The Variational Bayes Method For Inverse Regression Problems With an Application To The Palaeoclimate Reconstruction

Richa Vatsa and Simon Wilson
Dept. of Statistics, Trinity College Dublin, Dublin, Ireland
vatsar@tcd.ie and simon.wilson@tcd.ie

Abstract

The palaeoclimate reconstruction problem is described as an example of inverse regression problems. In the reconstruction problem, past climate is inferred using pollen data. Modern data is used to build a regression model of how pollen responds to climate. The inverse problem is to infer climate from data on ancient pollen prevalence. The inverse inference presents a challenging and computationally intensive problem. It is demonstrated that Variational Bayes (VB), that assumes conditional independence, provides quick solutions to the reconstruction problem. The advantage of the use of the VB method is that many more climate variables can be included in the estimation without imposing a huge burden to the reconstruction problem. We explore the accuracy of the VB method, and comment on its usefulness more generally in inverse inference problems.

Keywords: Inverse problem, palaeoclimate reconstruction, Variational Bayes method.

1 Introduction

Inverse problems form an important class of statistical inference problems, from geology to medical image processing and financial mathematics. The definition of an inverse problem is subject to some interpretation but broadly it refers to problems where we have indirect observations of an object (a function) that we want to reconstruct. From a mathematical point of view, this usually corresponds to the inversion of some operator.

A particular case of an inverse problem, and the motivating example for this paper, is that of ancient climate reconstruction from so-called 'proxy' data, such as ancient pollen that can be recovered from lake sediment. In this case, it is natural to model the response of pollen as a function of climate, and then fit this model through modern data on both pollen and climate. The inference task is to invert the fitted function using ancient pollen data in order to infer the climate.

In common with many inverse problems, both the model fitting and inversion involve a considerable computational burden. Bayesian approaches to this problem have used MCMC (Haslett et al., 2006), importance sampling for cross validation (Bhattacharya, 2004) and functional approximations (Salter-Townshend, 2009). All of these approaches are not without limitations. The MCMC approach suffered from poor mixing. The approach of Salter-Townshend (2009) used the INLA method of Rue. et al. (2009), and so was restricted to models with a limited number of parameters. They also split the problem into two distinct stages: fitting to modern data, and inversion.

In this paper we describe an alternative approach to implementing Bayesian inference for this inverse problem. This makes use of the variational Bayes (VB) approximation. We believe that it avoids the limitations of previous methods, and is able to jointly fit the model and infer past climate, although it is not without its own problems. We apply it to the response surface model used in the previous work. We illustrate the approach with several different simulated data sets. There is little existing work on the use of VB in inverse problems. Nakajima and Watanabe (2006) uses VB in in image processing problem. To our knowledge this is the first attempt at implementing VB to an inverse problem outside that application.

1.1 Climate proxies

Ancient pollen is one of several proxies that are used to learn about past climate; others include other fauna and flora, as well as O^{18} isotope levels. Huntley (1991) advocates the use of several different pollen taxa as a good description of climate over the past 15,000 years, since pollen responds smoothly to climate, and different vegetation types respond differently to climate. Figure (1) shows typical data in the relative proportions of pollen of different types that were collected from a column of lake sediment at Glendalough, Ireland. The age of the pollen is determined by radiocarbon dating and clear changes in the composition of pollen is seen over time. In particular, the large change that occurred with the end of the last ice age at about 9,000 years ago is clearly seen.

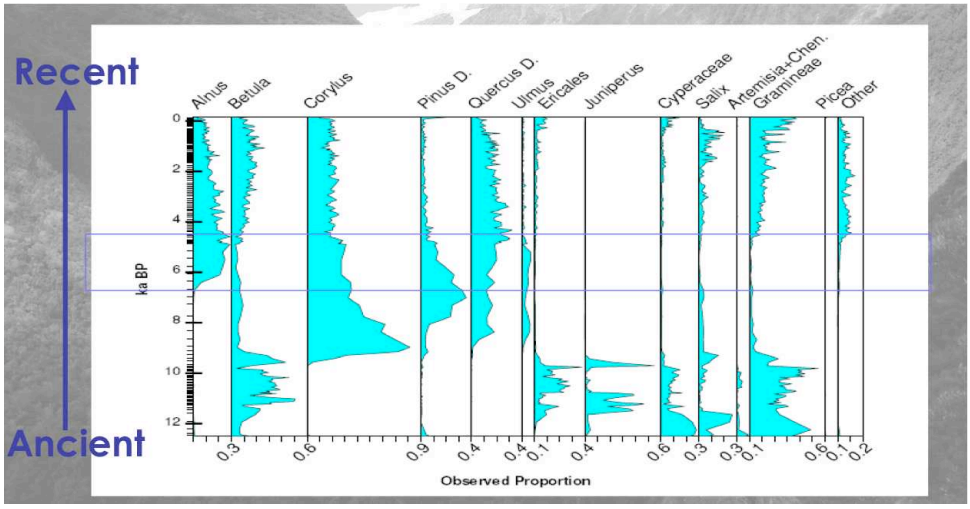

Figure 1: Proportions of each of the 14 different categorizations against radiocarbon years BP (vertical axis) at Glendalough (Haslett et al., 2006).

Huntley et al. (1993) also proposed the use of a response surface model, where each pollen taxon has a latent response that is a function of climate, and observed counts are a noisy function of the response. This is the model used in Haslett et al. (2006) and Salter-Townshend (2009), and we also adopt it.

1.2 Data and Notation

The RS10 data set is taken from Allen et al. (2000) which describes the nature of the palaeoclimate data on pollen and climate. The data set consists of two types of data. One is called the modern data that includes the modern count data on pollen and the modern data on climate variables observed at 7742 locations in the northern hemisphere. There may be several possible climate variables to study the climate behaviour. Haslett et al. (2006) consider mainly two climate variables MTCO (Mean Temperature of the Coldest month) and GDD5 (Growing Degree Days above $5°C$). Second is the ancient data, which has count data on fossil pollen collected from 150 ancient sites (lake sediments). We aim to infer the missing ancient climate for the fossil data. Following are some notations on the modern and the ancient data;

$$
\begin{aligned}
P &: \text{Total number of climate locations in the palaeoclimate study,} \\
T &: \text{Total number of taxa considered for the study of the ancient climate,} \\
K &: \text{Total number of climate variables in the study,} \\
N^m &: \text{total number of samples in modern data,} \\
N^f &: \text{total number of samples in ancient data,} \\
\mathbf{y}^m &= y^m_{tj};\ t = 1, \ldots, T;\ j = 1, \ldots, N^m;\ \text{modern counts on taxa,} \\
\mathbf{C}^m &= C^m_{kj};\ k = 1, \ldots, K;\ j = 1, \ldots, N^m;\ \text{modern data on climate variables} \\
\mathbf{X} &= \mathbf{X}_{ki}; k = 1, \ldots, K; i = 1, \ldots, P;\ \text{response surfaces of taxa.}
\end{aligned}
$$

Superscript m and f stand for modern and fossil respectively. Notations without any superscript or subscript stand for both the types of variables, modern and ancient.

Unlike that in Figure (1), the reported data on different taxa at various location in climate space are typically in counts. Due to huge variation in the tolerance limit of the taxa to different climate at different location, the count data on taxa are highly over-dispersed with an excess of zero counts in them. The zero-inflated behaviour of the count data can be understood with diagrams shown in Figure (2). The diagram represents the histograms of the count data on the taxa, Alnus and Corylus which shows that the most of the data points are zero.

1.3 Paper layout

In Section 2, we describe the model used for palaeoclimate reconstruction. Section 3 outlines the Bayesian inference approach, and implementation of the inference by variational Bayes is described in Section 4. Section 5 illustrates the approach with some examples.

2 Model description

For the demonstration of the Variational Bayes approximation to the palaeoclimate reconstruction problem, two types of the models are discussed in this paper;

1. one taxon and one climate model,
2. more than one taxon and one climate model.

