A Bayesian design pattern for regression problems

Kari T Vasko[*],
Research Seminar on Intelligent Systems (spring 2001),
University of Helsinki

February 22, 2001

Abstract

We outline a generic Bayesian mixture response model for regression problems. It is supposed that response variables are counting data. Statistical nature of other attributes may be practically arbitrary.

We demonstrate the use of the model by applying it to a paleoecological reconstruction problem.

1 Introduction

Preliminaries: Suppose we have observed n datapoints $d = (d_1, \ldots, d_n)$, where $d_i = (x_i, y_i)$, $x_i \in \Omega_X$ and $y_i = (y_{i1}, \ldots, y_{im})^T \in \mathbb{R}^m$. Further, suppose $d_0 = (x_0, y_0)$ consists of the observation $y_0 \in \mathbb{R}^m$ and unknown state of affairs $x_0 \in \Omega_X$. We use following abbreviations: $[a] = \{0, 1, \ldots, a\}$, $a_{i+} = \sum_j a_{ij}$ and $a_{+j} = \sum_i a_{ij}$.

Regression model with additivie noise model is

$$\mathbf{E}(y|x, \theta) + \epsilon, \tag{1}$$

where ϵ is noise, $\mathbf{E}(\cdot|\cdot)$ is the expectation operator and θ represents the model parameters. Regression is called, alternatively, classification, if Ω_X is nominal.

Role of prior knowledge in the presence of uncertainty: Predicting values outside the training data implies making inductive generalizations. Basically, all generalization is based on the prior knowledge. The prior knowledge provides the link between the training data and the not yet measured future samples x_0. Thus, one has to assume a non-empty set of *a priori* restrictions $\mathcal{K}$ in order to extrapolate outside the training data [Wol96, WMa96]. Let us denote inductive generalization conditioned on restrictions $\mathcal{K}$ by

$$f(x|d, y_0, \mathcal{K}), \text{ where } x \in \Omega_X. \tag{2}$$

Assuming Cox-Jaynes axioms [Cox46, Jay94] the only consistent inference mechanism is up to a constant Bayesian inference and, consequently, the full probability model setting. Here we adopt Cox-Jaynes axioms and, accordingly, the full probability modeling. It follows that object (2) is some probability density distribution and we will call it the posterior probability density.

[*]email: Kari.Vasko@csc.fi

Bayesian modeling is an approach to statistics in which all forms of uncertainty are expressed in terms of probability densities, i.e., a full probability model $f(d, y_0, x_0, \theta)$ for both observed (d, y_0) and unobserved (x_0, θ) quantities is defined, where $\theta \in \Omega_\Theta$ represents the model parameters. The most fundamental difference between Bayesian and classical statistics is that in the classical statistics observed data is considered random and model parameters fixed, whereas in the Bayesian approach unobserved state of affairs, e.g., model parameters and missing data, are considered random and observed data fixed. This issue is related to the interpretation of probability. For a profound discussion on this subject see [Jay94, GCS95, BSm94].

The posterior distribution is obtained by applying Bayes rule

$$f(x|d, y_0, \mathcal{K}) = \int_{\Omega_\Theta} f(x, \theta|d, y_0, \mathcal{K}) d\theta \underset{\theta}{\propto} \int_{\Omega_\Theta} f(x, \theta|\mathcal{K}) f(y|x, \theta, \mathcal{K}) d\theta \tag{3}$$

and marginalizing over the domain of θ. Here $f(x, \theta|\mathcal{K})$ is the prior probability distribution and it embodies our judgement (before seeing any data d) for the parameters θ and missing values x, of how plausible it is that the parameters could have values in the various regions of the parameter space. Further, $f(y|x, \theta, \mathcal{K})$ is the likelihood function that manifests also background knowledge concerning, e.g., sampling arrangement and conditional independence assumptions. In regression problems likelihood can be seen as a response function, specifying how observations y_i depend statistically from the observations x_i.

In general, Bayesian modeling procedure consists of three steps [GCS95]:

1. Build up a full probability model. A full probability model defines a joint distribution for all parameters of the problem domain (also for observed data).

2. Evaluate joint probability distribution by conditioning on the observed data, that is, evaluate the posterior probability density (3).

3. Assess whether the model fits the data. How robust the model is w.r.t. model assumptions?

2 Model specification

In this section we will consider regression problems in the following setting: Suppose y_i's are response variables and counting data of y_{i+} trials, i.e.,

$$y_i = \sum_{t=1}^{y_{i+}} \mathbf{e}_{q(t)}, \tag{4}$$

where $\mathbf{e}_i$ is ith standard base vector in $\mathbb{R}^m$ and $q(t) \in [m]$ is a realization of the tth trial ($1 \leq t \leq t_i$ for all $i \in [n]$). Further, suppose that y_i and y_j are mutually independent for all $i \neq j$ and $i, j \in [n] \times [n]$.

In the following subsections we will introduce a Bayesian mixture response model for regression problems based on the counting data. A general view of the model structure is illustrated in figure 1.

