Computational Statistics and Data Analysis 71 (2014) 241–261

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection

Shirley Pledger *, Richard Arnold

School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, Wellington, New Zealand

ARTICLE INFO

Article history:
Received 21 June 2012
Received in revised form 15 May 2013
Accepted 15 May 2013
Available online 25 May 2013

Keywords:
Association analysis
Biclustering
Biplots
Cluster analysis
Correspondence analysis
Data visualisation
Dimension reduction
Finite mixture
Fuzzy clustering
Multidimensional scaling

ABSTRACT

Matrices of binary or count data are modelled under a unified statistical framework using finite mixtures to group the rows and/or columns. These likelihood-based one-mode and two-mode fuzzy clusterings provide maximum likelihood estimation of parameters and the options of using likelihood ratio tests or information criteria for model comparison. Geometric developments focused on pattern detection give likelihood-based analogues of various techniques in multivariate analysis, including multidimensional scaling, association analysis, ordination, correspondence analysis, and the construction of biplots. Illustrative examples demonstrate the effectiveness of these visualisations for identifying patterns of ecological significance (e.g. abrupt versus slow species turnover).

1. Introduction

There is widespread use of matrices of binary and count data across many disciplines; for example incidence and abundance matrices in ecological communities where the rows are species and the columns are samples, or binary item response analysis with respondents in the rows and questions in the columns. Major objectives include summarising the multivariate data in fewer dimensions, giving an overview of relationships and patterns among rows and columns. Existing analysis techniques for these matrices include one-mode or two-mode cluster analysis (unsupervised learning), multidimensional scaling, association analysis, correspondence analysis and ordination (see e.g. Manly, 2005; Quinn and Keough, 2002). Many of these methods are mathematical, using either distance-based algorithms or matrix decomposition and eigenvalues, with dimension reduction achieved by transformations which load most of the information from the data onto the first few eigenvalues. Often there is no underlying probability model and hence statistical inference is unavailable, although (i) resampling methods (see e.g. Manly, 2007; Gotelli and Graves, 1996) provide some statistical tests, and (ii) clustering using finite mixtures gives a likelihood-based foundation for cluster analysis (McLachlan and Basford, 1988; Everitt et al., 2001; McLachlan and Peel, 2000; Böhning et al., 2007).

In this article we propose a group of likelihood-based models for a data matrix with binary or count data, using basic Bernoulli or Poisson building blocks. Standard generalised linear models are too restrictive; as well as overdispersion there may be redundancy, with groups of associated rows (or columns) having similar incidence or abundance patterns over the columns (or rows). Mixture models provide a flexible way of allowing for both heterogeneity and redundancy, giving likelihood-based statistical inference.

* Corresponding author. Tel.: +64 4 4636788; fax: +64 4 4635045.
E-mail address: shirley.pledger@vuw.ac.nz (S. Pledger).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.013

The use of finite mixtures for a one-mode fuzzy cluster analysis, with the variables in the columns being used to cluster the entities in the rows, is well documented and frequently used (McLachlan, 1982; McLachlan and Basford, 1988; Everitt et al., 2001; Wu et al., 2008; McLachlan and Peel, 2000). There are developments in many aspects of finite mixture modelling, including estimating the number of components (see e.g. Schlattmann, 2003), in speeding up the algorithm and avoiding being trapped in a local maximum (see e.g. O'Hagan et al., 2012), and in finding asymptotic results (van der Geer, 2003). Papers applying one-mode clustering of binary data include (Pledger, 2000; Arnold et al., 2010; Dunstan et al., 2011).

Biclustering (or two-mode or block clustering) is the simultaneous clustering of rows and columns into row clusters and column clusters. Recent developments in cluster analysis, including biclustering, have employed various distance measures and coefficients of similarity, association or correlation, together with substantial computing power (Wu et al., 2008). Biclustering using mixtures has been proposed for binary data (Pledger, 2000; Govaert and Nadif, 2003; Arnold et al., 2010), and for count data (Nadif and Govaert, 2005; Govaert and Nadif, 2005, 2010). We outline a unified suite of models, some new and some previously published, and show that new geometric insights provide likelihood-based analogues of multidimensional scaling, association analysis, correspondence analysis, pattern detection, ordination and biplots.

Section 2 has definitions of the models, while model fitting, statistical analysis and graphical displays are in Section 3, including comparisons with traditional multivariate techniques. Real examples are given in Section 4, and we conclude with a discussion in Section 5.

2. The models

In this section we give the data, assumptions and likelihoods (Section 2.1), followed by a suite of basic models (Section 2.2). A modification providing pattern-detection models is next (Section 2.3), followed by a summary (Section 2.4). Some of the models have been previously published, as noted in the text. We combine existing models with some new ones, putting them in a single framework with a unified notation. We emphasise the interpretation of the pattern-detection models.

2.1. Data, assumptions and likelihoods

The data are an $n \times p$ matrix $\mathbf{Y}$ of binary or count values, with value y_{ij} a realisation of a random variable Y_{ij} ($i = 1, \ldots, n, j = 1, \ldots, p$). We assume $Y_{ij} \sim$ Bernoulli(θ_{ij}) for binary data and $Y_{ij} \sim$ Poisson(θ_{ij}) for count data. Row sums, column sums and the overall data sum are denoted by y_{i+}, y_{+j} and y_{++} respectively.

We temporarily assume independence of Y_{ij} conditional on θ_{ij}. However, for some models this is changed to local independence. Failure of independence is addressed in the discussion (Section 5).

All our models for the $n \times p$ data matrix are based on full likelihood and fuzzy (probabilistic) clustering rather than the classification likelihood (Banfield and Raftery, 1993). If ϕ is the vector of all the parameters in the model, the saturated model has likelihood:

$$L(\phi \mid Y) = \prod_{i=1}^{n} \prod_{j=1}^{p} f\left(\theta_{ij}; y_{ij}\right), \tag{1}$$

where $f\left(\theta_{ij}; y_{ij}\right) = \theta_{ij}^{y_{ij}} \left(1 - \theta_{ij}\right)^{1-y_{ij}}$ for binary data, and $f\left(\theta_{ij}; y_{ij}\right) = e^{-\theta_{ij}} \theta_{ij}^{y_{ij}} / y_{ij}!$ for count data.

2.2. Basic models

In our suite of basic models, the rows and/or columns of the data matrix may be modelled as (i) homogeneous (forming a single group, denoted by $rR1$ or $cC1$), (ii) all different (denoted by rn or cp), or (iii) coming from finite mixture groups with membership unknown (rR or cC). Cases (i) and (ii) are just generalised linear models (GLMs), in which logit θ_{ij} or $\log \theta_{ij}$ (the transformed expected values in cell i, j for binary and count data respectively) are modelled by a linear predictor. The GLMs and extensions to models using mixtures are specified in Table 1. There are identifiability constraints on the parameters, e.g. $\sum_{i=1}^{n} \alpha_i = 0$. The subscript r is used if row i is in row group r ($i \in r$), and c if $j \in c$.

Details of two of the basic mixture-based models follow. The other models are similar or simpler.

2.2.1. The row-clustered model {rR, cp}:

The rows are assumed to come from a finite mixture with R components or row groups, yielding a model-based fuzzy clustering of the rows of $\mathbf{Y}$, which are p-vectors. The p columns are kept separate. McLachlan (1982) proposed this model for any distribution of Y_{ij}, and it has been widely applied using p-dimensional normal, Poisson or Bernoulli distributions (see, e.g. McLachlan and Peel, 2000; Pledger, 2000; Böhning et al., 2007; Wu et al., 2008; Arnold et al., 2010; Dunstan et al., 2011). For p-variate Bernoulli or Poisson distributions respectively the expected value of Y_{ij}, θ_{ij}, is given by

$$\text{logit}\, \theta_{rj} \quad \text{or} \quad \log \theta_{rj} = \mu + \alpha_r + \beta_j + \gamma_{rj}, \tag{2}$$

if $i \in r$. The parameters are $\mu, \alpha_r, \beta_j, \gamma_{rj}$ and π_r (the proportion of rows in each row group). Constraints, e.g. $\sum \alpha_r = 0, \sum \beta_j = 0, \forall r, \sum_j \gamma_{rj} = 0$ and $\forall j, \sum_r \gamma_{rj} = 0$, imply there are Rp independent parameters θ_{rj}. A further $R-1$ independent

Table 1
A summary of the suite of models. GLM = generalised linear model, FMR and FMC = finite mixture clustering of rows and columns respectively. Models with + in the labels are additive, with no interaction term. Cases A, B and D are indicated. The number of independent parameters includes the parameters in both the linear predictor and the group membership probabilities.

Model label	Row groups	Col. groups	Model equation for logit θ_{ij} or log θ_{ij}	No. parameters	Model type
Basic models:					
$\{rR1,\ cC1\}$	1	1	μ	1	GLM, Null model
$\{rR1,\ cC\}$	1	C	$\mu+\beta_c$	$2C-1$	FMC
$\{rR1,\ cp\}$	1	p	$\mu+\beta_j$	p	GLM, column effects
$\{rR,\ cC1\}$	R	1	$\mu+\alpha_r$	$2R-1$	FMR
$\{rR+cC\}$	R	C	$\mu+\alpha_r+\beta_c$	$2R+2C-3$	FMR + FMC
$\{rR,\ cC\}$	R	C	$\mu+\alpha_r+\beta_c+\gamma_{rc}$	$RC+R+C-2$	FMR, FMC
$\{rR+cp\}$	R	p	$\mu+\alpha_r+\beta_j$	$2R+p-2$	FMR
$\{rR,\ cp\}$	R	p	$\mu+\alpha_r+\beta_j+\gamma_{rj}$	$Rp+R-1$	FMR
$\{rn,\ cC1\}$	n	1	$\mu+\alpha_i$	n	GLM, row effects
$\{rn+cC\}$	n	C	$\mu+\alpha_i+\beta_c$	$n+2C-2$	FMC
$\{rn,\ cC\}$	n	C	$\mu+\alpha_i+\beta_c+\gamma_{ic}$	$nC+C-1$	FMC
$\{rn+cp\}$	n	p	$\mu+\alpha_i+\beta_j$	$n+p-1$	GLM, no interaction
$\{rn,\ cp\}$	n	p	$\mu+\alpha_i+\beta_j+\gamma_{ij}$	np	GLM, saturated
Pattern-detection models: (Number of row and column groups refers to the γ term only)					
$\{rn+cp\}$	1	1	$\mu+\alpha_i+\beta_j$	$n+p-1$	PD-null, as above
$\{rR,\ cC,\ PD\}$	R	C	$\mu+\alpha_i+\beta_j+\gamma_{rc}$	$n+p+RC-2$	FMR, FMC, Case D
$\{rR,\ cp, PD\}$	R	p	$\mu+\alpha_i+\beta_j+\gamma_{rj}$	$Rp+n-1$	FMR, Case A
$\{rn,\ cC, PD\}$	n	C	$\mu+\alpha_i+\beta_j+\gamma_{ic}$	$nC+p-1$	FMC, Case B
$\{rn,\ cp\}$	n	p	$\mu+\alpha_i+\beta_j+\gamma_{ij}$	np	Saturated, as above

parameters π_r are also estimated, with constraint $\sum\pi_r = 1$. Choosing $R \ll n$ ensures that the number of independent parameters in this model, $Rp+R-1$, is less than np, the number of bits of information.

The likelihood for this model sums over all possible allocations of rows to row groups:

$$L(\phi \mid \mathbf{Y}) = \sum_{r_1=1}^{R}\cdots\sum_{r_n=1}^{R}\pi_{r_1}\ldots\pi_{r_n}\prod_{i=1}^{n}\prod_{j=1}^{p} f\left(\theta_{r_i,j}; y_{ij}\right). \tag{3}$$

Assuming independence among rows and, conditional on the rows, independence over the columns, this simplifies to

$$L(\phi \mid \mathbf{Y}) = \prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r\prod_{j=1}^{p} f\left(\theta_{rj}; y_{ij}\right)\right]. \tag{4}$$

The row-based conditional independence assumption causes any apparent correlations between columns to be attributed to similarities among the rows, which we model by clustering the rows. When fitting a model we specify the number of components in the mixture. A model with (say) three row clusters ($R=3$), is labelled $\{rR3,\ cp\}$. An additive version of this model, $\{rR3+cp\}$, omits the γ_{rj} term in Eq. (2). This entails a plot of logit or log θ_{rj} versus j having parallel traces for the three row groups, with the rows being clustered by their row sums, high, medium and low. The interactive model (which includes the γ_{rj} term) allows for different slopes and possible crossings (turnover patterns).

