An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks

An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks

Dalila Failli[a], Maria Francesca Marino[a], and Francesca Martella[b]

[a]Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze; dalila.failli@unifi.it, mariafrancesca.marino@unifi.it

[b]Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Roma; francesca.martella@uniroma1.it

Abstract

Network data analysis has received increasing attention recently. Bipartite networks represent a specific type of network data describing the relationships between disjoint sets of nodes, called sending and receiving nodes. We extend the Mixture of Latent Trait Analyzers (MLTA) specifically tailored for the analysis of bipartite networks to achieve a twofold goal. First, the aim is to perform a joint clustering of sending and receiving nodes, thus partitioning the data matrix into homogeneous blocks, as in the biclustering approach. In addition, a latent trait is used to model the dependence between receiving nodes, as in the latent trait framework. The proposal also admits the inclusion of nodal attributes on the latent layer of the model to understand how they affect cluster formation. An EM algorithm with Gauss Hermite approximation is proposed to estimate the model parameters.

Keywords: Model-based clustering, Network data, Two-mode networks, Nodal attributes, EM algorithm

1. Introduction

Over the years, many social, technological, and biological processes have been represented as networks. These are collections of interconnected units (nodes) that can capture interactions within a system. In this context, bipartite networks are a special type of networks that represent the relationships between two disjoint sets of nodes, formally called sending and receiving nodes. A primary characteristic of this type of network is that connections exist only between nodes belonging to different sets, as illustrated in Figure 1.

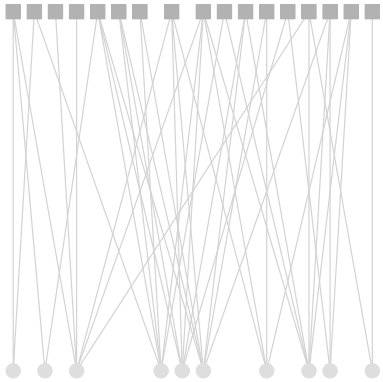

Figure 1: Example of bipartite networks

A relevant aspect of network analysis concerns the simultaneous clustering of sending and receiving nodes aiming at partitioning the data matrix into homogeneous blocks, called biclusters. An example of a block structure is shown in Figure 2, where rows (sending nodes) and columns (receiving nodes) of the data matrix are reordered according to the corresponding class membership, thus returning blocks of sending nodes that connect similarly with subsets of receiving nodes.

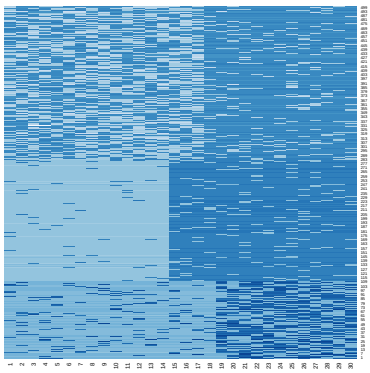

Figure 2: Example of block structure

A common example of application concerns the field of genetics, where the biclustering approach can be used to identify groups of genes which are co-expressed under subsets of experimental conditions.

Different biclustering approaches are available in the literature, such as the model-based ones (13; 18; 3; 15). In this specific context, several methods based on finite mixtures have been proposed (9; 10; 20; 12; 14; 19; 16).

We start from the MLTA model, introduced by (7; 8). Here, the aim is of clustering sending nodes via a finite mixture specification, while accounting for the dependence between receiving nodes via a continuous latent variable, as in the latent trait framework. Our proposal is to modify the MLTA in two ways. First, allowing for a joint clustering of sending and receiv-

ing nodes, where sending nodes are partitioned into clusters called components and, in each of them, receiving nodes are partitioned into clusters called segments. Furthermore, we also allow for the inclusion of nodal attributes on the latent layer in order to understand how they influence component formation.

The paper is organized as follows: in Section 2. we extend the MLTA model, also describing model assumptions, parameter estimation, and model selection. Section 3. shows the results of a simulation study conducted in order to verify the efficacy of the proposed approach. Section 4. contains concluding remarks and details further extensions of the approach.