2.1 The regression function

It is a requisite that there is some *a priori* knowledge concerning relationship between y_i and corresponding conditions x_i. Suppose we have M competing response models $r_1, \ldots, r_M$ representing background knowledge related to this interaction.

We denote parameters of the response model j for the observations of class k $(y_{ik})_{i\in[n]}$ as θ_{jk}, and denote $\theta_k = (\theta_{1k}, \ldots, \theta_{Mk})$. Response for the observations of class k $(y_{ik})_{i\in n}$ given model j is

$$r_j(x_i;\theta_{jk}) = \lambda^j_{ik}. \tag{5}$$

Suppose random variable ϕ^j_{ik} denotes probability of the response model j for the class k given conditions x_i, and $\phi_{ik} = (\phi^1_{i1}, \ldots, \phi^M_{im})$ (remark $\sum_{l=1}^M \phi^l_{ik} \equiv 1$).

We use regression model

$$\mathbf{E}(\hat{y}_{ik}|\Theta, x_i) = \frac{\sum_{l=1}^M \phi^l_{ik}\lambda^l_{ik}}{\sum_{l'=1}^M \sum_{k'=1}^m \phi^{l'}_{ik}\lambda^{l'}_{ik'}} \tag{6}$$

for the proportion $\hat{y}_{ik}$ of observation y_{ik}, where $\Theta = (\phi^1_{i1}, \ldots, \phi^M_{im}, \theta_1, \ldots, \theta_m)$.

Example 1. *Suppose* $(y_1, \ldots, y_n)$ *are representing occurences of taxon abundances in different environments* $(x_1, \ldots, x_n)$*. That is,* y_{ik} *represents abundance of taxon* k *in environmental conditions* x_i*. An example of response model is, for instance,*

$$\lambda^1_{ik} = r_1(x_i;\theta_{1k}) = \alpha_k exp\Big[-\Big(\frac{\beta_k - x_i}{\gamma_k}\Big)^2\Big], \tag{7}$$

where $\theta_{1k} = (\alpha_k, \beta_k, \gamma_k)$ *are parameters related to taxon (class)* k*. Another example of a response model is*

$$\lambda^2_{ik} = r_2(x_i;\theta_{2k}) = 0 \textit{ for all arguments} \tag{8}$$

($\theta_{2k} = ()$*).*

2.2 The likelihood function

Due to assumptions given in section 2 y_i's are distributed multinomially

$$(y_{i1}, \ldots, y_{im}|y_{i+}, p_i) \sim \mathrm{Mu}(y_{i+}, p_i) \tag{9}$$

with the parameter vector $p_i = (p_{i1}, \ldots, p_{im})$, where p_{ik} is the probability that a random individual is of class k given conditions x_i.

2.3 The prior distributions

Priors for ϕ_{ik}'s: The natural distribution of random variables on a simplex with a strictly positive density, that satisfy certain natural mutual conditional independence assumptions, is the Dirichlet distribution [GHe97]. Consequently, we assume that ϕ_{ik}'s are Dirichlet distributed

$$(\phi^1_{ik}, \ldots, \phi^M_{ik}|\kappa_1, \ldots, \kappa_M) \sim \mathrm{Di}(k_1, \ldots, k_M), \tag{10}$$

where κ_i's are global hyperparameters representing general evidence for the response model i. A non-informative (uniform) choice is to set $\kappa_i := 1$ for all $1 \leq i \leq M$. The Dirichlet distribution is a multi-dimensional generalization of the Beta distribution (see, e.g., [GCS95] pages 481–482).

Priors for p_i's: p_i's should be considered stochastic in order to leave room for variability between observations. It is again obvious that $p_{i+} = 1$. Following above mentioned reasoning we assume that p_i's are Dirichlet distributed

$$(p_{i1}, \dots, p_{im} | x_i, \Theta) \sim \mathrm{Di}\Big(\sum_{l=1}^{M} \phi_{1k}^l \lambda_{1k}^l, \dots, \sum_{l=1}^{M} \phi_{mk}^l \lambda_{mk}^l \Big) \tag{11}$$

given mixture responses. In addition, the Dirichlet distribution is a natural conjugate prior of multinomial distribution, so

$$\begin{aligned} f(y_0, y_1, \dots, y_n | p_1, \dots, p_n) f(p_1, \dots, p_n | x_0, x_1, \dots, x_n, \Theta) = \\ \prod_{i=0}^{n} \mathrm{Di}\Big(y_{i1} + \sum_{l=1}^{M} \phi_{1k}^l \lambda_{1k}^l, \dots, y_{im} + \sum_{l=1}^{M} \phi_{mk}^l \lambda_{mk}^l \Big). \end{aligned} \tag{12}$$

The Dirichlet distribution satisifies assumption Eq.6, since

$$\mathbf{E}(p_{ik} | \Theta, x_i) = \frac{\sum_{l=1}^{M} \phi_{ik}^l \lambda_{ik}^l}{\sum_{l'=1}^{M} \sum_{k'=1}^{m} \phi_{ik'}^{l'} \lambda_{ik'}^{l'}} \text{ for all } i \text{ and } k. \tag{13}$$

Alternatively functionally determined responses could be used, i.e., $p_{ik} := \mathbf{E}(\hat{y}_{ik} | \Theta, x_i)$ (recall $E(c|\Theta) = c$, where c is some constant).