Model $\{rn,\ cC\}$ is similar, with clustering of columns but not rows. It assumes column-based conditional independence. Models $\{rR,\ cC1\}$ and $\{rR1,\ cC\}$ are simplifications of the above models, with (respectively) all columns alike or all rows alike.

2.2.2. The biclustered model $\{rR,\ cC\}$

The model with two-mode clustering has simultaneous finite-mixture groupings of both the rows and the columns (into R, C groups respectively). This fuzzy model-based biclustering by mixtures was proposed for binary data in Pledger (2000) and Arnold et al. (2010), and for more general data types in Govaert and Nadif (2003) and Nadif and Govaert (2005).

If $i \in r$ and $j \in c$, the (interactive) model for binary or count data respectively is

$$\text{logit}\,\theta_{rc} \quad \text{or} \quad \log\theta_{rc} = \mu+\alpha_r+\beta_c+\gamma_{rc}. \tag{5}$$

Proportions are π_r for row groups, and κ_c for column groups, with $\sum\pi_r = \sum\kappa_c = 1$. Constraints are also imposed on α_i, β_j and γ_{rc}, giving $RC+(R-1)+(C-1)$ independent parameters, of types θ, π and κ respectively. With unknown cluster membership, the likelihood is obtained by summing over all possible partitions of the rows into R clusters and the columns into C clusters.

$$L(\phi \mid \mathbf{Y}) = \sum_{c_1=1}^{C}\cdots\sum_{c_p=1}^{C}\kappa_{c_1}\ldots\kappa_{c_p}\sum_{r_1=1}^{R}\cdots\sum_{r_n=1}^{R}\pi_{r_1}\ldots\pi_{r_n}\prod_{i=1}^{n}\prod_{j=1}^{p} f\left(\theta_{r_i,c_j}; y_{ij}\right). \tag{6}$$

Assuming row-based conditional independence gives some simplification to the likelihood:

$$L(\phi \mid \mathbf{Y}) = \sum_{c_1=1}^{C} \cdots \sum_{c_p=1}^{C} \kappa_{c_1} \cdots \kappa_{c_p} \times \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} f\left(\theta_{r,c_j}; y_{ij}\right) \right], \tag{7}$$

which is summed only over the possible column cluster partitions. Alternatively, a column-independence assumption gives a similar likelihood, summing only over the possible row clusters.

The biclustered model gives a broad overview of groupings in the data if we specify only a few clusters in the rows and columns (e.g. $\{rR4,\ cC3\}$). The rows of the estimated $R \times C$ matrix $\mathbf{\Theta} = (\theta_{rc})$ give profiles of the centroids of the row groups over the column groups, and vice versa for the columns of $\mathbf{\Theta}$. Omission of the γ_{rc} term in Eq. (5) specifies an additive model $\{rR + cC\}$, with row-group profiles parallel and column-group profiles parallel on the logit or log scale. This prohibits turnover and enforces grouping driven by the row sums and column sums.

2.3. Pattern detection (PD) models

The interactive finite mixture models in Section 2.2 are capable of detecting different types of patterns, which are exhibited in the estimated matrix $\mathbf{\Gamma} = (\gamma_{rj})$ or (γ_{ic}) or (γ_{rc}). However, if the row (or column) sums of $\mathbf{Y}$ have a high variance, the row (column) clusters are determined by these sums, splitting the rows (columns) into those with high and low sums. To prevent this feature from dominating the clustering, we firstly adjust for the differing sums by using the additive generalised linear model with linear predictor $\mu + \alpha_i + \beta_j$ as the new null model, called the PD-null model ($\{rn + cp\}$, see Table 1). With count data, the PD-null model is the no-association model used in contingency table analysis. We set sum-to-zero constraints on the row parameters α_i and the column parameters β_j. For the count data model we specify nonlinear constraints on $\mathbf{\Gamma}$ which take advantage of the multiplicative structure of the Poisson likelihood. These constraints, given in detail below, ensure that the maximum likelihood estimates (MLE) of μ, α_i and β_j under the PD models are the same as under the PD-null model. This means that, for the count data PD models, $\mathbf{\Gamma}$ is directly interpretable as a residual pattern matrix added to the simpler PD-null structure. The binary data PD models do not have such neat properties and we impose simple sum-to-zero constraints.

There are three PD models interpolated between the PD-null model $\{rn + cp\}$ which has no association patterns and the saturated model $\{rn,\ cp\}$ which has all the patterns in the data (Table 1).

2.3.1. The row-clustered PD model: Case A

After allowing for different row sums and column sums, the linear predictor for model $\{rR,\ cp, PD\}$ (labelled Case A) has an $R \times p$ pattern matrix, $\mathbf{\Gamma}^{(A)} = (\gamma_{rj})$. If $i \in r$,

$$\text{logit}\,\theta_{ij} \quad \text{or} \quad \log\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{rj}.$$

With constraints $\sum \pi_r = 1$ and $\sum \alpha_i = \sum \beta_j = 0$ and one constraint on each row and column of $\mathbf{\Gamma}^{(A)}$, this model has $(R-1)(p-1) + (R-1)$ more parameters than the PD-null model. Possible constraints on $\mathbf{\Gamma}^{(A)}$ are

$$\sum_j \gamma_{rj} = 0 \quad \forall r \quad \text{and} \quad \sum_r \gamma_{rj} = 0 \quad \forall j, \tag{8}$$

with any one of these constraints being determined from the remaining $R - p - 1$. However for the count data model we impose

$$\sum_j e^{\beta_j + \gamma_{rj}} = \sum_j e^{\beta_j} \quad \forall r \quad \text{and} \quad \sum_r e^{\alpha_i + \sum_r \widehat{z}_{ir}\gamma_{rj}} = \sum_r e^{\alpha_r} \quad \forall j, \tag{9}$$

where $\widehat{z}_{ir}$ is the posterior probability that row i is in row group r. An explicit formula for $\widehat{z}_{ir}$ is given in Eq. (13) of Appendix A. These parameter constraints are somewhat unusual, being data dependent. For the binary data model, we may choose constraints such as (8) or (9) for a full maximisation of the likelihood, or we may choose to match all μ, α_i and β_j estimates with estimates from the PD-null model. This latter is over-constrained, and while the conditional maximum likelihood will not attain the full maximum, the γ_{rj} estimates are directly interpretable as the pattern of deviations from the PD-null model. With the many data sets we tried, the suboptimal estimates were very close to the optimal.

2.3.2. The column-clustered PD model: Case B

When only the columns are clustered we have model $\{rn,\ cC, PD\}$ (labelled Case B), where if $j \in c$

$$\text{logit}\,\theta_{ij} \quad \text{or} \quad \log\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ic},$$

with constraints $\sum \kappa_r = 1$ on group proportions, $\sum \alpha_i = \sum \beta_j = 0$ and one constraint on each row and column of the pattern matrix $\mathbf{\Gamma}^{(B)} = (\gamma_{ic})$.

2.3.3. The biclustered PD model: Case D

The biclustered model {*rR*, *cC*, *PD*} (labelled Case D for "dual" clustering) gives broad, overall patterns using few parameters. This model was introduced by Govaert and Nadif (2010). If $i \in r$ and $j \in c$, the linear predictor is

$$\text{logit}\,\theta_{ij} \quad \text{or} \quad \log\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{rc}.$$

With constraints $\sum \pi_r = \sum \kappa_c = 1$ on the proportions, $\sum \alpha_r = \sum \beta_c = 0$ and one constraint on each row and column of the $R \times C$ pattern matrix $\mathbf{\Gamma}^{(D)} = (\gamma_{rc})$, this model has $(R-1)(C-1) + (R-1) + (C-1)$ more independent parameters than the PD-null model. For count data we show in Section 3.3 and Appendix B that the constraints ensuring that $\mathbf{\Gamma}^{(D)}$ are the residuals from the PD-null model are

$$\sum_j e^{\beta_j + \sum_c \widehat{x}_{jc}\gamma_{rc}} = \sum_j e^{\beta_j} \quad \forall r \quad \text{and} \quad \sum_i e^{\alpha_i + \sum_r \widehat{z}_{ir}\gamma_{rc}} = \sum_i e^{\alpha_i} \quad \forall c, \tag{10}$$

where $\widehat{z}_{ir}$ and $\widehat{x}_{jc}$ are respectively the posterior probabilities that row i is in row group r, and column j is in column group c (see Eq. (16) in Appendix A). For count data the constraints in (10) above ensure the estimates of μ, α_i and β_j exactly match the estimates from the PD-null model. As before (Section 2.3.1) for binary data we may choose to use constraints (10) or sum-to-zero constraints analogous to Eqs. (8) on the rows and columns of $\mathbf{\Gamma}^{(D)}$. Alternatively, a conditional likelihood assuming all values of μ, α_i and β_j match the PD-null model leads to a more interpretable $\mathbf{\Gamma}^{(D)}$ at the cost of a less-than-optimal likelihood. In practice we found the suboptimal estimates were close to optimal.

2.4. Summary

Table 1 summarises the models defined so far. Each model may be used for either binary or count data. The PD models are interpolated between the no-association model ($\mu + \alpha_i + \beta_j$) and the saturated model ($\mu + \alpha_i + \beta_j + \gamma_{ij}$). While the $n \times p$ pattern matrix $\mathbf{\Gamma}^{(S)} = (\gamma_{ij})$ from the saturated model captures all of the association patterns in the data table, our smaller matrices $\mathbf{\Gamma}^{(A)}$, $\mathbf{\Gamma}^{(B)}$ and $\mathbf{\Gamma}^{(D)}$ show some of the patterning—that which is explained by clustering similar rows or columns or both. These groupings in the PD models recognise associations and near-redundancies in the data table and have fewer parameters than the saturated model. A comparison of likelihoods indicates how much of the detailed patterning in the table may be explained by just a few broad groupings.

In the special case of Poisson-based PD models for count data (two-way contingency tables), the row and column groupings are determined by $y_{ij}/y^{\star}_{ij}$ where $y^{\star}_{ij} = y_{i+}y_{+j}/y_{++}$ (the expected value under the PD-null model). On the log scale, a positive estimate of γ_{rc} indicates that rows in group r are (on average) positively associated with columns in group c, while a negative value suggests a negative association between those rows and columns. A similar interpretation holds for binary data, although the Bernoulli distribution does not yield an explicit formula for $y^{\star}_{ij}$.

Of particular interest are the biclustered PD models with $R = C$. These models partition both rows and columns into (say) M*modules*($R = C = M$). With suitable relabelling of row groups and/or column groups, it may be possible to make each module contain a collection of rows and columns which are on average positively associated with each other but negatively associated with rows and columns in a different module. Such a partition clearly separates out different modules, for example distinct groups of species occurring in different habitats, or groups of people with distinct patterns of responses to a questionnaire.

3. Analysis, interpretation and data visualisation

In this section, we summarise existing results on model fitting and model comparison for finite mixture models, and then develop our new reduced-dimension data visualisation methods which provide a means of understanding and interpreting the multivariate data.

3.1. Model fitting and comparisons

All the models are likelihood-based and may be fitted by maximum likelihood, yielding parameter estimates from the optimisation, and their estimated asymptotic standard errors from the observed information matrix. Possible multimodality of the likelihood surface necessitates trying multiple starting points, with simulated annealing (Kirkpatrick et al., 1983; Zhou and Lange, 2010) a good option to avoid being locked into a local maximum.

The mixture models may be fitted by the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997), with the missing data being the group membership of each row and/or column. Indicator random variables for group membership are Z_{ir} for row i in row group r ($i \in r$) and X_{jc} for $j \in c$. Appendix A gives the formulae for the log likelihood under complete knowledge, denoted by ℓ_c, and formulae for estimates in the E and M steps. The EM algorithm also provides estimates of the posterior probabilities of allocations to clusters. $E(Z_{ir}) = \widehat{z}_{ir}$ is the posterior probability that $i \in r$, and $E(X_{jc}) = \widehat{x}_{jc}$ is the posterior probability that $j \in c$. Note that $\forall i$, $\sum_{r=1}^{R} \widehat{z}_{ir} = 1$ and $\forall j$, $\sum_{c=1}^{C} \widehat{x}_{jc} = 1$.

