). This is the preferred approach as the computations associated with a once off full inversion are less than that associated with solving Equation (4.38) *and* finding, via recursion, the marginal variances. The posterior mean, corrected for a left-out point is then given directly by a simple

matrix-vector product:

$$\mu_{-j} = \Sigma_{-j} b_{-j} \tag{4.40}$$

In the context of the GMRF approximation the γ_j required in Equation (4.29) is given by

$$\gamma_j = \frac{c_j}{c_j \Sigma_{jj} - 1} \tag{4.41}$$

Hyperparameters

Leaving out data at a single point will have minimal effect on the posterior for the hyperparameters. Therefore, the approximation that there is no effect on the hyperparameters is adopted here. The accuracy of this approximation may be tested by leaving out a datapoint and refitting the Laplace approximation for the posterior of the hyperparameters. Comparison with the saturated hyperparameter posterior will show no (negligible) difference if this approximation is accurate. In the limit of infinite (large amounts of) data, this approximation is exact.

4.2.4 More Computational Savings

Given the above methods for augmenting the posterior moments of the latent parameters, the main computational burden in performing inverse cross-validation is in constructing the inverse posterior predictive distributions. As these are often multimodal in shape (see Section 2.5.1), they cannot be well approximated with a deterministic distribution. MCMC is the most obvious choice, as per Haslett et al. (2006) and Bhattacharya and Haslett (2008); however a faster alternative is presented here.

The approximation to the posterior of the latent parameters requires the imposition of a discrete grid on the locations space. The inverse predictive distributions are therefore discrete. Hence, Monte-Carlo may again be avoided; the un-normalised posterior predictive distribution for the locations are calculated at all points on the grid. Dividing by the sum normalizes the mass function.

The multimodal nature of these distributions is a fundamental challenge to any MCMC sampling algorithm (see for example Bhattacharya and Haslett (2008)) as

the Markov chain may become stuck in one mode and fail to explore others. The fine grid necessitated by the GMRF approximation fortunately eliminates this problem.

4.2.5 Summary Statistics of Model Fit

Having performed inverse cross-validation, a summary statistic is required to report the "fit" of the model to the data. Bhattacharya (2004) advocates using reference distributions. A discrepancy measure is computed between samples drawn from the cross-validation leave-one-out posteriors and the corresponding observations. Four such discrepancy measures are presented in that work; the first three are measures of the "distance" between the mode of the posterior predictive distribution and the observed value, standardised by the variance of the predictive distribution. The sum of these values is labeled the observed discrepancy. The reference distribution is then the distribution of the discrepancy measure with the modal values replaced with samples from the posterior predictive distributions. The model is said to fit the data if the observed discrepancy is within the 95% highest density region of the reference distribution. Bhattacharya (2004) also notes, however, that "there seems to exist no easily computable reference distribution" for this statistic.

Although the reference distributions themselves are unimodal, the discrepancy measures used in Bhattacharya (2004) are a poor summary of the fit to data of the multimodal predictive distributions. The percentage of observations falling outside the corresponding 95% highest posterior predictive distribution is a useful statistic here. This will be approximately 5% if the model fits the data, regardless of the shape of the predictive distribution.

Definition 5 Δ *is the percentage of locations that fall outside the* 95% *highest posterior density region of their leave-one-out cross-validation inverse posterior predictive distribution.*

4.2.6 Toy Problem Example

Repeated runs of a toy model illustrate the cross-validation summary statistic given in Definition 5. Counts data are simulated from a known distribution function, defined on a 15×15 regular lattice. Leave-one-out cross-validation is performed in

the inverse sense and the percentage of points falling outside the 95% highest posterior density region of the predictive distribution is calculated. Cross-validation is performed using the fast updates to the saturated posterior moments derived in Section 4.2.3; however, these fast corrections are exact, given the GMRF approximation. Once the modal values of the hyperparameters are found using the Laplace approximation, they are considered fixed for the cross-validation procedure.

The toy example algorithm is:

1. for i in $1, \ldots, M$ do

   (a) Generate a random GMRF defined on the lattice.

   (b) Generate zero-inflated Binomial counts using Equation (3.18), with the GMRF as the logit of the Binomial parameters.

   (c) "Forget" the GMRF values, and the hyperparameters (smoothness of the GMRF and α in Equation (3.18).

   (d) Fit the correct zero-inflated model using the INLA method; call this the ZI model.

   (e) Fit a non-zero-inflated model using the INLA method; call this the non-ZI model.

   (f) Perform fast inverse cross-validation for both models and find the number of points falling outside their 95% HPD predictive regions; store these as Δ^i_{ZI} and $\Delta^i_{non\text{-}ZI}$.

2. Plot the sample density of Δ_{ZI} and $\Delta_{non\text{-}ZI}$.

The results for this exercise are shown in Figure 4.7 with M = 300.

4.3 Conclusions

This chapter has introduced the application of approximation techniques (INLA) to inverse problems. The approximations apply to the forward problem, negating the need for MCMC type inference and returning closed form posteriors. This represents a speed up of several orders of magnitude in the fitting the forward model to the training data.

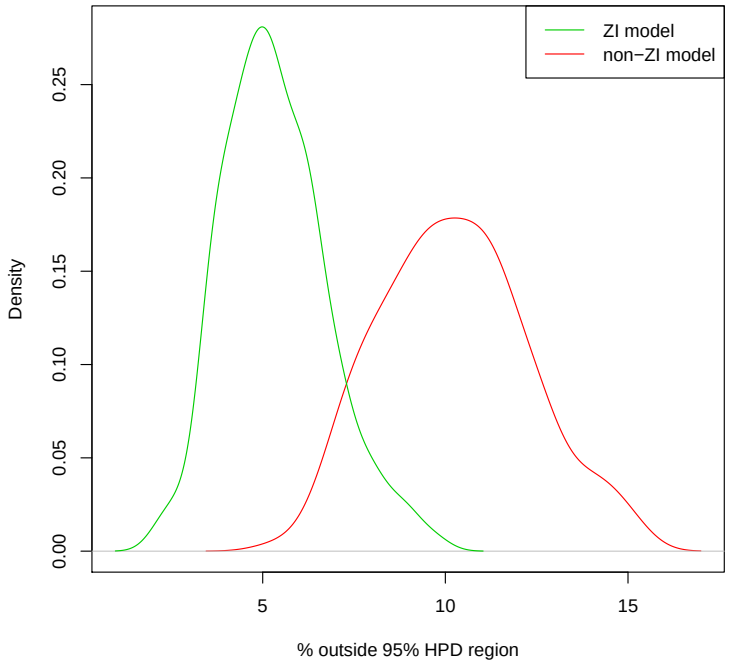


Fig. 4.7: **Multiple runs and cross-validations. Toy data is simulated for a zero-inflated Binomial model and inverse cross-validation using the correct model and a similar non-zero-inflated model is performed. This is repeated and the sample density of Δ (the percentage of points falling outside the 95% predicitve density region) is plotted. The incorrect model (non-ZI) tends to have a higher number of points falling outside the 95% predicitve density region.**

Spatial zero-inflated models in which there are two entirely separate latent processes are not compatible with the INLA method. However, a single-process model for zero-inflated counts data in which probability of presence and abundance when present are modelled as functions of a single underlying process is well suited to the technique. In fact, this model was developed before the INLA method was considered as it will lead to faster and more accurate inferences than the two-process model under MCMC, provided the model is correct.

Although model validation is the subject of a very extensive literature for the forward problem, the inverse problem has rarely been considered. MCMC based cross-validation is computationally intensive in the extreme; Bhattacharya and Haslett (2008) sets the standard for this research by augmenting MCMC based validations with importance resampling to greatly speed up calculations. However, this approach still requires much computation; although running times for an example palaeoclimate problem are reduced by several orders of magnitude, the cross-validation still takes hours or even days to run.

This computational burden hampers the comparison of multiple models. A method for performing inverse cross-validation is therefore introduced. This method fits the saturated posterior for the forward model using the INLA method. The hierarchical hyperparameters are fixed at their posterior modal values and a new, very fast method for correcting the posterior for the parameters is applied. This allows for fast calculations of the leave-one-out posterior for the forward problem. The inverse predictive distributions are also found without resorting to sampling methods; this is due to the imposition of a discrete grid in the forward fitting stage.

Observing that corrections to the saturated posterior are only necessary in a local region speeds up the cross-validation further. For the motivating palaeoclimate problem, inverse cross-validation now takes around one hour using these methods. The estimated running time using brute-force MCMC re-runs is of the order of many years; even the fast importance sampling method that set the current standard takes about two weeks (although most of that is taken in running the initial regular MCMC chain). These vast improvements in speed are coupled with improvements in the method also; hyperparameters are no longer fixed a-priori but are estimated from the data.

The advancements in the modelling of the pollen data since Haslett et al. (2006) are therefore as follows:

1. Estimation of model hyperparameters. These were fixed a-priori in Haslett et al. (2006).

2. Highly accurate closed-form posterior distributions on the latent X are delivered via INLA. This is achieved with a speed-up of several orders of magnitude, thus allowing for richer models to be built such as zero-inflated counts likelihoods.

3. Fast updates to the saturated posterior for the latent response to deliver the leave-one-out posteriors for the inverse problem. The MCMC based methodology used in Haslett et al. (2006) does not allow for such corrections to be made.

4. Model goodness-of-fit metrics, tailored to the inverse problem, are presented for the fast leave-one-out cross-validations. Such metrics were not considered in Haslett et al. (2006).

Chapter 5

Inference Methodology

Highly multivariate data is often challenging to model due to the "curse of dimensionality" Bellman (1957). Much of the work in this thesis is concerned with reducing the computational burden associated with modelling the response of vegetation to climate using the modern training data (see Section 1.1.1).

This is achieved through separate analysis of the marginal responses of individual plant taxa to climate. The approach necessarily ignores between taxa dependencies but allows for reduction of overall computation. The separate, marginal analyses can then be brought together *post-hoc*. This is referred to as the **inference-via-marginals** posterior and is an approximation to the joint posterior of the full model. Situations where the approximation is poor and where it is excellent, or even exact, are identified.

Details related to working on a discrete grid across the location space are also investigated. This chapter serves as a review and assessment of the preceding chapters methodology. The methods are brought together in preparation for the application to the pollen dataset in Chapter 6.

In particular, the concern here is with inference on multiple latent spatial Gaussian processes defined on a lattice. The INLA methodology introduced in Section 4.1 is unsuitable for this task as it can only deal with one such spatial process at a time (see Section 5.1.4). If the model does not disjoint-decompose (Section 3.2) exactly then approximations to the model that do decompose must be sought. The accuracy of these approximations must then be tested, as per Section 5.3.

The goal is inference on latent random variables X, given counts data Y defined

at discrete locations C. X is composed of N_T processes which are Gaussian (or approximately so) given the data. Thus at each value of C, there are N_T counts or proportions arising from N_T potentially dependent X values. As the posterior is GMRF due to the methods introduced in Section 4.1, it is expressed via the posterior mean vector μ and precision matrix Q. i.e. the posterior distribution of X is given approximately by:

$$\pi(X)\prod_{j=1}^{C}\pi(Y_j|X_j) \simeq GMRF(X|\mu,Q) \tag{5.1}$$

The goal is therefore inference on the μ and Q terms. By disjoint-decomposition into marginals defined as the product of independent multivariate $X_i, i = \{1,\ldots,N_T\}$ across locations C becomes:

$$GMRF(X|\mu,Q) = \prod_{i=1}^{N_T}\pi(X_i|Y_i) \simeq \prod_{i=1}^{N_T} GMRF(X_i|\mu_i,Q_i) \tag{5.2}$$

The INLA method then delivers the terms μ_i and Q_i that define the independent spatial X_i given the data. The error incurred due to this decomposition is the subject of Section 5.3.

5.1 Reasons for Disjoint-Decomposition

5.1.1 Parallelisation

If the overall problem disjoint-decomposes (see Section 3.2) then the exploitation of parallel programming resources is trivial. Inference on each disjoint module may be performed entirely independently of the others and thus may be done at the same time on multiple processors. No specialist code or hardware is required and joint summary statistics are simple and quick calculations that are made on a single processor.

5.1.2 Memory Usage

When responses are modelled non-parametrically (see Section 1.1.1) as for the palaeoclimate dataset, the number of parameters is of the order of 10^3 for a two

dimensional climate space. Ultimately, the goal is to work with at least three climate variables (leading to the order of millions of parameters) and also to integrate out the unknown hyperparameters. The memory required simply to store such a burdensome model may quickly overwhelm even modern, high specification personal computers.

Memory usage is therefore one of the most difficult issues associated with performing inference on this dataset. It is crucial to reduce both the computational overhead and the size the model takes up in memory. Breaking the model down into a series of approximately disjoint / conditionally independent marginals negates the need to store and manipulate all of the parameters at once on a single machine.

5.1.3 Inverse Problem

Another large saving in terms of computation is in the inversion of the forward model; a high dimensional integral (to find the joint marginal likelihood) is replaced with a product of uni-dimensional integrals. These may be evaluated without resorting to Monte-Carlo based sampling methods that are required for the estimation of multidimensional integrals.

5.1.4 Compatibility with the INLA Method

If the joint model does not decompose, then inference must be performed for the entire model at once. Interaction terms between marginals are non-zero and must be modelled explicitly. For problems consisting of multiple smooth surfaces giving rise to vectors of counts, such as the motivating palaeoclimate problem, this results in difficulties for the INLA inference method.

Interaction may occur at one of two levels in the model hierarchy; either in the prior or the likelihood. The former models smooth surfaces that are not independent across locations; the latter models independent smooth surfaces that jointly give rise to non-independent counts. This is a continuation of the discussion in Section 3.2.2.

Prior and Hyperparameters

If the interactions between marginals are modelled at the latent variable stage, then the joint prior must contain these terms. Specifically, the joint prior precision matrix

must contain non-zero terms for the interactions. If they are not known a-priori then unknown hyperparameters must be introduced to model these inter-process precisions.

The INLA method deals with hyperparameters via numerical integration. The entire vector of hyperparameters is set on a discrete grid, as per Section 4.1.3. This approach requires that the number of hyperparameters is low; any more than five or so and the method runs into computational difficulty. The motivating palaeoclimate problem has 28 plant taxa; even crude modelling of the interactions with a single hyperparameter governing each taxon-taxon interaction results in 378 additional hyperparameters; this is far more than the INLA method can cope with (Rue et al. (2008)).

Likelihood and Taylor Expansions

Modelling the interactions at the data level eliminates the need for additional interaction hyperparameters. Interaction terms are placed in the likelihood precision matrix and are thus parameters rather than hyperparameters; however the data are not conditionally independent. The Taylor series expansions in the GMRF approximation are no longer univariate, resulting in a massive increase in the computation required to fit the approximation (see Section 4.1.2). This is a fundamental challenge to the existing INLA methodology.