Priors for the response model parameters: Response models are fixed by an application domain expert. In general, there are two approaches to determine hyperparameters for the prior distributions of the response model parameters; (i) parameters may be estimated from the observed data (empirical Bayes approach) or; (ii) they may be estimated by an application domain expert (pure Bayes approach).

3 Example: Organism-based environmental reconstruction

3.1 Background

Paleoecology is a science that studies environmental conditions in the past. In the beginning, paleoecological study was based on the qualitative characterization of the past environment, but in the early 1970s the paleoecologists Imbrie and Kipp introduced the study of quantitative statistical methods for paleoecological reconstructions. After that, paleoecology has experienced an explosion of new ideas, approaches and techniques. In recent years, the whole discipline has been changing from a predominantly description-oriented subject to a quantitative science.

Methodological studies in this application domain are well-motivated, since they have considerable relevance with issues like global warming. The state-of-the-art methodology for paleoenvironmental reconstruction has been developed through an interdisciplinary interaction between researchers representing methodological sciences (computer science, mathematics and statistics) and paleoecology.

The time span of paleoecological reconstructions typically covers the Holocene, i.e., the last 10.000 years, and climate history reconstructed for some hundreds

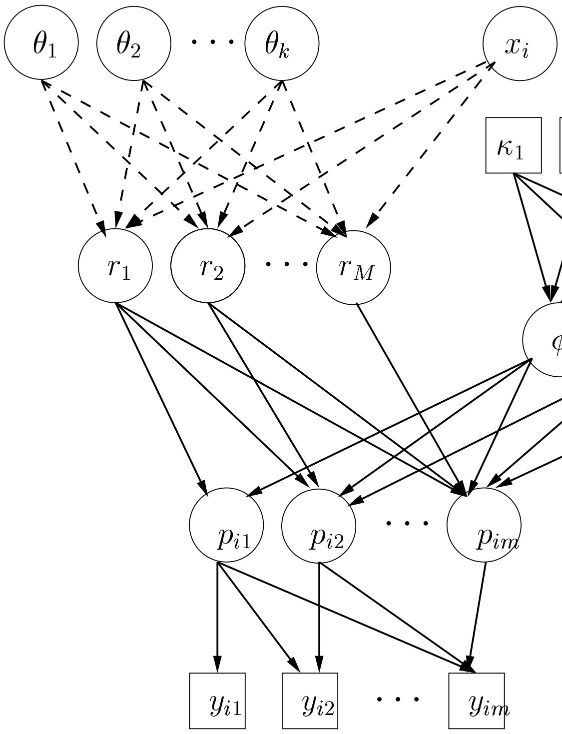

Figure 1: Bayesian network structure for the model specified in the section 2.

Modern data

Lake	Temperature	Species 1	Species 2
1	13.74 C	20	80
2	9.21 C	60	40

Fossil data

Layer	Species 1	Species 2
12	24	76
20	50	50
45	15	85

Table 1: Example data for reconstruction.

of points in this period. The modern training data is collected from a number of spatially distributed sites, but reconstructions are usually based on fossil data from few sites only, due to the laborious efforts needed in collecting and analyzing the fossil data.

As a simple example of the reconstruction task, consider two species and two present-day lakes, with the observations given in table 1. A reasonable guess is now that warm temperatures are more suitable for species 2 than for species 1, and vice versa for cold temperatures. Given the fossil assemblage in the right-hand part of the table, layer 12 looks warm, layer 20 cold, and 40 warm.

3.2 Embedding ecological background knowledge

Numerous environmental data-sets show that the distribution of the taxa abundances tend to occur over a characteristic, but limited environmental range. Further,

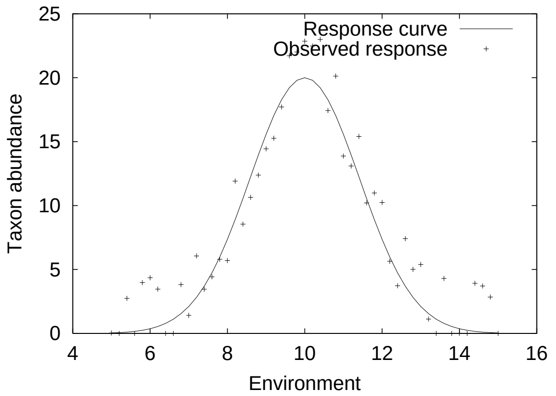

Figure 2: An example of response curve and corresponding (environment,abundance) observations.