Once the likelihood-based models are fitted, they may be compared by likelihood ratio tests (LRTs), or by an information criterion, for example AIC (Akaike's Information Criterion) or its small-sample modification AICc (Akaike, 1973; Burnham

and Anderson, 2002), BIC (Bayes' Information Criterion, Schwarz, 1978) or ICL (Integrated Classification Likelihood, Biernacki et al., 2000). The formulae for these criteria are

$$\text{AIC} = -2\ell + 2K,$$
$$\text{AICc} = -2\ell + 2K\left(\frac{np}{np-K-1}\right),$$
$$\text{BIC} = -2\ell + K\log(np),$$
$$\text{ICL} = -2\ell_c + K\log(np),$$

where ℓ is the log likelihood, ℓ_c is the log of the complete likelihood, K is the number of independent parameters in the model, np is the sample size (in our case, the number of elements in the $n \times p$ data matrix), and ICL is in the ICL-BIC version (see ICL-BIC in McLachlan and Peel, 2000). Note that BIC has a stronger penalty, and so tends to select fewer clusters, which can be an advantage for interpretation.

A standard LRT may be used to compare models with the same number of clusters. However, there is a failure of some boundary conditions for LRTs if the comparison is between models with different numbers of clusters. In these cases we may use the theory in Self and Liang (1987) or randomisation tests (McLachlan, 1987; Manly, 2007; Gotelli and Graves, 1996) to obtain the distribution of the test statistic under the null hypothesis. Estimation of standard errors using bootstrap methods remains valid under the boundary condition failure.

We used the statistical package **R** 2.11.1 for all our computing (R Development Core Team, 2010).

3.2. Dimension reduction and simple plots

We now show that mixture models provide more than clustering. They also give dimension reduction, as clustering the rows yields a low-dimensional plot of the columns, and vice versa. From this dimension reduction, we show how to construct informative plots showing the main features of the data. These plots give likelihood-based analogues of various plots used in multivariate analysis, including multidimensional scaling and correspondence analysis plots.

The plots arise from three types of mixture models, Case A from row clustering, Case B from column clustering, and Case D (for **D**ual clustering) from biclustering. A test data set which will be used to illustrate the data visualisation is given (with its clusters) in Fig. 1.

In *Case* A, the row-clustered PD model $\{rR,\ cp, PD\}$, the $R \times p$ pattern matrix is $\mathbf{\Gamma}^{(A)} = (\gamma_{rj})$. Two plots arise from this matrix (see Fig. 2).

1. A *profile plot of the row groups* has R line plots using the rows of $\mathbf{\Gamma}^{(A)}$ versus $j = 1, \ldots, p$. These row-group centroids trace the changing patterns of the row groups over the different columns. The columns may be ordered to minimise crossings and clarify any turnover patterns. A value $\gamma_{rj} > 0$ ($\gamma_{rj} < 0$) indicates a positive (negative) association between row group r and column j.
2. A *scatterplot of the columns* in R dimensions uses the columns of $\mathbf{\Gamma}^{(A)}$ as points. With $R = 2$, the two-dimensional scatterplot giving a broad overview of similarities and associations among the p columns. If the plot is approximately linear, this suggests that the columns may be ordered along a single axis to assist in interpretation of results.

In the case of *pattern detection models with count data*, some special geometry emerges which reduces the dimension of the scatterplot, as follows. This is caused by the multiplicative nature of the model with the Poisson distribution building block. In Appendix B we show that the parameters in a model satisfying (9) lead to the same probability model as those satisfying (8), though we prefer (9) due to the simpler interpretation of $\mathbf{\Gamma}^{(A)}$. The model is reparameterised in the original (not logarithmic) scale:

$$\theta_{ij} = \zeta\omega_i\nu_j\psi_{rj} \quad \text{if } i \in r,$$

where $\zeta = \exp\mu$, $\omega_i = \exp\alpha_i$, $\nu_j = \exp\beta_j$ and $\psi_{rj} = \exp\gamma_{rj}$. We adopt the estimates in Appendix A.2 Eqs. (12)–(14). This solution maximises ℓ_c and satisfies constraints (9). Since our estimates of ζ, ω_i and ν_j match those in the PD-null model, the ψ_{rj} parameters may be interpreted directly as modelled associations, being deviations from the no-association model.

Geometrically, for all j, the pattern in column j of $\mathbf{Y}$ is represented by the point $P_j = (\widehat{\psi}_{1j}, \ldots, \widehat{\psi}_{Rj})$ in R dimensions. Property (i) in Appendix A.2 Eq. (15) ensures that all the points P_j lie on the hyperplane $\sum_{r=1}^{R} a_r\psi_{rj} = y_{++}$, where $a_r = \sum_{i=1}^{n}\widehat{z}_{ir}y_{i+}$, the total count in row group r. Since all $\widehat{\psi} \geq 0$, the scatterplot of columns lies on the simplex in the non-negative orthant with rth vertex $A_r = (0, \ldots, 0, \frac{y_{++}}{a_r}, 0, \ldots, 0)^T$, on the rth axis. We label this Simplex A, as it arises from Case A.

Hence using $R = 3$, model $\{rR3,\ cp, PD\}$ for count data gives a 2-D plot on Triangle A in the non-negative octant, with vertices $A_1(\frac{y_{++}}{a_1}, 0, 0)$, $A_2(0, \frac{y_{++}}{a_2}, 0)$ and $A_3(0, 0, \frac{y_{++}}{a_3})$. A column with perfect association with row group r would be shown at the vertex A_r. A column equally associated with all row groups would appear at the centroid $G_A(1, 1, 1)$. It is a weighted centroid, weighted by the total counts a_r in the different row groups. The vector from the origin to the centroid G is given by

$$\vec{OG} = \frac{1}{y_{++}}(a_1\vec{OA}_1 + a_2\vec{OA}_2 + a_3\vec{OA}_3).$$

This centroid is on the A triangle, since $\sum_r a_r = y_{++}$. Vectors from G_A to the vertices A_1, A_2 and A_3 represent directions of the row groups. A general column appears as a point on Triangle A. Close proximity of two points in Triangle A indicates

	a	b	c	d	e	f	g	h	i	j
A	3	4	2	9	4	10	13	2	9	10
B	9	14	9	3	4	6	6	1	10	16
C	2	4	0	8	0	4	11	3	9	6
D	4	3	4	3	5	4	2	10	6	5
E	3	4	6	10	2	15	5	2	7	5
F	6	11	14	2	8	8	7	9	7	10
G	4	12	6	10	2	2	8	3	9	15
H	9	3	8	0	9	0	4	9	2	3

(a) Raw count data.

	h	e	c	a	b	j	i	g	d	f
A	2	4	2	3	4	10	9	13	9	10
B	1	4	9	9	14	16	10	6	3	6
C	3	0	0	2	4	6	9	11	8	4
D	10	5	4	4	3	5	6	2	3	4
E	2	2	6	3	4	5	7	5	10	15
F	9	8	14	6	11	10	7	7	2	8
G	3	2	6	4	12	15	9	8	10	2
H	9	9	8	9	3	3	2	4	0	0

(c) Col-clustered data.

	a	b	c	d	e	f	g	h	i	j
D	4	3	4	3	5	4	2	10	6	5
H	9	3	8	0	9	0	4	9	2	3
F	6	11	14	2	8	8	7	9	7	10
B	9	14	9	3	4	6	6	1	10	16
G	4	12	6	10	2	2	8	3	9	15
C	2	4	0	8	0	4	11	3	9	6
A	3	4	2	9	4	10	13	2	9	10
E	3	4	6	10	2	15	5	2	7	5

(b) Row-clustered data.

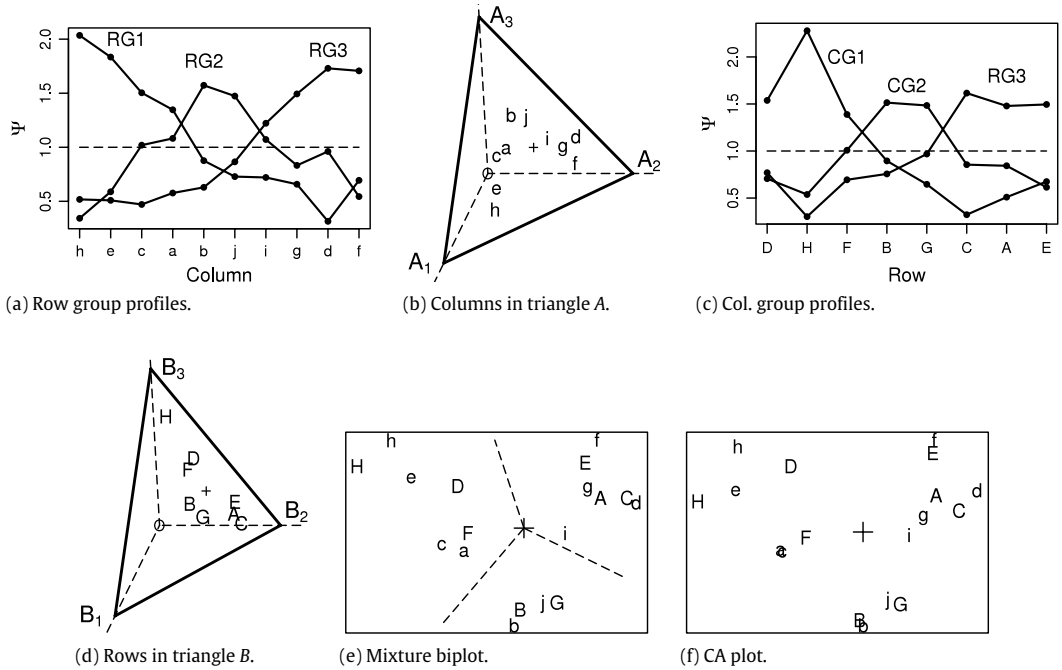
	h	e	c	a	b	j	i	g	d	f
D	10	5	4	4	3	5	6	2	3	4
H	9	9	8	9	3	3	2	4	0	0
F	9	8	14	6	11	10	7	7	2	8
B	1	4	9	9	14	16	10	6	3	6
G	3	2	6	4	12	15	9	8	10	2
C	3	0	0	2	4	6	9	11	8	4
A	2	4	2	3	4	10	9	13	9	10
E	2	2	6	3	4	5	7	5	10	15

(d) Biclustered data.

Fig. 1. Test count data and clusters. (a) The data table, (b) the data with rows clustered, model {$rR3$, cp}, (c) the data with columns clustered, model {rn, $cC3$}, and (d) the data with biclustering model {$rR3$, $cC3$}. With biclustering a diagonal pattern of high counts (above the average of 6.2) is found, while the off-diagonal corner-cell counts average below 6.2.

pattern similarities of those two original data columns. A column near vertex A_r is strongly associated with row group r, while a column near G_A is not specially associated with any row group. Positive (negative) association between row group r and column j is indicated by $\psi_{rj} > 1$ ($\psi_{rj} < 1$). For the test data, the profile plot and Triangle A from row clustering are shown in Fig. 2(a) and (b).

Similarly with count data and Case B, model {rn, cC, PD}, we may draw a profile plot of the column groups over rows $i = 1, \ldots, n$ and a scatterplot of the rows in C dimensions, using posterior probabilities $\widehat{x}_{jc}$ and the $n \times C$ estimated pattern matrix $\mathbf{\Gamma}^{(B)}$. Again the scatterplot dimension is reduced by one, so that points representing the rows of the data matrix are on Simplex B in the non-negative orthant of a C-dimensional space. If $C = 3$, they are on Triangle B in the non-negative octant, with vertices $B_1(\frac{y_{++}}{b_1}, 0, 0)$, $B_2(0, \frac{y_{++}}{b_2}, 0)$ and $B_3(0, 0, \frac{y_{++}}{b_3})$, where $b_c = \sum_{j=1}^{p} \widehat{x}_{jc} y_{+j}$, the total count in column-group c. The vertices represent possible rows fully associated with the three column groups, and point $G_B(1, 1, 1)$ in Triangle B is a weighted centroid representing a possible row equally associated with all the column groups. This profile plot and Triangle B for the test data are shown in Fig. 2(c) and (d).



(a) Row group profiles. (b) Columns in triangle *A*. (c) Col. group profiles.

(d) Rows in triangle *B*. (e) Mixture biplot. (f) CA plot.

Fig. 2. Low-dimensional plots for the test data. Plots (a) and (b) arise from row clustering, with (a) being the profile plot of the pattern-detection matrix and (b) being Triangle *A* from row clustering. Individual columns are shown as points on Triangle *A*, with similar columns occurring close together. Plots (c) and (d) arise from column clustering, where (c) is the profile plot of three column groups over the individual rows and (d) is Triangle *B* from column clustering. Individual rows occur as points on Triangle *B*, with similar rows occurring close together. Plot (e) is the mixture-based biplot from superimposing Triangles *A* and *B*, while (for comparison purposes) plot (f) is the traditional correspondence analysis plot. In (a) and (c) the dashed line shows the expected value of $\psi = 1$, in (b), (c), (e) and (f) the centroid is shown by +, and in (e) the row and column clusters are shown by the dashed lines. In (e) the biplot algorithm has rotated the triangles about their centroids and Triangle *A* has been turned over. The vertices from (b) and (d) are outside the plotting area in (e).