5.2 Multivariate Normal Model

Much of this chapter focusses on the marginal analysis of data in a Gaussian setting. The reasons for exploration of such a problem in a purely Gaussian framework are as follows:

1. For simple exploration and illustration of key points; the availability of the posterior in closed form is the primary motivation.

2. For synthesis with the Gaussian approximation technique described in Section 4.1. This technique applies a Gaussian prior and approximates the posterior with a Gaussian.

Given a multivariate normal prior $\pi(X)$ and multivariate normal likelihood $\pi(Y|X)$, the posterior $\pi(X|Y)$ is multivariate normal with mean and precision matrix given by

$$\begin{aligned} \mu &= (Q_X + Q_Y)^{-1}(Q_X\mu_X + Q_Y\mu_Y) \\ Q &= Q_X + Q_Y \end{aligned} \tag{5.3}$$

where the prior precision matrix is Q_X and the likelihood precision matrix is Q_Y.

5.2.1 Conditions for Perfect Disjoint-Decomposition

If the joint posterior expresses zero precision between processes then a fully joint inference may be done exactly via the marginals. There are in fact two situations in which this occurs:

1. The underlying processes are truly independent of each other
2. The joint model creates a posterior with conditional independence across the latent variables, regardless of any dependency structure suggested by the data [1]

The second situation for a perfect inference-via-marginals is clearly illustrated with two common examples. These are discussed in Section 5.2.2 and Section 5.2.

The terminology used in this thesis will be that the joint model **disjoint-decomposes** exactly (see Section 3.2).

It can immediately be seen from Equation (5.3) that the condition for exact disjoint-decomposition of the posterior is that both the prior and the likelihood disjoint-decompose. The form of the precision matrix shows whether a density will decompose; if it is block-diagonal, then the blocks each represent a disjoint part of the full model. See Table 5.1 for illustration.

The assumption that the precision of process i_1 at location j_1 given process i_2 at location $j_2 \neq j_1$ is zero is logical; there will not be interaction between a plant taxon at one climate location and a different taxon at another, disparate location in climate space. This results in a banded overall precision matrix and is

[1] All models so far used in the palaeoclimate project have in fact been of this type.

Table 5.1: **Joint prior precision matrix entries, using the indexing system $^{i_2,j_2}_{i_1,j_1}$ means precision between processes i_1 and i_2 at locations j_1 and j_2, respectively (processes here are the latent fields). Intra-process precision is highlighted in yellow and inter-process precision is unhighlighted. If the processes are conditionally independent given the locations then all values in the non-highlighted sub-matrices will be zero in the prior precision matrix, so that it is block-diagonal. This table shows the precision matrix for number of processes $N_T = 4$ and number of discrete locations $N_L = 4$.**

1,1 1,1	1,2 1,1	1,3 1,1	1,4 1,1	2,1 1,1	2,2 1,1	2,3 1,1	2,4 1,1	3,1 1,1	3,2 1,1	3,3 1,1	3,4 1,1	4,1 1,1	4,2 1,1	4,3 1,1	4,4 1,1
1,1 1,2	1,2 1,2	1,3 1,2	1,4 1,2	2,1 1,2	2,2 1,2	2,3 1,2	2,4 1,2	3,1 1,2	3,2 1,2	3,3 1,2	3,4 1,2	4,1 1,2	4,2 1,2	4,3 1,2	4,4 1,2
1,1 1,3	1,2 1,3	1,3 1,3	1,4 1,3	2,1 1,3	2,2 1,3	2,3 1,3	2,4 1,3	3,1 1,3	3,2 1,3	3,3 1,3	3,4 1,3	4,1 1,3	4,2 1,3	4,3 1,3	4,4 1,3
1,1 1,4	1,2 1,4	1,3 1,4	1,4 1,4	2,1 1,4	2,2 1,4	2,3 1,4	2,4 1,4	3,1 1,4	3,2 1,4	3,3 1,4	3,4 1,4	4,1 1,4	4,2 1,4	4,3 1,4	4,4 1,4
1,1 2,1	1,2 2,1	1,3 2,1	1,4 2,1	2,1 2,1	2,2 2,1	2,3 2,1	2,4 2,1	3,1 2,1	3,2 2,1	3,3 2,1	3,4 2,1	4,1 2,1	4,2 2,1	4,3 2,1	4,4 2,1
1,1 2,2	1,2 2,2	1,3 2,2	1,4 2,2	2,1 2,2	2,2 2,2	2,3 2,2	2,4 2,2	3,1 2,2	3,2 2,2	3,3 2,2	3,4 2,2	4,1 2,2	4,2 2,2	4,3 2,2	4,4 2,2
1,1 2,3	1,2 2,3	1,3 2,3	1,4 2,3	2,1 2,3	2,2 2,3	2,3 2,3	2,4 2,3	3,1 2,3	3,2 2,3	3,3 2,3	3,4 2,3	4,1 2,3	4,2 2,3	4,3 2,3	4,4 2,3
1,1 2,4	1,2 2,4	1,3 2,4	1,4 2,4	2,1 2,4	2,2 2,4	2,3 2,4	2,4 2,4	3,1 2,4	3,2 2,4	3,3 2,4	3,4 2,4	4,1 2,4	4,2 2,4	4,3 2,4	4,4 2,4
1,1 3,1	1,2 3,1	1,3 3,1	1,4 3,1	2,1 3,1	2,2 3,1	2,3 3,1	2,4 3,1	3,1 3,1	3,2 3,1	3,3 3,1	3,4 3,1	4,1 3,1	4,2 3,1	4,3 3,1	4,4 3,1
1,1 3,2	1,2 3,2	1,3 3,2	1,4 3,2	2,1 3,2	2,2 3,2	2,3 3,2	2,4 3,2	3,1 3,2	3,2 3,2	3,3 3,2	3,4 3,2	4,1 3,2	4,2 3,2	4,3 3,2	4,4 3,2
1,1 3,3	1,2 3,3	1,3 3,3	1,4 3,3	2,1 3,3	2,2 3,3	2,3 3,3	2,4 3,3	3,1 3,3	3,2 3,3	3,3 3,3	3,4 3,3	4,1 3,3	4,2 3,3	4,3 3,3	4,4 3,3
1,1 3,4	1,2 3,4	1,3 3,4	1,4 3,4	2,1 3,4	2,2 3,4	2,3 3,4	2,4 3,4	3,1 3,4	3,2 3,4	3,3 3,4	3,4 3,4	4,1 3,4	4,2 3,4	4,3 3,4	4,4 3,4
1,1 4,1	1,2 4,1	1,3 4,1	1,4 4,1	2,1 4,1	2,2 4,1	2,3 4,1	2,4 4,1	3,1 4,1	3,2 4,1	3,3 4,1	3,4 4,1	4,1 4,1	4,2 4,1	4,3 4,1	4,4 4,1
1,1 4,2	1,2 4,2	1,3 4,2	1,4 4,2	2,1 4,2	2,2 4,2	2,3 4,2	2,4 4,2	3,1 4,2	3,2 4,2	3,3 4,2	3,4 4,2	4,1 4,2	4,2 4,2	4,3 4,2	4,4 4,2
1,1 4,3	1,2 4,3	1,3 4,3	1,4 4,3	2,1 4,3	2,2 4,3	2,3 4,3	2,4 4,3	3,1 4,3	3,2 4,3	3,3 4,3	3,4 4,3	4,1 4,3	4,2 4,3	4,3 4,3	4,4 4,3
1,1 4,4	1,2 4,4	1,3 4,4	1,4 4,4	2,1 4,4	2,2 4,4	2,3 4,4	2,4 4,4	3,1 4,4	3,2 4,4	3,3 4,4	3,4 4,4	4,1 4,4	4,2 4,4	4,3 4,4	4,4 4,4

similar to the Separable Models described in Finkenstdt et al. (2006). However, their approach cannot be exploited here. Separable models involve modelling two separate precision matrices; typically a spatial matrix and a temporal matrix. The spatio-temporal precision matrix is then taken as the Kronecker product of these two matrices. Taking a Kronecker product of an intra-process precision matrix and an inter-process precision matrix would impose two modelling aspects that are unsuitable for this work. Specifically, (i) implementation of a separable model would require using a common spatial precision matrix for all processes and (ii) would lead to non-zero $\substack{i_2,j_2 \\ i_1,j_1}$ terms in the precision matrices for $i_1 \neq i_2$ and $j_1 \neq j_2$.

5.2.2 Compositional Independence

Although it would seem that compositional data analysis must necessarily model interaction between the components of the composition (in the pollen dataset, the components are the various plant taxa), Aitchison demonstrates that any statistical analysis making use of the Dirichlet distribution is in fact imposing a strong implied independence structure (Aitchison (1986) 3.4). Similarly, any logistic-normal (Section 3.5.4) distribution with diagonal precision matrix will impose the same strong implied independence structure.

For example, in a Bayesian setting, joint inference done on the latent vector of probability parameters P governing a compositional counts vector Y might proceed as follows:

Likelihood is Multinomial:

$$\pi(Y; n, P) = \frac{n!}{\prod_{i=1}^{N_T} y_i!} \prod_{i=1}^{N_T} p_i^{y_i} \tag{5.4}$$

where $n = \sum_i Y_i$

Prior is Dirichlet

$$\pi(P; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{N_T} p_i^{\alpha_i - 1} \tag{5.5}$$

where α is a vector of hyperparameters

The posterior is then Dirichlet due to conjugacy

$$\pi(P;\alpha+Y) = \frac{1}{B(\alpha+Y)}\prod_{i=1}^{N_T} p_i^{\alpha_i+y_i-1} \tag{5.6}$$

The vector P is subject to the constraint that $\sum_i P_i = 1$ and the vector of counts Y is subject to the constraint that $\sum_i Y_i = n$.

However, the Multinomial and the Dirichlet actually enforce conditional independence given the constraints.

The Multinomial can be expressed as a product of Poisson distributions, with parameters equal to $n \times P$, conditioned on the sum being equal to the total count n. This sum ($\sum_i y_i = n$) itself follows a Poisson distribution with rate parameter n.

$$\pi(Y;n,P) = \frac{\prod_{i=1}^{N_T}\text{Poisson}(y_i;\lambda_i)}{\text{Poisson}(n;n)} \tag{5.7}$$

with $\lambda_i = n \times p_i$ and $\text{Poisson}(y_i;\lambda_i) = \frac{\lambda_i^{y_i}e^{-\lambda}}{y_i!}$

Similarly, a Dirichlet with parameter vector η may be expressed as a product of Gamma distributions, with shape parameters η and rate parameters all equal to $\sum_i \eta_i$, conditioned on the sum = 1 following a Gamma distribution with shape and rate both equal to $\sum_i \eta_i$.

$$\pi(P;\eta) = \frac{\prod_{i=1}^{N_T}\text{Gamma}(p_i;\eta_i,\sum_k \eta_k)}{\text{Gamma}(1;\sum_k \eta_k,\sum_k \eta_k)} \tag{5.8}$$

Therefore, to perform joint inference given a vector of counts and using a Dirichlet prior and a Multinomial likelihood, no accuracy is lost in performing marginal inferences on each part of the composition and then conditioning on the sum. This result is regardless of the value of the parameters of the distributions.

If sampling from the Dirichlet posterior is required, this can be achieved by sampling from the Gamma marginals and then rescaling such that the sum is one. In fact, this is the usual algorithm for sampling from a Dirichlet distribution. The post-hoc conditioning or rescaling is the only step that requires joint knowledge of the marginals.

5.3 Sensitivity to Inference via Marginals

If the joint model disjoint-decomposes then the inference-via-marginals approximation is exact. If not, the approximation amounts to setting non-zero terms in the

overall precision matrix to be zero. This is equivalent to breaking some links in the graph of the model, specifically the inter-process links.

The accuracy of the inference-via-marginals approximation depends on several factors. The magnitude of the non-zero terms that are set to zero to facilitate decomposition are the most obvious of these. The further these are from zero, the greater the interaction and hence the worse the inference-via-marginals approximation will be. However, given non-zero interactions, several other factors will impact the level of accuracy.

Ultimately, the interest is in the differences between a full joint model and the inference-via-marginals model in terms of the inverse problem. Of course, if the forward model disjoint-decomposes exactly, then the inverse predictive distributions will be identical for the two models.

The worst case scenario is presented in Figure 5.1. A single surface is replicated T times; random counts are generated at various points in the location space. Inverse cross-validation predictive distributions are formed and the joint predictive distribution is found by taking the product and normalising (see Section 3.2). However, this model treats the surfaces as independent; they are in fact fully correlated. This results in a linear increase with Δ, the percentage of points lying outside their 95% cross-validation highest inverse predictive density with T.

Thus the inference-via-marginals model is a poor approximation; correlation of 1 between the surfaces (as they are all identical) represents the upper bound of inaccuracy of the inference-via-marginals approximation. Even if the data are simulated for each replication of the surface independently, the model does not disjoint-decompose.

To see this, take $s_1(c) = s_2(c)$ both the same smooth function of c. Random samples $x_1 \sim N(s_1, \epsilon)$ and x_2 $N(s_2, \epsilon)$ will have high $corr(x_1, x_2)$ but low $corr(x_1, x_2|c)$. So inference on the forward model may be completed marginally. Thus the result in Figure 5.1 is, at first glance, counter-intuitive. Independent counts should yield a tighter predictive density on the correct location. Δ should converge to zero as the predictive cross-validation densities converge to Dirac distributions on the correct locations.

The reason why this is not so is a peculiarity of the inverse problem; the shape of

the response surface in this case is such that for any given location, there is typically one or more other locations with the same (or a very similar) value for the response. When a point is left out, the posterior variance at that point is greater than the locations for which there is data (for illustration, see Figure 2.3(b)). Thus the marginal likelihood is higher for these other locations; in turn the inverse predictive density will give greater weight to those locations with data and a similar response to the correct location. For a single response surface, this does not cause a problem. It is when analysing multiple surfaces that the issue arises.

Although the correct, left-out location has non-zero predictive density for a single surface (and is within the bounds of the 95% highest posterior predictive density 95% of the time), it will tend to zero as T increases. Multiplication of the density with itself T times will cause the location with the *single* highest density to gain *all* the mass as the number of replications of the surface increases. Thus, the correct location loses predictive probability mass and Δ increases.

In the simplest example, suppose inference on a single surface leads to the (correct) location predictive density of 75% probability mass a location A and 25% at location B, with location B the correct location. This is ok in terms of the Δ statistic; B is within the 95% HPD region. Now suppose that we are presented with new data from the same response, but treat it as independent. We might find, again without error, that this data suggests $P(A) = 70\%$ and $P(B) = 30\%$. Again, this is ok. Assumption that the data are independent however, leads to a multiplication and re-normalisation of these values so that $P(A) = 87.5\%$ and $P(B) = 12.5\%$. A third dataset yields $P(A) = 80\%$ and $P(B) = 20\%$. The resultant multiplicative joint predictive distribution then has $P(A) = 96.55\%$ and $P(B) = 3.45\%$. Now B is *outside* the 95% HPD region.