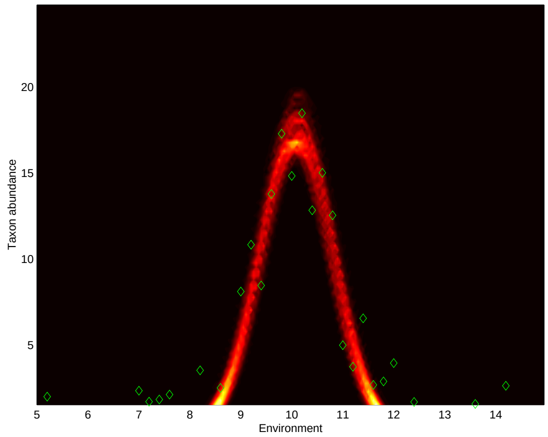

Figure 3: An approximation of the posterior response distribution. Lighter colors indicate higher posterior density. Diamonds are representing observed data.

within this range, the abundances seem to occur roughly symmetrically around an optimum [Bir95]. Consequently, we assume that the response curve for the actual abundances has a unimodal Gaussian shape controlled by parameters α_k (scaling factor), β_k (optimum) and γ_k (tolerance):

$$\lambda_{ik} := \alpha_k \exp\Big[-\Big(\frac{\beta_k - x_i}{\gamma_k}\Big)^2\Big], \tag{14}$$

where $\theta_k = (\alpha_k, \beta_k, \gamma_k)$ (recall example 1). Gaussian response curves can also reprent practically linear and sigmoidal responses within the observed gradient, but with a more plausible extrapolation.

Assuming no more response models the regression function is according to section 2.1

$$\mathbf{E}(\hat{y}_{ik}|x_i, \Theta) = \frac{\lambda_{ik}}{\lambda_{i+}}, \tag{15}$$

where $\Theta = (\theta_1, \ldots, \theta_k)$, using (14) as a model for absolute abundances of taxon k at site i.

The environmental variable is *a priori* normally distributed (hyperparameters μ and σ). Environmental variables are affected by multiple factors, so the central limit theorem gives support to this assumption.Similarly, the optima β_k are assumed to be normally distributed (hyperparameters a and b). The tolerances γ_k have a gamma distributed prior (hyperparameters c and d), for it is a natural conjugate prior of normal distribution w.r.t tolerance parameter. Finally, the taxon maxima α_k is assumed to be *a priori* uniformly distributed (hyperparameters e and f). Now the total prior distribution is

$$f(\Theta|h) = \prod_{i=1}^{m} \mathrm{No}(\beta_i|a,b)\mathrm{Ga}(\gamma_i|c,d)\mathrm{Un}(\alpha_i|e,f) \prod_{k=0}^{n} \mathrm{No}(x_k|\mu,\sigma), \tag{16}$$

where $h = (a, b, c, d, e, f, \sigma, \mu)$.

3.3 Calculational procedure

We carried out evaluation of the posterior distribution by applying MCMC-based Metropolis-Hastings-algorithm [MRR53], since we used only one response model (14). In total we integrated over the domain of more than 11000 nuisance parameters. In addition, there are non-linear deterministic as well as complex stochastic dependencies between the model parameters. In this particular application domain computation that takes few days is not a bottle neck of the data analysis, since collecting and analyzing the data by an expert biologist takes several months.

Computing a reconstruction for 150 data points is a question of one or two days on a present day PC for the Bummer model. whereas an increase in the number of taxa can lead to more severe performance problems. This is due to fact that in the multinomial response model all species must be fitted simultaneosly.

3.4 Predictive results

We developed prediction models for July mean air temperature by means of six commonly used, 'state-of-the art' calibration techniques [Bir95, TJu93, Bir98, Ter95], and the previous Bayesian model, Bum [TMK00], and then compared the outcomes to our model (Bummer) with local (taxon specific) and global tolerance, i.e. $\gamma_i \equiv \gamma_j$ for all $1 \leq i \leq j \leq m$. The methods used were: (1) simple weighted averaging with an 'inverse' deshrinking regression (WA-I), (2), WA with taxon-tolerance inverse weighting and an 'inverse' deshrinking regression (WAtol-I), (3) simple WA with a 'classical' deshrinking regression (WA-C), (4) WA with taxon-tolerance inverse weighting and a 'classical' deshrinking regression (WAtol-C), (5) partial least squares (PLS), and (6) WA partial least squares (WA-PLS) (for the methods, see [MNa89, Bir95, TJu93, Bir98, Ter95]).

The performance of each method was assessed on the basis of (i) the squared correlations (r^2) between the measured and chironomid-inferred temperature values, (ii) the root mean square error of the prediction (RMSEP), obtained by leave-one-out-cross-validation [TJu93, Bir95], (iii) the linear least-squares error slope parameter (η) (abbreviated LLSESP) between the residuals and the observed values, (iv) the maximum bias along the temperature gradient [TJu93], and (v) the 95% quantile of the leave-one-out cross-validation error. For calculating the maximum bias statistic, we followed the procedure given in ter Braak & Juggins [TJu93] where the temperature gradient was subdivided into 10 equal intervals and the largest absolute value of mean bias for an interval was used as a measure of maximum bias.