2. Mixture of latent trait analyzers

Let $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$ denote the set of sending nodes and $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$ the set of receiving nodes. In this framework, bipartite networks can be formally described by a random incidence matrix $\boldsymbol{Y} = \{Y_{ik}\}$, with elements

$$Y_{ik} = \begin{cases} 1 \text{ if sending node } n_i \text{ is connected with receiving node } r_k, \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

To obtain a clearer picture of the data at hand, (8) propose to extend the Mixture of Latent Trait Analyzers (MLTA) (7) in the context of bipartite networks. The model combines latent class and latent trait analysis by assuming that the set of N sending nodes can be divided into G distinct classes (or groups) and that the propensity of each sending node to be connected with the R receiving nodes depends also on a multidimensional continuous latent trait. Our contribution is to further extend the MLTA model by performing a joint clustering of sending and receiving nodes, also taking into account nodal attributes in the latent model structure.

2.1 The MLTA model

The MLTA model assumes that every sending node belongs to one of G unobserved groups identified by the latent random variable $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})' \sim \text{Multinomial}(1, (\eta_1, \ldots, \eta_G))$, whose generic element is

$$z_{ig} = \begin{cases} 1 \text{ if sending node } n_i \text{ belongs to group } g, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

The parameter η_g denotes the probability that a randomly selected sending node belongs to group g, with $g = 1, \ldots, G$, under the constraints that $\sum_{g=1}^{G} \eta_g = 1$ and $\eta_g \geq 0$, $g = 1, \ldots, G$. Furthermore, the model assumes the existence of a D-dimensional continuous latent trait $\boldsymbol{u}_i$ distributed according to a Gaussian density with null mean vector and identity covariance matrix, i.e. $\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, which captures the heterogeneity of the connections between sending and

receiving nodes. Thus, response variables contained in the $\boldsymbol{y}_i$ vector are assumed to be independent Bernoulli random variables with parameters $\pi_{gk}(\boldsymbol{u}_i)$, $k = 1, \ldots, R$, modelled through the following logistic function:

$$\pi_{gk}(\boldsymbol{u}_i) = p(y_{ik} = 1 \mid \boldsymbol{u}_i, z_{ig} = 1) = \frac{1}{1 + \exp[-(b_{gk} + \boldsymbol{w}'_{gk}\boldsymbol{u}_i)]}. \tag{3}$$

Here, the parameter b_{gk} represents the attractiveness of the k-th receiving node for sending nodes belonging to group g, while $\boldsymbol{w}_{gk}$ represents the influence of the latent trait $\boldsymbol{u}_i$ on the probability of a connection between sending nodes belonging to the g-th group and receiving node r_k.

2.2 Extending the MLTA model

To perform a joint clustering of sending and receiving nodes, we follow an approach similar to that proposed by (16) and modify the logistic function in (3) as:

$$\pi_{gk}(\boldsymbol{u}_i) = p(y_{ik} = 1 \mid \boldsymbol{u}_i, z_{ig} = 1) = \frac{1}{1 + \exp[-(b_g + \boldsymbol{a}'_{gk}(\boldsymbol{\mu} + \boldsymbol{u}_i)]}. \tag{4}$$

Here, b_g is a component-specific latent effect, $\boldsymbol{\mu}$ is a D-dimensional vector of fixed effects, and $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$ is a D-dimensional continuous latent trait capturing the residual heterogeneity of connections between sending nodes belonging to the g-th component and receiving nodes belonging to the d-th segment. Moreover, $\boldsymbol{a}_{gk}$ is a D-dimensional row stochastic vector ($D \leq R$) with

$$a_{gkd} = \begin{cases} 1 \text{ if receiving node } r_k \text{ belongs to segment } d, \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

This allows us to select the membership of the k-th receiving node to one of the D segments for those sending nodes belonging to component g.

