Overview of Mixture Models for Ordination

Shirley Pledger

30 April, 2014

These files contain R code for fitting mixture models to discrete data, as in Hui *et al.* 2014.

The data are assumed to occur as a matrix with n sites (rows) and p species (columns). The data are either presence/absence (binary, 0/1) or counts of the numbers of each species at each site. Ordination of the sites into a 2D scatterplot is the main objective. There are assumed to be no covariates, making this unconstrained ordination.

There are three basic probability distributions used as building blocks for the models, Bernoulli for presence/absence (binary) data, and either Poisson or negative binomial for count data. With the negative binomial, the dispersion parameter is assumed to be species-specific, to allow for the different amounts of spatial clustering over different species. The Poisson model assumes spatial independence, and the Bernoulli model allows for spatial clustering by simply recording presence, regardless of the actual number of species j present at site i.

For each probability distribution, the same suite of models is used. Each model has a linear predictor, on the logit scale for binary data, and on the log scale for count data.

If Y_{ij} is the random variable for the observation in row (site) i and column (species) j, we write $E(Y_{ij}) = \mu_{ij}$. Then either logit(μ_{ij}) or log(μ_{ij}) is modelled by a linear predictor.

Generalized Linear Models

Firstly, four generalized linear models are fitted, with linear predictors as follows.

Generalized linear models

Model	Linear Predictor for log(μ_{ij}) or logit(μ_{ij})	
Null	θ	constant
A	α_i	row effects only
B	β_j	column effects only
AB	$\alpha_i + \beta_j$	additive row and column effects, no association, assume $\sum_{i=1}^{n} \alpha_i = 0$.

Mixture Models

Next, three different mixture models are fitted. By clustering the columns (species) into two groups, two centroids in n dimensions are obtained. The $n \times 2$ matrix with centroids in the columns may also be viewed the other way, as n points (the sites) in two dimensions.

The EM algorithm is used to fit the mixture models, as in Pledger and Arnold (2014). This provides not only the parameter estimates needed for the plot of sites, but also information on the clustering of species. If we specify C clusters of species, κ_c is the prior probability that any species is in cluster c ($c = 1, \ldots, C$), and x_{jc} is the posterior probability that species j is in cluster c, given the data in column j ($j = 1, \ldots, p$, $c = 1, \ldots, C$). If species j is in cluster c, the linear predictor for logit(μ_{ij}) or log(μ_{ij}) is modelled as follows.

Mixture models when $j \in c$

Model	Linear Predictor for log(μ_{ij}) or logit(μ_{ij})	
C	γ_{ic}	An $n \times C$ matrix, with $C = 2$ giving a 2D plot of sites.
BC	$\beta_j + \gamma_{ic}$	If $C = 2$, matrix (γ_{ic}) gives a 2D plot of sites after allowing for common versus rare species. Assume $\sum_{j=1}^{p} \beta_j = 0$.
ABC	$\alpha_i + \beta_j + \delta_{ic}$	Assume $\sum_{j=1}^{p} \beta_j = 0$ and for each i, $\sum_{c=1}^{C} \delta_{ic} = 0$. The matrix (δ_{ic}) gives a site ordination after allowing for both species and site effects.

Ordination from Model C is driven by species commonness, site richness and species composition (turnover).

Model BC has allowed for common versus rare species, which is similar to standardising the data before ordination. Here the ordination is determined by rich versus poor sites and by species turnover, so two sites are similar if they have similar richness and similar species composition.

Model ABC has also allowed for rich versus poor sites, which is similar to a double standardisation. In this case the ordination is driven solely by species composition, so two sites are different if they have different species composition.

Model ABC is obtained from Model BC by decomposing the matrix (γ_{ic}) into the vector α_i of row sums of (γ_{ic}) and a matrix of deviations (δ_{ic}) found by subtracting the row means from (γ_{ic}):

$$\delta_{ic} = \gamma_{ic} - \frac{1}{C} \sum_{c=1}^{C} \gamma_{ic}.$$

This implies that δ_{ic} has row sums all zero. Hence if $C = 2$, any row of δ_{ic} plotted in 2D is a point on the line $x + y = 0$. There is a loss of one dimension. In order to obtain a 2D plot of sites, we fit the ABC model with $C = 3$. The points in 3D are all on the plane $x + y + z = 0$, and a rotation places them into an $x - y$ plane.

Model ABC3 is the same as Model BC3; it just has a different parameterisation.

Model Comparison

Since all the models are likelihood-based, they may be compared by likelihood ratio tests (which may be non-standard), or by information criteria (AIC = Akaike's Information Criterion, AICc = modification of AIC for small samples, BIC = Bayes' Information Criterion, etc.)

Comparison of models with differing numbers of clusters indicates if it is reasonable to represent the data in two dimensions. This is similar to the stress measures used in non-metric multidimensional scaling (nMDS).

Comparing model BC2 (two clusters) with ABC3 (three clusters, equivalent to BC3) shows if there is any need to allow for different sites richness separately. If model BC3 is preferred, we have a choice of two ordinations of sites: a 3D ordination driven by site differences which include both site richness and species composition (using (γ_{ic}) from BC3), or a 2D ordination driven only by species composition differences (using (δ_{ic}) from ABC3).

Running the Models

The zip file `MIXORD.zip` contains files for fitting the models. For each type of probability distribution (Bernoulli, Poisson or negative binomial), open the "run" file and paste code from that file into R. There is an associated "fun" file containing the functions.

The Bernoulli models use files `runBERN.R`, `funBERN.R` and `TikusData.csv`.

The Poisson models use files `runPOIS.R`, `funPOIS.R`, `spiderAbund.csv` and `spiderin.R`.

The negative binomial models use files `runNB.R`, `funNB.R`, `spiderAbund.csv` and `spiderin.R`.

References

Hui, Francis K.C., Sara Taskinen, Shirley Pledger, Scott D. Foster and David I. Warton (2014). "Model-Based Approaches to Unconstrained Ordination". Submitted to *Methods in Ecology and Evolution.*

Pledger, Shirley and Arnold, Richard (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics and Data Analysis*, **71**, 241–261.