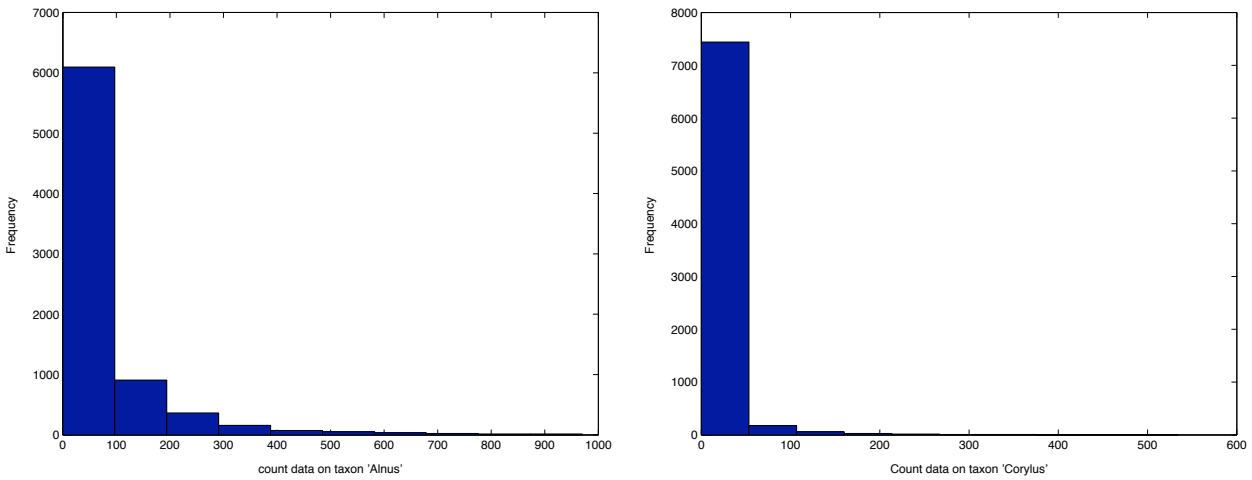

Figure 2: Histogram of the count data on the taxon (left) Alnus and (right) Corylus.

The possible choices of the likelihoods of data on taxa and the prior distributions over the unknown parameters (ancient climate, response surface and hyper-parameters) are discussed for both the models.

2.1 The one taxon and one climate model

To explain the VB inference for the palaeoclimate reconstruction problem at a basic level, we first discuss the one taxon and one climate model.

1. **Likelihood:** It is assumed that the data on taxon are independent across the climate locations given the latent response surface. The ignorance of the dependence in data across the locations makes the inference problem simpler. It also favors the independent assumption of the VB method. Three standard distributions (likelihoods) are discussed for modelling data on taxa.

   (a) **Gaussian Likelihood:** For a simple understanding of the inference problem, data on taxa are first modelled with a Gaussian distribution. The VB approximation over the response surface is also Gaussian for a Gaussian prior distribution.

   $$L(\mathbf{y}|\mathbf{X},\theta) = \prod_{j=1}^{N} N(\mathrm{y}_j; \mathbf{X}(\mathrm{C}_j), r^2). \tag{1}$$

   As described earlier, data on taxon $\mathbf{y}$ are the function of the climate $\mathbf{C}$ through the latent responses $\mathbf{X}$. At any climate location C_j, the mean of the Gaussian density of a data point y_j is equal to the response surface, $\mathbf{X}(\mathrm{C}_j)$, a function of the climate at the same location;

   $$\mathbb{E}(\mathrm{y}_j) = \mathbf{X}(\mathrm{C}_j). \tag{2}$$

   The parameter r^2 is the unknown variance in the likelihood.

(b) **Poisson Likelihood:** A Gaussian density is not suitable for modelling count data on taxa. A Poisson distribution is one of the simplest choice among the discrete distributions to model count data.

$$L(\mathbf{y}|\mathbf{X},\theta) = \prod_{j=1}^{N} Poiss(\mathrm{y}_j; \exp(\mathbf{X}(\mathrm{C}_j))). \tag{3}$$

In the Poisson likelihood $L(\mathbf{y}|\mathbf{X},\theta)$, the mean is equal to $\exp(\mathbf{X}(\mathbf{C}))$. Since the responses of the taxa to the climate can be both positive and negative, allowing the latent response surface raging from a negative value to a positive value in its space. The mean of the data point $\mathbf{y}_j$ at C_j^{th} climate location is an exponential function of the response surface $\mathbf{X}$ at the same location.

$$\mathbb{E}(\mathrm{y}_j) = \exp(\mathbf{X}(\mathrm{C}_j)). \tag{4}$$

(c) **Zero-inflated Poisson Likelihood:** The real data on taxa exhibit their zero-inflated behaviour (as shown in Figure 2). If zero-inflated data are modelled with a Poisson distribution, a non-zero-inflated distribution, the mean of the density would be underestimated and so the variance. The excess zeros in the data cannot even be ignored as they have potential information about the corresponding.

Ridout et al. (1998) discuss a zero-inflated Poisson distribution for modelling counts with excess zeros;

$$L(\mathbf{y}|\mathbf{X},\theta) = \prod_{j=1}^{N} ZIP(\mathrm{y}_j; \exp(\mathbf{X}(\mathrm{C}_j))), \tag{5}$$

where $ZIP(\mathrm{y}_j; \exp(\mathbf{X}(\mathrm{C}_j))) = \begin{cases} 1 - q_j + q_j \ \exp(-e^{\mathbf{X}(\mathrm{C}_j)}), & \text{if } \mathrm{y}_j = 0; \\ q_j Poiss(\mathrm{y}_j; \exp(\mathbf{X}(\mathrm{C}_j))), & \text{if } \mathrm{y}_j > 0 \end{cases}$

where, $(1 - q_j);\ j = 1 : N$ is the probability of observing essential zero counts. Salter-Townshend (2009) uses a power law functional relationship to define the probability q_j as

$$q_j = \left(\frac{\lambda_j}{1+\lambda_j}\right)^{\gamma}, \tag{6}$$

where, $\lambda_j = \exp(\mathbf{X}(\mathrm{C}_j))$ is the rate parameter in the likelihood. The power index γ takes values from 0 to ∞ such that a big value of γ induces many zero counts in data.

For a tractable VB approximation, we shall assume the probabilities of zeros inflation $(1 - q_j);\ j = 1 : N$, as known. The intractability issue of VB method with the unknown zeros inflation probabilities is explained in the Appendix section.

2. **Prior Distributions:** The two common unknowns to be inferred are the responses $\mathbf{X}$ and the ancient climate C^f. The climate space is assumed to be discrete and so climate values are taken on equally spaced discrete grid points. The discrete climate grid values are treated as the indexes for the response surfaces.

   (a) **Prior distribution over ancient climate C^f:** It is assumed that no particular information is available on the ancient climate to translate it into an informative prior

density. A non-informative prior is taken over the ancient climate assumed over discrete grid points;

$$P(\mathrm{C}^f) \propto 1. \tag{7}$$

(b) **Prior distribution over response surface X:** An intrinsic Gaussian Markov Random Field (IGMRF) by Rue and Held (2005) is assumed as a joint prior distribution over responses **X** across all the location in the climate space;

$$P(X|Q) = \mathrm{IGMRF}(Q). \tag{8}$$

IGMRF is an improper distribution with a low ranked precision matrix. The mean vector in the distribution is not specified. The markovian property of the GMRF allows the precision matrix Q to be sparse.

$$Q = \kappa_1 R_{P\times P}, \tag{9}$$

where, κ is an unknown smoothing parameter. The matrix R is a tri-diagonal matrix the first order random walk behaviour of the responses.

In the case of a single climate variable, the IGMRF is specified as a Gaussian random walk indexed by climate:

$$P(\mathbf{X}) \approx \prod_{i=2}^{P} N_{\mathrm{X}_i}(\mathrm{X}_{i-1}, \kappa^{-1}). \tag{10}$$

Generally, a multivariate Gaussian prior distribution is assumed as a prior distribution over a spatial variable. But often, the dense form of the covariance function in the distribution makes the computation slow. The advantage of a (intrinsic) GMRF over a multivariate Gaussian distribution is that it reduces the computation time, because the precision matrix in (intrinsic) GMRF is sparse.

(c) **Prior distribution over other unknown parameters:** The prior distribution assumed over a smoothing parameter κ in an IGMRF prior, is a conjugate gamma prior with known hyper-parameters;

$$\kappa \sim \mathrm{Gamma}(\kappa; a, b). \tag{11}$$

In the Gaussian likelihood, the variance parameter r^2 is assumed to be unknown. An inverse Gamma prior is considered over r^2.

$$r^2 \sim \mathrm{Inv\ Gamma}(r^2; \alpha, \beta). \tag{12}$$

2.2 The more than one taxa and one climate model

As the palaeoclimate reconstruction is a multi-proxy problem, many taxa should be added to the model. We assume two taxa and one climate model. For the computational ease, the responses of each taxon are also considered as independent of each other. We add some unknown random effects to the latent responses in the model to induce dependence between taxa. For each data point (in modern and ancient ancient data set) on taxa, there is assumed a bivariate random effects (number

of dimensions equal to the number of taxa). Random effects also capture the over-dispersion in data. They are assumed to be independent of the climate in the model.

$$\begin{aligned} \mathbf{U} &= \{\mathbf{U}_1, \mathbf{U}_2\}, \\ \mathbf{U}_k &= ((\mathrm{U}_{kj}));\ k = 1, 2;\ j = 1 : N. \end{aligned}$$

1. **Likelihood:** The Poisson likelihoods are considered to model count data on taxa. The taxa are assumed to be independent across climate location given the responses and are allowed to be independent of each other given the random effects.

$$L(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{X}_1, \mathbf{X}_2, \mathrm{C}, \mathrm{U}_1, \mathrm{U}_2) = \prod_{k=1}^{2} \prod_{j=1}^{N} Poiss(\mathrm{y}_{kj}; \exp(\mathbf{X}_k(\mathrm{C}_j) + \mathrm{U}_{kj})), \tag{13}$$

   where, y_{kj} denotes a data point on k^{th} taxon at C_j^{th} climate location, $\mathbf{X}_k$ is the corresponding unknown responses and $U_j = (U_{1j}, U_{2j})$ is a common latent response at climate C_j that allows for dependence between taxa given climate, as described below.