In order to estimate whether the data are in accordance with the assumptions of the Bummer model, one can compare the observed data to model predictions in the modern environmental conditions. Figure 4 contrasts Bummer point predictions and 95% credibility intervals ("error bars") to the corresponding observed proportions in Tsuolbmajavri, the key site of our reconstruction. The 95% credibility intervals $[a_{ik}, b_{ik}]$ are defined as

$$P(p_{ik} < a_{ik}| \text{ model assumptions}) = P(p_{ik} > b_{ik}| \text{ model assumptions}) = 0.025. \tag{17}$$

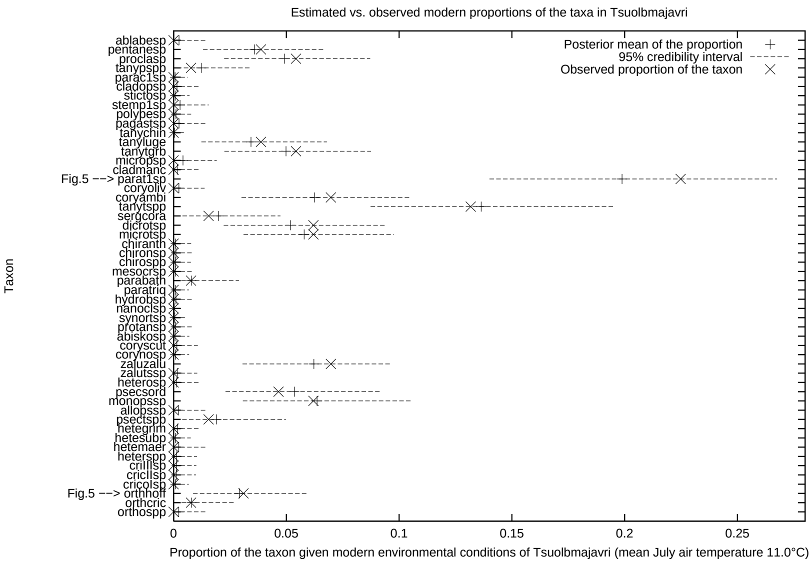

Figure 4: Estimated 95% credibility intervals and posterior means for the taxon proportions in modern environmental conditions of Tsuolbmajavri and the observed proportions.

Consequently,

$$P(p_{ik} \in [a_{ik}, b_{ik}] \Big| \text{ model assumptions }) = 0.95 \tag{18}$$

for all $1 \leq i \leq n$ and $1 \leq k \leq m$. Figure 4 illustrates that the Bummer model fits quite well to the taxon composition of the training data of Tsuolbmajavri, for all the observations are within the 95% credibility interval. Figure 5 gives more details for two taxa. It shows the posterior density functions of the proportions for the taxa *Orthocladius* and *Paratanytarsus sp.* Observe that the posterior density curve for the proportions given the environment of Tsuolbmajavri is unimodal and roughly symmetric for both taxa in Figure 5, altough it is not an inevitable consequence. For example, given two sites with similar environmental conditions, but inconsistent species compositions would lead us to a bimodal posterior density curves for the proportions.

Cross-validation results (Table 2) show clear differences in the prediction performance of the methods for this data-set. Bummer yields the smallest root mean squared error of prediction, 0.80°C, followed by PLS and WA-PLS (0.95°C), then by WAtol-I, WA-I, and finally by WAtol-C and WA-C (1.12 and 1.21°C). Prediction error in terms of the 95% quantiles gives similar results.

To give a visual impression of the performance of some of the techniques, Figure 6 illustrates the predicted and observed mean July air temperatures and corresponding residuals and LLSE-line achieved by WA-PLS, Bum and Bummer.

Bummer seems to be more powerful than Bum and WA-PLS especially at the extremes: it makes smaller prediction errors and has a lower biase in both the warm and the cold end of the gradient.

According to our results with this chironomid training set from northern Fennoscandia, Bummer shows promise as a reconstruction model. Bummer produced the

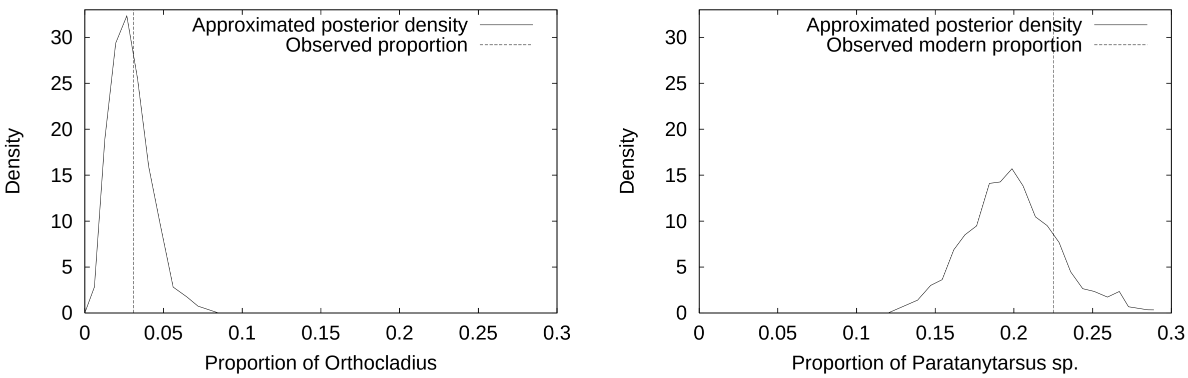

Figure 5: Estimated posterior densities of the proportions for taxa *Orthocladius sensu Hofmann* (abbreviated as *orthhof* in Figure 4 and *Paratanytarsus sp.* (abbreviated as *parat1sp* in Figure 4 in modern environmental conditions of Tsuolbmajavri (observed mean July air temperature = 11 celsius).

smallest root mean squared error of prediction (0.80°C vs. 0.95°C for the best non-Bayesian method), the highest r^2 (0.74 vs. 0.61), the smallest systematic bias or LLSESP (-0.12 vs. -0.14), the lowest maximum bias (0.98°C vs. 1.71°) and the smallest 95% quantile of leave-one-out cross-validation error (1.50°C vs. 1.86°C) among all the models used.