This issue will not arise when there is no other region of the location space with a very similar counts data vector. Using the exact same model and code used to produce Figure 5.1, but generating a new random response surface for each replication results in a Δ statistic of exactly zero for 20 surfaces. This is because each surface carries independent information on the location given counts data. Figure 5.2 shows a plot of $\Delta(T)$ for random independent surfaces.

Note that this convergence to a probability of one at the correct location will

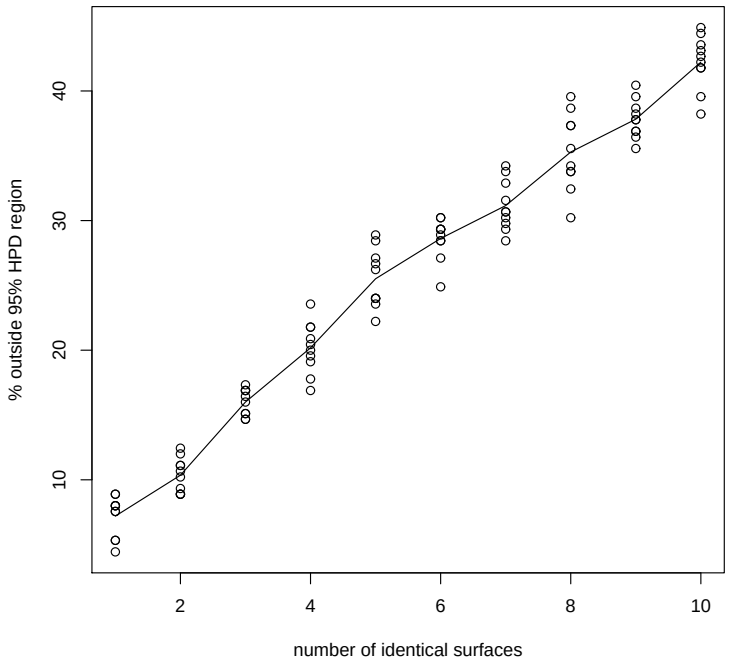


Fig. 5.1: **A single smooth surface gives rise to counts data at locations. Cross-validation in the inverse sense (location given count) is used to find Δ, the percentage of points falling outside their 95% highest predictive distribution region. As the data are simulated and the model fitted is the same as the model used to generate the data, Δ is about 5%.**
Taking T replications of the surface and taking the normalised product of the cross-validation inverse predictive distributions is equivalent to fitting the inference-via-marginals model to a joint model that has maximum inter-process correlation: the more replications of the single surface, the worse the approximation and the higher the Δ statistic. The above graph depicts 10 replications with randomly generated data for each value of T (scatter plot). The line is a mean across these 10 replications.

also occur for the case of replicated identical surfaces when the cross-validation is performed using the saturated posterior.

5.3.1 Discrete HPD Regions

The reason why Δ does not converge to 5% and instead converges to 0% as the number of conditionally independent counts increases is due to the discrete grid. As per Section 3.1.3, the HPD region contains 95% **or more** of the total probability mass. Therefore, the expected value of Δ is $\leq 5\%$ **for each independent surface**.

As more of these conditionally independent components are brought together, each with $\Delta \leq 5\%$, the predictive distributions become increasingly peaked. They are centred on the correct location (under the correct model) and so the probability mass becomes focused at the correct location. Eventually, the 95% HPD region becomes smaller than the grid spacing so that all the mass is concentrated on a single gridpoint. Using the algorithm in Section 3.1.3 for constructing discrete 95% HPD regions then results in selection of the single gridpoint that contains all this mass. As this is the correct location, none of the points lie outside their corresponding 95% (or more) HPD region and Δ tends to zero. A graphic illustration of this phenomenon is presented in Figure 5.3.

5.3.2 Nested Constrained Models

Nested constrained models represent an interesting opportunity to apply disjoint-decomposition to a joint model that does not decompose exactly as the product of its marginals (see Section 3.5.6).

Given knowledge of the nesting structure, the joint model may be exactly expressed as the product of the marginals across all levels of the nesting hierarchy, where the dependencies within levels are accounted for in the likelihood using knowledge of the counts and responses of higher levels in the structure.

The advantage of such models is that they will disjoint-decompose iff nesting structure is known a-priori. If the nesting is not known, the marginals model at the lowest level only will be a poor fit to the data. Figure 5.4 shows results for inference-via-marginals models fitted to the same data-sets used in Figure 3.14, but with the added task of inference of the forward model. It is clear that attempting to

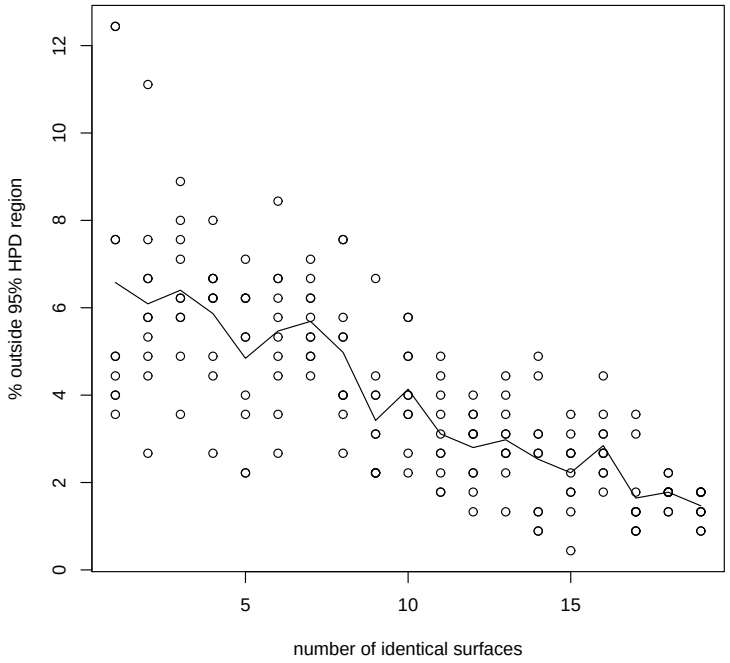
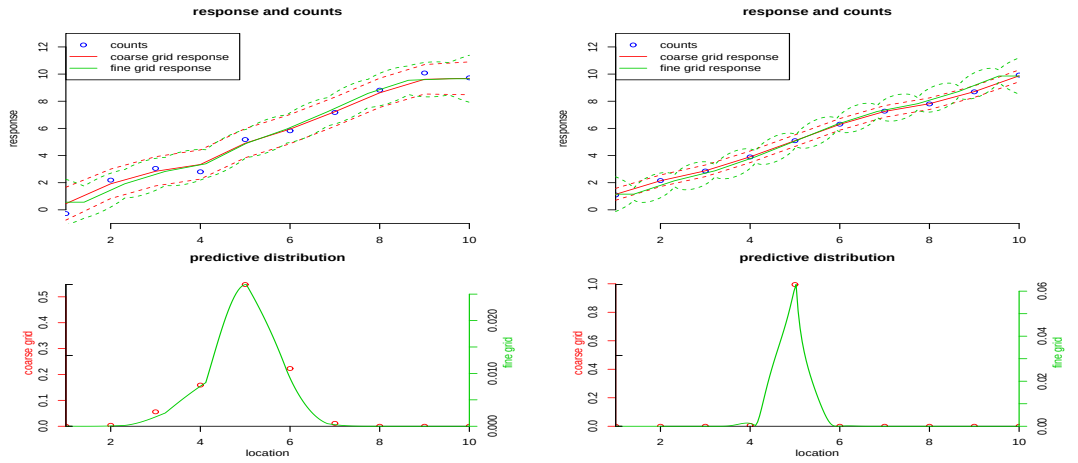


Fig. 5.2: **A single smooth surface gives rise to counts data at locations. Cross-validation in the inverse sense (location given count) is used to find Δ, the percentage of points falling outside their 95% highest predictive distribution region.**

Generating T such independent surfaces and taking the normalised product of the cross-validation inverse predictive distributions is equivalent to fitting the inference-via-marginals model to a joint model that has zero inter-process correlation. The inference-via-marginals approximation is exact and as the independent information increases, the inverse predictive densities converge towards Dirac distributions on the correct locations (see Section 5.3.1).



(a) 2 counts per location

(b) 20 counts per location

Fig. 5.3: **UPPER PANELS: A linear response gives rise to Gaussian counts (blue circles). If there are n independent counts at each location, this is analagous to multiple conditionally independent counts. These counts data are used to inform the mean posterior response on a coarse grid (red line) and fine grid (green line) and posterior variances (the broken lines depict the mean response $\pm 2\sigma$ with σ the posterior standard deviation).**

LOWER PANELS: As the number of independent observations increases, the inverse predictive distribution becomes sharper; a mean count of 5 given 2 counts (Figure (a)) gives rise to a broader predictive distribution than a mean count of 5 given 20 counts (Figure (b)).

The probability mass on the coarse grid is depicted without joining the points (red circles). The fine grid approximates continuous space and so a line is used.

Although the coarse and fine grids deliver similar results at the coarse gridpoints (subject to the difference in scale due to the differing number of evaluation points), Figure (b) shows how the 95% HPD region will differ. The coarse grid now concentrates all mass at a single point. Hence the discrete 95% or greater HPD region will actually converge to a 100% HPD region and the Δ statistic converges to 0%.

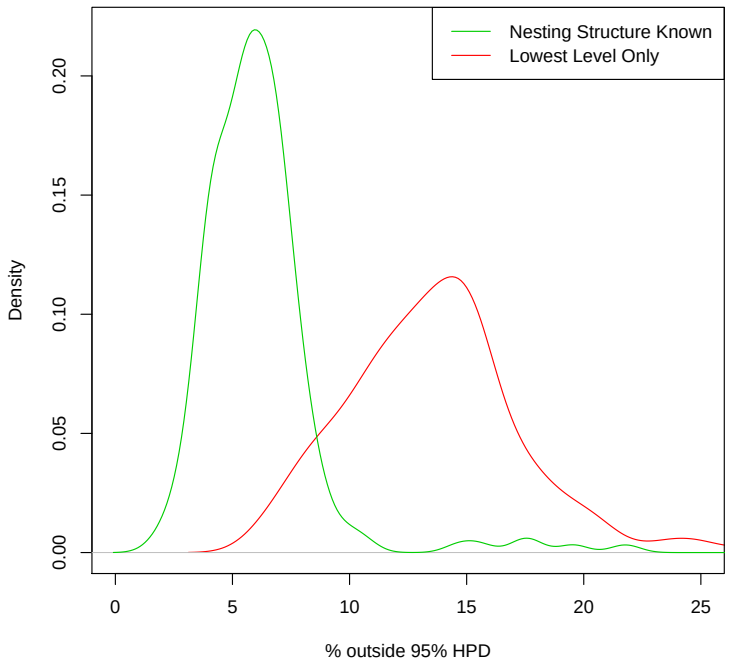


Fig. 5.4: **Sample density for the Δ fit statistic, which is 5% in theory, under the correct model. The correct model is nested with two levels. If the nesting structure is unknown then the decomposed model (red density) is a poor fit. The sample Δs were obtained via multiple random runs. The inference via-marginals model is a poor approximation to the joint model, resulting in many more points lying outside their 95% HPD predictive region.**

model the joint model as the product of the marginals without exploitation of the nesting structure results in a poor model fit. In contrast, knowledge of the nesting structure allows for exact disjoint-decomposition of the model.

5.4 Conclusions

Inference on multivariate models comprised of multiple spatial processes may be performed in disjoint modules; provided there is no interaction between these modules in either the prior or the likelihood. Decomposition is into the marginals for each process.

The multivariate normal setting delivers insight into more general models, where the posteriors may be approximated with a GMRF. This approximation requires that the model decomposes. If it does not, then use of a disjoint-decomposable model is equivalent to a model with the interaction terms set to zero. This is an approximation, the accuracy of which is determined by the degree of correlation between the marginals. When such correlation is non-zero, the loss in accuracy increases linearly with the number of non-independent processes. Accuracy here is measured by the percentage of points lying outside their 95% discrete HPD region under leave-one-out cross-validation inverse predictive distributions.

Compositional models, in which both the data and the likelihood parameters are constrained under summation, represent a class of model that do not decompose. However, many compositional models do decompose for inference on the model parameters, subject to a post-hoc conditioning or rescaling. Compositional models which do not disjoint-decompose may in do so given knowledge of a nesting structure. Such structures must be known a-priori to facilitate decomposition of the model.

Chapter 6

Application: the Palaeoclimate Reconstruction Project

The motivating problem is palaeoclimate reconstruction from pollen data. This chapter applies methods and models established in earlier chapters to this problem. No fossil reconstructions are presented; the focus here is in model fit and validation of the inverse problem. The inverse problem here is to predict (reconstruct) climate variables given a pollen assemblage.

Models are evaluated using cross-validation of the modern dataset in the inverse sense and the use of the inference-via-marginals approximation is evaluated. The inverse cross-validation is achieved using the fast updates derived in Section 4.2.3. The nested model is novel in this application.

6.1 Bayesian Palaeoclimate Reconstruction Project

The work contributed by this thesis to the ongoing Bayesian palaeoclimate reconstruction project described in Haslett et al. (2006) is presented. The main crux of the methodology up to and including the publication of that paper was acknowledged to be computational. Approximation of the posterior for the parameters of the forward model with closed form expressions via INLA greatly reduces the computations (Section 4.1).

6.1.1 The RS10 Dataset

The dataset comprises 28 pollen taxa proportions. There are 7742 sampling locations in the modern training dataset, each of which has physical variables (longitude, latitude and altitude) and climate recorded. The climate is measured here as the growing degree days above 5^oC, GDD5 and the mean temperature of the coldest month, MTCO. The former is a temperature sum and is a measure of the growing season.

The data are reported as counts, with total equal to around 400. In fact, many of the sampling locations do not have the total count reported; only the proportions are reported. In these cases, the somewhat unsatisfactory step of assuming a total count of 400 is taken. The reported climates are typically not in fact from precisely the same location in physical space as the lake from which the pollen grains are taken. The nearest meteorological station provides data on the climate. Thus an error term should be appended to these climatic observations. Expert opinion is used to inform these and a post-hoc method for correcting the inverse predictive distributions is used in Section 6.4.3.

The ultimate goal is to reconstruct these climate variables given fossil pollen counts. As this task cannot be assessed directly, inverse cross-validation on the modern training data, for which climate is known, is presented as a best-available model validation tool. This task could only be performed approximately in Haslett et al. (2006); the MCMC methodology was too labourious to cope with re-fitting the model for each left-out datum. The saturated posterior was therefore re-used for each approximate cross-validation step.

The INLA methodology allows approximations to be fitted quickly to the posteriors for the response surfaces and assorted hyperparameters, given the counts data. This thesis presents one of the first large scale tests of the INLA technique. The RS10 dataset is not only large, but some details present extra challenge to the INLA method that are not addressed in Rue et al. (2008): There are multiple, potentially interacting, counts at each sampling location; each of these are subject to overdispersion *and* zero-inflation relative to standard counts likelihood models. The counts vectors are constrained by the data collection method (count until a pre-chosen total is sorted) and are thus compositional in nature. The climate space should in fact

be $3D$; although alluded to in Rue et al. (2008), large scale problems such as this pose problems for the INLA methodology.