2. **Prior distributions:** Independent IGMRF priors as described in the previous section are assumed over the responses $\mathbf{X}_1$ and $\mathbf{X}_2$ with precision matrix Q_1 and Q_2 and unknown smoothing parameters κ_1 and κ_2. Conjugate Gamma priors are assumed over κ_1 and κ_2 with known hyper-parameters $a_k, b_k;\ k = 1, 2$. The same non-informative prior distribution (as discussed for the one taxa and one climate) is assumed over the ancient climate C^f.

   A bivariate Gaussian distribution with zero mean and unknown covariance is assumed as an independent prior distribution over the random effects $\mathbf{U}$ for each of the data points on the taxa.

$$P(\mathbf{U}|Q_{\mathbf{U}}) \equiv BVN_{\mathbf{U}} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, Q_{\mathbf{U}}^{-1} \right). \tag{14}$$

   A Wishart distribution is considered as a prior distribution over the unknown precision $Q_{\mathbf{U}}^{-1}$;

$$Q_{\mathbf{U}} \sim \text{Wishart}(df, SS), \tag{15}$$

   where df is the degree of freedom and SS is a symmetric positive definite matrix. The dimension of the matrix SS is same as the one of the precision matrix $Q_{\mathbf{U}}$.

3 Bayesian Inference

The aim of the reconstruction problem is to first, model the response surface and second, to infer the ancient climate. The response surfaces are modelled through their indirect observation (the count data) on the pollen across all the climate location considered for the study of the ancient climate. Then, the unknown ancient climate is inferred through the knowledge obtained from the modelling of the response surface. In the Bayesian paradigm, the inference on an unknown variable is obtained through its posterior distribution.

In the Bayesian approach to the palaeoclimate reconstruction problem, the joint posterior distribution of the response surfaces $\mathbf{X}$ the ancient climate $\mathbf{C}^f$ and other unknown parameters in the

model can computed by Bayes' law, as

$$P(\mathbf{X}, \mathbf{C}^f, \theta|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f,) = \frac{L(\mathbf{y}^m|\mathbf{C}^m, \mathbf{X}, \theta)L(\mathbf{y}^f|\mathbf{C}^f, \mathbf{X}, \theta)P(\mathbf{X}|\mathbf{C}^m, \mathbf{C}^f, \theta)P(\mathbf{C}^f)P(\theta)}{\int_{\mathbf{X},\mathbf{C}^f} L(\mathbf{y}^m|\mathbf{C}^m, \mathbf{X}, \theta)L(\mathbf{y}^f|\mathbf{C}^f, \mathbf{X}, \theta)P(\mathbf{X}|\mathbf{C}^m, \mathbf{C}^f, \theta)P(\mathbf{C}^f)P(\theta)d\mathbf{X}d\mathbf{C}^f}. \tag{16}$$

All the unknown parameters (and hyper-parameters) other than $\mathbf{X}$ and $\mathbf{C}^f$ are denoted by θ. In the above expression for the joint posterior distribution $P(\mathbf{X}, \mathbf{C}^f, \theta|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f)$ given $\mathbf{y}^m$ and $\mathbf{C}^m$, the term $L(\mathbf{y}^m|\mathbf{C}^m, \mathbf{X}, \theta)$ is the likelihood of $\mathbf{y}^m$ given $\mathbf{C}^m$ and $\mathbf{X}$, whereas $L(\mathbf{y}^f|\mathbf{C}^f, \mathbf{X}, \theta)$ is the likelihood of $\mathbf{y}^f$. A prior density over ancient climate $\mathbf{C}^f$ is denoted by $P(\mathbf{C}^f)$. The term $P(\mathbf{X}|\mathbf{C}^m, \mathbf{C}^f, \theta)$ is a joint prior over responses $\mathbf{X}$ and $P(\theta)$ is a prior density over the parameters θ.

The expression for the marginal posterior distribution of $\mathbf{X}$ after integrating out $\mathbf{C}^f$ and θ, is given as

$$P(\mathbf{X}|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f) = \int_{\mathbf{C}^f,\theta} P(\mathbf{X}, \mathbf{C}^f|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f)P(\theta)d\mathbf{C}^f d\theta, \tag{17}$$

Similarly, the marginal posterior distribution of $\mathbf{C}^f$ is computed as

$$P(\mathbf{C}^f|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f) = \int_{\mathbf{X},\theta} P(\mathbf{X}, \mathbf{C}^f|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f)P(\theta)d\mathbf{X}d\theta. \tag{18}$$

The marginal inference on $\mathbf{C}^f$ shown by above expression, is equivalent to the inverse inference on $\mathbf{C}^f$ discussed in Section 1. An inference on $\mathbf{C}^f$ requires the knowledge of $\mathbf{X}$ obtained from the marginal posterior distribution $P(\mathbf{X}|\mathbf{y}^m, \mathbf{C}^m, \mathbf{y}^f)$. Both the ancient climate $\mathbf{C}^f$ and the responses $\mathbf{X}$ depend on each other. The response surfaces $\mathbf{X}$ are high dimensional matrices (dimension equal to the total number of the climate locations in the study). For each taxon, there is a response surface across all the climate locations. The evaluation of the integrals over multi-dimensional responses surface makes the evaluation of the posterior distributions (joint and marginal) computationally intensive and intractable.

The goal is to find a tractable approximation to Equation (16). In this paper the Bayes (VB) method is used for a tractable approximation of the posterior distributions. A brief introduction to the VB method is presented in the next section.

4 The Variational Bayes method

4.1 A Brief Introduction

The Variational Bayes (VB) is a functional approximation for a posterior distribution that has an advantage of fast computation. In the VB method, we try to find a distribution $q_\theta(\theta)$ as an approximation to the intractable density $P(\theta|y)$ by minimizing a Kullback-Leibler divergence $KL(q_\theta(\theta) \parallel P(\theta|\mathbf{y}))$:

$$KL(q_\theta(\theta) \parallel P(\theta|\mathbf{y})) = \int_\theta q_\theta(\theta) \log \frac{P(\theta|\mathbf{y})}{q_\theta(\theta)} d\theta. \tag{19}$$

Kullback-Leibler (KL) divergence is a measure of discrepancy from one distribution to another. It is always a non-negative quantity. Šmídl and Quinn (2006) have discussed the KL divergence and its properties. Detailed discussion on KL divergence can be found in Kullback (1997).

In practice, it is not possible to find a minimum of $KL(q_\theta(\theta) \parallel P(\theta|\mathbf{y}))$ with respect to $q_\theta(\theta)$, ie;

$$q_\theta(\theta) = \arg\min_{q_\theta(\theta)} KL(q_\theta(\theta) \parallel P(\theta|\mathbf{y})), \tag{20}$$

as the posterior distribution $P(\theta|\mathbf{y})$ is known only up to a proportionally constant. But, a solution to the minimum of $KL(q(\theta) \parallel P(\theta|\mathbf{y}))$ may be found assuming the (posterior) independence between the components of the parameter θ in $q_\theta(\theta)$ i.e.

$$q_\theta(\theta) = \prod_{i=1}^{P} q_{\theta_i}(\theta_i) \tag{21}$$

where $q_{\theta_i}(\theta_i)$ is a marginal approximation of posterior distribution over the i^{th} component of θ, θ_i, also called the VB marginal (of θ_i). With the assumption of posterior independence in θ, a minimum solution of $KL(q(\theta) \parallel P(\theta|\mathbf{y}))$ with respect to $q_{\theta_i}(\theta_i)$ as;

$$q(\theta_i) \propto \exp[\mathbb{E}_{q(\theta_{-i})}[\log P(\mathrm{y}, \theta)]], \qquad i = 1, \cdots, d, \tag{22}$$

(see Theorem 3.3.1 of Šmídl and Quinn (2006)). In the above expression, the VB approximation over the components (other than θ_i) of θ, $q_{\theta_{-i}}(\theta_{-i}) = \prod_{\substack{j=1 \\ j\neq i}}^{d} q_{\theta_j}(\theta_j)$ is kept fixed. The term $q(\theta_i)$ is called the VB-marginal of θ_i.