Following the strategy adopted in (6), we account for the effect that nodal attributes may have on group membership by letting the parameter η_g vary across sending nodes. This is done by considering a latent class regression model based on the vector of nodal attributes $\boldsymbol{x}_i$, as follows:

$$\eta_{ig} = \frac{\exp\{\boldsymbol{x}'_i\boldsymbol{\beta}_g\}}{1 + \sum_{g'=2}^{G} \exp\{\boldsymbol{x}'_i\boldsymbol{\beta}_{g'}\}}, \quad g = 2, \ldots, G, \tag{6}$$

where $\boldsymbol{\beta}_g$ denotes the model coefficient vector for the g-th group.

2.3 Parameter estimation

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_G, b_1, \ldots, b_G, \boldsymbol{a}_{11}, \ldots, \boldsymbol{a}_{GR}, \mu_1, \ldots, \mu_D)$ represent the vector of all free model parameters. Given the assumptions described in the previous section, the log-likelihood

function of the model can be written as:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \Big(\sum_{g=1}^{G} \eta_{ig} \int \prod_{k=1}^{R} p(y_{ik} \mid \boldsymbol{u}_i, z_{ig} = 1) p(\boldsymbol{u}_i) d\boldsymbol{u}_i \Big), \tag{7}$$

where $p(y_{ik} \mid \boldsymbol{u}_i, z_{ig} = 1) = (\pi_{gk}(\boldsymbol{u}_i))^{y_{ik}} (1 - \pi_{gk}(\boldsymbol{u}_i))^{1-y_{ik}}$. The integral to be solved in equation (7) cannot be computed analytically, therefore an EM algorithm with a Gauss-Hermite approximation of the log-likelihood function is proposed.
In the Gauss-Hermite framework, integrals of the form $\int h(\boldsymbol{u}_i) e^{-||\boldsymbol{u}_i||^2} d\boldsymbol{u}_i$ are approximated by

$$\sum_{q_1 \dots q_D} h(\boldsymbol{u}_{q_1 \dots q_D}) \prod_{l=1}^{D} \omega_{ql},$$

where $\boldsymbol{u}_{q_1 \dots q_D}$ are the roots of the Gauss-Hermite polynomial and ω_{ql} are the corresponding weights.
Since $p(\boldsymbol{u}_i) = (2\pi)^{-D/2} \exp\{-\frac{1}{2} \boldsymbol{u}_i \boldsymbol{u}_i'\}$, the integral in (7) should be rewritten as a function of standardized variables $\tilde{\boldsymbol{u}}_i = \frac{\boldsymbol{u}_i}{\sqrt{2}}$. By applying a change of variable, the integral in (7) is modified to

$$\sqrt{2^D} \int p(\boldsymbol{y}_i \mid \sqrt{2} \tilde{\boldsymbol{u}}_i, z_{ig} = 1) e^{-||\tilde{\boldsymbol{u}}_i||^2} d\tilde{\boldsymbol{u}}_i$$

Thus, the log-likelihood function is approximated as:

$$\tilde{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \Big(\sum_{g=1}^{G} \eta_{ig} \sqrt{2^D} \sum_{q_1 \dots q_D} \prod_{k=1}^{R} p(y_{ik} \mid \boldsymbol{u}^*_{q_1 \dots q_D}, z_{ig} = 1) \; f(\boldsymbol{u}^*_{q_1 \dots q_D}) \; e^{||\boldsymbol{u}_{q_1 \dots q_D}||^2} \prod_{l=1}^{D} \omega_{ql} \Big)$$

where $\boldsymbol{u}^*_{q_1 \dots q_D} = \sqrt{2} \boldsymbol{u}_{q_1 \dots q_D}$.
Model parameters can be estimated via an EM algorithm (2), which represents a standard solution when dealing with latent variables. The complete data log-likelihood function is

$$\ell_c(.) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} [\log p(\boldsymbol{y}_i \mid z_{ig} = 1, \boldsymbol{u}_i)] + \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log \eta_{ig} + \sum_{i=1}^{N} \log f(\boldsymbol{u}_i).$$