Although additional assumptions and / or approximations have to be made to allow the application of INLA type inference to the dataset, the method performs well. Running times to fit the forward model approximations are around 4 orders of magnitude faster than for MCMC based inference. This is on a 50×50 size grid across $2D$ climate space; each non-parametric response surface is thus described by a latent field of 2500 random variables. Leave-one-out cross-validation for the inverse problem is achieved using fast updates to the saturated posterior in around an hour. The approximations' most appealing characteristic is the closed form expressions for the posterior distributions; this allows such manipulations as the fast updates of posterior moments used for inverse cross-validation.

Application of these methods to the RS10 pollen dataset does, however, reveal some drawbacks to the INLA method:

- All hyperparameters had to be given initial values. If these were poorly chosen, the iterative search algorithm did not converge. Trial and error is necessary for each new forward model fitted to the data to find sensible initial values.

- The data are *highly* overdispersed. It was not possible to fit GMRF approximations to the response surfaces posteriors for models without overdispersed likelihoods.

- The GMRF approximation is computationally incompatible with fitting multiple latent fields unless they are conditionally independent. If they are not, inference on the degree of dependence must be modelled via a hyperparameter for each pair of responses; this quickly swamps the numerical integration step in the INLA method. Thus, only disjoint-decompositional models will be considered.

- Zero-inflation; although the model with probability of potential presence and abundance when present is compatible with the GMRF approximation, models in which there are two distinct processes driving these responses are not compatible, as shown in Section 4.1.2. Although evidence has been presented that

supports the single-process model for the RS10 dataset, fitting the two-process model would be of interest.

- $3D$ (and above) climate space poses a real challenge to the INLA method. Sparsity of the precision matrices is much reduced in higher dimensions and therefore the fast numerical routines are slowed. This remains an outstanding issue.

6.1.2 Software and Hardware

The time taken to fit a GMRF approximation to the response surface of a single taxon on a 50×50 $2D$ grid is of the order of 30 seconds for a 50×50 size grid on the climate space. Finding the modal value for the hyperparameters takes about 2 minutes for a model with 3 hyperparameters (for each iteration of the search for the modal configuration, a full GMRF approximation to the parameters given the data and the hyperparameters must be fitted). Exploration of the $3D$ grid of hyperparameters takes up to 5 minutes.

In all cases, the GMRFLib C library was used. This library of C functions is available as a free download at `http://www.math.ntnu.no/~hrue/GMRFLib/`
Other freely available C libraries used were GSL, LAPACK, BLAS, an ATLAS. The free statistical software language R was used for post-processing results and for creating all images and plots appearing in this thesis.

The hardware used is a dedicated Beowulf Linux cluster, consisting of 3 machines each of which has 2 $3.4GHz$ processors and $4GB$ of RAM. This allows for the parallel implementation of the INLA method on up to 6 taxa at a time. Running the code on a single node machine of similar speed as the cluster nodes (such as a laptop or PC) involves running each INLA fit sequentially. Thus run times are increased approximately 5-fold for the 28 taxa dataset when parallel resources are not available (6 processors handle 6 taxa at a time; this requires 5 such groups of 6 for the 28 taxa problem). As all models considered disjoint-decompose, no message passing between nodes is necessary. The algorithms are perfectly parallelizable and specialist parallel computing code is not strictly required.

The methodology for the pollen dataset in Haslett et al. (2006) suffers from the usual issues related to MCMC based inference. Mixing is poor and convergence is

far from assured, even for runs lasting several weeks on multiple machines. Cross-validation, requiring many re-runs, is therefore all but impossible; for this reason the model validation results obtained here using the INLA method are not contrasted with MCMC based cross-validations.

Although the model in Haslett et al. (2006) appears to perform joint inference of all plant taxa at once, the underlying Dirichlet prior has an implied independence structure, given the constraint. Had inference been performed on each taxon separately and the results scaled post-hoc, the outcome would have been identical.

6.2 Model Description

In the forward problem, the responses are modelled a-priori as independent GMRFs, via an appropriate link function to the $[-\infty, \infty]$ range. Two models are considered in this chapter; a model that is disjoint-decomposed by taxon (referred to as the *by-taxon* marginals model) and a nested model. A grid size of 50×50 was used to model the GMRFs. The granularity of this grid was selected based on expert opinion of the measurement accuracy of the reported climates.

The forward model for each taxon is fitted independently of the others. Inversion of the model (for cross-validation) is also performed separately for each taxon, with scaling to induce the constraint (see Section 3.5.2) performed via Monte-Carlo. The joint predictive distributions are then formed as the normalised product all 28 taxon-specific inverse predictive distributions.

For all models in this chapter, the prior on the latent field is an intrinsic GMRF. The prior is specified through the precision matrix, which has a second order Markov structure matrix S given by

$$S_{ij} = \begin{cases} 20 & d^2(i,j) = 0 \\ -8 & d^2(i,j) = 1 \\ 2 & d^2(i,j) = 2 \\ 1 & d^2(i,j) = 4 \\ 0 & d^2(i,j) > 4 \end{cases} \tag{6.1}$$

where $d^2(i,j) = (i_r - j_r)^2 + (i_c - j_c)^2$ is the squared distance between node i and node j on the $2D$ discrete grid. The subscripts r and c stand for the row index and

column index of a node on the $2D$ grid.

This is the precision matrix derived from a second-order random walk in 2-dimensions that appears in Rue and Held (2005). (Appropriate corrections for the boundaries are made so that the rank of the matrix is $N - 2$, where $N = 50 \times 50$; these appear in Kneib (2006)).

The prior precision matrix Q is then the structure matrix S scaled by the positive hyperparameter κ

$$Q = \kappa S \tag{6.2}$$

The second order Markov structure ensures a stochastically smooth response surface with higher κ values for smoother surfaces.

Unless otherwise specified, all likelihood models are zero-inflated and overdispersed. Furthermore, all counts likelihoods are overdispersed Poisson (Negative-Binomial) or overdispersed Binomial (Beta-Binomial). Where required, the normalisation of the product-of-Poissons type models is performed via Monte-Carlo sampling as per Section 3.5.7.

A maximum of three hyperparameters are included in the model for each taxon:

1. κ; the smoothness of the latent surface.

2. δ; the degree of overdispersion. $\lim_{\delta \to \infty} \Rightarrow$ no overdispersion.

3. α; the power index of the single-process zero-inflation model.

Models with non zero-inflated likelihoods have only the first two hyperparameters.

Fitting of the forward model is performed using the INLA methodology. A GMRF approximation for the saturated posterior of the transformed latent surfaces is made. The Laplace approximation for the hyperparameters is used; however they are then fixed at their posterior modal values, for simplicity. This results in a large computational saving, with seemingly no substantive loss in accuracy (see Section 6.4.1).

To summarise, the goal is to build closed-form posteriors on latent random variables X, which describe response-to-climate surfaces for all plant taxa. i.e. the

posterior $\pi(X|Y,L)$ is sought. There are N_T counts Y associated with locations in the climate space L. A-priori, the X are modelled as independent across the N_T taxa and stochastically smooth across discretized L via intrinsic GMRF specification. The posterior is found approximately using the INLA methodology detailed in Section 4.1.

The independence is due to the disjoint-decomposition; this is first done taxon-by-taxon and the results are shown to have an error associated with this decomposition, using the methods developed in Section 5.3. A nested compositional model, as introduced in Section 5.3.2 is then invoked to reduce this error. This involves a re-ordering and scaling of the data, with inference then proceeding as per the by-taxon model.

Thus, the goal is to fit the decomposed forward model to the modern data Y

$$\begin{aligned} \pi(X,\theta|Y) &= \pi(X|\theta,Y)\pi(\theta|Y) \\ &= \prod_{i=1}^{N_T} \pi(X_i|\theta_i,Y_i)\pi(\theta_i|Y_i) \end{aligned} \tag{6.3}$$

where the right hand side is fitted using the INLA method. Thus, given GMRF priors on $\pi(X_i|\theta_i)$ and univariate, zero-inflated counts likelihoods $\pi(y_{i,j}|x_{i,j},\theta_i)$, the by-taxon posteriors are formed

$$\begin{aligned} \prod_{i=1}^{N_T} \pi(X_i|\theta_i,Y_i)\pi(\theta_i|Y_i) &\simeq \prod_{i=1}^{N_T} \pi_G(X_i|\theta_i,Y_i)\pi_L(\theta_i|Y_i) \\ &= \prod_{i=1}^{N_T} \pi_G(X_i|\theta_i,Y_i) C \frac{\pi(\theta_i)\pi(X_i|\theta_i)\prod_{j=1}^{N_L}\pi(y_{i,j}|x_{i,j},\theta_i)}{\tilde{\pi}_G(X_i|\theta_i,Y_i)}\Bigg|_{X_i=X_i^*(\theta_i)} \end{aligned} \tag{6.4}$$

with subscript G denoting the GMRF approximation and subscript L denoting the laplace approximation. The constant C is determined by normalisation on the grid upon which the $\pi_L(\theta_i|Y_i)$ are defined (see Section 4.1.3). N_L is the total number of locations in L at which there are data. $\pi_G(X_i|\theta_i,Y_i)$ are formed as per Section (4.1.1).

All likelihoods in the by-taxon model are zero-inflated Negative-Binomials:

$$\pi(y_{i,j}|x_{i,j},\alpha_i,\delta_i) = \begin{cases} 1-q_{i,j}+q_{i,j}p_{i,j}^{\delta_i} & y_{i,j}=0 \\ q_{i,j}\frac{\Gamma(\delta_i+y_{i,j})}{y_{i,j}!\Gamma(\delta_i)}p_{i,j}^{\delta_i}(1-p_{i,j})^{y_{i,j}} & y_{i,j}>0 \end{cases} \tag{6.5}$$

where the parameters α_i and δ_i are part of θ_i, $p_{i,j} = \frac{\delta_i}{\delta_i + e^{x_{i,j}}}$ and $q_{i,j} = \left(\frac{e^{x_{i,j}}}{1+e^{x_{i,j}}}\right)^{\alpha_i}$

For those modules in the nested compositonal model that have only one other term in the same nest zero-inflated Beta-Binomial likelihoods are employed:

$$\pi\left(y_{i,j}|x_{i,j},\alpha_i,\delta_i\right) = \begin{cases} 1 - q_{i,j} + q_{i,j}\frac{\Gamma(\delta_i)\Gamma(n_j+\delta_i(1-p_{i,j}))}{\Gamma(n_j+\delta_i)\Gamma(\delta_i(1-p_{i,j}))} & y_{i,j} = 0 \\ q_{i,j}\binom{n_j}{y_j}\frac{\Gamma(\delta_i)\Gamma(y_{i,j}+\delta_i p_{i,j})\Gamma(n_j-y_{i,j}+\delta_i(1-p_{i,j}))}{\Gamma(n_j+\delta_i)\Gamma(\delta_i p_{i,j})\Gamma(\delta_i(1-p_{i,j}))} & y_{i,j} > 0 \end{cases} \tag{6.6}$$

with $p_{i,j} = \frac{e^{x_{i,j}}}{1+e^{x_{i,j}}}$, $q_{i,j} = \left(\frac{e^{x_{i,j}}}{1+e^{x_{i,j}}}\right)^{\alpha_i}$ and n_j is the total count at location j. All other likelihoods in the nested compositional model are zero-inflated Negative-Binomials, as per Equation (6.5) above.

6.2.1 Cross-Validation

Fast updates to the saturated posteriors in Equation (6.3) using the method developed in Section 4.2.3 delivers a method for inverse leave-one-out cross-validation without recourse to re-fitting the model with one less datapoint. Inversion of these leave-one-out posteriors is straightforward due to the discretization of L to a lattice.

The Δ statistic (percentage of points lying outside their leave-one-out cross-validation inverse predictive distribution 95% highest posterior distribution region) is computed across all training data and all taxa. Values much larger than 5% indicate a poor model fit to the data.

The sample density and mean value for $D(l_j)$ (the expected squared distance to the observed climate l_j, under cross-validation) are used to compare models. This is a summary statistic (introduced in Section 2.6.1) on the predictive precision of the model in the inverse problem. In a $2D$ climate space across $GDD5$ and $MTCO$, the D metric for climate $l_j = \{GDD5_j, MTCO_j\}$ is given by

$$\begin{aligned} D(l_j) &= E_{\pi(l|Y,L_{-j})}|l - l_j|^2 \\ &= \sum_{i=1}^{N} \pi\left(l_i|Y,L_{-j}\right)\left(\left(GDD5_i - GDD5_j\right)^2 + \left(MTCO_i - MTCO_j\right)^2\right) \end{aligned} \tag{6.7}$$

where N is the total number of gridpoints.

6.2.2 Fast Inversion of the Forward Model

Inversion of the forward model was performed in Haslett et al. (2006) using MCMC. One-at-a-time sampling based inference of the inverse problem may be avoided altogether when the climates are constrained to lie on a discrete grid. The GMRF approximation in Section 4.1.1 already requires the use of such a grid. The posterior predictive probability mass function for the inverse problem is calculated at all discrete points on the grid and hence may be normalised.

6.2.3 Buffer Zone for Inverse Problem

In order to speed up the numerical inverse predictive distribution algorithm, those points lying outside a region of support were eliminated from the computation. This is possible as the modern training data all lie within a zone defined as the observed modern climate space. Although, theoretically at least, prehistoric climates may have occurred that were outside this zone they are not of interest as the response surface method requires that the palaeoclimates have some modern analogue. The zone, with buffer, is shown in Figure 6.1 for a 50×50 regular grid. Use of this buffer zone is equivalent to placing a prior of zero on points outside the buffer and a uniform prior inside for the inverse predictive distributions.

The response surfaces interpolate / extrapolate the counts data, but only within this bounded region. Areas outside this buffer zone are deemed to be outside the support of the data and are not considered. All internal points are interpolated.

6.3 Zero-Inflation

Many of the counts / proportions data are exactly zero. For the 28 plant taxon dataset a full 63.16% of the modern counts are zero. The data is highly zero-inflated and models not accounting for this will underestimate the expected proportions (see Salter-Townshend and Haslett (2006)).

Zero-modified distributions (Section 2.4.1) provide a way to account for these additional zeros explicitly in the model. However, the zero-modified distribution has two parameters; to facilitate compatibility with the GMRF approximation of Section 4.1.1, likelihoods must be functions of a single parameter.

