ELSEVIER

Ecological Modelling 160 (2003) 235–248

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Bayesian model-based cluster analysis for predicting macrofaunal communities

Cajo J.F. Ter Braak [a,*], Herbert Hoijtink [b], Wies Akkermans [a], Piet F.M. Verdonschot [c]

[a] *Biometris, Wageningen UR, P.O. Box 100, NL-6700 AC Wageningen, The Netherlands*
[b] *Department of Methodology and Statistics, University of Utrecht, P.O. Box 80140, NL-3508 TC Utrecht, The Netherlands*
[c] *Alterra Green World Research, P.O. Box 47, NL-6700 AA Wageningen, The Netherlands*

Abstract

To predict macrofaunal community composition from environmental data a two-step approach is often followed: (1) the water samples are clustered into groups on the basis of the macrofauna data and (2) the groups are related to the environmental data, e.g. by discriminant analysis. For the cluster analysis in step 1 many hard, seemingly arbitrary choices have to be made that nevertheless influence the solution (similarity measure, clustering strategy, number of clusters). The stability of the solution is often of concern, e.g. in clustering by the TWINSPAN program. In the discriminant analysis of step 2 it can occur that a water sample is misclassified on the basis of the environmental data but on further inspection happens to be a borderline case in the cluster analysis. One would then rather reclassify such a sample and iterate the two steps. Bayesian latent class analysis is a flexible, extendable model-based cluster analysis approach that recently has gained popularity in the statistical literature and that has the potential to address these problems. It allows the macrofauna and environmental data to be modelled and analyzed in a single integrated analysis. An exciting extension is to incorporate in the analysis prior information on the habitat preferences of the macrofauna taxa such as is available in lists of indicator values. The output of the analysis is not a hard assignment of water samples to clusters but a probabilistic (fuzzy) assignment. The number of clusters is determined on the basis of the Bayes factor. A standard feature of the Bayesian method is to make predictions and to assess their uncertainty. We applied this approach to a data set consisting of 70 water samples, 484 macrofauna taxa and four environmental variables for which previously a five cluster solution had been proposed. The standard for Bayesian estimation, the Gibbs sampler, worked fine on a subset with only 12 selected taxa but did not converge on the full set with 484 taxa. This is due to many configurations in which the assignment probabilities are all very close to either 0 or 1. This convergence problem is comparable with the local optima problem in classical cluster optimization algorithms, including the EM algorithm used in Latent Gold, a Windows program for latent class analysis. The convergence problem needs to be solved before the benefits of Bayesian latent class analysis can come to fruition in this application. We discuss possible solutions.

* Corresponding author. Tel.: +31-317-476929; fax: +31-317-418094
E-mail address: c.j.f.terbraak@plant.wag-ur.nl (C.J.F. Ter Braak).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00256-9

Keywords: Community composition; Macrofauna; Latent class analysis; Cluster analysis; Gibbs sampling; Species-environment relationships

1. Introduction

Ecological water quality management aims to contribute to the fascinating diversity of biological communities. Such management requires an understanding of how these communities are related to the environment. This paper takes communities of macrofauna taxa as an important example. To learn about the community–environment relationships, a data-analytical approach is often followed that consists of 2 steps: (1) sites are clustered into groups on the basis of the macrofauna data and (2) the groups are related to the environment data, for example by discriminant analysis. This approach is used in RIVPACS (Moss et al., 1999). The cluster analysis methods commonly used in step 1 are rather heuristic (Van Tongeren, 1995). Many hard, seemingly arbitrary choices have to be made that nevertheless influence the solution (similarity measure, clustering strategy, number of clusters, etc.). The stability of the solution is often of concern in that small changes in the input may have considerable impact on the resulting clusterings. Van Groenewoud (1992), Tausch et al. (1995), Oksanen and Minchin, (1997) discuss stability issues in TWINSPAN (Hill, 1979), a clustering program that is popular among ecologists.

In the traditional approach the environmental data play no part in the cluster analysis of step 1. This may be unfortunate. If the environmental data already show distinct groups as a result of water chemistry processes, it is a shame not to use this information. There is also a statistical reason. If discriminant analysis is used in step 2 of the analysis, it can occur that a site is misclassified on the basis of the environmental data but on further inspection happens to be a borderline case in the cluster analysis. One would then rather reclassify such a site and iterate the 2 steps.

A popular rival method for studying community–environment relationships is ordination instead of cluster analysis (Ter Braak, 1995). By ordination the macrofauna data are reduced to continuous gradients rather than to groups. But groups have a simplicity that helps to communicate the results to water managers. Groups, when well described in a typology, can get meaning—reified, as if they already existed. Examples are the typologies developed by Wright et al. (1984), Moss et al. (1984) and Braukmann, (1984) for running waters and by Kansanen et al. (1984) and Johnson and Wiederholm (1989) for stagnant waters. The cenotypes of Verdonschot (1990) for fresh waters are another example.

For these reasons we want to investigate the potential benefits of a model-based cluster analysis for ecological applications. After an introduction to the theory (Section 2), we apply the new methods to four test data sets (Sections 3 and 4). We encountered a number of problems in our application. We discuss these problems in Section 5 and conclude the paper with possible solutions.

2. Model-based clustering by latent class analysis

2.1. Model definition

The model behind latent class analysis for a fixed number of classes (clusters) is a finite mixture model that can be defined as follows (McLachlan and Peel 2000).

1) There are C classes and each statistical unit—the site in our case—belongs to one and only one of these classes, but it is unknown to which one.
2) The variables—in our case the environmental variables and/or the taxon counts—have probability distributions that differ between classes. Following Gelman et al. (1995), we denote the conditional probability density of the vector-variable $\mathbf{y}$ given class c by $p(\mathbf{y}|c)$.
3) The marginal distribution is a mixture of these distributions with mixing proportions $\{\pi_c, c = 1,\ldots, C\}$, i.e.

$$p(\mathbf{y}) = \sum_c \pi_c p(\mathbf{y}|c). \quad (1)$$

The mixing proportions must sum to unity. The likelihood is the product of $p(\mathbf{y})$ over all N statistical units.

The class memberships of the units are unknown. But having observed the data vector $\mathbf{y}$ at a particular site, the model allows one to calculate the class membership probability $p(c|\mathbf{y})$ that the site belongs to a particular class (Eq. (6) or Eq. (A3) of Appendix A). The output of the analysis is thus not a hard assignment of the units to classes but a probabilistic (fuzzy) assignment.

