Modelling Zero Inflation of Compositional Data

Michael Salter-Townshend[1] and Prof. John Haslett[1]

[1] Dept. of Statistics, School of Computer Science and Statistics, Trinity College Dublin

Abstract: In their paper on a computationally intensive MCMC method for palaeo-climate reconstruction from multivariate pollen counts, Haslett at al (2006) encountered substantial zero inflation; a database of 7815 14-dimensional counts contained 36% zeroes. This paper introduces a new model for such zero-inflated multivariate count data in a spatial context. Specifically, we show that a spatial Gaussian process can be conveniently used to create a zero-inflated mixture of the Multinomial.

Keywords: Bayesian inference; prior; zero-inflation; overdispersion; compositional data.

1 Introduction

1.1 The Data

The data analysed in our reconstructions of the Holocene palaeoclimate are fossil pollen counts data, pollen surface sample counts data and modern climatic data. The latter two components together provide what are often referred to as the 'training data'. The climatic data consists of measures of three aspects of climate, two of which are temperature related. For further details see Haslett et al (2006).

1.2 Reconstructing the Palaeo-Climate

The basis for palaeoclimate estimation from data such as this is straightforward. Variations in climate drive variations in vegetation, in turn leading to changes in the pollen assemblage accumulating in the sediments. Individual plant taxa have their 'preferred' climates: thus changes in past climate can be estimated from changes in the pollen assemblages.
Reconstructing the palaeoclimate may be considered as analogous to locating the fossil pollen assemblage in climate space due to a set of response surfaces and is motivated by the approach used in Huntley (1993).

1.3 Response Surfaces

The response surface approach is used to model the way that individual plant taxa respond to changes in climate. The interpretation of the response here is the propensity to contribute pollen to the assemblage sample in given climatic conditions. The responses are constrained to sum to one over all taxa and are thus proportions. Climate-induced changes to these proportions reflect changes in the vegetation composition local to the pollen sediment sample. The surfaces are constructed from the training data.
We adopt a non-parametric approach to modelling these surfaces. Climate space is discretized onto a regular grid. The components of the surfaces are thus the unobserved latent proportions of the plant taxa at that point in climate space. The responses at the grid-points are modelled stochastically and the data-points which lie off the grid-points are calculated deterministically from weighted averages of the neighbouring grid-point values. For further details see Haslett et al (2006).

2 Statistical Models and Algorithms

2.1 Simple Bayesian Model for Response Surface Components

The model for the proportions p is obtained via the Bayesian paradigm: posterior $\propto$ likelihood $\times$ prior

$$\pi(p|\text{data}) \propto \pi(\text{data}|p) \times \pi(p)$$

where the data are the counts y and the modern climate measurements. We sample from the posterior probability for the proportions using a Metropolis-Hastings Markov Chain Monte Carlo algorithm.
In the simplest model, the likelihood for the counts given the proportions is Multinomial. We then choose a Gaussian spatial process prior to model the responses as smooth functions in climate space.
In order to specify an appropriate prior on the p values, we model independent Gaussian processes x in d dimensional Real space $\Re^d$(or the positive subset thereof) and use a link function $f(x)$ (see Section 2.2) to transform to the $d+1$ dimensional simplex space ℓ^{d+1}.

$$x \sim \mathbf{N}(\mu, \Sigma) \qquad \text{and} \qquad p = f(x)$$

As the Gaussian processes are taken to be independent, the covariance matrix Σ is diagonal. However, the transformation to the simplex space constrains the proportions to sum to one and thus does introduce some dependence across p. The model is then:

$$\pi(p|\text{data}) \propto \text{Multinomial}(y|n, p) \times \pi(p) \tag{1}$$

where n is the total count at a particular site in climate space and $\pi(p) = \text{Jacobian}(x|p)\pi(x)$, depending on the choice of f.
To model extra-climatic variation (a much richer approach), the Gaussian processes would not be modelled as independent. However this would introduce too many extra parameters into the covariance matrix Σ and would dominate the already burdensome computational overhead.