3.3. Biplots from biclustering

If both rows and columns are clustered (model $\{rR,\ cC,\ PD\}$, Case D, biclustering), the $R \times C$ pattern matrix $\mathbf{\Gamma}^{(D)} = (\gamma_{rc})$ provides profiles of the row groups over the different column groups and vice versa, as well as scatterplots of the row groups in C dimensions and the column groups in R dimensions.

Again, special geometry arises with *count data*. Parameters in the untransformed model

$$\theta_{ij} = \zeta \omega_i \nu_j \psi_{rc} \quad \text{if } i \in r \text{ and } j \in c,$$

are estimated using Eqs. (12) and (17) (Appendix A.2), which satisfy constraints (10). In Appendix B it is shown that these constraints do not restrict the likelihood from achieving its maximum. The posterior probabilities for the rows and columns are $\widehat{z}_{ir}$ and $\widehat{x}_{jc}$ respectively. The properties in Eq. (18) of Appendix A show that all row groups r (rows of $\mathbf{\Gamma}^{(D)}$) lie in the hyperplane $\sum_{c=1}^{C} b_c \psi_{rc} = y_{++}$ in a C-dimensional space, where $b_c = \sum_j \widehat{x}_{jc} y_{+j}$. Similarly all column groups c (columns of $\mathbf{\Gamma}^{(D)}$) lie in the hyperplane $\sum_{r=1}^{R} a_r \psi_{rc} = y_{++}$ in an R-dimensional space, where $a_r = \sum_i \widehat{z}_{ir} y_{i+}$. Because all $\widehat{\psi}_{rc} \geq 0$, the points are on simplices in the non-negative orthants of each space. Further, we are able to place the individual columns in the column-group simplex (A), and the rows in the row-group simplex (B), by using the ψ estimators from singly-clustered models together with the posterior probabilities and a_r, b_c from the biclustering. These ψ estimates are

$$\widehat{\psi}_{rj} = \frac{y_{++} \sum_i \widehat{z}_{ir} y_{ij}}{y_{+j} a_r} \quad \text{and} \quad \widehat{\psi}_{ic} = \frac{y_{++} \sum_j \widehat{x}_{jc} y_{ij}}{y_{i+} b_c}. \tag{11}$$

Then using the row clustering, $\forall j$

$$\sum_r a_r \widehat{\psi}_{rj} = \sum_r \frac{y_{++}}{y_{+j}} \sum_i \widehat{z}_{ir} y_{ij} = \frac{y_{++}}{y_{+j}} \sum_i \sum_r \widehat{z}_{ir} y_{ij} = y_{++},$$

and so the point representing column j is on the A simplex defined by the column groups in R-dimensional space. Similarly individual rows are represented by points on the B simplex in C-dimensional space.

By reducing to three dimensions ($R = C = 3$), we may construct a biplot. Triangles *A* and *B* are formed as above, and standard biplot methodology (Gabriel, 1971) is applied to the biclustering matrix $\mathbf{\Psi}^{(D)}$ to superimpose the triangles (possibly with rotation, reflection and rescaling) in order to illustrate associations between row groups and column groups. The

following algorithm uses biclustering to produce a biplot associating row groups with column groups. Biplot methodology is in steps 3–6. Individual rows and columns are carried along with the transformations.

Biplot algorithm for count data:

1. Fit the biclustering three-group model $\{rR3,\ cC3,\ PD\}$. This provides the 3×3 pattern detection matrix $\boldsymbol{\Psi}^{(D)}$ and the posterior probabilities $\widehat{z}_{ir}$ and $\widehat{x}_{jc}$ for allocating rows and columns to their groups.
2. Use Eqs. (11) to compute a $3 \times p$ matrix $\boldsymbol{\Psi}^{(A)}$ and an $n \times 3$ matrix $\boldsymbol{\Psi}^{(B)}$. (Note that these estimates will match values from the singly-clustered models only if the posterior probabilities from one-mode clustering match those in the biclustering. This will not necessarily happen as mixture-based clustering is not hierarchical.)
3. Centre the $\boldsymbol{\Psi}$ matrices by subtracting the centroid $G(1, 1, 1)$: let $\boldsymbol{\Delta}^{(D)} = \boldsymbol{\Psi}^{(D)} - \mathbf{1}_{3\times 3}$, $\boldsymbol{\Delta}^{(A)}_{3\times p} = \boldsymbol{\Psi}^{(A)} - \mathbf{1}_{3\times p}$ and $\boldsymbol{\Delta}^{(B)}_{n\times 3} = \boldsymbol{\Psi}^{(D)} - \mathbf{1}_{n\times 3}$, where $\mathbf{1}$ is an appropriately-sized matrix of ones.
4. Do a singular value decomposition (SVD) of $\boldsymbol{\Delta}^{(D)}$ into $\mathbf{USV}^T$, where $\mathbf{U}$ is orthonormal, $\mathbf{V}$ is orthogonal and $\mathbf{S}$ is a diagonal matrix of non-negative singular values.
5. Separate $\mathbf{USV}^T$ into $\mathbf{GH}^T$, where $\mathbf{G} = \mathbf{US}^{\frac{1}{2}}$ and $\mathbf{H}^T = \mathbf{S}^{\frac{1}{2}}\mathbf{V}$. (We use $\mathbf{S}^{\frac{1}{2}}$ for equal absorption of $\mathbf{S}$ into the row groups and column groups. Alternatives of absorbing $\mathbf{S}$ entirely into $\mathbf{G}$ or entirely into $\mathbf{H}$ favours rows over columns or vice versa in the final biplot.)
6. In the biplot, the first two columns of $\mathbf{G}_{3\times 3}$ give coordinates for the three row groups, and the first two columns of $\mathbf{H}^T_{3\times 3}$ give coordinates for the column groups.
7. Do the same row transformation to the individual data rows, i.e. compute $\mathbf{G}_n = \boldsymbol{\Delta}^{(B)}\,\mathbf{V}\,\mathbf{S}^{-\frac{1}{2}}$ and use the first two columns of $\mathbf{G}_n$ $(n \times 3)$ as the coordinates of the data rows on the biplot.
8. Similarly apply the column group transformation to the individual data columns, calculating $\mathbf{H}^T_p = \mathbf{S}^{-\frac{1}{2}}\,\mathbf{U}^T\,\boldsymbol{\Delta}^{(A)}$ and using the first two columns of $\mathbf{H}_p$ $(p \times 3)$ as coordinates of the data columns on the biplot.

On the biplot, proximity represents associations among rows, row groups, columns and column groups. Usually biplot methodology is used to simultaneously reduce dimensions and draw the plot. Our biplot method differs, in that we have done the dimension reduction first using mixtures, so all that remains is superimposing the triangles. In our method, we can use model selection to assess the validity of the dimension reduction first, after which the biplot shows the association patterns in the data which can be explained by mixture-based biclustering.

Fig. 2(e) shows the biplot for the test data, while (f) shows the standard correspondence analysis plot for comparison. We note the similarity of the plots.

3.4. Analogues of multivariate analyses

Mixture-based fuzzy *clustering* of the rows of the data matrix is already well documented, and compares favourably with traditional clustering methods based on distance metrics (McLachlan and Basford, 1988; Everitt et al., 2001). Fuzzy biclustering using mixtures is also already developed (Pledger, 2000; Arnold et al., 2010; Govaert and Nadif, 2010). Since fuzzy clustering by mixtures is not hierarchical, the user is not locked into an undesirable solution at the early stages as the number of clusters is increased. (Note the shift of column cluster boundaries when row clusters are introduced for the test data, Fig. 1.)

We now specify how our mixture models provide model-based analogues of several other techniques in multivariate analysis.

Allowing for differing row sums and column sums in the PD models is analogous to *standardising* the data before analysis, which is frequently done in multivariate analysis to prevent any one row or column from dominating the analysis.

The row-clustered PD model provides a *profile plot* for the row groups and a scatterplot for the individual columns. By fitting a model with few row groups, we may reduce the column points from n dimensions to a two- or three-dimensional scatterplot; this is *dimension reduction*, with the same objective as in *multidimensional scaling* (MDS). There are various stress measures used in MDS, indicating how much information has been lost in compressing the data into fewer dimensions. We may measure the *stress* of mixture-based dimension reduction by comparing likelihoods of the fitted models as the number of clusters decreases. A choice of the effective dimensionality of the data could use an information criterion or non-standard LRTs.

A 2-*D plot* of the rows is found in Triangle *B*, while Triangle *A* does the same for the columns. The 2-D plot provides an *association analysis*, with proximity in the plot indicating a high association.

With count data a 1-D plot of the columns is provided by $\{rR2,\ cp,\ PD\}$, giving an *ordination* of the columns based on information in the rows. The usefulness of the ordination is evaluated by comparing models with two, three or more row groups. Similarly the columns may be grouped to give an ordination of the rows.

An *indirect gradient analysis* is sometimes used for ecological community data (species by samples) when the samples are assumed to differ but no useful covariates are known or measured. Fitting two row clusters to count data gives a 1-D representation of the columns as an ordering along a single axis or gradient, with the spacings indicating the similarity between columns. Higher dimensions of indirect gradients are obtained using more clusters.

In *correspondence analysis* (CoA), usually applied to count data, there is a simultaneous dimension reduction of the rows and columns, giving rise to a plot (generally in 2-D) showing associations between any two rows, two columns, or a row and a column. The patterns are those remaining after allowing for differing row sums and column sums. Our graphical analogue of CoA arises naturally from the biclustered model $\{rR3,\ cC3,\ PD\}$ as a *biplot*, and gives similar results (Fig. 2). Our *A* and *B* triangle plots are analogous to the "stretched simplex" plots in correspondence analysis geometry (Greenacre and Hastie, 1987).

Similarly, the biclustered PD model applied to binary data provides an analogue of *twinspan* (two-way indicator species analysis, Hill, 1979) used in vegetation analysis.

The most useful of our biclustered models are likely to be those with equal numbers of row and column groups, $R = C = M$ say. If the group labels can be reassigned to put the M highest values of $\boldsymbol{\Psi}^{(D)}$ on the main diagonal (e.g. as a Robinson matrix, Robinson, 1951; Wu et al., 2007), we have detected *M modules*. Each module is a set of rows and columns; there are high associations within modules and lower associations between modules. Using H, M and L to indicate high (positive), medium (near zero) and low (negative) estimated values of γ_{rc}, patterns of (i) hard clustering, and (ii) fuzzy clustering follow:

Row group	(i)			(ii)		
1	H	L	L	H	M	L
2	L	H	L	M	H	M
3	L	L	H	L	M	H

For example, with species by sample data, (i) shows abrupt turnover of three groups of species over three discrete communities, while (ii) shows a gradual turnover.

Further data may be available in the form of *covariates*, either continuous measurements or categories (factors), for the row entities and/or the column entities. These may be incorporated into the linear predictor as for generalised linear models. For example a row-clustered model with covariates was applied by Dunstan et al. (2011) to coral reef fish species with the latitude of each sample as a covariate. For ecological communities, this use of covariates is sometimes called *direct gradient analysis*, which uses known temporal, spatial or environmental gradients (e.g. soil pH). Covariates may also be used with our biclustered models to give an analogue of *canonical correspondence analysis*, but we do not develop those models here.

For all our multivariate analogues, we may monitor the model choice, using information criteria or LRTs to justify our choice of the effective number of dimensions in the data.

4. Real data examples

The following data sets from community ecology illustrate model fitting, fuzzy clustering, basic and pattern-detection models, binary and count data, and the analogues of ordination, multidimensional scaling and correspondence analysis. Our models are not, of course, limited to ecological applications.

4.1. Example 1: trees in the great smoky mountains

In Bullhead Creek in the Great Smoky Mountains, R. H. Whittaker (1956) recorded 32 tree species in eight plots ranging from moist (Plot 1) to dry (Plot 8). The (binary) incidence data are shown in Table 2.