For our case, the mixture model is specialized as follows. The J environmental variables are assumed to be quantitative and to follow within each class a multivariate normal distribution with unknown mean vector $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}$, i.e.

$$\mathbf{y}_{\text{env}}|c \sim \text{Normal}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}). \quad (2)$$

For simplicity we will assume in this paper a common but otherwise unspecified covariance matrix. For other possible specifications of the covariance matrix see Banfield and Raftery (1993), Celeux and Govaert (1995) and Bensmail et al. (1997). Often environmental data are non-normal. However, after a data-transformation such as taking logarithms, the assumption of normality within classes is not unreasonable. In the future, more advanced models may be considered.

The counts of the $k = 1\ldots, K$ taxa are assumed to follow independent Poisson distributions within each class, i.e.

$$\mathbf{y}_k|c \sim \text{Poisson}(\lambda_{ck}) \quad (3)$$

with $\mathbf{y}_k|c$ and λ_{ck} the count and Poisson mean of the k-th taxon in the c-th class, respectively. We notice that this assumption may be unrealistic as counts are often overdispersed in hydrobiology. However, the Poisson distribution is a good starting point and may be applicable if the observed counts are transformed prior to the analysis, as we will do in the test data sets.

Many authors have previously mentioned the joint analysis of normal and non-normal variables in a single analysis, but to our knowledge there are no real ecological applications of this joint model. For the joint analysis of $\mathbf{y} = (\mathbf{y}_{\text{env}}, \mathbf{y}_1, \ldots, \mathbf{y}_K)$ we assume in addition that, the taxon counts are independent of the environment variables within each class.

The model so formulated is a non-normal mixture model with a number of independence assumptions.

2.2. Role of taxa and environment: symmetric or asymmetric

As specified in the previous subsection, the role of environmental variables and taxa is symmetric: both are used for the clustering, and the classes simply determine the means of both the environmental variables and the taxon counts. In this sense, they are both response variables (Fig. 1a). However, in the usual two-step approach their role is often asymmetric: the taxon counts are used for the clustering whereas the environmental data are not. If discriminant analysis is used in the second step to predict the clusters from the environmental variables, the environmental variables have the role of predictors (Fig. 1b). In this subsection we show that the symmetric model allows an asymmetric interpretation and integrates the 2 steps of

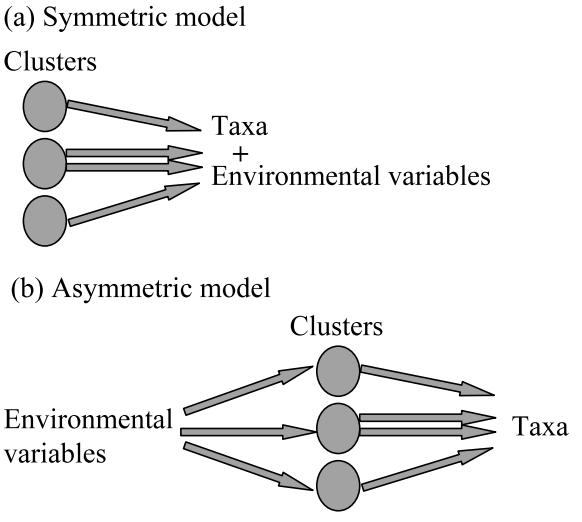

Fig. 1. Schematic representation of the role of taxa and environment in latent class analysis, with arrows indicating supposed causal pathways. (a) Symmetric model, (b) Asymmetric model.

the two-step approach into one integrated analysis.

From now onwards, we turn to a notation that is usual in regression analysis by using the letters y and x for response variable and predictor variable, respectively. Both y and x can be vector-valued. The taxon counts will take the roles of response variables and are thus denoted by the letter y. The environmental variables will take the role of predictors and are thus denoted by the letter x. This notation stresses the possible asymmetry.

The symmetric model of the previous section is a joint model for x and y. It can be written as

$$p(x,y) = \sum_c \pi_c p(x,y|c) = \sum_c \pi_c p(x|c)p(y|c) \qquad (4)$$

because of the conditional independence assumption of x and y. An asymmetric latent class model would be obtained by modelling $p(y|x)$ as follows

$$p(y|x) = \sum_c p(c|x)p(y|c).. \qquad (5)$$

In this model, the mixing proportions π_c of the standard latent class model are replaced by unit-specific probabilities $p(c|x)$. It is called a latent class model with covariates x (Dayton and Macready, 1988; Heinen, 1996; Vermunt and Magidson, 2000) and yields simultaneously both latent classes and an estimate of $p(c|x)$. With the latter estimate, classes can be predicted from the environment. In this sense the asymmetric model integrates the two-step approach of cluster analysis followed by discriminant analysis.

In the asymmetric model the relation between x and y is channeled through the C latent classes. Another way of viewing the model is that the covariate values x of a statistical unit determine the prior probability that a unit belongs to each particular class.

We now show that the symmetric model allows an asymmetric interpretation as well. In the symmetric model we have, from Bayes theorem, a relation between $p(c|x)$ and $p(x|c)$, namely

$$p(c|x) = \pi_c p(x|c) \Big/ \sum_c \pi_c p(x|c) = \pi_c p(x|c)/p(x). \qquad (6)$$

From this an expression for $p(x|c)$ is obtained. By inserting this expression in the right-hand side of Eq. (4) we obtain

$$p(x,y) = p(x)\sum_c p(c|x)p(y|c). \qquad (7)$$

Comparison with the asymmetric model in Eq. (5) shows that the symmetric model is equal to the asymmetric model multiplied by the marginal distribution of x. The symmetric model thus implies a conditional model that has the same structure as the asymmetic model.

We conclude this subsection by comparing the symmetric model with the asymmetric model. The asymmetric model is a regression-type model in that it conditions on the environmental variables x. The environmental values are considered as given fixed values; the environmental variables are not modelled at all. In contrast, the symmetric model does model the environmental variables x in addition to the taxon counts y. The symmetric model can thus use the distribution of x, in particular its possible multimodality, to help form the latent classes, whereas the asymmetric model cannot.

In the asymmetric model the researcher can explicitly choose the functional form of $p(c|x)$, whereas in the symmetric model the functional form follows from the joint specification. In the particular case of multivariate normal distributions for $p(x|c)$ with a common covariance matrix, the model for $p(c|x)$ is well-known to be a multinomial logistic-linear model (Efron, 1975) which is a natural starting point in the asymmetric model. The issues in choosing between the symmetric and the asymmetric model for normal environmental data with a common covariance matrix are thus expected to be similar to those in choosing between linear discriminant analysis and logistic discriminant analysis, methods for known class memberships. If the distributional assumptions hold true, linear discriminant analysis is the most efficient (Efron, 1975) and so will be the analysis based on the symmetric model, even more so because the multimodality in x will help define the class memberships. This symmetric model can thus be viewed as integrating a latent class analysis of the taxa followed by a linear discriminant analysis of classes with respect to the environmental data into one integrated analysis.