The E-step of the algorithm consists in computing the expectation of the complete data log-likelihood function conditional on the observed data and the current parameter estimates. This is equivalent to compute the posterior probabilities of z_{ig} and $\boldsymbol{u}_i$ as follows

$$1) \quad p(z_{ig} = 1 \mid \boldsymbol{y}_i) \approx \frac{\eta_{ig} \Big(\sum_{q_1 \dots q_D} p(\boldsymbol{y}_i \mid \boldsymbol{u}^*_{q_1 \dots q_D}, z_{ig} = 1) \phi(\boldsymbol{u}^*_{q_1 \dots q_D}) \sqrt{2^D} \; e^{||\boldsymbol{u}_{q_1 \dots q_D}||^2} \prod_{l=1}^{D} \omega_{ql} \Big)}{f(\boldsymbol{y}_i)} \tag{8}$$

$$2) \quad p(z_{ig} = 1, \boldsymbol{u}_i \mid \boldsymbol{y}_i) \approx \frac{p(\boldsymbol{y}_i \mid \boldsymbol{u}^*_{q_1 \dots q_D}, z_{ig} = 1) \; \eta_{ig} \; \phi(\boldsymbol{u}^*_{q_1 \dots q_D}) \sqrt{2^D} \; e^{||\boldsymbol{u}_{q_1 \dots q_D}||^2} \prod_{l=1}^{D} \omega_{ql}}{f(\boldsymbol{y}_i)} \tag{9}$$

$$3) \qquad p(\boldsymbol{u}_i \mid \boldsymbol{y}_i) \approx \frac{\sum_{g=1}^{G} \eta_{ig}\, p(\boldsymbol{y}_i \mid z_{ig}=1, \boldsymbol{u}^*_{q_1\ldots q_D})\phi(\boldsymbol{u}^*_{q_1\ldots q_D})\sqrt{2^D}\, e^{||\boldsymbol{u}_{q_1\ldots q_D}||^2} \prod_{l=1}^{D} \omega_{ql}}{f(\boldsymbol{y}_i)}, \qquad (10)$$

where $f(\boldsymbol{y}_i) \approx \sum_{g=1}^{G} \eta_{ig} \Big(\sum_{q_1\ldots q_D} p(\boldsymbol{y}_i \mid \boldsymbol{u}^*_{q_1\ldots q_D}, z_{ig}=1)\phi(\boldsymbol{u}^*_{q_1\ldots q_D})\sqrt{2^D}\, e^{||\boldsymbol{u}_{q_1\ldots q_D}||^2} \prod_{l=1}^{D} \omega_{ql} \Big)$.
The M-step of the algorithm consists in updating the model parameters by maximizing the expected complete data log-likelihood function with respect to $\boldsymbol{\theta}$ according to the following steps:

- update b_g and $\boldsymbol{\mu}$ via a standard Newton-Raphson algorithm with augmented data, with weights provided by eq. (9);
- update $\boldsymbol{a}_{gk}$ via a classification step, following a similar strategy to that proposed by (16):
  - compute the log-likelihood values ℓ_{gkd} for each receiving node r_k and the couple (g, d);
  - compute the maximum ℓ_{max} of this set $\{\ell_{gkd}\}$ over $d = 1, \ldots, D$;
  - for each component, allocate the k-th receiving node into the d-th segment iff ℓ_{gkd} = ℓ_{max};
- update $\boldsymbol{\beta}_g$ via a Newton-Raphson step and update η_{ig} accordingly.

The procedure is repeated until convergence, which occurs when

$$|| \ \ell_c^{(t+1)}(.) - \ell_c^{(t)}(.) \ || < \epsilon,$$

where t is the current iteration and $\epsilon > 0$ denotes a given tolerance level. In the following, we set $\epsilon = 10^{-4}$.
At convergence, each sending node can be assigned to the g-th component via a Maximum a Posteriori (MAP) rule and each receiving node can be assigned to the d-th segment according to the vector $\boldsymbol{a}_{gk}$.