The selected model (using either AIC or AICc) was $\{rR3,\ cp\}$. There was not enough variation in row sums or columns sums for any pattern-detection model to be chosen. Fig. 3(a) shows the profiles of the three *clusters* of species from this model, indicating a classification into groups preferring moist, moderate or dry conditions, even though we did not use soil moisture information in the analysis. Fig. 3(b) shows the *associations* among samples obtained from grouping the rows, with their distances apart determined by their species compositions. This is an analogue of *multidimensional scaling*. (Samples 7 and 8 coincide, having identical species compositions.) Projection onto a single axis provides a one-dimensional *ordination* of the samples which essentially sorts the plots to match Whittaker's moist to dry gradient, even though our analysis was based on neither the order of the columns in the data matrix, nor prior knowledge of covariates such as distance along the transect or soil moisture.

Model $\{rR3,\ cC3\}$, although not selected by AIC, shows the following estimated θ_{rc} matrix by row group (RG) and column group (CG):

$$\Theta = \begin{array}{c} \\ \text{RG1} \\ \text{RG2} \\ \text{RG3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{CG1} & \text{CG2} & \text{CG3} \end{array} \\ \begin{pmatrix} 1.000 & 0.036 & 0.000 \\ 0.582 & 0.819 & 0.104 \\ 0.090 & 0.145 & 0.700 \end{pmatrix} \end{array}.$$

The turnover pattern is shown by the high presence probabilities on the main diagonal, falling off to the smallest numbers in the other corners. This broad overview has provided *pattern detection*, identifying three habitats with some overlap of species. Row group 2 contains species tolerating a wide range of environments (CG1 and CG2), as opposed to RG1 and RG3 which each have a clear preference for a single column group. Our results accord well with the discussion and proposed grouping of species and sites in Whittaker (1956) where he used prior knowledge of the samples.

Table 2
Bullhead Creek data (Whittaker, 1956). Incidences of 32 tree species are given over eight plots from moist (1) to dry (8). (a) shows the raw data, and (b) has the species sorted into three groups by model {*rR*3, *cp*}. The diagonal pattern of 1's in (b) shows the species turnover from moist to dry plots.

(a)										(b)								
Tree species		1	2	3	4	5	6	7	8	Sorted	1	2	3	4	5	6	7	8
1	*A.pen*	0	1	1	1	1	0	0	0	6	1	0	0	0	0	0	0	0
2	*A.rub*	1	0	1	1	1	1	1	1	13	1	0	0	0	0	0	0	0
3	*A.sac*	1	1	1	1	0	0	0	0	17	1	0	0	0	0	0	0	0
4	*A.oct*	1	1	1	0	0	0	0	0	18	1	0	0	0	0	0	0	0
5	*A.arb*	0	0	0	0	0	1	0	0	19	1	0	0	0	0	0	0	0
6	*B.all*	1	0	0	0	0	0	0	0	14	1	0	1	0	0	0	0	0
7	*B.len*	0	1	1	1	1	1	0	0	4	1	1	1	0	0	0	0	0
8	*C.cor*	1	1	1	0	0	0	0	0	8	1	1	1	0	0	0	0	0
9	*C.gla*	0	1	1	0	0	0	0	0	3	1	1	1	1	0	0	0	0
10	*C.den*	0	0	0	1	1	1	1	1	11	1	1	1	1	0	0	0	0
11	*C.lut*	1	1	1	1	0	0	0	0	31	1	1	1	1	0	0	0	0
12	*C.acu*	0	0	0	0	1	1	0	0	32	1	1	1	1	0	0	0	0
13	*F.gra*	1	0	0	0	0	0	0	0	15	1	1	1	1	1	0	0	0
14	*F.ame*	1	0	1	0	0	0	0	0	22	0	1	1	1	0	0	0	0
15	*H.mon*	1	1	1	1	1	0	0	0	1	0	1	1	1	1	0	0	0
16	*H.vir*	0	0	1	1	1	0	0	0	9	0	1	1	0	0	0	0	0
17	*I.opa*	1	0	0	0	0	0	0	0	26	0	1	1	0	0	0	0	0
18	*L.tul*	1	0	0	0	0	0	0	0	20	1	1	0	0	1	0	0	0
19	*M.acu*	1	0	0	0	0	0	0	0	7	0	1	1	1	1	1	0	0
20	*M.f.a*	1	1	0	0	1	0	0	0	16	0	0	1	1	1	0	0	0
21	*N.syl*	0	0	0	0	1	1	1	1	2	1	0	1	1	1	1	1	1
22	*O.vir*	0	1	1	1	0	0	0	0	29	0	0	1	1	0	1	1	1
23	*O.arb*	0	0	0	0	1	1	1	1	12	0	0	0	0	1	1	0	0
24	*P.pun*	0	0	0	0	0	0	1	1	5	0	0	0	0	0	1	0	0
25	*P.rig*	0	0	0	0	0	0	1	1	10	0	0	0	1	1	1	1	1
26	*Q.bor*	0	1	1	0	0	0	0	0	21	0	0	0	0	1	1	1	1
27	*Q.coc*	0	0	0	0	0	0	1	1	23	0	0	0	0	1	1	1	1
28	*Q.pri*	0	0	0	0	1	1	1	1	28	0	0	0	0	1	1	1	1
29	*R.pse*	0	0	1	1	0	1	1	1	24	0	0	0	0	0	0	1	1
30	*S.alb*	0	0	0	0	0	0	1	1	25	0	0	0	0	0	0	1	1
31	*T.het*	1	1	1	1	0	0	0	0	27	0	0	0	0	0	0	1	1
32	*T.can*	1	1	1	1	0	0	0	0	30	0	0	0	0	0	0	1	1

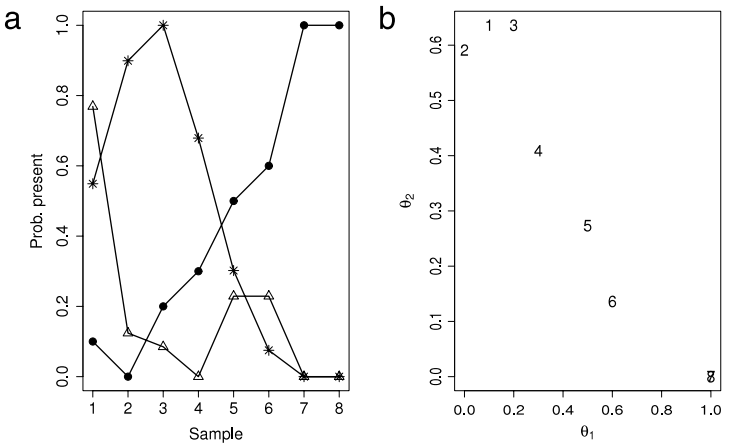


Fig. 3. Tree data plots. (a) Profiles of the three tree species groups from model {*rR*3, *cp*}, ranging from moist (Δ) to medium ($*$) to dry ($\bullet$). (b) A scatterplot of the samples using the θ_{rj} estimates from model {*rR*2, *cp*}.

4.2. Example 2: funghi in Liphook Forest

In an experiment in the Liphook Forest, UK, Peter Shaw counted ten species of funghi in seven fixed quadrats over five years, 1986–1990 (Shaw, 2003; Shaw et al., 2003). Table 3 shows the data for the five years, pooled over the (similar) quadrats.

Because of the varying row sums and column sums, AICc strongly favoured the PD models. Among the class of modules, AICc selected {*rR*3, *cC*3, *PD*}, giving the four pattern-detection matrices Ψ shown in Fig. 4.

Table 3
Liphook Forest data 1986–1990 (Shaw, 2003). The ten toadstool species are Bf = *Boletus ferrugineus*, Cs = *Cortinarius semisanguineus*, Gr = *Gomphidius roseus*, Il = *Inocybe lacera*, Lp = *Laccaria proxima*, Lr = *Lactarius rufus*, Pi = *Paxillus involutus*, Sb = *Suillus bovinus*, Sl = *Suillus luteus*, Sv = *Suillus variegatus*.

Species	Year					Sum
	86	87	88	89	90	
Bf	0	7	18	17	7	49
Cs	0	0	12	45	151	208
Gr	0	99	430	896	222	1 647
Il	0	0	16	1	6	23
Lp	567	2759	3868	182	266	7 642
Lr	0	0	0	2	47	49
Pi	117	131	20	21	40	329
Sb	0	219	2982	5823	2427	11 451
Sl	0	0	11	12	23	46
Sv	0	2	151	534	154	841
Sum	684	3217	7508	7533	3343	22 285

In $\boldsymbol{\Psi}^{(S)}$ the values are $y_{ij}/y^{\star}_{ij}$, where $y^{\star}_{ij}$ is the expected value under the PD-null (no association) model. Values above or below one indicate positive or negative association respectively. The other $\boldsymbol{\Psi}$ matrices are centroids after row and/or column clustering. For this data set the clusterings from the singly-clustered models matched those from the biclustering. These clusters are shown by the partition lines in Fig. 4. Module 1 is the "early" module, {{*Pi*}, {86, 87}}, Module 2 is the "mid" module, {{*Lp*}, {88}} and Module 3, {{Bf, Cs, Gr, Il, Lr, Sb, Sl, Sv}, {89, 90}} is "late" in the succession. The values in the pattern matrices are multiplicative adjustments to the expected values under the no-association model. For example, the value in Case A of $\psi_{23} = 1.50$ shows that with model {*rR*3, *cp*, *PD*} the fitted values for rows in row group two at sample three are 1.5 times that expected under the PD-null model. The matrix $\boldsymbol{\Psi}^{(D)}$ does not show a completely clearcut turnover pattern because of the high value 2.49 in RG2, CG1, which indicates that species Lp (row group 2) is strongly associated with years 86–87 as well as 88.

The data visualisation plots are in Fig. 5. Plots (a) and (b) show the profile plot and scatterplot from row clustering, while plots (c) and (d) show the profile plot and scatterplot from column clustering. Plots (b) and (d) illustrate association analysis on the columns and rows respectively. We note the similarity of years 89 and 90, following large differences over the earlier years, perhaps indicating that the climax species composition has been reached. The mixture-based biplot (e) and the correspondence analysis plot (f) are providing similar information about associations between rows and columns. Both exhibit the arch effect commonly found in correspondence analysis if the data are from a latent long gradient over the columns.

5. Discussion

We have proposed analysis of binary or count data matrices using finite mixtures to cluster the rows and/or columns. The pattern detection (PD) models first allow for varying row sums and column sums, which is our analogue of standardising the data before analysis. The remaining variation in the data, modelled as PD matrices, focuses on association patterns between rows and columns. The models are all fitted by maximum likelihood, with its attendant advantages of providing model comparison, hypothesis testing and likelihood-based confidence intervals for parameters. Interpretable plots are also provided by the PD matrices. A disadvantage of mixture modelling is possible multimodality of the likelihood surface, necessitating a comprehensive search over different starting points in order to find the global maximum. Computation of the full likelihood for dually-clustered models can be slow with large data matrices, and we have substantially reduced the time required by calling **C** from **R**. Free software will be made available as an **R** package.

The likelihoods of the PD models indicate how much of the full pattern in the data (the saturated model) can be attributed to the association of each row with certain columns, or vice versa. Examples of failure of this independence assumption from ecological communities may occur if (i) two species are in competition for the same resources, so presence (abundance) of one lowers probability (abundance) of the other, (ii) two species have a predator–prey relationship, so presence of the prey enhances presence of the predator, or (iii) there are cooperative relationships (mutualism) such as symbiosis. Correlations between residuals after fitting our models may provide evidence of such relationships, and extra covariates (e.g. presence/absence of another species) in the model could be tried (Pledger, in prep.).

The fuzziness or crispness of the clustering as seen in the posterior allocation to groups ($\widehat{z}_{ir}$ and $\widehat{x}_{jc}$) should not be used to assume the model is a good fit; instead we recommend the theoretically-justified information criteria or LRT. We found data sets with crisp clustering (e.g. each row allocated with probability one to its row group) while AIC showed there were several more row groups.

Overall, we have shown how a few finite mixture models applied to binary or count data can encompass applications in clustering, ordination, correspondence analysis and pattern detection. The models exploit redundancies in the row or column patterns, thereby reducing the number of parameters, enabling broad patterns to be identified and tested for statistically.