The assumption of posterior independence allows the expectation in the R.H.S of Equation (22) being expressed as a tractable product of $(d-1)$ 1-dimensional expectation over the log-joint density $\log P(\mathrm{y}, \theta)$, with a condition that the log-joint density should factorize over the components of θ. For example, it is common to have factorized log-joint density over the components of the parameter for the exponential family of distributions.

It can be seen from Equation (22) that the VB marginal $q_{\theta_i}(\theta_i)$ is a function of the moments of $\theta_j,\ j \neq i$, with respect to $q_{\theta_{-i}}(\theta_{-i})$, which leads to the fact the VB-marginals interact with each other via their moments, or in other words, the moments of one VB-marginal is a function of the moments of other VB-marginals. The interdependence of the VB marginals present an obstacle in their evaluation. This requires an iterative (VB) algorithm for a local minima of $KL(q(\theta) \parallel P(\theta|\mathbf{y}))$ (Šmídl and Quinn, 2006).

4.2 Implementation of the method

In this section, the VB approximation to the reconstruction problem is explained. As described in Section 3, the aim of the problem is to find a tractable solution to the marginal posterior densities over latent response surface $\mathbf{X}$ and the unknown ancient climate $\mathbf{C}^f$. The VB approximation for the joint posterior distribution $P(\mathbf{X}, \mathbf{C}^f, \theta)$ is given as

$$\begin{aligned} P(\mathbf{X}, \mathbf{C}^f, \theta) &\approx q(\mathbf{X}, \mathbf{C}^f, \mathbf{U}\theta), & (23)\\ &= q_{\mathbf{X}}(\mathbf{X}) q_{\mathbf{C}^f}(\mathbf{C}^f) q_{\mathbf{U}}(\mathbf{U}) q_\theta(\theta). & (24) \end{aligned}$$

The VB marginal $q_\theta(\theta)$ may be further decomposed into the VB marginals over the components of θ.

From Equation 22, the VB marginals $q_{\mathbf{X}}(\mathbf{X}), q_{\mathbf{C}^f}(\mathbf{C}^f)$ and $q_\theta(\theta)$ are expressed as

$$\begin{aligned} q_{\mathbf{X}}(\mathbf{X}) &\propto \exp[\mathbb{E}_{q_\theta(\theta)q_{\mathbf{C}^f}(\mathbf{C}^f)q_{\mathbf{U}}(\mathbf{U})}[\log P(\mathbf{y}^m, \mathbf{y}^f, \mathbf{C}^f, \mathbf{X}, \mathbf{U}, \theta|\mathbf{C}^m)]], \quad (25)\\ q_{\mathbf{C}^f}(\mathbf{C}^f) &\propto \exp[\mathbb{E}_{q_\theta(\theta)q_{\mathbf{X}}(\mathbf{X})q_{\mathbf{U}}(\mathbf{U})}[\log P(\mathbf{y}^m, \mathbf{y}^f, \mathbf{C}^f, \mathbf{X}, \mathbf{U}, \theta|\mathbf{C}^m)]], \quad (26)\\ q_{\mathbf{U}}(\mathbf{U}) &\propto \exp[\mathbb{E}_{q_\theta(\theta)q_{\mathbf{C}^f}(\mathbf{C}^f)q_{\mathbf{X}}(\mathbf{X})}[\log P(\mathbf{y}^m, \mathbf{y}^f, \mathbf{C}^f, \mathbf{X}, \mathbf{U}, \theta|\mathbf{C}^m)]], \quad (27)\\ \text{and, } q_\theta(\theta) &\propto \exp[\mathbb{E}_{q_{\mathbf{X}}(\mathbf{X})q_{\mathbf{C}^f}(\mathbf{C}^f)q_{\mathbf{U}}(\mathbf{U})}[\log P(\mathbf{y}^m, \mathbf{y}^f, \mathbf{C}^f, \mathbf{X}, \mathbf{U}, \theta|\mathbf{C}^m)]], \quad (28) \end{aligned}$$

where, the joint likelihood

$$P(\mathbf{y}^m, \mathbf{y}^f, \mathbf{C}^f, \mathbf{X}, \mathbf{U}, \theta|\mathbf{C}^m) = L(\mathbf{y}^m|\mathbf{C}^m, \mathbf{X}, \mathbf{U}, \theta)L(\mathbf{y}^f|\mathbf{C}^f, \mathbf{X}, \mathbf{U}, \theta)P(\mathbf{X}|\mathbf{C}^m, \mathbf{C}^f, \theta)P(\mathbf{C}^f)P(\mathbf{U}|\theta)P(\theta). \quad (29)$$

The prior distributions over $\mathbf{X}$ and $\mathbf{U}$ are not conjugate to the likelihoods of $\mathbf{Y}$. Hence, the VB marginal $q_{\mathbf{X}}(\mathbf{X})$ and $q_{\mathbf{U}}(\mathbf{U})$ may not be recognized as the standard distributions.

A numerical integration method is used to compute the normalizing constant and the required moments of the intractable VB marginals. However, the method will be too time consuming for intractable multi-dimensional VB marginals. Rue and Held (2005) describes a quick way of finding Gaussian approximations to the intractable posterior distributions with GMRF or Gaussian priors. For a Gaussian approximation to an intractable posterior distribution, a quadratic approximation to the log-joint likelihood is considered. If the corresponding prior density is already a quadratic function (Gaussian or GMRF), we only need to find a quadratic approximation to the log-likelihood.

We use the same idea of Rue and Held (2005) for Gaussian approximation of intractable VB marginals $q_{\mathbf{X}}(\mathbf{X})$ and $q_{\mathbf{U}}(\mathbf{U})$. The only difference is that we have VB-expectation of the log-joint likelihood with respect to the VB-marginals unknown parameters (presented in the Equation 28), which can further be decomposed in the log of prior term and the log of the likelihood terms depending on the VB moments of the other parameters.

4.3 VB solution to the one taxon and one climate model

VB solutions for the one taxon and one climate model with the likelihoods (discussed in Section 2.1) is presented. We present the VB marginals with their functional forms only. The functions are defined in detail in the Appendix section.

4.3.1 VB solution for the model with the Gaussian Likelihood:

The VB marginals over the responses surface $\mathbf{X}$, the ancient climate C^f and the unknown variance r^2 are given as follows:

1. VB marginal over $\mathbf{X}$ obtained, is a Gaussian density with mean $\mu_{\mathbf{X}}$ and variance $\sigma_{\mathbf{X}}$ and is expressed as

$$\begin{aligned} q_{\mathbf{X}}(\mathbf{X}) &\equiv N_{\mathbf{X}}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}}), \quad (30)\\ \Sigma_{\mathbf{X}} &= Q_{\mathbf{X}}^{\star\,-1}, \quad (31)\\ \mu_{\mathbf{X}} &= Q_{\mathbf{X}}^{\star} B_{\mathbf{X}}, \quad (32) \end{aligned}$$

where $B_{\mathbf{X}}$ is a function of likelihood terms involving y, data on pollen and the VB-moments of r^2 and C^f. The structure of the posterior (VB) precision matrix $Q^{\star}_{\mathbf{X}}$ is given as

$$\begin{aligned} Q^{\star}_{\mathbf{X}} &= \mathbb{E}_q(Q_{\mathbf{X}}) + \mathrm{diag}(V_{\mathbf{X}}), & (33)\\ &= \mathbb{E}_q(\kappa)R + \mathrm{diag}(V_{\mathbf{X}}), & (34) \end{aligned}$$

where, $Q_{\mathbf{X}}$ is the prior precision. The term $V_{\mathbf{X}}$ is the precision obtained from the likelihood and is a function of the data on pollen y and the VB-moments of r^2 and C^f.

2. The VB marginal over C^f, $q_{\mathrm{C}^f}(\mathrm{C}^f)$ is not a standard distribution. As the ancient climate C^f is univariate, a numerical approximation method can be applied to compute the normalizing constant of un-normalized $q_{\mathrm{C}^f}(\mathrm{C}^f)$.

3. The VB marginal over the variance parameter r^2 is an inverse Gamma density.

$$q_{r^2}(r^2) = \text{Inverse Gamma}(r^2; \alpha^{\star}, \beta^{\star}), \quad (35)$$

where, the hyper-parameters $\alpha^{\star}$ and $\beta^{\star}$ are functions of the data y and the VB-moments of $\mathbf{X}$ and C^f.

4. The VB marginal κ, $q_{\kappa}(\kappa)$ is a Gamma density.

$$q_{\kappa}(\kappa) \equiv \text{Gamma}(\kappa; a^{\star}, b^{\star}). \quad (36)$$

The hyper-parameters $a^{\star}$ and $b^{\star}$ in the VB marginal are functions of the VB moments of $\mathbf{X}$ and C^f. The smoothing parameter κ is independent of the likelihood of the data y. It appears only in the prior density of $\mathbf{X}$. Therefore, the VB hyper-parameters do not involve any term related to the data y.