For these reasons, we will use the symmetric model from now onwards and exploit its asymmetric properties when desired.

We conclude by remarking that the asymmetric model can be made non-linear in a flexible way by using splines and additive models for $p(c|x)$ (Hastie and Tibshirani, 1990). In the symmetric model $p(c|x)$ becomes a quadratic multinomial model if the covariance matrices are allowed to differ among classes.

2.3. Estimation: maximum likelihood or bayesian

By far the most commonly used approach to fitting mixture models is to maximize the likelihood using the EM algorithm (McLachlan and Peel, 2000) and latent class analysis is no exception (Vermunt and Magidson, 2000). More recently fully Bayesian approaches have been developed using Markov Chain Monte Carlo (MCMC) methods (McLachlan and Peel, 2000) of which the Gibbs sampler has been the most popular for the fixed number classes case (e.g. Bensmail et al., 1997; Hoijtink, 1998). The case with a variable number of classes requires more advanced MCMC methods (reversible jump in Richardson and Green, 1997 and birth–death MCMC in Stephens, 2000) but this case has been attempted so far only for one- and two-dimensional problems, and will not be considered in this paper.

The EM algorithm and the Gibbs sampler have much in common (Gelman et al., 1995). Both use the concept of data-augmentation: adding to the data missing data, which in our case concerns the unknown class memberships of the statistical units. Both methods are iterative and require calculation of the class membership probabilities of site given the then current estimates of the model parameters. In the EM algorithm the missing data must be imputed by expected values, i.e. the class membership probabilities, whereas in the Gibbs sampler the class memberships are randomly drawn using these probabilities. The newly imputed values are then used to update the model parameters, by maximizing the likelihood in EM, and by sampling from it in the Gibbs sampler. The EM algorithm converges to a point estimate of the parameters, whereas the Gibbs sampler converges to the posterior distribution of the parameters. The output of EM is one set of parameter values, whereas the Gibbs sampler results in n sets of sampled parameter values with n the number of saved iterations. These sampled parameter values together represent the posterior distribution, which can then be numerically summarized by means and 95% confidence intervals or by other statistics of interest.

It is instructive at this point to relate these methods to more traditional clustering algorithms. First, note that most non-hierarchical methods also use an iterative optimization scheme. Typically, units are reallocated among clusters using predefined move types so as to optimize the clustering. In the same vein, the Gibbs sampler considers different allocations by drawing memberships that are likely, namely using the then current membership probabilities. Second, note that membership probabilities can be considered as the fuzzy memberships coefficients in a fuzzy c-means clustering (Bezdek, 1974). Hathaway (1986) has shown that the EM algorithm can indeed by interpreted as a fuzzy clustering algorithm. There is an important distinction, however. In fuzzy clustering the fuzzy memberships are free parameters that are optimized, whereas in mixture modelling, hence in EM the membership probabilities are not free; they are functions of the parameters of the component distributions and the mixing proportions (Eq. (6) or Eq. (A3)). These parameters are the ones that are optimized in mixture modelling. As in c-means clustering there is no standard method for parameter initialization and the methods can get trapped in local maxima. For normal mixtures, greedy EM is an attempt to overcome some of these difficulties (Vlassis and Likas, 2002).

In this paper we use the user-friendly computer program Latent Gold (Vermunt and Magidson, 2000) for the maximum likelihood approach. In Latent Gold the EM algorithm is complemented by the Newton–Raphson method in the final stage of the maximization to speed up convergence and to calculate standard errors. For the Bayesian approach the Gibbs sampler was implemented in a special purpose FORTRAN program. Details are given in Appendix A. In the Bayesian approach

prior distributions must be specified for all parameters of the model. In this paper we will use proper priors that are believed to be fairly non-informative (Appendix A). Our program allows user-defined initial values. Latent Gold uses a random initialization scheme.

We wanted to invest in the Bayesian approach for the following reasons. First, it is more flexible. Parameters do not need to be equal or unequal; they can be 'a bit unequal' by assuming a prior distribution for them that hold them together (Gelman et al., 1995). Second, problems in the maximum likelihood approach with near zero λ parameters for the taxa (when some taxa are absent in a cluster) can be avoided by defining proper prior distributions. Third, the Bayesian approach is extendable in many ways. Details of the field sampling could be incorporated. We can use prior information, for example prior information on the habitat preferences of the taxa, such as available in published lists of indicator values. The use of prior information would allow the methods to accumulate knowledge rather than being a one-off exercise. Fourth, in the Bayes factor the Bayesian approach has a standard way of comparing one model to another in which model fit is nicely balanced against model complexity. We will use Bayes factors to determine the number of clusters. Fifth, there are ways for model checking and, last but not least, using the model for prediction is straightforward. In the Bayesian approach the uncertainty of predictions can be assessed in a natural way by integrating out all relevant sources of uncertainty.

2.4. Number of clusters

The Bayes factor is a standard way of comparing one model to another. In the Bayes factor, model fit is balanced against model complexity (Kass and Raftery, 1995). We will use the Bayes factor to determine the optimal number of clusters. The Bayes factor is the ratio of the marginal likelihood of the data under one model to that under a second model. The essential ingredient is thus the marginal likelihood,

$$p(D|\text{model}) = \int p(D|\varphi)p(\varphi)\mathrm{d}\varphi \tag{8}$$

with $p(D|\varphi)$ the likelihood of the data for given parameters φ, and $p(\varphi)$ the prior density of the parameters of the model. In words, the marginal likelihood is the likelihood integrated over the prior density of the parameters. In high dimensional problems like ours, this integral cannot be calculated analytically but can be estimated from the Gibbs sampler output. We use the fourth estimator in Newton and Raftery (1994) and Kass and Raftery (1995). Essentially, it is the harmonic mean of the likelihood of the sampled parameter values, but modified so as to avoid problems with occasional extremely small likelihood values. The modification involves a notional fraction δ of samples taken from the prior. We took $\delta = 0.05$. The estimator is iterative and was programmed in C. We will report the marginal likelihood in terms of the information statistic $-2\log p(D|\text{model})$.