$\Psi^{(S)}$

	86	87	88	89	90
Pi	11.59	2.76	0.18	0.19	0.81
Lp	2.42	2.50	1.50	0.07	0.23
Bf	0.00	0.99	1.09	1.03	0.95
Cs	0.00	0.00	0.17	0.64	4.84
Gr	0.00	0.42	0.77	1.61	0.90
Il	0.00	0.00	2.06	0.13	1.74
Lr	0.00	0.00	0.00	0.12	6.39
Sb	0.00	0.13	0.77	1.50	1.41
Sl	0.00	0.00	0.71	0.77	3.33
Sv	0.00	0.02	0.53	1.88	1.22

$\Psi^{(B)}$

	CG1	CG2	CG3
Pi	4.31	0.18	0.38
Lp	2.49	1.50	0.12
Bf	0.82	1.09	1.00
Cs	0.00	0.17	1.93
Gr	0.34	0.78	1.39
Il	0.00	2.06	0.62
Lr	0.00	0.00	2.05
Sb	0.11	0.77	1.48
Sl	0.00	0.71	1.56
Sv	0.00	0.53	1.68

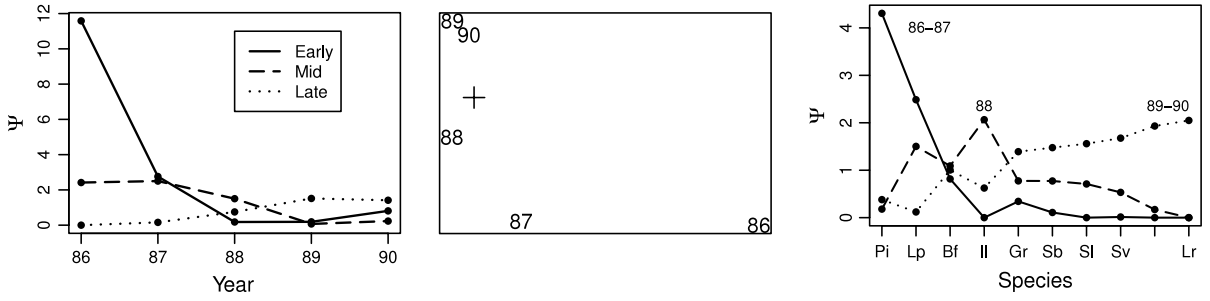
$\Psi^{(A)}$

	86	87	88	89	90
RG1	11.59	2.76	0.18	0.19	0.81
RG2	2.42	2.50	1.50	0.07	0.23
RG3	0.00	0.16	0.75	1.52	1.41

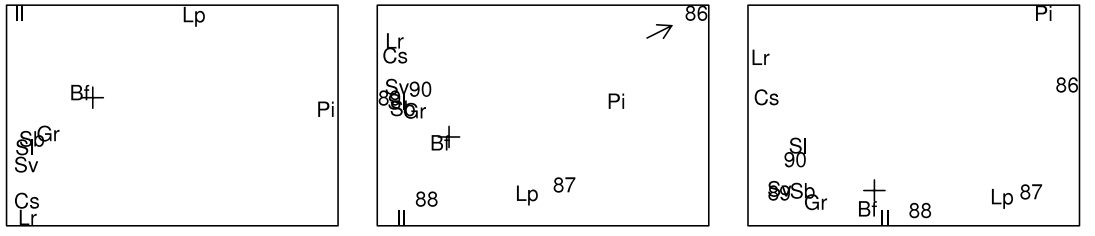
$\Psi^{(D)}$

	CG1	CG2	CG3
RG1	4.31	0.18	0.38
RG2	2.49	1.50	0.12
RG3	0.13	0.75	1.48

Fig. 4. The pattern-detection matrices Ψ for the Liphook Forest count data using Case S (the saturated model), Case B (model $\{rn,\ cC3, C, PD\}$), Case A (model $\{rR3,\ cp, C, PD\}$) and Case D (model $\{rR3,\ cC3, C, PD\}$). The row groups in Cases A and D are $\{Pi\}$, $\{Lp\}$, and {Bf, Cs, Gr, Il, Lr, Sb, Sl, Sv}, while the column groups in B and D are {86, 87}, {88} and {89, 90}.



(a) Row group profiles. (b) Columns scatterplot. (c) Column group profiles.



(d) Rows scatterplot. (e) Mixture-based biplot. (f) Corresp. analysis 2D plot.

Fig. 5. Plots from Liphook Forest funghi data. Plots (a) and (b) are from model $\{rR3,\ cp, C, PD\}$, with species groups Δ = early, $*$ = mid and $\bullet$ = late. Plots (c) and (d) are from model $\{rn,\ cC3, C, PD\}$, with year groups $\Delta = \{86, 87\}$, $* = \{88\}$ and $\bullet = \{89, 90\}$. The mixture-based biplot is in (e) and (f) is the traditional correspondence analysis 2D plot for comparison. Plots (b), (d) and (e) do not show the triangle vertices, which are outside the plotted region. The centroids are marked $+$, and the arrow in the biplot indicates that Year 86 is an outlier with true position twice the plotted distance from the centroid.

There are numerous applications for these models, for example in item response analysis and in contingency table analysis. Many disciplines use correspondence analysis and related multivariate techniques, and our models may be used to address problems in these areas, with the substantial advantage of having a likelihood-based foundation. We also anticipate extensions to detecting modularity in ecological networks (Pocock and Pledger, in prep.), to other count data (e.g. the geometric, negative binomial and multinomial distributions), to ordinal data, to higher dimensions and with Bayesian analyses.

Acknowledgement

We thank Peter Shaw for permission to use the Liphook Forest data set.

Appendix A. EM algorithm formulae

The mixture models are fitted using the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997). The missing information is the actual membership of the row groups and/or column groups.

For the row-clustered models we use the procedure in McLachlan and Peel (2000), Sections 1.9 and 2.8.2–2.8.4. The missing data vector has components $\mathbf{Z}_i$ $(i = 1, \ldots, n)$ which are realised values of random vectors $\mathbf{Z}_i$ of length R. These are indicators, with $Z_{ir} = 1$ if row i is in row group r $(i \in r)$, otherwise zero. *A priori* independence of the rows in the observed data matrix leads to assuming unconditional i.i.d. multinomial distributions for $\mathbf{Z}_i$,

$$\mathbf{Z}_i|\phi \sim \text{Multinomial}(1, \boldsymbol{\pi}) \quad \text{for } i = 1, \ldots, n.$$

Hence for parameter vector ϕ, observed data matrix $\mathbf{Y}$ $(n \times p)$ and unobserved data matrix $\mathbf{Z}$ $(n \times R)$, the complete log likelihood is

$$\ell_C(\phi \mid Y, Z) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R} z_{ir} \log f(\theta_{rj}; y_{ij}) + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir} \log \pi_r,$$

where $f(\theta; y) = \theta^y(1-\theta)^{1-y}$ for binary data and $f(\theta; y) = \frac{e^{-\theta}\theta^y}{y!}$ for count data. *A posteriori* the distribution of $\mathbf{Z}_i$ is still multinomial, but with modified probabilities:

$$\mathbf{Z}_i|\mathbf{Y}, \phi \sim \text{Multinomial}(1, \widehat{\mathbf{z}}_i) \quad \text{for } i = 1, \ldots, n,$$

where

$$\widehat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p} f(\theta_{rj}; y_{ij})}{\sum_{a=1}^{R} \pi_a \prod_{\ell=1}^{p} f(\theta_{a\ell}; y_{i\ell})}.$$

In the E-step of the EM algorithm, the estimate of $E(Z_{ir}|\mathbf{Y}, \phi)$ is updated to be the value of $\widehat{z}_{ir}$ obtained using the most recent updates of π_r and θ_{rj}. In the M-step, the most recent estimated expected values of Z_{ir} $(\widehat{z}_{ir})$ are used in the ℓ_C formula, which is maximised to provide new estimates of parameters θ_{rj} and π_r.

The column clustered models are just transposed versions of the row clustered models.

For biclustered models there is an extra missing data vector of indicator vector random variables $\mathbf{X}_j$ where $X_{jc} = 1$ if $j \in c$, otherwise zero. *A priori* independence of the columns in the observed data matrix leads to assuming unconditional i.i.d. multinomial distributions for $\mathbf{X}_j$,

$$\mathbf{X}_j|\phi \sim \text{Multinomial}(1, \boldsymbol{\kappa}) \quad \text{for } j = 1, \ldots, p,$$

and $\mathbf{Z}_i$ has the same multinomial distribution *a priori* as in the row clustered models. Hence the complete log likelihood for biclustered data is

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C} z_{ir}x_{jc} \log f(\theta_{rc}; y_{ij}) + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir} \log \pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C} x_{jc} \log \kappa_c,$$

with $f(\theta; y)$ for binary or count data as before.

A posteriori the full conditional distributions of $\mathbf{Z}_i$ and $\mathbf{X}_j$ are still multinomial, but with modified probabilities:

$$\mathbf{Z}_i|\mathbf{Y}, \mathbf{X}, \phi \sim \text{Multinomial}(1, \tilde{\mathbf{z}}_i) \quad \text{for } i = 1, \ldots, n$$
$$\mathbf{X}_j|\mathbf{Y}, \mathbf{Z}, \phi \sim \text{Multinomial}(1, \tilde{\mathbf{x}}_j) \quad \text{for } j = 1, \ldots, p$$

where

$$\tilde{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p}\prod_{c=1}^{C} f(\theta_{rc}; y_{ij})^{x_{jc}}}{\sum_{a=1}^{R} \pi_a \prod_{\ell=1}^{p}\prod_{b=1}^{C} f(\theta_{ab}; y_{i\ell})^{x_{\ell b}}},$$

$$\tilde{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n}\prod_{r=1}^{R} f(\theta_{rc}; y_{ij})^{z_{ir}}}{\sum_{a=1}^{C} \kappa_a \prod_{\ell=1}^{n}\prod_{b=1}^{R} f(\theta_{ba}; y_{\ell j})^{z_{\ell b}}}.$$

Integrating out the dependence of $\mathbf{Z}_i$ on $\mathbf{X}$ and the dependence of $\mathbf{X}_j$ on $\mathbf{Z}$ the marginalised *a posteriori* distributions of $\mathbf{Z}_i$ and $\mathbf{X}_j$ are

$$\mathbf{Z}_i|\mathbf{Y}, \phi \sim \text{Multinomial}(1, \widehat{\mathbf{z}}_i) \quad \text{for } i = 1, \ldots, n$$
$$\mathbf{X}_j|\mathbf{Y}, \phi \sim \text{Multinomial}(1, \widehat{\mathbf{x}}_j) \quad \text{for } j = 1, \ldots, p$$

where

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \widehat{\kappa}_c f(\widehat{\theta}_{rc}; y_{ij}) \right\}}{\sum_{a=1}^{R} \widehat{\pi}_a \prod_{\ell=1}^{p} \left\{ \sum_{b=1}^{C} \widehat{\kappa}_b f(\widehat{\theta}_{ab}; y_{i\ell}) \right\}},$$

$$\widehat{x}_{jc} = \frac{\widehat{\kappa}_c \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \widehat{\pi}_r f(\widehat{\theta}_{rc}; y_{ij}) \right\}}{\sum_{a=1}^{C} \widehat{\kappa}_a \prod_{k=1}^{n} \left\{ \sum_{b=1}^{R} \widehat{\pi}_r f(\widehat{\theta}_{ba}; y_{kj}) \right\}}.$$

The E-step of the EM algorithm calls for the expectation of the complete data log likelihood over the missing data ($\mathbf{Z}$, $\mathbf{X}$) conditional on the observed data ($\mathbf{Y}$):

$$E[\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) | \mathbf{Y}, \phi] = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} \sum_{c=1}^{C} E[z_{ir} x_{jc} | \mathbf{Y}, \phi] \log f(\theta_{rc}; y_{ij}) + \sum_{i=1}^{n} \sum_{r=1}^{R} E[z_{ir} | \mathbf{Y}, \phi] \log \pi_r + \sum_{j=1}^{p} \sum_{c=1}^{C} E[x_{jc} | \mathbf{Y}, \phi] \log \kappa_c.$$

The expectations in the latter two terms are simply $\widehat{z}_{ir}$ and $\widehat{x}_{jc}$. However the lack of *a posteriori* independence of the z_{ir} and x_{jc} makes the evaluation of $E[z_{ir}x_{jc}|\mathbf{Y}, \phi]$ computationally expensive: it involves a sum either over all possible allocations of rows to row groups, or over all possible allocations of columns to column groups.

The variational approximation employed by Govaert and Nadif (2005) is a solution to this problem: and we approximate the expectation as follows

$$E[z_{ir}x_{jc}|\mathbf{Y}, \phi] \simeq E[z_{ir}|\mathbf{Y}, \phi] E[x_{jc}|\mathbf{Y}, \phi] = \widehat{z}_{ir}\widehat{x}_{jc}.$$

Using this approximation we evaluate $\widehat{z}_{ir}$ and $\widehat{x}_{jc}$ using the most recent estimates of the parameters θ_{rc}, π_r and κ_c, and substitute them into $E[\ell_C|\mathbf{Y}, \phi]$. In the M-step the ℓ_C formula is maximised to provide updated estimates of the θ, π and κ parameters.