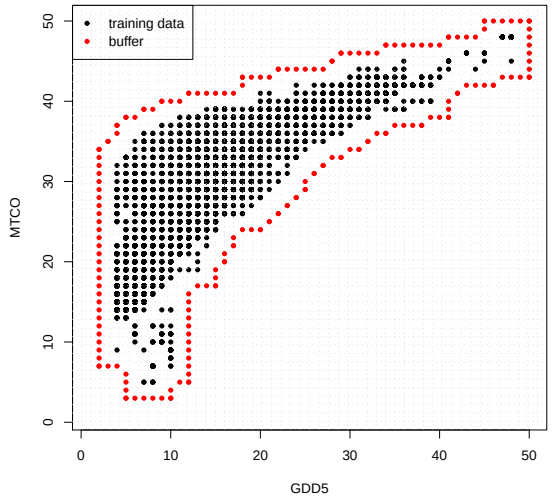


Fig. 6.1: **The modern climate data, when constrained to lie on a regular 50×50 grid are depicted as black dots. Eliminating the points of the grid that lie outside the red buffer cuts down on the number of discrete points at which the inverse predictive density must be computed. In effect, climates of no modern analogue are not considered and are given a-priori weights of zero.**

Models of the type introduced in Salter-Townshend and Haslett (2006) and Section 2.4.1 provide a solution to this problem; in fact this type of model was motivated by the pollen dataset and was developed as a single-process model before compatibility with the GMRF approximation was considered. This model does not apply to every zero-inflated dataset; it was motivated by analysis of the pollen dataset specifically and applies to any dataset in which the probability of potential presence and the abundance when present are governed by a single latent process.

The motivation for using Equation (2.22) in the pollen data analysis is based on logical conclusions on the nature of ecological counts data. Martin et al. (2005) identify four sources of zeros in observation of counts data in ecological datasets: the first two are essential / structural zeros, which they refer to as "true zeros" and the last two are sources of non-essential / sampling zeros which they refer to as "false zeros".

The four sources in Martin et al. (2005) are:

1. Species does not occur at sample site because of the ecological process, or effect under study; habitat unsuitable.

2. Species does not saturate its entire suitable habitat; absent at sample site by chance.

3. Species occurs at the site, but is not present during the survey period.

4. Species occurs at sample site and is present during survey. However, the observer fails to detect it.

Applying the four sources of zero count identified by Martin et al. (2005) to the RS10 pollen dataset informs a model for the extra zeros.

Modelling the data with the response surface technique (Section 2.5.1) shows the first source to occur when the response tends to zero. Therefore this source will be accounted for by any response surface model where the response tends to zero in areas of unsuitable climatic conditions. The second source is due to non-climatic (and unmeasurable) factors, such as plant migration, soil type and topology. The third source is a factor of the finite size of the sample; the plant taxon is present in the location of the sample, but not in the sample itself.

The second and third sources become one and the same (from a climatic viewpoint) if there is a uniform pollen rain in regions where the plant taxon occurs. This is one of the principles of pollen analysis introduced in Birks and Birks (1980) and is due to atmospheric turbulence mixing the pollen and spores (see also Smol et al. (2001)). This single source of extra zero is then referred to as the non-climatic source and is the primary concern here.

The final source of zero count is accounted for by the counts likelihood function. For example, using a Binomial likelihood for the counts, given there are no zeros arising from sources one to three, will deliver this final source of zero. These are effectively non-extra zeros and as such are part of the non-zero-inflated model.

Therefore, a zero-inflated model is required to account for a single source of extra zeros, the non-climatic source. The single-process model in Section 2.4.1 is advocated for the pollen and climate dataset. The justification for such an approach is that the environmental pressure exerted by climate influences the probability of presence and the expected proportion if present in a related fashion.

This is not to say the two are the same; all that is required to justify the single-process zero-inflated model is that a single process governs both presence and abundance when present. This is the case if the random extra zeros due to non-climatic factors are more likely to occur in regions where the climate is less suitable. This follows as a natural assumption; non-climatic factors promoting absence (such as unsuitable soil type) will be more likely to cause a plant type to fail in climatic regions where the plant is already struggling than in conditions under which it thrives.

The power link in Equation (2.22) is motivated for the pollen dataset by several simple observations:

1. The response r (propensity to produce pollen as a function of climate) and the probability of potential presence q both have limits of zero and one.

2. As the climate approaches total unsuitability, the probability of potential presence must tend to zero

$$\lim_{r \to 0} q = 0 \tag{6.8}$$

3. In the limit of the response tending to unity, that taxon *must* be observed (propensity to produce pollen is absolute). Therefore, the probability of potential presence tends to unity

$$\lim_{r \to 1} q = 1 \tag{6.9}$$

The simplest functional relationship between the response and the probability of potential presence is that q is some power of r. Limiting this power to to be positive ensures that q increases monotonically with r (see Section 2.4.1).

The validity of this simple, monotonic relationship between probability of potential presence and abundance when present is demonstrated in Figure 6.2. Plots such as these suggest a positive relationship between probability of potential presence and abundance when present.

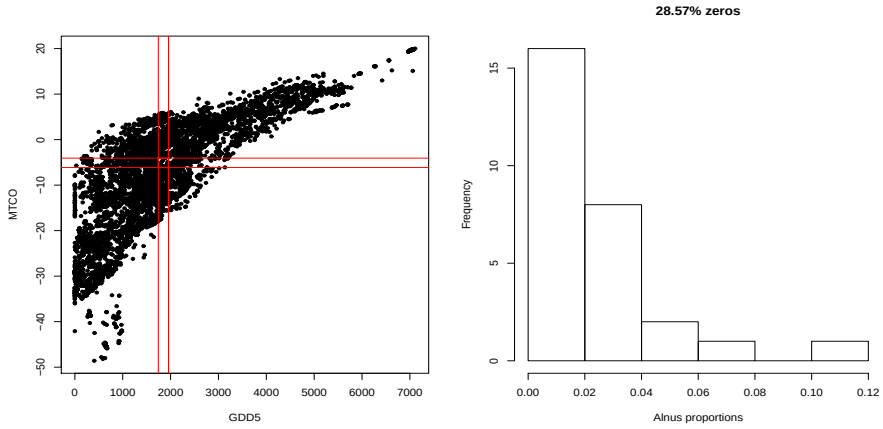
Further evidence from the data that the power-law monotonic relationship between probability of potential presence and abundance when present is given by Figure 6.3. Neither probability of potential presence or abundance when present are directly estimable from the data. However, the probability of non-zero proportion and the positive abundances act as crude proxies for these values. These proxies are plotted in Figure 6.3.

6.4 Results

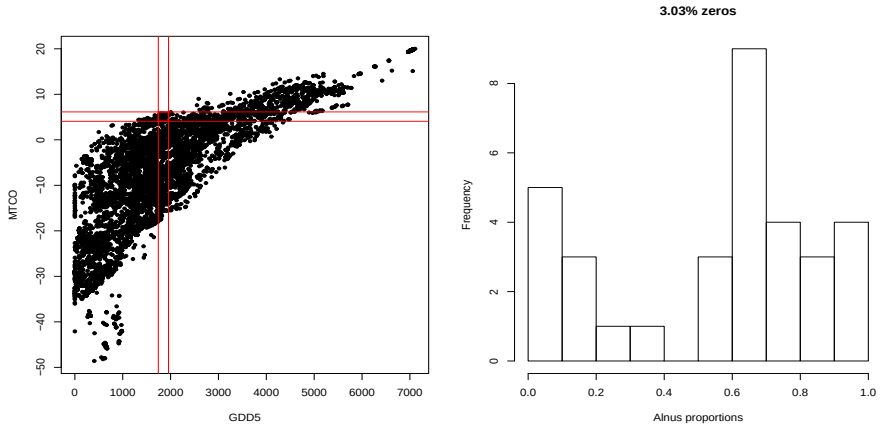
6.4.1 Treatment of Hyperparameters

There are three hyperparameters specified in the model for pollen response to climate. These are; (i) the smoothness of the response across climate space, (ii) an overdispersion parameter and (iii) the power index for the zero-inflation. In order to integrate out the hyperparameters, the model is fitted for each discrete triplet and a weighted average is calculated, as per Equation (4.22).

Computation time and memory usage are of real concern in performing the cross-validation. Evaluating the inverse predictive distributions only at the modal value of the posterior for the hyperparameters reduces both run-time and memory usage by a factor of the number of discrete triplets. Even a coarse grid across the hyper-



(a)



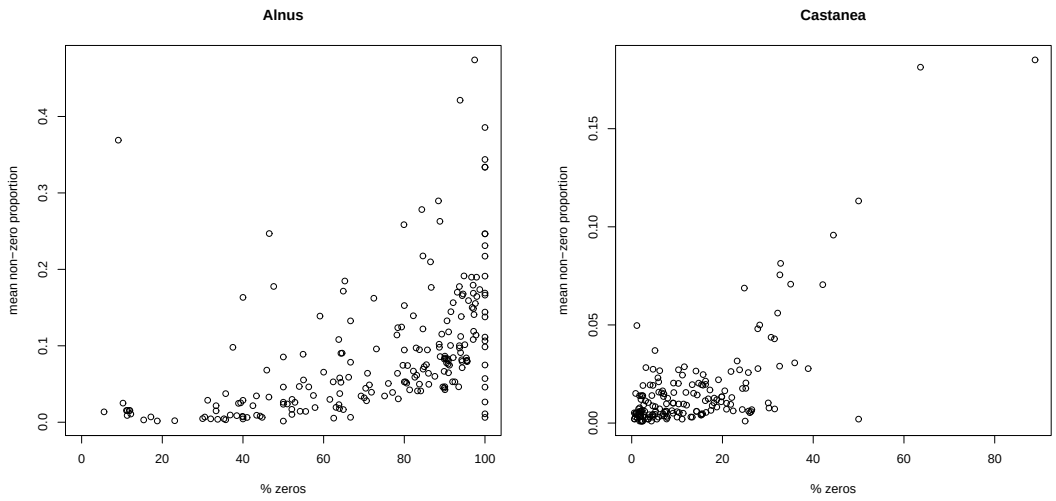
(b)

Fig. 6.2: **Example histograms of the raw proportions data for the plant taxon *Alnus* (Alder).**

The left panels show the selected local region of climate space in a red box within red-cross-hairs. The black dots are sampling locations for the modern data.

The accompanying right panels show the histogram of proportions data for *Alnus* within the (red) boxed region. Figure (a) shows the proportions data in a region of low abundance; many of the proportions are exactly zero in this region of climate space.

Figure (b) shows the proportions data in a region of high (and highly variable) abundance. Fewer of the proportions are exactly zero in this region of climate space than in Figure (a), however there are still additional zeros.



(a) (b)

Fig. 6.3: **Example plots of mean positive proportion against percentage of zero-proportions for *Alnus* (Alder) and *Castanea* (Chestnut). The local probability of presence is estimated by the percentage of non-zero proportions in a region. Similarly, positive abundance is estimated by the mean of the non-zero proportions in a region. Moving across all climatic regions and calculating these two numbers in the local region (locality size as per Figure 6.2) indicates that abundance and probability of presence are positively related. The form of the relationship between mean non-zero proportion and probability of a zero is a positive relationship for both taxa above. The two taxa differ in the details of this relationship.**

parameter space results in an average of 54 for the number possible configurations of the three hyperparameter values for each taxon.

Evaluation at the mode represents an approximation to the integration over the hyperparameters. The data are not overly sensitive to the hyperparameters as they depend on them indirectly. However, an evaluation on the impact of using the model fitted at the modal hyperparameters (modal-hyperparameters approximation) must be conducted to justify the saving in computation.

Figure 6.4 shows the effect on the predictive power in the inverse sense for a model evaluated using the modal hyperparameters and the same model having integrated out the hyperparameters (the model used is that in Section 6.4.5). The predictive power of a model is summarised using the performance statistic introduced in Section 2.6.1. The expected squared distance to the true left-out observation is calculated, with expectation taken w.r.t. the posterior predictive inverse cross-validation distribution. The kernel density estimate of the expected squared distance D to the left-out observation l_j (climate) appears to be insensitive to the use of the modal-hyperparameters approximation. This result, along with the large computational saving of about two orders of magnitude, justifies the use of the modal-hyperparameters approximation.

6.4.2 Marginals Model

Inference using the decomposed by-taxon model was applied to the pollen dataset. Under the marginals model (Section 3.2.1) each taxon is taken as independent of all the others. The individual taxa returned cross-validation Δ statistics that were consistent with theory; i.e. about 5% of points fell outside their 95% highest predictive distribution region (see Figure 6.5).

The mean value of Δ is 4.15% (see Figure 6.5). This is around the theoretical value of 5% or less, given a model that fits the data (see Section 5.3.1).

However, bringing the cross-validation predictive densities for all taxa together reveals an error. For 2 taxa there are $\binom{28}{2} = 378$ possible choices of 2 taxa; for 3 taxa there are $\binom{28}{3} = 3276$ unique combinations; etc. Taking a random 10 of these $\binom{28}{T}$ possible combinations for each of $T = \{1, \ldots 27\}$ and computing the $\Delta(T)$ statistic for each gives an indication of the relationship between T and Δ.

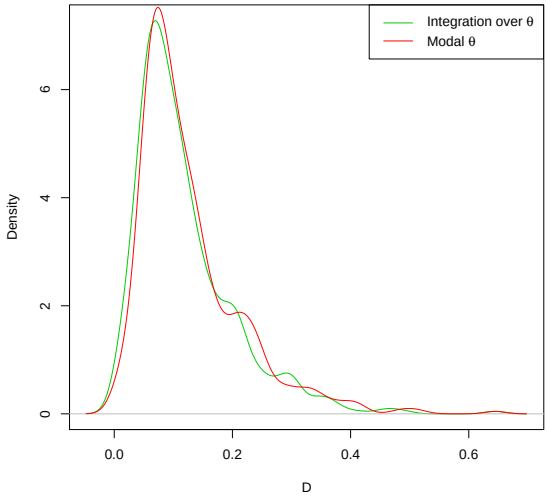


Fig. 6.4: **The kernel density estimate for $D(l_j) = E_{\pi(l|L_{-j},Y)}[(l - l_j)^2]$, across all 7742 j.**

The green density is the correct method in which the hyperparameters are averaged over, weighting the inverse cross-validation posterior predictive distributions by the posterior for the hyperparameters. The red line represents results using a far faster approximation to this; the inverse cross-validation is performed only once, using the posterior modal hyperparameters. This approximation is found to be both excellent and far cheaper. Run times are reduced by almost two orders of magnitude.