In the maximum likelihood approach the Bayes factor or the marginal likelihood is not available. Fortunately, an approximation of $-2\log p(D|\text{model})$ is available which is the Bayesian Information criterion (BIC; Kass and Raftery, 1995; McLachlan and Peel, 2000), defined as

$$\text{BIC} = -2\log p(D|\varphi_{\text{ML}}) + p\log(N) \tag{9}$$

with φ_{ML} the maximum likelihood estimator of the parameters, p the number of parameters of the model and N the number of statistical units. BIC explicitly trades the complexity of the model (p) against the fit of the model, as in related information criteria of which Akaike's information criterion (AIC) is perhaps the best known. In BIC the penalty per parameter is $\log(N)$ whereas it is 2 in AIC. The problem with AIC is that it does not account for the uncertainty about the parameter values. In our specification of the latent class model

$$p = J(J+1)/2 - 1 + (J+K+1)C. \tag{10}$$

The BIC is among the standard output in Latent Gold (Vermunt and Magidson, 2000). McLachlan and Peel (2000) presented an empirical comparison of AIC, BIC and related information criteria to select the number of clusters using Normal mixture

models. AIC led to a higher number of clusters than BIC. If the Normal mixture model was fitted to non-normal clusters, both AIC and BIC tended to overestimate the number of clusters.

3. Test data sets

To test the feasibility and utility of latent class analysis for ecological applications a subset of 70 sites was used from the EKOO database of Verdonschot (1990) consisting of all samples from ponds and small lakes in the province of Overijssel, The Netherlands. From the available environmental data we selected four quantitative environmental variables, namely width of the water body, acidity (pH), total phosphate and vegetation cover. As in the original analysis (Verdonschot, 1990), the following data transformations were applied. Width and total phosphate were transformed by taking logarithms, and the vegetation cover by applying the logistic transformation, so as to better comply with the assumptions of a normal mixture distribution. To avoid problems with zeroes, half the minimum non-zero value was added before taking logarithms. Counts of individuals of macrofauna taxa tend to show large overdispersion. For this reason, the counts of macrofauna taxa were transformed to Preston classes (Preston, 1962), which is $\log_2(\text{count}+1)$ truncated to an integer value.

In the original analysis by Verdonschot (1990), the 70 sites were classified on the basis of the taxa and environment data into five cenotypes, which were coded and summarized as follows:

- P1: temporary acidified moorland pools (9 sites)
- P2: permanent acid moorland pools (15 sites)
- P3: slightly acid moorland pools (11 sites)
- P7: large mesotrophic deep stagnant waters (7 sites)
- P8: medium-sized, eutrophic stagnant waters (28 sites)

An ordination biplot (Fig. 2) of the environmental data based on a linear discriminant analysis with respect to the five cenotypes shows that the cenotypes P1, P2 and P3 are well separated from

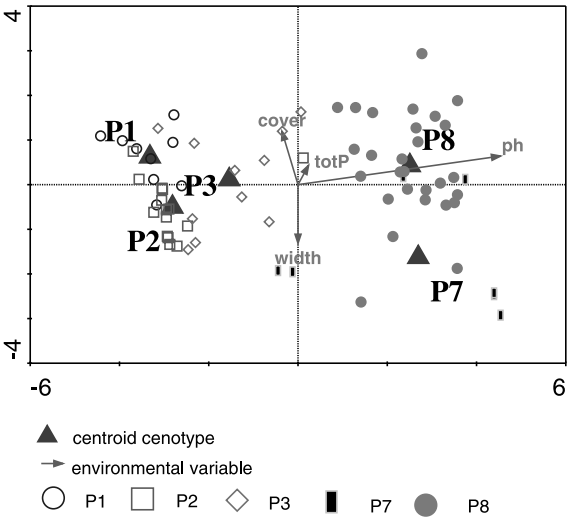

Fig. 2. Biplot of the environmental data of the real test data sets based on linear discriminant analysis with respect to the cenotypes P1, P2, P3, P7 and P8. Distances between the cenotype centroids in the plot approximate Mahalanobis distances in the data. Together with the environmental arrows, centroids and site points approximate cenotype means and individual data values. Environmental variables have been rescaled to unit-within cenotype variance (width = width of the water body, pH = acidity; totP = total phoshate, cover = vegetation cover).

the cenotypes P7 and P8 along the horizontal axis. The horizontal axis mainly represents differences in acidity. Along the vertical axis there is a more subtle subdivision of these two major groups in terms of width, vegetation cover and total phosphate.

The set of 70 sites contained in total 484 macrofauna taxa (large real data set). In this data set a site contained on average 55 taxa; 88.7% of the values in the taxon data was 0. The maximum count after transformation was 12. For a better understanding of the Gibbs sampler, a second smaller test set was created in which only 12 of the 484 taxa were retained (small real data set). The taxa were selected manually. For each coenotype, 2 to 3 taxa were selected that had high relative frequency within the coenotype and near zero frequency outside the coenotype. The resulting set showed a clear block structure against the original classification. This set is thus designed to show five classes on the basis of real data of a

small number of taxa. On average each site contained 2.8 taxa; 77% of all values was 0 and the maximum value was 9.

In addition, two simulated data sets were derived from the large and small real data sets. The simulations used the taxon and environment means per cenotype and the within-cenotype covariance matrix of the environmental variables. There are thus five real classes in these simulated data sets and the data followed all the assumptions of our latent class model (multivariate normality of the environmental variables and independent Poisson counts for the taxa). In the small simulated data set, each site contained on average 3.5 species; 71% of all values was 0 and the maximum value was 10.

4. Results

Fig. 3, Fig. 4 and Fig. 5 summarize the results of latent class analyses with varying number of clusters. The maximum likelihood analyses resulted in the curves labeled ML (minus twice the maximized loglikelihood) and BIC. The Gibbs samples of the posterior in the Bayesian analyses are summarized by the curves labeled BF, minG and maxG, which are the marginal likelihood and minimum and maximum of the sampled likelihood, respectively, all converted to information statistics as in the case of ML.

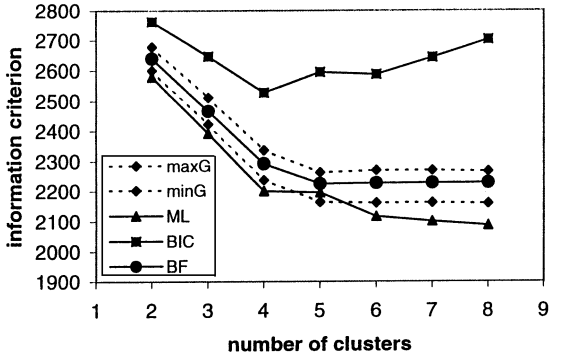

Fig. 3. Small simulated data set: information criteria in latent class analysis plotted against the number of clusters. The curves resulting from the maximum likelihood approach are minus twice the maximized loglikelihood (ML; triangles) and BIC (rectangles). The curves resulting from the Bayesian approach are the marginal likelihood (BF; solid circles), minimum and maximum of the sampled loglikelihood (minG and MaxG; diamonds in dashed lines), all converted to information statistics as in the case of ML. The Bayesian results are based on 10 000 Gibbs samples obtained by taking a 1/100 systematic sample from 1 million Gibbs iterations after a burn-in period of 100 000 iterations.