2.4 Standard errors and model selection

To evaluate the standard errors of the estimates obtained with the EM algorithm, several methods are available, such as the jackknife method (5; 7). Given an incidence matrix $\boldsymbol{Y}$ with N sending nodes and R receiving nodes, this method consists in extracting N samples of size $(N-1) \times R$, obtained by removing one sending node at a time from the original data matrix. However, we found that a more efficient strategy for deriving standard errors relies on the use of a non-parametric bootstrap (4), which consists in extracting with repetition N rows of the incidence matrix, so that each sending node can appear multiple times.
Since the number of components G and the number of segments D are considered as fixed quantities, it is possible to estimate the model for different values of G and D, then selecting the optimal model as the one corresponding to the smallest value of the chosen information criterion, such as the Bayesian Information Criterion (BIC) (17) or the Akaike's Information Criterion (AIC) (1).

3. Simulation study

The performance of the model in terms of parameters' recovery and clustering is evaluated through a simulation study with a different number of nodes, as described below.

3.1 Simulation setup

We simulated 100 samples in six different scenarios based on a varying number of sending nodes ($N = 50,\ N = 100,\ N = 500$) and receiving nodes ($R = 20,\ R = 30$). Furthermore, we consider a fixed number of segments $D = 2$ and components $G = 3$. As regards the latent class variable, block membership is defined via a single nodal attribute x_i which is drawn from a Gaussian distribution with mean and variance equal to 1. The latent structure is defined by setting $\boldsymbol{\beta} = [1, -0.4, 1.5, -0.9]$, while the block structure is defined by setting $\boldsymbol{b} = [-1.7, 0, 1.7]$ and $\boldsymbol{\mu} = [-2, 0.5]$. In each scenario, a multi-start strategy based on 100 random starts is adopted.

3.2 Simulation study: clustering recovery

The ability of the proposal in correctly classifying sending and receiving nodes is evaluated through the Adjusted Rand Index (ARI) (11), which measures the agreement between the true partition and the estimated partition. Results are shown in Table 1. Looking at this table, we note that, as the number of sending and receiving nodes increases, the classification improves for both rows (sending nodes) and columns (receiving nodes). Moreover, for higher values of N and R, the ARI values tend to 1.

Table 1: Adjusted Rand Index mean (median) across samples for varying N and R

	R=20		*R*=30	
	Row	Col.	Row	Col.
N=50	0.65 (0.66)	0.63 (0.69)	0.76 (0.76)	0.72 (0.75)
N=100	0.71 (0.72)	0.83 (0.93)	0.82 (0.83)	0.91 (0.96)
N=500	0.75 (0.75)	0.96 (1.00)	0.83 (0.83)	1.00 (1.00)

3.3 Simulation study: model parameters

Table 2 shows the Mean Squared Error (MSE) values across samples for b_g, $\boldsymbol{\mu}$ and $\boldsymbol{\beta}_g$ by varying the number of sending and receiving nodes. Looking at this table, it is evident that, as the size of the network increases, we are more and more able to identify the true values of model parameters. Moreover, for $N = 500$ and $R = 30$, MSE values are very low for all model parameters.

Table 2: MSE values across samples for b_g, μ_d and $\boldsymbol{\beta}_g$

	R=20			R=30		
	$\boldsymbol{b}$	$\boldsymbol{\mu}$	$\boldsymbol{\beta}_g$	$\boldsymbol{b}$	$\boldsymbol{\mu}$	$\boldsymbol{\beta}_g$
N=50	$[0.43, 0, 0.34]$	$[0.10, 0]$	$[10.5, 2.04, 4.20, 1.11]$	$[0.29, 0, 0.18]$	$[0.09, 0]$	$[14.6, 8.76, 5.76, 3.26]$
N=100	$[0.25, 0, 0.09]$	$[0.03, 0]$	$[1.85, 2.23, 0.70, 2.13]$	$[0.11, 0, 0.06]$	$[0.03, 0]$	$[0.51, 0.40, 0.14, 0.13]$
N=500	$[0.05, 0, 0.01]$	$[0.02, 0]$	$[0.22, 0.19, 0.03, 0.03]$	$[0.02, 0, 0.01]$	$[0.01, 0]$	$[0.06, 0.06, 0.02, 0.03]$