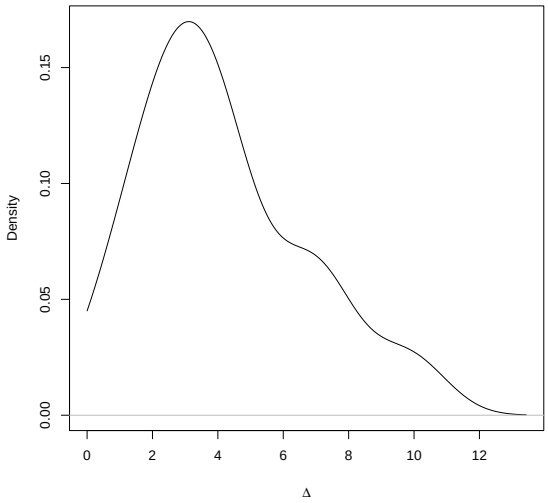


Fig. 6.5: **The percentage of points falling outside their corresponding** 95% **highest cross-validation posterior predictive density for taxon** i **is** Δ_i**. This density estimate of** Δ **is based 28** Δ_i **values. The mean value for** Δ **is** 4.15%**.**

If the joint model (all taxa) disjoint-decomposes exactly, then $\Delta(T)$ should stay around 5% (while $\overline{D}(T)$ goes down). This is clearly not the case, as shown in Figure 6.6. Δ increases with T, showing that the plant taxa are not conditionally independent given climate. Thus the joint model does not disjoint-decompose exactly. Figure 5.1 indicates that this occurs when the response surfaces are correlated, given climate.

Predictive Power of the Model

The $\Delta(T)$ plot in Figure 6.6 clearly demonstrates that the disjoint-decomposition of the model by taxon is not a good fit to the data as error rate increases with the more taxa modelled. Even still, the inverse predictive power of this model as measured with $\overline{D}$ increases with increased T. This result is shown in Figure 6.7. $\overline{D}$ (the expected squared distance to the observed normalised climate) decreases as T increases. It shows that as more taxa are modelled, the precision of the inverse predictive densities becomes higher. This relationship is not linear.

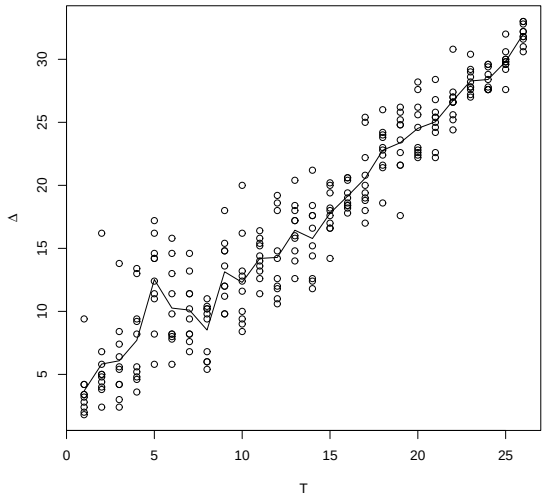


Fig. 6.6: **Plot of Δ (percentage of points lying outside their corresponding leave-one-out inverse cross-validation 95% highest predictive region).**
The forward model for each pollen taxon is fitted independently. Inverse predictive densities are computed for each point in the training dataset for cross-validation. For each of $T = \{1, \ldots 27\}$, 10 of the $\binom{28}{T}$ possible combinations are chosen at random and the joint Δ value is computed for each. These are shown as a scatter plot of Δ against T; there are 10 values at each value of T. The mean across the 10 values is shown with a line. This is the mean value for $\Delta(T)$ and it shows that Δ increases linearly with T to a value of 34.54% for all 28 taxa.

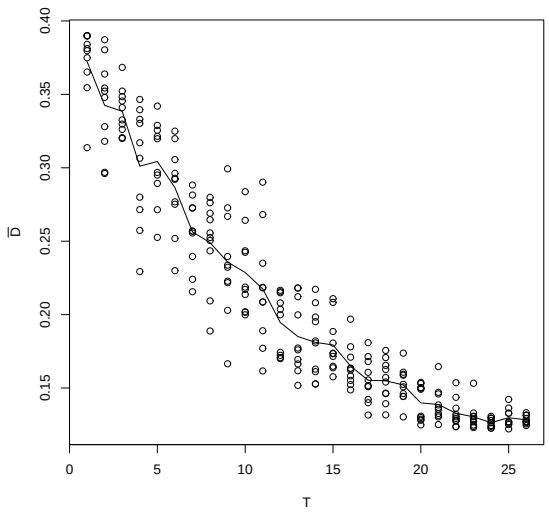


Fig. 6.7: **Plot of the expected squared distance to the observed normalised climate is $\overline{D}$.**

The forward model for each pollen taxon is fitted independently. Inverse predictive densities are computed for each point in the training dataset for cross-validation. For each of $T = \{1, \ldots 27\}$, 10 of the $\binom{28}{T}$ possible combinations are chosen at random and the mean $\overline{D}$ value is computed for each. These are shown as a scatter plot of $\overline{D}$ against T; there are 10 values at each value of T. The mean across the 10 values is shown with a line. This is the mean value for $\overline{D}(T)$ and it shows that $\overline{D}$ decreases with T.

6.4.3 Uncertainty in Climate Measurements

There are errors associated with the modern climate data measurements. Expert opinion [1] suggests that GDD5 measurements may be taken as plus or minus 100 degree days and MTCO as plus or minus 2 degrees. A crude method for correcting for these uncertainties post-hoc is to convolve the cross-validation inverse posterior predictive distributions with a kernel that has a width on this scale.

A $2D$ truncated Gaussian kernel convolution was applied to each of the leave-one-out posterior predictive distributions for the inverse cross-validation. This has the effect of increasing the variance of the inverse predictive distributions. For example, in the marginals by-taxon model below, convolution with a Gaussian kernel of variance 3 and truncation distance equal to 3 grid spacings was used.

The resulting predictive distributions are still defined on the regular grid. The value at each gridpoint becomes a weighted average of its neighbours, with weights determined by the Gaussian kernel. The truncation of this kernel is for computational reasons; beyond 3 grid spacings the weights are extremely low (0.248% of the central weight) and therefore not computed.

This Gaussian blurring causes the Δ statistic to drop from 34.54% (very poor fit) to 14.2%. This is around half of the non-blurred version. In fact, the slope of the line in Figure 6.6 roughly halves. At the same time, the predictive power statistic $\overline{D}$ increases from 0.14 to 0.148, representing a loss in accuracy of just 5%. However, this is a crude and somewhat ad-hoc method to account for the uncertainty in the reported climates, the form of which is unknown.

6.4.4 Zero-Inflated Model

The results presented above include an explicit modelling of the zero-inflation of the data, as per Section 6.3. If the data is not modelled with a zero-inflated likelihood and an overdispersed likelihood is used, as per Haslett et al. (2006), then the results will differ.

Analysis of the impact of explicit modelling of the extra zeroes was performed through comparison with a model in the spirit of Haslett et al. (2006). Specifically, two models with identical priors for the latent surfaces but with differing likelihoods

[1] obtained via correspondence with Brian Huntley

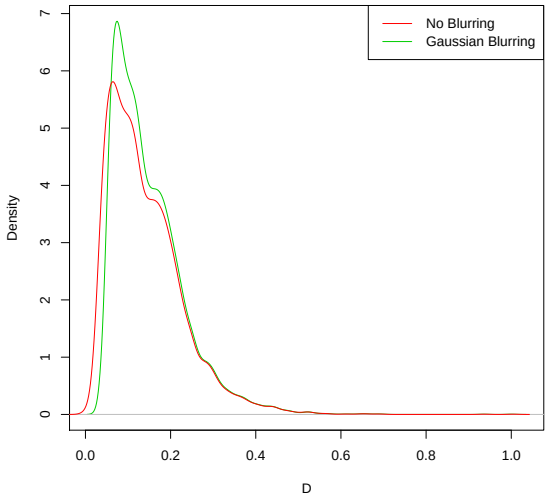


Fig. 6.8: **The effect on the sample density of the expected squared distance $D(l_j)$ to the left-out observation l_j in inverse cross-validation. Although the kernel convolution with a truncated Gaussian kernel results in a halving of the number of points falling outside the 95% HPD region of the associated cross-validation inverse predictive distribution, the $D(l_j)$ statistic is not dramatically effected. The mean $\overline{D}$ for the non-blurred version is 0.14 and for the Gaussian blurring, $\overline{D} = 0.148$.**

were fitted to a subset of the data (500 points).

The likelihood for the non zero-inflated model is an overdispersed scaled Poisson for each of the individual taxa. Mixing the Poisson with a Gamma function leads to a negative-Binomial likelihood. Thus, for taxon i at climate location index j

$$\pi(y_{ij}|x_{ij},\delta_i) = \frac{\Gamma(\delta_i + y_{ij})\delta_i^{\delta_i}(n\lambda_{ij})^{y_{ij}}}{y_{ij}!\Gamma(\delta_i)(n\lambda_{ij}+\delta_i)^{\delta_i+y_{ij}}} \tag{6.10}$$

with $\lambda_{ij} = e^{x_{ij}}$

The zero-inflated model uses a mixture of this likelihood with a point mass at zero. The size of the point mass at zero is $1 - q_{ij}$

$$\pi(y_{ij}|x_{ij},\delta_i,\alpha_i) = \begin{cases} 1 - q_{ij} + q_{ij}NB(0;x_{ij},\delta_i) & y_{ij} = 0 \\ q_{ij}NB(y_{ij};x_{ij},\delta_i) & y_{ij} > 0 \end{cases} \tag{6.11}$$

with $q_{ij} = \left(\frac{e^{x_{ij}}}{1+e^{x_{ij}}}\right)^{\alpha_i}$

and $NB(y_{ij};x_{ij},\delta_i)$ given by Equation (6.10).

Thus, for each of the 28 taxa, there are 2 hyperparameters (κ_i and δ_i) to be estimated for the first model and 3 hyperparameters (κ_i, δ_i and α_i) for the zero-inflated model. This zero-inflated negative-Binomial likelihood is the same used to produce the results in this section thus far. Comparison between the results for the two models using the Δ and $\overline{D}$ inverse cross-validation summary statistics is shown in Table 6.4.4. Also shown is the mean posterior modal value for the overdispersion hyperparameter δ_i across all taxa $i = 1,\ldots,28$. These results indicate that the zero-inflated model is a better fit to the data and that the likelihood variance (controlled by δ) is reduced. This leads to higher predictive power, as revealed by a lower mean expected squared distance to the observations $\overline{D}$.

6.4.5 Nested Compositional Model

Section 3.5.6 introduced the concept of nesting structures within compositional counts data. It is demonstrated in that section that lack of appropriate modelling of the nesting will lead to erroneous inferences and that this may manifest as an increased inverse predictive density error as measured by the statistic of the percentage of points lying outside their corresponding 95% highest predicitve density region.

Model	Δ	$\overline{D}$	$\overline{\delta}$
Negative-Binomial	10.8	0.179	0.252
Zero-inflated Negative-Binomial	8.0	0.146	1.40

Table 6.1: **Summary statistics for model fit and comparison for an overdispersed model (Negative-Binomial) and an overdispersed and zero-inflated model (zero-inflated Negative-Binomial). The overdispersed and zero-inflated model is a better model fit to the data (lower Δ statistic) and has a higher precision in inverse predictive ability, as measured with $\overline{D}$. The degree of overdispersion δ is reduced, as shown by a higher average δ. In fact, δ_i was higher for each taxon; results shown here are across all 28 taxa. Gaussian blurring as per Section 6.4.3 is employed.**

Applying the nesting structure depicted in Figure 6.9 leads to markedly different cross-validation results for the RS10 pollen dataset. This is referred to as the *nested model*. Although it is disjoint-decomposable, the taxa are no longer modelled as conditionally independent given the climate; the individual components of a nest still are.

Given the nesting structure, errors associated with the inference-via-marginals models are minimized; this is because down to all but the final level in each nest there are only two categories as the preceding nest is divided in a binary split. Thus each nest may be modelled as with Binomial type likelihood and is thus univariate.

The sum-to-unity constraint forces the two components of the split to be perfectly negatively correlated. Any other form of interaction is swamped and cannot therefore impact the model. As only one branch of each split need be modelled (the response of the other is exactly 1 minus the first), normalisation is not required.

For the 28 taxa there are 32 components of the nested model. The overall Δ statistic improved from 14.2% for the by-taxon model to 6.12% for the nested model. Based on this cross-validation statistic, the nested model makes considerably fewer errors for the inverse problem on this dataset.

Fortunately, the reduction in error does not appear to come with a reduction of predictive accuracy. The sample density of the D values for the by-taxon marginals model and the nested model is shown in Figure 6.10. The nested model produces

lower values for the expected squared distance to the observation, showing greater predictive power in the inverse problem. The mean values for D were found to be

- $\overline{D} = 0.148$ for the by-taxon model.
- $\overline{D} = 0.125$ for the nested model.

Figure 6.12 shows three example leave-one-out inverse cross-validation predictive densities and the associated left-out points for both the marginals-by taxon model and the nested model, with nested structure as per Figure 6.9.

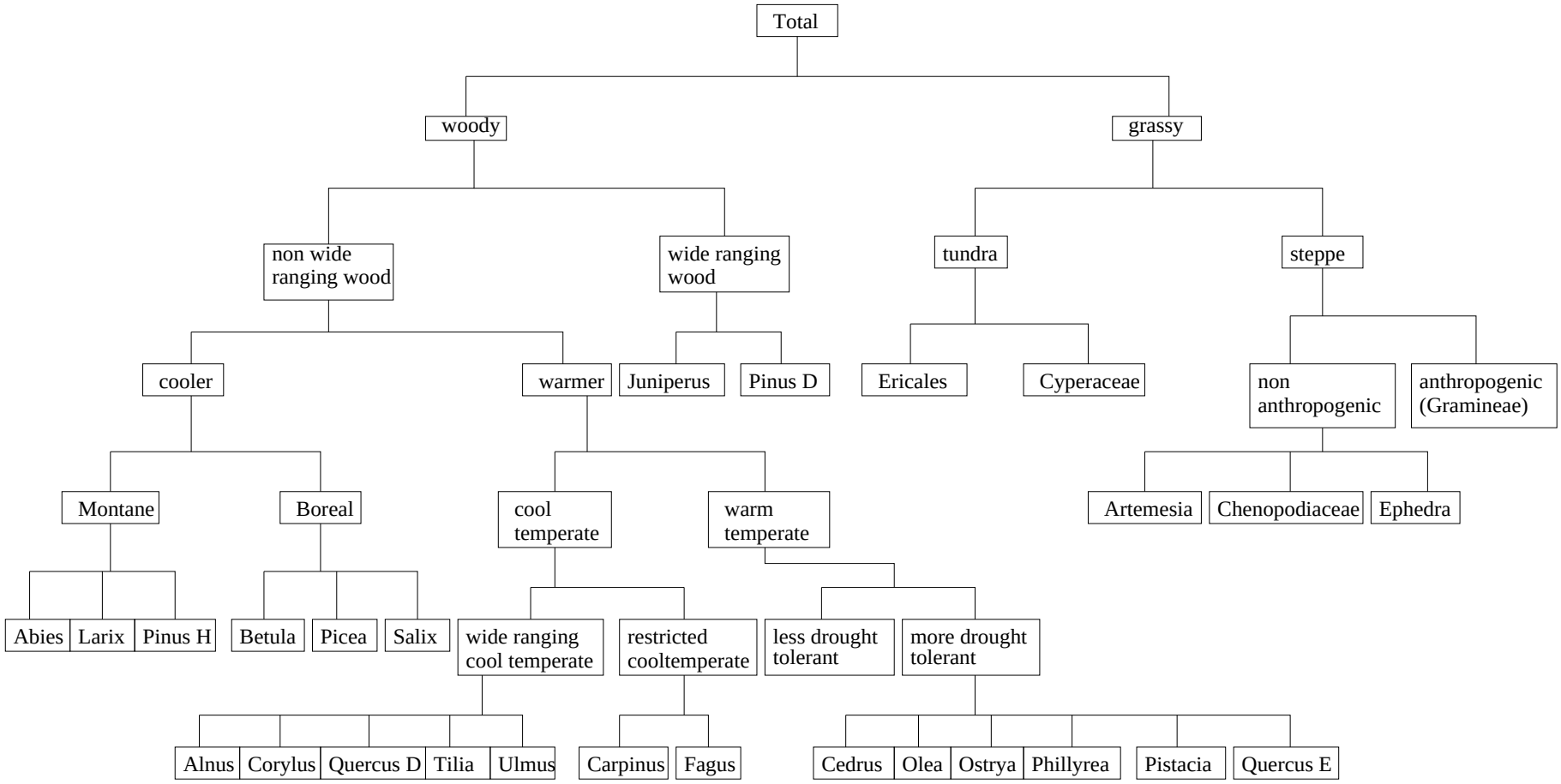


Fig. 6.9: **The layout of a compositional nesting structure in the pollen data. The taxa are grouped according to a progressively finer discrimination by botanic similarity. Although a seemingly very complicated model, it facilitates decomposition of the model, and thus the inference, into smaller independent tasks that may be run in parallel. This is certainly not the only such structure that could be imposed on this dataset. (Obtained through correspondence with Brian Huntley.)**

Conditional Independence of Lowest Levels

If the nesting structure given by Figure 6.9 is correct then the components of each nest should be conditionally independent, given the climates and the constraint (where there are only two components they are fully negatively correlated and the question is non-applicable).