4.3.2 VB solution for the models with the Poisson and the zero-inflated Poisson Likelihood:

With IGMRF prior and a non-Gaussian likelihood (Poisson and zero-inflated Poisson Likelihoods), the VB marginal over the response surface $\mathbf{X}$, $q_{\mathbf{X}}(\mathbf{X})$ is not a tractable density. We apply a Gaussian approximation for a tractable approximation to $q_{\mathbf{X}}(\mathbf{X})$.

$$\begin{aligned} q^{\star}_{\mathbf{X}}(\mathbf{X}) &\equiv N_{\mathbf{X}}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}}), & (37)\\ \Sigma_{\mathbf{X}} &= Q^{\star\,-1}_{\mathbf{X}}, & (38)\\ \mu_{\mathbf{X}} &= Q^{\star}_{\mathbf{X}} B_{\mathbf{X}}, & (39)\\ Q^{\star}_{\mathbf{X}} &= \mathbb{E}_q(Q_{\mathbf{X}}) + \mathrm{diag}(V_{\mathbf{X}}), & (40)\\ &= \mathbb{E}_q(\kappa)R + \mathrm{diag}(V_{\mathbf{X}}), & (41)\\ B_{\mathbf{X}} &= ((B_{\mathbf{X}_i}));\ i = 1:P. & (42) \end{aligned}$$

Any $(r,s)^{th}$ element in the precision matrix $Q^{\star}_{\mathbf{X}}$ has the following form;

$$Q^{\star}_{1\,rs} = \begin{cases} 2\mathbb{E}_q(\kappa) + V_{\mathbf{X}r}, & \text{if } r = s;\\ -\mathbb{E}_q(\kappa), & r \sim s;\\ 0, & \text{otherwise.} \end{cases}$$

where, $\mathbb{E}_q(\kappa)$ is the VB mean of the smoothing parameters κ, $V_{\mathbf{X}r}$ and $B'_{\mathbf{X}_i}$s are the functions of the first and the second derivatives of the VB expectation of log-likelihood at the posterior mode of $\mathbf{X}$ with respect to the other VB marginals.

The VB-marginal over C^f, $q_{\mathrm{C}^f}(\mathrm{C}^f)$ is not a tractable density (as earlier in the Gaussian case). For the tractable inference to the problem, a numerical approximation is applied to compute the normalizing constants and so the moments of intractable uni-dimensional density $q_{\mathrm{C}^f}(\mathrm{C}^f)$. The VB marginal over κ is again a Gamma density with hyper-parameters as functions of $\mathbf{X}$ and C^f.

4.4 VB solution to the more than one taxon and one climate model:

In this section, we present the VB approximations to the marginal posterior distributions over $\mathbf{X}_k$, κ_k, $\mathbf{U}_j$ and C^f for $k = 1, 2$.

The VB marginals $q_{\mathbf{X}_k}(\mathbf{X}_k)$ over $\mathbf{X}_k$; $k = 1, 2$ with the GMRF priors and the Poisson likelihood, are not tractable. The structure of the Gaussian approximation to $q_{\mathbf{X}_k}(\mathbf{X}_k)$ is same as to the VB marginal of $\mathbf{X}$ described for the one taxa model.

The VB marginals over the smoothing parameters κ_1 and κ_2 are Gamma densities. The hyper-parameters of the density are functions of the VB-moments of the responses and of the ancient climate. The VB marginal over uni-dimensional ancient climate is intractable as for the one taxa model.

The VB marginals over the bivariate random effects $\mathbf{U}_j; j = 1 : N$ are also not tractable densities. The Gaussian approximation to the intractable VB marginals over $\mathbf{U}_j$; $j = 1 : N$ is obtained as;

$$
\begin{aligned}
q^{\star}_{\mathbf{U}_j}(\mathbf{U}_j) &= BVN(\mu^{\star}_{\mathbf{U}_j}, Q^{\star}_{\mathbf{U}_j}) && (43)\\
Q^{\star}_{\mathbf{U}_j} &= \mathbb{E}_q(Q_{\mathbf{U}}) + \mathrm{diag}(V_{\mathbf{U}_j}) && (44)\\
\mu^{\star}_{\mathbf{U}_j} &= Q^{\star\,-1}_{\mathbf{U}_j} B_{\mathbf{U}_j}, && (45)
\end{aligned}
$$

where, the functions $V_{\mathbf{U}_j} = ((V_{\mathbf{U}_j k}))$; $k = 1, 2$ and $B_{\mathbf{U}_j} = ((B_{\mathbf{U}_j k}))$; $k = 1, 2$ depend on the first and second derivatives of the VB-expectation of the log-likelihood with respect to the other VB marginals.

VB marginal over the prior precision $Q_{\mathbf{U}}$ of the random effects $\mathbf{U}$ is a conjugate Wishart distribution.

$$
q_{Q_{\mathbf{U}}}(Q_{\mathbf{U}}) = \mathrm{Wishart}(df^{\star}, SS^{\star}), \tag{46}
$$

where, the hyper-parameters $df^{\star}$ and $SS^{\star}$ are the function of the VB-moments of the random effects $\mathbf{U}_j; j = 1 : N$.

5 Example

For the illustration of the VB approximation to the reconstruction problem, values are simulated from the likelihoods and the priors described in Section 2. A hundred regular spaced discrete values from 1 to 100 are assumed on climate location in the location space. Ten values are simulated on the climate from discrete uniform distribution, rages from 1 to 100. Vectors of 100 values on the response surfaces are generated from IGMRF prior densities. For the two taxa and one climate model, ten values on the bivariate random effects are derived from a bivariate Gaussian distribution.

Corresponding to each of the ten values on the climate, a data point on taxa is generated from the likelihoods described in Section 2. Nine of the ten values on the climate and the data on taxa are treated as the modern data and the tenth value on the data on taxa is used as ancient pollen data. The remaining one data point on the climate is assumed the true value on the ancient climate which is further used for the validation of the inference on the ancient climate.

We present the results on the VB approximation to the two types of the models discussed in the paper. The results are presented separately for different choices of the likelihoods.

1. **Result on one taxon and one climate model:** The results on the modelling of the responses and the estimation of the ancient climate are shown in Figure (3) for different likelihoods. The figure shows the comparison between the VB means and the true values (simulated) of the response surface. The data on taxon are also displayed in the figure to represent the indirect observations on the true (latent) responses. In Figure 3 for the Poisson and the Zero-Inflated Poisson likelihoods, the log of the data points are shown as the indirect observations on the responses, as the responses are assumed to be the log-function of the count data in the likelihoods (rate parameters are the exponential function of the latent responses).

   The VB marginal over the ancient climate is also presented in the figure. The data points having the same at different climate locations result in the multi-modal VB approximation (VB marginal) over the ancient climate.

2. **Result on the two taxa and one climate model:** Figure (4) shows the VB approximations to the modelling of the responses for two taxa and that to the estimation of the ancient climate. The true responses and VB estimated responses are shown in the figure for the comparison of the result with the true value. The log of the data points minus estimated (VB) random effects as the indirect observations on the latent responses are also displayed in the figure (as the rate parameters assumed in the model is the exponential of the addition of the responses and the random effects at the observed locations).

In Figures (3) and (4), the VB-means of the response surface pass through each of the modern data points, so are not a very smooth function of the data on taxa as they are desired. The VB marginal of the ancient climate is multi-modal due to the non-monotonic behaviour of the responses. The spiky modes of the VB marginal show the loss of some uncertainty in the estimation. The loss of accuracy in the estimation due to the posterior independence assumption of the VB method.

It is clear from the figures that the VB-variance of the responses are small at the locations where data on taxa is available. For the Poisson and ZI-Poisson likelihoods the variance of the responses are much smaller than those for the Gaussian likelihood. Making Gaussian approximations for the tractable VB approximation forces us to lose some accuracy (the uncertainty). It also emphasizes on the fact that the VB method works well only for the exponential family of distributions and the conjugate priors.

Though the Zero-Inflated Poisson model is capable of modelling the extra zero counts in data, the VB approximation over the ancient climate for the model is not very satisfactory (modes in the approximation are not close to the true value). The VB approximation requires some further approximations in the ZI-Poisson likelihood for the tractable solution which deteriorates the approximation on the inference on the ancient climate.