2.2 Link Functions from Real to Simplex Space

The link function used in Haslett et al (2006) was the linear rescaling of the positive $x \in \Re_+$ values to transform to the simplex space. This method does not necessitate the calculation of a Jacobian, however it does involve constraining all x to be positive. A seemingly more natural class of alternatives is suggested in Aitchison (1986). These are the logistic normal class and the primary example is the additive logistic normal distribution for p:

$$p_i = \frac{e^{x_i}}{1 + \sum_j^d e^{x_j}} \qquad \text{for } i = 1, \ldots, \text{d} \tag{2}$$

and 'fill up value' $p_{d+1} = 1 - \sum_i^d p_i$ where x are unconstrained Gaussian processes in $\Re$ space. This leads to the Jacobian $\text{jac}(x|p) = \prod_i (p_i)^{-1}$

2.3 Zero-Inflation of the Data

However, this model for the compositional random variables puts no probability mass on zero values for elements of p. The data is massively zero-inflated, so such a model will therefore be subject to errors in the estimation of the model parameters. The model must be augmented to account for these extra zeros in a meaningful way.
In order to model these zeros they are categorised into two groups based on the source of the zero count, specifically *structural* and *sampling* zeros (see Ridout et al (1998).) Structural zeros occur when a taxon is absent from a sampling point in climate space: i.e. the plant simply does not grow in that climate. Sampling (or counting) zeros occur when the taxon is present but not observed. i.e. it does grow in that climate but by chance was not present in the particular sample of pollen counts.

2.4 Computational Issues

For computational convenience Haslett et al (2006) adopted the Dirichlet Multinomial (or Compound Multinomial) distribution (Dey and Maiti (2002)), often used to model overdispersed multivariate count data. This model arises naturally in multi-level models as a Dirichlet mixture for the main compositional parameter p of the Multinomial. In that paper the main Dirichlet parameter itself was treated as spatially varying, modelled

via a Gaussian spatial process. Other models for spatially varying compositional data (e.g. Tjelmeland and Lund (2003)) are also based on Gaussian processes. The logistic normal distribution of Aitchison (1986) provides a natural alternative to the Dirichlet.
However, such models for compositional random variables put no probability mass on zero values for elements of p. They provide a crude model for massively zero-inflated data such as this. More focused alternatives (such as Lee et al (2006)) use latent binary variables for the mixture and are natural, but dramatically increase the computational load as the distributions are not conjugate; in the context of Haslett et al (2006) a further 7815 x 14 binary random variables are introduced. Marginalising over these dominates the computation.

2.5 Alternative Models

Alternatives for modelling these zeros without resorting to many additional latent variables were therefore explored.
Three new models for zero-inflated data without the use of binary latent variables were built and an intercomparison between the three and the binary latent variables approach to mixing was performed. The first two are close cousins and involve an augmentation of the prior for the $x \in \Re$ values and the third is similar in spirit to a detection limit approach, involving the addition of a single extra 'threshold' parameter.

1. A discontinuous spike at zero is added to the prior for x. The prior is now of 'spike and slab' form, with the slab of Gaussian form and positive only for positive x. Negative values of x are impossible. The linear link function is used to construct p as per Haslett et al (2006). The distribution is now semi-continuous.

2. Negative x are allowed, but the linear link function is changed so that only positive values of x are used to construct p. Negative x imply that p is zero. This is equivalent to an absence. i.e. the taxon has no propensity to produce pollen at this climate. The link function is now $p_i = I(x \in \Re_+)\frac{x_i}{\sum_i (x_i)}$ The interpretation is that there is an underlying process governing vegetation response to climate that can be negative but we only observe positive counts. Negative x are then being sampled from the prior as the likelihood is flat for these values. The x values are modelled rather than the ps.

3. Detection limit approach. The additive logistic normal link function is used and one one extra variable is added. This represents a threshold value that the response by the plant taxon must exceed before it contributes pollen to the assemblage. The link function now becomes:

$$p_i = (1+a)\frac{e^{x_i}}{1+\sum_j^d e^{x_j}} - a \qquad \text{for } i = 1, \ldots, \text{d}$$

TABLE 1. Toy Problem fit Results: Model Comparison

MODEL	Error	Computation Time
Spike and slab prior on $x \in \Re+$	3.74e-4	21.14
Unconstrained $x \in \Re$ modelled	3.96e-4	22.35
Detection Limit	FLAG	FLAG
Binary Latent Variables	3.73e-4	83.52

3 Investigation via a Toy Problem

A toy problem was constructed in order to test if any of the three new models could compete with the accepted binary in terms of accuracy and to ensure that convergence and running time were superior. The advantage of using a toy problem is that the values of p used to generate the toy data are known and so true errors may be computed.