4. Conclusions

The mixture of latent trait analyzers is modified to achieve a twofold objective for the analysis of bipartite networks: i) performing a joint clustering of sending and receiving nodes; ii) including nodal attributes to study how nodes' characteristics influence the component membership probability. Furthermore, the model allows the dependence between receiving nodes to be modelled via a multi-dimensional continuous latent trait.

The simulation study shows that the model can be effectively employed for biclustering bipartite networks. In detail, when the number of sending and receiving nodes increases, the proposal is able to correctly identify the model parameters and the classification of nodes is good. However, the simulation study needs to be further extended by letting the number of partitions vary. A further development may involve the application of the proposal for the analysis of a large real-world data set, such as a gene-experimental condition network.

References

[1] Akaike, H.: A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. **19**, 716–23 (1974)

[2] Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), **39** (1), 1–38 (1977)

[3] Dhollander, T., Sheng, Q., Lemmens, K., De Moor, B., Marchal, K., Moreau, Y.: Query-driven module discovery in microarray data. Bioinformatics. **23**, 2573–2580 (2007)

[4] Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. **7**, 1–26 (1979)

[5] Efron, B.: Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika. **68** 589–599 (1981)

[6] Failli, D., Marino, M.F., Martella, F.: Extending finite mixtures of latent trait analyzers for bipartite networks. In: Balzanella A., Bini M., Cavicchia C. and Verde R. (eds.) Book of short Paper SIS 2022, pp. 540–550. Pearson (2022)

[7] Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. **24**, 569–588 (2014)

[8] Gollini, I.: A mixture model approach for clustering bipartite networks. In: Ragozini, G.,

Vitale, M.P. (eds.) Challenges in Social Network Research: Methods and Applications, pp. 79–91. Springer International Publishing (2020)

[9] Govaert, G., Nadif, M.: Clustering with block mixture models. Pattern Recognit. **36(2)**, 463–473 (2003)

[10] Govaert, G., Nadif, M.: Block clustering with Bernoulli mixture models: comparison of different approaches. Comput. Statist. Data Anal. **52**, 3233–3245 (2008)

[11] Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

[12] Keribin, C., Brault, V., Celeux, G., et al.: Estimation and selection for the latent block model on categorical data. Stat. Comput. **25**, 1201–1216 (2014)

[13] Lazzeroni, L., Owen, A.B.: Plaid models for gene expression data. Statist. Sinica. **12**, 61–86 (2002)

[14] Martella, F., Alfò M., Vichi, M.: Biclustering of gene expression data by an extension of mixtures of factor analyzers. The Int. J. Biostat. (2008) doi:10.2202/1557-4679.1078

[15] Martella, F., Vichi, M.: Clustering microarray data using model-based double K -means. J. Appl. Stat. **39(9)**, 1853–1869 (2012)

[16] Martella, F., Alfò, M.: A finite mixture approach to joint clustering of individuals and multivariate discrete outcomes. J. Stat. Comput. Simul. 87:11, 2186–2206 (2017)

[17] Schwarz, G.: Estimating the Dimension of a Model. Ann. Stat. **6**, 461–464 (1978)

[18] Sheng, Q., Moreau, Y., De Moor, B.: Biclustering microarray data by Gibbs sampling. Bioinformatics. **19**, 196–205 (2003)

[19] Vicari, D., Alfò, M.: Model based clustering of customer choice data. Comput. Statist. Data Anal. **71**, 3–13 (2014)

[20] Wyse, J., Friel, N.: Block clustering with collapsed latent block models. Stat. Comput. **22**, 415–428 (2012)