We now give more explicit formulae for the various models.

A.1. Binary data

Model {rR, cp, B} row clustering

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} \widehat{z}_{ir} \left\{ y_{ij} \log \theta_{rj} + (1 - y_{ij}) \log(1 - \theta_{rj}) \right\} + \sum_{i=1}^{n} \sum_{r=1}^{R} \widehat{z}_{ir} \log \pi_r.$$

E-step:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \widehat{\theta}_{rj}^{y_{ij}} (1 - \widehat{\theta}_{rj})^{1 - y_{ij}}}{\sum_{a=1}^{R} \left\{ \widehat{\pi}_a \prod_{\ell=1}^{p} \widehat{\theta}_{a\ell}^{y_{i\ell}} (1 - \widehat{\theta}_{a\ell})^{1 - y_{i\ell}} \right\}}.$$

M-step:

$$\widehat{\theta}_{rj} = \frac{\sum_{i=1}^{n} \widehat{z}_{ir} y_{ij}}{\sum_{i=1}^{n} \widehat{z}_{ir}}.$$

Model $\{rR,\ cC, B\}$ *biclustering*

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C}\widehat{z}_{ir}\widehat{x}_{jc}\left\{y_{ij}\log\theta_{rc} + (1-y_{ij})\log(1-\theta_{rc})\right\} + \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log\pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C}\widehat{x}_{jc}\log\kappa_c.$$

E-steps:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r\prod_{j=1}^{p}\left\{\sum_{c=1}^{C}\widehat{\kappa}_c\widehat{\theta}_{rc}^{y_{ij}}(1-\widehat{\theta}_{rc})^{1-y_{ij}}\right\}}{\sum_{a=1}^{R}\widehat{\pi}_a\prod_{\ell=1}^{p}\left\{\sum_{b=1}^{C}\widehat{\kappa}_b\widehat{\theta}_{ab}^{y_{i\ell}}(1-\widehat{\theta}_{ab})^{1-y_{i\ell}}\right\}},$$

$$\widehat{x}_{jc} = \frac{\widehat{\kappa}_c\prod_{i=1}^{n}\left\{\sum_{r=1}^{R}\widehat{\pi}_r\widehat{\theta}_{rc}^{y_{ij}}(1-\widehat{\theta}_{rc})^{1-y_{ij}}\right\}}{\sum_{a=1}^{C}\widehat{\kappa}_a\prod_{k=1}^{n}\left\{\sum_{b=1}^{R}\widehat{\pi}_r\widehat{\theta}_{ba}^{y_{kj}}(1-\widehat{\theta}_{ba})^{1-y_{kj}}\right\}}.$$

M-step:

$$\widehat{\theta}_{rc} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{z}_{ir}\widehat{x}_{jc}y_{ij}}{\left(\sum_{i=1}^{n}\widehat{z}_{ir}\right)\left(\sum_{j=1}^{p}\widehat{x}_{jc}\right)}.$$

Model $\{rn + cp, B, PD\}$ *PD-null.*

The binary PD-null model has logit $\theta_{ij} = \mu + \alpha_i + \beta_j$, using constraints $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$. Since there are no explicit solutions we use numerical maximisation of the log likelihood

$$\ell(\phi \mid \mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left\{y_{ij}\log\theta_{ij} + (1-y_{ij})\log(1-\theta_{ij})\right\}.$$

Model $\{rR,\ cp, B, PD\}$ *row clustering with PD.*

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\widehat{z}_{ir}\left\{y_{ij}\log\theta_{ijr} + (1-y_{ij})\log(1-\theta_{ijr})\right\} + \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log\pi_r,$$

where logit $\theta_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{rj}$.

E-step:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r\prod_{j=1}^{p}\left\{\widehat{\theta}_{ijr}^{y_{ij}}(1-\widehat{\theta}_{ijr})^{1-y_{ij}}\right\}}{\sum_{a=1}^{R}\widehat{\pi}_a\prod_{\ell=1}^{p}\left\{\widehat{\theta}_{i\ell a}^{y_{i\ell}}(1-\widehat{\theta}_{i\ell a})^{1-y_{i\ell}}\right\}}.$$

M-step: Numerically maximise $\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z})$ over μ, α_i, β_j, γ_{rj} and π_r.

Alternatively, to make $\mathbf{\Gamma}$ represent a deviation from the PD-null model, fix μ, α_i and β_j at their estimates from that model and maximise $\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z})$ over γ_{rj} and π_r only. This yields suboptimal but meaningful estimates.

Model $\{rR,\ cC, B, PD\}$ *biclustering with PD.*

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C}\widehat{z}_{ir}\widehat{x}_{jc}\left\{y_{ij}\log\theta_{ijrc} + (1-y_{ij})\log(1-\theta_{ijrc})\right\} + \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log\pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C}\widehat{x}_{jc}\log\kappa_c,$$

where logit $\theta_{ijrc} = \mu + \alpha_i + \beta_j + \gamma_{rc}$.

E-steps:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \widehat{\kappa}_c \widehat{\theta}_{ijrc}^{y_{ij}} (1-\widehat{\theta}_{ijrc})^{1-y_{ij}} \right\}}{\sum_{a=1}^{R} \widehat{\pi}_a \prod_{\ell=1}^{p} \left\{ \sum_{b=1}^{C} \widehat{\kappa}_b \widehat{\theta}_{i\ell ab}^{y_{ij}} (1-\widehat{\theta}_{i\ell ab})^{1-y_{i\ell}} \right\}},$$

$$\widehat{x}_{jc} = \frac{\widehat{\kappa}_c \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \widehat{\pi}_r \widehat{\theta}_{ijrc}^{y_{ij}} (1-\widehat{\theta}_{ijrc})^{1-y_{ij}} \right\}}{\sum_{a=1}^{C} \widehat{\kappa}_a \prod_{k=1}^{n} \left\{ \sum_{b=1}^{R} \widehat{\pi}_b \widehat{\theta}_{kjba}^{y_{kj}} (1-\widehat{\theta}_{kjba})^{1-y_{kj}} \right\}}.$$

M-step: Numerically maximise $\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X})$ over μ, α_i, β_j, γ_{rc}, π_r and κ_c.

Alternatively to make $\mathbf{\Gamma}$ represent a deviation from the PD-null model, fix μ, α_i and β_j at their estimates from that model and maximise $\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X})$ only over the other parameters.

A.2. Count data

With count data, the M steps yield explicit formulae for the parameter estimates. The following likelihood formulae omit a factorial constant dependent only on the data.

Model $\{rR,\ cp, C\}$ *row clustering.*

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} \widehat{z}_{ir} \left(-\theta_{rj} + y_{ij} \log \theta_{rj} \right) + \sum_{i=1}^{n} \sum_{r=1}^{R} \widehat{z}_{ir} \log \pi_r.$$

E-step:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} e^{-\widehat{\theta}_{rj}} \widehat{\theta}_{rj}^{y_{ij}}}{\sum_{a=1}^{R} \widehat{\pi}_a \prod_{\ell=1}^{p} e^{-\widehat{\theta}_{a\ell}} \widehat{\theta}_{a\ell}^{y_{i\ell}}}.$$

M-step:

$$\widehat{\theta}_{rj} = \frac{\sum_{i=1}^{n} \widehat{z}_{ir} y_{ij}}{\sum_{i=1}^{n} \widehat{z}_{ir}}.$$

Model $\{rR,\ cC, C\}$ *biclustering.*

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} \sum_{c=1}^{C} \widehat{z}_{ir} \widehat{x}_{jc} \left(-\theta_{rc} + y_{ij} \log \theta_{rc} \right) + \sum_{i=1}^{n} \sum_{r=1}^{R} \widehat{z}_{ir} \log \pi_r + \sum_{j=1}^{p} \sum_{c=1}^{C} \widehat{x}_{jc} \log \kappa_c.$$

E-steps:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \left(\sum_{c=1}^{C} \widehat{\kappa}_c e^{-\widehat{\theta}_{rc}} \widehat{\theta}_{rc}^{y_{ij}} \right)}{\sum_{a=1}^{R} \widehat{\pi}_a \prod_{\ell=1}^{p} \left(\sum_{b=1}^{C} \widehat{\kappa}_b e^{-\widehat{\theta}_{ab}} \widehat{\theta}_{ab}^{y_{i\ell}} \right)}, \qquad \widehat{x}_{jc} = \frac{\widehat{\kappa}_c \prod_{i=1}^{n} \left(\sum_{r=1}^{R} \widehat{\pi}_r e^{-\widehat{\theta}_{rc}} \widehat{\theta}_{rc}^{y_{ij}} \right)}{\sum_{a=1}^{C} \widehat{\kappa}_a \prod_{k=1}^{n} \left(\sum_{b=1}^{R} \widehat{\pi}_b e^{-\widehat{\theta}_{ba}} \widehat{\theta}_{ba}^{y_{kj}} \right)}.$$

M-step:

$$\widehat{\theta}_{rc} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} \widehat{z}_{ir} \widehat{x}_{jc} y_{ij}}{\left(\sum_{i=1}^{n} \widehat{z}_{ir} \right) \left(\sum_{j=1}^{p} \widehat{x}_{jc} \right)}.$$

Model $\{rn + cp, C\}$ *PD-null.*

This PD-null model (the no-association loglinear model for a two-way contingency table) has $\log\theta_{ij} = \mu + \alpha_i + \beta_j$, using constraints $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$. The solutions are

$$\widehat{\mu} = \log\frac{G_R G_C}{y_{++}}, \qquad \widehat{\alpha}_i = \log\frac{y_{i+}}{G_R} \quad \text{and} \quad \widehat{\beta}_j = \log\frac{y_{+j}}{G_C}, \tag{12}$$

where G_R is the geometric mean of the row sums y_{i+}, and G_C is the geometric mean of the column sums y_{+j}.

Using our choice of constraints on $\mathbf{\Gamma}$ (Eqs. (9) and (10)) the PD models have the following EM algorithms.

Model $\{rR,\ cp, C, PD\}$ *row clustering with PD.*

If row i is in row-group r, $\log\theta_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{rj}$ and

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\widehat{z}_{ir}\left(y_{ij}\log\theta_{ijr} - \theta_{ijr}\right) + \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log\pi_r.$$

E-step:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} e^{-\widehat{\theta}_{ijr}}\widehat{\theta}_{ijr}^{y_{ij}}}{\sum_{a=1}^{R}\widehat{\pi}_a \prod_{\ell=1}^{p} e^{-\widehat{\theta}_{i\ell a}}\widehat{\theta}_{i\ell a}^{y_{i\ell}}}. \tag{13}$$

M step: The constraints are $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$ and Eqs. (9). We adopt the solution which uses the estimates in Eqs. (12) together with

$$\widehat{\psi}_{rj} = \frac{y_{++}\sum_{i=1}^{n}\widehat{z}_{ir}y_{ij}}{y_{+j}\,a_r}, \tag{14}$$

where $a_r = \sum_{i=1}^{n}\widehat{z}_{ir}y_{i+}$, the total abundance in row-group r. In Appendix B we show that any other solution satisfying the constraints has the same likelihood; this ensures we have not chosen a suboptimal solution. Note that our choice of solution has properties

$$\text{(i) } \forall j = 1,\ldots,p, \quad \sum_{r=1}^{R} a_r\widehat{\psi}_{rj} = y_{++}, \qquad \text{(ii) } \forall r = 1,\ldots,R, \quad \sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{z}_{ir}\widehat{\theta}_{ijr} = a_r. \tag{15}$$

Equation (ii) is that expected = observed total count within each row-group r. Equation (i) is used in the geometrical representation of the columns in Triangle A (Section 3.2).