3.1 Generation of Toy Data

In the toy problem, a one dimensional 'climate' space was used. Random Gaussian response curves for two taxa were created with negative values mapped to zero (absence / no response). Counts from a binomial distribution were then generated from these curves. 50 datapoints and 50 equally spaced grid-points were used. The total count at any location was taken to be 400. An example reconstruction of the response curve used to generate the data using the unconstrained x model is shown in Figure 1.

4 Results

A summary of results for the toy problem is presented in Table 1. The measure of error is an average sum of squared distances to the true underlying response curve for 1000 realisations from the posterior for each model, collected with a separation of 100 iterations following a burn in of 200,000 iterations. The computation time is in seconds.
It is important to note that the models only differ from each other at the points where one taxon is contributing all the counts and the other is absent. A comparison at these locations in climate space of the trace plots for the MCMC output is shown in Figure 2.

5 Conclusions

The model with a discrete spike at zero in the prior for the proportions is slightly faster to implement than the uncontrained modelling of the x

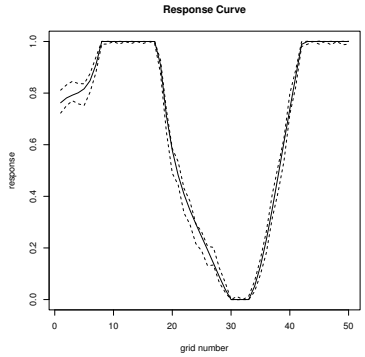

FIGURE 1. Example response curve with fitted 95% HPD regions for the unconstrained x model.

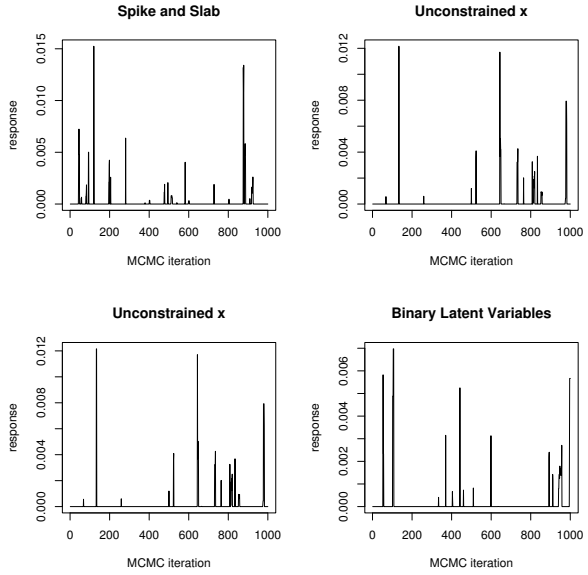

FIGURE 2. Example trace plots for the 4 models for a taxon absent in this region.

processes and gives a more accurate result. The binary latent variables model is the slowest but gives the most accurate result. However, the results contain just one measure of error.

Perhaps the best way to gain insight into the differences between the models is to examine the MCMC output directly. The trace plots for the binary latent variables model for zero response are the most accurate with the fewest and smallest excursions from zero.
Although the implementation of Metropolis-Hastings updates for the semi-continuous prior in the spike and slab type model is straightforward, the MCMC theory is undeveloped. Other future work includes applying all models to the real dataset.

Acknowledgments: Acknowledgements Michael Salter-Townshend and Prof. John Haslett have been supported variously by Enterprise Ireland grant SC/2001/171 and SFI grant 04/BR/M0049 The concept of modelling the x processes was suggested by Tony O'Hagan.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. *Monographs on Statistics and Applied Probability* Chapman and Hall Ltd.

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

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

Lambert, D. (1992). Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics* **34**, 1-14.

Lee AH, Wang K, Scott JA, Yau, KKW, McLachlan, GJ. (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros *Statistical Methods in Medical Research* **15 (1)**, 47-61.

Ridout, M., C. G. B. Demetrio, J. Hinde (1998). Models for Count Data with Many Zeros *International Biometric Conference*, Cape Town.

Tjelmeland H, Lund KV (2003). Bayesian modelling of spatial compositional data *Journal of Applied Statistics*, **30 (1)**, 87-100.