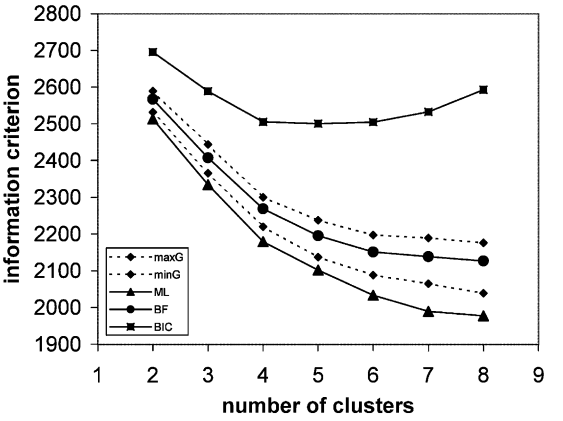

Fig. 4. Small real data set: information criteria in latent class analysis plotted against the number of clusters. For explanation see Fig. 3.

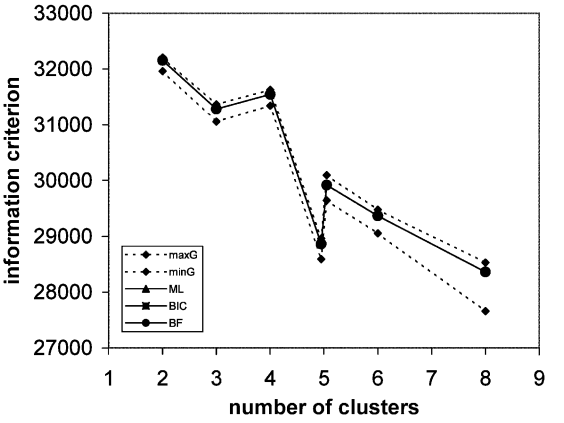

Fig. 5. Large real data set: information criteria in Bayesian latent class analysis plotted against the number of clusters. The results are based on 1000 Gibbs samples obtained by taking a 1/100 systematic sample from 100 000 Gibbs iterations after a burn-in period of 10 000 iterations, except for eight clusters, for which the results are based on 900 Gibbs samples obtained by taking a 1/10 000 systematic sample from 9 million Gibbs iterations. For explanation see Fig. 3.

For the small simulated data set (Fig. 3), the BIC-curve reaches a minimum at four clusters, whereas the minimum of the BF-curve is at five clusters. The Bayesian analysis thus recovers the true number of clusters in these data. The clusters found correspond with the original cenotypes P1–P8. In the maximum likelihood solution with four clusters that is indicated as best by BIC, the clusters P2 and P3 are fused. Fig. 3 shows a problem in the maximum likelihood analyses; at five clusters, the maximized likelihood is lower than the maximum of likelihoods in the Gibbs sample. Apparently, the Latent Gold program found a local maximum, rather than the global maximum. By default the Latent Gold program starts from 10 random configurations. In 10 reanalyses with 1000 restarts each, we found six other (local) maxima. The maximum and minimum of these 10 maximized likelihoods slightly exceeded the range of sampled likelihood indicated in Fig. 3 at both sides, thus resolving the anomaly in Fig. 3. The minimum BIC found for five clusters (2545) remained somewhat larger than the BIC at four clusters.

For the small real data set (Fig. 4), BIC has a minimum at five clusters with marginally larger values at four and six clusters, whereas the BF-curve decreases over the entire range of numbers of clusters tried. The maximum likelihood approach thus recovered the intended number of clusters in these data, whereas the Bayesian approach suggested a much higher number. We looked in detail at the Bayesian solution with six clusters. This solution is essentially the original cenotype classification with one extra cluster comprising three sites. The three sites had high values for one of the three taxa typical for cluster P8 (and not for the other two) and for one of the two taxa typical for P3. The environmental data analyzed alone led to two clusters as judged on the basis of BIC. These two clusters in essence consisted of the unions of P1, P2 and P3 and of P7 and P8.

The large data sets with 484 taxa could not be analyzed with the Latent Gold program. With 70 sites, the Latent Gold program could analyze up to about 60 taxa. Fig. 5 shows the Bayesian results for the large real data set. There are local minima in the BF-curve at three and five clusters, whereas the overall minimum is at eight clusters. For five clusters Fig. 5 actually shows results of two different runs of the Gibbs sampler, one run that was initialized from an uninformative all-equal configuration as were all previous runs (Appendix A) and another that was initialized from the original cenotype classification. The ranges of sampled likelihoods did not even overlap between these two runs. Apparently, the Gibbs sampler did not converge for these data. Reanalyses with other starting configurations and other random seeds led to essentially different solutions. For example, solutions in which P1 and P2 or P2 and P3 were merged and small groups are split off P7 or P8. In terms of the marginal likelihood the original classification was best. In the different solutions the class membership probabilities $\{q_{ci}\}$ of equation Eq. (A3) of Appendix A were all very close to 0 or 1 (being within 10^{-6} from either 0 or 1). In 100 000 Gibbs iterations any change in the cluster memberships was rarely observed. The consequence is that the Gibbs sampler did not move between rival cluster configurations. Even with two clusters, three different solutions were obtained by starting from four different initial values. The three solutions were small variations on the classification consisting of the group of P1-, P2- and P3-sites and the group of P7- and P8-sites. In 100 000 iterations the Gibbs sampler did not move between these solutions. In these runs all class membership probabilities were again very close to either 0 or 1.

The same convergence problem of the Gibbs sampler was observed in the large simulated data set, but not in the small data sets. In the small data sets the different runs led to nearly identi cal BF-values and to solutions that differed by less than 0.01 in the posterior class membership probabilities. In the individual Gibbs samples at least some class membership probabilities $\{q_{ci}\}$ of equation Eq. (A3) of Appendix A were not close to 0 or 1 (being more than 0.1 from either 0 or 1), and consequently class memberships of most sites varied considerably across Gibbs samples.