The results of applying the same method of plotting Δ against T as in Section 6.6 for the nest labeled "more drought tolerant" is shown in Figure 6.11. There is not a wide range of values across which to evaluate $\Delta(T)$ (maximum T is 6 taxa) and therefore it is not straightforward to assess whether Δ varies with T. For all other nests, such as that labeled "wide ranging cool temperate", there are even fewer values of T.

The figure shows that Δ does not increase as more taxa are modelled. This results suggests that the taxa within the "more drought tolerant" nest are conditionally independent, given climate.

6.4.6 Outliers

The $\overline{D}$ statistic and the distribution of $D(l_j)$ may be used to compare the inverse predictive power of two or more models. Analysis of the individual expected squared distances $D(l_j)$ is also used to detect outliers as these will have larger than expected D values, under the fitted model.

This approach was applied to the pollen training dataset and the highest D value for the nested model was found to be $D = 0.661$. Figure 6.13 shows the inverse predictive distributions for this datum given the nested model. Examination of the data at that location revealed there to be only two taxa present, *Artemesia* (sage-brush and wormwood) and *Chenopodiaceae*, both of which are non-anthropogenic grasses. The reported climate is GDD5 = 128 and MTCO = −168, a cold-in-winter climate with a very short growing season. The fitted responses for these taxa show that these grasses thrive in temperate climates. However they are indigenous to the physical region and are hardy. Although they may in fact grow sparsely at that location, they will nevertheless dominate the assemblage. Most interestingly, this outlier is the sampling location with the highest altitude; 5100 metres above sea level, in the mountains in Kashmir.

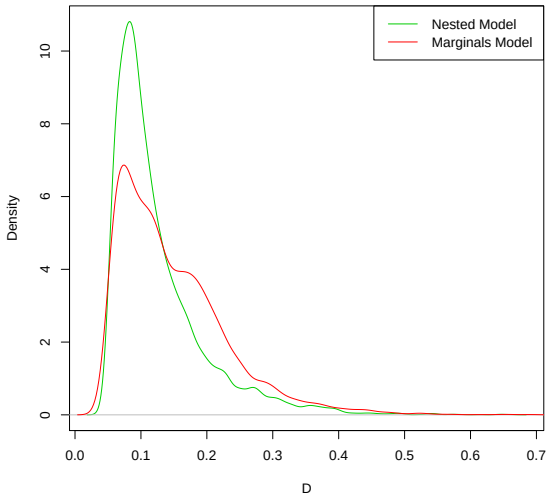


Fig. 6.10: **Sample densities of the expected squared distance to the true climates under the leave-one-out inverse cross-validation predictive distributions. The nested model has a lower mean expected squared distance $\overline{D} = 0.125$ than the by-taxon marginals model $\overline{D} = 0.148$, showing a greater ability to give accurate inverse predictions, given the training data.**

Investigation of the effect of extreme altitude on outliers expands on this result. Defining those data that have a D metric greater than the 95% quantile for D (i.e. the top 5% least well predicted data) to be outliers and plotting the sample density of the altitudes associated with those outlying data reveals thicker tails than the sample density of the entire dataset. This is shown in Figure 6.14.

A similar analysis for another climate variable for which data are available is shown in Figure 6.15. In this example, the ratio of actual to potential evapotranspiration (AET/PET) is used. This figure suggests that low values of AET/PET - indicating arid conditions - tend to be less well predicted by the model as the red line (sample density for AET/PET of the subset of the data with the top 5% worst predicted climates) shows a sample with a lower proportion of high values of AET/PET. This indicates that modelling of these AET/PET values should be carried out.

Figure 6.14 demonstrates that there is a relationship between the altitude of the sampling location and the probability that the datum will be an outlier in

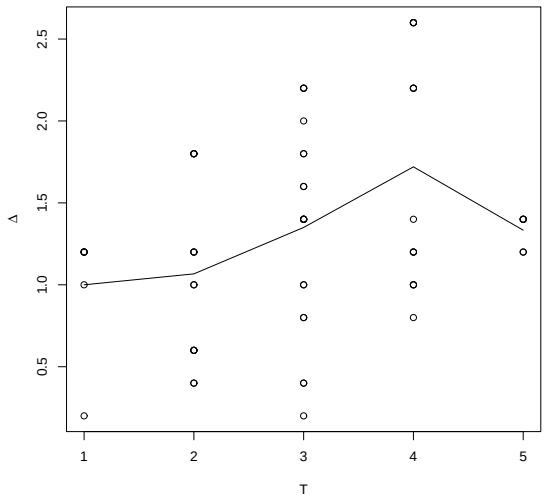


Fig. 6.11: **Plot of Δ (percentage of points lying outside their corresponding leave-one-out inverse cross-validation 95% highest predictive region). The forward model for each pollen taxon in the nest labeled "more drought tolerant" is fitted independently. Inverse predictive densities are computed for each point in the training dataset for cross-validation. For each of $T = \{1, \ldots 5\}$, the joint Δ value is computed for all $\binom{6}{T}$ possible combinations. These are shown as a scatter plot of Δ against T. The mean across the $\binom{6}{T}$ values is shown with a line. This is the mean value for $\Delta(T)$ and appears to show that Δ does not vary substantially with T.**

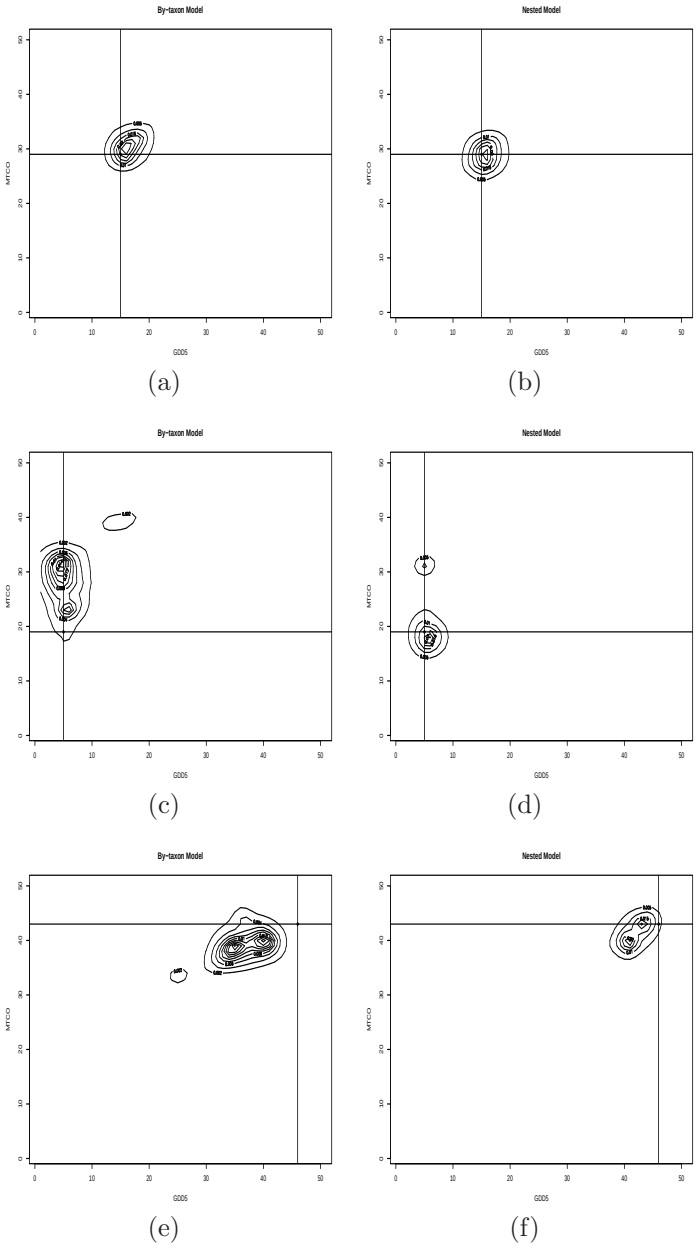


Fig. 6.12: **Some examples of inverse cross-validation predictive distributions for the marginals-by-taxon model and the nested model associated with Figure 6.9. Despite delivering sharper predictive densities, the nested model makes fewer errors, as measured by the Δ statistic.**
The true climate locations are marked with a dot; the intersection of two lines highlights the location.

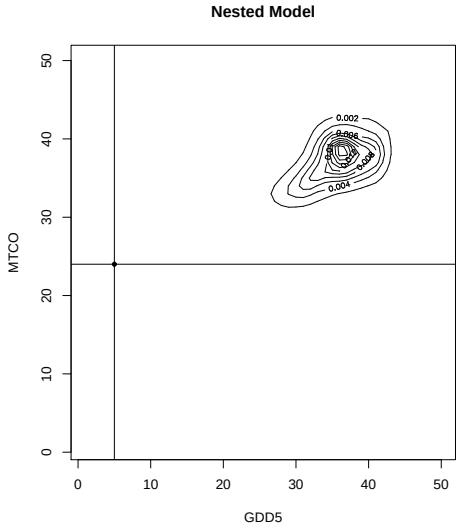


Fig. 6.13: **The outlier identified by the greatest expected squared distance between the observation and its associated leave-one-out inverse cross-validation predictive distribution ($D = 0.661$). Inspection of the data reveal that there are only two taxa present at this sampling location and that it is the sampling location with the highest altitude; 5100 metres above sea level. The observed location in climate space is marked with a dot and cross hairs.**

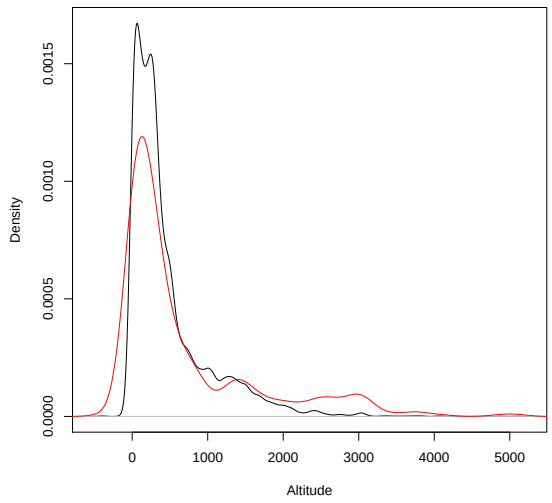


Fig. 6.14: **The sample density of 2 sets of altitudes from the RS10 dataset. The black density represents the entire dataset. The red line represents the sample density of those altitudes whose expected squared distance $D(l_j)$ to the observed climate location l_j is in the top 5% of all D. This density has thicker tails, suggesting that extreme altitude, both high and low, has a negative effect on the predictive power of the model for the inverse problem.**

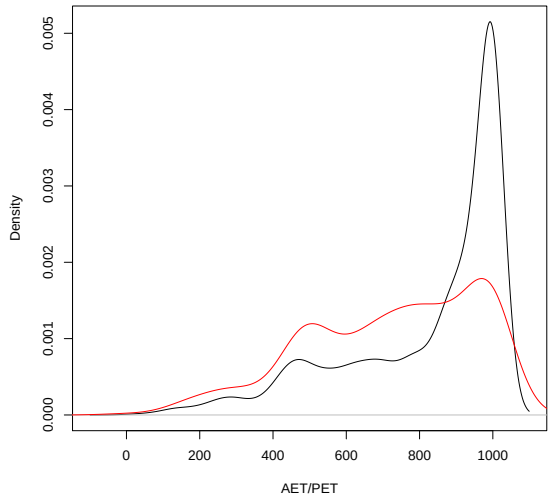


Fig. 6.15: **The sample density of 2 sets of AET/PET from the RS10 dataset.**

The black density represents the entire dataset. The red line represents the sample density of those AET/PET values for which the expected squared distance $D(l_j)$ to the observed climate location l_j is in the top 5% of all D. $2D$ climate l_j is defined in terms of $GDD5$ and $MTCO$.

the inverse (climate predictive) sense. There are only 9 datapoints whose cross-validation climate predictive distributions place exactly zero probability mass at the correct location. The 9 altitudes, in metres, associated with these locations are

$$\{3030, 3050, 2000, 3085, 2790, 3030, 3050, 15, 168\}$$

7 out of 9 of these have altitudes that are above the 98% quantile. This again suggests a strong link between extreme altitude and lack of fit. The final two values suggest that this is not the only factor.

6.5 Conclusions

6.5.1 Advances

The cross-validation model fit statistic Δ is a simple measure of fit; if the model fits the data, then the theory dictates that Δ should be around 5% or less. Simulated data, for which the model is known, in Chapter 4 shows that values as high as 10% are possbile. Values above this have not been observed for toy problems that are on the same scale as the pollen problem.

This allows for a simple, albeit crude, check of model fit to be made. For example, the disjoint-decomposition by-taxon of the model did not appear to be a good model fit to the data. More importantly, a positive linear relationship between the number of taxa used in the model and the Δ statistic was established empirically. This relationship suggests that the disjoint-decomposition by-taxon is a poor model due to a non-zero dependence between the taxa, given the climate.