It is clear from Figure (4) that the (VB) variance of the responses for the two taxa model are larger than those obtained for the one taxon model. So we gain in the uncertainty in the VB

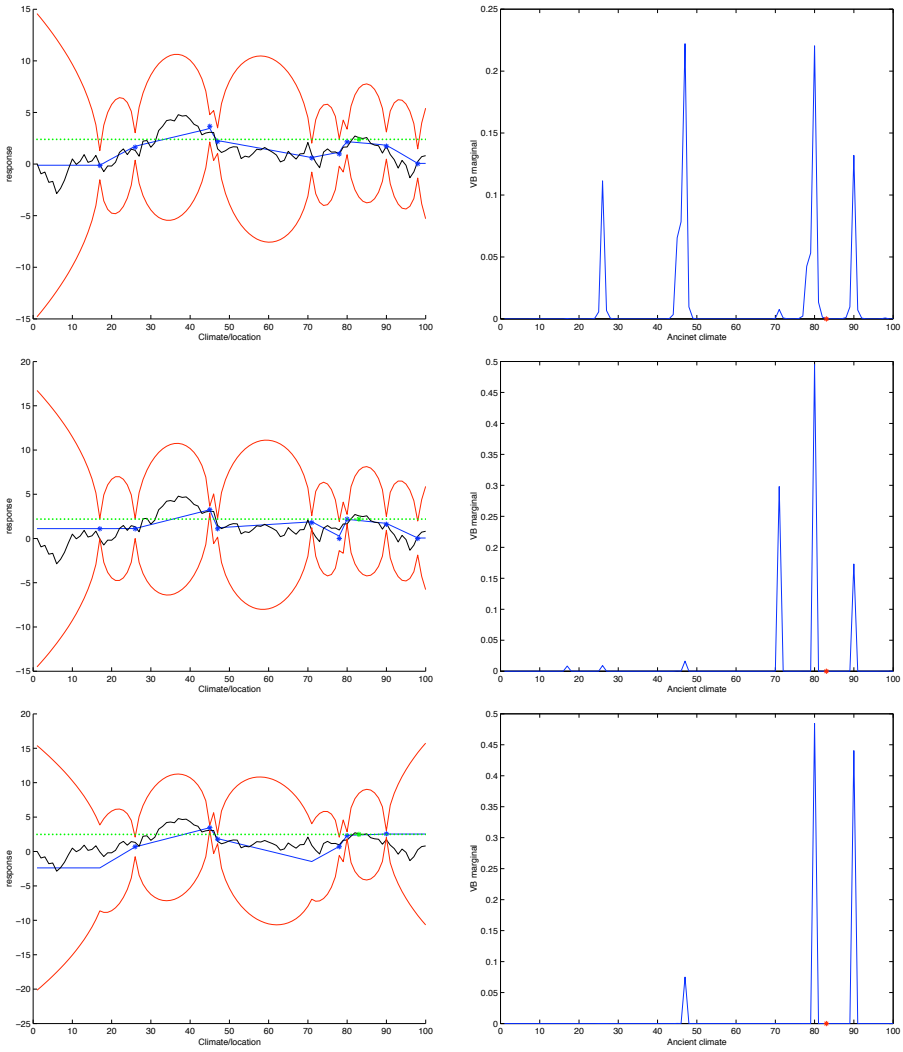

Figure 3: (Left) the black line denotes the true value of the response for the taxon, blue line is for the estimated response, red lines show the 95% HPD region, the blue asterisks represent the observations on taxon, the green asterisk stands for the data point on the taxon assumed as a single observation on the fossil pollen, (right) the VB marginal of the ancient climate given the observation on fossil pollen, the red asterisk shows the true (assumed) value of the ancient climate, the first block of the figures are for the Gaussian likelihood, second block, for the Poisson and the third is for the Zero-Inflated Poisson likelihood.

marginal over the ancient by introducing the random effects in the two taxa and one climate model.

6 Discussion

VB provides a relatively fast functional approximation to a posterior distribution, through making assumptions of independence in the posterior. It presents a trade-off between ease of computation, via more independence assumptions, and accuracy of the approximation. Palaeoclimate reconstruction is a good example of its use in an inverse regression problem.

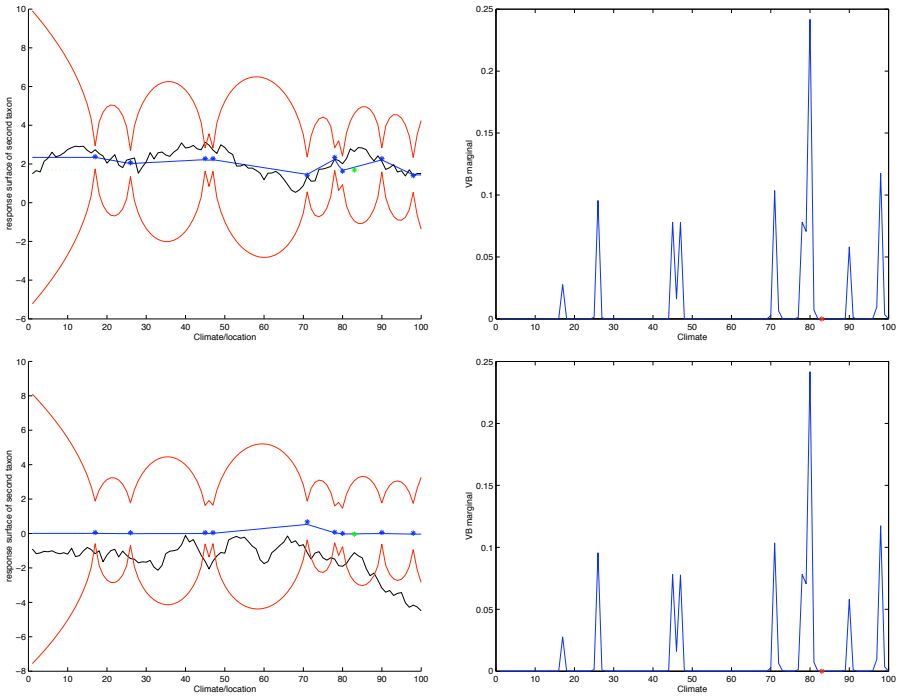

Figure 4: (Left) the black line denotes the true value of the response for the first taxon, blue line is for the estimated response, red lines show the 95% HPD region, the blue asterisks represent the observations on taxon, the green asterisk stands for the data point on the taxon assumed as a single observation on the fossil pollen, (right) the VB marginal of the ancient climate given the observation on fossil pollen, the red asterisk shows the true (assumed) value of the ancient climate.

From the results for the model with Gaussian likelihood, it can be understood that the method performs well for the conjugate family of distributions. It provides tractable approximations to the intractable posterior distributions over unknown parameters for conjugate prior distributions. We can include many unknown parameters for estimation (assumed with the conjugate prior distributions) in the model.

Unlike the MCMC method, the multi-modal behaviour of climate does not create any problem for a tractable VB solution. We learn (from Figures 3 and 4) that the VB method does have problems in convergence due to the multi-modal behaviour of climate. It performs quickly and provides a multi-modal VB-approximation to the multi-modal posterior distribution of the ancient climate. However, the resulting posterior distributions on climate appear to have smaller variance than desirable, in the sense that the true climate value is often not supported by the VB approximation. This is a commonly observed feature of VB approximations with some theoretical support (see Wang and Titterington (2004a), and Wang and Titterington (2004b)).

VB approximations for intractable posterior distributions with non-conjugate distributions (likelihoods and priors) may not be recognized as standard distributions, which requires further approximations to the multi-dimensional VB approximations known up to proportionality constants. Gaussian approximations to intractable VB approximations with GMRF or Gaussian priors and non-conjugate likelihoods are presented in the paper.

The VB approximation can be made more accurate by introducing the dependence between

taxa. But it may induce computational complexity in the inference problem. The assumption of random effects in the model, though increases the number of unknowns to estimate, but helps in gaining accuracy in the VB approximation without making the problem very complex.

The zero-inflated Poisson distribution is the most appropriate of the three likelihoods for real data, but the VB implementation requires that the zero-inflation parameter is known. How to tractably incorporate inference for this parameter is clearly important.

Since the responses of taxa to climate are latent and are modelled through the data on taxa, considering many more data points may smoothen the response surfaces. Making the inference problem not very complex and more accurate, the dependency between taxa can be allowed in nested structure, see (Salter-Townshend, 2009), keeping the intractability issue of the VB method in mind.

7 Bibliography

References

Allen, R. M. J., W. A. Watts, and B. Huntley (2000). Weichselian palynostratigraphy, palaeovegetation and palaeoenvironment; the record from lago grande di monticchio, southern italy. *Quaternary International 73–74*, 91–110.

Antonsson, K., S. J. Brooks, H. Seppä, R. J. Telford, H. John, and B. Birks (2006). Quantitative palaeotemperature records inferred from fossil pollen and chironomid assemblages from Lake Gilltjärnen, northern central Sweden. *21*(8), 831–841.