5. Discussion

Ecological water management can benefit from exciting new statistical methodology for delineating clusters in ecological data. Such methodology exists in the form of latent class analysis, in particular Bayesian latent class analysis. We have shown that latent class analysis is a natural, extendable framework for ecological cluster analysis problems. The new method can use community and environmental data jointly to create the clusters. It also allows predictive usage, for example the prediction of cluster membership and community composition from environmental data. In this sense the new method integrates the traditional two-step approach of cluster analysis of community data and discriminant analysis to predict communities from environmental data into a single integrated analysis. The methodology comes with an information criterion for choosing the number of clusters.

We have applied this new methodology to two ecological data sets with 12 and 484 taxa, respectively. The small data set and the large data set contained the same four environmental variables. In addition we created two simulated data sets of the same dimension and structure. Using these data we empirically compared the Bayesian approach to latent class analysis with the somewhat older maximum likelihood approach. In the Bayesian approach Gibbs sampling was used to estimate the model. In the maximum likelihood approach the EM algorithm was used. For Gibbs sampling we created our own computer program. For EM we used the commercially available, user-friendly Windows program Latent Gold. In applying this new methodology we experienced a number of problems.

With the small data set, repeated runs of Latent Gold led to different answers, because of the many local maxima in the likelihood. A partial solution to this problem was to increase the number of random initial configurations in the program. However, this was no panacea; for the small simulated data set, Latent Gold failed to recover the true five clusters. The Bayesian Gibbs sampler worked fine for these data. There were no convergence problems or problems with local maxima and the sampler did recover the true cluster configuration. For the small real data, the optimal number of clusters as indicated by the Bayes factor was much higher than the intended number of clusters. Apparently, the five cluster model could not account for the additional detailed structure in these data. The assumption in the model that the counts are independent Poisson draws, may be too strict for practical purposes.

The large data sets could not be analyzed with Latent Gold. In applying the Bayesian Gibbs sampler we detected severe convergence problems. The posterior distributions obtained depended strongly on the initial configuration. In general, the Gibbs sampler may fail to move between different regions of higher probability, when these regions are separated by regions of very low probability. In our particular case this happened because all class membership probabilities were very close to either 0 or 1. This essentially prevented the sampler to consider other cluster configurations.

We considered two criteria, the Bayes factor and BIC, to decide on the number of clusters. Theoretically BIC is an approximation to the Bayes factor (McLachlan and Peel, 2000). We would therefore expect that the gap between the BF-curve and the ML-curve would increase with the number of clusters in the same manner as the gap between the BIC-curve and the ML-curve (Eqs. (9) and (10)). Yet Fig. 3 and Fig. 4 show that the former gap increases far less. We therefore checked the stability of the estimator of the Bayes factor by rerunning the five cluster problem of the small simulated data set using 8 million Gibbs iterations, but this did not show any noticeable differences in BF and the range of sampled likelihoods with our previous results. We conclude that the BF-estimator is very stable. It is a little suspect, however, that the range of sampled likelihoods did not increase much with the number of Gibbs iterations. We therefore cannot rule out that the estimator is biased, but we find it unlikely that the bias would be cluster number dependent.

Another way to compare the BF and BIC is in terms of number of degrees of freedom. BF does not explicitly use the notion of number of parameters. But it may be instructive to try and derive

the effective number of parameters that BF uses. This can be done by assuming that the BF-curve relates to the ML-curve in the same way as the BIC-curve. The effective number of degrees of freedom can then be obtained by regressing the difference between the BF- and ML-curve, divided by $\log(N)$, against the number of clusters. The effective numbers of extra degrees of freedom per cluster so obtained from Fig. 3 and Fig. 4 are 3.8 and 3.9 (with standard errors 1.8 and 0.33). We tentatively interpret this result as being that the BF-curve only accounts for the $J=4$ extra environmental parameters and barely, if at all, for the $K=12$ taxon parameters. The explanations for this may be that the λ parameters that are close to zero do not really contribute to BF, whereas the number of clearly 'non-zero' λ parameters (about 12 in these data) does not depend on the number of clusters, at least in the range from two to five clusters. This is due to the strong block structure in the data. The same procedure applied to Fig. 5 of the large real data set leads to 18 effective degrees of freedom per additional cluster. BIC uses a penalty based on 489 degrees of freedom. The conclusion remains that the Bayes factor, or at least our estimator thereof based on the Gibbs sample of the posterior distribution, penalizes each additional extra cluster far less than the BIC.

But even without taxa ($K=0$), we would have expected from Eq. (10) $J+1=5$ instead of $J=4$ extra degrees of freedom per cluster. The one degree of freedom extra makes the distinction between latent class analysis and discriminant analysis. (In discriminant analysis with its known class memberships there are J extra degrees of freedom per cluster). To investigate this distinction further, we compared the variance of the sampled mean parameters in the Gibbs output with the variance to be expected if the cluster memberships and thus the cluster sizes were known. These variances turn out to be fully comparable. Also the variances of the taxon mean parameters λ in the Gibbs sample are close to the means divided by the cluster size, as expected under Poisson sampling in fixed classes. These comparisons were made using the three and five cluster solutions for the small simulated data set. The conclusion is that in our application there is no noticeable excess variance in the Gibbs sample that was expected due to the observed switches in class memberships between Gibbs samples.

With cluster memberships close to either 0 or 1, sites are 'hard' assigned to clusters. This may seem attractive from a user point of view (crisp clusters) but, as we have seen in the large data sets, it can be a nuisance for the Gibbs sampling approach. If there is more than one crisp cluster solution, as in our large data sets, the Gibbs sampler cannot move between these solutions and therefore does no converge. These problems hampered the EM algorithm even more and already cropped up in the small data sets.

We now discuss possible solutions to the numerical problems that we came across.

(1) One could modify the model in such a way that the clusters are less crisp. The crispness is certainly partly due to the Poissonian assumption for the taxon counts. The Poisson distribution implies a known variance within clusters (namely equal to the taxon mean). The clusters thus have fixed heterogeneity, which is undesirable. It may pay to relax this assumption and replace the Poisson distribution by, for example, the negative binomial.

(2) Gibbs sampling is just one of a number of Bayesian sampling algorithms. One could attempt other algorithms that give better mixing of the Markov chain. The algorithms that have been proposed for the model with a variable number of clusters are known to give a better mixing (reversible jump in Richardson and Green, 1997 and birth–death MCMC in Stephens, 2000). It is a challenge to apply these algorithms beyond the 1- and 2-dimensional problems to which they have been applied so far.

(3) One could reduce the number of parameters of the model. With fewer parameters optimization problems are often easier. One appealing way in our application is to use prior knowledge for the taxa. Prior knowledge is available in published lists of indicator values or other lists of species traits. We are currently working on a way to use this knowledge as soft constraints on the taxon parameters.