To give an impression of the final product of the reconstruction process Figure 7 illustrates reconstructed mean July temperatures in Finnish Lapland during the last 10000 years.

4 Discussion

We emphasize that Occam's razor should be followed (MDL). Consequently, number of response models M should be small, when there are scantly data available.

In general case evaluation of the posterior distribution of the model described in the section 2 cannot be carried out analytically. Thus, randomized algorithms, e.g., MCMC or EM, should be used in order to estimate desired quantities of the posterior distribution.

Markov chain Monte Carlo (MCMC) refers to a collection of randomized algorithms for closely approximating integrals with respect to awkward and high-dimensional probability distributions [GCS95, GRS96, Gam97]. MCMC is useful, when (i) there are no real-time requirements for the output of computation; (ii) the inherent structure of the problem domain is complex, which implies typically an intractable integral (2); (iii) the modelled process has a probability density which is only known up to a proportionality constant; (iv) and a full information (not only point estimates) concerning the posterior distribution (see, e.g., figure 7) is required. For instance, common deviation and mean measures, e.g., variance, entropy, expectation and median, are not always sufficient information. In addition, previously mentioned measures describe well only a certain restricted set of probability distributions, e.g., unimodal roughly symmetric distributions such as the normal distribution that occurs frequently in statistical analysis due to the central limit

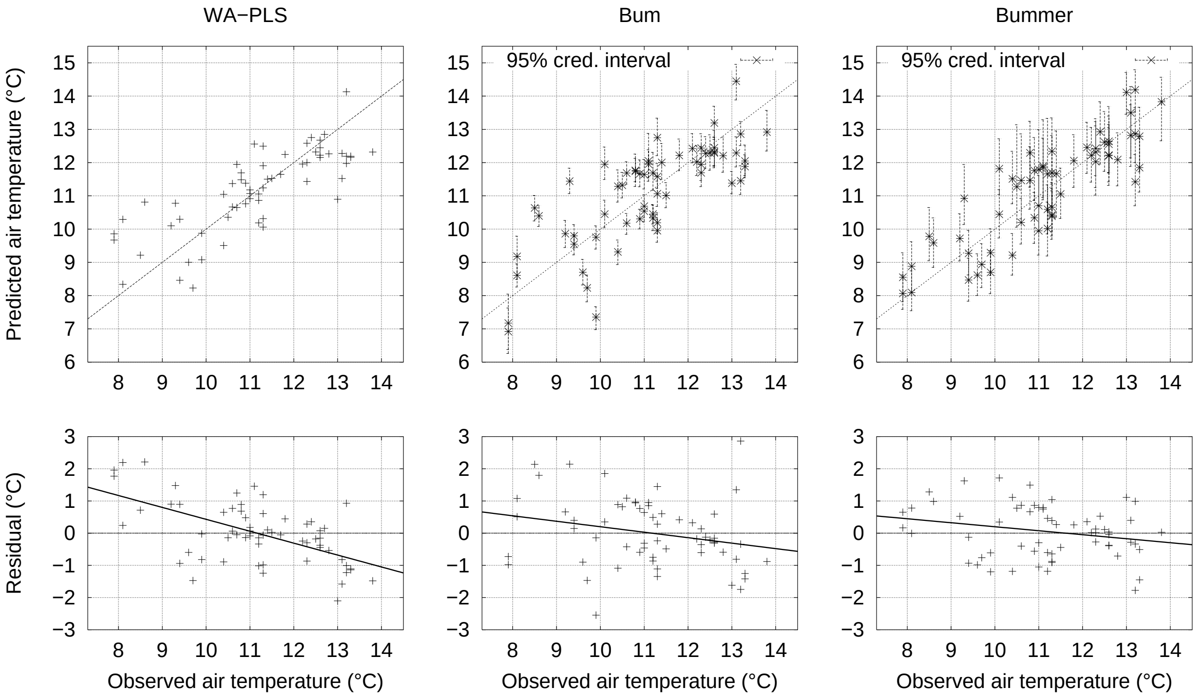

Figure 6: Plots of the relationship between observed and leave-one-out-cross-validated chironomid-inferred mean July air temperatures for the models WA-PLS (2-components), Bum and Bummer, respectively. The lower pictures present residuals and their corresponding linear least-squares lines.

theorem [Wil63].

If the response model parameters are partially or completely different from each other Reversible Jump-MCMC (abbreviated RJMCMC) should be used [Gre95]. RJMCMC is an algorithm for the Bayesian model determination problem among several complex models. RJMCMC jumps between parameter subspaces of differing dimensionality satisfying required ergodicity properties.