Bhattacharya, S. (2004). *Importance Resampling MCMC: A Methodology for Cross–Validation in Inverse Problems and Its Applications in Model Assessment*. Ph. D. thesis, Trinity College Dublin.

Haslett, J., M. Whiley, S. Bhattacharya, M. Salter-Townshend, S. P. Wilson, J. R. M. Allen, B. Huntley, and F. J. G. Mitchell (2006, July). Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society : Series A (Statistics in Society) 169*(3), 395–438.

Huntley, B. (1991). How plants respond to climate change: , Migrate rates, individualism and consequences for plant communities. *Annals of Botany 67*, 15–22.

Huntley, B., R. A. Spicer, W. G. Chaloner, and E. A. Jarzembowski (1993, August). The Use of Climate Response Surfaces to Reconstruct Palaeoclimate from Quaternary Pollen and Plant Macrofossil Data.

Kullback, S. (1997, Oct). *Information Theory and Statistics*. Dover Publications Inc.

Nakajima, S. and S. Watanabe (2006). Analytic Solution of Hierarchical Variational Bayes in Linear Inverse Problem. In *ICANN* (2), pp. 240–249.

Ridout, M., G. B. Demetrio, and J. Hinde (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, Invited Papers, pp. 179–192.

Rue, H. and L. Held (2005, February). *Gaussian Markov random fields. Theory and applications (Monographs on Statistics and Applied Probability)* (1 ed.). Chapman & Hall/CRC.

Rue., H., S. Martino, and N. Chopin (2009). Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B 71(2)*, 1–35.

Salter-Townshend, M. (2009). *Fast Approximate Inverse Bayesian Inference in non-parametric Multivariate Regression with application to palaeoclimate reconstruction*. Ph. D. thesis, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland.

Šmídl, V. and A. Quinn (2006). *The Variational Bayes Method in Signal Processing*. Signals and Communication Technology. Springer.

Wang, B. and D. M. Titterington (2004a). Convergence and Asymptotic Normality of Variational Bayesian Approximations for Expon. In D. M. Chickering and J. Y. Halpern (Eds.), *UAI*, pp. 577–584. AUAI Press.

Wang, B. and D. M. Titterington (2004b). Variational Bayes Estimation of Mixing Coefficients. In J. Winkler, M. Niranjan, and N. D. Lawrence (Eds.), *Deterministic and Statistical Methods in Machine Learning*, Volume 3635 of *Lecture Notes in Computer Science*, pp. 281–295. Springer.

8 Appendix

Gaussian Likelihood:
For the VB marginal over $\mathbf{X}$:
for $i = 1 : P$;

$$\begin{aligned} V_{\mathbf{X}_i} &= [n_i + q(c^f = i)]\mathbb{E}_q\left(\frac{1}{r^2}\right), \\ n_i &= \text{No. of times } i \text{ equals to } C^m, \\ B_{\mathbf{X}_i} &= \left[\sum_{\substack{j=1 \\ C_j^m = i}}^{N^m} \mathrm{y}_j^m + q(C^f = i)\mathrm{y}^f\right] \mathbb{E}\left(\frac{1}{r^2}\right). \end{aligned}$$

For the VB marginal over κ:

$$\begin{aligned} a^\star &= a + 0.5(P-1), \\ b^\star &= b + \frac{1}{2}\Bigg[2\sum_{i=1}^{P} q(C^f = i)\mathbb{E}_q(\mathbf{X}_i^2) + \sum_{i=2}^{P} q(C^f = i)\mathbb{E}_q(\mathbf{X}_{i-1}^2) + \sum_{i=1}^{P-1} q(C^f = i)\mathbb{E}_q(\mathbf{X}_{i+1}^2) \\ &\quad -2\sum_{i=1}^{P-1} q(C^f = i)\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i-1}) - 2\sum_{i=2}^{P} q(C^f = i)\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i+1})\Bigg] \\ &\quad +\frac{1}{2}\left[2\sum_{i=2}^{P-1}\mathbb{E}_q(\mathbf{X}_i^2) + \mathbb{E}_q(\mathbf{X}_1^2) + \mathbb{E}_q(\mathbf{X}_P^2) - 2\sum_{i=2}^{P}\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i-1})\right], \end{aligned}$$

For the VB marginal over r^2:

$$\begin{aligned}
\alpha^\star &= \alpha + \frac{N^m}{2}, \\
\beta^\star &= \beta + \frac{1}{2}\sum_{j=1}^{N^m}\Big[\mathrm{y}_j^{m\,2} - 2\mathrm{y}_j^m \mathbb{E}_q(\mathbf{X}(\mathrm{C}_j)) + \mathbb{E}_q(\mathbf{X}^2(\mathrm{C}_j))\Big] \\
&\quad + \frac{1}{2}\Big[\mathrm{y}^{f\,2} - 2\mathrm{y}^f \sum_{i=1}^{P} q(C^f = i)\mathbb{E}_q(\mathbf{X}_i) + \sum_{i=1}^{P} q(C^f = i)\mathbb{E}_q(\mathbf{X}_i^2)\Big],
\end{aligned}$$

For VB marginal over C^f: for $i = 2 : P-1$,

$$\begin{aligned}
q(C^f = i) \;\propto\; \exp\Bigg[&-\frac{1}{2}\mathbb{E}_q\left(\frac{1}{r^2}\right)\left(\mathrm{y}^f \mathbf{X}_i + \mathbf{X}_i^2\right) \\
&-\frac{1}{2}\mathbb{E}_q(\kappa)\Big\{2\mathbf{X}_i^2 + \mathbf{X}_{i-1}^2\mathbf{X}_{i+1}^2 - 2\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i-1}) - 2\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i+1})\Big\}\Bigg]
\end{aligned}$$

Poisson Likelihood:
For the VB marginal over $\mathbf{X}$:
for $i = 1 : P$;

$$\begin{aligned}
V_{\mathbf{X}_i} &= [n_i + q(C^f = i)]e^{\mathbf{X}_i^M}, \\
B_{\mathbf{X}_i} &= \left[\sum_{\substack{j=1 \\ C_j^m = i}}^{N^m} \mathrm{y}_j^m + q(C^f = i)\mathrm{y}^f\right] - (1 - \mathbf{X}_i^M)V_{\mathbf{X}_i}, \\
\mathbf{X}_i^M &= \text{Mode of the marginal posterior distribution over } \mathbf{X}_i.
\end{aligned}$$

The functional forms of the hyper-parameters of the VB marginal over κ are the same as those defined for Gaussian likelihood.

For the VB marginal over C^f: for $i = 2 : P-1$,

$$\begin{aligned}
q(C^f = i) \;\propto\; \exp\Bigg[&-\frac{1}{2}\mathbb{E}_q(\kappa)\Big\{2\mathbf{X}_i^2 + \mathbf{X}_{i-1}^2\mathbf{X}_{i+1}^2 - 2\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i-1}) - 2\mathbb{E}_q(\mathbf{X}_i\mathbf{X}_{i+1})\Big\} \\
&+\Big\{-\mathbb{E}_q(e^{\mathbf{X}_i}) + \mathrm{y}^f\mathbb{E}_q(\mathbf{X}_i)\Big\}\Bigg]
\end{aligned}$$

Zero-Inflated Poisson Likelihood:
For the VB marginal over $\mathbf{X}$:

for $i = 1 : P$

$$
\begin{aligned}
V_{\mathbf{X}_i} &= \Bigg[\sum_{\substack{j=1 \\ C_j^m = i}}^{N^m} (1 - Z_j^m) + q(C^f = i)(1 - Z^f) \Bigg] e^{\mathbf{X}_i^M} \\
&\quad + e^{-e^{\mathbf{X}_i^M}} e^{\mathbf{X}_i^M} \sum_{\substack{j=1 \\ C_j^m = i}}^{N^m} \Bigg[e^{-e^{\mathbf{X}_i^M}} e^{\mathbf{X}_i^M} q_j^{m\,2} Z_j^{m\,2} \Bigg\{ 1 - q_j^m + q_j^m e^{-e^{\mathbf{X}_i^M}} \Bigg\}^{-2} \\
&\quad + (1 - e^{\mathbf{X}_i^M}) q_j^m Z_j^m \Bigg\{ 1 - q_j^m + q_j^m e^{-e^{\mathbf{X}_i^M}} \Bigg\}^{-1} \Bigg] \\
&\quad + e^{-e^{\mathbf{X}_i^M}} e^{\mathbf{X}_i^M} q(C^f = i) \Bigg[e^{-e^{\mathbf{X}_i^M}} e^{\mathbf{X}_i^M} q^{f\,2} Z^{f\,2} \Bigg\{ 1 - q^f + q^f e^{-e^{\mathbf{X}_i^M}} \Bigg\}^{-2} \\
&\quad + (1 - e^{\mathbf{X}_i^M}) q^f Z^f \Bigg\{ 1 - q^f + q^f e^{-e^{\mathbf{X}_i^M}} \Bigg\}^{-1} \Bigg],
\end{aligned}
$$