Acknowledgements

We are grateful to Drs John Birks, Chris Maliepaard and Rebi Nijboer for their comments on the manuscript.

Appendix A: The Gibbs sampler

This appendix contains a more detailed description of our implementation of the Gibbs sampler. First the prior distributions are given, then the process of sampling from the posterior is described. In the sequel, N denotes the total number of sites, and C the number of latent classes. Furthermore, $\mathbf{X}_i$ contains the scores on each of the $j=1,\ldots, J$ environmental variables for sample i, $\mathbf{Y}_i$ contains the number of occurrences of each of the $k=1,\ldots, K$ taxa in sample i, and θ_i denotes the class membership of sample i.

Prior distributions: The parameters in the model are $\boldsymbol{\pi}$, $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_C$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_C$. Their joint prior distribution is assumed to be the product of the individual prior distributions:

$$p(\boldsymbol{\pi},\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_C,\boldsymbol{\Sigma},\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_C)$$

$$= p(\pi)p(\boldsymbol{\mu}_1|\boldsymbol{\Sigma})\cdots p(\boldsymbol{\mu}_C|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma})p(\boldsymbol{\lambda}_1)\cdots p(\boldsymbol{\lambda}_C). \quad \text{(A1)}$$

To avoid problems (Hobert and Casella, 1996) a proper prior distribution is specified for all of these non-random parameters. The priors actually used are not always completely uninformative:

(1) $p(\boldsymbol{\pi}) \sim \text{Dirichlet}\ (v_1,\ldots,v_C)$, where v_c-1 denotes the number of prior observations in class c. Here an uninformative prior has been used; such a prior is obtained with $v_c=1$ for $c=1,\ldots, C$ (Gelman et al., 1995, p. 76).

(2) $p(\boldsymbol{\mu}_c|\boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}_{c\mathbf{0}},\ \boldsymbol{\Sigma}/N_{c0})$, where N_{c0} denotes the number of prior measurements with mean $\boldsymbol{\mu}_{c0}$ on the $\boldsymbol{\Sigma}$ scale. A uninformative prior is obtained with $N_{c0}=0$ for $c=1,\ldots, C$ (Gelman et al. 1995 p. 81). The value that has been used in this study for $\boldsymbol{\mu}_{c0}$ is the overall mean of the data $\mathbf{X}$, for $c=1,\ldots, C$; the prior sample size has been taken to be $N_{c0}=1$ for $c=1,\ldots, C$.

(3) $p(\boldsymbol{\Sigma}) \sim \text{Inv-Wish}(N_0, S_0)$, where N_0 denotes the number of prior measurements, and S_0 denotes the prior Sums of Squares matrix. An improper uninformative prior distribution for $\boldsymbol{\Sigma}$ is obtained using $S_0=\mathbf{0}$ and $N_0=-1$ (Gelman et al., 1995, p. 81). However, since this may lead to an improper posterior, we will use the overall sample covariance matrix for S_0, and $N_0=1$;

(4) $p(\boldsymbol{\lambda}_c)=\prod_{k=1}^{K}(\lambda_{ck})$, where $p(\lambda_{ck}) \sim \text{Gamma}(\alpha_{ck},\beta_{ck})$, which has expectation α_{ck}/β_{ck} and variance $\alpha_{ck}/\beta^2{}_{ck}$. An uninformative prior is obtained using $\alpha_{ck}=1$ and $\beta_{ck}=0$ (Gelman et al., 1995, p. 49); in this study the values $\alpha_{ck}=0.2$ and $\beta_{ck}=1$ have been used.

Sampling from the posterior distribution: The parameters of the model will be sampled from their posterior distribution using an application of the Gibbs sampler as described by Zeger and Karim (1991), in a six step procedure. Iterations of the Gibbs sampler are indicated by the superscript t, where $t=0$ denotes the initial values:

(1) Assign initial values to each of the parameters of the model. The initial values used in the applications described in this paper are: $\pi_c^0=1/C$, for $c=1,\ldots, C$; $\boldsymbol{\mu}_c^0=\bar{\mathbf{X}}$, for $c=1,\ldots, C$, where $\bar{\mathbf{X}}$ denotes the average of $\mathbf{X}$ in the sample; $\boldsymbol{\Sigma}^0=$ the overall sample covariance matrix of $\mathbf{X}$; and $\pi_{ck}^0=1$, for $c=1,\ldots, C$, and $k=1,\ldots, K$.

(2) Sample θ_i for $i=1,\ldots, N$ from

$$p\left(\theta_i|\mathbf{X}_i,\mathbf{Y}_i,\pi^{t-1},\boldsymbol{\mu}_1^{t-1},\ldots,\boldsymbol{\mu}_C^{t-1}\sum^{t-1},\boldsymbol{\lambda}_1^{t-1},\ldots,\boldsymbol{\lambda}_C^{t-1}\right)$$

$$\sim \text{Multinomial}(1,q_{1i},\ldots q_{Ci}), \quad \text{(A2)}$$

where, for $c=1,\ldots,C$,

$$q_{ci}=\frac{p(\mathbf{X}_i|\theta_i=c)p(\mathbf{Y}_i|\theta_i=c)\pi_c^{t-1}}{\sum_{c=1}^{C}p(\mathbf{X}_i|\theta_i=c)p(\mathbf{Y}_i|\theta_i=c)\pi_c^{t-1}}, \quad \text{(A3)}$$

where $p(\mathbf{X}_i|\theta_i=c) \sim N(\boldsymbol{\mu}_c,\boldsymbol{\Sigma})$, and $p(\mathbf{Y}_i|\theta_i=c)=\prod_{k=1}^{K}p(y_{ik}|\theta_i=c)$, with $p(y_{ik}|\theta_i=c) \sim \text{Poisson}(\lambda_{ck})$. These formulae reflect our assumptions that, given θ_i, first, $\mathbf{X}$ and $\mathbf{Y}$ are independent, and second, for $k=1,\ldots, K$ the y_{ik} are independent.

(3) Sample $\boldsymbol{\pi}$ from

$$p(\boldsymbol{\pi}|\boldsymbol{\theta}^{t-1}) \sim \text{Dirichlet}(v_1+N_1^{t-1},\ldots,v_C+N_C^{t-1}), \quad \text{(A4)}$$

where N_c^{t-1} denotes the number of samples with $\theta_i=c$ in iteration $t-1$ of the Gibbs sampler.