The time complexity for our regression model scheme with respect to number of classes m, observations n, response models M in the MCMC-setting is as follows. Within one iteration we need $\mathcal{O}(m^2M)$ calculation primitives for one observation. The time complexity is linear linear with respect to the number of observations, so we are constrained to carry out $\mathcal{O}(nm^2M)$ within on iteration. Consequently, the total time complexity is $\mathcal{O}(qnm^2M)$.

5 Conclusions

We have introduced a Bayesian model for regression problems. The model is based on the counting data. The model takes into consideration observations as a whole

Method	RMSEP (°C)	r^2	LLSESP	Maximum bias (°C)	95% CV-error (°C)
Inverse calibration techniques					
WA-C	1.21	0.54	-0.14	2.13	2.29
WAtol-C	1.12	0.57	-0.32	1.71	2.29
WA-I	1.03	0.53	-0.44	2.19	2.19
WAtol-I	1.00	0.56	-0.42	1.97	2.00
PLS (1 component)	0.95	0.61	-0.37	1.81	1.86
WA-PLS (2 components)	0.95	0.61	-0.37	2.21	1.99
Bayesian techniques					
Bum	1.05	0.60	-0.17	1.80	2.14
Bummer	**0.80**	**0.74**	**-0.12**	**0.98**	**1.50**

Table 2: Cross-validated prediction performance statistics of nine different reconstruction models for mean July air temperature: root mean square error of prediction (RMSEP), coefficient of determination (r^2), linear least-squares error slope parameter between residuals and observed values (LLSESP), maximum bias, and 95% quantile of error (95% CV-error). The data-set consists of 62 lakes from Finnish Lapland; the length of the temperature gradient is 5.9°C.

and allows variability betewen observations. Further, the model takes into account latent variables by modeling probabilities of competing response models for each observation.

We demonstrate the use of the model scheme by applying it to a paleoecological reconstruction problem. We compared our model empirically to several reconstruction techniques on a chironomid data-set. It turned out that at least in this specific application domain our model has a better degree of reliability than the current state-of-the-art methods. In particular, the marked reduction in the systematic bias and maximum bias are of great importance for the future environmental reconstructions.

Further, we would like to propose two related advantages of Bayesian modeling over most other methods: (i) the conceptual structure of the probability model is easily conceivable, and (ii) scientific background knowledge can be embedded in the models in an mathematically elegant, consistent, and natural way.