Model $\{rR,\ cC, C, PD\}$ *biclustering with PD.*

If $i \in r$ and $j \in c$, the model is $\log\theta_{ijrc} = \mu + \alpha_i + \beta_j + \gamma_{rc}$ and

$$\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C}\widehat{z}_{ir}\widehat{x}_{jc}\left(y_{ij}\log\theta_{ijrc} - \theta_{ijrc}\right) + \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log\pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C}\widehat{x}_{jc}\log\kappa_c.$$

E steps:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p}\left(\sum_{c=1}^{C}\widehat{\kappa}_c e^{-\widehat{\theta}_{ijrc}}\widehat{\theta}_{ijrc}^{y_{ij}}\right)}{\sum_{a=1}^{R}\widehat{\pi}_a \prod_{\ell=1}^{p}\left(\sum_{b=1}^{C}\widehat{\kappa}_b e^{-\widehat{\theta}_{i\ell ab}}\widehat{\theta}_{i\ell ab}^{y_{i\ell}}\right)}, \tag{16}$$

$$\widehat{x}_{jc} = \frac{\widehat{\kappa}_c \prod_{i=1}^{n}\left(\sum_{r=1}^{R}\widehat{\pi}_r e^{-\widehat{\theta}_{ijrc}}\widehat{\theta}_{ijrc}^{y_{ij}}\right)}{\sum_{a=1}^{C}\widehat{\kappa}_a \prod_{k=1}^{n}\left(\sum_{b=1}^{R}\widehat{\pi}_b e^{-\widehat{\theta}_{kjba}}\widehat{\theta}_{kjba}^{y_{kj}}\right)}.$$

M step: The constraints are $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$ and Eqs. (10). We adopt the solution which uses the estimates in Eq. (12) together with

$$\widehat{\psi}_{rc} = \frac{y_{++}\sum_{i=1}^{n}\sum_{j=1}^{p}\widehat{z}_{ir}\widehat{x}_{jc}y_{ij}}{a_r\,b_c}, \tag{17}$$

where $a_r = \sum_{i=1}^{n}\widehat{z}_{ir}y_{i+}$ and $b_c = \sum_{j=1}^{p}\widehat{x}_{jc}y_{+j}$, the total abundances in row-group r and column-group c respectively. In

Appendix B we show that no other solution has a higher likelihood. Note that our choice of solution has properties

$$\text{(i) } \forall r = 1, \ldots, R, \quad \sum_{c=1}^{C} b_c \widehat{\psi}_{rc} = y_{++}, \quad \text{and} \quad \text{(ii) } \forall c = 1, \ldots, C, \quad \sum_{r=1}^{R} a_r \widehat{\psi}_{rc} = y_{++}. \tag{18}$$

These constraints are used in the low-dimensional representation of row groups and column groups (Section 3.2).

Appendix B. Constraints for PD models and count data

Model {*rRcp*, *C*, *PD*}.

For arbitrarily chosen constants μ, α_i, β_j and γ_{rj} (with $i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}$ and $r \in \{1, \ldots, R\}$) and a classification matrix $\{z_{ir}\}$ such that $z_{ir} \in \{0, 1\}$ and $\sum_r z_{ir} = 1\ \forall i$, we define

$$\eta_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{rj}, \qquad \tau_{ij} = \sum_r z_{ir}\gamma_{rj}, \qquad \overline{\alpha} = \frac{1}{n}\sum_i \alpha_i \quad \text{and} \quad \overline{\beta} = \frac{1}{p}\sum_j \beta_j,$$

so that $\log\theta_{ijr} = \eta_{ijr}$. Then we can find a set of constants $\mu', \alpha'_i, \beta'_j$ and γ'_{rj} that satisfy the constraints

$$\sum_i \alpha'_i = 0, \qquad \sum_j \beta'_j = 0, \qquad \sum_j e^{\beta'_j + \gamma'_{rj}} = \sum_j e^{\beta'_j} \quad \forall r,$$
$$\sum_i e^{\alpha'_i + \sum_r z_{ir}\gamma'_{rj}} = \sum_i e^{\alpha'_i} \quad \forall j \quad \text{and} \quad z_{ir}\eta_{ijr} = z_{ir}\eta'_{ijr}.$$

The final constraint ensures that the complete likelihood under both parameterisations is the same:

$$\begin{aligned}
\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}) &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R} z_{ir}\left(y_{ij}\eta_{ijr} - \exp\eta_{ijr}\right) + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij}\left(\sum_{r=1}^{R} z_{ir}\eta_{ijr}\right) - \exp\left(\sum_{r=1}^{R} z_{ir}\eta_{ijr}\right)\right) + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij}\left(\sum_{r=1}^{R} z_{ir}\eta'_{ijr}\right) - \exp\left(\sum_{r=1}^{R} z_{ir}\eta'_{ijr}\right)\right) + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r \\
&= \ell_C(\phi' \mid \mathbf{Y}, \mathbf{Z}).
\end{aligned}$$

Expressions for $\mu', \alpha'_i, \beta'_j$ and γ'_{rj} are as follows:

$$\mu' = \mu + \bar{\alpha} + \bar{\beta} - \log\left(\sum_{k\ell} e^{\alpha_k + \beta_\ell + \tau_{k\ell}}\right) + \frac{1}{n}\sum_k \log\left(\sum_\ell e^{\beta_\ell + \tau_{k\ell}}\right) + \frac{1}{p}\sum_\ell \log\left(\sum_k e^{\alpha_k + \tau_{k\ell}}\right),$$
$$\alpha'_i = \alpha_i - \bar{\alpha} + \log\left(\sum_\ell e^{\beta_\ell + \tau_{i\ell}}\right) - \frac{1}{n}\sum_k \log\left(\sum_\ell e^{\beta_\ell + \tau_{k\ell}}\right),$$
$$\beta'_j = \beta_j - \bar{\beta} + \log\left(\sum_k e^{\alpha_k + \tau_{kj}}\right) - \frac{1}{p}\sum_\ell \log\left(\sum_k e^{\alpha_k + \tau_{k\ell}}\right),$$
$$\gamma'_{rj} = \gamma_{rj} - \log\left(\sum_\ell e^{\beta_\ell + \tau_{i\ell}}\right) - \log\left(\sum_k e^{\alpha_k + \tau_{kj}}\right) + \log\left(\sum_{k\ell} e^{\alpha_k + \beta_\ell + \tau_{k\ell}}\right).$$

Model {*rRcC*, *C*, *PD*}.

For arbitrarily chosen constants μ, α_i, β_j and γ_{rc} (with $i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}, r \in \{1, \ldots, R\}, c \in \{1, \ldots, C\}$) and classification matrices $\{z_{ir}\}$ and $\{x_{jc}\}$ such that $z_{ir}, x_{jc} \in \{0, 1\}$, $\sum_r z_{ir} = 1\ \forall i$ and $\sum_c x_{jc} = 1\ \forall j$, we define

$$\eta_{ijrc} = \mu + \alpha_i + \beta_j + \gamma_{rc}, \qquad q_{rj} = \sum_c x_{jc}\gamma_{rc}, \qquad s_{ic} = \sum_r z_{ir}\gamma_{rc} \quad \text{and} \quad \tau_{ij} = \sum_{rc} z_{ir}x_{jc}\gamma_{rc},$$

so that $\log\theta_{ijrc} = \eta_{ijrc}$. Then we can find a set of constants $\mu', \alpha'_i, \beta'_j$ and γ'_{rc} that satisfy the constraints

$$\sum_i \alpha'_i = 0, \qquad \sum_j \beta'_j = 0, \qquad \sum_j e^{\beta'_j + \sum_c x_{jc}\gamma'_{rc}} = \sum_j e^{\beta'_j} \quad \forall r,$$
$$\sum_i e^{\alpha'_i + \sum_r z_{ir}\gamma'_{rc}} = \sum_i e^{\alpha'_i} \quad \forall c \quad \text{and} \quad z_{ir}x_{jc}\eta_{ijrc} = z_{ir}x_{jc}\eta'_{ijrc}.$$

The final constraint ensures that the complete likelihood under both parameterisations is the same:

$$\begin{aligned}
\ell_C(\phi \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}) &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C} z_{ir}x_{jc}\left(y_{ij}\eta_{ijrc} - \exp\eta_{ijrc}\right) + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C} x_{jc}\log\kappa_c \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij}\left(\sum_{r=1}^{R}\sum_{c=1}^{C} z_{ir}x_{jc}\eta_{ijrc}\right) - \exp\left(\sum_{r=1}^{R}\sum_{c=1}^{C} z_{ir}x_{jc}\eta_{ijrc}\right)\right) \\
&\quad + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C} x_{jc}\log\kappa_c \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_{ij}\left(\sum_{r=1}^{R}\sum_{c=1}^{C} z_{ir}x_{jc}\eta'_{ijrc}\right) - \exp\left(\sum_{r=1}^{R}\sum_{c=1}^{C} z_{ir}x_{jc}\eta'_{ijrc}\right)\right) \\
&\quad + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r + \sum_{j=1}^{p}\sum_{c=1}^{C} x_{jc}\log\kappa_c \\
&= \ell_C(\phi' \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}).
\end{aligned}$$

With the redefined τ_{ij}, the expressions for μ', α'_i, and β'_j are the same as for model $\{rRcp, C, PD\}$ above. The expression for γ'_{rc} is:

$$\gamma'_{rc} = \gamma_{rc} - \log\left(\sum_j e^{\beta_j + q_{rj}}\right) - \log\left(\sum_i e^{\alpha_i + s_{ic}}\right) + \log\left(\sum_{ij} e^{\alpha_i + \beta_j + \tau_{ij}}\right).$$

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Academiai Kiado, pp. 267–281.

Arnold, R., Hayakawa, Y., Yip, P., 2010. Capture–recapture estimation using finite mixtures of arbitrary dimension. Biometrics 66, 644–655.

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

Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Learning 22, 719–725.

Böhning, D., Siedel, W., Alfó, M., Garel, B., Patilea, V., Günther, W., 2007. Editorial: advances in mixture models. Computational Statistics and Data Analysis 51, 5205–5210.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, second ed. Springer, New York.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39, 1–38 (with discussion).

Dunstan, P.K., Foster, S.D., Darnell, R., 2011. Model based grouping of species across environmental gradients. Ecological Modelling 222, 955–963. http://dx.doi.org/10.1016/ecolmodel2010.11.030.

Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis, fourth ed. Arnold, London.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453–3467.

Gotelli, N.J., Graves, G.R., 1996. Null Models in Ecology. Smithsonian Institution Press, Washington DC.

Govaert, G., Nadif, M., 2003. Clustering with block mixture models. Pattern Recognition 36, 463–473.

Govaert, G., Nadif, M., 2005. An EM algorithm for the block mixture model. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (4), 643–647.

Govaert, G., Nadif, M., 2010. Latent block model for contingency table. Communications in Statistics—Theory and Methods 39, 416–425.

Greenacre, M., Hastie, T., 1987. The geometric interpretation of correspondence analysis. Journal of the American Statistical Association 82, 437–447.

Hill, M.O., 1979. TWINSPAN—A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-Way Table by Classification of the Individuals and Attributes. In: Section of Ecology and Systematics, Cornell University, New York, NY, USA, p. 90.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Manly, B.F.J., 2005. Multivariate Methods: A Primer, third ed. Chapman & Hall/CRC Press, Boca Raton, FL.

Manly, B.F.J., 2007. Randomization, Bootstrap and Monte Carlo Methods in Biology, third ed. Chapman and Hall, London.

McLachlan, G.J., 1982. The classification and mixture maximum likelihood approaches to cluster analysis. Handbook of Statistics 2, 199–208.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics 36, 318–324.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. M. Dekker, New York, NY.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley Interscience, New York.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley Interscience, New York.

Nadif, M., Govaert, G., 2005. Block clustering of contingency table and mixture model. In: Proceeding of: Advances in Intelligent Data Analysis VI, 6th International Symposium on Intelligent Data Analysis.

O'Hagan, A., Murphy, T.B., Gormley, I.C., 2012. Computational aspects of fitting mixture models via the expectation–maximization algorithm. Computational Statistics and Data Analysis 56, 3843–3864.

Pledger, S., 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. Biometrics 56, 434–442.

Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0.

Robinson, W.S., 1951. A method for chronologically ordering archaeological deposits. American Antiquity 16, 293–301.

Schlattmann, P., 2003. Estimating the number of components in a finite mixture model: the special case of homogeneity. Computational Statistics and Data Analysis 41, 441–451.

Schwarz, G.E., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Self, S.G., Liang, K.-Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82, 605–610.
Shaw, P.J.A., 2003. Multivariate Statistics for the Environmental Sciences. Arnold, London.
Shaw, P.J.A., Kibby, G., Mayes, J., 2003. Effects of thinning treatment on an ectomycorrhizal succession under Scots pine. Mycological Research 107, 317–328.
van der Geer, S., 2003. Asymptotic theory for maximum likelihood in nonparametric mixture models. Computational Statistics and Data Analysis 41, 453–464.
Whittaker, R.H., 1956. Vegetation of the great smoky mountains. Ecological Monographs 26, 1–80.
Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14, 137.
Wu, H.-M., Tzeng, S., Chen, C.-H., 2007. Matrix visualization. In: Handbook of Data Visualization. Springer, Berlin, pp. 681–708.
Zhou, H., Lange, K.L., 2010. On the bumpy road to the dominant mode. Scandinavian Journal of Statistics 37, 612–631.