(4) Sample $\boldsymbol{\Sigma}$ from (Bensmail et al., 1997)

$$p\left(\sum|\mathbf{X}_1,\ldots,\mathbf{X}_N,\theta^{t-1}\right)$$
$$\sim \text{Inv-Wish}(N_0+N,S), \quad \text{(A5)}$$

where

$$S = S_0 + \sum_{c=1}^{C}\sum_{i\in c}(\mathbf{X}_i-\bar{\mathbf{X}}_c^{t-1})(\mathbf{X}_i-\bar{\mathbf{X}}_c^{t-1})^T$$
$$+\sum_{c=1}^{C}\frac{N_c^{t-1}N_{c0}}{N_c^{t-1}+N_{c0}}(\bar{\mathbf{X}}_c^{t-1}-\boldsymbol{\mu}_{c0})(\bar{\mathbf{X}}_c^{t-1}-\boldsymbol{\mu}_{c0})^T, \quad \text{(A6)}$$

and where $\bar{\mathbf{X}}_c^{t-1}$ denotes the average of $\mathbf{X}$ in class c in iteration t−1.

(5) For $c=1,\ldots,C$ sample $\boldsymbol{\mu}_c$ from

$$p\left(\boldsymbol{\mu}_c|\mathbf{X}_1,\ldots\mathbf{X}_N,\boldsymbol{\theta}^{t-1},\sum{}^{t-1}\right)$$
$$\sim N\left(\mathbf{m}_c,\frac{\sum^{t-1}}{N_{c0}+N_c^{t-1}}\right), \quad \text{(A7)}$$

where

$$\mathbf{m}_c,=\frac{N_{co}}{N_{c0}+N_c^{t-1}}\boldsymbol{\mu}_{c0}+\frac{N_c^{t-1}}{N_{c0}+N_c^{t-1}}\bar{\mathbf{X}}_c^{t-1}. \quad \text{(A8)}$$

(6) For $c=1,\ldots,C$ and $k=1,\ldots,K$ sample λ_{ck} from

$$p(\lambda_{ck}|\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\boldsymbol{\theta}^{t-1})$$
$$\sim \text{Gamma}(\alpha_{ck}+N_c^{t-1}\bar{\mathbf{Y}}_c^{t-1},\beta_{ck}+N_c^{t-1}), \quad \text{(A9)}$$

where $\bar{\mathbf{Y}}_c^{t-1}$ denotes the average of $\mathbf{Y}$ in class c in iteration $t-1$.

References

Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821.

Bensmail, H., Celeux, G., Raftery, A.E., Robert, C.P., 1997. Inference in model-based cluster analysis. Statist. Comput. 7, 1–10.

Bezdek, J.C., 1974. Numerical taxonomy with fuzzy sets. J. Math. Biol. 1, 57–71.

Braukmann, U., 1984. Biologischer Beitrag zu einer allgemeiner regionalen Bachtypologie. Thesis, Justus Liebig Universität, 1–474, Giessen.

Celeux, G., Govaert, G., 1995. Gaussian parsimonious clustering models. Pattern Recog. 28, 781–793.

Dayton, C.M., Macready, G.B., 1988. Concomintant-variable latent-class models. J. Am. Statist. Ass. 83, 173–178.

Efron, B., 1975. The efficiency of logistic regression compared to normal discriminant analysis. J. Am. Statist. Ass. 70, 892–898.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall, London, p. 526.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London, p. 335.

Hathaway, R.J., 1986. Another interpretation of the EM algorithm for mixture distributions. Statist. Prob. Let. 4, 53–56.

Heinen, T., 1996. Latent Class and Discrete Latent Trait Models. Similarities and Differences. Sage, London, p. 220.

Hill, M.O., 1979. TWINSPAN-a FORTRAN Program for detrended correspondence analysis and reciprocal averaging. Cornell University, Ithaca.

Hobert, J.P., Casella, G., 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. J. Am. Statist. Ass. 91, 1461–1473.

Hoijtink, H., 1998. Constrained latent class analysis using the Gibbs sampler and posterior predictive P-values: applications to educational testing. Stat. Sinica 8, 691–711.

Johnson, R.K., Wiederholm, T., 1989. Classification and ordination of profundal macro-invertebrate communities in nutrient poor, oligo-mesohumic lakes in relation to environmental data. Freshwat. Biol. 21, 375–386.

Kansanen, P.K., Aho, J., Paasivirta, L., 1984. Testing the benthic lake type concept based on chironomid associations in some Finnish lakes using multivariate statistical methods. Ann. Zool. Fenn. 21, 55–76.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Statist. Ass. 90, 773–795.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley, New York, p. 419.

Moss, D., Wright, J.F., Furse, M.T., Clarke, R.T., 1999. A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britian. Freshwater Biology 41, 167–181.

Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). J. R. Statist. Soc. Ser. B 56, 3–48.

Oksanen, J., Minchin, P.R., 1997. Instability of ordination results under changes in input data order: explanations and remedies. J. Veg. Sci. 8, 447–454.

Preston, F.W., 1962. The canonical distribution of commonness and rarity. Ecology 43, 185–215.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. Roy. Statist. Soc. Ser. B59, 731–792.

Stephens, M., 2000. Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. Ann. Statist. 28, 40–74.

Tausch, R.J., Charlet, D.A., Weixelman, D.A., Zamudio, D.C., 1995. Patterns of ordination and classification instability resulting from changes in input order. J. Veg. Sci. 6, 897–902.

Ter Braak, C.J.F., 1995. Ordination. In: Jongman, R.H.G., ter Braak, C.J.F., van Tongeren, O.F.R. (Eds.), Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, pp. 91–173.

Van Groenewoud, H., 1992. The robustness of correspondence, detrended correspondence, and TWINSPAN analysis. J. Veg. Sci. 3, 239–246.

Van Tongeren, O.F.R., 1995. Cluster analysis. In: Jongman, R.H.G., ter Braak, C.J.F., van Tongeren, O.F.R. (Eds.), Data analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, pp. 174–212.

Verdonschot, P.F.M., 1990. Ecological characterization of surface waters in the province of Overijssel (The Netherlands). Ph.D. thesis, Wageningen, 255 pp.

Vlassis, N., Likas, A, 2002. A greedy EM algorithm for Gaussian mixture learning. Neural Processing Letters 15, 77–87.

Vermunt, J.K., Magidson, J., 2000. Latent Gold User's Guide. Statistical Innovations, Belmont, p. 186.

Wright, J.F., Moss, D., Armitage, P.D., Furse, M.T., 1984. A preliminary classification of running water sites in Great Britain based on macro-invertebrate species and the prediction of community type using environmental data. Freshwat. Biol. 14, 221–256.

Zeger, S.L., Karim, M.R., 1991. Generalized linear models with random effects: a Gibbs sampling approach. J. Am. Statist. Ass. 86, 79–86.