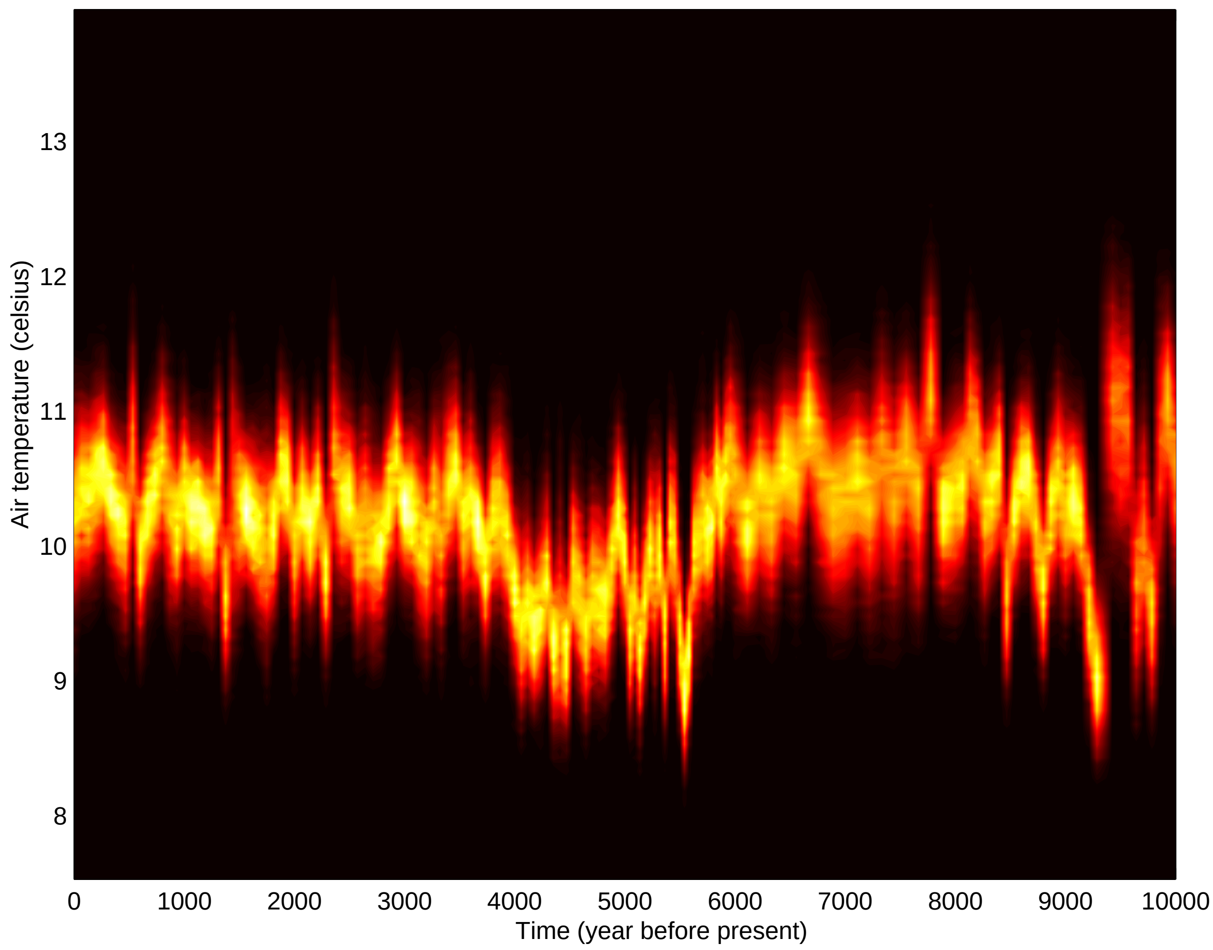

Figure 7: Reconstructed mean July temperatures (Bummer model) in Finnish Lapland during the last 10000 years. Lighter colors indicate higher posterior probability.

References

[Arn90] Arnold, F., *Mathematical Statistics*, Prentice-Hall, Inc., New Jersey, 1990.

[BSm94] Bernardo, J.M., Smith, A.M., *Bayesian theory*, J. Wiley & Sons, N.Y., 586pp.

[Bir95] Birks, H., *Quantitative Palaeoenvironmental Reconstructions*, Statistical Modelling of Quaternary Science Data, editors Maddy, D., Brew, J., Cambridge, 161-255.

[Bir98] Birks, H., *Numerical tools in quantitative palaeolimnology – progress, potentials and problems*, Journal of Paleolimnology 20:307-332, 1998.

[Cox46] R. T. Cox, *Probability, Frequency and Reasonable Expectation*, American Journal of Physics, 17:1-13, 1946.

[Gam97] Gamerman, D., *Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference*, Chapman&Hall, 1997.

[GCS95] Gelman, A., Carlin, J., Stern, H., Rubin, D., *Bayesian Data Analysis*, Chapman&Hall, 1995.

[GHe97] Geiger, D. & Heckerman, D., *Characterization of the Dirichlet distribution through global and local parameter independence*, Annals of Statistics 25:1344-1369, 1997.

[Gre95] Green, P., *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika, 82:711-732, 1995.

[GRS96] Gilks, W., Richardson, S., Spiegelhalter, D., *Markov Chain Monte Carlo in Practice*, Chapman&Hall, 1996.

[IKi71] Imbrie, J., Kipp, N., *A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core*, The Late Cenozoic Glacial Ages, editori Turekian, K., Yale University Press, New Haven, 1971, 77-181.

[Jay94] Jaynes, E., *Probability Theory: The Logic of Science*, <URL: http://bayes.wustl.edu/etj/prob/book.pdf.tar.gz>.

[MNa89] Martens, H., Naes, T., *Multivariate Calibration*, J. Wiley & Sons, Chichester, 1989.

[MRR53] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. *Equation of state calculations by fast computing machines*, Journal of Chemical Physics, 21:1087-1092, 1953.

[OBK99] Olander, H., Birks, J., Korhola, A., & Blom, T., *An expanded calibration model for inferring lake water and air temperatures from fossil chironomid assemblages in northern Fennoscandia*, The Holocene, 9:279-294, 1999.

[TMK00] Toivonen, H., Mannila, H., Korhola, A., Olander, H., *Applying Bayesian Statistics to Organism-Based Environmental Reconstruction*, Ecological Applications, to appear.

[TJu93] ter Braak, C., Juggins, S., *Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages*, Hydrobiologia 269/270:485-502, 1993.

[Ter95] ter Braak, C., *Non-linear methods for multivariate statistical calibration and their use in palaeecology: a comparison of inverse (k-nearest neighbours, partial least squares and weighted partial least squares) and classical approaches*, Chemometrics Intell. Lab. Syst. 28:165-180, 1995.

[TVM99] Toivonen, H., Vasko, K., Mannila, H., Korhola, A., Olander, H. *Bayesian modeling in paleoenvironmental reconstruction*, ACAI Workshop proceedings on Intelligent Techniques for Spatio-Temporal Data Analysis in Environmental Applications, pp. 76-85.

[VTK00] Vasko, K., Toivonen, H. & Korhola, A., *A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction*, Journal of Paleolimnology (rapid communications), 24:243-250, 2000.

[Wil63] Wilks, S., *Mathematical Statistics*, Wiley (2nd print), New York, 1963.

[Wol96] Wolpert, D. H., *The existence of a priori distinctions between learning algorithms*, Neural Computation, 8(7):1391-1420, 1996.

[WMa96] Wolpert, D. H., Macready, W.G., *The lack of a priori distinctions between learning algorithms*, Neural Computation, 8(7):1341-1390, 1996.