$$
\begin{aligned}
B_{\mathbf{X}_i} &= -\mathbf{X}_i^M V_{\mathbf{X}_i} + \sum_{\substack{j=1 \\ C_j^m = i}}^{N^m} \Bigg[(1 - Z_j^m) \Bigg\{ e^{\mathbf{X}_i^M} + \mathrm{y}_j \Bigg\} \\
&\quad - e^{-e^{\mathbf{X}_i^M}} e^{\mathbf{X}_i^M} q_j^m Z_j^m \Bigg\{ 1 - q_j^m + q_j^m e^{-e^{\mathbf{X}_i^M}} \Bigg\}^{-1} \Bigg] \\
&\quad + q(C^f = i) \Bigg[(1 - Z^f) \Bigg\{ e^{\mathbf{X}_i^M} + \mathrm{y}^f \Bigg\} \\
&\quad + q^f Z^f e^{-e^{\mathbf{X}_i^M}} e^{\mathbf{X}_i^M} \Bigg\{ 1 - q^f + q^f e^{-e^{\mathbf{X}_i^M}} \Bigg\}^{-1} \Bigg], \\
\mathbf{X}_i^M &= \text{Mode of the marginal posterior distribution over } \mathbf{X}_i.
\end{aligned}
$$

Z^f and Z_j^m's are the known auxiliary variables defined as $Z_j^m = \begin{cases} 1, & \text{if } \mathrm{y}_j^m = 0; \\ 0, & \text{if } \mathrm{y}_j^m > 0 \end{cases}$

$Z^f = \begin{cases} 1, & \text{if } \mathrm{y}^f = 0; \\ 0, & \text{if } \mathrm{y}^f > 0 \end{cases}$

The functional forms of the hyper-parameters of the VB marginal over κ are the same as those defined for Gaussian likelihood.

For the VB marginal over C^f:
for $i = 2 : P - 1$,

$$
\begin{aligned}
q(C^f = i) &\propto \exp \Bigg[-\frac{1}{2} \mathbb{E}_q(\kappa) \Bigg\{ 2\mathbf{X}_i^2 + \mathbf{X}_{i-1}^2 \mathbf{X}_{i+1}^2 - 2\mathbb{E}_q(\mathbf{X}_i \mathbf{X}_{i-1}) - 2\mathbb{E}_q(\mathbf{X}_i \mathbf{X}_{i+1}) \Bigg\} \\
&\quad + Z^f \mathbb{E}_q \Bigg[\log \Bigg\{ 1 - q^f + q^f e^{-e^{\mathbf{X}_i}} \Bigg\} \Bigg] + (1 - Z^f) \Bigg\{ -\mathbb{E}_q \left(e^{\mathbf{X}_i} \right) + \mathrm{y}^f \mathbb{E}_q(\mathbf{X}_i) \Bigg\} \Bigg]
\end{aligned}
$$

Second order Taylor's expansion of $\log\left\{1 - q^f + q^f e^{-e^{\mathbf{X}_i}}\right\}$ around $\mathbf{X}_i = 0$ is equal to

$$-q^f(1-q^f)e^{-e^{\mathbf{X}_i}} + \frac{1}{2}q^{f\,2}e^{-2e^{\mathbf{X}_i}}.$$

Considering further approximations to $e^{-e^{\mathbf{X}_i}}$ and $e^{-2e^{\mathbf{X}_i}}$ as

$$e^{-e^{\mathbf{X}_i}} \approx -e^{\mathbf{X}_i} + \frac{1}{2}e^{2\mathbf{X}_i} - \frac{1}{6}e^{3\mathbf{X}_i},$$
$$e^{-2e^{\mathbf{X}_i}} \approx -2e^{\mathbf{X}_i} + 2e^{2\mathbf{X}_i} - \frac{4}{3}e^{3\mathbf{X}_i}$$

$$\begin{aligned}\mathbb{E}_q\left[\log\left\{1 - q^f + q^f e^{-e^{\mathbf{X}_i}}\right\}\right] &= q^f\left[-\mathbb{E}_q(e^{\mathbf{X}_i}) + \frac{1}{2}\mathbb{E}_q(e^{2\mathbf{X}_i}) - \frac{1}{6}\mathbb{E}_q(e^{3\mathbf{X}_i})\right] \\ &\quad + \frac{1}{2}q^{f\,2}\left[\mathbb{E}_q(e^{2\mathbf{X}_i}) - \mathbb{E}_q(e^{3\mathbf{X}_i})\right]\end{aligned}$$

More than one taxa an done climate model: For the VB marginal over $\mathbf{X}_k$ $k = 1:2$:
for $i = 1:P$

$$\begin{aligned}V_{\mathbf{X}_{ki}} &= \sum_{\substack{j=1 \\ C_j^m = i}}^{N^m}\left[\mathbb{E}_q\left(e^{\mathrm{U}_{kj}^m}\right) + q(C^f = i)\mathbb{E}_q\left(e^{\mathrm{U}_k^f}\right)\right]e^{\mathbf{X}_{ki}^M};\ i = 1:P, \\ B_{\mathbf{X}_i} &= \left[\sum_{\substack{j=1 \\ C_j^m = i}}^{N^m} \mathrm{y}_{kj}^m + q(C^f = i)\mathrm{y}_k^f\right] - (1 - \mathbf{X}_{ki}^M)V_{\mathbf{X}_{ki}}, \\ \mathbf{X}_i^M &= \text{Mode of the marginal posterior distribution over } \mathbf{X}_i.\end{aligned}$$

The functional forms of the hyper-parameters of the VB marginal over κ_1 and κ_2 are the same as those defined for Gaussian likelihood.

The VB marginal over the random effects

$$q_{\mathbf{U}}(\mathbf{U}) = \prod_{j=1}^{N^m}\left[q_{\mathbf{U}_j^m}(\mathbf{U}_j^m)\right]q_{\mathbf{U}^f}(\mathbf{U}^f)$$

For the VB marginal over $\mathbf{U}_j^m$:
for $j = 1:N^m$

$$\begin{aligned}V_{\mathrm{U}_{kj}^m} &= e^{\mathrm{U}_{kj}^M}\mathbb{E}_q(e^{\mathbf{X}_k(\mathrm{C}_j^m)}), \\ B_{\mathrm{U}_{kj}^m} &= \mathrm{y}_{kj}^m - (1 - \mathrm{U}_{kj}^M)V_{\mathrm{U}_{kj}^m}, \\ \mathrm{U}_{kj}^M &= \text{Mode of the marginal posterior distribution over } \mathrm{U}_{kj}^m\end{aligned}$$

and for the VB marginal over $\mathbf{U}^f$:

$$\begin{aligned} V^f_{\mathrm{U}_k} &= e^{\mathrm{U}^f_k{}^M} \sum_{i=1}^{P} \Big\{ q(C^f = i) \mathbb{E}_q(e^{\mathbf{X}_{ki}(\mathrm{C}^f)}) \Big\}, \\ B^f_{\mathrm{U}_k} &= \mathrm{y}^f_k - (1 - \mathrm{U}^{Mf}_k) V^f_{\mathrm{U}_k}, \\ \mathrm{U}^{f\,M}_k &= \text{Mode of the marginal posterior distribution over } \mathrm{U}^f_k \end{aligned}$$

For the VB marginal over $Q_{\mathbf{U}_j}$

$$\begin{aligned} df^\star &= df + N^m + 1, \\ SS^\star &= SS^{-1} + \mathbb{E}_q(\mathbf{U}'^m_j \mathbf{U}^m_j) + \mathbb{E}_q(\mathbf{U}'^f_j \mathbf{U}^f). \end{aligned}$$

For the VB marginal over C^f: for $i = 2 : P - 1$

$$\begin{aligned} q(C^f = i) \;\propto\; \exp\Bigg[\sum_{k=1}^{2} \Bigg[-\frac{1}{2} \mathbb{E}_q(\kappa_k) \Big\{ 2\mathbf{X}^2_{ki} + \mathbf{X}^2_{k(i-1)} \mathbf{X}^2_{k(i+1)} - 2\mathbb{E}_q(\mathbf{X}_{ki} \mathbf{X}_{k(i-1)}) - 2\mathbb{E}_q(\mathbf{X}_{ki} \mathbf{X}_{k(i+1)}) \Big\} \\ + \Big\{ -\mathbb{E}_q(e^{\mathbf{X}_{ki}}) + \mathrm{y}^f_k \mathbb{E}_q(\mathbf{X}_{ki}) \Big\} \Bigg] \Bigg]. \end{aligned}$$