Comparison between models is performed using the cross-validation statistic $\overline{D}$. $D(l_j)$ is the expected squared distance to the climate observation l_j, under the model and given all other data. The lower this value, the more accurately the trained model can predict or reconstruct climates, given a counts vector. Thus, models with lower $\overline{D}$ are preferred.

Δ and $\overline{D}$ are used constructively to illustrate the importance of modelling advances made since Haslett et al. (2006) such as zero-inflated counts likelihoods. However, the greatest contribution to modelling the RS10 dataset is the delivery of working methods to compare and validate such models, using the INLA based

approximation techniques. MCMC based methods fail to achieve this objective due to the computational burden associated with sampling based methods.

Estimation of the various hyperparameters of the model is also performed quickly using approximation methods. This is difficult in the extreme using MCMC based methodology due to issues of mixing.

The discrete grid on which the data are made to lie facilitates fast inversion of the forward model. Normalisation of the posterior for climate, given count is performed numerically. This further speeds up cross-validation in the inverse sense.

A new nested compositional model is introduced. Assessment of the particular nesting structure reported on here shows a marked improvement in this model over the marginals by-taxon model. Most of the apparent conditional dependence between the various plant taxa has been accounted for via the nesting. This is achieved without incurring large computational overhead or developing additional code.

6.5.2 Shortcomings

Work has not yet been carried out to determine what value of $\overline{D}$ determines a model that may be said to fit the data. One method for performing such a calculation is to use Monte-Carlo methods to simulate data from the fitted model; the $\overline{D}$ statistic based on the simulated data then represents the $\overline{D}$ value for a model that fits the data.

Although the nested model defined by the diagram in Figure 6.9 leads to a welcome decrease in both Δ and $\overline{D}$, it is not clear from these statistics that the model may be said to fit the pollen data. Further criticism of the nesting structure is certainly required; alternative nesting structures should also be explored. The structure in Figure 6.9 is based on expert opinion, but comes with caveats; it is by no means a final statement of definite nesting structure.

Results for models with additional climate dimensions, for which data is available, have not been presented here. Perhaps the by-taxon decomposition of the model is a satisfactory assumption for a $3D$ climate. Taxa that are not conditionally independent given 2 climate dimensions may be conditionally independent given 3 or more climate variables.

The post-hoc Gaussian-blurring of the climate predictive distributions is ad-hoc and not thoroughly investigated. It is an attempt to account for the uncertainty in the reported climates for the training data in a cheap and simple way. Further thought is required in this area.

Both Δ and $\overline{D}$ were used with success to identify problematic data. Sampling locations associated with extreme altitudes were found to have an association with being outliers with respect to the fitted model. However, the definition of outlier is not firmly established here and incorporating the altitudes into the model has not been addressed.

Chapter 7

Conclusions and Further Work

Application of a cutting edge statistical methodology (INLA) to a large scale palaeoclimate reconstruction project has delivered two important research contributions. Firstly, the modelling associated with the palaeoclimate problem has been advanced. In particular, inference is now performed quickly, without recourse to MCMC. Secondly, the INLA methodology itself is challenged and extended. A method for fast cross-validation in the inverse sense is introduced. The richer models developed for the palaeoclimate problem are guided by the fast model validation procedure.

Both of these contributions are the subject of ongoing research. No claim is made to have developed a finished model for the palaeoclimate project; indeed the imperfection of the models contained in this thesis is demonstrated. Further extensions of the INLA method to cope with higher dimensions and modelling of data that are not conditionally independent is desirable.

7.1 Conclusions

The motivating pollen dataset is massively zero-inflated. In Haslett et al. (2006), this over-abundance of zero counts was dealt with via an overdispersion model. This method underestimates the mean of the response surfaces and overestimates the variance.

Zero-modified distributions are a flexible class of model that can account for zero-inflation of counts data. Typically, these models require an extra parameter to model the probability of potential presence. For spatial data, modeling of non-parametric

response surfaces for zero-inflated counts data in this way requires doubling the number of latent parameters. MCMC inference may slow drastically due to this large increase in random variables.

A single-process zero-modified distribution is therefore developed that requires a single extra (hyper)parameter. This model is valid for data in which the probability of extra zeros and the abundance, given presence, are functions of a single underlying process. Justification of such a model is provided for the pollen dataset and a simple, yet flexible model for this data is constructed. Synergy of this model with the emerging INLA inference procedure is demonstrated, as is incompatibility of INLA with traditional zero-inflation models.

Multivariate counts data, constrained to sum to a total, may exhibit high correlation, even after taking the constraint into account. Modelling such dependencies is inherently difficult. Nested models may provide a solution for breaking down such dependencies in compositional counts data. The main advantage of such models is that, given the nesting structure, the joint model will disjoint-decompose exactly. This means that inference may be performed on many separate, smaller problems in parallel. More importantly, decomposition of the joint model is required in order to apply the fast INLA inference methods.

Gaussian and Laplace approximations are fitted to the posteriors for the parameters and hyperparameters of the pollen dataset. This results in a dramatic reduction in both the forward and inverse stages of the non-parametric inference.

The forward fitting stage using approximations takes approximately 40 minutes for all 32 pollen types in the nested model. For a similar, yet cruder (non-zero inflated), model MCMC runs were previously ran for up to several weeks. Cross-validation was impracticable and hyperparameters had to be fixed a-priori.

The marginals model required for use of the INLA methods may break down with large numbers of correlated taxa. This can be tested for using inverse cross-validation statistics such as the percentage of training data points that lie outside their corresponding 95% highest posterior predictive density region.

If the data are in fact nested, then this enables decomposition of the problem. This cannot be achieved without knowledge of the structure of the nesting. These nested models are a novel aspect of compositional data analysis. One such nesting

structure, given by expert opinion, is applied to inference on the pollen dataset, with promising results.

Cross-validation is an important tool in model validation. A contribution in this thesis is a method for performing fast cross-validation in the inverse sense, using the INLA inference procedure. Approximating the saturated posterior for the latent variables with a multivariate normal permits fast updates to be made to correct for leaving out a single datum. This procedure has application in any spatial context where the interest is in the inverse problem and the forward model posterior is approximately Gaussian.

7.2 Further Work

This is perhaps the most important section in the thesis as there are several outstanding challenges in this project. Some of these are already the subject of preliminary investigation.

7.2.1 3 Dimensional Climate Space

The RS10 pollen dataset includes more than just two climate variables. Expert opinion in the botany community advocates using at least three in building the response surfaces.

In theory, the Gaussian approximation works in the same way in any number of dimensions; however, a second order intrinsic GMRF prior (such as used in Chapter 6) is far less sparse in $3D$ than in $2D$. Thus, the fast sparse-matrix algorithms employed in fitting a GMRF approximation to the posterior will be disproportionally more labourious. In addition, the grid size G scales as G^D where D is the dimension. Even moving up a single climate dimension from $2D$ to $3D$ can cause memory to become an issue as a single realisation of a response surface goes from taking around 50 kilobytes to around 2 megabytes for a grid with sides of length $G = 50$. Cross-validation, involving a unique inverse predictive distribution for each datum, thus takes up 7742 times more for the RS10 pollen dataset.

Preliminary results in $3D$ climate space are encouraging, but inference is slow. There are further coding issues, such as how to create a buffer zone in 3 dimensions.

7.2.2 Covariates

The outlier of highest $\overline{D}$ value was found to be from the sampling location of highest altitude. Altitude measurements are in fact available for all samples. Careful modelling of the altitude as a covariate to account for its effect should eliminate this problem. There are several options for dealing with such nuisance covariates; they may be fully modelled as the climate variables are, leading to an increase in the dimensionality of the problem. A more crude treatment, such as fitting a linear smooth of the counts data to the altitudes might suffice.

7.2.3 Inference Procedures

For compatibility with fast approximate inference procedures, the inference-via-marginals model is required. This can lead to errors in the inverse problem if the model does not decompose exactly. Although the imposition of the nested structure greatly reduced these errors for the pollen dataset, the problem did not disappear altogether.

It should be noted that the nesting structure used here is based on the opinion of a single expert. Other nesting structures may well lead to a further reduction or even elimination of the dependence induced error. These structures can either be selected a-priori based on expert opinion, as here, or perhaps inferred from the data. The latter may prove to be a very interesting problem in itself.

Weighting the marginal predictive distributions with the inverse of the correlation between counts is one ad-hoc method to reduce the dependency / correlation of the response surfaces given climate. Another option might be to use the nest as specified, but only down to the lowest binary splittings.

The hyperparameters must be pre-specified at sensible values for the iterative search algorithm in the INLA method to converge to the mode. Trial and error is the current practise; this could be replaced by a crude but fast method based on the model and the data, which would further automate the inference procedure.

7.2.4 Model Validation

Two model validation summary statistics are used in the inverse sense; the number of training data that fall outside the 95% highest density region of the leave-one-out cross-validation posterior predictive distributions Δ and the mean expected squared distance to the observed climate $\overline{D}$.

These statistics still require a scale to determine what values are significant. Monte-Carlo simulation could be used to simulate data from the fitted forward model; the test statistics Δ and $\overline{D}$ would then be calculated on this toy data. These inform what values these statistics take when the model fits the data. Repetition of this exercise should be used to build upper and lower bounds and confidence intervals for the statistics. The values pertaining to the real data would then be compared with these theoretical ranges.

Other cross-validation summary statistics should be developed for the inverse problem. Distance metrics that are commonly applied in forward cross-validation methods may be unsuitable as they frequently assume a unimodal predictive distribution. Inverse predictive distributions are commonly multimodal.

Preliminary work has begun on information criteria such as the Deviance Information Criterion for model comparison. The difficulty in the inverse setting is that the normalising constant for the inverse model is not known. In fact, unlike the forward model (likelihood), it is a function of the model parameters and thus the mean of the deviance is difficult to compute.

Bibliography

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Monographs on statistics and applied probability, Chapman and Hall, London.

Aitchison, J. and Egozcue, J. (2005), 'Compositional data analysis: Where are we and where should we be heading?', *Mathematical Geology* **37**(7), 829–850.

Allen, J., Watts, W. and Huntley, B. (2000), 'Weichselian palynostratigraphy, palaeovegetation and palaeoenvironment: the record from lago grande di monticchio, southern italy', *Quaternary International* **73**(74), 91–110.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, CRC, Boca Raton.

Bartlein, P., Prentice, I. and Webb III, T. (1986), 'Climatic response surfaces from pollen data for some eastern north american taxa', *Journal of Biogeography* **13**, 35–57.

Bellman, R. (1957), *Dynamic Programming*, Princeton University Press, Princeton, NJ.

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

Bhattacharya, S. (2004), Importance Resampling MCMC: a methodology for cross-validation in inverse problems and its applications in model assessment, PhD thesis, University of Dublin, Trinity College, Dept. of Statistics, Trinity College Dublin, Dublin 2, Ireland.

Bhattacharya, S. and Haslett, J. (2008), 'Importance re-sampling mcmc for cross-validation in inverse problems', *Bayesian Analysis* **2**(2), 385–408.

Birks, H. and Birks, H. H. (1980), *Quaternary Palaeoecology*, University Park Press, Baltimore, USA.

Connor, R. and Mosimann, J. (1969), 'Concepts of independence for proportions with a generalization of the dirichlet distribution', *Journal of the American Statistical Association* **64**, 194–206.

Finkenstdt, B., Held, L. and Isham, V. (2006), *Statistical Methods for Spatio-Temporal Systems*, Monographs on statistics and applied probability, Chapman and Hall, London.

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

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

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

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds (1996), *Markov chain Monte Carlo in practice*, Chapman & Hall, London.

Haslett, J., Bhattacharya, S., Michael, M. W., Salter-Townshend, Wilson, S. P., Allen, J., Huntley, B. and Mitchell, F. (2006), 'Bayesian palaeoclimate reconstruction', *Journal of the Royal Statisticcal 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.

Heilbron, D. C. (1994), 'Zero-altered and other regression models for count data with added zeros', *Biometrical Journal* **36**, 531–547.

Holden, P. B., Mackay, A. W. and Simpson, G. L. (2008), 'A bayesian palaeoenvironmental transfer function model for acidified lakes', *Journal of Paleolimnology* **39**(4), 551–566.

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

Kneib, T. (2006), Mixed model based inference in structured additive regression, PhD thesis, LMU München, Faculty of Mathematics, Computer Science and Statistics, LMU Mnchen.

Korhola, A., Vasko, K., Toivonen, H. T. and Olander, H. (2002), 'Holocene temperature changes in northern fennoscandia reconstructed from chironomids using bayesian modelling', *Quaternary Science Reviews* **21**(16–17), 1841–1860.

Krutchkoff, R. (1967), 'Classical and inverse regression methods of calibration', *Technometrics* **9**, 425–439.

Lambert, D. (1992), 'Zero-inflated poisson regression, with an application to defects in manufacturing', *Technometrics* **34**, 1–14.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. and Possingham, H. P. (2005), 'Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations', *Ecology Letters* **8**, 1235–1246.

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

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *Journal of Econometrics* **33**, 341–365.

Oakley, J. and O'Hagan, A. (2002), 'Bayesian inference for the uncertainty distribution of computer model outputs', *Biometrika* **89**, 769–784.

Prentice, I. C., Bartlein, P. J. and Webb, T. I. (1991), 'Vegetation and climate change in eastern north america since the last glacial maximum', *Ecology* **72**, 2038–2056.

Ridout, M., Demetrio, C. G. B. and Hinde, J. (1998), 'Models for count data with many zeros', *Proceedings of the XIXth International Biometric Conference* **Invited Papers**, 179–192.

Rougier, J. (2008), 'Comment on article by Sansó et al', *Bayesian Analysis* **3**(1), 45–56.

Rue, H. and 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. and Chopin, N. (2008), 'Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations', *Journal of the Royal Statistical Society: Series B* **71**(4), 1–35.

Salter-Townshend, M. and Haslett, J. (2006), 'Zero-inflation of compositional data', *Proceedings of the 21st International Workshop on Statistical Modelling* pp. 448–456.

Smol, J. P., Last, W. M. and Birks, H. J. B. (2001), *Tracking Environmental Change Using Lake Sediments: Terrestrial, Algal, and Siliceous Indicators*, Springer.

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

Tian, G.-L., Wang, K. and Geng, Z. (2003), 'Bayesian computation for contingency tables with incomplete cell counts', *Statistica Sinica* **13**, 189–206.

Toivonen, H. T. T., Mannila, H., Korhola, A. and Olander, H. (2001), 'Applying bayesian statistics to organism-based environmental reconstruction', *Ecological Applications* **11**(2), 618–630.

Vasko, K., Toivonen, H. T. and Korhola, A. (2000), 'A bayesian multinomial gaussian response model for organism-based environmental reconstruction', *Journal of Paleolimnology* **24**, 243–250.

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

Wong, T.-T. (1998), 'Generalized dirichlet distribution in bayesian analysis', *Applied Mathematics and Computation* **97**(2–3), 165–181.